

https://dspace.org/

Longitudinal modelling of water levels of the Okavango River

Unandapo, Lazarus Pendapala

2016

http://hdl.handle.net/10539/21273

Unandapo, Lazarus P (2016) Longitudinal modelling of water levels of the Okavango River, University of Witwatersrand, Johannesburg, <http://wiredspace.wits.ac.za/handle/10539/21273>

Downloaded from a DSpace repository. https://dspace.org/

University of the Witwatersrand



A dissertation presented to the faculty of science in fulfilment
of the requirement for the degree of
Master of Science

in the

School of Statistics and Actuarial Science

Longitudinal Modelling of Water Levels of the Okavango River

Author:
Lazarus, P. Unandapo

Supervisor:
Anna Kaduma

Submitted on: May 30, 2016

Declaration

I, Lazarus P. Unandapo, hereby declare that this dissertation is a true reflection of my own work, has never been submitted for a degree at any other institution of higher learning.

This dissertation can be reproduced or stored in retrieval system under the permission of the author or the University of the Witwatersrand.

I, Lazarus P. Unandapo, give the University of the Witwatersrand the right to reproduce this dissertation in any manner and format which they may prefer for any institute or individual requesting it for research purpose given that acknowledgement will be given.



Lazarus, P. UNANDAPO May 30, 2016

Dedication

This dissertation is dedicated to:

My great parents, brothers and sisters who never stop supporting me in so many ways. My friends who supported and always encouraged me, and more importantly my son and girlfriend who have been there for me every single day.

I dedicate this research.

Acknowledgements

I owe especial thanks to organisations and several people who made it possible for me to complete this dissertation. Special thanks to Deutscher Akademischer Austauschdienst, National Council for Higher Education, University of Namibia and the Namibia student financial assistant fund for providing me with necessary financial support over the entire research period. My girlfriend and my son who were understanding for the many nights that I spent working on this project. Special thanks to my colleagues, especially Mr. Hugh Marera, M.Sc (student) who spent sleepless nights helping me with some programming codes. A big thanks to my supervisor who worked hard with me from the beginning until the completion of this dissertation. I appreciate her guidance as well as her generous support and the support I received from everyone involved. Thanks guys.

Abstract

In statistics, a model is as good as the data fed to it. Data about hydrological events continues to grow rapidly over the years, with different variables being recorded on a continuous scale. These variables can be interpreted and used in a different manner among disciplines. Thus, choosing the right variables and interactions among variables is an important statistical step in building a good and accurate model.

This dissertation involved the development of a statistical model which can be used to predict weekly water level within the Okavango river in northern Namibia. The parameters of the statistical mixed model were estimated based on two methods for longitudinal data, the Generalised Estimating Equations (GEE) which is a well known method of parameter estimation in longitudinal data analysis when the observed variables are correlated, and the Restricted Maximum Likelihood Estimation (REML) which is a likelihood based approach method, unlike the GEE. Using cross-validation and a simulation study, the GEE method of estimation was found to be less accurate and inconsistent in terms of prediction of parameter estimation of water level while the well known REML was found to predict the water level with a good degree of accuracy, consistency and with lower variance. Parameters from a simulation study have also shown less bias in REML method and predicted the cross-validation test-set with less bias.

List of Acronyms

AIC	Akaike's Information Criterion
ACF	Autocorrelation function
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
CIC	Correlation Information Criterion
GEE	Generalised Estimating Equation
GLMM	Generalised Linear Mixed Model
IRWLS	Iteratively Re-Weighted Least Squares
MAR	Missing at Random
MCAR	Missing Completely at Random
MLE	Maximum Likelihood Estimation
PDF	Probability Density Function
QIC	Quasi Information Criterion
REML	Restricted Maximum Likelihood Estimation
PRNG	Pseudorandom number generation
SADC	Southern African Development Community
OKACOM	Okavango River Basin Water Commission
ORB	Okavango River Basin

Contents

1 INTRODUCTION **1**

1.1 INTRODUCTION 1

1.2 LONGITUDINAL DATA ANALYSIS 1

1.3 HYDROLOGICAL STUDY AND STATEMENT OF THE PROBLEM 3

1.3.1 Geographical Background Of Study Area 5

1.4 AIM AND STUDY OBJECTIVES 7

1.4.1 Aim 7

1.4.2 Study Objectives 7

1.5 BRIEF SUMMARY AND SOFTWARE 8

1.6 KEY CONTRIBUTIONS 8

1.7 REPORT LAYOUT 9

2 LITERATURE REVIEW **10**

2.1 INTRODUCTION 10

2.2 REVIEW OF LITERATURES CONCERNING HYDROLOGICAL STUDIES . 10

2.3 REVIEW OF LITERATURES CONCERNING DATA IMPUTATION AND SIMULATION STUDIES 12

2.3.1 Data Imputation 12

2.3.2 Simulation Studies 13

2.4 THEORETICAL BACKGROUND OF REML 14

2.4.1 Review of Literature Concerning MLE and REML 16

2.5 THEORETICAL BACKGROUND OF GEE 17

2.5.1 Review of Literature Concerning GEE 20

3 METHODOLOGY **23**

3.1 INTRODUCTION 23
3.2 DATA 23
3.2.1 Data Imputation 25
3.2.2 Cross Validated Sample 26
3.3 THE MODEL 27
3.4 GEE AND REML ESTIMATION 29
3.5 MODEL SELECTION 32
3.6 SIMULATION: BUILDING A MODEL 33
3.7 MODEL DIAGNOSTICS AND GOODNESS OF FIT 36

4 ANALYSIS, RESULTS AND DISCUSSION 38

4.1 INTRODUCTION 38
4.2 DATA IMPUTATION 38
4.2.1 Discussion 38
4.3 DESCRIPTIVE STATISTICS 40
4.3.1 Discussion 47
4.3.2 Residual Analysis 49
4.4 EMPIRICAL (ROBUST) GEE AND GEE MODELS WITH DIFFERENT CORRELATION STRUCTURES 50
4.4.1 Linearity On The Link Scale 52
4.4.2 Assessing for Mean-Variance Relationship 53
4.4.3 Assessing for Non-independence in Model Residuals 54
4.4.4 Discussion 54
4.4.5 GEE Models Plots 58
4.5 REML MODEL 60
4.5.1 Linearity On The Link Scale 62
4.5.2 Assessing for Mean-Variance Relationship 63
4.5.3 Assessing for Non-independence in Model Residuals 64
4.5.4 Discussion 64
4.6 REML MODEL PLOTS 67
4.7 SIMULATION AND MODELLING 69

4.7.1 Pseudo-Monte Carlo Simulation 69
4.7.2 Test Set Estimation 73

5 CONCLUSION AND RECOMMENDATION **77**
5.1 CONCLUSION 77
5.2 RECOMMENDATION 78

References **80**

Appendix A (Table and figures) **85**

Appendix B (R-codes) **95**

List of Figures

Figure 1.1: ACRU model 4
Figure 1.2: Okavango river basin 6

Figure 2.1: Design flood estimation methods 11

Figure 3.1: Flow diagram for creation of id and zTime Variable 25
Figure 3.2: Simple flow chart representation of a GLMM 27

Figure 4.1: Longitudinal plot of weekly water levels at Rundu and Mukwe 41
Figure 4.2: Symmetry plots of weekly water levels at Rundu and Mukwe for dry and wet seasons during (1950-2007) 42
Figure 4.3: Scatter plots of weekly water levels and weekly water flows at Rundu and Mukwe for wet and dry seasons 42
Figure 4.4: Scatter plots of weekly water levels and weekly water flows at Rundu and Mukwe for wet and dry seasons 43
Figure 4.5: Histogram and boxplot of water levels 44
Figure 4.6: Longitudinal weekly profile plot of Water levels 45
Figure 4.7: Lag1 correlation plots of weekly water levels for time period (1950-2007) 45
Figure 4.8: Longitudinal profile (Empirical growth) plots of weekly water levels at Rundu and Mukwe 46
Figure 4.9: GLM correlation of residuals 49
Figure 4.10: Relationship between, QIC, MSE, R^2 and p-values with the sample size for the GEE models 51
Figure 4.11: Observed vs Predicted values for different GEE models 52
Figure 4.12: Mean-variance relationship for different GEE models 53
Figure 4.13: Non independence in model residuals for different GEE models 54
Figure 4.14: Longitudinal plots of predicted and observed values for GEE model . . . 59

Figure 4.15: Residual plots for Robust GEE model 60
Figure 4.16: Relationship between, AIC, MSE, R^2 and p-values with the sample size 61
Figure 4.17: Observed vs Predicted values for REML model 62
Figure 4.18: Mean-variance relationship for REML model 63
Figure 4.19: Non independence in model residuals for REML model 64
Figure 4.20: Longitudinal plots of predicted and observed values for REML model . . 67
Figure 4.21: Partial autocorrelations of REML model residuals 68
Figure 4.22: Relationship between AIC and sample size 71
Figure 4.23: Plot of observed testdata weekly water levels vs predicted weekly water levels from the simulated model 73
Figure 4.24: Histogram plot for observed water levels value vs water levels value predicted using the simulated model parameters 74
Figure 4.25: Plot of observed testdata weekly water levels vs predicted weekly water levels from the simulated model, where simulated weekly water flows was done using location distribution as appose to monthy distributions as given in Table 3.2 76

Figure 5.1: Gamma distribution with different shape and scale parameters 86
Figure 5.2: Correlation plots of water levels at Rundu 87
Figure 5.3: Correlation plots of water levels at Mukwe 88
Figure 5.4: Reciprocal transformation of water levels for wet and dry seasons 89
Figure 5.5: Observed vs (Predicted values+residuals) for different GEE models . . . 90
Figure 5.6: Residuals for Exchangeable and AR1 GEE models 90
Figure 5.7: Residuals vs Predicted values for different GEE models 91
Figure 5.8: Predicted and observed Water values against water flow values for GEE model 92
Figure 5.9: Residual pattern with fitted valued within yearly weeks and differences between yearly weeks for GEE model 93
Figure 5.10: Residual pattern with fitted valued within yearly weeks and differences between yearly weeks for REML model 93

Figure 5.11: Predicted and observed Water values against water flow values for REML model 94

List of Tables

Table 3.1: Bins spline basis matrix for water flows 31
Table 3.2: Shape and scale parameters of the assumed distribution of water flows in different months 34

Table 4.1: K-nearest neighbour data imputation 39
Table 4.2: Descriptive statistics of the distribution of water levels before and after imputation 40
Table 4.3: Water level test statistics 41
Table 4.4: Spearman rank correlation coefficients of water levels and water flows . . 49
Table 4.5: Parameter estimates of a GEE models 51
Table 4.6: Table of estimated weekly water levels by Robust GEE model 58
Table 4.7: REML parameter estimates 62
Table 4.8: Table of estimated water levels by REML model 65
Table 4.9: Significance of model coefficient parameters 69
Table 4.10: REML parameter estimates for simulated model 71
Table 4.11: Table of estimated water levels by REML model 73
Table 4.12: Descriptive statistics of the distribution of testset weekly water levels and predicted weekly water levels 74
Table 4.13: REML parameter estimates for simulated model where water flows was simulated using Locations distributions instead of Months distributions . 75

Table 5.1: Summary statistics of the Okavango river data 86

Chapter 1

INTRODUCTION

1.1 INTRODUCTION

This section presents a brief introduction to longitudinal data analysis, as well as hydrological studies. It further gives a description of the study area, the aim of this study and key contributions to existing knowledge.

1.2 LONGITUDINAL DATA ANALYSIS

Most of the research in epidemiology and medicine are based on longitudinal designs where repeated measurements of any variable of interest for each individual or household are taken (Wassertheil-Smoller, 2004). These types of measurements might or might not take time to obtain, depending on the focus of the study. For example, laboratory based settings that involves simulations might take a relatively short period of time as compared to real life non-laboratory settings. Longitudinal designs can be both statistically as well as scientifically powerful as they can enable one to study changes within individuals or household over time and under diverse conditions. This design methodology has attracted the interest of many researchers since the beginning of the 20^{th} century (Hand and Crowder, 1996; Meyer and Hill, 1997; Twisk, 2003; Wassertheil-Smoller, 2004; Liang and Zeger, 1986; Zeger, Liang, and Albert,

1988; Zorn, 2001).

Researchers like Hand and Crowder (1996), and Meyer and Hill (1997), refer to cases whereby characteristics of interest such as body weight, crime record, and income are continuously collected from the same individual or household at different times points as longitudinal data. Such data records are commonly collected in public health and epidemiology (Wassertheil-Smoller, 2004; Twisk, 2003; Dahmen and Ziegler, 2004), but not necessarily restricted to these fields. (Twisk, 2003, p. 1) defines studies in which an "outcome variable is measured in the same individual or household at several points in time " as longitudinal studies. This type of studies can allow for the differences within and between individuals or households to be followed as function of time.

Statistical methods like those of cross sectional studies which many people are aware of (e.g, census), assume that each primary observation in a study is independent of all the other observations. Thus, a sequence of recurrent cross sectional data can give a wide understanding of the trend for the variable of interest for individuals or households over time but will fail to show variations experienced by each individual or household over a certain time period. When repeated measurements are taken for each individual or household, the above assumption of independence no longer holds as correlation between repeated measurements starts playing a big role. Wassertheil-Smoller (2004) state that examination of longitudinal data leads to different inferences compared to those of cross sectional studies. She motivates this with an example that if we consider cross sectional datasets to be pictures of say an individual income or health, then longitudinal data provides a moving picture of income or health showing the path that each individual went through how and how much their live have changed as time goes, while cross sectional data will just show pictures of individual income or health taken at different time points. Thus, cross sectional data only give general information about the population income or health and fails to trace population income or health as they progress in life.

Hand and Crowder (1996) describe longitudinal records as records that consist of ordered nature of the observations, where measurements which are closer together in time tend to show higher dependency or correlation, than measurements that are further apart in time. This dependency associated with longitudinal data introduces complications in the analysis (Liang and Zeger, 1986). Due to this dependency, it will be wrong to use standard statistical techniques for cross

sectional studies, such as simple linear regression or the chi-square test of association e.t.c, that rely on data that is independent and identically distributed. In order to yield conclusions that are reliable and consistent, dependency of measurements needs to be taken into account. Therefore, it becomes necessary to use appropriate methods for longitudinal data analysis, like the method proposed by (Liang and Zeger, 1986).

1.3 HYDROLOGICAL STUDY AND STATEMENT OF THE PROBLEM

As mentioned earlier, longitudinal data are not only collected in public health and epidemiology. One of the areas in which longitudinal data can be captured is in hydrology, where repeated measurements on river water levels, as well as river water flows and rainfall records are captured hourly or daily, continuously over a long period of time. Atiya, El-Shoura, Shaheen, and El-Sherif (1999) mention that forecasting of river flows is of importance as it significantly helps in predicting water supply for agricultural irrigation projects as well as potential flood damages, among others. Flood estimation is commonly studied in hydrology, as mentioned by Smithers, Chetty, Frezghi, Knoesen, and Tewolde (2013).

Some of the mathematical models used in hydrological research for flood estimations include Continuous Simulation Modelling (CSM), Soil Conservation Service model (SCS), General Circulation Models (GCM), Regional Climate Models (RCM), Probability Distributed Models (e.g. Generalized Extreme Value models) among others. Smithers et al. (2013); Chetty and Smithers (2005); Leith and Chandler (2005) refer to such models as ACRU models, The acronym ACRU was derived from the "Agricultural Catchments Research Unit" within the faculty of engineering at the University of Kwazulu-Natal in South Africa. The ACRU are agro-hydrological models which operate on a daily time step according to Chetty and Smithers (2005). ACRU models make the assumption that the frequency of estimated flood is the same as that of rainfall. Many studies have indicated that in general this was not the case as the antecedent soil moisture conditions prior to rainfall event was a significant factor in determining the run off responses (Smithers et al., 2013). Another major limitation of ACRU models was the inability to account

for precursor soil moisture condition prior to flood events which results in unrealistic estimates of run off. This assumptions was most likely to introduce bias in estimating flood frequency. They also stated that it becomes challenging to estimate floods in an area when there is limited information.

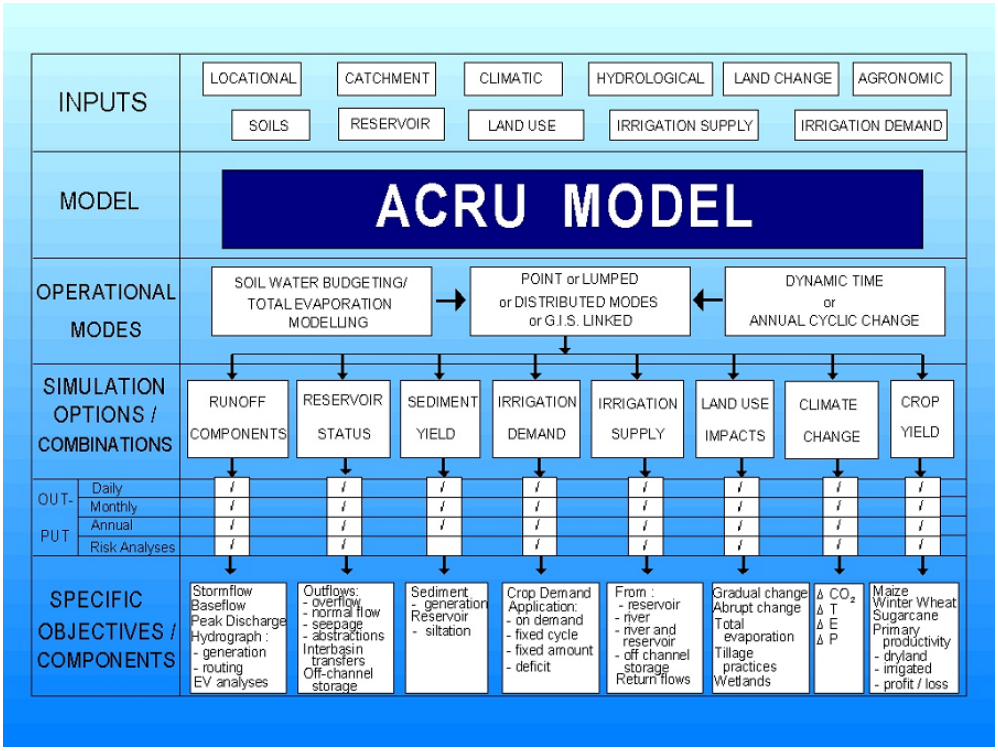


Figure 1.1: ACRU model

Smithers et al. (2013) mention that ACRU models as models shown in Figure 1.1, are not models adjusted to fit parameters of observed data as the physical features of the catchment area determines estimates of the variables. They also stated that ACRU does not account for variability within individual storm events but only for spatial variabilities of rainfall and other variables like land use e.t.c. It was however not clear as to what level these models address correlation over time as we have not so far come across literature that discuss correlation over time when applying these models. Longitudinal models are known for addressing correlation and unlike ACRU models, they are parameter fitting models and can also account for variability within and between individuals (Liang and Zeger, 1986; Hedeker and Gibbons, 2006; Hand and Crowder, 1996; Zorn, 2001; Zeger et al., 1988). Hydrological data are captured on a longitudinal basis and correlation is usually a common factor in such data. Thus, modelling hydrological data using longitudinal models might contribute significantly to the existing knowledge as it

can give a picture of how much hydrological event changed over time.

1.3.1 Geographical Background Of Study Area

Namibia is one of the biggest sub-humid zone countries in the Southern African Development Community (SADC) region, with an approximate population of 2.1 million people according to the Namibia Population and Housing Census report (NPHC) (Namibia, 2011). The report indicates that the country is divided into 13 geographical regions with more than half of the country being covered by desert and rocky mountainous areas. Namibia shares borders with Angola in the north, Zambia in the north east, Botswana in the east, South Africa in the south and its western border is the Atlantic ocean. The report indicates that shortage of fresh water is commonly encountered in most parts of the country. It also states that perennial rivers in Namibia are only found along side the country's northern and southern borders.

The Okavango river is one of the few perennial rivers in the country, and has a catchment area that originates from the rivers of Cubango and Cuito in central Angola, which then merge on the downstream of south eastern Angola to become what is called the Okavango river (Okavango Delta Management Plan (OMD)report, (Botswana, 2008)). The report states that the river water then flows through the Zambezi strip in the north eastern Namibia, and drains in north western Botswana, on what is called the Okavango river delta in the Kalahari desert. Mbaiwa (2004) indicates that this delta covers approximately 3% of the Botswana land, and that 50% of the delta land is flooded on a permanent basis. The Botswana (2008) OMD report mentions that the basin covers a hydrologically active area of about 323 192 km^2 (see Figure 1.2) which is shared by Angola, Namibia and Botswana. It also states that 95% of the basin water is generated by the "headwaters of Cuito and Cubango", which have different topographical response to rainfall; Cuito gives an early peak while Cubango gives a late peak to the Okavango river after rainfall.

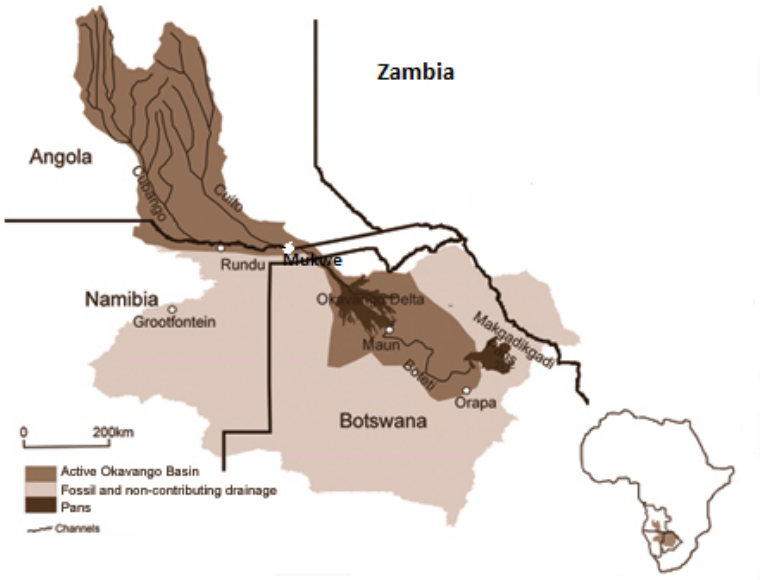


Figure 1.2: Okavango river basin

The Okavango river is one of the longest rivers in SADC, stretching for about 1 100 km, with an estimated basin population of about 122 064 people living in Botswana, 179 000 living in Namibia and a rough estimate of about 204 024 people living in Angola, who depend heavily on the river for survival (Porto and Clover, 2003). The three major settlements in the basin are, Menongue in Angola, Rundu in Namibia and Maun in Botswana. Botswana (2008) OMD report as well as Milzow, Kgotlhang, Bauer-Gottwein, Meier, and Kinzelbach (2009) mentions that the estimated annual stream flow to the river is about 11 billion cubic meters of which about 96% is lost due to evaporation. Namibia has fresh water shortage, but in Namibia, the river is characterized by flat terrain, with features that make it difficult to develop a deep storage dam as mention by Mbaiwa (2004). He further mentions that little is known about water developmental projects in Angola. The basin supports the socio-economic activities of the people living in it, as well as support to biodiversity which is an important aspect for tourism (Botswana, 2008) OMD report.

As a source of water that supports sustainability of the basin region, it becomes of interest to be able to carry out scientific studies on the river, one of the most important being to describe the river water levels given available information such as river water flows and locations among

others. Studying whether there have been changes in river levels over time, and hence model the information using statistical methods so that water levels is monitored as accurately as possible is of importance. Hydrological studies have been done in the past to model floods, extreme events and design flood estimates at non-gauged location, but none of those studies, to the best knowledge of the current author, seem to address the issue of dependency within observed data. Thus, it is expected that conducting a longitudinal study to model water levels of the Okavango may be useful in informing decisions and policy making on water management of this resource.

1.4 AIM AND STUDY OBJECTIVES

Modelling the levels and flows of water in the Okavango river can have a significant impact on Namibia's economy; it can help in agricultural, hydrological and commercial water management, thus, making policies that can protect the country from possible water shortages and flood damage, for example, damage to infrastructure like bridges may impact on transportation in the region, while water shortage if not planned has a direct agricultural and economic implications in the region.

1.4.1 Aim

The main aim of this study was to develop a statistical model, based on longitudinal techniques, that best describes the river water levels in the Okavango river over time.

1.4.2 Study Objectives

The objectives of this study were therefore to:

- investigate water levels of the Okavango river as well as changes in levels over time,
- identify factors that may significantly determine water levels over time,

- establish the variability pattern of the levels over time,
- point out limitations encountered when modelling the river levels in the absence of information from the catchment area,
- develop a model to describe water levels using a simulation study.

1.5 BRIEF SUMMARY AND SOFTWARE

The method of estimation used to fit the model in this research was the Restricted Maximum Likelihood Estimation (REML). The well known Generalized Estimating Equations (GEE) method in longitudinal data was also fitted to the original data with the aim of making comparisons with the results of REML which is commonly used in mixed models and was also the better model for the used data set in this project. The diagnostic procedures used for this methods was the well known AIC and QIC for likelihood and non-likelihood (GEE) method respectively. The coefficients of determination R^2 was used to check for model fit together with bias.

All Statistical Analysis was performed using R programming language Version 3.1.3 (R Core Team, 2015). Distribution fitting and goodness of fit was done using Easy-Fit software which is specifically designed for fitting distributions of data sets. Microsoft Excel was also used for data cleaning. The project report was typed using LaTeX.

1.6 KEY CONTRIBUTIONS

Most past studies done on water levels have used models different from longitudinal models (Smithers et al., 2013; Chetty and Smithers, 2005; Leith and Chandler, 2005). This study however, can give an important step in singling out the problem, but can hardly identify the course of the problem as they are not longitudinal. Longitudinal studies on the other hand goes beyond identifying and describing problems by understanding how and why problems occur and what is most likely to be the best solution. Longitudinal models incorporate correlation

in the data within their model. When water level is measured on a daily basis at a location, correlation will exist, as measurements which are observed closed together in time will most likely be closer in observed values. Thus, it was therefore very important to model correlation when modelling hydrological longitudinal data as ACRU does not account for within individual variations (Smithers et al., 2013). Since correlation within measured data is the primary focus of longitudinal studies, it is hoped that this study has a significant contribution to the way we model longitudinal data in hydrology.

It was mentioned in the Okavango delta management plan (ODMP) report of 2008 that 95% of water in the Okavango river is generated from the river upstream in Cuito and Cubango which are in the higher lands of Angola. Access to recorded hydrological data from Angola was not available to the researcher. The results of this study do not directly link any hydrological events downstream to the catchment of Angola at this current time.

1.7 REPORT LAYOUT

This dissertation consists of 5 chapters namely; Introduction, Literature reviews and theoretical background on methods of estimation, Methodology, Analysis, and finally Conclusions and recommendations. Chapter 1 consists of a general introduction to modelling of longitudinal data and a brief overview of existing models in hydrology. Chapter 2 consists of derivation of theoretical models and their applications found in literature. Chapter 3 gives a detailed overviews of the method and procedure used in this study. Chapter 4 consists of results and discussions, and Chapter 5 summarises the results of the study and gives recommendations for future research.

Chapter 2

LITERATURE REVIEW

2.1 INTRODUCTION

This Chapter presents a brief summary of derivation of Generalised Estimating Equations (GEE) and Restricted Maximum Likelihood Estimation (REML) methods as well as literature reviews on hydrological studies, simulation studies, Maximum Likelihood Estimation (MLE), REML and GEE.

2.2 REVIEW OF LITERATURES CONCERNING HYDROLOGICAL STUDIES

As was mentioned in Chapter 1, Chetty and Smithers (2005) together with Smithers et al. (2013) and Smithers (2012) studied the estimation of floods in the Thukela catchment area in South Africa using the Continuous Simulation Modelling (CSM). They mention that CSM has many advantages when it comes to flood estimation and that it can overcome many limitations as compared to other methods of flood estimations like the one in Figure 2.1 as explained in details by Smithers (2012). Chetty and Smithers (2005); Smithers (2012) also mention that design flood estimation changes according to availability of data, where they state that at areas where there is enough data of stream-flow, the choice between the rainfall based methods and

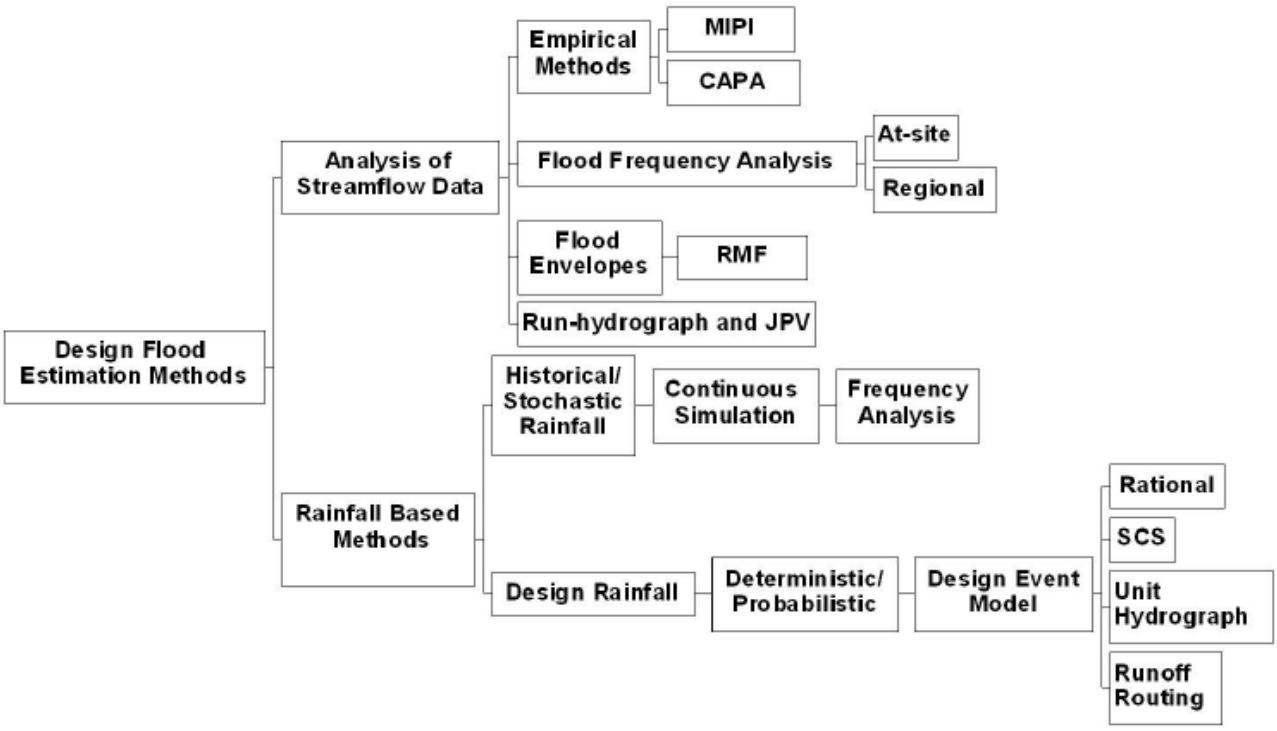


Figure 2.1: Design flood estimation methods

the flood frequency analysis need to be made to estimate floods. They also mention that the rainfall based methods are probabilistic by nature and perform better at estimating floods. The regional statistical analysis are used in estimating flooding in areas with insufficient data as well as un-gauged locations using data from nearby areas. This was however criticized by Boughton and Droop (2003) as they stated that to extend the use of gauged catchment model to non-gauged catchment will require a relationship to exist and the current relationship yield results from which the flood estimation methods can not be applied with confidence. According to Chetty and Smithers (2005), CSM is not a parameter fitting model, but a model which represents major processes that aim at converting rainfall input into run-off.

Chetty and Smithers (2005); Smithers et al. (2013) as well as Atiya et al. (1999) mention that flood estimations are usually a necessity especially for engineers when designing structures such as dams or bridges, and can also help protect community from possible water crises. They mention that designing a very strong structure (e.g. dam or bridge) results in wastage of resources which could have been used for other projects or to help the community in other ways, and designing a weak structure might result in damage to property, infrastructures or

even loss of life when they break down due to heavy floods. They said that it is assumed that variables tend to have the same distribution if they are from the same region (catchment area) and data from the same region can be combined to form one hydro-graph of that region. On the other hand, Atiya et al. (1999) used neural networks to estimate river flow of the Nile river where different neural networks was compared to time series forecasting methods. None of this papers have talked about longitudinal or correlation effects on any of these models.

2.3 REVIEW OF LITERATURES CONCERNING DATA IMPUTATION AND SIMULATION STUDIES

2.3.1 Data Imputation

Jonsson and Wohlin (2004); Mullan, Daraganova, and Baker (2015); Engels and Diehr (2003) mentioned that the exclusion of missing variables or cases from longitudinal dataset reduces the power of statistical analysis, and includes bias in the results. They stated that different methods of data imputation are available but not all of them are suitable for longitudinal data due to correlations in dataset, e.g. Engels and Diehr (2003) mentioned that in longitudinal data missing values of a subject can be assumed to be primarily related to to that subject only and should be imputed based on information from that subject. Some of the methods used for data imputation includes regression imputation, mean imputation, ratio imputation, donor (hot deck) imputation and multivariate imputation (Israëls, Kuyvenhoven, van der Laan, Pannekoek, and Nordholt, 2011). None of these methods are suitable for longitudinal data as longitudinal data have correlation within measurements. For example, the mean imputation replaces the missing values with mean of non missing cases which is not appropriate for longitudinal data. However, the donor method has been extensively used in literature and was proved to be a very good method of longitudinal data, example, Mullan, Daraganova, and Baker (2015) used the k-nearest neighbour (donor method) in a longitudinal Australian study of children to impute income. The k here was the chosen number of nearest neighbours. Jonsson and Wohlin (2004) mentioned that the k must be chosen so that it is equal to the odd of the square root of number of complete cases. However, as k gets bigger, the mean of the distance to donor become wide

which will results in a replacement value being less accurate. This will then produce the same results as the mean imputation method. Thus, k must be as big as possible to give reliable estimates and small as possible to minimise the euclidean distance of donor to the missing case.

2.3.2 Simulation Studies

According to Burton, Altman, Royston, and Holder (2006), simulation studies are severe techniques of computer to evaluate the performance of different statistical methods in relation to a known truth. They stated that such performance can not be accomplished alone by the use of real data. Designing a simulation study that reflect the real situation in practice is a complicated process and there have been little literature on past design simulation study. Burton et al. (2006) stated that some of the issues to look at when designing simulation studies includes;

- defining aims and objectives of the simulation study,
- clearly stating the simulations procedures, method for generating random numbers and number of simulation to be done,
- scenarios to be investigated and methods for evaluations of simulation estimates,
- evaluating performance of statistical methods and stating how simulation results will be presented.

Burton et al. (2006) mentioned that the number of simulations done vary from of 100 to 100 000 with 1000 and 10 000 being the most common choice in most research. Simulations can be done based on existing data set or some other factors. Conducting a large number of simulation has its consequences as it can cause a simulation machine to fail or stop running. Alternately to avoid this, simulation can be done in steps, for example, if a simulation of 100 000 data sets is to be carried out, then one can simulate 10 000 data sets at a time for ten times. This is not the best way to carry out simulation as the results of this simulation might not be reproducible. Ševčíková (2004) proposed a method of statistical simulations on parallel computers where he mentioned that a cluster of personal computers can be put together to work as one which will

then improve the process speed as compared to having only one single central process unit (CPU). This have an advantage as it can allow codes to run in parallel instead of series.

2.4 THEORETICAL BACKGROUND OF REML

MLE method was originally developed by Fisher and states that "the desired probability distribution is the one that makes the observed data most likely" (Myung, 2003). Given that the observed data $\mathbf{y_i}$, is normally distributed with unknown mean, μ_i, and variance, σ_i^2, then let f be a probability density function (PDF) of $\mathbf{y_i}$ which is a vector response variable y_i, given as;

$$f(\mathbf{y_i}; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right] \tag{2.4.1}$$

where $\theta = (\mu, \sigma^2)$. Assuming independence, the function of observed variable $\mathbf{y_i}$, given a set of parameter values is modified for longitudinal data by Oehlert (2014) as;

$$f(\mathbf{y_i}; \theta) = \left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]^m exp\left[-\sum_{i=1}^{m} \frac{(y_i - \mu)^2}{2\sigma^2}\right] \tag{2.4.2}$$

where m is the total number of observed data, and the likelihood of observing the data is given as a function of the parameter, μ and σ^2;

$$L(\theta; \mathbf{y_i}) = \left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]^m exp\left[-\sum_{i=1}^{m} \frac{(y_i - \mu)^2}{2\sigma^2}\right] \tag{2.4.3}$$

Oehlert (2014); Meyer and Hill (1997) and Meyer (1991) derived the log-likelihood from the likelihood function as they notice that it is easier to work with log-likelihood. The log-likelihood can be derived from the likelihood as;

$log(L(\theta; \mathbf{y_i})) = log\left[\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]^m exp\left[-\sum_{i=1}^{m} \frac{(y_i-\mu)^2}{2\sigma^2}\right]\right]$

$= m \times log(\frac{1}{\sqrt{2\pi\sigma^2}}) - \sum_{i=1}^{m} \frac{(y_i-\mu)^2}{2\sigma^2}$

$= m \times \log(2\pi\sigma^2)^{-1/2} - \sum_{i=1}^{m} \frac{(y_i-\mu)^2}{2\sigma^2}$

$= -\frac{m}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{m} \frac{(y_i-\mu)^2}{2\sigma^2}$

Now, let C$=-\frac{m}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{m} \frac{(y_i-\mu)^2}{2\sigma^2}$, then

$$\log(L(\theta; y_i)) = l(\theta; y_i) = C \tag{2.4.4}$$

The unknown parameters μ and σ^2, can be estimated using calculus or optimization algorithms (Oehlert, 2014). In terms of the calculus approach, if the ML estimate exists, then the log-likelihood is differentiable at θ and the 1^{st} and 2^{nd} derivatives are equal to zero and less than zero respectively. The first derivative is a necessary condition for a local maxima or minima to exist when the function is evaluated at θ, while the second derivative is a sufficient condition for the existence of a local maxima. The MLE estimate will then be the value that generates maximum of $l(\theta; y_i)$, as mentioned by Oehlert (2014).

MLE for variances mostly underestimates the variance parameters (Meyer, 1985; Oehlert, 2014). Equation (2.4.4) is build up of a linear combination of unknown parameters (fixed effects) and unknown random variables (random effects). MLE produces biased estimators of variance components in mixed effect models due to ignoring the loss in degree of freedom as a result of fitting the fixed effects. This can however, be solved by maximizing the likelihood independent of fixed effects, this is called restricted maximum likelihood estimation (REML) (Meyer, 1985; Patterson and Thompson, 1971). The REML estimation method works by first obtaining residuals of the observation modelled by the fixed effects, ignoring all variance component (Patterson and Thompson, 1971). Taking all residuals then remove the fixed effect part and only the random effect and error part will remain in the model. They state that when the model for residuals is obtained, MLE can then be done on the residuals to get estimates of variance components.

2.4.1 Review of Literature Concerning MLE and REML

As mentioned in the above section, REML is a modified MLE procedure that accounts for loss in degrees of freedom due to the fixed effect in the mixed effect models. Unlike the MLE, REML does not base its estimates on maximum likelihood fit of all informations available but rather uses the likelihood function so that nuisance parameters have no effects. Myung (2003) states that MLE procedure uses all available data and the method is considered to have less bias for model selection as compared to the least-square estimation (LSE) method and Analysis of Variance (ANOVA), this is due to the fact that MLE requires little or no assumption about distribution of data and has many optimal properties in estimation such as, sufficiency; efficiency; consistency and parameterization invariance which can not be said about the LSE method. He stated that most of the inference methods in Statistics such as, inference for missing data; random effect models; chi-square test and selection criterion like the Akaike information criterion (AIC) and the Bayesian information criteria (BIC) which accounts for model selections was developed based on the MLE method. REML method is more preferred in mixed effect modelling. Meyer (1985); Meyer and Hill (1997); Olori, Hill, McGuirk, and Brotherstone (1999) as well as Johnson and Thompson (1995) used the REML to estimation the variances and covariance components for animal breeding applications.

Meyer (1991) applied the REML method to animal models and used the derivative free approach to the multivariate analysis with some missing records to obtain the MLE of the variance and covariance components. He used the direct maximization of the likelihood by means of derivative free optimization methods. His main interest was to model the mode of inheritance of traits (clusters) in addition to the correlation between traits. He states that "univariate analysis implicitly assumes that all correlations are zero", hence he used the multivariate approach which uses all available information from traits to get estimates of some specific trait, which he then mentioned that its most likely to yield more accurate results. In conclusion he mentioned that this is of importance especially when some of the records are missing due to sampling.

Arango, Cundiff, and Van Vleck (2004) studied the change in cow weight over time in beef cattle by use of the REML method using random effects models and covariance function which is an equivalence of the covariance matrices, when traits have too many records. They noticed

that the estimates of all the variances in weight increase with age, but only up to a certain age, and that there is a fluctuating seasonal pattern. Meyer and Hill (1997) stated that an advantage of random effect model as compared to multiple trait model is that it reduces the number of parameters that needs to be estimated. Arango et al. (2004) as well as Meyer and Hill (1997) have proposed the use of covariance function when dealing with longitudinal data where they used a linear mixed model to describe covariance between any two measurements taken at different age, especially when it comes to growth data as growth reach a peak at maturity. In case of Arango et al. (2004), they assume that environmental effects to cow weight was identically distributed and they analysed it in two ways. Firstly they assume constant variances for all ages and secondly assume heterogeneous variances for each age in the data. Arango et al. (2004) concluded that the heterogeneous error variance was significantly better as compared to the single residual error variance model and that the REML was used to analyse the covariance function using the derivative free algorithm which is an option in D_xM_{RR} software program. This software also has an option of average information REML method in addition to the derivative free REML method.

Olori et al. (1999) applied REML method to the random effect animal model to estimate the variance components of daily records of milk. Their aim was to get coefficient estimates of covariance function and the variance component for average weekly milk consumed. They have also investigated the effect of genetic and environmental covariances to the animal models.

2.5 THEORETICAL BACKGROUND OF GEE

A quasi-likelihood estimator, as defined by Zeger and Liang (1986) is a solution to the score-likelihood equation system given below:

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right]^{T} V(\hat{\alpha})_i^{-1}(y_i - \mu_i(\boldsymbol{\beta})) = 0 \tag{2.5.1}$$

where $\mathbf{y} = (y_1, ..., y_n)$ is a vector of outcomes (y_i) variable decomposed into n strata with μ_i an expected value of y_i given as:

$$\mathsf{E}(\mathbf{y_i}) = \mathsf{h}^{-1}(\mathbf{X_i}\boldsymbol{\beta}) \tag{2.5.2}$$

$\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_\mathsf{n})$ is an $n \times p$ design covariate matrix of predictor variables decomposed into n strata and $\boldsymbol{\beta}$ its a $k \times 1$ vector of regression parameters. Here p is the dimension of each of the strata and k is the dimension of the vector of regression parameters. According to (Zorn, 2001; Zeger and Liang, 1986), h is a link function, which specifies the relationship between $\mathsf{E}(\mathbf{y_i})$ and the $\mathbf{X_i}$. This function transforms the expectation of the response variable μ_i to linear predictors, e.g. $\mathsf{h}(\mu_\mathsf{i}) = \mathbf{X_i}\boldsymbol{\beta}$. $V(\hat{\alpha})_\mathsf{i}$ is the variance of $\mathbf{y_i}$ given as a known function g of $\mathsf{E}(\mathbf{y_i})$, e.g. $V(\hat{\alpha})_\mathsf{i} = \mathsf{g}(\mu_\mathsf{i})\phi$ where ϕ is a scale parameter and $\hat{\alpha}$ is a consistent estimate of α (Zorn, 2001). The solution to Equation (2.5.1) can be obtained by the method of iteratively re-weighted least squares (IRWLS) as stated by (Zorn, 2001; Zeger and Liang, 1986; Millar, 2011). According to (Crowder, 1995) specifications of the correlation between the $\mathbf{y_i}$ can be avoided by assuming a prior working correlation matrix (working correlation structure) $\mathbf{R}(\hat{\boldsymbol{\alpha}})$ when repeated measurements are analysed using GEE models. Here $\mathbf{R}(\hat{\boldsymbol{\alpha}})$ is a fully specified vector of unknown regression parameters (Weiss, 2005). The choices of working correlation matrix include independent working correlation matrix, exchangeable working correlation matrix, first order auto-regressive (AR1) working correlation matrix and unstructured working correlation matrix among others. Each $\mathbf{R}(\hat{\boldsymbol{\alpha}})$ has its own assumptions, for example, the independent $\mathbf{R}(\hat{\boldsymbol{\alpha}})$ assumes zero correlation between the subsequent measurements, exchangeable $\mathbf{R}(\hat{\boldsymbol{\alpha}})$ assumes constant correlation across all observations in a strata (in this case seasons), while AR1 $\mathbf{R}(\hat{\boldsymbol{\alpha}})$ assumes that two measurements taken one time point away within a strata tend to be highly correlated than two observations taken far apart in the same strata. See (Weiss, 2005; Pan and Connett, 2002; Zorn, 2001; Cui and Qian, 2007; Wang and Carey, 2003) for more choices of $\mathbf{R}(\hat{\boldsymbol{\alpha}})$.

Given $\mathbf{R}(\hat{\boldsymbol{\alpha}})$ for response vector $\mathbf{y}$, Pan and Connett (2002); Zeger and Liang (1986); Zorn (2001) expressed the covariance matrix $\boldsymbol{V}(\hat{\boldsymbol{\alpha}})$ in terms of the correlation matrix $\boldsymbol{V}(\hat{\boldsymbol{\alpha}})$ as:

$$\boldsymbol{V} = \mathbf{V}(\hat{\boldsymbol{\alpha}}) = \mathbf{A^{1/2}R}(\hat{\boldsymbol{\alpha}})\mathbf{A^{1/2}}\phi \tag{2.5.3}$$

where $\mathbf{A} = \mathtt{diag}(V(y_1), V(y_2), ..., V(y_p))$ better link A and $V(y_i)$ is a diagonal matrix with $V(y_i) = V(\mu_i)$. The extension of Equation (2.5.1) to longitudinal data is expressed as:

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} D_i^T V(\alpha)_i^{-1}(y_i - \mu) = 0 \tag{2.5.4}$$

with $D_i = D_i(\beta)$ the partial derivative of μ_i with respect to β. When $n = 1$, Zeger and Liang (1986) note that Equation (2.5.4) reduces to the quasi-likelihood estimation. They further state that when the link function h is correctly specified, the GEE (2.5.4) give consistent regression coefficients. Equation (2.5.4) is a score equation for $\boldsymbol{\beta}$, and depends on both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (Zorn, 2001; Zeger and Liang, 1986).

Zeger and Liang (1986) replaced $\boldsymbol{\alpha}$ with some $K^{1/2}$ consistent estimator, $\hat{\boldsymbol{\alpha}}(\mathbf{y}, \boldsymbol{\beta}, \phi)$, in Equation (2.5.3) and (2.5.4) to express the two equations as functions of $\boldsymbol{\beta}$ only. They also replaced the scale parameter ϕ in $\hat{\boldsymbol{\alpha}}$ by $K^{1/2}$ consistent estimator, $\phi(\mathbf{y}, \boldsymbol{\beta})$, so that the estimate $\hat{\beta}$ of $\boldsymbol{\beta}$ is expressed as a solution to:

$$\sum_{i=1}^{n} \mathbf{U}_i \left\{ \boldsymbol{\beta}, \hat{\alpha}[\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})] \right\} = 0. \tag{2.5.5}$$

with $\mathbf{U}_i = \mathbf{D^T}_i \mathbf{V^{-1}}_i \mathbf{S}_i$ as a function of both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. When K increases to infinity, $\hat{\beta}$ becomes a consistent estimator of $\boldsymbol{\beta}$ and $K^{1/2}(\hat{\beta} - \boldsymbol{\beta})$ becomes a multivariate Gaussian with covariate matrix V_β, which consistently estimate the variance (Zeger and Liang, 1986; Oh, Carriere, and Park, 2008):

$$V_\beta = \lim_{K \to \infty} K \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i \right)^{-1} \left[\sum_{i=1}^{n} D_i^T V_i^{-1} \mathtt{cov}(y_i) V_i^{-1} D_i \right] \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i \right)^{-1} \tag{2.5.6}$$

solving the GEE for $\hat{\beta}$, one first has to solve for the regression coefficients, the correlation $\boldsymbol{\alpha}$ and scale parameter ϕ. If we are given an estimate of working correlation matrix $\mathbf{R}(\hat{\boldsymbol{\alpha}})$ and scale parameters ϕ, then $\boldsymbol{\beta}$ can be calculated by IRWLS method. If the V_i is reasonably approximated, then the estimates of $\boldsymbol{\beta}$ is efficiently relative to ML estimates.

2.5.1 Review of Literature Concerning GEE

GEEs are also called quasi-likelihood equations as their parameters are estimated by quasi-likelihood estimator rather than Ordinary least square or Maximum likelihood estimation (Zorn, 2001; Zeger and Liang, 1986; Zeger et al., 1988; Millar, 2011), are a form of Generalized Linear Models (GLM) that can assess for correlation within observed data $\mathbf{y_i}$, which is a vector of responses y_i in this case (water levels). GEE models are used to estimate parameters of a GLM with a possible unknown correlation within observed y_i. Twisk (2003) states that; with GEE, "the relationships between the variables of the model at different time-points are analysed simultaneously". GEE is different from the standard maximum likelihood analyses that require specification of the full conditional distribution of the dependent variable. Zorn (2001) mentioned that: "quasi-likelihood requires only that we postulate the relationship between the expected value of the outcome variable and the covariates, and between the conditional mean and variance of the response variable".

GEEs assume that, the observed variable $\mathbf{y_i}$ are correlated within strata but not necessary between strata. Its covariates are not restricted to a linear form, variances do not need to be homogeneous and that the errors terms are correlated. It also assumes a priori working correlation matrix. In addition to this, GEE works at its best when the observation within a strata are small and the number of strata are large.

Some of past work done on GEE includes Liang and Zeger (1986) where they proposed a GEE methodology as an extension to the GLM for the analysis of longitudinal data. The extension was a class of estimating equations which gives consistent estimates of the variance and regression parameters on assumption of time dependency or joint distribution. The main interest with GEE is to model the pattern of change over time, which can be investigated by solving the first two moment, $\mathtt{E}(y_i)$ and $\mathtt{Var}(y_i)$. The approach used by Liang and Zeger (1986) was related to the quasi likelihood methodology. When there was only one observation to analyse per subject, GEE was used to obtain a description of the outcome variables, however, when repeated measurements are taken per subject, the correlation within subject values are taken into consideration. The GEE model of Liang and Zeger (1986) models the marginal distribution instead of the conditional distribution given past observations. They further state

that the GEE method reduces to MLE when the observed values y_i are multivariate Gaussian. Zeger et al. (1988); Zorn (2001) used the GEE to model longitudinal data where, they applied two methods; the subject specific models that take heterogeneity into account, and on the population average models which focus on aggregate response of the population. They also mentioned that for GEE to give consistent estimates, the working correlation matrix needs to be correctly specified.

Hanley, Negassa, Edwardes, and Forrester (2003) studied the analysis of correlated data using the GEE. They used small hand calculated worked examples to compare results of GEE method, in addition to one real data set, of which both binary and quantitative response data where used to help end user appreciate and understand the method. Their approach used a weighted combinations of observation to obtain the appropriate amount of information from the correlated data. The variability of statistics derived from both correlated and uncorrelated observations was also discussed, for example they mentioned that if observations are uncorrelated then its only the diagonal element in the correlation matrix that will be non-zero, so that the variance of y_i is $Var(\hat{y_i}) = (\frac{1}{m})^2\sigma^2 + (\frac{1}{m})^2\sigma^2 + ... + (\frac{1}{m})^2\sigma^2 = \frac{\sigma^2}{m}$ and standard deviation, $SD(\hat{y_i}) = \frac{\sigma^2}{\sqrt{m}}$ given that each of the observation has a weight of $\frac{1}{m}$, with m being the number of measurements.

The choice of working correlation matrix for the GEE was studied in detail by Ziegler and Vens (2010), where they mention that one of the strengths of the GEE is that it does not "require the correct specification of the multivariate distribution but only the mean structure". Their main objective was to apply the GEE to dichotomous dependent variables, more importantly, the choice of working correlation matrix in the case of dichotomous dependent variables that will lead to a minimal loss in efficiency. Ziegler and Vens (2010) mentioned that the working correlation should be chosen such that it is closest to the true correlation, however, the true correlation is not known but can only be assumed based on some theoretical background or estimated based on available data. Some statistical criteria used to select the working correlation matrix are discussed in Pan and Connett (2002) where they suggest a method of bootstrapping to select the best correlation structure. The Bootstrap method make inference about population data from sample data by assigning a measure of accuracy to sample estimates at every re-sampled data point. Cui and Qian (2007) proposed the Quasi-likelihood Information

Criterion (QIC), as a method to select the best working correlation structure,

$$\widehat{QIC}(R) = -2\hat{\varphi}Q(\hat{\mu}) + 2tr(\hat{Q})$$

with $\hat{Q} = \hat{A}_p^{-1}C_R$ and $\hat{\mu} = g^{-1}(X\hat{\beta})$ with g^{-1} being the inverse link function. Here $\hat{\beta}$ and $\hat{C}_R = \frac{1}{n}\sum_{i=1}^{n}(\hat{D}_i\hat{\sum}_i^{-1}\hat{D}_i)$ are coefficients estimates and robust variance estimator respectively obtained from a model with general working covariance matrix. A_p is another covariance estimator obtained from a model with an assumption of independent working covariance matrix, $\hat{\sum}_i$ is the estimated diagonal covariance matrix, $\hat{\varphi}Q(\hat{\mu})$ being the quasi-likelihood function and $\hat{\varphi}$ a dispersion parameter. The QIC function assumes that all observations are independent and the best working correlation structure is the one with the smallest QIC. Another alternative was also mentioned by Ziegler and Vens (2010) and also by Shults, Sun, Tu, Kim, Amsterdam, Hilbe, and Ten-Have (2009), where the selection criterion such as Correlation Information Criterion (CIC) and the Rotnitzky Jewell Criterion (RJC) was used to select the best working correlation structure. The RJC is defined as;

$$\widehat{RJC}(R) = \sqrt{\left(1 - \frac{C_1}{p}\right)^2 + \left(1 - \frac{C_2}{p}\right)^2}$$

with $C_2 = tr(\hat{Q}^2)$, $C_1 = tr(\hat{Q})$ and $CIC = 2tr(\hat{Q})$ where Q is defined as before. The working correlation matrix with small CIC and $\widehat{RJC}(R)$ will be the best structure. Wang and Carey (2003) suggested that a working correlation structure should be chosen based on either biological or statistical reasons.

Kenward, Lesaffre, and Molenberghs (1994) have applied the GEE together with Maximum likelihood estimation (MLE) to ordinary longitudinal data where some of the missing data do not occur completely at random. They concluded that for subjects with no missing data, the GEE and MLE provide similar results, while for whole data set of subjects (with missing data), the results of the two estimation method differ.

Chapter 3

METHODOLOGY

3.1 INTRODUCTION

This section presents the discussions on the development of the models, methods of estimation, model diagnostics as well as data used for analysis. The developed model used in this research was a generalised linear mixed model which can be used to estimate water levels of the Okavango river over a period of time.

3.2 DATA

The main variable of interest used in this study was daily observations of water levels and water flows of the Okavango river collected from two distinct locations (Rundu and Mukwe) from October 1943 to December 2013. Due to the dataset having a lot of missing values at certain time points, the observations were aggregated to weekly averages which resulted in the reduction of the data. This was also done as the recorded number of records of water levels and water flows was not consistent from day to day. The part of the dataset where aggregation of data could be done as information was records available for each week between 1950 and 2007 inclusive. This was done due to the fact that at Mukwe location there was no recorded information available before 1950 and there were no further records from 2007 on ward. Thus,

the total number of observations in the dataset used for analysis consisted of 2 784 repeated weekly average observations of water levels and water flows at each of the two locations. Thus, the sample sizes were each 2 784. Water level was measured in meters while water flows was measured in cubic meters per second $\mathfrak{m}^3\mathfrak{s}^{-1}$. The numbers of missing cases in final dataset were 145 for water levels of which 72 and 73 was observed at Rundu and Mukwe locations respectively. For water flows the missing records were 29 of which 14 and 15 were observed at Rundu and Mukwe locations respectively. This missing values were assumed to be missing at random (MAR) and they were due to reasons like (but not restricted to);

- recording instrument at river site was not working properly,
- the recording instrument was affected by mud and was unable record changing water levels,
- the battery in the recording instrument die as it was not replaced on time
- technical errors or difficulties when entering the data into the data base.

Thus, to avoid further loss of information, the remaining missing values were imputed using the method of k-nearest neighbour (k-NN) for longitudinal data (Jonsson and Wohlin, 2004; Mullan, Daraganova, and Baker, 2015; Engels and Diehr, 2003). Other variables of interest in the dataset were Season and time measured in weeks for a total of 58 years. The Season variable was a binary coded variable based where month of June to November was considered dry season (code=0) and the remaining was considered wet season (code=1). The total number of years in the study was 58 years, starting from 1950 to 2007, each year having a total of 48 weekly records. Other variables used which were added to the dataset were the quota variable, id variable and the Time variable. The quota variable was derived from the season variable to help with the creation of clusters and was coded as; month of December, January and February = first quota; March, April and May = second quota; June, July and August = third quota; while September, October and November = fourth quota. The id variable was a variable which was used as an identification of clusters given the year, location and quota. It was coded as a discrete combination of Year/Location/Quota, e.g. first quota at Rundu during 1950 and 2005 were coded as **195011** and **200511** respectively. Thus, total number of clusters in the dataset

was $58(\texttt{years}) \times 4(\texttt{quota}) \times 2(\texttt{locations}) = 464$ with each cluster having a maximum cluster size of 12 observations.

The Time variable was another coded time variable that defines the longitudinal time of the whole study. The Time variable were coded to represent time as follow: Time=**19501** gives a row whose records was taken on the first week of 1950 while **20052** gives a row whose record was taken on the second week of 2005. For model convergence and better results however, it is better to be working with small Time representations in numbers. Thus, instead of having 1950 to 2007 as year codes , another code variable (coded 1=1950 to 58=2007) was created which represent years, and hence Time code was then adjusted accordingly, e.g. **19501** becomes **11** while **20052** becomes **562**.

Figure 3.1 gives graphical representation of how the id and Time (zTime) variable was created.

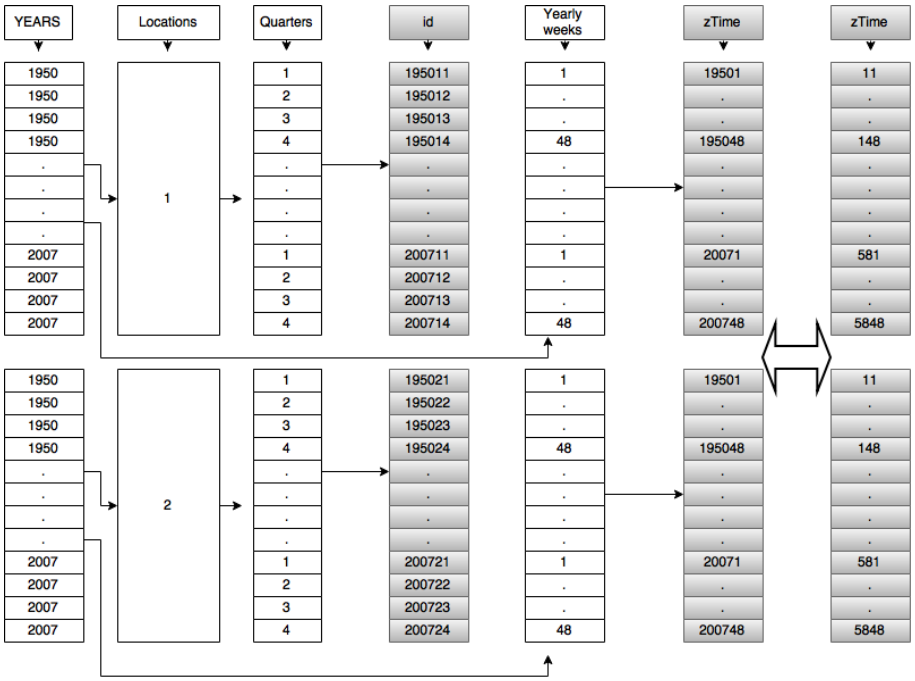


Figure 3.1: Flow diagram for creation of id and zTime Variable

3.2.1 Data Imputation

The hot deck (k-NN) method was used to impute data in this research. The k-NN used here was a weighted average method where the weight was calculated as the exponent of negative

of the euclidean distance (dist(k,x)) between the case with missing value (x) and its neighbour (k), weight=$e^{-\mathtt{dist(k,x)}}$. k=5 was chosen for this project, for which the weighted mean of values for the 5 nearest neighbour was used to replace the missing x value. k=5 was appropriate for this longitudinal study due to the fact that values which are closer together in time are more related than values that are further apart in time. Some other options of commonly used k are k=3 and k=10. A very small value of k like k=3 is good but could be heavily affected by an outlier while a very big value of k like k=10 could give an estimated value far from the closest neighbours. Thus k=5 was chosen as it minimises the trade off between being affected by outliers and producing an estimate further away from the nearest neighbours. There is no standard way to follow when choosing the number of neighbours to use in the imputation. In this study, a very big k will not be ideal as water levels and flows changes based on time and a bigger k might result in the imputed valued bigger compared to its nearest neighbours. The k=3 was also avoided so that if technical errors exist in the dataset, imputed values will not be significantly affected.

3.2.2 Cross Validated Sample

The imputed dataset was then partitioned into two parts (training set and the test set) using a method of cross validation. Cross validation is a technique for model validation in statistics that help asses the results of a statistical model by testing it on an unseen data and therefore generalising the obtained results. The training set consisted of a random 80% of the observed dataset which was then used in building of the generalised linear mixed model (GLMM) using GEE and REML method of estimation. These are models from which the parameters used for simulation were obtained based on criteria mentioned in Section 3.6. The test set data (remaining 20% of the observed dataset) was used to validate the model built from a simulation study. A simulation model was built in order to assess its parsimony and ability to describe the data compared to the GLMM model. Simulation study was conducted as described in Section 3.6.

3.3 THE MODEL

The GLMM model used to obtain the parameter estimates used in simulation in this research was a GLMM. The GLMM was fitted to the training data set using both GEE and REML methods of estimation in order to asses which of the two methods will be the most appropriate for modelling these data. The appropriate model is the model with the smallest AIC/QIC, largest coefficient of determination and small bias. The parameters of the best model were then used as a basis for the simulation. The GLMM used in this study is similar to the models proposed by (Tiwari and Shukla, 2011; Verbeke and Molenberghs, 2009), where they stated that given a dataset with observed values from different clusters to be used in this longitudinal study where $c = 464$ to be the total number of clusters each with a maximum cluster size of $w = 12$ observations, where total number of weekly water level observations for all clusters $Y=\sum_{i=1}^{c}\sum_{k=1}^{w} y_{aw}$ =5568. Our GLMM is given in Equation (3.3.1) with its symbolic flow chart representation in Figure 3.2.

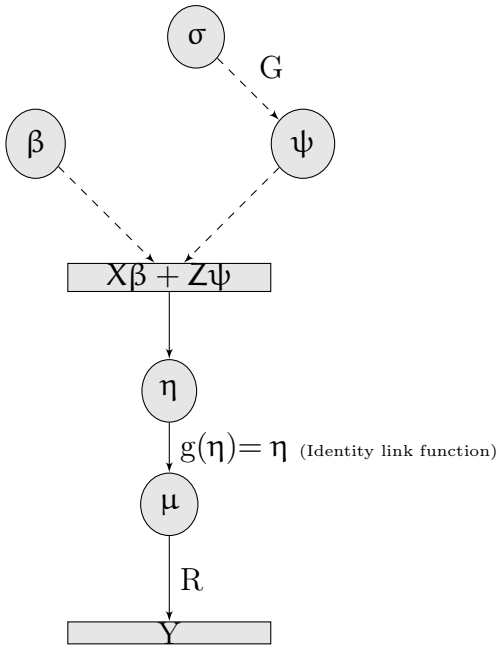


Figure 3.2: Simple flow chart representation of a GLMM

$$\mathbf{Y} = g(E(Y)) + \boldsymbol{\epsilon} = \eta + \boldsymbol{\epsilon} = \mathbf{X\beta} + \mathbf{Z\psi} + \boldsymbol{\epsilon} \tag{3.3.1}$$

where G and R in Figure 3.2 are the covariance matrices for the random and fixed effects respectively, which may depend on a set of some unknown variance component. $\eta = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\psi}$ is a vector of linear predictors. The GLMM link function was an Identity link function which map the η_i values to the conditional mean of μ_i. The mean and variance was assumed to be $g(E(Y)) = E(Y) = \eta = \mu$ and $g(Var(Y)) = Var(Y) = \sigma^2$ respectively. The mixed effect model (3.3.1) consists of three parts. Firstly, the linear predictor part η which can be sub-divided into fixed effect part $\mathbf{X}\boldsymbol{\beta}$ and the random effect part $\mathbf{Z}\boldsymbol{\psi}$ which are used to obtain the mean model, secondly the link function, which model relationship between η and the conditional mean, and finally the variance function which can model residuals variabilities within clusters (Kachman, 2000). $\mathbf{Z}\boldsymbol{\psi} + \boldsymbol{\epsilon}$ is called the covariance structure or the covariance model for the mixed effect model (3.3.1). In the mixed-effect model (3.3.1), η is a $(c_i w_k \times 1)$ vector of linear predictor variables $y_{j(c_i)}$ during the j^{th} time in the c_i^{th} cluster. $\mathbf{X}$ is a $(c_i w_k \times (r+1))$ known constant design matrix that depends only on the observed variables (e.g, Time, Location and Quota). $\boldsymbol{\beta}$ is a $((r+1) \times 1)$ fixed vector of regression coefficients. $\mathbf{Z}$ is a $(c_i \times s)$ design matrix of random effects for $\boldsymbol{\psi}$ which is a vector of dimension $(s \times 1)$ of unknown random effects. $\boldsymbol{\epsilon}$ is a $(c_i w_k \times 1)$ vector of errors.

The models (3.3.1) is a generalised linear mixed model. The model parameters were estimated using the GEE and REML methods which are used to fit GLM objects. It is a mixed effect model since the clusters at different locations were treated as random. For the GEE method; GEE need a pre-specification of the working correlation matrix (Zorn, 2001; Zeger et al., 1988). From the literature, it is recommended to start with a more complicated covariance structure for example; the unstructured model, and then move to a less complicated one like the independence covariance and the empirical (robust) covariance structures. The covariance structures for GEE used in this study were restricted to Unstructured, Exchangeable, First-order autoregressive, Independence and the Empirical (robust) covariance structures. Thus, based on mixed effect model (3.3.1), the actual model for Rundu location during dry season can be expressed mathematically by;

$$\mathbf{Y}_{L_1 S_k} = \mathbf{X}_{L_1 S_k}\boldsymbol{\beta}_{L_1 S_k} + \mathbf{Z}_{L_1 S_k}\boldsymbol{\psi}_{L_1 S} + \boldsymbol{\epsilon}_{L_1 S_k} \tag{3.3.2}$$

Where L_1 and S_k means that information fitted to the model came from location 1 during season k, were in this research report, location 1 was defined to be Rundu location, location 2 was defined to be Mukwe location, season 0 was defined as dry season and season 1 was defined as wet season.

3.4 GEE AND REML ESTIMATION

GEE: The model fitted (estimated) using this method is a GLMM as given in Equation (3.3.1). GLMM model estimates under GEE was fitted using the "geepack" package (Højsgaard, Halekoh, and Yan, 2014) in R statistical software. The uncertainty of model parameters was estimated in two different ways;

1. by using the variance and correlation matrix assumed under the model and assume that they are realistic using the naive estimator,
2. by using the robust or empirical method which give reliable standard errors (sandwich estimator).

The robust method works well when it is used for balanced longitudinal data where every cluster has the same number of measurements, where the number of clusters needs to be much higher compared to the number of repeated measurements in each cluster (Mackenzie and Scott-Hayward, 2015). This is because the cluster size determines the covariance matrix to be estimated. The "geepack" has four options of working correlation structures (matrices) when using the naive estimator; the Exchangeable, Independence, Unstructured, and the First Order Auto-Regressive (AR1) correlation structures. When a correlation structure is not specified, the error term is modelled by robust method which uses the independence working correlation structure. The unstructured covariance matrix becomes difficult to model when the repeated number of observations per cluster is very large as it assumes different correlations between measurements, for example, if repeated measurements within a cluster is of size 1000, then unstructured GEE model will produce a covariance matrix of 1000×1000 of which the covariances between measurements are in no way related. GEE are GLMM objects, unlike linear models

which assume a linear relationship between errors and predicted values, constant variance, independence and normality of the error term, the GLMM make only the following assumptions about the dependent variable;

1. linearity on the link scale (in this case , the "Identity link"), linear relationship between observed values and their linear predictor (η)
2. mean-variance relationship (constant variance),
3. independence of error.

Thus, GLMM does not make an assumption of normality and it only assume that the error terms are correlated only within a cluster but independent between clusters (Kachman, 2000; Mackenzie and Scott-Hayward, 2015). The GEE model was fitted in the following way: firstly, an id variable which groups dataset into clusters of similar observations was created as described in Section (3.2). Section (3.2) identifies another time variable zTime, which uniquely identify time was also added to the dataset. GEE does not require a time component to be present in the model, but only requires the dataset to be ordered according to time; if a dataset is not ordered, GEE gives different results every time the order of the data set changes. A piecewise polynomial functions called bin-splines (B-splines) for water flow was calculated. This was done since transforming the water values could not yield a better model outputs. This was probably due to having very large numbers in the basis Water flow column. The B-spline can help avoid this problem as B-spline columns only contain values in the interval (0;1), where the number of columns depends on the degree ($\mathtt{d}$) of the basis function and the number of knots where the knots are values that divide the fitted values in some portions of polynomials, for example, a quadratic basis with 2 knots has ((3-1)+2=4) basis matrix columns. Commonly used $\mathtt{d}$ is $\mathtt{d}$=2 and $\mathtt{d}$=3 (Mackenzie and Scott-Hayward, 2015) while knots could be mean, quantiles or percentiles of the smoothed vector. In the GEE model for this study, $\mathtt{d}$=3 was used at knots equal to mean of the smoothed vector (Water flow), this was done at this intervals to avoid possibilities of overfitting or under-fitting the model. B-spline basis for water flow with $\mathtt{d} = 3 - 1 = 2$ and knots=2=mean(Water.flows) has the matrix as given in Table 3.1 below;

Table 3.1: Bins spline basis matrix for water flows

	BS.1	BS.2	BS.3	BS.4
1	0.2481015	0.004162394	2.267332e−05	0
2	0.4644504	0.018058774	2.208793e−04	0
3	0.6928853	0.058974128	1.491657e−03	0
4	0.7533144	0.085053652	2.771828e−03	0
5	0.7957303	0.134466105	6.243907e−03	0
6	0.7307253	0.247226875	2.109961e−02	0
7	0.6185271	0.335120498	4.630744e−02	4.499357e−05
8	0.4551249	0.430154698	1.125581e−01	2.162299e−03
9	0.4631758	0.426563254	1.083447e−01	1.916233e−03
10	0.4444386	0.434718454	1.183175e−01	2.525489e−03
11	0.2951559	0.469156540	2.207369e−01	1.495062e−02
12	0.2589093	0.466951047	2.525252e−01	2.161453e−02
13	0.2457517	0.464803789	2.648044e−01	2.464011e−02
14	0.1751187	0.437916841	3.376471e−01	4.931734e−02
15	0.3107989	0.468593420	2.079152e−01	1.269247e−02
16	0.5052978	0.405746143	8.800251e−02	9.535101e−04
17	0.6451928	0.315801194	3.899555e−02	1.050686e−05
18	0.7476992	0.231261663	1.819656e−02	0
19	0.7920681	0.170485097	9.763938e−03	0
20	0.7814209	0.107555196	4.175137e−03	0

Note: BS.1, BS.2, BS.3 and BS.4 are the calculated smoothing B-splines basis of water flows for degree equal to 3-1=2 and knots equal to 2.

A lot of knots could results in model over-fit as the fitted line will be too smooth while few number of knots could sometime results in under-fit which usually produces a very small coefficient of determination (R^2). Another aspect that can capture overfitting and under fitting is Mean Squared error $\mathsf{MSE} = \mathsf{Var}(\mathsf{x}) - \mathsf{bias}^2(\mathsf{x})$, where under-fit can results in higher variance and low bias while over-fit can results in low variance and high bias. The number of d used was decided based on the MSE and QIC of the model; the $\mathsf{d} = 3$ value that yielded the smallest QIC was chosen (see results in Table 4.5). Thus, the formulae used for fitted GEE mixed model was then;

$$y_{j(c_i w_k)} = model + \epsilon_{j(c_i w_k)} \tag{3.4.1}$$

where in Equation (3.4.1),

$model = (\hat{\beta}_0 + \hat{\beta}_1 * BS.1_{j(c_i w_k)} + \hat{\beta}_2 * BS.2_{j(c_i w_k)} + \hat{\beta}_3 * BS.3_{j(c_i w_k)} + \hat{\beta}_4 * BS.4_{j(c_i w_k)} + \hat{\beta}_5 * S_{j(c_i w_k)} + \hat{\beta}_6 * T_{j(c_i w_k)} + \hat{\beta}_7 * (S:T)_{j(c_i w_k)})$.

Here $BS.1(X_{j(c_i w_k)}), ..., BS.4(X_{j(c_i w_k)})$ represents matrices of B-spline basis function terms for water flows at time j within cluster c_i during month w_k. The $y_{j(c_i w_k)}$ is a vector of linear predictors while $S_{j(c_i w_k)}$, $zTime_{j(c_i w_k)}$, $(S : zTime)_{j(c_i w_k)})$ are vectors of predictor variables, and $\epsilon_{j(c_i w_k)}$ a vector of residuals as defined in Section 3.3. The Robust and GEE with Unstructured, Exchangeable, AR1 and Independence structures was fitted and the one with the smallest QIC was selected as the best model among the 5 models fitted using GEE method.

REML: The GLMM model was also fitted in R using the "nlme" package for linear and nonlinear mixed effect models (Pinheiro, Bates, DebRoy, and Sarkar, 2014), where the model structure was the same as the one fitted using the GEE method as given in Equation (3.4.1). The parameter estimation was done using the REML method. The assumptions of the REML method are similar to that of the GEE method since they are both methods used to estimate parameters of GLMM objects. Hence, AIC was used as the information criterion in selecting the best REML model.

The two models from GEE and REML methods of parameter estimations were then compared. The best model that had given parameter estimates with lowest MSE and higher coefficient of determination was then selected as the one whose parameter were then used in the simulation of the model described in section 3.3

3.5 MODEL SELECTION

The Akaike's Information Criterion (AIC) and Quasi Information Criterion (QIC), which is the GEE equivalent of the AIC, were used as model selection criteria. The AIC is a well established

goodness of fit statistic for selecting models that are likelihood-based (Pan, 2001; Hardin and Hilbe, 2003). AIC is expressed mathematically by:

$$AIC = -2L(\hat{\beta}; \mathbf{y}) + 2K \tag{3.5.1}$$

where $L(\hat{\beta}; \mathbf{y})$ is the likelihood function, $2K$ is a penalty term where K is number of parameters in model. Since GEE is a non-likelihood, under the independence $\mathbf{R}_i(\boldsymbol{\alpha})$ model, Pan (2001) modified (3.5.1) to:

$$QIC = -2Q(h^{-1}(X\hat{\beta})) + 2\mathtt{trace}(\Omega V_r) \tag{3.5.2}$$

and generalised the penalty term by re-calculating $2K$ in (3.5.1) to become $2\mathtt{trace}(\Omega V_r)$, where $\Omega = \sum_{i=1}^{n} D_i^T V_i^{-1} D_i$ is the variance matrix of the independent model and $V_r = \mathtt{cov}(\hat{\beta})$ the covariance estimator of $\hat{\beta}$. $Q(h^{-1}(X\beta))$ is the quasi-likelihood value and trace refers to the summation of the diagonal element of the matrix. Hardin and Hilbe (2003) state that when evaluating quasi-likelihood, $\hat{\mu} = h^{-1}(X\hat{\beta})$ is used in place of μ where $\mu = h^{-1}(X\beta)$, with h being an inverse link function for the model. The $2\mathtt{trace}(\Omega V_r)$ is also referred to as the CIC, which can also be used in many cases to select the best working correlation matrix. Millar (2011) mention that the rule of thumb for the AIC is that if two models have an AIC with a difference of 2 or more between, then the model with small value of AIC should be given preference but if is difference is less than 2 then it can be argued that both model are worth of consideration.

3.6 SIMULATION: BUILDING A MODEL

Water flow of the imputed dataset was left-skewed. Many of the distributions for non-negative values that are also left-skewed include the Gamma, Frechet, Burr, Dagum, Weibull, and Generalized Extreme Value (GEV) among many others. However, based on the goodness of fit of the Kolmogorov-Smirnov (K-S) and the Anderson-Darling (A-D) tests, none of the known

distributions were found to be fitting the weekly water flow data. K-S statistic works by quantifying the distance between the Empirical Distribution Function (EDF) of the tested sample and the Cumulative Distribution Function (CDF) of the reference distribution, while the D-S statistics assumes absence of parameters to be estimated in the tested distribution, where the test and critical values are said to be distribution free (Schittkowski, 2002). In the Easy-Fit documentation, it was mentioned that D-S gives more weight to the tail of the distribution. When the distribution of water flow in different months was investigated for the two location. These where found to have known distributions. Their shape and scale parameters, together with their probability density functions (PDF ($\mathsf{f(x)}$) are summarised in Table 3.2:

Table 3.2: Shape and scale parameters of the assumed distribution of water flows in different months

Month	Mukwe Parameters	Distribution	Rundu Parameters	Distribution
Dec	$\alpha 2$=0.73655, α=6.8639 , β=167.19	Burr	$\alpha 2$=0.8493, α=3.02 , β=62.63	Burr
Jan	$\alpha 2$=0.66227, α=5.7573 , β=219.52	Burr	$\alpha 2$=0.94054 , α=2.3768 , β=97.671	Burr
Feb	$\alpha 2$=0.77458, α=3.972 , β=243.61	Burr	α=1.4768 , β=133.36	Gamma
Mar	α=7.4244 , β=66.504	Gamma	α=4.2004 , β=84.34	Gamma
Apr	α=8.7731 , β=66.008	Gamma	α=5.4313 , β=76.036	Gamma
May	$\alpha 2$=1.8336, α=2.4782 , β=457.39	Burr	$\alpha 2$=1.0641 , α=3.5507 , β=249.43	Burr
Jun	α=7.0593 , β=38.997	Gamma	α=4.3846 , β=27.582	Gamma
Jul	α=10.887 , β=20.982	Gamma	$\alpha 2$=1.4013 , α=2.979 , β=100.72	Burr
Aug	$\alpha 2$=0.72656, α=8.0721 , β=181.23	Burr	$\alpha 2$=1.6748 , α=4.2757 , β=87.648	Burr
Sep	$\alpha 2$=0.59466, α=9.219 , β=157.48	Burr	$\alpha 2$=4.2679 , α=3.5446 , β=102.72	Burr
Oct	$\alpha 2$=0.61422, α=8.0028 , β=139.67	Burr	α=8.444 , β= 9.2459	Gamma
Nov	$\alpha 2$=0.87037, α=6.0401 , β=147.47	Burr	$\alpha 2$=0.98118 , α=4.2158 , β=37.566	Burr

$\text{Gamma } f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x.\beta^{-1}}$ and $\text{Burr } f(x) = \frac{\alpha.\alpha 2\,(x.\beta^{-1})^{\alpha-1}}{\beta\,(1+(x.\beta^{-1})^{\alpha})^{\alpha 2+1}}$

Key: shape parameter=α, scale parameter=β, shape parameters no.2=$\alpha 2$, Location parameter=$\gamma \equiv 0$(default) where $\alpha,\ \alpha 2,\ \beta \in (0,\infty),\ \gamma \in (-\infty,\infty)$

The simulation of water flow data was therefore done from the non-negative distributions, mainly the Gamma and the Burr distributions whose parameters and PDF are given in Table 3.2. The Gamma and Burr distributions are the ones that were found to have lower rank of goodness of fit of the K-S and A-D test among many tested distributions. The simulations were done separately for each month. Simulations were done in the following steps:

Step 1

1. the imputed dataset were grouped by locations and then by month. Descriptive statistics were then obtained for each month;
2. monthly field data for each location were then imported into the Easy-Fit statistical software (Schittkowski, 2002) where the goodness of tests fit using the K-S and A-D statistics were done for multiple distributions;
3. a distribution with lowest rank on K-S and A-D test was adopted to be the distribution which a particular monthly water flow data follows (or at-least close to the correct distribution), and parameter estimates were then obtained;
4. a simulation of water flow data for each month at each location was then done using R statistical software (R Core Team, 2015);
5. each monthly simulated water flow was column bound to the data set from which its parameter was obtained in order to keep good track of id variable and;
6. monthly data sets were then row bound together to obtain matrix with simulated data;

Step 2

7. the $(c_i w_k \times 4)$ matrix of smoothing B-spline for water flows was calculated;
8. the regression β coefficients from the selected model were used to generate simulated water levels values using simulated water flow data as summarised in model below;

$$simulated\ Water.level = \hat{\beta}_0 + \hat{\beta}_1 * BS.1 + \hat{\beta}_2 * BS.2 + \hat{\beta}_3 * BS.3 + \hat{\beta}_4 * BS.4 + \hat{\beta}_5 * zTime + \hat{\beta}_6 * Season + \hat{\beta}_7 * (zTime : Season)$$

9. simulated data parameter estimates were obtained using REML and tested using MSE 10 000 times in an R loop for accuracy, efficiency and consistency (and the generated dataset was saved as a list in a hard drive).

Crowther and Cox (2005); Khu and Werner (2003) mentioned that in order to check how well a simulated model performs, it is necessary to test it on an unknown data sample, in this case it was tested on the 20% of the randomly imputed field test data set and the MSE was also calculated for simulation assessment.

3.7 MODEL DIAGNOSTICS AND GOODNESS OF FIT

According to Park and Lee (2004), regression models used to analyse longitudinal data also need model diagnostics for detecting outliers as well as checking for model fit and influential observations.

Diagnostics: Model diagnostics were analysed using the coefficients of determination (R^2) as well as the AIC (QIC). R^2 was calculated by dividing the explained sum of square (mss) of model by total sum of square (TSS) where TSS=mss+residual sum of square (rss). Thus, $R^2 = \frac{mss}{TSS}$. A model with good prediction power usually have bigger R^2 value where $0 \leqslant R^2 \leqslant 1$. Models with $R^2 \in [0, 0.4)$ are considered weak and thus give unreliable estimates. $R^2 \in [0.4, 0.7)$ can be considered moderate but still not so good, while $R^2 \in [0.7, 1)$ is considered strong and thus gives reliable estimates. $R^2 = 1$ defines a perfect model.

Goodness of fit: The goodness of the model fitted was assessed using the bias (or percentage of bias) and the MSE, where $bias = \bar{\hat{\varphi}} - \bar{\varphi}$ (or percentage bias is $(\frac{bias}{\bar{\varphi}}) \times 100$) and MSE=$\overline{residuals^2}$ where φ is the parameter of interest. The bias was used to check how far the predicted values are from the observed values and also to asses the model parameters estimated by the simulated model as compared to those estimated by the non simulated model. The MSE was used to assess the accuracy, efficiency and consistency of the model especially during simulation or when sample size changes. Furthermore, the Kolmogorov-Smirnov (K-S) test and Anderson-Darling (A-D) test were also used to analyse the goodness of fit on model residuals, to test if the residuals came from the hypothesised continuous distribution. The results were based on the empirical cumulative distribution function (CDF) given as $F_n(\epsilon) = \frac{1}{n}[numbers\ of\ observations \leqslant \epsilon]$ where ϵ was the vector of residuals of n dimension from a distribution with CDF $F(\epsilon)$. The K-S and A-D statistics are given as;

K-S statistics$=MAX_{1\leqslant i\leqslant n}\left[F(\epsilon_i)-(i-1)n^{-1}, n^{-1}-F(\epsilon_i)\right]$

A-D statistics$=-n-n^{-1}\sum_{i=1}^{n}(2i-1)\left[InF(\epsilon_i)+In(1-F(\epsilon_{n-i+1}))\right]$

these tests based their calculations on maximum difference between theoretical and empirical CDF where A-D gives more weight to the tail as compared to K-S (see section 3.6). Distributions with lowest rank on K-S and A-D statistics give best fit of the residuals.

Chapter 4

ANALYSIS, RESULTS AND DISCUSSION

4.1 INTRODUCTION

This chapter consists of four main sections. These include data imputation of missing cases in the data set, descriptive data analyses which explored the existing data set in more detail, multivariate data analysis which dealt with the development of statistical models and the simulation and modelling part which consisted of 10 000 Pseudo-Monte Carlo simulations of random data to help develop a statistical model that is not based on observed field data.

4.2 DATA IMPUTATION

4.2.1 Discussion

This dataset used in this study had 174 missing cases. Doing analysis on such data set would have meant the missing cases are deleted, this could have resulted in loss of information (when missing cases are deleted, an entire row, containing a missing value will be deleted). Restricting our regression analysis to complete cases only may have resulted in bias in our regression

estimates and the power of a statistical test could also be reduced. In this dataset, deleting rows with missing values will be worse as it could result in the deletion of other useful informations whose some of the cells in a row might be empty.

Different k values for the k-nearest neighbour (k-NN) imputation method can be used on this dataset, where k is the number of neighbours used to impute a missing value. In this dataset, k=5 was used. There is always a trade off between the number of k used. Jonsson and Wohlin (2004) mention that the estimation of missing cases gets worse as the distance between donor (complete case) and recipient (missing value) gets large. In this research k=5 was deemed as appropriate, as it was necessary to get the weighted mean of as many points as possible without increasing the distance between donor and recipient significantly. If a smaller k, say k=3, were to be considered appropriate in this research, then there would have been a problem in finding neighbours for some missing cases since some of the cells had two or three consecutive missing values. The advantage of the k-NN method is that the imputed point is only influenced by the nearest points than by all points. This is very important for longitudinal data as correlation between cases will not be jeopardised. The variance between measurements did not significantly change after imputation as the difference in relative variance of observed and imputed values 0.2%, 0.04%, 0% and 0.02% for Rundu dry, Rundu wet, Mukwe dry and Mukwe wet respectively (see Table 4.2).

Table 4.1: K-nearest neighbour data imputation

	k=3 nearest neighbours				k=5 nearest neighbours				k=10 nearest neighbours			
YEAR	RL	RF	ML	MF	RL	RF	ML	MF	RL	RF	ML	MF
1992	3.87	54.25	3.45	143.74	3.87	54.25	3.45	143.74	3.87	54.25	3.45	143.74
1992	3.91	57.84	3.34	143.58	3.91	57.84	3.34	143.58	3.91	57.84	3.34	143.58
1992	3.94	62	3.23	146.69	3.94	62	3.23	146.69	3.94	62	3.23	146.69
1993	**4.26094**	66.8	3.11	152.75	**4.21522**	66.8	3.11	152.75	**4.18055**	66.8	3.11	152.75
1993	**4.05941**	74.04	**2.85909**	154.23	**4.02829**	74.04	**2.81246**	154.23	**4.11553**	74.04	**2.82084**	154.23
1993	4.22	81.52	**2.85177**	155.95	4.22	81.52	**2.79791**	155.95	4.22	81.52	**2.86037**	155.95
1993	4.25	90.34	3.09	167.45	4.25	90.34	3.09	167.45	4.25	90.34	3.09	167.45
1993	4.43	97.34	3.07	195.66	4.43	97.34	3.07	195.66	4.43	97.34	3.07	195.66
1993	4.35	119.33	3.03	220.24	4.35	119.33	3.03	220.24	4.35	119.33	3.03	220.24
1993	4.38	111.05	2.98	239.82	4.38	111.05	2.98	239.82	4.38	111.05	2.98	239.82
1993	4.39	114.09	2.94	235.23	4.39	114.09	2.94	235.23	4.39	114.09	2.94	235.23
1993	4.37	113.73	2.89	239.31	4.37	113.73	2.89	239.31	4.37	113.73	2.89	239.31
1993	4.68	115.57	2.81	230.21	4.68	115.57	2.81	230.21	4.68	115.57	2.81	230.21
1993	5.35	142.18	2.75	236.7	5.35	142.18	2.75	236.7	5.35	142.18	2.75	236.7
1993	**5.3763**	243.86	2.7	237.14	**5.30119**	243.86	2.7	237.14	**5.29209**	243.86	2.7	237.14
1993	**5.33229**	240.23	**2.82743**	273.34	**5.45653**	240.23	**2.78072**	273.34	**5.36677**	240.23	**2.80923**	273.34
1993	4.75	219.09	**2.89883**	330.21	4.75	219.09	**2.80931**	330.21	4.75	219.09	**2.76699**	330.21
1993	4.34	199.22	2.67	367.79	4.34	199.22	2.67	367.79	4.34	199.22	2.67	367.79
1993	**4.56118**	185.74	2.66	375.96	**4.6253**	185.74	2.66	375.96	**4.6639**	185.74	2.66	375.96
1993	**4.55149**	174.55	**2.86799**	349.14	**4.45931**	174.55	**2.94258**	349.14	**4.51912**	174.55	**2.85389**	349.14
1993	**4.38707**	164.28	**2.91523**	326.19	**4.46678**	164.28	**2.89712**	326.19	**4.52201**	164.28	**2.80908**	326.19
1993	**4.30331**	153.55	**2.70825**	307.62	**4.32751**	153.55	**3.01446**	307.62	**4.37026**	153.55	**3.02215**	307.62
1993	4.61	143.04	**2.98704**	291.77	4.61	143.04	**3.13744**	291.77	4.61	143.04	**3.12124**	291.77

Note: RL, RF, ML and MF stands for Rundu water levels, Rundu water flows, Mukwe water levels and Mukwe water flows respectively. The column (YEAR), give the year at which the given records was obtained

Table 4.2: Descriptive statistics of the distribution of water levels before and after imputation

	Before imputation				After imputation			
statistics	Rundu		Mukwe		Rundu		Mukwe	
	Dry season	Wet season	Dry season	Wet season	Dry season	Wet season	Dry season	Wet season
Mean (in meters)	3.98	5.34	3.01	3.04	3.98	5.33	3.01	3.05
Median (in meters)	3.9	5.21	2.93	2.95	3.9	5.19	2.94	2.95
Std. dev	0.41	1.1	0.44	0.44	0.4	1.1	0.44	0.44
% of relative variance	4.22	22.66	6.43	6.37	4.02	22.7	6.43	6.35
observed values	1344	1368	1341	1370	1392	1392	1392	1392
% of missing values	3.45	1.72	3.66	1.58	0	0	0	0

Table 4.1 shows a snapshot of dataset of water levels and water flows from two locations along the Okavango river in Namibia where missing values were imputed using the hot deck k-NN method (Jonsson and Wohlin, 2004; Engels and Diehr, 2003). The bold values are imputed water levels and water flows values. RL, RF, ML and MF in Table 4.1 corresponds to Rundu water levels, Rundu water flows, Mukwe water levels and Mukwe water flows respectively. Table 4.2 gives summary descriptive statistics of water levels before and after imputations was done.

4.3 DESCRIPTIVE STATISTICS

Table 4.3 summarises the tests done on the dataset; Shapiro-Wilk normality and Anderson-Darling normality tests to check for normality in the distribution of the water level data. The results of these tests were used to help choose the appropriate summary statistics to use in describing these data. The results of the Shapiro-Wilk and Anderson-Darling normality tests are further supported by the symmetry plots in Figure 4.2 as well as the histograms with kernel densities and box plot in Figure 4.5. A symmetry plot is used to determine whether sample data come from a symmetric distribution. A distribution is symmetric if data on both sides of the median are distributed the same way, that is, the tails of the distribution are mirror images of each other. A symmetry plot graphs the upper distance to the median on the x-axis vs. the lower distance to the median on the y-axis, for each data point. A reference line on

Table 4.3: Water level test statistics

Test Conducted	Season	Okavango River estimates	Rundu estimates	Mukwe estimates
Shapiro-Wilk normality test	Dry season	$p-value < 2.2e^{-16}$	$p-value < 2.2e^{-16}$	$p-value < 2.2e^{-16}$
	Wet season	$p-value < 2.2e^{-16}$	$p-value < 2.2e^{-16}$	$p-value < 2.2e^{-16}$
Anderson-Darling normality test	Dry season	$p-value < 2.2e^{-16}$	$p-value < 2.2e^{-16}$	$p-value < 2.2e^{-16}$
	Wet season	$p-value < 2.2e^{-16}$	$p-value < 2.2e^{-16}$	$p-value < 2.2e^{-16}$
Note:	Alternative hypothesis (H_1): the water levels sample did not come from a normally distributed population			
Wilcoxon Rank-Sum test	Dry season median	3.496508	3.979712	3.013303
	Wet season median	4.187847	5.329189	3.046506
		$p-value < 2.2e^{-16}$	$p-value < 2.2e^{-16}$	$p-value = 0.03902$
Note:	Alternative hypothesis (H_1): true difference in median water levels is not equal to 0 at 5 % level of significance			

the plot represents a perfectly symmetric sample. The more symmetric the data, the closer the sample data points will be to the line. The histograms in Figure 4.5 display the shape of the distributions. It should be noted that many statistical procedures assume that data come from a normal distribution. However, many statistical procedures are robust to violations of normality, so having data from a symmetric distribution is often adequate. In this case the median was used to test for difference in water levels for the two location for each season. The Wilcoxon Rank-Sum test indicated that difference in median water levels were highly significant in the two seasons. This further validated season as an important variable to be included in the model. The longitudinal plots of the water levels in (Figure 4.1) shows the weekly seasonal water levels, where differences for the two locations are clearly visible.

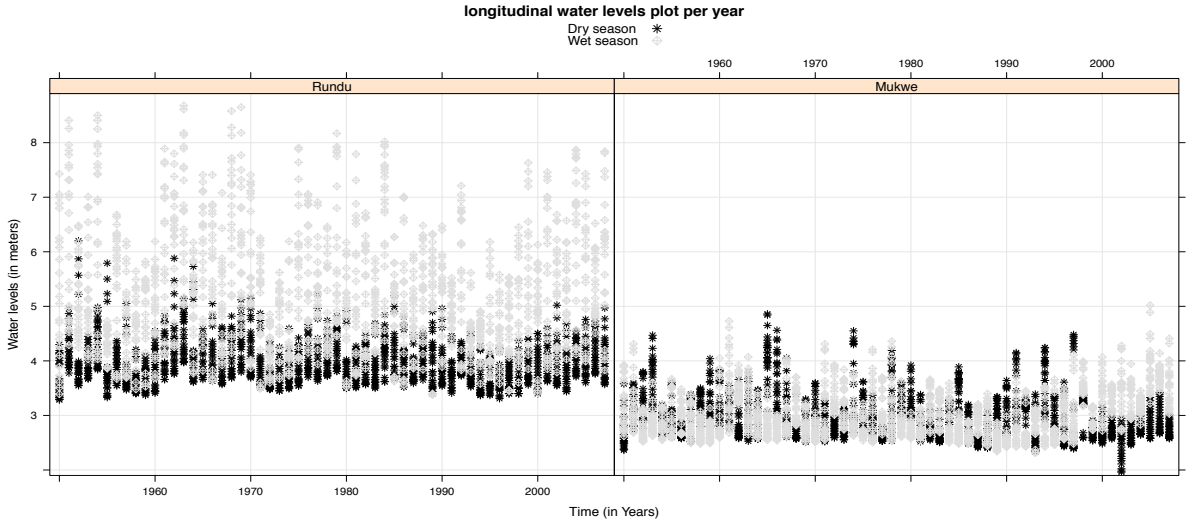


Figure 4.1: Longitudinal plot of weekly water levels at Rundu and Mukwe

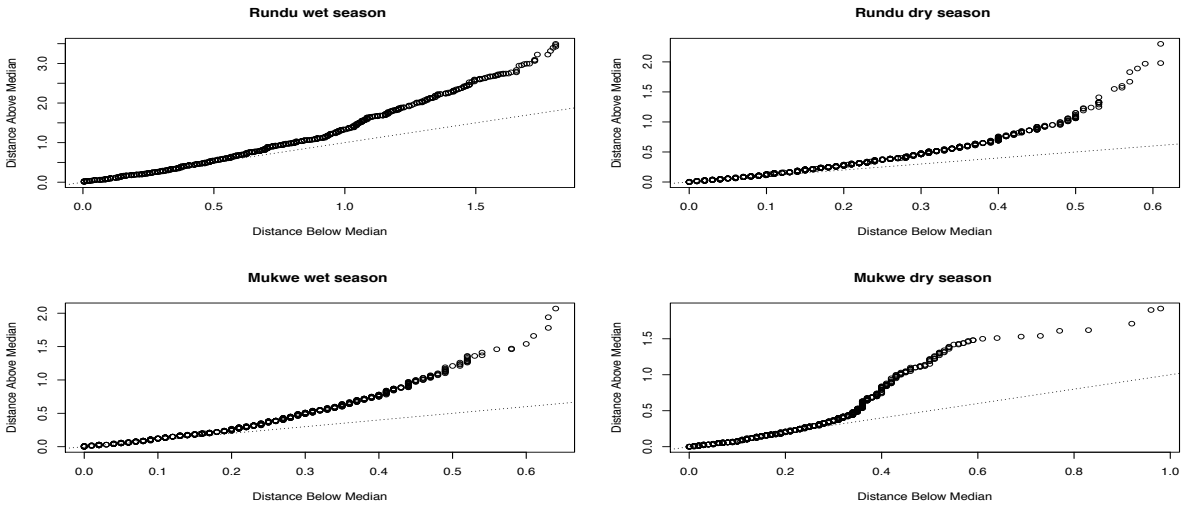


Figure 4.2: Symmetry plots of weekly water levels at Rundu and Mukwe for dry and wet seasons during (1950-2007)

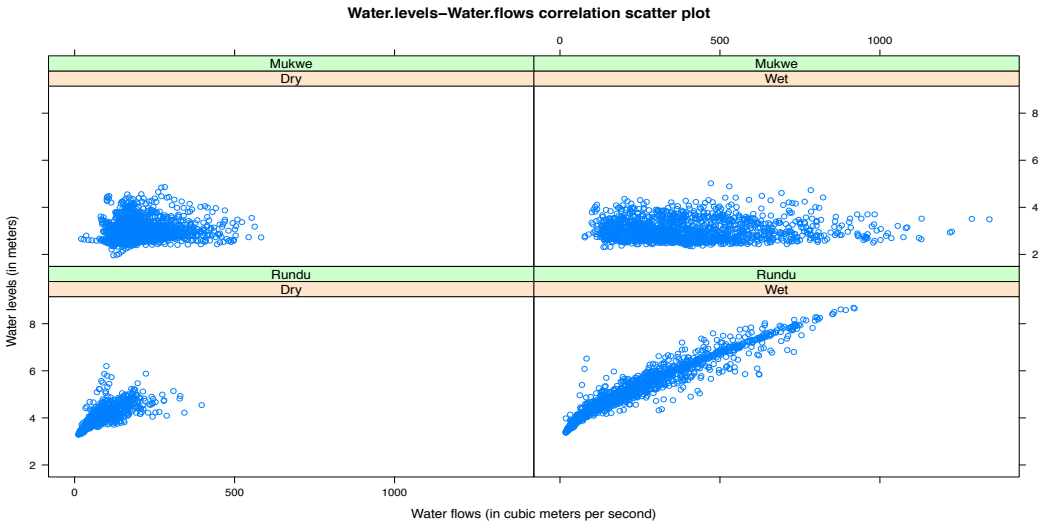


Figure 4.3: Scatter plots of weekly water levels and weekly water flows at Rundu and Mukwe for wet and dry seasons

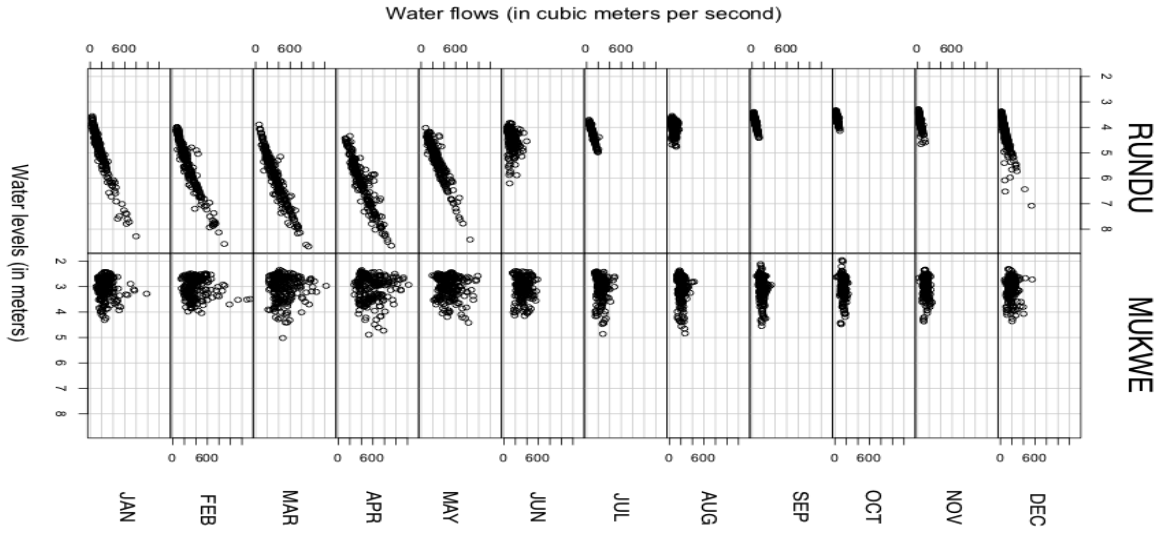


Figure 4.4: Scatter plots of weekly water levels and weekly water flows at Rundu and Mukwe for wet and dry seasons

Figure 4.3 and Figure 4.4 show the relationships between the weekly water levels and the predictor weekly water flow in different months regardless of years (see Figure 4.4) as well as in the two locations for the two seasons (plot of Figure 4.3). Note that these plots show relationships in a cross-sectional manner. Spearman correlation coefficients were calculated for the relationship in Figure 4.3 and Figure 4.4, and presented in Table 4.4.

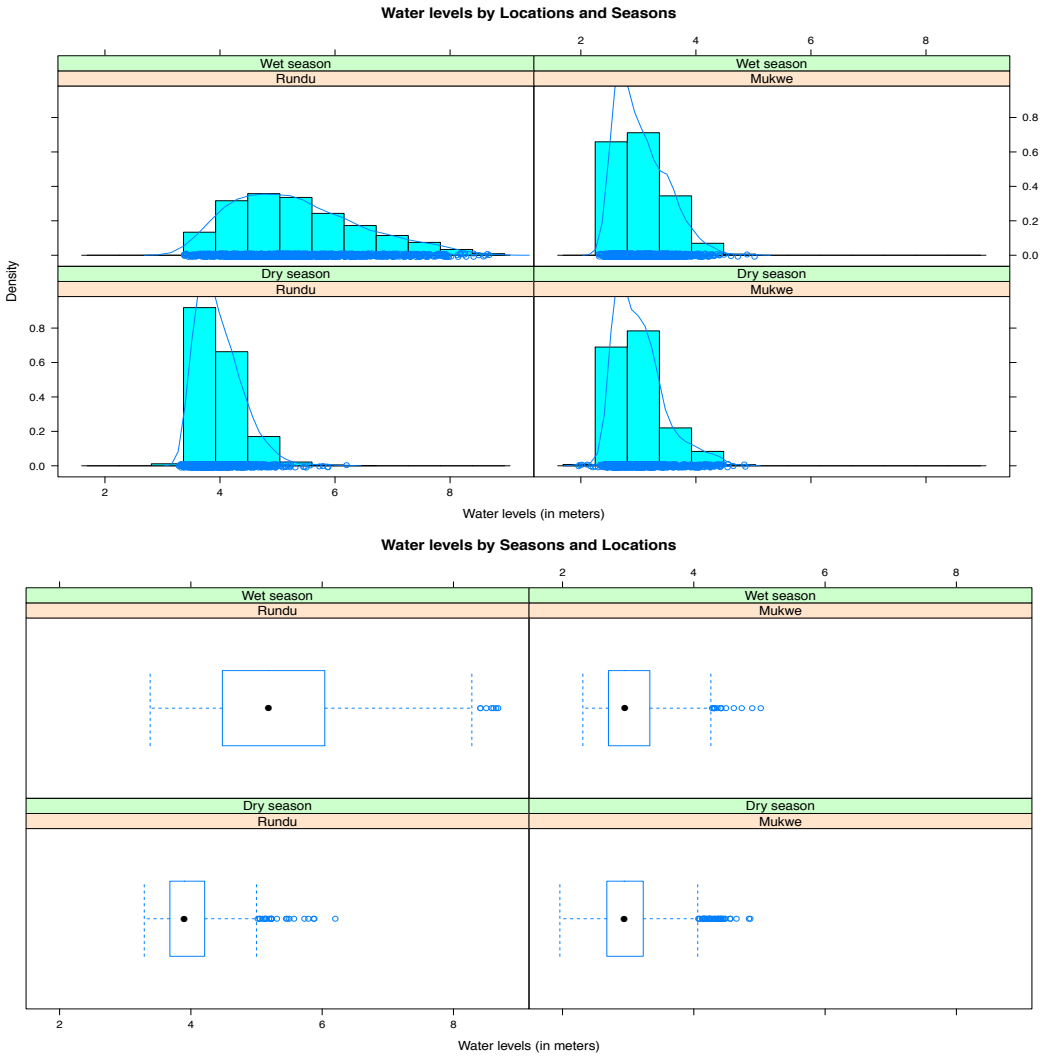


Figure 4.5: Histogram and boxplot of water levels

The variability of the weekly water levels within a season for each location is represented by Figure 4.6 and 4.8. In Figure 4.6 each line of the plot represents one of the 48 weeks in a year, which has then been followed for the study period (1950-2007). The lag 1 correlation of weekly water levels between any two subsequent years is also plotted in Figure 4.7. Autocorrelation matrices are found in the appendices in Figure 5.2 and Figure 5.3.

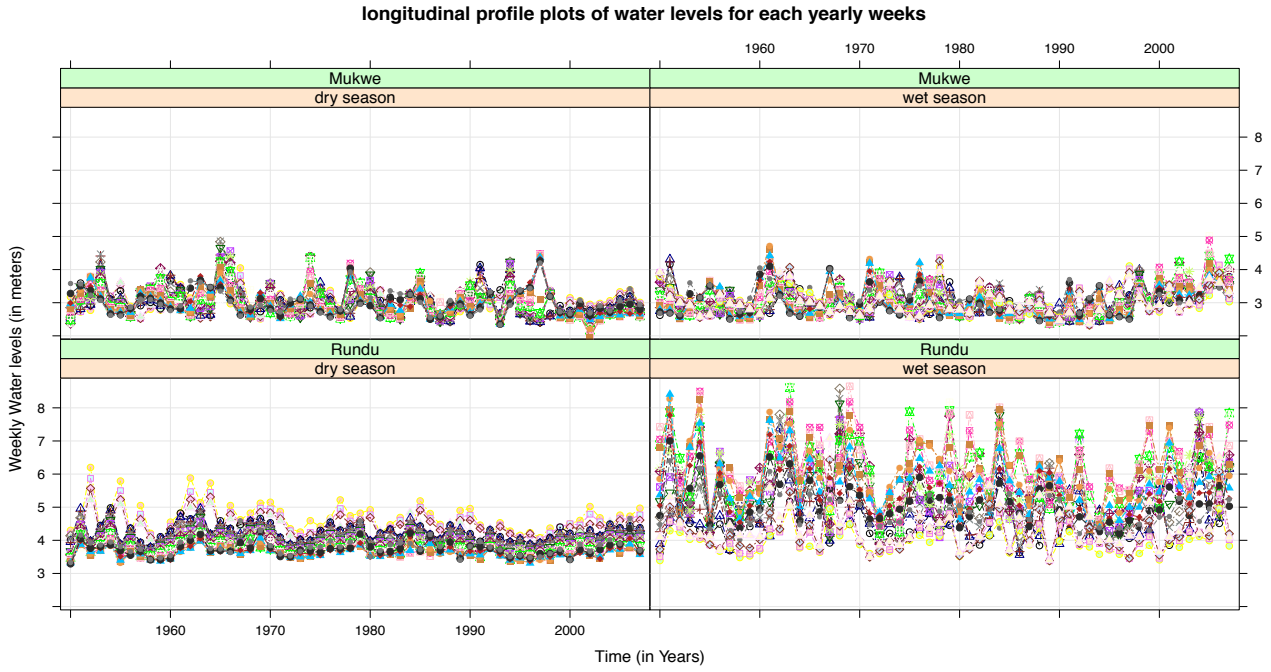


Figure 4.6: Longitudinal weekly profile plot of Water levels

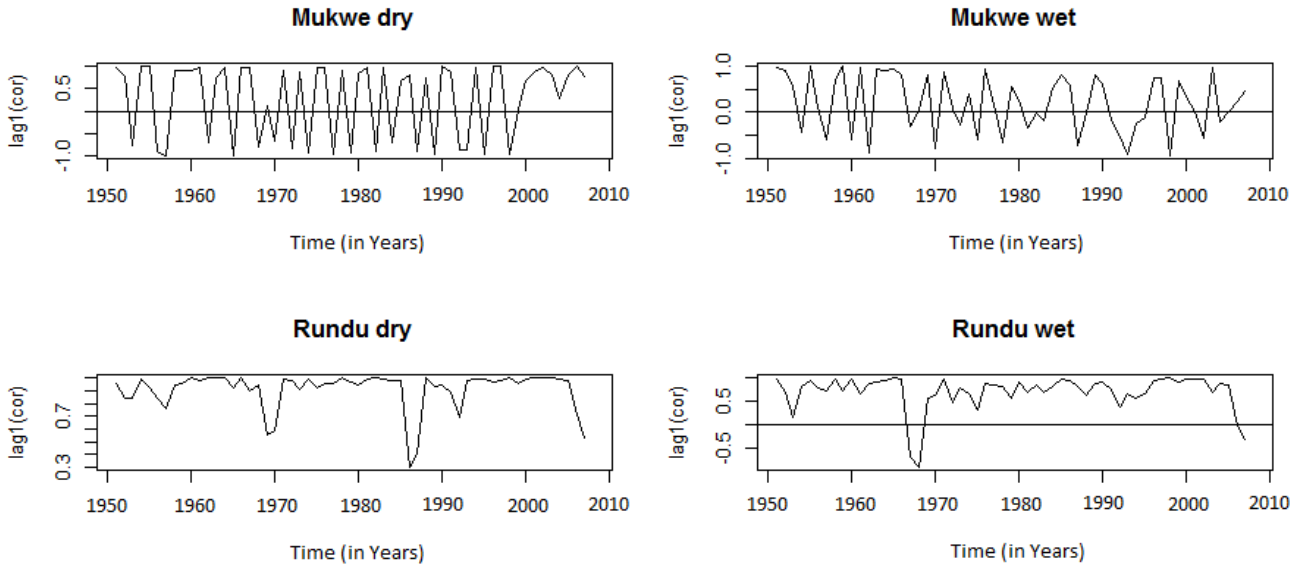


Figure 4.7: Lag1 correlation plots of weekly water levels for time period (1950-2007)

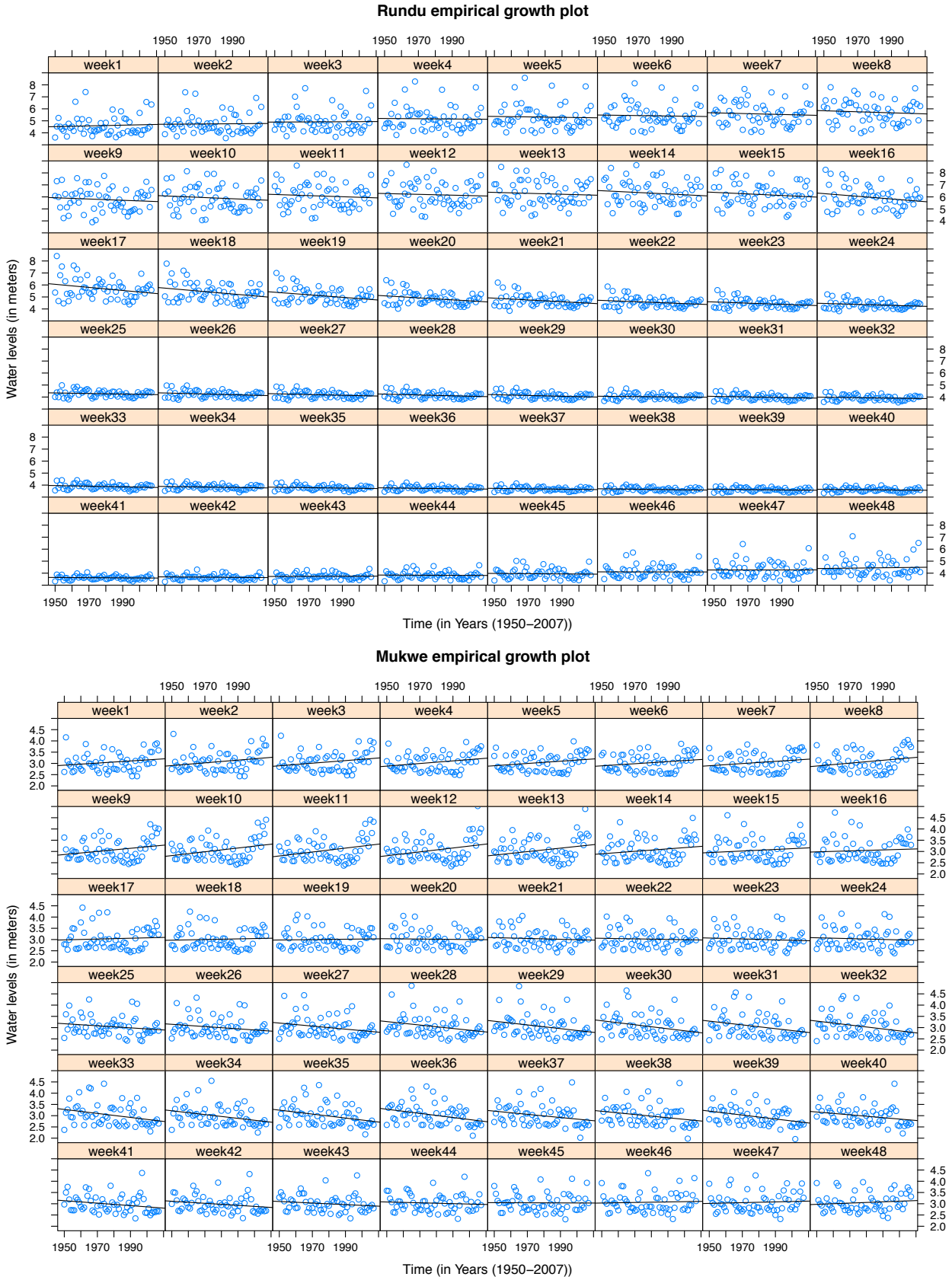


Figure 4.8: Longitudinal profile (Empirical growth) plots of weekly water levels at Rundu and Mukwe

4.3.1 Discussion

The results of Wilcoxon Rank-Sum test showed significant difference in median water levels for the two seasons (p-value < 0.001 and p-value $= 0.03902$ for Rundu and Mukwe locations, respectively). These observations were also observed when data was combined for both locations with p-value < 0.001. It is to be noted that these results do not include the time component (or treat data as cross-sectional observations). Test for difference in median weekly water levels for two locations in each year under study for each season are presented in Table 4.3.

Graphical representation of the weekly water levels using box plots, symmetry plots as well as the histograms with kernel densities indicate a suggesting that normality assumptions of these data may not be correct. This is observed deviation from symmetry, for distributions of weekly water levels for both Rundu and Mukwe during the two seasons. All these graphs indicate (show) a right-skewed distribution of data. Further numerical tests were carried out to check (confirm) these observations.

Shapiro-Wilk normality test, which is regarded by Razali and Wah (2011) to have better power as compared to other numerical statistical tests of normality like the Anderson-Darling, Kolmogorov-Smirnov and Lilliefors tests at any given significance level, was done to verify the results shown by the graphical summaries. Both the Shapiro-Wilk normality test and Anderson-Darling normality test, showed a significant departure from normality for the weekly water levels data in both locations for both seasons(see Table 4.3 where p-value < 0.001 for all tests). All the graphical summaries show a right-skewed distribution for the weekly water levels. Gamma distributions with different scale parameters are right-skewed (Figure 5.1 in the appendices). However, the weekly water level did not pass the goodness of fit for gamma distribution when tested, and neither did they pass the goodness of fit of other non-negative distributions that the data were tested against: Generalized Gamma, Log-Gamma, Chi-squared, Burr, Dagum, Erlang, Exponential, Fatique life, Frechet, Inverse Gaussian, Levy, Log-Logistic, Log-Normal, Nakagami, Pareto, Pearson, Rayleigh, Rice and the Weibull distribution. Kolmogorov-Smirnov test, Anderson-Darling test and the Chi-squared test were used as goodness of fit tests. Transformation of weekly water levels data still yielded skewed distributions as shown by Figure 5.4 in the appendices.

A linear relationship was also observed between the weekly water flows and the weekly water levels where for the months of December, January, February, March, April and May at Rundu location, the water levels increases as the water flows start rising while for Mukwe location the water level does not increase significantly based on water flows but variability in observed water levels values does. This is also shown in Table 4.4 where the Spearman rank correlation coefficients are highly positive for Rundu location and very weak for Mukwe location during the above given months.

The longitudinal and empirical growth plots in Figure 4.6 and 4.8 show how the weekly water levels change over time. Looking at the pattern in the Rundu plot, it is clearly seen that there is more variability in the water levels for the weeks 1-20 and again in the weeks 45-48 over time. The variability is seen to reduce from week 21 and remains very low through weeks 21-44, over time. The variability in the Mukwe plots however is seen to have a similar pattern throughout. The distinct pattern difference in observed water levels for Rundu were used to establish a factor variable season; two seasons dry and wet, were established. This was due to the fact that the pattern coincided with the rainfall (wet) and those months were there was no rainfall in the area of study.

The lag 1 correlation plot in Figure 4.7 shows the correlations of seasonal data between subsequent years. This plot helped in choosing the working correlation matrix for GEE models. Lag 2 to lag 10 was also plotted to check if there are changes in correlation between subsequent years (see Figures 5.2 and 5.3 in appendices). The results shows that for Rundu location the lag 1 correlations are high and positive for almost all years, while for Mukwe the correlation fluctuates between positive and negative values depending on year. The complete correlation matrices with all information are given in Figures 5.2 and 5.3 in the appendices. From these matrices it is observed that at Rundu correlations are continuously positive for most subsequent years, while at Mukwe these change from year to year. Thus, the correlation of water levels at Rundu seems to be similar to that of the Exchangeable correlation matrix while that at Mukwe seems to be similar to that of an Unstructured correlation matrix.

Table 4.4: Spearman rank correlation coefficients of water levels and water flows

	Wet season							Dry season						
Locations	DEC	JAN	FEB	MAR	APR	MAY	season corr coef.	JUN	JUL	AUG	SEP	OCT	NOV	season corr coef.
Rundu	0.9415	0.9628	0.9433	0.9579	0.9356	0.9212	0.9637	0.3392	0.9798	0.331	0.9728	0.9522	0.9348	0.9113
Mukwe	-0.0194	-0.0513	-0.1069	0.0027	0.0406	0.0893	-0.0265	0.0877	0.1827	0.2365	0.2226	0.1722	0.105	0.1306

4.3.2 Residual Analysis

Since a GEE is a GLM object, GLM was fitted to the dataset in order to analyse the residuals of the GLM. The above Figure 4.9 is a plot of correlation from GLM to asses the correlation of residuals in the dataset which was later on used to helped fit a good GEE model. From the plot it can be observed that the residuals have a cyclic pattern which repeat approximately every 12^{th} (quarter) week of the 48 weeks in a year. Thus, a quarter variable was coded to the data set in order for GEE models to produce better results.

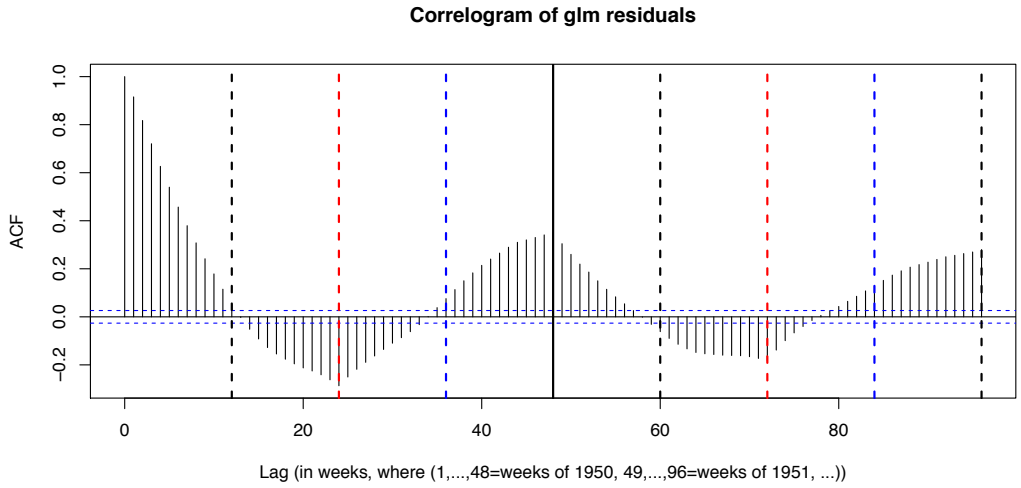


Figure 4.9: GLM correlation of residuals

4.4 EMPIRICAL (ROBUST) GEE AND GEE MODELS WITH DIFFERENT CORRELATION STRUCTURES

Table 4.5 shows parameter estimates from a GEE model. Different working correlation structures were used to get estimates of parameters. Assessment of model bias, accuracy, model fit and model selection was evaluated using the bias, MSE, coefficient of determination (R^2) and QIC respectively. Here bias was a process of getting predicted water level values that over or under estimate the true observed water level values. QIC and QIC_u were used to select the best GEE model. The model with a lower value of QIC is usually the preferred model while QIC_u can be used to compare models with the same correlation structures but different mean structures (Cui and Qian, 2007). Assessment of bias was necessary in this study as some of the models produced a lower QIC but had bigger bias values, this implies that QIC alone in this case was not a good indicator of choosing the best model but must be used together with other methods of model selection (e.g, MSE, R^2 and standard errors) and vice versa.

Figure 4.10 shows a plot of relationship between GEE models, QIC, MSE, R^2 and significance of the coefficients for water flow when sample size changes. It can be seen that the QICs for models with Exchangeable, AR1 and Independence correlation structures increase faster as sample size increases as compared to the QIC of the robust GEE model. The MSEs for these GEE models also rise with increasing sample size, but seem to reach a peak as sample size gets larger. The R^2 seem to be constant when sample size is big enough but when the sample size is small there seem to be a small variation especially for the GEE with exchangeable and AR1 correlation structure. The observed QIC were however, relatively moderate for all tested GEE models. The p-values for 2^{nd}, 3^{rd} and 4^{th} bin splines of water flows were very significant for all the tested sample sizes, while the 1^{st} bin spline produced non-significant coefficients at large sample sizes as given in Table 4.5. The generalised linear models makes assumption of linearity on link scale, constant variance of residuals and independence of error term (for a GEE object the error terms are independent between clusters but correlated within a cluster).

Table 4.5: Parameter estimates of a GEE models

GEE model estimates: dependent variable is y= *Water levels*

Model parameters	Robust	(Std.err)	Unstr	(Std.err)	Exch	(Std.err)	AR1	(Std.err)	Ind	(Std.err)
(Intercept)	3.88***	($9.55e^{-02}$)	3.138365	(12.949652)	3.33***	($1.74e^{-01}$)	3.26***	($1.14e^{-01}$)	3.94***	($9.60e^{-02}$)
BS.1(flows)	$-9.30e^{-02}$	($1.81e^{-01}$)	0.688167	(0.943852)	$5.68e^{-01}$***	($7.78e^{-02}$)	$4.95e^{-01}$***	($6.86e^{-02}$)	$-1.85e^{-01}$	($1.69e^{-01}$)
BS.2(flows)	-1.30**	($4.38e^{-01}$)	2.376621	(1.567235)	2.17***	($1.76e^{-01}$)	1.84***	($1.55e^{-01}$)	-1.36***	($3.79e^{-01}$)
BS.3(flows)	2.40·	(1.30)	2.306242	(5.098660)	2.58***	($3.40e^{-01}$)	2.23***	($2.64e^{-01}$)	1.79	(1.12)
BS.4(flows)	-3.91**	(1.52)	1.738537	(2.634341)	2.04***	($3.18e^{-01}$)	1.96***	($1.87e^{-01}$)	-2.55**	($9.87e^{-01}$)
zTime(T)	$-6.09e^{-05}$**	($2.21e^{-05}$)	-0.000126	(0.002801)	$-7.67e^{-05}$·	($4.22e^{-05}$)	$-6.61e^{-05}$*	($3.02e^{-05}$)	$-6.11e^{-05}$**	($2.25e^{-05}$)
Seasons(S)	$6.53e^{-01}$***	($1.26e^{-01}$)	-0.431022	(12.944807)	$-4.36e^{-01}$*	($1.77e^{-01}$)	$-2.28e^{-01}$·	($1.18e^{-01}$)	$6.93e^{-01}$***	($1.30e^{-01}$)
T : S	$1.21e^{-05}$	($4.57e^{-05}$)	0.000125	(0.003112)	$5.32e^{-05}$	($4.37e^{-05}$)	$8.14e^{-05}$**	($3.11e^{-05}$)	$1.87e^{-05}$	($4.58e^{-05}$)
ψ	1.16	(0.071)	0.0909	(0.0605)	0.0828	(0.00283)	0.0792	(0.00276)	0.0708	(0.00355)
φ	0		-	-	0.941	(0.0215)	0.979	(0.005)	0	
$\hat{y}$ bias	$-5.77e^{-15}$		-0.0635		0.0945		0.0439		1.6e-06	
R^2	0.531		0.561		0.552		0.547		0.531	
MSE	1.16		1.16		1.16		1.16		1.16	
RMSE	1.08		1.08		1.08		1.08		1.08	
QIC	5250		8501		13561		13700		13784	
QIC_u	5181		14230		13543		13697		13722	

Signif. codes: ***p<0.001; **p<0.01; *p<0.05; ·p<0.1

BS(flows)=coefficients of the matrix of smoothing basis-splines of water flows; ψ=scale(dispersion) parameter; φ=correlation parameter

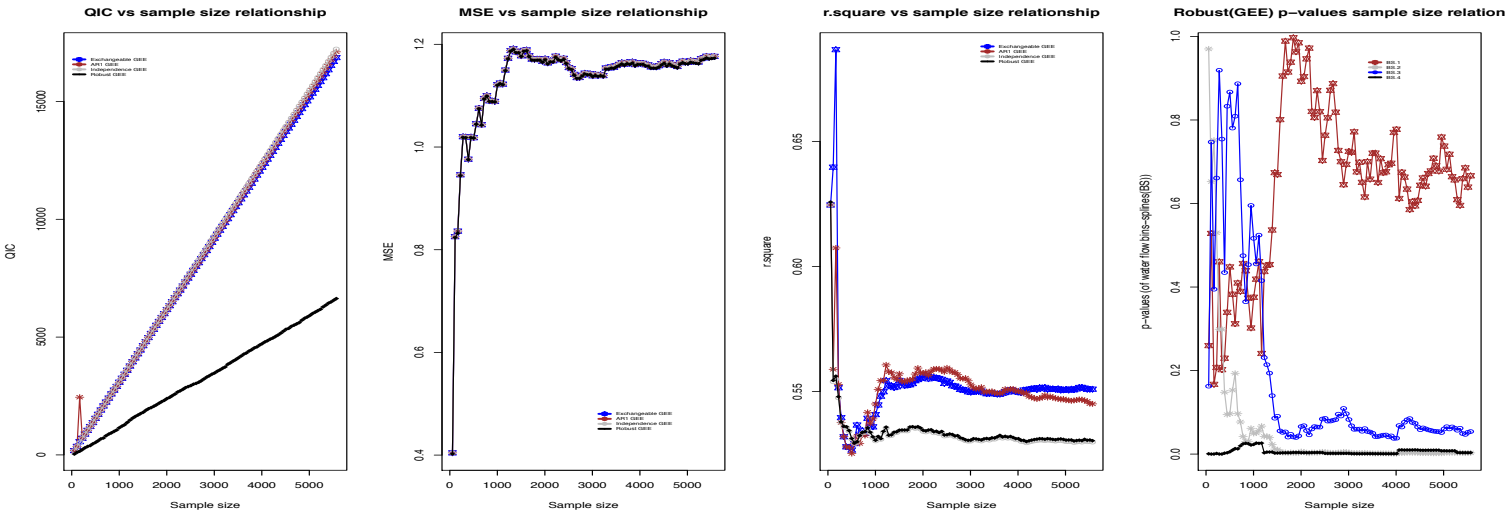


Figure 4.10: Relationship between, QIC, MSE, R^2 and p-values with the sample size for the GEE models

4.4.1 Linearity On The Link Scale

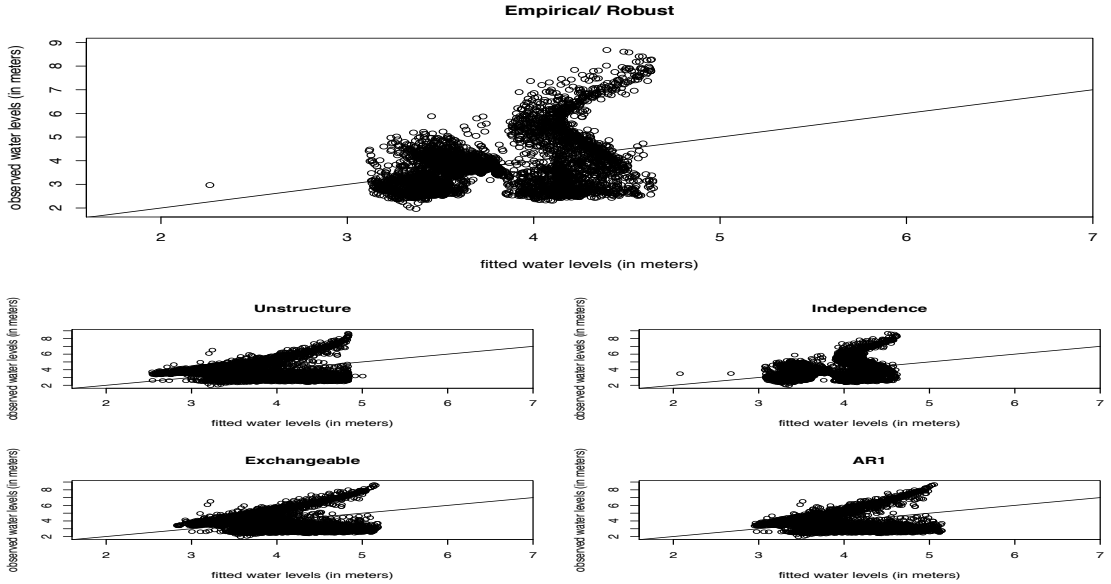


Figure 4.11: Observed vs Predicted values for different GEE models

Linearity on link scale, where the link function in this case is the identity, can be assessed by a plot of observed values plotted against the predicted values (linear predictor η as given in Equation (3.3.1)). An ideal model has a one to one relationship of the observed values and the linear predictor η. Figure 4.11 shows that none of the fitted GEE models have a perfect linear relationship between the observed water levels and their linear predictor η, however, for GEE with unstructured, exchangeable and ar1 correlation structure, the relationship was a little linear with values deviating in two different direction as water levels get bigger. On the other hand, GEE with independence structure the linear relationship was clearly violated. When the residuals was added to the predicted values, the Robust GEE and GEE with independence structure predicted the observed values accurately and hence produced a perfect observed-predicted value relationship as shown in Figure 5.5, a good plot was also produced by the GEE with Unstructured, Exchangeable and AR1 correlation structured but could not give a perfect plot as compared to that from robust GEE and GEE with independence correlation structure. This indicated that the Robust and GEE with independence correlation structures had the accurate residuals as compared to the other three. it is also worth noting that under the robust GEE and GEE with independence structure, outliers were detected but this was not the case

for the other fitted GEE models.

4.4.2 Assessing for Mean-Variance Relationship

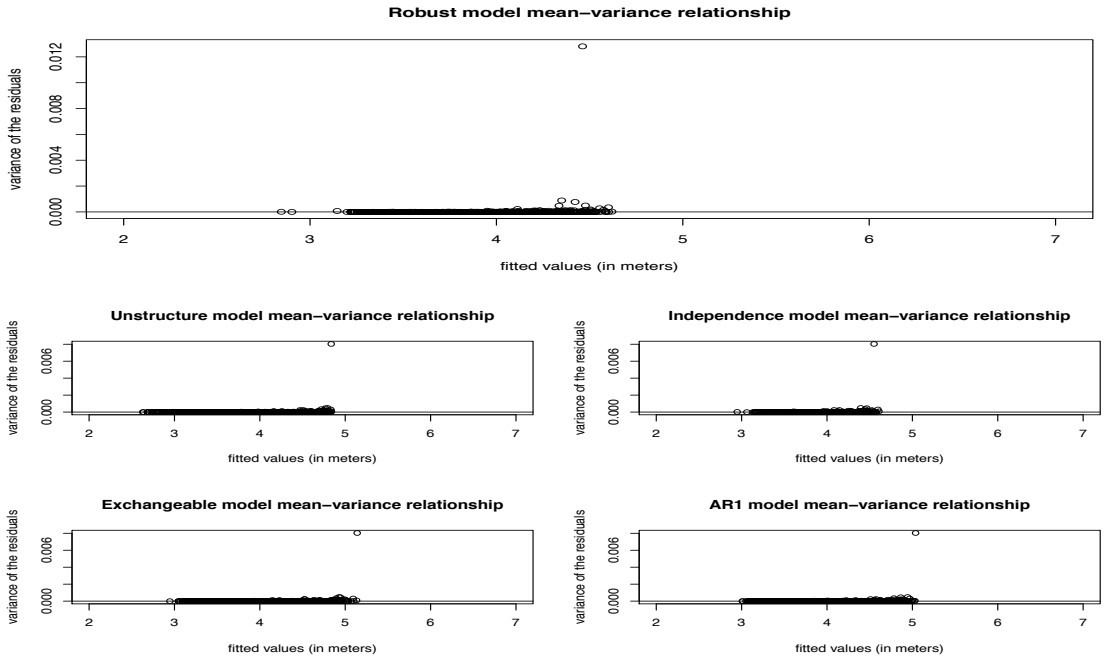


Figure 4.12: Mean-variance relationship for different GEE models

The mean-variance relationship can be assessed through a plot of residuals plotted against predicted values. The ideal model has residuals scattered around zero for every predicted value or variance of zero for all predicted values (Mackenzie and Scott-Hayward, 2015). The plot of residuals Figure 5.7 in the appendices shows that the residuals are scattered around zero for all the tested GEE models in this study. In Figure 4.12 the residuals and predicted values was divided into bins of 2000 equal data points of which the variance of residuals was calculated for each bin (this was done since the mean and variance of the whole residuals are fixed point and can not be used to analyse the relationship) and the mean of each bin was calculated for the fitted values. For each bin, the variance of residuals was then plotted against the mean of the predicted values. A constant mean-variance relationship was observed for all of the tested GEE models in Table 4.5 which is not so different from that of an ideal model.

4.4.3 Assessing for Non-independence in Model Residuals

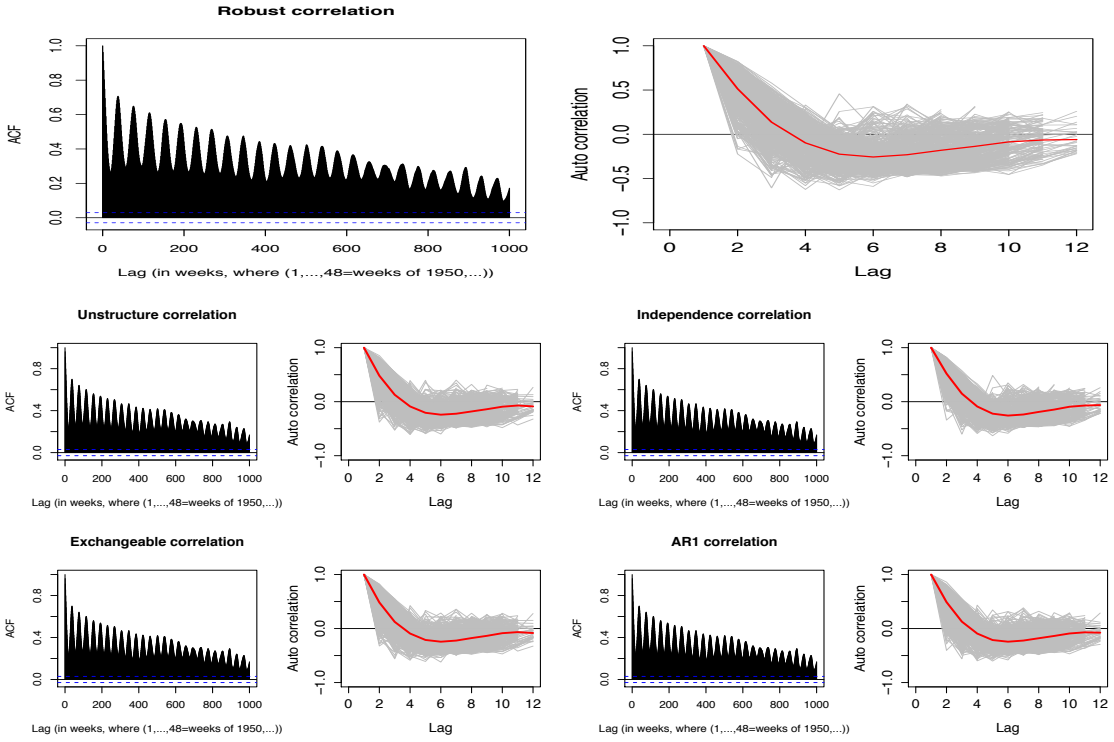


Figure 4.13: Non independence in model residuals for different GEE models

The GEE error terms are correlated within a block (cluster) but are independent between blocks. The Auto correlation function (ACF) is one way to check for residuals correlation. The ACF in Figure 4.13 can not give information on the expressing of doubt that the residuals are correlated since for GEE the ACF are only correlated within a block. However, the auto-correlation plot next to the ACF plot in Figure 4.13 shows that the correlation between blocks (clusters) is about zero as shown by the mean block correlation (solid line) while within clusters is given by the spaghetti plots which clearly indicate that the residuals are correlated within their clusters.

4.4.4 Discussion

In addition to the empirical (robust) GEE model, four working correlation structures, the Un-structured, Exchangeable, First Order Autoregressive (AR1) and the Independence working

correlation matrices were selected to model the GEE models in Table 4.5. From preliminary analysis as supported by Figure 5.2 and 5.3, the weekly water levels was estimated to have a fixed correlation over time, which is consistent with what is theoretically observed under an Exchangeable working correlation structures for weekly water levels at Rundu location while the weekly water levels at Mukwe location was estimated to have an Unstructured working correlation structure. Mackenzie and Scott-Hayward (2015) mentioned that robust GEE model is usually preferred in the case of not wanting to rely on making a choice on selecting a correlation structure. They also pointed out that robust GEE only depends on the blocking (clustering) structure and is very accurate when having balanced longitudinal data like in this case, and that robust GEE returns very reliable standard errors as compared to the other GEE models.

Based on the GEE models in Table 4.5, the Robust GEE model had the smallest QIC=5 250 which is quite smaller than the QIC of other GEE models in Table 4.5, thus according to Cui and Qian (2007), Robust GEE model is the preferred model for this longitudinal data. The R^2 of the Robust GEE model was also moderate (R^2= 0.531), while that of other tested correlation structures range from $R^2 \in [0.531, 0.561]$. The Robust GEE and GEE with independence structure has very small bias as compared to GEE with unstructured, exchangeable and AR1 correlation structure. On the other hand, the standard errors of the $\hat{\beta}$ coefficients are very small on the AR1 model as compared to the robust and any other GEE model tested. Furthermore, the AR1 model has also picked up all $\hat{\beta}$ coefficients to be significant in the model while other GEE models could not pick up some $\hat{\beta}$ coefficients, especially the interaction of time and season as significant.

In addition to this, a Wilcoxon signed rank test was done to check the median difference between the predicted and observed weekly water levels of this GEE models where under the Robust GEE model, GEE model with Independence, Exchangeable and AR1 correlation structure gave a $p - value = 1e^{-06}$, $p - value = 2e^{-07}$, $p - value = 5e^{-13}$ and $p - value = 3e^{-07}$ respectively, which implies that there is enough evidence that the medians of the observed and predicted weekly water levels differ, while the GEE model with Unstructured correlation matrix gave a $p - value = 0.06$ which indicates that the observed and predicted weekly water level have the same median. The robust GEE model does its calculation based on the Independence correlation matrix. Thus the correlation parameter of the robust GEE and

GEE with independence structure was zero while the correlation matrices of the GEE with Unstructured, Exchangeable and AR1 structures were as given below.

$$\begin{bmatrix} 1 & 0.972 & 0.999 & 1.011 & 1.006 & . & . & . & 0.808 \\ 0.972 & 1 & 1.036 & 1.042 & 1.036 & . & . & . & 0.828 \\ 0.999 & 1.036 & 1 & 1.086 & 1.036 & . & . & . & 0.824 \\ 1.011 & 1.042 & 1.086 & 1 & 1.101 & . & . & . & 0.2.68 \\ 1.006 & 1.036 & 1.036 & 1.101 & 1 & . & . & . & 0.792 \\ . & . & . & . & . & . & & & . \\ . & . & . & . & . & & . & & . \\ . & . & . & . & . & & & . & . \\ 0.808 & 0.828 & 0.824 & 0.811 & 0.792 & . & . & . & 1 \end{bmatrix}$$

Unstructured correlation

;

$$\begin{bmatrix} 1 & 0.941 & 0.941 & 0.941 & . & . & . & 0.941 \\ 0.941 & 1 & 0.941 & 0.941 & . & . & . & 0.941 \\ 0.941 & 0.941 & 1 & 0.941 & . & . & . & 0.941 \\ 0.941 & 0.941 & 0.941 & 1 & . & . & . & 0.941 \\ . & . & . & . & . & & & . \\ . & . & . & . & & . & & . \\ . & . & . & . & & & . & . \\ 0.941 & 0.941 & 0.941 & 0.941 & . & . & . & 1 \end{bmatrix}$$

Exchangeable correlation

$$\begin{bmatrix} 1 & 0.979 & 0.979^2 & 0.979^3 & . & . & . & 0.979^{11} \\ 0.979 & 1 & 0.979 & 0.979^2 & . & . & . & 0.979^{10} \\ 0.979^2 & 0.979 & 1 & 0.979 & . & . & . & 0.979^9 \\ 0.979^3 & 0.979^2 & 0.979 & 1 & . & . & . & 0.979^8 \\ . & . & . & . & . & & & . \\ . & . & . & . & & . & & . \\ . & . & . & . & & & . & . \\ 0.979^{11} & 0.979^{10} & 0.979^9 & 0.979^8 & . & . & . & 1 \end{bmatrix}$$

$First-order\ Autoregressive\ correlation$

The linear mixed model for analysis of the longitudinal weekly water levels of the Okavango river data was fitted as given in Equation (4.4.1) where a snapshot of the fitted weekly water level are summarised in Table 4.6;

$$\hat{\mathbf{y}}_{it} = \mathtt{model} + \epsilon_{it} \tag{4.4.1}$$

where in Equation (4.4.1), $\mathtt{model} = (\hat{\beta}_0 + \hat{\beta}_1 * \mathsf{BS}.1_{it} + \hat{\beta}_2 * \mathsf{BS}.2_{it} + \hat{\beta}_3 * \mathsf{BS}.3_{it} + \hat{\beta}_4 * \mathsf{BS}.4_{it} + \hat{\beta}_5 * \mathsf{T}_{it} + \hat{\beta}_6 * \mathsf{S}_{it} + \hat{\beta}_7 * (\mathsf{T} : \mathsf{S})_{it})$.

$\hat{\mathbf{y}}_{it}$ is the value of estimated weekly water levels at a any given point in time (t) while ϵ_{it} is the error term for the i^{th} predicted weekly water level at time t. The $\mathsf{BS}.1_{it}, \mathsf{BS}.2_{it}, \mathsf{BS}.3_{it}, \mathsf{BS}.4_{it}$ represents the continuous smoothing bin-splines of weekly water flow covariate for observation i at time t, while the discrete covariates T_{it}represents zTime, S_{it} represents Season, $(\mathsf{T} : \mathsf{S})_{it}$ represents interaction of zTime with Season. The linear combination of $(\hat{\beta}_0 + \hat{\beta}_1 * \mathsf{BS}.1_{it} + \hat{\beta}_2 * \mathsf{BS}.2_{it} + \hat{\beta}_3 * \mathsf{BS}.3_{it} + \hat{\beta}_4 * \mathsf{BS}.4_{it} + \hat{\beta}_5 * \mathsf{L}_{it} + \hat{\beta}_6 * \mathsf{S}_{it} + \hat{\beta}_7 * (\mathsf{L} : \mathsf{S})_{it})$ might describe the relationship between weekly water level and the covariates well, but in practice it will never describe the weekly water level exactly. Thus, the difference between the observed weekly water level and predicted weekly water level was considered to be the model error (ϵ_{it}). According to

Twisk (2003), each regression coefficients $(\hat{\beta}_1, ..., \hat{\beta}_7)$ for a particular predictor variable "relates the vector of weekly water levels over time to the vector of predictor variable over time". Twisk (2003) mention that the $(\hat{\beta}_1, ..., \hat{\beta}_7)$ can not be interpreted straightforward as it combines the between cluster and within cluster relationship which then resulted in a single $(\hat{\beta}_1, ..., \hat{\beta}_7)$ regression coefficient. The intercept coefficient $\hat{\beta}_0 = 3.88$ gives an average estimate of weekly water levels when other regression coefficients $(\hat{\beta}_1, ..., \hat{\beta}_7)$ are assumed to be zero.

The error of the Robust GEE model was modelled using the Dagum (4p) distribution of 4 parameters, $\epsilon \sim \text{Dagum}(k, \alpha, \lambda, \gamma)$, where $k > 0$ and $\alpha > 0$ are continuous shape parameters, $\lambda > 0$ is a continuous scale parameter and γ is a location parameter. For $\gamma \equiv 0$ one gets the Dagum (3p) distribution. The error term for the Robust GEE model had $k = 0.38887$, $\alpha = 4.8405$, $\lambda = 2.7093$ and $\gamma = 2.0702$. The probability density function (PDF) of Dagum (4p) is given as;

$$f(x) = \frac{\alpha k \left(\frac{x-\gamma}{\lambda}\right)^{\alpha k - 1}}{\lambda \left(1 + \left(\frac{x-\gamma}{\lambda}\right)^{\alpha}\right)^{k+1}}; \; k, \alpha, \lambda > 0, -\infty < \gamma < \infty$$

The Rammsey's RESET test which is used to test if there are powers of some predictor variables left out in the model which might result in a better model if taken into consideration gave a $p - value = 2e^{-16}$ for all GEE models which implies strong evidence that there might be some powers of predictor variables not accounted for. However, adding the n^{th}-root power or a power of 2^{nd}, 3^{rd} and 4^{th} order term to these GEE models did not make any marginal statistical contribution.

In the end, it is worth noting that although the Robust GEE had the smallest QIC, it had bigger standard errors as compared to the GEE model with an AR1 correlation structure. However, Mackenzie and Scott-Hayward (2015) mentioned that the robust GEE model produce more reliable standard errors as it does not estimate its model parameters based on correlation matrix as compared to the GEE models with AR1, Unstructured, and Exchangeable correlation structures.

Table 4.6: Table of estimated weekly water levels by Robust GEE model

```
> print(Latex_Data_sample , row.names = F)
 Year Months Seas W.levels W.flows Loc Loc_Time_Order     id zTime  BS.1    BS.2       BS.3         BS.4 fitted(y)
 1961     11    0     3.66    38.7   1            569 196114  1241 0.305 0.00650 0.00004465 0.0000000000       3.77
 1983      2    1     3.27   262.9   2           1589 198321   345 0.643 0.31773 0.03974237 0.0000158250       4.14
 1967      9    0     3.85    57.3   1            849 196714  1833 0.470 0.01837 0.00022550 0.0000000000       3.70
 1999      7    0     4.40   117.2   1           2377 199913  5025 0.761 0.08769 0.00290584 0.0000000000       3.40
 1981     11    0     3.51    28.6   1           1529 198114  3241 0.197 0.00248 0.00001022 0.0000000000       3.66
 1995      8    0     3.73    35.1   1           2192 199513  4632 0.811 0.00488 0.00002874 0.0000000000       3.57
 1968      4    1     6.25   409.7   1            879 196812  1915 0.414 0.44600 0.13573434 0.0039241690       4.13
 1979      9    0     3.36   211.8   2           1425 197924  3033 0.737 0.24202 0.02007722 0.0000000000       3.36
 1961      4    1     3.95   629.5   2            541 196122  1213 0.185 0.44315 0.32662878 0.0452034451       4.49
 1990      6    0     2.71   286.2   2           1943 199023  4123 0.602 0.34664 0.05143675 0.0000981794       3.25
 1989      9    0     3.80    67.2   1           1906 198914  4034 0.542 0.02682 0.00041105 0.0000000000       3.55
 1969      8    0     3.01   283.1   2            941 196923  2029 0.607 0.34306 0.04981976 0.0000816718       3.37
 2000     11    0     3.47    25.3   1           2441 200014  5141 0.159 0.00157 0.00000506 0.0000000000       3.55
 1991     11    0     2.91   133.5   2           2010 199124  4242 0.788 0.11227 0.00448358 0.0000000000       3.41
 1971      2    1     2.89   500.0   2           1016 197121   228 0.305 0.46856 0.21241515 0.0136905761       4.34
 1973      5    1     2.74   316.2   2           1124 197322  2420 0.552 0.37886 0.06873116 0.0003923856       4.03
 1989      7    0     3.12   239.4   2           1899 198923  4027 0.685 0.28497 0.02963288 0.0000000585       3.27
 1960     12    1     2.65   208.4   2            528 196021  1148 0.743 0.23649 0.01906909 0.0000000000       4.15
 1983      1    1     4.65    90.6   1           1587 198311   343 0.672 0.05192 0.00120052 0.0000000000       4.39
 1972      7    0     2.75   195.9   2           1084 197223  2328 0.763 0.21578 0.01564639 0.0000000000       3.42
>>
```

Note: Table of estimated water level values using parameters from the model and some observed informations. Seas is short for season, W.levels is short for water levels, W.flows is short for water flows, Loc is short for Location, Loc_Time_Order is short for time order within a location. BS.1, BS.2, BS.3 and BS.4 are the calculated smoothing B-splines of water flows.

4.4.5 GEE Models Plots

Figure 4.14 shows a plot of longitudinal observed weekly water level plotted together with their predicted weekly water levels for Rundu and Mukwe where predicted values was done using the Robust GEE model. The top plot has model residuals added to the model predicted weekly water levels which makes the observed weekly water levels and (predicted weekly water levels + residuals) exactly the same, the middle plot is a plot of observed weekly water levels and predicted weekly water levels while the bottom plot is also a plot of predicted weekly water levels against observed weekly water levels where the predicted weekly water levels was manually calculated using model $(\hat{\beta}_0, ..., \hat{\beta}_7)$ coefficients and observed informations as given in Equation (4.4.1). The residuals and partial autocorrelation of the residuals for the Robust model are given in Figure 4.15.

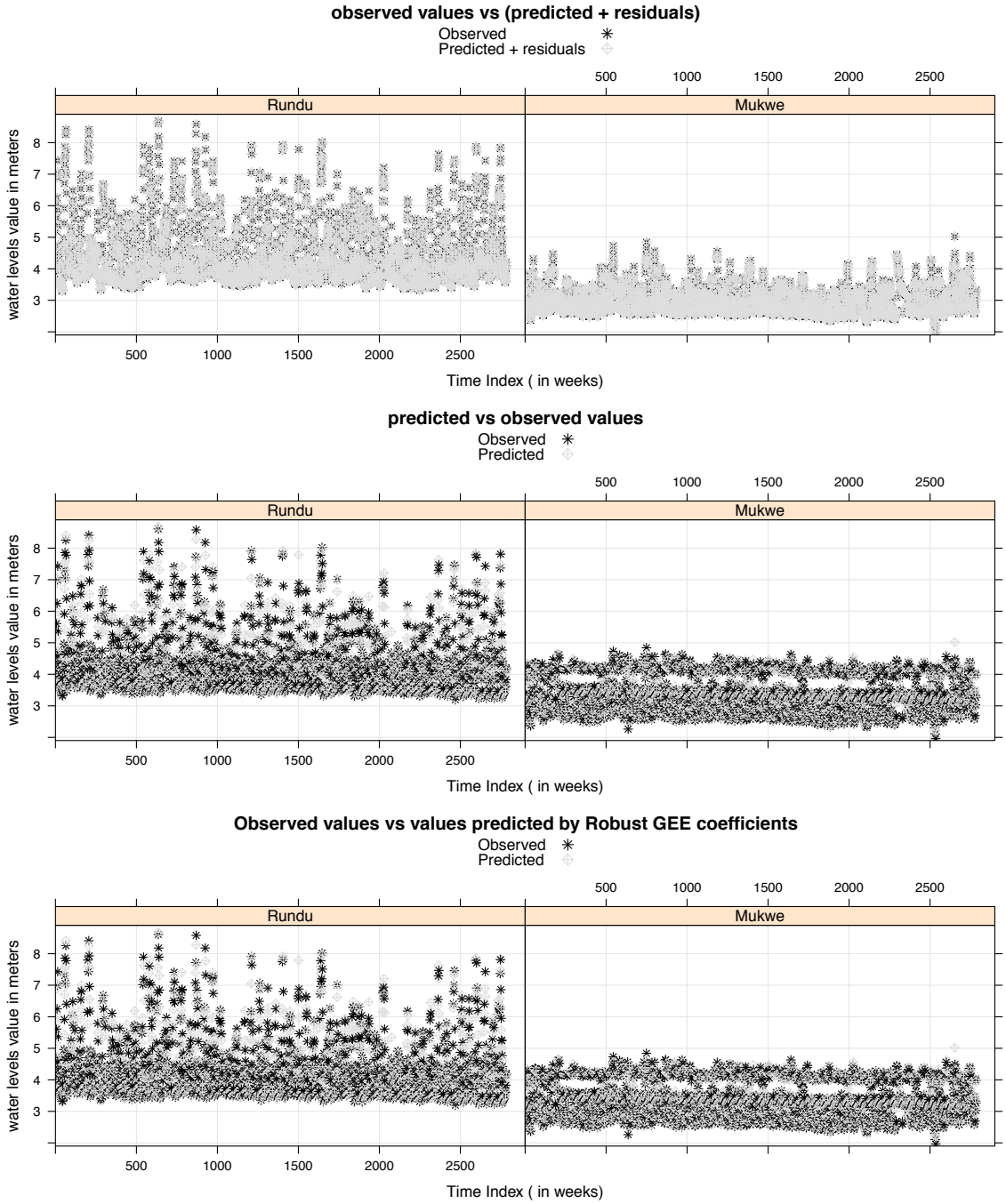


Figure 4.14: Longitudinal plots of predicted and observed values for GEE model

The residual plot in Figure 4.15 shows that the difference in predicted weekly water levels and observed weekly water levels is mostly centred around zero. An ideal model has residuals with zero mean and residual values that are well scattered (centred) near zero. Positive values of residuals indicate under prediction of fitted weekly water levels and negative residuals indicate over prediction of fitted weekly water levels. Since our residuals was centred around zero and

very small, we would expect the predicted weekly water levels from our model to be very close to the observed weekly water levels. The test for randomness in the residuals for the Robust GEE model produced a $p - value < 2e^{-16}$ where the alternative hypothesis was non-randomness in the residuals. This implies that our GEE model was not misspecified. The partial autocorrelation is also showing a trend close to what is usually expected in an ideal model where the dependency among the correlations is about zero, for example residuals at time lag 3 does not depend on residuals at previous past time lags. The residuals of GEE with an AR1 and Exchangeable correlation structures do not seem to differ significantly as those of the Robust GEE model (see Figure 5.6 in the appendices) for comparison. However, based on Figure 5.5, the GEE with Unstructured, Exchangeable as well as AR1 structured seem to produce wrong residuals as when added to the weekly predicted water level under such model does not add up to the observed weekly water levels.

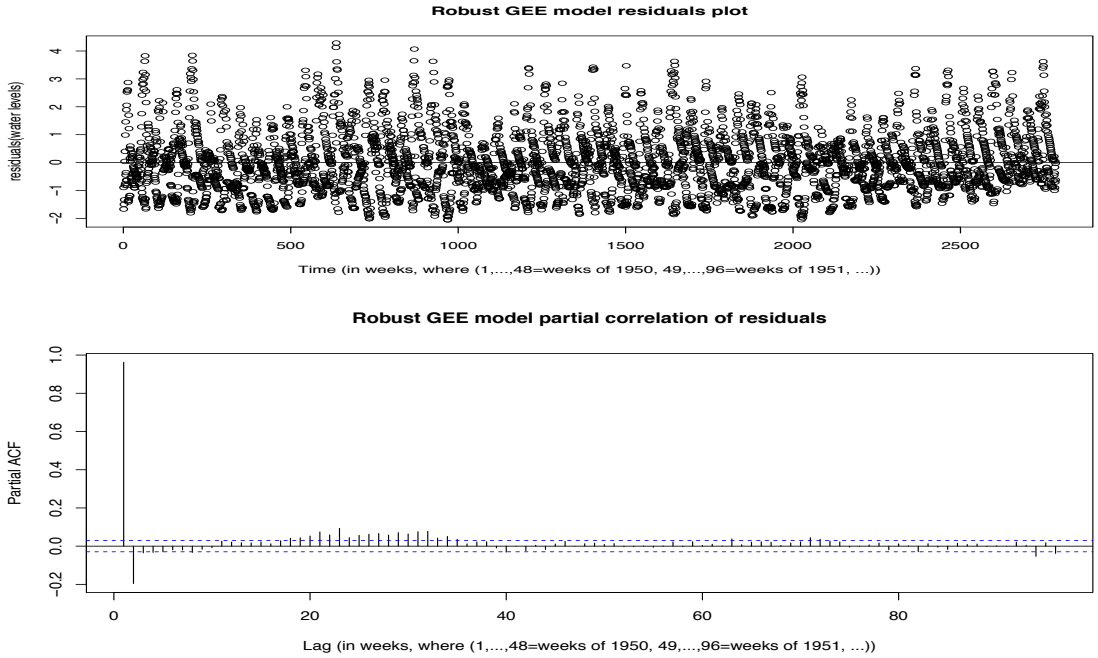


Figure 4.15: Residual plots for Robust GEE model

4.5 REML MODEL

Table 4.7 shows a generalised linear mixed model estimated using REML method for the weekly water levels. The assessment of bias, accuracy and model fit was evaluated using the Bias, MSE

and R^2 respectively. The model selection was done using the AIC (BIC), which is an equivalence of the QIC for model which are likelihood based. Model with small values of AIC are better at estimation. Figure 4.16 show plots of AIC (BIC), MSE, R^2 and water flows p-values at different samples sizes. It can be observed that for the REML model, when sample size increases, so is the AIC (BIC) and the MSE where AIC (BIC) increases linearly until it reachs a peak and start deceasing as sample size gets bigger. R^2 on the other hand decrease as the sample size increases and seem to reach its base of $R^2 = 0.975$ when sample size was not more than 3977.

The p-values for water flows was very significant for all the tested sample sizes, indicating that water flows was not only significant at large sample size as given in Table 4.7. MSE seem to converge to its peak at approximately MSE=0.034 when sample size was 4773. Generalised linear mixed models (GLMM) makes assumptions of linearity on link scale, constant variance of error and independence of the error term. Unlike the general linear model that require the response variable to be normally distributed, GLMM extends the general linear models by allowing the response variable to follow different distributions and use link function as the transform when fitting a GLMM.

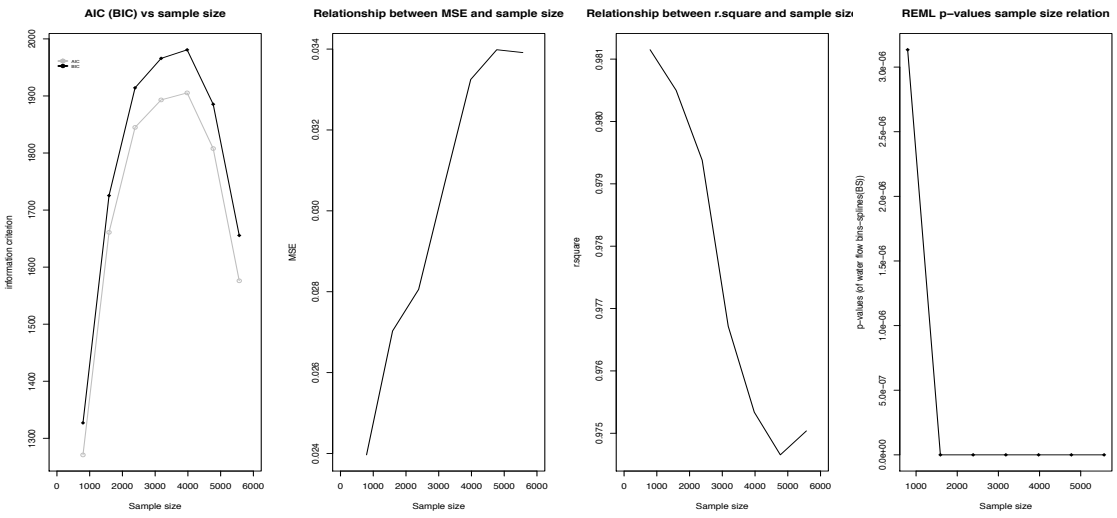


Figure 4.16: Relationship between, AIC, MSE, R^2 and p-values with the sample size

Table 4.7: REML parameter estimates

	depended variable:Water levels	
Model Parameters	REML Constants	(Std.err)
Intercept	3.21***	($1.10e^{-01}$)
BS.1(flows)	0.686***	($5.50e^{-02}$)
BS.2(flows)	1.79***	($1.33e^{-01}$)
BS.3(flows)	1.9***	($2.64e^{-01}$)
BS.4(flows)	2.69***	($3.80e^{-01}$)
zTime	-0.0001***	($3.00e^{-05}$)
Seasons	-0.191*	($1.13e^{-01}$)
zTime : Seasons	0.0001**	($3.00e^{-05}$)
$\widehat{y}$ bias	0	
R^2	0.975	
MSE	0.0338	
RMSE	0.184	
Log Likelihood	-907.000	
AIC	1837.00	
BIC	1914.00	

Signif.code *p<0.1; **p<0.05; ***p<0.01; Note: BS(flows)=coefficients of the matrix of smoothing basis-splines of water flows;

4.5.1 Linearity On The Link Scale

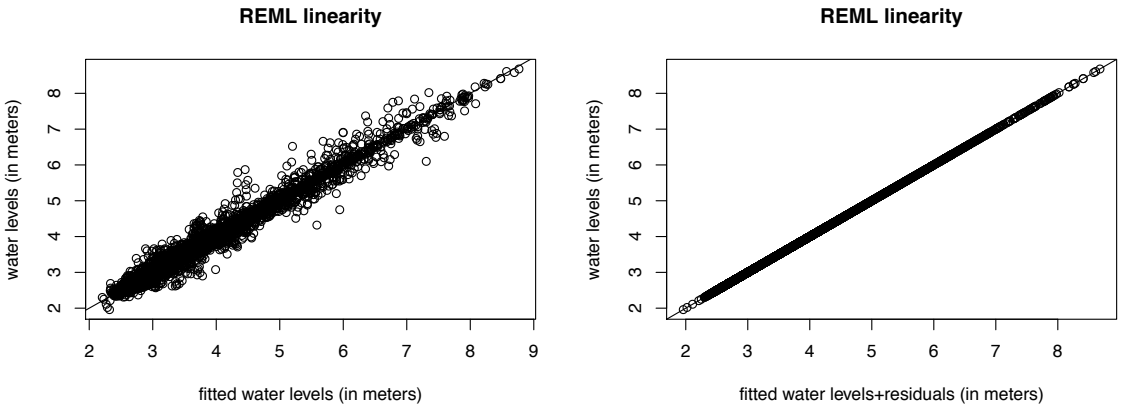


Figure 4.17: Observed vs Predicted values for REML model

Linearity on link scale for the REML model, where the link was the identity link, was assessed by a plot of observed weekly water levels against the predicted weekly water levels (linear predictor(η)) as shown in Figure 4.17. An ideal model has a one to one relationship between the observed and predicted weekly water levels (like the right plot of Figure 4.17), note: some of the GEE models as shown in Figure 5.5 did not show this perfect relationship when residuals was added to the fitted weekly water levels. Figure 4.17 (left) shows a plot of observed weekly water levels against fitted weekly water levels while the plot on right have REML model residuals added to the fitted values. The plot on the left shows that our REML model was very good as the predicted and observed weekly water levels plot follows a perfect linear trend. Thus, linearity assumption was not violated for this REML model.

4.5.2 Assessing for Mean-Variance Relationship

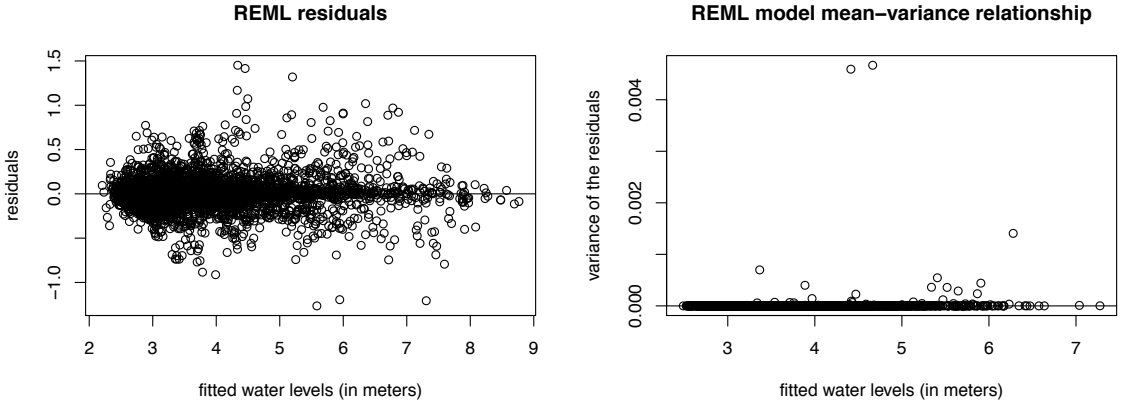


Figure 4.18: Mean-variance relationship for REML model

Figure 4.18 is a plot of REML residuals and variance of the residuals plotted against the fitted weekly water levels. A good GLMM has zero mean for its model residuals with its residuals values well scattered around zero and the variance of the residuals constant at zero. The mean-variance relationship indicates that the variance of the residuals is constant and very low. The variance given in Figure 4.18 is the variance of bins of data where the predicted weekly water levels was sub-divided into bins of 2000 equal data points of which the variance of the residuals was then calculated for each bin and plotted against the mean of predicted weekly water levels

for each of the 2000 bins (as described in subsection 4.4.2).

4.5.3 Assessing for Non-independence in Model Residuals

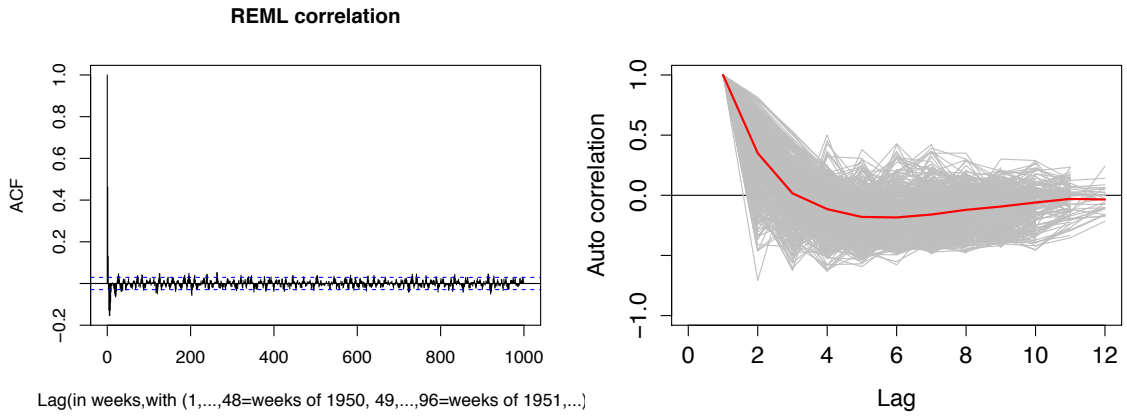


Figure 4.19: Non independence in model residuals for REML model

One of the last assumptions of GLMM is that of independence in the residuals. The ACF plot in Figure 4.19 (left plot) shows that there is no auto-correlation in the residuals as the correlations at different time lags is almost within the ACF significance bound, unlike the one given by the GEE models in Figure 4.13. The only correlation that seems to exist however is that within the a block (cluster) as shown on the left plot of 4.19 where the mean correlation (correlation between) clusters is about zero (solid line) while within clusters is given by the spaghetti plots. Thus, the REML error terms are correlated within a cluster but are independent between clusters.

4.5.4 Discussion

REML model fit in Table 4.7 was fitted in order to see if the commonly used non-likelihood GEE method for longitudinal data provides better estimates as compared to the likelihood method REML. From literatures it was mentioned that GEE method works better when there are a lot of random component (clusters) with fewer data points in each cluster. In this project however, we had the opposite where the random components were only two and the sample

within a random component was very large (about 80% of 2784 measurements at each location). For model convergence sake, new cluster variable was created as defined in subsection 3.2. The REML bias for weekly water levels variable $\widehat{y}$ was zero. For the GEE model the bias for $\widehat{y}$ was $-5.77e^{-15}$ for the Robust GEE and 0.0439 for the AR1 GEE model. The coefficient of determination (R^2) from the REML model was also better compared to the one from the Robust GEE nor AR1 GEE model. R^2 was 0.975 for REML as compared to $R^2 = 0.531$ and $R^2 = 0.547$ for Robust GEE and AR1 GEE models respectively. The error terms from the REML was also found to be smaller than those of Robust GEE model which implies much better estimation for the REML model.

Table 4.8: Table of estimated water levels by REML model

```
> print(Latex_Data_sample , row.names = F)
```

Year	Months	Seas	W.levels	W.flows	Loc	Loc_Time_Order	id	zTime	BS.1	BS.2	BS.3	BS.4	fitted(y)
1961	11	0	3.66	38.7	1	569	196114	1241	0.305	0.00650	4.47e-05	0.00e+00	3.68
1983	2	1	3.27	262.9	2	1589	198321	345	0.643	0.31773	3.97e-02	1.58e-05	3.35
1967	9	0	3.85	57.3	1	849	196714	1833	0.470	0.01837	2.25e-04	0.00e+00	3.86
1999	7	0	4.40	117.2	1	2377	199913	5025	0.761	0.08769	2.91e-03	0.00e+00	4.24
1981	11	0	3.51	28.6	1	1529	198114	3241	0.197	0.00248	1.02e-05	0.00e+00	3.54
1995	8	0	3.73	35.1	1	2192	199513	4632	0.268	0.00488	2.87e-05	0.00e+00	3.72
1968	4	1	6.25	409.7	1	879	196812	1915	0.414	0.44600	1.36e-01	3.92e-03	6.25
1979	9	0	3.36	211.8	2	1425	197924	3033	0.737	0.24202	2.01e-02	0.00e+00	3.30
1961	4	1	3.95	629.5	2	541	196122	1213	0.185	0.44315	3.27e-01	4.52e-02	4.07
1990	6	0	2.71	286.2	2	1943	199023	4123	0.602	0.34664	5.14e-02	9.82e-05	2.80
1989	9	0	3.80	67.2	1	1906	198914	4034	0.542	0.02682	4.11e-04	0.00e+00	3.83
1969	8	0	3.01	283.1	2	941	196923	2029	0.607	0.34306	4.98e-02	8.17e-05	2.94
2000	11	0	3.47	25.3	1	2441	200014	5141	0.159	0.00157	5.06e-06	0.00e+00	3.46
1991	11	0	2.91	133.5	2	2010	199124	4242	0.788	0.11227	4.48e-03	0.00e+00	2.98
1971	2	1	2.89	500.0	2	1016	197121	228	0.305	0.46856	2.12e-01	1.37e-02	2.69
1973	5	1	2.74	316.2	2	1124	197322	2420	0.552	0.37886	6.87e-02	3.92e-04	3.06
1989	7	0	3.12	239.4	2	1899	198923	4027	0.685	0.28497	2.96e-02	5.85e-08	3.12
1960	12	1	2.65	208.4	2	528	196021	1148	0.743	0.23649	1.91e-02	0.00e+00	2.71
1983	1	1	4.65	90.6	1	1587	198311	343	0.672	0.05192	1.20e-03	0.00e+00	4.49
1972	7	0	2.75	195.9	2	1084	197223	2328	0.763	0.21578	1.56e-02	0.00e+00	2.76
1970	8	0	3.59	242.5	2	989	197023	2129	0.680	0.28952	3.09e-02	2.99e-07	3.38

Note: Table of estimated water level values using parameters from the model and some observed informations. Seas is short for season, W.levels is short for water levels, W.flows is short for water flows, Loc is short for Location, Loc_Time_Order is short for time order within a location. BS.1, BS.2, BS.3 and BS.4 are the calculated smoothing B-splines of water flows.

Table 4.8 shows the estimated water values using the following equation;

$$\hat{\mathbf{y}}_{it} = \mathbf{model} + \epsilon_{it} \tag{4.5.1}$$

where in Equation (4.5.1), $\mathbf{model} = (\hat{\beta}_0 + \hat{\beta}_1 * \mathrm{BS.1}_{it} + \hat{\beta}_2 * \mathrm{BS.2}_{it} + \hat{\beta}_3 * \mathrm{BS.3}_{it} + \hat{\beta}_4 * \mathrm{BS.4}_{it} + \hat{\beta}_5 * \mathrm{T}_{it} + \hat{\beta}_6 * \mathrm{S}_{it} + \hat{\beta}_7 * (\mathrm{T} : \mathrm{S})_{it})$.

$\hat{y}_{it}$ is the value of estimated weekly water levels by REML method at a any given point in time while ϵ_{it} is the error term for the i^{th} predicted weekly water level at time t. The $BS.1_{it}, BS.2_{it}, BS.3_{it}, BS.4_{it}$ represents the continuous smoothing B-splines of weekly water flow covariate for observation i at time t, while the discrete covariates T_{it} represents zTime, S_{it} represents Season, $(T:S)_{it}$ represents interaction of zTime with Season. The linear combination of $(\eta = \hat{\beta}_0 + \hat{\beta}_1 * BS.1_{it} + \hat{\beta}_2 * BS.2_{it} + \hat{\beta}_3 * BS.3_{it} + \hat{\beta}_4 * BS.4_{it} + \hat{\beta}_5 * T_{it} + \hat{\beta}_6 * S_{it} + \hat{\beta}_7 * (T:S)_{it})$ might describe the relationship between water levels and the covariates well, but in practice it will never describe the water levels exactly. Thus, the difference between the observed water levels and linear predictor (η) was considered to be the model error (ϵ_{it}). Twisk (2003) state that each regression coefficients $(\hat{\beta}_1, ..., \hat{\beta}_7)$ for a particular predictor variable "relates the vector of weekly water levels over time to the vector of predictor variable over time". Twisk (2003) also mention that the $(\hat{\beta}_1, ..., \hat{\beta}_7)$ can not be interpreted straightforward as it combines the between cluster and within cluster relationship which then resulted in a single $(\hat{\beta}_1, ..., \hat{\beta}_7)$ regression coefficient. The intercept coefficient $\hat{\beta}_0 = 3.21$ gives an average estimate of weekly water levels when other regression coefficients $(\hat{\beta}_1, ..., \hat{\beta}_7)$ are assumed to be zero.

The ϵ for the GLMM fitted under REML was modelled to follow the Cauchy distribution, $\epsilon \sim$ Cauchy (σ, μ) with probability density function given as;

$$f(x) = f(x|\mu, \sigma) = \pi\sigma\Big(1 + \Big(\frac{x - \mu}{\sigma}\Big)^2\Big)^{-1} \tag{4.5.2}$$

where $\sigma > 0$ and μ are scale and location parameters respectively. The Cauchy distribution in Equation (4.5.2) is a reminiscent of normal distribution where the domain of $f(x) \in (-\infty, \infty)$. The PDF of the normal distribution is given by the squared difference from the mean while that of the Cauchy is expressed as the ratio of two independent standard normal variables. The distribution of ϵ here had $\sigma = 0.06006$ and $\mu = -0.00513$. An ideal Cauchy distribution has $\sigma = 1$ and $\mu = 0$.

The fitted weekly water levels $(\hat{y})$ were the values of estimated weekly water levels at any given point in time with ϵ_{it} being the error term for the i^{th} predicted weekly water level at time (t). Equation (4.5.1) indicates that in the absence of all random and fixed effects, the predicted weekly water level will be 3.210 otherwise it changes as a multiplicative factor of

model coefficients and the given predictor variables.

4.6 REML MODEL PLOTS

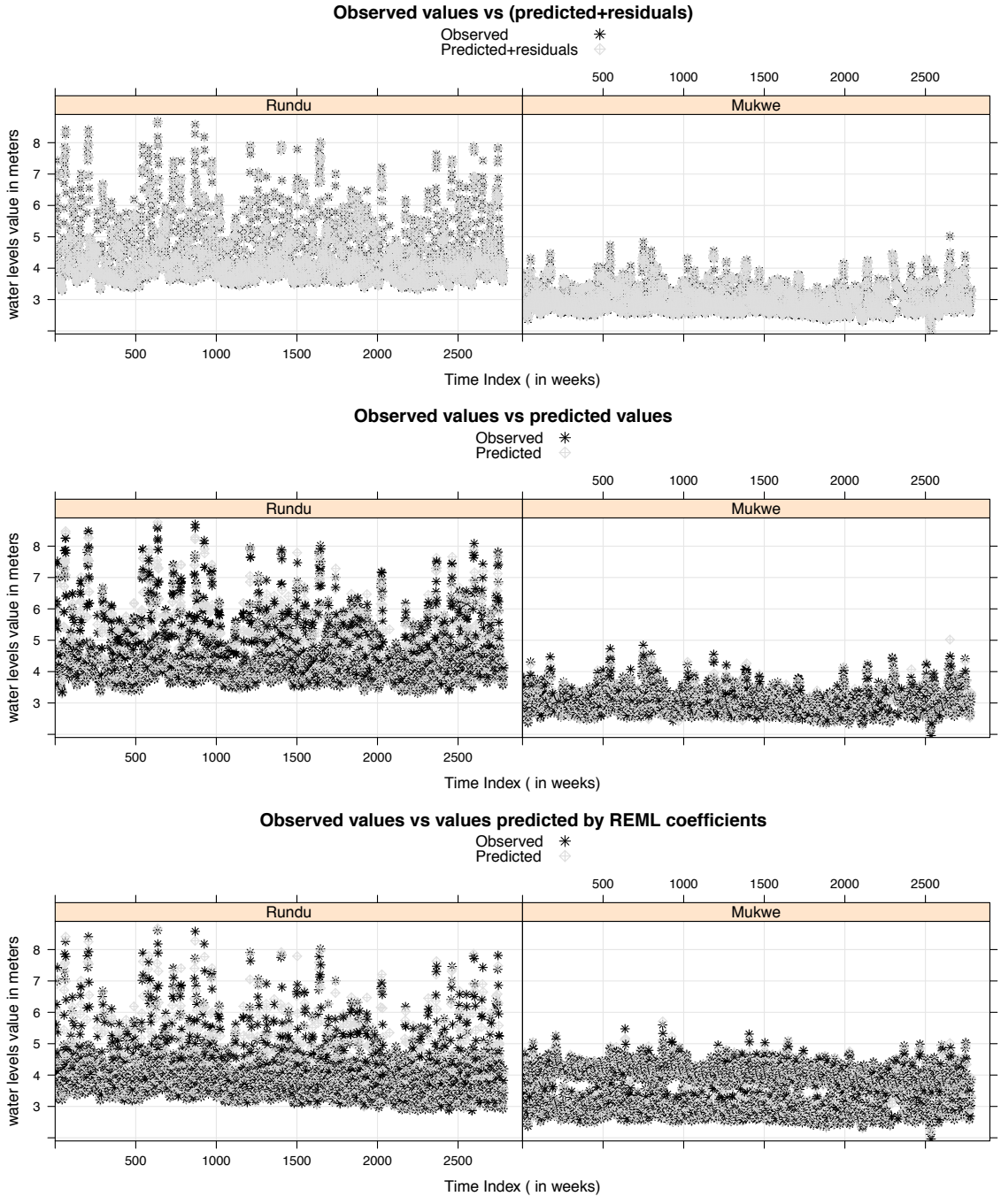


Figure 4.20: Longitudinal plots of predicted and observed values for REML model

Figure 4.20 shows a plot of observed weekly water level plotted together with their predicted weekly water level for Rundu and Mukwe where the fitted weekly water level was extracted from the REML model. The top plot has model residuals added to the model fitted weekly water level which makes the observed and (fitted weekly water level + REML model residuals) exactly the same, the middle plot is a plot of observed weekly water level and predicted weekly water level as produced by the REML model, while the bottom plot is also a plot of observed weekly water level with their predicted weekly water level where the predicted weekly water level was manually calculated using the $(\hat{\beta}_0, ..., \hat{\beta}_7)$ of the REML model and their corresponding predictor variables.

The partial autocorrelation of the residuals for the REML model is given in Figure 4.21 which shows no significant partial autocorrelation in the residuals. Thus, the is no dependency of residuals over time when this model is used.

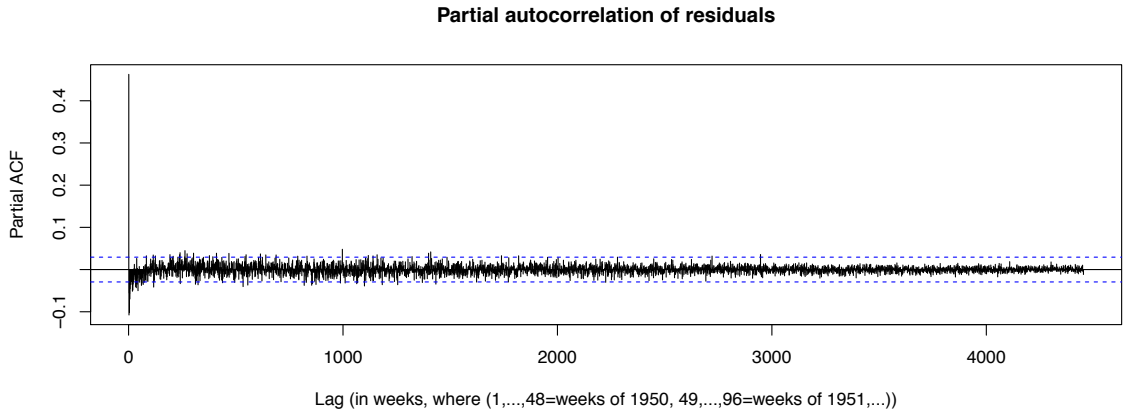


Figure 4.21: Partial autocorrelations of REML model residuals

Based on the R^2, MSE and the bias of the better GEE model (Robust GEE) and the REML, the REML model out performed the GEE model. In addition to this, the REML model was also found to have small standard errors as compared to any of the GEE models given in Table 4.5. The significance of the $(\hat{\beta}_0, ..., \hat{\beta}_7)$ model coefficients are summaries in Table 4.9, where the REML again seem to have better significant coefficients at both 10%, 5% and 1% level of significance, followed by the AR1 GEE model. Testing for the significance of $(\hat{\beta}_0, ..., \hat{\beta}_7)$ coefficients was very important, especially in the next section where simulation of weekly water levels was performed based on model coefficients.

Table 4.9: Significance of model coefficient parameters

	Significance of REML model parameters as compare to GEE models											
	Significance at 10 %				Significance at 5 %				Significance at 1 %			
Model parameters	REML	Robust GEE	AR1 GEE	EXCH GEE	REML	Robust GEE	AR1 GEE	EXCH GEE	REML	Robust GEE	AR1 GEE	EXCH GEE
(Intercept)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
BS.1(flows)	✓	×	✓	✓	✓	×	✓	✓	✓	×	✓	✓
BS.2(flows)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
BS.3(flows)	✓	✓	✓	✓	✓	×	✓	✓	✓	×	✓	✓
BS.4(flows)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
zTime(T)	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	×	×
Seasons(S)	✓	✓	✓	✓	×	✓	×	✓	×	✓	×	×
T : S	✓	×	✓	×	✓	×	✓	×	×	×	✓	×

Note: ✓ implies that the parameter was significant while × implies that the parameter was not significant at the given level of signifcance.

4.7 SIMULATION AND MODELLING

R software uses the Mersenne Twister algorithm for pseudorandom number generator (PRNG) (Venables and Smith, 2003). The Mersenne Twister algorithm is reported to have passed many statistical tests of randomness including the Diehard tests, commonly used to measure the quality of sets of random numbers. Most statistical software for simulations, e.g. Easy-fit software, R , Python, Matlab, SAS, SPSS among others, uses the Mersenne Twister algorithms as a default for random number generations.

4.7.1 Pseudo-Monte Carlo Simulation

Monte Carlo (MC) simulation consists of statistical techniques that can be used to give approximate solutions to quantitative analytical problems. This may include the formulation of imitated real data that can then be used to model real scenario when there is absence of data.

Correlation is expected in any longitudinal data, which includes hydrological data. When simulating data for a model this should be one of the main factor that needs to be addressed. The simulation conducted in this research was a pseudo MC simulation of weekly water levels

and weekly water flows. In order to simulate random values with a specific correlation, the simulation of weekly water flow was done using the distribution of weekly water flow for each month, based on the goodness of fit of the Kolmogorov-Smirnov test and the Anderson-Darling test. On the other hand, simulation of weekly water level was done using $(\hat{\beta}_0, ..., \hat{\beta}_7)$ coefficients from the REML model together with the simulated weekly water flows and other respective observed predictor variables like season and time. Different distributions were used to simulate weekly water flow data as there was no known overall distribution of weekly water flows per season, location or whole combined data. Weekly water flows had unique distributions for each month however, e.g. month of May, June and July weekly water flows followed the Burr and Gamma distributions (see Table 3.2). The goodness of fit was tested at 20%, 10%, 5%, 2% and 1% level of significance and the distribution that ranked low on the Kolmogorov-Smirnov test and the Anderson-Darling test was selected as the desired distribution for that month. The shape, scale and location parameters together with the desired distributions of weekly water flows are given in Table 3.2.

Based on Burton, Altman, Royston, and Holder (2006), average number of simulations to be carried out in a simulation study varies from 100 to 100 000 for most simulation studies, with 1 000 and 10 000 replication being more common choice as compare to the others. Díaz-Emparanza (2002) stated that "when probability distribution is approximated by means of simulation it is obvious that the bigger the number of replications is, the better the approach will be". Thus accuracy and reliability depends on numbers of simulations where, higher numbers of simulations produce higher accuracy and very reliable results. Díaz-Emparanza (2002) also stated that a small number of replications does not produce high levels of precisions. Due to the availability of resources, simulations were repeated 10 000 times, where the sample size range from 100%, 62%, 25% and 18% of train data set. This was done to analyse the effect of sample size on consistency of the model coefficients as the sample size changes.

Table 4.10: REML parameter estimates for simulated model

REML model estimates at different sample size: *dependent variable is* y= *Water levels*												
	Simulated sample size (n)											
	n = 4454 ≡ 100% of trainset			n = 2761 ≡ 62% of trainset			n = 1113 ≡ 25% of trainset			n = 801 ≡18% of trainset		
Model parameters	coef.	Bias	MSE	coef.	Bias	MSE	coef.	Bias	MSE	coef.	Bias	MSE
Intercept	3.214922***	-2.670e^{-05}	1.116e^{-09}	3.309821***	9.487e^{-02}	9.001e^{-03}	3.702527 ***	4.876e^{-01}	2.377^{-01}	3.789141 ***	5.742e^{-01}	3.297e^{-01}
BS.1(flows)	6.856087e^{-01}***	1.348e^{-08}	1.272e^{-14}	6.055746e^{-01}***	-8.003e^{-02}	6.405e^{-03}	1.451677e^{-01}***	-5.404e^{-01}	2.921^{-01}	6.248482e^{-02}***	-6.231e^{-01}	3.883e^{-01}
BS.2(flows)	1.795368***	7.297e^{-08}	2.434e^{-13}	1.632704***	-1.627e^{-01}	2.646e^{-02}	9.986366e^{-01}***	-7.967e^{-01}	6.348^{-01}	4.579832e^{-01}***	-1.337384	1.788597
BS.3(flows)	1.899535***	-7.539e^{-07}	3.040e^{-10}	2.016553***	1.170e^{-01}	1.369e^{-02}	1.733646***	-1.659e^{-01}	2.752^{-02}	2.158113 ***	2.586e^{-01}	6.686e^{-02}
BS.4(flows)	2.688091***	-6.095e^{-07}	1.568e^{-11}	2.62725***	-6.084e^{-02}	3.702e^{-03}	2.506595***	-1.815e^{-01}	3.294^{-02}	2.562859***	-1.252e^{-01}	1.568e^{-02}
zTime(T)	-7.907611e^{-05}***	1.096e^{-08}	1.667e^{-16}	-8.248942e^{-05}***	-3.402e^{-06}	1.158e^{-11}	-7.663308e^{-05}***	2.454e^{-06}	6.022^{-12}	-6.993868e^{-05}***	9.148e^{-06}	8.369e^{-11}
Season(S)	-1.908033e^{-01}***	1.276e^{-04}	1.663e^{-08}	-1.766528e^{-01}***	1.428e^{-02}	2.039e^{-04}	-1.263519e^{-01}***	6.458e^{-02}	4.170^{-03}	-3.166716e^{-02}***	1.593e^{-01}	2.536e^{-02}
T : S	-1.096110e^{-08}***	-7.278e^{-05}	5.296e^{-09}	-2.723795e^{-09}***	-7.276e^{-05}	5.295e^{-09}	-5.90681e^{-12}***	-7.276e^{-05}	5.294e^{-09}	-4.063338e^{-12}***	-7.276e^{-05}	5.294e^{-09}
$\hat{y}$ bias		1.009e^{-16}			4.188e^{-17}			-3.242e^{-17}			2.727e^{-17}	
R^2	1	0.025		1	0.025		1	0.025		1	0.025	
MSE	6.0004e^{-13}			1.1187e^{-14}			1.7986e^{-15}			7.3857e^{-16}		
RMSE	7.723e^{-07}			7.040e^{-08}			2.040e^{-08}			1.369e^{-08}		
AIC	-107 099.8			-72 232.49			-27 213.05			-19 298.07		
BIC	-107 055.9			-72 161.41			-27 152.87			-19 241.84		

Signif.code *p<0.1; **p<0.05; ***p<0.01;

Note: BS(flows)= coefficients of the matrix of smoothing basis-splines of simulated water flows. The bias column show the difference in simulated parameters to that obtained by the observed data REML model (e.g, (real REML constant) - (simulated REML constant)).

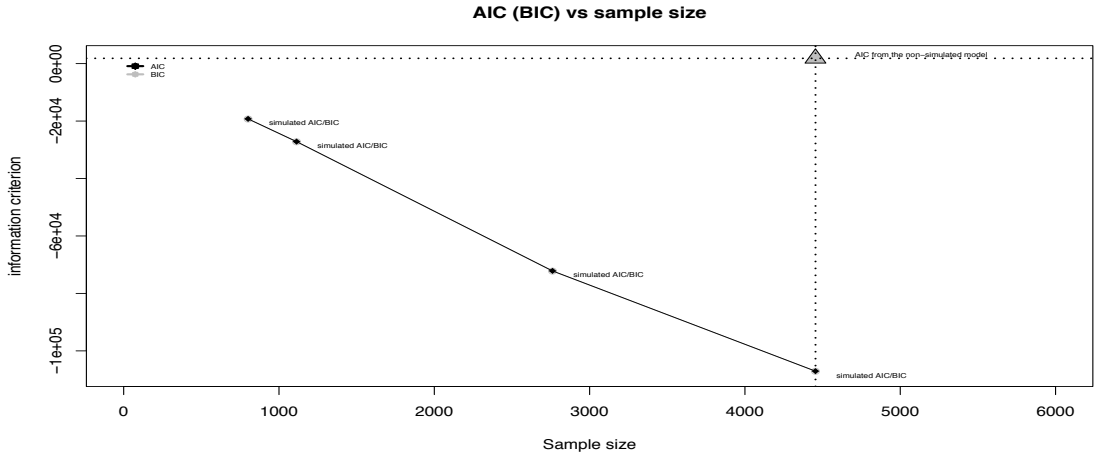


Figure 4.22: Relationship between AIC and sample size

Table 4.10 shows the mean parameter constants of the simulated REML model at different sample sizes. The MSE, RMSE, bias, R^2 and AIC are also summarised in Table 4.10. The RMSE, MSE and bias for the ($\hat{y}$) from the REML simulated model was very small, significantly not different from zero, while the R^2 was equal to one, similar to that of an ideal model. The parameters given in Table 4.10 shows that the simulated model was better as compared to the observed data model as given in Table 4.7; as it has a smaller value of MSE and R^2=1. The

AIC ranges from AIC=-107 099.8 when the sample size was 4445 to AIC= -19 298.07 when the sample size was 801. Based of Figure 4.22, when sample size was less than 4 445, the relationship between AIC and sample sizes was linear and negative, a similar pattern was also observed for the BIC. When sample size was 4 445, the AIC obtained by simulated model was 58 times smaller than the one obtained by the observed data model which was equal to AIC=1 837 (represented by the grey triangle on upper right corner of Figure 4.22). Unlike for the observed data REML model, the simulated data produced MSE for $\hat{\mathsf{y}}$ and R^2 values that seem to be constant and not necessarily depending on the sample size as compared to those given in Figure 4.16. The bias and MSE of estimated coefficients $(\hat{\beta}_0, ..., \hat{\beta}_7)$ was also very small when sample size was 4 445. However, when sample size was decreased, the bias and MSE was observed to be increasing. When the sample size was large, as compared to the observed weekly data, the simulated model in Table 4.10 was almost the same as the model in Table 4.7. However, the simulated model in Table 4.10 was better in terms of accuracy, bias and information criterion selection (AIC) as compared to that in Table 4.7. The simulated model equation was the same as the one in Equation (4.5.1).

In Equation (4.5.1), outcome $(\hat{\mathbf{y}})$ was the values of the estimated simulated weekly water levels by REML method while ϵ was the model error vector. The $\mathsf{BS.1}_{\mathsf{it}}, \mathsf{BS.2}_{\mathsf{it}}, \mathsf{BS.3}_{\mathsf{it}}, \mathsf{BS.4}_{\mathsf{it}}$ represents the continuous smoothing B-splines of the simulated weekly water flow covariate for observation i at time t, while the discrete covariates $z\mathsf{Time}_{\mathsf{it}}$ represents Time, S_{it} represents Season, $(\mathsf{T:S})_{\mathsf{it}}$ represents interaction of zTime with Season at which the distribution of the simulated weekly water flows and weekly water levels was taken. Like with the observed data REML model, the linear combination of $(\hat{\beta}_0 + \hat{\beta}_1 * \mathsf{BS.1}_{\mathsf{it}} + \hat{\beta}_2 * \mathsf{BS.2}_{\mathsf{it}} + \hat{\beta}_3 * \mathsf{BS.3}_{\mathsf{it}} + \hat{\beta}_4 * \mathsf{BS.4}_{\mathsf{it}} + \hat{\beta}_5 * z\mathsf{Time}_{\mathsf{it}} + \hat{\beta}_7 * \mathsf{S}_{\mathsf{it}} + \hat{\beta}_8 * (\mathsf{T:S})_{\mathsf{it}})$ might describe the relationship between simulated weekly water levels and the covariates well, but in practice it might never describe the simulated weekly water levels exactly at all the time. Thus the error (ϵ) component for the simulated REML model was modelled using the Cauchy distribution, $\epsilon \sim \text{Cauchy}$ $(\sigma = 4.1584e^{-8}, \mu = -1.2757e^{-9})$. The PDF of a Cauchy distribution is given in Equation (4.5.2). The simulated model had residuals which was not significantly different from zero, which was an indication that the model predict simulated weekly water levels almost accurately (as shown in Table 4.11 below).

Table 4.11: Table of estimated water levels by REML model

```
> print(Latex_Data_sample , row.names = F)
```

Year	Months	Seas	Loc	Loc_Time_Order	id	zTime	Sim.flows	Sim.levels	BS.1	BS.2	BS.3	BS.4	fitted(y)
1961	11	0	1	569	196114	1241	23.5	3.27	0.221	0.00226	7.50e-06	0.00e+00	3.27
1983	2	1	2	1589	198321	345	239.2	3.97	0.739	0.24138	1.99e-02	4.41e-06	3.97
1967	9	0	1	849	196714	1833	81.6	3.62	0.699	0.03958	6.66e-04	0.00e+00	3.62
1999	7	0	1	2377	199913	5025	63.8	3.27	0.595	0.02394	2.95e-04	0.00e+00	3.27
1981	11	0	1	1529	198114	3241	32.9	3.19	0.327	0.00538	2.84e-05	0.00e+00	3.19
1995	8	0	1	2192	199513	4632	105.3	3.51	0.789	0.06475	1.52e-03	0.00e+00	3.51
1968	4	1	1	879	196812	1915	610.2	4.35	0.317	0.46106	2.05e-01	1.72e-02	4.35
1979	9	0	2	1425	197924	3033	122.0	3.70	0.825	0.08493	2.43e-03	0.00e+00	3.70
1961	4	1	2	541	196122	1213	689.4	4.49	0.256	0.45904	2.56e-01	2.97e-02	4.49
1990	6	0	2	1943	199023	4123	160.8	3.72	0.841	0.13657	5.80e-03	0.00e+00	3.72
1989	9	0	1	1906	198914	4034	38.3	3.17	0.382	0.00772	4.96e-05	0.00e+00	3.17
1969	8	0	2	941	196923	2029	195.4	3.96	0.804	0.18452	1.07e-02	0.00e+00	3.96
2000	11	0	1	2441	200014	5141	31.2	3.03	0.309	0.00472	2.32e-05	0.00e+00	3.03
1991	11	0	2	2010	199124	4242	260.7	3.89	0.708	0.26649	2.58e-02	2.79e-05	3.89
1971	2	1	2	1016	197121	228	373.7	4.19	0.558	0.37142	6.90e-02	1.12e-03	4.19
1973	5	1	2	1124	197322	2420	490.8	4.17	0.428	0.43560	1.31e-01	5.86e-03	4.17
1989	7	0	2	1899	198923	4027	142.0	3.68	0.843	0.11103	3.93e-03	0.00e+00	3.68
1960	12	1	2	528	196021	1148	262.2	3.95	0.706	0.26822	2.62e-02	3.08e-05	3.95
1983	1	1	1	1587	198311	343	181.2	3.87	0.823	0.16499	8.44e-03	0.00e+00	3.87
1972	7	0	2	1084	197223	2328	336.2	4.16	0.605	0.34160	5.26e-02	5.04e-04	4.16
1970	8	0	2	989	197023	2129	183.0	3.93	0.821	0.16742	8.69e-03	0.00e+00	3.93

Note: Table of estimated simulated water level values using parameters from the model and some observed informations. Seas is short for season, Sim.levels is short for simulated water levels, Sim.flows is short for simulated water flows, Loc is short for Location, Loc_Time_Order is short for time order within a location. BS.1, BS.2, BS.3 and BS.4 are the calculated smoothing B-splines of simulated water flows.

4.7.2 Test Set Estimation

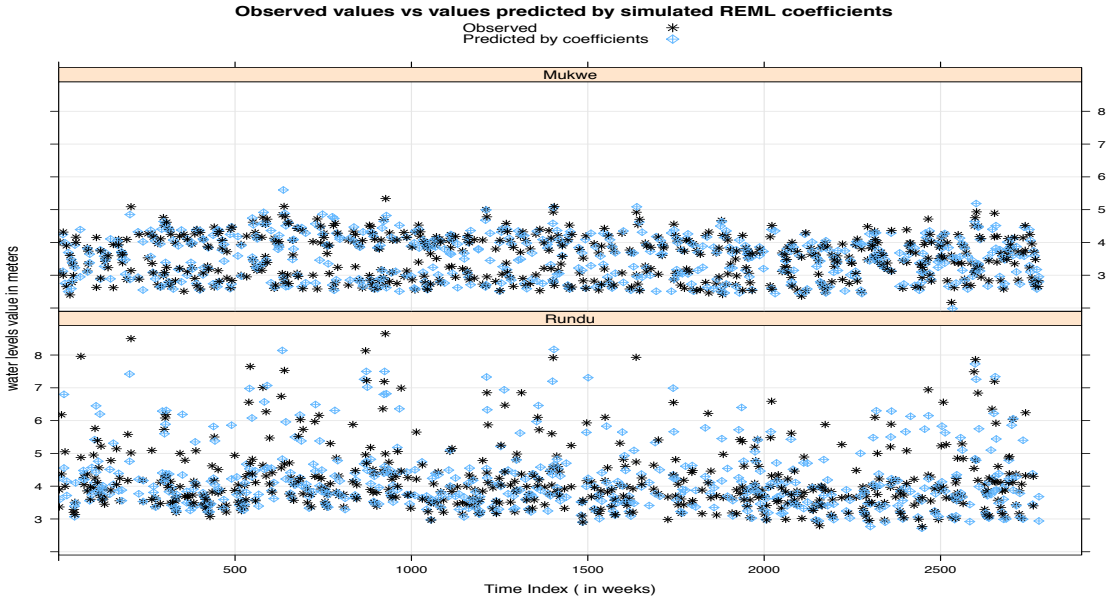


Figure 4.23: Plot of observed testdata weekly water levels vs predicted weekly water levels from the simulated model

In order to see how well the simulated model performs, it was tested on the test data which was 20% hold sample of the observed data set, Figure 4.23 shows the plots of observed weekly water levels plotted together with predicted weekly water levels, calculated using the simulated REML model coefficients $(\hat{\beta}_0, ..., \hat{\beta}_7)$ together with their respective predictor variables from the test set data. The Wilcoxon Rank-Sum test gave a p-value = $6.27e^{-12}$ which indicates that there was evidence at 5% level of significance that the median of the observed weekly water levels and the weekly water levels predicted by the simulated model coefficients differ in median (see Table 4.12 as well as histograms in Figure 4.24). However, the mean weekly water level of the test set data and mean weekly water level predicted by simulated model coefficients was the same with mean=3.86. The MSE=1.46 and RMSE=1.21 was also observed for comparisons.

Table 4.12: Descriptive statistics of the distribution of testset weekly water levels and predicted weekly water levels

statistics	Observed weekly water level	Predicted weekly water level
Mean (in meters)	3.85977	3.860403
Median (in meters)	3.61	3.868833
Std. dev	1.169929	0.4662443
Number of observations	1 114	1 114

Note: The Wilcoxon rank sum test for the two variable gave W = 516140 and p-value = $6.27e^{-12}$ with 95 percent confidence interval for the difference in median = (-0.2986773, -0.1714502)

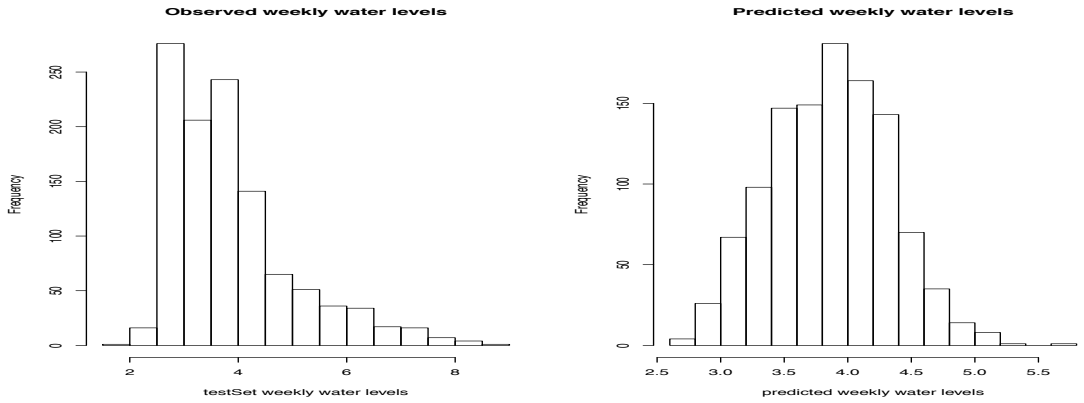


Figure 4.24: Histogram plot for observed water levels value vs water levels value predicted using the simulated model parameters

Table 4.25 shows parameter estimates of the simulated REML model where the weekly water flow values was simulated using the burr distribution of weekly water flow at Rundu and Mukwe locations as oppose to the monthly distributions at Rundu and Mukwe locations as given in Table 3.2. Based on the Bias and MSE of model coefficients in Table 4.10 and Table 4.13, the is no significant difference between this two model coefficients. This could also be observed in Figure 4.25 as compared to Figure 4.23. However, the AIC and BIC are a litle better in model presented in Table 4.10. Also the Kolmogorov-Smirnov test and the Anderson-Darling test did not show that the distribution of weekly water flows at Rundu or Mukwe Location followed the Burr distribution, which was used to generate weekly water flows at the two locations. The two test could also not further confirm which distribution best fit our weekly water flow as all the tested distributions rejected the null hypothesis that the distribution of weekly water flows at Rundu or Mukwe location came from the tested distribution. Hence, model in Table 4.10 was used as the simulated model in this research. The mean and median of 3.868812 and 3.860395 respectilelyfor simulated weekly water flow was observed for model in Table 4.13.

Table 4.13: REML parameter estimates for simulated model where water flows was simulated using Locations distributions instead of Months distributions

		Bias and MSE with respect to model in Table 4.7		Bias and MSE with respect to model in Table 4.10	
Model parameters	coef.	Bias	MSE	Bias	MSE
Intercept	3.214945^{***}	$-3.97377e^{-06}$	$2.072679e^{-10}$	$-2.272313e^{-05}$	$1.115627e^{-09}$
BS.1(flows)	$6.856089e^{-01***}$	$1.909233e^{-07}$	$7.874775e^{-14}$	$-1.774394e^{-07}$	$8.691082e^{-14}$
BS.2(flows)	1.795367^{***}	$-2.105008e^{-08}$	$5.543194e^{-12}$	$9.4018e^{-08}$	$5.775184e^{-12}$
BS.3(flows)	1.899537^{***}	$1.608719e^{-06}$	$1.517974e^{-07}$	$-2.362598e^{-06}$	$1.520979e^{-07}$
BS.4(flows)	2.688093^{***}	$1.104973e^{-06}$	$4.214961e^{-10}$	$-1.714469e^{-06}$	$4.407738e^{-10}$
zTime(T)	$-7.908448e^{-05***}$	$2.580513e^{-09}$	$2.925741e^{-17}$	$8.377323e^{-09}$	$1.39877e^{-16}$
Season(S)	$-1.908224e^{-01***}$	$1.085506e^{-04}$	$1.195495e^{-08}$	$1.906449e^{-05}$	$8.833324e^{-10}$
T:S	$-2.580319e^{-09***}$	$-7.276369e^{-05}$	$5.294555e^{-09}$	$-8.380785e^{-09}$	$1.399725e^{-16}$
$\hat{y}$ bias	$-9.285905e^{-17}$				
R^2	1				
MSE	$6.977808e^{-13}$				
RMSE	$8.289360e^{-07}$				
AIC	-104 490.6				
BIC	-104 413.8				

Signif.code *p<0.1; **p<0.05; ***p<0.01;

Note: BS(flows)= coefficients of the matrix of smoothing basis-splines of simulated water flows. The bias column show the difference in simulated parameters to that obtained by the observed data REML model (e.g, (real REML constant) - (simulated REML constant)).

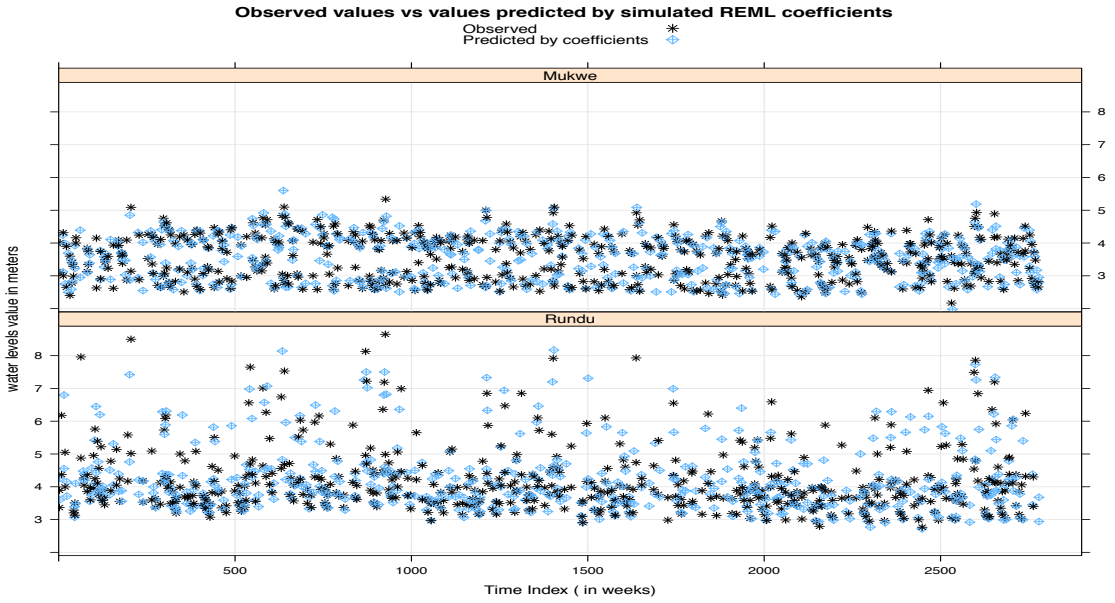


Figure 4.25: Plot of observed testdata weekly water levels vs predicted weekly water levels from the simulated model, where simulated weekly water flows was done using location distribution as appose to monthy distributions as given in Table 3.2

Chapter 5

CONCLUSION AND RECOMMENDATION

5.1 CONCLUSION

The main aim of this research was to develop a statistical model that best describe the levels of weekly water levels passing through the Okavango river every year. Two methods of parameter estimation, the GEE and the REML, were used to estimate parameters of the GLMM. The parameters of the best method were then used in the simulated model development.

The GEE method based its estimation on pre-specified covariance matrices while the REML method worked by giving more weight to the random effects and the error terms. The model fitted was a generalised linear mixed effect model. The random effect, location at which the data were observed, and the fixed effects were the covariates used in the model (e.g, water flow, time and season). Five GEE models were analysed the empirical GEE as well as the GEE with the Unstructured, Exchangeable, Independence and First-order autoregressive (AR1) correlation matrices to estimate variability. The empirical GEE were found to have lower QIC and MSE as sample size changes, but had higher standard errors as compared to the AR1 GEE model. The R^2 for the AR1 and Exchangeable GEE models were also higher in relation to the sample size as compared to the other GEE models. Even though the R^2, MSE and QIC of the empirical GEE

and AR1 GEE were better compared to the other GEE models, the REML produced a much better model in terms of R^2 and MSE. The standard errors obtained under the REML method of estimation were also small compared to those obtained under the GEE method of estimation. In addition, REML model produced better fitted weekly water level values as compared to those of GEE models, see Figure 4.17 vs. Figure 4.11 where the linearity assumption were violated by the GEE models while under REML model linearity assumption was not violated. Furthermore, the REML model had zero bias on fitted observed weekly water levels values which then resulted in a better mean-variance relationship and small residuals values as compared to the GEE models.

The Monte Carlo simulations for the best performing model (REML model) were done using different sample sizes for the aim of analysing the effect of AIC, MSE, R^2 and bias. It was observed that the R^2=1, bias=0 and MSE=0 remained relatively the same for different sample sizes, while the AIC were linearly decreasing as the sample size increases (see Figure 4.22), which implies that simulated model becomes stronger as sample size gets bigger. One of the main reasons of carrying out simulations was to help come up with a model that works better when estimating water levels as compared to the pre-established REML model in Table 4.16. From simulation results, it was observed that the simulated model performed better due to its higher R^2, and lower MSE and bias. It was also noted that simulated models had small value of AIC which were inversely linearly related to the sample size.

5.2 RECOMMENDATION

The developed model was only tested on data set from the Okavango river. It would have been very desirable to have data from other rivers tested on this model to see its effectiveness. It would have also been desirable if multiple variables like amount of rainfall in catchment areas were captured and included in the developed model. The use of spatial variables would be recommended for future research.

It was noted that when REML and GEE models were fitted to the observed field data, the AIC/QIC and MSE were increasing as sample size was increased. However, when these models

were fitted to the simulated data, AIC/QIC were decreasing as sample size was increasing and MSE were zero. It could not be established as to what caused this behaviour in AIC/QIC as sample size changes, but further investigations for future research on similar datasets is highly recommended

It was also understood that the Ministry of Agriculture Water and Forestry in Namibia had a method of dealing with missing data and some of the data were imputed for missing values while some were left as observed from the field. It was not clearly explained which statistical methods were used to impute data. In this research however, missing data was imputed using the k-nearest neighbour for longitudinal data. Thus recommendation were also given to the Ministry of Agriculture Water and Forestry in Namibia to keep a file of field data as obtained from field, in order to allow researchers to use their own techniques of imputation.

References

Arango, J. A., Cundiff, L. V., and Van Vleck, L. D. (2004). Covariance functions and random regression models for cow weight in beef cattle. *Journal of Animal Science*, 82(1):54–67.

Atiya, A. F., El-Shoura, S. M., Shaheen, S. I., and El-Sherif, M. S. (1999). A comparison between neural-network forecasting techniques-case study: river flow forecasting. *Neural Networks, IEEE Transactions on*, 10(2):402–409.

Botswana (2008). Okavango delta management plan (odmp). Technical report, Government of the Republic of Botswana (Ministry of Environment, Wildlife and Tourism).

Boughton, W. and Droop, O. (2003). Continuous simulation for design flood estimation a review. *Environmental Modelling & Software*, 18(4):309–318.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine*, 25(24):4279–4292.

Chetty, K. and Smithers, J. (2005). Continuous simulation modelling for design flood estimation in south africa: Preliminary investigations in the thukela catchment. *Physics and Chemistry of the Earth*, 30(11):634–638.

Crowder, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika*, 82(2):407–410.

Crowther, P. S. and Cox, R. J. (2005). *Knowledge-based intelligent information and engineering systems*, chapter A method for optimal division of data sets for use in neural networks, pages 1–7.

Cui, J. and Qian, G. (2007). Selection of working correlation structure and best model in gee analyses of longitudinal data. *Communications in Statistics Simulation and Computation*, 36(5):987–996.

Dahmen, G. and Ziegler, A. (2004). Generalized estimating equations in controlled clinical trials: hypotheses testing. *Biometrical Journal*, 46(2):214–232.

Díaz-Emparanza, I. (2002). Is a small monte carlo analysis a good analysis? *Statistical Papers*, 43(4):567–577.

Engels, J. M. and Diehr, P. (2003). Imputation of missing longitudinal data: a comparison of methods. *Journal of clinical epidemiology*, 56(10):968–976.

Hand, D. J. and Crowder, M. J. (1996). *Practical longitudinal data analysis*, volume 34. CRC Press.

Hanley, J. A., Negassa, A., Edwardes, M. D., and Forrester, J. E. (2003). Statistical analysis of correlated data using generalized estimating equations: an orientation. *American journal of epidemiology*, 157(4):364–375.

Hardin, J. W. and Hilbe, J. M. (2003). *Generalized estimating equations*. USA: Chapman & Hall.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal data analysis*, volume 451. John Wiley & Sons.

Højsgaard, S., Halekoh, U., and Yan, J. (2014). Geepack: Generalized estimating equation package.

Israëls, A., Kuyvenhoven, L., van der Laan, J., Pannekoek, J., and Nordholt, E. S. (2011). *Statistical methods*, chapter Imputation. Statistics netherlands.

Johnson, D. L. and Thompson, R. (1995). Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *Journal of dairy science*, 78(2):449–456.

Jonsson, P. and Wohlin, C. (2004). An evaluation of k-nearest neighbour imputation using likert data. *Proceedings of the Symposium*, pages 108–118.

Kachman, S. D. (2000). An introduction to generalized linear mixed models. In *Proceedings of a symposium at the organizational meeting for a NCR coordinating committee on Implementation Strategies for National Beef Cattle Evaluation, Athens*, pages 59–73.

Kenward, M. G., Lesaffre, E., and Molenberghs, G. (1994). An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics*, pages 945–953.

Khu, S. T. and Werner, M. G. F. (2003). Reduction of monte-carlo simulation runs for uncertainty estimation in hydrological modelling. *Hydrology and Earth System Sciences*, 7(5):680–692.

Leith, N. and Chandler, R. (2005). Using generalised linear models to simulate daily rainfall under scenarios of climate change. Technical Report FD2113_rpt2.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Mackenzie, M. and Scott-Hayward, L. (2015). *Spatial Modelling Methods for Correlated Data (Centre for Research into Ecological and Environmental)*.

Mbaiwa, J. E. (2004). Causes and possible solutions to water resource conflicts in the okavango river basin: The case of angola, namibia and botswana. *Physics and Chemistry of the Earth, Parts A/B/C*, 29(15):1319–1326.

Meyer, K. (1985). Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices. *Biometrics*, pages 153–165.

Meyer, K. (1991). Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. *Genet. Sel. Evol*, 23(1):67–83.

Meyer, K. and Hill, W. G. (1997). *Livestock Production Science*, 47(3):185–200.

Millar, R. B. (2011). *Maximum likelihood estimation and inference: with examples in R, SAS and ADMB*, volume 111. John Wiley & Sons.

Milzow, C., Kgotlhang, L., Bauer-Gottwein, P., Meier, P., and Kinzelbach, W. (2009). *Hydrogeology Journal*, 17(6):1297–1328.

Mullan, K., Daraganova, G., and Baker, K. (2015). Imputing income in the longitudinal study of australian children.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100.

Namibia (2011). Namibia 2011 population and housing census main report. Technical report, Namibia Statistics Agency.

Oehlert, G. W. (2014). *A few words about REML*.

Oh, S., Carriere, K. C., and Park, T. (2008). Model diagnostic plots for repeated measures data using the generalized estimating equations approach. *Computational Statistics & Data Analysis*, 53(1):222–232.

Olori, V. E., Hill, W. G., McGuirk, B. J., and Brotherstone, S. (1999). Estimating variance components for test day milk records by restricted maximum likelihood with a random regression animal model. *Livestock Production Science*, 61(1):53–63.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1):120–125.

Pan, W. and Connett, J. E. (2002). Selecting the working correlation structure in generalized estimating equations with application to the lung health study. *Statistica Sinica*, 12(2):475–490.

Park, T. and Lee, S. Y. (2004). Model diagnostic plots for repeated measures data. *Biometrical journal*, 46(4):441–452.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554.

Pinheiro, J., Bates, D., DebRoy, S., and Sarkar, D. (2014). R core team (2014) nlme: linear and nonlinear mixed effects models. r package version 3.1-117.

Porto, J. G. and Clover, J. (2003). The peace dividend in angola: Strategic implications for okavango basin cooperation.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Razali, N. M. and Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33.

Schittkowski, K. (2002). Easy-fit: a software system for data fitting in dynamical systems. *Structural and Multidisciplinary Optimization*, 23(2):153–169.

Ševčíková, H. (2004). Statistical simulations on parallel computers. *Journal of Computational and Graphical Statistics*, 13(4):886–906.

Shults, J., Sun, W., Tu, X., Kim, H., Amsterdam, J., Hilbe, J. M., and Ten-Have, T. (2009). A comparison of several approaches for choosing between working correlation structures in generalized estimating equation analysis of longitudinal binary data. *Statistics in medicine*, 28(18):2338–2355.

Smithers, J. C. (2012). Methods for design flood estimation in south africa. *Water SA*, 38(4):633–646.

Smithers, J. C., Chetty, K. T., Frezghi, M. S., Knoesen, D. M., and Tewolde, M. H. (2013). Development and assessment of a daily time-step continuous simulation modelling approach for design flood estimation at ungauged locations: Acru model and thukela catchment case study. *Water SA*, 39(4):467–475.

Tiwari, P. and Shukla, G. (2011). Approach of linear mixed model in longitudinal data analysis using sas. *Reliability Statistical Studies*, 4:73–84.

Twisk, J. W. R. (2003). *Applied longitudinal data analysis for epidemiology: a practical guide*. Cambridge University Press.

Venables, W. N. and Smith, D. M. (2003). "an introduction to r (version 1.8.1)," r foundation for statistical computing.

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer Science & Business Media.

Wang, Y. G. and Carey, V. (2003). Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. *Biometrika*, 90(1):29–41.

Wassertheil-Smoller, S. (2004). *Biostatistics and epidemiology: a primer for health and biomedical professionals*, volume 1. Springer Science & Business Media.

Weiss, R. E. (2005). *Modeling Longitudinal Data: With 72 Figures*. Springer Science & Business Media.

Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130.

Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060.

Ziegler, A. and Vens, M. (2010). Generalized estimating equations: Notes on the choice of working correlation matrix. *Biometrics*, 49(1):421–425.

Zorn, C. J. W. (2001). Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, pages 470–490.

Appendix A (Table and figures)

DESCRIPTIVE ANALYSIS

Table 5.1: Summary statistics of the Okavango river data

Statistic	N	Mean	St. Dev.	Min	Max
Time	5,568	24.500	13.855	1	48
Year	5,568			1950	2007
Months	5,568			1	12
Water.levels in $\mathfrak{m}$	5,423	3.847	1.161	1.960	8.680
Water.flows in $\mathfrak{m}^3.s^{-1}$	5,539	236.996	186.268	11.530	1,342.290

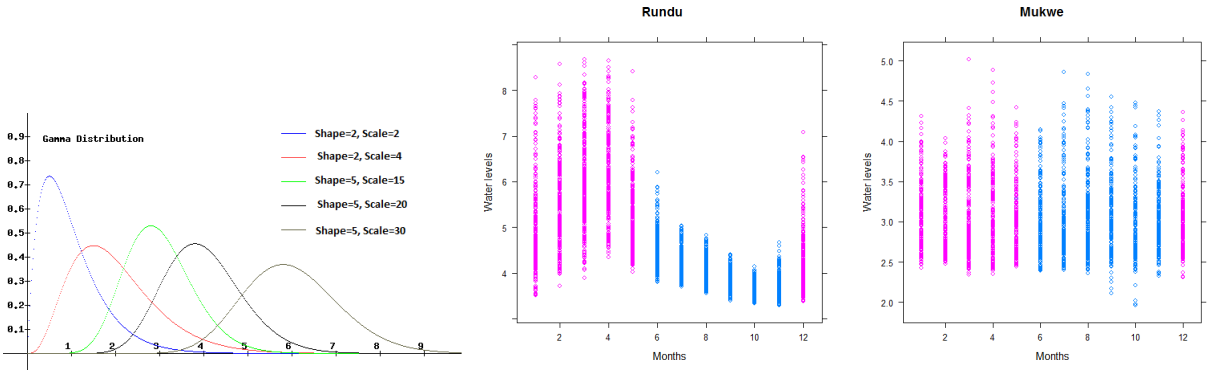


Figure 5.1: Gamma distribution with different shape and scale parameters

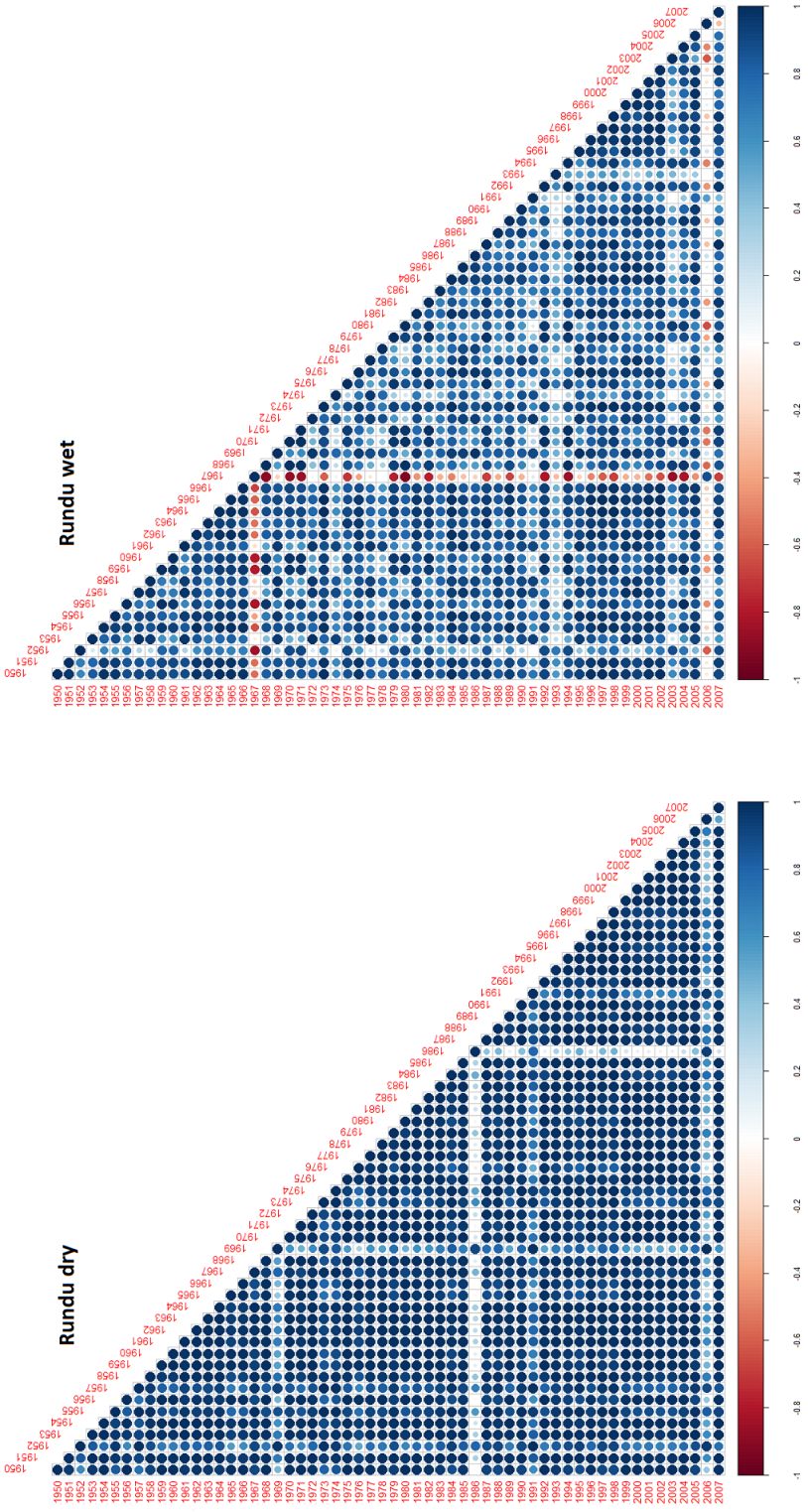


Figure 5.2: Correlation plots of water levels at Rundu

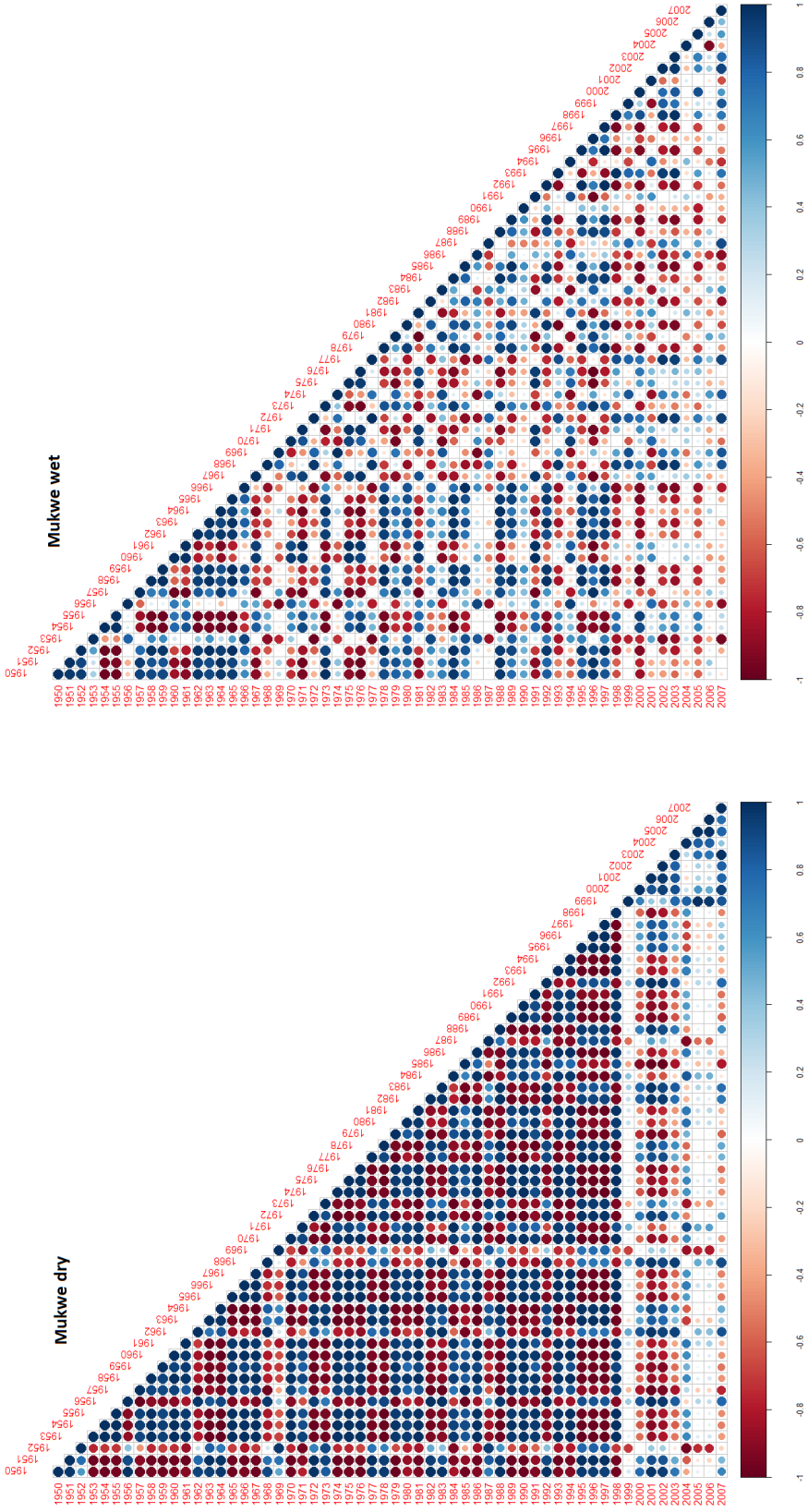


Figure 5.3: Correlation plots of water levels at Mukwe

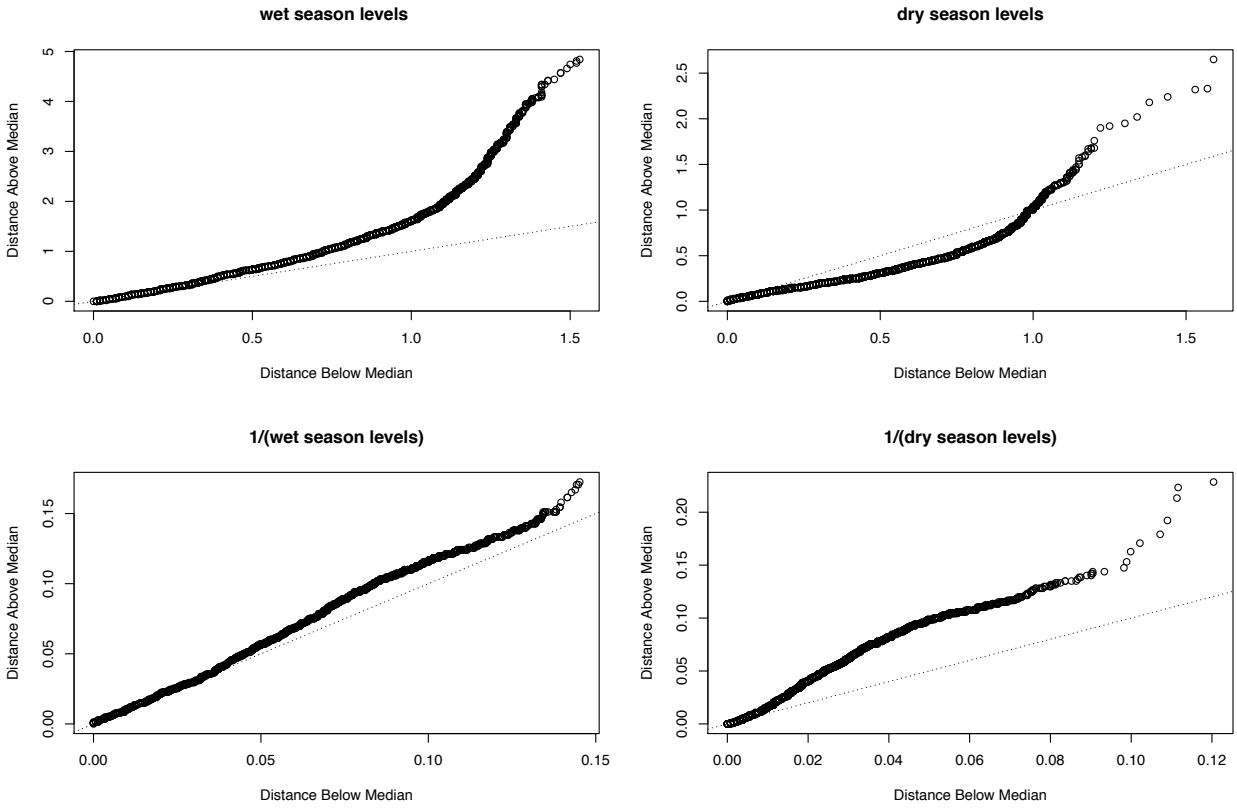


Figure 5.4: Reciprocal transformation of water levels for wet and dry seasons

RESIDUAL ANALYSIS

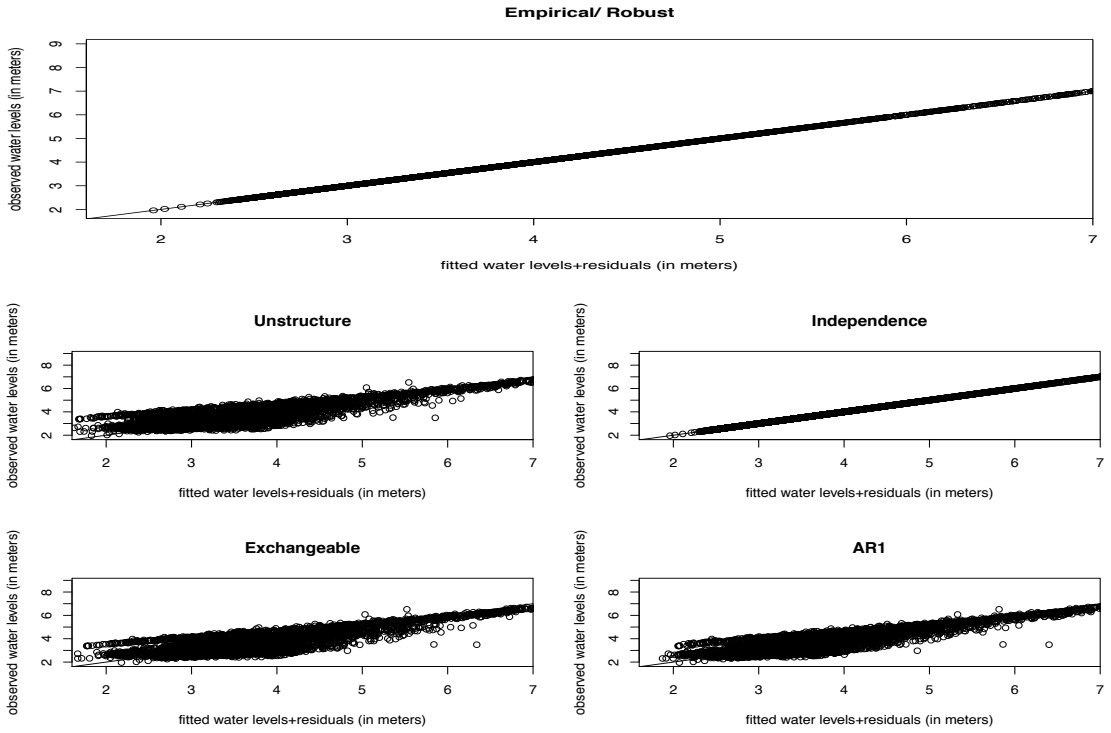


Figure 5.5: Observed vs (Predicted values+residuals) for different GEE models

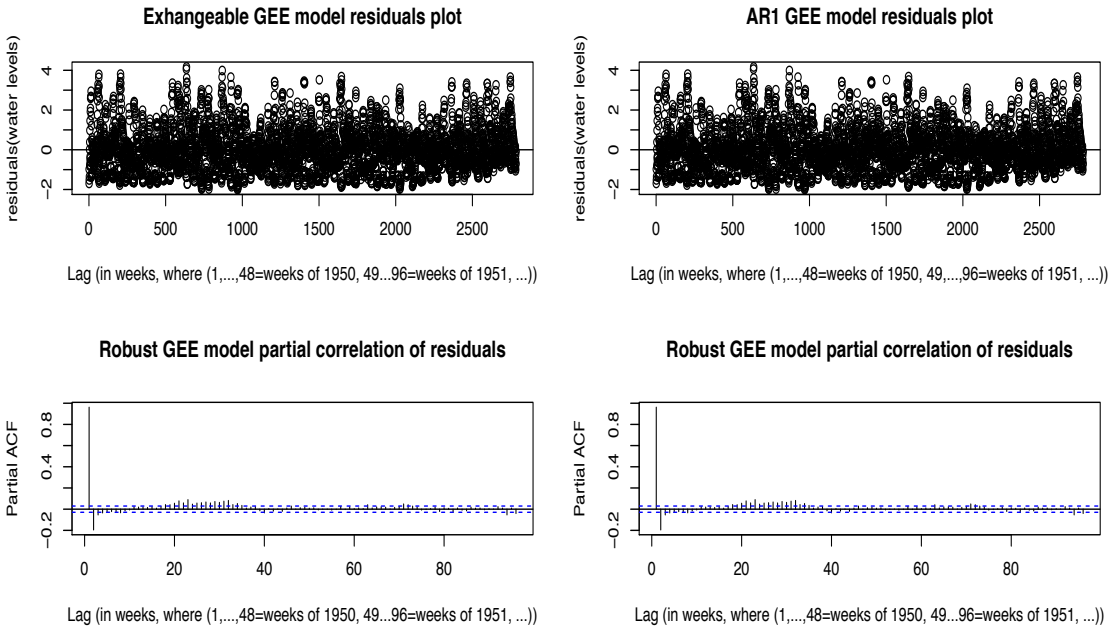


Figure 5.6: Residuals for Exchangeable and AR1 GEE models

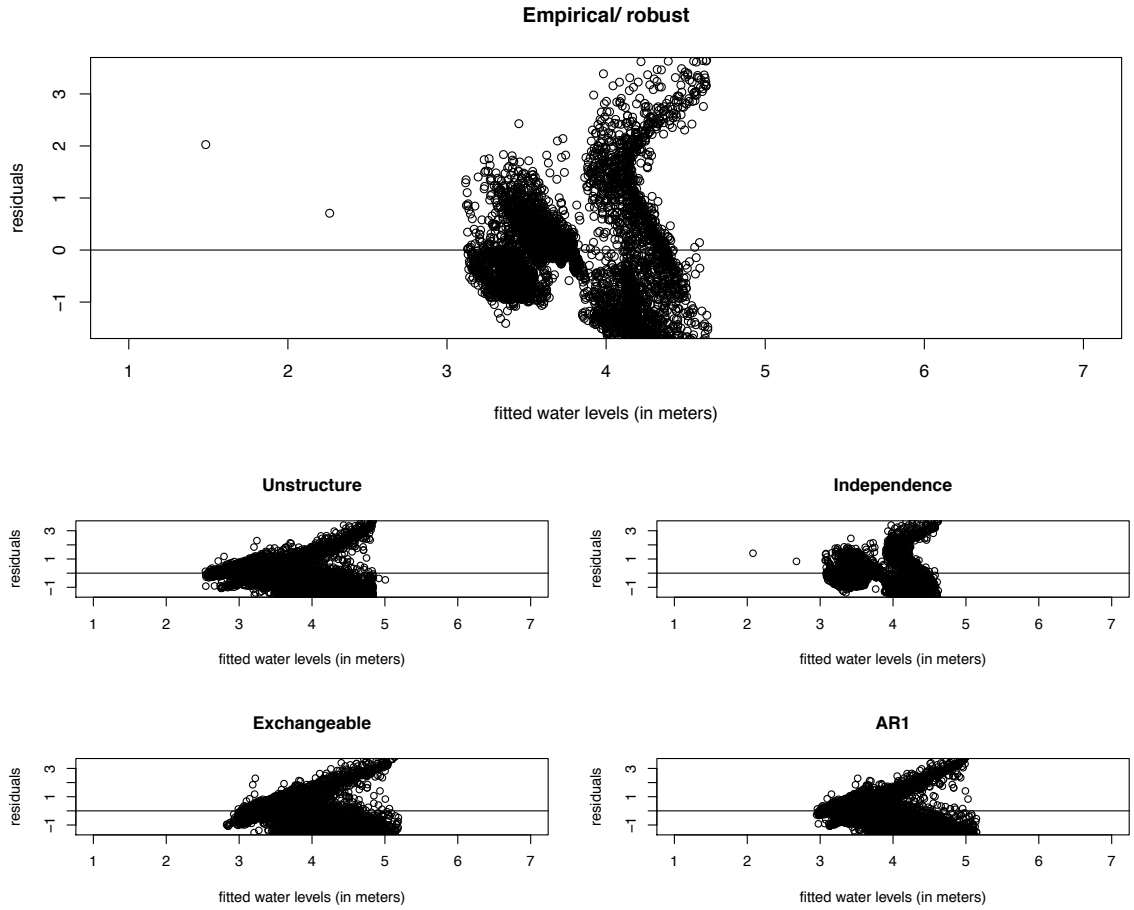


Figure 5.7: Residuals vs Predicted values for different GEE models

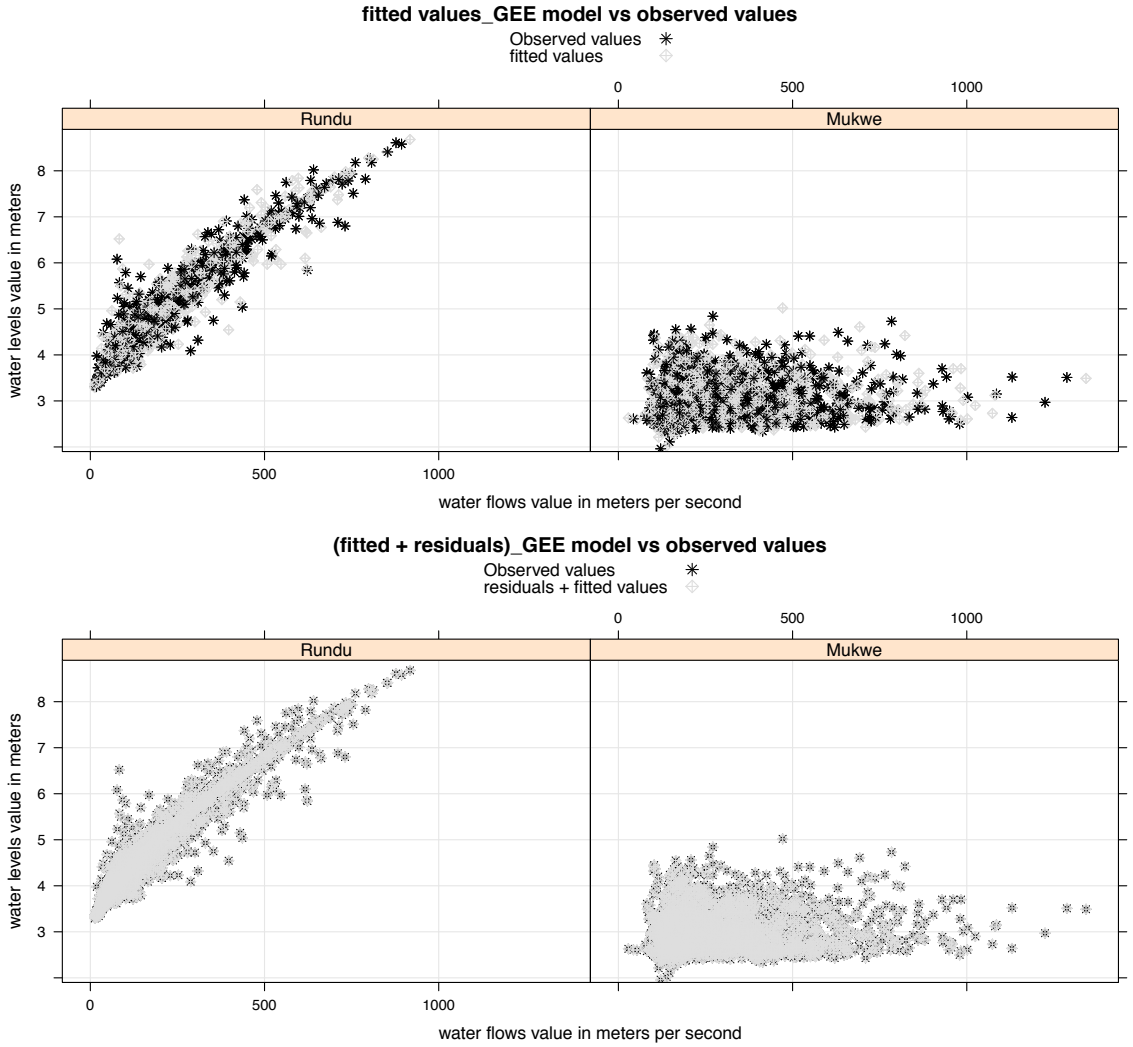


Figure 5.8: Predicted and observed Water values against water flow values for GEE model

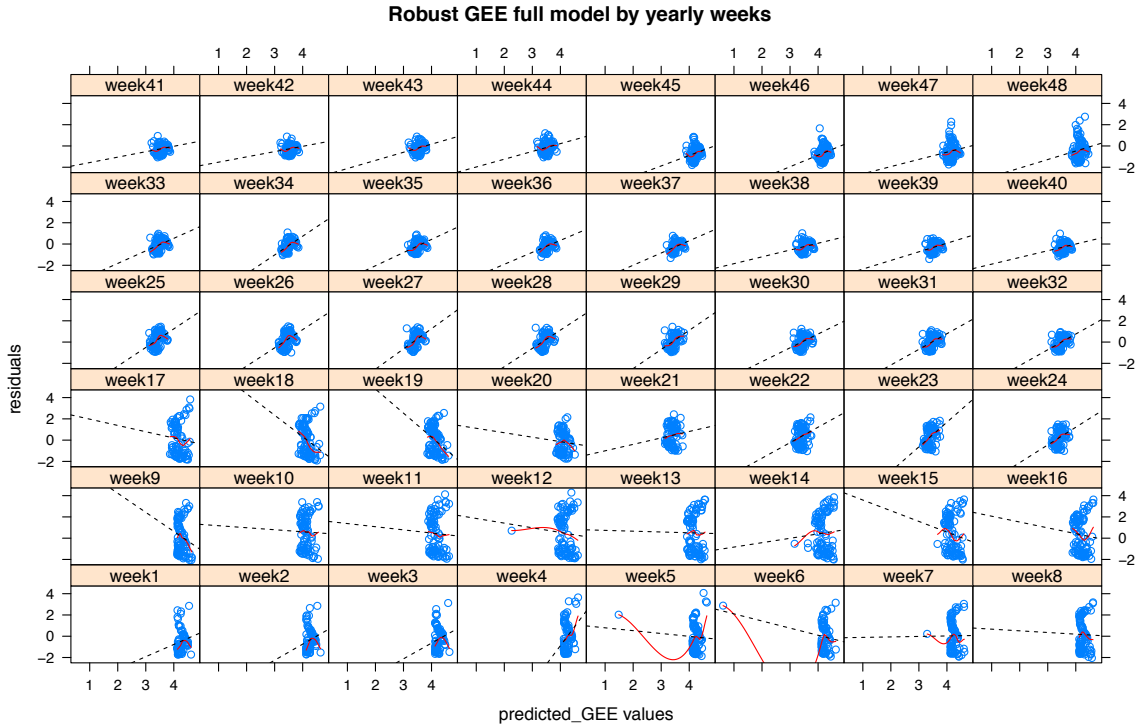


Figure 5.9: Residual pattern with fitted valued within yearly weeks and differences between yearly weeks for GEE model

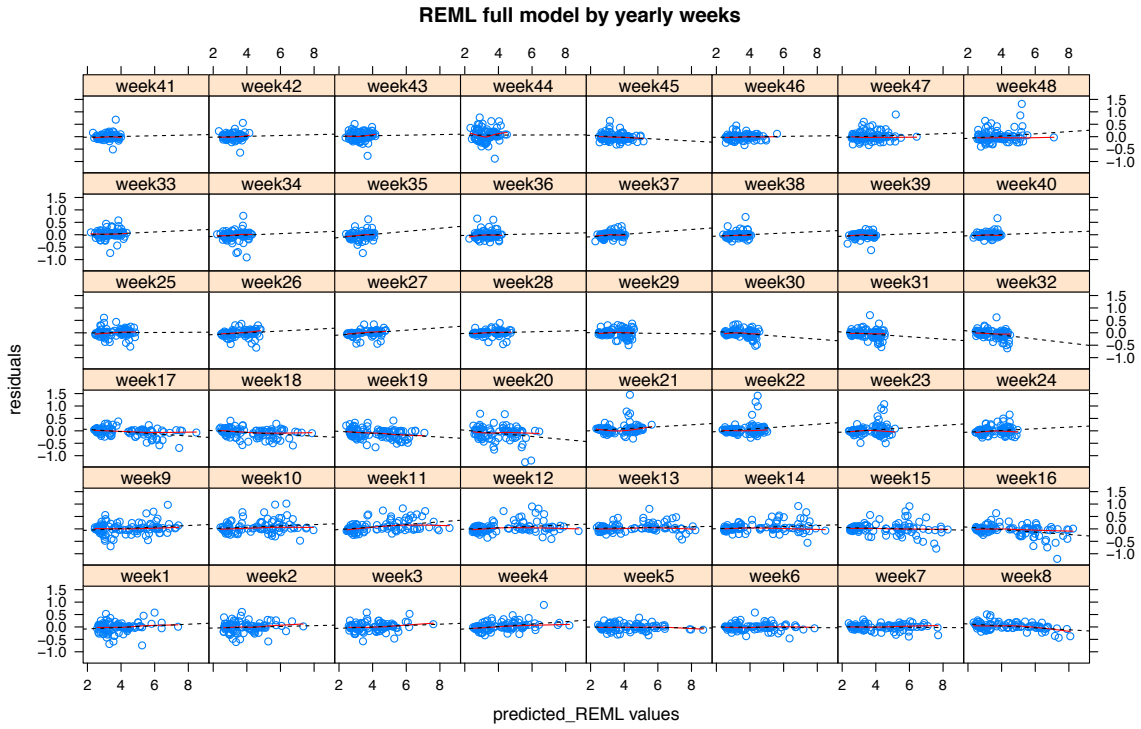


Figure 5.10: Residual pattern with fitted valued within yearly weeks and differences between yearly weeks for REML model

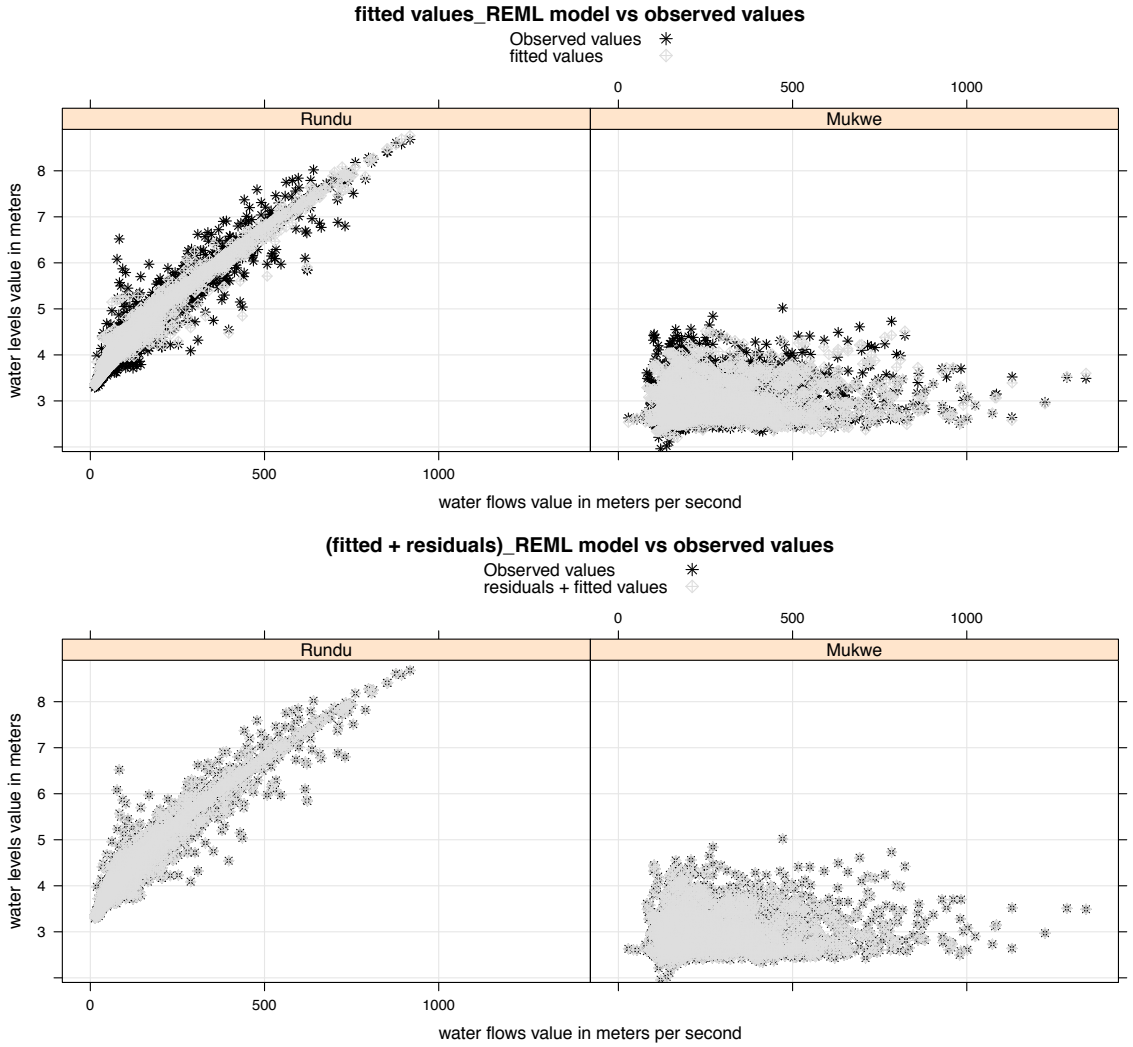


Figure 5.11: Predicted and observed Water values against water flow values for REML model

Appendix B (R-codes)

Import r-packages, data and partition data

```
install.packages(c("dplyr","VGAM","actuar","pracma","geepack","MASS","MESS","lattice","latticeExtra","stargazer","psych","nlme
    ","lme4","lmtest","doBy","beepr","car","randtests","DMwR","grid","corrplot","nortest","plyr", "calibrate"))
library(dplyr);library(actuar);library(MRSea);library(ggplot2);library(VGAM);library(actuar);library(pracma);library(geepack);
    library(MASS);library(MESS);library(lattice);library(latticeExtra);library(stargazer);library(psych);library(nlme);
    library(lme4);library(lmtest);library(doBy);library(beepr);library(car);library(randtests);library(DMwR);library(grid);
    library(corrplot);library(nortest);library(mgcv);library(splines);library(plyr);library(calibrate)
#
setwd("/Users/unandapo/OneDrive/Master of science report") # setting my working directorate
MScData<-readRDS("MScData.rds") # importing my dataset
NewMScData<-knnImputation(data=MScData, k=5, meth="weighAvg") # impute missing values using k-nearest neighbour method
NewMScDataLoc1<-NewMScData[NewMScData$Locations=="1", ];NewMScDataLoc2<-NewMScData[NewMScData$Locations=="2", ];
    NewMScDataSeas0<-NewMScData[NewMScData$Seasons=="0", ];NewMScDataSeas1<-NewMScData[NewMScData$Seasons=="1", ] # data
    sampling
NewMScDataSeas0Loc1<-NewMScDataLoc1[NewMScDataLoc1$Seasons=="0", ];NewMScDataSeas1Loc1<-NewMScDataLoc1[NewMScDataLoc1$Seasons
    =="1", ];NewMScDataSeas0Loc2<-NewMScDataLoc2[NewMScDataLoc2$Seasons=="0", ];NewMScDataSeas1Loc2<-NewMScDataLoc2[
    NewMScDataLoc2$Seasons=="1", ] # data sampling
CorrMukweDry<-readRDS("CorrMD.rds");CorrMukweWet<-readRDS("CorrMW.rds");CorrRunduDry<-readRDS("CorrRD.rds");CorrRunduWet<-
    readRDS("CorrRW.rds") # import data for covariance
```

Descriptive statistics

```
t.test(Water.levels ~ Seasons, data=NewMScData); t.test(Water.levels ~ Seasons, data=NewMScDataLoc1); t.test(Water.levels ~
    Seasons, data=NewMScDataLoc2) # t test for difference in mean
#
wilcox.test(Water.levels~Seasons, mu=0, alt="two.sided", conf.int=T, conf.level=0.95, pared=F, exact=F, correct=T, data=
    NewMScData);wilcox.test(Water.levels~Seasons, mu=0, alt="two.sided", conf.int=T, conf.level=0.95, pared=F, exact=F,
    correct=T, data=NewMScDataLoc1);wilcox.test(Water.levels~Seasons, mu=0, alt="two.sided", conf.int=T, conf.level=0.95,
    pared=F, exact=F, correct=T, data=NewMScDataLoc2) # Wilcoxon sum rank test
shapiro.test(NewMScDataSeas0$Water.level);shapiro.test(NewMScDataSeas1$Water.level);shapiro.test(NewMScDataSeas0Loc1$Water.
    level);shapiro.test(NewMScDataSeas1Loc1$Water.level);shapiro.test(NewMScDataSeas0Loc2$Water.level);shapiro.test(
    NewMScDataSeas1Loc2$Water.level) # shapiro test for normality
ad.test(NewMScDataSeas0$Water.level);ad.test(NewMScDataSeas1$Water.level);ad.test(NewMScDataSeas0Loc1$Water.level);ad.test(
    NewMScDataSeas1Loc1$Water.level);ad.test(NewMScDataSeas0Loc2$Water.level);ad.test(NewMScDataSeas1Loc2$Water.level) #
    anderson test for normality
#
pdf("SymetryPlot.pdf", width = 13, height=7);par(mfrow=c(2,2))
symplot = function(x) { n = length(x);n2 = n %/% 2; sx = sort(x); mx = median(x); plot(mx - sx[1:n2], rev(sx)[1:n2] - mx, main
    ="Rundu wet season ", xlab = "Distance Below Median", ylab = "Distance Above Median");abline(a = 0, b = 1, lty = "dotted
    ")};symplot(NewMScDataSeas1Loc1$Water.levels)
```

```
symplot = function(x) { n = length(x);n2 = n %/% 2; sx = sort(x); mx = median(x); plot(mx - sx[1:n2], rev(sx)[1:n2] - mx, main
     ="Rundu dry season ", xlab = "Distance Below Median", ylab = "Distance Above Median");abline(a = 0, b = 1, lty = "dotted
     ")};symplot(NewMScDataSeas0Loc1$Water.levels)
symplot = function(x) { n = length(x);n2 = n %/% 2; sx = sort(x); mx = median(x); plot(mx - sx[1:n2], rev(sx)[1:n2] - mx, main
     ="Mukwe wet season ", xlab = "Distance Below Median", ylab = "Distance Above Median");abline(a = 0, b = 1, lty = "dotted
     ")};symplot(NewMScDataSeas1Loc2$Water.levels)
symplot = function(x) { n = length(x);n2 = n %/% 2; sx = sort(x); mx = median(x); plot(mx - sx[1:n2], rev(sx)[1:n2] - mx, main
     ="Mukwe dry season ", xlab = "Distance Below Median", ylab = "Distance Above Median");abline(a = 0, b = 1, lty = "dotted
     ")};symplot(NewMScDataSeas0Loc2$Water.levels) # Symmetryplots
dev.off()
#

coplot(Water.flows~Water.levels|factor(Locations,labels=c("Rundu","Mukwe")) + factor(Months),data=NewMScData, ylab="Water
     flows (in meters per second)", xlab = "Water levels (in meters)") # Level_Flow scatter plot by Location and Months
pdf("flowLevelsScatter.pdf", width = 13, height=7);
xyplot(Water.levels~Water.flows| factor(Seasons, label=c("Dry","Wet"))+factor(Locations, label=c("Rundu", "Mukwe")), xlab="
     Water flows (in meters per second)", ylab = "Water levels (in meters)", main="Water.levels-Water.flows correlation
     scatter plot", data=NewMScData) # Level_Flow scatter plot by Location and Season
dev.off()
#
pdf("kernelDensity.pdf", width = 13, height=7);
histogram( ~ Water.levels |factor(Locations,labels=c("Rundu","Mukwe")) + factor(Seasons,labels=c("Dry season","Wet season")),
     data=NewMScData, xlab = "Water levels (in meters)", main="Water levels by Locations and Seasons", type = "density", panel
      = function(x, ...) { panel.histogram(x, ...); panel.densityplot(x)} ) # Kernel density and histogrames of water levels
dev.off()
pdf("boxplots.pdf", width = 13, height=7);
bwplot(~Water.levels|factor(Locations,labels=c("Rundu","Mukwe")) + factor(Seasons,labels=c("Dry season","Wet season")), data=
     NewMScData,main="Water levels by Seasons and Locations",xlab="Water levels (in meters)") # box plot of water levels
dev.off()
#
MD <- cor(newCorrMukweDry, use="complete.obs"); MW <- cor(newCorrMukweWet, use="complete.obs"); RD <- cor(newCorrRunduDry, use
     ="complete.obs"); RW <- cor(newCorrRunduWet, use="complete.obs") # correlation matrix
corrplot(MD, type = "lower"); corrplot(MW, type = "lower"); corrplot(RD, type = "lower"); corrplot(RW, type = "lower") #
     correlation matrix plot
par(mfrow=c(2,2)); plot.ts(diag(MD[-1,]), ylab = "lag1(corrrelation)", main="Mukwe dry"); abline(0,0); plot.ts(diag(MW[-1,]),
     ylab = "lag1(corrrelation)", main="Mukwe wet"); abline(0,0); plot.ts(diag(RD[-1,]), ylab = "lag1(corrrelation)", main="
     Rundu dry"); abline(0,0); plot.ts(diag(RW[-1,]), ylab = "lag1(corrrelation)", main="Rundu wet"); abline(0,0)# plots of
     lag1 correlation
#
pdf("longProfile.pdf", width = 15, height=7);
xyplot(Water.levels ~ Year| factor(Locations,labels=c("Rundu","Mukwe")), groups = Seasons,NewMScData, key = list(text = list(c
     ("Dry season", "Wet season")), points = list(pch = 8:9, col = c("black","gray87"))),pch = 8:9, col = c("black","gray87"),
      panel = function(...)panel.xyplot(lty = 6,grid=TRUE,...), ylim = 1.9:9, xlim = 1949:2008.4 , xlab="Time (in Years)",
     ylab="Water levels (in meters)",main="longitudinal water levels plot per year ") # longitudinal plot
dev.off()
#
pdf("profileplot.pdf", width = 13, height=7);
xyplot(Water.levels ~ Year| factor(Seasons,labels=c("dry season","wet season"))+factor(Locations,labels=c("Rundu","Mukwe")),
     groups = Time,NewMScData, pch = 1:24, col = c("black","blue4","brown4","azure4","bisque4","darkgreen","darkorchid1","
     darkolivegreen1","deeppink4","gray100","green","ivory","maroon1","pink","peru","tan2","deepskyblue","firebrick","gray17
     ","gray48","yellow","violet","tomato4","thistle1"), panel = function(...)panel.xyplot(lty = 6,grid=TRUE,...), ylim =
     1.9:9, xlim = 1949:2008.4 ,type="o", xlab="Time (in Years)",ylab="Weekly Water levels (in meters)",main="longitudinal
     profile plots of water levels for each yearly weeks") # Longitudinal weekly profile plots
dev.off()
#
pdf("xyplotRunduLev1.pdf", width = 12, height=8)
xyplot(Water.levels~Year|factor(Time,label=c("week1", "week2", "week3", "week4", "week5", "week6","week7", "week8", "week9",
     "week10", "week11", "week12","week13", "week14", "week15","week16", "week17", "week18","week19", "week20", "week21","
     week22", "week23", "week24","week25", "week26", "week27",  "week28", "week29", "week30","week31", "week32", "week33","
     week34", "week35", "week36", "week37", "week38", "week39","week40", "week41", "week42","week43", "week44", "week45" ,"
     week46","week47", "week48" )),data=NewMScDataLoc1,as.table=T,ylab="Water levels (in meters)", type="o", panel = function
     (x, y){ panel.xyplot(x, y); panel.lmline(x, y)}, ylim=c(3, 9), xlab="Time (in Years (1950-2007))", main="Rundu empirical
     growth plot") # Rundu empirical growth plot
dev.off()
```

```
pdf("xyplotMukweLev1.pdf", width = 12, height=8)
xyplot(Water.levels~Year|factor(Time,label=c("week1", "week2", "week3", "week4", "week5", "week6","week7", "week8", "week9",
     "week10", "week11", "week12","week13", "week14", "week15","week16", "week17", "week18","week19", "week20", "week21","
     week22", "week23", "week24","week25", "week26", "week27",  "week28", "week29", "week30","week31", "week32", "week33","
     week34", "week35", "week36", "week37", "week38", "week39","week40", "week41", "week42","week43", "week44", "week45" ,"
     week46","week47", "week48" )),data=NewMScDataLoc2,as.table=T,ylab="Water levels (in meters)", type="o", panel = function
     (x, y){ panel.xyplot(x, y); panel.lmline(x, y)}, ylim=c(1.8, 5), xlab="Time (in Years (1950-2007))", main="Mukwe
     empirical growth plot") # Mukwe empirical growth plot
dev.off()
#
pdf("transformation.pdf", width = 12, height=8)
par(mfrow=c(2,2))
symplot = function(x) { n = length(x);n2 = n %/% 2; sx = sort(x); mx = median(x); plot(mx - sx[1:n2], rev(sx)[1:n2] - mx, main
     ="wet season levels ", xlab = "Distance Below Median", ylab = "Distance Above Median");abline(a = 0, b = 1, lty = "dotted
     ")};symplot(NewMScDataSeas1$Water.levels)
symplot = function(x) { n = length(x);n2 = n %/% 2; sx = sort(x); mx = median(x); plot(mx - sx[1:n2], rev(sx)[1:n2] - mx, main
     ="dry season levels", xlab = "Distance Below Median", ylab = "Distance Above Median");abline(a = 0, b = 1, lty = "dotted
     ")};symplot(NewMScDataSeas0$Water.levels)
symplot = function(x) { n = length(x);n2 = n %/% 2; sx = sort(x); mx = median(x); plot(mx - sx[1:n2], rev(sx)[1:n2] - mx, main
     ="1/(wet season levels)", xlab = "Distance Below Median", ylab = "Distance Above Median");abline(a = 0, b = 1, lty = "
     dotted")};symplot(1/NewMScDataSeas1$Water.levels)
symplot = function(x) { n = length(x);n2 = n %/% 2; sx = sort(x); mx = median(x); plot(mx - sx[1:n2], rev(sx)[1:n2] - mx, main
     ="1/(dry season levels)", xlab = "Distance Below Median", ylab = "Distance Above Median");abline(a = 0, b = 1, lty = "
     dotted")};symplot(1/NewMScDataSeas0$Water.levels)
dev.off()
```

GEE and REML models based on Real data

```
setwd("/Users/unandapo/OneDrive/Master of science report")
MScData<-readRDS("MScData.rds")
NewMScData<-knnImputation(data=MScData, k=5, meth="weighAvg") # impute missing values using k-nearest neighbour method
#
attach(NewMScData)
m<-1;m<-ifelse(Months<3,1,m);m<-ifelse(Months>2 & Months<6,2,m);m<-ifelse(Months>5 & Months<9,3,m);m<-ifelse(Months>8 & Months
     <12,4,m);NewMScData$m<- m # creating quota for the data set
NewMScData$id<- as.numeric(paste(NewMScData$Year, NewMScData$Locations, NewMScData$m, sep=""))
NewMScData$zTime<- as.numeric(paste( NewMScData$Year_1_69, NewMScData$Time, sep=""))
View(NewMScData)
#
pdf("glmcorrelation.pdf", width = 10, height=5)
fit2<-glm(Water.levels~ Water.flows + as.factor(Months)+Locations*Seasons,data=NewMScData)
acf(fit2$residuals,96, main="Correlogram of glm residuals", xlab="Lag (in weeks, where (1-48=weeks of 1950, 49-96=weeks of
     1951, ...))");
abline(v=12, lwd=2, lty=2);abline(v=24, lwd=2, lty=2, col="red");abline(v=36, lwd=2, lty=2, col="blue");abline(v=48, lwd=2,
     lty=1);abline(v=60, lwd=2, lty=2);abline(v=72, lwd=2, lty=2, col="red");abline(v=84, lwd=2, lty=2, col="blue");abline(v
     =96, lwd=2, lty=2)
dev.off()
#
################ Cross validated sample ##############
splitdf1 <- function(dataframe, seed=NULL) {
  if (!is.null(seed)) set.seed(seed)
  index <- 1:nrow(dataframe)
  trainindex <- sample(index, trunc(length(index)*8/10)) # making 80 % of data as training data and the remaining as test data
  trainset <- dataframe[trainindex, ]
  testSet <- dataframe[-trainindex, ]
  list(trainset=trainset,testSet=testSet)
} # cross validated sample
splits1 <- splitdf1(NewMScData, seed=10112014) # train and test set sample
trainSet <- splits1$trainset;#trainSet<- orderBy(~TimeCode_weeks, trainSet);
```

```
trainSetLoc1<-trainSet[trainSet$Locations=="1", ];trainSetLoc1<- orderBy(~TimeCode_weeks, trainSetLoc1);trainSetLoc2<-trainSet
     [trainSet$Locations=="2", ];trainSetLoc2<- orderBy(~TimeCode_weeks, trainSetLoc2);
trainSet<-rbind(trainSetLoc1,trainSetLoc2) # Ordering the train set data based on time and location
#View(trainSet)
testSet <- splits1$testSet;#testSet<- orderBy(~TimeCode_weeks, testSet);
testSetLoc1<-testSet[testSet$Locations=="1", ];testSetLoc1<- orderBy(~TimeCode_weeks, testSetLoc1);testSetLoc2<-testSet[
     testSet$Locations=="2", ];testSetLoc2<- orderBy(~TimeCode_weeks, testSetLoc2);
testSet<-rbind(testSetLoc1,testSetLoc2) # Ordering the test set data based on time and location
#View(testSet)
################################################
#
strt<-Sys.time()
Robust_gee<-geeglm(Water.levels~bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons, data=trainSet, id=id) # robuste gee
     model
geeUNSTR<-geeglm(Water.levels~bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons, id=id,data=trainSet, family=Gamma(link="
     identity"),corstr="unstructure");
geeEXCH<-geeglm(Water.levels~bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons, id=id,data=trainSet, family=Gamma(link="
     identity"),corstr="exchangeable");
geeAR1<-geeglm(Water.levels~bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons, id=id,data=trainSet, family=Gamma(link="
     identity"),corstr="ar1");
geeIND<-geeglm(Water.levels~bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons, id=id,data=trainSet, family=Gamma(link="
     identity"),corstr="independence");
Sys.time()-strt
summary(Robust_gee);summary(geeUNSTR);summary(geeEXCH);summary(geeAR1);summary(geeIND)
QIC(Robust_gee, geeUNSTR, geeEXCH, geeAR1, geeIND)
PRESS(Robust_gee)
#library(qpcR)
hist(Robust_gee$residuals)
#
f1<-Robust_gee$fitted.values+Robust_gee$residuals;
r1<-Robust_gee$residuals;
mss1<-if(Robust_gee$coefficient[1]) sum((f1 - mean(f1))^2) else sum(f1^2);rss1<-sum(r1^2);
r.squared1 <-mss1/(mss1 + rss1);
r.squared1;
bias<-mean(f1)-mean(Robust_gee$y);
bias;
biasPerc<-(bias/mean(Robust_gee$y))*100;
biasPerc # caculating the R square and bias
#
#############################################################################
trainSet$fitRobust_gee<-Robust_gee$fitted.values;
trainSet$fitUNSTR<-geeUNSTR$fitted.values;
trainSet$fitIND<-geeIND$fitted.values;
trainSet$fitEXCH<-geeEXCH$fitted.values;
trainSet$fitAR1<-geeAR1$fitted.values #predicted values
View(trainSet)
#
trainSet$fitRobust_gee<-Robust_gee$fitted.values+Robust_gee$residuals;
trainSet$fitUNSTR<-geeUNSTR$fitted.values+geeUNSTR$residuals;
trainSet$fitIND<-geeIND$fitted.values+geeIND$residuals;
trainSet$fitEXCH<-geeEXCH$fitted.values+geeEXCH$residuals;
trainSet$fitAR1<-geeAR1$fitted.values+geeAR1$residuals # predicted values+residuals
View(trainSet)
#
trainSet$fitRobust_gee_resid<-Robust_gee$residuals;
trainSet$fitUNSTR_resid<-geeUNSTR$residuals;
trainSet$fitIND_resid<-geeIND$residuals;
trainSet$fitEXCH_resid<-geeEXCH$residuals;
trainSet$fitAR1_resid<-geeAR1$residuals # model residuals
View(trainSet)
#############################################################################
pdf("geestructures_Robust.pdf", width = 11, height=4.5)
par(mfrow=c(1,1));plot(trainSet$fitRobust_gee, trainSet$Water.levels, ylim = range(1.9:9), xlim = range(1.8:7), main = "
     Empirical/ Robust", xlab = "fitted water levels (in meters)", ylab = "observed water levels (in meters)");abline(0,1)
```

```
dev.off()
pdf("geestructures.pdf", width = 11, height=5);
par(mfrow=c(2,2));
plot(trainSet$fitUNSTR, trainSet$Water.levels, ylim = range(1.9:9), xlim = range(1.8:7), main = "Unstructure", xlab = "fitted
     water levels (in meters)", ylab = "observed water levels (in meters)");abline(0,1);
plot(trainSet$fitIND, trainSet$Water.levels, ylim = range(1.9:9), xlim = range(1.8:7), main = "Independence", xlab = "fitted
     water levels (in meters)", ylab = "observed water levels (in meters)");abline(0,1);
plot(trainSet$fitEXCH, trainSet$Water.levels, ylim = range(1.9:9), xlim = range(1.8:7), main = "Exchangeable", xlab = "fitted
     water levels (in meters)", ylab = "observed water levels (in meters)");abline(0,1);
plot(trainSet$fitAR1, trainSet$Water.levels, ylim = range(1.9:9), xlim = range(1.8:7), main = "AR1", xlab = "fitted water
     levels (in meters)", ylab = "observed water levels (in meters)");abline(0,1) # check for model linearity
dev.off()
#
pdf("geestructures+error1.pdf", width = 11, height=4.5)
par(mfrow=c(1,1));plot(trainSet$fitRobust_gee, trainSet$Water.levels, ylim = range(1.9:9), xlim = range(1.8:7), main = "
     Empirical/ Robust", xlab = "fitted water levels+residuals (in meters)", ylab = "observed water levels (in meters)");
     abline(0,1)
dev.off()
pdf("geestructures+error.pdf", width = 11, height=5);
par(mfrow=c(2,2));
plot(trainSet$fitUNSTR, trainSet$Water.levels, ylim = range(1.9:9), xlim = range(1.8:7), main = "Unstructure", xlab = "fitted
     water levels+residuals (in meters)", ylab = "observed water levels (in meters)");abline(0,1);
plot(trainSet$fitIND, trainSet$Water.levels, ylim = range(1.9:9), xlim = range(1.8:7), main = "Independence", xlab = "fitted
     water levels+residuals (in meters)", ylab = "observed water levels (in meters)");abline(0,1);
plot(trainSet$fitEXCH, trainSet$Water.levels, ylim = range(1.9:9), xlim = range(1.8:7), main = "Exchangeable", xlab = "fitted
     water levels+residuals (in meters)", ylab = "observed water levels (in meters)");abline(0,1);
plot(trainSet$fitAR1, trainSet$Water.levels, ylim = range(1.9:9), xlim = range(1.8:7), main = "AR1", xlab = "fitted water
     levels+residuals (in meters)", ylab = "observed water levels (in meters)");abline(0,1) # check for model linearity
dev.off()
#
#
pdf("geestructures_linearity1.pdf", width = 11, height=4.5)
par(mfrow=c(1,1));plot(Robust_gee$fitted.values , Robust_gee$residuals, ylim = range(-1.5:3.5), xlim = range(1:7), main = "
     Empirical/ robust", xlab = "fitted water levels (in meters)", ylab = "residuals");abline(0,0);
dev.off()
pdf("geestructures_linearity.pdf", width = 11, height=4.5)
par(mfrow=c(2,2));
plot(geeUNSTR$fitted.values , geeUNSTR$residuals, ylim = range(-1.5:3.5), xlim = range(1:7), main = "Unstructure", xlab = "
     fitted water levels (in meters)", ylab = "residuals");abline(0,0)
plot(geeIND$fitted.values , geeIND$residuals, ylim = range(-1.5:3.5), xlim = range(1:7), main = "Independence", xlab = "fitted
      water levels (in meters)", ylab = "residuals");abline(0,0);
plot(geeEXCH$fitted.values , geeEXCH$residuals, ylim = range(-1.5:3.5), xlim = range(1:7), main = "Exchangeable", xlab = "
     fitted water levels (in meters)", ylab = "residuals");abline(0,0);
plot(geeAR1$fitted.values , geeAR1$residuals, ylim = range(-1.5:3.5), xlim = range(1:7), main = "AR1", xlab = "fitted water
     levels (in meters)", ylab = "residuals");abline(0,0) # check for constant variance
dev.off()
#
#
trainSet$predicted_robust<-Robust_gee$fitted.values;
trainSet$resid_robust<-Robust_gee$residuals;
trainSet$predicted_UNSTRD<-geeUNSTR$fitted.values;
trainSet$resid_UNSTR<-geeUNSTR$residuals;
trainSet$predicted_IND<-geeIND$fitted.values;
trainSet$resid_IND<-geeIND$residuals;
trainSet$predicted_EXCH<-geeEXCH$fitted.values;
trainSet$resid_EXCH<-geeEXCH$residuals;
trainSet$predicted_AR1<-geeAR1$fitted.values;
trainSet$resid_AR1<-geeAR1$residuals
#
trainSet$resid_bins<-as.numeric(cut_number(trainSet$resid_robust,2000))# divind data into 2000 bins of equal data points
trainSet$resid_bins1<-as.numeric(cut_number(trainSet$resid_UNSTR,2000))# divind data into 2000 bins of equal data points
trainSet$resid_bins2<-as.numeric(cut_number(trainSet$resid_IND,2000))# divind data into 2000 bins of equal data points
trainSet$resid_bins3<-as.numeric(cut_number(trainSet$resid_EXCH,2000))# divind data into 2000 bins of equal data points
trainSet$resid_bins4<-as.numeric(cut_number(trainSet$resid_AR1,2000))# divind data into 2000 bins of equal data points
```

```
#
resid_robust<-tapply(trainSet$resid_robust, trainSet$resid_bins, var);predicted_robust<-tapply(trainSet$predicted_robust,
     trainSet$resid_bins, mean) #calculating the variance and mean of the bins
resid_UNSTR<-tapply(trainSet$resid_UNSTR, trainSet$resid_bins1, var);predicted_UNSTR<-tapply(trainSet$predicted_UNSTR,
     trainSet$resid_bins1, mean) #calculating the variance and mean of the bins
resid_IND<-tapply(trainSet$resid_IND, trainSet$resid_bins2, var);predicted_IND<-tapply(trainSet$predicted_IND,
     trainSet$resid_bins2, mean) #calculating the variance and mean of the bins
resid_EXCH<-tapply(trainSet$resid_EXCH, trainSet$resid_bins3, var);predicted_EXCH<-tapply(trainSet$predicted_EXCH,
     trainSet$resid_bins3, mean) #calculating the variance and mean of the bins
resid_AR1<-tapply(trainSet$resid_AR1, trainSet$resid_bins4, var);predicted_AR1<-tapply(trainSet$predicted_AR1,
     trainSet$resid_bins4, mean) #calculating the variance and mean of the bins
#
pdf("mean_variance_Robust.pdf", width = 11, height=4.5)
par(mfrow=c(1,1));plot(predicted_robust , resid_robust, xlim = range(2:7),xlab="fitted values (in meters)", ylab="variance of
     the residuals",main="Robust model mean-variance relationship");abline(0,0);
dev.off()
pdf("mean_variance.pdf", width = 11, height=5)
par(mfrow=c(2,2));
plot(predicted_UNSTR , resid_UNSTR, xlim = range(2:7),xlab="fitted values (in meters)", ylab="variance of the residuals",main
     ="Unstructure model mean-variance relationship");abline(0,0);
plot(predicted_IND , resid_IND, xlim = range(2:7),xlab="fitted values (in meters)", ylab="variance of the residuals",main="
     Independence model mean-variance relationship");abline(0,0);
plot(predicted_EXCH , resid_EXCH, xlim = range(2:7),xlab="fitted values (in meters)", ylab="variance of the residuals",main="
     Exchangeable model mean-variance relationship");abline(0,0);
plot(predicted_AR1 , resid_AR1, xlim = range(2:7),xlab="fitted values (in meters)", ylab="variance of the residuals",main="AR1
      model mean-variance relationship");abline(0,0) # plots of mean-variance relashipnship
dev.off()
#
pdf("non_indepenence_Robust.pdf", width = 11, height=5)
par(mfrow=c(1,2));acf(Robust_gee$residuals, 1000,main="Robust correlation");runACF(block=trainSet$id, model=Robust_gee);
dev.off()

pdf("non_indepenence.pdf", width = 11, height=5)
par(mfrow=c(2,4));
acf(geeUNSTR$residuals, 1000, main="Unstructure correlation");runACF(block=trainSet$id, model=geeUNSTR);
acf(geeIND$residuals, 1000, main="Independence correlation");runACF(block=trainSet$id, model=geeIND);
acf(geeEXCH$residuals, 1000,main="Exchangeable correlation");runACF(block=trainSet$id, model=geeEXCH);
acf(geeAR1$residuals, 1000,main="AR1 correlation");runACF(block=trainSet$id, model=geeAR1) # residual plots of autocorrelation
      function and autocorrelation of bins : checking for independence
dev.off()
#
runs.test(Robust_gee$residuals) # Test for randomness in residuals
resettest(Robust_gee, power=2:5, type="regressor", data=trainSet);resettest(Robust_gee, power=1^(1/2), type="regressor", data=
     trainSet) # Ramsey RESET Test
t.test(trainSet$Water.levels, Robust_gee$fitted.values, paired=TRUE)# t test for difference in the mean of predicted and
     observed values
wilcox.test(trainSet$Water.levels, Robust_gee$fitted.values, paired=T)
#
trainSet$predicted<- Robust_gee$fitted.values;
res <- stack(data.frame(Observed = trainSet$Water.levels, Predicted = trainSet$predicted));
ress1 <- cbind(res, xW = rep(trainSet$TimeCode_weeks, 2));ress12 <- cbind(res, xW2 = rep(trainSet$Water.flows, 2))

pdf("fitVSobsD.pdf", width = 11, height=4.3)
xyplot(values ~ xW | factor(trainSet$Locations, label=c("Rundu","Mukwe")),   data = ress1, key = list(text = list(c("Observed
     ", "Predicted")),points = list(pch = 8:9, col = c("black","gray87"))),pch = 8:9, col = c("black","gray87"), panel =
     function(...)panel.xyplot(lty = 6,grid=TRUE,...), ylim = 1.9:9, xlim = 0:2900 , main="predicted vs observed values", ylab
     ="water levels value in meters", xlab="Time Index ( in weeks)") # longitudinal plot of observed and predicted values
dev.off()
#
trainSet$predicted<-Robust_gee$residuals+Robust_gee$fitted.values # add error to predicted values
res <- stack(data.frame(Observed = trainSet$Water.levels, Predicted = trainSet$predicted));
ress1 <- cbind(res, xW = rep(trainSet$TimeCode_weeks, 2));ress12 <- cbind(res, xW2 = rep(trainSet$Water.flows, 2))

pdf("fitVSobsW.pdf", width = 11, height=4.3)
```

```
xyplot(values ~ xW | factor(trainSet$Locations, label=c("Rundu","Mukwe")),  data = ress1, key = list(text = list(c("Observed",
      "Predicted + residuals")),points = list(pch = 8:9, col = c("black","gray87"))),pch = 8:9, col = c("black","gray87"),
     panel = function(...)panel.xyplot(lty = 6,grid=TRUE,...), ylim = 1.9:9, xlim = 0:2900 , main="observed values vs (
     predicted + residuals)", ylab="water levels value in meters", xlab="Time Index ( in weeks)") # longitudinal plot of
     observed and predicted values
dev.off()
#
pdf("levels_vs_flows_residuals_GEE.pdf", width = 11, height=5)
trainSet$predicted<-Robust_gee$residuals+Robust_gee$fitted.values;res <- stack(data.frame(Observed = trainSet$Water.levels,
     Predicted = trainSet$predicted));ress12 <- cbind(res, xW2 = rep(trainSet$Water.flows, 2))
xyplot(values ~ xW2 | factor(trainSet$Locations, label=c("Rundu","Mukwe")),group=ind, data = ress12, key = list(text = list(c
     ("Observed values", "residuals + fitted values")),points = list(pch = 8:9, col = c("black","gray87"))),pch = 8:9, col = c
     ("black","gray87"), panel = function(...)panel.xyplot(lty = 6,grid=TRUE,...), ylim = 1.9:9, main="(fitted + residuals)
     _GEE model vs observed values", ylab="water levels value in meters", xlab="water flows value in meters per second") #
     longitudinal plot of observed and predicted values
dev.off()
pdf("levels_vs_flows_GEE.pdf", width = 11, height=5)
trainSet$predicted<-Robust_gee$residuals+Robust_gee$fitted.values;res <- stack(data.frame(Observed = trainSet$Water.levels,
     Predicted = trainSet$predicted));ress12 <- cbind(res, xW2 = rep(trainSet$Water.flows, 2))
xyplot(values ~ xW2 | factor(trainSet$Locations, label=c("Rundu","Mukwe")), data = ress12, key = list(text = list(c("Observed
     values", "fitted values")),points = list(pch = 8:9, col = c("black","gray87"))),pch = 8:9, col = c("black","gray87"),
     panel = function(...)panel.xyplot(lty = 6,grid=TRUE,...), ylim = 1.9:9, main="fitted values_GEE model vs observed values
     ", ylab="water levels value in meters", xlab="water flows value in meters per second") # longitudinal plot of observed
     and predicted values
#trainSet$predicted<-Robust_gee$fitted.values;res <- stack(data.frame(Observed = trainSet$Water.levels, Predicted =
     trainSet$predicted));ress12 <- cbind(res, xW2 = rep(trainSet$Water.flows, 2))
#xyplot(values ~ xW2 | factor(trainSet$Locations, label=c("Rundu","Mukwe")), data = ress12, key = list(text = list(c("Observed
      values", "fitted values")),points = list(pch = 8:9, col = c("black","gray87"))),pch = 8:9, col = c("black","gray87"),
     panel = function(...)panel.xyplot(lty = 6,grid=TRUE,...), ylim = 1.9:9, main="fitted values_GEE model vs observed values
     ", ylab="water levels value in meters", xlab="water flows value in meters per second") # longitudinal plot of observed
     and predicted values
dev.off()
#
pdf("EXCH_Resid.pdf", width = 11, height=5)
par(mfrow=c(1,1));plot(Robust_gee$residuals ~ TimeCode_weeks , data = trainSet, main="Robust GEE model residuals plot", ylab="
     residuals(water levels)",xlab="Time (in weeks, where (1?48=weeks of 1950, 49?96=weeks of 1951, ...))");abline(0,0) # plot
      of residuals
dev.off()
#
pdf("acfGEE.pdf", width = 11, height=5)
par(mfrow=c(1,1));pacf(Robust_gee$residuals, main="Robust GEE model partial correlation of residuals",96,xlab="Lag (in weeks,
     where (1?48=weeks of 1950, 49?96=weeks of 1951, ...))") # plot of partial autocorelation of residuals
dev.off()
#
pdf("EXCH_AR1_Resid_PACF.pdf", width = 11, height=5)
par(mfrow=c(2,2))
plot(geeEXCH$residuals ~ TimeCode_weeks , data = trainSet, main="Exhangeable GEE model residuals plot", ylab="residuals(water
     levels)",xlab="Lag (in weeks, where (1?48=weeks of 1950, 49?96=weeks of 1951, ...))");abline(0,0) # plot of residuals
plot(geeAR1$residuals ~ TimeCode_weeks , data = trainSet, main="AR1 GEE model residuals plot", ylab="residuals(water levels)",
     xlab="Lag (in weeks, where (1?48=weeks of 1950, 49?96=weeks of 1951, ...))");abline(0,0)  # plot of residuals
pacf(geeEXCH$residuals, main="Robust GEE model partial correlation of residuals", 96,xlab="Lag (in weeks, where (1?48=weeks of
      1950, 49?96=weeks of 1951, ...))") # plot of partial autocorelation of residuals
pacf(geeAR1$residuals, main="Robust GEE model partial correlation of residuals",96,xlab="Lag (in weeks, where (1?48=weeks of
     1950, 49?96=weeks of 1951, ...))") # plot of partial autocorelation of residuals
dev.off()
#
pdf("residPattern.pdf", width = 11, height=7)
xyplot(Robust_gee$residuals ~Robust_gee$fitted.values | factor(trainSet$Time,label=c("week1", "week2", "week3", "week4", "
     week5", "week6","week7", "week8", "week9",  "week10", "week11", "week12","week13", "week14", "week15","week16", "week17",
      "week18","week19", "week20", "week21","week22", "week23", "week24","week25", "week26", "week27",  "week28", "week29", "
     week30","week31", "week32", "week33","week34", "week35", "week36", "week37", "week38", "week39","week40", "week41", "
     week42","week43", "week44", "week45" ,"week46","week47", "week48" )),ylab="residuals", xlab="predicted_GEE values",main="
     Robust GEE full model by yearly weeks", panel=function(x,y){panel.xyplot(x,y); panel.loess(x,y,span=0.75, col="red");
     panel.lmline(x,y,lty=2) })  # residual pattern within and between yearly weeks
```

```
dev.off()
#
z=bs(trainSet$Water.flows, knots = mean(trainSet$Water.flows));head(z)
trainSet$predicted<-Robust_gee$coefficients[1]+Robust_gee$coefficients[2]*z[,1]+Robust_gee$coefficients[3]*z[,2]+
      Robust_gee$coefficients[4]*z[,3]+Robust_gee$coefficients[5]*z[,4]+  Robust_gee$coefficients[6]*trainSet$zTime+
      Robust_gee$coefficients[7]*trainSet$Seasons+Robust_gee$coefficients[8]*trainSet$zTime*trainSet$Seasons  # estimating
      Water levels values using model parameters
#
res <- stack(data.frame(Observed = trainSet$Water.levels, Predicted = trainSet$predicted));
ress1 <- cbind(res, xW = rep(trainSet$TimeCode_weeks, 2));ress12 <- cbind(res, xW2 = rep(trainSet$Water.flows, 2))

pdf("fitVSobsD_coef.pdf", width = 11, height=4.3)
xyplot(values ~ xW | factor(trainSet$Locations, label=c("Rundu","Mukwe")),  layout=c(2,1), data = ress1, key = list(text =
      list(c("Observed", "Predicted")),points = list(pch = 8:9, col = c("black","gray87"))),pch = 8:9, col = c("black","gray87
      "), panel = function(...)panel.xyplot(lty = 6,grid=TRUE,...), ylim = 1.9:9, xlim = 0:2900 , main="Observed values vs
      values predicted by Robust GEE coefficients", ylab="water levels value in meters", xlab="Time Index ( in weeks)") #
      longitudinal plot of observed and predicted values
#xyplot(values ~ xW, data = ress1, key = list(text = list(c("Observed", "Predicted")),points = list(pch = 8:9, col = c("black
      ","gray87"))),pch = 8:9, col = c("black","gray87"), panel = function(...)panel.xyplot(lty = 6,grid=TRUE,...), ylim =
      1.9:9, xlim = 0:2900 , main="Observed values vs values predicted by Robust GEE coefficients", ylab="water levels value in
       meters", xlab="Time Index ( in weeks)") # longitudinal plot of observed and predicted values
dev.off()
#xyplot(values ~ xW2 |  factor(trainSet$Locations, label=c("Rundu","Mukwe")), data = ress12,  group = ind,auto.key = TRUE,main
      ="predicted vs observed values", ylab="water levels value", xlab="water flows") # predicted values vs water flow values
#
#
#
strt<-Sys.time()
crossval_list<-list()
for ( iter in 1:100)
{
  splitdf1 <- function(dataframe, seed=NULL) {
    if (!is.null(seed)) set.seed(seed)
    index <- 1:nrow(dataframe)
    trainindex <- sample(index, trunc(length(index)*iter/100))
    trainset <- dataframe[trainindex, ]
    testSet <- dataframe[-trainindex, ]
    list(trainset=trainset,testSet=testSet)
  } # cross validated sample
  splits1 <- splitdf1(NewMScData, seed=10112014) # train and test set sample
  trainSet <- splits1$trainset;#trainSet<- orderBy(~TimeCode_weeks, trainSet);
  trainSetLoc1<-trainSet[trainSet$Locations=="1", ];trainSetLoc1<- orderBy(~TimeCode_weeks, trainSetLoc1);trainSetLoc2<-
       trainSet[trainSet$Locations=="2", ];trainSetLoc2<- orderBy(~TimeCode_weeks, trainSetLoc2);
  trainSet<-rbind(trainSetLoc1,trainSetLoc2) # Ordering the train set data based on time and location
  Robust_gee1<-geeglm(Water.levels~bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons, data=trainSet, id=id)
  Robust_gee2<-geeglm(Water.levels~bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons, id=id,data=trainSet, family=Gamma(
       link="identity"),corstr="exchangeable");
  Robust_gee3<-geeglm(Water.levels~bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons, id=id,data=trainSet, family=Gamma(
       link="identity"),corstr="ar1");
  Robust_gee4<-geeglm(Water.levels~bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons, id=id,data=trainSet, family=Gamma(
       link="identity"),corstr="independence");
  # Robust_gee<-geeglm(Water.levels~bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons, id=id,data=trainSet, family=Gamma(
       link="identity"),corstr="unstructure");
  #REML<-lmer(Water.levels~ bs(Water.flows, knots=mean(Water.flows))+zTime+zTime*Seasons+(1|id),data=trainSet)
  ##
  sample<- (iter/100)*5568;coefs1<-coef(summary(Robust_gee1))[,4];f11<-Robust_gee1$fitted.values+Robust_gee1$residuals;r11<-
       Robust_gee1$residuals; mss11<-if(Robust_gee1$coefficient[1]) sum((f11 - mean(f11))^2) else sum(f11^2);rss11<-sum(r11^2)
       ;r.squared11 <-mss11/(mss11 + rss11);r.squared11;ifelse(iter==1,Sample<-sample,Sample<-rbind(Sample, sample));ifelse(
       iter==1,MSE1<-mean(Robust_gee1$residuals^2),MSE1<-rbind(MSE1, mean(Robust_gee1$residuals^2))); ifelse(iter==1,QIC1<-QIC
       (Robust_gee1),QIC1<-rbind(QIC1,QIC(Robust_gee1)));ifelse(iter==1,RSQR1<-r.squared11,RSQR1<-rbind(RSQR1, r.squared11));
       ifelse(iter==1,pvalues1<-coefs1,pvalues1<-rbind(pvalues1, coefs1))
  #
  coefs2<-coef(summary(Robust_gee2))[,4];f12<-Robust_gee2$fitted.values+Robust_gee2$residuals;r12<-Robust_gee2$residuals;mss12
       <-if(Robust_gee2$coefficient[1]) sum((f12 - mean(f12))^2) else sum(f12^2);rss12<-sum(r12^2);r.squared12 <-mss12/(mss12
```

```
      + rss12);r.squared12;ifelse(iter==1,MSE2<-mean(Robust_gee2$residuals^2),MSE2<-rbind(MSE2, mean(Robust_gee2$residuals^2)
      )); ifelse(iter==1,QIC2<-QIC(Robust_gee2),QIC2<-rbind(QIC2,QIC(Robust_gee2)));ifelse(iter==1,RSQR2<-r.squared12,RSQR2<-
      rbind(RSQR2, r.squared12));ifelse(iter==1,pvalues2<-coefs2,pvalues2<-rbind(pvalues2, coefs2))
  #
  coefs3<-coef(summary(Robust_gee3))[,4];f13<-Robust_gee3$fitted.values+Robust_gee3$residuals;r13<-Robust_gee3$residuals;mss13
      <-if(Robust_gee3$coefficient[1]) sum((f13 - mean(f13))^2) else sum(f13^2);rss13<-sum(r13^2);r.squared13 <-mss13/(mss13
      + rss13);r.squared13;ifelse(iter==1,MSE3<-mean(Robust_gee3$residuals^2),MSE3<-rbind(MSE3, mean(Robust_gee3$residuals^2)
      )); ifelse(iter==1,QIC3<-QIC(Robust_gee3),QIC3<-rbind(QIC3,QIC(Robust_gee3)));ifelse(iter==1,RSQR3<-r.squared13,RSQR3<-
      rbind(RSQR3, r.squared13));ifelse(iter==1,pvalues3<-coefs3,pvalues3<-rbind(pvalues3, coefs3))
  #
  coefs4<-coef(summary(Robust_gee4))[,4];f14<-Robust_gee4$fitted.values+Robust_gee4$residuals;r14<-Robust_gee4$residuals;mss14
      <-if(Robust_gee4$coefficient[1]) sum((f14 - mean(f14))^2) else sum(f14^2);rss14<-sum(r14^2);r.squared14 <-mss14/(mss14
      + rss14);r.squared14;ifelse(iter==1,MSE4<-mean(Robust_gee4$residuals^2),MSE4<-rbind(MSE4, mean(Robust_gee4$residuals^2)
      ));ifelse(iter==1,QIC4<-QIC(Robust_gee4),QIC4<-rbind(QIC4,QIC(Robust_gee4)));ifelse(iter==1,RSQR4<-r.squared14,RSQR4<-
      rbind(RSQR4, r.squared14));ifelse(iter==1,pvalues4<-coefs4,pvalues4<-rbind(pvalues4, coefs4))
  ##
  crossval_list[[iter]]<-trainSet
};Sys.time()-strt;beep(8)
Datasample<-data.frame(Sample);Datasample$MSE_robust<-MSE1; Datasample$RSQR_robust<-RSQR1; Datasample$QIC_robust<-QIC1
Datasample$MSE_EXCH<-MSE2; Datasample$RSQR_EXCH<-RSQR2; Datasample$QIC_EXCH<-QIC2
Datasample$MSE_AR1<-MSE3; Datasample$RSQR_AR1<-RSQR3; Datasample$QIC_AR1<-QIC3
Datasample$MSE_IND<-MSE4; Datasample$RSQR_IND<-RSQR4; Datasample$QIC_IND<-QIC4
#pval_sample<-data.frame(Sample);pval_sample<-cbind(pval_sample, pvalues1);
View(Datasample)
saveRDS(Datasample , "Data_sample_AIC_MSE_RSQR.rds")
Datasample<-readRDS("Data_sample_AIC_MSE_RSQR.rds")
#
#
#
pdf("GEE_QIC_MSE_rsqr.pdf", width = 13, height=7);
par(mfrow=c(1,4))
range(Datasample$QIC_robust[,1]);range(Datasample$QIC_EXCH[,1]);range(Datasample$QIC_AR1[,1]);range(Datasample$QIC_IND[,1])
yrange<-c(35.5,17231); xrange<-range(Datasample$Sample)
plot(xrange, yrange, type="n", xlab="Sample size", ylab="QIC", main="QIC vs sample size relationship")
lines(Datasample$Sample, Datasample$QIC_EXCH[,1], col="Blue", type="o", pch=11)
lines(Datasample$Sample, Datasample$QIC_AR1[,1], col="brown", type="o", pch=8)
lines(Datasample$Sample, Datasample$QIC_IND[,1], col="grey", type="o", pch=21)
lines(Datasample$Sample, Datasample$QIC_robust[,1], col="Black", type="o", pch=18)
# add a legend
legend(0,17000, c("Exchangeable GEE", "AR1 GEE", "Independence GEE", "Robust GEE"), lty=c(1, 1, 1, 1), cex=0.5, bty="n", pch=c
     (11, 8, 21, 18), lwd=c(2.5,2.5), col=c("blue", "brown", "grey", "black"))
#
range(Datasample$MSE_robust);range(Datasample$MSE_EXCH);range(Datasample$MSE_AR1);range(Datasample$MSE_IND)
yrange<-c(0.402,1.191);xrange<-range(Datasample$Sample)
plot(xrange, yrange, type="n", xlab="Sample size", ylab="MSE", main="MSE vs sample size relationship")
lines(Datasample$Sample, Datasample$MSE_EXCH, col="Blue", type="o", pch=11)
lines(Datasample$Sample, Datasample$MSE_AR1, col="brown", type="o", pch=8)
lines(Datasample$Sample, Datasample$MSE_IND, col="grey", type="o", pch=21)
lines(Datasample$Sample, Datasample$MSE_robust, col="Black", type="o", pch=18)
# add a legend
legend(3100,0.49, c("Exchangeable GEE", "AR1 GEE", "Independence GEE", "Robust GEE"), lty=c(1, 1, 1, 1), cex=0.5, bty="n", pch
     =c(11, 8, 21, 18), lwd=c(2.5,2.5), col=c("blue", "brown", "grey", "black"))
#
range(Datasample$RSQR_robust);range(Datasample$RSQR_EXCH);range(Datasample$RSQR_AR1);range(Datasample$RSQR_IND)
yrange<-c(0.525, 0.687); xrange<-range(Datasample$Sample)
plot(xrange, yrange, type="n", xlab="Sample size", ylab="r.square", main="r.square vs sample size relationship")
lines(Datasample$Sample, Datasample$RSQR_EXCH, col="Blue", type="o", pch=11)
lines(Datasample$Sample, Datasample$RSQR_AR1, col="brown", type="o", pch=8)
lines(Datasample$Sample, Datasample$RSQR_IND, col="grey", type="o", pch=21)
lines(Datasample$Sample, Datasample$RSQR_robust, col="Black", type="o", pch=18)
# add a legend
legend(3200,0.69, c("Exchangeable GEE", "AR1 GEE", "Independence GEE", "Robust GEE"), lty=c(1, 1, 1, 1), cex=0.5, bty="n", pch
     =c(11, 8, 21, 18), lwd=c(2.5,2.5), col=c("blue", "brown", "grey", "black"))
#
```

```
range(pval_sample[,1]);range(pval_sample[,3]);range(pval_sample[,4]);range(pval_sample[,5]);range(pval_sample[,6])
yrange<-c(0.000037,0.97009); xrange<-range(pval_sample[,1])
plot(xrange, yrange, type="n", xlab="Sample size", ylab="p-values (of water flow bins-splines(BS))", main="Robust(GEE) p-
     values sample size relation")
lines(pval_sample[,1], pval_sample[,3], col="brown", type="o", pch=11)
lines(pval_sample[,1], pval_sample[,4], col="grey", type="o", pch=8)
lines(pval_sample[,1], pval_sample[,5], col="Blue", type="o", pch=21)
lines(pval_sample[,1], pval_sample[,6], col="Black", type="o", pch=18)
# add a legend
legend(3400,0.95, c("BS.1", "BS.2", "BS.3", "BS.4"), lty=c(1, 1, 1, 1), cex=0.5, bty="n", pch=c(11, 8, 21, 18), lwd=c(2.5,2.5)
     , col=c("brown", "grey", "Blue", "black"))
dev.off()
#
#
##### Restricted maximum likelihood model_CODES #####
#REML<-lmer(Water.levels~ bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons+(1|id),data=trainSet);
#REML<-lmer(Water.levels~ bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons+(Time|id),data=trainSet);
#REML<-lmer(Water.levels~ bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons+(zTime|id),data=trainSet);
REML<-lmer(Water.levels~ bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons+(Water.flows|id),data=trainSet) # (Water.flows
     |id)Correlated random intercept and slope

ranef(REML)
AIC(REML)
summary(REML)
Anova(REML)
stargazer(REML, type="text")
stargazer(REML)
#
runs.test(resid(REML)) # Test for randomness in residuals
resettest(REML, power=2:5, type="regressor", data=trainSet);
t.test(trainSet$Water.levels, fitted(REML), paired=TRUE)# t test for difference in the mean of predicted_REML and observed
     values
#
trainSet$predicted_REML<-fitted(REML)
trainSet$resid_REML<-resid(REML);
pdf("REML_Linearity.pdf", width = 11, height=4.3);
par(mfrow=c(1,2));plot(fitted(REML), trainSet$Water.levels, main = "REML linearity", xlab = "fitted water levels (in meters)",
      ylab = "water levels (in meters)");abline(0,1);plot(fitted(REML)+resid(REML), trainSet$Water.levels, main = "REML
     linearity", xlab = "fitted water levels+residuals (in meters)", ylab = "water levels (in meters)");abline(0,1); # check
     for model linearity
dev.off()

trainSet$resid_bins<-as.numeric(cut_number(trainSet$resid_REML,2000))# divind data into 2000 bins of equal data points
resid_REML<-tapply(trainSet$resid_REML, trainSet$resid_bins, var);predicted_REML_REML<-tapply(trainSet$predicted_REML,
     trainSet$resid_bins, mean) #calculating the variance and mean of the bins
pdf("mean_varianceREML.pdf", width = 11, height=4.3);
par(mfrow=c(1,2));plot(fitted(REML) , resid(REML), main = "REML residuals", xlab = "fitted water levels (in meters)", ylab = "
     residuals");abline(0,0);plot(predicted_REML_REML , resid_REML, xlab = "fitted water levels (in meters)", ylab="variance
     of the residuals",main="REML model mean-variance relationship");abline(0,0);# check for constant variance
dev.off()
pdf("non_indepenenceREML.pdf", width = 11, height=4.3);
par(mfrow=c(1,2));acf(resid(REML),1000, main="REML correlation");runACF(block=trainSet$id, model=REML);
dev.off()
#
res <- stack(data.frame(Observed = trainSet$Water.levels, predicted_REML = fitted(REML) ));
ress1 <- cbind(res, xW = rep(trainSet$TimeCode_weeks, 2));
ress12 <- cbind(res, xW2 = rep(trainSet$Water.flows, 2))

pdf("fitVSobsREML.pdf", width = 11, height=4.3);
xyplot(values ~ xW | factor(trainSet$Locations, label=c("Rundu","Mukwe")), group = ind, data = ress1, key = list(text = list(c
     ("Observed", "Predicted")),points = list(pch = 8:9, col = c("black","gray87"))),pch = 8:9, col = c("black","gray87"),
     panel = function(...)panel.xyplot(lty = 6,grid=TRUE,...), ylim = 1.9:9, xlim = 0:2900 , main="Observed values vs
     predicted values", ylab="water levels value in meters", xlab="Time Index ( in weeks)") # longitudinal plot of observed
     and predicted values
```

```
dev.off()
#
pdf("levels_vs_flows_residuals_REML.pdf", width = 11, height=5)
trainSet$predicted<-resid(REML)+fitted(REML);res <- stack(data.frame(Observed = trainSet$Water.levels, Predicted =
     trainSet$predicted));ress12 <- cbind(res, xW2 = rep(trainSet$Water.flows, 2))
xyplot(values ~ xW2 | factor(trainSet$Locations, label=c("Rundu","Mukwe")),  group = ind, data = ress12, key = list(text =
     list(c("Observed values", "residuals + fitted values")),points = list(pch = 8:9, col = c("black","gray87"))),pch = 8:9,
     col = c("black","gray87"), panel = function(...)panel.xyplot(lty = 6,grid=TRUE,...), ylim = 1.9:9, main="(fitted +
     residuals)_REML model vs observed values", ylab="water levels value in meters", xlab="water flows value in meters per
     second") # longitudinal plot of observed and predicted values
dev.off()
pdf("levels_vs_flows_REML.pdf", width = 11, height=5)
trainSet$predicted<-fitted(REML);res <- stack(data.frame(Observed = trainSet$Water.levels, Predicted = trainSet$predicted));
     ress12 <- cbind(res, xW2 = rep(trainSet$Water.flows, 2))
xyplot(values ~ xW2 | factor(trainSet$Locations, label=c("Rundu","Mukwe")),  group = ind, data = ress12, key = list(text =
     list(c("Observed values", "fitted values")),points = list(pch = 8:9, col = c("black","gray87"))),pch = 8:9, col = c("
     black","gray87"), panel = function(...)panel.xyplot(lty = 6,grid=TRUE,...), ylim = 1.9:9, main="fitted values_REML model
     vs observed values", ylab="water levels value in meters", xlab="water flows value in meters per second") # longitudinal
     plot of observed and predicted values
dev.off()
#
pdf("acfREML.pdf", width = 11, height=4.3)
par(mfrow=c(1,1));pacf(resid(REML),5000 ,main="Partial autocorrelation of residuals") # PACF plots
dev.off()

pdf("residPatternREML.pdf", width = 11, height=7)
xyplot(resid(REML) ~ fitted(REML) | factor(trainSet$Time,label=c("week1", "week2", "week3", "week4", "week5", "week6","week7",
      "week8", "week9",  "week10", "week11", "week12","week13", "week14", "week15","week16", "week17", "week18","week19", "
     week20", "week21","week22", "week23", "week24","week25", "week26", "week27",  "week28", "week29", "week30","week31", "
     week32", "week33","week34", "week35", "week36", "week37", "week38", "week39","week40", "week41", "week42","week43", "
     week44", "week45" ,"week46","week47", "week48" )),ylab="residuals", xlab="predicted_REML values",main="REML full model by
      yearly weeks", panel=function(x,y){panel.xyplot(x,y); panel.loess(x,y,span=0.75, col="red"); panel.lmline(x,y,lty=2) })
      # residual pattern within and between yearly weeks
dev.off()
#
#
f1<-fitted(REML)+resid(REML);
r1<-resid(REML);
mss1<-if(summary(REML)$coefficient[1]) sum((f1 - mean(f1))^2) else sum(f1^2);
rss1<-sum(r1^2);
r.squared1 <-mss1/(mss1 + rss1);
r.squared1;
bias<-mean(f1)-mean(trainSet$Water.levels);
bias;
biasPerc<-(bias/mean(trainSet$Water.levels))*100;
biasPerc # R square and bias
mse = mean( (resid(REML))^2, na.rm = TRUE);mse
sqrt(mse)
#
z=bs(trainSet$Water.flows, knots = mean(trainSet$Water.flows));head(z);
trainSet$predicted<-summary(REML)$coefficients[1]+summary(REML)$coefficients[2]*z[,1]+summary(REML)$coefficients[3]*z[,2]+
     summary(REML)$coefficients[4]*z[,3]+summary(REML)$coefficients[5]*z[,4]+summary(REML)$coefficients[6]*trainSet$zTime+
     summary(REML)$coefficients[7]*trainSet$Seasons+summary(REML)$coefficients[8]*trainSet$zTime*trainSet$Seasons # predicted
     values from REML model coefficients

res <- stack(data.frame(Observed = trainSet$Water.levels, predicted_by_coefficients= trainSet$predicted ));
ress1 <- cbind(res, xW = rep(trainSet$TimeCode_weeks, 2));
ress12 <- cbind(res, xW2 = rep(trainSet$Water.flows, 2))

pdf("fitVSobsREML_coef.pdf", width = 11, height=4.3);
xyplot(values ~ xW | factor(trainSet$Locations, label=c("Rundu","Mukwe")) , data = ress1, key = list(text = list(c("Observed",
      "Predicted")),points = list(pch = 8:9, col = c("black","gray87"))),pch = 8:9, col = c("black","gray87"), panel =
     function(...)panel.xyplot(lty = 6,grid=TRUE,...), ylim = 1.9:9, xlim = 0:2900 , main="Observed values vs values predicted
      by REML coefficients", ylab="water levels value in meters", xlab="Time Index ( in weeks)") # longitudinal plot of
```

```
     observed and predicted values
#xyplot(values ~ xW, data = ress1, key = list(text = list(c("Observed", "Predicted")),points = list(pch = 8:9, col = c("black
     ","gray87"))),pch = 8:9, col = c("black","gray87"), panel = function(...)panel.xyplot(lty = 6,grid=TRUE,...), ylim =
     1.9:9, xlim = 0:2900 , main="Observed values vs values predicted by REML coefficients", ylab="water levels value in
     meters", xlab="Time Index ( in weeks)") # longitudinal plot of observed and predicted values
dev.off()
View(trainSet)
#xyplot(values ~ xW2 |  factor(trainSet$Locations, label=c("Rundu","Mukwe")), data = ress12,  group = ind,auto.key = TRUE,main
     ="predicted_REML vs observed values", ylab="water levels value", xlab="water flows");
#
#
#set.seed(10112014)
strt<-Sys.time()
crossval_list<-list()
for ( iter in 1:7)
{
  splitdf1 <- function(dataframe, seed=NULL) {
    if (!is.null(seed)) set.seed(seed)
    index <- 1:nrow(dataframe)
    trainindex <- sample(index, trunc(length(index)*iter/7))
    trainset <- dataframe[trainindex, ]
    testSet <- dataframe[-trainindex, ]
    list(trainset=trainset,testSet=testSet)
  } # cross validated sample
  splits1 <- splitdf1(NewMScData, seed=10112014) # train and test set sample
  trainSet <- splits1$trainset;#trainSet<- orderBy(~TimeCode_weeks, trainSet);
  trainSetLoc1<-trainSet[trainSet$Locations=="1", ];trainSetLoc1<- orderBy(~TimeCode_weeks, trainSetLoc1);trainSetLoc2<-
       trainSet[trainSet$Locations=="2", ];trainSetLoc2<- orderBy(~TimeCode_weeks, trainSetLoc2);
  trainSet<-rbind(trainSetLoc1,trainSetLoc2) # Ordering the train set data based on time and location
  REML<-lmer(Water.levels~ bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons+(Water.flows|id),data=trainSet)
  sample<- (iter/7)*5568
  coefs<-Anova(REML)[,3]
  f1<-fitted(REML)+resid(REML);
  r1<-resid(REML);
  mss1<-if(summary(REML)$coefficient[1]) sum((f1 - mean(f1))^2) else sum(f1^2);
  rss1<-sum(r1^2);
  r.squared1 <-mss1/(mss1 + rss1);
  r.squared1;
  ifelse(iter==1,Sample<-sample,Sample<-rbind(Sample, sample))
  ifelse(iter==1,MSE<-mean(residuals(REML)^2),MSE<-rbind(MSE, mean(residuals(REML)^2)))
  ifelse(iter==1,AIC<-AIC(REML),AIC<-rbind(AIC,AIC(REML)))
  ifelse(iter==1,BIC<-BIC(REML),BIC<-rbind(BIC,BIC(REML)))
  ifelse(iter==1,RSQR<-r.squared1,RSQR<-rbind(RSQR, r.squared1))
  ifelse(iter==1,pvalues<-coefs,pvalues<-rbind(pvalues, coefs))
  crossval_list[[iter]]<-trainSet
}
Sys.time()-strt;beep(8);
pval_sample<-data.frame(Sample);pval_sample<-cbind(pval_sample, pvalues);Datasample<-data.frame(AIC);Datasample$MSE<-MSE;
     Datasample$RSQR<-RSQR; Datasample$Sample<-Sample;Datasample$BIC<-BIC
View(pval_sample)
View(Datasample)
#
pdf("AIC_MSE_rsqr.pdf", width = 13, height=7)
par(mfrow=c(1,4))
range(Datasample$AIC);range(Datasample$BIC);yrange<-c(1271,1981); xrange<-range(0,6000)
plot(xrange, yrange, type="n", xlab="Sample size", ylab="information criterion", main="AIC (BIC) vs sample size")
lines(Datasample$Sample, Datasample$AIC, col="grey", type="o", pch=21)
lines(Datasample$Sample, Datasample$BIC, col="Black", type="o", pch=18)
# add a legend
legend(10,1970, c("AIC", "BIC"), lty=c(1, 1, 1, 1), cex=0.5, bty="n", pch=c(11, 8, 21, 18), lwd=c(2.5,2.5), col=c("grey", "
     black"))
plot(Datasample$Sample, Datasample$MSE, type="l", pch=11, xlab="Sample size", ylab="MSE", xlim = range(0:6000), main="
     Relationship between MSE and sample size")
#points(4454.4, 0.0999, pch=24, bg="black", cex = 2)
```

```
#abline(h=0.0951, v=4454.4, lwd=2, lty=2)
plot(Datasample$Sample, Datasample$RSQR, type="l", pch=11, xlab="Sample size", ylab="r.square", xlim = range(0:6000), main="
     Relationship between r.square and sample size")
#points(4454.4, 0.93, pch=24, bg="black", cex = 2)
#
plot(pval_sample[,1], pval_sample[,2], type="o",pch=18, xlab="Sample size", ylab="p-values (of water flow bins-splines(BS))",
     main="REML p-values sample size relation")
dev.off()
#
```

Codes for Pseudo-Monte-Carlo simulation of data sets

```
setwd("/Users/unandapo/OneDrive/Master of science report")
set.seed(10112014) # fixing the random number generator for replications of results
REML<-lmer(Water.levels~ bs(Water.flows, knots=mean(Water.flows))+zTime*Seasons+(Water.flows|id),data=trainSet)
strt<-Sys.time()
SimData_List<-list()
trainSet<-trainSet[, !(colnames(trainSet) %in% c("X","Year_1_69","weeks", "m"))]
Rundu<-trainSet[trainSet$Locations=="1", ]; Mukwe<-trainSet[trainSet$Locations=="2", ]
for ( iter in 1:10000) # replicate loop results 10 000 times   replace(232, 234, 230,)
{
  R_Jan<-as.matrix(Rundu[Rundu$Month=="1", ]);R_Feb<-as.matrix(Rundu[Rundu$Month=="2", ]);R_Mar<-as.matrix(Rundu[Rundu$Month
       =="3", ]);R_Apr<-as.matrix(Rundu[Rundu$Month=="4", ]);R_May<-as.matrix(Rundu[Rundu$Month=="5", ]);R_Jun<-as.matrix(
       Rundu[Rundu$Month=="6", ]);R_Jul<-as.matrix(Rundu[Rundu$Month=="7", ]);R_Aug<-as.matrix(Rundu[Rundu$Month=="8", ]);
       R_Sep<-as.matrix(Rundu[Rundu$Month=="9", ]);R_Oct<-as.matrix(Rundu[Rundu$Month=="10", ]);R_Nov<-as.matrix(Rundu[
       Rundu$Month=="11", ]);R_Dec<-as.matrix(Rundu[Rundu$Month=="12", ])
  M_Jan<-as.matrix(Mukwe[Mukwe$Month=="1", ]);M_Feb<-as.matrix(Mukwe[Mukwe$Month=="2", ]);M_Mar<-as.matrix(Mukwe[Mukwe$Month
       =="3", ]);M_Apr<-as.matrix(Mukwe[Mukwe$Month=="4", ]);M_May<-as.matrix(Mukwe[Mukwe$Month=="5", ]);M_Jun<-as.matrix(
       Mukwe[Mukwe$Month=="6", ]);M_Jul<-as.matrix(Mukwe[Mukwe$Month=="7", ]);M_Aug<-as.matrix(Mukwe[Mukwe$Month=="8", ]);
       M_Sep<-as.matrix(Mukwe[Mukwe$Month=="9", ]);M_Oct<-as.matrix(Mukwe[Mukwe$Month=="10", ]);M_Nov<-as.matrix(Mukwe[
       Mukwe$Month=="11", ]);M_Dec<-as.matrix(Mukwe[Mukwe$Month=="12", ])
  WR_flows1 <- matrix(rburr(dim(R_Jan)[1], shape1=0.94054, shape2 = 2.3768, scale=97.671));WR_flows2 <- matrix(rgamma(dim(
       R_Feb)[1], shape = 1.4768, scale=133.36));WR_flows3 <- matrix(rgamma(dim(R_Mar)[1], shape = 4.2004, scale=84.34));
       WR_flows4 <- matrix(rgamma(dim(R_Apr)[1], shape = 5.3413, scale=76.036));WR_flows5 <- matrix(rburr(dim(R_May)[1],
       shape1=4.2679, shape2 = 3.5446, scale=102.72));DR_flows6 <- matrix(rgamma(dim(R_Jun)[1], shape = 4.3846, scale=27.582))
       ;DR_flows7 <- matrix(rburr(dim(R_Jul)[1], shape1=1.4013, shape2 = 2.979, scale=100.72)); DR_flows8 <- matrix(rburr(dim(
       R_Aug)[1], shape1=1.6748, shape2 = 4.2757, scale=87.648));DR_flows9 <- matrix(rburr(dim(R_Sep)[1], shape1=4.2679,
       shape2 = 3.5446, scale=102.72));DR_flows10 <- matrix(rgamma(dim(R_Oct)[1], shape = 8.444, scale=9.2459)); DR_flows11 <-
        matrix(rburr(dim(R_Nov)[1], shape1=0.98118, shape2 = 4.2158, scale=37.566));  WR_flows12 <- matrix(rburr(dim(R_Dec)
       [1], shape1=0.8493, shape2 = 3.02, scale=62.63))
  R_Jan<-cbind(R_Jan,WR_flows1);R_Feb<-cbind(R_Feb,WR_flows2);R_Mar<-cbind(R_Mar,WR_flows3);R_Apr<-cbind(R_Apr,WR_flows4);
       R_May<-cbind(R_May,WR_flows5);R_Jun<-cbind(R_Jun,DR_flows6);R_Jul<-cbind(R_Jul,DR_flows7);R_Aug<-cbind(R_Aug,DR_flows8)
       ;R_Sep<-cbind(R_Sep,DR_flows9);R_Oct<-cbind(R_Oct,DR_flows10);R_Nov<-cbind(R_Nov,DR_flows11); R_Dec<-cbind(R_Dec,
       WR_flows12)
  Rundu.Sim<-rbind(R_Jan, R_Feb, R_Mar, R_Apr, R_May, R_Jun, R_Jul, R_Aug, R_Sep, R_Oct, R_Nov, R_Dec)
  Rundu.Sim<-as.data.frame(Rundu.Sim) %>% arrange(TimeCode_weeks) # Simulated flows at Rundu location
  #
  WM_flows1 <- matrix(rburr(dim(M_Jan)[1], shape1=0.66227, shape2 = 5.7573, scale=219.52));WM_flows2 <- matrix(rburr(dim(M_Feb
       )[1], shape1=0.77458, shape2 = 3.972, scale=243.61));WM_flows3 <- matrix(rgamma(dim(M_Mar)[1], shape = 7.4244, scale
       =66.504));WM_flows4 <- matrix(rgamma(dim(M_Apr)[1], shape = 8.7731, scale=66.008));WM_flows5 <- matrix(rburr(dim(M_May)
       [1], shape1=1.8336, shape2 = 2.4782, scale=457.39));DM_flows6 <- matrix(rgamma(dim(M_Jun)[1], shape = 7.0593, scale
       =38.997));DM_flows7 <- matrix(rgamma(dim(M_Jul)[1], shape = 10.887, scale=20.982));DM_flows8 <- matrix(rburr(dim(M_Aug)
       [1], shape1=0.72656, shape2 = 8.0721, scale=181.23));DM_flows9 <- matrix(rburr(dim(M_Sep)[1], shape1=0.59466, shape2 =
       9.2190, scale=157.48));DM_flows10 <- matrix(rburr(dim(M_Oct)[1], shape1=0.61422, shape2 = 8.0028, scale=139.67));
       DM_flows11 <- matrix(rburr(dim(M_Nov)[1], shape1=0.87037, shape2 = 6.0401, scale=147.47));WM_flows12 <- matrix(rburr(
       dim(M_Dec)[1], shape1=0.73655, shape2 = 6.8639, scale=167.19))
  M_Jan<-cbind(M_Jan,WM_flows1);M_Feb<-cbind(M_Feb,WM_flows2);M_Mar<-cbind(M_Mar,WM_flows3);M_Apr<-cbind(M_Apr,WM_flows4);
       M_May<-cbind(M_May,WM_flows5);M_Jun<-cbind(M_Jun,DM_flows6);M_Jul<-cbind(M_Jul,DM_flows7);M_Aug<-cbind(M_Aug,DM_flows8)
       ;M_Sep<-cbind(M_Sep,DM_flows9);M_Oct<-cbind(M_Oct,DM_flows10);M_Nov<-cbind(M_Nov,DM_flows11);M_Dec<-cbind(M_Dec,
       WM_flows12)
  Mukwe.Sim<-rbind(M_Jan, M_Feb, M_Mar, M_Apr, M_May, M_Jun, M_Jul, M_Aug, M_Sep, M_Oct, M_Nov, M_Dec)
  Mukwe.Sim<-as.data.frame(Mukwe.Sim) %>% arrange(TimeCode_weeks) # Simulated flows at Mukwe location
  #
  Sim_Data<-rbind(Rundu.Sim, Mukwe.Sim);
  names(Sim_Data) = c("Time", "Year","Months", "Seasons","Water.levels", "Water.flows", "Locations", "TimeCode_weeks", "id",
       "zTime","Sim_flows")
  z<-data.frame(bs(Sim_Data$Sim_flows, knots = mean(Sim_Data$Sim_flows)))
  #
  Sim_Data$Sim_levels<-summary(REML)$coefficients[1]+summary(REML)$coefficients[2]*z$X1 + summary(REML)$coefficients[3]*z$X2 +
        summary(REML)$coefficients[4]*z$X3 + summary(REML)$coefficients[5]*z$X4 + summary(REML)$coefficients[6]*Sim_Data$zTime
        +summary(REML)$coefficients[7]*Sim_Data$Seasons + summary(REML)$coefficients[8]*Sim_Data$Locations*Sim_Data$Seasons#
       Simulated water levels at Rundu and mukwe location
```

```
  #
  SimData_List[[iter]]<-Sim_Data
}
#print(head(SimData_List[[2]]));#for(i in 1:10)print(head(SimData_List[[i]]))
Sys.time()-strt;beep(8) # produce noise after loop complete running
#
length(SimData_List)
#saveRDS(SimData_List , "SimData_List.rds")
#SimData_List<-readRDS("SimData_List.rds")
#
#
```

Replicating the REML model results 10 000 times

```
sort( sapply(ls(), function(x){object.size(get(x))})) # check whats in R memory
rm(list=(ls()[ls()!=""])) # clear all datasets in R memory excluding the one in quotation
#
setwd("/Users/unandapo/OneDrive/Master of science report");getwd()
SimData_List<-readRDS("SimData_List.rds")
#

str<-Sys.time(); for(i in 1:10000){
  REML<-lmer(Sim_levels~bs(Sim_flows, knots=mean(Sim_flows)) + zTime*Seasons + (Sim_flows|id),as.data.frame(SimData_List[i]))
  #SimData_List[[i]]$fitted_sim<-fitted(REML)
  #SimData_List[[i]]$resid_sim<-resid(REML)
  mss<-if(summary(REML)$coefficients[1]) sum((fitted(REML)-mean(fitted(REML)))^2) else sum(fitted(REML)^2)
  rss<-sum(residuals(REML)^2)
  r.sqr<-mss/(mss+rss) #coefficients of determinations
  ifelse(i==1,coefficients<-summary(REML)$coefficients[,1], coefficients<-rbind(coefficients, summary(REML)$coefficients[,1]))
  ifelse(i==1,MSE<-mean(residuals(REML)^2),MSE<-rbind(MSE, mean(residuals(REML)^2)))
  ifelse(i==1,RMSE<-sqrt(mean(residuals(REML)^2)),RMSE<-rbind(RMSE,sqrt(mean(residuals(REML)^2))))
  ifelse(i==1,AIC<-AIC(REML),AIC<-rbind(AIC,AIC(REML)))
  ifelse(i==1,RSQR<-r.sqr,RSQR<-rbind(RSQR, r.sqr))
  ifelse(i==1,bias<-mean(SimData_List[[i]]$Sim_levels)-mean(fitted(REML)), bias<-rbind(bias, mean(SimData_List[[i]]$Sim_levels
      )-mean(fitted(REML))))
  ifelse(i==1,table<-stargazer(REML, type="text"),table<-rbind(table, stargazer(REML, type="text")))
   };print(Sys.time()-str) ;beep(8)
#dim(SimData_List);length(SimData_List)
#for(i in 1:10)print(head(SimData_List[[i]]))
#View(SimData_List)
#saveRDS(SimData_List, "SimData_List_REML.rds")
#SimData_List<-readRDS("SimData_List_REML.rds")
#
SimData_coef<-as.data.frame(coefficients);SimData_coef$AIC<-AIC;SimData_coef$MSE<-MSE;SimData_coef$RMSE<-RMSE;SimData_coef$r.
     sqr<-RSQR;SimData_coef$bias<-bias
names(SimData_coef)<-c("const", "BS1","BS2","BS3","BS4","zTime","Seasons","zTime*Season","AIC","MSE","RSME","rsqrt","bias")
#saveRDS(SimData_coef, "SimData_coef.rds")
apply(SimData_coef, 2, mean); #apply(coefficients ,2, mean );apply(MSE ,2, mean );apply(AIC ,2, mean );apply(RMSE ,2, mean );
     apply(RSQR ,2, mean );apply(bias ,2, mean )# calculating column mean.
View(SimData_coef)
dim(SimData_coef)
#
SimData_coef<-readRDS("SimData_coef.rds")
z=bs(testSet$Water.flows, knots = mean(testSet$Water.flows));head(z);
testSet$predicted<-apply(SimData_coef ,2, mean )[1]+apply(SimData_coef ,2, mean )[2]*z[,1]+apply(SimData_coef ,2, mean )[3]*z
     [,2]+apply(SimData_coef ,2, mean )[4]*z[,3]+apply(SimData_coef ,2, mean )[5]*z[,4]+apply(SimData_coef ,2, mean )[6]*
     testSet$zTime+apply(SimData_coef ,2, mean )[7]*testSet$Seasons  # predicted values from REML model coefficients
View(testSet)
head(testSet)
```

```
#calculate mean sqr of error
res <- stack(data.frame(Observed = testSet$Water.levels, predicted_by_coefficients= testSet$predicted ));
ress1 <- cbind(res, xW = rep(testSet$TimeCode_weeks, 2));
ress12 <- cbind(res, xW2 = rep(testSet$Water.flows, 2))

pdf("pre_vs_obs_SIMULATED.pdf", width = 11, height=8.5);
xyplot(values ~ xW | factor(testSet$Locations, label=c("Rundu","Mukwe")), layout=c(1,2), data = ress1, key = list(text = list(
     c("Observed", "Predicted by coefficients")),points = list(pch = 8:9, col = c("black","gray87"))),pch = 8:9, col = c("
     black","gray87"), panel = function(...)panel.xyplot(lty = 6,grid=TRUE,...), ylim = 1.9:9, xlim = 0:2900 , main="Observed
     values vs values predicted by simulated REML coefficients", ylab="water levels value in meters", xlab="Time Index ( in
     weeks)") # longitudinal plot of observed and predicted values
dev.off()
#
SimData_coef<-readRDS("SimData_coef_1.rds")
bias<- apply(SimData_coef ,2, mean )[1]-summary(REML)$coefficient[1];biasPerc<-(bias/summary(REML)$coefficient[1])*100;   bias
     ;biasPerc
bias<- apply(SimData_coef ,2, mean )[2]-summary(REML)$coefficient[2];biasPerc<-(bias/summary(REML)$coefficient[2])*100;   bias
     ;biasPerc
bias<- apply(SimData_coef ,2, mean )[3]-summary(REML)$coefficient[3];biasPerc<-(bias/summary(REML)$coefficient[3])*100;   bias
     ;biasPerc
bias<- apply(SimData_coef ,2, mean )[4]-summary(REML)$coefficient[4];biasPerc<-(bias/summary(REML)$coefficient[4])*100;   bias
     ;biasPerc
bias<- apply(SimData_coef ,2, mean )[5]-summary(REML)$coefficient[5];biasPerc<-(bias/summary(REML)$coefficient[5])*100;   bias
     ;biasPerc
bias<- apply(SimData_coef ,2, mean )[6]-summary(REML)$coefficient[6];biasPerc<-(bias/summary(REML)$coefficient[6])*100;   bias
     ;biasPerc
bias<- apply(SimData_coef ,2, mean )[7]-summary(REML)$coefficient[7];biasPerc<-(bias/summary(REML)$coefficient[7])*100;   bias
     ;biasPerc
bias<- apply(SimData_coef ,2, mean )[8]-summary(REML)$coefficient[8];biasPerc<-(bias/summary(REML)$coefficient[8])*100;   bias
     ;biasPerc
bias<- 1-0.975; biasPerc<-(bias/0.975)*100;   bias;biasPerc
#
wilcox.test(testSet$Water.levels, testSet$predicted)
mean(testSet$Water.levels); mean(testSet$predicted)
par(mfrow=c(2,1));hist(testSet$Water.levels);hist(testSet$predicted)
#
AIC_val<-c(-107099.8, -72232.49, -27213.05, -19298.07)
BIC_val<-c( -107055.9, -72161.41, -27152.87, -19241.84)
sample<-c(4454, 2761, 1113, 801)
mydata<-data.frame(AIC_val, BIC_val, sample)

pdf("AIC_Sample.pdf", width = 11, height=6)
xrange<-c(0,6000); yrange<-c(-108000, 1850)
plot(xrange, yrange, type="n", xlab="Sample size", ylab="information criterion", main="AIC (BIC) vs sample size")
lines(mydata$sample, mydata$AIC_val, col="grey", type="o", pch=21)
lines(mydata$sample, mydata$BIC_val, col="Black", type="o", pch=18)
points(4454, 1837, pch=24, bg="grey", cex = 2)
names=c("simulated AIC/BIC", "simulated AIC/BIC", "simulated AIC/BIC", "simulated AIC/BIC" ); textxy(mydata$sample,
     mydata$AIC_val, labs=names)
name_real=c("AIC from the non-simulated model" ); textxy(4454, 1835, labs=name_real)
legend(10,1970, c("AIC", "BIC"), lty=c(1, 1, 1, 1), cex=0.5, bty="n", pch=c(11, 8, 21, 18), lwd=c(2.5,2.5), col=c("grey", "
     black"))
abline(h=1835, v=4454.4, lwd=2, lty=3)
dev.off()
#
```