Evaluating the use of neural networks to predict river flow gauge values

by
Wesley Michael Walford
Student Number: 10661990

Submitted in partial fulfilment of the requirements for the degree
MSc Geoinformatics

In the Faculty of Natural & Agricultural Sciences

University of Pretoria
Pretoria

(22 June 2017)

Supervisor: Professor Serena M Coetzee
Co-supervisor: Doctor Terence Van Zyl

Declaration

I, Wesley Michael Walford declare that the dissertation, which I hereby submit for the degree MSc Geoinformatics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature:

Date: 22 June 2017

Summary

Without improved water management the global population could be facing serious water shortages. River flow discharge rates are one factor that could contribute to improving water management, being able to predict a forecasted river flow value would provide support in the management of water resources.

This research investigates the use of an artificial neural network (ANN) to create a model that predicts river flow gauge values. The Driel Barrage monitoring station on the Thukela river in South Africa was used as a case study. The research makes use of data from the Department of Water and Sanitation (DWS) and weather forecast data from the European Center For Medium-Range Forecasts (ECMWF) to train the predictive model.

An evaluation of the ANN model identified that the model is highly sensitive to selected weather parameters and is sensitive to the initial weights used in the ANN. These were overcome using an ANN ensemble and selective scenarios to identify the best weather parameters to use as input into the ANN model. Five weather parameters and a correlation coefficient cut-off value produced the most accurate prediction by the ANN.

The research found that ANNs can be used for predicting river flow gauge values but to improve the results a greater ensemble, additional data and different ANN structures may create a better performing model. For the ANN model to be used in practice the research needs to be extended to evaluate the whole catchment area and a range of rivers in South Africa.

Keywords: River Flow gauge value, Artificial Neural Network, South Africa, Thukela
Degree: MSc Geoinformatics
University: University of Pretoria
Department: Faculty of Natural & Agricultural Sciences
Supervisor: Professor Serena M Coetzee
Co-supervisor: Doctor Terence van Zyl

Acknowledgements

This dissertation would not have been possible without support and guidance; I would like to express my gratitude to my supervisors Prof. Serena Coetzee and Dr. Terence van Zyl for their useful comments, remarks and engagement through the learning process of this master's dissertation. Further more I would like to thank them for their patience and understanding when planned deadlines were missed, and for guiding me to the completion of this dissertation.

A masters takes a lot more time and effort than I expected and I would like to acknowledge this for other students starting the process. It never seems as if you have read enough, I still feel I need to read more on various topics and that there is a lot more to understand. I feel a more structured start and a well defined approach would have made the process and experience a lot better. Prof. Serena Coetzee has assured me this is normal and there are too many unknown variables when one starts off on this journey. One area of importance I would suggest for any student starting this journey is to use the tools available for managing references and editing as they are extremely helpful and simplify the process.

I would like to thank my loved ones, who have supported me throughout this entire process. Your encouragement and motivation will always be remembered. I will be grateful forever for your love. I would like to specifically thank my wife, Melanie Walford, for her encouragement, understanding and support. I will forever be thankful for the time you sacrificed during this period and your loving support to the end of this journey.

Table of Contents

ABBREVIATIONS

- **AN** – Artificial Neuron
- **ANN** – Artificial Neural Network
- **DAFF** - Department of Agriculture, Forestry and Fisheries
- **DWS** – Department of Water and Sanitation
- **CF**- Climate and Forecast
- **ECMWF** – European Center for Medium-Range Weather Forecasts
- **GD** – Gradient Descent
- **RMSE** – Root mean square error
- **WMA** – Water management areas

LIST OF FIGURES

LIST OF TABLES

LIST OF GRAPHS

LIST OF EQUATIONS

1. INTRODUCTION

1.1. Background and context

Water is a valuable resource that is often taken for granted. It is estimated that by 2025 around 1.8 billion people across the globe will be living with extreme water scarcity (Blumenfeld et al. 2009). The main reasons for this are poor water management and an increasing population size. According to Reid (Reid 2014), the global population has increased from 2.5 billion in 1950 to 6 billion in 2000, with the consumption of natural resources soaring more than six-fold. To further compound the issue, only a very small amount (less than 1%) of all water on the planet is fresh water and accessible to humans. The ever increasing population on Earth has access to a very small amount of water. Proper water management is crucial as water is a valuable but limited resource. The problem is that accessible fresh water, such as rivers, do not naturally sustain the human population or else there would be no need for dams and reservoirs (McCartney et al. 2013). There is an increasing need to manage this accessible fresh water to ensure continued access for the human population.

There are many different approaches to managing water availability, and water flow rates play a central role in these approaches. The South African Department of Agriculture, Forestry and Fisheries (DAFF) (www.daff.gov.za) formerly known as the South African Department of Water and Sanitation (DWS) (www.dwa.gov.za), acknowledge this as a part of their integrated planning and management approach. The management approach is referred to as Integrated Water Resource Management (IWRM)(Department of Water and Forestry 2004). One of the processes of the IWRM is to find models that represent the flow of river systems. A study by O'Keeffe (1989) alluded to the fact that river flow drastically affects farming and that farmers required access to a more constant water source, but also highlighted that management of water flow can be used to restrict erosion. The same research also highlighted pest problems associated with changing river flow and that water management needs to be done in such a way as to simulate natural flow variations to avoid colonization of pest species.

River flow predictions allow better planning of water management as they provide a means for future preparation. Traditional river flow values were manually measured on site in the river. These values were then extended to a concept called flow duration curves (McCartney et al. 2013). Newer approaches include intelligent systems (Steinfeld et al. 2015; Aichouri et al. 2015) and satellite altimetry data (Tarpanelli et al. 2013) to predict and simulate river flow. This research will be looking at using an intelligent system approach by using an artificial neural network (ANN) with weather parameter data to ascertain how accurate ANNs can predict river flow. River flow discharge rates can assist with flood warning systems to mitigate loss of life and reduce damage during flooding, flooding in South Africa can be extensive (Mkamba 2013; Hill 2014; SAPA 2014). Neural networks are adaptive statistical models and a predictive ANN model making use of weather data could provide a new approach within the South African context. Neural networks are statistical tools used in fields such as psychology, statistics, physics and engineering (Kröse & van der Smagt 1991). Using neural networks could allow river flow values to be predicted in areas where it is too expensive or dangerous to physically collect data onsite and incorporates erratic changing weather.

This research studies the Upper Thukela river in South Africa (KwaZulu Natal), as this river has significant ties to water availability in both the province of KwaZulu Natal and Gauteng due to a water transfer scheme. Weather forecasting is a global activity performed by many

organizations for many different reasons. A prediction, designed and built based on weather forecasting, could provide a meaningful approach to predicting river flow discharge rates.This research looks at what weather parameters are required for such a prediction and with what accuracy the river flow gauge value predictions can be performed.

1.2. Problem statement

Proper water resource planning and flood control requires accurate predictions of river flow gauge values. South Africa needs an appropriate method to predict river flow gauge values and the method must make use of available data and provide a set of predicted river flow gauge values. If river flow gauge values cannot be predicted water management is hampered which in turn increases the probability of water stress in the future.

1.3. Objectives

I propose to research the use of ANs to predict river flow gauge values. I will evaluate the performance of an ANN with weather data sets to predict river flow gauge values. The research will provide a model that can predict river flow gauge values and then analyze the accuracy of these predictions against measured river flow gauge values.

1.4. Scope

This research investigates the use of intelligent systems in the form of ANNs to form a predictive model for river flow gauge values. The research and the model will be limited to a single river flow station in South Africa. The research will not generalize the model but will evaluate the use of ANNs to understand their benefits and limitations as a predictive model for river flow gauge values. The research will make use of existing neural network structures and libraries and does not aim to improve on these exiting methods but will evaluate their use in this specific problem area.

1.5. Hypothesis

This research will attempt to evaluate the following hypothesis:

> ***By developing an artificial neural network model based on inputs from predicted weather parameters, it will be possible to evaluate the model identifying critical aspects and improvements for using an artificial neural network to predict river flow gauge values.***

1.6. Assumptions

This research makes the assumptions that:

1. The code libraries used in the research are correct and without error based on the information provided by the library owners and their use in previous research.
2. The data downloaded for use in this research has the correct quality rating and has been provided according to the specifications by the data owner.

1.7. Significance of the research

River discharge rates are used to manage water resources across the globe and South Africa is no different. Improved predictions of river flow discharge rates can assist in ensuring water availability in the future. River flow discharge rates can also assist with flood warning systems to mitigate loss of life and reduce damage during flooding, flooding in South Africa can be extensive (Mkamba 2013; Hill 2014; SAPA 2014).

There are currently difficulties and barriers to manually monitoring river discharge rates, it can be dangerous and expensive to physically collect data onsite at the river (Tarpanelli et al. 2013). In South Africa there are also sections of rivers that flow from neighboring countries or sections of rivers that are inaccessible where river discharge rates cannot be physically gathered. There need to be approaches available where river flow discharge rates can be estimated or calculated rather than physically monitoring the river. With the erratic weather being experienced in South Africa (Dube et al. 2016), frequency and extent of floods and the water shortages being experienced there is definitely a need for a more accurate predictive model for river discharge rates (The Water Project 2016). A predictive ANN model making use of weather data could provide a new approach within the South African context. The approach takes into consideration the South African context with regards to data availability, financial constraints, climate and necessary future growth for water resource management. If water management improved there could be a direct effect on the ability to predict floods and ensure a continued good quality water supply. A more secure water supply in the future improves food security and helps stabilize industrial outputs which could in turn improve the supply of exportable products (Ejaz Qureshi et al. 2013).

1.8. Ethical considerations

This research deals directly with data sets and the use of the data in a predictive manner. The research does not include any surveys, interviews or direct respondents where confidentiality and anonymity needs to be ensured. This research will consider and be aligned with the following ethical considerations.

1.8.1. Quality and integrity

The research needs to consider the quality and integrity of the tools, code libraries, data and results. The research makes use of data from recognized and independent sources, which is then validated, reanalyzed and given quality ratings by independent sources. A sufficient data sample size will be used within the dataset available. This ensures the data used in the research is of a recorded quality and that the data is not being edited, or mistreated in such a way as to influence the outcome of the research results. The research will be using recognized code libraries and software tools that have previously been used in research. The libraries and software tools are used to minimize coding bugs and to ensure the data and research is done with recognized code to ensure the quality and integrity of the research outputs.

To ensure the quality of the research outputs and findings the research also makes use of a very careful river selection. River selection was not based on a best case scenario to ensure good results. The river selection was done based on adequate knowledge of the area and ensuring there are natural influences on the river and external influences on the river such as man-made infrastructures; for example, dams.

1.8.2. Rights of use

The research makes use of only legally obtained data, tools and code libraries. The data comes from openly available data sources that the public can access. No access contracts or agreements were required to access the data. The code libraries and the software tools are all open source and have been released under various open source licenses. The legal obligations of the licenses will be held as required and all references will be given back to the relevant providers. Ethics approval has been received from the NAS Ethics Committee at the university of Pretoria, Figure 26 (Annex) provides the NAS Ethics Committee approval letter.

1.8.3. Research and findings are independent and impartial

The research reports both negative and positive findings. There are no direct third parties that have a direct interest in the outcome of the research. The research can report any positive or negative finds impartially as the research is independent of any 3rd party obligations.

1.9. Conclusion and overview of remaining chapters

To ensure water resource availability in the future there needs to be an approach to improve water management. River flow discharge rates are important in ensuring the quality and availability of water. South Africa is no different and there are a few key rivers in South Africa that provide a large percentage of the water required by the South African population. Chapter 1 of this research has provided a background context to the situation, what this research aims to investigate and the limitations of the research. It gives insight, at a high level, to the need for this research and the rest of this dissertation will provide the details of the research. Chapters 2 and 3 provide insight and context to the research while chapter 4 provides the detailed research plan. Chapters 5, 6 and 7 are direct outputs from the research providing graphs, results and discussions around the criteria and hypothesis of the research:
By developing an artificial neural network model based on inputs from predicted weather parameters, it will be possible to evaluate the model identifying critical aspects and improvements for using an artificial neural network to predict river flow gauge values
The chapters in this research are as follows:

- Chapter 2: The Thukela River
  This chapter provides an insight into the the river studied in this research. The chapter describes information about the area in which the river flows, how the river is managed and why this river was chosen for this research.
- Chapter 3: Literature review
  This chapter outlines all the necessary background information required to do the research. The chapter provides an insight into previous research done with regards to river flow discharge rates and the suitability of neural networks for predicting river flow gauge values.
- Chapter 4: Methodology
  This chapter describes the methodology used during this research. It includes information about the selected libraries, the data used in the research, naive predictions and the ANN setup. The methodology describes a performance scorecard to record and compare criteria used to ascertain the accuracy of the ANN predictions.

- Chapter 5: Results
  In chapter 5 the results from the ANN's output are presented. The results are explained and graphed. These results are then presented in the scorecard defined in the methodology, chapter 4. The scorecard criteria drive the discussions of the various scenarios and allow comparison of results across scenarios.
- Chapter 6: Evaluation of using ANN's
  This chapter provides the main output of the research. It evaluates the use of ANN's for predicting river flow gauge values. The chapter discusses the performance and limitations for using ANNs to predict the river flow gauge values against the determined criteria in Chapter 4.
- Chapter 7: Conclusion
  Chapter 7 concludes the research by summarizing the research problem, the research approach and evaluation discussed in chapter 6 to draw a conclusion to the research. This chapter provides the bigger picture view of the research and interprets the results. The chapter makes final comments and concludes the research with drawing a conclusion against the hypothesis from Chapter 1.

2. THE THUKELA RIVER

2.1. Introduction

This research is focused on the South African context, and so the research only considered one river in South Africa that plays an important role in the water availability for the country. Chapter 1 has given a high level overview of the problem the research investigated and the scope of the research. The Thukela river in KwaZulu Natal was selected as the case study river for the research. The Thukela plays an important role in ensuring water availability in South Africa. It provides a complex but realistic case study with the river discharge rates being affected by both natural and man-made factors. Chapter 2 provides detailed information with regards to the population, climate and water related infrastructure to describe the context of the research area. The chapter is broken into various sections, firstly section 2.2 provides a description of the Thukela river and the surrounding area. The chapter will then explain why the Thukela river is appropriate for this research and why it was selected in section in 2.3. Section 2.4 describes how the Department of Water and Sanitation (DWS) manages the Thukela river. The final section of the chapter will draw some conclusions around the Thukela river and the selection of this river as the research case.

2.2. The Thukela river and catchment area

2.2.1. Overview

An increasing growth in the world population is placing severe stress on water resources around the globe and South Africa is not excluded (Wilson 2001a). With a rainfall of less than 500mm per annum, South Africa is defined as an arid country and in 2001 it was estimated that water resources in South Africa could be exhausted within 20 to 30 years. The evidence for this is already becoming visible at some local levels where water demand cannot be met.

The Thukela river is one of the major rivers and sources of water in South Africa. The Thukela river starts in the Drakensberg mountains close to Lesotho and creates a funnel shaped catchment area towards the Indian ocean near Mandini. On 1st October 1999 in the Government Notice No. 1190 , 19 water management areas (WMA) across South Africa were established and boundaries defined (Staatskoerant 2012). The Thukela WMA was defined as one of the 19 WMA's in this Government Notice. The Thukela WMA is referred to as the 'V' hydrological drainage region and consists of the entire Thukela catchment area (Staatskoerant 2012). Figure 1 provides a high-level view of the study area with relation to the Drakensberg mountains, Lesotho and Mandini. In July 2012, the minister of Water and Environmental Affairs proposed the establishment of 9 new WMA's in South Africa replacing the 19 WMA's defined in 1999 , calling for comments on the proposal (Staatskoerant 2012). The newly defined WMA's had the Thukela river and catchment area falling into a WMA referred to as the Pongola-Mtamvuna WMA along with the Pongola, Mfolozi and Umgeni rivers to mention a few. On 16 September 2016, the Minister of Water and Sanitation announced that the 9 new WMA's were established and that each WMA would have a catchment management agency (CMA) established (Staatskoerant 2016). This officially moved the Thukela river and catchment area into the Pongola-Mtamvuna WMA.

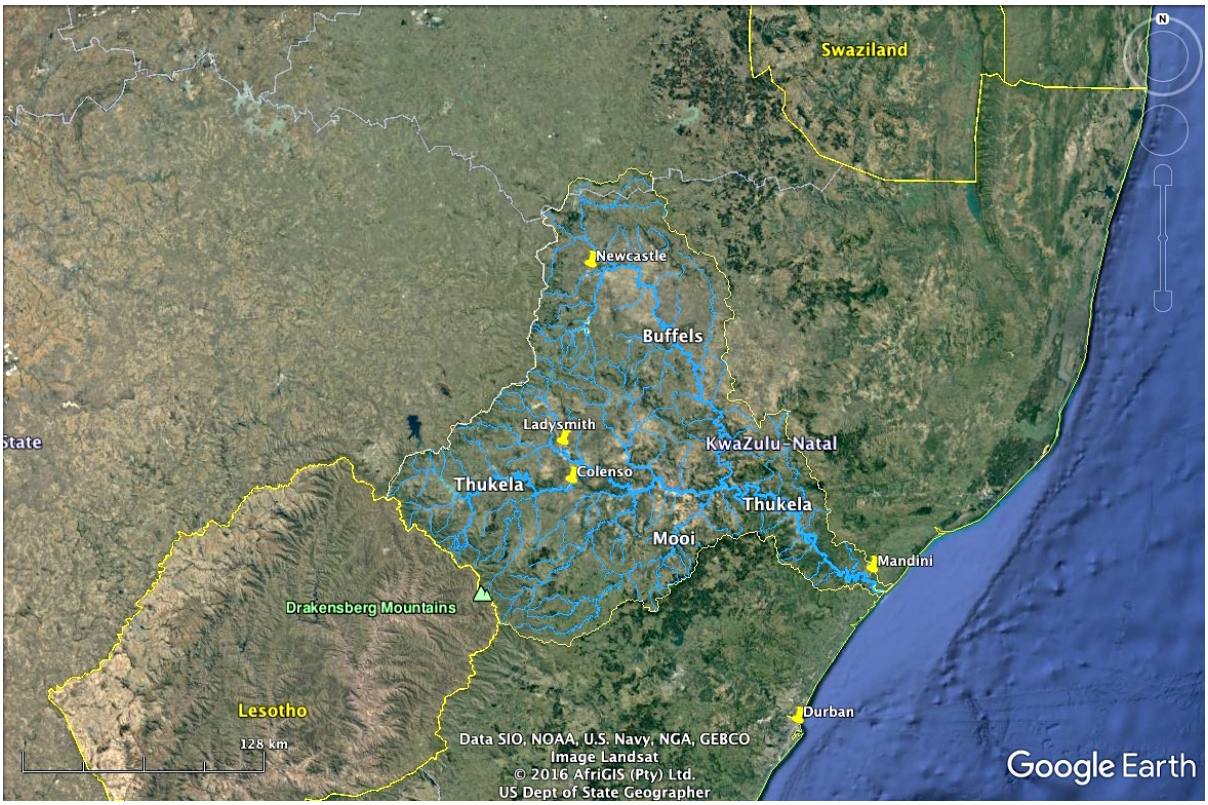


Figure 1: High-level map of the Thukela study area (Google Earth 7.1 2016)

The Thukela river has a catchment area of approximately 30,000 km^2 and has a relatively high rainfall by South African Standards (Department of Water and Forestry 2004). The upper mountainous area of the catchment receives over 1500mm of rainfall per annum while the central catchment area receives around 650mm of rainfall per annum. This high rainfall translates directly into high runoff with a mean annual runoff (MAR) estimated at 3700.42×10^6 m^3/a (Department of Water and Forestry 2002). The Thukela also has a steep topography with the river starting at approximately 3000 meters in the Drakensberg, flowing down to sea level. Figure 2 presents an example of an elevation profile for a single tributary which starts in the Drakensberg. In 26.4 km the rivers elevation drops 1854 meters. Figure 3 provides a zoomed out view of the Drakensberg with the yellow line representing the border between South Africa and Lesotho. A number of tributaries for the Thukela river start in the Drakensberg mountain with steep elevation profiles.

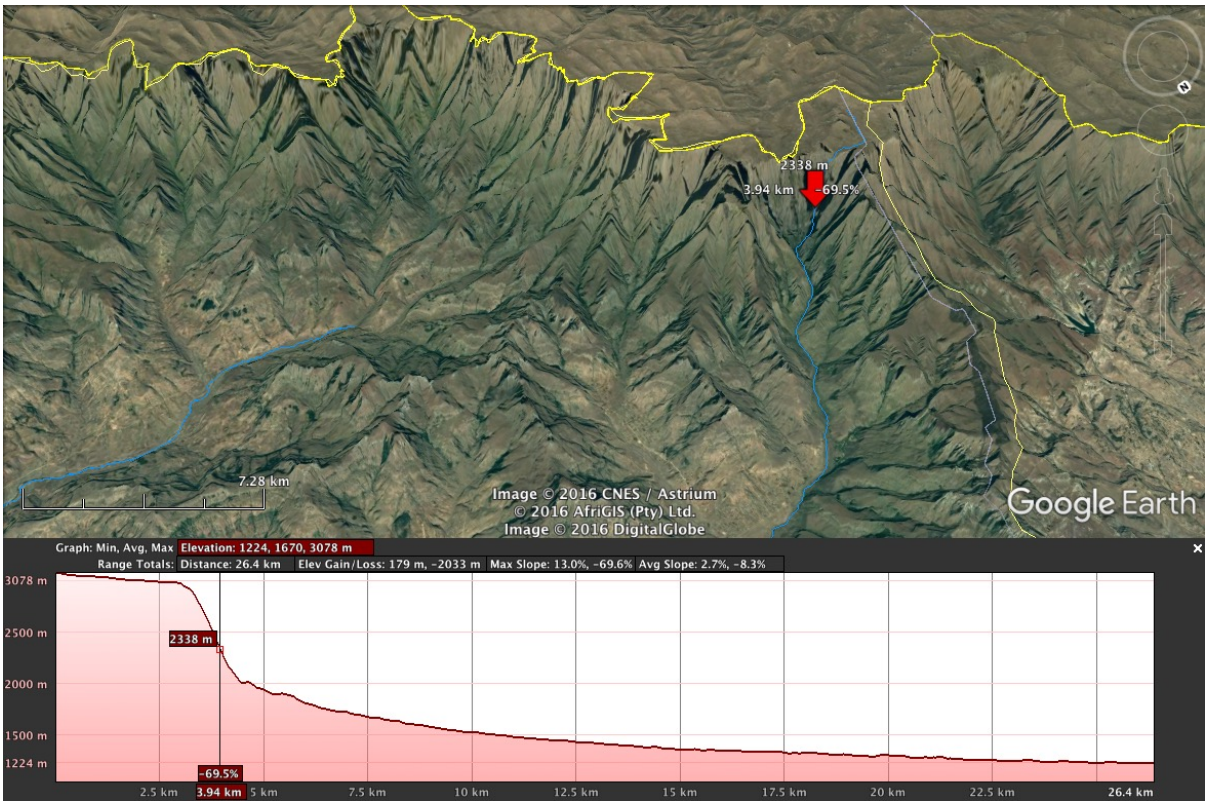


Figure 2: Example elevation profile for a single tributary (Google Earth 7.1 2016)

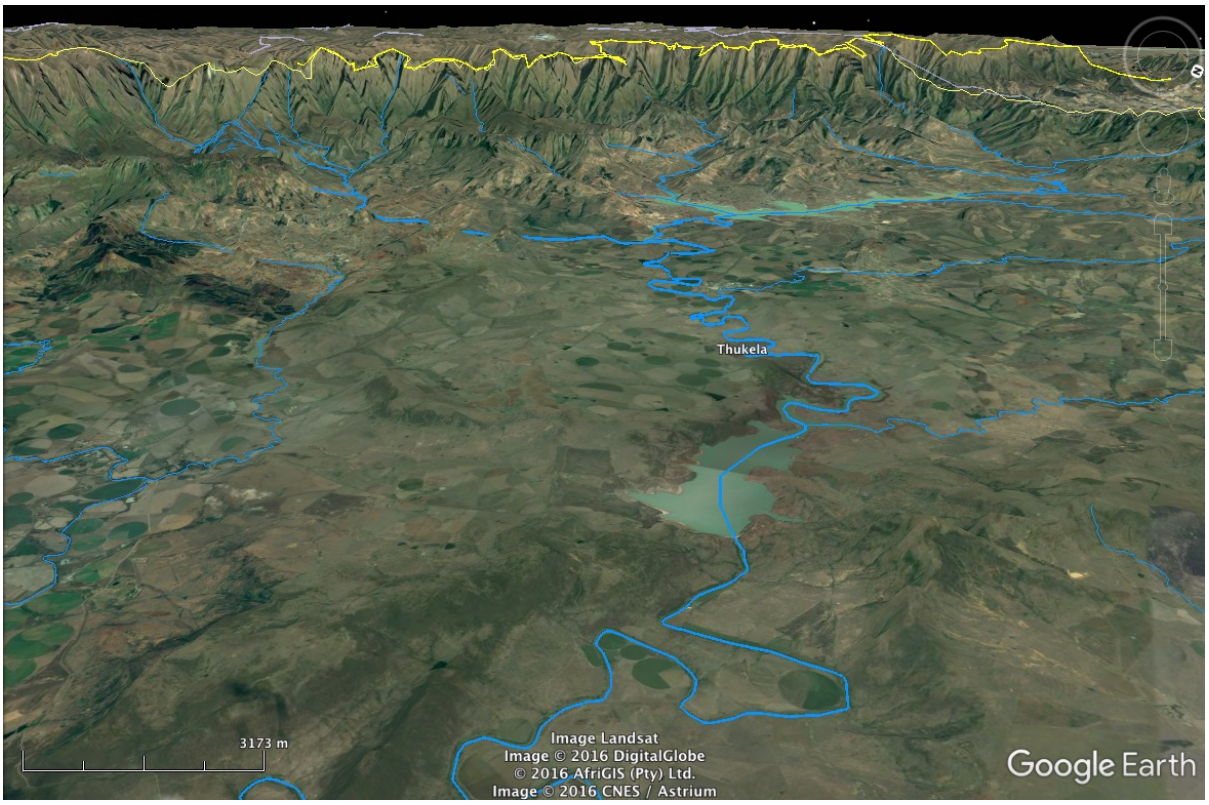


Figure 3 : View of the steep elevation due to the Drakensberg Mountains (Google Earth 7.1 2016)

The Thukela river can be divided up into key areas, which assist with analysis and reporting. Table 1 describes the key areas and the tertiary catchments (Department of

Water and Forestry 2004), while Figure 4 provides a detailed map of the key areas in the original Thukela WMA (Google Earth 7.1 2016). The same primary drainage region V is still defined in newly established Pongola-Mtamvuna WMA (Staatskoerant 2016). According to the provincial growth and development strategy for KwaZulu-Natal water has not only been identified as an important resource, but a competitive advantage for the province (Wilson 2001a).

Table 1: Key Catchment areas – Thukela WMA (Department of Water and Forestry 2004)

Key Area	**Tertiary Catchments**
Upper Thukela	Tertiary catchments V11, V12, V14 and quaternaries, V60G, H and J
Little Thukela	Tertiary catchment V13
Bushmans	Tertiary catchment V70
Sundays	Quaternary catchments V60A, B, C, D, E and F
Mooi	Tertiary catchment V20
Buffalo	Tertiary catchments V31, V32 and quaternaries, V33A and B
Lower Thukela	Tertiary catchments V40, V50 and quaternaries, V33C, D and V60K

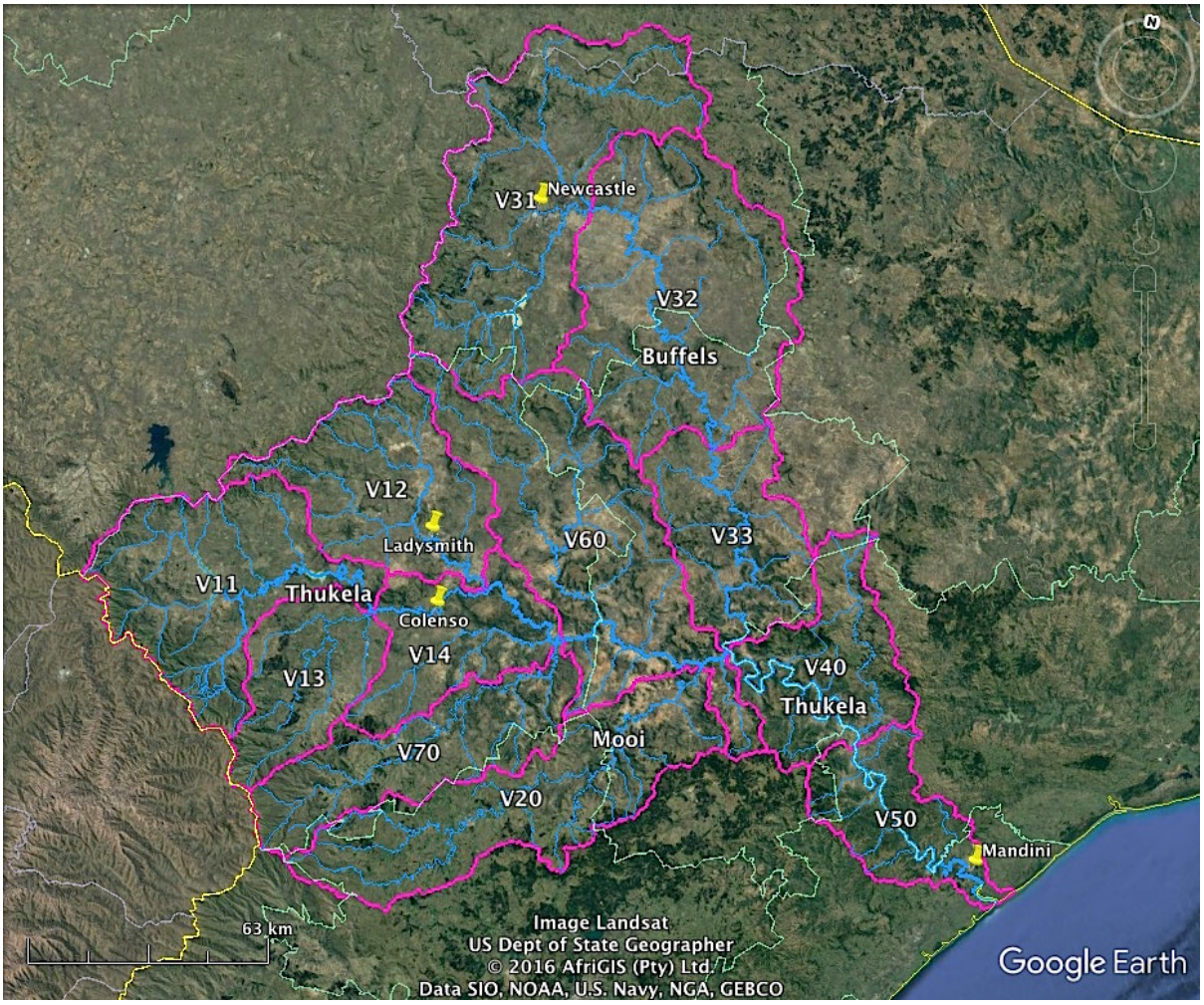


Figure 4: View of the Key Areas of the Thukela WMA (Google Earth 7.1 2016)

2.2.2. Climate

Across the Thukela catchment a variety of climates exist, from the generally colder and wetter Drakensberg mountains to the hot and dry section from Colenso towards the coast referred to as the Thukela valley (Department of Water and Forestry 2004). Along the coast the Thukela catchment experiences hot and humid weather. Snow in the Drakensberg mountain peaks is fairly common but melts quickly. In January (summer), the mean temperature for the Thukela catchment is 21,9ºC, with the mean inland temperatures ranging between 12ºC to 24ºC and the mean coastal temperatures ranging between 22ºC to 26ºC (Wilson 2001c). In July (winter), the mean temperature for the Thukela WMA is 13,3ºC, with the mean inland temperatures ranging between 10ºC to 16ºC and the mean coastal temperatures ranging between 14ºC to >16ºC (Wilson 2001c). In the mountainous areas of the Thukela catchment, the average rainfall reaches 1500mm, while the central parts of the catchment experience about 650mm per annum.

2.2.3. Geology

According to the internal strategy perspective by the Department of Water and Forestry (Department of Water and Forestry 2004) "The upper and middle Thukela River flows eastwards through a succession of sedimentary strata of the Karoo Supergroup, ranging from the younger rocks of the Triassic System (situated just below the Drakensburg volcanics) to the base of the Karoo succession in the Tugela Ferry area". The average annual sediment load at the Thukela mouth has been estimated at 5.5million tons. Figure 5 below presents a geological map of South Africa , Figure 6 provides the Legend for Figure 5 (Thiéblemont 2016).

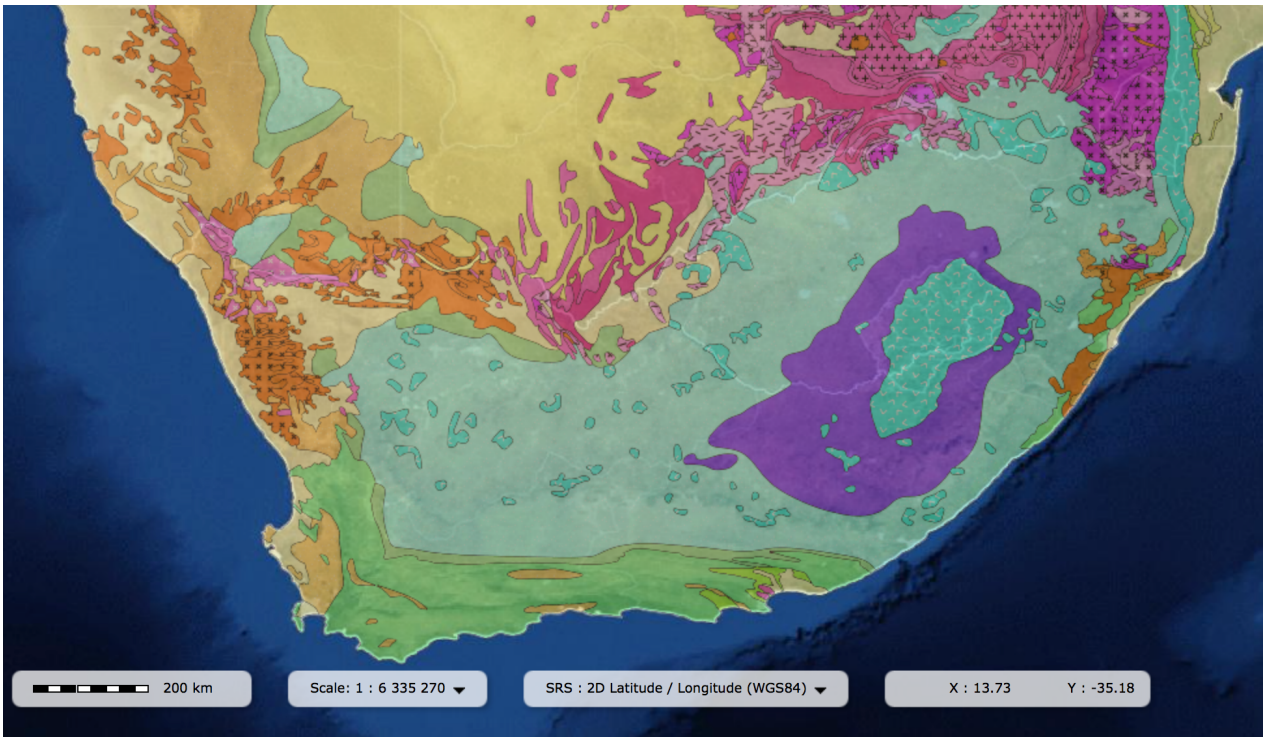


Figure 5 : Geological Map of South Africa at 1:10 M scale (Thiéblemont 2016)

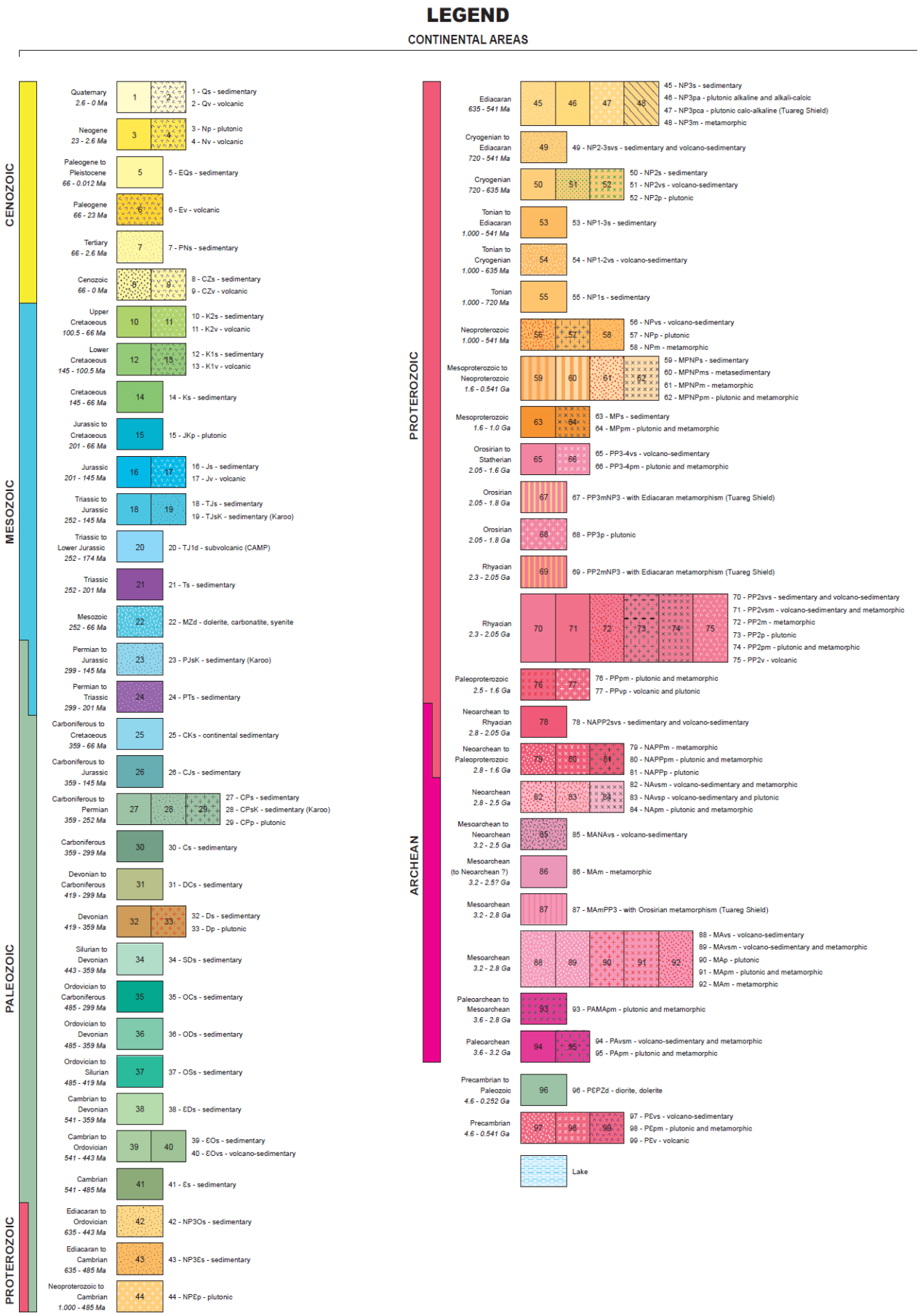


Figure 6 : Legend for Figure 5 (Thiéblemont 2016)

2.2.4. Vegetation

The vegetation in the Thukela catchment falls into similar area groupings as the geology of the catchment area. The higher lying upper parts of the catchment area are grasslands, with invading Acacia sieberiana savanna as the river moves towards Colenso (Department of Water and Forestry 2004). The vegetation then slowly becomes valley bushveld, while closer to the coast the vegetation changes to coastal grasslands and sugar cane farming. The Mooi river system, shown in Figure 1, supports pastoral farming, which has modified the river system to support the farming practices. The steeper parts of the sub-catchment maintain subtropical forests. Overgrazing together with the arid and erosive landscape has created conditions for sedimentation, erosion and loss of vegetation cover (Department of Water and Forestry 2004).

Alien plant species have added additional strain to the water availability and have added an additional water resource management problem to overcome. A large reduction in available water has been seen due to water consumption by alien plant species such as black and silver wattle (Wilson 2001a) (Department of Water and Forestry 2004). The National Working for Water Programme and local organizations have been actively tackling the problem of alien plant species (Department of Water and Forestry 2004). The efforts have been successful in controlling invasive plant species specifically in the upper Thukela river catchment area where much of the black and silver wattle infestations have been cleared.

2.2.5. Population

The Thukela WMA falls into the province of Kwazulu natal which in 2011 had a total population of 10,267,300 in 2011 (Statistics South Africa 2011b), an increase from the total population of 9,426,017 in 2001 and 8,417,021 in 1996 (Statistics South Africa 2001). In 2001, the total population of the Thukela catchment was approximately 1,56 million (Wilson 2001a). This is less than the Mvoti/Mzinkulu WMA to the South but it is still relatively high. The Thukela catchment falls across multiple municipal districts including Uthukela, Amajuba, Umzinyathi, UMgungundlovu and iLembe. The largest municipal area in the Thukela WMA is the KZN252: Newcastle municipality. This is followed closely KZN232: Ladysmith (Statistics South Africa 2011a). Both of these municipalities are shown in Figure 1. In 1996 Newcastle had a total population of 287,659; this grew by 2.9% to 332,982 by 2001, and the last census in 2011 showed a total population in Newcastle of 363 236 (Statistics South Africa 2011a). Ladysmith also experienced growth from 178,514 in 1996 to 225,459 in 2001 and 237 437 by 2011.

The Uthukela municipal district falls completely within the Thukela catchment area. The Uthukela district municipality experienced a population growth of 3.3% from 1996 to 2001, this slowed to 0.2% from 2001 to 2011 when the last census was conducted (Statistics South Africa 2011a). The average household size has shown a decline in the Uthukela district municipality, from an average household size of 5.9 in 1996 down to 4.5 in 2011. The area also experienced a growth in formal types of main dwellings. In 1996, 45.1% of the population lived in a formal main dwelling with 52.2% living in traditional dwellings and 2.4% in informal dwellings. In 2011 the formal dwellings had increased to 65.9% with traditional dwellings declining to 32.2%. The average household income has increased in the Uthukela district municipality, from R22 542 in 2001 to R56 316 in 2011 (Statistics South Africa 2011a). The district municipality has improved water resources in the area by increased the number of households receiving

piped tap water. In 1996, 34.9% of the households had piped water inside either the yard or dwelling, by 2011 this number increased to 50.2% of households.

The areas in the Thukela catchment with the lowest population are Mooi river and Weenan. Mooi river sits within the Mpofana municipality, which falls into the larger UMgungundlovu municipality district (Wilson 2001a)(Statistics South Africa 2011a). In 2011 the Mpofana municipality had a total population of 38,103, up from the 25,512 in 1996. Like the Uthukela district municipality the UMgungundlovu district municipality saw an increase in average household income, an increase in the number of households receiving piped water and a decrease in the average household size. This is the general trend across most of the Thukela catchment area.

2.2.6. Land use

The land in the Thukela catchment is mainly used for agriculture which includes beef and dairy pastures, sugar cane, vegetables, nuts and citrus fruit (Wilson 2001b). Areas are dedicated to game reserves and national parks, including the well known uKhahlamba Drakensberg Park World Heritage Site (UNESCO 2016). There are urban areas but these are generally spread out and the majority being minor urban settlements. The majority of the smaller urban areas are used to support farming. The Thukela also includes a large number of rural settlements with some being densely populated such as settlements outside the large urban areas of Ladysmith and Newcastle. In general, there is limited industry and mining near Ladysmith. Newcastle is the only major industrial center with the exception of a large paper mill near Mandini (Statistics South Africa 2010). Overall the Thukela WMA has its main activity defined as forestry, agriculture and ecotourism.

2.2.7. Organisations

Regional organisations drive change within a catchment area and have an influence on how water is used and managed within an area (Wilson 2001a). Table 2 outlines a few identified organisations within the Thukela catchment area, these regional organisations should be consulted when decisions and issues within the Thukela catchment arise. The organisations also promote participation from businesses and organisations within the catchment area. They are the forum and leaders to enforce and regulate water use within the catchment areas.

Table 2: Important regional organizations

Organization	Purpose
Kwanalu (The KwaZulu-Natal Agricultural Association)	Provides leadership to commercial farmers and agricultural organizations in KZN, specifically on key agricultural issues (Kwanula 2016). Website: http://www.kwanalu.co.za
KwaZulu-Natal Business Chambers Council	Provides a voice for business within KZN and provides a flow of information relating to business in the province. The organization also provides input to provincial government in terms of economic growth and strategies (KZN Business Chambers Council 2016). Website: http://www.kznchamber.co.za

Organization	Purpose
WESSA (The Wildlife and Environmental Society of Southern Africa)	Aims to implements environmental and conservation initiatives in Southern Africa to promote public participation in caring for the Earth. WESSA strives to shape environmental policy and is a founder member of the World Conservation Union (WESSA 2016). Website: http://www.wessa.org.za
South African Sugar Association	Responsible for administering the South African sugar industry and promotes the agricultural activities of sugarcane cultivation (SASA 2016). Website: http://www.sasa.org.za

2.2.8. Water related infrastructure

The Thukela catchment area is an important water source for South Africa and due to this there are a large number of dams in the Thukela catchment. Figure 7 shows the dams that make up the Thukela-Vaal transfer scheme. The largest Dam on the Thukela river is the Woodstock Dam shown in Figure 9, which releases its water to the Driel Barrage shown Figure 10 (Department of Water and Forestry 2004). From here water is pumped towards Sterkfontein Dam, the water is pumped from Driel Barrage into canals towards Kilburn dam and from Kilburn dam the water is pumped over the escarpment into Driekloof Dam at the upper end of Sterkfontein Dam. The Sterkfontein Dam will then release the water into the Vaal Dam which is essential to the industrial and residential areas in the Gauteng region (Ewisa n.d.). The Thukela-Vaal transfer scheme works on the operating policy that water should be pumped into the Sterkfontein Dam until it is full (Wilson 2001d; Department of Water and Forestry 2002). This can run almost uninterrupted at between approximately $15m^3/s$ and $25m^3/s$ as the dam is rarely full.

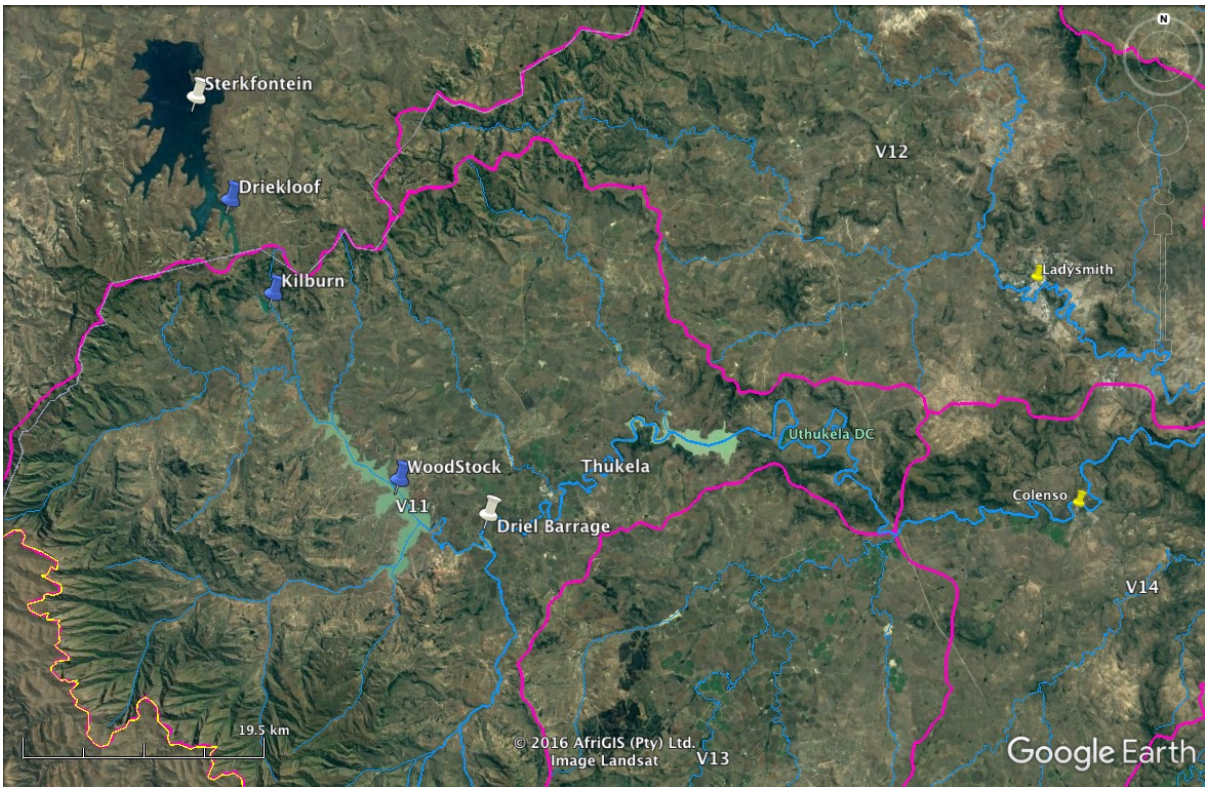


Figure 7: Dams in the Thukela-Vaal Transfer Scheme (Google Earth 7.1 2016)

Other important infrastructure in the Thukela catchment area, shown in Figure 8, includes the Zaaihoek Dam on the Slang river (which is a tributary to the Buffalo river) which is a major tributary for the Thukela river (Department of Water and Forestry 2004). The Zaaihoek Dam makes up part of the transfer scheme to transfer water to Eastern Vaal sub-system. Other major dams in the Thukela catchment include Spioenkop Dam (Figure 11), Ntshingwayo Dam (formerly Chelmsford Dam) and Wagendrift Dam.

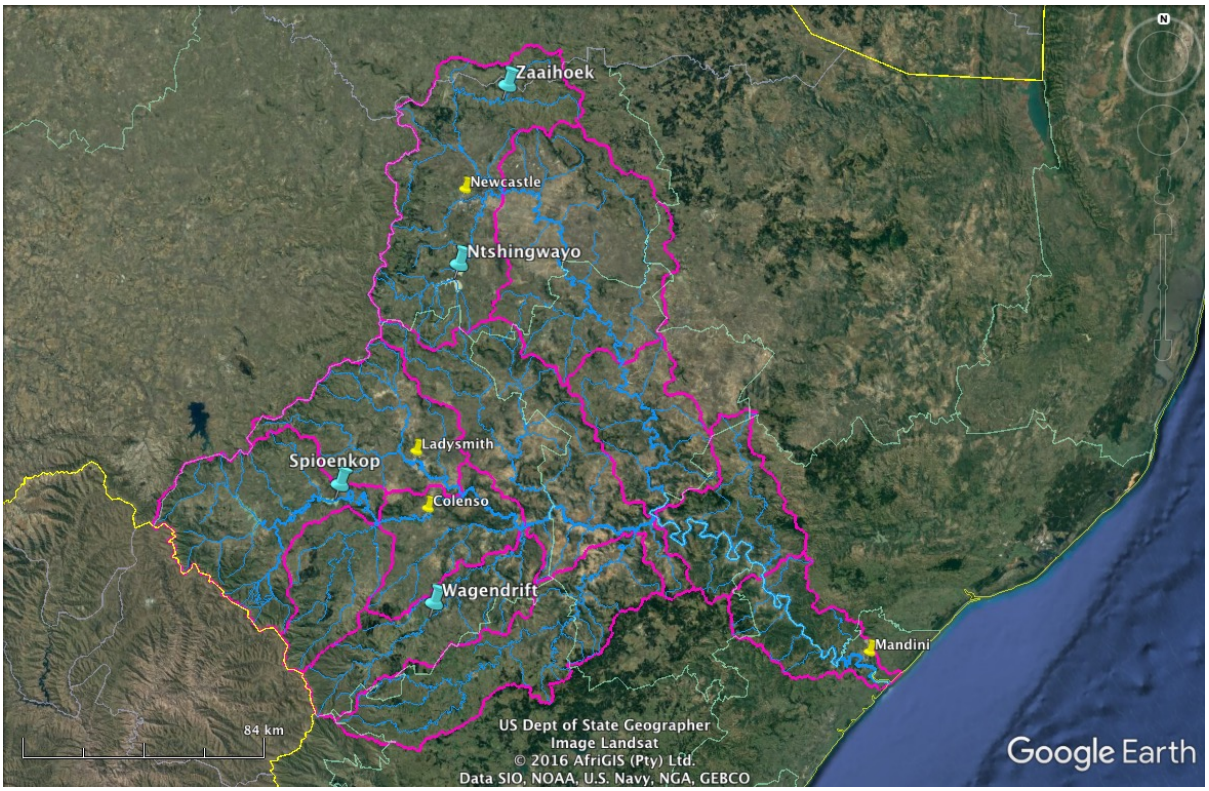


Figure 8: Other important infrastructure in the Thukela catchment area (Google Earth 7.1 2016)

Figure 9: Woodstock Dam (Department of Water and Sanitation n.d.)

Figure 10: Driel Barrage (Department of Water and Sanitation n.d.)

Figure 11: Spioenkop Dam (Department of Water and Sanitation n.d.)

2.3. Why the Thukela River

The Thukela river provides a representative river to be studied in the South African context. Like many of the rivers in South Africa the hydrology of the Thukela river is more variable than the rivers in the northern hemisphere (O'Keeffe 1989). The coefficient of variation for river flow in Southern Africa averages 0,7 which is more than 3 times more than rivers in Europe. This implies extreme episodes of flood and drought that are less predictable than rivers in Europe (Braune 1985).

The Thukela river provides evidence of human influences and is frequently affected by man-made structures, specifically by the Tugela-Vaal inter-basin transfer scheme (Bourblanc & Blanchon 2014). This scheme makes up a significant portion (about 30%) of the available water in the upper Vaal river system which supplies water to Gauteng, the economic center of South Africa (Department of Water and Forestry 2004). Gauteng is responsible for 33.8% of South Africa's Economy, with the main contributor being the finance, real estate and business service industry (Statistics South Africa 2015). Gauteng also has the highest level of construction adding to the Economy, with manufacturing also having a considerable contribution. The water resource provided from the Thukela catchment supports these industries and supports the economy of South Africa.

There has also been a lot of emphasis on the quality, as the quantity of water gets scarce and water is reused the quality of the water becomes a concern (Department of Water and Forestry 2004). The affects of irrigation, fertilizers, domestic run off and industrial waste all have an affect on the water quality. There are active strategies for flushing out the Thukela River Mouth from Spioenkop Dam and releases from the Vaal barrage to bring water quality back to an acceptable level. The Thukela river becomes important in ensuring safe usable water resources for the populations of both KwaZulu-Natal and Gauteng.

The Thukela river is clearly important to South Africa, and presents an important river system to study. The Thukela river offers a great example to study and a good case for using artificial neural networks (ANN's). The river has natural river flow variance, man-made variance through dams and even inter-basin transfers. The ANN approach of river flow gauge value predicting caters for various types of external influences and with these variations the Thukela river is a suitable case for explaining and assessing the use of ANN for predicting river flow gauge values, and and it is for these reasons the Thukela river was selected for this research.

2.4. Custodian of the Thukela River

A national department of the South African Government is the custodian of South Africa's water resources (Department of Water and Sanitation 2016a). The department is known as the South African Department of Agriculture, Forestry and Fisheries (DAFF), formerly known as the South African Department of Water and Sanitation (DWA). The country's water resources are managed and developed through regulations put in place by the DAFF (South African Government 2016). The department is important to this research as it is responsible for various dams and water schemes, such as the Thukela-Vaal transfer scheme. The departments role includes monitoring and storing river discharge values; data which is critical for evaluating the use of neural networks to predict river flow gauge values. Understanding how the DAFF manages the Thukela catchment provides context of the study area and the department providing critical data used in the research.

The DAFF monitors and manages the Thukela river, its catchment area and its dams through regional and national offices. According to the Thukela assessment done in 2001 (Wilson 2001a), water management in the area includes many activities and these are generally split between regional and national offices. Table 3 shows the allocation of responsibilities between local and national offices in the DAFF. It is part of the DAFF's responsibilities to ensure that these activities are carried out. The DAFF runs initiatives to promote effective and efficient water use including projects for alien plant eradication in the catchment area (Wilson 2001d). The DAFF is tasked with ensuring the water quality in the Thukela. One main responsibility that the DAFF has, is to ensure cost recovery. Managing the river, the catchment area and ensuring water availability is costly and these costs need to be recovered. It is the DAFF regional office's responsibility to recover the costs and this is done through billing of water users such as water boards, regional councils, irrigation boards and farmers (Wilson 2001d). In 1999/2000 it was reported that 100% of costs were recovered from these customers, yielding R70 million.

Table 3: Allocation of responsibilities in the DWS

Function	Regional	National
Dam Safety		✓
Water resources planning		✓
Pollution control	✓	
Forestry regulation	✓	
Hydrographic survey	✓	
Hydrology: Data Collection	✓	
Hydrology: analysis		✓
Water drilling services	✓	
Geotechnical drilling	✓	
Geohydrology	✓	
Monitoring of Water User Associations	✓	
Environmental monitoring and rehabilitation of abandoned mines	✓	
Abstraction control	✓	
Water user licenses (and streamflow reduction licenses)	✓	
Water demand management		✓
Working for Water programme	✓	
Operation of government schemes	✓	
Betterments		✓

The DAFF collects data for the Thukela river and makes this data available through the governmental website (Department of Water and Sanitation 2016c). The DAFF has a monitoring station below the Driel barrage on the Thukela river, the DAFF station number is V1H058. The DAFF has a recorded catchment area for this station of 1664 km^2, and their coordinates for the station are latitude -28.75889 and longitude 29.29389 (Department of Water and Sanitation 2016b). This research will be making use of this data as an input data set. The reason this monitoring station was selected is because it sits in the Upper Thukela catchment and specifically in the tertiary catchments V11. V11 experiences natural variance from run-off, and evaporation but also variance from man-made structures such as dams. More importantly this is tertiary catchment that supports the Thukela-Vaal inter-basin transfer scheme meaning this tertiary catchment plays an important part in the South African economy and water resource management. By selecting a monitoring station in the upper Thukela, and

specifically a single tertiary catchment the study area is restricted. If a monitoring station in the lower Thukela or at the Thukela mouth was selected the study area and influences on the river flow value greatly increases. Selecting tertiary catchment V11 provides a restricted study area but an area that has many influences that the neural network needs to be able to cater for when predicting river flow gauge values.

Figure 12: DWS monitoring station V1H058 (Department of Water and Sanitation n.d.)

2.5. Conclusion

The Thukela river is managed by the DAFF, who has the responsibility for ensuring future water resource availability. There are a number of organizations, as well as small and large populations in the Thukela catchment area that rely on this water source. One of the most important factors of the Thukela river is the water supply it provides to the capital and economic hub in Gauteng, South Africa. A large number of industries and residential areas depend on the water from the Thukela that is provided to them through the Thukela-Vaal transfer scheme. The river itself lies in a catchment area with diverse vegetation and varying climate. The Thukela river in KwaZulu Natal, South Africa, provides a prime example of an important river in the South African context and a good case study for evaluating the use of neural networks to predict river flow gauge values.

The Thukela river and specifically tertiary catchment V11 experiences natural variation in river flow values, and variations from man-made dams and inter-basin transfer schemes. This variation presents a complex river system to predict river flow gauge values, allowing the evaluation of neural networks ability to predict river flow gauge values. The Thukela river represents a complex case study and therefore ensures the evaluation of the use of neural networks to predict river flow gauge values is not on a best case scenario.

Chapter 2 provided an overview of the Thukela river, the catchment area and the context of the Thukela river within South Africa. The chapter highlighted the DAFF monitoring station at the Driel barrage, the tertiary catchment area V11 and described why this monitoring station was selected for this study. Chapter 4 describes the use of the DAFF data in the research and how it will be utilized within the neural network.

3. LITERATURE REVIEW

3.1. Introduction

River flow discharge rates play an important part in water resource management around the globe. Being able to assess the amount of water flowing through a river at various points can ensure water availability and water quality. The concept of river flow discharge monitoring falls into the bigger scheme called 'Water resource quality monitoring' (Grobler & Ntsaba 2004). River flow discharge data has been referred to as being 'data-rich but information-poor' since the 1970's.

This chapter will outline the the various approaches to river flow analysis and specifically river discharge rates. The chapter starts by giving an overview of instantaneous discharge **Q** and the need for predictive approaches in section 3.2. Then in section 3.3 more traditional approaches of river flow analysis are explained. Section 3.4 will then outline newer different approaches to river flow analysis. Section 3.5 will cover a more specific topic of river flow analysis using artificial neural networks (ANN's) similar to this research. Finally, a conclusion will be drawn in section 3.6 based on the literature review and its relevance to this research.

3.2. Overview

River flow discharge rates are a measurement of the total instantaneous discharge **Q** in the channel cross section of a river (Smith & Pavelsky 2008). The instantaneous discharge **Q** is measured in $\mathbf{m^3/s}$ or $\mathbf{ft^3/s}$, and traditionally the value of **Q** is then averaged from measurements taken across the stream. Currently there are limited methods of using the data to provide information that can be fed into strategic planning policies and decisions around water resource management, and it is for this reason river flow discharge rates are referred to as being 'data-rich but information-poor' (Grobler & Ntsaba 2004).

While analysis methods are improving, the traditional monitored data is not being used, instead newer predictive algorithms and models are being generated. This move towards predictive modelling is necessary for a number of reasons, with the major driving force being a decrease in hydraulic monitoring networks in conjunction with physical monitoring being difficult and expensive in inaccessible and remote areas (Tarpanelli et al. 2013). In some countries there is also a difficulty associated to data sharing due to technological and political reasons, which is where predictive models may be able to provide an approach less hampered by these restrictions. Not only the physical attributes of river flow discharge rate monitoring need to be considered, the significance of the available information also needs to be considered. Predictive algorithms can be designed for a user centric approach to ensure the output from the model is meaningful and adds value during analysis (Grobler & Ntsaba 2004). This feature of predictive models assists in moving river flow discharge rate data away from the stigma of 'data-rich but information-poor'. The use of predictive models opens up the ability to include additional information into the data for analysis. There are many different influences that could affect a rivers discharge rate such as sedimentation of flood plains, evaporation rates, land use, topology and soil characteristics. Certain types of predictive models can incorporate and try to cater for these influences in their analysis and predictions (Potter et al. 2010; Tarpanelli et al. 2013). One such influence is weather attributes, for example precipitation, temperature, humidity and cloud cover all could have an effect on the river discharge rate.

3.3. Traditional approaches

The core concept of traditional approaches to river flow analysis lies in the calculation of the river current, which relies heavily on the width of the river and the depth of the river. The traditional and more accepted approach is the "summation of the products of the partial areas of a stream cross section and their respective average velocities" (Buchanan & Somers 1969). Equation 1 represents the total discharge, and the partial discharge at any partial section x is computed using Equation 2. The partial discharge is calculated at locations 1,2,3,4,… n , where the area of each section extends from half the distance to the previous location and half the distance of the next location. In doing so the river is divided using the mid-section approach and the calculation will cover the full river cross-section as depicted in Figure 13 (Buchanan & Somers 1969).

$$Q = \sum_{x=1}^{n} q_x$$

Where

$$q_x = discharge\ through\ partial\ section\ x$$

Equation 1: Summation of the total Discharge

$$q_x = v_x \left[\frac{(b_x - b_{x-1})}{2} + \frac{b_{x+1} - b_x}{2} \right] d_x$$
$$= v_x \left[\frac{b_{x+1} - b_{x-1}}{2} \right] d_x$$

Where

$$q_x = discharge\ through\ partial\ section\ x$$
$$v_x = mean\ velocity\ at\ location\ x$$
$$b_x = distance\ from\ initial\ point\ to\ location\ x$$
$$b_{x-1} = distance\ from\ initial\ point\ to\ proceeding\ location$$
$$b_{x+1} = distance\ from\ initial\ point\ to\ next\ location$$
$$d_x = depth\ of\ water\ at\ location\ x$$

Equation 2: Partial discharge at location x

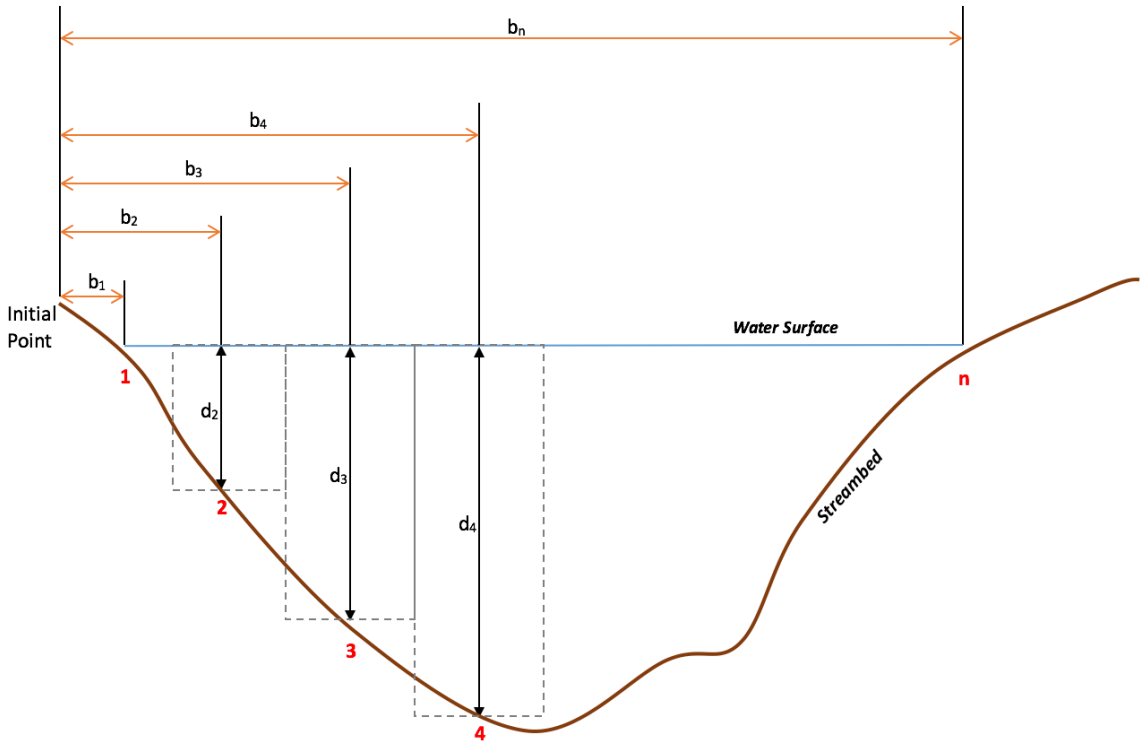


Figure 13: Midsection method of computing cross-section area for discharge measurement (Buchanan & Somers 1969)

There are a number of different procedures and instruments used in the traditional approaches and these are not restricted to wading into the river to take measurements. In 1969, Buchanan and Somers (1969) described the approaches that were used during that period. The equipment and instruments could be broken down into three groups namely current meters, sounding equipment and width-measuring equipment. A current meter was an instrument that measured the velocity of flowing water. The current meter was based on the angular velocity of the meter rotor caused by the velocity of the water flowing over the rotor. The proportional velocity was calculated from the number of revolutions during a specific measurement time. There were two types of current meters, namely vertical-rotor axis meters with cups or vanes and horizontal-rotor axis meters with vanes. The vertical-rotor axis meter is the more common of the two. In the same publication Buchanan and Somers (1969) made reference to some of the newer approaches at the time, specifically research from the California Department of Water resources which had developed an optical current meter that overcame some of the flaws of submersible current meters, such as entangled debris.
Using the current meters described above a number of procedures were used to take the current measurement, these include (Buchanan & Somers 1969):

- Current-meter measurements from wading – this approach was preferred if the conditions permitted. This approach adds the benefit of selecting the best cross section measurements. This approach could have accuracy issues as the wading rod needs to be held at the correct angles.
- Current-meter measurements from cableways – this approach makes use of weights to lower the equipment into the river when wading is not possible. Debris and ice have been known to move the sounding line and give inaccurate readings. The approach has been known to give erratic results.

- Current-meter measurements from bridges – using bridges is another approach when wading is not possible, but cable way sections are usually preferred. Handlines or sounding reels are generally used with the bridge approach. The bridge itself can change horizontal angles and can also cause scouring. More cross sections are generally ensured with the bridge approach.
- Current-meter measurements from ice cover – discharge measurements under ice cover are done in extreme conditions and are dangerous as the water can melt the underside of the ice. Drilling holes into the ice and taking current measurements presents complexities as the measurements are similar to those in a pipe where the velocity of the water is lower directly under the ice.
- Current-meter measurements from boats (stationary or moving) – A boat approach is used when no cableway or suitable bridge exists and where the river is too deep to wade. Personal safety needs to be considered especially when rivers have a high velocity. Large streams and estuaries can be costly and impractical to be measured with the other conventional approaches and a moving boat approach provides a flexible and quick means of measuring the discharge.

The second group of instruments and equipment described by Buchanan and Somers (1969) was sounding equipment. Sounding equipment was used to determine the depth of a river which is crucial in the midsection method for computing cross-section discharge measurements. A wide range of instruments existed for sounding such as wading rods, sounding weights, sounding reels, hand lines and sonic sounders (Buchanan & Somers 1969). The wading rod was used by placing the rod in the stream with the base plate on the streambed, and the water depth is read off the marked main rod. When the stream is too deep or the velocity is to fast to wade into the stream with a wading rod then a sounding weight would be used to measure the depth of the river. The sounding weight was suspended in the water from an appropriate position such as a boat, bridge or cableway. The cable with the sounding weight is either controlled by a reel or a hand line. The sonic sounder approach allows depth to be measured without needing to lower any sounding weights or wade into the river. The sonic sounder makes use of a narrow beam angel of 6° to reduce error on steep streambeds or near other obstacles. The sonic sounder can be affected by temperature but the error is limited to around 2% in fresh water and can be eliminated through adjusting the sonic sounder using initial correct average depths.

The third group of instruments and equipment described by Buchanan and Somers (1969) was width-measuring equipment. Measuring the width of the river was generally done with steel or metallic tapes or tag lines. A tagline was made of galvanized steel and has solder beads at measurement intervals. With long taglines, tagline reels are used particularly when being used from a boat. When measurements are commonly taken around a bridge the bridge often has painted marks to indicate the various intervals.

In some circumstances the common current-meter measurement procedures were not possible and other traditional approaches were needed. For example in shallow, low velocity streams the approaches previously described provide inaccurate measurements (Buchanan & Somers 1969). A 90°V-notch portable weir plate offered a more appropriate method in these situations. The most accurate method in these situations was volumetric measurements. Volumetric measurements observed the time needed to fill or part fill a container of known volume and from these the current discharge can be calculated.
The approaches, equipment and instruments described above can still be found in use. Gravelle (2015) explains the use of mechanical current meters and describes them as either horizontal

or vertical axis current meters. The manufacturers and materials used to produce the equipment may have changed since 1969 but the equipment is the same design. It is still possible to purchase the type AA meter Buchanan and Somers (1969) described, such as the USGS Type AA current meter from Rickly Hydrological Company (2016). Buchanan and Somers (1969) also describe pygmy meters used in shallow depths, these can also still be purchased for example from Rupson Industries (2016).

This section has described the more common traditional approaches to measuring current discharge. There are newer approaches that overcome some of the difficulties associated with the traditional approaches such as cost, debris and accuracy. Section 3.4 and 3.5 outlines some of these approaches.

3.4. Newer approaches

There are many approaches to improving river flow discharge approaches. Some take the approach of improving the ability to analyse the river discharge data collected with conventional approaches, while others look to estimate the river flow discharge. One of the analytical approaches of river discharge is called flow duration curves (otherwise known as stream flow duration curve). "A streamflow duration curve illustrates the relationship between the frequency and magnitude of streamflow" (Vogel & Fennessey 1995). Flow duration curves have been used in flood control, water quality management, river sediment monitoring and water use planning. Flow duration curves can be used to illustrate the impact of a number of external influences to river flow such as geology and climate. One of the main advantages to using flow duration curves is that they provide a graphical representation of flow. Flow duration curves are ideal for summarizing complex data into a format that can be understood and shown graphically.

Vogel and Fennessey (1995) describe an example of using flow duration curves to construct a weighted usable habitat duration curve, shown in Figure 14. The example made use of the physical habitat simulation system introduced by the Cooperative Instream Flow Group of the U.S Fish and Wildlife service. The simulation system could develop a rating curve which 'relates total habitat area to river discharge for a particular species during a particular stage in its life'(Vogel & Fennessey 1995). As Figure 14 shows this rating curve is then combined with a flow duration curve to produce a habitat-duration curve. This is one example of how the flow duration curve can be used.

There are always some disadvantages associated with summarizing complex data in that there is a loss of specific complex details. Flow duration curves are no different and are often considered to over simplify the data. The interpretation of the flow duration curve is also dependent on the period of record and this must be considered when analyzing the flow duration curve. Flow duration curves also present a severe drawback in that they ignore serial structure and this limits their use in time dependent problems.

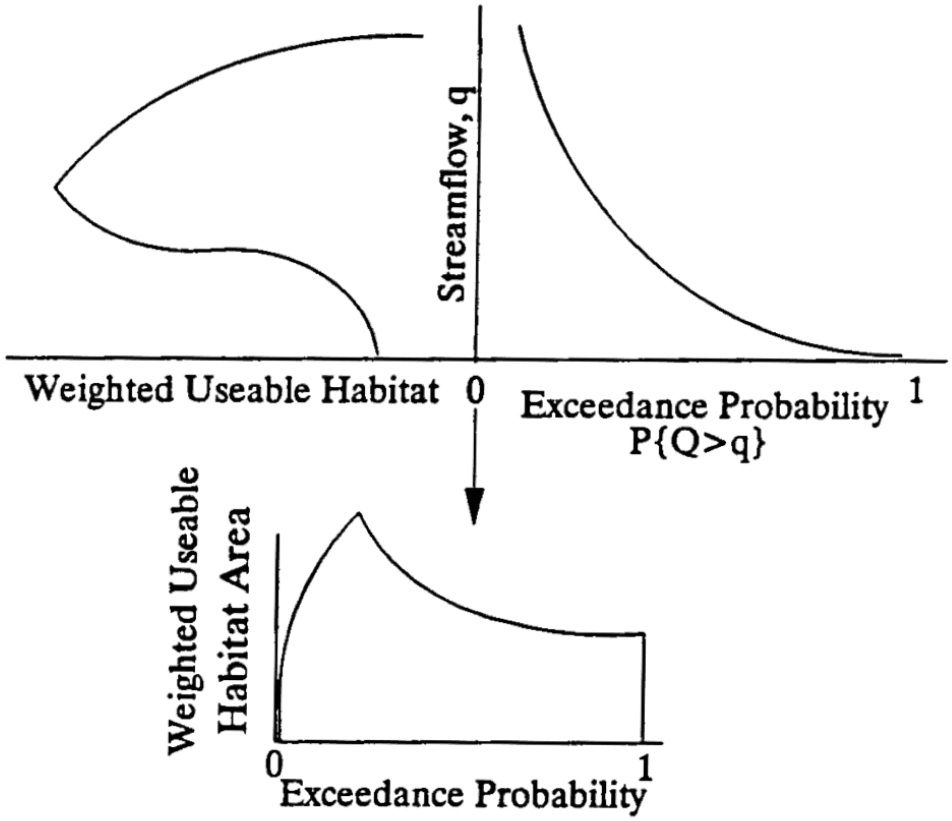


Figure 14: Development of a usable habitat duration curve by Vogel and Fennessey (Vogel & Fennessey 1995)

There are approaches that also try to improve on the more traditional river flow discharge estimates. In 2013, Tarpanelli et al. (2013) proposed an approach for estimating discharge that made use of satellite altimetry on the Po River. The approach was based on rating curve models (RCM). The RCM approach allows an estimation of the river flow discharge rate based on measurements taken at another location (Barbetta et al. 2012). The RCM approach was developed for downstream estimates based on upstream measurements. By using the RCM approach, Tarpanelli et al. (2013) defined an approach that would only require water levels at a specific point in the river and the discharge observed further upstream. The proposed method derived the water levels of the river from satellite altimetry, and then made use of existing water levels and discharge measurements upstream (or downstream). The accuracy of the approach was evaluated against an empirical formula using remotely sensed hydraulic information. Tarpanelli et al. (2013) showed that while it is possible to use satellite altimetry to estimate river discharge the satellite sensor effects the results. In the study the ERS-2 satellite provided less accurate estimations than the ENVISAT satellites. The research also indicated that two outstanding issues would need to be addressed to improve the accuracy of the estimations. Firstly the assessment of the cross-section geometry, and secondly the estimation of the elevation of the cross-section bottom.

Research has been done previously into predictive models for river discharge rates using various different data inputs and different predictive algorithms. Research has shown that predictive models can perform better than current traditional approaches, when the correct knowledge is used and the correct intelligent predictive models are used with a defined

expected output. For example Chang et al. (2005) made use of genetic algorithms (GA) and a fuzzy rule base (FRB) to improve reservoir operations at the Shihmen reservoir, Taiwan. The need for an intelligent model was to handle the complexity of reservoir operations due to the variability of natural stream flow and water demand. The Shihmen reservoir makes use of operating rule curves, called the M-5 operating rule curve, to manage the demand, upper limits, lower limits and critical limits of the reservoir (F.-J. Chang et al. 2005). The M-5 operating rule curve was defined by trial and error and provides the policy under which the reservoir is guided and managed. The approach made use of genetic algorithms to extract knowledge from historical inflow data and a fuzzy rule base to extract knowledge from current operating rule curves. Chang et al. (2005) then used this knowledge in an adaptive network-based fuzzy inference system (ANFIS) to estimate the optimal reservoir operations. The output from the ANFIS was then compared against the current operating approach which was making use of an M-5 operating rule curve. The research concluded in finding that an ANFIS model making use of GA to draw knowledge from the input-output patterns and a FRB to draw knowledge from the operating rule curves can out-perform the more traditional M-5 curve approach.

Another approach was presented by Miller and Jinwon (Miller & Jinwon 1995) which made use of the coupled atmospheric-river flow simulation (CARS) system that was developed by the University of California Lawrence Livermore National Laboratory. The approach linked the Mesoscale Atmospheric simulation (MAS) model, the Automated Land Analysis System (ALAS) and a modified version of the hydrology model TOPMODEL. The CARS was able to simulate river flow within 10% of the observed river flow at the Hopland gauge station on the Russian river during the flood stage. In this particular study it was found that soil texture, topography and initial soil water saturation deficit were the most important surface properties when simulating the river flow.

3.5. River flow analysis using artificial neural networks

Some of these newer approaches are based on ANNs. Research by Aichouri et al. (2015) showed that neural networks can provide superior predictions of river flow to multiple linear regression-based models. This approach used as input data for the preceding 7 days for both rainfall and runoff as well as the expected rainfall for the day. The output from the ANN represented the expected runoff value for the day. The research results allowed Aichouri to draw the conclusion that ANNs are capable of modelling runoff in area's of highly fluctuating rainfall and runoff (Aichouri et al. 2015). The research also showed that ANNs could be more suitable for predicting runoff than the traditional regression model approach. ANNs have been used for more than just river flow research and have been proven to be successful in other areas of research related to river characteristics. A study by DeWeber and Wagner (2014) used neural networks to predict mean daily water temperatures in 197,402 individual stream reaches. The approach was able to predict the mean daily water temperature with good accuracy according to the research, but with the fact that it could generalize to new stream reaches and years being an important feature of the approach. The research highlighted a few key points: the first was that the mean temperature was taken from an ensemble of 100 ANNs, giving a good prediction median. A second key point was the important predictors that were used in the ANN's namely the daily air temperature, prior 7 day mean air temperature and the network catchment area. An interesting consideration was the connection between two of the input sets, namely the network catchment land cover and the network forest. The two inputs were not included into the same ANN model as there was a high correlation. The research provides a good example of an ANN model that can provide good accuracy with low overall bias. ANNs can also be used in flood warning systems, Elsafi (2014) used an ANN approach to predict flooding of the

Nile River in Sudan. The research aimed to provide a a model for detecting the flood hazard by making use of river flow at upstream locations. The research made use of data from various rivers including the Blue Nile, White Nile and Main Nile between 1965 and 2003. This research pointed out that an ANN approach provided a viable option for flood forecasting as it was needed in an area where information was lacking or difficult to obtain and that the analytical costs were reduced.

The Nile is a well known research area and has a number of complex cases for study, including the use of ANNs for river characteristics estimations. In 2010 research was conducted to investigate the use of ANNs to forecast river flow (Shamseldin 2010). The research used the Blue Nile in Sudan as a test case for an ANN rainfall-runoff model. The ANN was constructed using a multi-layer perceptron (MLP) structure with the rainfall index amongst other external inputs to estimate discharges. The research made use of a rainfall index to represent the rainfall in the area, it made use of a linear transformation of the rainfall values. The research then considered a seasonality input in the form of a seasonal expected discharge and the seasonal rainfall index. The research had therefore limited the maximum inputs into the ANN to three inputs. With four different variations of these inputs, the research created four different ANN models to assess which inputs delivered the best performance. It was found that the model with the most inputs (all three external inputs) performed the best, while the model with the least number of inputs performed the worst (Shamseldin 2010). The research drew three important conclusions regarding the use of an ANN for estimating river discharge. Firstly, that based on the results for the Blue Nile, neural network models have a good potential for estimating the river discharge in developing countries. Secondly that appropriate input selection is extremely important to ensure the success and accuracy of the estimates, and may only be possible through trial and error. Finally, the use of different neural network model structures and different updating procedures may lead to improved results.

This research does not build on a traditional approach, it does not take an existing model and use new techniques to gather the data for the traditional approach. This research is similar to the research by Shamseldin (2010) in that it takes a number of inputs and feeds this into an ANN to produce a river flow estimation. As Shamseldin (2010) found the appropriate input into the ANN is important, and while Shamseldin restricted the input to rainfall index values this research considers a much larger number of inputs in the form of weather parameters.

3.6. Conclusion

This overview of previous research into predicting river characteristics shows that there have been improvements on the traditional approaches of river characteristic estimations. It also clearly shows that for a number of years neural networks have been used and studied as successful candidates for predicting river characteristics such as runoff, water temperature, flooding and river flow rates. Previous research has shown though that it is important to consider the weaknesses of each approach such as the sensor type and physical cross section geometry in satellite altimetry approaches. The research has also shown that when using ANNs, input data must be selected carefully and managed correctly as this affects the prediction accuracy. An ANN can provide a cost effective way of analyzing and predicting river flow discharge rates. Chapter 4 describes the methodology used in this research and how the information from this literature review feeds into this research. The chapter describes the input data in this research, where it is from and how it should be handled. Then neural networks are explained in more detail, outlining what an ANN is, how it is structured and how the accuracy of the neural network can be determined. Finally, the chapter describes the exact neural

network model used in this research and the naive prediction the neural network aims to improve on.

4. RESEARCH DESIGN

4.1. Introduction

This chapter of the research defines the methodology used to implement the artificial neural network (ANN) model required for the research. This research makes use of two methods, firstly a literature review which is presented in Chapter 2 and Chapter 3. Secondly, the research uses an experiment methodology which Chapter 4 outlines in detail. The chapter starts off by describing the literature review method and the experiment method as well as how they form the structure of this research. The chapter then outlines in section 4.3 the programming libraries used. Data is one of the most important requirements for this experiment as the ANN model requires data inputs. This chapter in sections 4.4 and 4.5 explains and discusses in detail the data used from both the South African Department of Agriculture, Forestry and Fisheries (DAFF) and the European Center for Medium-Range Forecasts (ECMWF). Section 4.6 of the Chapter then concentrates on the ANN model itself, firstly explaining what neural networks are and how they work, how their performance is measured and problems associated with them in section 4.5. The ANN model used in the experiment is defined in section 4.7, with the limitations to the research explained in section 4.8. The chapter then concludes in section 4.9.

4.2. Methods

4.2.1. Literature review

This research makes use of a literature review in Chapters 2 and 3 to present the context and background to the research. The literature review is used to investigate the Thukela river and the surrounding catchment area. This showed that the Thukela river is an important river for both industries and residential areas in South Africa. The single role the Thukela river plays is twofold in that it supplies water to the province of KwaZulu-Natal through its natural flow, but also to Gauteng through the Thukela-Vaal transfer Scheme. The literature review in Chapter 3 provides the required knowledge to understand river flow, how it is measured and the difficulties related to the measurements. The Chapter also provides key insight into a few of the newer approaches to estimating river flow, specifically the use of ANN's. This experiment uses this information and is steered by the findings in the literature review.

4.2.2. Model

A model is a "simplified representation of a portion of reality, expressed in conceptual, physical, graphical or computational form" (Montello & Sutton 2006). A model is essentially an information bearing entity, and in this research that entity is an ANN ensemble. Originally models were seen to simplify reality, but this is not always true such as certain mathematical, computational or adaptive statistical models. The ANN in this research is an adaptive statistical model that is a nonlinear mapping of real numbers from the input space to the target or output space (Kröse & van der Smagt 1991; Engelbrecht 2007). In this specific model the real numbers from the input space are weather characteristic values such as temperature and precipitation. The models real number output space is the river flow gauge value at the DWS monitoring station V1H058 in the Thukela river downstream from the Driel Barrage. The model created in this research creates an output that is treated as if it were a measurement. The model

is generating simulated data based on the portion of reality that the model refers to, the weather and river flow data. Feature selection is an important step in creating the model and forms the first part of the experiment. Section 4.2.3 describes the feature selection used in this research.

4.2.3. Feature selection

This research makes use of a multi-scenario design for the feature selection in an attempt to evaluate the use of neural networks in many different setups. A number of the scenarios make use of a naive prediction. In this research the naive river flow gauge value estimate is taken as the previous day's river flow gauge value (Pappenberger et al. 2015). This is seen as a completely naive prediction and makes the assumption that the current river flow gauge value will be the same as the following day. The naive prediction is explained in more detail in section 5.2.

The scenarios are broken down based on the input features. The multi-scenario design allows the effect of a single variable to be evaluated. Many input variables are being used in the research and there are a number of different variables that could have an effect on the performance of the neural network and in turn its ability to be used to predict river flow gauge values. The scenarios in this research are described in Table 4, and all the feature selection scenarios will use a set scorecard provided in Section 4.6 to evaluate the performance of the ANN model with the subset of input features. Figure 15 shows the flow of information between scenarios, with scenario D being the final scenario.

Table 4: Methodology – Scenarios

Scenario	**Description**
Scenario A *Naive vs Weather data as ANN ensemble inputs*	This scenario analyses the effect of using weather data vs naive data as input data into the neural network ensemble. This is done by running three neural network ensemble predictions. The first ANN ensemble is with only naive data as input data. The second ANN ensemble is with only weather data as input data. The last ANN ensemble is with both naive data and weather data as input.
Scenario B *Effect of individual weather characteristics as ANN ensemble input*	This scenario analyses the effect of individual weather characteristics on the performance of the ANN ensemble. This is done through using only one weather characteristic as input into the ANN ensemble at a time. Each ANN ensemble has the naive data, month indicators and a single weather characteristic data as input. This shows the effect of each of weather characteristics on the ability of the neural network ensemble to predict river flow gauge values.
Scenario C *Effect of correlation filtering on the ANN ensemble input data.*	This scenario analyses the effect of filtering inputs based on correlation with the actual river flow. This is done through running multiple ANN ensembles but each time making the data input selection stricter based on correlation. The correlation requirement will be incrementally increased by five from zero until there are no input values meeting the correlation filter. Each ANN ensemble will use the naive data and all weather characteristics, both subject to the filtering requirements.
Scenario	**Description**

Scenario D *Best case scenario for the ANN ensemble input data*	This scenario only involves one ANN ensemble with the best case scenario selection from the other scenarios. Only those weather characteristic values that positively affected the ANN ensemble in scenario B will be selected. All input data into the ANN ensemble are filtered based on correlation filters found in scenario C. This scenario should represent the best case scenario using the data available and the filtering available.

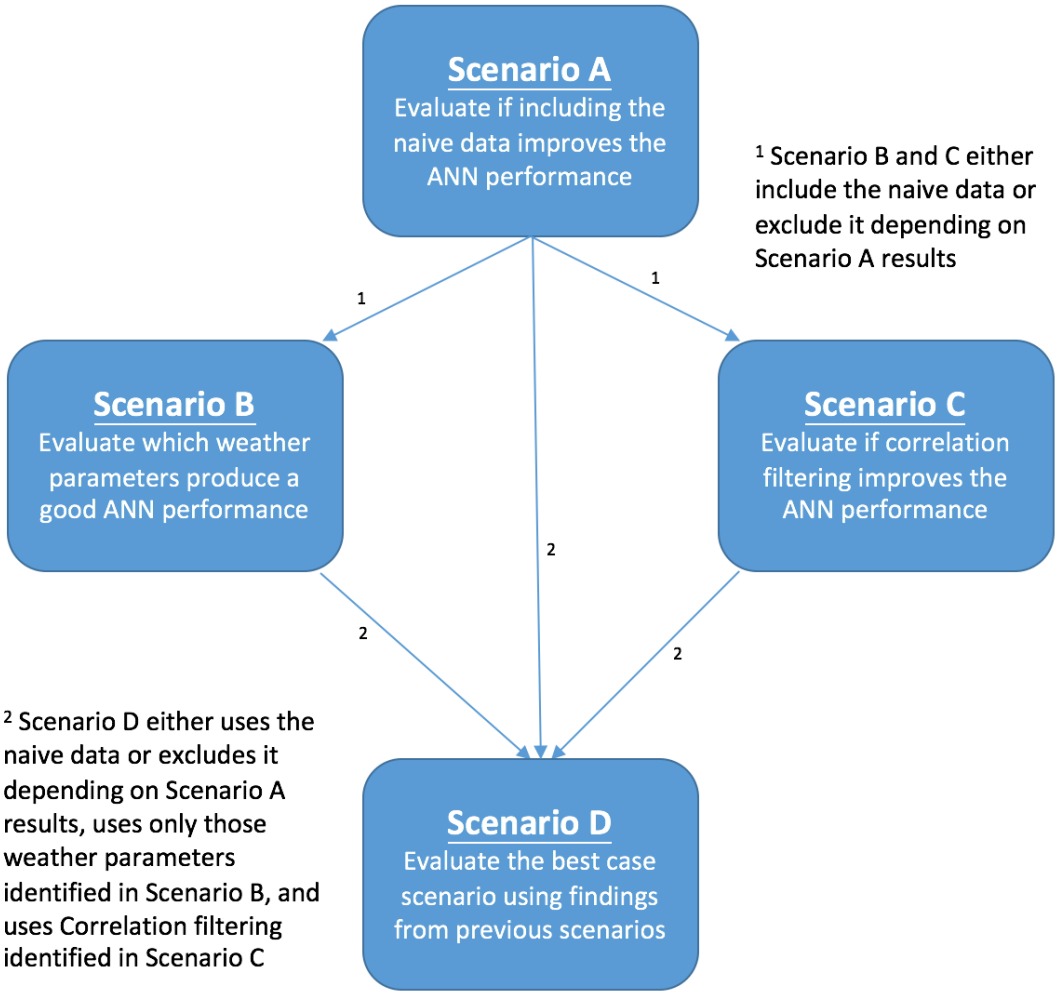


Figure 15: Scenario flow, showing flow of information

The results of the multi-scenario feature selection evaluations are presented in Chapter 5. The rest of this chapter outlines the libraries used in the model, the structure and setup of the ANN and the naive prediction.

4.3.Selected Libraries

4.3.1. Introduction

It is important to minimize the possibility of programming bugs and to ensure an efficient accurate programming of the model. The research made use of two main coding libraries to implement the extraction of data and the ANN to minimize the possibility of programming bugs. The first library, 'Network Common Data Form' (NetCDF)[1] was used to extract weather data retrieved from the ECMWF. The second library used was Encog[2] for creating the ANN. Sections 4.3.2 and 4.3.3 provides the background to the libraries, where they have been used before and how they were used in this research.

4.3.2. NetCDF

NetCDF is created and maintained by Unidata which is part of the University Corporation for Atmospheric Research (UCAR) community programs (UCP) funded by the National Science Foundation of the United States. Unidata provide data services and tools to help researchers and educators use earth related data through a number of software packages (Unidata 2016c). NetCDF is one of the packages offered by Unidata, and is a set of libraries and machine independent data formats that assist with managing array-oriented data. Climate and forecast (CF) metadata conventions are designed in such a way as to promote the use and sharing of data files created through the NetCDF API. The CF conventions are used to ensure data has a description through structured metadata that describe the data in each variable, the spatial properties and the temporal properties of the data variable. This allows comparison of data from varying sources. The CF conventions have been accepted by a number of projects and have been endorsed by a number of standards bodies. Portions of the NetCDF format have been endorsed by standards bodies such as NASA Earth Science Data Systems (ESDS) Standards Process Group, the US Steering Committee of the Federal Geographical Data Committee (FGDC) and the Open Geospatial Consortium (OGC) (Unidata 2016b). NetCDF has not only been endorsed by a number of standards groups, but has also been actively used by a large number of organizations and on a wide variety of projects (Unidata 2016d). Organizations from all around the world make use of NetCDF such as NASA Jet propulsion laboratory (United States of America), CSIRO Marine and Atmospheric Research remote sensing (Australia), AVISO satellite data (France), German Aerospace agency (Germany) and European Climate Assessment (Netherlands) to mention a few. A few examples of projects where NetCDF is used include (Unidata 2016d):

- NASA's Halogen Occultation Experiment (HALOE) which makes data available in NetCDF format.
- Generic Mapping tools (GMT) which is a set of command-line tools for data manipulation using PostScript.
- NCAR's research data program makes use of NetCDF as its primary archiving file format in the Zebra display and analysis system.

[1] Source for the NetCDF library: http://www.unidata.ucar.edu/software/netcdf/ (Last accessed 09/10/2016)

[2] Source for the Encog library : http://www.heatonresearch.com/encog/ (Last accessed 09/10/2016)

- The Amber project, in the field of molecular dynamics and biomolecule simulations, uses NETCDF conventions for trajectory files.
- The Federal Waterways Engineering and research Institute (BAW), NetCDF is used for storing water levels, current velocity, salinity, tidal highs, tidal range and low water in addition to other similar characteristics.

NetCDF format and libraries have not only been used in organizations and project groups but also in published research over the years. Research by (Gaustad et al. 2014) discusses the atmospheric radiation measurement data integrator (ADI) framework, and how the use of ADI can decrease time and cost of implementing scientific algorithms. The research describes in detail how the NetCDF format fits into the framework and the use of Unidata's software libraries for file handling and manipulation. The research also explains that the ADI software package that produces NetCDF CF compliant NetCDF files (Gaustad et al. 2014). Research done by Su et al describes that NetCDF "can store, manage, obtain and distribute grid data efficiently" (Su et al. 2016). The research notes that NetCDF file format can obtain target data efficiently in massive amounts of data that is dynamic and multi-dimensional. The research highlights the importance that NetCDF file format provides in the ability to realize unified management and share marine data. In some cases, research into NetCDF has produced new software packages that leverage the NetCDF file format. As an example in 2010 a Python package called PyPnetCDF, was developed to provide parallel access to NetCDF files (Galiano et al. 2010). This package was then listed on the Unidata Software page as useful software for manipulating NetCDF data.

The data in NetCDF format can be defined as self-describing, portable, scalable, appendable, sharable and archivable. To achieve this, one of the NetCDF programming interfaces must be used. NetCDF offer a wide variety of programming interfaces and this research makes use of their NetCDF-Java library. The NetCDF-Java library provides a java interface for writing and reading CFD data. This research made use of the library to open a dataset from the European Center for Medium-Range Weather Forecasts (ECMWF) and extract the data specific to the catchment area of the Thukela river above the Driel barrage. Table 5 (Unidata 2016a) provides a list of the main classes/functions used from the NetCDF-Java library in this research.

Table 5: Main NetCDF-Java library classes/functions used (Unidata 2016a)

Name	Class/Function	Description
NetcdfDataset	Class	The NetcdfDataset class is used to extend the NetCDF NetcdfFile object. The NetcdfDataset opens with all functions turned on. The functions used include ScaleMissing, CoordSystem and ConvertEnums.
ScaleMissing	Function	This function processes scale/offset/missing attributes while reading the data file and automatically converts the data.
CoordSystem	Function	This function extracts the CoordinateSystem value using the CoordSysBuilder plug-in.
ConvertEnums	Function	This function converts enum values to their corresponding String values.
NetcdfFile	Class	NetcdfFile is used to read scientific datasets that are accessible through the NetCDF API.

Variable	Class	A Variable is a logical container for data that has a specific dataType, a set of dimensions that define the array shape and a set of Attributes. The data stored in Variable is a multi dimensional array of primitive types. To access the data you call the Read() function to read the data in the Variable.

The NetCDF format was used in this research as it is a well known library that has been used before in previous research. The library was also used because ECMWF publishes data in this format. The NetCDF-Java library was used as it is provided by Unidata who maintain the NetCDF format. The library was used to reduce the possibility of any programming bugs and improve the quality of the outputs from this research.

4.3.3. Encog

Encog is a framework for advanced machine learning that has been in active development since 2008 (Heaton 2015). The framework supports machine learning algorithms such as support vector machines, ANN's, Bayesian networks, Hidden Markov models and genetic algorithms. The Encog API is available in both Java and C#. This research makes use of the Java library to implement an ANN. One of the major features of the Encog framework is that a large number of the training algorithms are multi-threaded and scale well to multi-core hardware which improves performance and can reduce training time during research.

The Encog framework provides a wide range of propagation learning algorithms for ANN's such as back propagation, resilient propagation, Levenberg-Marquardt, quickpropagation and scaled conjugate gradient. Further features have been added into the framework to improve the performance by finding the optimal network architecture such as neural network pruning and model selection. The neural network architecture can even be automatically built by genetic algorithms using NEAT and HyperNEAT (Stanley & Miikkulainen 2002). This research makes use of the neural network pruning functionality provided by the library, resilient propagation learning algorithm and the library's K-Fold cross validation functionality.

An important part of using the library is to understand how Encog stores and uses data. Encog breaks down data object handling into three interfaces, as shown in Figure 16 (Heaton 2015):

1. **MLData/IMLData** – This holds a vector that will be either the input data into a model or output from a model.
2. **MLDataPair/IMLDataPair** – This holds a pair of MLData vectors for supervised training. The pair consists of a MLData vector as input, and a MLData vector as the output.
3. **MLDataSet/IMLDataSet** – This holds lists of MLDataPair objects to provide inputs to training algorithms.

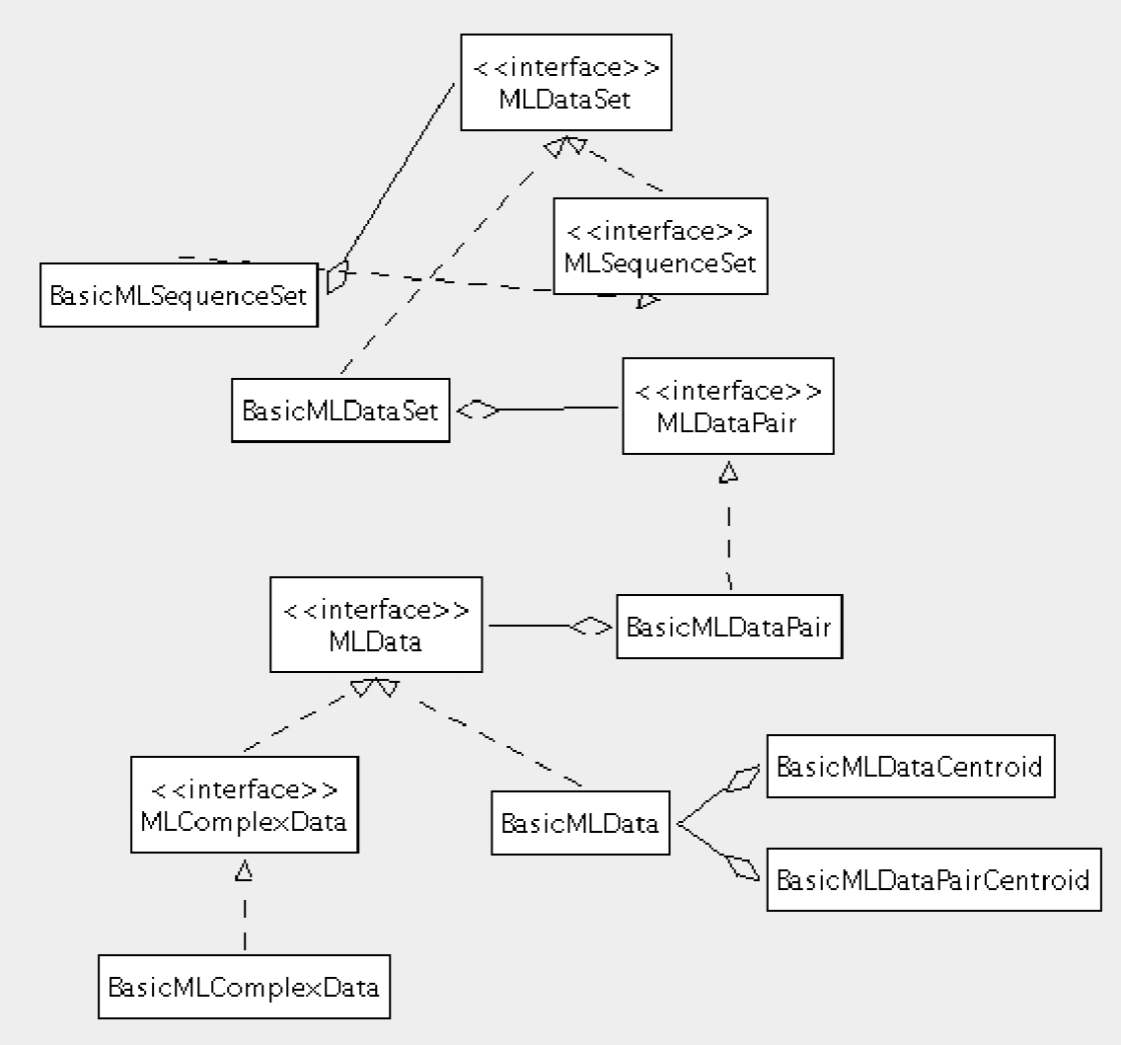


Figure 16: Encog MLData Classes, Source: (Heaton 2014)

The Encog library provides access to a recognized, previously used set of ANN frameworks. This reduces the possibility of programming bugs and through the use of the multi-threaded learning algorithms improves run duration during the research.

4.4.Input data

4.4.1. Introduction

An adaptive statistical model such as the ANN requires data to generate the non-linear mapping from the input space to the output space (Kröse & van der Smagt 1991; Engelbrecht 2007). This research requires input data in the form of weather characteristics and river flow gauge values. This research makes use of data from the DAFF and the ECMWF. The data was downloaded from these sources and then processed to ensure the data could be used in the model. The data needed to be accessible to allow use in this research and any future research. The period of data availability was also a consideration in selecting data.

This section describes the data sources and the data available for this research. Section 4.4.2 describes the DWS data, section 4.4.3 describes the ECMWF data. After

understanding the data sources and data availability section 4.4.4 discusses the data cleaning and rolling window approach in the research. Section 4.4.5 gives an indication of the correlations between the data, as this is required in feature selection scenario C and feature selection scenario D in Chapter 5. A brief summary of the selected data for this research is provided in the conclusion, section 4.4.6.

4.4.2. DWS data

The South African DAFF (www.daff.gov.za) formally known as the South African Department of Water and Sanitation (www.dwa.gov.za) is responsible for ensuring that South Africa's water resources are managed correctly and developed through regulations and supporting water delivery across the country (South African Government 2016). Section 2.4 highlighted the role the DAFF plays in managing the Thukela River in KwaZulu Natal, South Africa. Part of this role includes monitoring and storing river discharge values. The DAFF makes this data accessible through their website (https://www.dwa.gov.za/Hydrology/ - last accessed 12 June 2016).

The DAFF provides data for a number of different monitoring stations within the Thukela catchment area. This research makes use specifically of the Driel monitoring station V1H058. The monitoring station was installed in 1985 and the DAFF provides data from 1985-05-30 till 2016-01-18. The monitoring station is still active and recording data. The DAFF provides a number of different data sets and information on the station data through the webpage (Department of Water and Sanitation 2016b). The DAFF states on the monitoring stations data that catchment area is 1664 km^2 and the monitoring stations coordinates are latitude -28.75889 and longitude 29.29389. The DAFF provides a photo for the monitoring station, see Figure 12 in section 2.4.

The DAFF provides a number of datasets, namely monthly volume values in m^3; a primary data set; daily average flow in m^3/s; monthly flood peaks and annual flood peaks. This research makes use of the daily average flow in m^3/s, an example of the data is shown in Figure 17. The DWS provided a set of quality ratings, which are shown in the annex section in Table 20. These quality ratings can also be seen on the example daily average flow data provided in Figure 17. These quality ratings are very important when using the data as they can be used to assist with the data cleaning. This is discussed in section 4.4.4 but ideally the quality rating for this research is 1= Good continuous data or 2 = Good edited data.

Data are continuously updated and reviewed.
The format of this file is as follows:
POS. 1-8 = Date of daily flow CCYYMMDD
POS. 10-18 = Daily avg flow rate in cubic metres/sec 99999.999
POS. 20-24 = Quality code

V1H058
Variable 100.00 Surface Water Level

DATE	D AVG F/R	QUAL
19860101	8.403	1
19860102	1.023	1
19860103	1.035	1
19860104	1.089	1
19860105	4.311	1
19860106	10.536	1
19860107	3.020	1
19860108	0.915	1
19860109	0.803	1
19860110	0.792	1

Figure 17: DWS Daily Average Flow m^3/s example, Source: (Department of Water and Sanitation 2016b)

4.4.3. ECMWF data

The European Centre for Medium-Range Weather Forecasts is an operational service provider and research institute (European Centre for Medium-Range Weather Forecasts 2016b). The ECMWF was established in 1975 and is independent but relies on intergovernmental organisations supported by 34 member states. The ECMWF is closely linked to a number of other organisations such as the North Atlantic Treaty Organisation (NATO) and European Space Agency (ESA) through their membership in the co-ordinated organisations. At its core, the ECMWF produces weather forecasts and monitors the earth-system (European Centre for Medium-Range Weather Forecasts 2016a). Their mission includes research and meteorological data archival to improve on there ability to produce these weather forecasts. The ECMWF therefore has a large amount of data. One particular data set is the ERA-Interim data set. The ERA-Interim data can be downloaded from a public access web interface (http://apps.ecmwf.int/datasets/data/interim-full-daily/ - last accessed 12 June 2016). To access the data a user account needs to be setup but the use of the ERA-Interim data is provided free of charge. There is a license agreement that is accepted when creating a user account. The license states that 'The Licensee is authorised to use on a non-exclusive basis the Archive Products for its own purposes, including research' (European Centre for Medium-Range Weather Forecasts 2016d).

The ERA-Interim project was produced to replace a previous dataset called ERA-40 and the ERA-Interim dataset is a global atmospheric reanalysis (Dee et al. 2011). The dataset was produced using the 2006 release of the ECMWF forecast model IFS (release CY21R2) (European Centre for Medium-Range Weather Forecasts 2016c). The forecast includes a 4-dimensional variational analysis, with a 12 hour analysis window, with one update per month and a delay of 2 months allowing for quality assurance. The dataset is provided at a spatial resolution of approximately 80 km on 60 vertical levels from the surface. The dataset provides data global reanalysis from

January 1979 till present. For this research all data available for the catchment area was downloaded for the period 1986-01-01 to 2014-12-31. Figure 18 shows the full request, all the parameters available were downloaded as feature selection scenario B is used to identify which parameters influence the performance of the ANN. A full list of all downloaded parameters including their short name, unit of measure and data type is provided in Table 21 in the annex. There is a total of 82 parameters downloaded. This list is then refined during the data cleaning process which is discussed in section 4.5. The data is retrieved in NetCDF format, and is read using the NetCDF library described in section 4.3.2.

netcdf

Final request

Stream:	Atmospheric model
Area:	28.0°S 28.0°E 30.0°S 30.0°E
Parameter:	10 metre U wind component, 10 metre V wind component, 10 metre wind gust since previous post-processing, 2 metre dewpoint temperature, 2 metre temperature, Boundary layer dissipation, Boundary layer height, Charnock, Clear sky surface photosynthetically active radiation, Convective available potential energy, Convective precipitation, Convective snowfall, Downward UV radiation at the surface, Eastward gravity wave surface stress, Eastward turbulent surface stress, Evaporation, Forecast albedo, Forecast logarithm of surface roughness for heat, Forecast surface roughness, Gravity wave dissipation, High cloud cover, Ice temperature layer 1, Ice temperature layer 2, Ice temperature layer 3, Ice temperature layer 4, Instantaneous eastward turbulent surface stress, Instantaneous moisture flux, Instantaneous northward turbulent surface stress, Instantaneous surface sensible heat flux, Large-scale precipitation, Large-scale precipitation fraction, Large-scale snowfall, Low cloud cover, Maximum temperature at 2 metres since previous post-processing, Mean sea level pressure, Medium cloud cover, Minimum temperature at 2 metres since previous post-processing, Northward gravity wave surface stress, Northward turbulent surface stress, Photosynthetically active radiation at the surface, Runoff, Sea surface temperature, Sea-ice cover, Skin reservoir content, Skin temperature, Snow albedo, Snow density, Snow depth, Snow evaporation, Snowfall, Snowmelt, Soil temperature level 1, Soil temperature level 2, Soil temperature level 3, Soil temperature level 4, Sunshine duration, Surface latent heat flux, Surface net solar radiation, Surface net solar radiation, clear sky, Surface net thermal radiation, Surface net thermal radiation, clear sky, Surface pressure, Surface sensible heat flux, Surface solar radiation downwards, Surface thermal radiation downwards, TOA incident solar radiation, Temperature of snow layer, Top net solar radiation, Top net solar radiation, clear sky, Top net thermal radiation, Top net thermal radiation, clear sky, Total cloud cover, Total column ice water, Total column liquid water, Total column ozone, Total column water, Total column water vapour, Total precipitation, Volumetric soil water layer 1, Volumetric soil water layer 2, Volumetric soil water layer 3, Volumetric soil water layer 4
Dataset:	interim_daily
Step:	12
Version:	1
Type of level:	Surface
Time:	12:00:00
Date:	19860101 to 20141231
Grid:	0.25° x 0.25°
Type:	Forecast
Class:	ERA Interim

The status of the request is: **active**

Figure 18: ERA-Interim dataset download request (European Centre for Medium-Range Weather Forecasts 2016c)

4.5. Data preparation and rolling windows

Data preparation is crucial when using ANN's, and could be considered one of the most important steps in the process of developing a neural network to solve real-world problems (Engelbrecht 2007). The first step involves deciding what data is used as inputs and outputs in the ANN. In this research the river flow gauge value is the output and the inputs include the naive prediction described in section 4.7 in conjunction with the weather forecast data from ECMWF. Finding the best scenario data inputs are handled through the various scenarios in chapter 5, this implies all data needs to be prepared upfront even though it may not be used in

the best case scenario in scenario D. Data preparation then involves removing outliers, handling missing values, transferring non-numeric data to numerical data and lastly scaling the data to the active range for the selected activation function (Engelbrecht 2007). Each of these steps are outlined in sections 4.5.1 to 4.5.4.

4.5.1. Handling outliers

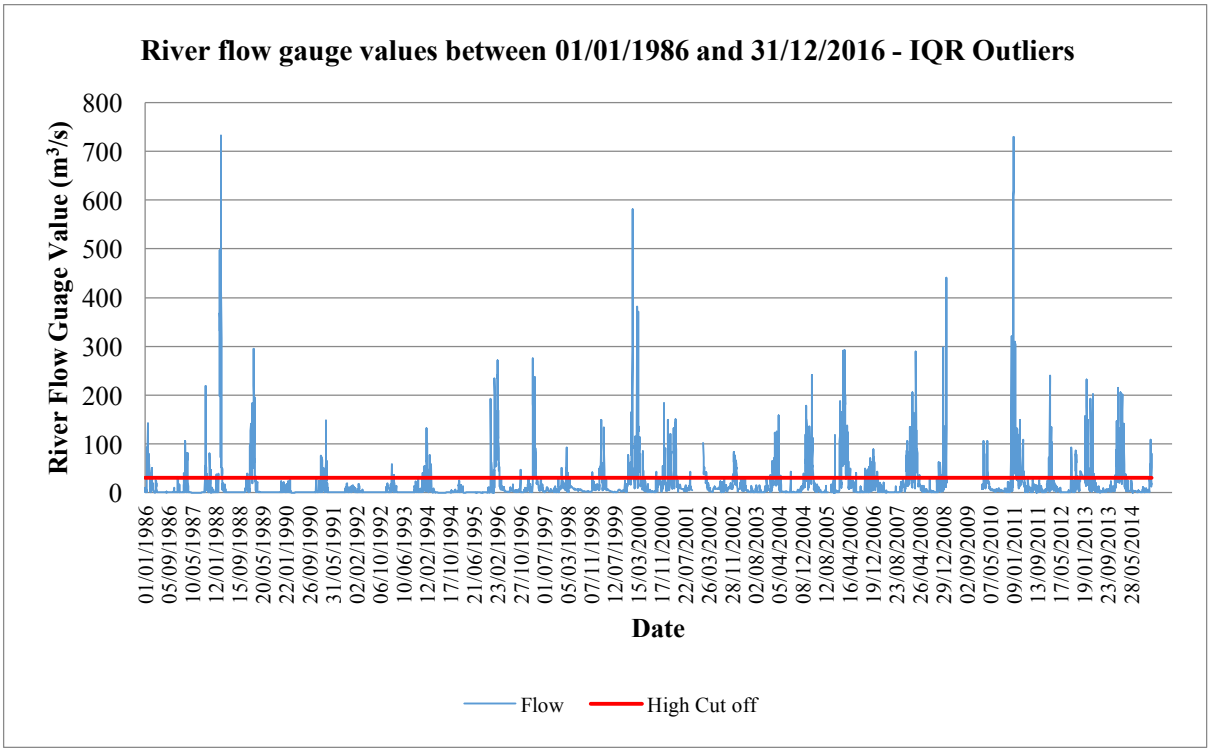
Outliers were identified in the river flow gauge values as this was the output space and would be used in performance calculations. Two approaches were needed to handle outliers. The first approach was the Box plot approach, which is used to visually identify possible outliers (Seo & Bae 2013). The approach makes use of interquartile ranges to identify possible outliers. The method uses five values to generate a lower and upper cut off value. The method requires the minimum value, the maximum value, lower quartile value (Q1), upper quartile value (Q3) and the median value (Q2). Then from these values an interquartile range IQR is calculated as IQR= Q3 – Q1. Outliers are then detected by using the following rule:

1. Data points higher than Q3 + 1.5*IQR are considered outliers.
2. Data points lower than Q1 – 1.5*IQR are considered outliers

Table 6 provides the calculated values for this approach. The approach was found to be extremely strict and classified a large number of values as outliers. Between 01 January 1986 to 31 December 2014 there are 10592 data points. Using the above Box plot interquartile ranges 1286 values would be classified as outliers which is over 12% of the available data. Graph 1 shows a plot of river flow gauge values over time, the red horizontal line shows the upper cut off identifying outliers.

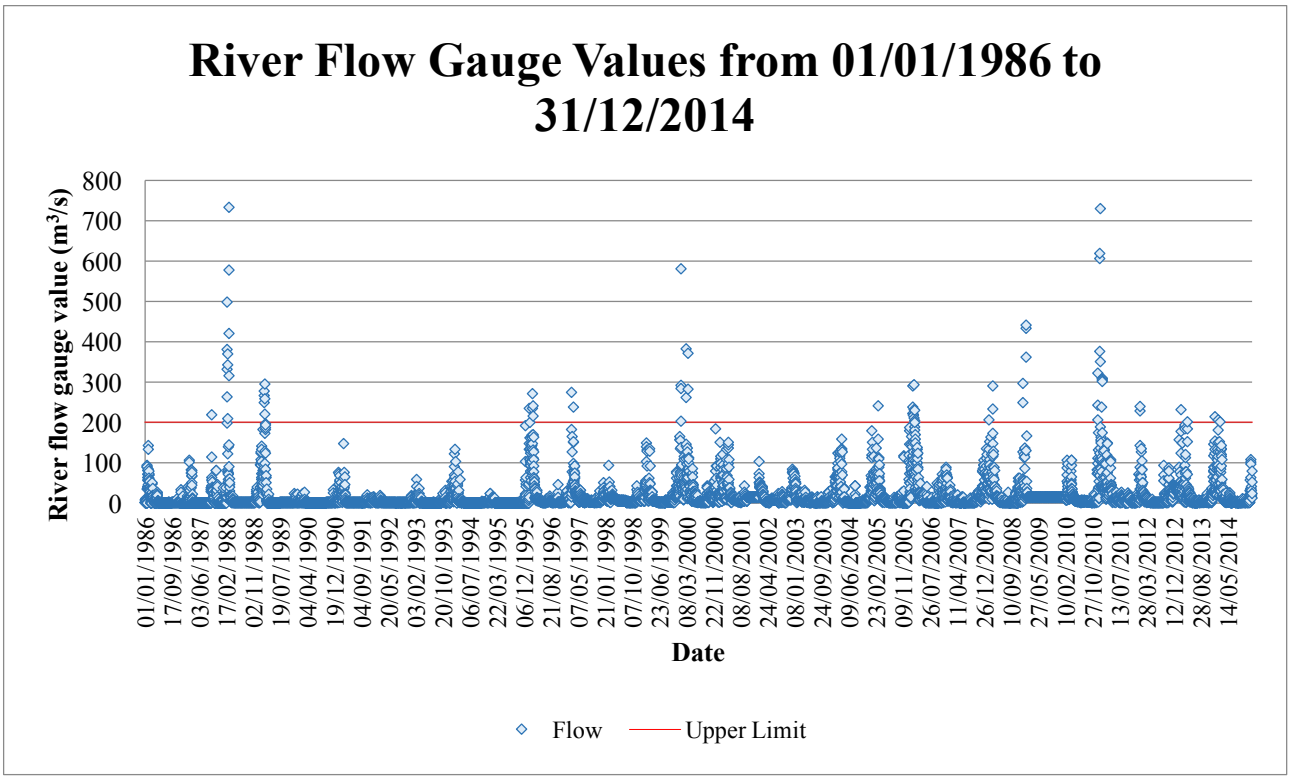
Table 6: Box Plot IQR Values

Description	**Value**
Minimum	*0,027m^3/s*
Maximum	*732,074 m^3/s*
Quartile 1 (Q1)	*1,076 m^3/s*
Median (Q2)	*3,375 m^3/s*
Quartile 3 (Q3)	*13,136 m^3/s*
IQR	*12,06 m^3/s*
Lower Outlier Cut-off	*-17,014 m^3/s*
Upper Outlier Cut-off	*31,226 m^3/s*



Graph 1: River flow gauge values from 01/01/1986 to 31/12/2016 - IQR Outliers

The Box plot approach was too strict and classified too many values as outliers. A more graphical discretionary approach was required to identify outliers. A basic approach was made by producing a scatter plot of river flow gauge values between 01/01/1986 and 31/12/2014. From this plot a discretionary upper limit of 200 m^3/s was selected. Graph 2 shows the scatter plot with the red horizontal line representing the discretionary cut-off for outliers. Using a discretionary upper limit cut off of 200m^3/s means that 73 values are classified as outliers. That represents less than 0,7% identified as outliers. This approach only removes the extreme values in the data set such as 723,074 m^3/s on 11/03/1988, which is more than 18 500% more than the data set median. These outliers were removed from the dataset to ensure they did not increase the error during training and testing. After removing these values, the dataset was reduced to 10519 data points.



Graph 2: River flow gauge values from 01/01/1986 to 31/12/2014 - Discretionary Outliers

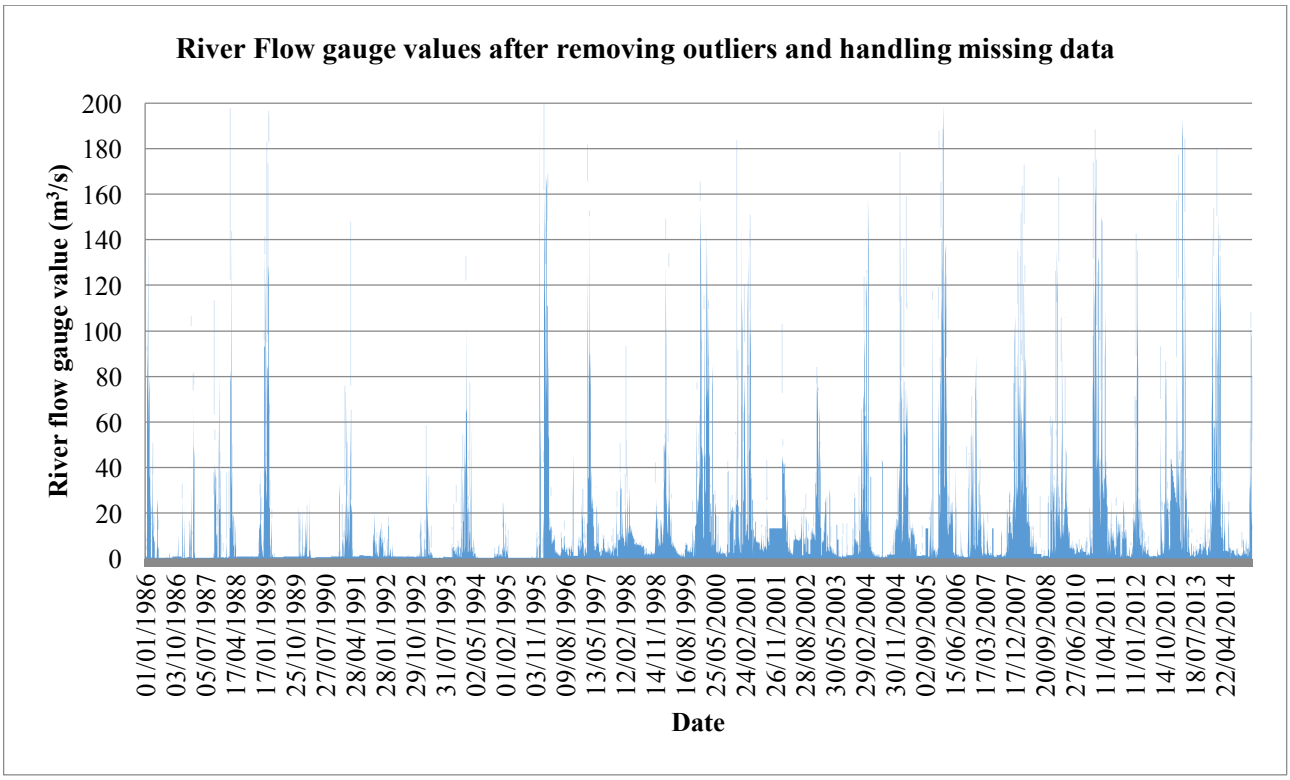
4.5.2. Missing values

The DWS and ECMWF datasets were processed differently when handling missing data. The ECMWF was the easiest to handle as the NetCDF library used a provided method called 'ScaleMissing' which processes missing attributes while reading the data file and automatically converts the data (Unidata 2016a). The NetCDF library makes use of a 'FillValue' to fill the missing values as part of the ScaleMissing process (Unidata 2013). This 'FillValue' is defined for a variable and will be the same type as the variable. The 'FillValue' can be set in code, but can also be left for the library to handle through the default fill value for the type of variable. In this research the default fill value for the variable was utilized. Once the data is read out of the NetCDF file into java the library handles missing values and therefore there is no need to manually cater for missing values in the ECMWF dataset. The only missing values that needed to be considered were those ECWMF parameters that have no value for the catchment area. Of the 82 possible parameters, 7 parameters were not applicable for the catchment area:

- **ci:** Sea-Ice cover
- **sst**: Sea surface temperature
- **istl1**: Ice temperature layer 1
- **istl2**: Ice temperature layer 2
- **istl3**: Ice temperature layer 3
- **istl4**: Ice temperature layer 4
- **chnk**: Charnock

The DWS data set required additional work to handle missing values. The dataset had a few missing values. Before removing the outliers, the data set had data with both quality ratings of '60' and '170'. As per Table 20, the values with a quality rating of '60' means "Above Rating" and a quality rating of '170' means "Permanent Gap" . The original downloaded data from the DWS included 527 values with a quality rating of '170' and 3 values with a quality rating of '60'. After removing the outliers from the data set as described in section 4.5.1 the dataset of 10519 points included only the 527 values with a quality rating of '170'. The 527 values in the dataset make up approximately 5% of the dataset and should not be completely removed from the training set. The approach used in this research was to replace each of the missing values with the average of the remainder of the dataset as this will introduce no bias into the training set (Engelbrecht 2007).

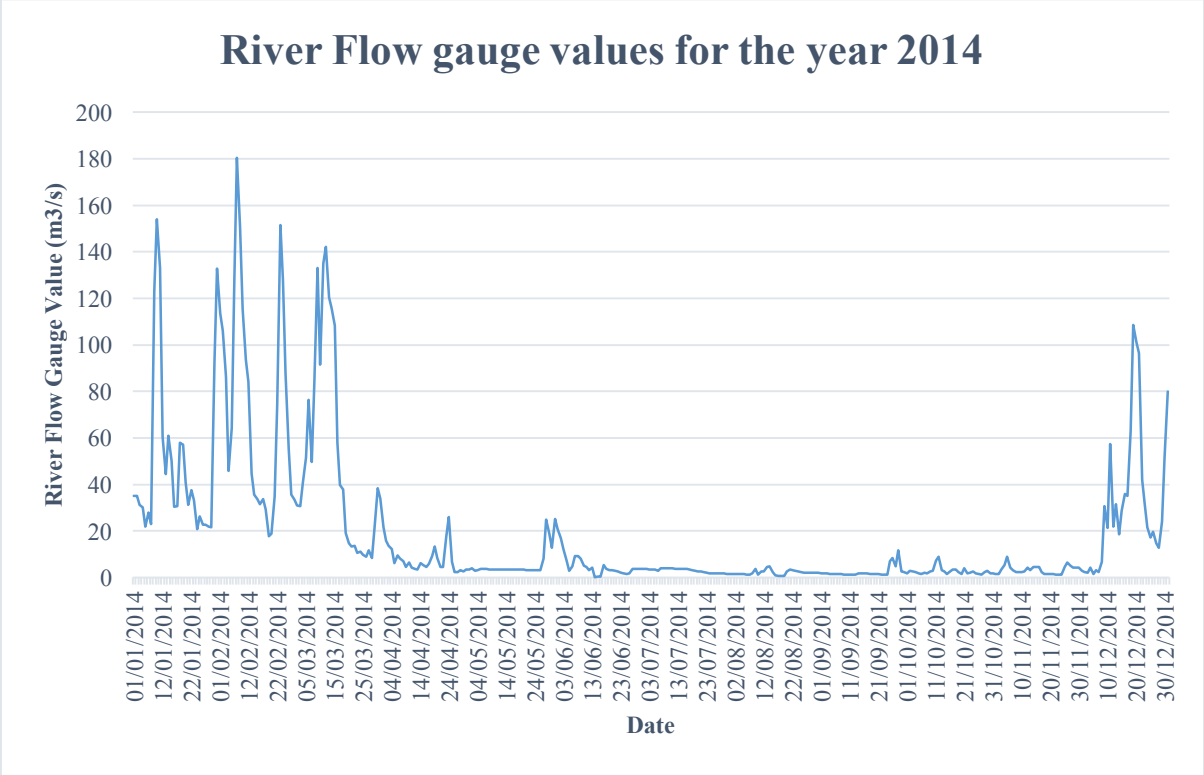
After further investigation of the dataset there was a long period in 2009/2010 where there is a permanent gap in the data, from 10/02/2009 to 09/02/2010. This could introduce an issue in training if the values are replaced with an average. The decision was made to remove these 365 values from the dataset, leaving 162 missing values throughout the dataset. These 162 values were replaced with the average of the full dataset. The average was calculated as 13,498 and all 162 missing values were replaced with this average. After removing the outliers and then handling missing values, the DWS dataset had 10154 days of river flow gauge values. The ECMWF data would be mapped to the DWS dataset therefore automatically dropping values that had been removed from the DWS data in this step. Graph 3 represents the full period of river flow gauge values after removing outliers and processing missing data.



Graph 3: River Flow gauge values after removing outliers and handling missing data

4.5.3. Transferring non-numeric data

Non-numeric data needs to be transformed into numeric input data for the ANN. In this research the data from both data sources, namely the DWS and ECMWF are already numeric values. Additional non-numeric data will be added to the input pattern in the form of a month indicator. Looking at Graph 4 (the year of 2014) it is clear to see that seasons have a major impact on the river flow gauge value. For this reason, a season indicator in the form of a month indicator has been added to the neural network input pattern. According to Engelbrecht (2007) a nominal input parameter (with n different values) should be transformed to n different input patterns. Using this method, the month indicator adds 12 additional input values into the dataset, and will be transformed as shown in Table 7.



Graph 4: River Flow gauge values for the year 2014

4.5.1. Rolling windows

The neural network needs to be able to have a view of the input history. As an example rainfall in the catchment area today may not increase the river flow gauge value today. The effects may be delayed and only increase the river flow gauge values at a later stage. Research done on other rivers, such as the Zambezi river, have shown that floodwaters can take 4-6 weeks to pass through some of the flood plains before peaking down stream discharge rates (Beilfuss 2012). To handle this historical data view of the input values the concept of rolling windows was implemented. The rolling windows approach takes the average of a specific number of days and uses this as a possible input into the neural network. This research made use of the rolling window periods of 7 days, 30 days, 90 days, 180 days, 365 days and 1095 days. At 1095 days this provides 3 years' worth of historical data, which would be able to detect a phenomenon such as a long term drought. A window of 30 to 180 days would be able to contain data to detect

a shorter term drought or delayed river flow impact. Figure 19 represents how each input value, apart from the month indicator, generates 7 actual inputs into the neural network. When creating these rolling windows you reduce the size of the dataset as three years of data is condensed into one value. Since the longest rolling window is 1095 days you essentially condense 1095 of the 10154 days of river flow gauge values. The information is still included in the dataset though, just in an averaged format. After creating these rolling windows due to missing values and the rolling windows the number of values is reduced to 9059.

Table 7: 12 Additional inputs to transform the non-numeric month values

Month	**Input 1**	**Input 2**	**Input 3**	**Input 4**	**Input 5**	**Input 6**	**Input 7**	**Input 8**	**Input 9**	**Input 10**	**Input 11**	**Input 12**
January	***1***	0	0	0	0	0	0	0	0	0	0	0
February	0	***1***	0	0	0	0	0	0	0	0	0	0
March	0	0	***1***	0	0	0	0	0	0	0	0	0
April	0	0	0	***1***	0	0	0	0	0	0	0	0
May	0	0	0	0	***1***	0	0	0	0	0	0	0
June	0	0	0	0	0	***1***	0	0	0	0	0	0
July	0	0	0	0	0	0	***1***	0	0	0	0	0
August	0	0	0	0	0	0	0	***1***	0	0	0	0
September	0	0	0	0	0	0	0	0	***1***	0	0	0
October	0	0	0	0	0	0	0	0	0	***1***	0	0
November	0	0	0	0	0	0	0	0	0	0	***1***	0
December	0	0	0	0	0	0	0	0	0	0	0	***1***

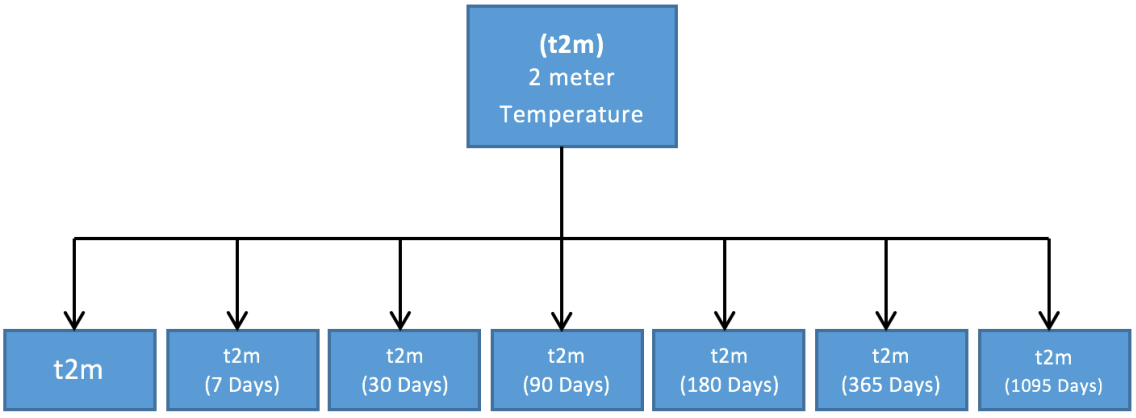


Figure 19: 2 Meter Temperature rolling window example

The 75 ECMWF parameters along with the one naive river flow gauge value, means a total of 76 input values. Using the rolling window approach these 76 input values create 532 possible input values. To these 532 possible input values, the month indicators are added as described in section 4.5.3, to give a total of 544 values per pattern. With 9059 values, and each day having 544 points, the computational complexity increases dramatically as there are 4,928,096 (544 points × 9059 values) unique data points . The

feature selection scenarios in Chapter 5 aim at reducing this computational complexity by identifying the ECMWF weather parameters that improve the performance of the ANN model.

4.5.2. Scaling

One of the most important steps in data preparation for neural networks is to scale the data to the active domain and range of the activation function (Engelbrecht 2007). The most important data to scale is the output/desired values as these could fall outside the range of the activation function. Scaling the input data to the active domain of the activation function is not mandatory as it is for the output values, but performance can be improved if the input data is scaled. This research scales both the input data and the output data to ensure the data is scaled for the active domain of the activation functions. This is to ensure the best possible performance of the ANN model. While the active domain of the sigmoid function is $[-\sqrt{3}, \sqrt{3}]$ (Engelbrecht 2007), testing showed that in this research scaling the input data to (0,1) gave better results through trial and error, see Table 22 in the annex. The output data in this research is scaled to the output range of the sigmoid function, $(0,1)$. Scaling is done using Equation 3.

$$t_s = \frac{t_u - t_{u,min}}{t_{u,max} - t_{u,min}} \left(t_{s,max} - t_{s,min}\right) + t_{s,min}$$

where

$$\boldsymbol{t_u} = unscaled\ value$$
$$\boldsymbol{t_{u,max}} = maximum\ value\ of\ the\ unscaled\ data$$
$$\boldsymbol{t_{u,min}} = minimum\ value\ of\ the\ unscaled\ data$$
$$\boldsymbol{t_s} = new\ scaled\ value$$
$$\boldsymbol{t_{s,max}} = required\ maximum\ value\ of\ the\ scaled\ data$$
$$\boldsymbol{t_{s,min}} = required\ minimum\ value\ of\ the\ scaled\ data$$

Equation 3: Scale data to a [min, max] range

4.6. Neural network

4.6.1. Introduction

The brain has the ability to recognize patterns, learn, memorize and yet still generalize in a wide variety of circumstances. The ability for the brain to do this drove research into ANN's which attempts to model biological neural systems (Engelbrecht 2007). It is accepted that there is currently no way of truly modelling the human brain due to the complexity of how neurons are arranged. Research has shown though that it is possible to solve single objective problems using ANN's such as speech recognition, data mining, image processing, classifications and even composing music. It is the neural networks ability of time series modeling that this research takes advantage of, in order to predict river flow gauge values. Neural networks are adaptive statistical models and are used as statistical tools in many fields such as psychology, statistics, physics, engineering and many more (Kröse & van der Smagt 1991). This section will provide the background into the structure of neural networks and the related equations in section 4.5.2. In section 4.5.3 accuracy and performance measures will be explained, while

section 4.5.4 will outline some concerns and common problems associated with neural networks. Finally, section 4.5.5 will describe the neural network used in this research.

4.6.2. Artificial neural networks

The biological neural system is made up of billions of neurons interconnected by synapse (Engelbrecht 2007). Signals propagate from a neuron to all interconnected neurons, and this signal can either excite or inhibit the signal to all its connected neurons and so on. In ANN's these neurons are represented by an artificial neuron (AN). The input signals in an AN are excited or inhibited by weighted connections. The AN collects all weighted input signals and calculates an exiting signal strength via an activation function which it passes onto connected AN's. This structure then represents the biological neuron, signal propagation and replicates neurons inhibiting or exciting the signal it passes on. Figure 20 represents an AN.

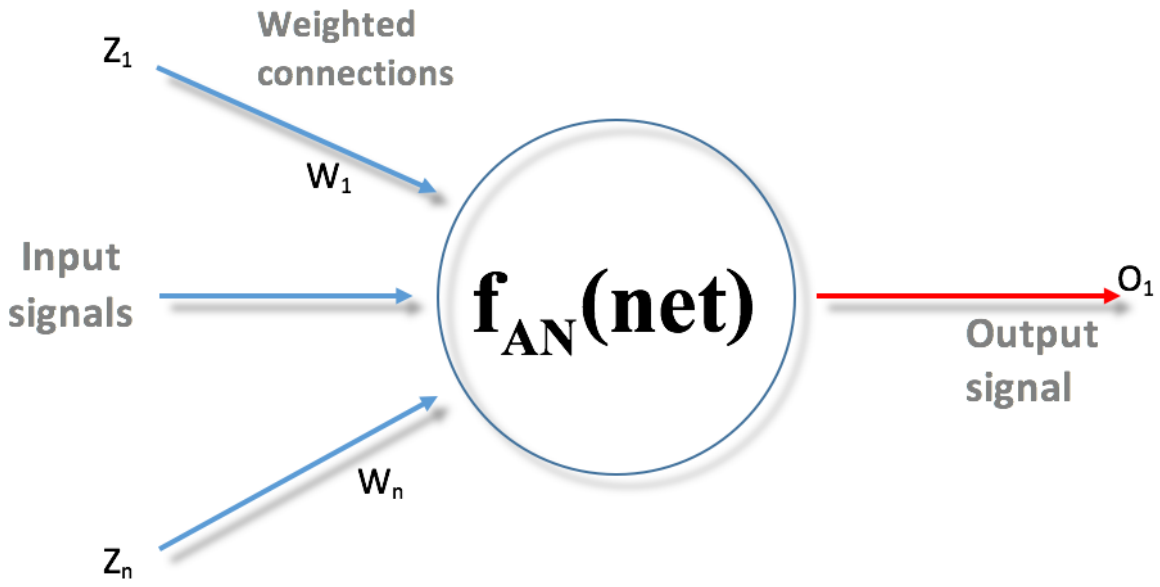


Figure 20: Single Artificial Neuron

As can be seen in Figure 20 a neuron has an activation function $f_{AN}()$ to produce the output signal based on the net input signal into the neuron. A neuron can have any number of input signals and the net input signal into the neuron's activation function is typically calculated as the weighted sum of all the input signals (Equation 4 where z is the input signals, w is the weighted connections and n is the number of inputs).

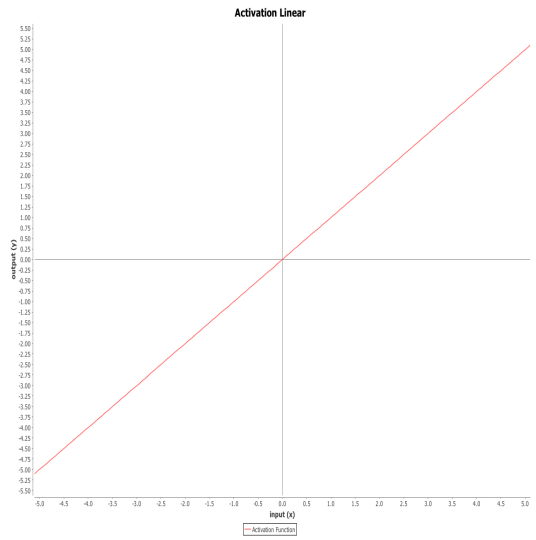
$$net = \sum_{i=1}^{n} z_i w_i$$

Equation 4: Artificial neuron net input signal

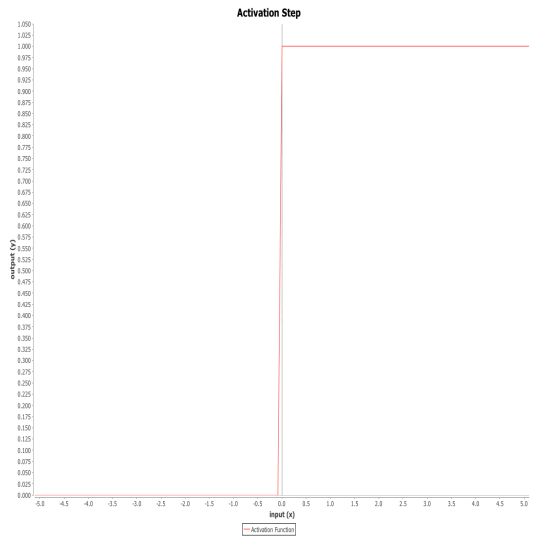
This net input signal is passed into the activation function $f_{AN}()$ to produce the output signal (O_1 in Figure 20). There are a large number of activation functions that can be

used. Most activation functions produce a mapping (or output) to [0,1] or [-1,1] (with the exclusion of the linear function), meaning $f_{AN}(-\infty) = 0$ or $f_{AN}(-\infty) = 1$ and $f_{AN}(\infty) = 1$. Some of the more frequently used activation functions are shown in Table 8 and include linear activation function, Graph 5; step activation function, Graph 6; sigmoid activation function, Graph 7 and Gaussian activation function, Graph 8.

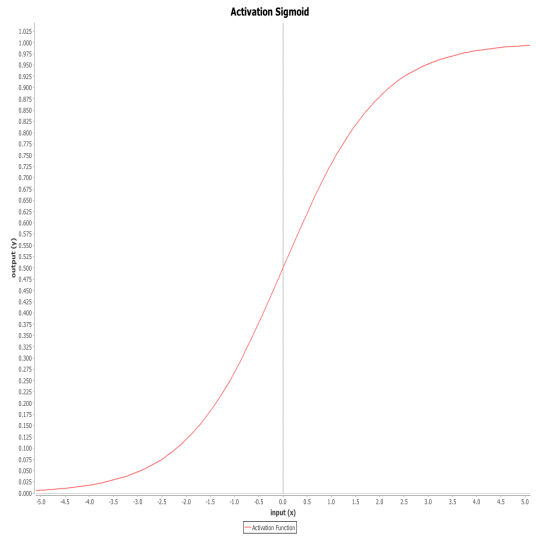
Table 8: Activation functions



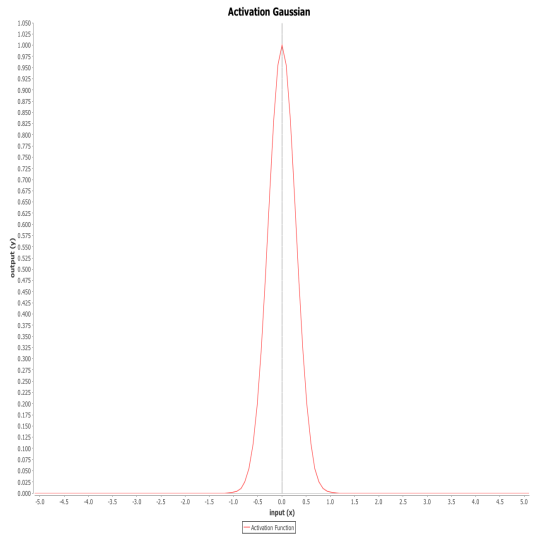
Graph 5: Linear activation function



Graph 6: Step activation function



Graph 7: Sigmoid activation function



Graph 8: Gaussian activation function

AN's in a layered network structure make up what is called an ANN (Engelbrecht 2007). The network consists of a number of layers, with each layer having a set number of AN's. Early neural networks were restricted to linear problems, but it was found that by adding additional layers this could be overcome (Kröse & van der Smagt 1991). ANN's consist of an input layer, a number of hidden layers and an output layer. The

neurons in one layer are either fully or partially connected to the neurons in the next layer and in some network structures feedback connections to previous layers can also be added. Figure 21 below depicts a simple ANN structure with only 1 hidden layer and no feedback connection. As described previously each connection is weighted and each neuron has an activation function to produce an output signal.

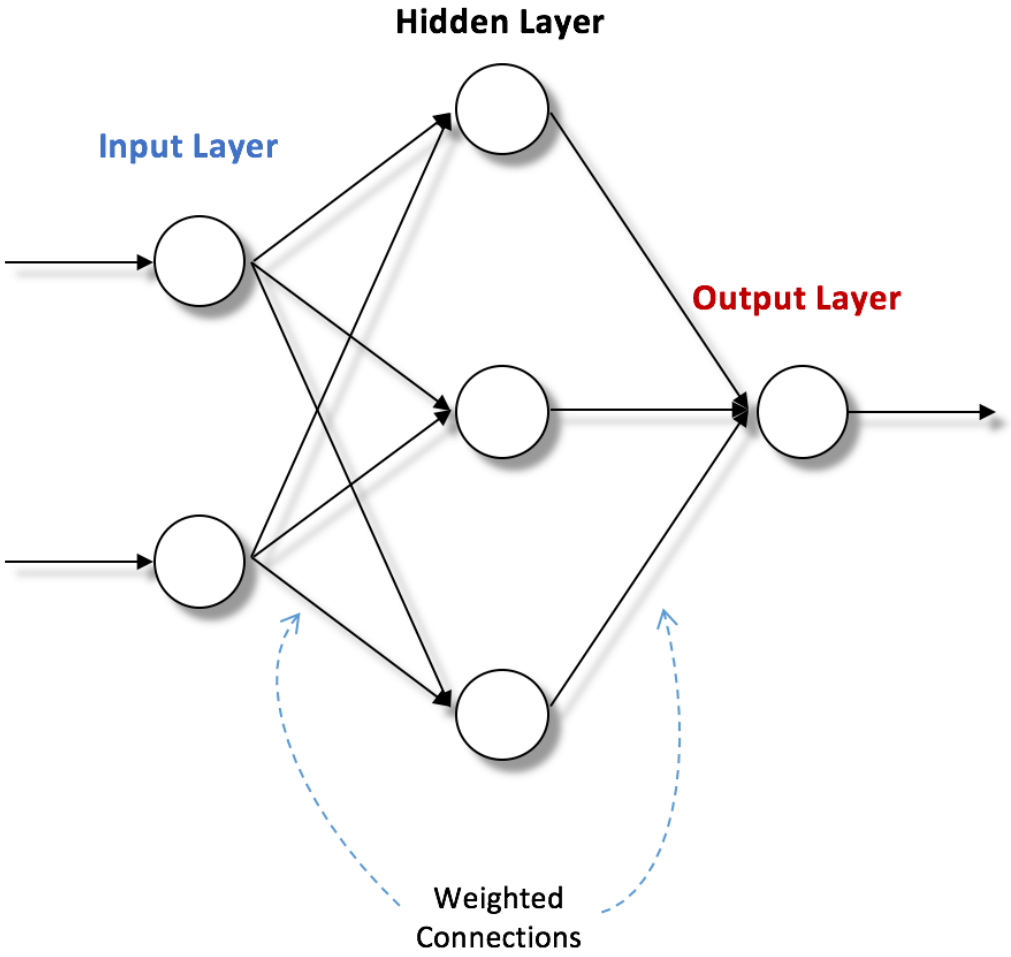


Figure 21: Artificial Neural Network Layers

There are many different ANN types that have been developed for different reasons with different advantages and disadvantages. Some examples include multilayer feedforward neural networks, temporal neural networks, self-organizing neural networks and combined supervised and unsupervised neural networks (Engelbrecht 2007). Figure 21 shows a simple feedforward neural network, in comparison to this Figure 22 shows a simple recurrent neural network. The simple recurrent neural network has feedback connections to the input layer, this allows the neural network to learn temporal characteristics of the data set.

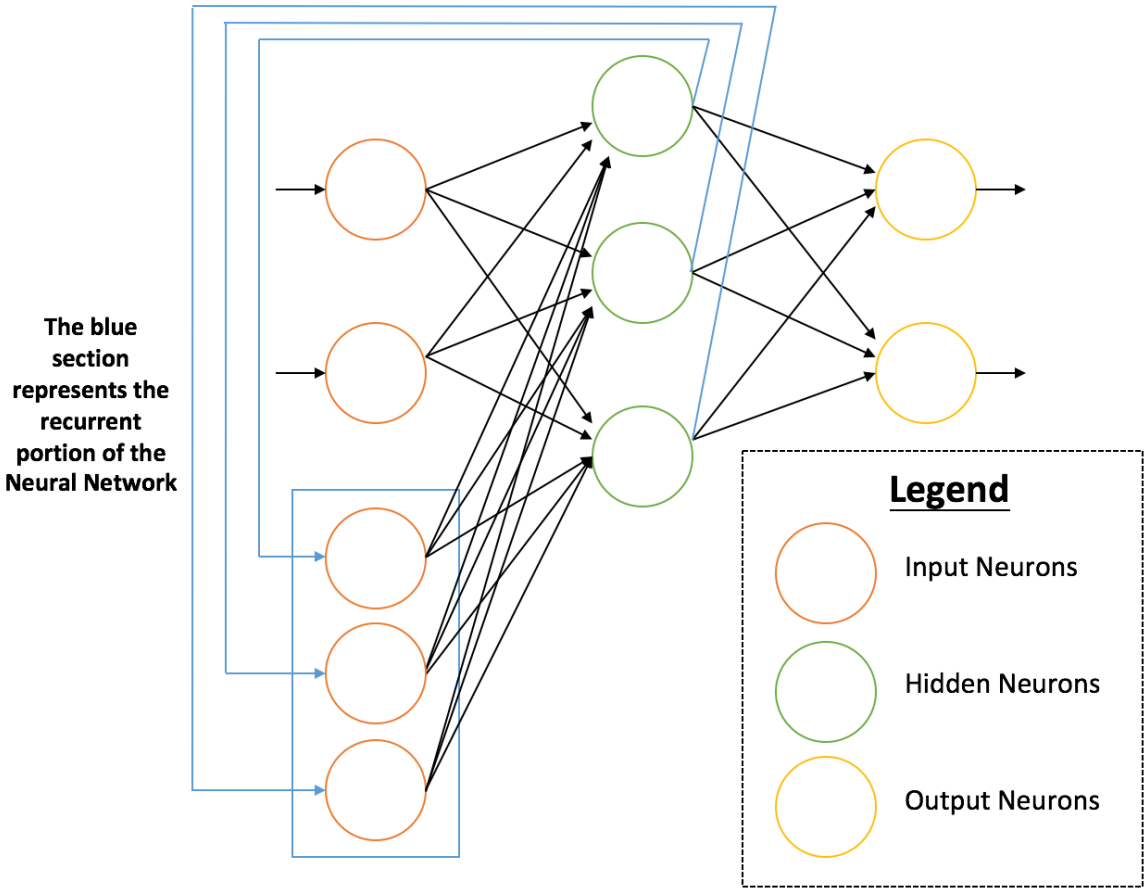


Figure 22: Simple recurrent neural network

When considering the architecture of inputs and outputs in an ANN, an ANN is a nonlinear mapping of real numbers from the input space to the target or output space. In this research the neural network maps the inputs of river flow gauge values and weather values to the output space of future river flow gauge values. Once the structure of a neural network has been defined then a neural network must adjust the values of the weighted connections to excite certain inputs and inhibit other inputs to best map the input space to the output space. This process is referred to as learning, the ANN must learn the best value for the weights (Engelbrecht 2007). This process is done by constantly adjusting the weights until a set number of criteria are satisfied. There are three main types of learning. These include supervised learning, unsupervised learning and reinforcement learning. Supervised learning makes use of a training data set that has a number of inputs and their target outputs, the weights are adjusted to minimize the difference between the target output in the training set, and the ANN mapping output. Unsupervised learning is used to discover patterns with no additional support from external influences. Reinforcement learning makes use of a reward and penalty approach where parts of a neural network are rewarded for good performance and penalized for poor performance.

The structure of the ANN has now been discussed and the different approaches to learning have been discussed but what is still to be explained is the approaches for

updating the weighted connections. There are many different approaches to updating the connection weights and this is an area where constant improvement and research is being done (Leerink et al. 1995; Engelbrecht & Ismail 1999; van den Bergh & Engelbrecht 2001) . One of the most common approaches is called gradient descent (GD), which is based on the error approximating the target output by using an error function. A common error function used is the sum of squared, errors Equation 5 (Engelbrecht 2007). At a high level the goal of the GD is to adjust the weights to minimize the error, this is done by calculating the gradient of the error calculated in Equation 5 , and moving along the negative gradient. The minimum point on the error function is then the optimum weights, Figure 23.

$$\varepsilon = \sum_{i=1}^{p} (t_i - o_i)^2$$

Where

p *= total number of input-output pairs in the training set*
t_i *= target output (i-th pair)*
o_i *= actual output (i-th pair)*

Equation 5: Sum of squared error

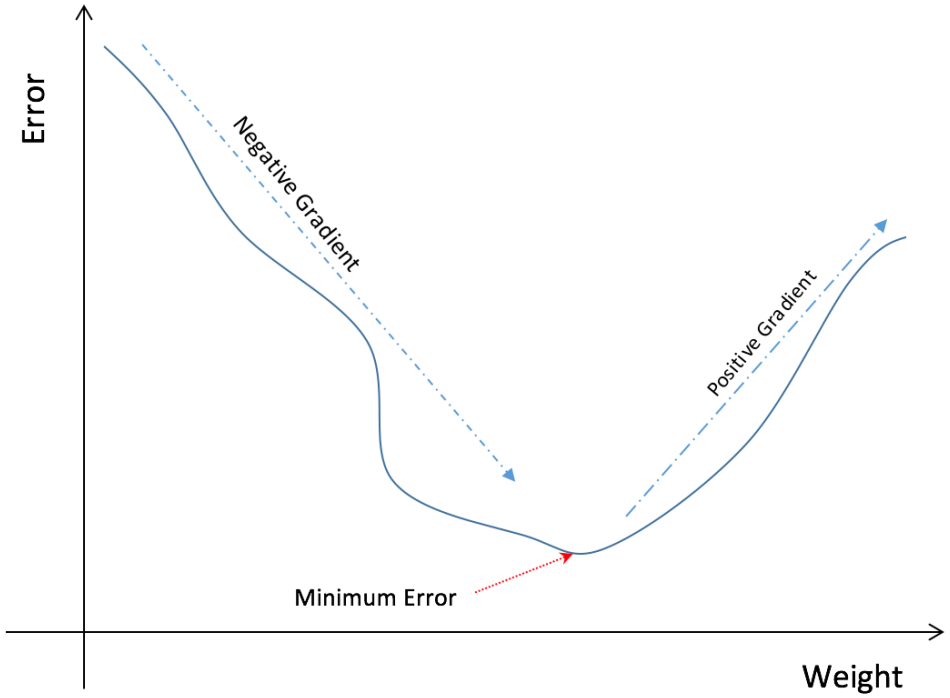


Figure 23: Gradient descent learning rule

GD is one of the more common approaches there are for weight updating, many other approaches, for example the Widrow-Hoff learning rule and the Error-Correction learning rule. This research made use of an approach called Resilient Propagation.

Resilient propagation (RPROP) is a local adaptive learning approach that was introduced by Riedmiller and Braun (Riedmiller & Braun 1992; Riedmiller & Braun

1993). The RPROP approach adjusts the weight step by using the local gradient and is done in the form of reward or punishment (Engelbrecht 2007). The weights are adjusted based on the gradient calculated using partial derivatives. If the partial derivatives of the weights change sign (for example a negative to a positive) then the algorithm could have jumped over a local minimum error. In this case the weights are penalized. On the other hand, if the sign of the partial derivative stays the same then the weights are moving in the correct direction and they are rewarded to speed up step sizes to accelerate convergence. This approach is to overcome the harmful influences of the partial derivatives in the backpropogation algorithm where the weight step is then too large or too small (Santra et al. 2009). RPROP does have a few values to configure and these include (Engelbrecht 2007):

1. Step increase/decrease (reward/penalize) amount
2. First weight step, but it has been shown that the performance of RPROP is insensitive to the first weight step value (Riedmiller & Braun 1993).
3. Upper and lower limits on the step size, which basically sets the maximum and minimum step values that can be set.

RPROP is an efficient algorithm that rewards convergence but penalizes the algorithm for stepping over local minimums ensuring efficient steps to convergence. This training algorithm is offered by the Encog framework, and the training algorithm is multithreaded which reduces the run time required during training.

This section has described the basics of an ANN, this includes the structure and learning. In summary a neural network is made up of layered AN's that attempt to represent the brain's neurons and synapses. The ANN attempts to map an input space to an output space by adjusting the weights of the connections between the neurons to perform this mapping as accurately as possible. In this research a multilayered feedforward neural network is used, with the sigmoid activation function and resilient propagation as the learning algorithm. The next section will describe how the accuracy and performance of a neural network can be monitored.

4.6.3. Accuracy and performance measures

The performance of ANNs is not only defined by how well it can map the input space to the output space. The performance of the neural network must be based on how well it can map points not used during the training, this is referred to as generalization (Engelbrecht 2007). A neural network may have a low training error, but when presented with values not included in the training data set it cannot perform the mapping. In this case the ANN is said to have bad generalization, and the main cause for this is the neural network memorizes the training data which is called overfitting. Overfitting is normally caused when either the ANN is trained for too long or the neural network architecture is too large with too many hidden neurons and irrelevant input values. There are a number of ways for reducing the risk of overfitting, the approaches used in this research are described in section 4.6. For this section what needs to be understood is that the accuracy of the neural network needs to be considered for both values in the training set and those not included in the training set. To consider both of these in the research the training error and the generalization error will be calculated. The most common accuracy calculation, and used in this research, is mean squared error (MSE) shown in Equation 6 (Engelbrecht 2007). MSE is calculated using the training set to calculate the training error, and the generalization error is calculated using a data set not used for training.

$$\varepsilon = \frac{\sum_{p=1}^{P}\sum_{k=1}^{K}\left(t_{k,p} - o_{k,p}\right)^2}{PK}$$

Where:

$P = Total\ number\ of\ training\ patterns/test\ patterns$

$K = Number\ of\ output\ units$

$t = target\ value$

$o = output$

Equation 6: Mean square error (MSE)

There are a number of additional performance measures that can provide insight into the performance and accuracy of the ANN. This research makes use of correlation coefficient (R), root mean squared error (RMSE), Nash-Sutcliffe efficiency coefficient (NS) and mean absolute relative error (MARE).

The correlation coefficient (Equation 7) quantifies the degree to which two variables are linearly related, and in the case of the neural network output the correlation coefficient measures the linear relationship between the output values and the true values (He et al. 2014).

$$R = \frac{\sum_{i=1}^{n}(Q_i^o - \overline{Q^o})\left(Q_i^p - \overline{Q^p}\right)}{\sqrt{\sum_{i=1}^{n}\left(Q_i^o - \overline{Q^o}\right)^2 \sum_{i=1}^{n}\left(Q_i^p - \overline{Q^p}\right)^2}}$$

Where

$n = number\ of\ input\ patterns$

$Q_i^o = observed\ river\ flow\ at\ i$

$Q_i^p = predicted\ river\ flow\ at\ i$

$\overline{Q^o} = mean\ observed\ river\ flow$

$\overline{Q^p} = mean\ predicted\ river\ flow$

Equation 7: Correlation Coefficient (R)

RMSE in Equation 8 provides the predictive ability of a neural network (He et al. 2014). RMSE provides a calculation that measures the 'goodness-of-fit'. RMSE has been used as a statistical metric for measuring model performance in a number of study areas including climate research studies and has been shown to be a good metric for analyzing model performance (Chai & Draxler 2014).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_i^o - Q_i^p\right)^2}$$

Where

$n = number\ of\ input\ patterns$

$Q_i^o = observed\ river\ flow\ at\ i$

$Q_i^p = predicted\ river\ flow\ at\ i$

Equation 8: Root mean squared error (RMSE)

Mean absolute relative error (MARE) in Equation 9 also provides a measure of predictive ability, but also includes a measure of the distribution of the prediction error (He et al. 2014). MARE provides an indication if a model overestimates or underestimates.

$$MARE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Q_i^o - Q_i^p}{Q_i^o}\right| \times 100$$

Where

$$n = number\ of\ input\ patterns$$
$$Q_i^o = observed\ river\ flow\ at\ i$$
$$Q_i^p = predicted\ river\ flow\ at\ i$$

Equation 9: Mean absolute relative error (MARE)

The last measure is Nash-Sutcliffe efficiency coefficient (NS) and is given in Equation 10 (He et al. 2014). NS measures the resemblance between the actual observed river flow gauge value and the predicted river flow gauge values. Varying values are considered satisfactory normally from 0.6 – 1 (Santos et al. 2014; DeWeber & Wagner 2014).

$$NS = 1 - \frac{\sum_{i=1}^{n}\left(Q_i^o - Q_i^p\right)^2}{\sum_{i=1}^{n}\left(Q_i^o - \overline{Q^o}\right)^2}$$

Where

$$n = number\ of\ input\ patterns$$
$$Q_i^o = observed\ river\ flow\ at\ i$$
$$Q_i^p = predicted\ river\ flow\ at\ i$$
$$\overline{Q^o} = mean\ observed\ river\ flow$$

Equation 10: Nash-Sutcliffe efficiency coefficient (NS)

This section describes the performance measures used in this research. The MSE is used during training and after training of the neural network to compare the ability of the ANN to generalize. Then the other 4 statistical measures are used to analyze the performance of the neural networks and provide the base for a statistical evaluation criteria scorecard. The best performing neural network would give the following values: R=1, RMSE=0, NS=1 and MARE =0. This would be highly unlikely in most circumstances, especially one with a high computation complexity. For statistical analysis this perfect case is provided as this drives the ability to compare neural networks.

4.6.4. Problems associated with artificial neural networks

ANN's provide an approach for estimating the river flow gauge value as required by this research, but they do pose a number of critical problems and concerns. The first major concern is with the structure itself, how does one find the optimal structure. If

the ANN structure has too many neurons or too many layers then the neural network memorizes the input and output value patterns and will not be able to provide a suitable output value (Engelbrecht 2007). This process is what leads to overfitting and a low generalization ability. This is not the only problem with neural networks that lead to overfitting, if the neural network is trained for too long there will be a point where the training error still reduces but the generalization error for unseen data gets worse (as depicted in Figure 24). Training should be stopped at this point, if not it will be over trained and overfitted (Engelbrecht 2007).

Another problem associated with the structure of the ANN is the weight initialization. The performance of the ANNS is very sensitive to the initial weights. Research has found that setting optimal initial weight parameters has been found to enhance the accuracy of the ANN (Venkatesan et al. 2009; Mulia et al. 2013). Some research makes use of genetic algorithms to try find the optimal initial weights, while other research has utilized an ensemble technique (Kim & Seo 2015). The ensemble technique can be divided into two steps, firstly the creation of the ensemble and secondly producing the most appropriate output from the ensemble. To create the ensemble, a set number of ANN's are created with different initial weights. These ANN's make up the ensemble and each have the same structure, the same data sets and the same learning algorithms. After each has been trained, the output from the ANN ensemble needs to be determined. The most common approaches include, selecting the the ANN in the ensemble with the best generalization performance or calculating the average of the outputs from all the ANN's in the ensemble (Engelbrecht 2007). Using ensemble ANN's the aim is to optimize the results and minimize the effect of the initial weights.

These are not the only problems associated with ANN's, there are also computational complexities that need to be considered. These computational complexities affect hardware requirements for running the neural network, and the number of training iterations required for the neural network to converge. The computational complexity is directly affected by the network architecture that is chosen (Engelbrecht 2007). Basically the more neurons in the network the more calculations are required to update weights and calculate outputs per neuron. This value is also directly related to the number of training input patterns that will need to be passed through the neural network calculations. This is not only the number of patterns but also the number of inputs in each pattern. The final characteristic directly affecting the computational complexity is the optimization method. Very sophisticated calculations and approaches have been made to improve the accuracy and convergence characteristics. These sophisticated approaches can come with heavy computational complexity (Engelbrecht 2007).

In summary the problems associated with ANNs are related to structure, training duration and training set size. When designing the neural network structure, it is important to consider these characteristics. The research made use of strategies such as K-Fold cross-validation, pruning, early stopping and ensemble techniques to overcome these problems. Section 4.6 will describe the ANN configuration for this research and how these strategies were used to handle these problems.

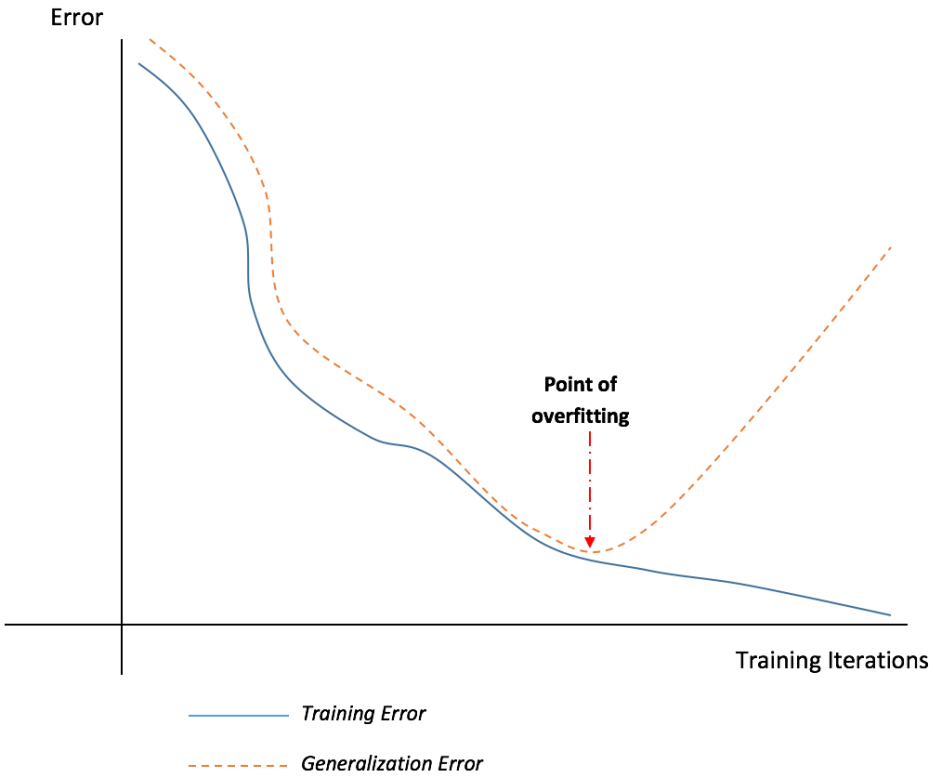


Figure 24: Point of Overfitting

4.7. Research model

The importance of defining the ANN structure, considering the computational complexity and the possibility of overfitting has been explained in section 4.6. This section explains how the model for this research has been defined. This research implemented a resilient propagation, feed forward ANN with a sigmoid activation function and makes use of network pruning, K-Fold cross validation and early stopping strategies. Section 4.6.2 explained that resilient propagation (RPROP) is an adaptive learning approach that adjusts the weight step through a form of reward and punishment (Engelbrecht 2007; Riedmiller & Braun 1992). Section 4.6.2 also explained activation functions, this research makes use of a sigmoid activation function Graph 7. The sigmoid function was selected as an activation function, and as was explained in section 4.6.2 the input and output data was scaled to the active domain of the activation function. For the sigmoid activation function the input values were scaled between $[-\sqrt{3}, \sqrt{3}]$, and the output values scaled between $(0,1)$ (Engelbrecht 2007). A number of ANN strategies are used in this research to handle common concerns and problems associated with ANN's as described in section 4.6.4.

The first strategy used is called network pruning. Network pruning is an approach used to regulate and determine the optimal ANN architecture. The main aim of network pruning is to remove unnecessary network parameters such as individual weights, hidden layers or neurons and in some cases even input units (Engelbrecht 2007). The model in this research made use of the pruning functionality provided by the Encog 3.3 Java library (Heaton 2015). The pruning method in the library is aimed at optimal configuration of the hidden network layers in the

ANN. The pruning method in the library accepts a list of maximum and minimum neurons per hidden layer, and will then attempt to train the neural network at all configurations to find the best neural network architecture. The pruning method keeps a set number of lowest error rate neural network configurations and will then select the one with the fewest connections to be the best neural network architecture. The number of networks to be considered for the 'best' is configurable and in the research, the top 3 performing networks are kept and considered for the 'best' neural network architecture.

The second strategy used in the research is K-Fold cross validation. Cross-validation is used to evaluate the predictive ability of a model and can be used to help with the generalization ability of a neural network (Zhang & Yang 2015). K-Fold cross validation basically makes use of the cross-validation method but the approach splits the training data set into 'k' disjoint folds and in turn each fold is used to measure the predictive ability of the ANN (Wong 2015). The performance of the artificial neural network is calculated as the average of the k accuracies from each k-fold validation. The model used in this research was configured using the Encog 3.3 java library 'CrossValidationKFold' functionality (Heaton 2015). The library implements K-Fold cross-validation by training using a K-Fold approach. Each training iteration will train a set number of times equal to the number of Folds-1. Each sub-iteration will train with all the data minus the fold, and the fold is then used to validate. When using the K-Fold functionality provided by the Encog 3.3 library the error value returned during training always reflects data that was not part of training due to the fold validation on each iteration. Figure 25 graphically shows how the data is split into folds, with multiple sub-iterations and providing a single error value including the validation data from each fold. This does add computational complexity and require, more hardware or longer to train as there are more complex calculations. This added computational complexity has been considered and is deemed necessary to ensure the ANN can generalize. Research has shown that generally 5-Fold cross validation is used (He et al. 2014; Zhang & Yang 2015; Wong 2015).This research will use the same configuration setting as the Encog 3.3 java library to use a 5-Fold implementation.

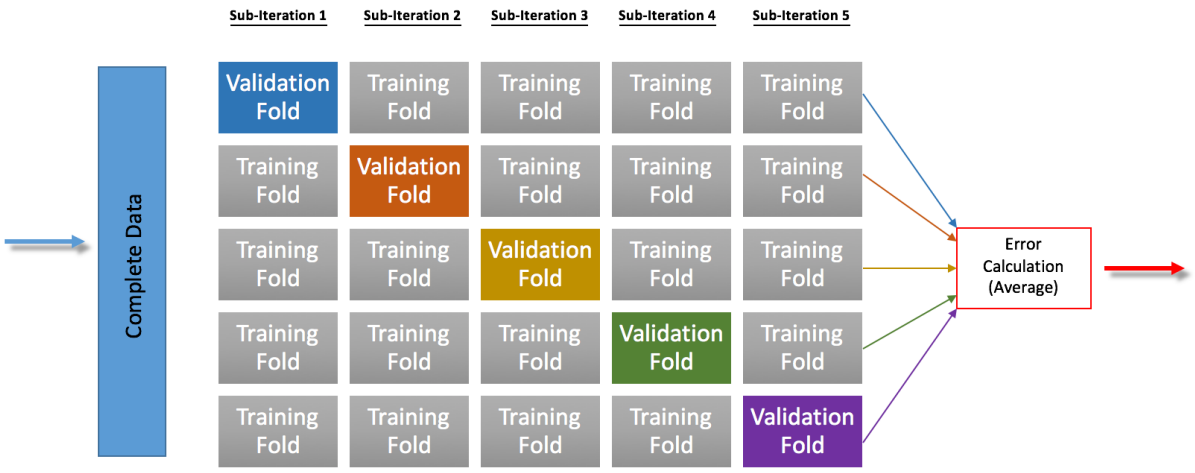


Figure 25: K-Fold Cross validation

The third strategy used in the research model is early stopping, which is used to identify when a neural network is becoming overfitted and will lose the ability to generalize (Engelbrecht 2007). Section 4.6.4 discussed overfitting and Figure 24 graphically represents the point of over fitting that needs to be identified. Early stopping is an approach that stops the neural network before it reaches a convergence point, its maximum iterations or its desired error value.

Early stopping will restrict the neural network from training further and stops the neural network from losing its ability to generalize. The model in this research makes use of an average validation set error and the current validation set error. Overfitting is detected when the current MSE of the validation set is greater than the average MSE of the validation set from the start of training, this is shown in Equation 11. The implemented model makes use of this validation every 5 training iterations to see if the criteria for early stopping has been met.

$$\varepsilon_V > \bar{\varepsilon}_V$$

Where

$$\varepsilon_V = Current\ MSE\ of\ the\ validation\ set$$
$$\bar{\varepsilon}_V = average\ MSE\ of\ the\ validation\ set\ since\ training\ started$$

Equation 11: Early stopping validation - Overfitting

The final strategy used in this research is the ANN ensemble. This technique helps overcome the problem finding the optimal initial weights. The research creates an ensemble of 100 ANN's with the same structure, the same data and the same learning algorithms but different initial weights. The research then selects the ANN that generalizes the best as the output for the ensemble.
The high level programming logic for the model is provided below:

1. ***Load the data into the model***
2. ***Split the data in training, validation and testing data sets***
3. ***Find the best artificial neural network architecture using pruning***
4. ***Create the ensemble of 100 artificial neural networks using the structure found in 3 but with different initial weights***
5. ***Train all 100 artificial neural networks using K-Fold cross validation***
    a. ***Train until either early stopping conditions are met or a maximum of 100 training iterations have been performed***
6. ***Select the artificial neural network from the ensemble with the best generalization ability***
7. ***Generate report providing the require statistical measures of performance***

To be able to asses the accuracy and performance a set number of performance measures need to be calculated for each neural network model produced. These performance measures need to be stored in a scorecard for easy comparison and to ensure the same performance calculations are gathered for each neural network. A scorecard like Table 9 containing information is recorded for each ANN run and can be found in the various scenarios presented in Chapter 5.

This section has described the setup for the ANN model in this research. It has explained a number of implementation strategies and decisions to deal with problems associated with ANN's and has described the programming logic in implementing the model. This section has included a scorecard that will be used for recording the performance of the model for discussion in chapter 5 and chapter 6. Table 10 provides a summary of the configurations settings and strategies used in the implementation of this model. The next section provides a base line prediction, a measurement that the ANN model should improve on.

Table 9: Neural network performance scorecard

Variation	Optimized Structure	Training Set MSE	Validation Set MSE	Test Set MSE	Test Set (Year: 2014)			
					Correlation Coefficient	RMSE	MARE	NS

Table 10: Summary of the model

Summary of the model	
Programming Language	Java (using NetBeans IDE)
Libraries	Encog 3.3 NetCDF
Neural Network	Resilient Propagation
Neural Network activation function	Sigmoid
Neural Network Strategies	Pruning K-Fold Cross-validation Early stopping Ensemble Network
Neural Network values	Max Iterations: 100 Pruning: 3 best Architectures (with max hidden neurons 100) Max Hidden Layers: 1 K-Fold: 5-Fold Early Stopping Ensemble size: 100 artificial neural networks
Performance Measures	MSE Correlation Coefficient RMSE MARE NS
Data Sources	DWS ECMWF
Data Scaling	Input: (0,1) Output: (0,1)
Data Sets	1. Data between 1989 and 2013 is used as training and validation data. 2. Data in 2014 is used as the test data set, and the artificial neural network will not 'see' the data in 2014 during training.

4.8. Limitations of the Research

This research is currently limited to evaluating the use of ANN's to predict river flow gauge values in South Africa by making use of weather parameters such as those provided by the European Center for Medium-Range Weather forecasts (ECMWF). This research will not provide a flood forecasting model. The research will not provide a model to predict river flow on rivers where there are currently no gauges. The research is limited to a single river and a single river flow gauge station, the research is not considering ungauged points along the river. The research is only limited to a single river flow gauge, located high up the Thukela river with a reasonably small catchment area. This allows the research to focus on which weather parameters are required, if additional data is required and to what accuracy a neural network can predict the river flow. This research will not provide a prediction model suitable for daily operations, but will provide a model that can drive discussion around the performance and suitability of neural networks to predict river flow in South Africa using ECMWF data. Finally, this research will also not provide a new type of neural network or new neural network structure, it will make use of already known, tested and accepted neural network libraries and structures.

4.9. Conclusion

This chapter has provided a detailed explanation of the methodology for this research. In summary this research makes use of a model in the form of an ANN to predict river flow gauge values. The model makes use of data from two data sources, DWS for the Thukela river flow gauge values at the Driel gauge station and the ECMWF dataset for meteorological data relevant to the catchment area. The model makes use of two main code libraries namely the NetCDF library to read the ECMWF data set and the Encog java library to create the neural network. These two datasets went through a process of data cleaning and rolling window generation. The data cleaning process removed outliers, handled missing data values and transferred non-numeric monthly indicators into numeric inputs. This cleaned data was then used to generate rolling windows of 1 day, 7 days, 30 days, 90 days, 180 days, 365 days and 1095 days. The rolling windows will provide an historical view for the ANN. All the input and output data went through a scaling process which scaled input data to the active domain of $(0,1)$ and output data to the active range of $(0,1)$. The model makes use of a number of different strategies to overcome intrinsic problems associated with ANN's. One of the strategies used is pruning, to find the best ANN architecture, this makes use of the pruning function provided in the Encog library. The ensemble strategy is then used to create an ensemble of 25 ANN's to over come the sensitivity of the initial weights. These 25 neural networks then are trained using a k-fold strategy to ensure the resulting ANN's are able to generalize. The research model makes use of a 5-fold k-fold strategy. During training of the ANN there is an early stopping strategy employed to ensure the model does not become overfitted losing its ability to generalize. Once all the training has taken place the ANN that can generalize the best is used from the ensemble. The best generalizing ANN is then used to predict the river flow gauge values for 2014. The chapter also provided a performance scorecard that is used during performing the scenarios in Chapter 5. The scorecard includes measures such as MSE, correlation coefficient, RMSE, MARE and NS. Chapter 4 provides the required equations to calculate these performance and accuracy measures. Chapter 5 provides the results from the scenarios using this described model, coding libraries and data sets.

5. RESULTS

5.1. Introduction

The results chapter presents the output results for a number of scenarios using the artificial neural network (ANN) model as per section 4.6. The scenarios are designed to show the affect of the naive input, the individual characteristics and the effect of correlation filtering. Each feature selection scenario allows only one parameter to vary in order to see the effect of that parameter. Section 5.2 describes what is called a naive prediction, and is used as a comparative measure in the discussion of each scenario. Section 5.3 describes and presents results for scenario A where the naive input is the parameter that varies. Section 5.4 varies the individual weather parameters as inputs into the ANN in scenario B. Scenario B highlights the effect of each weather parameter and the results show which weather parameters provide the best performance for inputs into the neural network. Scenario C is presented in section 5.5 and shows the effect of correlation filtering on the input parameters. In scenario C the parameter that varies is the correlation cut-off value. Section 5.6 presents input filtering based on findings in scenario A, B and C. It proposes to find an optimum set of inputs for this research, not necessarily the best case scenario but a logical set of input filtering to get a well filtered data set to get a well performing neural network. These scenarios are used to identify the most effective filters, and parameters to use when selecting inputs into the ANN in order to predict river flow gauge values. Chapter 5 provides the results, and a number of plots to show the effect of the various parameters. Chapter 6 provides detailed analysis of the overall results and the performance of using an ANN to predict river flow gauge values.

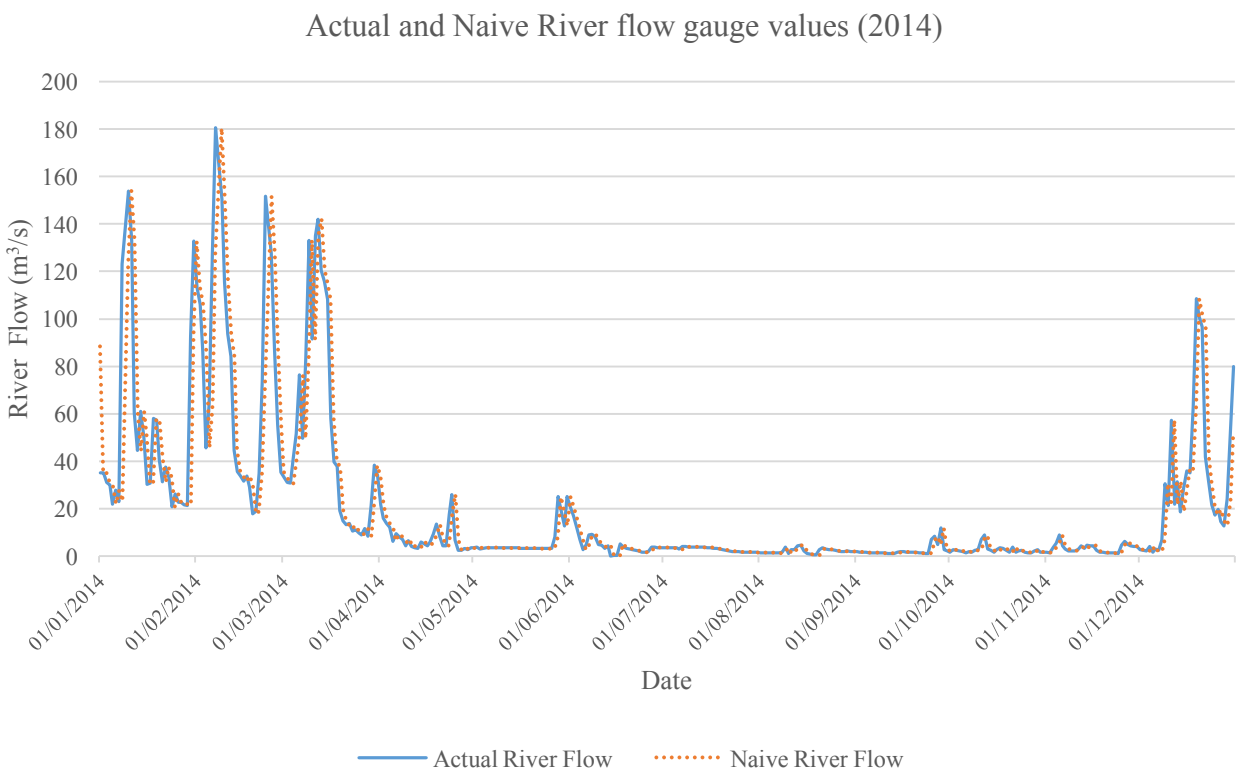
5.2. Naive prediction

The research needs a base line to compare the performance of the ANN against. A naive prediction provides this base line. In this research the naive river flow gauge value estimate is taken as the previous day's river flow gauge value (Pappenberger et al. 2015). This is seen as a completely naive prediction and makes the assumption that the current river flow gauge value will be the same as the following day. This naive prediction provides a means to discuss the performance of the ANN and can be used to ascertain the suitability of ANN for predicting river flow gauge values in this research.

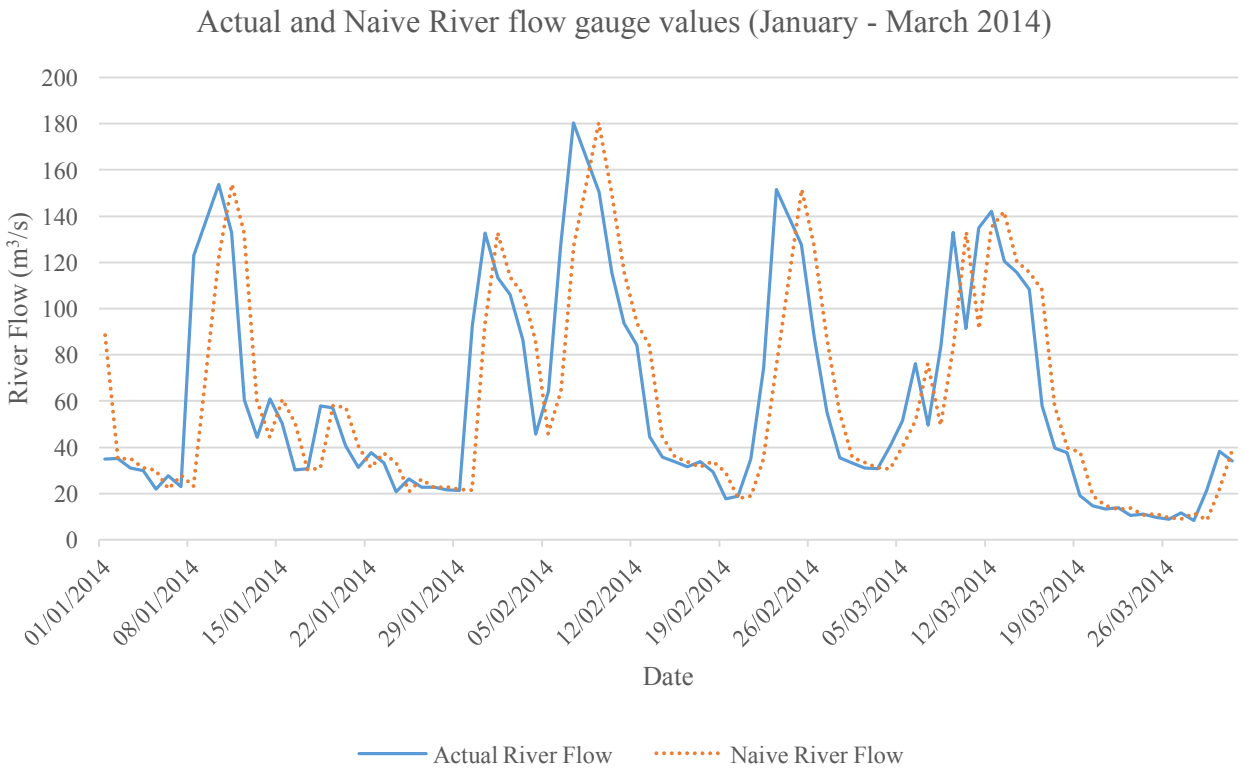
Section 4.6.3 describes a number of performance measures that are used to determine the performance of the neural network. These include MSE, the correlation coefficient, RMSE, MARE and NS. These same measurements are calculated for the Naive prediction and can then be directly compared. Table 11 below provides the scorecard for the naive prediction, this will be used in section 5.5 to compare the naive prediction and 'Scenario D: Best case Scenario' results. Graph 9 provides a view of the actual river flow gauge values and the naive prediction flow values for the test year, 2014. In Graph 9 one cannot clearly see the difference between the actual river flow gauge values and the naive prediction flow values. Graph 10 provides a zoomed in view of Graph 9, with only the flows between January to March 2014. In this view there is clearly a difference between the actual river flow gauge values and the naive prediction river flow gauge values. Graph 11 provides a scatter plot that represents the linear correlation between the actual river flow gauge values and the naive prediction river flow gauge values. The actual correlation value is shown in Table 11.

Table 11: Naive Prediction scorecard

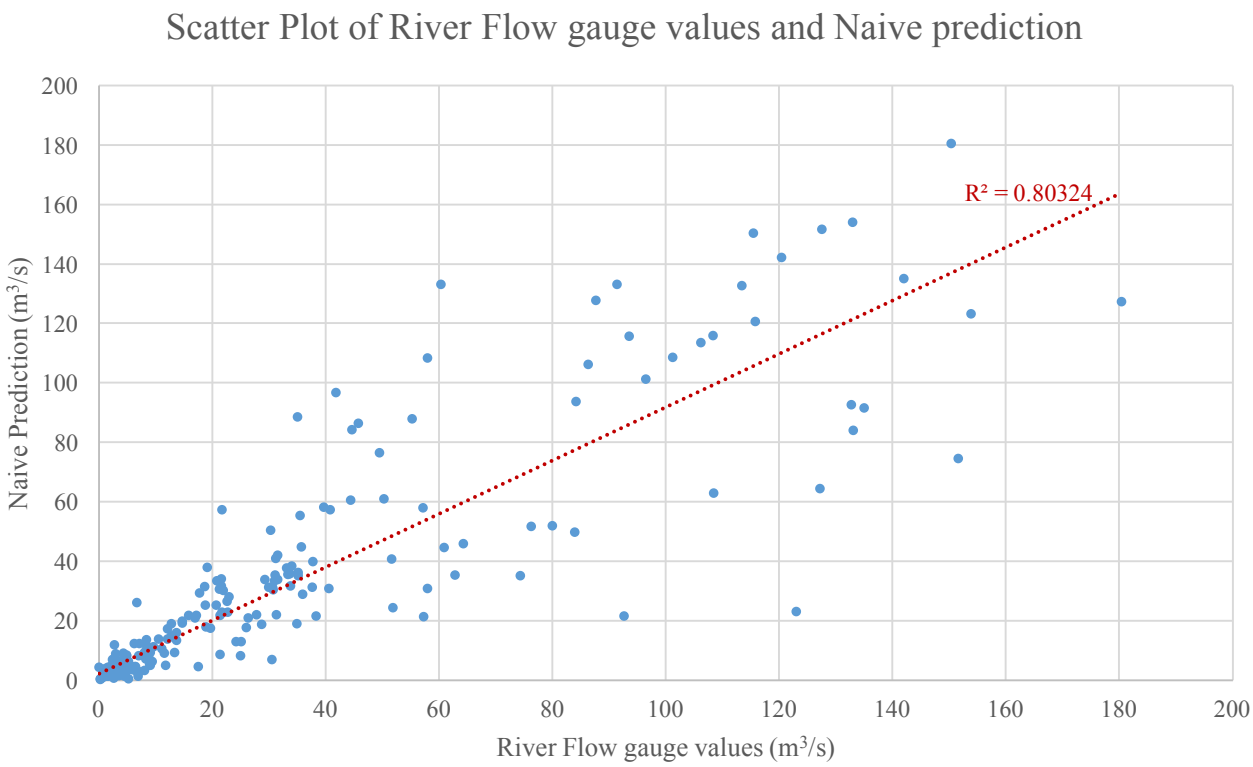
Variation	Optimized Structure	Training Set MSE	Validation Set MSE	Test Set MSE	Test Set (Year: 2014)			
					Correlation Coefficient	RMSE	MARE	NS
Naive Prediction	NA	182.10	103.88	223.48	0.8962	14.9493	42.6287	0.7922



Graph 9: Actual and Naive River flow gauge values (2014)



Graph 10: Actual and Naive River flow gauge values (January - March 2014)



Graph 11: Scatter Plot of River Flow gauge values and Naive prediction

5.3.Scenario A: Naive vs Weather data

5.3.1. Introduction

Scenario A shows the effect of including the naive prediction as an input into the ANN. The scenario runs three variations to show this effect. The first variation uses only the naive as an input into the neural network. The second variation uses only weather parameters as the input into the ANN. The third and final variation combines both the naive prediction and the weather parameters as input into the ANN. Section 5.3.2 describes the configuration of the ANN and a tabled set of ANN output results. Section 5.3.3 presents a number of graphs representing the best of the three variations. Section 5.3.4 discusses the ANN outputs and graphs. Scenario A will be summarized and concluded in section 5.5.5.

5.3.2. Neural network output

The three variations in this scenario were performed with the same ANN configurations. An ensemble of 100 neural networks were used, with the ANN with the lowest generalization error being selected as the best ANN. Each of the neural networks in the ensemble ran for a maximum of 100 iterations. Each ANN in the ensemble had the same structure (including hidden layers), but each ANN in the ensemble had different initial weights. The ANN model is exactly as described in Table 10 in section 4.7. Table 12 provides the summary performance of each of the three variations in scenario A. The inputs into the optimized structure are setup as follows:

ANN Naive Inputs only (19-9-1)

- 12-month indicator inputs
- 7 naive inputs (previous day, 7-day rolling window, 30-day rolling window, 90-day rolling window, 180-day rolling window, 365-day rolling window, 1095-day rolling window)

ANN Weather Inputs Only (537-8-1)

- 12-month indicator inputs
- 525 inputs = 75 ECMWF parameters each with 7 rolling windows (previous day, 7-day rolling window, 30-day rolling window, 90-day rolling window, 180-day rolling window, 365-day rolling window, 1095-day rolling window)

ANN Both Naive and Weather Inputs (544-9-1)

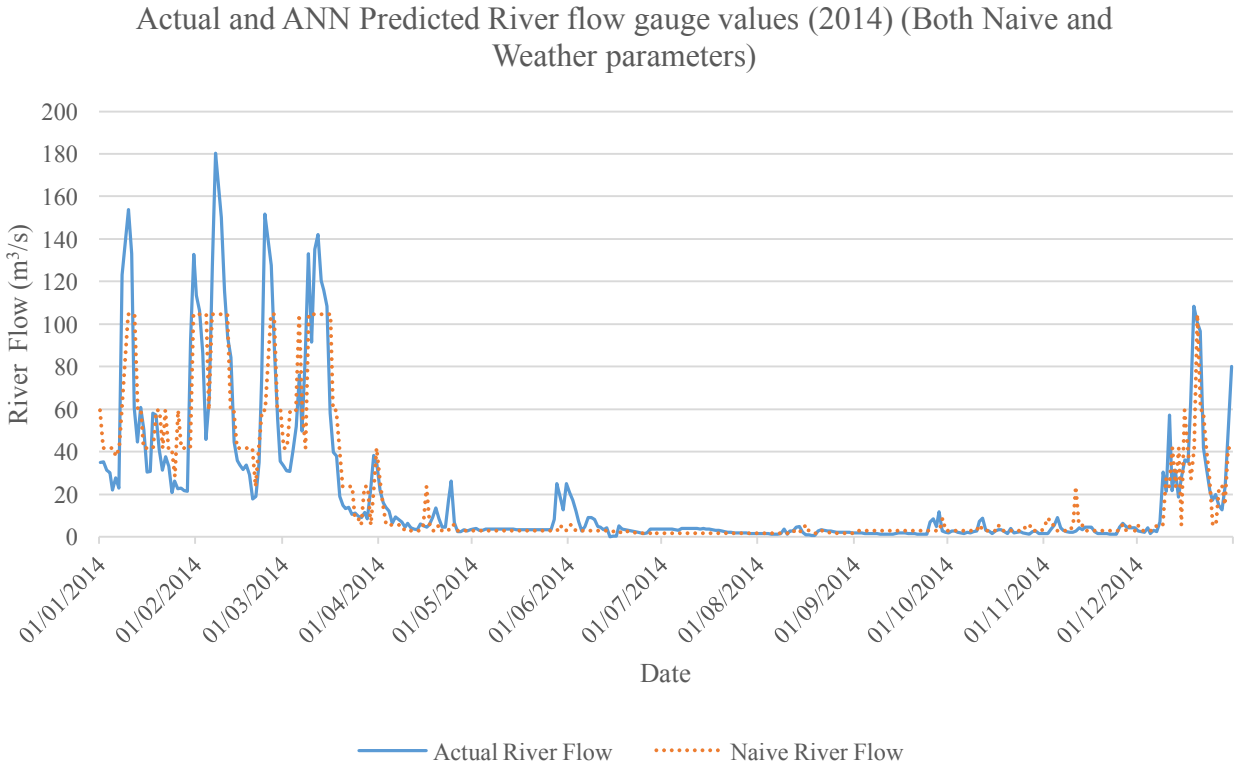
- 12-month indicator inputs
- 7 naive inputs (previous day, 7-day rolling window, 30-day rolling window, 90-day rolling window, 180-day rolling window, 365-day rolling window, 1095-day rolling window)
- 525 inputs = 75 ECMWF parameters each with 7 rolling windows (previous day, 7-day rolling window, 30-day rolling window, 90-day rolling window, 180-day rolling window, 365-day rolling window, 1095-day rolling window)

Table 12: Results - Scenario A: Naive vs Weather Data

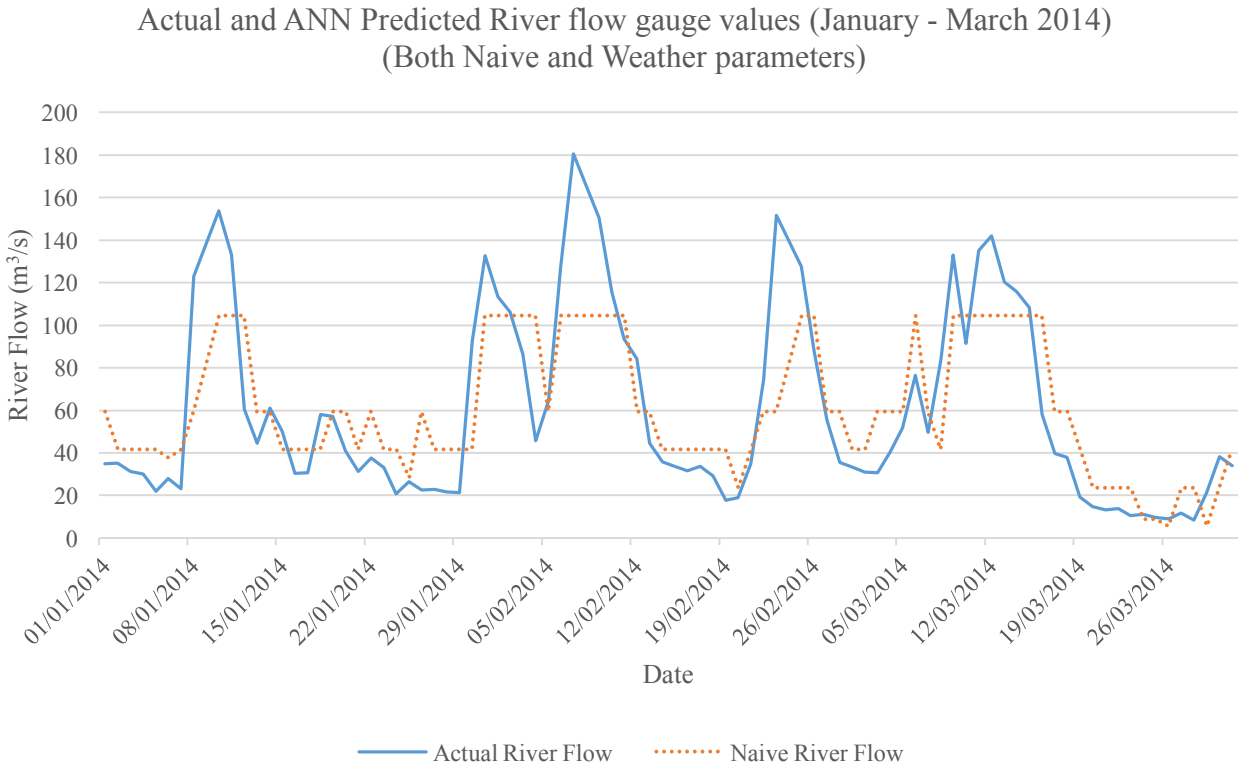
Variation	Optimized Structure	Training Set MSE	Validation Set MSE	Test Set MSE	Test Set (Year: 2014)			
					Correlation Coefficient	RMSE	MARE	NS
ANN Naive inputs only	19-9-1	159,92	95,31	212,21	0,9016	14,5673	75,4184	0,8027
ANN Weather inputs only	537-8-1	310,15	354,92	507.91	0.7314	22.5368	100.8129	0.5277
ANN both Naive and Weather inputs	544-9-1	174,80	107,47	205,57	0,8997	14,3378	65,3845	0,8088

5.3.3. Best result

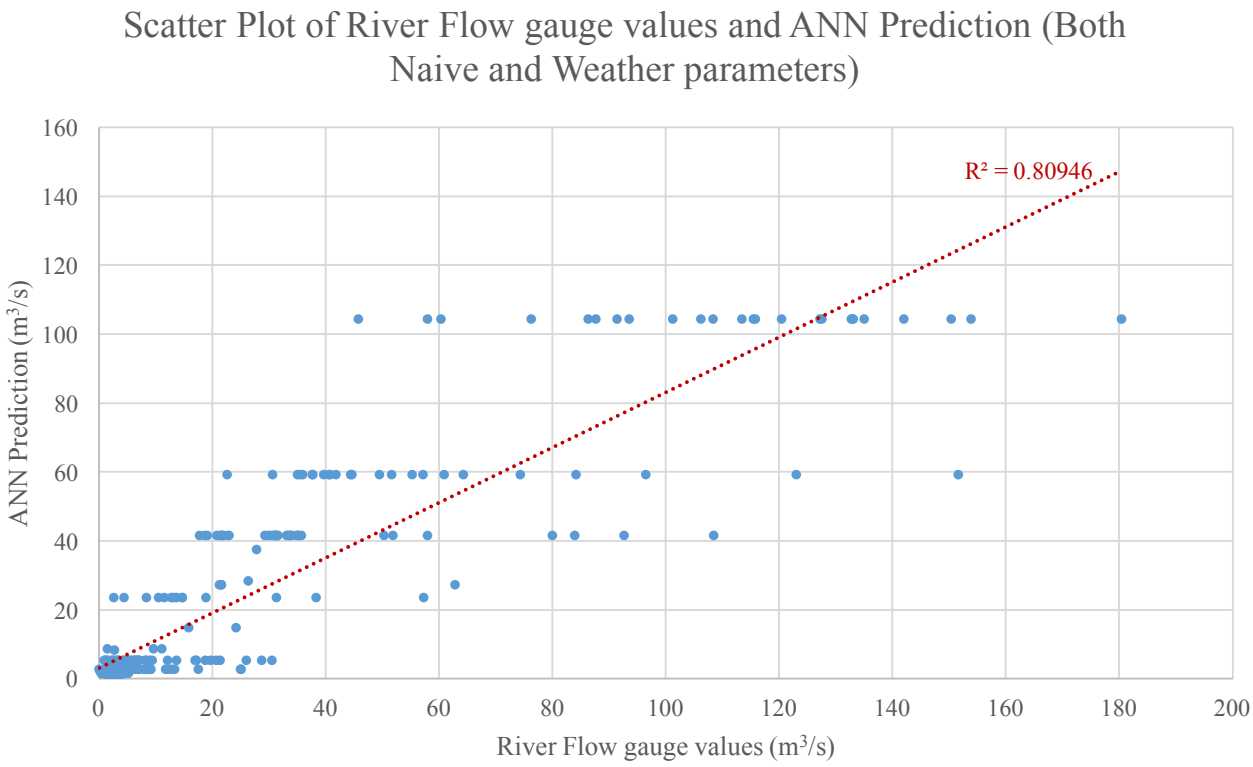
Of the three variations in scenario A, the best results were achieved when both the naive prediction data and the weather parameter data were used as input into the ANN. When both the naive data and weather parameter data were used as inputs, the ANN performed the best with a RMSE of 14,3378, NS value of 0,8088 and a MARE value of 65,3845. Graph 12 shows the full test year of 2014 and plots the actual river flow gauge values with the predicted values from the ANN. In this graph it is difficult to visually tell the difference in some areas, the only conclusion that can be drawn is that the ANN had difficulty predicting values higher than around 104 m^3/s. Graph 13 shows a zoomed in view of the same data but only for January to March 2014. In this graph the differences are more obvious, and the ANN's inability to predict more than around 104m^3/s is more obvious. The linear correlation between the actual river flow gauge values and the ANN's predicted river flow gauge values is represented in the scatter plot in Graph 14.



Graph 12: Actual and ANN Predicted River flow gauge values (2014) (Both Naive and Weather parameters)



Graph 13: Actual and ANN Predicted River flow gauge values (January - March 2014) (Both Naive and Weather parameters)



Graph 14: Scatter Plot of River Flow gauge values and ANN Prediction (Both Naive and Weather parameters)

5.3.4. Discussion

The results table (Table 12) shows that in general the best variation is when both naive prediction data and weather parameter data are used. This is closely followed by purely using naive prediction data as an input into the ANN. The third and worst performing variation is the variation where only weather parameters are used as inputs. This order can be drawn when considering the performance measure RMSE, NS and MARE. The only performance measure that is slightly different is the correlation coefficient, where the naive data as an input out performs the naive data and weather data as inputs by a very small margin. From these performance measures it can be established that adding weather data to the naive prediction data as inputs into the ANN improves the neural networks performance.

An interesting observation from these results is that the ANN structure tends to lean towards smaller hidden layer neurons. The pruning process does aim to find the least computational complex structure but all three variations found that around 8 or 9 neurons in the hidden layer would give the best results. The variation with only the naive prediction data, with 19 inputs, and the variation including both the weather parameters and the naive prediction data, with 544 inputs, both only had 9 neurons in their hidden layer.

Comparing the results in Table 11 and Table 12, using a neural network improves on the naive prediction. If only considering RMSE, the variation with only naive prediction data as an input into the neural network improved on the RMSE by 0,382. The variation with both naive data and weather parameter data as inputs into the neural network improved on the RMSE by 0,6115.

5.3.5. Conclusion

Scenario A provides a base for further investigation in this research. The main aim of scenario A is to show that an ANN can be used to improve on a Naive prediction of river flow gauge values. The results in Table 12 show firstly that a neural network model can improve on a naive prediction, secondly, the results show that adding weather parameter data can further improve the ability for an ANN to predict river flow gauge values. The next scenarios study various individual parameters to identify possible means of improving this capability of ANN's.

5.4. Scenario B: Effect of individual weather parameters

5.4.1. Introduction

Scenario B shows the effect of the individual weather parameters as an input into the ANN. This is done by only changing the weather parameters used as an input into the ANN, all other configuration settings are kept the same. Each weather parameter is a separate variation, and is used in conjunction with the month indicator and the naive river flow. As there are 75 weather parameters there are 75 variations in scenario B. Section 5.4.2 provides the performance measures for the 75 different variations in this scenario. Section 5.4.3 presents a number of graphs providing a graphical view of the best variation. Section 5.4.4 presents a number of graphs providing a graphical view of the worst variation. Section 5.4.5 discusses the various results and outputs from the scenario B variations. Section 5.4.6 summarizes the scenario.

5.4.2. Neural network output

All the variations in scenario B have the same number of inputs unlike scenario A in section 5.3. This is because all of the variations have the following inputs:

- 12-month indicator inputs
- 7 naive inputs (previous day, 7-day rolling window, 30-day rolling window, 90-day rolling window, 180-day rolling window, 365-day rolling window, 1095-day rolling window)
- 7 weather parameter inputs (current day, 7-day rolling window, 30-day rolling window, 90-day rolling window, 180-day rolling window, 365-day rolling window, 1095-day rolling window) as represented in Figure 19

Similarly, to scenario A all the ANN's in the ensemble have the same structure, only the initial weights are different. Each variation generates an ANN model as described in Table 10. Each variation can have a different number of neurons in the hidden layer as determined by the pruning in the ANN model. Table 13 provides the summary performance of each of the 75 variations in scenario B.

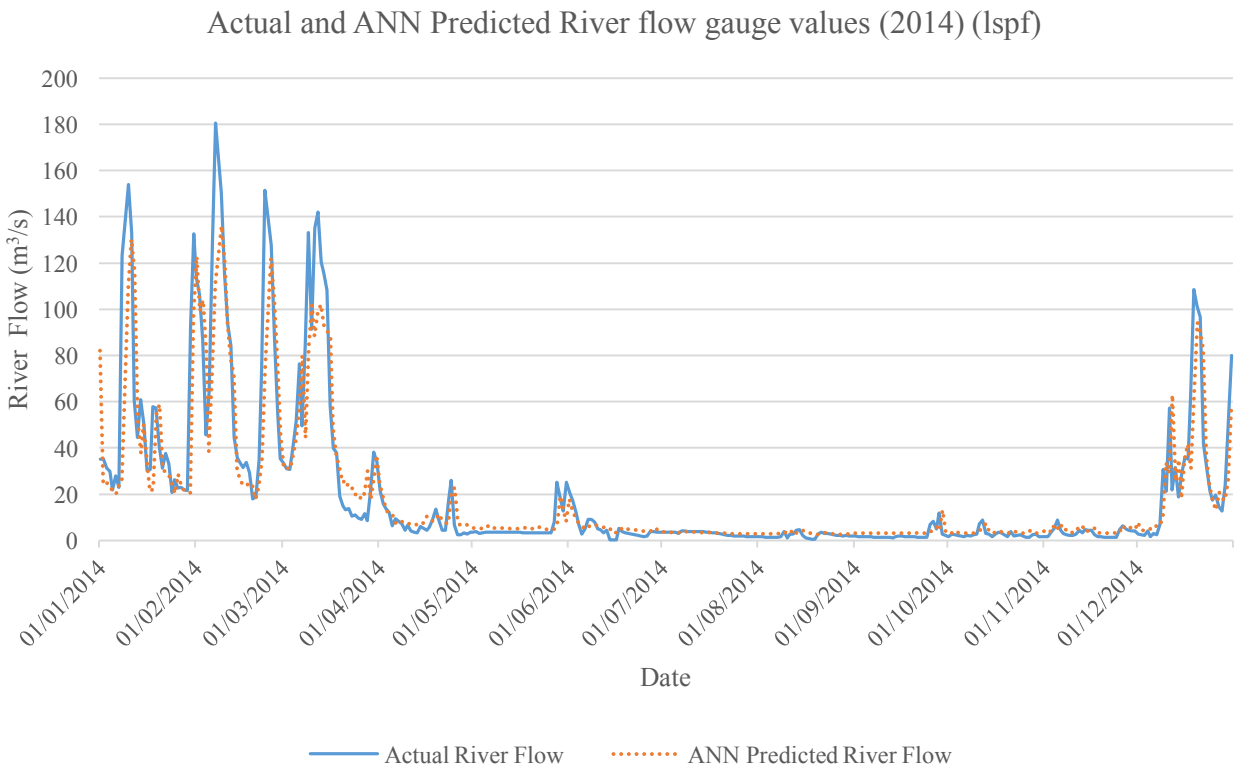
Table 13: Results – Scenario B: Effect of individual weather characteristics

Variation	Optimized Structure	Training Set MSE	Validation Set MSE	Test Set MSE	Test Set (Year: 2014)			
					Correlation Coefficient	RMSE	MARE	NS
asn	26-9-1	172,95	100,95	213,85	0,8958	14,6237	96,7383	0,8011
bld	26-7-1	182,86	111,30	291,35	0,8877	17,0689	68,5162	0,7291
blh	26-10-1	160,75	97,58	205,38	0,9008	14,3309	61,3909	0,8090
cape	26-4-1	186,82	126,72	236,55	0,8842	15,3801	73,2928	0,7800
cp	26-8-1	180,87	106,19	206,13	0,8994	14,3573	74,7496	0,8083
csf	26-8-1	182,62	104,26	218,29	0,8958	14,7747	46,4584	0,7970
d2m	26-13-1	179,56	94,10	226,14	0,8893	15,0379	78,2667	0,7897
e	26-12-1	175,77	105,27	214,85	0,8946	14,6578	64,1811	0,8002
es	26-6-1	203,53	106,04	241,72	0,8806	15,5474	81,7483	0,7752
ewss	26-15-1	174,19	92,88	214,33	0,8975	14,6402	54,4492	0,8007
fal	26-7-1	178,59	106,74	230,67	0,8964	15,1878	67,6875	0,7855
fg10	26-16-1	165,19	97,23	218,62	0,8934	14,7859	66,9653	0,7967
flsr	26-15-1	187,75	102,69	239,36	0,8923	15,4714	86,9390	0,7774
fsr	26-15-1	185,47	104,72	227,51	0,8895	15,0833	115,2717	0,7884
gwd	26-15-1	174,53	100,52	218,61	0,8931	14,7854	79,2341	0,7967
hcc	26-6-1	167,55	107,16	228,30	0,8878	15,1094	72,4701	0,7877
ie	26-4-1	189,67	103,36	236,62	0,8850	15,3823	67,9909	0,7800
iews	26-10-1	169,72	99,55	244,17	0,8941	15,6261	51,6415	0,7729
inss	26-4-1	187,82	113,02	262,24	0,8722	16,1938	63,8822	0,7561
ishf	26-7-1	177,38	107,01	253,83	0,8831	15,9319	55,5166	0,7640
lcc	26-16-1	160,93	95,71	206,95	0,8986	14,3857	75,0139	0,8076
lgws	26-8-1	170,87	101,61	244,55	0,8831	15,6380	55,8892	0,7726
lsf	26-17-1	177,99	101,99	254,38	0,8865	15,9494	55,1214	0,7634
lsp	26-15-1	169,34	100,96	205,51	0,8996	14,3357	70,7251	0,8089
lspf	26-10-1	168,36	100,03	197,53	0,9063	14,0546	80,8010	0,8163
mcc	26-8-1	151,01	113,26	207,38	0,8988	14,4008	57,2690	0,8071
mgws	26-10-1	171,58	94,26	213,50	0,8995	14,6116	72,3173	0,8015
mn2t	26-10-1	180,74	98,51	223,24	0,8911	14,9411	57,4966	0,7924
msl	26-13-1	174,50	101,74	223,01	0,8924	14,9334	64,9306	0,7926
mx2t	26-17-1	184,99	99,80	223,79	0,8907	14,9595	50,5645	0,7919
nsss	26-12-1	171,93	105,39	208,34	0,8984	14,4339	76,2088	0,8063
par	26-16-1	162,98	90,04	217,83	0,8986	14,7589	66,3315	0,7974
parcs	26-8-1	185,75	101,72	226,28	0,8910	15,0427	60,2281	0,7896
ro	26-9-1	177,62	103,18	232,75	0,8863	15,2560	112,8162	0,7836
rsn	26-10-1	175,29	107,94	234,29	0,8847	15,3066	62,4238	0,7821

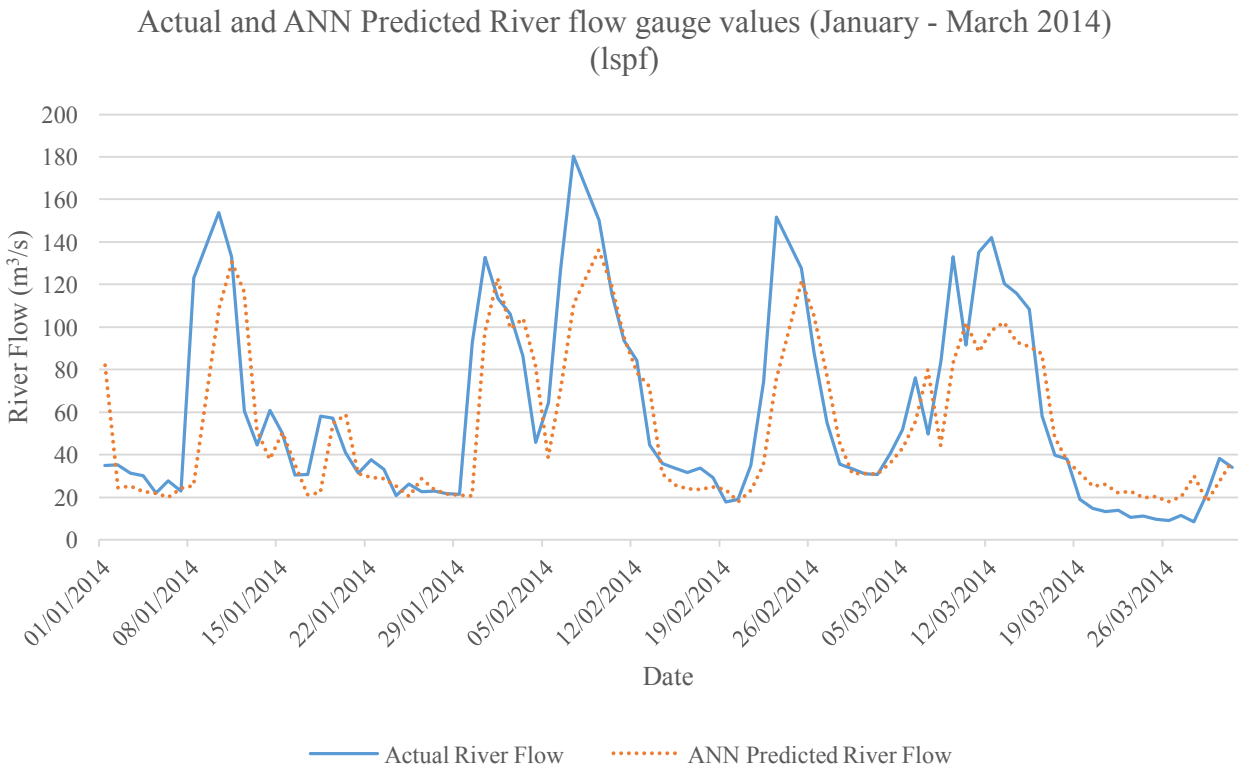
sd	26-11-1	177,80	108,75	213,21	0,8955	14,6015	77,7970	0,8017
sf	26-10-1	173,66	101,06	248,35	0,8828	15,7591	54,2740	0,7691
skt	26-11-1	201,05	99,00	238,55	0,8828	15,4452	63,8596	0,7782
slhf	26-13-1	165,72	109,12	201,72	0,9066	14,2029	60,7797	0,8124
smlt	26-5-1	192,82	114,60	239,47	0,8850	15,4748	141,9743	0,7773
sp	26-12-1	198,73	106,62	229,48	0,8875	15,1485	61,0069	0,7866
src	26-9-1	162,20	102,15	201,62	0,9017	14,1994	69,5094	0,8125
sshf	26-9-1	172,08	99,99	224,34	0,8995	14,9779	70,8411	0,7914
ssr	26-7-1	160,46	96,46	224,08	0,8959	14,9692	47,5653	0,7916
ssrc	26-16-1	168,52	102,25	244,60	0,8951	15,6395	45,1729	0,7725
ssrd	26-10-1	161,51	90,08	237,10	0,8862	15,3980	56,6057	0,7795
stl1	26-10-1	178,71	101,38	216,91	0,8935	14,7278	53,0800	0,7983
stl2	26-30-1	163,45	99,68	209,57	0,8990	14,4764	55,3784	0,8051
stl3	26-14-1	168,98	97,00	209,73	0,8972	14,4821	55,8391	0,8050
stl4	26-13-1	176,43	95,41	226,34	0,8890	15,0446	48,8784	0,7895
str	26-16-1	155,59	85,16	200,95	0,9023	14,1757	58,9998	0,8131
strc	26-16-1	168,52	102,25	244,60	0,8951	15,6395	45,1729	0,7725
strd	26-13-1	169,61	94,80	236,19	0,8960	15,3684	153,8186	0,7804
sund	26-8-1	180,30	102,13	226,54	0,8920	15,0512	60,8512	0,7893
swvl1	26-18-1	179,79	99,14	229,54	0,8901	15,1504	61,9868	0,7865
swvl2	26-15-1	170,29	96,83	205,37	0,8999	14,3306	85,4837	0,8090
swvl3	26-17-1	174,80	100,58	214,51	0,8951	14,6460	66,1528	0,8005
swvl4	26-12-1	180,11	100,98	216,03	0,8940	14,6978	58,2019	0,7991
t2m	26-9-1	184,60	99,46	222,57	0,8913	14,9187	62,7712	0,7930
tcc	26-10-1	171,95	95,15	210,01	0,8971	14,4916	64,7367	0,8047
tciw	26-8-1	182,94	103,17	235,40	0,8881	15,3428	52,6170	0,7811
tclw	26-6-1	174,58	101,58	225,09	0,8908	15,0031	59,2464	0,7907
tco3	26-4-1	200,82	124,72	254,48	0,8738	15,9523	78,3538	0,7634
tcw	26-10-1	203,94	108,63	233,69	0,8849	15,2869	88,6753	0,7827
tcwv	26-9-1	176,83	97,80	214,27	0,8949	14,6381	59,8084	0,8007
tisr	26-5-1	188,80	112,11	230,99	0,8878	15,1983	49,8800	0,7852
tp	26-9-1	168,96	104,57	213,50	0,8962	14,6115	70,5776	0,8015
tsn	26-14-1	171,57	93,07	222,55	0,8907	14,9181	69,4298	0,7930
tsr	26-13-1	160,70	98,60	214,33	0,9003	14,6402	74,7791	0,8007
tsrc	26-9-1	168,59	101,53	215,59	0,8968	14,6829	116,7892	0,7995
ttr	26-28-1	174,37	102,16	225,18	0,8899	15,0061	59,8379	0,7906
ttrc	26-13-1	164,04	93,26	202,27	0,9027	14,2221	67,4071	0,8119
u10	26-13-1	163,42	98,41	210,16	0,8992	14,4969	56,1971	0,8046
uvb	26-11-1	157,12	86,36	209,69	0,9013	14,4807	67,4401	0,8050
v10	26-10-1	176,13	94,75	219,75	0,8923	14,8241	75,7464	0,7956

5.4.3. Best result

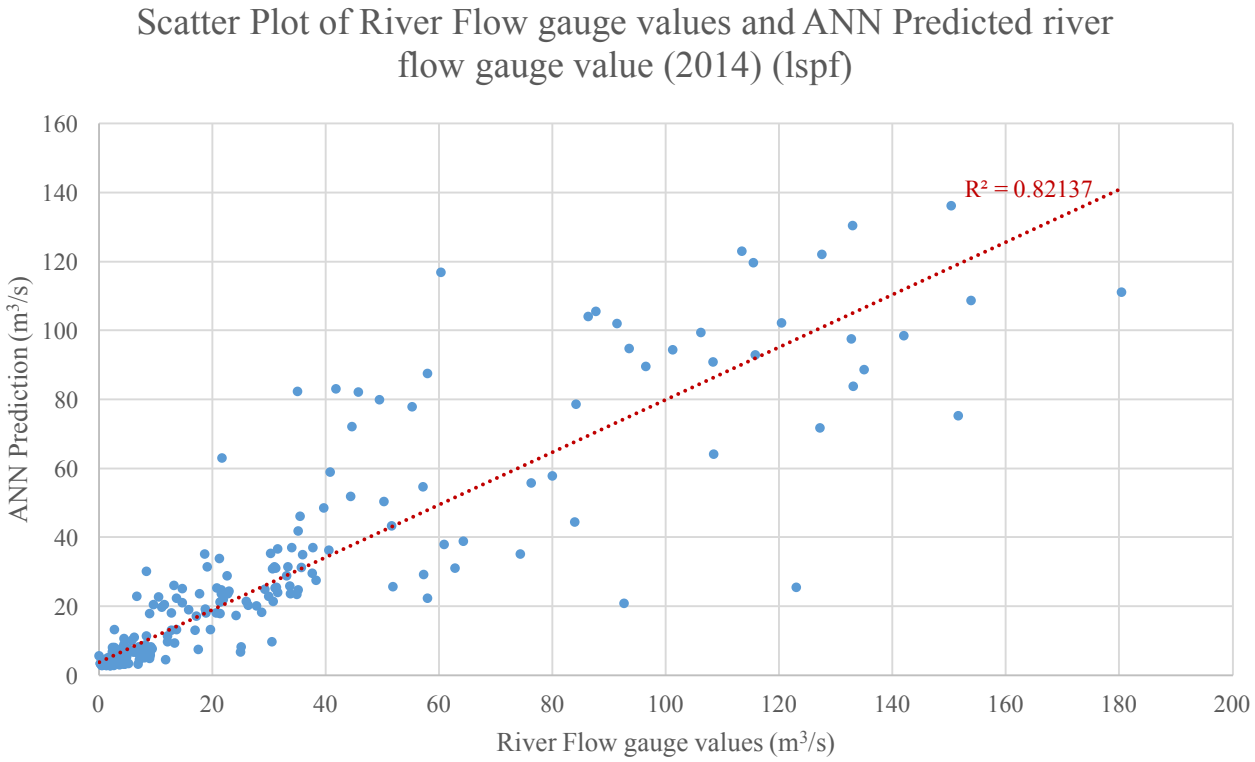
After running all 75 variations in scenario B the weather parameter 'Large-scale precipitation fraction' (lspf) gave the best individual parameter result. This was determined based on the RMSE value 14,0546, the NS value of 0,8163 and correlation coefficient of 0,9063. The weather parameter 'lspf' had the best RMSE value, the best NS value and the second best correlation coefficient value. Graph 15 provides an overview of the river flow gauge values results for 2014. The graph plots both the actual river flow gauge values and the river flow gauge values predicted by the ANN using the weather parameter 'lspf' as an input. Graph 12 provides a full year's values and it is difficult to see any differences in the actual vs predicted values. Graph 16 shows a zoomed in view of the results for January 2014 through to March 2014. The linear correlation between the actual river flow gauge values and the ANN predicted river flow gauge values is represented on a scatter plot in Graph 17.



Graph 15: Actual and ANN Predicted River flow gauge values (2014) (lspf)



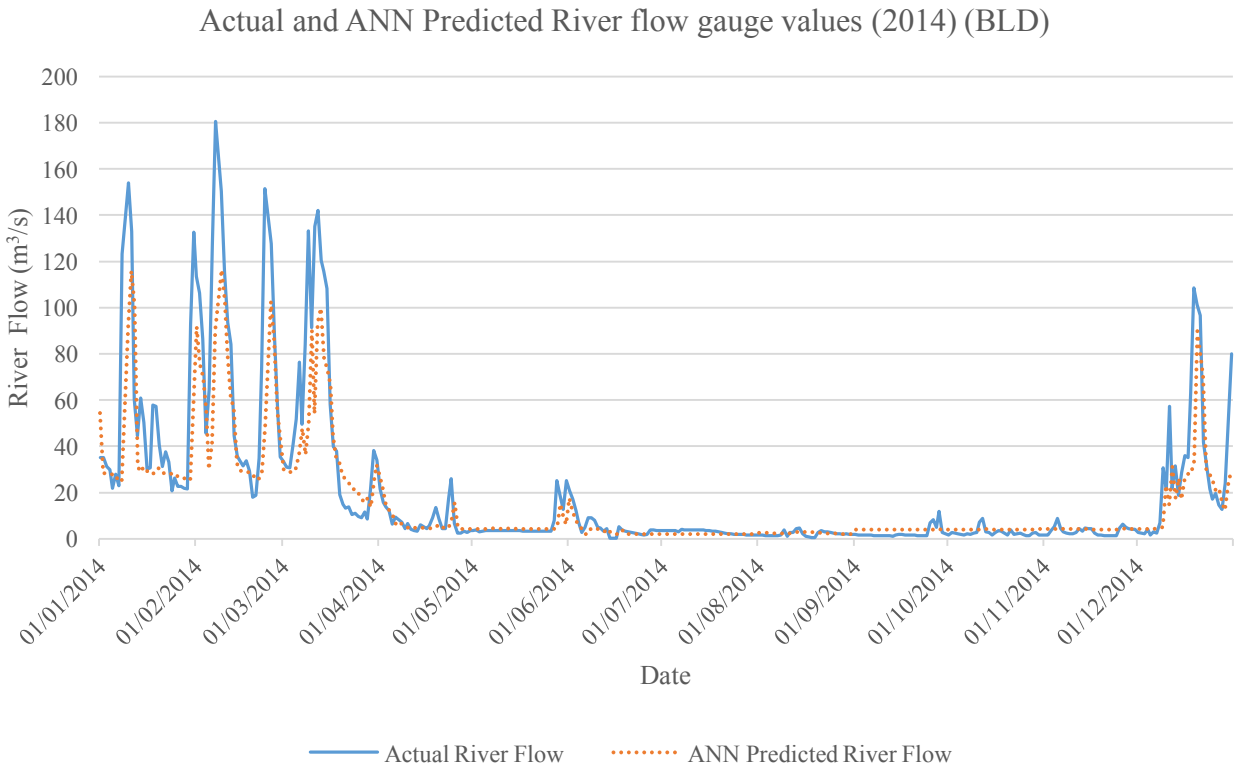
Graph 16: Actual and ANN Predicted River flow gauge values (January - March 2014) (lspf)



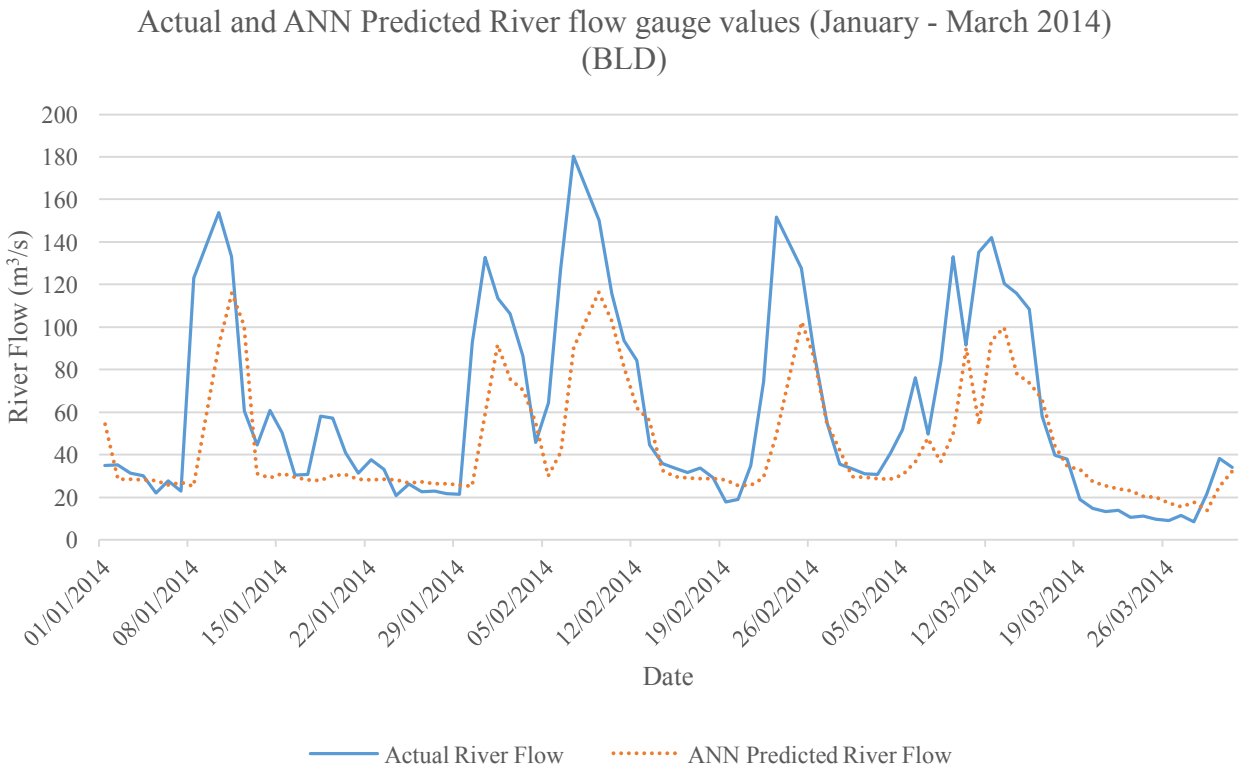
Graph 17: Scatter Plot of River Flow gauge values and ANN Predicted river flow gauge value (2014) (lspf)

5.4.4. Worst result

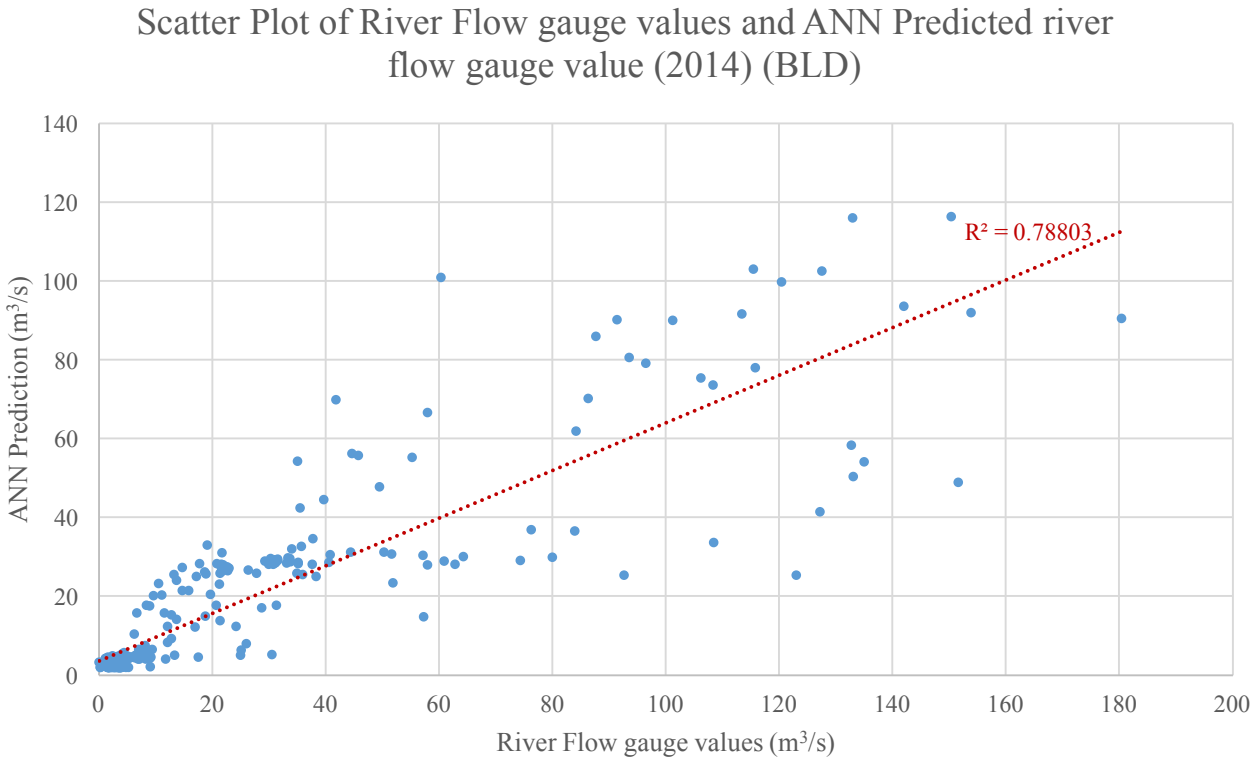
As a comparison to the best performing weather parameter, the weather parameter 'Boundary layer dissipation' (bld) gave the worst individual parameter results. In comparison to the RMSE value 14,0546 for the 'lspf' weather parameter, 'bld' gave an RMSE value of 17,0689. There is also a noticeable difference in the other performance measures; the correlation coefficient value was 0,8877 and the NS value is 0,7291. Similarly to Graph 15 for 'lspf', Graph 18 shows the full predicted year of 2014 using the weather parameter 'bld' against the actual river flow gauge values. The difference in performance is noticeable when comparing the zoomed in views for January to March 2014 for the two weather parameters, Graph 16 and Graph 19. The correlation coefficient for the 'bld' variation is represented in the scatter plot in Graph 20.



Graph 18: Actual and ANN Predicted River Flow gauge values (2014) (BLD)



Graph 19: Actual and ANN Predicted River Flow gauge values (January - March 2014) (BLD)



Graph 20: Scatter Plot of River Flow gauge values and ANN Predicted river flow gauge value (2014) (BLD)

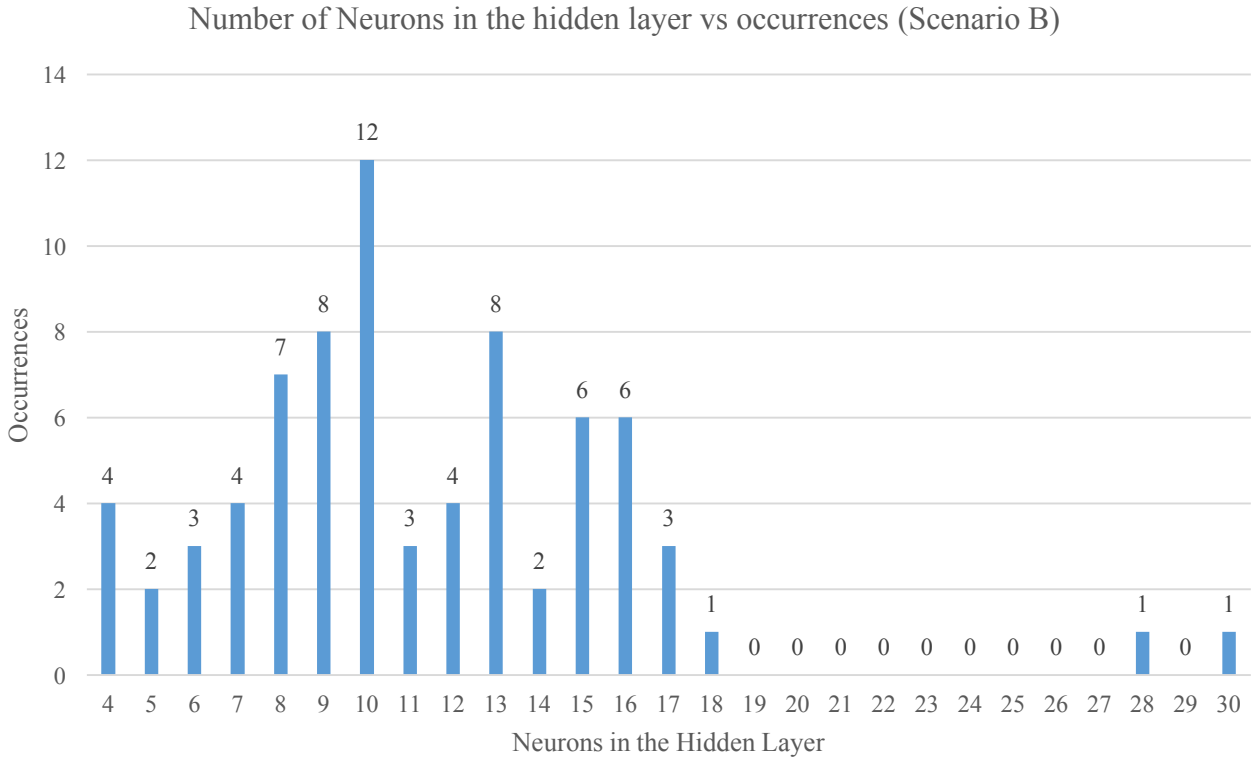
5.4.5. Discussion

The most obvious result from scenario C is the variation in performance of weather parameters as inputs into the ANN. This is clearly seen when comparing the best result (section 5.4.3) and the worst result (section 5.4.4). Table 14 shows the performance of the individual weather parameters when compared to the naive prediction in section 5.2. The most interesting measurement in Table 14 is that no weather parameter can improve on the relative error percentage of 42.6287% for the naive weather prediction. The closest weather parameters were 'ssrc' and 'strc' at 45,1729%.

Table 14: Performance of individual weather parameters compared to the Naive Prediction

Measurement	**Superior Prediction**	**Inferior Prediction**
Correlation Coefficient	24	51
RMSE	38	37
MARE	0	75
NS	38	37

The structure of the ANN's for all the variations also highlights the fact that a large hidden layer is not required. The minimum number of neurons used in the hidden layer was 4 neurons for the weather parameters 'cape','ie','inss' and 'tco3'. The maximum number of neurons used in the hidden layer was 'stl2'. The average neurons in the hidden layer was 11, while the most common number of neurons was 10 with 12 occurrences. The occurrence of each number of neurons in the hidden layer are shown in Graph 21. The change in neurons on the hidden layer is handled through the pruning feature of the Encog Library (section 4.3.2), the pruning is done using the weather parameter and should provide the best performing neural network for that weather parameter.



Graph 21: Number of Neurons in the hidden layer vs occurrences (Scenario B)

It is important to note that these results are not purely weather parameter inputs, the inputs into the ANN include the naive prediction data and the monthly indicators as well. Each variation only varies the weather parameter. The monthly indicators and the naive prediction data stays the same. It is possible to then draw the conclusion that the difference in the performance is related to the weather parameter data.

5.4.6. Conclusion

In summary the performance of the individual weather parameter varies greatly with a number of them out performing the naive prediction from section 5.2, but a large number of them not being able to improve or match the naive prediction of the river flow gauge values. The top 5 based on RMSE are Large-scale precipitation fraction (**lspf**), Surface net thermal radiation (**str**), Skin reservoir content (**src**), Surface latent heat flux (**slhf**) and Top net thermal radiation clear sky (**ttrc**). Scenario B has shown that the selection of weather parameters can drastically affect the performance of the ANN. The results in Table 13 are fundamental to identifying those weather parameters that have a positive impact on the performance of the ANN. These results are referenced in Section 5.6.

5.5.Scenario C: Effect of correlation filtering

5.5.1. Introduction

Scenario C makes use of correlation filtering to understand the effect of removing those weather parameters with low linear correlation. The scenario starts with all weather parameters and then makes a cutoff value. Only those weather parameters with a higher

correlation than the cutoff are included in the variation. As the correlation cut off value increases, so the number of inputs decrease. The correlation cut off value starts at 0,00 and increases by 0,05 with each variation until reaching 0,50. Section 5.5.2 describes correlations in more details and the calculations of the correlation for each weather parameter. Section 5.5.3 provides more details on the ANN model used and tables the variation performance results. Section 5.5.4 briefly describes the best case variation and graphs the predicted river flow gauge values. Section 5.5.5 gives a more detailed discussion relating to using correlation filtering, if it is worthwhile and significant findings. Section 5.5.6 concludes scenario C by summarizing section 5.5.

5.5.2. Correlations

Two variables could appear to be closely related. A simple example would be when there is more rain the river flow discharge increases and when there is less rain the river flow discharge decreases. This is a simple example, and there is no numerical value to represent how closely related these two variables are. The correlation coefficient (R) quantifies the degree to which two variables are linearly related (He et al. 2014). In section 4.6.3 the correlation coefficient is used as a measurement of the performance of the neural network. In this section the correlation coefficient is used to determine the linear correlation between European Center for Medium-Range Weather Forecasts (ECMWF) data parameters and the DWS river flow gauge values. This is done in an attempt to identify those ECMWF parameters that have a linear impact on the river flow gauge value, as per the previous example where an increase in rain possibly increases the river flow gauge value.

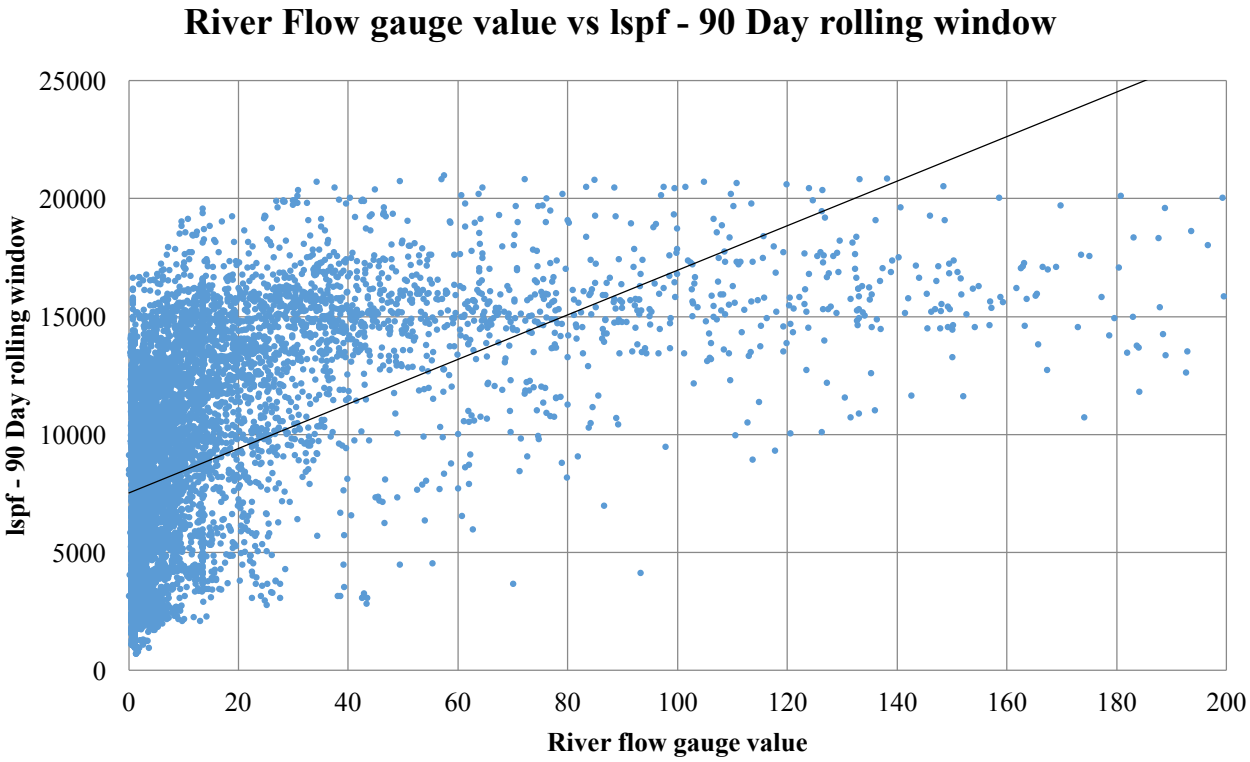
A correlation calculation was performed using the cleaned data between each ECMWF parameter and the DWS river flow gauge value. These correlation values are then used during scenario C to filter data parameters in and out of the ANN inputs. The correlation coefficient calculation is done using Equation 12. Table 23 in the annex shows the calculated correlation for the 532 input values (excludes the monthly indicator). Two examples of positive correlation coefficient scatter plots are shown in Graph 22 and Graph 23. Two examples of negative correlation coefficient scatter plots are shown in Graph 24 and Graph 25.

$$R = \frac{\sum_{i=1}^{n}\left(Q_i^{IN} - \overline{Q^{IN}}\right)\left(Q_i^{RF} - \overline{Q^{RF}}\right)}{\sqrt{\sum_{i=1}^{n}\left(Q_i^{IN} - \overline{Q^{IN}}\right)^2 \sum_{i=1}^{n}\left(Q_i^{RF} - \overline{Q^{RF}}\right)^2}}$$

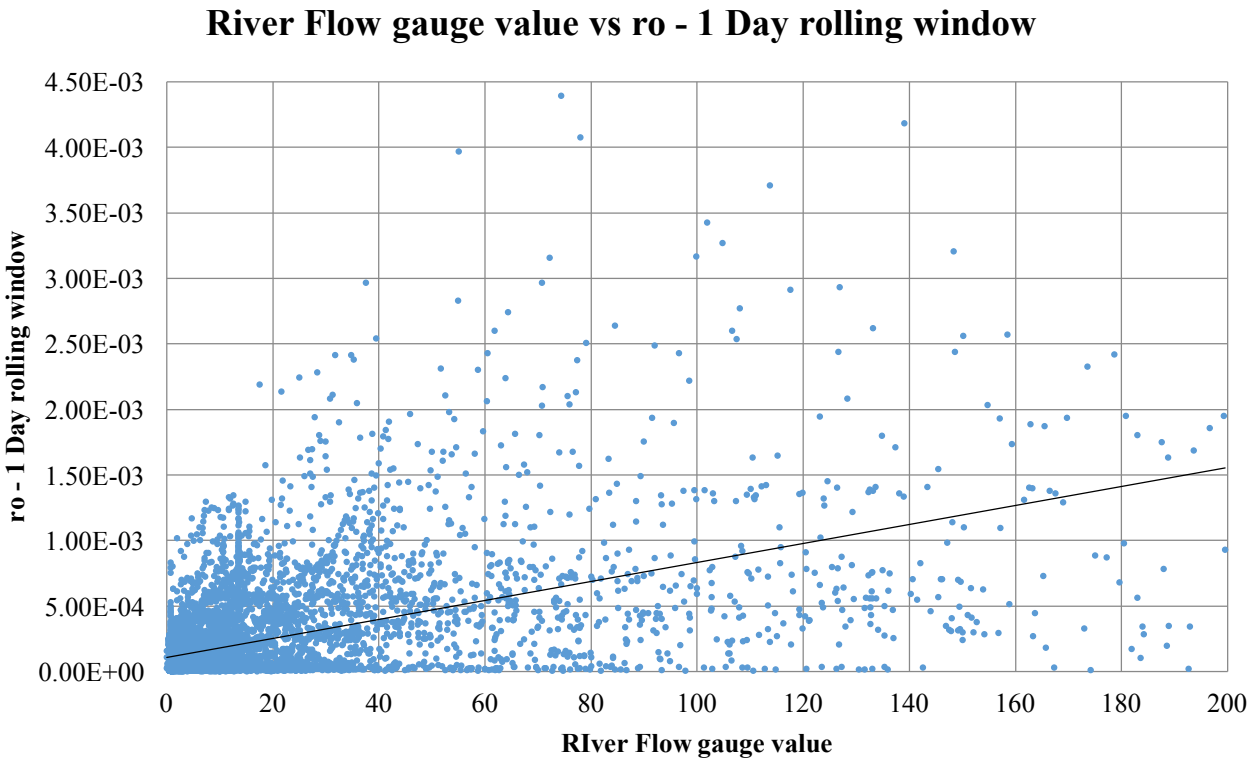
Equation 12: Correlation Coefficient (R) for input values

Where

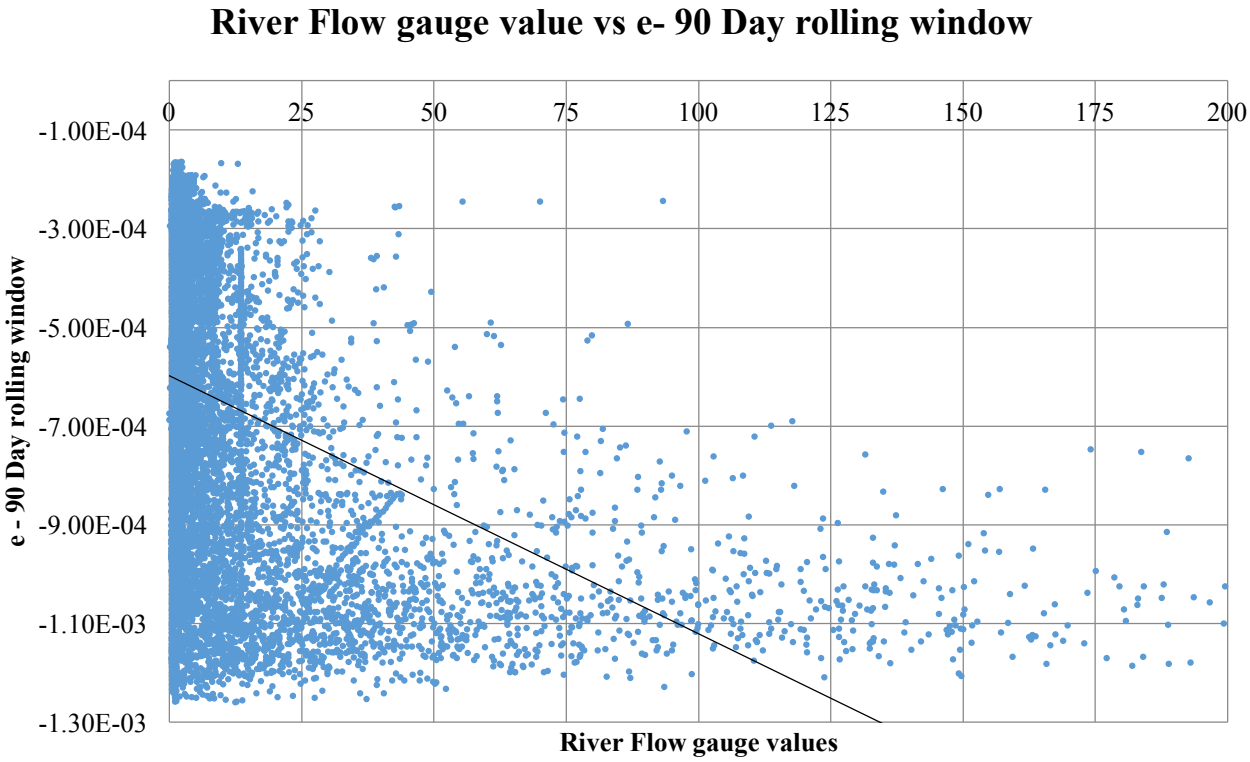
$$n = number\ of\ inputs$$
$$Q_i^{IN} = input\ value\ at\ i$$
$$Q_i^{RF} = river\ flow\ at\ i$$
$$\overline{Q^{IN}} = mean\ input\ value$$
$$\overline{Q^{RF}} = mean\ river\ flow$$



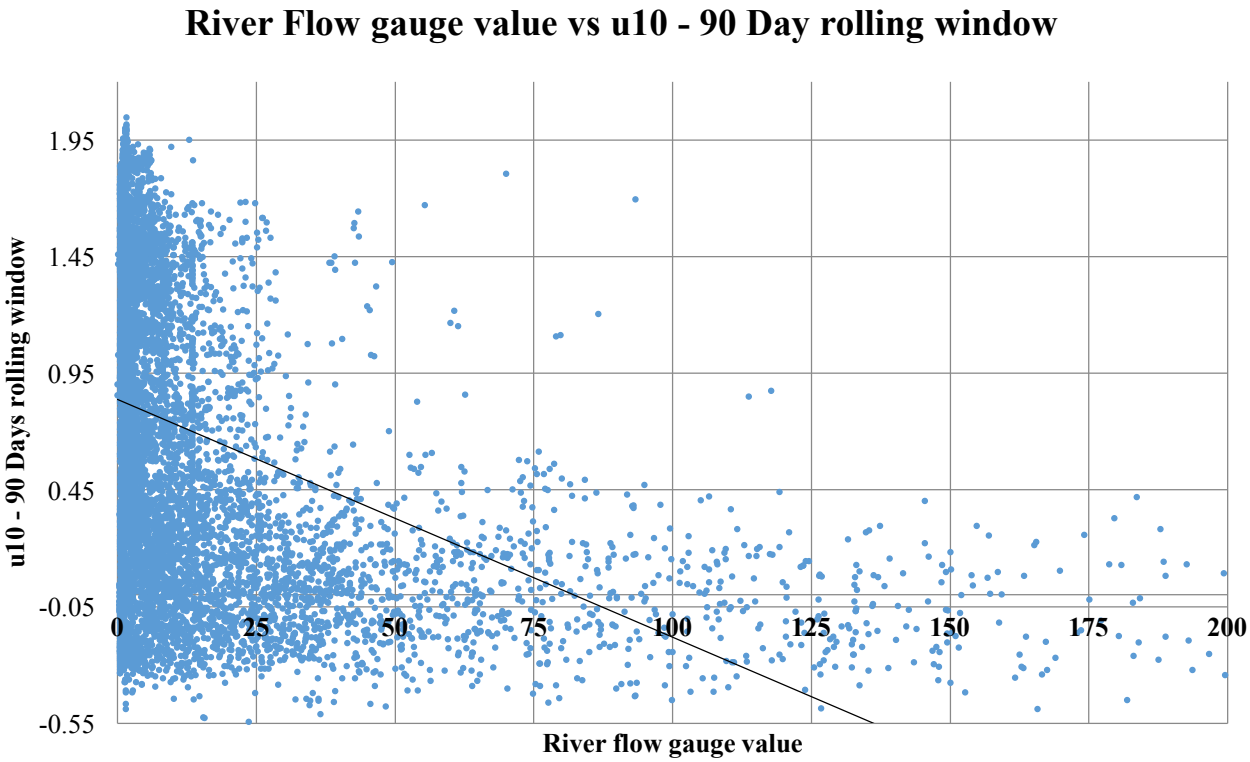
Graph 22: River Flow gauge value vs lspf - 90 Day rolling window



Graph 23: River Flow gauge value vs ro - 1 Day rolling window



Graph 24: River Flow gauge value vs e- 90 Day rolling window



Graph 25: River Flow gauge value vs u10 - 90 Day rolling window

5.5.3. Neural network output

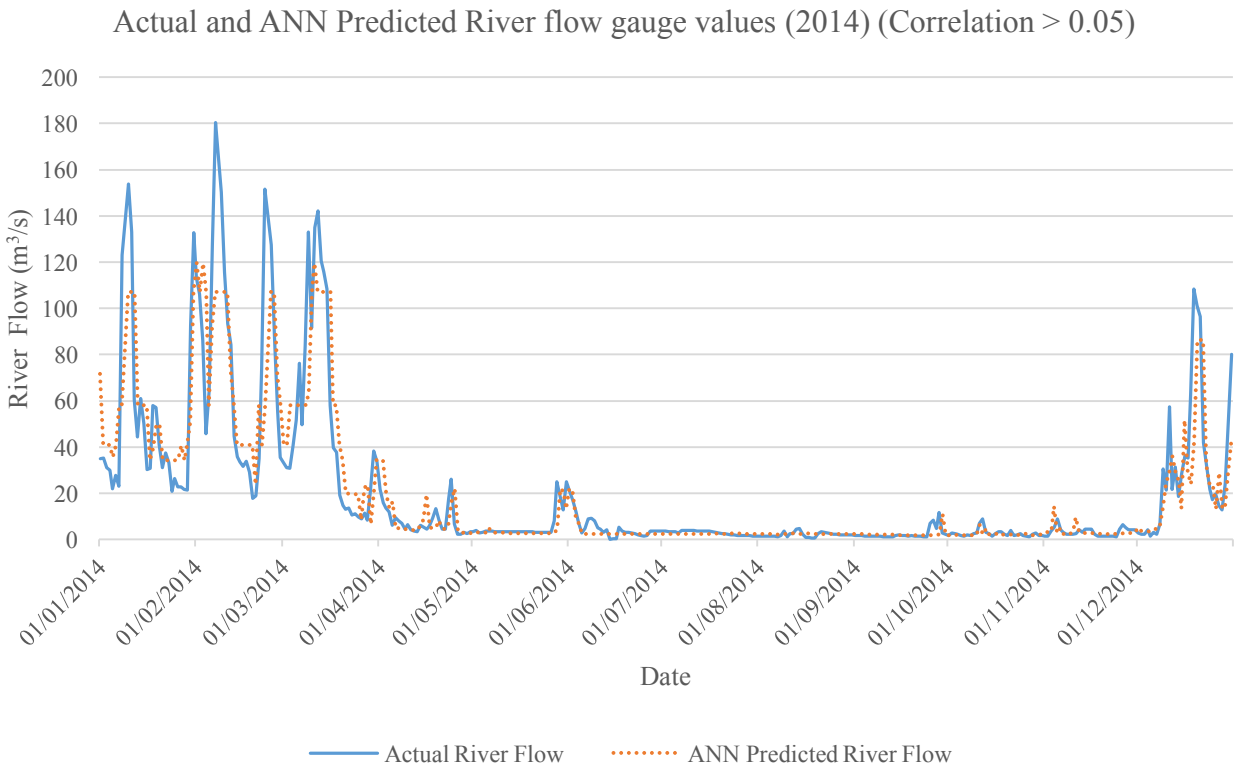
In this scenario C, all the weather parameters and naive prediction inputs are considered but filtered based on their individual correlation coefficient. Not only the main weather parameter data is considered, rather each individual rolling window is treated as an individual input irrespective of the correlation coefficient values of other rolling windows. This scenario attempts to filter out those inputs that have limited linear correlation to the river gauge values. The absolute value of the correlation coefficient is used for filtering. In this way, both positively correlated and negatively correlated input parameters are considered. The filtering is done by increasing the correlation coefficient cut-off value by 0,05 for each subsequent variation. The variations stop at the cutoff value 0,50 as there are only 26 input values with an absolute correlation coefficient value higher than 0,50. The only input values that were not filtered based on correlation coefficient values was the monthly indicator as these are constant throughout all the scenarios and all the variations. The ANN was configured similarly to scenario A and scenario B. The ANN is configured as described in the model in section 4.7. Table 15 presents the results for this scenario. Section 5.5.4 graphs the best results for this scenario and section 5.5.5 will discuss the findings for this scenario.

Table 15: Results – Scenario C: Effect of correlation filtering

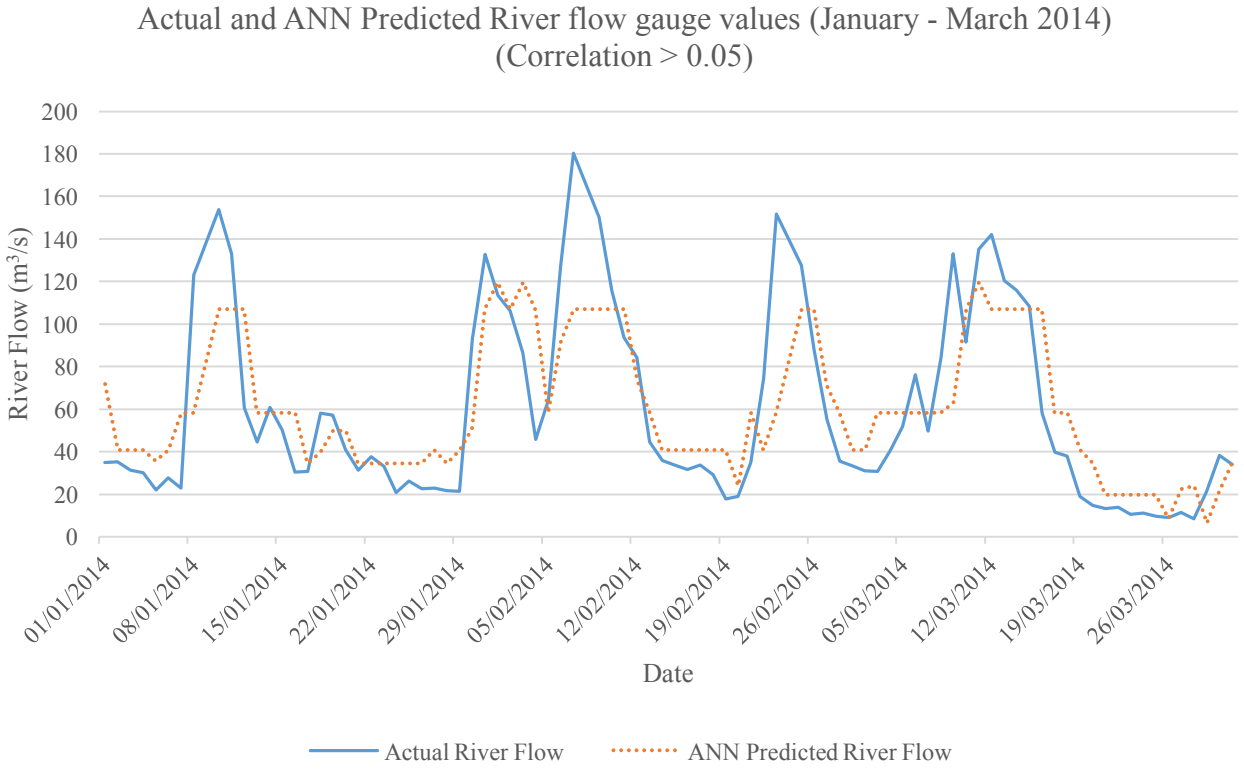
Variation	Optimized Structure	Training Set MSE	Validation Set MSE	Test Set MSE	Test Set (Year: 2014)			
					Correlation Coefficient	RMSE	MARE	NS
|Correlation| > 0.00	544-9-1	174,80	107,47	205,57	0,8997	14,3378	65,3845	0,8088
|Correlation| > 0.05	449-19-1	165,44	106,33	205,06	0,8997	14,3201	55,0866	0,8093
|Correlation| > 0.10	386-9-1	165,63	107,43	205,79	0,9025	14,3454	71,1086	0,8086
|Correlation| > 0.15	332-20-1	167,90	110,17	231,57	0,8867	15,2175	70,1960	0,7847
|Correlation| > 0.20	290-11-1	163,17	115,34	217,09	0,8948	14,7339	87,5913	0,7981
|Correlation| > 0.25	261-10-1	173,95	106,38	234,17	0,8845	15,3025	71,5425	0,7822
|Correlation| > 0.30	219-9-1	168,90	107,77	211,85	0,8962	14,5549	62,0176	0,8030
|Correlation| > 0.35	169-8-1	181,68	108,11	267,60	0,8696	16,3584	68,9202	0,7512
|Correlation| > 0.40	101-11-1	196,80	108,50	238,46	0,8833	15,4422	50,0409	0,7782
|Correlation| > 0.45	57-17-1	183,09	110,41	228,95	0,8882	15,1311	69,4587	0,7871
|Correlation| > 0.50	26-12-1	172,47	102,80	214,50	0,8956	14,6457	56,4621	0,8005

5.5.4. Best result

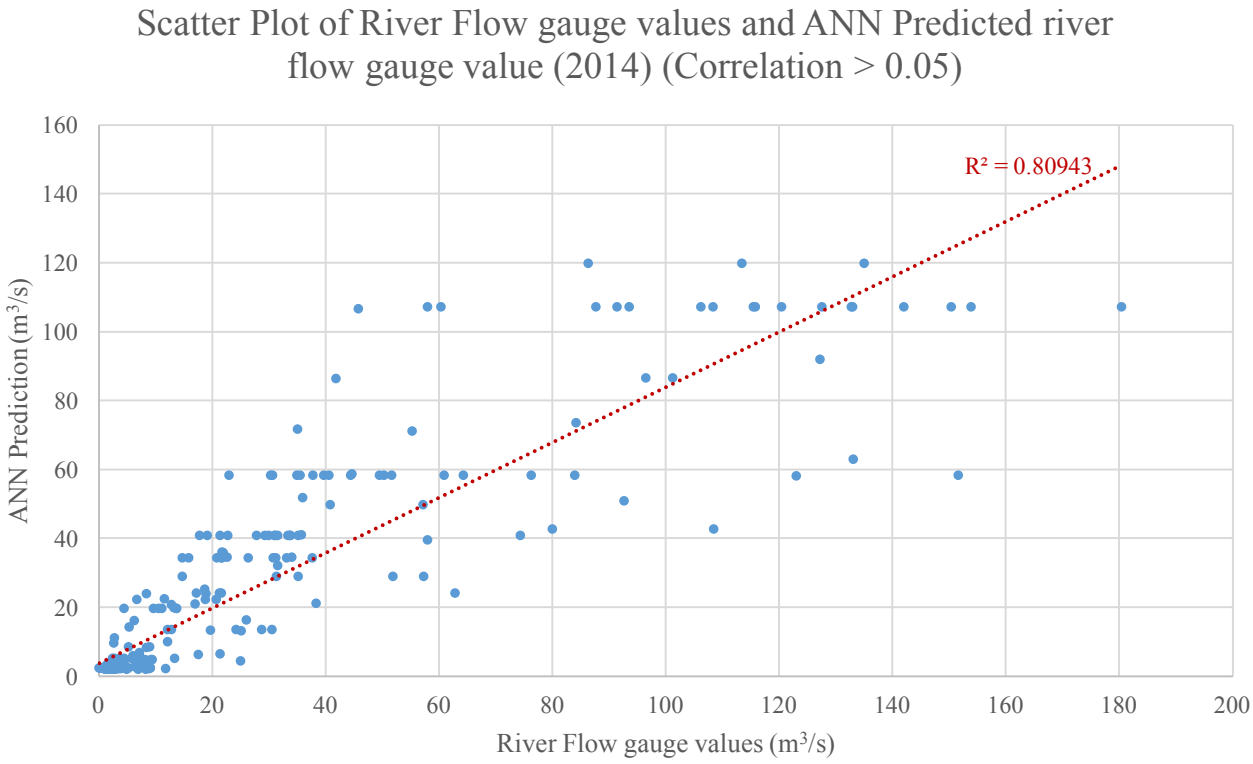
The best result in scenario C was the variation '|Correlation| > 0,05'; any input with an absolute correlation coefficient value less than or equal to 0,05 were removed from the input data. This removed 95 input values from the available 544, leaving a total of 449 input values. With these 95 inputs removed, the variation '|Correlation| > 0,05' gave an RMSE value of 14,3201, a MARE value of 55,0866 and a NS value of 0,8093. Graph 26 shows the predicted year of 2014 for this variation against the actual river flow gauge values for the same period. Graph 26 in some areas is too high level and the difference in predicted values and actual river flow gauge values is difficult to visually, Graph 27 provides the zoomed-in period of January to March 2014. The correlation between the ANN predicted river flow gauge values and the actual river flow gauge values is represented in the scatter plot in Graph 28.



Graph 26: Actual and ANN Predicted River flow gauge values (2014) (Correlation > 0.05)



Graph 27: Actual and ANN Predicted River flow gauge values (January - March 2014) (Correlation > 0.05)



Graph 28: Scatter Plot of River Flow gauge values and ANN Predicted river flow gauge value (2014) (Correlation > 0.05)

5.5.5. Discussion

The variation in scenario C show an interesting finding, in that correlation filtering does not make a linear improvement in performance. There is also no identifiable point where the inputs have been over filtered losing required input data. If this were the case, any variation where correlation filtering occurred above 'X' would perform poorly. Variation '|Correlation > 0,45|' and '|Correlation > 0,5|' are both highly filtered with limited inputs, yet these variations did not perform the worst, showing there is no identifiable point of over filtering.

In this scenario, the variation '|Correlation| > 0,05' gave the best result while variation '|Correlation > 0,35|' gave the worst result. When considering the performance of the ANN all 4 performance measures need to be considered. When looking at the other performance measures it is also not possible to draw any conclusions that increased correlation filtering increases the performance of the ANN. Table 16 shows the order of performance based on each performance measure. The table shows there is no definitive sequence of filtering to improve the performance of the ANN in its ability to predict river flow gauge values. What the table does show is that variation '|Correlation| > 0,05' performs the best overall, it is not always at the top of the performance measure ratings, but it is consistently in the top 3.

Table 16: Scenario C – Variation performance rankings

Correlation Coefficient	RMSE	MARE	NS
1. |Correlation| > 0.10	1. |Correlation| > 0.05	1. |Correlation| > 0.40	1. |Correlation| > 0.05
2. |Correlation| > 0.00	2. |Correlation| > 0.00	2. |Correlation| > 0.05	2. |Correlation| > 0.00
3. |Correlation| > 0.05	3. |Correlation| > 0.10	3. |Correlation| > 0.50	3. |Correlation| > 0.10
4. |Correlation| > 0.30	4. |Correlation| > 0.30	4. |Correlation| > 0.30	4. |Correlation| > 0.30
5. |Correlation| > 0.50	5. |Correlation| > 0.50	5. |Correlation| > 0.00	5. |Correlation| > 0.50
6. |Correlation| > 0.20	6. |Correlation| > 0.20	6. |Correlation| > 0.35	6. |Correlation| > 0.20
7. |Correlation| > 0.45	7. |Correlation| > 0.45	7. |Correlation| > 0.45	7. |Correlation| > 0.45
8. |Correlation| > 0.15	8. |Correlation| > 0.15	8. |Correlation| > 0.15	8. |Correlation| > 0.15
9. |Correlation| > 0.25	9. |Correlation| > 0.25	9. |Correlation| > 0.10	9. |Correlation| > 0.25
10. |Correlation| > 0.40	10. |Correlation| > 0.40	10. |Correlation| > 0.25	10. |Correlation| > 0.40
11. |Correlation| > 0.35	11. |Correlation| > 0.35	11. |Correlation| > 0.20	11. |Correlation| > 0.35

5.5.6. Conclusion

It is difficult to draw any conclusive findings about the result of correlation filtering. There is no linear relationship between correlation filtering and the performance of the ANN in its ability to predict river flow gauge values. The best scenario '|Correlation| > 0,05' does give the best result and will be used in scenario D, but purely in an attempt to reduce computational complexity rather than to directly improve performance. By reducing the number of inputs slightly the neural network trains quicker and has less computational complexity.

5.6. Scenario D: Filter inputs based on scenario A, B & C

5.6.1. Introduction

Scenario D is a combination of scenario A, scenario B and scenario C. This scenario takes lessons learnt and findings from the previous scenario and attempts to find an optimum set of inputs for this research. It may not be the best case variation and may not be the best results possible, but it looks to produce an ANN that can generalize. The scenario looks at using a correlation filter from scenario C, since the variation '|Correlation| > 0,05' was consistently in the top 3 for all the performance measures, this is the correlation filtering used in scenario D. The individual weather parameters performance from scenario B will be used to create selected input parameters based on the best performing weather parameters. scenario A shows that using the naive input data and the weather parameter together improves the performance of the ANN. In scenario D both the naive input data and the weather parameter data will be used as inputs. Section 5.6.3 briefly discusses the ANN model and the inputs before presenting the performance scorecard of the various variations. Section 5.6.3 graphs the best performing variation. Section 5.6.4 describes the results and findings in this scenario before being concluded in section 5.6.5.

5.6.2. Neural network output

Scenario D includes three variations, all three variations include the naive data, a selection of the weather parameter data, month indicators and use a correlation coefficient filter of '|Correlation| > 0.05'. Each variation uses a selection of the top performing weather parameters, these weather parameters with the naive inputs are filtered based on the correlation filter and only those with a correlation > 0.05 are used as an input into the ANN. The three variations use the top 5, top 10 and top 15 weather parameters as inputs. Table 17 shows the three variations and the selected weather parameters based on results from scenario B. Like the previous scenarios the ANN model is the same as discussed in section 4.7 and shown in Table 10. Table 18 shows the results for each of the three variations in the same scorecard used in previous scenarios and used to represent the naive prediction in section 5.2.

5.6.3. Best result

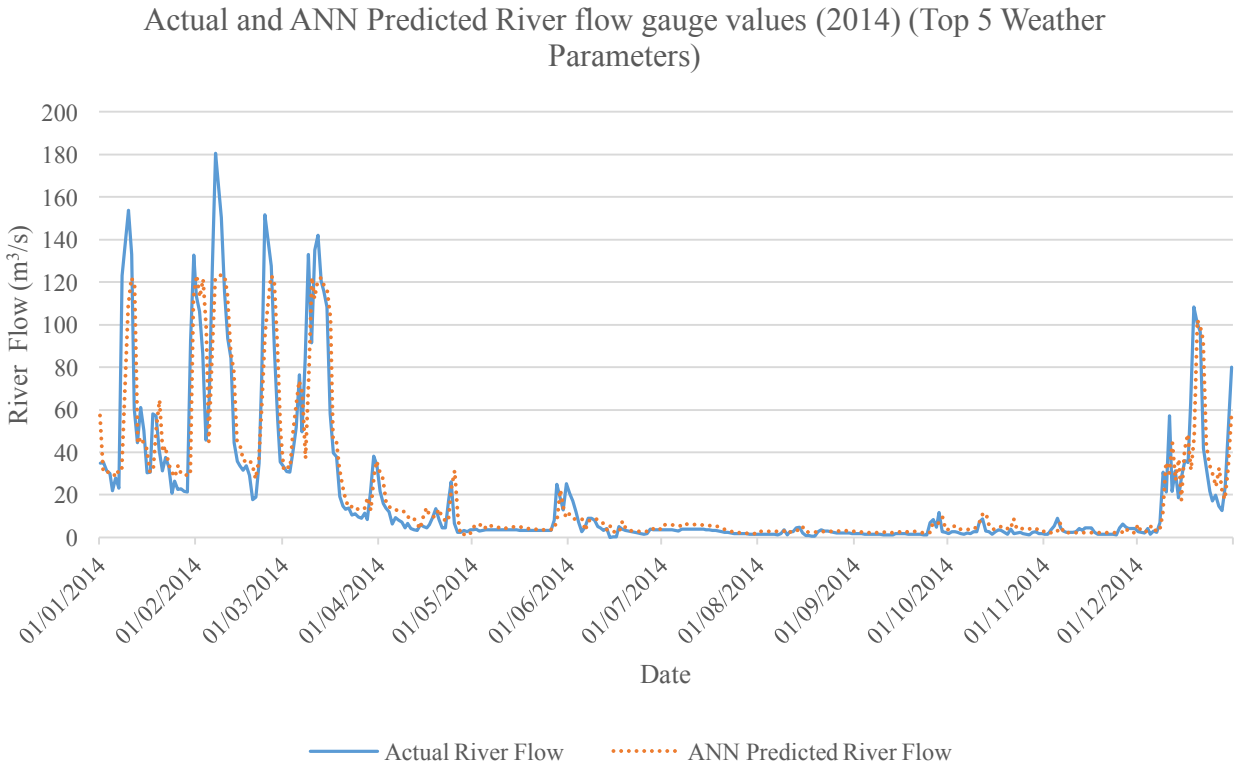
The best variation in scenario D was 'Top 5 weather parameters'. In this variation the top 5 weather parameters from scenario B are used in conjunction with the naive prediction data based on findings in scenario A. These inputs are finally filtered using the correlation filtering found in scenario C, variation '|Correlation| > 0,05'. Using this to identify the inputs into the ANN the variation had 51 input values, and using the pruning feature (in the Encog library described in section 4.3.2) had 10 neurons in the hidden layer. This input selection gave the best results across all the scenarios. Graph 29 shows the full predicted year of 2014 against the actual river flow gauge values for the year. As with the previous scenarios, Graph 30 shows the zoomed in period of January to March 2014 making it easier to see the difference between the predicted river flow gauge value and the actual river flow gauge value. Graph 31 shows a scatter plot representing the linear correlation between the ANN predicted river flow gauge values and the actual river flow gauge values for the year 2014.

Table 17: Top Weather parameters for scenario D

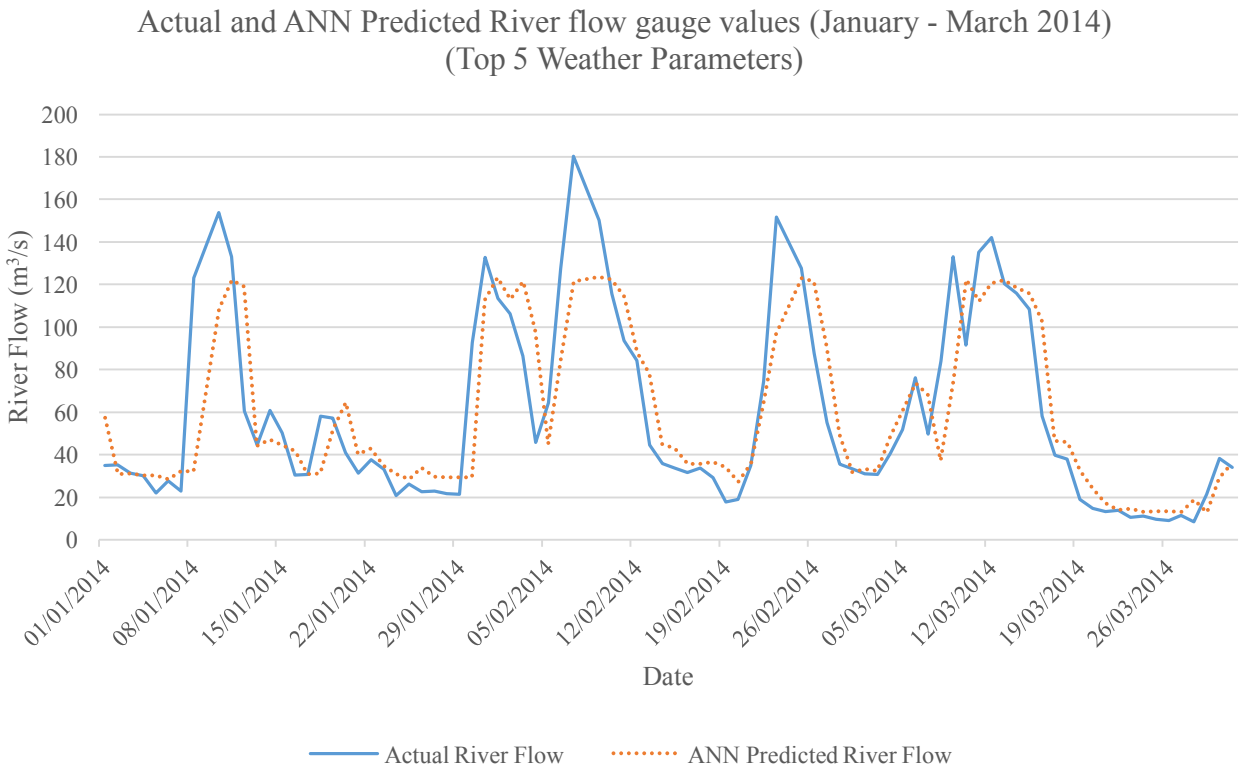
Variation	Weather Parameters
Top 5 Weather parameters	• Large-scale precipitation fraction (**lspf**) • Surface net thermal radiation (**str**) • Skin reservoir content (**src**) • Surface latent heat flux (**slhf**) • Top net thermal radiation clear sky (**ttrc**)
Top 10 Weather parameters	• Large-scale precipitation fraction (**lspf**) • Surface net thermal radiation (**str**) • Skin reservoir content (**src**) • Surface latent heat flux (**slhf**) • Top net thermal radiation clear sky (**ttrc**) • Volumetric soil water layer 2 (**swvl2**) • Boundary layer height (**blh**) • Large-scale precipitation (**lsp**) • Convective precipitation (**cp**) • Low cloud cover (**lcc**)
Top 15 Weather parameters	• Large-scale precipitation fraction (**lspf**) • Surface net thermal radiation (**str**) • Skin reservoir content (**src**) • Surface latent heat flux (**slhf**) • Top net thermal radiation clear sky (**ttrc**) • Volumetric soil water layer 2 (**swvl2**) • Boundary layer height (**blh**) • Large-scale precipitation (**lsp**) • Convective precipitation (**cp**) • Low cloud cover (**lcc**) • Medium cloud cover (**mcc**) • Northward turbulent surface stress (**nsss**) • Soil temperature level 2 (**stl2**) • Downward UV radiation at the surface (**uvb**) • Soil temperature level 3 (**stl3**)

Table 18: Results - Scenario D: Filter inputs based on Scenario A,B & C

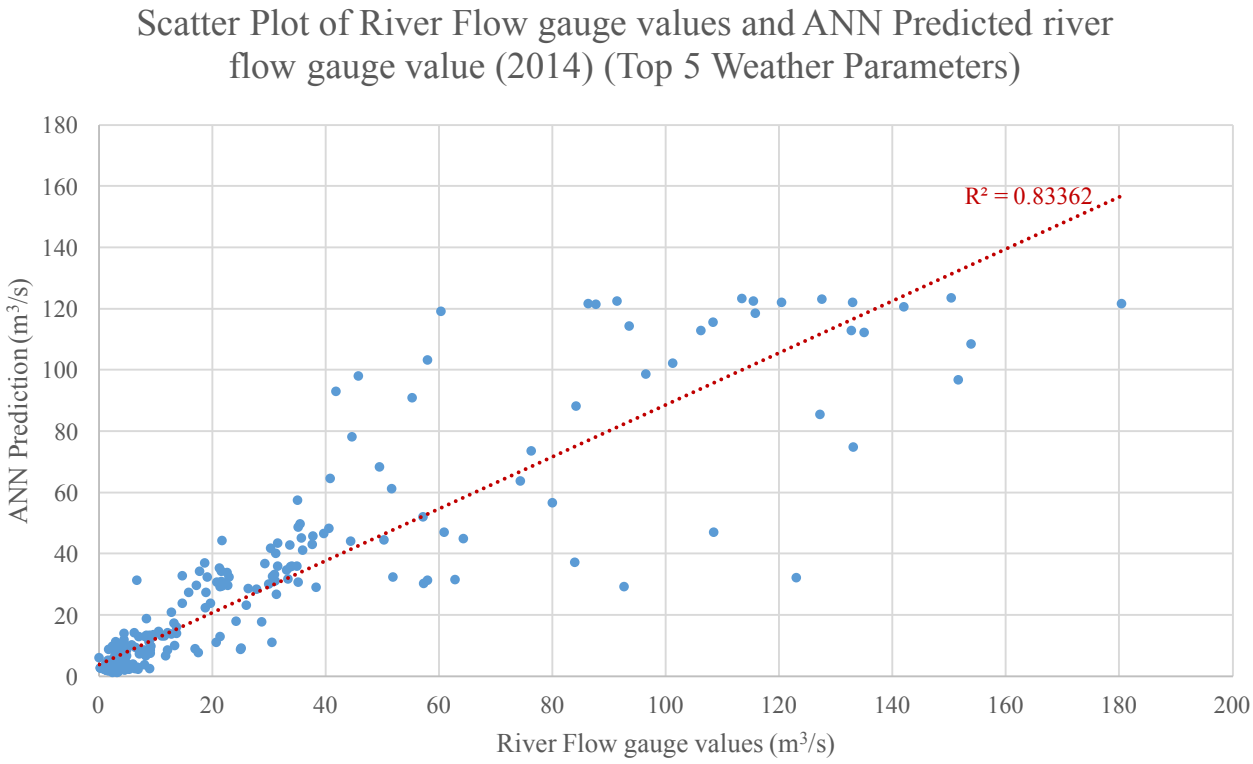
Variation	Optimized Structure	Training Set MSE	Validation Set MSE	Test Set MSE	Test Set (Year: 2014)			
					Correlation Coefficient	RMSE	MARE	NS
Top 5 weather parameters	51-10-1	170,82	99,06	179,82	0,9130	13,4096	78,3980	0,8328
Top 10 weather parameters	85-19-1	200,75	100,28	209,75	0,8986	14,4828	72,4697	0,8049
Top 15 weather parameters	115-15-1	170,00	105,30	213,28	0,8956	14,6040	62,8952	0,8017



Graph 29: Actual and ANN Predicted River flow gauge values (2014) (Top 5 Weather Parameters)



Graph 30: Actual and ANN Predicted River flow gauge values (January - March 2014) (Top 5 Weather Parameters)



Graph 31: Scatter Plot of River Flow gauge values and ANN Predicted river flow gauge value (2014) (Top 5 Weather Parameters)

5.6.4. Discussion

The first noticeable difference between the different variations in scenario D is the increase in inputs. All the variations had the same correlation filtering, the same naive inputs and the same month indicators. The only variation was the weather parameters included prior to correlation filtering. The number of inputs increased from 51 to 85 and then lastly to 115. The increase in inputs increases the computational complexity, but from the results, also decreased the performance. With an increase in inputs there was a similar increase in RMSE and a decrease in the Correlation Coefficient and NS. The only performance measure that did not follow the trend was MARE that actually showed a better performance the more inputs that were included. The scatter plot in Graph 31 shows a good linear relationship between the predicted river flow gauge values and the actual river flow gauge values with few outliers.

The second discussion point is around the hidden neurons in the hidden layer, there is no increase linked to the increased number of inputs. The variation 'Top 5 weather parameters' which has the least number of inputs also has the least number of neurons in the hidden layer, 10 neurons. The second variation 'Top 10 weather parameters' has 19 neurons in the hidden layer which is the most, yet it does not have the most inputs as the variation 'Top 15 weather parameters' has the most inputs. Graph 21 in scenario B found that the most common number of neurons in the hidden layer was 10 neurons, this incidentally is the same number of neurons in the best performing variation in scenario D.

5.6.5. Conclusion

Scenario D presents a summarized view of scenario A, scenario B and scenario C. It takes the findings of each scenario and combines this into a filtered variation, giving the ANN model the most logical data possible in an attempt to get the best predicted river flow gauge values. While it is not possible to state this is the best possible performance using an ANN, it is possible to see that scenario D has given the best variation results across all the scenarios. Section 5.6 summarizes the chapter and the findings while chapter 6 evaluates the findings, scenarios, and approach of this research to be able to draw conclusions in chapter 7.

5.7. Conclusion

Chapter 5 provides the results for this research, it looks at four scenarios and various resulting river flow gauge value predictions. Section 5.2 presented the naive prediction which just makes use of yesterdays river flow gauge value and assumes today will be the same. This presented a reasonably good prediction with a RMSE value of 14,9493 and a NS value of 0.7922. A value greater than 0,8 for NS is considered accurate (He et al. 2014), so the naive prediction NS value in section 5.2 did not leave much room for improvement. Scenarios A, B and C looked to identify those weather parameters that improved the ability of the ANN to predict accurately. The scenarios also looked to reduce the computational complexity of the ANN by removing weather parameters that did not perform well and further filtered based on linear correlation between the inputs and the river flow gauge values. scenario D incorporated all these various conclusions from scenario A,B and C to present a logically filtered data input set. This reduced the computational complexity and presented only the previous well performing data inputs.

The output from scenario D found a neural network using only the top 5 weather parameters and filter based on correlations greater than 0,05 gave the best results. The best results found in this scenario showed the neural network gave a better prediction than that of the naive prediction in section 5.2. The result from scenario D improved on the naive predictions RMSE by more than 10% giving an RMSE value of 13,4096. The RMSE performance measure was not the only measure that the ANN improved on. The ANN improved on the NS performance measure of the naive prediction by 0,0406, giving an NS value of 0,8328 which pushes it over the 0,8 value to be considered accurate. The ANN also gave an improved correlation coefficient value of 0,9130, improving on the naive predictions correlation coefficient value of 0.8962.

Overall the findings in chapter 5 show that the ANN gave a better prediction of the river flow gauge values for 2014 than the naive prediction. Chapter 6 evaluates the use of ANN's, highlighting the difficulties experienced, the advantages of the ANN and the overall performance from the various scenarios presented in chapter 5. Chapter 7 concludes the research by looking at and summarising the research, research area, the data used, the use of ANN's and the performance of those ANN's. Chapter 7 also draws a conclusion from the research as a whole.

6. EVALUATION OF USING ARTIFICIAL NEURAL NETWORKS FOR RIVER FLOW GAUGE VALUES

6.1. Introduction

The purpose of this chapter is to take a higher level analysis of the use of artificial neural networks (ANN's) for predict river flow gauges. Chapter 5 looked in-depth at various scenarios and the content was heavily weighted towards multiple variations and the outcomes of those variations. Chapter 6 looks into the use of neural networks and evaluates their use by doing high level analysis of the ANN results against the naive prediction. This analysis is done in section 6.2. The evaluation of using ANN's is not only limited to the results achieved in chapter 5 but also to the approached used in this research. A higher level analysis is presented discussing the criteria used in chapter 5, this is done in section 6.3. This research is aimed at evaluating the use of ANN's for predicting river flow gauge values. Section 6.4 discusses the appropriateness of using ANN's in light of the analysis in section 6.2 and 6.3. This chapter is concluded in section 6.5.

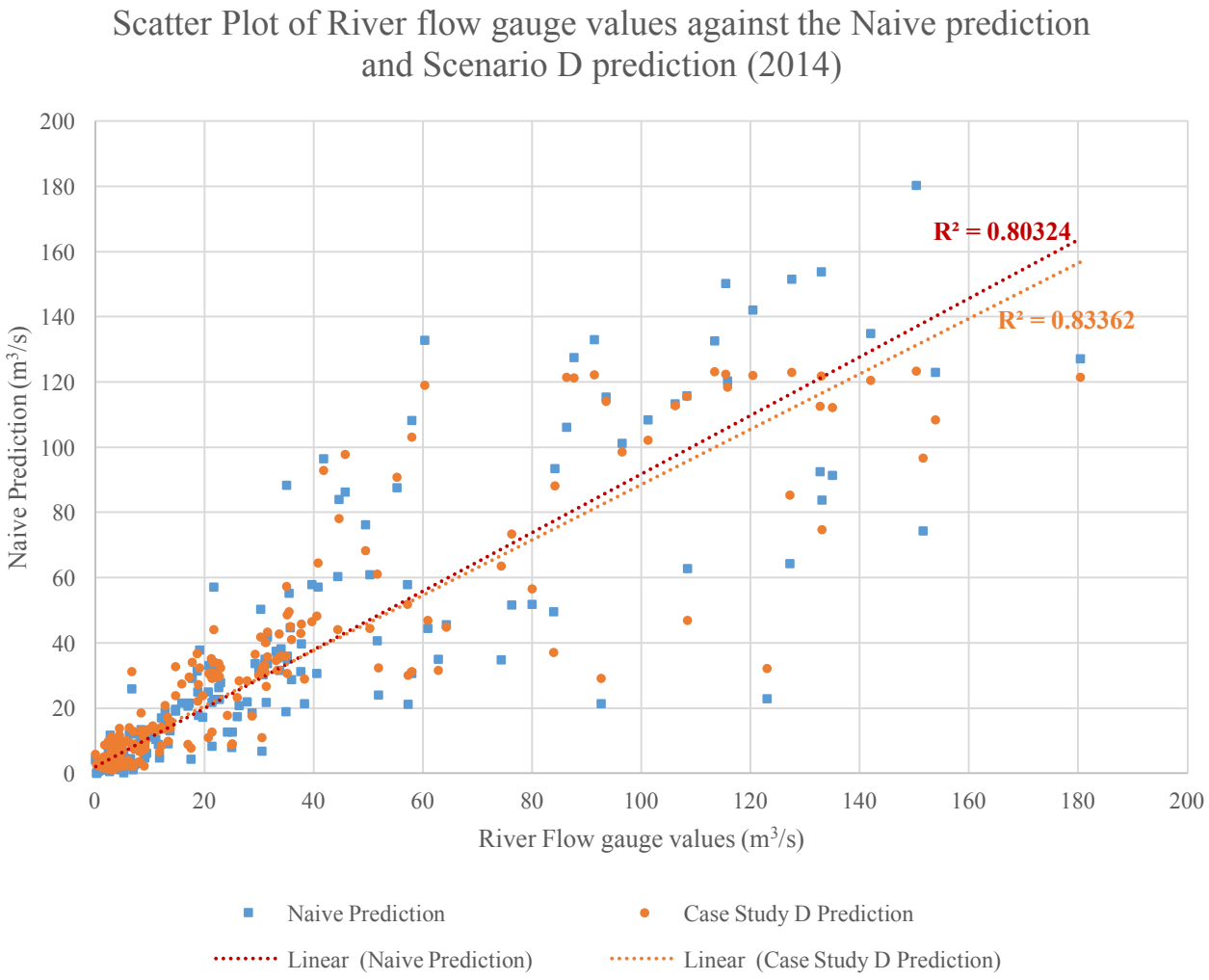
6.2. Analysis of scenario variations

Analysis of the individual scenarios needs to be done against the naive prediction presented in section 5.2. As mentioned in chapter 5, each scenario examines a certain aspect of the ANN. Each scenario has a number of variations. The findings from each of these scenarios is fed into the final scenario, scenario D, to get a logical set of input data. When considering the various scenarios, it is interesting to see that all the scenarios improved on the naive prediction. From this, it is possible to surmise that the ANN can predict the river flow gauge values better than the naive prediction. Intriguingly in scenario A even the ANN with only the naive prediction data can improve on the basic naive prediction. Table 19 summarizes the top results from each scenario with the naive prediction to clearly see the improvement the ANN made in predicting the river flow gauge value for 2014.

Table 19: Top scenario results against the naive prediction

Scenario	Optimized Structure	Training Set MSE	Validation Set MSE	Test Set MSE	Test Set (Year: 2014)			
					Correlation Coefficient	RMSE	MARE	NS
Naive Prediction	NA	182.10	103.88	223.48	0.8962	14.9493	42.6287	0.7922
Scenario A	544-9-1	174,80	107,47	205,57	0,8997	14,3378	65,3845	0,8088
Scenario B	26-10-1	168,36	100,03	197,53	0,9063	14,0546	80,8010	0,8163
Scenario C	449-19-1	165,44	106,33	205,06	0,8997	14,3201	55,0866	0,8093
Scenario D	51-10-1	170,82	99,06	179,82	0,9130	13,4096	78,3980	0,8328

If scenario D is taken as the best result by the ANN, the improvements can be seen in the correlation coefficient, RMSE and NS measurements. The side-by-side linear correlation is represented in Graph 32. The MARE performance measure is the only performance measure where the ANN could not improve on the naive prediction and the ANN over-predicts the river flow gauge value.



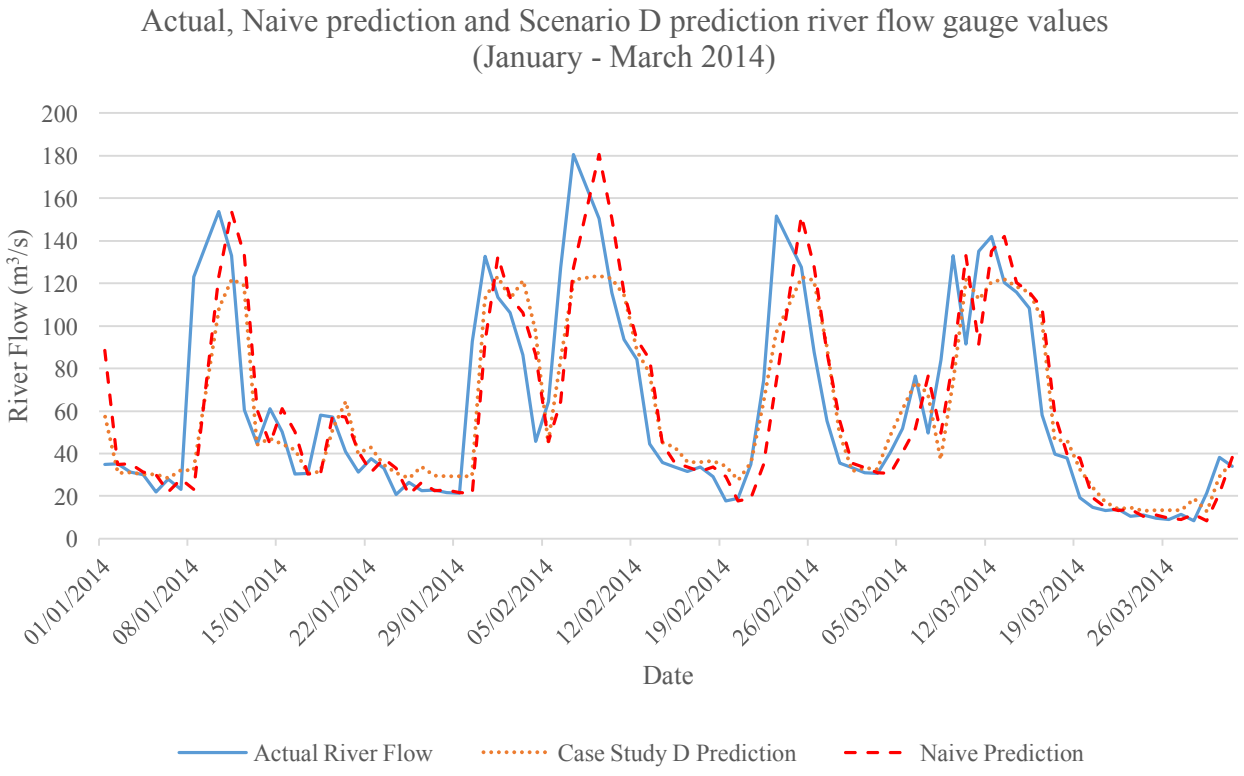
Graph 32:Scatter Plot of River flow gauge values against the Naive prediction and Scenario D prediction (2014)

When looking at individual days, a number of revealing numbers can also be revealed between the naive prediction and the ANN prediction in scenario D. The highest river flow gauge value in 2014, after data cleaning was 180,433 m^3/s on 7th February 2014. The naive prediction predicted 127,236 m^3/s and the scenario D prediction predicted 121,586 m^3/s. In this case the naive prediction was more accurate than the ANN. The lowest actual river flow gauge value was on the 14th June 2014 at 0,086 m^3/s. The naive prediction predicted 4,291 m^3/s and the ANN in scenario D predicted 5,844 m^3/s. In this variation the naive prediction again was more accurate than the ANN.

For the full test year of 2014 the naive prediction over predicted 67% of the days and under predicted 33% of the days. The ANN from scenario D over predicted 75% of the days and under predicted 25% of the days. The largest over prediction by the naive prediction was on the 12th January 2014 where it predicted the river flow gauge value would be 132,925 m^3/s,

meanwhile the actual river flow was 60,372 m^3/s. On this day the ANN was more accurate than the naive prediction by predicting a river flow gauge value of 118,984 m^3/s, which is still a large error. The largest under prediction by the naive prediction was on the 8th January 2014 where the actual river flow gauge value was 123,078 m^3/s but the naive prediction predicted 23,008 m^3/s. On this day again the ANN prediction was more accurate by predicting 32,154 m^3/s. Looking at the same situation with the scenario D prediction, the largest over prediction was on the 12th January 2014, the same day as the naive prediction as stated earlier. The largest under prediction by the ANN in scenario D was on the exact same day as the naive predictions under prediction, the 8th January.

Furthering this daily investigation one step further of the test year 2014, when the naive prediction over predicted it over predicted on average by 4,898 m^3/s. Similarly, when the naive prediction under predicted it under predicted by on average 9,687 m^3/s. The same analysis done for the scenario D prediction shows that when the ANN over predicted it over predicted by on average 4,820 m^3/s. When the ANN under predicted it under predicted by on average 11,274 m^3/s. This is a very important analysis because what we can determine from this is that while the ANN in scenario D over predicts a lot more than the naive prediction the average prediction error is smaller. It is also possible to determine that the ANN has smaller maximum over and under predictions than the naive prediction. This helps explain how the ANN in scenario D can give more accurate predictions than the naive prediction for 2014. Graph 33 presents the most fluctuating period of 2014, and shows the actual river flow gauge values, the naive predicted river flow gauge values and the river flow gauge values predicted by the ANN.



Graph 33: Actual, Naive prediction and scenario D prediction river flow gauge values (January - March 2014)

6.3. Analysis of scenario criteria

The research split the results and criteria up into a number of scenarios. Each criterion created a scenario, and each scenario had a number of variations based on the criteria. In chapter 5 there are 3 criteria that need to be analyzed. The first criterion looked at the use of weather parameter, and if they enhanced the ANN's ability to predict river flow gauge values. The second criterion was the individual weather parameters themselves, and which improved the ANN's ability to predict a river flow gauge value. The third and final criterion was the linear correlation between the weather parameters and the actual river flow gauge values.

Scenario A looked at isolating the effect of the weather parameters and identify if the neural network should have both the naive prediction data and the weather parameter data as inputs. As mentioned previously the scenario showed that the best approach was to include both the weather data and the naive prediction data as inputs into the follow-on scenarios. This criterion was clear in its results, firstly that the weather data by itself would not produce an accurate prediction, it needed a starting point close to the actual river flow gauge value. The naive prediction offered this. It also showed that when combining the weather data and the naive prediction the performance of the ANN would provide a better prediction than the naive prediction. The criterion was applicable to start the results section with as it fed into all the other scenarios. The criterion meant that three clearly distinct variations could be processed and there was limited possibility in ambiguity of the results. The outcome from the criterion for scenario A was as expected, by adding more data to the naive prediction it would be possible to improve the prediction of river flow gauge values.

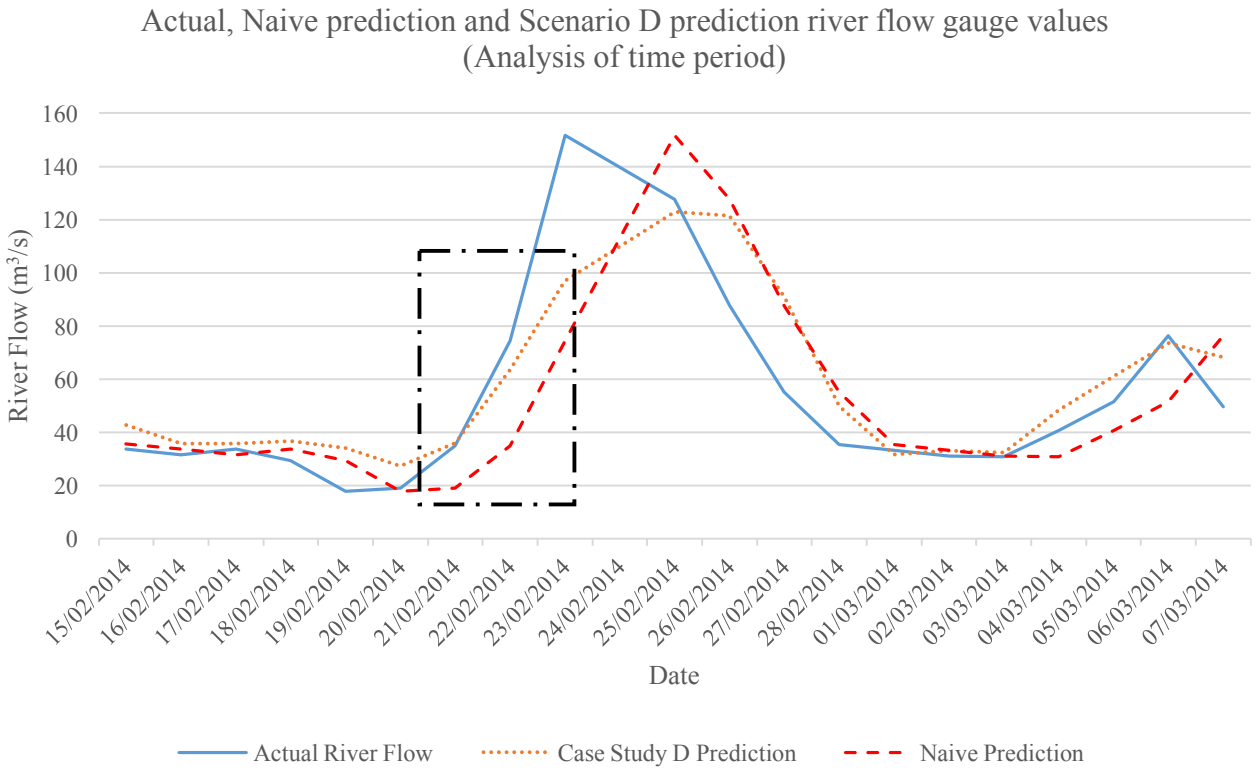
Scenario B isolated individual weather parameters, this is a more defined and isolated set of results than Scenario A. The criterion was defined to identify those individual weather parameters that improve the ANN. This is essentially a more detailed scenario of scenario A, and specifically the variation where both the naive prediction data and weather parameters are used. The criterion allowed for 75 distinct easily differentiable variations where only the weather parameter changed. The results from the criterion also allowed the findings to be fed into further scenarios. The results from this criterion also provided the expected result. Logically certain weather parameters have a larger impact on the river flow gauge value at that point in the river and it would be these parameters that improved the performance of the ANN. The identified weather parameters were not all expected, I had expected parameters such as Run-off ('ro') to have more of an impact but this was not one of the top performing weather parameter results. This criterion only varied one variable, that being the weather parameter, doing this identified the top 15 weather parameters that could be used in the following scenarios, specifically scenario D.

The final criterion was the correlation coefficient, and was used to define a number of variations in scenario C. This criterion was easy to separate various inputs into the different variations based on a correlation coefficient cut-off value. This criterion, while easy to separate, was less clear on the impact, the output was less conclusive and difficult to draw information or criteria to fed into the final scenario. It didn't give the expected results; it was expected that the accuracy would be improved as the correlation coefficient cut-off value increased. There was also an expectation that somewhere around the correlation coefficient value of 0,40 the performance would decrease as there would not be enough data being fed into the ANN. Neither of these expected results proved to be true. The output was not conclusive and the accuracy varied as the correlation cut-off value changed. A correlation-cut off of 0,05 did show to be generally the most accurate and this was fed into scenario D.

Scenario D used these three criteria to limit the input values and get a defined set of inputs for 3 variations. This reduced the computational complexity of the ANN and allowed a conclusive result to be found. The three criteria together formed a good definition and filtering mechanism for building the input data set and proved this by improving on the bench mark of the naive prediction from section 5.2.

6.4. Analysis of using artificial neural networks

This research has shown that by using an ANN it is possible to get a better prediction than the naive prediction. The naive prediction takes the previous days' river flow gauge value as todays river flow gauge value. When looking at this in section 5.2 it is possible to see that the naive prediction always lags behind the actual river flow gauge value. The results in chapter 5 and the discussion in section 6.2 highlight the fact the the ANN can improve on the naive prediction. Unlike the naive prediction the ANN in some area's of the test year correct the lag and move the prediction closer to the actual. This can be seen in Graph 34. At this point all the results and discussions have been towards ascertaining that the ANN gave a better predicted result than the naive prediction. This section considers the appropriateness of using an ANN to predict river flow gauge values.



Graph 34: Actual, Naive prediction and Scenario D prediction river flow gauge values (Analysis of time period)

The ANN was able to provide a better overall prediction for the year 2014, but in certain aspects the naive prediction did outperform the ANN. There are certain areas of the ANN which add complexity and could possibly reduce the ability for the ANN to predict the river flow gauge values. The aspects are:

- **The ANN is very sensitive to initial weights**
  This was overcome by using random weights and an ensemble. By using an ensemble and random weights the assumption is that one of the initialized weights would prove to generate a well trained but generalized ANN. This slows down training and requires a number of ANN to be used, while only one of them will be used. This was identified during the literature review and the model was setup with an ensemble to handle the sensitive initial weights.

- **Finding the best structure is done through a trial and error approach**
  This research made use of the Encog library which provided a form of pruning to determine the best structure. The best structure in this case was the number of hidden neurons in the hidden layer. If this was done manually or through the library, it involves trial and error to find the appropriate number of neurons in the hidden layer. The more neurons in the hidden layer the higher the computational complexity and in turn the longer it takes to train the ANN. Finding the best structure is key to ensuring a well trained, yet generalized ANN

- **Deviations from previous research**
  Previous research provides suggested configurations and proven means of ensuring a better performing ANN. The recommendations do not always provide the best results in every situation. In this research one of the main differences was the scaling of the input data to the domain of the activation function. This was discussed in section 4.5.5 where the input data was scaled between (0,1) instead of the recommended $\left[-\sqrt{3}, \sqrt{3}\right]$.

- **The model configuration is for a single study**
  The model provided in Table 10 in section 4.7 is the model that is applicable to this research. It is applicable to this input data, this river and possibly even the Driel barrage gauge station. It is therefore possible that this model could be improved, and that this model is not the most appropriate. This could only be confirmed through an indeterminate number of trial and error attempts with different model configurations.

- **When to stop training**
  When to stop training iterations is always a problem with ANN's as described in section 4.6. This research set the maximum number of iterations to 100, unless the early stopping condition was met which could stop the training before 100 iterations. Training could have continued to 1000 iterations or even 10000 iterations. This research model did not need to go to that many iterations to show ANN's can predict river flow gauge values more accurately than the naive prediction, but the results may have been more accurate had training continued.

The above aspects could be classified as the weaknesses of the the ANN approach to predicting river flow gauge values. These weaknesses and complexities were definitely experienced with this research. The use of ANN's also has a number of positives and the ability to extend and possibly improve the model is one of them. Due to the weakness in finding the best structure, iterations and inputs which are done through trial and error, there are a number of ways to over comes these weaknesses and possibly improve the results already found in this research. Possible improvements could include:

- **Ensemble**
  A possible improvement could be to increase the ensemble size. With a more powerful PC or more training time, additional neural networks could be included in the ensemble increasing the size from the current ensemble size of 100 ANN's. The ensemble could also

be managed differently. Using an average from all the ANN's rather than taking the ANN with the lowest generalization error. Lastly the ANN's in the ensemble could be different using a combination of different structures, different activation functions or different learning methods. By doing this the output from the ensemble could possibly improve.

- **Increase iterations**
  Currently the training iterations are limited to 100 iterations. This could stop the neural network from training further; increasing the training iterations and modifying the early stopping rules could possibly see an improvement in the accuracy of the prediction.

- **Additional data**
  The neural network was trained only using the weather parameters available from the European Center for Medium-Range Weather Forecasts (ECMWF). Additional data could either be in the form of more years of data or alternatively more types of data such as land use data or vegetation data. By adding additional data it may be possible to improve the accuracy of the prediction.

- **Different structure**
  The structure of the ANN plays an important role in the performance of the ANN. By changing the structure, the computational complexity could change, or the ANN may possibly be able to represent the river flow gauge values more accurately.

- **Different activation functions**
  This research made use of the sigmoid activation function for both the hidden layer and the output layer. Other activation functions may represent the river flow gauge values better or maybe able to transform the input data more appropriately. All activation functions have strengths, like the step function for binary data. There may be a more appropriate activation function for the river flow gauge value.

This research aimed at evaluating the use of ANN's and to determine if it is possible to use ANN's with weather parameter data to predict river flow gauge values, specifically in the Thukela river. From the results and discussions, it has been possible to highlight weaknesses and complexities with using ANN's to predict river flow gauge values. This section has also highlighted possible areas for improvement.

6.5. Conclusion

Chapter 6 evaluated the use of ANN's for predicting river flow gauge values at the Driel Barrage gauge station on the Thukela river. Scenarios were performed and recorded in chapter 5, each of these scenarios had a number of variations, where a single parameter was varied to see its effect on the performance of the ANN's ability to predict the river flow gauge value. Performing careful, detailed analysis of the outputs from the neural network against the naive prediction and the actual river flow gauge values, it is possible to conclude that the ANN generally over-predicted the river flow gauge values. The average amount that it over-predicted by was less than that of the naive prediction. Analysis of the largest over-predictions and under predictions showed that the naive prediction had much larger over and under predictions in the extreme cases. Since the research shows it is possible to predict the river flow gauge values more accurately than the naive prediction, an evaluation of the criteria and neural network needed to be done. The evaluation highlighted that the ANN used in this research had drawbacks and weaknesses but also highlighted the possible positive improvements that could

be done. Next, Chapter 7 concludes the research by summarizing the aim of the research, the research area, the findings of the research and the evaluations from this chapter before drawing final conclusions from this research.

7. CONCLUSION

7.1. Introduction

The Thukela river is a major river flowing through South Africa, it is vital to South Africa and provides water to both KwaZulu Natal and Gauteng. This river has a number of man-made water related infrastructure along its length in the form of dams and inter-basin transfer schemes. This research aimed to show that it would be possible to develop an artificial neural network (ANN) model based on predicted weather parameters, and so evaluate the benefits and criteria of this approach. The preceding chapters have laid out the context of the Thukela river and the significance of it to South Africa. The preceding chapters have also, in detail, examined the use of ANN's, presented a model for this research and documented a number of results based on the predictions by the ANN. Chapter 6 studied the results and analyzed the variations, selected criteria and lastly the use of ANN's for predicting river flow gauge values. This chapter provides a summarized overview of the all the previous chapters and draws final conclusions and remarks. Section 7.2 outlines the aim of the research, the context of the area being studied, high level comments on previous literature and the methodology used in the research. Section 7.3 discussed the research findings in light of the discussions and results within the overall research context. Section 7.5 considers possible future research, with section 7.4 concluding the research.

7.2. Literature review

The Thukela river plays an important part in ensuring continued water resources for South Africa. The South African Department of Agriculture, Forestry and Fisheries (DAFF) has the responsibility for ensuring future water resource availability and is responsible for the Thukela river. The Thukela river experiences natural variation in river flow values, variations from man-made dams and inter-basin transfer schemes. The Driel barrage in tertiary catchment V11 was selected as a study area as it provides a suitable case where river flow is affected by natural and external influences.

Intelligent systems in the form of ANN's are a form of predictive models. The literature review in Chapter 3 shows that neural networks have been previously been used in predicting river characteristics such as runoff, water temperature and river flow rates. Research did show that input data must be carefully selected and managed to improve the prediction accuracy when using ANN's. The research aimed to understand the benefits, limitations and accuracy with which an ANN can predict river flow gauge values.

7.3. Research design

The research made use of data from the European Centre for Medium-Range Weather Forecasts (ECMWF) and the South African DAFF. The ANN model was developed through the use of existing code libraries, namely the NetCDF and Encog libraries. These two libraries were used to generate the required ANN model. The research used a number of scenarios to assess the performance of the ANN model and such evaluate ANN models use and benefits in predicting river flow gauge values. The evaluation looked at identifying critical aspects and possible improvements.

7.4. Research findings

7.4.1. Results

The research findings were based on the concept of a naive prediction and the majority of the discussions and analysis was a comparison between the ANN results and the naive prediction. The scenarios for the results looked to identify the inputs which improved the ANN model and those that had a negative effect on the ANN performance. The result looked at predicting the river flow gauge values for the Driel barrage for the year of 2014.

The results showed that while the ANN with only the naive prediction data gave a RMSE value of 14,5673, the addition of the weather parameter data improved the performance of the ANN giving and RMSE value of 14,3378. The research then identified that not all weather parameters gave the same ANN performance. The weather parameters need to be filtered to remove those which negatively affected the ANN performance. This was with two approaches, firstly by removing the individual weather parameters that under performed. Secondly by using a linear relationship through a correlation coefficient. The values were filtered out using a correlation coefficient cut-off value, and any absolute value of the correlation coefficient that was less than the cut-off value was removed from the input data set.

The final scenario combined the findings of the previous scenarios at that point in the research to generate a few specific variations where only the top performing weather parameters were used as input and had the correlation filtering cut-off applied before training the ANN. From this scenario a best case variation for this research was achieved where the top 5 weather parameters were used as inputs into the ANN. In the variation the ANN model was able to achieve a RMSE value of 13,4096, correlation coefficient value of 0,9130, MARE value of 78,3980 and an NS value of 0,8328. All these performance measure were better than the naive prediction, other than MARE. These results showed that the ANN can predict the river flow gauge values for the Driel Barrage more accurately than the naive prediction.

7.4.2. Critical Aspects

After analysis of the model, the results and the approach used for the research, it is possible to identify critical aspects of using ANN and possible improvements to the model. The research identified the following critical aspects to using ANN's for predicting river flow gauge values:

- The ANN's model is **very sensitive to the initial weights** set in the ANN structure and a number of neural networks were required in the form on an ensemble to overcome this problem. This was identified during the research design through previous research (Engelbrecht 2007; Venkatesan et al. 2009; Mulia et al. 2013; Kim & Seo 2015) and the model was setup with an ensemble to handle the sensitive initial weights.
- The research design explained that it is difficult to determine **when to stop training** the ANN (Engelbrecht 2007). Without early stopping the performance of the model could be severely affected by overtraining or under training.

- The model was unable to use previous research recommendations when **scaling the input values**. The research made use of different scaling than previously suggested to achieve a better performing mode.
- The model is configured for this one research area with a relatively **limited data set**, additional data and changes to the configurations could improve or change the performance of the ANN

These critical aspects are areas within the research that directly affected the performance if the ANN model. Changes to these and improvements to these could possibly improve the performance of the model. The critical aspects also outline some of the model decisions made that affected the model design.

7.4.3. Improvements

The evaluation of the ANN also yielded a number of positive aspects and recommendations for predicting river flow using an ANN model. These positive aspects are areas that positively improved the performance of the ANN model:

- The **use of ensembles** did improve the performance of the ANN model and increasing the ensemble size or modifying the way the ensemble is used could give improved prediction accuracy.
- This research trained the ANN model to 100 iterations and this proved that the model can improve on the naive prediction's accuracy. **Increasing the iterations** or managing the stopping point in another way could provide more accurate predictions.
- The input data also proved to be a positive finding, this research identified the top weather parameters that gave a good result. Other weather parameter combinations could further improve the performance and **adding additional data** such as land use could further improve the predictive accuracy or the ANN.
- The **structure and activation functions** used in the ANN's could be changed or improved, there are many different structures and functions defined for ANN's. One of these ANN structures could provide a better river flow gauge value prediction.

These improvements present a number of aspects that can be used to improve the performance of the ANN. These could all still be used to even further improve the findings in this research and just because they have already been used improvements in these areas would still further improve the performance of the model.

7.5. Future research

The research used a single river in South Africa, a single catchment area and a single river flow gauge station. Future research on this topic would look at the following:

- Increasing the amount of input data, ensemble's and iterations of the model. This requires increased computing capabilities and additional data sources. Increasing these would allow the model to be trained longer and with more information improving the performance of the model.
- Comparing different neural network architectures, such as a recurrent neural network architecture to identify the best performing artificial neural network architecture.

- The hidden layer used a destructive approach to find assign the best neurons in the hidden layer. Constructive approaches are also possible and this could affect the construction of the hidden layer in turn effecting the performance of the artificial neural network.
- The artificial neural network made use of a sigmoid function for the output neuron with a wide range of other activation functions for the output possible, improvements to the neural network could be made by using different output activation functions. Some examples would be a limited ramp function or a linear activation function.
- Extending the research to other river flow gauge stations on the Thukela river. Possibly ones not affected by man-made structures or inter-basic transfer schemes. This would limit the influences on the river flow.
- Extend the research to the river flow at the Thukela river mouth, to assess the performance of the ANN model on an entire catchment area. This would evaluate the affect of weather variations across an entire catchment rather than a single small catchment area.
- Extend the research to other rivers. While the Thukela plays an important part in the South African context the research would need to consider more than one river system.
- The research would need to look at generalizing the model. Considering the river and catchment characteristics and create a model that could be used on other rivers. This would also need to consider removing the need for river flow gauge values as an input.

These future research ideas would look at building on the finding from this research, and as such allow a model to be developed that could be used in real world application to improve water management.

7.6. Conclusion

This research evaluated the use of an ANN model for predicting river flow gauge values. The research evaluated the ANN model and identified a number of critical aspects and areas of improvement within the ANN model approach. The research has shown that the ANN model can predict river flow gauge values but is heavily reliant in the data and model training.

An ANN model could present a powerful approach to river flow predictions. It removes some of the downfalls of other newer approach linked to satellite altimetry approaches and also presents an approach which makes use of already existing data. The approach is heavily influenced by available data though and data acquisition and feature selection is critical to improving the artificial neural model approach. Improved training, testing and parameterization would further improve the approach. The ever improving field of artificial intelligence would allow this approach to continually be improved and become more usable in a real world application, changing the data from being ‘data-rich but information-poor’ to ‘information-rich’, usable and accessible information drawn from rich data.

8. REFERENCES

Aichouri, I. et al., 2015. River Flow Model Using Artificial Neural Networks. *Energy Procedia*, 74, pp.1007–1014. Available at: http://www.sciencedirect.com/science/article/pii/S1876610215016008 [Accessed September 8, 2015].

Barbetta, S. et al., 2012. Enhancement and comprehensive evaluation of the Rating Curve Model for different river sites. *Journal of Hydrology*, 464–465, pp.376–387.

Beilfuss, R., 2012. *A risky climate for Southern African hydro: assessing hydrological risks and consequences for Zambezi river basin dams*, Berkeley, CA. Available at: internationalrivers.org.

van den Bergh, F. & Engelbrecht, A.P., 2001. Training Product Unit Networks using Cooperative Particle Swarm Optimisers. In *IEEE International Joint Conference on Neural Networks*. pp. 126–131.

Blumenfeld, S. et al., 2009. Water, Wetlands and Forests: A Review of Ecological, Economic and Policy Linkages. *Secretariat of the Convention on Biological Diversity and Secretariat of the Ramsar Convention on Wetlands, Montreal and Gland, CBD Technical Series No. 47*. Available at: https://www.cbd.int/doc/publications/cbd-ts-47-en.pdf [Accessed May 20, 2017].

Bourblanc, M. & Blanchon, D., 2014. The challenges of rescaling South African water resources management: Catchment Management Agencies and interbasin transfers. *Journal of Hydrology*, 519, pp.2381–2391. Available at: http://www.sciencedirect.com/science/article/pii/S0022169413005714 [Accessed November 5, 2015].

Braune, E., 1985. Aridity and hydrological characteristics: Chairman's summary. In *Perspectives in Southern Hemisphere Limnology*. Dordrecht: Springer Netherlands, pp. 131–136. Available at: http://www.springerlink.com/index/10.1007/978-94-009-5522-6_9 [Accessed October 6, 2016].

Buchanan, B.T.J. & Somers, W.P., 1969. Discharge measurements at gaging stations. *Techniques of Water-Resources Investigations of the United States Geological Survey - Book 3 - Applications of Hydarulics*, p.171. Available at: http://pubs.er.usgs.gov/browse/usgs-publications.

Chai, T. & Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), pp.1247–1250. Available at: http://www.geosci-model-dev.net/7/1247/2014/gmd-7-1247-2014.html [Accessed November 15, 2015].

Chang, F.-J., Chen, L. & Chang, L.-C., 2005. Optimizing the reservoir operating rule curves by genetic algorithms. *Hydrological Processes*, 19(11), pp.2277–2289. Available at: http://doi.wiley.com/10.1002/hyp.5674 [Accessed October 8, 2016].

Chang, Y.-T., Chang, L.-C. & Chang, F.-J., 2005. Intelligent control for modeling of real-time reservoir operation, part II: artificial neural network with operating rule curves. *Hydrological Processes*, 19(7), pp.1431–1444. Available at: http://doi.wiley.com/10.1002/hyp.5582 [Accessed May 2, 2016].

Dee, D.P. et al., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656), pp.553–597. Available at: http://doi.wiley.com/10.1002/qj.828 [Accessed June 12, 2016].

Department of Water and Forestry, 2002. Thukela water project feasibility study water resource evaluation and system analysis task. Available at: https://www.dwa.gov.za/thukela/Reports/Docs/pdf/Water Resource Evaluation and Systems Analysis Task.pdf.

Department of Water and Forestry, 2004. *Thukela WMA: Internal Strategic Perspective*, Available at: https://www.dwa.gov.za/Documents/Other/WMA/7/ThukelaISPNov04.htm.

Department of Water and Sanitation, 2016a. Department of Water and Sanitation : About Us. Available at: https://www.dwaf.gov.za/about.aspx#vision [Accessed June 12, 2016].

Department of Water and Sanitation, Driel Barrage. Available at: https://www.dwa.gov.za/hydrology/weekly/Photo.aspx?photo=V1R002.jpg.

Department of Water and Sanitation, 2016b. DWS: Hydrology Data Sets Tugela River - Driel V1H058. Available at: https://www.dwa.gov.za/Hydrology/HyDataSets.aspx?Station=V1H058 [Accessed November 16, 2015].

Department of Water and Sanitation, 2016c. DWS Hydrological Services - Surface Water (Data,Dams, Floods and Flows). Available at: https://www.dwaf.gov.za/Hydrology/ [Accessed June 9, 2016].

Department of Water and Sanitation, V1H058. Available at: https://www.dwaf.gov.za/Hydrology/HyImage.aspx?Station=V1H058.

Department of Water and Sanitation, Woodstock Dam. Available at: https://www.dwa.gov.za/hydrology/weekly/Photo.aspx?photo=V1R003.jpg.

DeWeber, J.T. & Wagner, T., 2014. A regional neural network ensemble for predicting mean daily river water temperature. *Journal of Hydrology*, 517, pp.187–200. Available at: http://www.sciencedirect.com/science/article/pii/S0022169414003990 [Accessed June 29, 2015].

Dube, T. et al., 2016. The Impact of Climate Change on Agro-Ecological Based Livelihoods in Africa: A Review. *Journal of Sustainable Development*, 9(1), p.256. Available at: http://www.ccsenet.org/journal/index.php/jsd/article/view/52039.

Ejaz Qureshi, M., Hanjra, M.A. & Ward, J., 2013. Impact of water scarcity in Australia on global food security in an era of climate change. *Food Policy*, 38, pp.136–145.

Elsafi, S.H., 2014. Artificial Neural Networks (ANNs) for flood forecasting at Dongola Station in the River Nile, Sudan. *Alexandria Engineering Journal*, 53(3), pp.655–662. Available at: http://www.sciencedirect.com/science/article/pii/S1110016814000660 [Accessed September 8, 2015].

Engelbrecht, A.P., 2007. *Computational Intelligence An Introduction* Second Edi., Pretoria, South Africa: John Wiley & Sons, Ltd.

Engelbrecht, A.P. & Ismail, A., 1999. Training Product Unit Neural Networks. *Stability and Control: Theory and Applications*, (2), pp.59–74.

European Centre for Medium-Range Weather Forecasts, 2016a. ECMWF: What We Do. Available at: http://www.ecmwf.int/en/about/what-we-do [Accessed June 12, 2016].

European Centre for Medium-Range Weather Forecasts, 2016b. *ECMWF: Who We Are*, Available at: http://www.ecmwf.int/en/about/who-we-are.

European Centre for Medium-Range Weather Forecasts, 2016c. ERA-Interim. Available at: http://www.ecmwf.int/en/research/climate-reanalysis/era-interim [Accessed June 12, 2016].

European Centre for Medium-Range Weather Forecasts, 2016d. ERA-Interim : Use of data from this server. Available at: http://apps.ecmwf.int/datasets/data/interim-full-daily/licence/ [Accessed June 12, 2016].

Ewisa, Thukela River the fearsome one. , pp.1–23. Available at: http://www.ewisa.co.za/misc/school/eBOOK_PDFs/THUKELARIVER.pdf.

Galiano, V. et al., 2010. PyPnetCDF: A high level framework for parallel access to netCDF files. *Advances in Engineering Software*, 41(1), pp.92–98. Available at: http://www.sciencedirect.com/science/article/pii/S0965997809001458 [Accessed March 13, 2016].

Gaustad, K. et al., 2014. A scientific data processing framework for time series NetCDF data. *Environmental Modelling & Software*, 60, pp.241–249. Available at: http://www.sciencedirect.com/science/article/pii/S1364815214001704 [Accessed April 14, 2016].

Google Earth 7.1, 2016. Google Earth with data from the South African Department of Water and Sanitation. Available at: https://www.dwa.gov.za/iwqs/wms/data/000key2data.asp.

Gravelle, R., 2015. Equipment Estimation : Techniques and Equipment. In *Geomorphological Techniques (Online Edition)*. London: British Society for Geomorphology, pp. 1–8. Available at: http://www.geomorphology.org.uk/sites/default/files/geom_tech_chapters/3.3.5_DischargeEstimation.pdf.

Grobler, D. & Ntsaba, M., 2004. *Strategic Framework for National Water Resource Quality Monitoring Programmes*, Pretoria, South Africa. Available at: https://www.dwa.gov.za/iwqs/wrmais/National_Water_Resource_Quality_strategy_ed01_dr05_final.pdf [Accessed April 24, 2016].

He, Z. et al., 2014. A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. *Journal of Hydrology*, 509, pp.379–386. Available at: http://www.sciencedirect.com/science/article/pii/S0022169413008780 [Accessed July 23, 2015].

Heaton, J., 2015. Encog: Library of Interchangeable Machine Learning Models for Java and C#. *Journal of Machine Learning Research*, 16, pp.1243–1247. Available at: http://jmlr.org/papers/v16/heaton15a.html [Accessed November 15, 2015].

Heaton, J., 2014. Encog 3.3: Development Guide. Available at: https://s3.amazonaws.com/heatonresearch-books/free/encog-3_3-devguide.pdf.

Hill, E., 2014. Flood Death Toll in South Africa Climbs to 32. *FloodList*. Available at: http://floodlist.com/africa/south-africa-floods-kill-32 [Accessed June 19, 2016].

Kim, S.E. & Seo, I.W., 2015. Artificial Neural Network ensemble modeling with conjunctive data clustering for water quality prediction in rivers. *Journal of Hydro-environment Research*, 9(3), pp.325–339. Available at: http://www.sciencedirect.com/science/article/pii/S1570644315000192 [Accessed April 30, 2015].

Kröse, B.J.A. & van der Smagt, P., 1991. *Neural Networks*, Thousand Oaks: SAGE Publications, Inc. Available at: http://0-srmo.sagepub.com.innopac.up.ac.za/view/neural-networks/SAGE.xml.

Kwanula, 2016. KwaZulu-Natal Agricultural Union - Kwanula. Available at: http://www.kwanalu.co.za [Accessed October 4, 2016].

KZN Business Chambers Council, 2016. KZN Business Chambers Council. Available at: http://www.kznchamber.co.za [Accessed October 4, 2016].

Leerink, L. et al., 1995. Learning with Product Units. *Advances in Neural Information Processing Systems*, (7), p.537.

McCartney, M., Cai, X. & Smakhtin, V., 2013. *Evaluating the Flow Regulating Functions of Natural Ecosystems in the Zambezi River Basin* (IWMI Rese., Colombo, Sri Lanka: International Water Management Institute (IWMI).

Miller, N. & Jinwon, K., 1995. Numerical prediction of precipitation and river flow over the Russian river watershed during the January 1995 California storms. *Bulletin of the American Meteorological Society*.

Mkamba, L., 2013. Tugela floods sweep away homes and families. *IOL*. Available at: http://www.iol.co.za/news/south-africa/kwazulu-natal/tugela-floods-sweep-away-homes-and-families-1469486 [Accessed June 19, 2016].

Montello, D. & Sutton, P., 2006. *An Introduction to Scientific Research Methods in Geography*, 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc. Available at: http://sk.sagepub.com/books/an-introduction-to-scientific-research-methods-in-geography [Accessed June 11, 2016].

Mulia, I.E. et al., 2013. Hybrid ANN–GA model for predicting turbidity and chlorophyll-a concentrations. *Journal of Hydro-environment Research*, 7(4), pp.279–299.

O'Keeffe, J.H., 1989. Conserving rivers in Southern Africa. *Biological Conservation*, 49(4), pp.255–274. Available at: http://www.sciencedirect.com/science/article/pii/0006320789900475 [Accessed November 5, 2015].

Pappenberger, F. et al., 2015. How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *Journal of Hydrology*, 522, pp.697–713.

Potter, C. et al., 2010. Modeling river discharge rates in California watersheds. *Journal of Water and Climate Change*. Available at: https://www.researchgate.net/profile/Vanessa_Genovese/publication/240792725

_Modeling_river_discharge_rates_in_California_watersheds/links/00b4952d449878e0cd000000.pdf [Accessed April 24, 2016].

Reid, V., 2014. Biodiversity, rewilding and human population growth. *Biodiversity*. Available at: http://www.tandfonline.com/doi/abs/10.1080/14888386.2014.898218 [Accessed November 15, 2015].

Rickly Hydrological Company, 2016. Rickly Hydrological Company. Available at: http://www.rickly.com/sgi/AA.htm [Accessed October 8, 2016].

Riedmiller, M. & Braun, H., 1993. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. *IEEE International Conference on Neural Networks*, pp.586–591.

Riedmiller, M. & Braun, H., 1992. A Fast adaptive learning algorithm. *Seventh International Symposium on Computer and Information Sciences*, pp.279–285.

Rupson Industries, 2016. Water Current Meter - Pygmy. Available at: http://www.indiamart.com/rupsonindustries/conventional-meteorology-equipments.html#water-current-meter-pygmy [Accessed October 8, 2016].

Santos, R.M.B. et al., 2014. The impact of climate change, human interference, scale and modeling uncertainties on the estimation of aquifer properties and river flow components. *Journal of Hydrology*, 519, pp.1297–1314. Available at: http://www.sciencedirect.com/science/article/pii/S0022169414006714 [Accessed September 8, 2015].

Santra, A.K., Chakraborty, N. & Sen, S., 2009. Prediction of heat transfer due to presence of copper–water nanofluid using resilient-propagation neural network. *International Journal of Thermal Sciences*, 48(7), pp.1311–1318. Available at: http://www.sciencedirect.com/science/article/pii/S1290072908002330 [Accessed May 29, 2016].

SAPA, 2014. Mpumalanga flood damage cost R535m. *SABC News*. Available at: http://www.sabc.co.za/news/a/3c5ae500437e4e219cce9da64eba5fdc/Mpumalanga-flood-damage-cost-R535m-20140204 [Accessed June 19, 2016].

SASA, 2016. South African Sugar Association. Available at: http://www.sasa.org.za [Accessed October 5, 2016].

Seo, Y.-S. & Bae, D.-H., 2013. On the value of outlier elimination on software effort estimation research. *Empirical Software Engineering*, 18(4), pp.659–698. Available at: http://link.springer.com/10.1007/s10664-012-9207-y [Accessed June 12, 2016].

Shamseldin, A.Y., 2010. Artificial neural network model for river flow forecasting in a developing country. *Journal of Hydroinformatics*, 12(1), p.22.

Smith, L.C. & Pavelsky, T.M., 2008. Estimation of river discharge, propagation speed, and hydraulic geometry from space: Lena River, Siberia. *Water Resources Research*, 44(3).

South African Government, 2016. Water and Sanitation. Available at: http://www.gov.za/about-sa/water-affairs [Accessed June 12, 2016].

Staatskoerant, 2012. National Water Act: Establishment of water management areas. , (35517), pp.25–29. Available at:

http://www.gov.za/sites/www.gov.za/files/35517_gon547.pdf [Accessed October 5, 2016].

Staatskoerant, 2016. National Water Act: New water management areas of South Africa. , (1056), pp.169–172. Available at: http://www.gov.za/sites/www.gov.za/files/40279_gon1056.pdf [Accessed October 5, 2016].

Stanley, K.O. & Miikkulainen, R., 2002. Evolving Neural Networks through Augmenting Topologies. *Evolutionary Computation*, 10, pp.99–127. Available at: http://nn.cs.utexas.edu/downloads/papers/stanley.ec02.pdf.

Statistics South Africa, 2001. 2001 Census. Available at: http://www.statssa.gov.za/?page_id=3892 [Accessed October 2, 2016].

Statistics South Africa, 2011a. *Census 2011 Municipal report – KwaZulu-Natal*, Available at: www.statssa.gov.za.

Statistics South Africa, 2010. National accounts: Water Management Areas in South Africa. *Discussion Document: D0405.8*. Available at: https://www.statssa.gov.za/publications/D04058/D04058.pdf [Accessed October 5, 2016].

Statistics South Africa, 2015. *Statistical release Gross domestic product*, Pretoria, South Africa. Available at: http://www.statssa.gov.za/publications/P0441/P04413rdQuarter2014.pdf.

Statistics South Africa, 2011b. Statistics by place. Available at: http://www.statssa.gov.za/?page_id=964 [Accessed October 2, 2016].

Steinfeld, C.M.M. et al., 2015. A simulation tool for managing environmental flows in regulated rivers. *Environmental Modelling & Software*, 73, pp.117–132. Available at: http://www.sciencedirect.com/science/article/pii/S1364815215300323 [Accessed September 8, 2015].

Su, T. et al., 2016. Multi-dimensional visualization of large-scale marine hydrological environmental data. *Advances in Engineering Software*, 95, pp.7–15. Available at: http://www.sciencedirect.com/science/article/pii/S0965997816300175 [Accessed March 25, 2016].

Tarpanelli, A. et al., 2013. River Discharge Estimation by Using Altimetry Data and Simplified Flood Routing Modeling. *Remote Sensing*, 5(9), pp.4145–4162. Available at: http://www.mdpi.com/2072-4292/5/9/4145/.

The Water Project, 2016. Water in crisis - South Africa. Available at: https://thewaterproject.org/water-crisis/water-in-crisis-south-africa [Accessed June 19, 2016].

Thiéblemont, D., 2016. *Geological Map of Africa at 1:10 M scale*, Available at: http://portal.onegeology.org/OnegeologyGlobal/.

UNESCO, 2016. Maloti-Drakensberg Park. Available at: http://whc.unesco.org/en/list/985 [Accessed October 4, 2016].

Unidata, 2013. Clarification on _FillValue, missing_value, valid_xxx. Available at: http://www.unidata.ucar.edu/mailing_lists/archives/netcdfgroup/2013/msg00291.html [Accessed May 20, 2017].

Unidata, 2016a. NetCDF-Java javadoc. Available at: http://www.unidata.ucar.edu/software/thredds/v4.5/netcdf-java/javadoc/index.html [Accessed May 28, 2016].

Unidata, 2016b. Network Common Data Form (NetCDF). Available at: http://doi.org/10.5065/D6H70CW6 [Accessed November 15, 2015].

Unidata, 2016c. Unidata - About Us. Available at: http://www.unidata.ucar.edu/about/ [Accessed May 28, 2016].

Unidata, 2016d. Where is NetCDF Used? Available at: http://www.unidata.ucar.edu/software/netcdf/usage.html [Accessed May 28, 2016].

Venkatesan, D., Kannan, K. & Saravanan, R., 2009. A genetic algorithm-based artificial neural network model for the optimization of machining processes. *Neural Computing and Applications*, 18(2), pp.135–140. Available at: http://link.springer.com/10.1007/s00521-007-0166-y [Accessed June 19, 2016].

Vogel, R.M. & Fennessey, N.M., 1995. Flow duration curves II: a review of applications in water resources planning. *Journal of the American Water Resources Association*, 31(6), pp.1029–1039. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1752-1688.1995.tb03419.x/pdf.

WESSA, 2016. WESSA. Available at: http://www.wessa.org.za/ [Accessed October 4, 2016].

Wilson, A.J., 2001a. *Thukela Water Management Area Situational Assessment : Part 1*, Available at: https://www.dwa.gov.za/Documents/Other/CMA/Thukela/thukela.htm.

Wilson, A.J., 2001b. *Thukela Water Management Area Situational Assessment : Part 2*, Available at: https://www.dwa.gov.za/Documents/Other/CMA/Thukela/thukela.htm.

Wilson, A.J., 2001c. *Thukela Water Management Area Situational Assessment : Part 3*, Available at: https://www.dwa.gov.za/Documents/Other/CMA/Thukela/thukela.htm.

Wilson, A.J., 2001d. *Thukela Water Management Area Situational Assessment : Part 4*, Available at: https://www.dwa.gov.za/Documents/Other/CMA/Thukela/thukela.htm.

Wong, T.-T., 2015. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9), pp.2839–2846. Available at: http://www.sciencedirect.com/science/article/pii/S0031320315000989 [Accessed May 30, 2016].

Zhang, Y. & Yang, Y., 2015. Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, 187(1), pp.95–112. Available at: http://www.sciencedirect.com/science/article/pii/S0304407615000305 [Accessed May 10, 2016].

9. ANNEX

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Natural and Agricultural Sciences
Ethics Committee

E-mail: ethics.nas@up.ac.za

Date: 09 September 2016

ETHICS SUBMISSION: LETTER OF APPROVAL

Name of Applicant	Prof S Coetzee
Department	Geography, Geoinformatics and Meteorology
Reference number	EC160814-066
Title	Evaluating the use of neural networks to predict river flow gauge values

Dear Prof S Coetzee

The submission conforms to the requirements of the NAS EC. Any amendments must be submitted to the NAS EC on a relevant application form as used for the original application quoting the reference number and detailing the required amendment. An amendment would be for example differentiating within the research target population.

You are required to submit a progress report no later than two months after the anniversary of this application as indicated by the reference number. The progress report document is accessible of the NAS faculty's website: Research/Ethics Committee.

You are required to notify the NAS EC upon the completion or ending of the project using the form Project Completed. Completion will be when the data has been analysed and documented in a postgraduate student's thesis or dissertation, or in a paper or a report for publication.

The digital archiving of data is a requirement of the University of Pretoria. The data should be accessible in the event of an enquiry or further analysis of the data.

The NAS EC wishes you well with your research project.

Yours sincerely,

Chairperson
NAS Ethics Committee

Figure 26: NAS Ethics Committee approval letter

Table 20: DWS quality codes for river flow gauge values

Quality code	Print code	Description
1		Good continuous data
2		Good edited data
3		Preserved historical data
4	Q	Unaudited
5		Height derived from flow
6	D	Drops
7	Q	Good edited unaudited
25	Q	Unaudited Gauge Plate Readings / dip level readings
26	$	Audited Gauge Plate Readings / dip level readings
27	&	Good monthly reading
50	S	Gap filled data
58	H	Hb not available - assume no submergence
59	V	Static or reverse flow due to backwater submergence conditions
60	A	Above Rating
64	E	Audited Estimate
65	E	Unaudited Estimate
66	*	Program Estimate
70	?	Unknown
78	U	Not Accumulated Unreliable
79	%	Accumulated Unreliable
80	+	Accumulated Reliable
81	#	Wet day within accumulated rainfall period
90	<	GW: Water level below instrument
91	>	Minimum Value
92	<	Maximum Value
93	<	Dry borehole
94	>	Artesian borehole level
95	<	Borehole Seepage
100	?	Flag Boshielo Dam under construction. New FSL not yet implemented.
130	E	Used previous week's level as an estimate for this week
140	!	Data not yet checked
150	^	Rating table extrapolated - flows estimated
151	M	Data Missing
152	~	Negative
153	F	No height data - Flow data only
154	[	Reversal start
155	[	Reversal end
160	Z	No info for stage/discharge determination (zero dt loaded)
161	T	Rating missing
162	R	Rating unreliable

163	G	Gate(s) in operation - no spillway discharge
164	B	Continuously variable submergence flow derivation DT in operation
165	P	Estuarine water level recording only - no flow calculated
170	M	Permanent Gap
172	M	Temporary Gap
173	?	Data Unreliable
201	[	Data not recorded or incomplete
245	V	Undefined submergence flow calc program exception
246	M	No cross-sectional area upstream of notch/structure
247	V	ha > rating table limit - no calculation performed
248	V	No hb data
249	V	No ha data
250	V	Structural submergence > 97.7%
251	V	Static or reverse flow possibilities
252	V	Froude number > 0.8 at inlet section
253	V	Flow not converging to constant val after max # iterations
254	A	Rating Table Exceeded
255	M	Data Missing

Table 21: ECMWF ERA-Interim parameter list

Description	Short Name	Units	Data Type
Clear sky surface photosynthetically active radiation	parcs	J m**-2	double
Snow albedo	asn	(0 - 1)	double
Snow density	rsn	kg m**-3	double
Volumetric soil water layer 1	swvl1	m**3 m**-3	double
Volumetric soil water layer 2	swvl2	m**3 m**-3	double
Volumetric soil water layer 3	swvl3	m**3 m**-3	double
Volumetric soil water layer 4	swvl4	m**3 m**-3	double
Snow evaporation	es	m of water equivalent	double
Snowmelt	smlt	m of water equivalent	double
10 metre wind gust since previous post-processing	fg10	m s**-1	double
Large-scale precipitation fraction	lspf	s	double
Downward UV radiation at the surface	uvb	J m**-2	double
Photosynthetically active radiation at the surface	par	J m**-2	double
Convective available potential energy	cape	J kg**-1	double
Total column liquid water	tclw	kg m**-2	double
Total column ice water	tciw	kg m**-2	double
Surface pressure	sp	Pa	double
Total column water	tcw	kg m**-2	double

Total column water vapour	tcwv	kg m**-2	double
Soil temperature level 1	stl1	K	double
Snow depth	sd	m of water equivalent	double
Large-scale precipitation	lsp	m	double
Convective precipitation	cp	m	double
Snowfall	sf	m of water equivalent	double
Boundary layer dissipation	bld	J m**-2	double
Surface sensible heat flux	sshf	J m**-2	double
Surface latent heat flux	slhf	J m**-2	double
Mean sea level pressure	msl	Pa	double
Boundary layer height	blh	m	double
Total cloud cover	tcc	(0 - 1)	double
10 metre U wind component	u10	m s**-1	double
10 metre V wind component	v10	m s**-1	double
2 metre temperature	t2m	K	double
2 metre dewpoint temperature	d2m	K	double
Surface solar radiation downwards	ssrd	J m**-2	double
Soil temperature level 2	stl2	K	double
Surface thermal radiation downwards	strd	J m**-2	double
Surface net solar radiation	ssr	J m**-2	double
Surface net thermal radiation	str	J m**-2	double
Top net solar radiation	tsr	J m**-2	double
Top net thermal radiation	ttr	J m**-2	double
Eastward turbulent surface stress	ewss	N m**-2 s	double
Northward turbulent surface stress	nsss	N m**-2 s	double
Evaporation	e	m of water equivalent	double
Soil temperature level 3	stl3	K	double
Low cloud cover	lcc	(0 - 1)	double
Medium cloud cover	mcc	(0 - 1)	double
High cloud cover	hcc	(0 - 1)	double
Sunshine duration	sund	s	double
Eastward gravity wave surface stress	lgws	N m**-2 s	double
Northward gravity wave surface stress	mgws	N m**-2 s	double
Gravity wave dissipation	gwd	J m**-2	double
Skin reservoir content	src	m of water equivalent	double
Maximum temperature at 2 metres since previous post-processing	mx2t	K	double
Minimum temperature at 2 metres since previous post-processing	mn2t	K	double
Runoff	ro	m	double
Total column ozone	tco3	kg m**-2	double
Top net solar radiation, clear sky	tsrc	J m**-2	double
Top net thermal radiation, clear sky	ttrc	J m**-2	double

Surface net solar radiation, clear sky	ssrc	J m**-2	double
Surface net thermal radiation, clear sky	strc	J m**-2	double
TOA incident solar radiation	tisr	J m**-2	double
Total precipitation	tp	m	double
Instantaneous eastward turbulent surface stress	iews	N m**-2	double
Instantaneous northward turbulent surface stress	inss	N m**-2	double
Instantaneous surface sensible heat flux	ishf	W m**-2	double
Instantaneous moisture flux	ie	kg m**-2 s**-1	double
Skin temperature	skt	K	double
Soil temperature level 4	stl4	K	double
Temperature of snow layer	tsn	K	double
Convective snowfall	csf	m of water equivalent	double
Large-scale snowfall	lsf	m of water equivalent	double
Forecast albedo	fal	(0 - 1)	double
Forecast surface roughness	fsr	m	double
Forecast logarithm of surface roughness for heat	flsr	~	double
longitude	longitude	degrees_east	float
latitude	latitude	degrees_north	float
time	time	hours since 1900-01-01 00:00:0.0	int

Table 22: Affect of input scaling

Input Variables	**RMSE with inputs scaled between $[-\sqrt{3}, \sqrt{3}]$**	**RMSE with inputs scaled between (0,1)**
Naive and parcs	16,0588	15,0445
Naive and fg10	18,3490	15.2644
Naive and twcv	16.2480	14.8393

Table 23: Correlation Coefficient of Input data

Parameter	**1 Day**	**7 Days**	**30 Days**	**90 Days**	**180 Days**	**365 Days**	**1095 Days**
naive	0,8654	0,7459	0,6460	0,4299	0,1828	0,1915	0,1865
parcs	0,3714	0,3914	0,4351	0,4511	0,2884	-0,2814	-0,3503
asn	0,0336	0,0503	0,0912	0,1317	0,0736	0,0168	0,0010
rsn	-0,0404	-0,0580	-0,1031	-0,1432	-0,0900	-0,0210	0,0053
swvl1	0,5091	0,4872	0,4709	0,4116	0,1474	0,0589	0,0277
swvl2	0,5290	0,4934	0,4716	0,3974	0,1261	0,0545	0,0256
swvl3	0,5263	0,4878	0,4528	0,3369	0,0590	0,0422	0,0167

swvl4	0,4263	0,3966	0,3318	0,1436	-0,0680	0,0341	0,0163
es	0,0079	-0,0075	-0,0419	-0,0841	-0,0265	-0,0562	-0,0870
smlt	0,0023	-0,0200	-0,0501	-0,0705	0,0167	0,0021	-0,0234
fg10	-0,1157	-0,2245	-0,2754	-0,1349	0,1757	0,0398	0,0415
lspf	0,3231	0,4690	0,5212	0,5301	0,3446	0,1721	0,1110
uvb	0,1019	0,1883	0,3166	0,3644	0,2465	-0,1623	-0,1064
par	0,1402	0,2228	0,3383	0,3813	0,2563	-0,1648	-0,1095
cape	0,1618	0,2656	0,3487	0,4163	0,2640	-0,0057	-0,0101
tclw	0,2134	0,3995	0,4713	0,5194	0,3526	0,1488	0,1022
tciw	0,1505	0,2944	0,3613	0,4364	0,3159	0,0328	-0,0308
sp	-0,1155	-0,1852	-0,2930	-0,4204	-0,3290	-0,0390	0,0001
tcw	0,4118	0,4898	0,4901	0,4496	0,2398	0,1732	0,1169
tcwv	0,4110	0,4885	0,4893	0,4483	0,2383	0,1722	0,1169
stl1	0,3416	0,3738	0,4005	0,3995	0,2495	0,1068	0,1018
sd	-0,0034	-0,0276	-0,0631	-0,0897	-0,0270	-0,0020	0,0020
lsp	0,2066	0,3808	0,4690	0,5489	0,3899	0,1480	0,1431
cp	0,2607	0,4202	0,4874	0,5234	0,3381	0,1457	0,0936
sf	0,0062	-0,0115	-0,0414	-0,0540	0,0029	-0,0059	0,0108
bld	-0,1721	-0,2775	-0,3236	-0,2065	0,1223	0,0025	-0,0070
sshf	-0,1949	-0,2625	-0,3183	-0,3487	-0,2459	0,0590	0,0660
slhf	-0,2611	-0,3243	-0,4131	-0,4249	-0,2311	0,0588	0,0238
msl	-0,1522	-0,2514	-0,3445	-0,4190	-0,3055	-0,0748	-0,0314
blh	0,1769	0,2854	0,3694	0,4645	0,3955	0,1534	0,1014
tcc	0,2438	0,3842	0,4354	0,4814	0,3325	0,1399	0,0850
u10	-0,1849	-0,3249	-0,3956	-0,4040	-0,2341	0,0489	0,1203
v10	-0,0590	-0,1528	-0,2364	-0,3564	-0,3792	-0,1083	-0,1211
t2m	0,3145	0,3658	0,3974	0,3978	0,2573	0,1303	0,1211
d2m	0,3732	0,4212	0,4394	0,4162	0,2115	0,0566	0,0201
ssrd	0,0675	0,1531	0,2925	0,3514	0,2505	-0,1582	-0,1026
stl2	0,3180	0,3479	0,3798	0,3787	0,2308	0,0472	0,0440
strd	0,3960	0,4360	0,4453	0,4379	0,2738	0,1441	0,0789
ssr	0,0855	0,1721	0,3066	0,3577	0,2443	-0,1594	-0,1049
str	0,3716	0,4660	0,4839	0,4658	0,2355	0,0900	0,0406
tsr	0,2117	0,2753	0,3682	0,4103	0,2837	-0,1598	-0,1066
ttr	0,1501	0,2298	0,2240	0,3455	0,2818	0,0211	-0,0362
ewss	-0,2113	-0,3222	-0,4103	-0,3497	-0,0566	0,0211	-0,0124
nsss	0,0967	0,1412	0,1793	0,0534	-0,2692	-0,0827	0,0380
e	-0,2611	-0,3243	-0,4131	-0,4249	-0,2311	0,0588	0,0239
stl3	0,3724	0,3829	0,3937	0,3512	0,1433	0,0636	0,0484
lcc	0,2536	0,3976	0,4646	0,4909	0,3232	0,1681	0,1071
mcc	0,2255	0,3867	0,4424	0,4737	0,3218	0,0437	-0,0553
hcc	0,1100	0,2547	0,3142	0,4240	0,3347	0,0832	0,0416

sund	0,0575	0,1657	0,3174	0,3858	0,2608	-0,1266	-0,0958
lgws	-0,1807	-0,2814	-0,3690	-0,3737	-0,2345	-0,0088	-0,0076
mgws	-0,0147	-0,0646	-0,1273	-0,2879	-0,3963	-0,1143	-0,0122
gwd	-0,0935	-0,1976	-0,2898	-0,2727	-0,1035	0,0193	0,0309
src	0,3066	0,4368	0,4758	0,5058	0,3368	0,1469	0,0628
mx2t	0,2890	0,3456	0,3821	0,3855	0,2532	0,0914	0,0903
mn2t	0,3128	0,3642	0,3958	0,3965	0,2570	0,1252	0,1136
ro	0,5350	0,4997	0,4383	0,2637	0,0378	0,0787	0,0439
tco3	-0,0963	-0,0968	-0,0516	0,1016	0,2778	-0,0130	-0,0257
tsrc	0,3643	0,3849	0,4303	0,4530	0,2992	0,0367	-0,0360
ttrc	-0,1108	-0,1822	-0,2996	-0,3118	-0,2022	-0,0422	-0,0774
ssrc	0,3640	0,3849	0,4316	0,4496	0,2892	-0,1723	-0,1634
strc	0,3502	0,4069	0,3910	0,3050	0,0382	0,0688	0,0338
tisr	0,3559	0,3772	0,4249	0,4524	0,3058	-0,3708	-0,3668
tp	0,2543	0,4317	0,4977	0,5459	0,3699	0,1582	0,1264
iews	-0,1866	-0,3151	-0,3903	-0,3451	-0,0967	0,0726	0,0856
inss	0,0343	0,0355	0,0074	-0,1598	-0,3720	-0,1182	-0,0717
ishf	-0,2400	-0,3775	-0,4186	-0,3947	-0,1759	0,0116	0,0376
ie	-0,0503	-0,0618	-0,1140	-0,2672	-0,3150	0,0042	-0,0026
skt	0,3455	0,3874	0,4113	0,4110	0,2622	0,1559	0,1272
stl4	0,3789	0,3691	0,3318	0,1906	-0,1149	0,0969	0,0633
tsn	0,3013	0,3451	0,3846	0,3896	0,2450	0,0926	0,0856
csf	0,0028	-0,0141	-0,0411	-0,0446	0,0085	0,0199	0,0182
lsf	0,0074	-0,0099	-0,0399	-0,0551	0,0017	-0,0143	0,0079
fal	-0,3079	-0,3636	-0,4114	-0,3785	-0,1569	0,0028	0,0101
fsr	0,2807	0,4134	0,4577	0,4951	0,3326	0,1523	0,0603
flsr	0,2190	0,3311	0,4081	0,4598	0,3025	0,1128	0,0518