# ASSESSING TECHNIQUES FOR SELECTING A CLIMATE DRIVER STATION FOR A STUDY CATCHMENT

**by**

**Thobeka Xolo**

**Submitted in fulfilment of the academic requirements of**

**Master of Science in Hydrology**

Centre for Water Resources Research

School of Agricultural, Earth and Environmental Science

College of Agriculture, Engineering and Science

University of KwaZulu-Natal

Pietermaritzburg, South Africa

**February 2025**

# PREFACE

The research contained in this thesis was completed by the candidate while based in the Centre for Water Resources Research at the University of KwaZulu-Natal in Pietermaritzburg, South Africa. The research was financially supported by the Water Research Commission.

The contents of this work have not been submitted in any form to another university and, except where the work of others is acknowledged in the text, the results reported are due to investigations by the candidate.

Signed: R Kunz

Date: February 2025

# DECLARATION 1: PLAGIARISM

I, Thobeka Xolo, declare that:

(i) the research reported in this dissertation, except where otherwise indicated or acknowledged, is my original work;

(ii) this dissertation has not been submitted in full or in part for any degree or examination to any other university;

(iii) this dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons;

(iv) this dissertation does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

   a. their words have been re-written but the general information attributed to them has been referenced;

   b. where their exact words have been used, their writing has been placed inside quotation marks and referenced;

(v) where I have used material for which publications followed, I have indicated in detail my role in the work;

(vi) this dissertation is primarily a collection of material, prepared by myself, published as journal articles or presented as a poster and oral presentations at conferences. In some cases, additional material has been included;

(vii) This dissertation does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged and the source being detailed in the dissertation and in the references sections.

Signed: Thobeka Xolo

Date: February 2025

# ABSTRACT

The accurate assessment and modelling of hydrological processes relies heavily on comprehensive and reliable climate data. In South Africa, the alarming decline in the number of climate monitoring stations and the poor quality of observed data (i.e. missing records) present a significant challenge to reliable hydrological modelling. In addition, it is crucial to select climate driver stations as representative as possible of the catchment being studied. Climate driver stations are essential for capturing representative climate conditions necessary for water resources management and planning. This study assessed existing techniques used for selecting climate driver stations for a catchment. Available daily rainfall and temperature datasets were infilled and extended to create a 70-year record for quaternary catchments C41A to C41E, which are situated in the Lejweleputswa District Municipality (Free State Province, South Africa). The Inverse Distance Weighting method was used to infill rainfall data, whilst the Mean Temperature Difference method, Difference in Standard Deviation method and a ranking algorithm method were used to infill missing temperature data. Rainfall driver stations were selected using the common Driver Station (DS) method as well as the Adjustment Factor (AF) method which is closely related to the *CalcPPTcor* approach. Pseudo temperature driver stations were selected for each selected rainfall driver station using a revised ranking algorithm. The Dent et al. (1989) median, Lynch (2004) median, Lynch (2004) mean and Pegram et al. (2016) mean gridded rainfall datasets were compared for their performance in estimating rainfall adjustment factors using the $R^2$, Nash-Sutcliffe Efficiency and Root Mean Square Error statistics. Each gridded dataset was then used to verify the methods for selecting a climate driver station DS and AF methods. The *ACRU* model was used to simulate inflow to the Erfenis Dam, which was then compared to a dam water balance as a means of verifying which method performed better. Key findings showed that the Pegram mean gridded datasets (monthly and annual) perform better in enhancing the representativeness of station rainfall for the study catchment. The results for the AF and DS methods were inconclusive due to various challenges, i.e. having no observed streamflow for the study catchment. It is recommended that the Pegram mean grids be considered when deriving rainfall adjustment factors, which are applied to the rainfall driver station to improve the representativity of catchment rainfall. It is recommended that the DS and the AF methods be re-evaluated in another study catchment with more climate stations and a reliable streamflow monitoring network.

# ACKNOWLEDGMENTS

First and foremost, I would like to thank God for granting me the perseverance, strength and wisdom to complete this study. “I can do all things, through Christ who strengthens me” (Philippians 4:13).

I would like to express my deepest gratitude to my supervisor, Mr Richard Kunz, for his continuous support, patience and immense knowledge throughout this research. His guidance assisted me throughout the research and writing of this dissertation.

Special thanks to the Water Research Commission for their financial support. Without their funding, this research would not have been possible. I am also grateful to the South African Weather Service, the Agricultural Research Council and the Department of Water and Sanitation for providing climate data.

I extend my appreciation to Mr Mark Horan, Dr David Clark and Dr Stefanie Schütte for their valuable advice and also for making time to assist with data issues and setup of the *ACRU* model. Their assistance and expertise were crucial for the completion of this research. I also appreciate the Centre for Water Resources Research (CWRR) for the resources they provided that made the research possible and the support given by other CWRR colleagues. I am also grateful to the College of Agriculture Engineering and Science student support team at UKZN; their availability helped me maintain my mental health during this time.

I am also grateful to my friends and family for supporting and tirelessly encouraging me during the challenging times of my study. Their love and faith motivated me to keep going. Finally, I acknowledge all the unnamed persons who contributed to this research in one way or another. Your support and efforts are sincerely appreciated.

# TABLE OF CONTENTS

## LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

The following abbreviations are frequently used in this dissertation:

| | |
|---|---|
| ACRU | Agricultural Catchments Research Unit |
| AF | Adjustment Factor |
| ALTF | Altitude factor |
| APAN | A-pan equivalent reference evaporation |
| ARC | Agricultural Research Council |
| AWS | Automatic Weather Station |
| CAES | College of Agriculture Engineering and Science |
| CHIRPS | Climate Hazards Group InfraRed Precipitation with Station data |
| CSAG | Climate Systems and Analysis Group |
| CWRR | Centre for Water Resources Research |
| DALT | Difference in Altitude between control and target station |
| DFFE | Department of Forestry Fisheries and Environment |
| DEM | Digital Elevation Model |
| DSTF | Distance Factor |
| DIST | Distance between station and quinary catchment |
| DM | District Municipality |
| DRE | Daily Rainfall Extraction |
| DSD | Difference in Standard Deviation |
| DWS | Department of Water and Sanitation |
| DS | Driver Station |
| EMA | Expectation Maximisation Algorithm |
| EO | Earth Observation |
| ET | Evapotranspiration |
| $ET_O$ | Evaporative Demand |
| FAO | Food and Agriculture Organisation |
| GIS | Geographic Information System |
| GWR | Geographical Weighted Regression |
| IDW | Inverse Distance Weighting |
| ISCW | Institute for Climate, Soil and Water |

| | |
|---|---|
| LR | Lapse Rate |
| LRR | Lapse Rate Region |
| MIT | Monthly Infilling Technique |
| MAP | Mean Annual Precipitation |
| ML | Mega Litres |
| MR | Median Ratio |
| MTD | Mean Temperature Difference |
| NSE | Nash-Sutcliffe Efficiency |
| PRESS | Predicted Error of Sum Squares |
| QC | Quaternary Catchment |
| QCDB | Quaternary Catchment Data Base |
| RF | Ranking Factor |
| RFL | Rainfall |
| RMSE | Root Mean Square Error |
| SANLC | South African National Land Cover |
| SASA | South African Sugar Association |
| SAWB | South African Weather Bureau |
| SAWS | South African Weather Service |
| STRMFL | Observed streamflow |
| $T_{MAX}$ | Maximum temperature |
| $T_{MIN}$ | Minimum Temperature |
| TRMM | Tropical Rainfall Measuring Mission |
| UCT | University of Cape Town |
| UN | United Nations |
| UKZN | University of KwaZulu-Natal |
| USFLOW | Unit Streamflow |
| WASA | Wind Atlas for South Africa |
| WMA | Water Management Areas |
| WRC | Water Research Commission |

# 1 INTRODUCTION

## 1.1 Background

Water resource availability in South Africa is of particular concern owing to growing water demands by, *inter alia*, agriculture, industry and a rapidly growing population. In addition, future water resource availability is affected by the potential impacts of climate change. Some catchments are already faced with demand exceeding supply, resulting in a need to transfer significant volumes of water from neighbouring catchments (Clark, 2015). Therefore, considering the projected changes in future climate, there is a need to focus more on how water resources are currently managed. To better manage water resources, one must first acknowledge that "you cannot manage what you cannot measure" (Patrinos, 2014).

To address these challenges, extensive climate databases have been developed, such as those by Dent et al. (1989), Lynch (2004), Pegram et al. (2016) and Schulze and Maharaj (2004). Dent et al. (1989) introduced one of the earliest spatial databases for rainfall in South Africa, employing rigorous data collection and quality control measures. This was followed by Lynch (2004), who refined the Dent dataset using advanced techniques to generate high-resolution rainfall surfaces. Pegram et al. (2016) further improved these methodologies by integrating updated data sources and introducing significant statistical approaches to handle mixed discrete-continuous nature of rainfall data. Schulze and Maharaj (2004) developed a comprehensive temperature dataset comprising daily minimum and maximum temperature for 973 stations across South Africa spanning 1950-2000. Despite their utility, the temporal limitation of these datasets necessitates extending and updating records to accommodate ongoing hydrological and climate modelling efforts.

In addition to these datasets, the delineation of catchments into quaternary and quinary catchments plays a pivotal role in understanding and managing water resources. South Africa's Department of Water and Sanitation (DWS) initially divided the country into 1 946 quaternary catchments for operational planning. Recognising the need for greater hydrological homogeneity, Schulze and Horan (2011) introduced the quinary catchments by sub-delineating quaternary catchments into 3 altitudinal zones (referred to as the upper, middle, and lower quinaries).

This refinement resulted in 5 838 hydrologically interlinked quinary catchments, which exhibit more uniform hydrological and agricultural responses compared to their quaternary counterparts. These advancements laid the groundwork for the development of the Quaternary Catchment Data Base (QCDB; Schulze et al., 2005) and its successor, the Quinary Catchment Data Base (QnCDB; Schulze et al., 2011).

The QCDB, initiated in the late 1980s, has undergone numerous revisions to improve its accuracy and applicability. The QnCDB expanded upon the QCDB, integrating data at the quinary catchment level to support refined hydrological modelling. Both databases are foundational for water resource assessments, with their standardised datasets providing critical inputs for climate and hydrological models. These databases, however, are limited to the 1950-1999 period, necessitating the identification of new "driver" stations to extend climate records and address the growing data gaps caused by station closures.

The above-mentioned datasets have been used to calculate rainfall adjustment factors and to apply elevation corrections. Rainfall adjustment factors were developed to address variations in observed rainfall due to geographic and altitudinal differences. Using Lynch's (2004) raster database of median monthly rainfall, monthly adjustment factors were calculated for each catchment by determining the ratio of spatially averaged rainfall over the catchment to the median rainfall at the station level. These factors were then applied to the driver station's daily rainfall records, resulting in corrected datasets that better represent the spatial variability of rainfall across catchments. The difference in elevation between selected temperature stations and the spatially averaged elevation of catchments was also corrected for. Regional lapse rates, as outlined by Schulze and Maharaj (2004), were applied to temperature data to adjust for these elevation differences, ensuring that temperature inputs more accurately reflected the catchment's climatic conditions. The adjustments utilised a 1' by 1' spatial resolution digital elevation model to calculate representative elevation for each catchment. By applying these corrections, the datasets provided more accurate temperature estimates, which are critical for calculating evapotranspiration, solar radiation, and vapour pressure deficits in hydrological modelling.

Furthermore, the DWS updated the quaternary catchment boundaries in 2018 (DWS, 2019). This resulted in the need to revise both the Schulze et al. (2005) and the Schulze and Horan (2011) catchment boundaries, which are needed for water resource assessments at different catchment scales. The quinary catchments were recently updated by Clark et al. (2024) and are now referred to as altitudinal zones. The “driver” stations previously selected for the quaternary catchments (QCs) are no longer operational, and thus new “driver” stations must be selected so that climate records can be extended beyond 1999. The closure of the stations is due to the decline in the number of operational stations.

Climate stations are essential for gathering local weather data, providing essential information for long-term climate monitoring, forecasting of weather events and scientific research. Climate data from weather stations are essential for understanding water availability within a catchment, providing crucial information such as rainfall, temperature, and evaporation data. To best represent each catchment, climate "driver" stations are selected using various techniques tailored to each variable (e.g. rainfall and temperature) (Schulze and Pike, 2004). However, no standard technique exists, and current methods are neither completely accurate nor error-free. This makes it vital to improve these techniques to enhance water resource management. While past methods benefited from a larger network of rain gauges and temperature stations, many stations have since closed due to issues such as insufficient funding (Erasmus, 2022), further underscoring the need to refine and adapt driver station selection methods. Throughout this document, temperature refers to air temperature, unless stated otherwise.

## 1.2 Justification

Faniriantsoa and Dinku (2022) stated that various parts of Africa have experienced a decrease in the number of meteorological stations over the past 50 years and that current stations are not evenly distributed spatially. Many of the current stations have missing data, making them unreliable for use until the records are infilled. According to Lynch (2004), the number of active stations in South Africa has declined since 1930 (**Figure 1.1a**).

Approximately 4 out of 10 South African Weather Services (SAWS) weather stations are closed every year due to a lack of funding to maintain these stations (Erasmus, 2022). **Figure 1.1b** also shows the significant decline in rain gauges in South Africa from 1970 to 2009, with an approximate decrease of 55 % from 3 261 to 1 450 (Pegram et al., 2016). Pitman and Bailey (2021) also stated that an extension of graphs like **Figure 1.1a** and **Figure 1.1b** could well show a further decline from 2009 onwards**. Figure 1.1c** shows the relatively sparse SAWS station coverage in the country, as there are approximately 164 Automatic Weather Stations (AWSs) covering 1.2 million km² (Sinclair and Pegram, 2010).

Temperature is typically measured using an AWS and there has been a noticeable decrease in the number of stations across South Africa in recent years, as shown in **Figure 1.1d**. With fewer temperature stations, there is insufficient spatial coverage, particularly in remote or less populated areas. This reduction in spatial coverage limits the availability of localised temperature data, which is crucial for understanding regional climate patterns. Furthermore, some of the "driver" stations previously selected for each quaternary catchment are no longer in operation, resulting in a need to select new "driver" stations for each catchment (to extend existing records from 1999 onwards).

## 1.3 Aims and objectives

The main aim of this study is to assess existing techniques for selecting representative climate driver stations for a study catchment, in order to possibly refine existing methods or even develop a new approach. This requires the following objectives to be met first: (i) selection of a suitable study catchment, (ii) development of extended daily rainfall datasets for this study catchment, and (iii) assessment of existing gridded rainfall datasets (i.e. Dent median, Lynch median, Lynch mean and Pegram mean) for estimating daily rainfall adjustment factors for the study catchment. Finally, the performance of each method for selecting a climate driver station was evaluated by comparing hydrological simulations of streamflow against observed data.

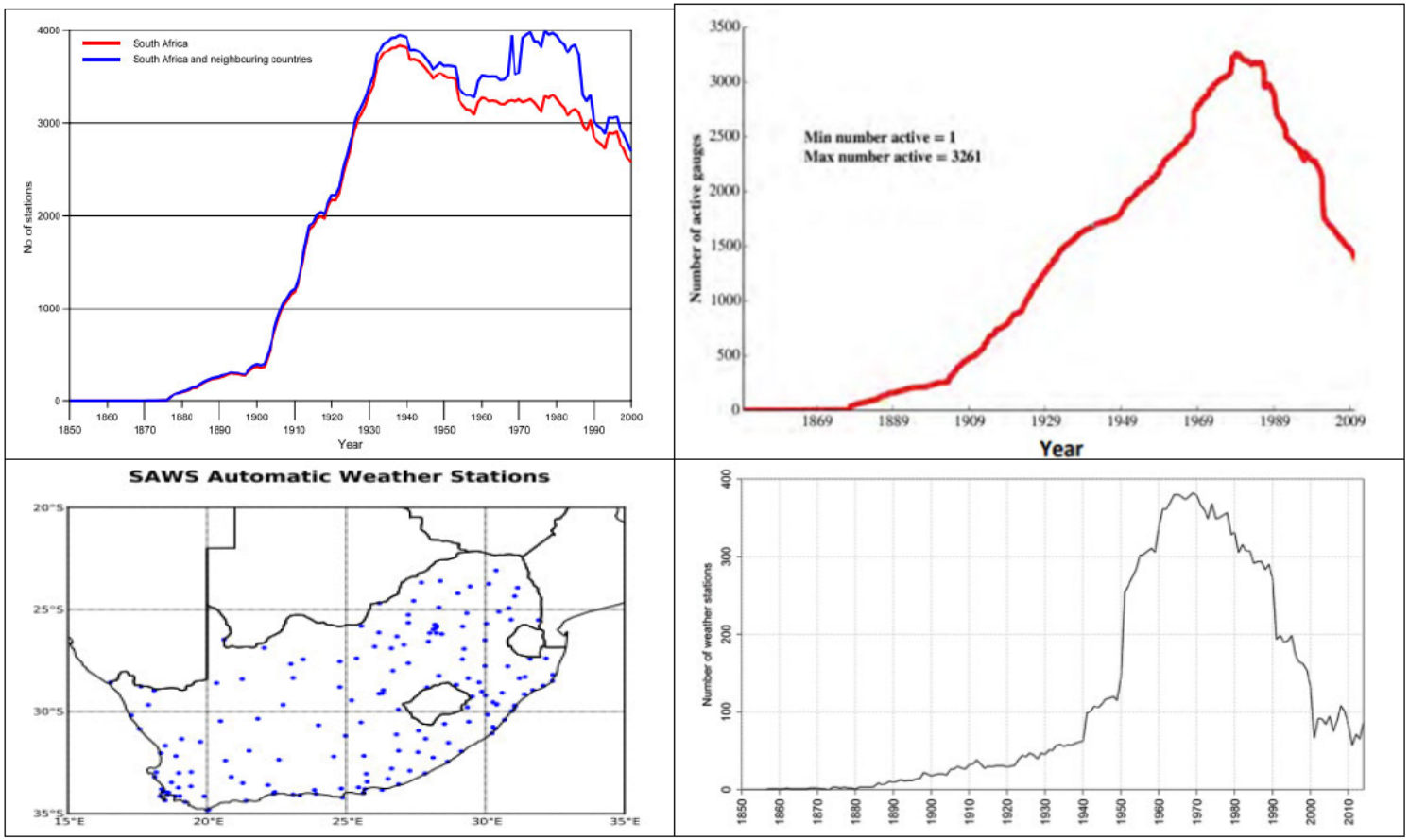


**Figure 1.1**: Decline in climate monitoring infrastructure in South Africa over time, highlighting the reduction in rainfall and temperature stations and the sparse spatial distribution of automatic weather stations (Lynch, 2004; Pegram et al., 2016; Davis-Reddy and Vincent, 2017)

## 1.4 Dissertation outline

This dissertation consists of 5 chapters, which are outlined as follows:

- **Chapter 1** provides a general introduction to the study, outlining the decline of reliable climate stations and the effect this has on the climate driver station that has been selected for each quaternary catchment.
- **Chapter 2** provides a literature review on climate data availability, rainfall and temperature data infilling methods as well as methods used to select a climate "driver" station.
- A detailed description of the methodologies used to achieve the aims and objectives is provided in **Chapter 3**.
- **Chapter 4** presents and discusses the results related to daily rainfall data extension, selection of representative driver stations and verification of methods for selecting a climate driver station.
- **Chapter 5** provides a synthesis of the approach followed, the main findings, including challenges and limitations. Recommendations for future studies are also included in this chapter.

# 2 LITERATURE REVIEW

## 2.1 Introduction

Hydrological models are essential tools for simulating the movement and storage of water within catchments, supporting applications such as water allocation, reservoir planning, flood forecasting, and climate change impact assessments. Widely used models in South Africa include *ACRU* (Schulze, 1995), *Pitman* (Midgley et al., 1994), and *SWAT* (Arnold et al., 1998), all of which rely heavily on accurate climate inputs, especially rainfall and temperature. Errors or gaps in these inputs can introduce significant uncertainty in model outputs (Clark et al., 2011). Given the ongoing decline in active climate stations across South Africa, improved methods for selecting representative driver stations and infilling missing data are increasingly critical to ensure reliable hydrological simulations for decision-making (Hrachowitz et al., 2013).

The accuracy of hydrological modelling in South Africa relies significantly on the quality and representativeness of climate data across diverse catchments (Suleman et al., 2020). Due to declining numbers of climate monitoring stations and data quality issues, selecting a climate "driver" station that is representative of a catchment has become more challenging (Dinku et al., 2011). This chapter reviews observed and interpolated climate datasets for rainfall, temperature and reference evaporation. A review of methods for selecting climate driver stations is also given. Additionally, this chapter describes utilities used to optimise station selection and discusses the implications of data quality on model outputs, thus providing context for the methodologies evaluated and developed in this study.

## 2.2 Observed climate datasets

### 2.2.1 Rainfall

Although rainfall is the most important input for hydrological and agricultural modelling, other climatic datasets (i.e. temperature and reference evaporation) are also required, and therefore must be obtained from existing weather stations. Each quaternary catchment therefore requires a climate driver station, where the station's data are deemed representative of the catchment's climatic conditions. At present, the hydrology of the 1 946 quaternary catchments is currently

driven by 1 240 rainfall stations selected from the rainfall database compiled by Lynch (2004). Hence, in some cases, the same driver station is selected for more than 1 quaternary catchment.

Lynch (2004) developed a daily rainfall database of > 300 million daily rainfall values for 12 153 stations. Daily and monthly rainfall datasets were first obtained from the former Computing Centre for Water Resources (CCWR) in early 2000, which housed datasets originally developed by Dent et al. (1989). The CCWR was hosted by the former University of Natal, which is now UKZN. In addition, daily and monthly rainfall datasets were collated from several organisations and private individuals in South Africa and neighbouring countries. These organisations included SAWS, the Agricultural Research Council (ARC) and the South African Sugar Association (SASA), as well as various municipalities and private individuals. Most of the rainfall was recorded at a daily timestep, but some stations had monthly data only (Lynch, 2004).

Different methods were utilised by Lynch (2004) to identify and correct suspect rainfall amounts. For example, one of the quality control steps involved flagging all rainfall values greater than 597 mm as suspect data. This represents South Africa's largest daily rainfall total, measured near St. Lucia Lake in January 1984 due to cyclone Domoina. Lynch (2004) stated that other errors in the rainfall database were highlighted when rainfall data were utilised in hydrological modelling exercises, where simulated streamflow was compared to measured streamflow data. A common error is due to the time rain gauges are read (i.e. 08h00 is the standard), but the rainfall total is incorrectly recorded on the day and not for the previous day. However, this type of “phase” error, which affects monthly and annual totals less, can be detected when daily rainfall values are compared to those from nearby stations.

### 2.2.2 Temperature

Schulze and Maharaj (2004) developed a database of 973 temperature stations with daily data from 1950 to 2000. Daily temperature datasets were obtained from 3 different custodians, *viz.*

- SAWS (which included stations from eSwatini and Lesotho),
- The Institute for Climate, Soil and Water (ISCW), and
- SASA.

The following checks were developed by Schulze and Maharaj (2004) and applied to the daily temperature data: (i) $T_{MAX} \leq T_{MIN}$, (ii) $T_{MAX} - T_{MIN} < 1.5°C$, and (iii) $T_{MAX} < 0°C$. Schulze and Maharaj (2004) also developed methods to infill missing temperature data, which are further explained in **Section 2.3.2**.

## 2.3 Infilling of missing data

Climate records with missing data limit their use for hydrological modelling as simulation models require a full dataset to function properly (Gao et al., 2018). Regardless of the technology used to take the recordings, whether manual measurements or the use of electronic sensors, some of the data stored will be either faulty or missing. The application of patching methods is dependent on the length of the gap. The gaps in the meteorological archives are caused mainly by absence of observers, vandalism, loss of records, data contamination, data-processing errors, effects of natural disasters like tornadoes or human-induced factors like wars, lack of funds for replacing broken instruments as well as instrument malfunctioning (Moeletsi et al., 2016). The climate data infilling process can also be time consuming and a difficult task (Pitman and Bailey, 2021). Some of the different rainfall and temperature infilling methods are explained next.

### 2.3.1 Rainfall

Lynch (2004) utilised various techniques to infill missing daily and monthly rainfall data, addressing gaps that commonly arise due to instrument failure, observer error or data telemetry issues. The following sub-sections outline the specific methods used to infill missing rainfall data. The rainfall infilling methods described next were applied to the dataset compiled by Lynch (2004) as outlined in **Section 2.2.1.** This dataset includes over 300 million values from more than 12 000 stations across South Africa and neighbouring countries.

#### 2.3.1.1 Expectation maximisation algorithm

The Expectation Maximisation Algorithm (EMA) was the primary method of infilling over 113 million rainfall values. This approach estimates a missing value by using available data to calculate the most probable value through an iterative process. Firstly, it estimates a missing value based on currently available data (Expectation step), then it updates the overall statistics of the dataset (Maximisation step). These steps repeat until the estimated values become stable, allowing EMA to infill missing data in a statistically reliable way.

#### 2.3.1.2 Inverse distance weighting

Inverse Distance Weighting (IDW) was used to infill 81 million missing values. This method estimates missing values by weighting nearby observations based on their distances, giving more weight to closer stations. IDW is useful in areas where rainfall patterns do not vary drastically over short distances.

#### 2.3.1.3 Median ratio

The Median Ratio (MR) method was used to infill over 40 million missing values. This method is effective in non-normal rainfall distributions, as it calculates ratios of median monthly rainfall to infill missing values, ensuring that infilled values align with the typical monthly profile.

#### 2.3.1.4 Monthly infilling

This technique was applied to ensure accurate representation of low-intensity rainfall events, which are still important for hydrological assessments. Specifically, a monthly infilling method was used to infill missing or minimal daily rainfall values (typically < 2 mm), thereby improving the completeness of the rainfall database across the full range of rainfall Intensities.

### 2.3.2 Temperature

Schulze and Maharaj (2004) utilised 2 methods to infill any missing daily maximum and minimum temperature values, which are described next. They found the second method to be more precise, which involves an algorithm to select 2 representative control stations.

#### 2.3.2.1 Mean temperature difference

For the Mean Temperature Difference (MTD) method, differences in means of daily maximum and minimum temperatures between the target station (station with missing data to be infilled) and the surrounding control stations (stations with no missing data to be used for infilling) are calculated for each month of the year. For each target station, the calculated monthly differences are then ranked from the lowest to the highest value to identify the most suitable control station for patching. This ranking process is based on the hypothesis that the control station with the smallest absolute difference is the most appropriate for infilling missing daily temperature data for the target station. The MTD method adjusts the target station's daily records by the smallest mean monthly difference. This adjustment assumes that the average monthly difference can be applied to daily values, providing a reasonable estimate of the target station's temperature. The MTD method implies that:

- Different control stations may be selected as most appropriate for patching in different months of the year.
- For a given month, distinct control stations may be chosen for maximum and minimum temperatures.
- Monthly mean differences can be applied to daily temperature values.

While conceptually simple, the MTD method does not account for the differences in deviation from mean values between stations, which may arise due to factors such as varying exposure. To address this limitation, Schulze and Maharaj (2004) developed another technique that is described next.

#### 2.3.2.2 Difference in standard deviation

The difference in standard deviation (DSD) method involves computing the standard deviation of the daily maximum and minimum temperatures for each month of the year, utilising data from the target station and the surrounding control stations. For each month, the difference in standard deviations between the target and controls are then ranked, identifying the most suitable control with the smallest difference. The best control station is selected for a given month and temperature variable (i.e. $T_{MAX}$ and $T_{MIN}$) to synthesise daily temperature values and the target's record is adjusted by the mean monthly temperature difference (Schulze, 2024; pers. comm.).

## 2.4 Interpolated climate datasets

The spatial datasets of rainfall, temperature and reference evaporation are essential for selecting an appropriate climate driver station for a particular catchment. Three gridded rainfall datasets are described next.

### 2.4.1 Rainfall

#### 2.4.1.1 Dent et al. (1989)

The Dent et al. (1989) study relied on numerous organisations and private individuals for rainfall data. The organisations included, *inter alia*, the following:

- the former South African Weather Bureau (SAWB; now known as SAWS); and
- the former Department of Agriculture and Water Supply, which has been split into 2 departments, namely the (i) Department of Agriculture, Forestry and Fisheries (DAFF), and the (ii) DWS; and
- the former Department of Environmental Affairs.

Other sources included the SASA, provincial park boards, municipalities and mines. Additional datasets were obtained by asking for rainfall records via announcements on the radio and in the press and by letters addressed to specific organisations.

Considering human errors might lead to inaccuracies or errors in rainfall data, the former SAWB developed quality control codes for daily rainfall data, as shown in **Table 2.1**. Dent et al. (1989) developed additional codes to flag monthly rainfall totals. Monthly and annual rainfall statistics (i.e. mean and median) used in mapping were calculated from reliable monthly totals, identified by quality control codes as either blank (representing reliable daily rainfall) or $ (indicating accumulated daily rainfall totals are considered reliable). Dent et al. (1989) interpolated spatial rainfall datasets from 9 409 stations.

There was a need to develop a primary set of physiographic data, which influences the spatial distribution of rainfall. This was done at a spatial grid resolution of 1 minute of a degree arc across southern Africa. Since this dataset was large (437 043 grid points of data), it was necessary to automate most of the data verifying, manipulating and management tasks. However, rigorous human error checking was also carried out. The country was manually divided into 34 rectangular regions (**Figure 2.1**), with a 15-arc minute overlap with neighbouring regions.

Dent et al. (1989) then used a multiple linear regression approach to generate gridded images of mean annual precipitation; **Figure 2.2**) and median monthly rainfall for South Africa. For each region, a regression model selection process was undertaken. Initially, potential models were identified using forward, backwards and stepwise regression techniques based on $R^2$ values. The final model was chosen using the Predicted Error of Sum Squares statistic, prioritising predictive accuracy over goodness-of-fit. This approach ensured that the selected model provided the best predictions for MAP rather than being the best-fit model.

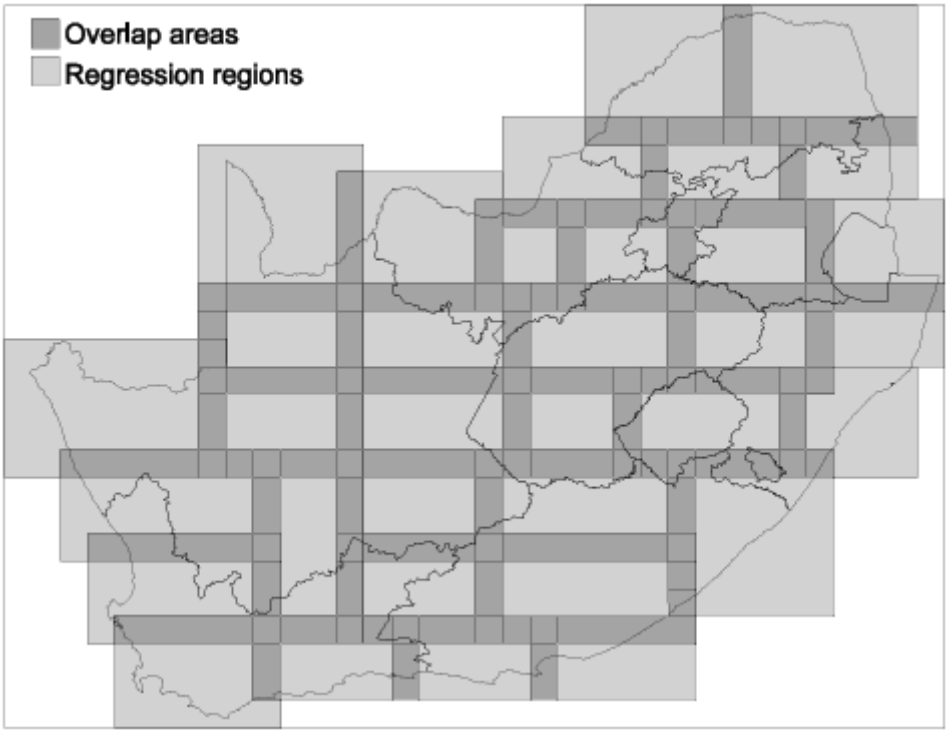


**Figure 2.1**: Regression regions developed by Dent et al. (1989) to estimate spatial MAP

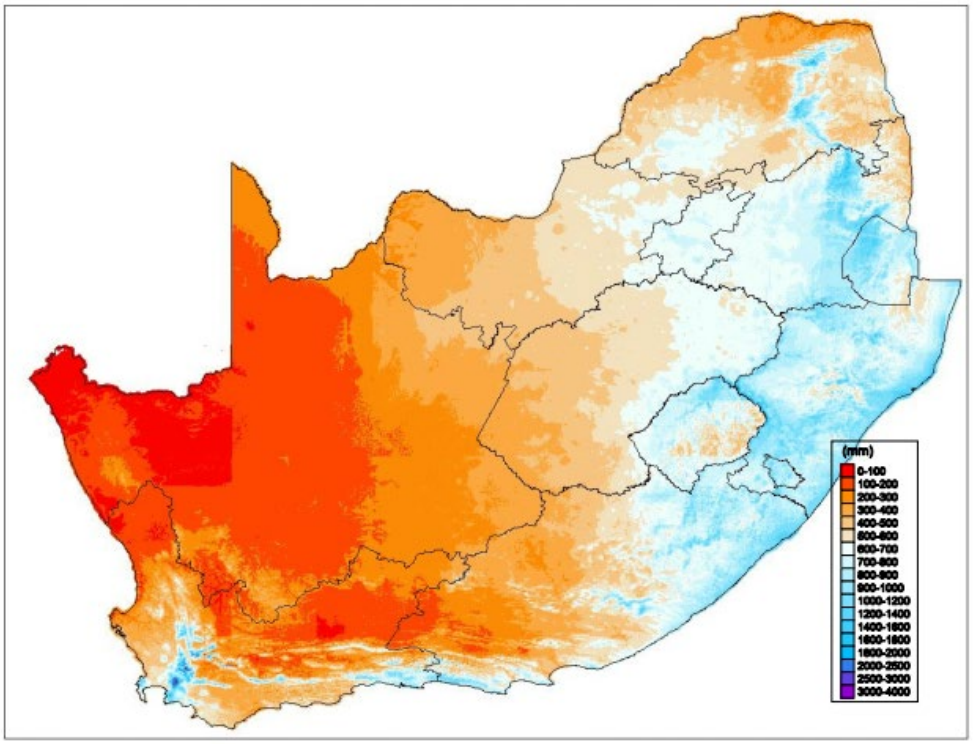


**Figure 2.2**: Map showing the mean annual precipitation surface determined by Dent et al. (1989) using multiple linear regression

**Table 2.1**: Monthly quality control codes developed for SAWB rainfall data (Dent et al., 1989)

| Monthly data code | Description of data quality | SAWB daily data code |
|---|---|---|
| blank | Reliable daily rainfall | No code |
| $ | Accumulated daily rainfall data (total reliable) | -6666<br>80000+<br>90000+<br>70000+ |
| * | Accumulated daily rainfall data (total unreliable) | 60000+<br>50000+<br>40000+ |
| # | Estimated daily rainfall data (not accumulated and not reliable) | 30000+ |
| M | Missing daily rainfall data | -9999 |

#### 2.4.1.2 Lynch (2004)

Dent et al. (1989) only generated median monthly precipitation due to time and computing power constraints. Lynch (2004) used a similar technique (i.e. multiple liner regression) (cf. **Section 2.4.1.1**) that expressed the median (and mean) monthly values as a ratio of the MAP values. These ratios were then interpolated onto a rectangular raster at a spatial resolution of one-by-one minute of a degree arc using the technique shown in **Figure 2.3**. This interpolated raster was then multiplied by the estimated MAP raster developed using Geographical Weighted Regression (GWR; **Figure 2.4**), which was repeated for each month. When compared to multiple regression, GWR is much faster in creating a MAP raster for a large region and also requires fewer explanatory variables. Spatial non-stationarity occurs when relationships between variables and their underlying processes change over space. GWR is an attempt to account for this spatial non-stationarity within a system by allowing the coefficients to fluctuate spatially.

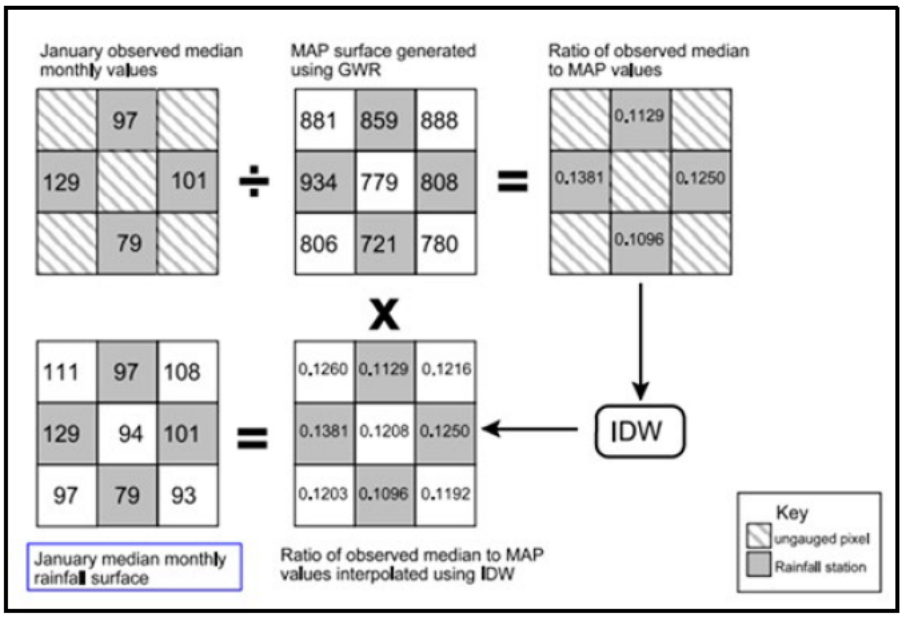


**Figure 2.3**: Procedure to calculate spatial datasets of interpolated mean and median monthly rainfall (Lynch, 2004)

Lynch (2004) used the following variables to generate reasonable estimates of MAP:

- **ialtCV**: coefficient of variation (CV) of altitude at a 5 arc minute spatial resolution;
- **latlong**: product of latitude and longitude coordinates (in degrees decimal) of the grid centroid;
- **xplusy**: sum of the latitude and longitude coordinates (in degrees decimal) of the grid centroid;
- **xx**: square of the longitude coordinates (degrees decimal) of the grid centroid; and
- **slope**: slope (degrees) of the 8 neighbouring grids surrounding each grid cell.

Lynch (2004) mentioned that the use of GWR to estimate monthly rainfall surfaces was not recommended because monthly rainfall is more variable than the annual rainfall. Furthermore, the process would be very time-consuming and computer intensive.

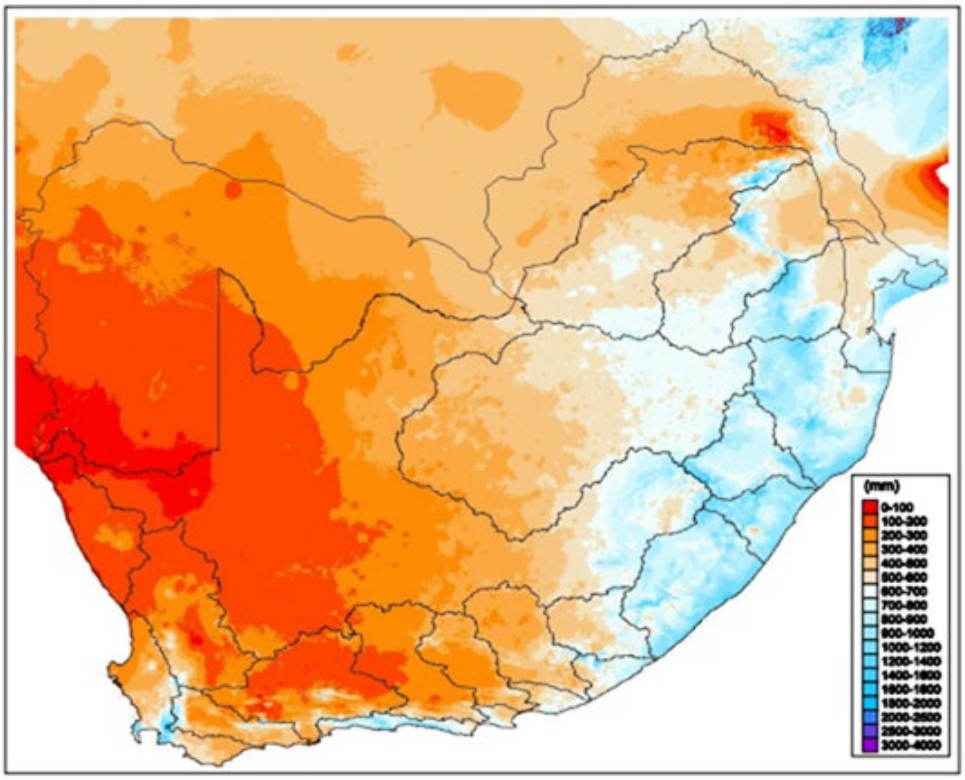


**Figure 2.4**: Map showing the mean annual precipitation surface determined by Lynch (2004) using geographically weighted regression

#### 2.4.1.3 Pegram et al. (2016)

Pegram et al. (2016) obtained daily rainfall datasets from the Climate Systems and Analysis Group based at the University of Cape Town. This dataset was developed during the Lennard et al. (2013) study and the authors quality-controlled the data to remove certain anomalies. The rainfall dataset integrates rainfall data from SAWS and ARC stations up to year 2000, sourced from Lynch (2004) and additional SAWS data from 2000 to 2010.

Pegram et al. (2016) proposed a robust methodology for creating spatially continuous rainfall surfaces using incomplete and scattered rainfall records in Southern Africa. Pegram's work addressed the limitations of traditional interpolation techniques by incorporating advanced statistical methods and handling the mixed nature of rainfall data.

Pegram et al. (2016) also developed updated MAP surfaces by interpolating the infilled annual rainfall data. The annual focus required different considerations due to the aggregated nature of annual totals, which smooth out short-term variations in precipitation but still reflect underlying spatial patterns driven by topography, proximity to the coast, and regional climate drivers. For MAP interpolation, the researchers employed similar techniques to those used for mean monthly rainfall including Gaussian copulas and quantile-quantile transforms. These methods ensured the reliable spatial representation of MAP, even in areas with sparse gauge coverage. The incorporation of uncertainty bounds for MAP estimates made the outputs particularly useful for hydrological modelling and water resource management applications.

To enhance the usability of MAP outputs, the results were summarised at the quaternary catchment scale, providing practitioners with ready-to-use data for regional-scale applications. The final MAP maps highlighted spatial rainfall variability, such as the pronounced west-to-east rainfall gradient and the influence of orography in mountainous regions. The variation in MAP between the Dent et al. (1989), Lynch (2004) and Pegram et al. (2016) surfaces were produced by Pegram et al. (2016). In comparison of the different interpolated MAP developed by Dent et al. (1989; **Figure 2.2**), Lynch (2004; **Figure 2.4**) and Pegram et al. (2016; **Figure 2.5**) a difference was noted in the northeast region of South Africa, where the Pegram et al. (2016) map shows a drier zone compared to the other 2 rainfall maps. Another difference is in the eastern part of eSwatini as there is high rainfall above the escarpment, which the Lynch (2004) map does not show.

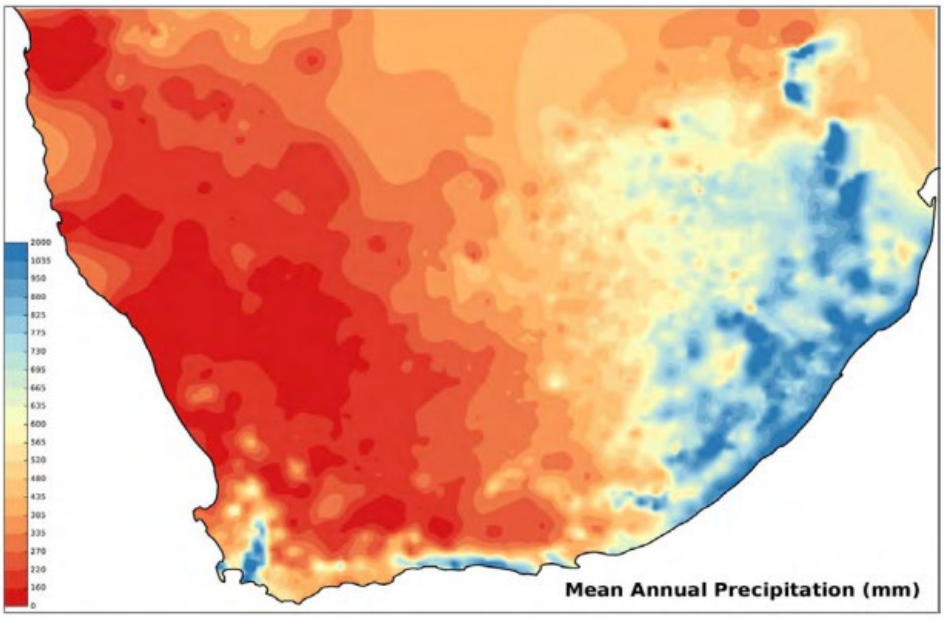


**Figure 2.5**: The gridded MAP, bi-linearly interpolated into a finer 1' by 1' grid (Pegram et al., 2016)

To generate spatially continuous mean monthly rainfall maps, Pegram et al. (2016) used advanced statistical interpolation techniques on infilled monthly rainfall data. The infilling process accounted for missing records at individual gauges, using Dynamic Copula Regression to address the mixed discrete-continuous nature of rainfall and provide confidence intervals for estimates. The monthly-scale focus leveraged the relatively strong spatial correlation of rainfall at this temporal resolution, ensuring more accurate reconstructions of missing data.

The interpolated mean monthly rainfall surfaces were derived by integrating the repaired gauge data with external factors such as topography and proximity to the coast. Techniques like Gaussian copulas and quantile-quantile transforms were used to account for the non-stationary and asymmetrical correlation structures typical of rainfall patterns. The resulting mean monthly rainfall maps included variability estimates, offering critical insights into monthly rainfall distributions across Southern Africa's diverse climatic regions.

### 2.4.2 Temperature

#### 2.4.2.1 Schulze and Maharaj (2004)

As noted in **Section 2.2.2**, Schulze and Maharaj (2004) developed a database of 973 temperature stations, each with 51 years of daily $T_{MAX}$ and $T_{MIN}$ values. This database was then used to derive interpolated temperature data for each 1' by 1' of a degree arc across southern Africa (437 043 grid points). For each grid point, 2 nearby temperature stations were selected using a simple algorithm. The algorithm is based on 2 factors representing the elevation difference and the distance between the temperature station and point of interest (e.g. quaternary catchment). The algorithm selected 2 representative stations with the highest weightings. The algorithm to select 2 representative stations for estimating temperature data at a particular grid point was modified by Kunz et al. (2015). Kunz et al. (2020) improved the algorithm by assigning more weighing to the altitude difference than distance. These stations were selected from different quadrants around the grid point to reduce the risk of bias from any single station.

For each day, maximum ($T_{MAX}$) and minimum ($T_{MIN}$) temperature values interpolated from the 2 selected stations were then averaged. This approach smoothed out any localised biases, resulting in temperature values that were more representative of the grid point conditions. For each temperature record from the 2 selected stations, adjustments were made according to these monthly LRs.

#### 2.4.2.2 Adiabatic lapse rates

Schulze (1995) published monthly Lapse Rates (LR) for maximum and minimum temperatures for 12 regions in South Africa, which were later revised by Schulze and Maharaj (2004), as seen in **Figure 2.6**. These LRs are essential to adjust temperature data from a station located at a different elevation relative to the target location (i.e. grid centroid). A lapse rate is a known decrease in temperature with increasing altitude. An average lapse of -6.5°C per 1000 m is commonly applied in climatological and meteorological studies (Dutra et al., 2020). In 1965, former SAWB identified distinct summer and winter LRs across 6 regions. These included an interior design, bordered by 5 coastal regions on the seaward side of the Great Escarpment (Schulze and Maharaj, 2004).

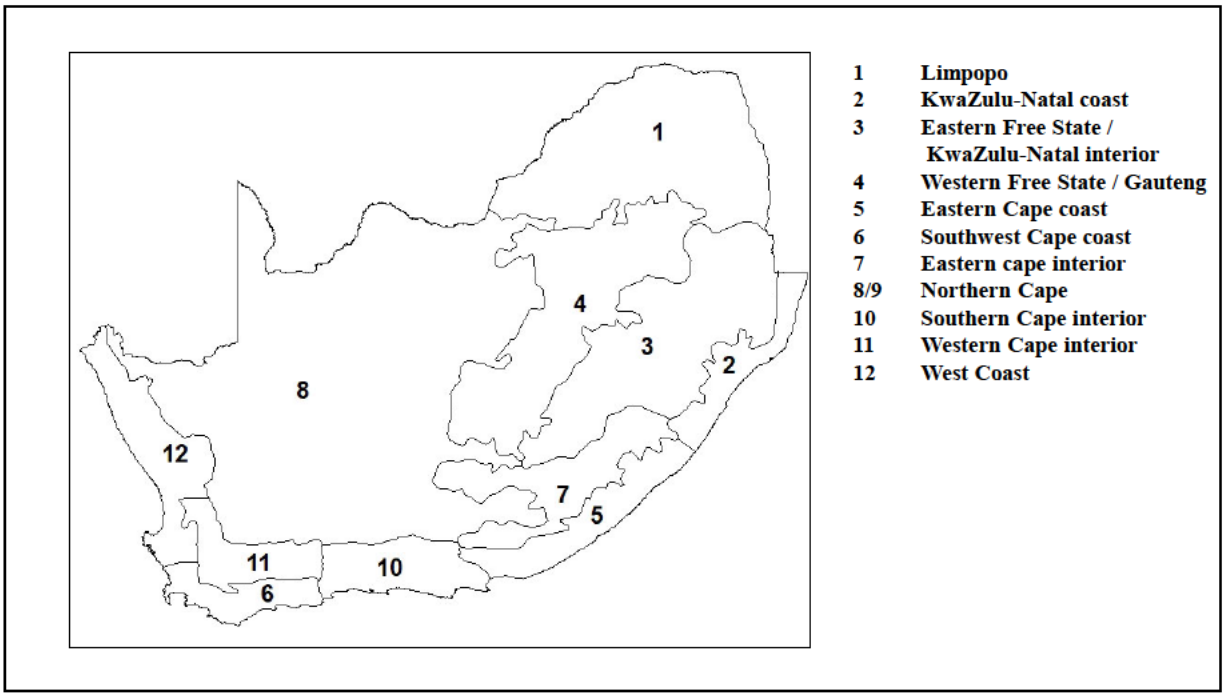


**Figure 2.6**: Map showing the modified lapse rate regions developed by Schulze (1995)

The regions were then modified in 2004, where regions 6 and 11 were swopped and regions 8 and 9 were joined to better reflect climatic and topographical similarities observed during further analysis (Schulze and Maharaj, 2004). To generate new lapse rates, Schulze and Maharaj (2004) determined the initially computed lapse rates for regions and assigned confidence ratings to these LRs. Each lapse rate plot was subjectively assigned a confidence rating as follows:

- **High**: tight and distinctly linear association between temperature and elevation. Therefore, no other adjustments/revisions to LRs were deemed necessary.
- **Medium**: a discernible linear relationship between temperature and elevation, but the relationship was more diffuse, frequently due to different plots of those stations with high and low Topographic Index, which would then result in a re-evaluation of the LRs.
- **Low**: no linear pattern evident in the scatter plot, necessitating a subsequent revision of lapse rates.

### 2.4.3 Reference evaporation

Evaporation has a significant influence on the yield of water supply reservoirs and on the economics of building reservoirs of different sizes (van Dijk and van Vuuren, 2015). According to Moeletsi et al. (2013), precise estimation of reference evapotranspiration ($ET_O$) holds significant importance in modelling water use, agricultural and ecological applications, as well as in natural resource management and various planning endeavours. The United Nations (UN) Food and Agriculture Organisation (FAO) advocates for the adoption of the Penman-Monteith equation to estimate $ET_O$ from meteorological data. $ET_O$ represents the maximum evapotranspiration (ET) from a hypothetical grass surface with specific characteristics, including height, surface resistance and albedo (Allen et al., 1998). In South Africa, there are significant evaporation losses from reservoirs. **Figure 2.7** shows that most of the country's evaporation rate is greater than 1 400 mm per annum

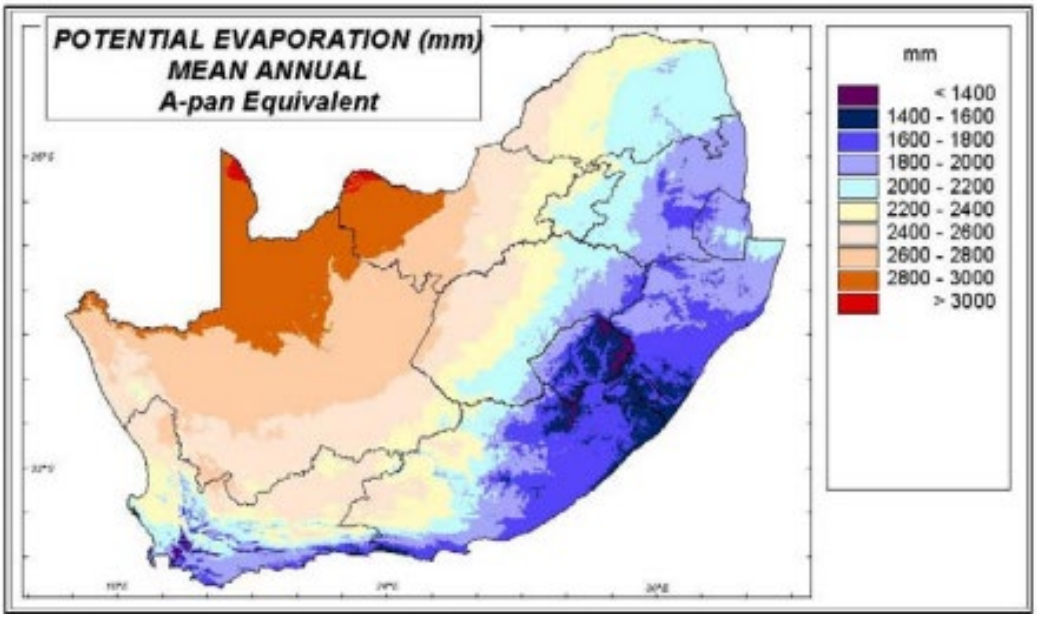


**Figure 2.7**: Map of mean annual potential evaporation (mm) across South Africa developed using the Penman-Monteith equation (Schulze et al., 2007b)

Despite its accuracy, the Penman-Monteith equation's dependence on multiple meteorological variables limits its applicability in data-scarce regions, including many parts of South Africa and other African countries (Moeletsi et al., 2013). In these areas, measurements are often available only for rainfall and temperature (minimum and maximum), while data for wind speed, vapour pressure (i.e. relative humidity) and radiation are scarce.

A constant wind speed value of 2 m $s^{-1}$, as recommended by Allen et al. (1998), has commonly been used to estimate $ET_O$ across South Africa (Kunz et al., 2020). Alternatively, Schulze et al. (2007b) suggested a wind speed of 1.6 m $s^{-1}$ for cases where measurements are unavailable. However, **Figure 2.8** highlights the value of the Wind Atlas for South Africa (WASA), which provides a detailed raster dataset for estimating wind speed across different regions. As the WASA wind speeds are provided at a height of 100 m above ground level, they would need to be adjusted to the standard 2 m height using a logarithmic wind profile equation before being used in $ET_O$ calculations. This offers a more accurate representation compared to relying on constant wind speed values.

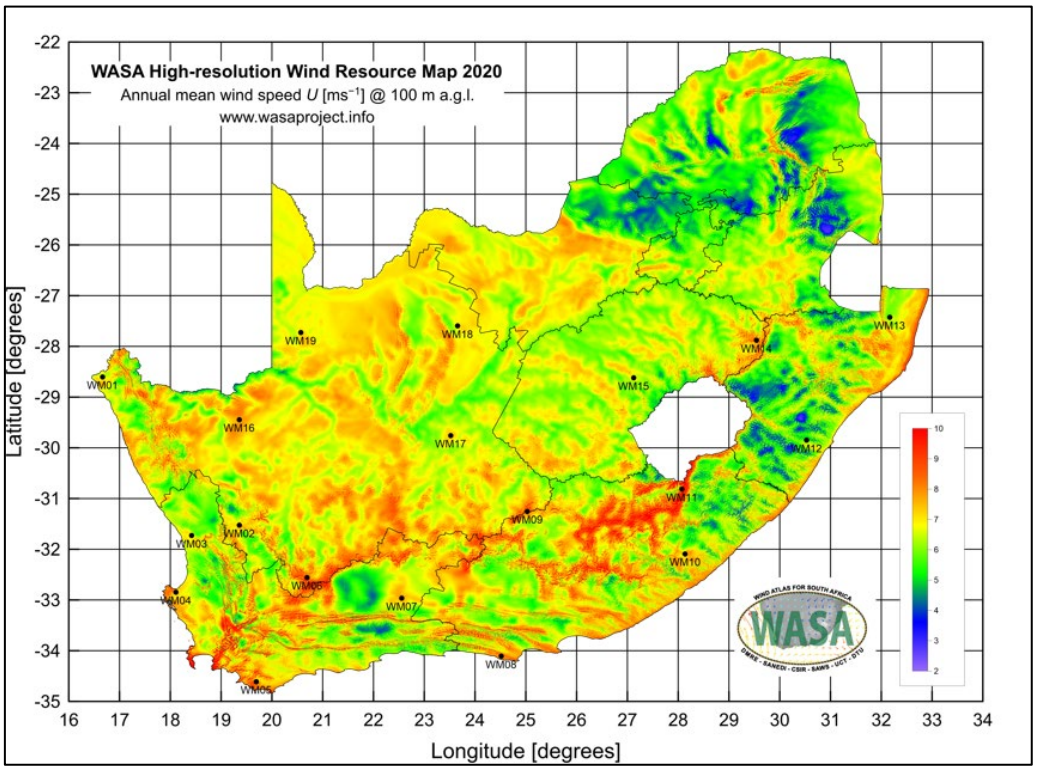


**Figure 2.8**: Map showing the South African annual mean wind speed atlas (WASA, 2020)

In regions where meteorological data are sparse, simpler, less data-intensive methods for estimating $ET_O$, such as the Hargreaves and Samani (1985) equation, can be employed. This method requires only long-term records of temperature and provides a practical alternative to the Penman-Monteith equation in data-limited environments (Moeletsi et al., 2013).

## 2.5 Useful utilities

According to Smithers et al. (2004), some of the time-consuming tasks in configuring hydrological models include, *inter alia*:

- the selection of the most suitable daily rainfall stations to "drive" the model, and
- the calculation of adjustment factors to ensure that the daily data from the selected point rainfall station (in or adjacent to the catchment) represents the catchment's areal rainfall.

Hence, 3 utilities were developed to assist users in selecting a rainfall driver station and computing rainfall adjustment factors.

### 2.5.1 *CalcPPTCor*

Pike (2004) developed a software utility called *CalcPPTCor* that helps the user to select driver rainfall stations automatically. It is also useful for estimating rainfall adjustment factors, defined as a ratio of catchment MAP (spatially averaged) to rainfall station MAP. *CalcPPTCor* requires spatially averaged mean or median monthly rainfall values for the catchment. The average statistic is affected by significant events (i.e. outliers) such as floods and droughts. Since the median statistic is less influenced by extreme events, median monthly rainfall values are deemed more appropriate for calculating adjustment factors (Schulze, 2010; Kunz et al., 2020). To ensure that topographical and climate influences are minimised and that the resulting data fairly represent the catchment's rainfall, the calculated monthly ratios are applied to each day's rainfall within that month.

### 2.5.2 *Grid Extractor*

The *Grid Extractor* utility can be used to determine the catchment's mean elevation, MAP and mean/median monthly rainfall totals. The ratio of the catchment's gridded median monthly rainfall to the observed median monthly rainfall of the station is then computed using *CalcPPTCor* (Smithers and Schulze, 2004).

With the aid of a Geographic Information System (GIS), each station's geographic coordinate can be used to obtain the station's elevation, MAP and mean/median monthly rainfall values using the raster datasets that are currently available. These values can then be checked against the station's elevation provided by the custodian and the station's MAP and mean/median monthly rainfall totals calculated from observed data.

### 2.5.3 *Daily Rainfall Extraction Utility*

The *Daily Rainfall Extraction* (*DRE*) utility is a software utility developed by Kunz (2004), which extracts daily, monthly or annual rainfall data from Lynch (2004) rainfall database (cf. **Section 2.4.1.2**). The utility requires the coordinate of interest for which rainfall data are required, i.e. the centroid of a catchment. The utility then provides a list of neighbouring stations that have data for the required period. The *DRE* utility also provides valuable statistics for the selected station(s), such as the start and end date of the record, record length, station MAP and the relative fraction of reliable, patched and missing data in the record.

The *DRE* utility derives the best station from the neighbouring 10 stations using a set of 10 criteria, each with a specific equation and weighting:

1. **Distance**: stations closer to the user's point of interest are ranked higher.
2. **Operational status**: stations that have data in 2000/01 (i.e. considered operational) are given higher scores.
3. **Start year comparison**: stations whose data start earlier than the required start year are scored higher.
4. **End-year comparison**: stations whose data end later than the required end year are favoured.
5. **Station record length**: stations with longer complete historical records are preferred.
6. **Estimated MAP comparison**: Stations with **gridded MAP values** (derived from the raster database) that are closer to the gridded MAP value for the point of interest receive higher scores. This comparison focuses on the consistency of gridded estimates between the station's location and the point of interest.

7. **Observed MAP comparison**: Stations with **observed MAP values** (calculated from historical rainfall records) that are similar to the gridded MAP value for the point of interest receive higher scores. This comparison evaluates how well the station's actual rainfall data aligns with the estimated MAP for the point.
8. **Reliable record portion**: a higher proportion of reliable (observed) data results in a better score.
9. **Patched record portion**: fewer patched daily values improve the station's suitability.
10. **Missing record portion**: a lower percentage of missing data increases the station's score. Hence, if more than 50 % of the data are missing, the weighting of this criterion decreases (Kunz, 2004).

Ultimately, the station with the highest cumulative score across all criteria is deemed the most representative, while others are ranked accordingly. This approach ensures the selection of the most suitable “driver” station for the catchment based on both spatial proximity and data integrity.

Although the utility can rank up to 20 neighbouring rainfall stations, it only analyses data from the 10 nearest stations. This limitation may lead to the exclusion of other potentially more suitable “driver” stations located within or near the quaternary catchment. As a result, stations with more reliable or extensive records that could better represent rainfall dynamics within the catchment might be overlooked. Furthermore, the methodology does not incorporate a “human experience” factor, which could leverage local knowledge or expert judgment to enhance the selection process. This absence means that important contextual insights, such as historical weather anomalies or known data collection inconsistencies, are not considered. While some topographical influences could be accounted for by comparing the MAP of the station to that of the catchment, additional environmental and observational factors may further refine the ranking process. To integrate experiential knowledge, the *DRE* utility could be modified to include a user-input mechanism, allowing experts to provide insights on station reliability, known anomalies, or other qualitative factors that may improve selection accuracy.

## 2.6 Driver station selection methods

### 2.6.1 Rainfall

Estimation of daily rainfall for a catchment using the DS approach is done as follows:

1. Station filtering: Various criteria are used to select suitable driver stations. For example, the station's location, start and end year of record, percentages of reliable and missing data and the MAP statistic are commonly used criteria (Smithers and Schulze, 1995).
2. Driver station selection: From the filtered list of suitable stations, other criteria are then used to select the best station, which includes, *inter alia*, distance from the station to the catchment centroid and the station's elevation in relation to the average elevation of the catchment. In addition, the station's MAP is compared to the catchment's spatially averaged MAP. The driver station and catchment's mean or median monthly rainfall values can also be compared (Schulze, 1995). For example, a catchment's median monthly precipitation can be calculated using the 1' by 1' grid of median monthly precipitation developed for southern Africa by Dent et al. (1989).

The above approach was followed to select a suitable driver station for each quaternary and quinary catchments. More detail is provided in the 2 sub-sections below.

#### 2.6.1.1 Quaternary catchments

To select a driver station for each of the 1 946 quaternary catchments, Warburton (2005) used a GIS to determine each catchment's centroid. For each catchment's centroid, the *DRE* utility (cf. **Section 2.5.3**) was then used to determine the 10 best rainfall stations which have reliable data from January 1950 to December 1999. The stations were then ranked from "best" to "worst", with the "best" station selected as the driver for each quaternary catchment.

The highest-ranked station identified by the *DRE* utility was not automatically selected as the driver station for the quaternary catchment. Instead, the selection process was subject to additional manual evaluation. For example, the process considered the experience of the research team working at the former School of Bioresources Engineering and Environmental Hydrology (Warburton, 2005).

This approach ensured that local environmental factors and expert judgment were considered alongside the automated ranking process. In addition, if the suitability of the "best" station was poor, the catchment was flagged and the station selection process was re-visited.

The *DRE* utility was then re-run and 20 rainfall stations (rather than 10) were extracted, from which the most representative rainfall station was selected. A GIS was also used to assist with checking the selected driver station, together with Lynch's (2004) MAP surface, a 200 m Digital Elevation Model (DEM; Schulze and Horan, 2011), the Quaternary catchment boundaries (DWS, 2019) and the rainfall station locations (Warburton, 2005).

#### 2.6.1.2 Quinary catchments

In the study by Schulze et al. (2010), rainfall driver stations were selected for each of the 5 838 quinary catchments by assuming that the driver station assigned to the parent quaternary catchment would also represent its three internal quinaries. Further investigation identified the need to improve the representative rainfall for 11 quaternary catchments, leading to changes in the driver stations selected for these catchments. Consequently, the total number of driver stations was reduced from 1 244 to 1 240.

These 1 240 stations were then utilised to generate daily rainfall data for the 5 838 quinary catchments. To enhance the accuracy of the rainfall data for each quinary catchment, multiplicative adjustment factors specific to each catchment were applied to the station's rainfall data. These adjustments ensured that the rainfall measurements were more representative of the unique characteristics of each quinary catchment. The monthly adjustment factors were derived by first calculating the spatial averages of all 1 arc minute (~1.7 by 1.7 km) gridded median rainfall values within a quinary catchment. The catchment averages of median monthly rainfall were then divided by the driver station's median monthly rainfall to calculate the 12 adjustment factors (Kunz et al., 2015).

### 2.6.2 Temperature

#### 2.6.2.1 Quaternary catchments

The gridded daily temperature dataset developed by Schulze and Maharaj (2004) was utilised in generating the QCs temperature dataset. This database facilitated the production of a 50-year historical series (1950-1999) at a spatial resolution of approximately 1.7 by 1.7 km, which was interpolated to account for regional variations in temperature. A 200 m DEM was used to calculate the average elevation for each QC. The elevation was then applied to adjust temperature values, using LRs to account for elevation differences. Daily temperature values were generated at the QC's centroid, creating spatially representative data specific to each catchment's elevation and location. Where gaps existed in temperature station data, 2 infilling techniques (i.e. MTD method and DSD method) were used to infill the missing data. This approach provided each QC with a complete, interpolated temperature record (Schulze et al., 2007a).

#### 2.6.2.2 Quinary catchments

The determination of temperature data for each quinary is explained below. However, different versions to determine quinary catchment temperature data exist, which are explained next.

Version 1 (Schulze et al., 2010): A procedure was developed to select a representative temperature dataset for each quinary using a temperature dataset developed by Schulze and Maharaj (2004). The mean elevation of each quinary and its centroid were used to find a representative grid point that was (i) of similar elevation (to the quinary mean) and (ii) closest to the quinary centroid. The mean elevation of each quinary was calculated from the 200 m DEM.

Version 2: Lumsden et al. (2011): The algorithm originally developed by Schulze and Maharaj (2004) that selects two2 stations to generate interpolated temperature data for any location (e.g. grid or catchment centroid) was revised to give more weighting to the altitude difference between each temperature station and quinary catchment's average value.

The algorithm identifies the 5 best temperature stations for a point of interest (centroid of each quinary catchment) by performing a preliminary suitability ranking of all neighbouring stations. The 2 factors used to rank the stations were as follows:

$$DSTF = 0.9 \cdot \left(1 - \frac{DIST}{350}\right) + 0.1 \quad \textbf{Equation 2.1}$$

$$AF = 0.9 \cdot \left(1 - \frac{DALT}{1500}\right) + 0.1 \quad \textbf{Equation 2.2}$$

Where *DSTF* is the distance factor and *DIST* is the distance between the station and quinary centroid (in minutes of a degree; relative to a maximum value of 350 minutes). Similarly, *AF* is the altitude factor and *DALT* is the altitude difference between the station and quinary's mean altitude (m; relative to a maximum value of 1500 m). The ranking factor (*RF*) for each station was calculated as follows:

$$RF = 10 \cdot DF + AF \quad \textbf{Equation 2.3}$$

As mentioned earlier, *DSTF* was given a higher weighting (10) compared to the *AF* (1). *DF* ranges from 0.1 (worst case, where the station is 350 degrees minutes or more away) to 1 (best case). Similarly, *AF* is 0.1 when the station's elevation differs by 1 500 m or more from the catchment's mean elevation.

The 5 stations with the highest *RF* values were initially chosen. Thereafter, the "best" 2 stations were determined by ranking these 5 stations. To accomplish this, the range of distances (relative to the quinary catchment centroid) and elevation differences (relative to the catchment mean elevation) among the 5 stations were factored into the *DF* and *AF* calculations as follows:

$$DF = 0.9 \cdot \left(1 - \left(\frac{DIST - MIND}{MAXD - MIND}\right)\right) + 0.1 \quad \textbf{Equation 2.4}$$

$$AF = 0.9 \cdot \left(1 - \left(\frac{DALT - MINA}{MAXA - MINA}\right)\right) + 0.1 \quad \textbf{Equation 2.5}$$

Where *MIND* is the distance between the closest station and the quinary catchment centroid (m) and *MAXD* is the distance between the most distant station and quinary catchment centroid (m). For the station closest to the catchment's centroid, *DF* becomes 1.0 (best). Similarly, *AF* is the calculated factor, *MINA* is the altitude difference between the station with a similar altitude to the quinary catchment mean altitude and *MAXA* is the largest altitude difference between the worst station and the quinary catchment mean altitude (m).

For stations at elevation similar to the catchment's mean altitude, *AF* approaches 1.0 (best). Again, all stations are ranked relative to the range (*MAXD* − *MIND*) and (*MAXA* − *MINA*). These equations compare each station to the "best" case; thus, this "relative" ranking technique outperforms other approaches used in the past. Since the initial ranking would exclude stations that were unsuitable from a distance perspective, more weighting was then given to *AF* in the final calculation of *RF* as follows:

$$RF = 10 \cdot DF + 3 \cdot AF \qquad \textbf{Equation 2.6}$$

Following the selection of the 2 "best" temperature stations to represent the catchment (i.e. the 2 highest *RF* values), daily data from these 2 stations was used to produce the catchment's final temperature record. The 2 *RF* factors were used to weight each station's data, i.e. *WF1 = RF1/(RF1 + RF2)* and *WF2 = 1 – WF1*. Adiabatic lapse rates were also applied to each station's data to account for *DALT*, which is further explained in the next section.

Version 3 (Kunz et al., 2020): The algorithm described by Lumsden et al. (2011) to select 2 representative stations for estimating temperature data for a catchment was modified once again to select a "pseudo" temperature station for each quaternary catchment's driver rainfall station. The distance between the driver rain gauge and each surrounding temperature station was calculated (*DIST*; in minutes of a degree) along with the altitude difference (*DALT*).

$$DIST = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad \textbf{Equation 2.7}$$

$$DALT = target\ station\ altitude - control\ station\ altitude \qquad \textbf{Equation 2.8}$$

Where *DIST* is the distance between target (i.e. rainfall) station to each of the neighbouring control (i.e. temperature) stations (minutes of a degree), *x is the* longitude and y is the latitude and *DALT* is the altitude difference between target and control station (m).

Temperature stations over 200 minutes from the rain gauge were assigned the same distance factor (*DF*) of 0.1 (unsuitable). A temperature station at the exact location of the rain gauge was assigned a *DF* of 1.0 (ideal):

$$DF = 0.9 \cdot \left(1 - \frac{DIST}{200}\right) + 0.1 \qquad \textbf{Equation 2.9}$$

The 200 minutes distance criterion was calculated using a "trial and error" approach until the algorithm provided the same station selection as was manually chosen. Setting a higher distance threshold (e.g. 250 minutes of a degree) placed more emphasis on the altitude difference (Kunz et al., 2020). The *AF* was calculated using **Equation 2.2**.

The 5 "best" stations were initially determined using the same methodology as Lumsden et al. (2011). The stations were then re-ranked relative to the station furthest away and the one with the greatest altitude difference. The "worst" of the 5 stations exhibited the lowest ranking. Every rain gauge location was then assigned the highest ranked (i.e. "best") temperature station. Of the 973 temperature stations (cf. **Section 2.2.2**), 543 stations were selected to represent the 1 240 driver rain gauges. There were 114 temperature stations with the same SAWS station ID as the rain gauges, indicating their close proximity. Therefore, these temperature stations were considered to be the "perfect match" for the rain gauge (Kunz et al., 2020).

The altitude difference between the temperature station and the average value for each was quinary was used to apply an adiabatic lapse rate adjustment to generate "unique" temperature values for each quinary (cf. **Figure 2.6**). As noted by Kunz et al. (2020), the 90 m DEM (Weepener et al., 2012) was used to update the mean elevation for each quinary, which was an improvement when compared to the 200 m DEM used by Schulze and Horan (2011).

## 2.7 Hydrological modelling in data-scarce regions

Many parts of the Global South face similar challenges to South Africa, including declining station networks and incomplete climate records. These limitations have prompted the development and application of alternative approaches to support catchment-scale hydrological modelling in data-scarce environments.

In recent years, Earth Observation (EO) rainfall products (e.g. CHIRPS; cf. **Section 2.7.1**) and reanalysis datasets have gained traction as alternative or supplementary sources to ground-based station data in hydrological modelling. These satellite-derived datasets offer consistent spatial coverage, making them particularly valuable in data-scarce or inaccessible regions (Huffman et al., 2020). However, their accuracy varies with factors such as topography, rainfall intensity, and the degree of calibration with in situ observations (Toté et al., 2015). While EO data can improve climate input quality in poorly monitored regions, they may underestimate extreme events and exhibit biases linked to satellite sensor limitations or spatial averaging effects. These limitations highlight the need for cautious application and validation before integration into operational hydrological models.

### 2.7.1 Remotely sensed rainfall products

The Climate Hazards Group InfraRed Precipitation with Station Data (CHIRPS) is a quasi-global rainfall dataset developed by the Climate Hazards Center at the University of California, Santa Barbara, USA. It spans the period 1981 to near-present and provides gridded precipitation estimates at a spatial resolution of 0.05° (~5 km) (Funk et al., 2015). CHIRPS integrates satellite imagery with in-situ station data to produce high-resolution precipitation estimates suitable for climate monitoring and hydrological applications, particularly in data-sparse regions (Funk et al., 2015). In sub-Saharan Africa, CHIRPS has been widely used to assess drought, agricultural planning, and rainfall variability due to its relatively long temporal coverage and availability at daily and monthly time steps (Dinku et al., 2019). However, despite its strengths, CHIRPS may underestimate extreme precipitation events and can be less accurate in areas with dense cloud cover or limited station data required for calibration (Toté et al., 2015).

While CHIRPS was not a primary dataset used in this study, its relevance as a globally accessible alternative to local datasets such as those described in **Section 2.4.1** is acknowledged. Future work should consider directly comparing CHIRPS with the interpolated products to assess its utility in South African hydrological contexts. However, satellite data requires bias correction against measured rainfall, as satellite data was not as reliable (Siddig et al., 2022).

Dinku et al. (2011), validated satellite rainfall products over Ethiopia and found CHIRPS to perform well when adjusted with ground observations. Gebrechorkos et al. (2018) similarly recommended blending satellite and gauge data to enhance rainfall estimates in East Africa. In addition, hybrid and multi-source methods have been developed that combine observed station data with satellite or interpolated grids. For example, Masih et al. (2010) used a combination of CHIRPS and ground station data to simulate streamflow in the Upper Indus Basin, while Wang et al. (2016) applied merged gauge-satellite datasets to improve spatial rainfall representation in the Andes and Mekong River Basins. These approaches collectively offer flexible solutions for hydrological modelling in regions with limited in-situ data.

### 2.7.2 Model calibration strategies

In the absence of high-quality streamflow data, ensemble modelling or indirect calibration approaches (e.g. using remote sensing-based evapotranspiration estimates or soil moisture) have been explored (Beck et al., 2017). These approaches account for uncertainty and allow for more robust model development in data-poor areas.

Together, these international experiences highlight the need for flexible, transparent, and reproducible methods that can overcome data limitations. This study builds on those insights by refining rainfall and temperature driver station selection for hydrological modelling in South Africa and demonstrating how model performance can be verified despite data scarcity.

## 2.8 Summary and conclusions

The reviewed literature underscores the critical importance of selecting representative climate driver stations for hydrological modelling and water resources management. Climate driver stations provide essential data inputs (e.g. rainfall, temperature and reference evaporation), which directly influence the accuracy of hydrological simulations. However, several challenges hinder the effective selection of appropriate stations. One of the main constraints is the decline in operational rainfall and temperature stations in South Africa, resulting in reduced spatial coverage. There is also a notable increase in gaps in climate data (i.e. missing data). As a result, few stations have complete datasets, necessitating infilling methods like IDW for rainfall and DSD for temperature. Although useful, these methods are less accurate in sparsely monitored regions, raising concerns about the reliability of infilled data.

**Table 2.2** compares gridded rainfall datasets by Dent et al. (1989), Lynch (2004), and Pegram et al. (2016), highlighting differences in methodology, coverage and record length. These datasets are critical for estimating rainfall adjustment factors, especially when using median values to reduce the influence of extreme events. Their evolution underscores the need for up-to-date datasets when selecting rainfall driver stations. Less attention has been given to temperature driver station selection, with existing algorithms needing refinement due to declining data availability. To verify the accuracy of selected climate driver stations, the study utilised a hydrological model to simulate inflows into a large dam. The simulated inflows were then compared to dam water balance data to evaluate the reliability of the different station selection methods.

**Table 2.2**: Summary of key gridded rainfall datasets used for climate driver station selection

| Study | Method | Number of stations | Last date of rainfall record |
|---|---|---|---|
| Dent et al. (1989) | Gridded rainfall datasets using multiple regression equations | 8 281 | Mid 1980s |
| Lynch (2004) | Raster datasets using GWR to improve spatial variability capture | 13 251 | 2000 |
| Pegram et al. (2016) | Combined datasets using Gaussian copulas for interpolation and infilling | ~1 000 | 2010 |

# 3 STUDY METHODOLOGY

This chapter details the methodology employed to accomplish the aims and objectives of this study. A comprehensive description of the study catchment, as well as climate data sourcing and simulation processes, are included in the methodology.

## 3.1 Study catchment selection

To ensure the accuracy and reliability of this study, specific criteria were established for selecting an appropriate study catchment. For example, the catchment should have a high density of climate stations, preferably with no more than 90 days (~1 %) of missing data in each dataset. Neighbouring stations with up to 10 % of missing record were used for patching the selected stations. The catchment also needed to have an active gauging weir with reliable data, situated downstream of the main catchment river system. Lejweleputswa met the required criteria and was therefore selected as the study area.

The process for the selection of the study catchment (watershed) also considered the direction of water flow. Ensuring that all catchment inlets were situated within the catchment boundary was also crucial for hydrological modelling. In the process of selecting a study catchment, several watersheds were created as shown in **Figure 7.5** (cf. **Section 7.4**). The watershed selected for this study catchment is depicted in **Figure 3.1**. This study catchment has 5 quaternary catchments, namely C41A to C41E. In total, this study area is 4 728 $km^2$.

Other watersheds were not selected for various reasons. For example, the catchment above weir gauge C4H004 was excluded despite having the least percentage of missing data. This decision was due to significant limitations, including the presence of only 2 manual rainfall stations located near the weir, with no stations situated within the catchment itself and the absence of nearby temperature stations. Furthermore, this catchment included a large dam, which would have introduced complexities for hydrological modelling. Similarly, certain catchments with gauging weirs were not selected due to low data availability or insufficient station density near the catchment boundary. For instance, 2 weirs were located at coordinates -28.11722°S; 26.71916°E and -28.11666°S; 26.72527°E. While data were obtained for 6 weirs, the proximity of these 2 weirs resulted in their combined representation on the map (**Figure 3.1**).

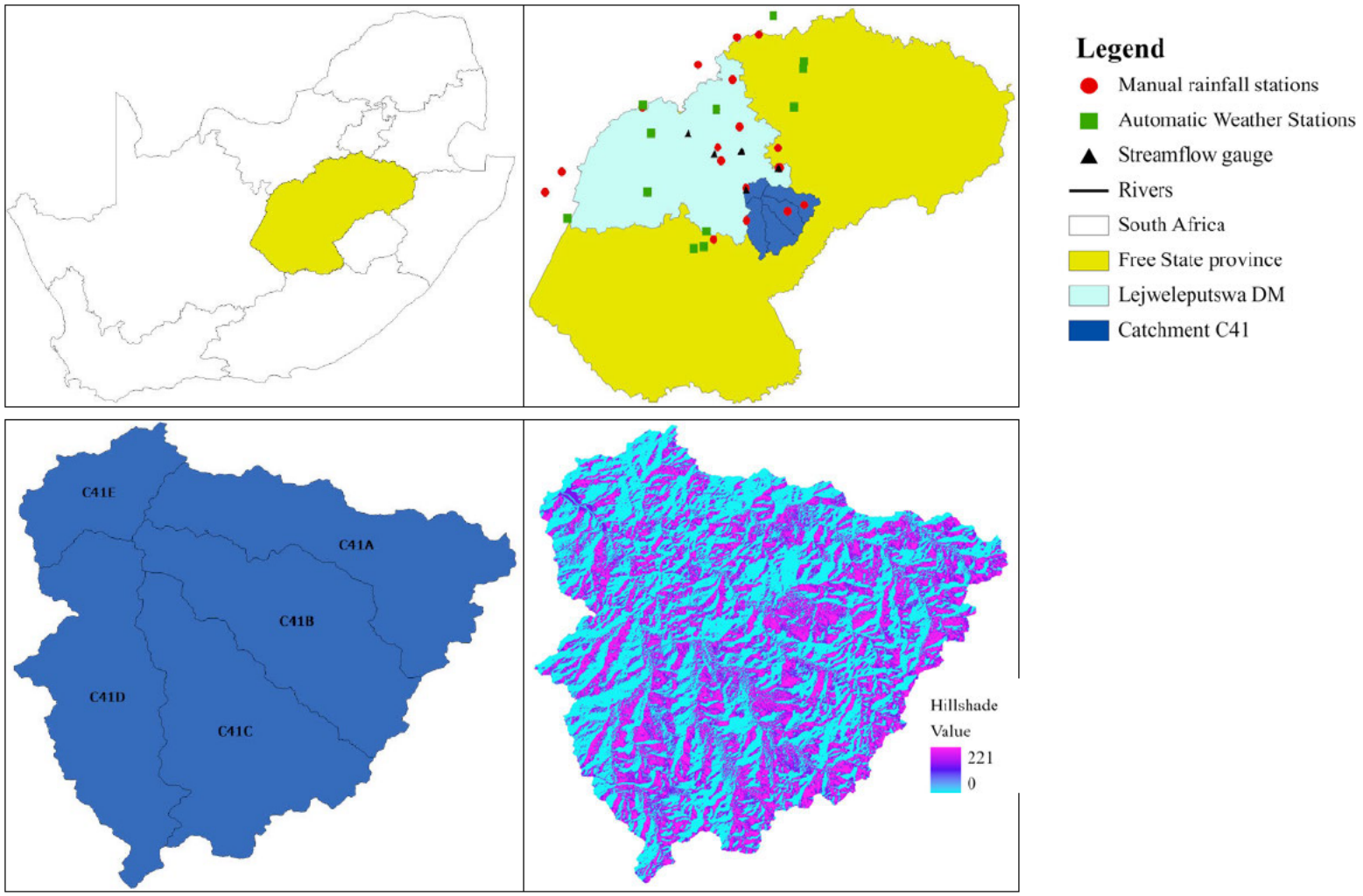


**Figure 3.1**: Tertiary catchment C41 with quaternary catchments C41A-C41E; as well as the location of the manual rainfall stations and AWSs, together with river flow directions for the selected tertiary catchment in the Lejweleputswa district (Free State province)

It is also important to note that data from the SAWS were available only for the northern half of the uMgungundlovu District. Consequently, this district was not considered as a study catchment. Furthermore, the lack of reliable temperature data from AWS for both the Vhembe and uMgungundlovu districts precluded their selection. These limitations highlighted the challenges of identifying an ideal study catchment. While the Lejweleputswa catchment may not be perfect, it represents the most suitable option given the study's objectives and data availability constraints. By addressing these criteria and limitations, this study ensures a robust foundation for hydrological modelling and analysis.

## 3.2 Study catchment description

An overview of land cover in the selected study catchment (C41) was extracted with ArcMap 10.8 (ESRI, 2020) using the South African National Cover (SANLC) raster dataset developed in 2020 by the Department of Forestry, Fisheries and the Environment (DFFE, 2024). As shown in **Figure 3.2** catchment C41 is mostly dominated by natural grassland (green) and irrigated farmlands (brown). This catchment also has a large dam (blue in **Figure 3.2**) located at the outlet of tertiary catchment C41.

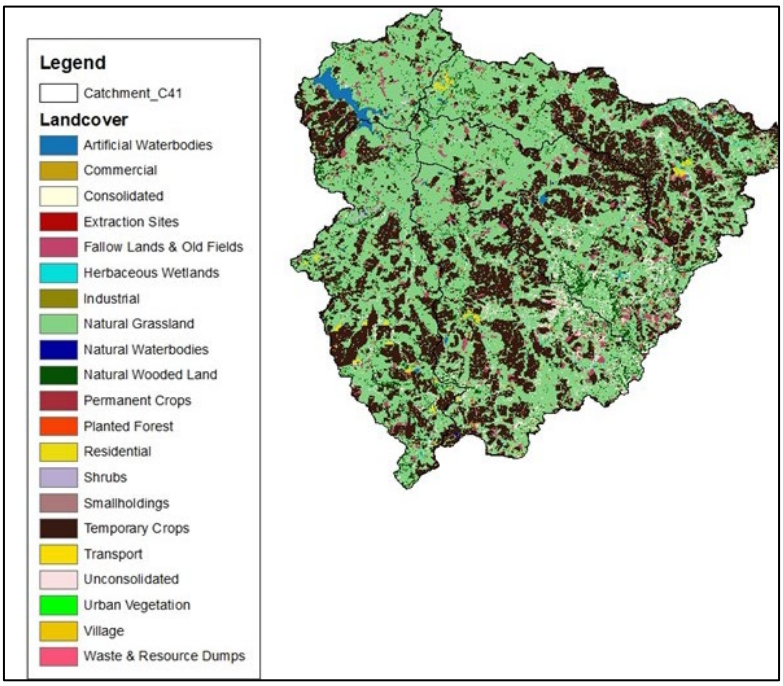


**Figure 3.2**: Overview of land cover in tertiary catchment C41 with a dominance of grassland (green) and irrigated crops (brown) (after DFFE, 2024)

## 3.3 Sourcing of climate data

The QCDB and QnCDB have been invaluable for water resources assessments but only have daily data for the period 1950/01/01-1999/12/31. However, extending these climate databases by an additional 20 years is critical given the climate change-related extreme events that have occurred from 2000 to 2019. Therefore, the intention was to extend existing datasets (available from the Lynch (2004) rainfall database; cf. **Section 2.2.1)** for the selected study catchment. For an independent research project conducted at UKZN, SAWS supplied daily climate data corresponding to the 3 District Municipalities described in **Section 3.3.1** Climate data (rainfall and temperature) were requested for the period 1996/01/01 to 2019/12/31, thus providing a 4-year overlap with existing rainfall records. This would ensure that new datasets could be joined to existing data with confidence. Data were requested for all quinary catchments, both within and upstream of each DM.

Climate data were obtained from multiple sources, including the SAWS, the ARC and the DWS. The locations of manual rainfall stations and AWS used in this study are shown in **Figure 7.1** in **Section 7.1.1** and **Figure 7.3** in **Section 7.1.2,** respectively. However, it is important to note that SAWS only provided data for the northern half of the district. Due to this limitation, this district could not be considered as a study catchment since the available data did not provide complete coverage for hydrological analysis.

### 3.3.1 Background on district municipalities

The Vhembe DM is 1 of the 55 districts in the Limpopo province **(Figure 3.4**). It is the northern most district of the country and lies south of the Zimbabwe border, i.e. the Limpopo River. In addition, Vhembe mainly falls within Primary Catchment A, whilst the south and eastern regions are situated in Primary Catchment B. In total, there are 192 quinary catchments (inclusive of upstream catchments).

The Lejweleputswa DM is 1 of 55 districts in the Free State province (**Figure 3.6**), with Welkom as its seat Lejweleputswa falls entirely within Primary catchment C. In total, there are 570 quinary catchments (inclusive of upstream catchments).

UMgungundlovu is 1 of 11 DMS in KwaZulu-Natal (**Figure 3.8**), with the seat being Pietermaritzburg. Most of the district falls within Primary Catchment U, whilst the north-western region is in Primary Catchment V. In total, there are 273 quinary catchments (inclusive of upstream catchments).

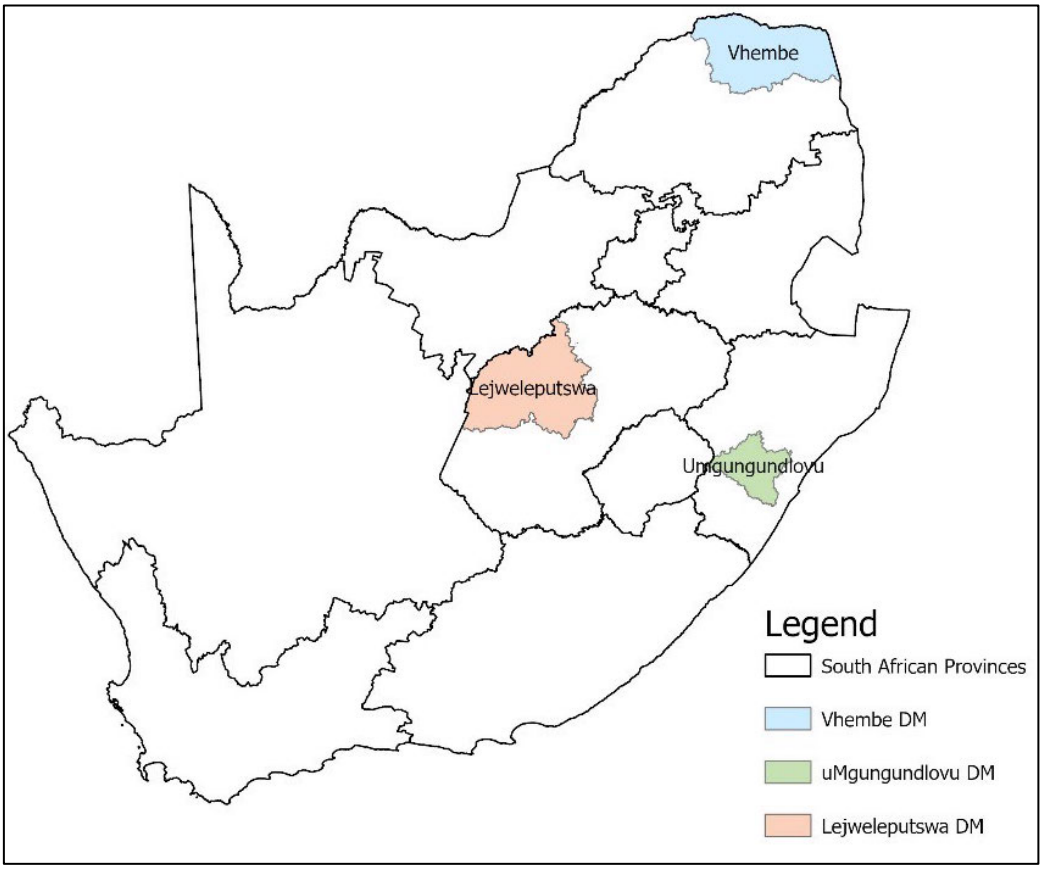


**Figure 3.3**: Location of the 3 district municipalities for which climate data was sourced: Vhembe in Limpopo (blue), Lejweleputswa in Free State (orange) and uMgungundlovu in KwaZulu-Natal (green)

### 3.3.2 Amount of data received

SAWS provided data for 175 rainfall (manual) stations in August 2022. Among these, 14 stations are located in Vhembe, 16 in Lejweleputswa and 1 in uMgundundlovu DM (cf**. Figure 7.1**, in **Section 7.1.1**). The three stations were identified as having good quality data. In October 2022, SAWS also provided daily climate data measured by 22 AWS, which included rainfall and temperature (maximum and minimum).

Of these, only 1, 22 and 0 stations are located within the Vhembe, Lejweleputswa and uMgungundlovu DM, respectively. In addition, 4 of the AWSs are situated in close geographical proximity to one another. As a result, data from "duplicate" stations were merged to form 1 dataset with a longer record (cf. **Table 3.1**), resulting in data for 20 AWSs.

**Table 3.1**: Two duplicate automatic weather stations owned by SAWS, whose data were merged to form 2 datasets with a longer record

| Station ID | Station Name | Latitude | Longitude | Start Year | End year | Merged station ID |
|---|---|---|---|---|---|---|
| 0240808A2 | Durban South | -29.97 | 30.95 | 1996 | 2014 | 0240808A2 |
| 0240837B7 | Merebank | -29.96 | 30.97 | 2014 | 2018 | |
| 0300630_8 | Estcourt | -29.01 | 29.86 | 2012 | 2022 | 0300630_8 |
| 0300690_1 | Estcourt | -29.02 | 29.87 | 2003 | 2013 | |

Due to the lack of good quality climate data within the 3 DMs, additional climate datasets were requested from the ARC and DWS. DWS provided daily rainfall for each district on the 2nd of March 2023. A total of 14 rainfall stations were received from DWS, where 3, 5 and 6 stations are located in Lejweleputswa, Vhembe and uMgungundlovu districts, respectively. In addition, on the 8th of March 2023 the ARC provided 14 AWS. Data received from the council included, *inter alia,* daily rainfall, daily maximum and minimum temperature and evaporation data. Overall, approximately 41, 135, and 40 climate stations are located within and upstream of the Vhembe, Lejweleputswa, and uMgungundlovu District Municipalities, respectively. Of these, 7 stations are situated in the Northern Cape Province. A summary of all stations is provided in **Table 3.2**.

**Table 3.2**: Total number of climate stations (manual and AWS) received from SAWS, ARC and DWS

| Organisation | Number of stations |
|---|---|
| SAWS | 195 |
| ARC | 14 |
| DWS | 14 |
| **Total** | **223** |

### 3.3.3 Assessment of data quality

Schulze et al. (1995) stated that rain gauges in and around a catchment are considered more representative of the catchment's topographical variability. The dataset received from SAWS, DWS and ARC was quality-checked for reliability and the criteria used to select a reliable station were as follows:

- Stations are as close as possible to, or within the district's boundary;
- the station has a long continuous record (1996/01/01 to 2019/12/31) and
- if the station has missing data, then must be less than 90 days (~1 %) of missing data

#### 3.3.3.1 Vhembe District Municipality

In the Vhembe DM, 27 manual stations within and near the DM's boundary record daily rainfall data. Analysis of the station data revealed a significant issue with incomplete daily rainfall records. As shown in **Figure 3.4**, for stations from SAWS (26) and DWS (1), there are notable percentages of missing data. Only 1 of the 27 stations (0723363_X) has a complete record and only 5 stations were considered to have a reliable dataset. In addition, 98 % of 1 station's record was missing (cf. **Table 7.1** in **Section 7.2.1**), which is concerning.

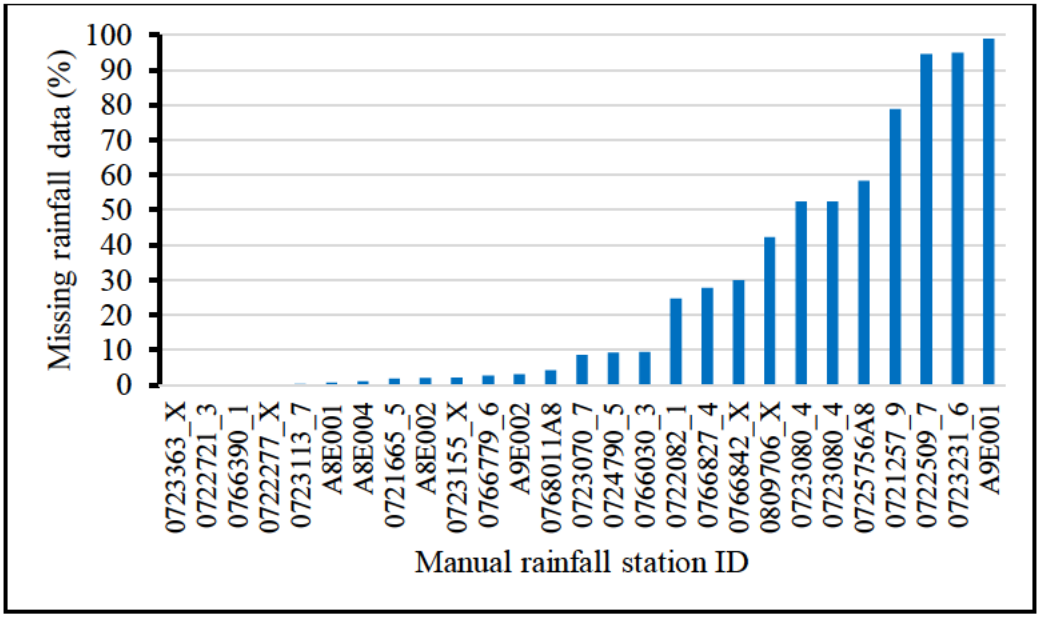


**Figure 3.4**: The portion (expressed as a % of total record length) of missing daily rainfall data in the SAWS and DWS datasets situated in the Vhembe District Municipality

The situation with missing data are similar for the automatic stations in this region. **Figure 3.5** depicts the percentage of missing rainfall and temperature data in datasets obtained from the ARC and SAWS. The Vhembe DM has 7 AWSs within and around its boundaries (cf. **Figure 7.4** in **Section 7.3.1**), with 5 monitored by the ARC and 2 by SAWS. None of the AWSs provide a reliable dataset of both rainfall and temperature, since the missing record exceeds 30 % (cf. **Table 7.4** in **Section 7.3.1**).

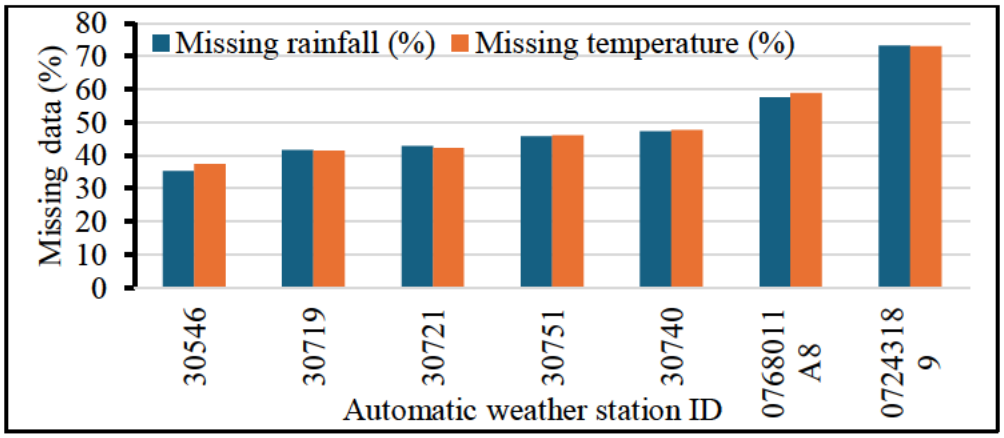


**Figure 3.5**: The portion (expressed as a % of total record length) of missing daily rainfall and temperature data in the ARC and SAWS datasets situated in the Vhembe District Municipality

### 3.3.3.2 Lejweleputswa District Municipality

From **Figure 3.6**, the Lejweleputswa District Municipality has 3 and 28 manual stations from DWS and SAWS, respectively, located within and near its boundaries (cf. **Figure 7.2** in **Section 7.1.1**), of which only 2 (C4E008 and 0261722_8) have no missing data. In total, ten stations were found to have a reliable dataset (cf. **Table 7.2** in **Section 7.2.2**). The district also has 12 AWS within and around its boundaries, as shown in **Figure 3.7**. The detailed reliable and missing data points are shown in **Table 7.5** (cf. **Section 7.3.2**).

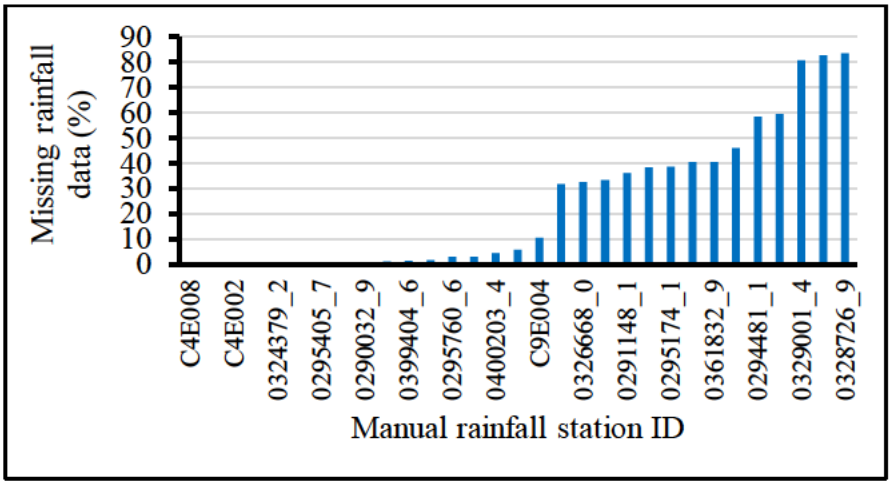


**Figure 3.6**: The portion (expressed as a % of total record length) of missing daily rainfall data in the SAWS and DWS datasets situated in Lejweleputswa

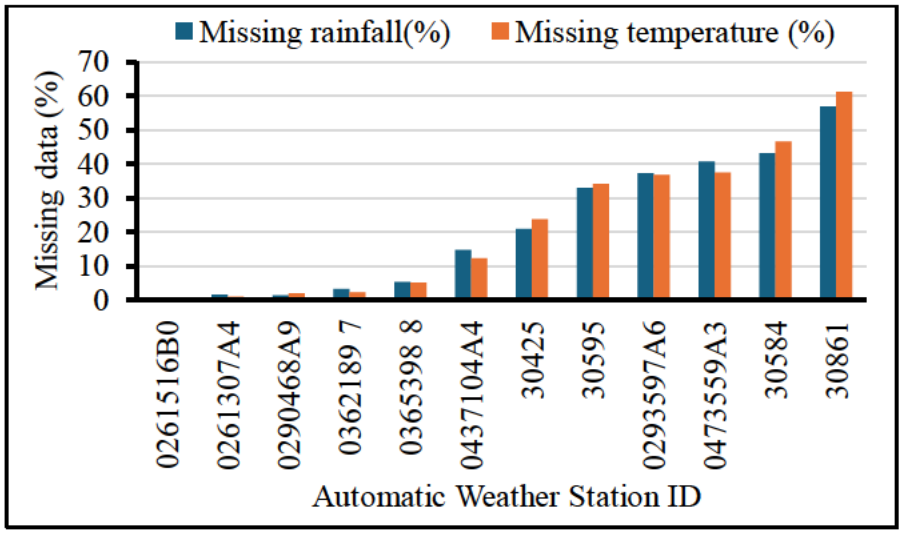


**Figure 3.7**: The portion (expressed as a % of total record length) of missing daily rainfall and temperature data in the ARC and SAWS datasets situated in the Lejweleputswa District Municipality

#### 3.3.3.3 uMgungundlovu District Municipality

**Figure 3.8** shows that uMgungundlovu DM has 6 and 5 manual stations from DWS and SAWS, respectively. Only 1 of the SAWS stations is within the DM's boundary (cf. **Table 7.3** in **Section 7.2.3**).

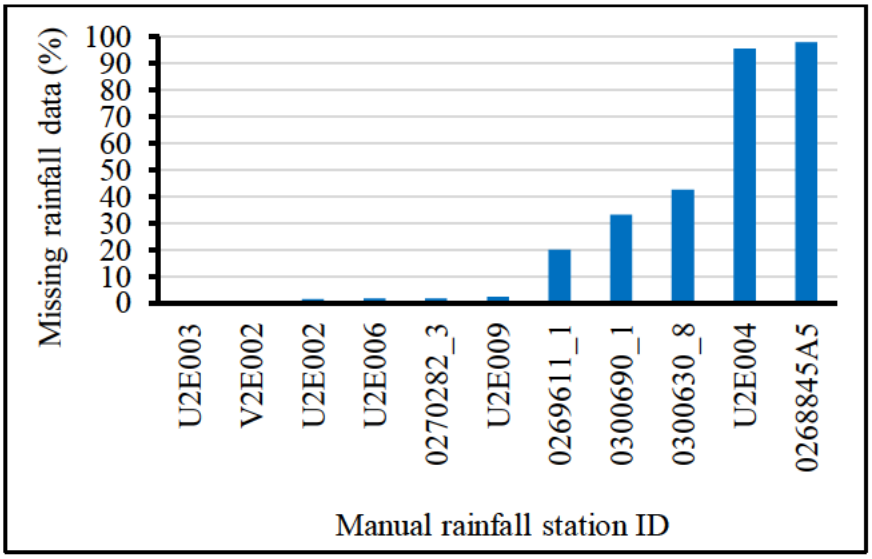


**Figure 3.8**: The portion (expressed as a % of total record length) of missing daily rainfall data in the SAWS and DWS datasets situated in the uMgungundlovu District Municipality

The graph below (**Figure 3.9**) depicts 9 AWSs from SAWS and the ARC located within and near the uMgungundlovu DM. None of the stations in the district have a full climate dataset as they all have missing data (cf. **Table 7.6** in **Section 7.3.3**).

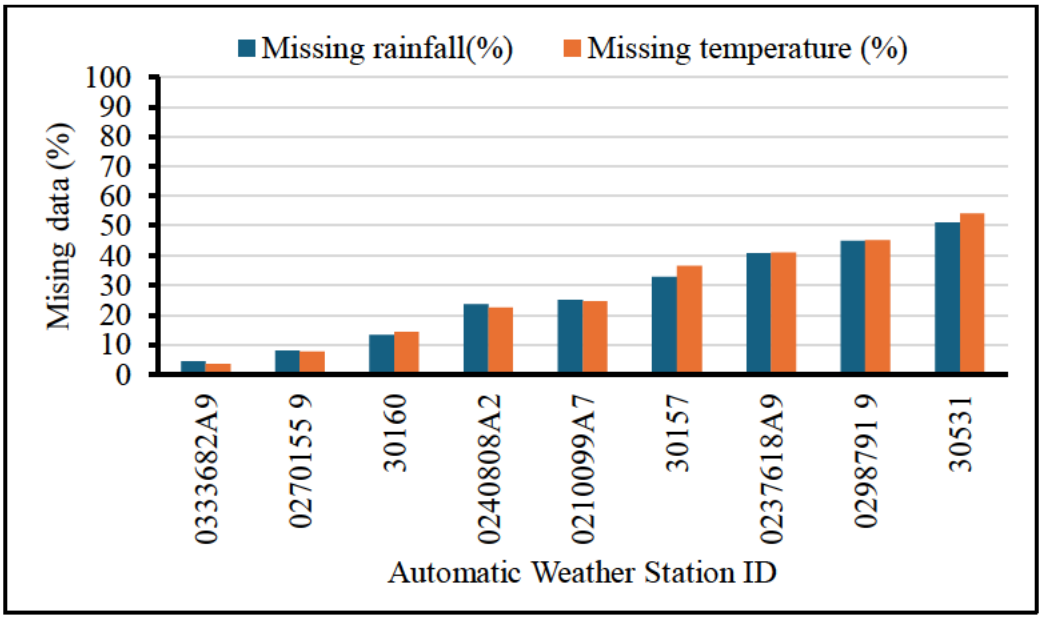


**Figure 3.9**: The portion (expressed as a % of total record length) of missing daily rainfall and temperature data in the ARC and SAWS datasets situated in the uMgungundlovu District Municipality

The high level of missing data from both manual and automatic stations greatly affects the catchment's ability to monitor climate patterns and perform reliable hydrological modelling. The sparse network of climate stations makes it difficult to infill missing data accurately, which reduces confidence in rainfall and temperature estimates and trends. Furthermore, stations with large gaps in climate data can lead to inaccurate conclusions about variability and long-term climate patterns. This inaccuracy can negatively affect important sectors such as agriculture, disaster management and infrastructure planning, which rely on accurate climate information for decision making.

In addition, the gaps in temperature and rainfall records reduce the reliability of selecting climate "driver" stations, which are essential for understating climate variability in the catchment. These issues further highlight the need for better data collection strategies, improved station maintenance and the use of advanced methods to estimate missing data.

## 3.4 Infilling of missing climate data

From **Section 3.3.2**, almost all of the climate stations sourced from custodians had missing data. Incomplete climate data can lead to significant errors in hydrological modelling and water resources management; this highlights the importance of filling any gaps in the climate data records. Methods used to infill the rainfall and temperature daily datasets are discussed next. Additionally, the sparse network of rainfall stations has led to difficulties in infilling missing rainfall data as stations are spread apart and some stations happen to have the same gaps on their records, making it difficult to infill data.

### 3.4.1 Rainfall

Various methods are used to infill missing daily rainfall data (cf. **Section 2.3.1**). This study used the IDW method to estimate missing rainfall. The IDW method is one of the techniques commonly used to estimate missing data in hydrology, due to its satisfactory results in filling rainfall data gaps (Ruezzene et al., 2021). The accuracy of this technique is dependent on the density of stations in an area, so the closer the stations, the more accurate the estimation will be. This method was applied as follows:

$$x^* = \frac{(wi * xi)}{\sum wi} \qquad \textbf{Equation 3.1}$$

where $x^*$ is the unknown target station rainfall (mm), *xi* is the known control station rainfall (mm). The term *wi* is the distance between the known point location *i* and the unknown point location and is calculated as follows:

$$wi = \frac{1}{(si)^2} \qquad \textbf{Equation 3.2}$$

where *si* is the distance between the target and control station (m). The distance $s_i$ between the target station and the control station was calculated using the Euclidean distance. The Euclidean distance between 2 points ($x_1$, $y_1$; coordinates of the target station) and ($x_2$, $y_2$; coordinates of the control station) is given by:

$$s_i = \sqrt{(x_2 - y_2)^2 + (x_1 - y_1)^2} \qquad \textbf{Equation 3.3}$$

### 3.4.2 Temperature

As mentioned in **Section 2.2.2**, Schulze and Maharaj (2004) found that the DSD method outperformed the MTD method. Hence, the DSD method was selected in this study for patching of missing temperature data. Furthermore, a new (i.e. third) method, called the ranking algorithm, was tested in this study to infill temperature datasets where the DSD method was not applicable (cf. **Section 3.4.2.3**).

#### 3.4.2.1 Difference in Standard Deviation Method

The differences in standard deviations between the target and surrounding control stations were computed for each month of the year. The differences were then ranked from lowest to highest, with the lowest highlighting the best control station to use for patching. From **Table 3.3** for differences in standard deviation for maximum temperature, the top-ranked (01) station was used to patch missing daily temperatures in that month.

**Table 3.3**: Ranking of maximum temperature ($T_{MAX}$ in °C) difference in standard deviation estimated monthly using various control temperature stations in the Lejweleputswa DM for target station 0261516B0

| Control stations | Monthly difference (target – control) in standard deviation (station ranking in brackets) | | | | | |
|---|---|---|---|---|---|---|
| | Jan | Feb | Mar | Apr | May | Jun |
| 0293597A6 | 0.03 (01) | -0.34 (04) | 0.38 (04) | -0.02 (01) | -0.04 (02) | -0.04 (01) |
| 0362189_7 | 0.05 (02) | 0.18 (02) | 0.76 (08) | -0.04 (05) | 0.001 (01) | -0.28 (07) |
| 0363792_A | -0.26 (06) | -0.36 (05) | 0.79 (09) | -0.11 (09) | 0.04 (03) | -0.11 (03) |
| 0261307A4 | 0.23 (04) | -0.02 (01) | -0.29 (03) | 0.05 (07) | -0.18 (07) | 0.10 (02) |
| 0290468A9 | 0.22 (03) | -0.40 (06) | 0.17 (02) | -0.03 (04) | -0.57 (11) | -0.55 (11) |
| 0365398_8 | -0.29 (09) | -0.55 (08) | 0.51 (05) | 0.05 (06) | 0.26 (09) | 0.26 (06) |
| 0437104A4 | -0.49 (12) | -0.75 (10) | 0.90 (10) | -0.22 (11) | 0.05 (04) | 0.34 (08) |
| 0473559A3 | -0.26 (07) | -0.76 (11) | 1.49 (11) | 0.02 (02) | 0.30 (10) | 0.44 (10) |
| 30584 | -0.30 (10) | -1.18 (12) | 2.04 (12) | 1.02 (12) | -0.86 (12) | 0.63 (12) |
| 30909 | 0.24 (05) | -0.27 (03) | -0.14 (01) | 0.16 (10) | 0.10 (05) | 0.13 (04) |
| 30425 | -0.28 (08) | -0.58 (09) | 0.72 (07) | -0.07 (08) | 0.19 (08) | 0.21 (05) |
| 30861 | -0.32 (11) | -0.42 (07) | 0.53 (06) | -0.03 (03) | 0.14 (06) | 0.38 (09) |

It is important to note that the sign of the difference was not considered in the ranking process. The sign served only as an indicator of whether the temperature would be adjusted upwards or downwards by adding (if positive) or subtracting (if negative) the adjustment factor, as seen in **Table 3.4**. Temperature data with no patching code are observed values and patched data are flagged as p-1 (cf. **Table 3.5**). The patched data are the data that were infilled, the patched data were thereafter adjusted as explained above.

#### 3.4.2.2 Mean temperature difference

For each month of the year, the difference in the monthly averages was calculated for minimum and maximum temperatures for all the control stations. The monthly mean differences estimated here were then used to adjust any infilled temperature dataset as done by Schulze and Maharaj (2004; cf. **Section 2.3.2.2**).

**Table 3.4**: Rankings of control stations using the mean temperature difference method for maximum temperature ($T_{MAX}$) from January to June for target station 0261516B0

| Control station | Mean monthly differences (target – control) in $T_{MAX}$ (°C) (station ranking) | | | | | |
|---|---|---|---|---|---|---|
| | **Jan** | **Feb** | **Mar** | **Apr** | **May** | **Jun** |
| 0293597A6 | -2.77 | -2.73 | -2.91 | -2.27 | -2.40 | -1.32 |
| 0362189_7 | -0.08 | -1.12 | -1.96 | -0.99 | -1.30 | -1.44 |
| 0363792_A | 0.32 | 0.02 | -0.70 | -0.73 | -1.11 | -1.58 |
| 0261307A4 | 0.71 | 0.66 | -0.60 | -0.42 | -0.57 | 0.54 |
| 0290468A9 | -2.62 | -1.87 | -2.62 | -1.18 | -1.25 | -0.70 |
| 0365398_8 | 1.50 | 0.89 | 0.50 | 0.74 | 0.25 | -0.74 |
| 0437104A4 | -0.93 | -2.38 | -1.84 | -2.55 | -2.93 | -3.87 |
| 0473559A3 | 2.40 | 1.09 | 0.57 | 0.44 | -0.79 | -2.14 |
| 30584 | 1.30 | 0.41 | -4.43 | -7.80 | -6.99 | -7.94 |
| 30909 | -1.51 | -1.46 | -2.33 | -1.44 | -1.55 | -1.11 |
| 30425 | 2.36 | 0.82 | 0.54 | 0.38 | 0.09 | -1.14 |
| 30861 | 1.41 | 0.47 | -0.07 | -0.27 | -0.80 | -1.89 |

**Table 3.5**: An example of how the maximum patched temperature in March 2013 was adjusted using the mean monthly differences estimated for the station

| Date | Observed/*patched* $T_{MAX}$ | Patching code | Adjusted $T_{MAX}$ |
|---|---|---|---|
| 2013/03/01 | 28.40 | | |
| 2013/03/02 | 29.00 | | |
| 2013/03/03 | 31.50 | | |
| 2013/03/04 | 33.00 | | |
| 2013/03/05 | *36.36* | p-1 | 34.0 |
| 2013/03/06 | *34.02* | p-1 | 31.7 |
| 2013/03/07 | *32.50* | p-1 | 30.2 |
| 2013/03/08 | *34.15* | p-1 | 31.8 |
| 2013/03/09 | *32.80* | p-1 | 30.5 |
| 2013/03/10 | *30.85* | p-1 | 28.5 |
| 2013/03/11 | *29.73* | p-1 | 27.4 |
| 2013/03/12 | 29.90 | | |
| 2013/03/13 | 31.70 | | |
| 2013/03/14 | 29.10 | | |
| 2013/03/15 | 31.10 | | |

#### 3.4.2.3 Temperature station ranking algorithm

The station ranking algorithm (cf. **Section 2.6.2.2**) was applied for the first time in this study as a patching method. It was applied for stations where DSD method was not applicable, i.e. when an entire month of daily data were missing, and thus standard deviations could not be computed. For this method, the equations modified by Kunz et al. (2020) with DIST normalised by 200 instead of 350 was used and DALT by 1 000 m instead of 1 500 m.

The computed rankings (cf. **Table 3.6**) were then used to determine the best patching (control) station. A lapse rate adjustment was applied to the control temperature data to account for the elevation difference between the target station and the selected control as follows:

$$AF = T + (DALT * \frac{LR}{1\,000}) \qquad \textbf{Equation 3.4}$$

The adjustment factor (*AF*) is determined using the maximum and minimum temperature (*T*, °C), the difference in elevation between the target and control station (*DALT*, m) and the lapse rate (*LR*, °C/m). Pegram et al. (2016) tested various methods to infill missing rainfall data and to validate each method. Up to 20 % of the observed record was artificially removed from the target station to mimic a portion of missing data. The "missing" data were infilled using data from the chosen control station, then compared to the observed values. A similar procedure was followed in this study to validate the ranking algorithm as a patching method. Station 0261516B0 was used as the target where 20 % of its observed data were removed. Station 0261307A4 was selected as a control since it was closest to the target station and had a relatively small elevation difference (cf. **Table 3.7**).

**Table 3.6**: Ranking of neighbouring control stations with the ranking algorithm method to select the best station to patch target station 0261516B0 using the distance (DIST in minutes of a degree) and difference in elevation (DALT in metres) from the control to the target station, as well as the distance factor (DF), altitude factor (AF), ranking factor (RF) and Lapse Rate Region (LRR) number

| | ***DIST (min)*** | ***DALT (m)*** | ***DF*** | ***AF*** | ***RF*** | ***LRR*** |
|---|---|---|---|---|---|---|
| 0261307A4 | 6.71 | -59 | 0.97 | 1.04 | 12.80 (01) | 3 |
| 0293597A6 | 9.77 | 50 | 0.96 | 0.97 | 12.47 (02) | 4 |
| 30909 | 50.63 | 59 | 0.77 | 0.96 | 10.62 (03) | 3 |
| 30584 | 78.17 | 34 | 0.65 | 0.98 | 9.42 (04) | 3 |
| 0363792_A | 84.96 | 65 | 0.62 | 0.96 | 9.06 (05) | 3 |
| 0290468A9 | 96.43 | 156 | 0.58 | 0.91 | 8.51 (06) | 3 |
| 0362189_7 | 96.09 | 124 | 0.57 | 0.93 | 8.45 (07) | 3 |
| 0365398_8 | 105.04 | -75 | 0.53 | 1.05 | 8.41 (08) | 4 |
| 30861 | 128.26 | -28 | 0.42 | 1.02 | 7.27 (09) | 4 |
| 30425 | 132.32 | -28 | 0.40 | 1.02 | 7.10 (10) | 4 |
| 0437104A4 | 149.13 | 5 | 0.33 | 1.00 | 6.28 (11) | 4 |
| 0473559A3 | 170.17 | -147 | 0.23 | 1.09 | 5.61 (12) | 4 |

**Table 3.7** shows a portion of results (January 2007) obtained after testing the ranking algorithm method on station 0261516B0. The method proved to infill $T_{MAX}$ and $T_{MIN}$ successfully, as there are slight differences between the observed and patched data. For the purpose of comparison, the same period of data as **Table 3.5** (i.e. infilled by DSD method) was infilled using the ranking algorithm (cf. **Section 4.1.2**).

**Table 3.7**: Results of target station 0261516B0 $T_{MAX}$ and $T_{MIN}$ (°C) patched using the algorithm method for January 2007

| | Maximum temperature (°C) | | | | Minimum temperature (°C) | | | |
|---|---|---|---|---|---|---|---|---|
| **Date** | **Observed** | **Patched** | **Adjusted** | **Difference (Observed − patched)** | **Observed** | **Patched** | **Adjusted** | **Difference (Observed − patched)** |
| 2007/01/01 | 25.5 | 27.5 | 26.2 | -0.7 | 7.0 | 10.6 | 11.4 | -4.4 |
| 2007/01/02 | 26.7 | 28.5 | 27.3 | -0.6 | 10.9 | 13.6 | 12.8 | -1.9 |
| 2007/01/03 | 30.8 | 32.5 | 29.3 | 1.5 | 12.4 | 16.3 | 13.8 | -1.4 |
| 2007/01/04 | 31.5 | 32.8 | 30.9 | 0.6 | 12.1 | 16.0 | 15.9 | -3.8 |
| 2007/01/05 | 31.9 | 33.0 | 31.8 | 0.1 | 17.1 | 20.7 | 15.1 | 2.0 |
| 2007/01/06 | 33.3 | 34.5 | 33.1 | 0.2 | 13.4 | 17.8 | 14.8 | -1.4 |
| 2007/01/07 | 29.0 | 29.9 | 33.9 | -4.9 | 13.4 | 15.3 | 18.4 | -5.0 |
| 2007/01/08 | 28.0 | 29.3 | 31.0 | -3.0 | 7.5 | 11.1 | 13.8 | -6.3 |
| 2007/01/09 | 31.3 | 32.7 | 32.1 | -0.8 | 12.4 | 15.0 | 13.7 | -1.3 |
| 2007/01/10 | 32.4 | 34.0 | 31.4 | 1.0 | 10.5 | 17.9 | 17.2 | -6.7 |
| 2007/01/11 | 32.6 | 33.9 | 33.1 | -0.5 | 12.8 | 17.8 | 15.2 | -2.4 |
| 2007/01/12 | 34.9 | 35.3 | 34.2 | 0.7 | 11.9 | 17.1 | 15.0 | -3.1 |
| 2007/01/13 | 35.4 | 36.1 | 35.7 | -0.3 | 11.0 | 16.4 | 18.5 | -7.5 |
| 2007/01/14 | 34.1 | 34.5 | 34.2 | -0.1 | 15.7 | 19.0 | 16.2 | -0.5 |
| 2007/01/15 | 31.3 | 31.4 | 28.8 | 2.5 | 11.9 | 16.4 | 15.6 | -3.7 |
| 2007/01/16 | 30.3 | 29.8 | 31.4 | -1.1 | 9.5 | 14.6 | 15.7 | -6.2 |
| 2007/01/17 | 32.5 | 32.3 | 31.7 | 0.8 | 10.7 | 15.4 | 16.6 | -5.9 |
| 2007/01/18 | 31.3 | 31.5 | 31.1 | 0.2 | 11.2 | 15.1 | 16.0 | -4.8 |
| 2007/01/19 | 32.1 | 33.0 | 33.2 | -1.1 | 11.8 | 15.9 | 16.0 | -4.2 |
| 2007/01/20 | 31.0 | 32.4 | 28.9 | 2.1 | 14.1 | 17.2 | 18.0 | -3.9 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2007/01/21 | 29.4 | 30.7 | 28.2 | 1.2 | 15.6 | 16.4 | 17.6 | -2.0 |
| 2007/01/22 | 30.1 | 30.3 | 30.9 | -0.8 | 13.5 | 15.4 | 13.9 | -0.4 |
| 2007/01/23 | 33.1 | 34.2 | 33.3 | -0.2 | 14.5 | 17.3 | 17.8 | -3.3 |
| 2007/01/24 | 32.6 | 34.6 | 32.3 | 0.3 | 16.0 | 18.6 | 17.9 | -1.9 |
| 2007/01/25 | 29.8 | 30.4 | 31.3 | -1.5 | 16.9 | 19.4 | 17.7 | -0.8 |
| 2007/01/26 | 29.3 | 29.5 | 32.6 | -3.3 | 12.9 | 16.5 | 17.3 | -4.4 |
| 2007/01/27 | 28.5 | 30.3 | 31.6 | -3.1 | 15.1 | 17.0 | 18.1 | -3.0 |
| 2007/01/28 | 30.0 | 32.0 | 29.9 | 0.1 | 14.9 | 17.2 | 16.3 | -1.4 |
| 2007/01/29 | 29.8 | 30.4 | 31.8 | -2.0 | 16.5 | 19.2 | 16.5 | 0.0 |
| 2007/01/30 | 28.0 | 29.2 | 33.8 | -5.8 | 16.2 | 18.5 | 16.2 | 0.0 |
| 2007/01/31 | 28.6 | 29.7 | 30.3 | -1.7 | 13.5 | 15.5 | 15.5 | -2.0 |

## 3.5 Driver station selection methods

Driver stations for each quaternary catchment within the study catchment were initially selected using the *DRE* utility (cf. **Section 2.5.3)**. Thereafter, driver stations were selected by the shortest distance (i.e. closest station) from the quaternary catchment centroid point to an active reliable station. Other techniques were also tested, such as choosing the rain gauge with the lowest set of monthly rainfall adjustment factors as a suitable driver station. Different rainfall grids were tested for calculating the adjustment factors.

### 3.5.1 Rainfall

#### 3.5.1.1 Driver station approach

For the selection of rainfall driver stations for each quaternary catchment within the study catchment, the distance was the main variable determined and used. The distance was measured from the quaternary catchment's centroid to the rainfall station (minutes of a degree). The closest station with no specific distance buffer was selected.

#### 3.5.1.2 Adjustment factor method

Applying monthly adjustment factors improve the driver station's representativeness of the catchment's average rainfall. Rainfall adjustment factors were computed for each month (*i*) as:

$$Rainfall\ adjustment\ factor_i = \frac{Catchment\ mean\ or\ median_i}{Station\ mean\ or\ median_i} \qquad \textbf{Equation 3.5}$$

The following rules developed by Schulze et al. (2011) were used to calculate the rainfall adjustment factors:

if *stn_rfl* = 0 and *cat_rfl* = 0, then *cor_fac* = 1.00

if *stn_rfl* = 0 and *cat_rfl* > 0, then *cor_fac* = 2.00

if *stn_rfl* > 0 and *cat_rfl* = 0, then *cor_fac* = 0.50

if *stn_rfl* > 0 and *cat_rfl* > 0, then *cor_fac* = *cat_rfl*/*stn_rfl*

if *cor_fac* < 0.5, then *cor_fac* = 0.50

if *cor_fac* > 2.0, then *cor_fac* = 2.00

Where station rainfall (*stn_rfl*) represents the measured rainfall at a specific station, catchment rainfall (*cat_rfl*) refers to the average rainfall over the catchment area and the calculated monthly correction factor (*cor_fac*) ranges between 0.5 and 2.0 to account for spatial variability in rainfall distribution.

The adjustment factors were estimated using 4 existing interpolated rainfall grids, namely Dent median, Lynch median, Lynch mean and Pegram mean (**Section 2.4.1**). This was done to determine which of the 4 datasets ensures that point rainfall data from the "driver" station are most representative of catchment's spatial rainfall. Furthermore, the monthly adjustment factors were utilised in each method for selecting rainfall driver station, where a rainfall station with a set of adjustment factors approaching one was deemed more representative of the quaternary catchment, and thus selected as the best driver station.

### 3.5.2 Temperature

The revised ranking algorithm (Kunz et al., 2020; cf. **Section 2.6.2.2**) was used to select temperature driver stations. This was the only method used to select temperature stations for this study.

## 3.6 Comparison of observed and gridded climate data

A comparison was made between point climate data (i.e. rainfall and temperature) for each selected driver station and spatially averaged climate data for the catchment. This comparison was done to assess the accuracy of the grids used in the calculation of adjustment factors. For rainfall, this was done using 2 gridded mean (monthly and annual) rainfall surfaces (Dent et al., 1989; Lynch, 2004) and 2 median grids (Lynch, 2004, Pegram et al., 2016).

For temperature, the mean monthly $T_{MAX}$ and $T_{MIN}$ gridded estimates developed by Schulze and Maharaj (2004) were used. The grid datasets were evaluated against observed data using the following 3 statistical measures:

- **Coefficient of determination ($R^2$)**: Measures the proportion of variance in the dependent variable that is predictable from the independent variable(s). Values closer to 1 indicate a better fit.
- **Nash-Sutcliffe Efficiency (NSE)**: Used to assess the predictive power of hydrological data values range from -∞ to 1, with 1 indicating a perfect match between observed and simulated data.
- **Root Mean Square Error (RMSE)**: Measures the differences between a model's predicted and observed values. Lower values indicate better spatial data representation. (Kouzehgar and Eslamian., 2023).

## 3.7 Verification of driver station selection methods

### 3.7.1 Sourcing of streamflow data

The Lejweleputswa DM was found to have more reliable data for both rainfall and temperature stations. For this DM, daily streamflow data from 1950 to 2019 was sourced from DWS. The streamflow data are required to verify the selection of climate driver stations using hydrological modelling. **Figure 3.10** shows the percentage of missing streamflow data for 6 gauging weirs, which ranged from 32-79 %. This is concerning since a full record of observed streamflow is required to compared against simulated streamflow.

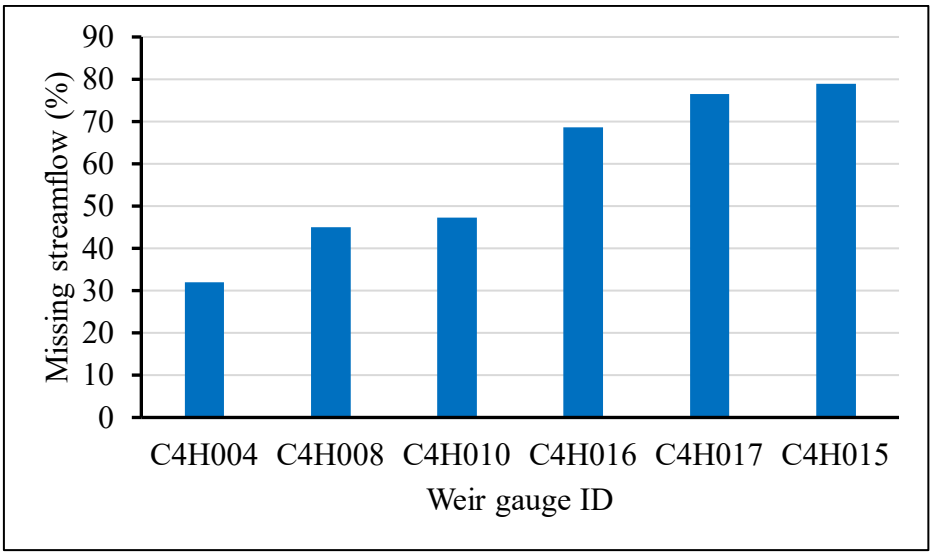


**Figure 3.10**: The portion (expressed as a % of total record length) of missing daily streamflow data in the DWS datasets situated in the Lejweleputswa District Municipality

### 3.7.2 Water balance data

The Erfenis Dam is located mainly in quaternary catchment C41E, with an upper portion in quaternary catchment C41D. The dam's area is 38 $km^2$ at fully supply capacity (Fourie et al., 2013). The dam holds approximately 212.2 million $m^3$ of water, with a 489 m long and 34 m high dam wall (ORASECOM, 2013). Adjacent to the dam is a reserve covering about 4 $km^2$ of grassland (Fourie et al., 2013).

The C4H010 gauging weir at the base of the dam is for water that is released into an irrigation canal that runs alongside the main river and there is no weir located on the main river below the dam wall. As a result, the weir data received from DWS were not useful.

According to Ginster et al. (2014), Erfenis Dam is a critical part of the Sand-Vet water scheme designed to supply water for irrigation and municipal use in parts of the Free State province. It supports about 76 $km^2$ of agricultural irrigation supplies water to towns such as Brandfort and Bultfontein. The dam is integral to Brandfort's water infrastructure, as the town's bulk water supply is drawn from the Sand-Vet canal about 16 km downstream of the dam, where it is then pumped to the town's water treatment works.

As shown in **Figure 3.11**, the Erfenis Dam has 3 main inlets, 2 of which are in sub-catchment C41D and 1 is in C41E. The presence of multiple inlets makes simulating the Erfenis Dam in a hydrological model very difficult. DWS was contacted to provide a water balance for the Erfenis Dam. An example is shown in **Table 3.8**, where the quality codes are explained as follows:

**Table 3.8**: The quality codes and their descriptions for the Erfenis Dam water balance data received from DWS

| Quality code | Description |
|---|---|
| & | Good monthly reading |
| * | Program estimate |
| E | Estimate (either audited or unaudited) |
| M | Data missing / period of no record / data missing (for system use only) |
| Q | checked - still unaudited / edited and checked - still unaudited / checked GPlate reading or dip level reading - still unaudited / good edited unaudited |

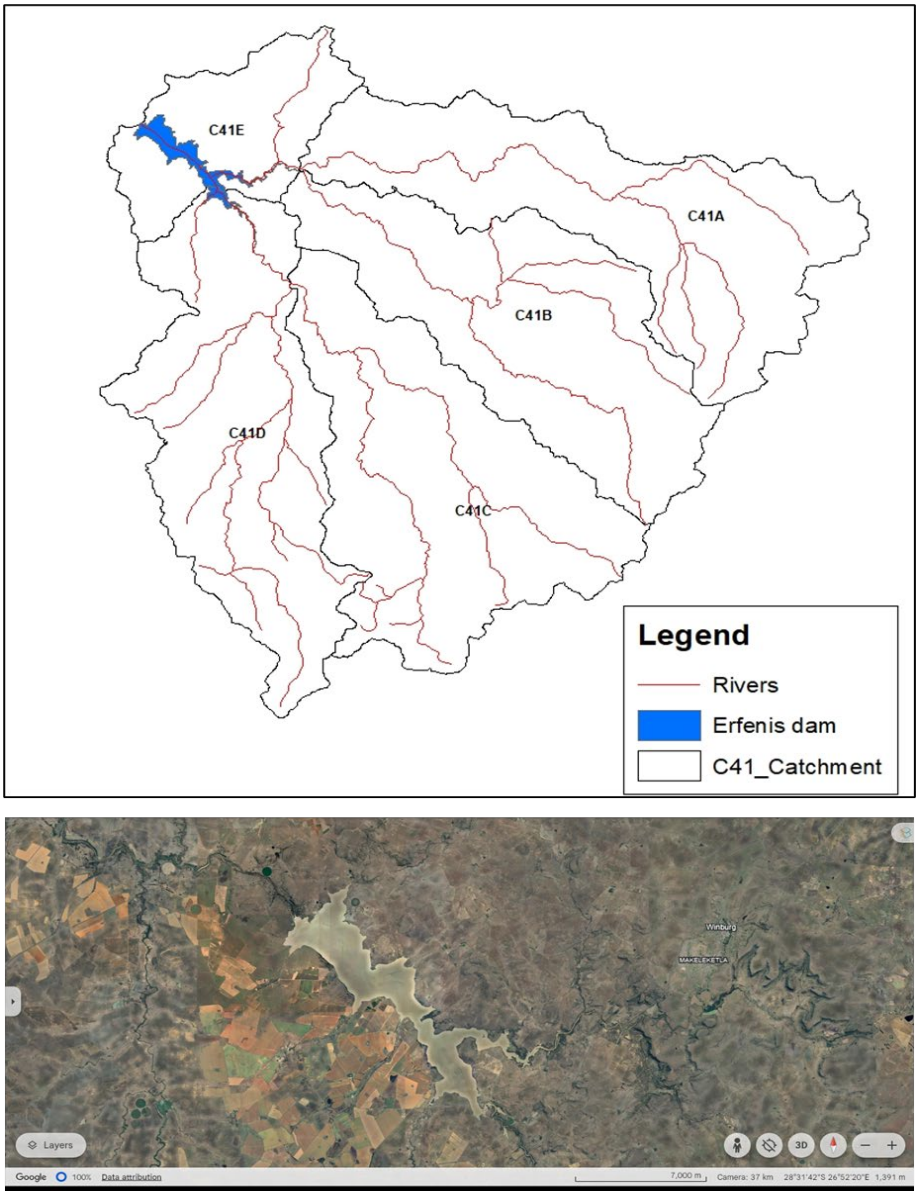


**Figure 3.11**: Map of the Erfenis dam and the river inlets to the dam, as well as an image of the dam from Google Earth

Using the above quality control codes (cf. **Table 3.8**), a reliability table was created, where E, * and M were deemed unreliable and codes &, Q and not coded data were deemed reliable (**Table 3.9**). The analysis highlights critical disparities in data reliability within the DWS water balance, with rain and irrigation data exhibiting high reliability, suggesting robust monitoring systems in these areas.

However, the relatively lower reliability of parameters such as river releases and streamflow underscores potential vulnerabilities in monitoring practices, which could affect the accuracy of water allocation and resource planning. These findings point to the need for targeted improvements in data collection and validation processes, particularly for parameters that directly influence water availability assessments, to enhance the overall reliability of hydrological models.

**Table 3.9**: Reliable and unreliable data from the DWS water balance using quality codes, expressed as % of the total number of months (744 months)

| Parameter | Reliable (%) | Unreliable (%) |
|---|---|---|
| River releases | 72.45 | 27.55 |
| Irrigation | 96.24 | 3.76 |
| Gross evaporation | 79.97 | 20.03 |
| Rain | 99.87 | 0.13 |
| Streamflow | 79.70 | 20.30 |

According to DWS, the dam inflow is estimated by considering the change in storage level with all measured releases, losses and gains into/from the dam. The equation used by DWS to calculate inflow in megalitres (ML) is as follows:

$$Calculated\ inflow = Change\ in\ storage\ level + total\ river\ releases + irrigation + evaporation - rainfall \quad \textbf{Equation 3.6}$$

**Table 3.10**: Sample of the monthly Erfenis Dam water balance provided by DWS, including data quality codes

| Date | Gauge reading (m) | Contents (ML) | Uncont. Spill (ML) | Total outflow (ML) | Total river releases (ML) | Irrigation (ML) | Gross evap. (ML) | Rainfall (ML) | Calculated inflow (ML) | Unaccounted losses (ML) |
|---|---|---|---|---|---|---|---|---|---|---|
| 2010/01 | 25.638 | 80 392 | 0 | 2 596 | 18 | 2 578 | 2 809 (*) | 3 325 (Q) | 74 590 (*) | 0 |
| 2010/02 | 28.759 | 152 903 | 0 | 3 757 | 16 | 3 741 | 3 183 (Q) | 338 (Q) | 43 630 (Q) | 0 |
| 2010/03 | 29.994 | 189 931 | 0 | 8 683 | 25 | 8 658 | 3 065 (*) | 852 (Q) | 4 742 (*) | 0 |
| 2010/04 | 29.797 | 183 777 | 0 | 2 121 | 14 | 2 108 | 2 990 (Q) | 389 (Q) | 3 166 (Q) | 0 |
| 2010/05 | 29.747 | 182 220 | 0 | 1 706 | 7 | 1 699 | 2 361 (Q) | 0 (Q) | 186 (Q) | 0 |
| 2010/06 | 29.620 | 178 339 | 0 | 2 253 | 5 | 2 248 | 2 522 (Q) | 0 (Q) | 379 (Q) | 0 |
| 2010/07 | 29.476 | 173 944 | 0 | 3 396 | 5 | 3 391 | 2 415 (Q) | 0 (Q) | 657 (Q) | 0 |
| 2010/08 | 29.306 | 168 790 | 0 | 4 811 | 6 | 4 806 | 2 732 (*) | 0 (M) | 0 (M) | 0 |
| 2010/09 | 29.076 | 161 979 | 0 | 9 697 | 13 | 9 684 | 2 852 (*) | 0 (Q) | 14 (*) | 0 |
| 2010/10 | 28.635 | 149 444 | 0 | 12 101 | 17 | 12 084 | 2 104 (Q) | 0 (Q) | 0 (Q) | 639 (Q) |
| 2010/11 | 28.085 | 134 600 | 0 | 6 353 | 9 | 6 344 | 2 065 (Q) | 1 307 (Q) | 1 199 (Q) | 0 |
| 2010/12 | 27.855 | 128 688 | 0 | 322 | 15 | 3 229 | 2 620 (Q) | 2 710 (Q) | 50 912 (Q) | 0 |

### 3.7.3 *ACRU* model description

The *ACRU* agro-hydrological model (Schulze, 1995; Smithers and Schulze, 1995) is a daily time-step, conceptual model developed in South Africa to simulate agro-hydrological processes. **Figure 3.12** represents important components of the hydrological cycle within the *ACRU* model. Notably, one of the significant advantages of the *ACRU* model is its ability to account for baseflow contributions to total runoff (Rowe and Smithers, 2019).

The model requires inputs of climate data, soil parameters, slope and elevation data, including parameters describing the vegetation cover. Flow alterations (e.g. irrigation, dams and water transfers) can also be included in the model setup. While *ACRU* permits direct calibration of certain input parameters, Schulze (1995) cautions against this approach, as the model is not designed for parameter fitting or optimisation purposes. Instead, parameter values are based on physical characteristics of the catchment, determined through field investigations. Model performance is assessed by comparing simulated streamflow against observed data.

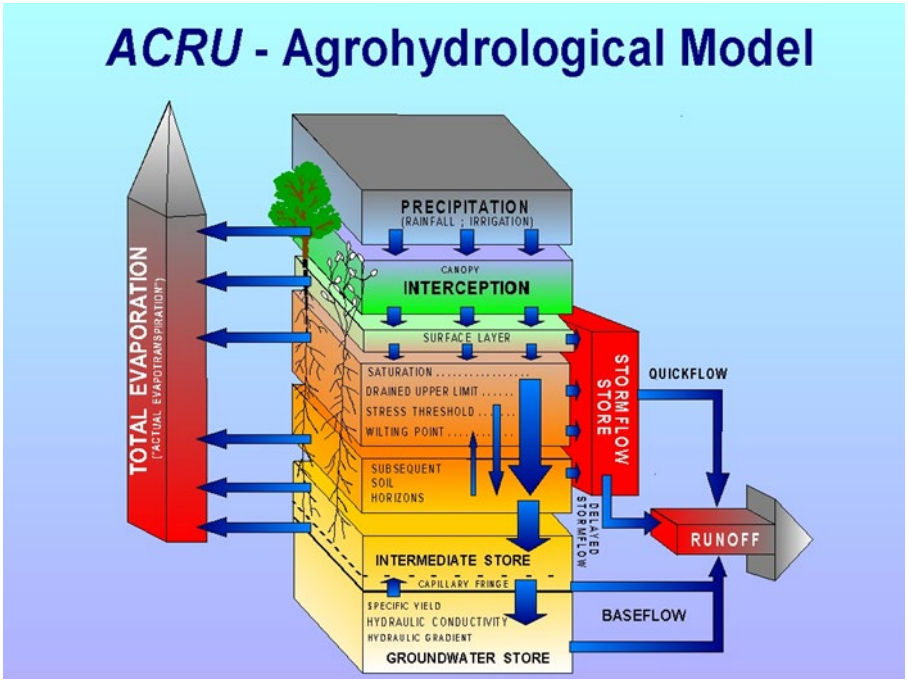


**Figure 3.12**: A schematic of *ACRU* processes represented in the *ACRU* model (Schulze, 1995)

### 3.7.4 *ACRU* model configuration

#### 3.7.4.1 Climate data

An *ACRU* menu developed by Clulow et al. (2023) was used to obtain soil and vegetation information for the study's quaternary catchments. The daily rainfall for the selected driver station was used as input for *ACRU*. To improve the driver station's representativeness of the catchment's rainfall, calculated adjustment factors were applied. A-pan evaporation was estimated in *ACRU* using the Hargreaves and Samani (1985) equation, which uses temperature as its sole input. The temperature for the selected driver stations was used as input for *ACRU* to estimate this A-pan evaporation. No adjustments were applied to the A-pan reference estimates.

#### 3.7.4.2 Soils data

Soil data for each horizon included the thickness and soil water retention values, such as the permanent wilting point, field capacity and saturation. Additionally, the model requires the fraction of saturated soil water above field capacity to be redistributed daily from the subsoil into the groundwater store. All soil parameters were obtained by area-weighting soil information for each of the 3 quinaries within each quaternary catchment.

#### 3.7.4.3 Vegetation data

Monthly vegetation parameters, including the crop coefficient, vegetation interception loss, fraction of roots in the topsoil and the coefficient of initial abstraction, were also obtained using the area-weighting approach from quinary to quaternary catchment scales. It is important to note that although the catchment consists of natural grassland, several irrigated crops and numerous small farm dams, the model was only configured for natural grassland. This limitation was due to inadequate information about the specific crops cultivated and the volume of water abstracted from rivers and farm dams for irrigation. Furthermore, essential *ACRU* inputs for irrigated areas, such as rooting depth and percentage of ground cover (cf. **Figure 3.13**), were also too difficult to obtain.

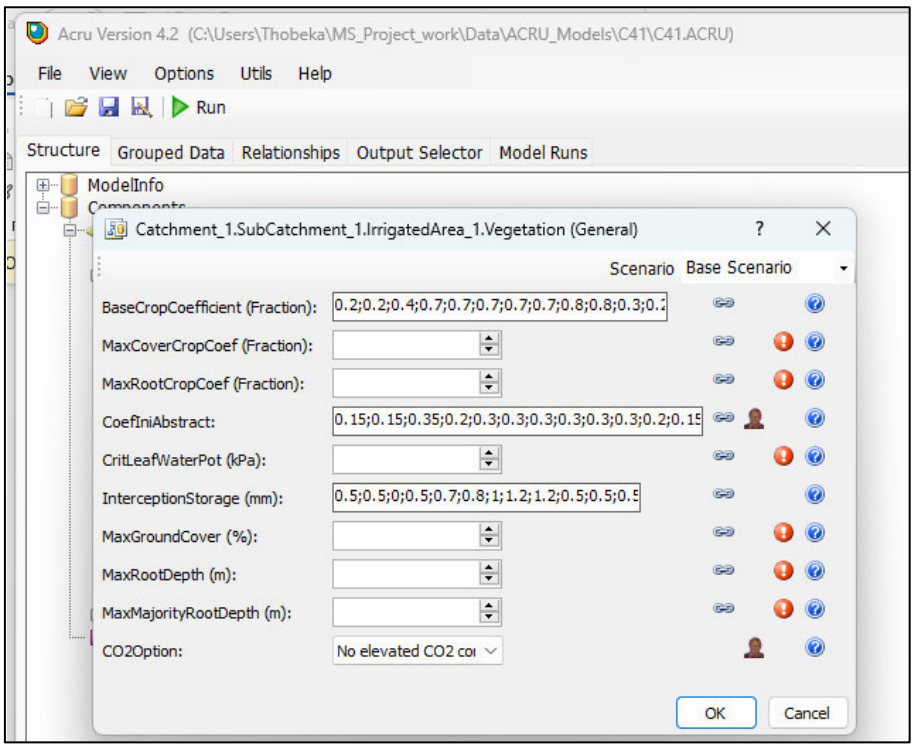


**Figure 3.13**: Information required by the *ACRU* model for an irrigated crop

#### 3.7.4.4 Catchment and streamflow data

For the tertiary catchment outflow, only quaternary catchments C41A to C41D were included, with the C41D exit node being used as the tertiary catchment exit node (i.e. C41A, C41B and C41C flow into C41D). The C41E quaternary catchment was excluded because the water balance for the catchment is provided by the dam inflow, which includes 3 inlets located in the upper quaternaries (cf. **Section 3.7.2**). Streamflow data comparison was conducted over a 10-year period from January 2010 to December 2019 at a monthly time step.

To assess the accuracy of the methods used for selecting a climate driver station, streamflow data simulated by *ACRU* was compared to inflow estimates by DWS for the Erfenis Dam. However, it is crucial to note that *ACRU* is best suited for modelling catchments smaller than 30 km² that contain small farm dams (Schulze, 1995). Modelling the dam located at the outlet of tertiary C41 is challenging due to its large size and the presence of 3 inlets. Insufficient data to accurately configure Erfenis Dam in *ACRU*, including information on water abstraction for irrigation and domestic use, further complicates modelling efforts. This challenge extends to the numerous farm dams in quaternaries C41A to C41D.

#### 3.7.4.5 Validation of *ACRU* model setup

The *ACRU* model was configured to simulate streamflow for the study catchment. Observed streamflow data were used to validate the *ACRU* model setup and verify the methods for selecting a climate driver station. This validation was conducted by comparing observed streamflow data to simulated data, as detailed in **Section 3.7.1.**

# 4 RESULTS AND DISCUSSION

This chapter provides the outcome of extending daily rainfall datasets obtained from Lynch (2004) using the *DRE* utility with datasets provided by 3 custodians, namely SAWS, ARC and DWS (**Section 4.1**). Additionally, the chapter provides results of rainfall driver stations selected using the existing DS approach and a new method developed in this study, namely the AF method. It also shows the outcome of the temperature stations selected using the temperature station ranking algorithm (**Section 2.6.2**).

ArcMap 10.8 (ESRI, 2020) was used to extract the Dent median, Lynch median, Lynch mean and Pegram mean gridded monthly and annual rainfall data for quaternary catchments C41A to C41E (from which the spatial average was calculated), as well as gridded data for the location of the selected driver stations. Values obtained from the gridded rainfall datasets were then compared to the station's observed values (**Section 4.3**).

Furthermore, *ACRU* outputs, including, *inter alia*, adjusted rainfall data, runoff, simulated streamflow and channel outflow) were used to verify the DS approach and the AF method for selecting a driver station. The performance of existing spatial rainfall datasets (Dent median, Lynch median, Lynch mean and Pegram mean) in improving the station's rainfall data to represent the catchment's spatial rainfall was also analysed (**Section 4.3**).

## 4.1 Climate data extension

### 4.1.1 Rainfall

The initial rainfall dataset for the Lejweleputswa District Municipality comprised 37 manual stations and 12 AWS, primarily sourced from SAWS. One manual station was obtained from DWS, while no ARC stations were included due to data quality issues. Despite the initial number, only 7 stations were ultimately used due to limited record lengths and high portions of missing or unreliable data. These constraints reflect systemic challenges in South Africa's rainfall monitoring infrastructure (Lynch, 2004; Erasmus, 2022).

To enable hydrological modelling, missing data were patched using the IDW method. However, this introduced additional uncertainty, especially for stations with large data gaps or few neighbouring stations. The exclusion of ARC data and the intensive nature of the infilling

process highlights the urgent need for improvements in data availability, quality, and infrastructure maintenance. **Table 4.1** lists the historical and updated station IDs for those used in this study.

**Table 4.1**: Station ID of the historical station and the station ID that was used to extend historical stations

| Historical station ID | Updated station ID |
|---|---|
| 0295760_W | 0295760_6 |
| 0295405_W | 0295405_7 |
| 0294500_W | 0294500_X |
| 0329166_W | 0329166_5 |
| 0294481_W | C4E002 |
| 0327883_W | 0327883_9 |
| 0261722_W | 0261722_8 |

As mentioned in **Section 3.3**, the daily climate data were requested from various custodians from January 1996 onwards. Rainfall data for the overlapping period (1996/01/01-1999/12/31) was used to ensure that existing datasets were correctly extended by an additional 20 years of daily data (up to 2019/12/31). This overlapping period was essential, particularly where the old and the new station IDs did not match.

**Figure 4.1** highlights the overlapping period, which was extended slightly to 31 August 2000, since the existing station (0261722_W) had data up to this date. The total rainfall recorded in the updated dataset is 1 276 mm higher than the historical dataset, with the updated data showing a cumulative total of 3 631 mm compared to 2 355 mm in the historical data. This discrepancy is primarily due to zero daily rainfalls in the historical dataset between 1999/05/01 and 2000/01/31, whereas the updated dataset obtained from SAWS provided actual (i.e. non-zero) rainfall measurements (**Figure 4.1**). Despite these differences, the correlation ($R^2$) between the datasets is 0.65, indicating a moderate positive relationship, suggesting that while the datasets generally follow similar trends, the historical data "gaps" affect the overall accuracy.

Several other stations also had the same problem where the historical ("old") and updated ("new") data differed for several periods. It was therefore decided to join the historical dataset from 1996 onwards to account for potential errors in the historical dataset. This finding represents an important contribution of the study, as it highlights a significant limitation in rainfall data obtained from the *DRE* utility, particularly for dates beyond 1999. Specifically, the presence of zero values in these records was found to often indicate missing data rather than actual rainfall observations, raising concerns about the reliability of the dataset for hydrological modelling and analysis. As such, this study underscores the need for thorough validation of rainfall datasets obtained from automated utilities, especially when extending historical data, to avoid incorporating erroneous or misleading values. This finding has broader implications for studies relying on similar datasets, emphasising the critical need for careful quality control and validation in hydrological research.

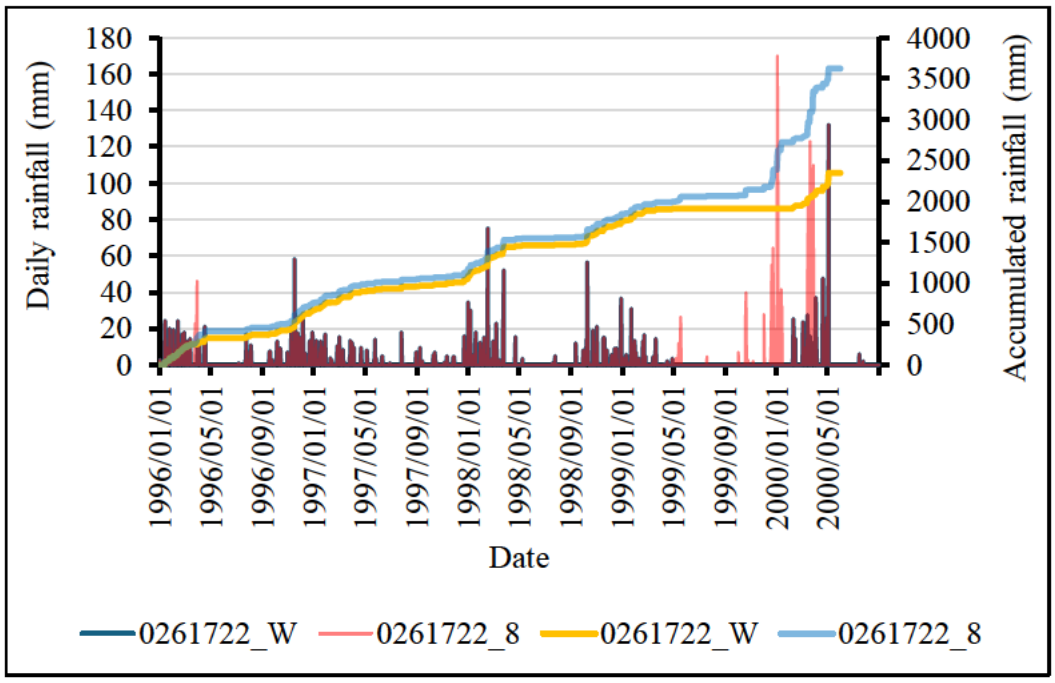


**Figure 4.1**: Overlap of daily and accumulated rainfall record for a period of 4 years and 8 months (1996/01/01 - 2000/08/31) for rain gauge 0261722_W (historical dataset) and 0261722_8 (updated dataset)

**Figure 4.2** shows the MAP of the historical dataset (1950-2000) and the MAP of the extended dataset (1950-2019) for 7 rainfall stations. There is no major difference between the MAP of these 2 records, with the highest being 23 mm for station 0294500_X. Hence, extending the record length of existing rainfall datasets by an additional 20 years slightly improved the accuracy of the station's MAP statistic.

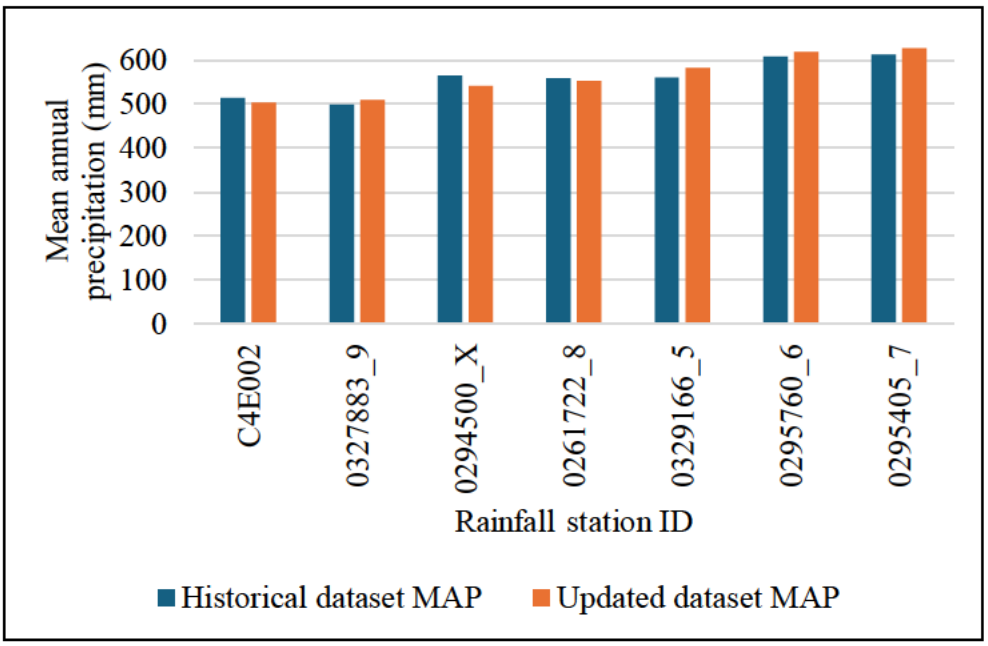


**Figure 4.2**: Mean annual precipitation (mm) of the historical dataset (1950-2000) and that of the updated dataset (1950-2019)

To assess how the length of a rainfall record influences the reliability of the MAP statistic, **Figure 4.3** presents the MAP calculated over increasing time intervals, ranging from 15 to 70 years in 5-year increments for station 0294500_X. The results show that using the longest available record length when estimating MAP is best practice. Although Lynch (2004) recommended a minimum record of 15 and 35 years for calculating MAP in wet and dry climates respectively, this was based on records ending in 2000/01. When only 35 years of data were used, the estimated MAP was 570 mm. However, when the record was extended to 70 years, the MAP decreased to 545 mm. From 55 to 70 years, the MAP steadily declined, indicating that recent decades have experienced drier conditions.

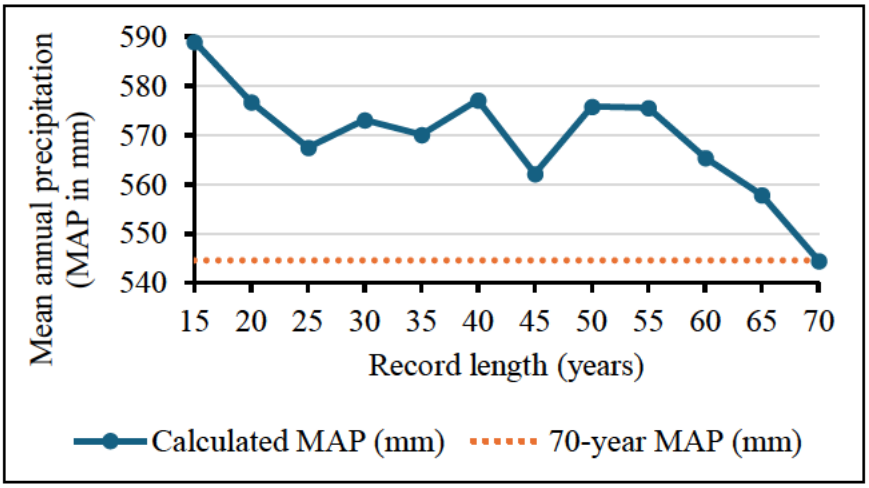


**Figure 4.3**: Mean annual rainfall (MAP in mm) for various record lengths from 15 to 70 years compared to the long-term (70-year) annual rainfall for station 0294500_X

### 4.1.2 Temperature

The Lejweleputswa DM had a total of 12 AWS from ARC and SAWS. **Table 4.2** shows the historical station ID (i.e. for the data developed by Schulze and Maharaj, 2004) as well as the station ID of the gauge that was used to extend the temperature data for the 3 temperature driver stations selected in this study. The missing data in these stations ranges from 0-34 %, the difficulty that came with patching the temperature data was mostly due to the sparse station network as 1 of the stations (i.e. 30595) had a high amount of missing data.

**Table 4.2**: Historical temperature station ID and updated station ID (i.e. stations used to extend the historical dataset)

| Historical station ID | Updated station ID |
|---|---|
| 0363792_A | 30595 |
| 0261516BW | 0261516B0 |
| 0365398_W | 0365398_8 |

To extend the temperature datasets from 1950-1999 all the missing datasets required infilling and the DSD method (cf. **Section 3.4.2.1**) and the ranking algorithm (cf. **Section 3.4.2.3**) were both used to infill and adjust the temperature dataset These 2 methods were compared as seen in **Table 4.3** where patching code P-1 indicated the data were infilled using the DSD method and P-R1 indicate the data were patched using the ranking algorithm. The DSD method and ranking algorithm show distinct adjustments to the observed data.

The DSD method generally produces lower adjusted values compared to the algorithm, which maintains values closer to the original observed data. This suggests that DSD method applies a stronger correction factor, possibly reducing potential overestimations but at the risk of underestimating actual temperatures. In contrast, the ranking algorithm provides more moderated adjustments, which could be beneficial for preserving the natural variability of temperature trends. Given these differences, the algorithm appears to be the preferable method as it maintains a closer alignment with observed values while still applying necessary corrections. However, further validation against independent temperature records would help confirm its accuracy and reliability.

**Table 4.3**: Comparison of methods to infill temperature data (i.e. DSD method and ranking algorithm) for a portion of 0291516B0 maximum temperature

| | **Difference in standard deviation method** | | | **Ranking algorithm** | | |
|---|---|---|---|---|---|---|
| Date | Observed/ *patched* data | Patching code | Adjusted data | Observed/*patched* data | Patching code | Adjusted data |
| 2013/03/01 | 28.40 | | | 28.40 | | |
| 2013/03/02 | 29.00 | | | 29.00 | | |
| 2013/03/03 | 31.50 | | | 31.50 | | |
| 2013/03/04 | 33.00 | | | 33.00 | | |
| 2013/03/05 | *36.36* | P-1 | 34.03 | *34.80* | P-R1 | 35.35 |
| 2013/03/06 | *34.02* | P-1 | 31.69 | *32.60* | P-R1 | 33.15 |
| 2013/03/07 | *32.50* | P-1 | 30.17 | *32.40* | P-R1 | 32.94 |
| 2013/03/08 | *34.15* | P-1 | 31.82 | *33.00* | P-R1 | 33.54 |
| 2013/03/09 | *32.80* | P-1 | 30.47 | *31.00* | P-R1 | 31.54 |
| 2013/03/10 | *30.85* | P-1 | 28.52 | *28.70* | P-R1 | 29.24 |
| 2013/03/11 | *29.73* | P-1 | 27.40 | *20.00* | P-R1 | 20.53 |
| 2013/03/12 | 29.90 | | | 29.90 | | |
| 2013/03/13 | 31.70 | | | 31.70 | | |
| 2013/03/14 | 29.10 | | | 29.10 | | |
| 2013/03/15 | 31.10 | | | 31.10 | | |

## 4.2 Representative driver station selection

Sections **4.2.1** and **4.2.2** provide a detailed outcome of the rainfall and temperature stations selected for the quaternary catchments C41A to C41E used in this study. It also compares observed vs gridded rainfall values.

### 4.2.1 Rainfall

Two methods were used to select driver stations for catchments C41A to C41E, namely the common DS approach and the AF method that was developed in this study (cf. **Section 3.5.1**).

#### 4.2.1.1 Driver station approach

The rain gauges selected using the DS approach were mainly based on the distance from their location to the quaternary catchment centroid. Therefore, 1 "driver" station was selected for each quaternary catchment. Furthermore, due to the low density of stations across the 5 quaternary catchments, C41B and C41C have the same rainfall driver station (**Table 4.4**). The record length was the same for all stations (i.e. 70 years), so this criterion could not be used to select the drier station for each quaternary catchment Additionally, **Table 4.5** presents the percentage of missing data that were patched for each station. The missing data percentage for the 7 stations ranges from 0.00 % to 2.84 %. While distance was the primary criterion, the reliability of each station's data were also taken into consideration so as to prioritise stations that required less data infilling.

**Table 4.4**: Rainfall driver stations selected for each quaternary catchment (C41A to C41E) using the driver station approach and the adjustment factor method

| Quaternary catchment | Driver station approach | Adjustment factor method | | | |
|---|---|---|---|---|---|
| | | Dent median | Lynch median | Lynch mean | Pegram mean |
| C41A | 0295760_6 | 0295405_7 | 0295760_6 | 0261722_8 | 0295405_7 |
| C41B | 0295405_W | 0295405_7 | 0295405_7 | 0261722_8 | 0295405_7 |
| C41C | 0295405_W | 0295760_6 | 0295405_7 | 0261722_8 | 0295405_7 |
| C41D | 0294500_X | 0329166_5 | 0329166_5 | C4E002 | 0261722_8 |
| C41E | C4E002 | 0294500_X | 0329166_5 | 0327883_9 | 0294500_X |

**Table 4.5**: Amount of missing data (%) for the 7 rainfall driver stations

| rainfall station ID | Infilled data (%) |
|---|---|
| 0261722_8 | 0.00 |
| C4E002 | 0.02 |
| 0329166_5 | 0.10 |
| 0294500_X | 0.27 |
| 0295405_7 | 0.29 |
| 0327883_9 | 0.65 |
| 0295760_6 | 2.84 |

#### 4.2.1.2 Adjustment factor method

Unlike the DS approach, the AF method allowed for the selection of different driver stations for each quaternary catchment based on the gridded rainfall dataset (i.e. Dent median, Lynch median, Lynch mean and Pegram mean) used to calculate the monthly adjustment factors. The rain gauge with monthly adjustment factors closest to 1 was selected (cf. **Table 7.7** in **Section 7.5**) as the driver station for each quaternary catchment (**Table 4.4**). Some stations are located outside the tertiary catchment C41 (e.g. stations 0261722_8, 0327883_9 and 0329166_5) since this method does not consider the distance from the station to the quaternary catchment centroid. The selected driver station's data requires the least adjustment to better represent the quaternary catchment spatially averaged rainfall. Hence, this new technique emphasises the representativity of the rainfall data more than proximity. Due to the scarcity of rainfall stations, some quaternary catchments also have the same driver station. In C41B, gauge 0295405_7 is the most selected station using 3 of the 4 gridded datasets.

The Pegram mean dataset performed better than the Dent and Lynch datasets for estimating the monthly adjustment factors. This shows that the Pegram dataset has good reliability and precision in improving the station's representativeness for the catchment. The median datasets were mainly affected by the *PPTCor* rules (cf. **Section 3.5.1.2**), especially in winter months. The *PPTCor* rules state that if the station rainfall is zero and the catchment rainfall is > 0, then the adjustment factor must be 2, which will result in the daily rainfall for that particular month being doubled.

The median grids are more likely to have zero rainfall during winter. Therefore, it is recommended that the mean grids should be used to calculate monthly rainfall adjustment factors instead of median grids. Compared to Lynch mean, Pegram mean shows a better performance, as show in **Section 4.3**.

### 4.2.2 Temperature

**Table 4.6** shows the "pseudo" temperature station that was selected for each rainfall driver station. Each temperature station was selected using the ranking algorithm method described in **Section 3.4.2.3**. This shows that of all the available reliable stations only 3 temperature stations met the algorithm's criteria for selection. In addition, all the selected temperature stations are situated outside of the tertiary catchment. Rainfall station [a]C4E002 is the only manual station monitored by DWS and [b]30595 is the only ARC AWS (cf. **Figure 3.1**).

**Table 4.6**: Pseudo temperature station selected for each rainfall driver station using the ranking algorithm

| Rainfall | Pseudo temperature |
|---|---|
| 0327883 | [b]30595 |
| [a]C4E002 | 0261516 |
| 0261722 | |
| 0294500 | |
| 0295405 | |
| 0295760 | 0365398 |
| 0329166 | |

The absence or scarcity of temperature stations within a catchment can result in models failing to adequately represent temperature dynamics within the catchment (Luo et al., 2011). In this study, the limited availability of AWSs that provide daily temperature data contributed to temperature data being unrepresentative of the quaternary catchment. This also negatively affected the estimation of reference evaporation from the temperature data. Furthermore, localised climatic variations such as microclimates can result in significant temperature differences over short distances.

## 4.3 Comparison of observed and gridded climate data

According to Fenta et al. (2018), relying only on station (i.e. point) data to assess rainfall fails to capture the spatial and temporal complexities due to inadequate station coverage. To overcome this limitation, gridded (i.e. interpolated) datasets incorporate various data sources to improve the accuracy and representation of precipitation patterns. This comparison aimed to evaluate the representativeness of the observed dataset relative to the gridded dataset, thereby assessing the accuracy of the interpolated grids. Three statistical metrics were used for comparison, *viz*. $R^2$, NSE and RMSE.

For each of the 7 rainfall stations, the long-term annual and monthly values calculated from the extended 70-year observed records were compared to values obtained from 3 gridded datasets derived by Dent et al. (1989), Lynch (2004) and Pegram et al. (2016). For simplicity, they are referred to as the Dent, Lynch and Pegram datasets. 3 statistical measures were used in this comparison (cf. **Section 3.6**). Duc and Sawada (2023) highlighted the importance of high NSE for showing good predictive performance. Similarly, Randrianasolo et al. (2010) discussed the significance of low RMSE values for ensuring precise rainfall predictions.

### 4.3.1 Mean annual rainfall

The 3 gridded MAP values were compared to observed MAP for all 7 stations and the statistics are shown in **Table 4.7**. All 3 interpolated datasets showed good agreement when compared to observed values. Dent had the lowest $R^2$ of 0.997, while Lynch and Pegram both had $R^2$ values of 0.999, indicating higher correlation between observed and gridded data. However, the Pegram dataset produced the highest NSE (0.858) and lowest RMSE (16.9 mm), which suggests the Pegram dataset best represented observations. This was expected since the Pegram dataset used data up to 2010, which is closer to the study's extended dataset (2019). In comparison, Dent and Lynch used data up to the 1980s and 2000/01, respectively. Based on these findings, it is recommended that the Pegram dataset be used to calculate the average MAP across all catchments, as it demonstrates the highest accuracy in representing observed data.

Additionally, to make a definitive recommendation, it is advisable to repeat this study in other catchments. Conducting similar analyses in different regions would help determine whether the Pegram dataset consistently outperforms the other datasets under varying climatic conditions, thereby strengthening its suitability for broader application.

**Table 4.7**: Statistical performance of 3 gridded mean annual rainfall datasets (Dent, Lynch and Pegram) in prediction of the observed MAP for the 7 driver rainfall stations

| Statistic | Dent | Lynch | Pegram |
|---|---|---|---|
| $R^2$ (n=7) | 0.997 | 0.999 | 0.999 |
| NSE | 0.453 | 0.791 | **0.858** |
| RMSE (mm) | 33.3 | 20.6 | **16.9** |

### 4.3.2 Mean monthly rainfall

Monthly rainfall from each gridded dataset was compared to observed rainfall for all 7 driver stations, as shown in **Table 4.8**. The Dent median, Lynch median, Lynch mean and Pegram mean rainfall datasets were evaluated for their effectiveness in estimating the station's observed values and allowing for the stations to be more representative of their respective catchments. Based on the RMSE statistic, the Pegram dataset performed slightly better than the Lynch dataset for estimating mean monthly rainfall.

For the estimation of median monthly rainfall, the evaluation was conducted using the sum of each statistical measure ($R^2$, NSE, and RMSE). The Dent median dataset achieved the highest overall sum for $R^2$ (6.899), NSE (6.762) and the lowest sum of RMSE (41.021 mm), making it the best-performing dataset for estimating median rainfall. These results indicate that the Dent median dataset provides the most accurate estimates for median-based rainfall analyses. For mean monthly rainfall estimation, the Pegram mean dataset had the highest sum of $R^2$ (6.967) and NSE (6.876), while also maintaining a relatively low RMSE sum (29.079 mm). This suggests that the Pegram mean dataset is the most reliable choice for mean rainfall estimation.

Overall, based on the sum of each statistic, the Pegram mean dataset outperformed the others in terms of combined accuracy across multiple measures. It demonstrated strong predictive capability, making it the most suitable dataset for general rainfall estimation.

Regarding the calculation of rainfall adjustment factors, it is recommended that the Pegram mean dataset be used as the reference grid. Currently, adjustment factors for quinary and quaternary catchments are based on the Lynch median dataset. Given the superior performance of the Pegram mean dataset, these adjustment factors should be recalculated to reflect a more accurate representation of rainfall distribution.

While the findings of this study provide strong evidence supporting the use of the Pegram mean dataset for rainfall estimation, it is advisable to repeat this analysis in other catchments. Local climatic and topographic variations may influence the performance of different datasets and further validation across diverse hydrological regions would ensure a more definitive recommendation for widespread application. The Dent, Lynch and Pegram maps may vary in their ability to adequately capture the spatial distribution of rainfall in different regions. Pitman and Bailey (2021) stated that about 70 % of quaternary catchments have less than a 10 % difference in Dent and Pegram MAP, except for some catchments, particularly in mountainous and remote regions. Schulze et al. (1995) also mentioned that it is incorrect to assume that a rain gauge provides actual rainfall at a site. Point rainfall amounts are, therefore, only estimates of actual rainfall due to catch deficiencies caused by the aerodynamic interaction of, *inter alia*, rainfall, wind, the rain gauge itself and local or regional topography.

**Table 4.8**: Statistical performance of mean and median monthly gridded rainfall (Dent median, Lynch median, Lynch mean and Pegram mean) for predicting observed values obtained from 7 climate stations

| Gridded dataset | Statistic | 0294500_X | C4E002 | 0295405_7 | 0295760_6 | 0291722_8 | 0329166_5 | 0327883_9 |
|---|---|---|---|---|---|---|---|---|
| **Dent median** | $R^2$ (n=12) | 0.985 | 0.996 | 0.998 | 0.999 | 0.992 | 0.989 | 0.941 |
| | NSE | 0.950 | 0.988 | 0.994 | 0.966 | 0.951 | 0.973 | 0.940 |
| | RMSE (mm) | 6.415 | 2.902 | 2.573 | 5.956 | 6.510 | 4.826 | 6.839 |
| **Lynch Median** | $R^2$ (n=12) | 0.988 | 0.982 | 0.993 | 0.993 | 0.990 | 0.985 | 0.936 |
| | NSE | 0.968 | 0.952 | 0.975 | 0.982 | 0.974 | 0.974 | 0.936 |
| | RMSE (mm) | 5.094 | 5.888 | 5.280 | 4.378 | 4.701 | 4.689 | 7.051 |
| **Lynch mean** | $R^2$ (n=12) | 0.997 | 0.994 | 0.995 | 0.997 | 0.995 | 0.996 | 0.992 |
| | NSE | 0.988 | 0.965 | 0.971 | 0.982 | 0.979 | 0.988 | 0.966 |
| | RMSE (mm) | 3.164 | 5.210 | 5.789 | 4.450 | 4.318 | 3.479 | 5.276 |
| **Pegram mean** | $R^2$ (n=12) | 0.997 | 0.993 | 0.997 | 0.996 | 0.993 | 0.995 | 0.996 |
| | NSE | 0.987 | 0.972 | 0.991 | 0.983 | 0.974 | 0.984 | 0.985 |
| | RMSE (mm) | 3.411 | 4.849 | 3.376 | 4.532 | 5.065 | 4.172 | 3.676 |

## 4.4 Comparison of driver station to quaternary catchment rainfall

### 4.4.1 Monthly rainfall

In **Table 4.9**, monthly mean and median datasets of gridded rainfall for each quaternary catchment were compared to monthly mean and median dataset of observed rainfall obtained from the driver station selected for each quaternary catchment. The driver station data were selected using the DS approach and AF method. Daily rainfall data from each driver station were adjusted using monthly adjustment factors derived from the 4 gridded datasets (i.e. Dent median, Lynch median, Lynch mean and Pegram mean). The grid that produced the lowest RMSE statistic is highlighted in bold in the table.

In comparison to other rainfall grids, while the Dent median grids produced high NSE values in other catchments, it generally has higher RMSE values. The Lynch median and Lynch mean datasets performed well in catchments such as C41A and C41E, respectively, but they are outperformed by the Pegram dataset in other catchments for both methods. This outcome underscores Pegram dataset robustness and reliability in accurately reflecting average rainfall conditions across different catchments, thereby improving hydrological assessments in these regions. These catchments might be influenced by more localised precipitation patterns, variations in land cover, or differences in hydrological response that are better captured by other datasets such as Lynch or Dent. Additionally, Pegram dataset grid resolution and interpolation techniques may be less suited for capturing fine-scale variations in these specific regions.

To conclusively determine whether driver station selection errors contributed to these results, further validation of station placement is required. This would involve assessing whether the selected stations accurately represent the spatial and temporal characteristics of rainfall and runoff in C41A and C41E. If discrepancies are found, then adjustments in station selection may be necessary. However, if the stations are correctly placed, the explanation likely lies in the hydrological and dataset-specific factors rather than an outright methodological error.

**Table 4.9**: Statistical measures ($R^2$, NSE and RMSE) for observed monthly rainfall from the driver station selected for each quaternary catchment monthly versus gridded quaternary catchment rainfall derived from datasets developed by Dent et al. (1989), Lynch (2004) and Pegram et al. (2016)

| | | **Driver station approach** | | | | **Adjustment factor method** | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Quaternary catchment** | **Statistic** | **Median grids** | | **Mean grids** | | **Median grids** | | **Mean grids** | |
| | | **Dent** | **Lynch** | **Lynch** | **Pegram** | **Dent** | **Lynch** | **Lynch** | **Pegram** |
| C41A | $R^2$ (n=12) | **0.996** | 0.990 | 0.984 | 0.995 | 0.989 | **0.997** | 0.986 | 0.996 |
| | NSE | **0.988** | 0.988 | 0.974 | 0.984 | 0.982 | **0.988** | 0.939 | 0.987 |
| | RMSE (mm) | **3.6** | 3.6 | 5.3 | 4.1 | 4.5 | **3.6** | 7.4 | 3.8 |
| C41B | $R^2$ (n=12) | 0.997 | 0.983 | 0.981 | **0.997** | 0.991 | 0.994 | 0.988 | **0.997** |
| | NSE | 0.984 | 0.974 | 0.966 | **0.989** | 0.984 | 0.974 | 0.946 | **0.989** |
| | RMSE (mm) | 4.2 | 5.4 | 6.3 | **3.5** | 4.2 | 5.4 | 7.0 | **3.5** |
| C41C | $R^2$ (n=12) | 0.992 | 0.969 | 0.975 | **0.995** | 0.989 | 0.989 | 0.989 | **0.996** |
| | NSE | 0.967 | 0.945 | 0.940 | **0.984** | 0.983 | 0.945 | 0.962 | **0.985** |
| | RMSE (mm) | 6.1 | 7.9 | 8.3 | **4.3** | 4.4 | 7.9 | 5.8 | **4.1** |
| C41D | $R^2$ (n=12) | 0.978 | 0.948 | 0.975 | **0.995** | 0.963 | 0.983 | 0.991 | **0.992** |
| | NSE | 0.933 | 0.948 | 0.968 | **0.984** | 0.963 | 0.945 | 0.968 | **0.969** |
| | RMSE (mm) | 7.4 | 6.5 | 5.2 | **3.7** | 5.7 | 6.7 | 5.3 | **5.0** |
| C41E | $R^2$ (n=12) | 0.995 | 0.978 | 0.980 | 0.989 | 0.963 | 0.985 | **0.994** | 0.994 |
| | NSE | 0.975 | 0.957 | **0.980** | 0.964 | 0.963 | 0.955 | **0.980** | 0.977 |
| | RMSE (mm) | 4.2 | 5.5 | **3.9** | 5.3 | 5.5 | 6.3 | **4.0** | 4.4 |

### 4.4.2 Daily rainfall

The water balance data were checked using **Equation 3.6** and the calculated inflow did not match the values given in **Table 3.10**. Further quality checking was done by comparing rainfall and reference evaporation data from the C4E002 station, with values listed in the table above. The C4E002 rainfall station is located near the dam and maintained by the DWS. Evaporation data were estimated using the Hargreaves and Samani (1985) equation, with temperature data obtained from a nearby AWS (0261516B0), which was 45.87 km away from the rain gauge.

The performance of the median dataset was assessed using the sum of $R^2$, NSE and RMSE values across the datasets. The Dent median generally performs better in both datasets, as it has higher NSE and $R^2$ values, which indicates a stronger correlation and predictive accuracy. Additionally, it has a lower RMSE (i.e. sum of 24.3 m) in both methods, meaning it produces a fewer errors. The Lynch median dataset had a slightly lower $R^2$ and NSE scores and higher RMSE, suggesting that it may introduce more variability and larger deviations from observed values.

The Pegram mean dataset was shown to have the highest performance compared to the Dent median, Lynch median and Lynch mean datasets. Adjustment factors calculated using mean monthly grids are better compared to median monthly grids. Therefore, the *ACRU* outputs (e.g. adjusted rainfall and streamflow) and statistics (e.g. NSE and RMSE) computed using the Pegram mean dataset were used to verify and choose the best method to select a driver station. These 2 methods used to select a climate driver station in the study are the widely used DS approach and the newly developed AF method.

**Table 4.10** shows the comparison of the observed daily data and the adjusted daily rainfall data for The DS approach and the AF method were compared using the observed (unadjusted) and adjusted daily rainfall data for each quaternary catchment (C41A to C41E) to determine the best climate driver station selection technique. The comparison between the new AF method against the existing DS approach using the Pegram dataset reveals notable differences in performance across various catchments, particularly in terms of NSE and RMSE.

In catchment C41A, the AF method achieved an NSE of 0.995 and RMSE of 0.445 for the Pegram dataset, slightly outperforming the DS approach with an NSE of 0.93 and an RMSE of 0.498 mm. For catchment C41B, the AF method has a lower RMSE of 0.446. In catchments C41C and C41D, the DS approach shows better performance (lower RMSE) compared to the AF method. Lastly, for C41E, the AF method has a higher NSE and a lower RMSE, indicating good performance. Overall, the AF method demonstrates a better performance in 3 out of 5 catchments (C41A, C41B and C41E), with higher NSE values and lower RMSE values, indicating is it better in selecting driver stations along with the Pegram rainfall dataset.

**Table 4.10**: Comparison of statistical metrics for observed and adjusted daily rainfall (1950/01/01-2019/12/31) for driver stations selected using DS approach and AF method for each quaternary catchment using different spatial datasets (Lynch mean, Pegram mean, Dent median, Lynch median)

| | | Driver station approach | | | | Adjustment factor method | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Quaternary catchment** | **Statistic** | **Dent median** | **Lynch median** | **Lynch mean** | **Pegram mean** | **Dent median** | **Lynch median** | **Lynch mean** | **Pegram mean** |
| **C41A** | $R^2$ (n=25566) | 0.978 | 0.983 | 0.992 | **0.994** | 0.990 | 0.982 | 0.987 | **0.995** |
| | NSE | 0.978 | 0.982 | 0.987 | **0.993** | 0.987 | 0.980 | 0.992 | **0.995** |
| | RMSE (mm) | 0.912 | 0.837 | 0.715 | **0.498** | 0.660 | 0.832 | 0.766 | **0.445** |
| **C41B** | $R^2$ (n=25566) | 0.986 | 0.975 | 0.993 | **0.995** | 0.987 | 0.975 | 0.987 | **0.995** |
| | NSE | 0.985 | 0.972 | 0.983 | **0.994** | 0.985 | 0.972 | 0.982 | **0.994** |
| | RMSE (mm) | 0.730 | 0.989 | 0.779 | **0.453** | 0.730 | 0.989 | 0.765 | **0.446** |
| **C41C** | $R^2$ (n=25566) | 0.973 | 0.974 | 0.994 | **0.997** | 0.982 | 0.974 | 0.990 | **0.996** |
| | NSE | 0.972 | 0.970 | 0.986 | **0.996** | 0.981 | 0.970 | 1.088 | **0.994** |
| | RMSE (mm) | 0.982 | 1.022 | 0.703 | **0.357** | 0.860 | 1.022 | 0.614 | **0.447** |
| **C41D** | $R^2$ (n=25566) | 0.964 | 0.961 | 0.990 | **0.994** | 0.972 | 0.961 | **0.989** | 0.988 |
| | NSE | 0.944 | 0.952 | 0.989 | **0.993** | 0.968 | 0.958 | **0.986** | 0.987 |
| | RMSE (mm) | 1.362 | 1.264 | 0.602 | **0.465** | 0.952 | 1.093 | **0.605** | 0.639 |
| **C41E** | $R^2$ (n=25566) | 0.970 | 0.965 | 0.992 | 0.994 | 0.978 | 0.973 | 0.994 | 0.992 |
| | NSE | 0.947 | 0.938 | **0.991** | 0.982 | 0.977 | 0.970 | 0.977 | **0.992** |
| | RMSE (mm) | 1.161 | 1.256 | **0.469** | 0.673 | 0.878 | 0.920 | 0.878 | **0.526** |

## 4.5 *ACRU* simulations of streamflow

Hydrological simulation modelling plays a critical role in verifying the effectiveness of climate driver station selection methods. In this study, the *ACRU* was used to assess the performance of 2 methods for selecting a driver station (i.e. DS approach and AF method). These techniques were evaluated based on their ability to provide reliable climate data inputs for hydrological modelling. The verification process involved comparing *ACRU*-simulated streamflow, generated using driver stations selected by both methods, against observed inflow data from the Erfenis Dam. The assessment aimed to determine which method more accurately captured streamflow variability and provided representative rainfall inputs for the catchment.

### 4.5.1 Adjustment factors based on mean vs median rainfall data

While challenges such as data scarcity, missing climate records and complexities in streamflow simulation influenced the verification outcomes, this section presents the results of hydrological modelling and discusses the implications of different driver station selection approaches in ensuring accurate climate representation for hydrological applications.

A comparison was made of simulated streamflow derived from adjustment factors calculated using the 4 grids (i.e. Dent median, Lynch median, Lynch mean and Pegram mean) The comparison was done for catchment C41D and the driver stations being 0329166_5 (Dent and Lynch median), C4E002 (Lynch mean) and 0261722_8 (Pegram mean). The comparison was done to highlight the difference in streamflow simulated during the winter months, where the median grids typically produce adjustment factors of 2 due to 0 values (cf. **Section 3.5.1.2**). From **Figure 4.4**, the median (Dent and Lynch) adjustment factors produced higher streamflow in July 2016 compared to streamflow produced using the mean (Lynch and Pegram) adjustment factors.

Adjustment factors based on median rainfall result in an over-estimation of water availability during the winter months. This discrepancy between mean and median adjustment factors highlights the need for careful consideration which grid to use for estimating monthly adjustment factors. At present, the monthly adjustment factors utilised in the quaternary catchment and quinary climate databases are derived from the Lynch median grid. However, reliance on median-based adjustment factors has been shown to overestimate water availability

during the winter months, primarily due to the influence of zero values, which result in an adjustment factor of 2. This discrepancy between adjustment factors calculated using mean and median values underscores the necessity for a reassessment of the methodology employed in estimating monthly adjustment factors. To enhance the accuracy of hydrological modelling and water resource assessments, it is recommended that the Pegram mean grid be used in place of the Lynch median grid. The use of mean-based adjustment factors, particularly from the Pegram mean grid, provides a more representative estimate of rainfall variability, reducing the risk of overestimation during drier months and improving the reliability of simulated streamflow.

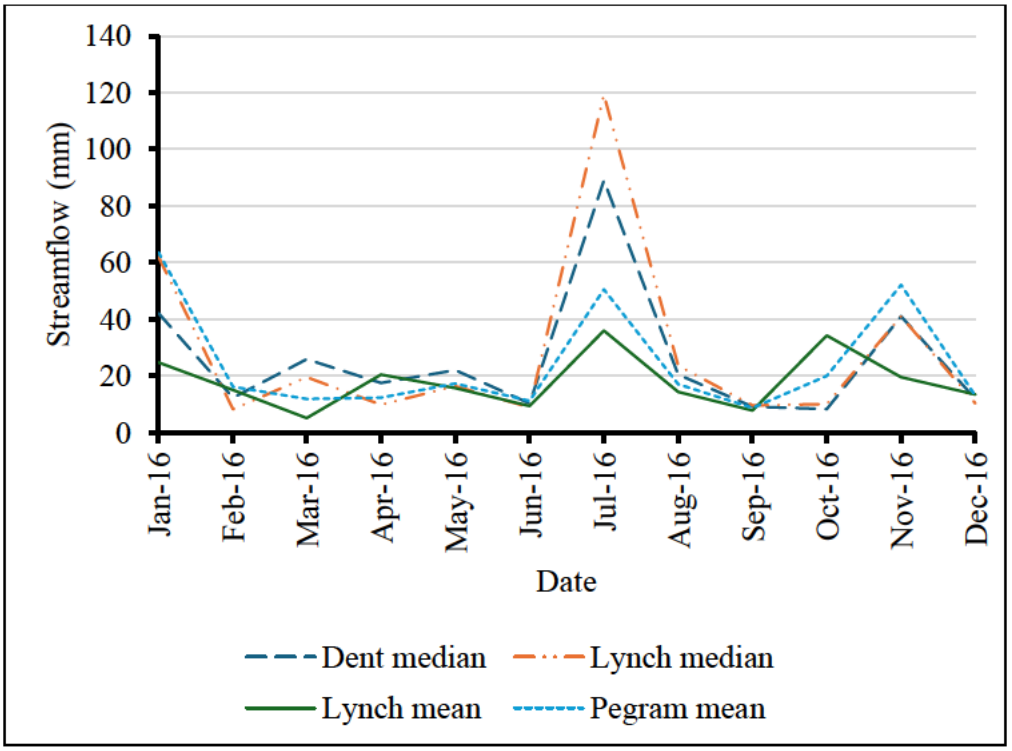


**Figure 4.4**: *ACRU* simulated streamflow using driver stations selected with AF method and adjusted with factors determined by the Dent median, Lynch median, Lynch mean and Pegram mean datasets

### 4.5.2 Verification of driver station selection methods

The monthly inflow estimated by DWS for the Erfenis Dam was compared to the inflow simulated by the *ACRU* model. It must be noted that only C41A to C41D were included in the comparison. The C41E quaternary catchment was not included, as explained in **Section 3.7.4**. The *ACRU* simulated inflow was obtained using driver stations selected by the DS approach

and the AF method. The rainfall adjustment factors were determined using the Pegram mean dataset, based on evidence provided in the previous section (**Section 4.3**).

In the AF method, station 0261722_8 was used in 3 catchments (C41A, C41B and C41C) under the Lynch mean grid, while station 0295405_7 was also used in 3 catchments (C41A, C41B, and C41C) under the Pegram mean grid. In contrast, the DS approach relied on station 0295405_7 for only 2 catchments (C41B and C41C), showing a reduced dependence on this station compared to the AF method (cf. **Table 4.11**). Station selection was influenced by the availability of reliable climate data. In this study, some stations were assigned to multiple catchments due to limited station availability. This means selection was based not only on hydrological representativity but also on data reliability. When monitoring stations are scarce, capturing spatial variability in climate and hydrology becomes more challenging, reducing the ability to fully represent all catchments

**Table 4.11**: Summary of driver stations selected for C41A-C41D use the DS approach and AF method

| Quaternary catchment | Driver station approach | Adjustment factor method | | | |
|---|---|---|---|---|---|
| | | Dent median | Lynch median | Lynch mean | Pegram mean |
| C41A | 0295760_6 | 0295405_7 | 0295760_6 | 0261722_8 | 0295405_7 |
| C41B | 0295405_W | 0295405_7 | 0295405_7 | 0261722_8 | 0295405_7 |
| C41C | 0295405_W | 0295760_6 | 0295405_7 | 0261722_8 | 0295405_7 |
| C41D | 0294500_X | 0329166_5 | 0329166_5 | C4E002 | 0261722_8 |

**Figure 4.5** shows that *ACRU* was better at simulating higher inflow than lower inflows. For example, February 1967 had the highest observed inflow in **Figure 4.5**, reaching 350 $Mm^3$and the AF method simulated this flow better with an inflow of 326 $Mm^3$ compared to DS approach with an inflow of 438 $Mm^3$. The AF method showed better agreement with the observed data, making it potentially more reliable for streamflow predictions than the DS approach.

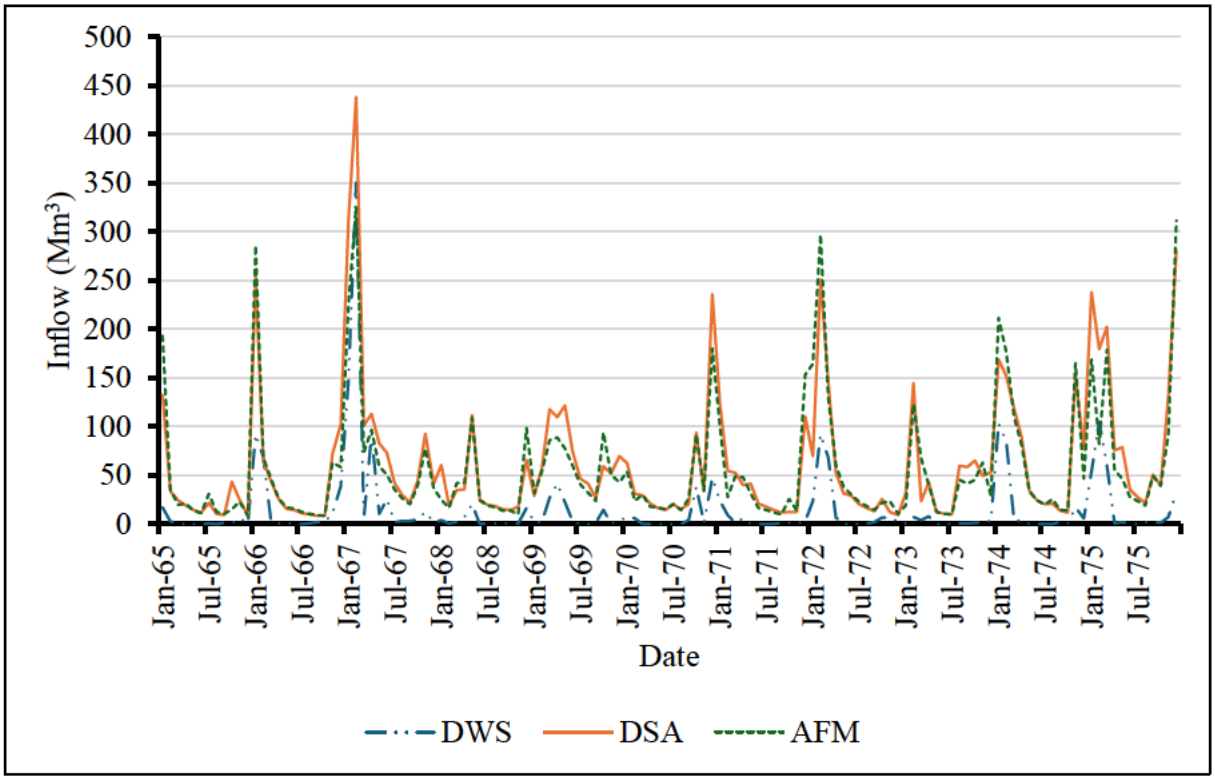


**Figure 4.5**: Comparison of Erfenis Dam inflow with *ACRU* simulated inflow using DS approach and AF method selected driver stations for quaternary catchments C41A-C41D

Several factors may contribute to this uncertainty. Configuring *ACRU* for complex quaternary catchments, where irrigation is a major land use and numerous small farm dams exist, presents significant challenges. These complexities may introduce inaccuracies in the simulated inflows. Additionally, the Erfenis Dam inflow estimated by DWS may be problematic, as discussed in the following section.

### 4.5.3 Erfenis Dam water balance

**Figure 4.6** shows the comparison of monthly rainfall from the Erfenis Dam water balance and rainfall from station C4E002. The DWS under-estimates rainfall input to the Erfenis Dam.

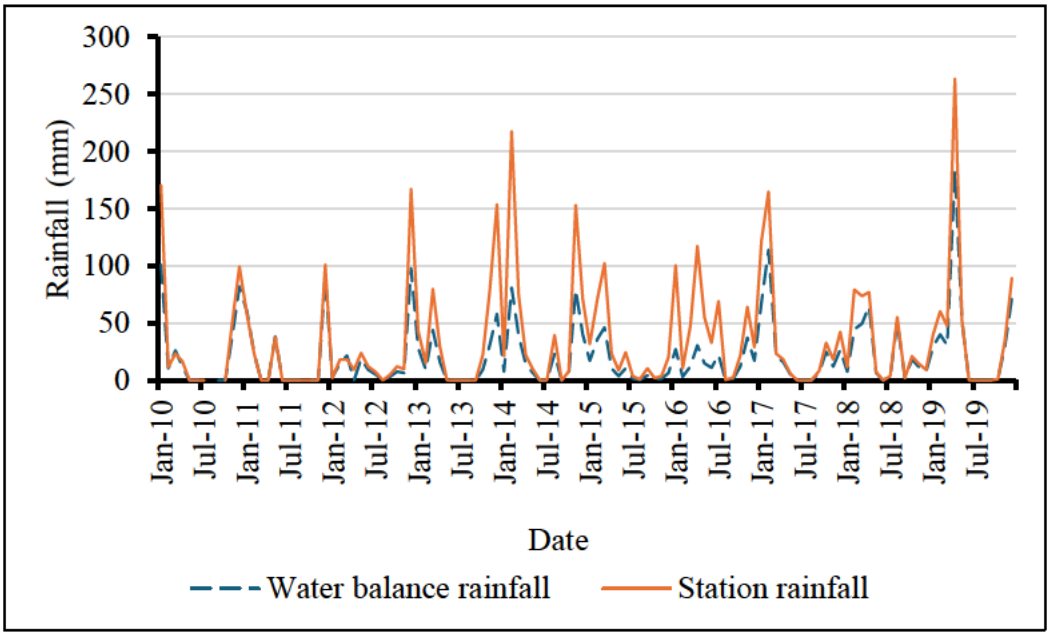


**Figure 4.6**: Comparison between the water balance rainfall and the monthly rainfall for station C4E002

**Figure 4.7** shows the comparison of the A-pan equivalent evaporation estimated in *ACRU* using the Hargreaves and Samani (1985) equation to the evaporation values used by DWS in the Erfenis Dam water balance. However, a coefficient of 0.65 was applied to the *ACRU* simulations of A-pan evaporation to estimate open water (i.e. dam) evaporation. DWS under-estimates evaporation loss from the Erfenis Dam.

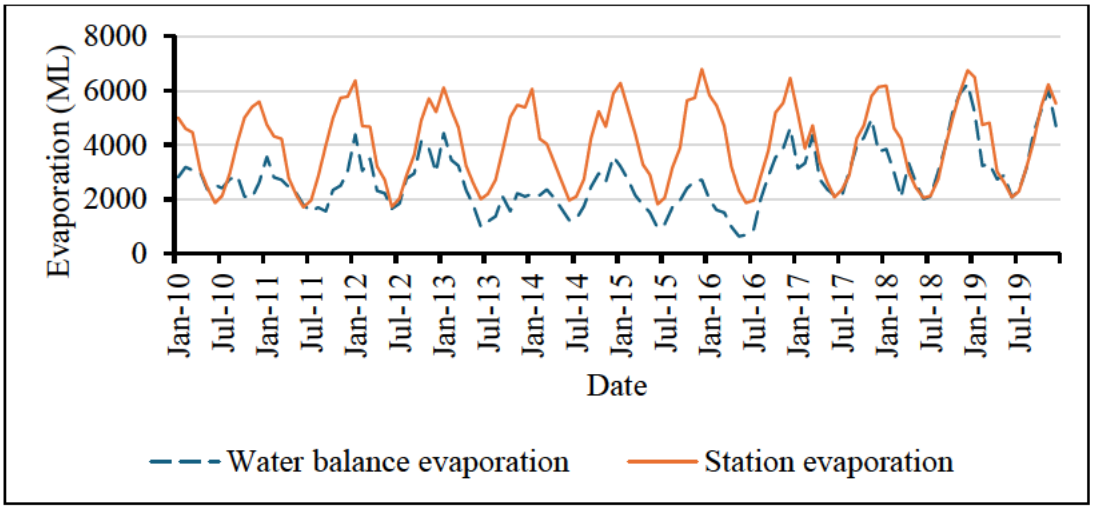


**Figure 4.7**: Comparison of C4E002 station *ACRU* simulated monthly gross evaporation (ML) and the DWS Erfenis Dam water balance gross evaporation (ML) for the last 10 years of the data (January 2010 to December 2019)

### 4.5.4 Streamflow frequency analysis

The comparison of annual streamflow exceedance probabilities is illustrated in **Figure 4.8** for the DS approach and **Figure 4.9** for the AF method. The base run serves as a reference for this analysis, utilising 50 years of data (1950-1999). However, this period is too short to generate the 5$^{th}$ and 95$^{th}$ percentile values, which are critical for understanding extreme events.

Streamflow data for the base run was simulated by Clulow et al. (2023) using the QnCDB (cf. **Section 2.6.2.1**), with rainfall adjustment based on the Lynch median. The inability of the base run to estimate these percentiles highlights the necessity of longer datasets. Razavi and Coulibaly (2016) emphasise that capturing low-probability events requires extended data periods, as these extreme occurrences are statistically uncommon. Similarly, Serinaldi and Kilsby (2014) assert that datasets exceeding 50 years offer more reliable estimates of extreme values, reducing susceptibility to random variability.

The frequency analysis for both the DS approach and the AF method was conducted on *ACRU*-simulated streamflow to estimate the percentage of exceedance for both driver station selection methods and the different gridded datasets used to calculate the rainfall adjustment factors. This comparison aimed to demonstrate the benefit of extending the model's input data by an additional 20 years while also considering potential drawbacks such as increased computational costs and the reliability of older datasets. Ensuring data consistency and accuracy over extended periods remains a challenge, particularly when integrating different sources with varying resolutions and methodologies.

The comparison between the DS approach and the AF method in streamflow frequency analysis revealed differences that varied with exceedance probability. The DS approach exhibited greater variability relative to the base run, particularly at lower exceedance probabilities, suggesting higher uncertainty in flow projections. However, deviations were smaller at higher exceedance probabilities. In contrast, the AF method produced less variation in streamflow, indicating a more stable and consistent representation of hydrological behaviour. This stability, however, may also reduce sensitivity to localised hydrological variations, which could be a limitation in cases where finer-scale changes are important. Additionally, the AF method relies on adjustment factors derived from external datasets, meaning its accuracy depends on the quality and representativeness of those datasets.

The Pegram dataset, used in the AF method, produced streamflow simulations that closely matched the base run, with a 50th percentile flow of 176.47 mm compared to the base run's 166.89 mm. This suggests that the Pegram dataset provides a reliable adjustment mechanism for rainfall inputs, minimising deviations from observed conditions. A dataset containing more than 50 years is essential for conducting frequency analysis, as it improves estimates of extreme hydrological events by facilitating the calculation of the 5th and 95th percentiles. The extended dataset (70 years) enhanced the robustness of the analysis by allowing for a more accurate characterisation of these conditions.

However, it is also important to consider whether there is an optimal dataset length beyond which additional years provide diminishing returns, as longer datasets may introduce inconsistencies due to changes in climate patterns, land use and data collection methods over time. Serinaldi and Kilsby (2014) noted that datasets longer than 50 years provide more reliable percentile estimates, reducing the influence of random variability. For direct comparison, the 10th (high flows), 50th (medium flows) and 90th (low flows) percentiles were used, representing probabilities of exceedance of 10 %, 50 % and 90 %, respectively.

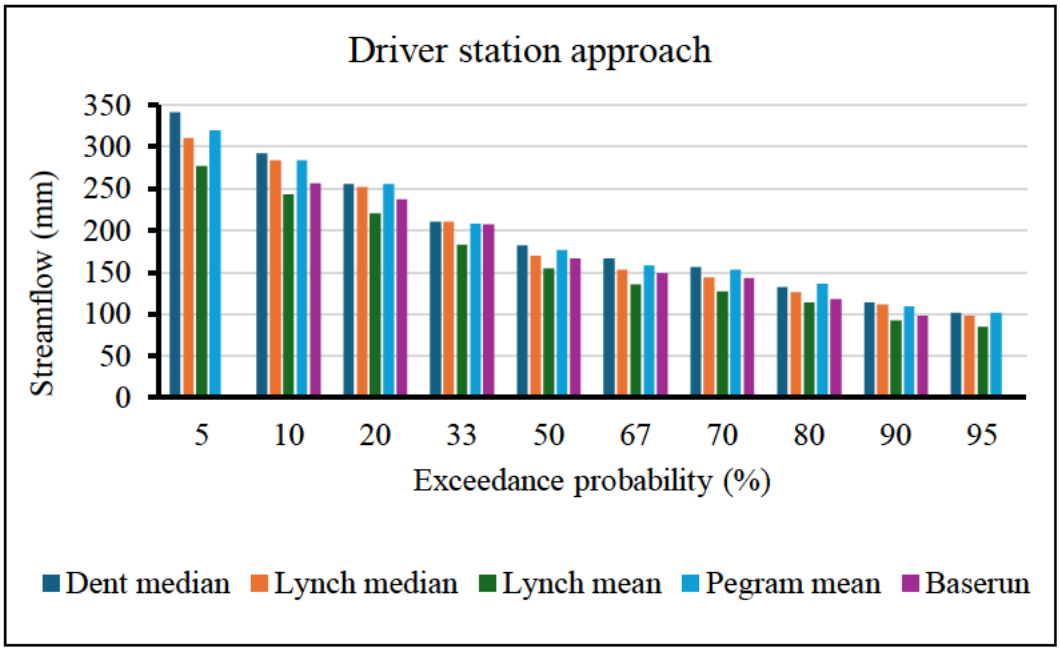


**Figure 4.8**: Comparison of annual streamflow exceedance probability using rainfall from stations selected using the driver station approach and adjustment factors calculated from the Dent median, Lynch median, Lynch mean and Pegram mean datasets

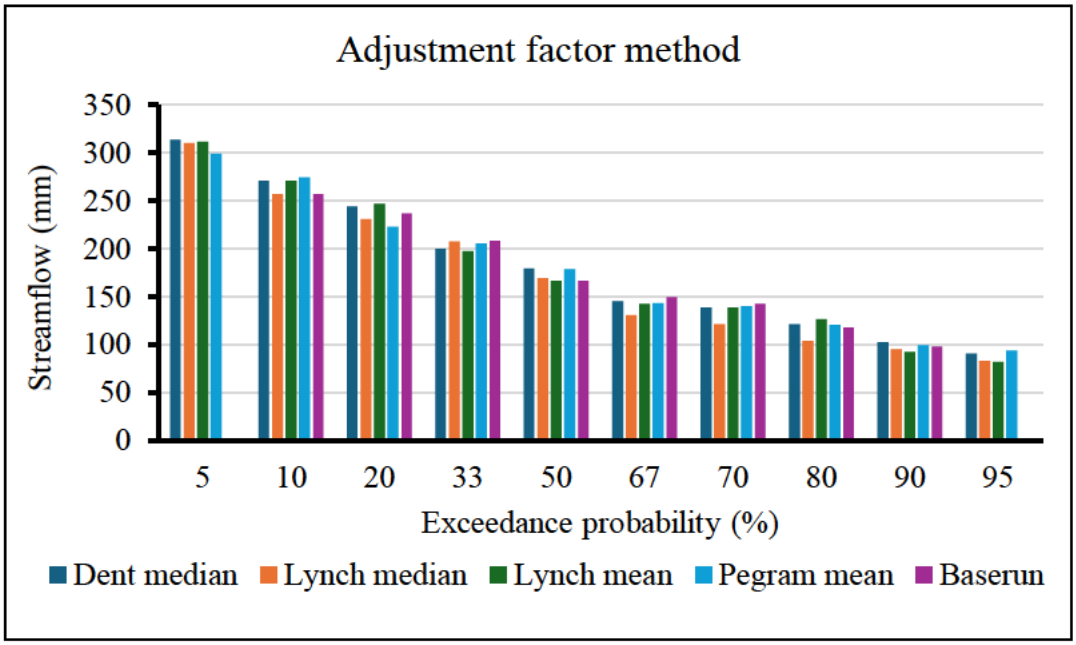


**Figure 4.9**: Comparison of annual streamflow exceedance probability using rainfall from stations selected using the adjustment factor method and adjustment factors calculated from the Dent median, Lynch median, Lynch mean and Pegram mean datasets

Since there is no clear evidence to determine which method is better, further research is recommended to assess their performance in different catchments with more available climate and streamflow data. Additionally, testing both methods in regions with well-monitored hydrological conditions could help refine their application. Future studies should also explore the sensitivity of each method to variations in rainfall and temperature datasets to better understand their impact on streamflow simulations.

# 5 CONCLUSIONS AND RECOMMENDATIONS

This chapter provides a summary of the approaches used to achieve the aims and objectives of this study. A summary of the main findings and recommendations for future research are included in this chapter.

## 5.1 Summary of approach

The study aimed to evaluate and assess techniques for selecting a climate driver station for a study catchment. Daily climate data were sourced from SAWS, ARC and DWS, which included rainfall and temperature data for the period of 01 January 1996 to 31 December 2019. Daily streamflow data were also sourced for the period of 01 January 1950 to 31 December 2019.

Missing rainfall data were infilled using IDW and missing temperature data were infilled using the DSD method and the ranking algorithm (where the DSD method could not be applied). The updated daily rainfall dataset was then used to extend existing Lynch (2004) datasets and the updated daily temperature data were used to extend existing Schulze and Maharaj (2004) datasets. The daily reference evaporation from 1950-2019 was estimated in *ACRU* using the extended temperature data as an input.

The selection of the study catchment (i.e. tertiary C41) was mainly based on the availability of reliable and continuous climate stations within and near the catchment's boundary. It also included the catchment having streamflow data at the outlet of the tertiary catchment. Representative rainfall driver stations were selected using the DS approach and AF method and for temperature, they were selected using the ranking algorithm.

The *ACRU* hydrological model was used for simulating streamflow in the study catchment. The driver stations selected using the 2 methods were used for climate data input in *ACRU*. To ensure the driver stations' climate data were representative of their quaternary catchments, 4 sets of monthly rainfall adjustment factors were determined using existing gridded datasets (i.e. Dent median, Lynch median, Lynch mean and Pegram mean). The temperature driver stations were selected using the ranking algorithm for each rainfall station. For temperature, station data were adjusted for the difference in elevation between the quaternary catchment and the driver station, which was done before the data were input in *ACRU*.

To identify the best rainfall dataset to estimate the adjustment factors, the gridded dataset was compared to the observed dataset at different timesteps (i.e. annual, monthly and daily) using statistical performance metrics, including $R^2$, NSE and RMSE, to assess the accuracy and reliability of the rainfall estimates.

The selected study catchment included a large dam used primarily for irrigation and domestic purposes, but inadequate water abstraction information was available. A dam water balance was obtained from DWS, which provided calculated monthly inflows to the dam. *ACRU* simulations of streamflow into the dam from upstream quaternaries (C41A to C41D) were then compared to the dam inflows obtained from DWS. This was done to evaluate the different techniques that were tested for selecting a climate driver station. Lastly, a frequency analysis for streamflow was done to emphasise the importance of having more than 50 years of data.

## 5.2 Summary of findings

A total of 223 manual and AWS stations were received from the 3 data custodians. However, many of these stations had significant missing data, making them unreliable for use. Additionally, rainfall data obtained from the *DRE* utility beyond 1999 was found to be unreliable due to several months consistently recording zero rainfall. The chosen study catchment had no temperature stations within its boundary and the number of stations with reliable data were low. As a result, some stations were selected for multiple rainfall stations each, for instance AWS 0291516B0 was selected for 4 rainfall stations. This highlights the consequences of the decline in climate stations, as fewer available stations make selecting representative driver stations increasingly challenging. Furthermore, the lack of AWS stations presents a major issue for obtaining temperature data, particularly if more stations are closed due to funding constraints.

When comparing different rainfall datasets, the Pegram mean dataset was found to be the most effective for estimating rainfall adjustment factors, leading to more representative station rainfall data for the catchment. Additionally, Pegram's gridded monthly datasets should be used for calculating adjustment factors. However, extending the record length of existing rainfall datasets by an additional 20 years did not improve the accuracy of the station's MAP statistic.

The adjustment factors were influenced by the *PPTCor* rules applied to median data. During winter months, an adjustment factor of 2 was applied, effectively doubling the rainfall. Since streamflow is highly sensitive to rainfall, this resulted in an overestimation of streamflow during that period.

In addition, verification results showed that *ACRU* overestimated dam inflows, likely due to unaccounted factors such as crop irrigation from rivers and river abstractions for domestic water supply. While *ACRU* performed better in simulating high dam inflows, the AF method provided a more accurate simulation of streamflow than the DS approach, particularly during high-flow periods. The AF method effectively captured seasonal fluctuations and demonstrated strong performance across various timeframes, making it a preferred choice for selecting climate driver stations in the catchment area.

The study also emphasised the need to use more than 50 years of data for accurate frequency analysis, especially for capturing extreme values in hydrological data. This is crucial for improved water resource management and infrastructure planning. Lastly, the study highlights the need for continued optimisation and validation of methods for selecting climate driver stations, particularly to address challenges in low-flow simulations. These improvements are essential for effective water resource management and planning in the study area.

## 5.3 Revisiting aims and objectives

The primary aim of this study was to evaluate existing for selecting climate driver stations. By achieving this aim, the research sought to enhance the accuracy and reliability of climate data inputs for hydrological modelling.

### 1. Selection of a suitable study catchment

Climate data were provided by SAWS for 3 DMs (i.e. Lejweleputswa, Vhembe and uMgungundlovu). For the uMgungundlovu DM, SAWS only provided manual station for the northern part of the district, and therefore, the focus shifted to the Vhembe and Lejweleputswa districts. The Vhembe district had temperature data from AWSs with a minimum of 38 % missing, which was impossible to infill due to the stations having some common periods of

missing data. The Lejweleputswa DM was selected for this study on the basis that it had a better station network with more reliable data compared to the other 2 DMs.

**2. Development of an extended daily rainfall dataset**

This objective was achieved through the successful application of the IDW method for rainfall data infilling, resulting in a comprehensive 70-year dataset for selected quaternary catchments. Data from 1996 to 2019 was obtained from custodians and existing datasets up to 1999 were extended by 20 years. This process was successfully applied to 16 rainfall stations in and near the Lejweleputswa DM, ensuring more complete and reliable datasets for analysis. Additionally, the objective was partially achieved for 11 temperature stations, though the limited availability of AWS stations remains a challenge for obtaining long-term temperature data and extending historical data.

**3. Assessment of gridded rainfall datasets**

The assessment of gridded rainfall datasets was conducted by comparing the Pegram, Lynch, and Dent datasets using statistical performance measures, including $R^2$ and RMSE. These datasets were evaluated based on their ability to estimate rainfall adjustment factors and improve the representativeness of rainfall data for a catchment. The methodology involved selecting driver stations using the DS approach and AF method and then adjusting daily rainfall data using monthly adjustment factors derived from each gridded dataset. The dataset that produced the lowest RMSE and highest NSE was identified as the most reliable. The Pegram mean dataset outperformed the others in estimating rainfall adjustment factors, leading to improved representation of rainfall data within the catchment. This enhanced the accuracy of hydrological modelling, as it provided more representative rainfall inputs for streamflow simulations.

**4. Evaluation of existing station selection techniques**

The evaluation of existing station selection techniques was conducted by comparing the commonly used DS approach and the newly developed AF method. Both methods were tested for their effectiveness in selecting representative climate driver stations for rainfall and temperature data. The DS approach relied on selecting the nearest station with reliable data,

while the AF method adjusted station data using monthly adjustment factors derived from gridded rainfall datasets to better match the catchment's rainfall characteristics. Additionally, rainfall and temperature data from the selected driver stations were compared to spatially averaged catchment values derived from gridded datasets, including the Dent, Lynch and Pegram datasets.

To assess the performance of these techniques, the selected driver stations were used as input for the *ACRU* agro-hydrological model. The *ACRU* model was applied to simulate streamflow for the study catchment and simulated inflows to the Erfenis Dam were compared against dam water balance data to verify the accuracy of the selected stations.

The accuracy of each approach was evaluated using 3 statistical performance metrics. However, the results were inconclusive due to various challenges, including the absence of observed streamflow data for the study catchment. The *ACRU* model was unable to adequately simulate inflow to the dam, due to various land uses in upstream catchments (especially irrigation since river abstraction volumes were unknown), making it difficult to verify which method performed better.

## 5. Development of alternative methods for selecting a representative driver station

Building on the finding of the assessment of the existing methods, the AF method was used to improve rainfall driver stations selection. The AF method used gridded rainfall datasets to derive adjustments factors, which were then applied to observed station data to enhance its representativeness at the catchment scale. The AF method aimed to address spatial inconsistencies in rainfall measurements and improve accuracy of climate data used in hydrological modelling.

## 6. Evaluating the performance of the methods for selecting a climate driver station using hydrological modelling

The performance of the DS approach and AF method was evaluated using hydrological modelling. The *ACRU* agro-hydrological model was applied to simulate streamflow for the study catchment and simulated inflows to the Erfenis Dam were compared with dam water balance data from DWS. However, due to unaccounted factors such as irrigation abstractions

and less reliable climate data, the model overestimated streamflow, making the verification process inconclusive.

## 5.4 Challenges and limitations

Due to the low number of climate stations in/near the study catchment, infilling of missing data were a challenge. The low number of stations having full data record after being infilled resulted in the same station being selected for multiple quaternary catchments. This was more complicated for the infilling of missing temperature data, as it required a different control station to be selected for each temperature variable (i.e. $T_{MAX}$ and $T_{MIN)}$ for each month of the year. The decision to use hydrological modelling to test the driver station selection methods was made to evaluate the effectiveness of different techniques in identifying representative climate driver stations. However, this approach was not successful due to the challenges explained next.

In the Lejweleputswa DM, a tertiary study catchment (i.e. C41) was selected as it had more stations with reliable data within and nearby the catchment boundary. A reliable streamflow gauge was required at the catchment outlet. However, the gauging weir only measured flow within an irrigation canal. The dominant land cover for the selected study catchment included irrigated farmlands. The catchment also included a large dam primarily used for irrigation and domestic purposes, but inadequate water abstraction information was available. Therefore, modelling this catchment using *ACRU* was complex, which negatively affected the verification of techniques to select climate driver stations.

## 5.5 Recommendations for future work

Future studies should focus on catchments with a higher density of climate stations, as this is crucial for improving the accuracy and reliability of hydrological and climate modelling. It is essential to address the issue of climate stations and consider alternative data sources, such as CHIRPS, which provides satellite-based measurements. CHIRPS offers continuous, gap-free data that are readily available, making it a valuable resource for climate analysis.

There is a need to re-assess the existing *PPTCor* rules. When calculated using median data, they create adjustment factors that either halve or double rainfall data during the dry months. In future hydrological modelling studies, it is recommended that the Pegram mean monthly rainfall grids are adopted to estimate monthly rainfall adjustment factors. The adjustment factors developed for the quaternary and quinary catchments were calculated using the Lynch median dataset, and thus the recommendation is that the monthly *PPTCor* values are re-calculated using the Pegram mean dataset.

Additionally, implementing an automated method for infilling missing rainfall and temperature data would be beneficial, as it would enhance data continuity, improve bias correction and reduce uncertainties in hydrological modelling. Methods such as machine learning algorithms, statistical interpolation, or geostatistical approaches could be explored to enhance data accuracy, particularly in regions with sparse climate station networks.

Finally, further validation and refinement of the methods for selecting a climate driver station should be pursued, particularly in other areas where a more reliable and denser network of climate stations is available. Additionally, the methods must be tested in a catchment with a reliable weir gauge and, preferably, pristine land cover. These requirements are necessary to adequately assess their applicability and accuracy, thereby improving effective water resource management and planning initiatives in the catchment area.

# 6 REFERENCES


Allen RG, Pereira LS, Raes D and Smith M (1998). Crop evapotranspiration-guidelines for computing crop water requirements. *Irrigation and Drainage Paper No. 56*, Chapter 6, Food and Agricultural Organisation (FAO), Rome Italy. https://www.fao.org/4/x0490e/x0490e00.htm

Arnold JG, Srinivasan R, Muttiah RS and Williams JR (1998). Large area hydrological modelling and assessment part I: model development. *Journal of the American Water Resources Association,* 34(1), 73-890. https://doi.org/10.1016/k.atmosres.2018.05.009

Beck HE, Van Djik AIJM, De Roo A, Dutra E, Fink G, Orth R and Schellekens J (2017). Global evaluation of runoff from 10 state-of-the-art hydrological models. *Hydrology and Earth System Sciences,* 21(6), 2881-2903. https://doi.org/10.5194/hess-21-2881-2017

Clark DJ (2015). *Development and assessment of an integrated water resources accounting methodology for South Africa*. WRC Report No. 2205/1/15, Water Research Commission, Pretoria, South Africa.

Clark DJ, Horan MJC, Kunz RP, Schütte, Schulze RE, Xolo T and Smithers JC (2024). *Development of datasets for multi-scale water resource assessments towards a water secure South Africa.* WRC Report No. xxxx/x/24, Water Research Commission, Pretoria, South Africa (in press).

Clark MP, Kavetski D and Fenicia F (2011). Pursuing the method of multiple working hypotheses for hydrological modelling. *Water Resources Research,* 47(9). https://doi.org/10.1029/2010wr009827

Clulow AD, Kunz RP, Gokool S, Toucher ML, Schütte S, Schulze RE, Horan R, Everson CE, Thornton-Dibb SLC, Horan MJC, Kapein N, Clark DJ and Germishuizen I (2023). *The expansion of knowledge on evapotranspiration and stream flow reduction of different clones/hybrids to improve the water use estimation of SFRA species (i.e. Pinus, Eucalyptus, and wattle species): improved water use estimation of SFRA species – Volume 2: SFRA assessment utility*. WRC Report No. TT898/2/22, Water Research Commission, Pretoria, South Africa.

Davis-Reddy CL and Vincent K (2017). *Climate Risk and Vulnerability: A handbook for Southern Africa,* (2nd Ed), CSIR, Pretoria, South Africa.

Dent MC, Lynch SD and Schulze RE (1989). *Mapping mean annual rainfall statistics over Southern Africa*. WRC Report No. 109/1/89, Water Research Commission, Pretoria, South Africa.

Department of Forestry, Fisheries and the Environment (DFFE) (2014). *South African National Land Cover.* Department of Forestry, Fisheries and the Environment, Pretoria, South Africa. https://egis.environment.gov.za

Department of Water and Sanitation (DWS) (2019). *Quaternary catchments of South Africa*. Department of Water and Sanitation, Pretoria, South Africa. https://www.dws.gov.za/SLIM/Systems/Metadata/metadata.aspx

Dinku T, Ceccato P and Connor SJ (2011). Challenges of satellite rainfall estimation over mountainous and arid parts off East Africa. *International Journal of Remote Sensing,* 32(21), 5965-5979. https://doi.org/10.1080/01431161.2020.499381

Dinku T, Thomson MC, Cousin R, del Corral J, Ceccato P, Hansen J and Connor SJ (2017). Enhancing National Climate Services (ENACTS) for development in Africa. *Climate and Development,* 10, 664-672. https://doi.org/10.1080/17565529.201.1405784

Duc L and Sawada Y (2023). A signal-processing-based interpretation of the Nash–Sutcliffe efficiency. *Hydrology and Earth System Sciences*, 27(9), 1827-1839. https://doi.org/10.5194/hess-27-1827-2023

Dutra E, Muñoz-Sabater J, Soubail BS, Komori T, Hirahara S and Blsamo G (2020). Environmental lapse rate for high-resolution land surface downscaling: An application to ERA5. *Earth and Space Science,* 7(5). https://doi.org/10.1029/2019ea000984

Erasmus D (2022). *Rapid loss of weather stations puts SA's disaster forecasting at risk*. Business LIVE. https://www.businesslive.co.za/bd/national/2022-05-16-rapid-loss-of-weather-stations-puts-sas-disaster-forecasting-at-risk/#:~:text=Vanetia%20Phakula%2C%20a%20meteorologist%20at

Environmental System Research Institute (ESRI) (2020). *What is new in ArcMap, Documentation*. https://desktop.arcgis.com/en/arcmap/latest/get-started/introduction/whats-new-in-arcgis.htm

Faniriantsoa R and Dinku T (2022). ADT: The automatic weather station data tool. *Frontiers in Climate*, 4. https://doi.org/10.3389/fclim.2022.933543

Fenta AA, Yasuda H, Shimizu K, Ibaraki Y, Haregeweyn N, Kawai T, Belay AS, Sultan D and Ebabu K (2018). Evaluation of satellite rainfall estimates over the Lake Tana basin at the source region of the Blue Nile River. *Atmospheric Research,* 212, 43-53. https://doi.org/10.1016/j.atmosres.2018.05.009

Fourie R, Haddad CR, Dippenaar-Schoeman AS and Grobler A (2013). Ecology of the plant-dwelling spiders (Arachnida: Araneae) of the Erfenis Dam Nature Reserve, South Africa. *Koedoe*, 55(1). https://doi.org/10.4102/koedoe.v55i1.1113

Funk C, Peterson P, Landsfeld M, Pedreros D, Verdin J, Shukla S, Husak G, Rowland J, Harrison L, Hoell A and Michaelsen J (2015). The climate hazards infrared precipitation with stations−a new environmental record for monitoring extremes. *Scientific Data,* 2(1). https://doi.org/10.1038/sdata.2015.66

Gao Y, Merz C, Lischeid G ans Schneider M (2018). A review on missing hydrological data processing. *Environmental Earth Sciences,* 77(2). https://doi.org/10.1007/s12665-018-7228-6

Gebrechorkos SH, Hülsmann S and Bernhofer C (2018). Evaluation of multiple climate data sources for managing environmental resources in East Africa. *Hydrology and Earth System Sciences,* 22(8), 4547-4564. https://doi.org/10.5194/hess-22-4547-2018

Ginster M, Tempelhoff JWN, van der Merwe L, Berner S, Motloung S and Rabali U (2014). Running on empty: investigating the 2013 Brandfort water supply. *Research Niche for the Cultural Dynamics of Water,* North West University, Vanderbijlpark, version 7.

Hargreaves GH and Samani ZA (1985). Reference crop evaporation from temperature. *Journal of Applied Engineering in Agriculture,* 1, 96-99.

Hrachowitz M, Saveniie HHG, Blöschl G, McDonnell JJ, Sivapalan M, Pomeroy JW, Arheimer B, Blume T, Clark MP, Ehret U, Fenicia F, Freer JE, Gelfan S, Gupta HV, Hughes DA, Hut RW, Montanari A, Pande S, Tetzlaff D and Troch PA (2013). A decade of Prediction in Ungauged Basins (PUB) - a review. *Hydrological Sciences Journal,* 58(6), 1198-1255. https://doi.org/10.1080/02626667.2013.803183

Huffman GJ, Bolvin DT, Braithwaite D, Hsu KL, Joyce R and Xie P (2020). NASA Global Precipitation Measurement (GPM) Integrated Multi-satellite Retrievals for GPM (IMERG). *Algorithm Theoretical Basis Document (ATBD),* Version 5.2.

Kunz R (2004). *Daily Rainfall Data Extraction Utility User Manual v.1.4.* Institute for Commercial Forestry Research, Pietermaritzburg, South Africa.

Kunz RP, Davis NS, Thornton-Dibb SLC, Steyn JM, Du toit ES and Jewitt GPW (2015). *Assessment of biofuel feedstock production in South Africa: Atlas of water use and yield of biofuel crops in suitable growing areas.* WRC Report No. TT 652/15, Water Research Commission, Pretoria, South Africa.

Kunz R, Masanganise J, Reddy K, Mabhaudhi T, Lembede L, Naiken V and Ferrer S (2020). *Water use and yield of soybean and grain sorghum for biofuel production*. WRC Report No. 2491/1/20, Water Research Commission, Pretoria, South Africa.

Lennard C, Coop L, Morison D and Grandin R (2013). *Extreme events: Past and future changes in the attributes of extreme rainfall and dynamics of their driving processes.* WRC Report No. 1960/1/12, Water Research Commission, Pretoria, South Africa.

Lumsden TG, Kunz RP, Schulze RE, Knoesen DM and Barichievy KR (2011). Methods 4: Representation of grid and point scale regional climate change scenarios for national and catchment level hydrological impacts assessments. In: Schulze RE, Hewitson BC, Barichevy KR, Tadross M, Kunz RP, Horan MJC and TG Lumsden, *Methodological approaches to assessing eco-hydrological responses to climate change in South Africa*, Chapter 9*,* 89-99. WRC Report No. 1562/1/10, Water Research Commission, Pretoria, South Africa.

Luo Y, Ogle K, Tucker C, Fei S, Gao C, LaDeau SL, Clark JH and Schimel DS (2011). Ecological forecasting and data assimilation in a data-rich era. *Ecological Applications,* 21(5), 1429-1442. https://do.org/10.1890/09-1275.1

Lynch SD (2004). *Development of a raster database of annual, monthly and daily rainfall for South Africa*. WRC Report No. 1156/1/04, Water Research Commission, Pretoria, South Africa.

Masih I, Maskey S, Uhlenbrook S and Smakhtin V (2010). Assessing the impact of areal precipitation input on streamflow simulations using the SWAT model1. *Journal of the American Water Resources Association,* 47(1), 179-195. https://doi.org/10.1111/j.1752-1688.2010.00502.x

Midgley DC, Pitman WV and Middleton BJ (1994). *Surface water resources of South Africa 1990*. WRC Report No. 298/1/94 to 298/6.2/94, Water Research Commission, Pretoria, South Africa.

Moeletsi ME, Walker S and Hamandawana H (2013). Comparison of the Hargreaves and Samani equation and the Thornthwaite equation for estimating dekadal evapotranspiration in the Free State Province, South Africa. *Physics and Chemistry of the Earth, Parts A/B/C*, 66, 4-15. https://doi.org/10.1016/j.pce.2013.08.003

Moeletsi ME, Shabalala ZP, de Nysschen G and Walker S (2016). Evaluation of an inverse distance weighting method for patching daily and dekadal rainfall over Free State Province, South Africa. *Water SA,* 42(3), 466. https://doi.org/10.4314/wsa.v4i3.12

ORASECOM (2013). *Infrastructure catalogue for the Orange-Senqu River Basin, (TR21).* https://wis.orasecom.org/contents/study/UNDP-GEF/general/Documents/Technical%20Reports/TR21_InnfrastructureCatalogue_lowres_Dec2013.pdf

Patrinos HA (2014). *You can't manage what you don't measure. World Bank Blogs*. https://blogs.worldbank.org/education/you-can-t-manage-what-you-don-t-measure#:~:text=But%20as%20the%20management%20guru,know%20how%20you%20are%20doing%3F

Pegram GGS, Sinclair S and Bardossy A (2016). *New methods of infilling Southern African raingauge records enhanced by annual, monthly and daily precipitation estimates tagged with uncertainty*. WRC Report No. 2241/1/15, Water Research Commission, Pretoria, South Africa.

Pike A, Schulze RE, Hallowes L, Thornton-Dibb S, Clark D, Horan M, Taylor V and WMA Consultants (2004). New development in and refinements to, supporting software, documentation, user support and promotion of the *ACRU* agrohydrological modelling system. In: Schulze RE and Pike A (Eds), *Development and evaluation of an installed hydrological modelling system.* WRC Report No. 1155/1/04, Water Research Commission, Pretoria, South Africa.

Pitman WV and Bailey AK (2021). Can CHIRPS fill the gap left by the decline in the availability of rainfall stations in Southern Africa. *Water SA,* 47(2), 162-171. https://doi.org/10.17159/wsa/2021.v47i2.10912

Randrianasolo A, Ramos MH, Thirel G, Andréassian V and Martin E (2010). Comparing the scores of hydrological ensemble forecasts issued by two different hydrological models. *Atmospheric Science Letters*, *11*(2), 100-107. https://doi.org/10.1002/asl.259

Razavi T and Coulibaly (2016). An evaluation of regionalization watershed classification schemes for continuous daily streamflow prediction in ungauged watersheds. *Canadian Water Resources Journal,* 42(1), 2-20. https://doi.org/10.1080/07011784.2016.1184590

Rowe T and Smithers J (2018). Review: Continuous simulation modelling for design flood estimation-a South Africa perspective and recommendation. *Water SA,* 44. https://doi.org/10.4314/wsa.v44i4.18

Schulze RE (1995). *Hydrology and agrohydrology: A text to accompany the ACRU 3.00 agrohydrological modelling system*. WRC Report No. TT69/9/95, Water Research Commission, Pretoria, South Africa.

Schulze RE (2010). Rainfall: Background. In: Schulze, RE, *Atlas of Climate Change and the South African Agricultural Sector: A 2010 Perspective,* Chapter 3.6, 85-89. Department of Agriculture, Forestry and Fisheries, Pretoria, South Africa.

Schulze RE (2024). Personal communication, Emeritus Professor of Hydrology, Centre for Water Resources Research, School of Agricultural, Earth and Environmental Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa, 28 May 2024.

Schulze RE and Horan MJC (2011). Methods 1: Delineation of South African, Lesotho and Swaziland into quinary catchments. In: Schulze RE, Hewitson BC, Barichievy KR, Tadross M, Kunz RP, Horan MJC and Lumsden TG, *Methodological Approaches to Assessing Eco-Hydrological Responses to Climate Change in South Africa*. Chapter 7, 635-74. WRC Report No. 1562/1/10, Water Research Commission, Pretoria, South Africa.

Schulze RE and Maharaj M (2004). *Development of a database of gridded daily temperatures for South Africa.* WRC Report No. 1156/2/04, Water Research Commission, Pretoria, South Africa.

Schulze RE and Pike A (2004). *Development and evaluation of an installed hydrological modelling system.* WRC Report No. 1155/1/04, Water Research Commission, Pretoria, South Africa.

Schulze RE, Dent MC, Lynch SD, Schäfer NW, Kienzle SW and Seed AW (1995). Rainfall. In Schulze RE (Ed.), *Hydrology and Agrohydrology: a text to accompany the ACRU 3.00 Agrohydrological modelling system,* Chapter 3, AT3-AT38. School of Bioresources Engineering and Environmental Hydrology, University of KwaZulu-Natal, Pietermaritzburg, South Africa.

Schulze RE, Horan MJC, Kunz RP, Lumsden TG and Knoesen DM (2010). The South African quinary catchments database. In: Schulze RE (Ed.), *Climate change and the South African sugarcane sector: a 2010 perspective*, *ACRU*cons Report 61, Chapter 4, 25-31. School of Bioresources Engineering and Environmental Hydrology, University of KwaZulu-Natal, Pietermaritzburg, South Africa.

Schulze RE, Hallowes LA, Horan MJC, Lumsden TG, Pike A, Thornton-Dibb S and Warburton ML (2007a). South African Quaternary Catchments Database. In: Schulze RE (Ed.), *South African atlas of climatology and agrohydrology,* Chapter 2.3. WRC Report No. 1489/1/06, Water Research Commission, Pretoria, South Africa.

Schulze RE, Maharaj M and Moult N (2007b). Reference crop evaporation by the penman-Monteith method. In: Schulze RE (Ed.), *South African atlas of climatology and agrohydrology,* Chapter 13.3. WRC Report No. 1489/1/06, Water Research Commission, Pretoria, South Africa.

Schulze RE, Horan MJC, Kunz RP, Lumsden TG, Knoesen DM (2011). Method 2: Development of the South African quinary catchments database. In: Schulze RE, Hewitson BC, Barichevy KR, Tadros M, Kunz RP, Horan MJC and Lumsden TG, *Methodological approaches to assessing eco-hydrological responses to climate change in South Africa*, Chapter 7, 635-74. WRC Report No. 1562/1/10, Water Research Commission, Pretoria, South Africa.

Schulze RE, Warburton ML, Lumsden TG and Horan MJC (2005). The Southern African Quaternary Catchments Database: Refinements to, and Links with, the ACRU System as a Framework for Modelling Impacts of Climate Change on Water Resources. In: Schulze RE, *Climate Change and Water Resources in Southern Africa: Studies on Scenarios, Impacts, Vulnerabilities and Adaptation*, Chapter 8, 111-137. WRC Report No. 1430/1/05, Water Research Commission, Pretoria, South Africa.

Serinaldi F and Kilsby CG (2014). Rainfall extremes: Towards reconciliation after the batter of distributions. *Water Resources Research,* 50(1), 336-352. https://doi.org/10.1002/2013wr014211

Siddig MSA, Ibrahim S, Yu Q, Abdalla A, Osman Y, Atiem IA, Hamukwaya SL and Taha MMM (2022). Bias adjustment of four-satellite-based rainfall products using ground-based measurements over Sudan. *Water,* 14, 1475. https://doi.org/10.3390/w14091475

Sinclair S and Pegram GGS (2010). A comparison of ASCAT and modelled soil moisture over South Africa, using Topkapi in land surface mode. *Hydrology and Earth System Sciences*, 14(4), 613-626. https://doi.org/10.5194/hess-14-613-2010

Smithers JC and Schulze RE (1995). *ACRU Agrohydrological modelling system: User Manual Version 3.00*. WRC Report No. TT70/95, Water Research Commission, Pretoria, South Africa.

Smithers, JC and Schulze, RE. (2004). *ACRU Agrohydrological Modelling System: User Manual Version 4.00*. School of Bioresources Engineering and Environmental Hydrology, University of KwaZulu-Natal, Pietermaritzburg, South Africa.

Smithers JC, Schulze RE, Lynch SD, Hallowes LA, Thornton-Dibb SLC, Pike A and Rivers-Moore N (2004). Preparation of hydroclimatic input files and the quaternary catchments database. In: Smithers JC and Schulze RE (Eds), *ACRU Agrohydrological modelling system: user manual version 4.00*. University of KwaZulu-Natal, Pietermaritzburg, South Africa.

Suleman S, Chetty K, Clark D and Kapangaziwiri E (2020). Assessment of satellite-derived rainfall and its use in the SCRU agro-hydrological model. *Water SA,* 46. https://doi.org/10.17159/wsa/2020.v46.i46.i4.9068

Toté C, Patricio D, Boogaard H, Van de Wiingaart R, Tarnavsky E and Funk C (2015). Evaluation of satellite rainfall estimates for drought and flood monitoring in Mozambique. *Remote Sensing,* 7(2), 1758-1776. https://doi.org/10.3390/rs70201758

van Dijk M and van Vuuren S (2012). Destratification induced by bubble plumes as a means to reduce evaporation from open impoundments, *Water SA*, 32(2), 158-167. https://doi.org/10.4314/wsa.v35i2.76731

Wang W, Lu H, Yang S, Sothea K, Jiao Y, Gao B, Peng X and Pang Z (2016). Modelling hydrological processes in the Mekong River basin using a distributed model driven by satellite precipitation and rain gauge observations. *PLOS ONE,* 11(3), e0152229. https://doi.org/10.1371/journal.pone.0152229

Warburton WL (2005). *Detection of changes in temperature and streamflow parameters over Southern Africa*. MSc dissertation, School of Bioresource Engineering and Environmental Hydrology, University of KwaZulu-Natal, Pietermaritzburg, South Africa.

Weepener HL, Van Den Berg HM, Metz M and Hamandawana H (2012). *The development of a hydrologically improved Digital Elevation Model and derived products for South Africa based on the SRTM DEM.* WRC Report No. 1908/1/11, Water Research Commission, Pretoria, South Africa.

Wind Atlas South Africa (WASA) (2021). WASA 3 Final Virtual Seminar: Programme with presentations. https://www.wasaproject.info/docs/WASA_3_Resource_Map_March_2021.png

# 7 APPENDIX A

## 7.1 Location of climate stations

### 7.1.1 Manual rainfall stations

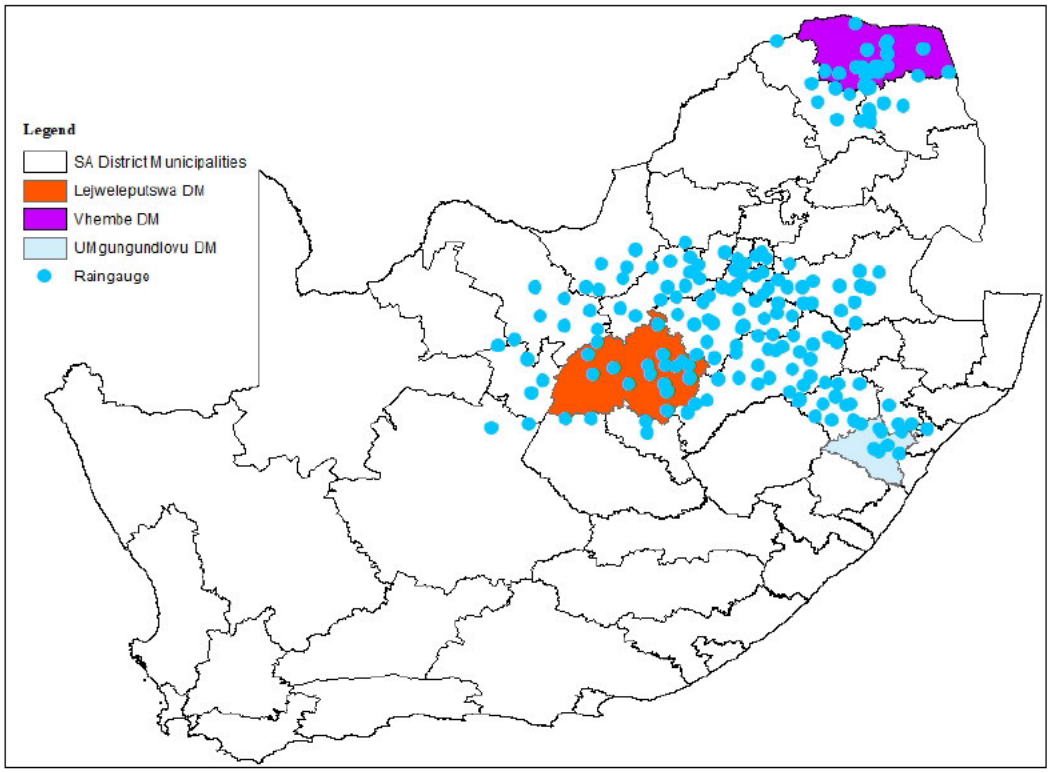


**Figure 7.1**: Location of all manual rainfall stations received from SAWS and DWS for the Vhembe, Lejweleputswa and uMgungundlovu district municipalities

**Figure 7.2**: Location of SAWS and DWS manual rainfall stations within and around the Vhembe (top), Lejweleputswa (middle) and uMgungundlovu (bottom) district municipalities

### 7.1.2 Automatic weather stations

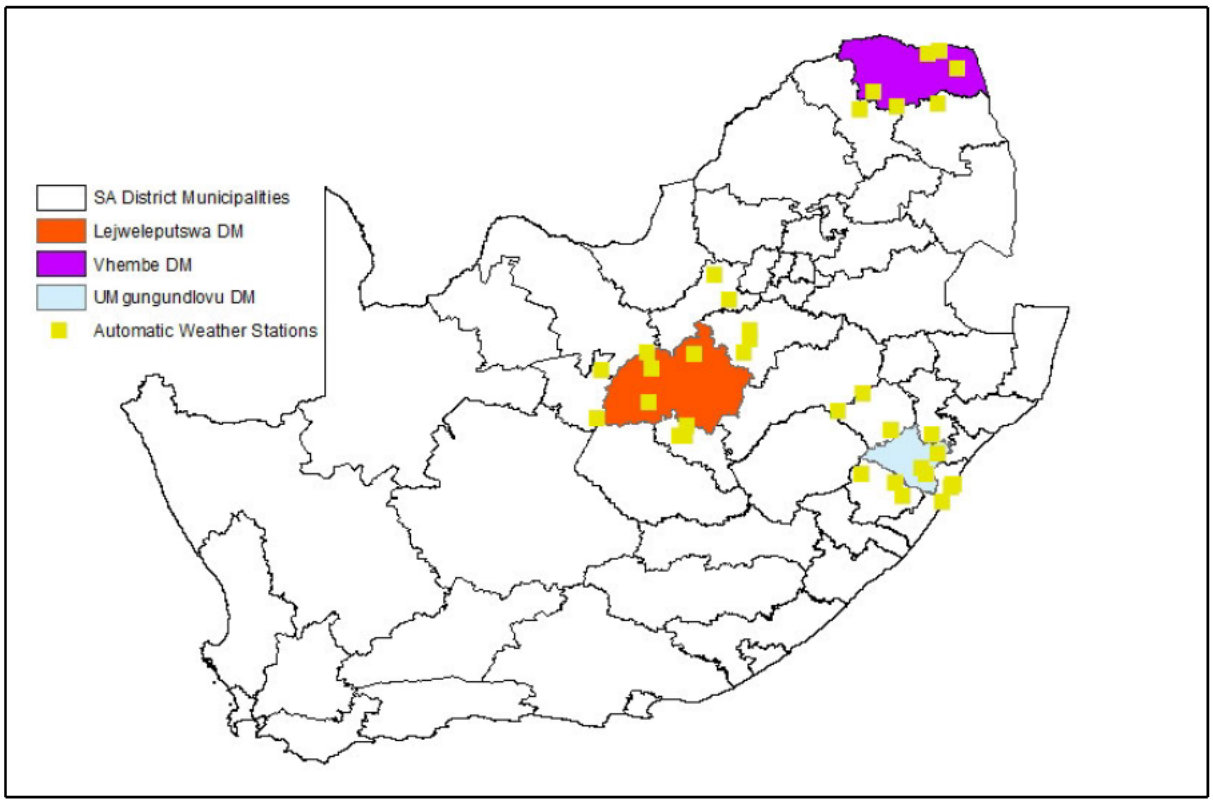


**Figure 7.3**: Location of SAWS and ARC automatic weather stations within and around Vhembe, Lejweleputswa and uMgungundlovu district municipalities

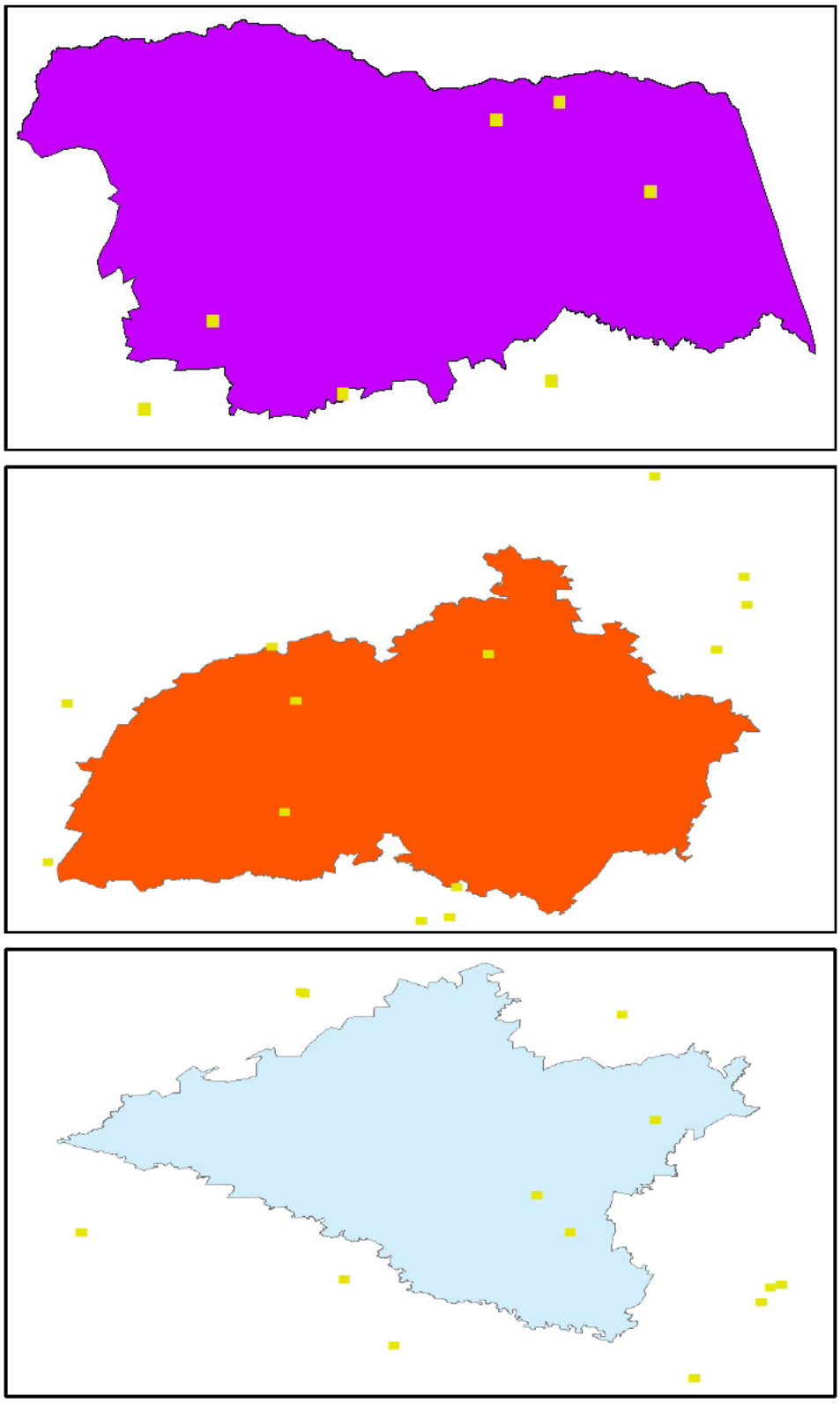

**Figure 7.4**: Location of SAWS and ARC automatic weather stations within and around the Vhembe (top), Lejweleputswa (middle) and uMgungundlovu (bottom) district municipalities

## 7.2 Quality of manual station data

### 7.2.1 Vhembe DM

**Table 7.1**: Portion of reliable and missing record for SAWS and DWS manual rainfall stations located within and around the Vhembe District Municipality

| Organisation | Station ID | Data end date | Reliable data (%) | Missing data (%) |
|---|---|---|---|---|
| SAWS | 0723363_X | 2019/12/31 | 100.00 | 0.00 |
| | 0722721_3 | 2019/12/31 | 99.99 | 0.01 |
| | 0766390_1 | 2019/12/31 | 99.98 | 0.02 |
| | 0722277_X | 2019/12/31 | 99.97 | 0.03 |
| | 0721665_5 | 2019/12/31 | 98.29 | 1.71 |
| | 0723155_X | 2019/12/31 | 97.81 | 2.19 |
| | 0766779_6 | 2019/12/31 | 97.19 | 2.81 |
| | 0768011A8 | 2019/12/31 | 95.80 | 4.20 |
| | 0723070_7 | 2019/12/31 | 91.40 | 8.60 |
| | 0766030_3 | 2019/12/31 | 90.58 | 9.42 |
| | 0766827_4 | 2014/08/13 | 72.20 | 27.80 |
| | 0766842_X | 2012/10/31 | 69.96 | 30.04 |
| | 0809706_X | 2009/11/27 | 57.73 | 42.27 |
| | 0723080_4 | 2008/12/31 | 47.55 | 52.45 |
| | | | | |
| **DWS** | A8E001 | 2019/12/31 | 99.33 | 0.67 |
| | A8E004 | 2019/12/31 | 98.94 | 1.06 |
| | A8E002 | 2019/12/31 | 98.29 | 1.71 |
| | A9E002 | 2019/12/31 | 97.19 | 2.81 |
| | A9E001 | 2017/01/01 | 1.06 | 98.94 |

### 7.2.2 Lejweleputswa DM

**Table 7.2**: Portion of reliable and missing record for SAWS and DWS manual rainfall stations located within and around the Lejweleputswa District Municipality

| Organisation | Station ID | End date | Reliable data (%) | Missing data (%) |
|---|---|---|---|---|
| SAWS | 0329166_5 | 2019/12/31 | 99.90 | 0.10 |
| | 0294500_X | 2019/12/31 | 99.73 | 0.27 |
| | 0329215_5 | 2019/12/31 | 99.62 | 0.38 |
| | 0327883_9 | 2019/12/31 | 99.00 | 1.00 |
| | 0364322_X | 2019/12/31 | 98.38 | 1.62 |
| | 0400203_4 | 2019/12/31 | 95.52 | 4.48 |
| | 0327784_2 | 2019/12/31 | 94.34 | 5.66 |
| | 0326668_0 | 2012/03/31 | 67.60 | 32.40 |
| | 0326073_X | 2012/01/31 | 66.60 | 33.40 |
| | 0328425_2 | 2010/08/31 | 59.66 | 40.34 |
| | 0361832_9 | 2010/09/30 | 41.67 | 58.33 |
| | 0294481_1 | 2005/12/31 | 40.54 | 59.46 |
| | 0329001_4 | 2000/08/31 | 19.45 | 80.55 |
| | 0327264_7 | 2000/02/29 | 17.35 | 82.65 |
| | 0328726_9 | 1999/12/31 | 16.47 | 83.53 |
| | | | | |
| DWS | C4E008 | 2019/12/31 | 100.00 | 0.00 |
| | C4E002 | 2019/12/31 | 99.98 | 0.02 |
| | C9E004 | 2019/12/31 | 89.61 | 10.39 |

### 7.2.3 uMgungundlovu DM

**Table 7.3**: Portion of reliable and missing record for SAWS and DWS manual rainfall stations located within and around the uMgungundlovu District Municipality

| Organisation | Station ID | End date | Reliable data (%) | Missing data (%) |
|---|---|---|---|---|
| **SAWS** | 0269611_1 | 2017/01/31 | 80.03 | 19.97 |
| | | | | |
| **DWS** | U2E003 | 2019/12/31 | 99.94 | 0.06 |
| | V2E002 | 2019/12/31 | 99.65 | 0.35 |
| | U2E002 | 2019/12/31 | 98.55 | 1.45 |
| | U2E006 | 2019/12/31 | 98.25 | 1.75 |
| | U2E009 | 2019/12/31 | 97.58 | 2.42 |
| | U2E004 | 2013/04/01 | 4.52 | 95.48 |

## 7.3 Quality of automatic station data

### 7.3.1 Vhembe DM

**Table 7.4**: The number of Automatic Weather Stations received from SAWS and ARC stations in and around the Vhembe District Municipality, as well as the portion of reliable and missing record

| | | | | Rainfall | | Temperature | |
|---|---|---|---|---|---|---|---|
| **Organ-isation** | **Station ID** | **Start date** | **End date** | **Reliable data (%)** | **Missing data (%)** | **Reliable data (%)** | **Missing data (%)** |
| **SAWS** | 0768011A8 | 2008/01/22 | 2019/12/31 | 42.35 | 57.65 | 40.09 | 59.01 |
| | 0724318_9 | 2010/06/22 | 2019/12/31 | 26.80 | 73.20 | 26.92 | 73.08 |
| | | | | | | | |
| **ARC** | 30546 | 2001/12/01 | 2019/12/31 | 65.69 | 35.31 | 62.63 | 37.37 |
| | 30719 | 2005/09/01 | 2019/12/31 | 58.37 | 41.63 | 58.53 | 41.47 |
| | 30721 | 2005.09/01 | 2019/12/31 | 57.15 | 42.85 | 57.71 | 42.29 |
| | 30751 | 2006/08/01 | 2019/12/31 | 54.33 | 45.67 | 53.79 | 46.21 |
| | 30740 | 2006/07/01 | 2019/12/31 | 52.74 | 47.26 | 52.28 | 47.72 |

### 7.3.2 Lejweleputswa DM

**Table 7.5**: The number of Automatic Weather Stations received from SAWS and ARC stations in and around the Lejweleputswa District Municipality, as well as the portion of reliable and missing record

| | | | | Rainfall | | Temperature | |
|---|---|---|---|---|---|---|---|
| **Organ-isation** | **Station ID** | **Start date** | **End date** | **Reliable data (%)** | **Missing data (%)** | **Reliable data (%)** | **Missing data (%)** |
| **SAWS** | 0261516B0 | 1996/01/01 | 2019/12/31 | 99.89 | 0.11 | 99.41 | 0.59 |
| | 0261307A4 | 1996/01/01 | 2019/12/31 | 98.43 | 1.57 | 98.92 | 1.08 |
| | 0290468A9 | 1996/01/01 | 2019/12/31 | 98.57 | 1.43 | 98.00 | 2.00 |
| | 0362189_7 | 1996/01/01 | 2019/12/31 | 96.73 | 3.27 | 97.55 | 2.45 |
| | 0365398_8 | 1996/01/01 | 2019/12/31 | 94.55 | 5.45 | 94.84 | 5.16 |
| | 043710A4 | 1997/05/01 | 2019/12/31 | 85.17 | 14.83 | 87.74 | 12.26 |
| | 0293597A6 | 2004/06/17 | 2019/12/31 | 62.67 | 37.33 | 63.15 | 36.85 |
| | 0473559A3 | 1997/04/01 | 2019/12/31 | 59.31 | 40.69 | 62.40 | 37.60 |
| | 0360597B0 | 2013/05/03 | 2019/12/31 | 27.02 | 72.98 | 26.64 | 73.36 |
| | | | | | | | |
| **ARC** | 30425 | 2000/06/01 | 2019/12/31 | 78.96 | 21.04 | 76.27 | 23.73 |
| | 30595 | 2003/06/01 | 2019/12/31 | 66.93 | 33.07 | 65.88 | 34.12 |
| | 30584 | 2002/11/01 | 2019/12/31 | 56.74 | 43.26 | 53.43 | 46.57 |
| | 30861 | 2009/08/01 | 2019/12/31 | 43.12 | 56.88 | 38.74 | 61.26 |
| | 30909 | 2012/09/01 | 2019/12/31 | 37.26 | 62.74 | 37.21 | 62.79 |

### 7.3.3 uMgungundlovu DM

**Table 7.6**: The number of Automatic Weather Stations received from SAWS and ARC stations in and around the uMgungundlovu District Municipality, as well as the portion of reliable and missing record

| | | | | Rainfall | | Temperature | |
|---|---|---|---|---|---|---|---|
| **Organ-isation** | **Station ID** | **Start Date** | **End date** | **Reliable data (%)** | **Missing data (%)** | **Reliable data (%)** | **Missing data (%)** |
| | | | | | | | |
| **SAWS** | 0333682A9 | 1996/01/30 | 2019/12/31 | 95.51 | 4.49 | 96.49 | 3.51 |
| | 0270155_9 | 1996/01/01 | 2019/12/31 | 91.95 | 8.05 | 92.28 | 7.72 |
| | 0240808A2 | 1996/01/01 | 2014/08/05 | 76.41 | 23.59 | 77.41 | 22.59 |
| | 0210099A7 | 2001/08/27 | 2019/12/31 | 74.87 | 25.13 | 75.38 | 24.62 |
| | 0237618A9 | 2003/05/20 | 2019/12/31 | 59.13 | 40.87 | 48.99 | 41.01 |
| | 0298791_9 | 2003/06/23 | 2019/12/31 | 55.12 | 44.88 | 54.94 | 45.06 |
| | 0238806_6 | 2006/02/06 | 2019/12/31 | 37.69 | 62.31 | 37.81 | 62.19 |
| | 0300630_8 | 2012/04/01 | 2019/12/31 | 31.14 | 68.86 | 30.81 | 69.19 |
| | 0300690_1 | 2003/07/05 | 2012/01/31 | 26.93 | 73.07 | 26.53 | 73.47 |
| | 0240837B7 | 2014/08/01 | 2018/06/03 | 15.75 | 84.25 | 15.67 | 84.33 |
| | 0240750_1 | 2018/05/24 | 2019/12/31 | 2.29 | 97.71 | 2.38 | 97.62 |
| | | | | | | | |
| **ARC** | 30160 | 1996/01/01 | 2019/12/31 | 86.71 | 13.29 | 85.67 | 14.33 |
| | 30157 | 2002/01/01 | 2019/12/31 | 67.10 | 32.90 | 63.43 | 36.57 |
| | 30531 | 2001/01/01 | 2014/04/31 | 48.94 | 51.06 | 45.87 | 54.13 |
| | 30844 | 2009/02/01 | 2019/12/31 | 36.95 | 63.05 | 36.85 | 63.15 |

## 7.4 Watersheds

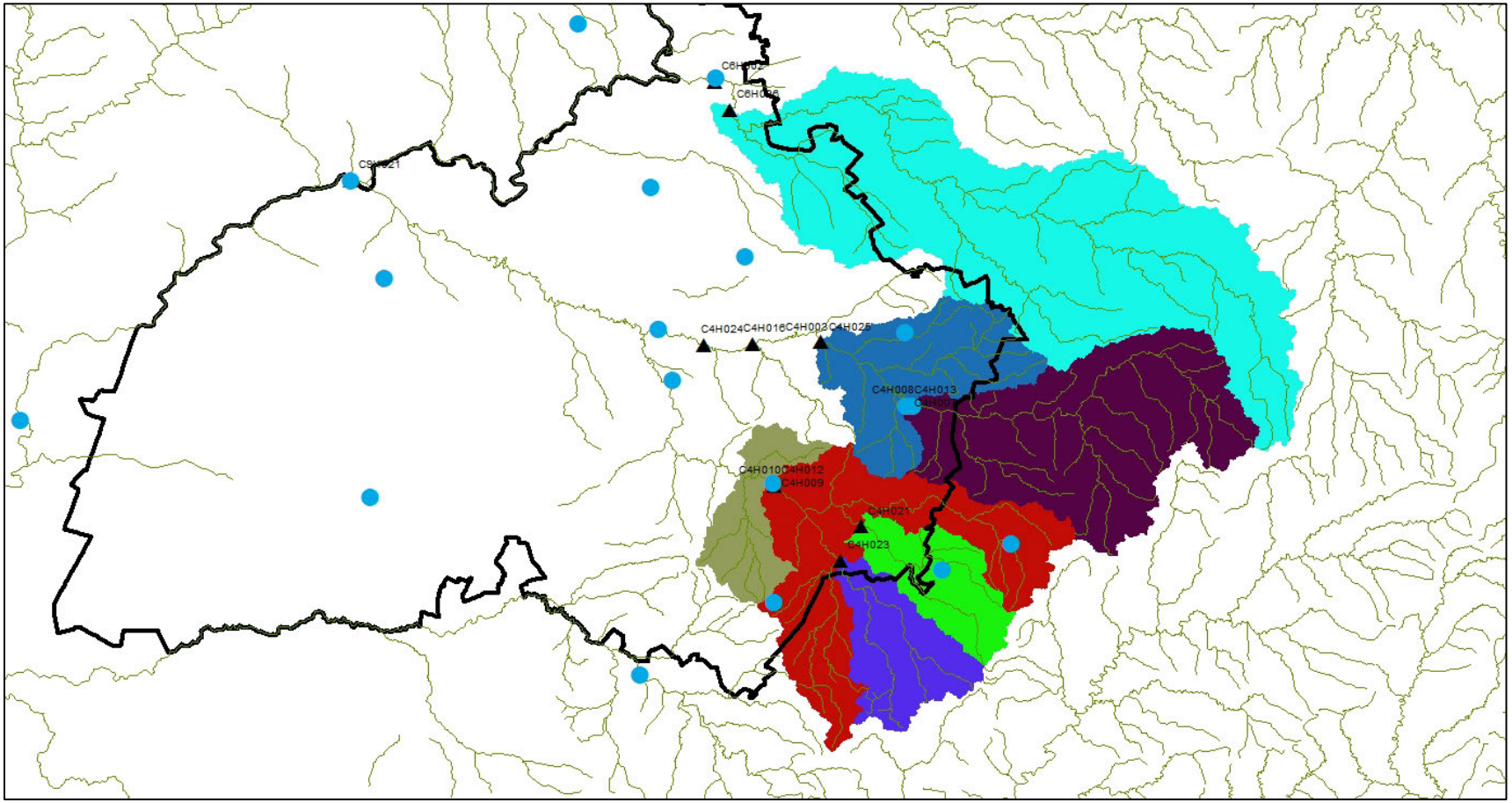


**Figure 7.5**: Location of catchments, climate stations and gauging weirs within and surrounding the Lejweleputswa DM

## 7.5 Rainfall adjustment factors

**Table 7.7**: Rainfall adjustment factors derived for driver stations selected by the AF method for quaternary catchments C41A-C41E

| Catchment | Driver station | Grid | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **C41A** | 0295405_7 | Dent median | 0.99 | 0.89 | 1.02 | 0.95 | 0.88 | 0.50 | 0.99 | 0.47 | 1.24 | 1.00 | 0.96 | 0.87 |
| | 0295760_6 | Lynch median | 0.96 | 0.90 | 0.97 | 0.87 | 1.04 | 0.50 | 2.00 | 1.21 | 0.97 | 1.05 | 1.01 | 1.02 |
| | 0261722_8 | Lynch mean | 1.11 | 0.88 | 1.15 | 0.86 | 1.07 | 0.69 | 1.06 | 0.90 | 1.01 | 1.22 | 1.14 | 1.12 |
| | 0295405_7 | Pegram mean | 0.97 | 0.98 | 1.05 | 0.90 | 1.06 | 0.88 | 1.06 | 1.06 | 1.18 | 1.02 | 1.12 | 0.94 |
| **C41B** | 0295405_7 | Dent median | 0.94 | 0.91 | 1.03 | 0.95 | 0.96 | 0.50 | 1.28 | 0.61 | 1.30 | 1.00 | 0.98 | 0.89 |
| | 0295405_7 | Lynch median | 0.92 | 0.93 | 1.06 | 0.87 | 0.92 | 0.50 | 1.67 | 0.80 | 1.42 | 1.00 | 0.98 | 0.84 |
| | 0261722_8 | Lynch mean | 1.10 | 0.89 | 1.16 | 0.86 | 1.05 | 0.73 | 1.10 | 0.91 | 1.06 | 1.18 | 1.13 | 1.12 |
| | 0295405_7 | Pegram mean | 0.96 | 1.01 | 1.07 | 0.91 | 1.04 | 0.89 | 1.06 | 1.05 | 1.13 | 1.00 | 1.07 | 0.93 |
| **C41C** | 0295760_6 | Dent median | 0.89 | 0.92 | 1.00 | 0.97 | 1.20 | 0.50 | 2.00 | 1.12 | 0.92 | 0.95 | 1.02 | 1.00 |
| | 0295405_7 | Lynch median | 0.85 | 0.88 | 1.09 | 0.90 | 1.05 | 0.50 | 1.67 | 1.20 | 1.42 | 0.94 | 0.92 | 0.78 |
| | 0261722_8 | Lynch mean | 1.02 | 0.86 | 1.16 | 0.90 | 1.08 | 0.76 | 1.18 | 1.05 | 1.05 | 1.12 | 1.08 | 1.05 |
| | 0295405_7 | Pegram mean | 0.95 | 1.00 | 1.04 | 0.92 | 1.04 | 0.97 | 1.02 | 1.18 | 1.07 | 0.96 | 1.02 | 0.86 |
| **C41D** | 0329166_5 | Dent median | 0.99 | 0.99 | 1.19 | 1.00 | 0.97 | 0.50 | 2.00 | 1.11 | 1.12 | 0.91 | 1.09 | 0.81 |
| | 0329166_5 | Lynch median | 0.95 | 0.94 | 1.15 | 0.85 | 0.92 | 0.53 | 2.00 | 1.78 | 1.26 | 0.92 | 1.02 | 0.75 |
| | C4E002 | Lynch mean | 1.01 | 1.01 | 1.14 | 0.97 | 0.94 | 0.85 | 1.20 | 1.20 | 1.52 | 1.06 | 1.08 | 0.85 |
| | 0261722_8 | Pegram mean | 0.99 | 0.83 | 1.07 | 0.88 | 1.03 | 0.88 | 1.16 | 1.07 | 1.05 | 1.08 | 1.04 | 0.98 |
| **C41E** | 0294500_X | Dent median | 0.85 | 0.98 | 1.15 | 1.01 | 1.03 | 0.50 | 1.18 | 0.50 | 0.92 | 0.93 | 1.21 | 1.01 |
| | 0329166_5 | Lynch median | 1.01 | 0.95 | 1.13 | 0.85 | 0.78 | 0.50 | 2.00 | 1.33 | 0.98 | 0.92 | 0.98 | 0.76 |
| | 0327883_9 | Lynch mean | 0.96 | 1.01 | 1.03 | 0.91 | 1.05 | 0.71 | 1.05 | 0.97 | 1.15 | 0.99 | 1.12 | 0.86 |
| | 0294500_X | Pegram mean | 1.01 | 0.92 | 1.02 | 0.88 | 1.02 | 0.80 | 0.89 | 1.05 | 1.00 | 1.03 | 1.20 | 1.03 |