Application of data mining and machine learning techniques for geohydrological datasets in South Africa

C de Bruyn

orcid.org/0000-0003-3011-8563

Dissertation accepted in fulfilment of the requirements for the degree *Master of Science in Environmental Sciences with Hydrology and Geohydrology* at the North-West University

Supervisor: Dr SR Dennis

Graduation October 2023

24963623

ACKNOWLEDGEMENTS

Firstly, I would like to thank YHVH, my Creator, and his Son, Yeshua, for giving me the strength to push onwards through this endeavour.

I am thankful to Dr Rainier Dennis and the Centre of Water Sciences and Management who granted me this opportunity and who guided me through this process, equipping me with the proper tools and knowledge.

To my parents, for all the love and support through this tough time. I would not have been able to pursue this degree, let alone finish it without them. Also, my brother for giving me advice on atopic which was completely new to me at the start of this study. And my fiancé, for encouraging me to finish what I started.

Finally, I am grateful for having a friend in Lohan Bredenhann, who supported me and gave technical advice regarding this academic pursuit.

ABSTRACT

A desktop study was conducted to research data-driven modelling techniques to classify relationships between borehole parameters and the relevant geological setting. Borehole surveying and drilling is a costly endeavour and by applying data mining and machine learning techniques to national groundwater databases and other available national datasets such as spatial data, better insight and improvements on management of groundwater resources can result.

Five machine learning algorithms were tested on a consolidated dataset and their performances compared in order to establish which algorithm yielded the most accurate results. It was established that Random Forest Regression and Classification could be used to model yield, and Support Vector Regression and Random Forest Classification could model static water levels. The algorithm was tested on three case study areas, based on Vegter regions.

The results indicated that static water levels could be modelled with high rates of accuracy, but yield modelling was not as successful, and a lot of uncertainty still remains as to the drivers behind water strike yield.

Keywords: data mining, machine learning, groundwater resource management, geohydrological datasets, data-driven modelling; water level modelling, yield modelling.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

ACC	Classification Accuracy
ANN	Artificial Neural Network
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSV	Comma Separated Value
DDM	Data-driven modelling
DT	Decision Tree
DWS	Department of Water and Sanitation
FN	False Negative
FP	False Positive
FRBS	Fuzzy Rule Based Systems
GIS	Geographic Information System
GRIP	Groundwater Resources Information Project
IBL	Instance-based learning
IDE	Integrated Development Environment
KDD	Knowledge Discovery in Databases
K-NN	K-Nearest Neighbour
MAE	Mean Absolute Error
MAP	Mean Annual Precipitation
MAPE	Mean Absolute Percentage Error
MLP	Multilayer Perceptron
MLR	Multiple Linear Regression
MODFLOW	Modular Finite-Difference Groundwater Flow Model
MS	Microsoft
MSE	Mean Square Error
NGA	National Groundwater Archive
NGDB	National Groundwater Database
RBF	Radial basis functions
RFC	Random Forest Classification
RFR	Random Forest Regression
RMSE	Root Mean Square Error
SVM	Support Vector Machines
SVR	Support Vector Regression
TN	True Negative
TP	True Positive
USGS	United States Geological Survey

CHAPTER 1: INTRODUCTION

This chapter introduces the research problem and the context in which the research took place, and it outlines the aims and objectives for the present project. It also serves as a 'road map' for topics that will be discussed.

1.1 Background

Water has been an indispensable resource since the dawn of time. Surface water was the first to be utilised, mainly for fishing and hunting. Upon the advent of agriculture and animal husbandry, ancient civilizations dating back to biblical times realised that more water would be required for the sustenance of an expanding agriculture. The first book of the Bible, Genesis, mentions that the patriarch Isaac dug wells with great success. Such interventions led to substantial growth in agriculture and irrigation, especially in the arid regions of southern Asia and northern Africa (Meinzer, 1934).

Villholth and Giordano (2007) note that surface water became more important over time and that the public was more occupied with surface water resources than groundwater. However, with the increase in awareness regarding water quality and quantity, the public interest in groundwater grew. Groundwater can be more accessible than surface water depending on geographic locations, and drilling and pumping techniques improved, making groundwater a favourable option in agriculture and industry. It should be noted, however, that the increase in usage of this resource generates the need for appropriate management practices to mitigate the plethora of groundwater problems that results from this (Villholth & Giordano, 2007).

Africa, especially sub-Saharan Africa (Taylor et al., 2009), is rapidly urbanising, and a tripling of the population is predicted from 2000 to 2050. This increase is directly proportional to the demand for easily accessible, potable water resources. Groundwater has become the primary source of water for most domestic households across Africa, since it offers a low-cost alternative, enjoys wide spatial distribution and has a generally potable quality. Taylor et al., (2009) note that reliable groundwater data are scarce in some parts, restricting the ability to formulate abstraction policies to manage the aquifers being abstracted from.

Large portions of Southern Africa fall within hyper-arid, arid, or semi-arid climate parameters (Xu & Beekman, 2019). Due to this, groundwater resources are of increasing importance, not least also since many regions in South Africa have been struck by severe droughts throughout history

and surface water resources have variable reliance. Groundwater ostensibly acts as a buffer between drought conditions and water supply, given that groundwater is generally reliable and exhibits resistance to hydrological droughts, generally has good quality, and is mostly free of organic matter, while it can typically be found in near proximity to where it is required (Allwright et al., 2013; Shirmohammadi et al, 2013).

Various sectors involved in development and management of the natural environment require yield estimates of boreholes to sustainably meet water demands and manage groundwater resources (Allwright et al, 2013). Shirmohammadi et al. (2013) note that groundwater overexploitation is a significant issue for developing countries and that groundwater level plays a key role in the sustainable yield of groundwater resources. The modelling of groundwater levels in South Africa is therefore vital for continuous management of sub-surface water resources.

Data generation is rapidly increasing in this digital era and a legion of discoveries are yet to be made in terms of this endless source. The process of computer-assisted analysis and extraction of new insights from large quantities of data are known as data mining or knowledge discovery (Babovic, 2005). Zaki & Meira Jr (2014) indicate that data mining is an interdisciplinary field that combines study areas such as database systems, machine learning, pattern recognition, and statistics. By combining data mining with the environmental sector, new patterns can be generated for analysis, and novel management techniques may be discovered for implementation. Zhu et al. (2022) discuss the application of machine learning in the water quality domain from modelling the movement of pollutants in surface and groundwater to management of water supply systems by predicting changes in water production given certain parameters as well as the monitoring of wastewater quality to streamline wastewater treatment plant (WWTP) management.

One infers that further insights may well be gained in die field of groundwater studies by utilising and mining massive databases such as the National Groundwater Archive (NGA) and other publicly available databases such as the Groundwater Resources Information Project (GRIP).

1.2 Problem statement

Physical surveying methods for groundwater supply in poorly understood regions are resource intensive and incur great costs, with high risk of being either unsuccessful, or not able to supply the water demand in full. This may lead to the need for additional borehole drilling. Drilling should primarily focus on areas with high probabilities of water bearing subsurface units with higher yields to ensure water demands will be met with minimal costs (Khan *et al*., 2023). Therefore, a need

exists to identify the most appropriate area easily and efficiently for successful water supply without the need to survey an extensive area.

Identifying potential surveying areas and drilling boreholes with a high rate of accuracy relative to the desired water level and water supply are key aspects when it comes to ensuring that this costly endeavour will have a valuable outcome. The NGA and GRIP databases are readily available for exploitation and, by researching data mining and machine learning techniques, a cost-effective analysis of data can be obtained that will aid decision making.

1.3 Aims and objectives

1.3.1 Aims

The principal aim of this study is to implement data mining and machine learning techniques to classify relationships between borehole parameters and their relevant geological settings. Drilling of boreholes is costly and, by mining the national databases, enhanced insight can be gained that will improve the management of boreholes and wellfields.

1.3.2 Objectives

The objectives of this study are as follows:

1. Compiling a single database containing borehole and other relevant information from the national groundwater datasets as well as data obtained from geographic information systems.
2. Testing data mining techniques and machine-learning algorithms on the created database.
3. Validating identified methods by applying these to predefined case studies where actual data are available for these.

1.4 Basic hypothesis

By applying data-mining and machine-learning techniques on borehole data, geohydrological characterisation of unexplored areas could be enhanced.

1.5 Scope of research

This study will mainly focus on predicting two aspects of groundwater, namely groundwater level and yield as these are the critical parameters when siting new boreholes. Although the prediction of actual aquifer parameters is not attempted, borehole yield and water level are related to the

afore mentioned parameters. Groundwater level and yield prediction as a result of in situ geological conditions is the main focus of this study. The study solely rely on exiting data in the national datasets and no additional data acquisition was undertaken.

1.6 Assumptions and limitations

The assumptions made in this study is that there exists an underlying relationship amongst parameters in the database that can be leveraged to successfully predict the parameters in question for this study, i.e., water level and yield.

The intrinsic limitation to any study is data. Since machine learning required an adequate dataset to extract relationships, a possible limitation is the completeness of the dataset used. In an attempt to reduce the impact of the identified limitation, the Limpopo dataset was targeted to test the methodology, as this is the most complete borehole dataset across a large area in South Africa.

1.7 Research contribution

Khan *et al*. (2023) notes that borehole drilling and surveying are very costly and optimal borehole location selection is key to ensure the sustainable management of this vital resource. The research is considered a building block in the development of a tool or a means of assistance to aid in the decision-making process to sustainably develop the groundwater supply. The application of machine learning to existing databases could expedite this process, delivering insights not previously known about areas, and ensure optimisation of field surveys and reducing costs associated with well field development.

1.8 Dissertation structure

The dissertation is structured as follows:

1. Chapter 1: Introduction
    a. General introduction describing the background of why the research was done, the problem statement, scope of work, and the aims and objectives of the study.
2. Chapter 2: Literature Review
    a. Data-mining, machine learning and geohydrological modelling background is briefly discussed to familiarise the reader with what data-mining can achieve in the context of geohydrology and the different types of modelling that is used in the sector.
    b. Various data-driven techniques are explored and discussed.

   c. Statistical evaluation and metrics of accuracy of models are discussed.
   d. Geohydrological characterisation is briefly discussed along with studies conducted relating to the topic of machine learning and the geohydrological study field.
3. Chapter 3: National Groundwater Datasets
   a. Three databases, namely the NGA, GRIP and GIS, were discussed on the context of data quality and data availability.
4. Chapter 4: Methodology
   a. A methodology was compiled based on the literature review findings and discussed along with assumptions and limitations.
5. Chapter 5: Case Studies
   a. Three study areas were selected, and each discussed in the following contexts:
      i. Background, locality, and groundwater specific data analysis.
      ii. Water level predictions based on the methodology.
      iii. Yield predictions based on the methodology.
6. Chapter 6: Results and Discussion
   a. The findings obtained during application of the methodology to each study area are discussed and validity of the hypothesis is examined.
7. Chapter 7: Conclusions and Recommendations
   a. A conclusion is reached based on the results and as it relates to other studies.
   b. Recommendations are discussed for future research.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The appropriate management of freshwater resources, especially groundwater, is critical in order to attain the maximum value from these assets without crippling their sustainability (Maliva, 2016). Maliva (2016) suggests that numerical groundwater modelling is a crucial tool for evaluating and managing groundwater. Kenda *et al*. (2018) note that the conventional process-based models, as stipulated by Maliva (2016), rely on prior and intricate knowledge of the aquifer dynamics so that extremely specific sets of data are required. These factors have caused a shift towards data-driven modelling.

Computer-assisted analysis such as machine learning and data mining are powerful and useful tools in any scientific field where vast amounts of data are available. Sahoo *et al*. (2017) comment on the intricacy of accurately modelling a system with complex underlying physical processes due to the substantial amounts data needed for development and calibration, as is the case with geohydrological modelling. Therefore, it has become popular to explore data-driven modelling techniques, that is, machine learning, to interpret large datasets without prior or deep knowledge about the subject matter. This literature review aims to investigate the topic of machine learning and the popular algorithms used for predictive analysis and how it has been and can be used in the field of geohydrological modelling.

2.2 Data mining

Vast amounts of new data are being generated every day as a result of increased internet usage, business related services, surveys, academic studies, and the progress in storage and connection of technology (García et al., 2015). These datasets are too massive for manual analyses, and this has led to a need for gathering useful information and structured knowledge through the utilization of data mining (García et al., 2015). Data mining is the practice of detecting underlying patterns in data through data acquisition and preparation as well as processing by means of mathematical or statistical techniques and, finally, analysis (Aggarwal, 2015; Larose, 2005). Hand et al. (2001) define data mining as ‘the analysis of (often large) observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner’.

Data mining examines data to discover previously unseen patterns and plays a critical role in knowledge discovery in databases (KDD) (Bramer, 2016; García et al., 2015; Hand et al., 2001). Associations and summaries engendered by data mining are referred to as models or patterns, which include linear equations, rules, clusters, graphs, tree structures, and recurrent patterns in time series (Hand et al. 2001). Data mining is normally applied to secondary data, meaning that the data used was primarily collected for another purpose (Hand, 2013).

Data mining is a pipeline process with approximately six phases (Larose, 2005; Aggarwal, 2015), the latter which will be discussed later. According to Larose (2005), data mining requires something analogous to a standard operating procedure. This is known as cross-industry standard process for data mining, or CRISP-DM, which is illustrated in Figure 2-1. García et al. (2015) however note that the steps of data mining are different for each individual or project. Nonetheless, CRISP-DM offers a good framework for structuring a data-mining project.

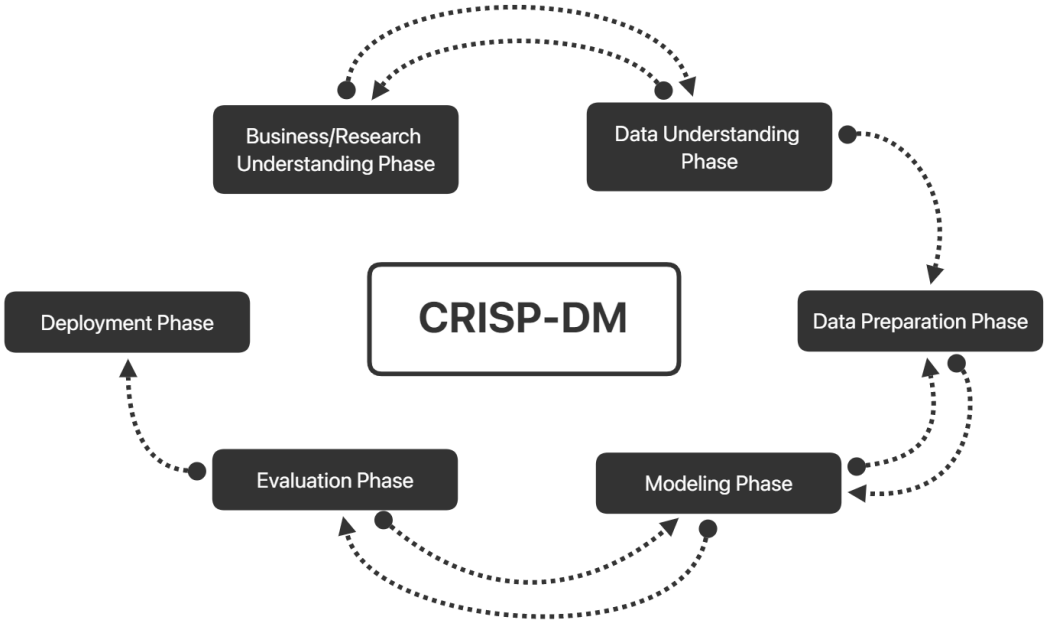


Figure 2-1: CRISP-DM standard process (adapted from Larose (2005)).

Based on Aggarwal (2015), who divided the workflow into three steps, García *et* al. (2015), who incorporated a hybridised version of the KDD process, and the CRISP-DM standards (Larose, 2005), an aggregated workflow is presented below.

Problem statement or understanding phase, the stage in which the problem is identified and the method of application determined. Proper understanding of relevant concepts is attained

through research and integrated into the problem-solving process. García *et* al. (2015) note that relevant prior knowledge, as obtained from experts, is vital to the reliability of the application.

Data collection is the process in which data are collected from various sources and arranged into databases for processing. Although this step might seem affable, it is very important to make good selections, as the data chosen will considerably impact the data-mining process.

Data preparation includes cleaning of noisy and erratic data, combining multiple data sources by means of a data dump, implementing data transformation where data are converted into standard useable formats for the chosen data-mining method, and data reduction. Aggarwal (2015) terms this phase feature extraction and data cleaning. Data are seldom sourced in a 'ready to use' form, that is, functional for data mining algorithms. It is a crucial step in the process so as to ensure that data used in the mining process are useful. Data-mining-appropriate formats include multidimensional, time series, or semi-structured data. Missing and incorrect data may be cleaned by estimation or correction or these may be omitted from the set.

Modelling is the stage where new information and patterns are derived from the data by using proper methods. These include choosing the appropriate data mining task, and Bramer (2016) notes four main types of data mining: numerical prediction, clustering, classification, or association. Consider that various models may be built. Once adequate techniques have been selected, parameter calibration can be performed and the model must continuously be validated.

Evaluation is a critical step for determining the quality of the results from the preceding stage and test the validity of the created models. In other words, circle back to the initial phase and insure that all objectives have been met by the model results (Larose, 2005).

Result exploitation is the direct application of the knowledge gained, integrating it for another purpose, or creating tools for others to use.

2.2.1 Datasets

Hand *et* al. (2001) define datasets as 'a set of measurements taken from some environment or process'. A dataset consists of a group of n objects (entities, individuals, cases or records) where, for each object, the same p measurement is structured in an $n \times p$ data matrix. Multiple p measurements could be taken and are known as variables or attributes (Hand *et al*. 2001). The form in which data is available will be specific to each situation: nevertheless, there are distinctions

to be made. Hand *et al*. (2001) and Aggarwal (2015), for instance, differentiate between quantitative and categorical data.

Quantitative attributes, also termed continuous or numeric ones, can be measured on a numerical scale and, depending on the type of measurement it represents, can be any value (Hand *et al*. 2001). Aggarwal (2015) notes that some numerical values are numeric in the sense of having a natural order. Quantitative data are the most common type and also the most useful to work with from a statistical standpoint, as numerous mathematical calculations can be done with such data. Any other type of data may not necessarily represent a numeric value which, in turn, makes it more difficult to incorporate these into a dataset that will be usable for an algorithm (Aggarwal, 2015).

Categorical attributes can only consist of discrete values (Hand *et al*., 2001). Santner and Duffy (1989) and Hand *et al*. (2001) comment that the measurement scales of discrete data find themselves on the ordinal or nominal scale. Ordinal scales categorise data into groups and orders within the group, meaning that the data can have a natural order such as low/ medium/ high) whereas nominal scales merely categorise data into groups where no particular order is present, that is, involving only discrete categories such as true of false (Santner & Duffy, 1989, Hand *et al*., 2001). Aggarwal (2015) notes that, more often than not, categorical data can be of a binary nature, meaning that only two categories are present. This can be converted into useable numeric values for an algorithm in the form of 0 or 1. Table 2-1 reflects these distinctions: columns in purple are quantitative variables, whereas columns in blue are examples of categorical variables.

Table 2-1: Difference between continuous and categorical data. Excerpt from the UCI Machine learning repository dataset 'Adult' (Kohavi & Becker, 1996)

age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income threshold
39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<= 50k
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<= 50k
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<= 50k
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<= 50k
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<= 50k
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<= 50k
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<= 50k
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	> 50k
31	Private	45781	Masters	14	Never-marries	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	> 50k
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	> 50k

Hand *et al*. (2001) note that data can occur in various relationships and configurations. Data could be arranged sequentially in time series or they can describe spatial relationships. In the case of the former, data mining might address the entire time series or just a section thereof, whereas the latter considers singular instances only in the context of others. Structures of datasets play an integral part in data mining. Complex data structures require complex models and algorithms (Hand *et al*., 2001).

2.2.2 Data-mining methods

Thus far, we have ascertained that data mining aims to establish patterns within large datasets in order to gain a deeper understanding of the particular data. Various means by which this can be accomplished exist and are used for different applications. Machine learning is one such method that will be explored. Figure 2-2 demonstrates the various disciplines intersecting within the computer science and statistics field as originally illustrated by Mitchell-Guthrie (2014).

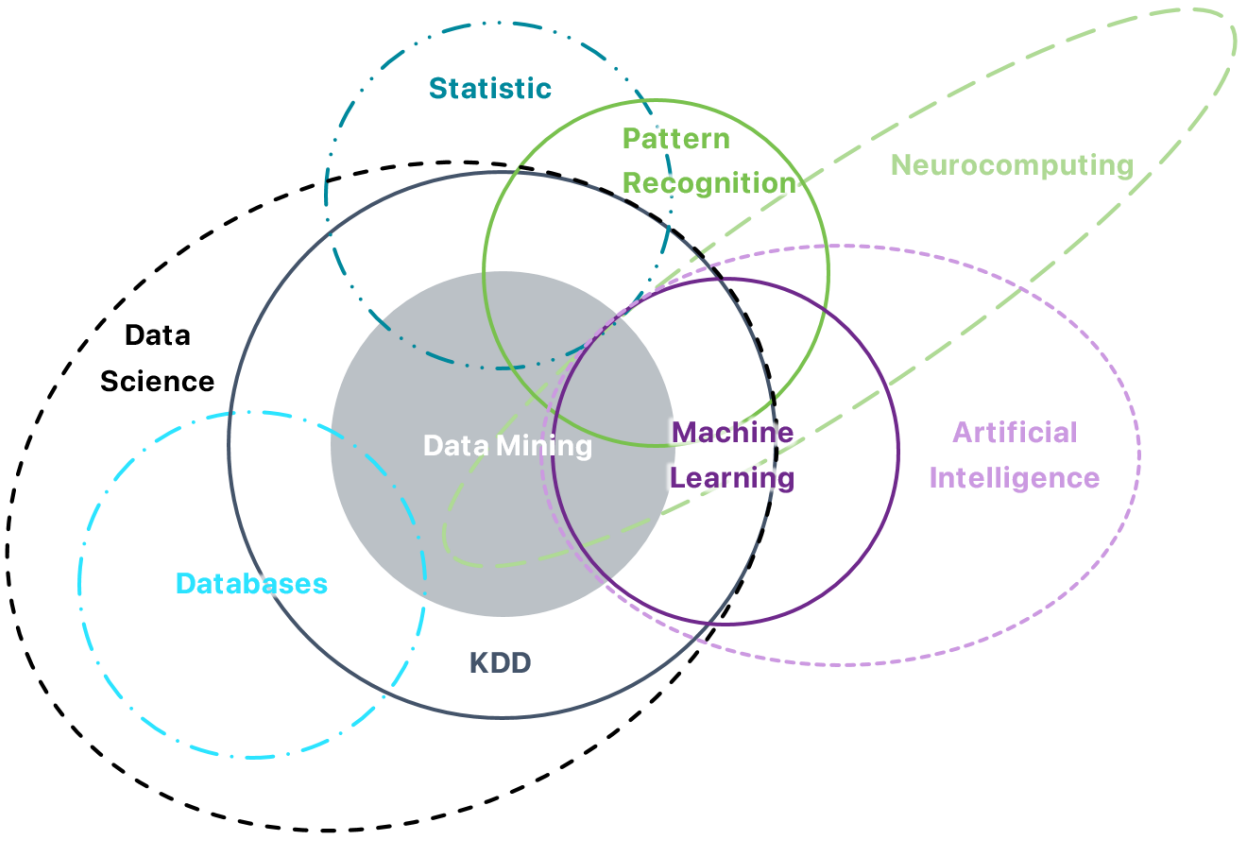


Figure 2-2: Intersections of disciplines that influence data mining and machine learning (adapted from Mitchell-Guthrie (2014)).

Teng and Gong (2018) define machine learning as learning techniques that automate the process of gaining knowledge. Therefore, data mining is the study of gaining knowledge and machine learning is the method used for the acquisition of said knowledge.

Data mining can be divided into two major method categories: those relating to prediction and those relating to description (García *et al*., 2015). Figure 2-3 below illustrates reflects the different techniques available for prediction and description.

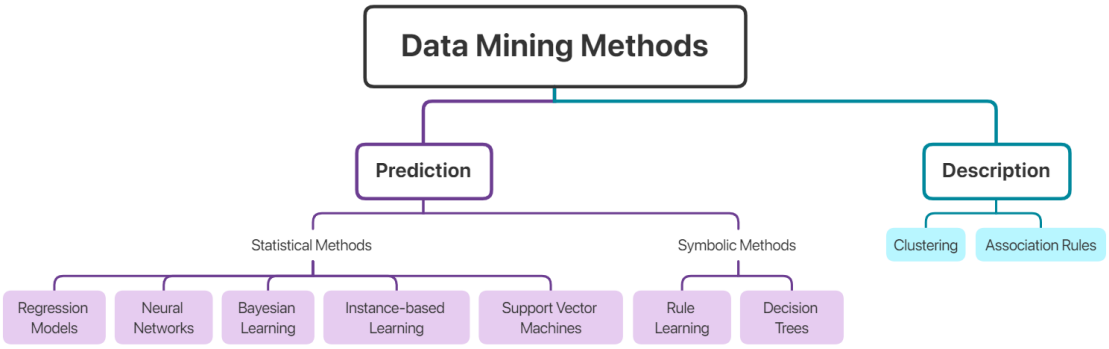


Figure 2-3: Data-mining methods (adapted from García *et al*., 2015).

For this study, the feasibility of the following prevalent machine learning techniques and algorithms will be researched:

1. Decision trees
2. Bayesian classifiers
3. Neural networks
4. K-Nearest neighbour
5. Support vector machines
6. Linear regression
7. Fuzzy logic

2.3 Modelling and forecasting of geohydrological settings

According to Wheater *et al*. (2007), a model is 'a simplified representation of a real world system'. Goltz and Huang (2017) define a model as an approximation of a system based on assumptions and simplified. Devi *et al*. (2015) state that the purpose of modelling is to better understand underlying processes of systems and predicting their behaviour, which is reiterated by Goltz and Huang (2017).

Goltz and Huang (2017) note that a system such as that of the subsurface is tremendously challenging to model due to the numerous unknowns and uncertainties that accompany its existence. Desirable models are those that simulate reality best with the minimum amount of parameters and model complexity. Therefore, if an understanding of governing processes can be gained, geohydrological modelling will be an essential tool for water-resource management in complex systems (Devi *et al*., 2015).

2.3.1 Model types

Hydrological and geohydrological models can be broadly categorised into empirical models, conceptual models, and physically based models (Wheater *et al*., 2007; Goltz and Huang, 2017). Physical models are small-scale representations of reality, whereas conceptual models are based on a theory or perceived logic of the system (Goltz and Huang, 2017). Elefteriadou (2014) notes that the difference between empirical models and mathematical models is that the former is based on field observations, whereas the latter is based on mathematical equations that describe relationships in the system. Solomatine and Ostfeld (2008) propose a condensed approach and classify models into process-based or data-driven ones, which will be discussed in brief further detail below. Oyebode *et al*. (2014) note that, due to commonly applied process-based techniques for modelling hydrological settings and scenarios, data-driven techniques have not been entirely incorporated into the field of hydrology. Sun *et al*. (2022) state that, although data-driven groundwater models have become increasingly popular at small scale, not enough research is being done at the local or regional spatial scales.

2.3.1.1 Process-based modelling

Process-based modelling, also referred to as ‘knowledge-driven’ modelling, is based on comprehensive descriptions of hydrological processes and first-order principles of physics. These are conceptual and physically based models (Solomatine & Ostfeld, 2008).

Wheater et al., (2007) discuss conceptual models, and notes that these are based on prior information in the form of conceptual representations of processes that are deemed to be important. The model must be calibrated in accordance with observed data of the catchment of interest to obtain a set of parameters that characterise the catchment being modelled. Physically based models rely on catchment processes and equations of motion that are numerically solved by using a grid (Wheater et al., 2007).

Modelling is extremely complex and requires wide-ranging levels of detail, causing the resultant models to simplify some processes; it will suffer from mis-calibration, over-parameterization, parameter instability, insensitivity or redundancy, high computational requirements, and huge data demand (Oyebode et al., 2014). The most popular process-based groundwater model is MODFLOW. The modular finite-difference groundwater flow model (MODFLOW) was developed by the U.S. Geological Survey (USGS), and its feature simulations include water flow of confined and unconfined aquifers and recharge from precipitation, evapotranspiration, rivers, and streams (Provost et al., 2009). There are a variety of versions of MODFLOW, where each focuses on a different area of specialisation (Kumar, 2019).

With a view to the massive amounts of data required for process-based modelling, Wheater et al. (2007) note that considerable computing power is needed to run the modelling task. Depending on the model used, a certain amount of computing power and time are needed. Therefore, the need arises to achieve the result of physically based models by means of quicker development and ease of use (Oyebode et al., 2014), and data-driven modelling could be a possible solution.

2.3.1.2 Data-driven modelling

Data-driven modelling (DDM) aims to establish correlations between input variables and output objective data by means of statistical regression, and it does not take into consideration any physical processes of the modelled system (Jing *et al*., 2022). Therefore, the input data are analysed to characterise a system with a limited assumption in order to establish connections between the input and output variables. These include statistical models and machine learning methods (Solomatine & Ostfeld, 2008; Oyebode *et al*., 2014). Solomatine and Ostfeld (2008) discuss the advantage contemporary methods over empirical modelling, noting that the former solves numerical prediction problems, allows for recreating nonlinear functions, classification and grouping of data, and the construction of rule-based systems.

Solomatine and Ostfeld (2008) note that the water resource community have reservations about the relevance of data-driven models as these are not associated to the physical principles of the system. While traditional statistical models are considered accurate enough, every situation is unique, and the most adequate model must be selected (Solomatine & Ostfeld, 2008).

DDMs include artificial neural networks (ANNs) comprising the multilayer perceptron (MLP), radial basis functions (RBFs), fuzzy rule-based systems (FRBSs), instance-based learning (IBL), tree-based methods, evolutionary computational methods (gene expression programming), and support vector machines (SVMs) (Oyebode et al., 2014; Solomatine & Ostfeld, 2008).

2.4 Data-driven modelling techniques

Sivakumar and Berndtsson (2010) note that one of the foremost reasons for the escalation of mathematical techniques in science is the advancement in computer and measurement technology. Increases in computational power and data of higher quality have turned data-driven modelling into a preferred approach (Sivakumar & Berndtsson, 2010). This section will discuss the different data mining techniques that are listed in Section 0 above.

2.4.1 Decision tree (model trees)

A decision tree (DT) uses the structure of a tree and its branches to represent possible decision paths and their respective outcomes (Grus, 2015). Larose and Larose (2019) defines a decision tree as containing a set of decision nodes, linked by branches spreading out towards a terminating leaf node, as depicted in Figure 2-4.

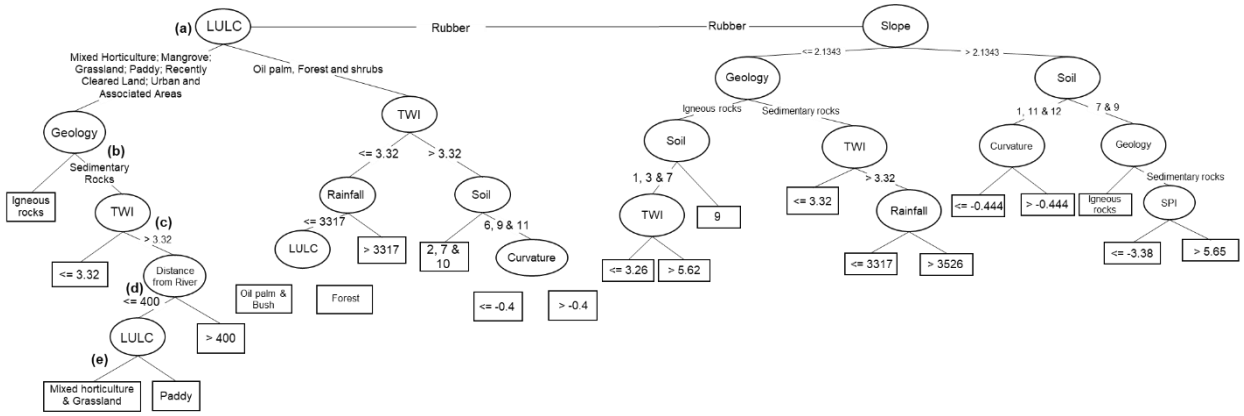


Figure 2-4: Example structure of a decision tree (Tehrany *et al*., 2013).

Larose and Larose (2019) explain that the purpose of a decision tree is to terminate in a set of leaf nodes where the records contained in each leaf node has the exact same classification. Root nodes are placed at the top of the decision tree, whereby variables are tested at the subsequent decision nodes, and each outcome results in a branch. Branches can either lead to another decision node or a terminating leaf node (Larose & Larose, 2019). Grus (2015) divides decision trees into classification trees and regression trees which return categorical and numerical outputs respectively.

Advantages in using decision trees is that the process by which they classify data is immediately apparent to the user and therefore make them easy to understand and interpret. Decision trees can handle mixed attributes, such as quantitative, categorical, and missing ones with ease (Grus,

2015; Hand *et al*., 2001). Hand *et al*. (2001) also note the speed by which they can classify new records and that they are powerfully predictive tools, since they are flexible.

Hand *et al*. (2001) discuss the problem of overfitting training data in decision trees. This occurs when the splitting of the decision nodes continues until a leaf node only contains a distinct data point or data points with identical input variables. García (2015) argues that noisy training data could impact the overfitting of a decision tree, and suggests an algorithm such as C4.5 that uses pruning strategies to reduce overfitting. C4.5 is one of many decision tree algorithms, and Saha (2018) indicates that it has significant advantages over other decision tree algorithms as it mitigates overfitting, can be used for both classification and regression, and can process incomplete data.

2.4.2 Naive Bayes / Bayesian Classifiers

The Bayes and naive Bayes classifiers are used for probabilistic classification tasks and makes use of the Bayes theorem (Zaki & Meira Jr., 2014). Joyce (2019) defines Bayes' theorem as a mathematical formula calculating conditional probabilities. Conditional probability could be described as 'the probability of a hypothesis *H* conditional on a given body of data *E* is the ratio of the unconditional probability of the conjunction of the hypothesis with the data to the unconditional probability of data alone' (Joyce, 2019). Zaki & Meira Jr. (2014) explain that the Bayes classifier uses the Bayes theorem to predict the class based on the label that maximises the probability. Grus (2015) explains the mechanism of Naive Bayes by example of spam filtering. Event *S* is 'the message is spam' and event *V* is 'the message contains the word *Earn $*'. Bayes' Theorem predicts the probability *P* that spam messages contain the word '*Earn $*' using Equation (2-1).

$$P(S \mid V) = [P(V \mid S)P(S)]/[P(V \mid S)P(S) + P(V \mid \neg S)P(\neg S)] \quad (2\text{-}1)$$

Zaki & Meira Jr. (2014) note that the full Bayes classifier ineffectually deals with datasets with large number of dimensions, while it suffers from estimation-related problems according to them. Naive Bayes is a surprisingly effective classifier due to the simple assumption that is made, namely that all the attributes of the dataset is independent. This is the key difference between the Bayes classifier and the naive Bayes classifier.

2.4.3 Artificial Neural Networks

Neural networks aim to replicate the data processing and non-linear learning abilities of a network of neurons (Rojas, 1996; Larose & Larose, 2019). Neural networks are models based on the neuron structure of the human brain and are used for cognitive tasks such as learning and optimisation (Müller *et al*., 1995). Larose and Larose (2019) explain that the basic function of a neuron is to gathering inputs from other neurons or, in case of artificial neural networks, a dataset, combining n inputs by means of a combination function, producing a non-linear response by means of an activation function, and sending it forward to other neurons. Grus (2015) notes that a neuron fires only when the calculation exceeds some threshold, that is: if the activation function produces an adequate response, it will send the response forward; otherwise, it will produce no output.

Neural networks are used on account of their forecasting and classification abilities, non-parametric nature, and capacity to generalize (Gaur, 2012). Although artificial neural networks are robust for using complicated non-linear data (Larose & Larose, 2019), they do not lend insight into how exactly they are solving the problem (Grus, 2015; Rojas, 1996).

2.4.3.1 Structure of an artificial neural network

Larose and Larose (2019) explain that the structure of an artificial neural network (ANN) consists of nodes, layers, connections, and weights (Figure 2-5). Layers contain nodes, where every node connects to every other node in the next layer. Nodes within the same layer, however, remain unconnected. Connections between nodes has an associated weight (W_{1A}) that is arbitrarily allocated a value between 0 and 1 at initialisation (Larose & Larose, 2019).

Müller *et al*. (1995) define neural network models as a directed graph with four distinct properties: 1) a variable (n_i) associated with each node i; 2) links (ik) between nodes i and k that have an associated real-value weight (w_{ik}); 3) a real-valued bias (ϑ_i) for each node i; and 4) a transfer function ($f_i[n_k, w_{ik}, \vartheta_i(k \neq i)]$) defined for each node i.

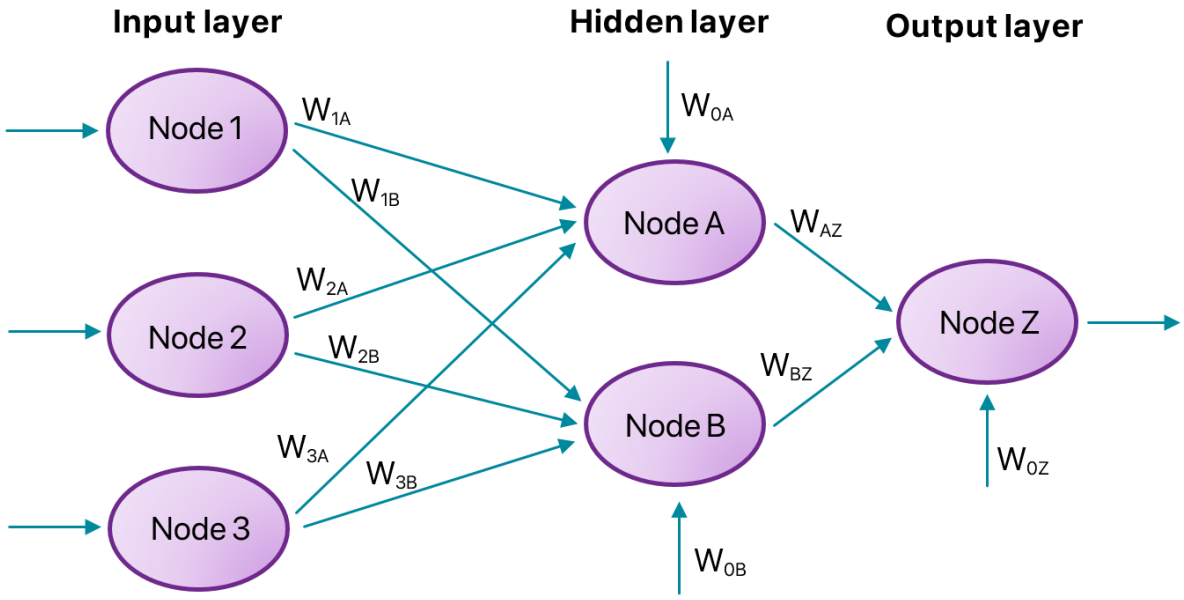


Figure 2-5: Basic structure of a neural network (Larose & Larose, 2019).

According to Gaur (2012), neural networks can be divided into three categories: feed-forward networks, feedback networks, and self-organisation networks. Feed-forward networks are predominantly used for prediction and pattern recognition (Gaur, 2012), and this is the focus of the present section. Feedback networks are largely used for associative memory and optimisation calculation, whereas self-organising networks are used for cluster analysis. Feed-forward networks only allow for single direction flow from input towards output, without the possibility of looping (Larose & Larose, 2019). Rojas (1996) indicates that, in the absence of cycling (looping), results of the computation is overt and no synchronisation of the computing units are necessary.

As mentioned, and as stated also by Mijwel (2018), ANN does not lend any insight into the behaviour of the system being modelled. Mijwel (2018) further explains the disadvantages of ANN: 1) the networks are excessively dependent on hardware and processing power, 2) the optimum result is not necessarily achieved due to the duration of the network, and 3) the experience of the user influences the integrity of the structure.

2.4.4 K-Nearest neighbours

K-Nearest neighbour (K-NN) methods belong to the class of instance-based learning (IBL) algorithms (Oyebode *et al*., 2014). Nearest-neighbour methods are fairly simple and rely on the principle of predicting a new data point by only considering those closest to it (Grus, 2015).

Oyebode *et al*. (2014) explain that IBL algorithms store information from training samples, while the information is subsequently applied to render a classification for a new instance. This is possible due to the retrieval of relevant information from the nearest neighbours (Oyebode *et al*., 2014). To classify a data point with similar input vectors as those of the adjacent points, the *k* nearest points are examined for a particular input vector and assign the new point to the class majority (Hand *et al*. 2001).

Closest data points are calculated by using Euclidean distance, which is the measurement of the proximity of a feature vector of a specified distance and a training samples' feature vector. Hand *et al*. (2001) explain that k-nearest neighbour is based on probabilities. In Figure 2-6, the centre circle represents a new data point, while squares and triangles are training data consisting of two distinct classes. The solid circle indicates $k = 3$, therefore the three nearest points are used to classify the new point. The dotted circle indicates $k = 5$, then the five nearest points will be used for classification (Alaliyat, 2008).

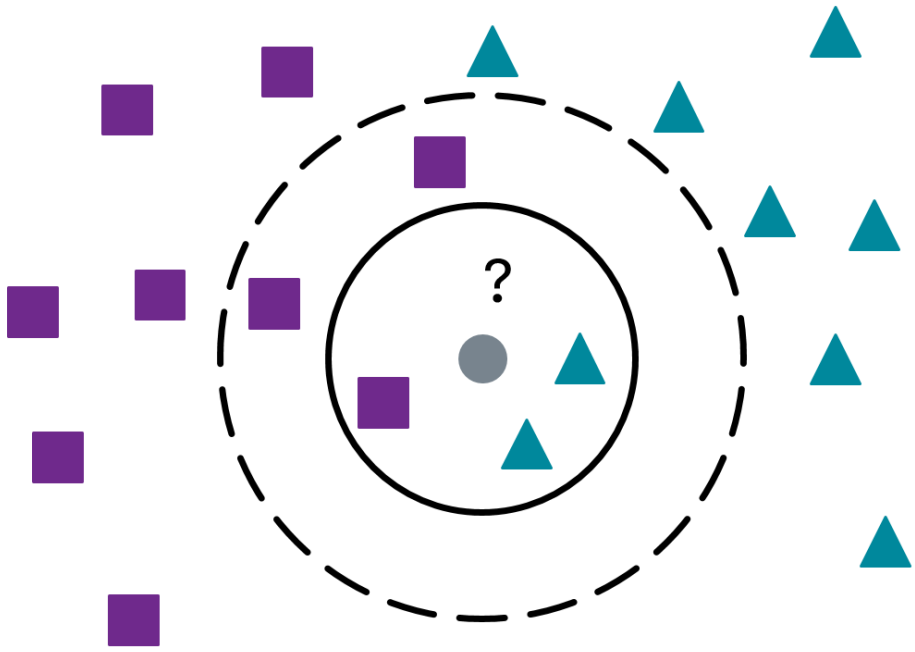


Figure 2-6: K-nearest neighbour illustration (Alaliyat, 2008).

Theoretically, a small portion of variables clustered around the new data point are used, with a radius equal to the distance to the k^{th} nearest neighbour. Subsequently, probability proportions are calculated for the likelihood of the point belonging to each possible class in the small portion.

The maximum probability class is then assigned to the new point (Hand *et al*., 2001). Class or labels can be true or false, which is predicated on there being a condition to fulfil; alternatively, they can be categorical (Grus, 2015).

Hand *et al*. (2001) explain the process of choosing a *k* value. At the very basic form *k* = 1, but this does not make for a stable classifier, as it has high variance. Reliable predictions can be made by steadily increasing *k*, keeping in mind the distance of points included with a higher *k* value, that is, it reduces variance but increases bias. Data-adaptive approaches appear to be the best technique for choosing *k*. Try several values, noting the misclassification rate of each, choosing *k* based upon the best performing value. The performance can then be verified on the testing data (Hand *et al*., 2001). Increasing dimensionality (adding variables) causes the data to become sparser which ultimately influences the true probability (Hand *et al*., 2001).

Grus (2015) notes that K-NN makes no mathematical assumptions: it only requires a distance aspect and the assumption that grouped points are similar. Therefore, it is easily programmable and requires no optimisation (Hand *et al*., 2001). Hand *et al*. (2001) also note that K-NN is well adapted to manage missing values.

High-dimensionality is a cause for concern in most models, and K-NN performs poorly with large amounts of variables (Hand *et al*., 2001; Grus, 2015). A potential consequence of using k-nearest neighbour is not knowing the drivers of the phenomenon or system being studied (Grus, 2015) as it does not build a model, but depends on recalling all training data (Hand *et al*., 2001). There are also the problems of computing time and storage requirements. For large training sets consisting of *n* data points, each data point is visited and *p* operations performed to calculate distance. This process requires considerable time and memory (Hand *et al*., 2001).

2.4.5 Support Vector Machines

In its simplest form, a support vector machine (SVM) is an algorithm that learns through example and assigns labels to new data points (Noble, 2006). Noble (2006) explains that an SVM is fundamentally a mathematical unit capable of capitalising on a mathematical function concerning a particular dataset. Neelamegam (2013) notes that the SVM is an effective method for classification, pattern recognition, and regression, due to its high generalisation capacity concerning input data with high dimensionality. Noble (2006) specifies that SVM classification consists of four basic concepts: 1) the separating hyperplane, 2) the maximum-margin hyperplane, 3) the soft margin, and 4) the kernel function.

The separating hyperplane is a linear separator passing through the dataset in order to separate two classes (Neelamegam, 2013; Noble, 2006). Russell and Norvig (2010) note that many separating hyperplanes can exist, as displayed in Figure 2-7a. Comparing logistic regression with SVM, one finds that logistic regression establishes a separating hyperplane based on all the data points, minimising the loss. Alternatively, SVM calculates the hyperplane to use based on a small selection of points that are considered to be more significant than the rest. Except for the support vectors, that is, the points closest to the hyperplane, all other points therefore have an associated weight of zero. This allows SVM to minimise generalisation loss and is known as the maximum-margin hyperplane (Russell & Norvig, 2010), as depicted in Figure 2-7b.

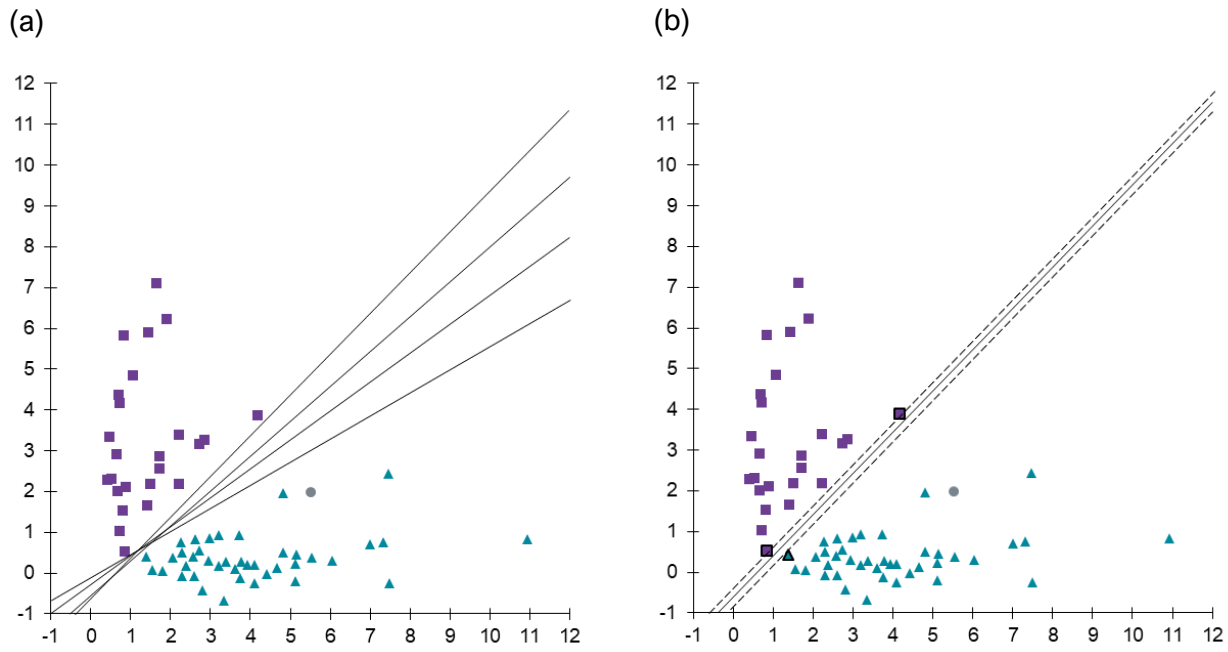


Figure 2-7: Support vector machine classification for a binary class problem. (a) Possible separating hyperplanes. (b) Maximum-margin hyperplane (Russell & Norvig, 2010).

Russell & Norvig (2010) argue that the purpose of the maximum-margin hyperplane is to minimise generalisation loss through selecting the hyperplane that is farthest away from the training data points. The margin is the width of the area between the dashed lines of Figure 2-7b. Choosing a hyperplane that is in close proximity to the purple squares, but further from the blue triangles, may result in a situation where testing data points of the purple class fall outside the decision boundary and are incorrectly classified as belonging to blue (Russell & Norvig, 2010).

However, this does not mean that all data can be faultlessly separated linearly without an anomalous example, as explained by Noble (2006). SVM algorithms can be adapted by means of

the use of a soft margin. A soft margin is a membrane of sorts that allows a certain number of data points to reside on the other side of the hyperplane without affecting the final result. The number of data points allowed across and the distance from the hyperplane must be specified by the user (Noble, 2006).

The final concept, and arguably the most useful, is the kernel function. Figure 2-8a illustrates a linearly inseparable dataset. The input data can be re-expressed and mapped to a new input space with appropriately higher dimension: subsequently, and two-dimensional data can be defined by three features, as indicated in Figure 2-8b (Russell & Norvig, 2010). Noble (2006) also notes that one-dimensional data (Figure 2-8c) can be mapped to a two-dimensional input space. This is illustrated in Figure 2-8d, where the original expression values have merely been squared, so that a linear distinction can be made between the purple and blue instances. Thus, the kernel function can mathematically project low-dimensional data into a high-dimensional input space (Noble, 2006).

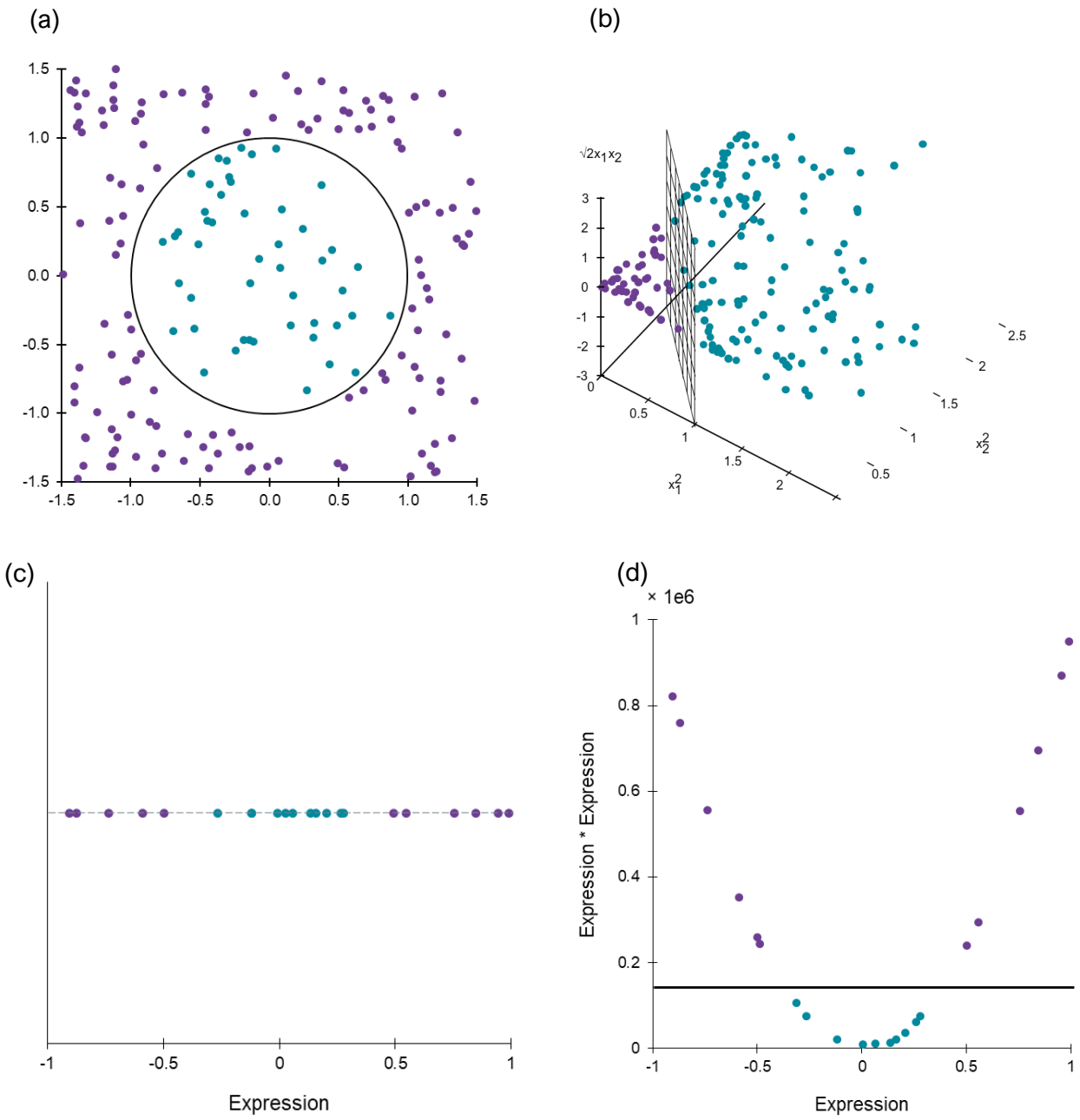


Figure 2-8: Support vector machine classification for a linear inseparable problem. (a) Two-dimensional dataset with a circular decision boundary. (b) The same dataset mapped into a three-dimensional space. The data takes on a cone shape and the circular decision boundary becomes linear. (c) One-dimensional dataset with no clear decision boundary. (d) Two-dimensional space due to applied kernel function (Russell & Norvig, 2010; Noble, 2006).

2.4.6 Linear regression

Montgomery *et al*. (2012) define regression analysis as 'a statistical technique for investigating and modelling the relationship between variables'. A model with unbent regression parameters is considered to be linear regression (Yan & Su, 2009). Linear-regression models can be either simple or multiple in nature, while other models include polynomial regression ones and nonlinear-regression ones. Simple linear-regression models are models with a single regressor and a response variable that form a straight line (Montgomery *et al*., 2012). Equation (2-2) is a simple linear-regression model.

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2-2}$$

Where β_0 = intercept
β_1 = slope
ε = random error component

Multiple linear-regression models are those that comprise two or more regressors, while their response variable may possibly be related to *k* regressors (Montgomery *et al*., 2012). Equation (2-3) is a multiple linear-regression model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{2-3}$$

2.4.7 Fuzzy Logic / Fuzzy rule-based systems (FRBS)

Zadeh (1988) states that 'fuzzy logic is concerned with the formal principles of approximate reasoning', as opposed to the exact reasoning used in classical logic systems, although precise reasoning could be used for limitation. In essence, fuzzy logic focuses on modelling imprecise reasoning in order to retrieve an estimated answer based on partial, inaccurate, or undependable knowledge. Fuzzy logic and fuzzy-logic-based process control has been implemented in a variety of ways, viz. automatic train operation, robot control, speech recognition, and stabilization control (Zadeh, 1988).

Classical logic systems fall short in two ways: firstly, they do not make available a structure in which the meaning of proposals articulated in the statement are characterised. Secondly, if meanings can be characterised by means of symbolic representation, there are no tools for interpretation. Fuzzy logic alleviates these problems by characterising variables from the

statement as elastic constraints, and the question is answered by inference from the propagation of the elastic constraints (Zadeh, 1988).

IF-THEN rules are central to fuzzy logic and fuzzy set theory, and information can be denoted in a system of IF-THEN rules where antecedents and outcomes are fuzzy rather than explicit (Aranibar, 1994). Set theory is a branch in mathematical logic pertaining to the theory of well-defined collections of objects (sets) where objects are called members of the set (Bagaria, 2019). Sammut and Webb (2017) define fuzzy sets as those that are distinguished by having a membership function that allocates a degree of membership to all objects in the set. Membership has a value in the range of [0, 1], where 0 is no certain membership and 1 is a certain membership, and all values in between signify partial membership. Thus, fuzzy logic can account for concepts that are more efficiently represented in a spectrum rather than a binary true-or-false classification (Sammut & Webb, 2017). Aranibar (1994) further notes that rules in the systems are activated relative to the membership function of the match between the antecedents and the input, allowing for basic interpolation due to the imprecise nature of the antecedents. This interpolation reduces the number of IF-THEN rules required to define the input-output relationship (Aranibar, 1994).

According to Sammut and Webb (2017), fuzzy systems are computing structures based on the concepts of fuzzy logic and fuzzy sets. These structures are partitioned into four main components: 1) a knowledge base, 2) a fuzzification interface, 3) an inference engine, and 4) a defuzzification interface. The knowledge base includes the fuzzy rules and a database defining the linguistic terms of each linguistic input and output variable. The fuzzification interface converts the precise input variables into imprecise fuzzy variables. This is achieved by assigning computed membership values to each variable according to the linguistic terms defined in the knowledge base. The inference engine computes the activation degree and the output of each rule defined in the knowledge base. The defuzzification interface does the inverse of the fuzzification interface by transforming the fuzzy variables into precise outputs (Sammut & Webb, 2017).

(Kapitanova, *et al*., 2012) note that a significant disadvantage of fuzzy logic is that it generates a large rule-base, which requires significant amounts of memory and processing power. The reason for this large rule-base is that the number of rules increases exponentially with the number of variables used.

2.5 Statistical evaluation/ model evaluation

Models have to be evaluated according to how accurately they perform and what the error rate is (Larose & Larose, 2019).

2.5.1 Metrics for regression

The most commonly used metrics to evaluate the performance of regression models include mean square error (MSE) or root mean square error (RMSE), mean absolute error (MAE) or mean absolute percentage error (MAPE), and R squared (R^2) or adjusted R squared (Wu, 2020; Brownlee, 2021).

2.5.1.1 Mean square error/ root mean square error

Mean square error is an outright measurement of how well the model fits the observed system (Wu, 2020). Mean square error is calculated by summing the square of the prediction error (observed – predicted) and dividing by the total number of data entries in the set (Wu, 2020). Melville and Sindhwani (2017) note that (R)MSE emphasises greater absolute errors.

Cichosz (2015) indicates that MSE has a slight disadvantage, namely the effect of changed scale due to squaring. This complicates that understanding of the errors. RMSE is simply the square root of the MSE value and is calculated by using equation (2-4) (Melville and Sindhwani, 2017). It is used more frequently due to its ease of interpretation, as the value is smaller (Wu, 2020). Cichosz (2015) explains that the monotonic nature of the root square function measures uniformly, therefore making RMSE and MSE virtually the same. The only difference occurs around the ease of interpretation (Cichosz, 2015).

$$RMSE = \sqrt{\frac{\Sigma_{\{i\}}(P_i - r_i)^2}{N}} \tag{2-4}$$

Where P_i = predicted value at i

r_i = observed value at i

N = total number of entries

2.5.1.2 Mean absolute error/ mean absolute percentage error

Melville and Sindhwani (2017) note that MAE is the most commonly used metric, which may be due to its straightforward nature (Cichosz, 2015). MAE is the averaged absolute difference between a set of observed values and predicted values and is given in equation (2-5) (Melville and Sindhwani, 2017). Wu (2020) states that MAE treats all errors in the same manner, as opposed to MSE, which squares errors to give larger penalisations to larger errors.

$$MAE = \frac{\Sigma_{\{i\}}|P_i - r_i|}{N} \quad (2\text{-}5)$$

Where P_i = predicted value at i

r_i = observed value at i

N = total number of entries

2.5.1.3 R square/ adjusted R square

R square, also known as the coefficient of determination, indicates the variance explained by the model (Akossou & Palm, 2013). Values closer to 1 indicate a perfect model fit, whereas lower to negative values indicate a poor to inadequate model (Cichosz, 2015). R squared is biased, as noted by Akossou and Palm (2013), because it gradually increases as new variables are added to the model. Adjusted R square was introduced because R square did not cater for overfitting; added independent variables were penalised to curb this (Wu, 2020).

2.5.2 Confusion matrix and associated metrics for classification

The confusion matrix is an appropriate starting point for evaluation, as it can accommodate 2-class or M-class classification problems and registers the associations between the classifier outputs and the actual label (Diez, 2018). Sirsat (2019) explains that the four outputs of a 2-class classification confusion matrix include true positive (TP), false positive (FP), true negative (TN), and false negative (FN), as illustrated in Figure 2-9a. TP represents the number of data points predicted correctly as positive, whereas false positives are the number of data points that have been incorrectly predicted as being positive where, in reality, they are negative. TN represents the number of data points predicted correctly as being negative, while a FN is the number of data points incorrectly predicted as negative where, in reality, they are positive (Sirsat, 2019).

(a)

Actual Class \ Predicted Class	Positive	Negative	
Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

(b)

Actual Class \ Predicted Class	1	2	3	4	Total
1	70	10	15	5	100
2	8	67	20	5	100
3	0	11	88	1	100
4	4	10	14	72	100

Figure 2-9: Confusion matrix structure (a) for a 2-class classification and (b) for a 4-class classification problem (Sirsat, 2019; Diez, 2018).

Diez (2018) explains that the M-class confusion matrix has elements n_{ij} – where i denotes row identifier and j the column identifier – that are indicative of cases correctly classified. Figure 2-9b depicts the correctly classified n_{ij} elements in diagonal shaded squares, whereas all other elements are misclassified. Even though the confusion matrix neatly displays all classifier output information, it is not convenient for comparison and discussion purposes (Diez, 2018). Therefore, additional metrics must be extracted from the confusion matrix.

The mathematically expressed metrics facilitate in-depth evaluation criteria for a model (Sirsat, 2019). Sirsat (2019) differentiates among sensitivity, specificity, accuracy, and precision. Sensitivity measures the TP rate, which is the positive data points labelled as positive. Ideally, the TP value should be greater than that of the FN value so as to ensure a high sensitivity (Figure 2-10a). Specificity measures the TN rate, which is the negative data points labelled as negative. As in the case of sensitivity, specificity should have a high value (Figure 2-10b). Precision is the proportion of the total number of correctly predicted positive data points and the total number of predicted positive data points (Figure 2-10c) (Sirsat, 2019).

Sirsat (2019) defines accuracy as the ratio of total number (probability) of predictions that are correctly predicted (Figure 2-10d), which can be calculated by using Equation (2-6) or Equation (2-7) as found in Diez (2018). Diez (2018) explains that classification accuracy (ACC) must be sensibly examined for the reason that it depends on the number of classes and cases. For instance, a 2-class classification problem will have a 50% chance of a case belonging in either class (Diez, 2018).

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (2\text{-}6)$$

$$ACC = \frac{\sum_{i=1}^{M} n_{ii}}{N} \quad (2\text{-}7)$$

Diez (2018) notes that Equation (2-7) has two drawbacks, namely that diagonal values are omitted and classes with reduced numbers of cases have a lower weight in the calculation. In contrast, Cohen's kappa coefficient (κ) utilises the entire confusion matrix and is calculated by using Equation (2-8) (Diez, 2018), where p_0 is classification accuracy (ACC).

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \quad (2\text{-}8)$$

Equation (2-9) is used to calculate p_e, a theoretical expected classification accuracy (Landis & Koch, 1977) where $n_{:i}$ is the sum of i-th column and $n_{i:}$ the sum of i-th row.

$$p_e = \frac{\sum_{i=1}^{M} n_{:i} n_{i:}}{N^2} \quad (2\text{-}9)$$

Landis and Koch (1977) partition the 0.00 to 1.00 scale of the kappa value into six categories, labelling each category as a measure of strength of agreement (Table 2-2).

Table 2-2: Kappa value partitioning and associated labels (Landis & Koch, 1977)

Kappa value	Strength of agreement
< 0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

Sirsat (2019) notes that the F1 score, that is, Equation (2-10), is a valuable measure to distinguish between models based upon their sensitivity and precision values and is calculated for each class or label.

$$F1\ Score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{2-10}$$

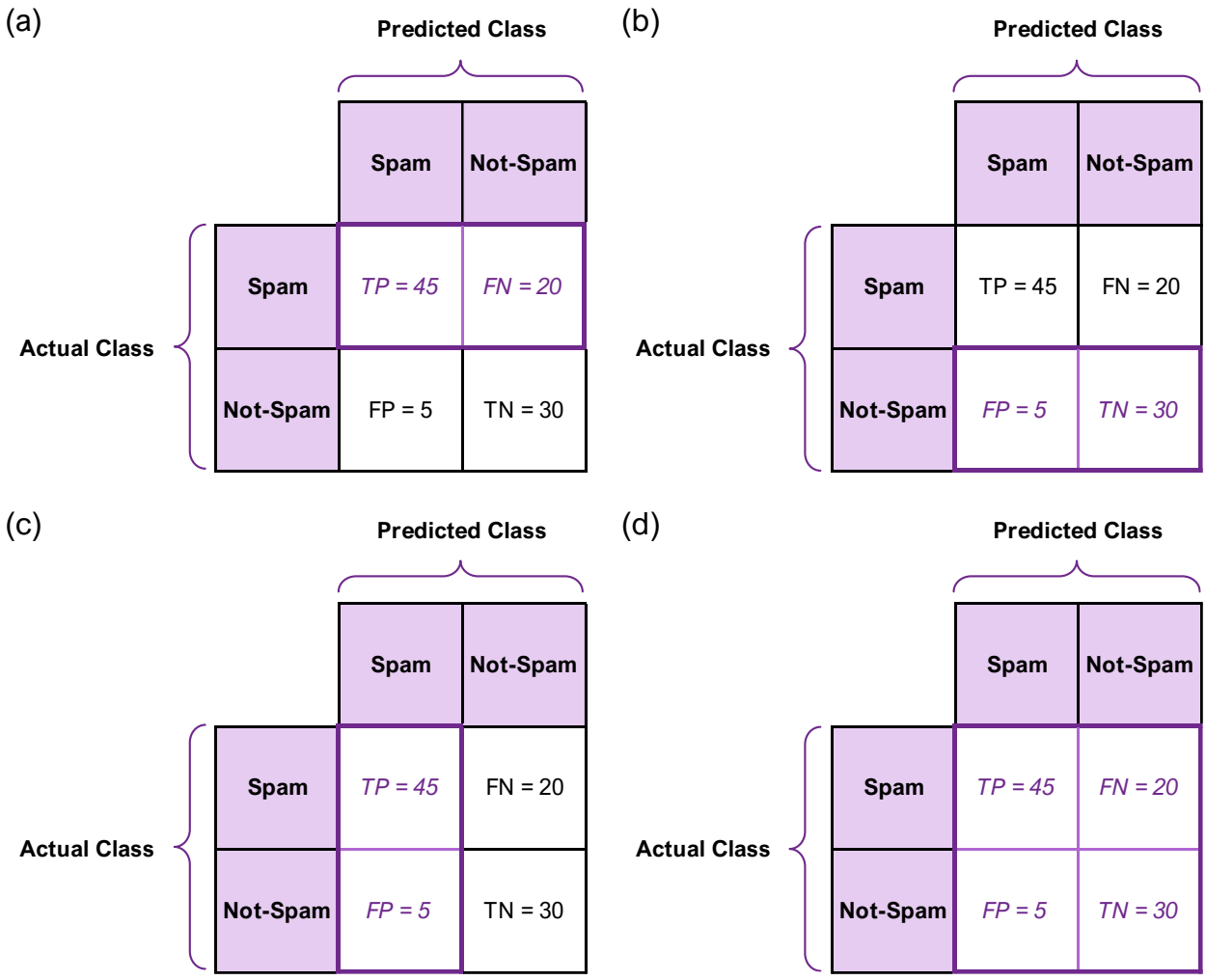


Figure 2-10: Example output confusion matrix of a spam filter. (a) Sensitivity, (b) specificity, (c) precision and (d) accuracy (Sirsat, 2019).

2.6 Borehole parameters / geohydrological characterisation

The present study focuses on water level and yield. Taylor and Alley (2001) and Kenda *et al*. (2018) state that water levels, particularly those found in observation wells, are primary indicators of hydrologic stresses that affect aquifers which, in turn, affect recharge, storage, and discharge. Sahoo *et al*. (2017) also comment on the importance of groundwater levels for sustainable planning and management.

Taylor and Alley (2001) note a variety of factors that may influence groundwater levels. These include aquifer types such as confined and unconfined ones, the balance of recharge, storage and discharge from the aquifer, and the physical characteristics of the aquifer-- the latter which include porosity, permeability, and the thickness and composition of the geological units. Climatic and other hydrological factors also play a significant role in groundwater level fluctuations, such as the intensity and duration of precipitation events, the extent to which baseflow contributes to surface water bodies, and evapotranspiration. Countless variables are present in the complex system of groundwater dynamics, and Sahoo *et al*. (2017) recognise this: therefore, the need arises to investigate possible solutions to gather more knowledge regarding the inner workings of and relationships among these variables.

One of the many imperative functions of a boreholes is water supply (Gaaloul *et al*., 2018). Yield is an important aspect to keep in mind in terms of a borehole used for water supply, which is dependent on the physical environment within which a borehole is located. Freeze and Cherry (1979) define groundwater yield as the maximum abstraction rate that may be allowed to safeguard against declining water levels where they are deemed to be unacceptable. Therefore, an indirect relationship exists between water levels and groundwater yield. As mentioned, fluctuations in water level indicate stresses on the system, which could impact water supply.

2.7 Geohydrological studies already conducted by using machine learning

Numerous studies focusing on different fields within hydrology and geohydrology have been conducted with the aim of modelling and forecasting groundwater level fluctuations by using data-driven techniques. Sahoo *et al*. (2017) made use of the Multilayer Perceptron network to model fluctuations of groundwater levels in agricultural regions in the United States. Kenda *et al*. (2018) used a variety of regression algorithms to predict groundwater level fluctuations in the Ljubljana Polje aquifer in central Slovenia. Arabameri *et al*. (2019) achieved good predictions of groundwater potential in Iran by using machine-learning algorithms and a variety of different input data.

Caté *et al*. (2017) used six machine-learning algorithms to predict the presence of gold in drill logs. Although this study does not interface directly with geohydrology, the same data-driven models may be used to characterise geological units that greatly influence groundwater. For instance, Bougher (2009) used gamma ray logs to train K-nearest neighbour classifiers to predict stratigraphic units.

The possibilities for the implementation of machine learning are nearly infinite when it comes to data-driven models. The input data can be altered to suit virtually any need, and the algorithms can be tweaked, adjusted, or completely changed to find the best fit for the purpose at hand. That is why research that employs machine learning and data mining is becoming popular.

2.8 Machine learning in the context of South African policy

Policies regarding water are widespread in South Africa. The research does not target any specific policy, but does support the National Water Act (RSA, 1998) in terms of sustainable use. Important principles of the National Water Act include the sustainable use water that allows for social and economic development, that every citizen of South Africa have access to water, and that the water being used must not be wasted.

The National Water Act (NWA) is the foundation of water management within South Africa, with a large focus on ensuring water availability to all users without stressing water resources, such as groundwater over abstraction. The NWA considers classification, reserve, international obligations, inter-basin transfers, strategic use, and future use before authorising any further water use. The machine learning techniques could assist in developing the groundwater supply in underdeveloped areas of good yielding aquifers to ensure the overall sustainability of the groundwater supply.

2.9 Conclusion

Extant literature evidences that cost-effective and timely predictions are key aspects when it comes to the management of water resources. Data-driven modelling in a geohydrological context is a worthy endeavour in this respect and could, with a respectable degree of probability, facilitate the surveying and drilling of new boreholes.

CHAPTER 3: NATIONAL GROUNDWATER DATASETS

This chapter examines the two main South African national groundwater datasets, namely the **National Groundwater Archive** (NGA) and the **Groundwater Resources Information Project** (GRIP). It is centred on data availability and discusses the available data and their quality. Spatial datasets that cover the whole of South Africa will also be discussed, as they may well contain critical data that are not necessarily captured by the other databases.

3.1 Data quality

Data quality is an essential part of research, as it determines the overall quality as well as the replicability of the results (Oliveira *et al*., 2005; Rosli *et al*., 2016). Batini and Scannapieco (2016) state that the term quality has been defined as the 'totality of characteristics of a product that bear on its ability to satisfy stated or implied needs' or 'fitness for intended use'. Rosli *et al*. (2018) note that any conclusions built upon poor quality data may be invalid. Therefore, it is crucial to gain understanding of that which data quality entails and how data can be scrutinised to ensure that these are of good quality.

Batini and Scannapieco (2016) observe that data quality is frequently associated exclusively with accuracy, while they point out that it also relies on data completeness, consistency, and currency. According to Rosli *et al*. (2016) and Rosli *et al*. (2018), countless research has been conducted on the basis of publicly available data repositories and that issues have been raised about the quality of these datasets and how to overcome these. Quality issues include noise, missing data, incorrect data, duplicate data, and inconsistent data. Pipino *et al*. (2002) define various data quality dimensions as presented in Table 3-1 below.

Table 3-1: Data quality dimensions (Pipino *et al*., 2002)

Dimensions	Definitions
Accessibility	the extent to which data are available, or easily and quickly retrievable
Appropriate amount of data	the extent to which the volume of data are appropriate for the task at hand
Believability	the extent to which data are regarded as true and credible
Completeness	the extent to which data are not missing and are of sufficient breadth and depth for the task at hand

Concise representation	the extent to which data are compactly represented
Consistent representation	the extent to which data are presented in the same format
Ease of manipulation	the extent to which data are easy to manipulate and apply to different tasks
Free-of-error	the extent to which data are correct and reliable
Interpretability	the extent to which data are presented in appropriate languages, symbols, and units, and clear definitions
Objectivity	the extent to which data are unbiased, unprejudiced, and impartial
Relevancy	the extent to which data are applicable and helpful for the task at hand
Reputation	the extent to which data are highly regarded in terms of their source or content
Security	the extent to which access to data are restricted appropriately to maintain their security
Timeliness	the extent to which the data are sufficiently up-to-date for the task at hand
Understandability	the extent to which data are easily comprehended
Value-Added	the extent to which data are beneficial and provide advantages from its use

Rosli *et al*. (2016) further explain the need for additional information known as metadata. These the purpose, meaning, and context of data, therefore facilitating a better understanding of the data in question. Metadata also aim to avoid any misinterpretation that could arise from data (Rosli *et al*., 2016).

3.1.1 Measuring Data Quality

Pipino *et al*. (2002) discuss three prevalent functional forms for the performance of objective assessments: simple ratio, minimum or maximum operation, and weighted average. The present project employs simple ratio, as the weighted average approach relies on a weighting factor assigned to a variable based on its overall importance, whereas the purpose of this study is to let the data mining algorithms detect the more important variables organically.

Simple ratio quantifies the ratio between desired or undesired outcomes and total outcomes, and are represented in the convention of 1 and 0 where 1 is most and 0 least desirable. Data quality

dimensions evaluated by using simple ratio are completeness, concise representation, consistent representation, ease of manipulation, free-of-error, and relevancy (Pipino et al., 2002).

Completeness can be categorised into three types: schema completeness, column completeness, and population completeness. *Schema completeness* is the most abstract perspective and measures the degree to which entries within rows and columns as a collective are complete and not absent. *Column completeness* is viewed from a data perspective and considers the absent values of individual columns within the table. *Population completeness* suggests that a column should contain a range of values entailing that, when values within the range are absent, the population is incomplete. Completeness of each of these three types can be calculated by using the ratio of incomplete items to total number of items, and subtracting the ration from 1, as found in Equation (3-1) (Pipino *et al*., 2002).

$$Completeness = 1 - \frac{Incomplete\ items}{Total\ number\ of\ items} \tag{3-1}$$

Consistent representation centres on identical data values that are represented in the same format throughout the entire table. Consistency can be measured by taking the ratio of violations (with regard to a consistency type) to total number of consistency checks and subtracting the ratio from 1, as found in Equation (3-2) (Pipino *et al*., 2002).

$$Consistency = 1 - \frac{Violations}{Total\ number\ of\ consistency\ checks} \tag{3-2}$$

The free-of-error metric portrays data accuracy and correctness and can be calculated by dividing the number of erroneous data units by the total number of data units and subtracting from 1, as found in Equation (3-3) (Pipino *et al*., 2002). Pipino *et al*. (2002) also note that clearly defined sets of criteria are required to establish that which is a data unit and subsequently that which will be an error for that data unit. Thus, a degree of precision must be specified to ensure a threshold so as to determine when a data unit is correct, erroneous, or tolerable in a certain circumstance.

$$Free\text{-}of\text{-}error = 1 - \frac{Erroneous\ data\ units}{Total\ number\ of\ data\ units} \tag{3-3}$$

3.2 National Groundwater Datasets and Data Availability

3.2.1 National Groundwater Archive

The web-enabled National Groundwater Archive (NGA), managed by the directorate of Surface and Groundwater Information, is a centralised database containing most of South Africa's boreholes, referred to as *geosites* by the NGA (NGA, s.a.(d)). The NGA was preceded by the National Groundwater Database (NGDB) comprising an estimated 225 000 boreholes (DWA, 2009) and registered users were allowed to capture information to further expand the database as well as update, view, and extract data (NGA, s.a.(d)). The NGA currently contains approximately 270 000 geosites. The purpose of the NGA is to assess national and regional groundwater resources with a view to the sustainable management of these valuable assets (NGA, s.a.(d)).

Consider that, according to a report by DWA (2009), it is not compulsory to capture geosite information on the database. Therefore, no incentive exists for users to upload valuable groundwater information that they might possess. The report by DWA (2009) notes that a common problem with the NGDB centres on the lack of quality. Few geosite records are complete, critical data such as pumping tests and aquifer information are scarce, and a noticeable decline is seen in geosite capture data (Figure 3-1). This is a major cause for concern regarding data mining, as vast amounts of data are preferable to conduct analysis.

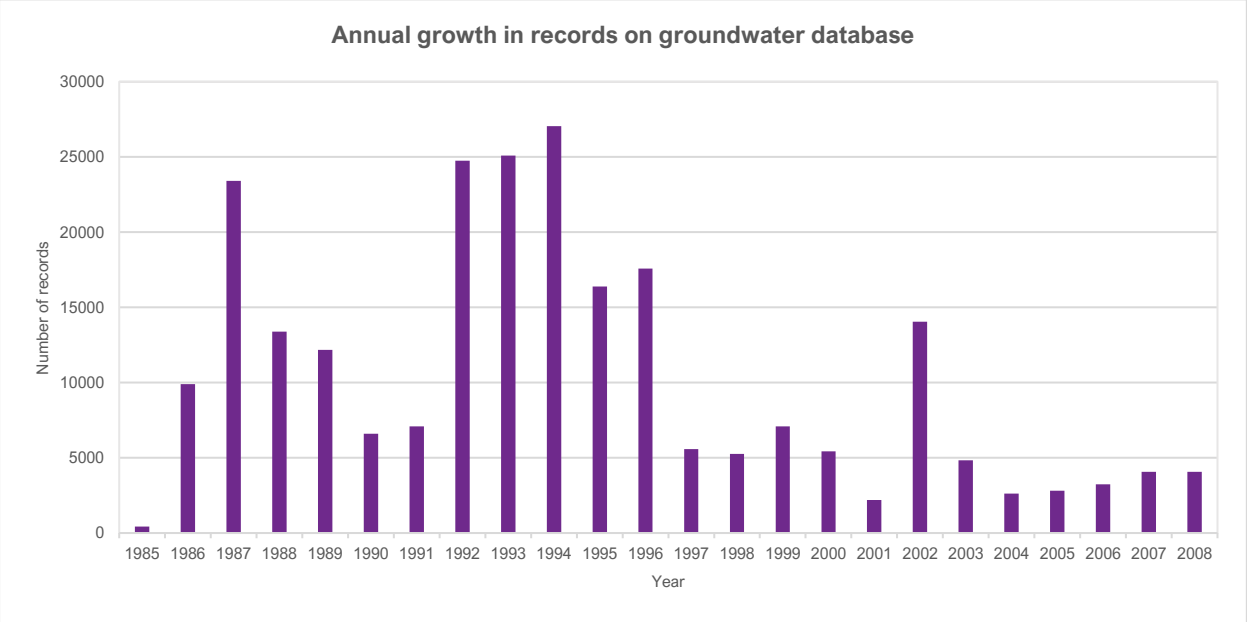


Figure 3-1: Annual growth in NGDB and NGA records from 1985 to 2008 as adapted from DWA (2009)

The distribution of the geosites contained within the NGA is depicted in Figure 3-2. It is evident that a majority of the higher density distribution areas occur in the Limpopo Province and neighbouring provinces, which is also the area covered by the GRIP database. The latter will be discussed in the subsequent section.

The NGA has a data disclaimer stating that the use of data is limited to academic, research, and personal purposes only. Permission should be requested from the Directorate: Surface and Groundwater Information if data are to be used for commercial purposes. The NGA also discloses that the data supplied has no implied warranty as to the suitability for purpose, accuracy, or completeness. Errors may be reported for corrections or enhancements to the Department of Water and Sanitation (DWS) (NGA, s.a.(a)). The NGA also provides a glossary that acts as metadata of a kind describing and clarifying the different attributes found in the database (NGA, s.a.(b)).

The export options are quite extensive and geosites can be selected based on many different criteria including, but not limited to, drainage region and farm name or map number. Drainage region filters in terms of geosites based on quaternary catchment names, but several other criteria are available to filter by. After selecting the desired geosites, various attributes can be exported, including geosite information, water levels, abstractions, lithology, and so on. The site map (NGA, s.a.(c)) reflects the attributes available for export, as depicted in Figure 3-3. A comma-separated-value file (CSV) will be emailed to the user upon processing. Data can also be requested by means of email by completing a data request form. These include groundwater data and groundwater-chemistry data.

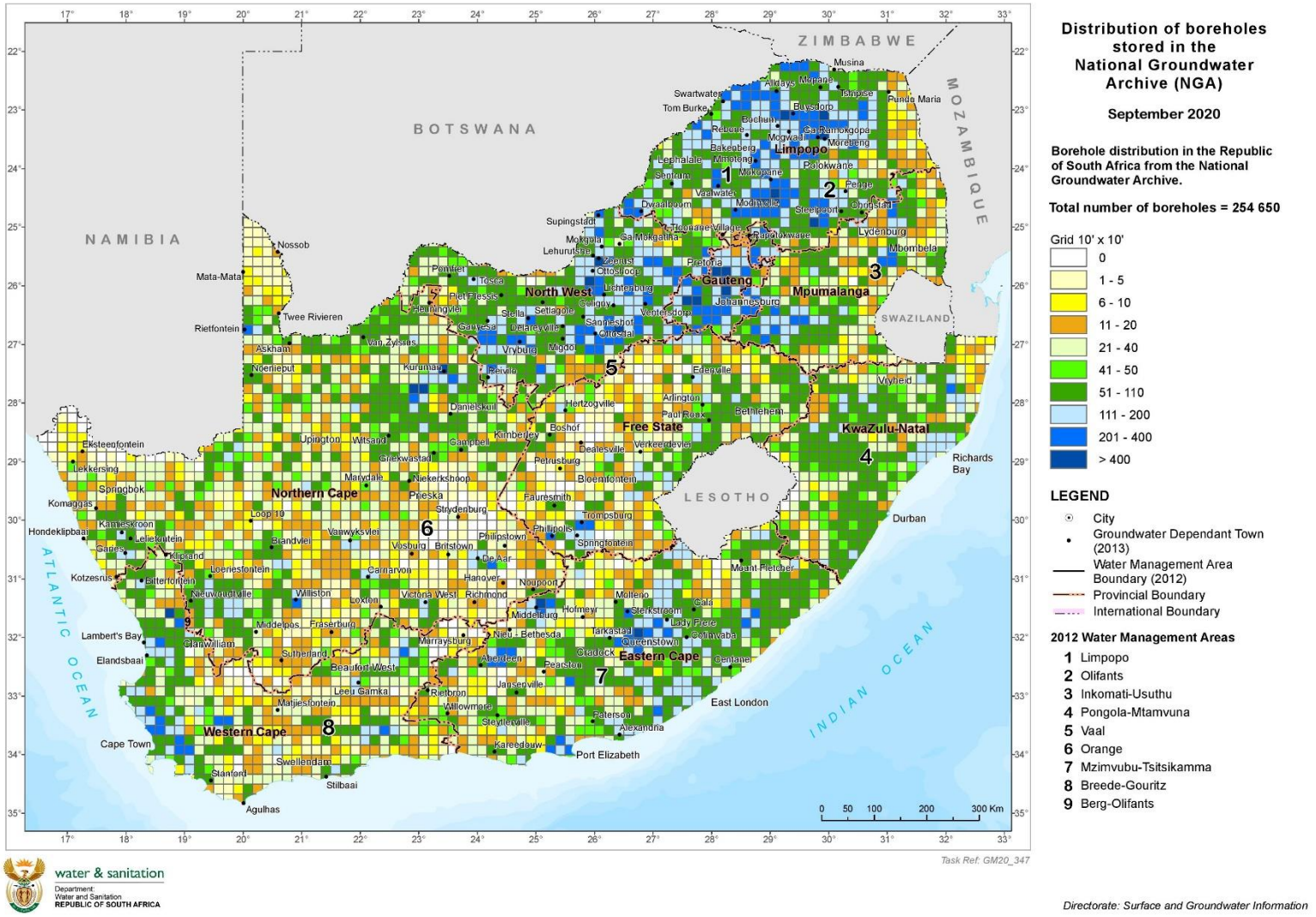


Figure 3-2: NGA borehole distribution and density per 10' x 10' grid (DWS, 2020).

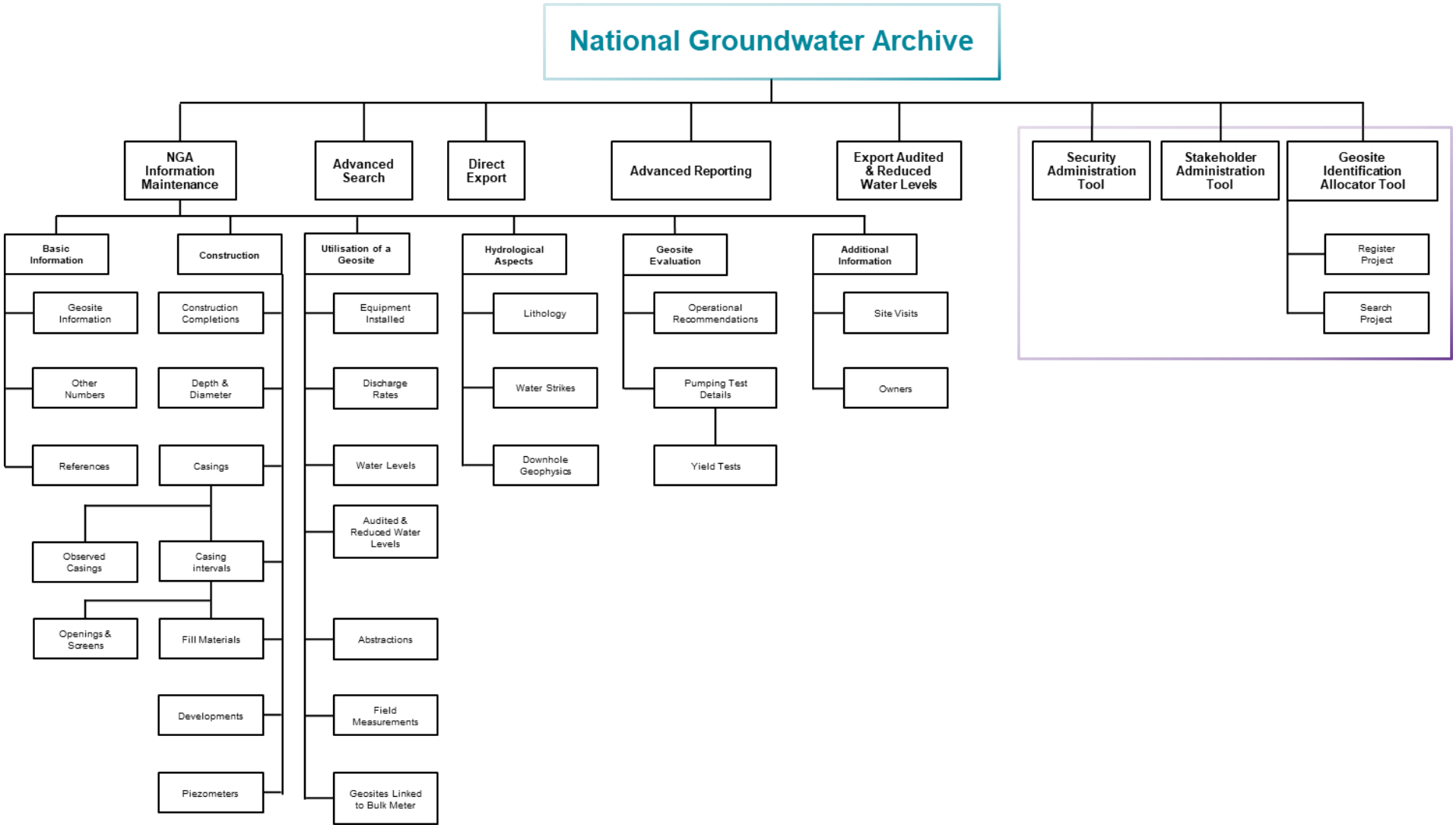


Figure 3-3: NGA Site Map (NGA, s.a.(c)).

3.2.2 Groundwater Resources Information Project

The Groundwater Resources Information Project (GRIP) was initially introduced in the Limpopo Province to collect and verify groundwater-related data with the goal of presenting these to engineers and planners for incorporation into studies (DWA, 2009). It aims to maintain a broad-gauged groundwater dataset consisting of verified data with the goal of making these freely available to its users (GRIP, s.a.) and ensuring the assimilation of groundwater data into the management of water resources.

The DWA report (DWA, 2009) suggests that a GRIP for every province is desirable, and it has been introduced in Kwazulu-Natal and the Eastern Cape. However, implementation is hampered due to a lack of resources. The report goes on to mention that the GRIP in the Eastern Cape Province has been severely hindered as a result of the lack of resources such as funding and human resources (DWA, 2009).

The GRIP database is divided into online data that are readily available for export and a request for further data. The online data acquisition process includes selecting an area such as a district municipality, H-area, local municipality or settlement, and exporting an Excel spreadsheet containing the desired information. The online data includes borehole name, alternative borehole names, coordinates, depth, latest water levels, yield, duty cycle, equipment, and water class. Data that are only available upon request include geology, borehole test data, chemical analysis, borehole construction logs, equipment, time series water levels, and photos.

The distribution of boreholes within the GRIP database is illustrated in Figure 3-4.

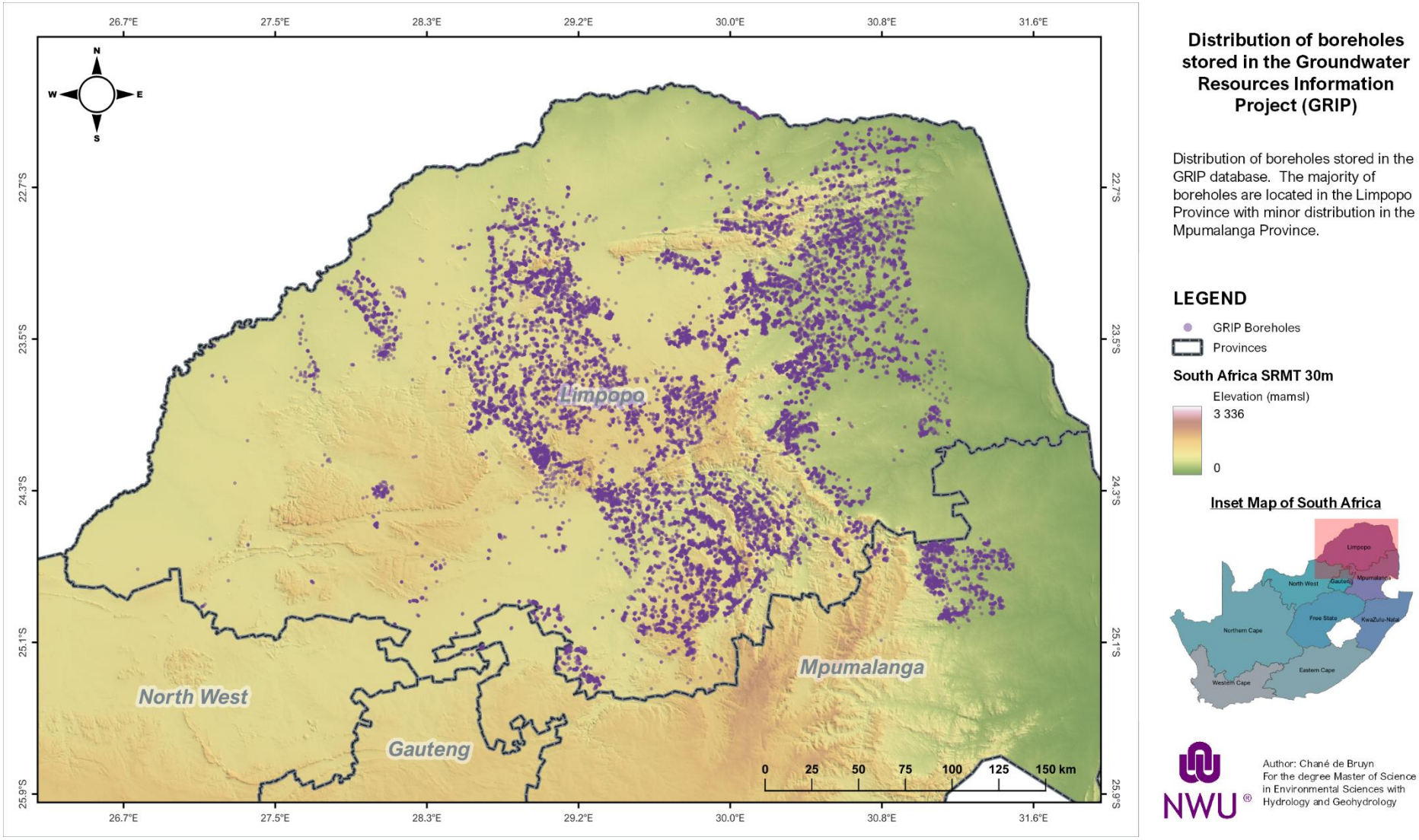


Figure 3-4: GRIP borehole distribution

3.3 Available Data Discussion

Data available from each database will be discussed in the context with the data quality measurements discussed in Section 0 - Rosli et al. (2016) further explain the need for additional information known as metadata. These the purpose, meaning, and context of data, therefore facilitating a better understanding of the data in question. Metadata also aim to avoid any misinterpretation that could arise from data (Rosli *et al*., 2016).

Measuring Data Quality, to gain insights into the quality of each database and how this might affect any analysis performed by using data-mining.

3.3.1 National Groundwater Archive

The NGA gives a data disclaimer (NGA, s.a.(a)) regarding the quality of the data contained in the database by stating that 'all data is supplied with no expressed or implied warranty as to its suitability for purpose, geometric accuracy or completeness'. This database is extremely large and contains numerous attributes that may be exported. Performing the data quality measures set out in Section 0 on the entire database will be laborious. Therefore, the data quality measures will only be applied to a selected number of attributes deemed to have a significant influence on the geohydrological setting. For the purpose of comparison to the GRIP, data from the primary catchment areas A, B, and X will be consolidated into a single dataset on which the data quality measures will be performed.

3.3.1.1 Completeness

3.3.1.1.1 Schema completeness

Within the confines of the Limpopo Province border, the total number of unique boreholes logged in the NGA database at the time of this study was 65 530. A total of 373 possible attributes exist for each borehole identifier in the database. Therefore, as mentioned, only a few critical attributes will be discussed. These attributes include lithology, water strike, chemistry, abstraction, discharge, depth, and water levels. For the selected attributes, a maximum of 1 507 190 data entries is possible. The total number of populated cells is 555 009. Therefore, the schema completeness factor equates to 0.37 for the selection of attributes when Equation (3-1) is applied, or 37%. The schema completeness statistics are summarised in Table 3-2.

Table 3-2: Schema completeness results for a selection of the NGA located in the Limpopo Province

Attribute	Total
Number of unique entries	65 530
Number of attributes per entry	22
Maximum number of possible data points	1 507 190
Number of functional data points	555 009
Schema Completeness Factor	0.37

3.3.1.1.2 Column completeness

Equation (3-1) was used to calculate column completeness of the selected attributes for each of the columns. The results are tabulated in Annexure A, Table 8-2, and are sorted from the most complete to the least complete attributes. In summary, the column completeness is very poor with regard to the chosen attributes. However, it should be taken into account that the completeness factor is determined by the number of boreholes and that an incredible number of boreholes are present. Therefore, it may appear that an attribute such as, say, lithology is incomplete compared to the 65 530 boreholes, but there are 25 154 boreholes that enjoy lithology information. Although a column completeness factor of 1 is desirable, one may be able to conduct a reasonably good analysis of the lithology if so desired. A total of 21.74% of the columns are complete at a factor of 1 and this is attributed to the most basic information regarding the borehole such as the name and location. Figure 3-5 below depicts and compares column completeness.

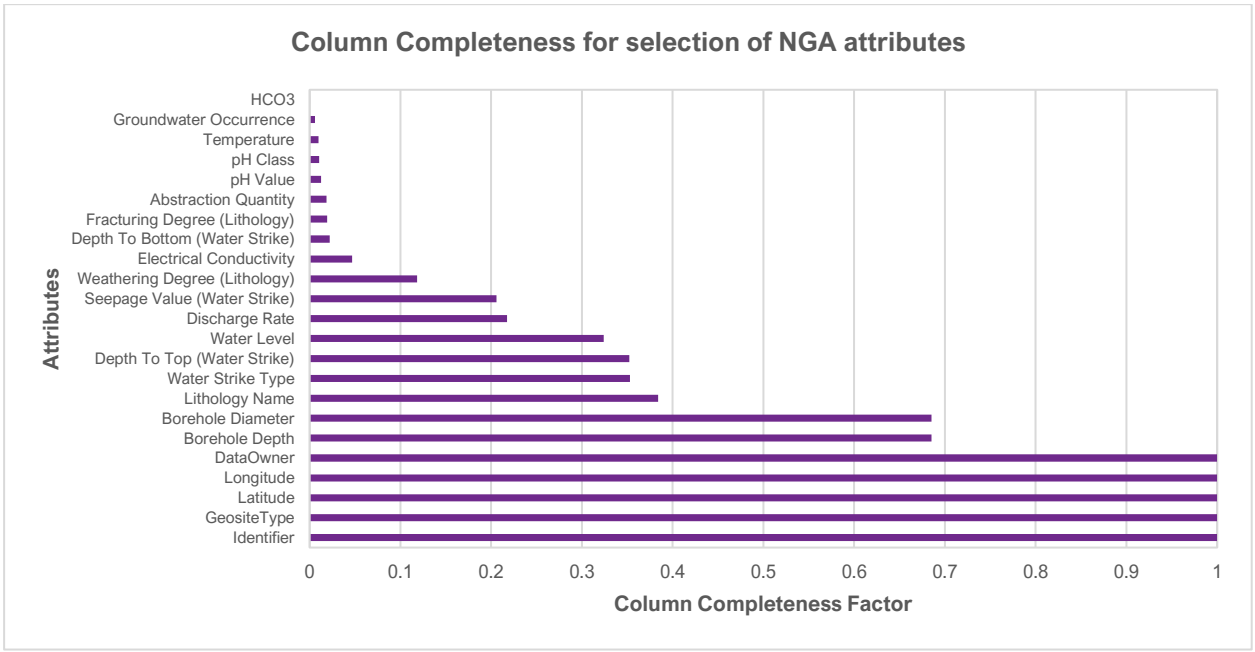


Figure 3-5: Column completeness overview for a selection of the NGA

3.3.1.2 Consistency

No obvious inconsistencies were noted in the selection. No typographical errors were detected in the borehole names, which is important for the sake of ascertaining that all the information can be retrieved for the relevant borehole. Suspect data occurred in the *depth to top* of water strikes, since nine had a reading of 999.99 and forty-four 9999.99. These values were used to flag the fact that no data were available and only constituted 0.19% of the total, while it would in all likelihood not have had a profound impact if removed from data mining.

Typographical errors play a major part in data inconsistency, and care should be taken to ensure their minimisation. A key component when it comes to avoiding errors is to confirm the language settings of the device being used, as this can cause unintended typographical errors. This may be the cause of faults around some entries in the diameter data, where entries such as 216 mm (8.5**ö**) are present and should rather read 216 mm (8.5**0**)".

3.3.1.3 Free-of-error

Section 0 notes that a specific criterion is required in order to classify a data point as erroneous or not. Currently, no such a criterion exists for the NGA. As noted in section 3.2.1, the NGA specifies that erroneous data may occur. It was impossible at the time of conducting this study to determine this parameter, albeit a very important one.

3.3.2 Groundwater Resources Information Project

Upon exporting the borehole data directly from the GRIP website, the Excel spreadsheets contain the following columns; GRIP site ID number, GRIP borehole number, H-area, quaternary catchment area, regional borehole number, alternative borehole number 1 & 2, farm name, farm number, province, district municipality, local municipality, settlement name, settlement ID, alternative settlement name, longitude, latitude, borehole depth, water level, water level date taken, depth to pump intake, discharge rate, duty cycle, daily abstraction, equipment, power, quality, and comments. An example of the spreadsheet can be found in Annexure B – GRIP database example.

The time frame within which water level measurement entries fall, spans from 1900-01-01 to 2015-05-15. Thus, GRIP covers a reasonable period of historic data, although up-to-date data will always be desirable. The date 1900-01-01 is used as a flag value to indicate that the date is unknown. Three other possible erroneous dates include 2088-08-13, 3005-03-09 and 1004-12-

09. These are most likely typographical errors, which are important causes for concern (Gardner, 1992). Lastly, a date occurs as 07/09/2009, which is not in accordance with the date style of the database. The majority of dates are in the format YYYY-MM-DD. This ties into the data quality parameter of consistent representation. There is also a plethora of zeros ('0') in the date column. No metadata are readily available to indicate the exact meaning of the zero, but it could be assumed that it is representative of a missing date for this column.

Columns with the same zero value abound. The assumption is made that this is representative of missing values or measurements. This is based upon the fact that, for attributes with numerical data types such as *borehole depth*, *water level*, *discharge rate*, *depth to pump intake*, *duty cycle* and *daily abstraction*, no unpopulated cells occur within the Excel spreadsheet, while this could also be indicative of a missing value. Also, in multiple instances, the attributes *depth to pump intake*, *discharge rate*, *duty cycle*, and *daily abstraction* are populated by a zero value, therefore indicating missing measurements.

3.3.2.1 Completeness

3.3.2.1.1 Schema completeness

The borehole data from both Limpopo and Mpumalanga yield a total borehole entry count of 25 431 unique boreholes. Each has 27 attributes, amounting to a maximum of 686 637 possible data points in the entire dataset: this number excludes the first column, as the latter contains only the number of boreholes. Of this, 573 898 cells are populated, which includes the zero values discussed in Section 0. There are 110 913 zero value data points. Hence, there are in actuality 462 985 data points and this equates to a dataset that is 67% complete in its unchanged original state. The GRIP dataset therefore has a completeness factor of 0.67 when calculated in terms of Equation (3-1). It should be noted that this completeness factor was calculated for all available data. The completeness factor will change in the event of removal of attributes that are deemed unimportant or redundant. The schema completeness statistics are summarised in Table 3-3.

Table 3-3: Schema completeness results for the GRIP

Attribute	Total
Number of unique entries	25 431
Number of attributes per entry	27
Maximum number of possible data points	686 637
Number of functional data points	462 985
Schema Completeness Factor	0.67

3.3.2.1.2 Column completeness

Equation (3-1) was also used for each of the columns in an unaltered GRIP dataset. The results are tabulated in Table 8-3 in Annexure A and are sorted from the most complete to the least complete attributes. A total of 46.43% of the columns are complete at a factor of 1. To visualise and compare column completeness, please see Figure 3-6.

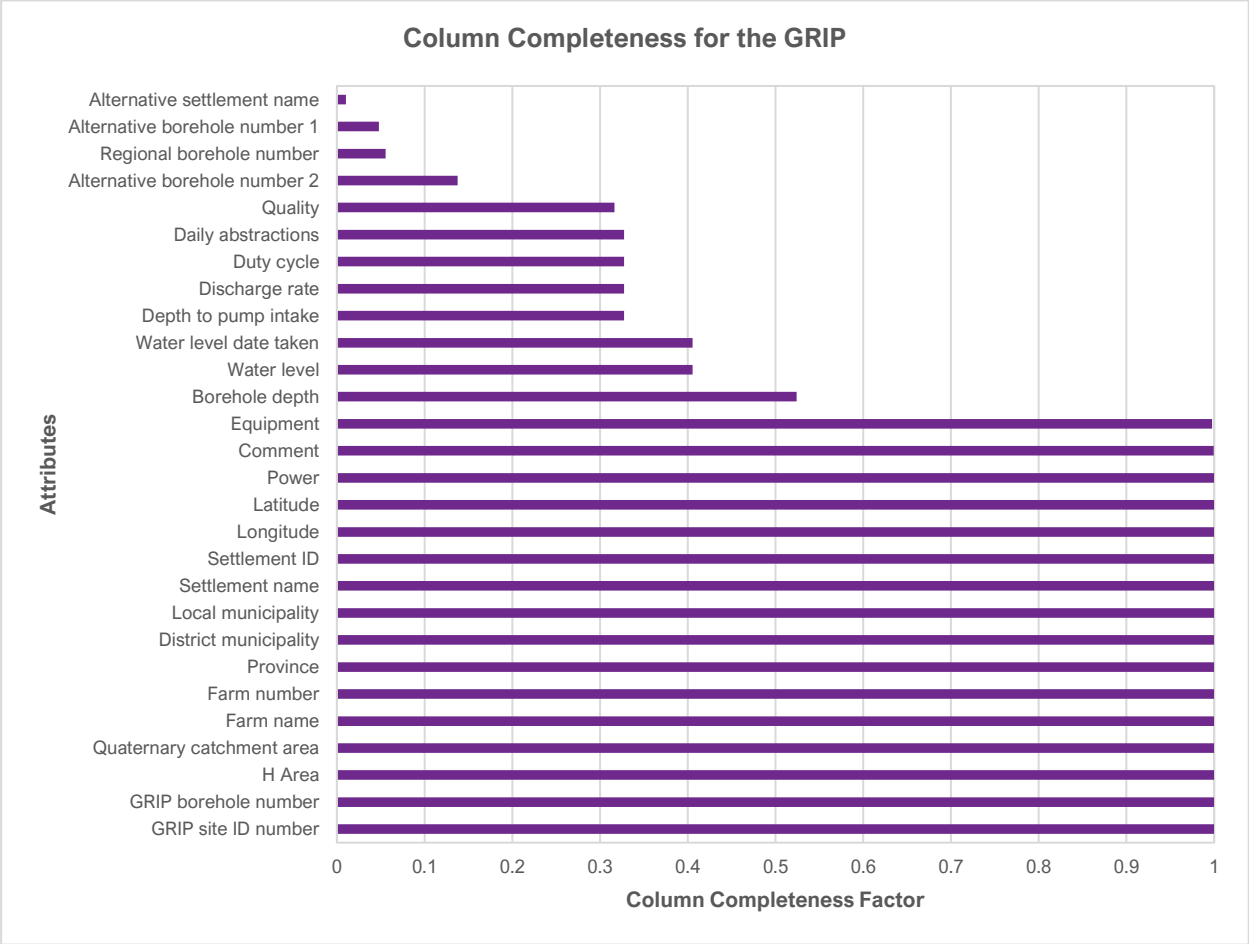


Figure 3-6: Column completeness overview for the GRIP

3.3.2.2 Consistency

GRIP contains several inconsistencies. Various columns contain text data where some values are placed between two quotation marks. For example, the majority of the *farm name* column entries do not contain quotation marks, but multiple data points do contain them. These values do not cause duplicates upon removing the quotation marks. The export engine adds the quotation marks around text strings if spaces are present. Other entries could be considered to be duplicates as a result of typographical errors such as spelling mistakes or different spelling of the same name. This is only the case where farm names are considered, though, and will likely not

affect the outcome of data mining in the present project. One example of this can be found in the *farm name* column where "ZEBEDIELA ESTATES" and "ZEBED**E**IELA ESTATES" are contained within the same column. Where alternative borehole names are concerned, inconsistencies occur regarding spacing within the names. On average, and in most cases, the alternative names consist of a combination of characters and letters with spaces absent. Values that once again contained quotation marks, also presented with spaces in the name. This did not cause duplicates but, as mentioned, it is worth noting this inconsistency. Spaces are considered to be characters within an entry within a cell in Excel, and this could influence formulas and analyses performed on a particular set of cells.

The problem of analysing consistency within datasets centres on not knowing whether an entry can be considered to be an inconsistency or not. If a space is present somewhere in the borehole ID name, the assumption could be made that, in all likelihood, it amounts to a typographical error and can be easily fixed. Where names are concerned, such as farm names, it can be assumed that, in all likelihood, two or more farms can exist with similar names spelled differently. Therefore, care needs to be exercised when analysing inconsistencies and attempting to fix them.

3.3.2.3 Free-of-error

Accuracy is difficult to examine on such a large scale since it is hard to be knowledgeable about the process of measurements that had been followed.

3.4 Spatial datasets

Spatially distributed data are key components in many research fields. Visualisation of data on a geospatial level may offer crucial insights that might not have been immediately apparent in terms of tabled or a two-dimensional structure. Spatial datasets usually cover a single area of interest such as evaporation. Consequently, data must be congregated from a variety of different sources, as no single dataset will contain each desired attribute. Many factors that could potentially influence the groundwater regime occur on surface level, such as evaporation, rainfall, and runoff, to mention a few. This is confirmed by Lerner and Harris (2009), who connect groundwater and the landscape, noting that anthropogenic activities such as urbanisation tampers with recharge. Therefore, it is essential to incorporate data from spatial datasets into that of the NGA and GRIP databases so as to gain a comprehensive dataset that would expectantly aid the prediction of groundwater levels and the establishment of relationships within the complex system.

Various spatial datasets are publicly available from government entities such as the Department of Forestry, Fisheries and the Environment, which established an E-GIS website containing multiple environmental geospatial datasets (DFFE, 2022). These datasets include land cover ranging from 1990 to 2020, the change in land cover, and other data such as protected areas. Land cover, whether natural or manufactured, influences recharge value (Lerner & Harris, 2009). Lerner and Harris (2009) note that urban surfaces are by and large impermeable, causing less rainfall to penetrate the earth beneath and lowering recharge values. However, it should be noted that the construction and maintenance of urban areas differ widely, also influencing recharge. Lerner and Harris (2009) further explain that vegetation plays a critical role in recharge, and so does agriculture. It has been surmised that croplands have higher recharge rates than those of native vegetation (Lerner & Harris, 2009; Kim & Jackson, 2012). Therefore, land cover is critical to incorporate into the final set used for data mining. The land cover data obtained from DFFE (2022) have been generated by using automated mapping models and Sentinel 2 satellite imagery (DFFE, 2021). The resolution is 20 m and has a calculated accuracy of 85.47% (DFFE, 2021).

The Water Research Commission authorised a study regarding the water resources of South Africa, Lesotho, and Swaziland, colloquially known as WR2012. A web-enabled system has been established to provide water resource specialists with all the data, including maps, water resource models, and other tools that resulted from the study (WRC, 2012). The GIS maps available include many geohydrological data such as transmissivity, recharge depth grids, groundwater volume of aquifers, and so on. Other important datasets include rainfall, evaporation, runoff, vegetation, and geology. Consider that many of the groundwater parameters made available by the WR2012 originate from the Groundwater Resources Assessment Phase II, also known as GRAII.

Finally, it has to be considered that not all GIS datasets necessarily cover the entirety of South Africa.

CHAPTER 4: METHODOLOGY

The purpose of the present study is to research the validity of data-based approaches, as discussed in the literature review, to substantiate a cause-and-effect relationship between borehole parameters and the geological and geographical settings. This will be achieved by means of a desktop study and investigating data mining and machine-learning techniques. As the name implies, data mining requires extensive amounts of quantitative and qualitative data.

For this study, data will be gathered from three different sources, namely the NGA, GRIP and spatial GIS data available for South Africa. Subsequently the data will be processed on an individual basis first to ensure that there are no duplicate entries and to clean up erroneous data. Entries with insufficient data will also be removed during this phase. These datasets will then be into a single dataset in a format that will be usable for the next phase.

The next phase is model building. The statistical language R will be used in and integrated development environment like RStudio to compile scripts, where the algorithms discussed in the literature review are implemented by using the data from the previous phase. The algorithms discussed can be classified as either regression or classification algorithms and, in some instances, they can be used for both.

Classification and regression model analyses and evaluation will be done. The results will be compared for each instance in order to establish the algorithm that is best suited for the purpose of predicting a continuous numerical water level or yield as well as the best-suited algorithm for predicting a class of water level or yield.

Once regression and classification algorithms have been established, the methodology will have been concluded, and implementation will be conducted on three case study areas in South Africa, as found in the next chapter.

4.1 Data acquisition

Publicly accessible data will be utilised in this study to ensure a degree of replicability. Data such as this are found in two major databases, namely the NGA which is developed and maintained by the Department of Water and Sanitation, and the GRIP database which is centred around borehole data specific to the Limpopo Province. Another data source used for this study is spatially distributed data. There are numerous repositories for Geographic Information System (GIS) data found online and were discussed in the previous section. The WR2012 was the predominantly

used source as this database contains numerous valuable geospatial data with a focus on geohydrological data.

The acquisition process of data from the various sources will be discussed in the sections below. Section 4.2 will focus on their consolidation into a single set.

4.1.1 NGA data acquisition process

For the purpose of consolidating data from various databases into one single dataset, only regions that would correspond geographically with the GRIP database were chosen for export. The GRIP database covers the Limpopo Province and a small area of the Mpumalanga Province. Therefore, primary catchment areas A, B, and X were chosen as geographic areas whereby borehole information was selected and exported. For the selection from the NGA, all quaternary catchments within the primary catchments A, B, and X were used. The exported boreholes and their distribution within the Limpopo Province are indicated in Figure 4-1. Figure 4-1 also depicts the overlap between the NGA boreholes and those of the GRIP database.

After selecting the geographic area from which boreholes were to be exported, attributes were selected based upon their relevance to the geohydrological setting. Thus, not every available feature was selected for export. It should be taken into account when exporting data from the NGA that it should be done conservatively with regard to the number of attributes chosen in comparison to the number of boreholes. That is, if the number of selected boreholes is large and, if all the desired attributes were to be selected within one export, it would not execute. It is assumed that the server could not send a file of compiled data that were too large. When this was the case, multiple exports were conducted. After retrieving all the desired data in CSV format, the files were imported into Microsoft Access on the basis of the sections in terms of which they were exported. The data processing phase will be discussed further in Section 4.2.

4.1.2 GRIP data acquisition process

The search function on the GRIP database webpage allows the user to export borehole data according to district municipality, H-area, and quaternary catchments. For the sake of thoroughness, all three categories were used. The data stretched across the Limpopo Province and parts of the Mpumalanga Province, as illustrated in Figure 4-1. The files were easily exported to a CSV file for further processing.

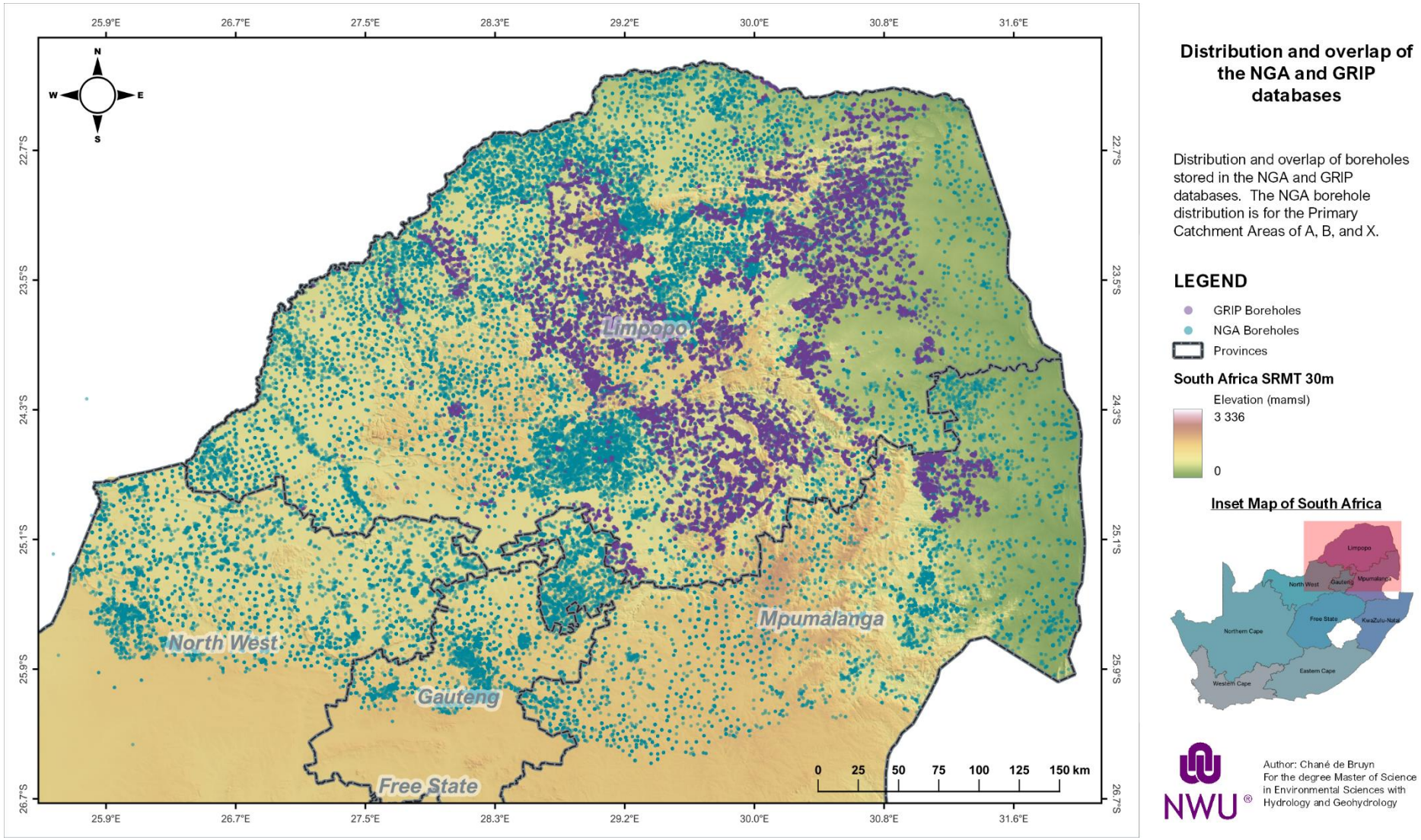


Figure 4-1: Distribution and overlap of boreholes from both the NGA and the GRIP databases

4.1.3 GIS data acquisition process

Additional data were collected from spatial datasets, as it was assumed that surface features such as rainfall and recharge, vegetation, and land use might have an impact on the borehole parameters. Various geological and geohydrological data were also captured within spatial datasets, aiming to capture subsurface attributes within a vector or raster file. Numerous geohydrological related maps were downloaded from the WR2012 website (WRC, 2012) and imported into GIS software such as QGIS (QGIS Development Team, 2021). The aim was to create an additional dataset with the combined boreholes of the NGA and GRIP which would be used to generate a GIS feature database.

All the relevant GIS data were loaded into QGIS along with the borehole locations. QGIS has a *join attributes by location* function in its *data management tools* menu that joins a vector layer to a base layer, that is, it joins the attributes of a desired layer to the attributes table of the borehole layer. The resulting output layer was then used as the base layer for appending subsequent attributes. This process was repeated until all the desired attributes had been appended. For raster layers, the *sample raster values* function was used where the same concept applies as that of the vector layers. A single layer would be the end result, which was exported as a CSV file. Figure 4-2 reflects the process of assigning raster and vector data from various layers to a specific borehole.

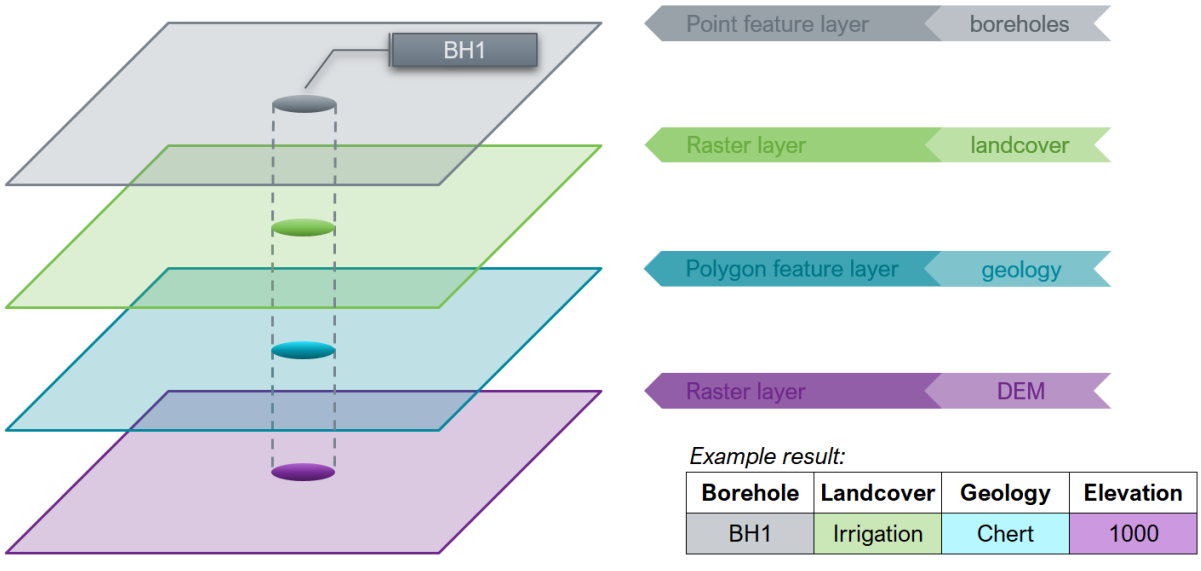


Figure 4-2: Assignment process of GIS data to a single borehole

4.2 Data processing

Data processing of the acquired data underwent several phases. Firstly, data of each base were processed separately to ensure that the final dataset from a specific source contained no duplicate entries. Subsequently, all three sources were assembled into one single dataset. Further processing took place to ensure that duplicate or similar attributes were discarded.

4.2.1 Data processing - Phase 1

Phase 1 involved an assemblage of data gathered from a source into a single representative dataset. This had to be done for each of the three sources. In the case of the NGA, 11 different datasets needed to be processed for multiple values per entry and assembled into one dataset. The GRIP and GIS files needed to be merged into one dataset as well.

4.2.1.1 NGA

Microsoft Access was used to create separate datasets for each of the features within the NGA. Microsoft Excel also has this capability, but the reason for using a database management system such as Microsoft Access was that it supports relational databases. Microsoft Access allows the user to easily run queries on a dataset to calculate the average of the values for a unique entry or other functions such as minimum or maximum values, first or last values, and so on. This proved useful for exported files like those of water levels and water strikes which had multiple entries for the same borehole. The following datasets were identified as containing multiple values; *abstractions*, *depth and diameter*, *discharge rate*, *field measurements*, *lithology*, *pumping test details*, *water strike*, and *water levels*. In each instance, the average was calculated to gain a single representative value for each borehole.

This however proved difficult for data such as lithology. In the case of static water level data, the data type is numeric. The number of water levels observed for a borehole is inconsequential, because any number of observed water levels can be averaged to gain a single representative value of the static water level. The case for the lithology and log data were not found to be a matter of numerical data which can be averaged, since it consists of character data or text strings. Furthermore, each borehole had a differing length of log entries, where one borehole could for example have five lithology bands noted and another twenty or more, whereas each lithology occurred at a different depth. This proved difficult to convert to a logical value for an algorithm to process and was therefore omitted from this study.

Although the other datasets did not need to have any queries run on them, they were also added to Microsoft Access for the ease of exporting one single dataset. After averaging all the necessary attributes, the *geosite* was used as the primary key whereby all the other attributes would be returned and exported into a single set that would be representative of the NGA set.

4.2.1.2 GRIP

The GRIP data acquisition generated separate datasets based upon district municipality, H-area, and quaternary catchments. By using Microsoft Excel, each of the files were imported into the same spreadsheet. The *remove duplicates* function was performed on all the imported data to ensure that no duplicate entries were present. Upon verifying this, the GRIP dataset consisted of 25 431 unique entries.

4.2.1.3 GIS

No processing was needed for the GIS data during this phase due to the fact that all duplicate boreholes were removed before importing into GIS.

4.2.2 Data processing - Phase 2

Phase 2 involved an assemblage of the three datasets from Phase 1 into one single dataset. Data from all three sources were consolidated into a spreadsheet upon which further processing was done. Elimination of similar features that occurred between the different datasets was necessary. This process was repeated for all instances of similar features. Furthermore, columns were examined for missing values, and those that indicated a poor completeness factor were removed, as there were too little data to use meaningfully in the data-mining process.

Some variables were found to be numerical in nature, such as elevation, mean annual precipitation (MAP), depth, and so on. Other variables are categorical or have a text value which must be assigned a number. However, these categories or assigned indexes must ideally not be interpreted as a numerical value but, rather, as a factor value. Care had to be taken during the assembling of scripts to ensure these values were assigned as factor data types.

4.2.3 Data processing - Phase 3

The aim of this study, as indicated, is to classify relationships between the relevant geological setting and borehole parameters such as water level and yield. Therefore, Phase 3 included the creation of a dataset for each dependent variable, such as water level and average yield of water

strike. For each dataset, the independent variable was placed in the last column for ease of use in the model scripts. A wide variety of variables were available to choose from for the final dataset of each independent variable that was predicted, but keeping the aim of the study in mind, only parameters relevant to the geological setting were considered.

Both regression and classification analysis were conducted for comparative reasons. Regression uses continuous data and classification requires the variable to be categorised. Therefore, the dependent variables were placed in appropriate classes.

For classification, the dependant variable had to be a class. Therefore, categorising the variable into distinct classes was required. This was done for water levels and for water strike yields. Water levels were tested in terms of unit metres above mean sea level (mamsl), which ranged from approximately 50 mamsl to 2600 mamsl. This is a very broad range, and the classes had to be of a lower resolution, such as 50 m and 100 m. Yield was classed by using the same classes as those used in the hydrogeological map series, that is, five classes that were divided as reflected in Table 4-1 below.

Table 4-1: Assigned yield classes

Yield range	Class
0.0 – 0.1 l/s	A
0.1 – 0.5 l/s	B
0.5 – 2.0 l/s	C
2.0 – 5.0 l/s	D
> 5.0 l/s	E

The yield classes were chosen in this manner with a view to the distribution in each class. If equal interval classes had been chosen, most of the yield value would have fallen into one class, as depicted in Figure 4-3. The maximum yield class that was considered was > 5 l/s, since this was the maximum yield class expressed in the Geohydrological Map Series for groundwater occurrence.

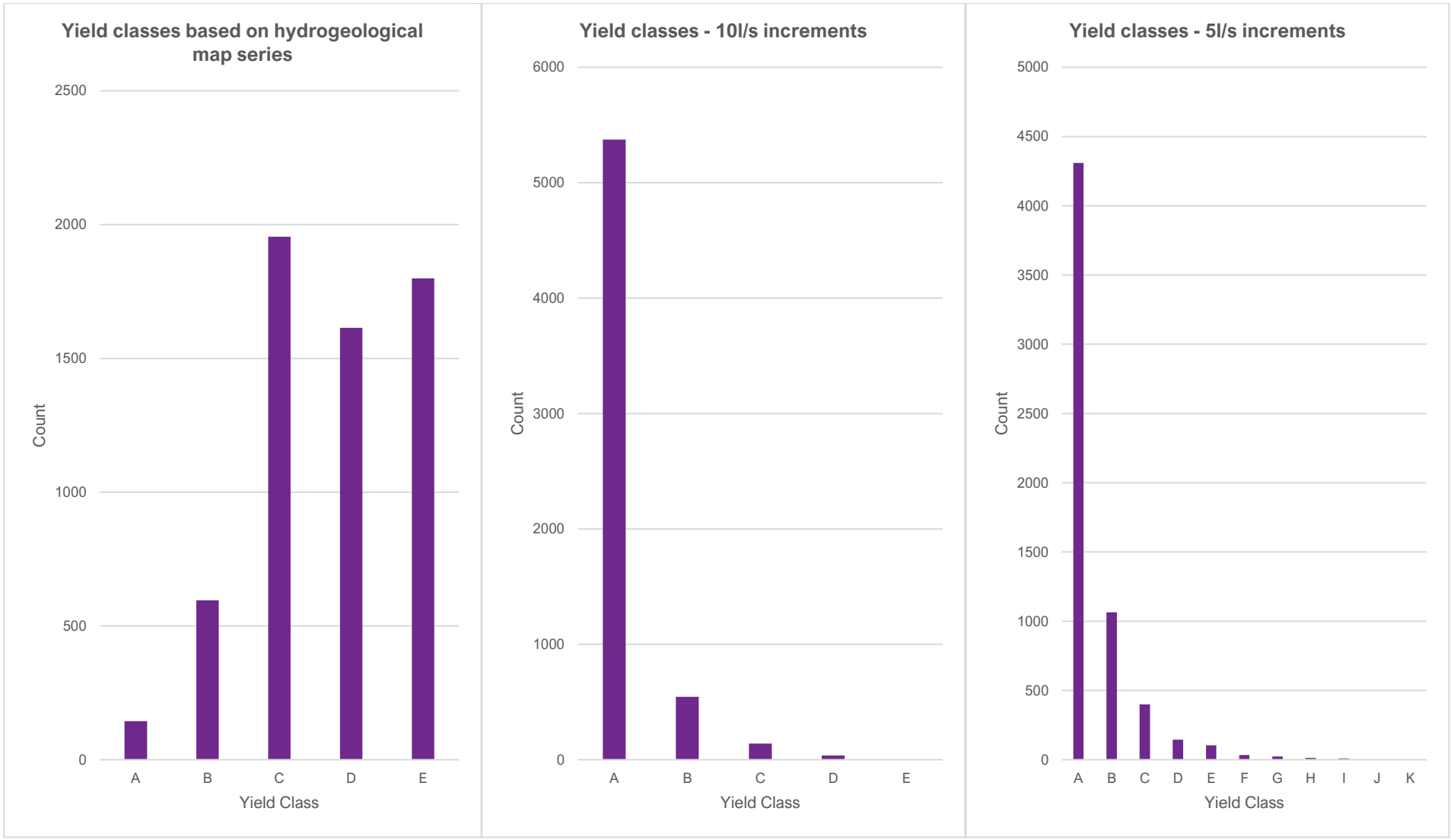


Figure 4-3: Distribution of yield values in different size classes

4.3 Computer methods

The data processing phase resulted in a final dataset – with a few variations for the classification algorithms - which would be used in all models. Five different algorithms were used to analyse and predict the dependent variables. The algorithms used were classified in terms of regression or classification or both. SVM and DT were found to be both regression and classification algorithms. Figure 4-4 below indicates the algorithms and their categorisation along with the R library used in the scripts.

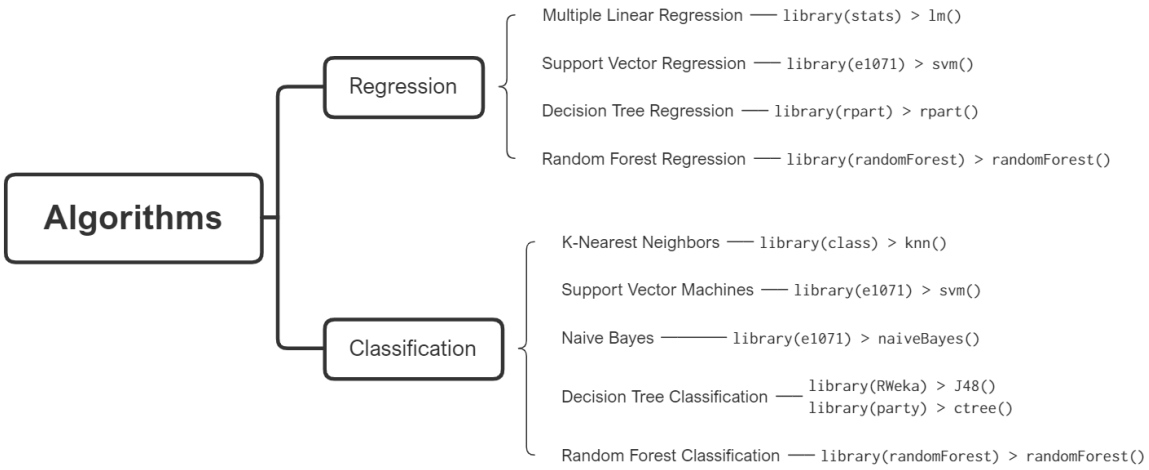


Figure 4-4: Types of machine-learning algorithms and the R libraries used in each

R is a statistical language, and it was used to compile scripts and build all the models that are mentioned in Figure 4-4. RStudio was the integrated development environment (IDE) used to develop the models and the scripts were sourced from SuperDataScience (2020). The specific scripts used for each algorithm will be laid out in Annexure C – Model Scripts along with plots or other results that had been generated.

4.4 Algorithms

A set of five algorithms was used to compare the accuracy of each and to determine which algorithm best suites the aim of this study, namely:

- Decision trees
- Baysian classifiers
- K-Nearest neighbour
- Support vector machines
- Linear regression.

Fuzzy logic and ANN were omitted from the methodology as elucidated by the literature review. According to extant literature, fuzzy logic uses immense amounts of memory to generate a rule-base, which increases exponentially for each independent variable in the database. Due to the size of the initial dataset to be used, this method was not suitable in this study. The same issue cropped up in the case of ANNs, while they also produce limited (if any) insight into the system studied.

Algorithms were run with the datasets split to an 8:2 ratio to ensure an adequate number of data observations to train the models on. Therefore, 80% of the dataset was chosen at random for the training set, and the remaining 20% used to test or validate the model.

Measurements of accuracy or error used here, as discussed in terms of the literature review (Section 0, included RMSE, MAE, MAPE, and Pearson correlation for regression models. For classification models, confusion matrices were used to illustrate the results of the model. The Kappa value and other metrics could be calculated by using the resultant matrices. The testing focused on static water levels and yield. The following generated sections indicated the model performance for each algorithm (for both static water level and yield), and briefly discusses how each model were set up within the parameters of the algorithms. Regression and classification models with the best performance were applied to three case studies.

The procedure for compiling the scripts started with the use of the entire dataset and the elimination of parameters that seemed to have a neutral to negative effect on accuracy. It was expected that this would result in parameters that were the drivers behind the independent variable that was being investigated.

4.4.1 Static Water Level

4.4.1.1 Regression

Starting off with the dataset containing all the available parameters, it was immediately apparent that the regression models predicted with high rates of accuracy. This was to be expected, since groundwater levels follow the natural topography and, therefore, elevation dominates the predictions. Omitting the elevation resulted in a lower Pearson correlation, validating the preceding statement. Nonetheless, the algorithms performed well without the elevation data.

4.4.1.1.1 Multiple Linear Regression

The multiple linear regression (MLR) model is very simple to create without having to tweak additional arguments. The dataset was used and the dependant variable was set although, if refinement was desired, arguments such as weights, contrast, and offset were used as options.

4.4.1.1.2 Support Vector Regression

For support vector regression (SVR) in terms of the `e1072` library, two types of regression could be performed, namely *nu* and *eps*. The *nu* regressions performed marginally better than those of the *eps* regression. Furthermore, the kernel could be specified as linear, polynomial, radial basis, or sigmoid. The linear kernel performed better than the others.

4.4.1.1.3 Decision Tree Regression

The *rpart* model facilitated control of the algorithm details such as *minsplit,* which is the minimum number of observations in a node before a split can be tried. The minimum split for the model was designated to start from 2, but there was no noticeable difference between a small and a large split.

4.4.1.1.4 Random Forest Regression

The algorithm for random forest regression (RFR) has the option for selecting the number of trees. Multiple runs were conducted by using different totals of trees. A selection of 100 trees resulted in the best performing model. Consider that the larger the number of trees, the longer the model runtime is.

4.4.1.1.5 Regression model selection

While initially omitting elevation, other variables were used singularly to establish the way in which the model reacted to variables such as geological parameters, transmissivity and storativity values, recharge, and annual precipitation. This one-on-one approach made it clear which variable had the greatest effect on the prediction. By using the variable with the greatest influence, other variables were added until the best correlation was attained. Variables that led to the best model performance in addition to elevation were mean annual precipitation, storativity grid values, and five sequential lithologies. Only these parameters were used in all four models to ensure that they could be compared on the basis of the same data. The final accuracy metrics are summarised in Table 4-2 reflecting that SVR was the model with the best performance and MLR

the second best. If elevation is omitted, RFR is the best performing model. For comparison purposes, the results were also compared with a Bayesian interpolator (Tripol[1]) and, in this case, the SVR performed better.

Table 4-2: Water level regression model performance metrics

	Elevation used				Elevation omitted			
Regression Algorithm	**Pearson (%)**	**RMSE**	**MAPE (%)**	**MAE**	**Pearson (%)**	**RMSE**	**MAPE**	**MAE**
Multiple Linear	99.90	14.27	1.25	10.39	70.17	229.06	22.62	170.44
Support Vector	99.90	14.41	1.22	10.04	69.97	231.99	22.68	166.32
Decision Tree	97.49	72.23	6.53	50.12	77.42	205.32	20.11	148.04
Random Forest	98.96	48.25	2.72	19.56	90.16	142.59	12.35	89.34
Bayesian Interpolation	99.84	23.24	1.60	15.03	-	-	-	-

4.4.1.1.6 Comparison with established geohydrological software

The Bayesian interpolator in Tripol, which interpolates water levels based on elevation data, was also used to interpolate water levels. This was done as a measure for comparing the results of the machine learning models to those of established methods. The Bayesian estimation was used, where all its parameters were left at default. The Bayesian interpolation only uses coordinates, elevation, and water level elevation. No other parameters can be added. It should be noted that the interpolation results are only valid when high correlations between elevation and water levels exist.

4.4.1.2 Classification

The parameters established during the regression model building phase were used to build the classification models in order to measure their performance in terms of the same set of parameters. The classification accuracy was excellent, albeit that, if a strong parameter such as elevation was omitted, some algorithms managed to predict results with fair accuracy.

[1] Tripol is an interpolation application that performs Inverse Distance, Kriging, and Bayesian interpolations.

4.4.1.2.1 K-Nearest neighbour classification

Use of the K-NN algorithm is fairly straightforward: the number of neighbours was the only significant argument to be tweaked. The number of neighbours (k) was iterated starting from 1 while increasing k with every iteration. It was found that k = 1 performed the best.

4.4.1.2.2 Support vector classification

As discussed in Section 0 above, SVM has different types of classification and kernel types. *C-classification* performed better than *nu-classification* when combined with a linear kernel type. The algorithm has numerous arguments to fine-tune the model, but arguments are sometimes dependent on the type of classification and kernel.

4.4.1.2.3 Naive Bayes classification

The naive Bayes classifier is also straightforward to use. Minimum arguments were to set the dataset as *x* and the class to be predicted as y.

4.4.1.2.4 Decision-tree classification

The C4.5 algorithm was used in terms of the *RWeka* library by using the J48 classification tree learner. J48 has very few arguments to tweak, but it performs better than the *rpart* or *ctree* algorithms that were also tested. Therefore, model creation is simple.

4.4.1.2.5 Random-forest classification

The *randomForest* algorithm has various arguments that can be tweaked to find the best-performing model. The number of trees used to give the best performance were 100 along with a node size of 1, which is recommended for classification.

4.4.1.2.6 Classification model selection

As stated, only the parameters established to be critical were used for building the models. The accuracy rates were nearly perfect, and elevation was the primary driver. For comparison, the elevation was omitted from models to evaluate the extent to which the performance would change. SVM and Naive Bayes struggled in this regard, whereas the other three algorithms handled the omitted elevation relatively well.

The final accuracy metrics of the confusion matrices are summarised in Table 4-3. Random-forest classification (RFC) was found to be the best performing model, with decision-trees ranking at a close second. RFC was expected to outperform the latter, since it is an ensemble method that combines the outcomes of multiple decision trees. With regard to the elevation omitted, K-NN performed the best and decision-trees again ranked at a close second.

Table 4-3: Water level classification model performance metrics

	Elevation used			Elevation omitted		
Regression Algorithm	**Correctly Classified (%)**	**Kappa value**	**Strength of Agreement**	**Correctly Classified (%)**	**Kappa value**	**Strength of Agreement**
K-Nearest Neighbour	89.94	0.89	Almost Perfect	67.21	0.64	Substantial
Support Vector Machines	88.77	0.88	Almost Perfect	36.77	0.30	Fair
Naive Bayes	83.86	0.82	Almost Perfect	31.98	0.26	Fair
Decision Tree	91.50	0.91	Almost Perfect	66.40	0.63	Substantial
Random Forest	91.52	0.91	Almost Perfect	58.75	0.55	Moderate

4.4.2 Average water strike yield

The models created for water level predictions were also used to predict yield. The same approach was taken, where all available parameters were used and tested on a one-on-one basis. The most notable parameters were those obtained during pumping tests, namely transmissivity, storage coefficient, and specific capacity. Other parameters that seemed to influence the prediction were recharge, mean annual precipitation and runoff, baseflow per quaternary, lithology, and the count of water strikes present in the borehole.

Interpolated transmissivity and storativity values were also tested, as opposed to the pumping test parameters, since many boreholes with yield data did not necessarily include these detailed parameters. In the case of these interpolated values, the model performance severely degraded, reinforcing the importance of the parameters that were established during pumping tests and their relation to yield.

4.4.2.1 Regression and model selection

Table 4-4 summarises the accuracy metrics obtained from the regression models. Random forest regression was found to perform best showing a fair Pearson correlation of 63% and the lowest

RMSE and MAE values. The accuracy metrics for the interpolated pumping test values have severely degraded in performance.

Table 4-4: Yield regression model performance metrics

	Pumping test parameters				Interpolated S and T values			
Regression Algorithm	**Pearson (%)**	**RMSE**	**MAPE (%)**	**MAE**	**Pearson (%)**	**RMSE**	**MAPE**	**MAE**
Multiple Linear	32.59	3.93	-	2.85	20.50	4.10	-	3.07
Support Vector	32.88	3.95	-	2.37	17.22	3.82	-	2.60
Decision Tree	57.24	3.51	-	2.19	18.86	4.09	-	3.10
Random Forest	62.91	3.44	-	2.01	22.54	4.69	-	3.23

4.4.2.2 Classification and model selection

Table 4-5 presents the accuracy metrics for the classification models. The random-forest algorithm again showed the best performance for yield classification, with a classification accuracy of 59% and a moderate Kappa value.

Table 4-5: Yield classification model performance metrics

	Pumping test parameters			Interpolated S and T values		
Regression Algorithm	**Correctly Classified (%)**	**Kappa value**	**Strength of Agreement**	**Correctly Classified (%)**	**Kappa value**	**Strength of Agreement**
K-Nearest Neighbour	49.94	0.29	Fair	38.95	0.13	Slight
Support Vector Machines	52.13	0.31	Fair	36.66	0.10	Slight
Naive Bayes	43.78	0.20	Slight	30.44	0.07	Slight
Decision Tree	55.65	0.38	Fair	36.33	0.12	Slight
Random Forest	58.76	0.42	Moderate	41.49	0.18	Slight

4.5 Assumptions and limitations

The following assumptions and limitation are noted for this study:

- The data gathered from the various databases are assumed to be comprehensive enough to achieve the objectives as presented in Chapter 1.
- The study is limited by disregarding temporal aspects of data, so that the study is not seasonally bound. This is due to the fact that limited temporal parameters are available.
- Only the most prevalent algorithms and associated libraries were considered for this study.
- Algorithm arguments were generally used with default values and only adjustments of major arguments were considered.

CHAPTER 5: CASE STUDIES

In order to test the methodology, it was applied to three case studies. It was expected that its validity would be apparent when the result of each case study was compared with the actual observed water levels.

Study areas were chosen according to Vegter groundwater regions. These regions were delineated based on lithostratigraphic units and geological structures (see Nell & van Huyssteen, 2014), given that geology is a critical driver behind the groundwater characteristics found in an area (Dennis & Dennis, 2020). Therefore, each region has similar geohydrological responses, and are ideal to use as the main selection criteria, as they data within each region is assumed to be representative of the drivers within that specific area. A total of 64 groundwater regions occur in South Africa. In order to test the validity of the methodology, two areas with an abundance of data were used as well as an area with sparse data. Table 5-1 indicates the chosen groundwater regions and their respective number of boreholes as present in each groundwater region, along with the total of boreholes with water level data and yield data.

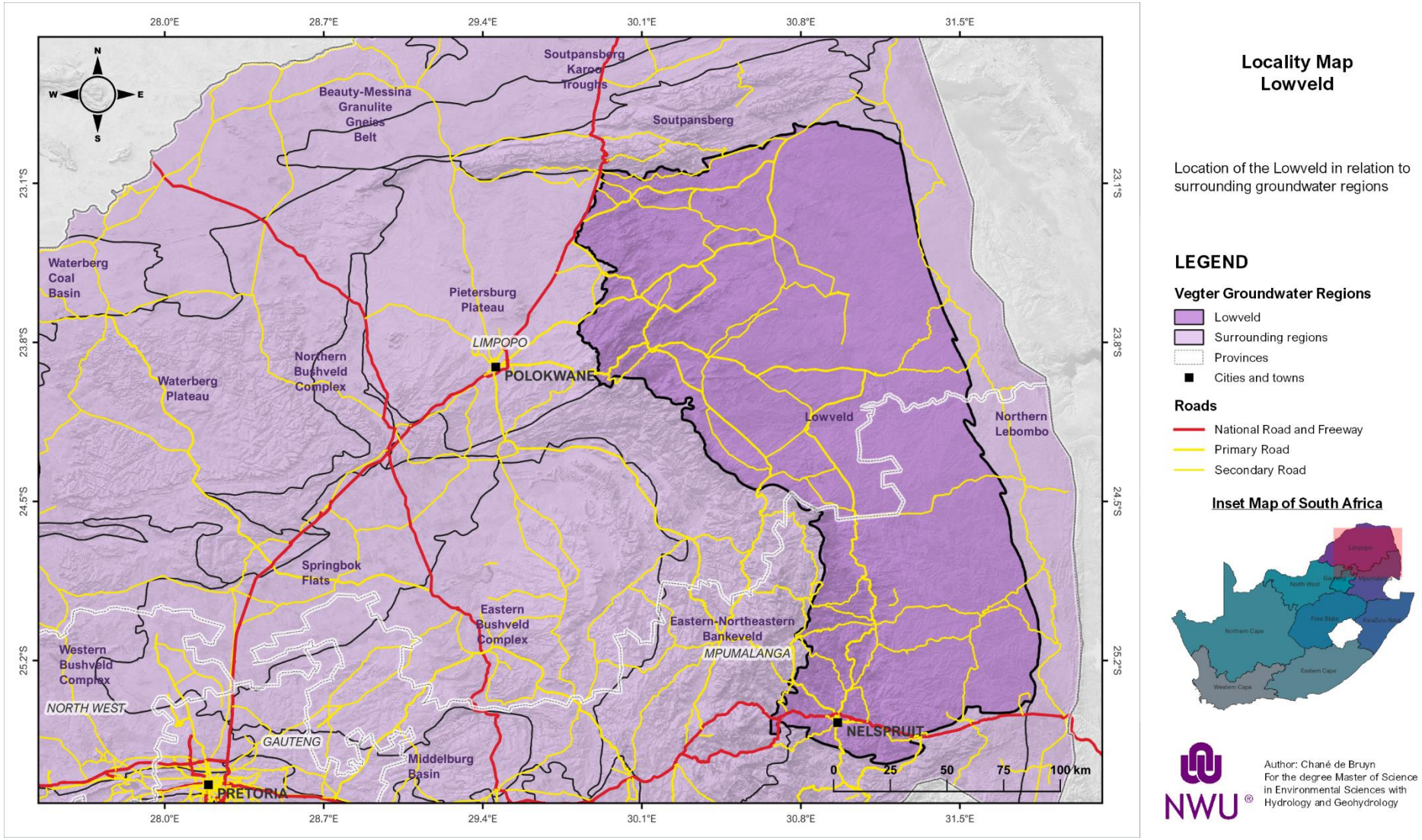
Table 5-1: Borehole data distribution for chosen Vegter regions

Groundwater Region	Total boreholes	Boreholes with static water level data	Boreholes with yield data
Lowveld	17 744	5 848	11 096
Eastern Bushveld Complex	11 656	3 156	7 225
Taung-Prieska Belt	1645	575	738

Each case study will be discussed separately.

5.1 Lowveld case study

The first case study was conducted within the Lowveld groundwater region in view of the large amounts of data available. The locality of the region is depicted in Figure 5-1.



Locality Map
Lowveld

Location of the Lowveld in relation to surrounding groundwater regions

LEGEND

Vegter Groundwater Regions

- Lowveld
- Surrounding regions
- Provinces
- Cities and towns

Roads

- National Road and Freeway
- Primary Road
- Secondary Road

Inset Map of South Africa

NWU®

Author: Chané de Bruyn
For the degree Master of Science in Environmental Sciences with Hydrology and Geohydrology

Figure 5-1: Locality map of the Lowveld groundwater region

5.1.1 Background

The Lowveld groundwater region spans from the Limpopo Province to Mpumalanga. It covers an area of 35 462 km^2 and comprises of 17 744 boreholes, of which 11 096 has water strikes and yield data, while 5 848 has static water level data (Figure 5-2). It should be taken into account that water level data are available for 7 476 boreholes, but the conditions under which the water levels have been captured are not all static. A majority of the water levels were captured during pumping tests at drawdown and recovery periods. These water levels were omitted from analysis as they are not representative of the naturally occurring water levels and would skew results. Useful information that could be applied from pumping test analysis included transmissivity, storage coefficient, and specific capacity estimations (see Figure 5-3). The density of the different borehole distributions is displayed in Table 5-2.

Table 5-2: Borehole density for the Lowveld region

Borehole with specific data	Total boreholes	Density (boreholes/km^2)
All boreholes	17 744	0.50
Boreholes with water level	7 476	0.21
Boreholes with static water level	5 848	0.16
Boreholes with yield	11 096	0.31
Boreholes with transmissivity	2 162	0.06
Boreholes with storage	2 260	0.06
Boreholes with Specific Capacity	2 276	0.06

Each borehole with static water level data has on average two water level entries. There are, however, boreholes with numerous entries. Most notably is borehole *2329BB00004*, which contains 1091 static water level entries spanning over a 126-month period. The change in water level data for this borehole is displayed in Figure 5-4 and it was found that the borehole experiences a maximum drawdown of 3 m. Overall, static water level entries span a time period from 1950/03/21 to 2018/10/12, amounting to approximately 68 years of data, whereas the data are not continuous.

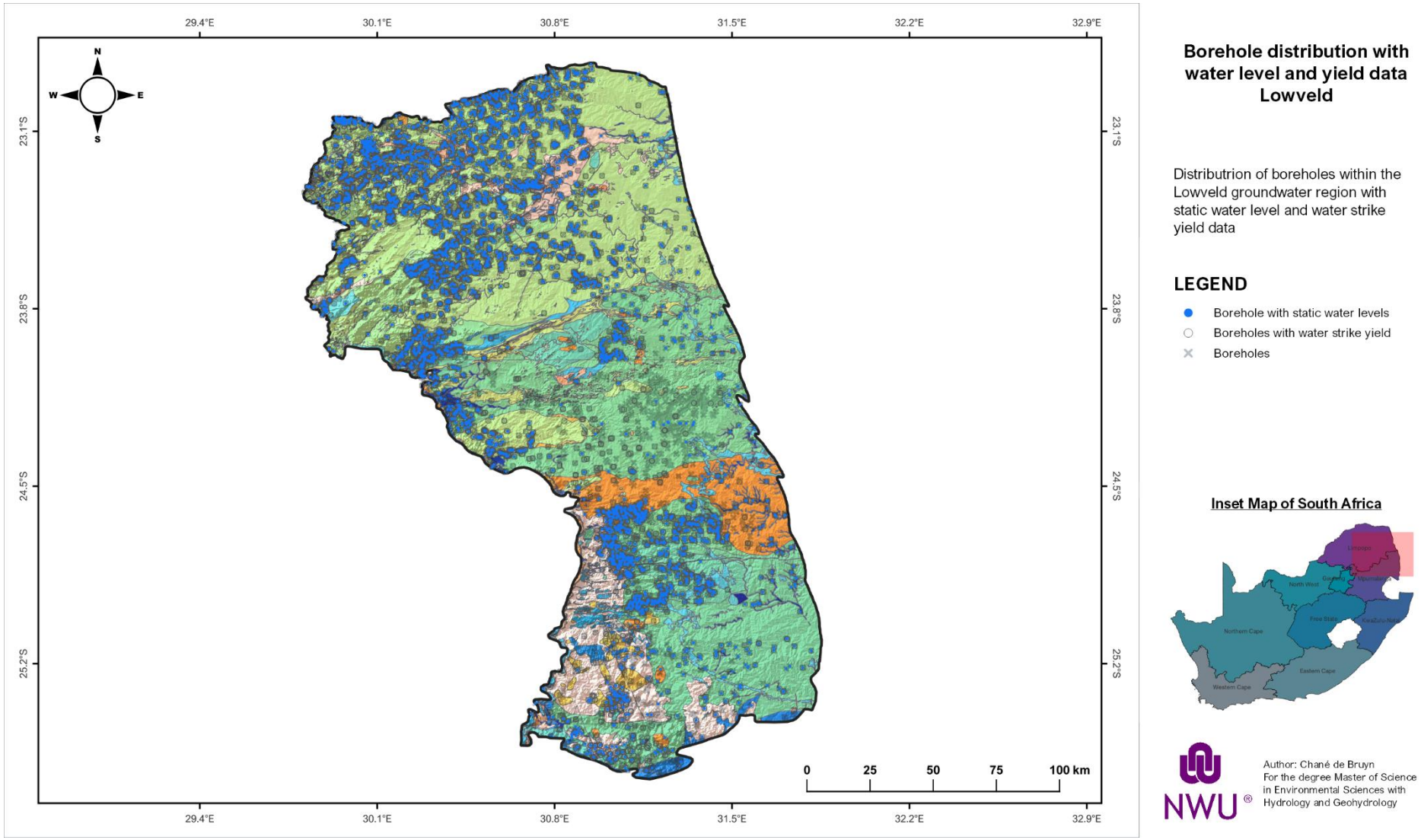


Figure 5-2: Borehole distribution in the Lowveld region – static water levels and yield

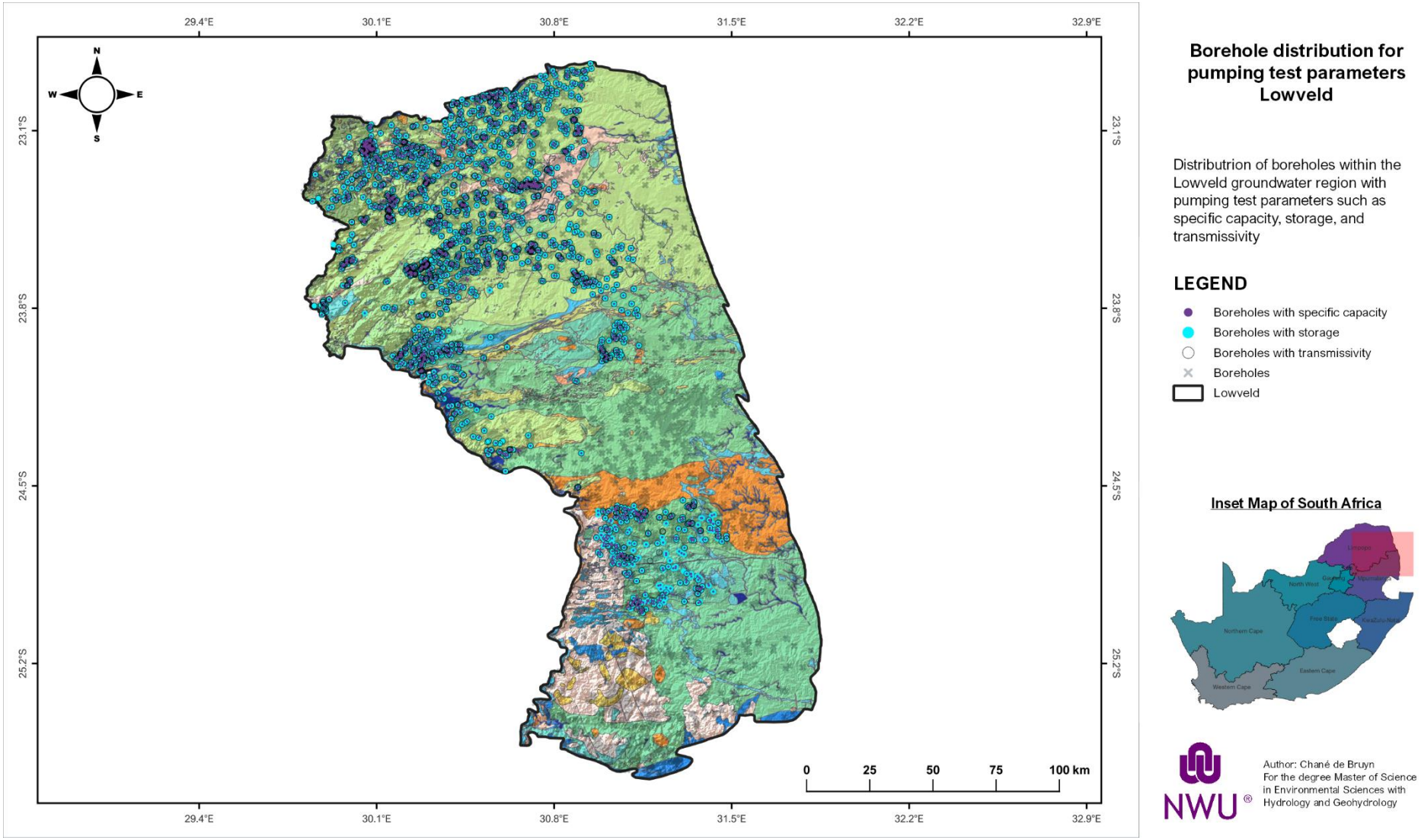


Figure 5-3: Borehole distribution in the Lowveld region – pumping test parameters

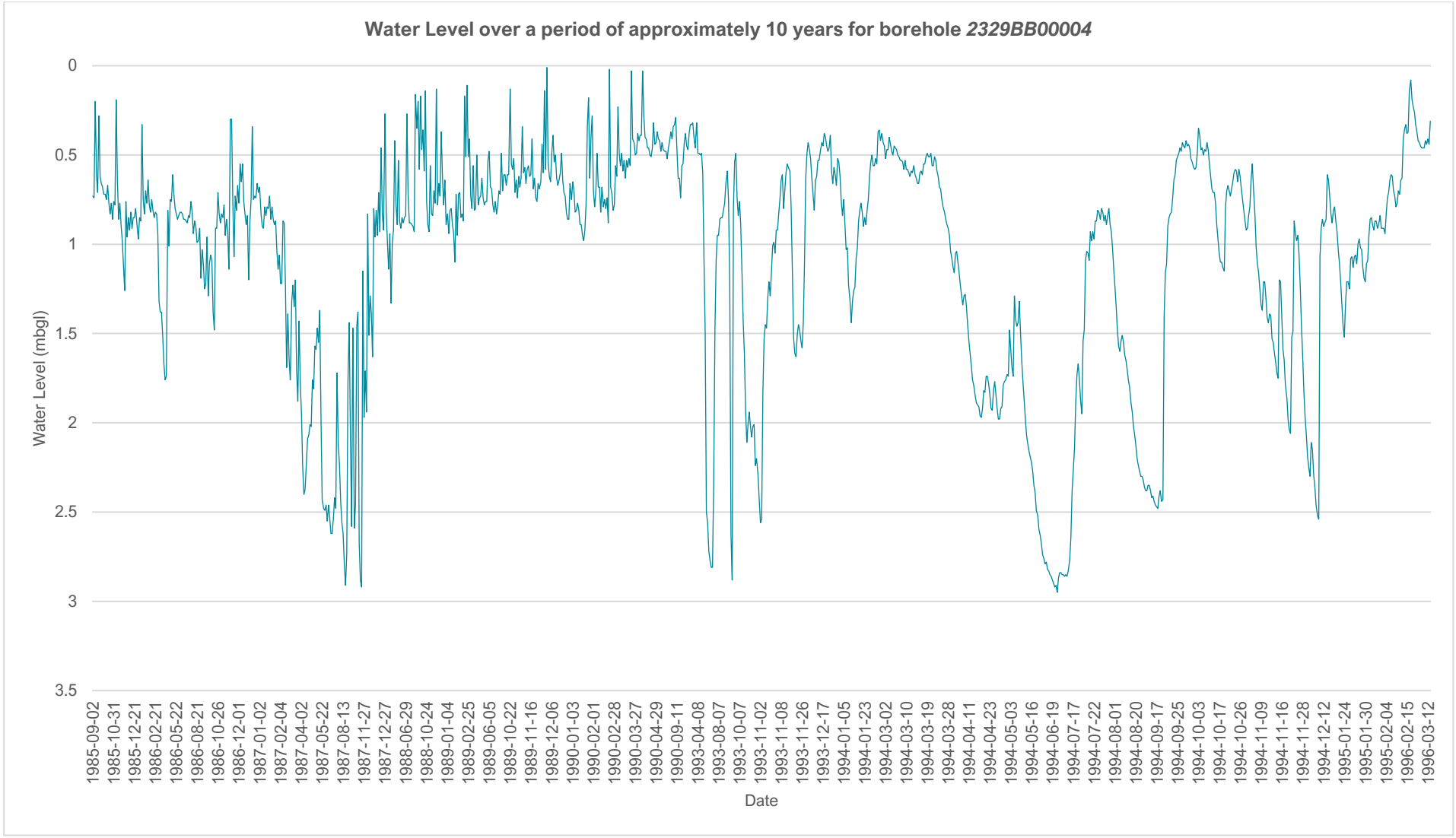


Figure 5-4: Time series water levels for borehole 2329BB00004

5.1.2 Water Level Predictions

Groundwater levels generally follow the topography of the area. The elevation of the Lowveld ranges from 142 mamsl in the lowest area towards the south-east of the Lowveld, and highest at 1 878 mamsl north-western towards mountainous area (Figure 5-6). The correlation between water level and elevation can be seen in Figure 5-5. Several boreholes do, however, deviate from the correlation. This could be caused by the aquifer type in which the boreholes are located or anomalies in the environment, such as geology.

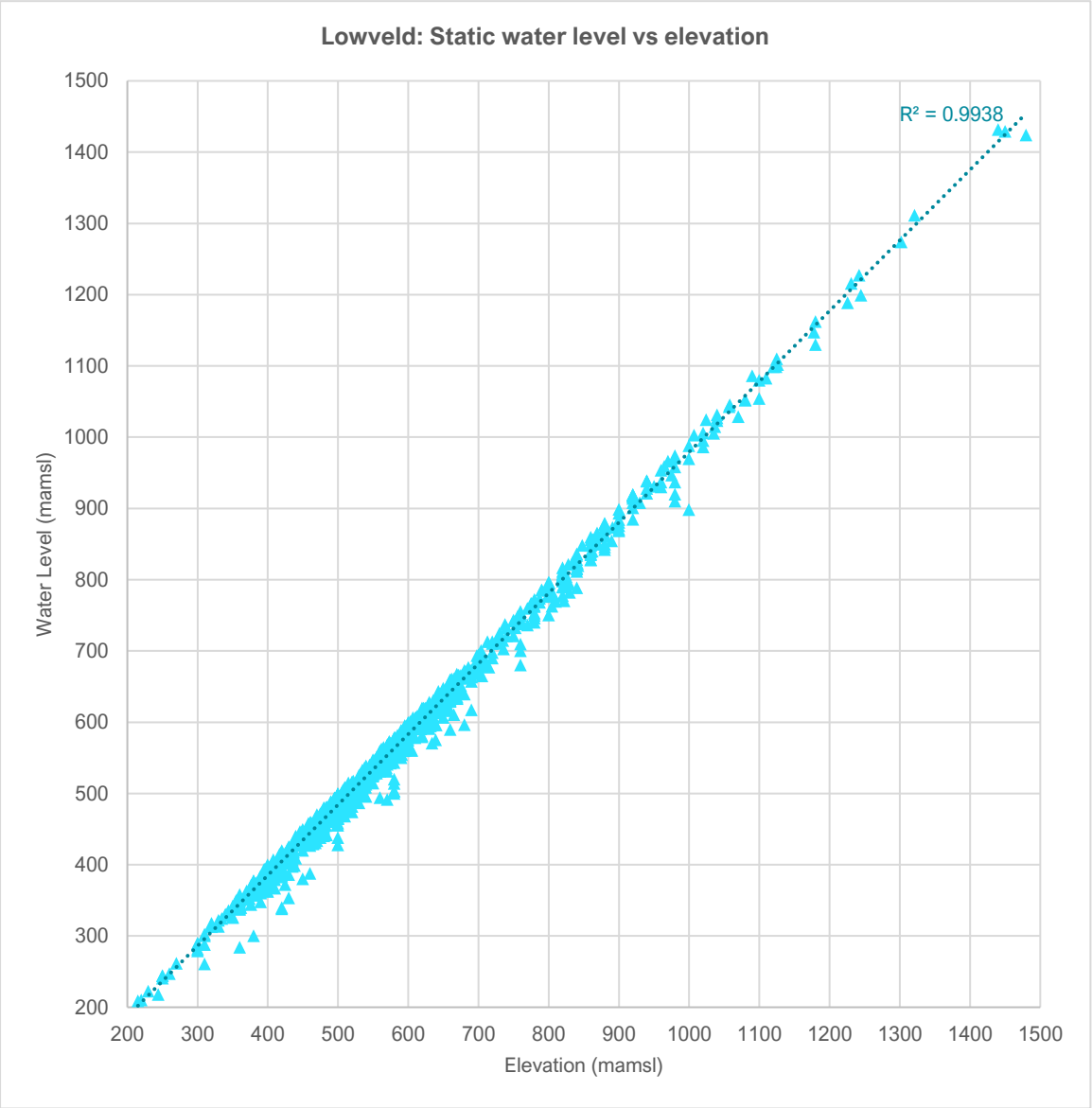
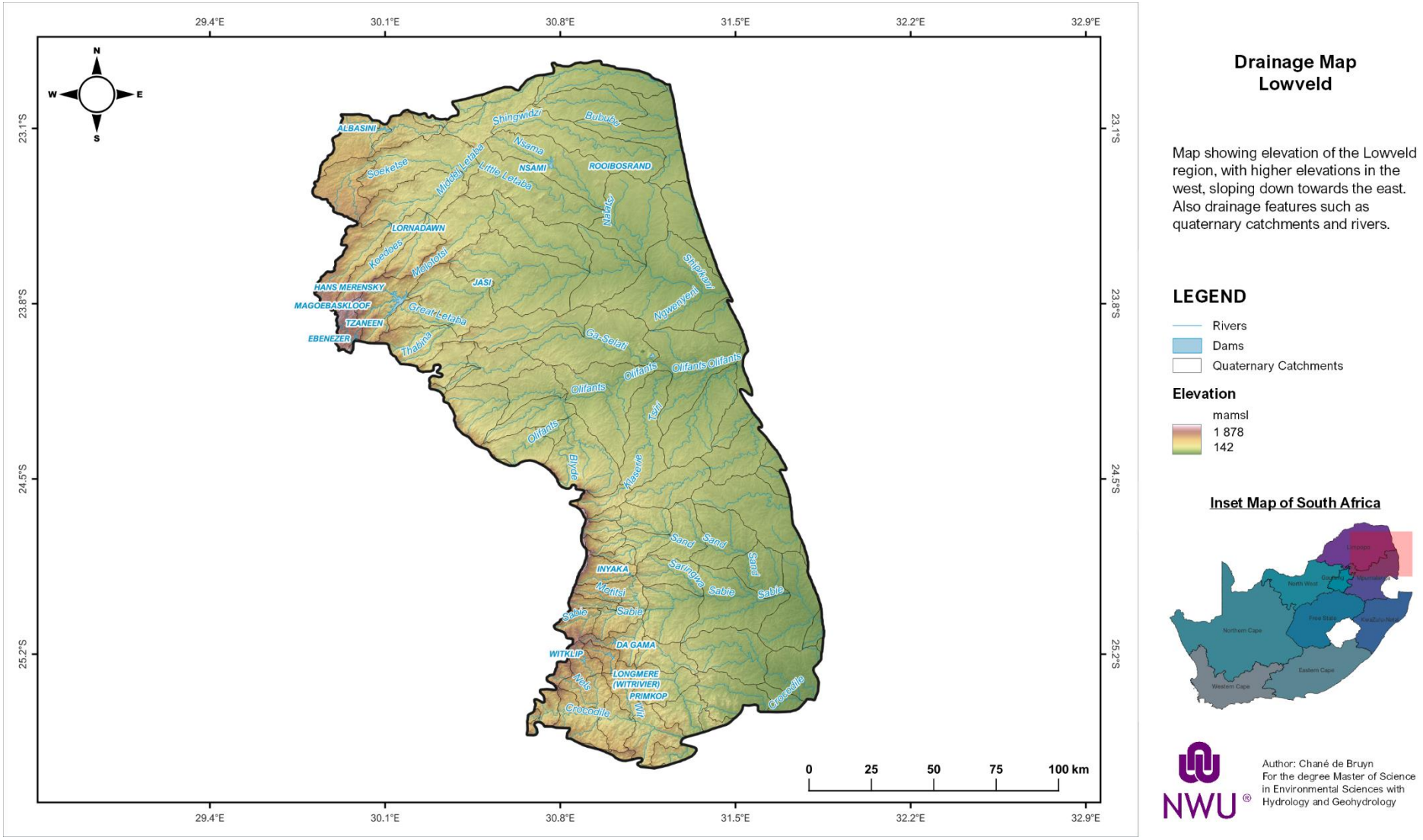


Figure 5-5: Lowveld static water level and elevation correlation



Drainage Map
Lowveld

Map showing elevation of the Lowveld region, with higher elevations in the west, sloping down towards the east. Also drainage features such as quaternary catchments and rivers.

LEGEND

- Rivers
- Dams
- Quaternary Catchments

Elevation

mamsl
1 878
142

Inset Map of South Africa

Author: Chané de Bruyn
For the degree Master of Science in Environmental Sciences with Hydrology and Geohydrology

Figure 5-6: Lowveld elevation and drainage map

During the conceptualisation of the methodology, four critical drivers behind water level could be established by using machine learning and data mining. Elevation is the biggest influencer for water level as established by the clear correlation between elevation and water levels. The three other factors include storativity, mean annual precipitation and geology. These four parameters were used to predict static water levels with high accuracy rates. Maps indicating the geospatial distribution of these parameters are presented in Annexure D – Maps.

Figure 5-7 below shows water levels predicted by using SVR, as well as interpolated water levels by using the Bayesian interpolation. Although both clearly predict water levels with good accuracy, that of the SVR is the closest to the observed water levels.

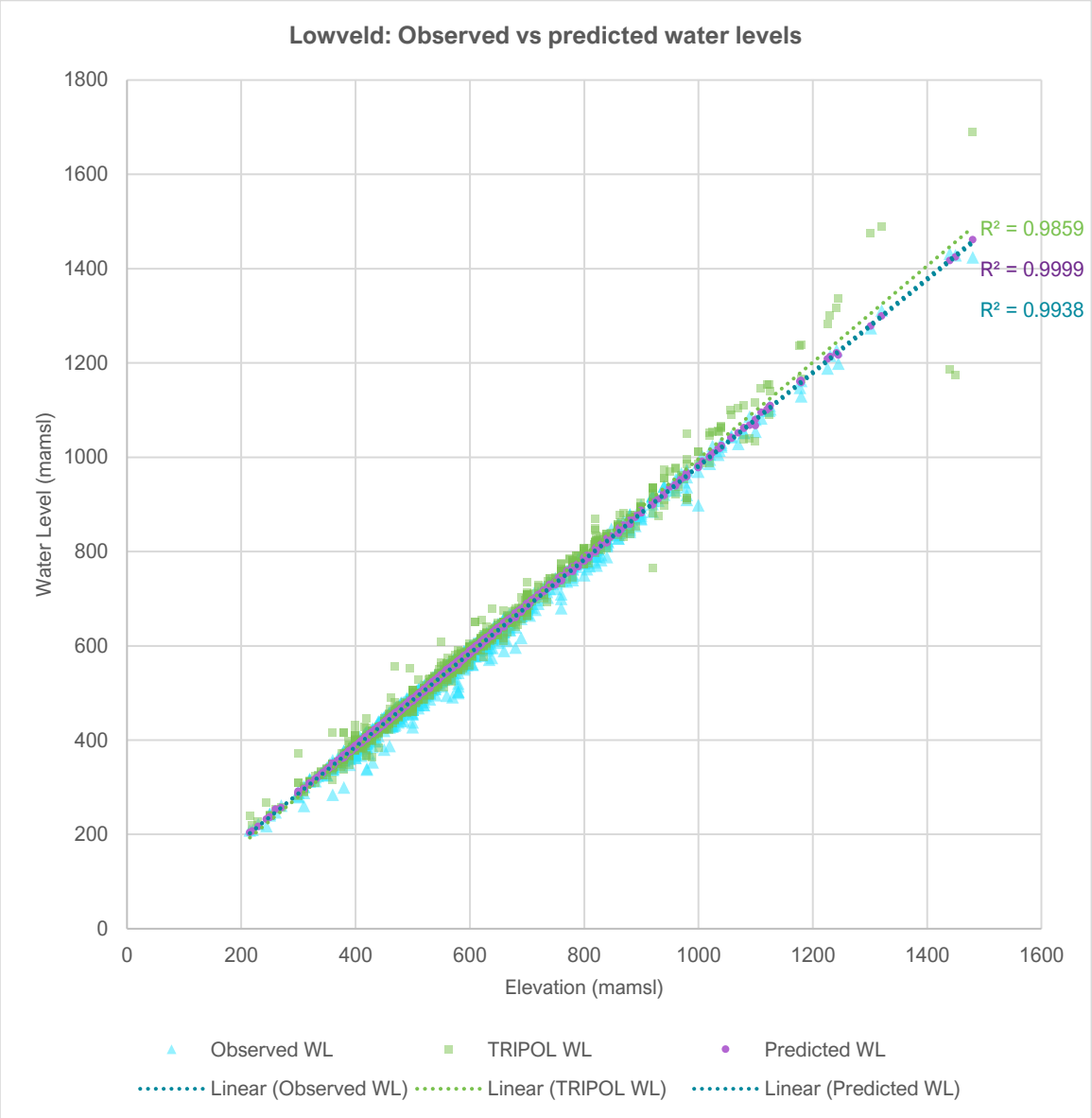


Figure 5-7: Lowveld predicted water level correlation

The SVR model predicts numerical static water levels with a Pearson correlation of 99.69% and an RMSE of 13.74. This clearly contrasts with the Bayesian interpolation, which has an RMSE of 25.11, indicating that the SVR model has the better fit. This may be due to the influence of the extra parameters such as storativity, mean annual precipitation, and lithology. The Bayesian interpolation showed stark deviations in the higher elevations of the Lowveld area. It interpolated numerous water levels being well above the elevation, instead of below the surface. The SVR model results tended to stay close to the trend line, and it could be assumed that elevation was the primary predictor in the model.

Figure 5-8 illustrates the water level elevation, observed water levels, and predicted water levels per entry of the test set. A section of the graph has been magnified due to the density of the data in order to examine the way in which different models performed. In the highlighted section, the variance of the Bayesian interpolation can be observed. The water level was predicted to be above the surface level on numerous occasions, whereas the SVR model water level predictions were never predicted to be above surface level. There were, however, instances where the observed water level was noted to be above the surface. This could have been caused by an error in the database. Alternatively, water levels could have been measured with respect to the borehole casing, which is situated well above ground level, while no correction was done for casing length above ground.

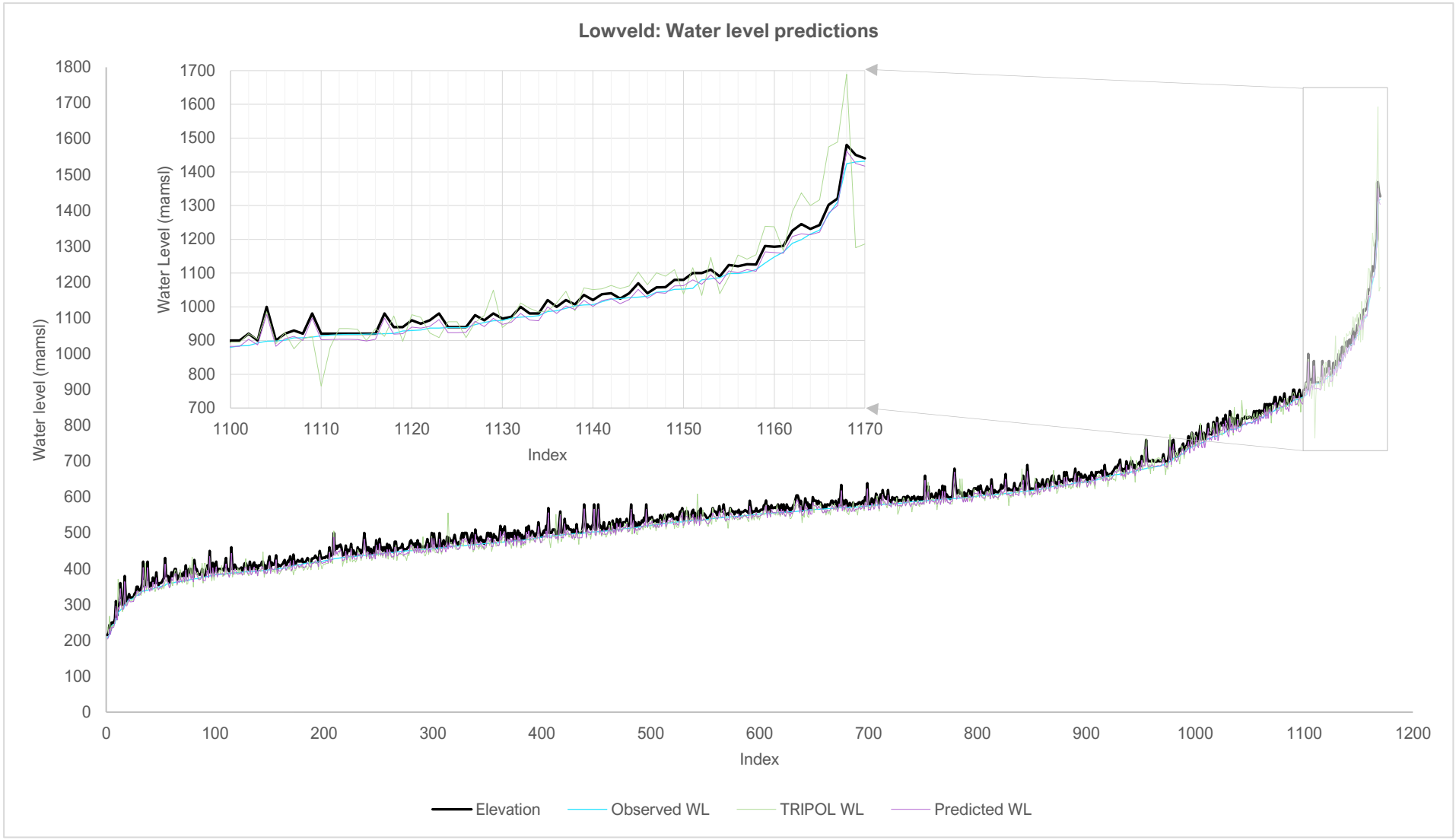


Figure 5-8: Lowveld numerical water level predictions

With regard to classification, the random-forest model predicted classes of 100 m intervals with an accuracy of 91.78% and a Kappa value of 0.90, which was considered perfect (Figure 5-9). For comparative purposes, the elevation was omitted in another run of the model to determine the extent to which the model would cope in the absence of this critical parameter. The model performed relatively well, although a significant decrease in performance was noted with an accuracy of 60.10% and a Kappa value of 0.49, which is moderate. The conclusion can be made that, although elevation is a primary driver, the parameters of storativity, mean annual precipitation, and five subordinate lithologies are also key parameters to be considered in unexplored areas.

1168		OBSERVED CLASS																		
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	R	Tot	CE	UA
PREDICTED CLASS	A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N/A	N/A
	B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N/A	N/A
	C	0	0	15	2	0	0	0	0	0	0	0	0	0	0	0	0	0	17	12%	88%
	D	0	0	1	120	13	0	0	0	0	0	0	0	0	0	0	0	0	134	10%	90%
	E	0	0	0	5	270	11	0	0	0	0	0	0	0	0	0	0	0	286	6%	94%
	F	0	0	0	0	12	308	15	0	0	0	0	0	0	0	0	0	0	335	8%	92%
	G	0	0	0	0	0	6	190	4	0	0	0	0	0	0	0	0	0	200	5%	95%
	H	0	0	0	0	0	0	1	59	3	0	0	0	0	0	0	0	0	63	6%	94%
	I	0	0	0	0	0	0	0	5	58	1	0	0	0	0	0	0	0	64	9%	91%
	J	0	0	0	0	0	1	0	0	4	28	1	0	0	0	0	0	0	34	18%	82%
	K	0	0	0	0	0	0	0	0	0	5	17	0	0	0	0	0	0	22	23%	77%
	L	0	0	0	0	0	0	0	0	1	0	3	3	0	0	0	0	0	7	57%	43%
	M	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	2	50%	50%
	N	0	0	0	0	1	0	0	0	0	0	0	0	0	3	0	0	0	4	25%	75%
	O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N/A	N/A
	P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N/A	N/A
	R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N/A	N/A
	Tot	0	0	16	127	296	326	206	68	66	34	22	3	1	3	0	0	0	1072		
	OE	N/A	N/A	6%	6%	9%	6%	8%	13%	12%	18%	23%	0%	0%	0%	N/A	N/A	N/A			
	PA	N/A	N/A	94%	94%	91%	94%	92%	87%	88%	82%	77%	100%	100%	100%	N/A	N/A	N/A			
	OCA	91.78%						p_o	91.78%												
	K	0.90		Perfect				p_e	0.19244567												

Figure 5-9: Lowveld water level classification prediction

5.1.3 Yield predictions

The yield parameters were more challenging to establish than those of the water levels. The parameters that did seem to have the most influence on the yield were transmissivity, storage coefficients, and specific capacity (especially those established during the pumping test of the borehole), recharge, mean annual precipitation, mean annual runoff, quaternary baseflow, surface lithology, and subordinate lithologies.

Starting off with the parameters gained from pump-tests such transmissivity, storage, and specific capacity, and by using a random-forest regression model, the Pearson correlation was 56.23% and the RMSE 2.93. Figure 5-10 illustrates the observed versus predicted yields.

Another script was run while omitting the pump-test parameters and using interpolated transmissivity and storage values instead, so as to observe the extent to which the model would predict successfully around interpolated data. A substantially greater number of observations were available for the model to use, since all the boreholes with yield could be included, instead of only those that had yield as well as pump-test parameters. Although the RMSE was roughly the same at 2.69, the Pearson correlation was only 30.68%.

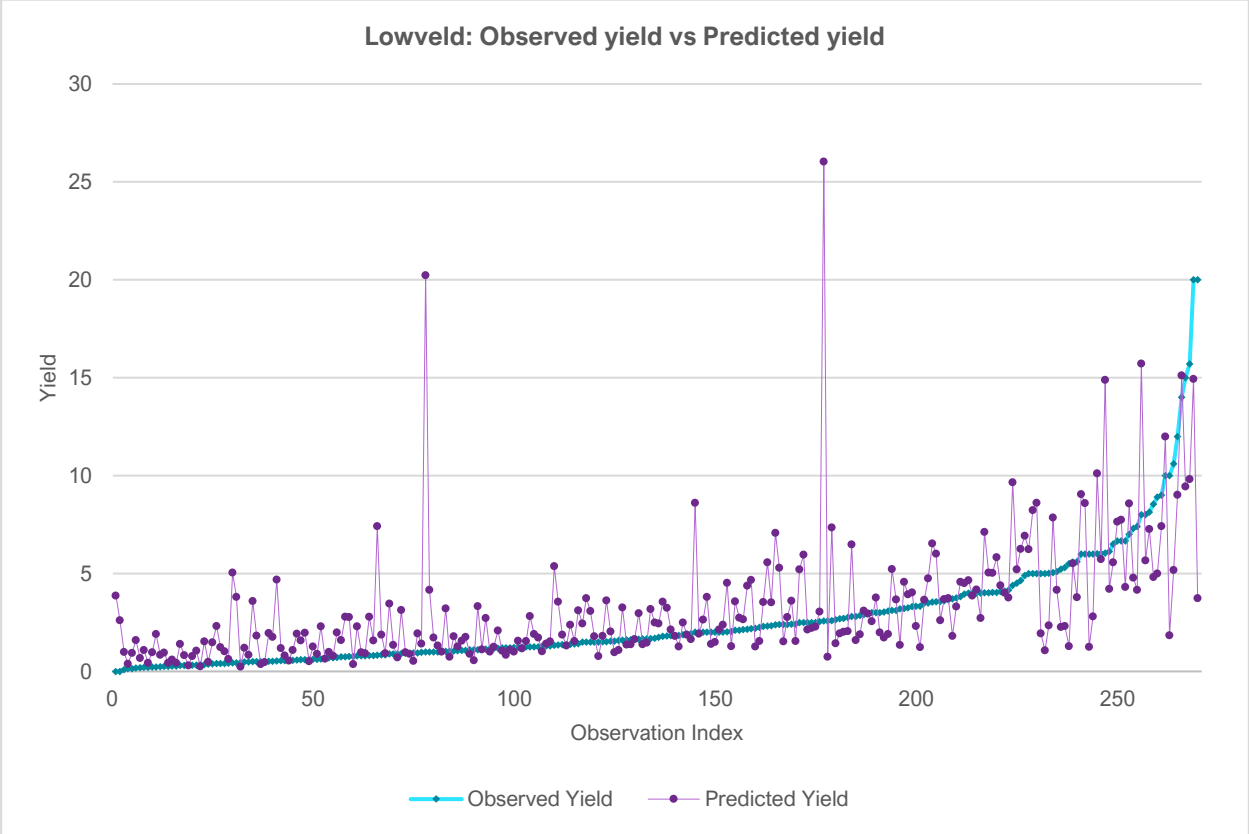


Figure 5-10: Lowveld predicted yield

The same thought process was followed for yield classification. By using the pump-test parameters, the random-forest classification model predicted results with an accuracy of 57.36% and a Kappa value of 0.40, which was considered to be fair, as presented in Figure 5-11. With regard to the use of interpolated transmissivity and storage values, the classification accuracy increased to 66.32%, whereas the Kappa value decreased to 0.18, which is slight. Therefore, the model clearly performed better when information of a greater accuracy representative of the environment was used, which was to be expected.

401		OBSERVED CLASS							
		A	B	C	D	E	Tot	CE	UA
PREDICTED CLASS	A	0	0	3	1	0	4	100%	0%
	B	0	13	20	3	0	36	64%	36%
	C	0	8	94	26	5	133	29%	71%
	D	0	0	34	59	28	121	51%	49%
	E	0	0	7	36	64	107	40%	60%
	Tot	0	21	158	125	97	230		
	OE	N/A	38%	41%	53%	34%			
	PA	N/A	62%	59%	47%	66%			
	OCA	57.36%						p_o	57%
	K	0.40		Fair				p_e	0.29

Figure 5-11: Lowveld yield classification confusion matrix

5.2 Eastern Bushveld Complex Case study

The second case study was conducted in the Eastern Bushveld Complex groundwater region, which also enjoys large amounts of data. The locality of the region is depicted in Figure 5-12.

5.2.1 Background

The Eastern Bushveld Complex stretches across three provinces, namely Limpopo, Gauteng, and Mpumalanga. The largest portion of the region is situated within the Limpopo Province south of

Polokwane. It has an area of approximately 16 807 km^2 and comprises 11 656 boreholes, of which 7 225 enjoys water strike and yield data. Water level data are available for 4 007 boreholes, but only 3 156 boreholes have static water level data (Figure 5-13). Data from pumping test analysis include specific capacity, storage coefficient, and transmissivity. Boreholes that have these data available are indicated in Figure 5-14. Table 5-3 summarises the borehole distribution regarding specific data and density.

Table 5-3: Borehole density for the Eastern Bushveld Complex region

Borehole with specific data	Total boreholes	Density (boreholes/km^2)
All boreholes	11 656	0.69
Boreholes with water level	4 007	0.24
Boreholes with static water level	3 156	0.19
Boreholes with yield	7 225	0.43
Boreholes with transmissivity	1 041	0.06
Boreholes with storage	1 013	0.06
Boreholes with specific capacity	1 045	0.06

On average, more than half of the boreholes with static water levels have one or two noted water levels. Some boreholes show a significantly greater number of static water level entries, but only a few have more than ten entries. The borehole with the most static water level entries in the Eastern Bushveld Complex is *2429BDC0001* with 204 entries. However, all 204 entries were noted during a three-day period. The same is true for *2429DBA001*, which has 180 entries. The fact that these water levels were noted as static may be erroneous and could potentially have been based on a pumping test. The change in water levels for *2429BDC0001* are displayed in Figure 5-15. Regarding all static water level entries for the region, the data spans a time period from 1911/08/12 to 2018/09/04, amounting to approximately 107 years.

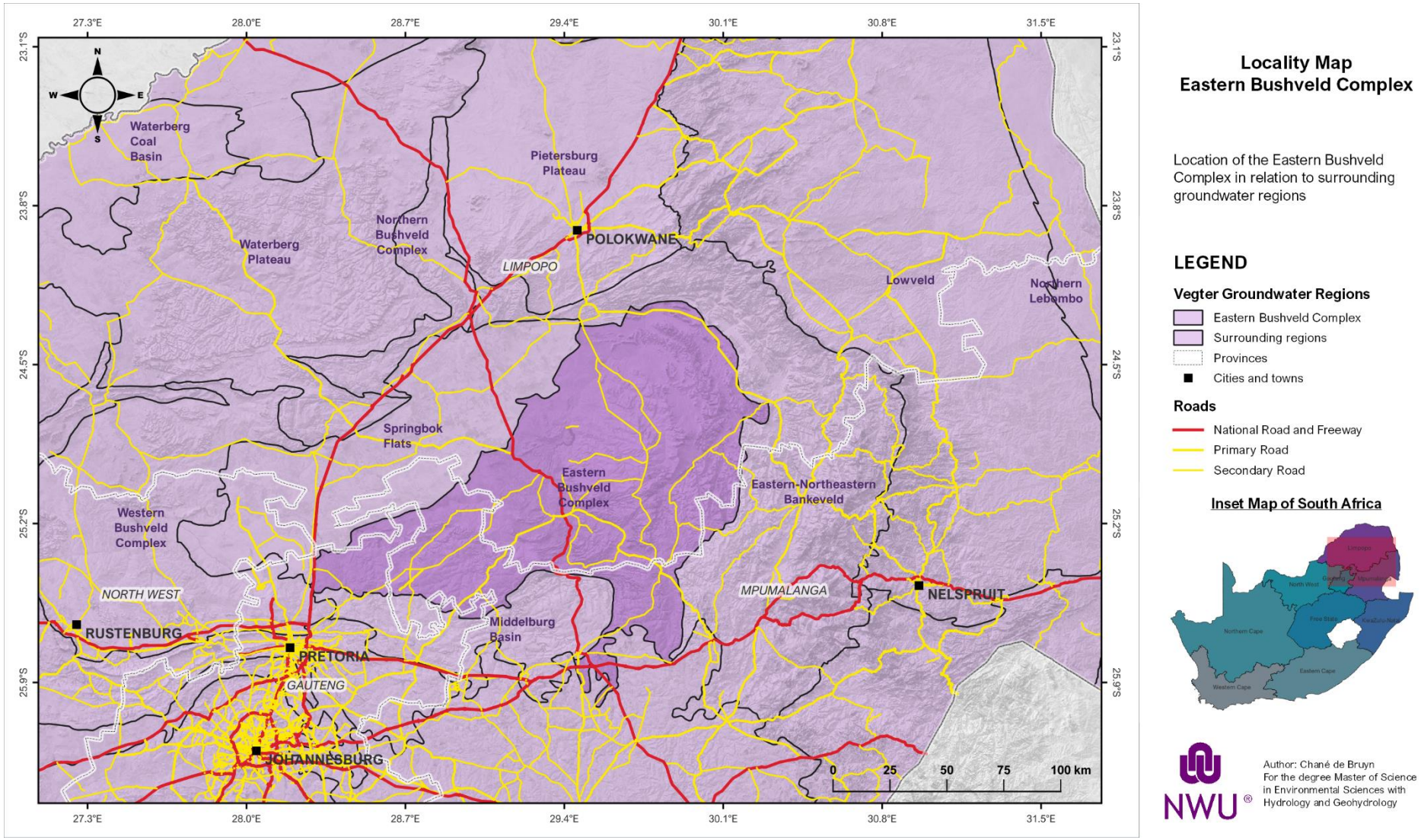


Figure 5-12: Locality map of the Eastern Bushveld Complex groundwater region

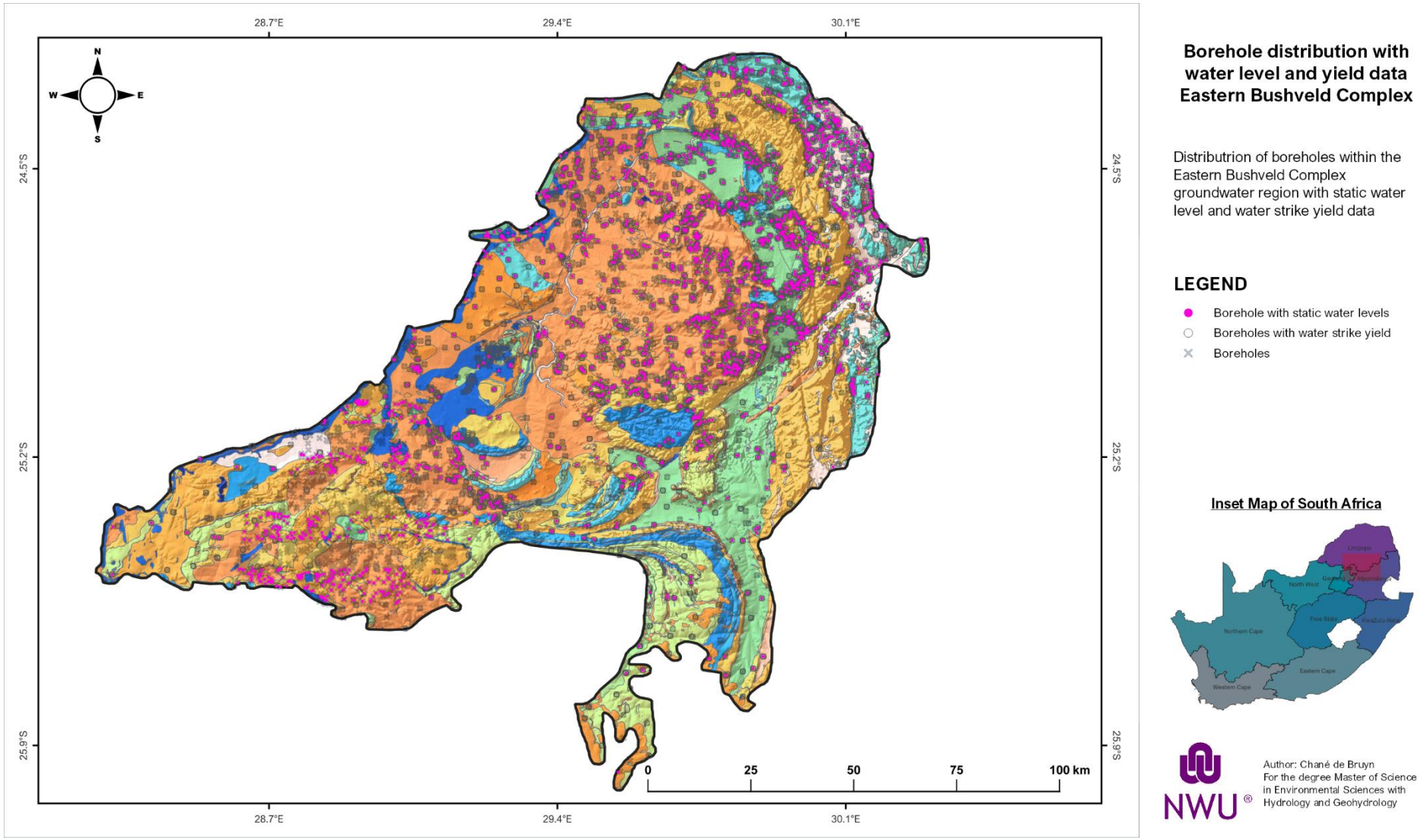


Figure 5-13: Borehole distribution in the Eastern Bushveld Complex region – static water levels and yield

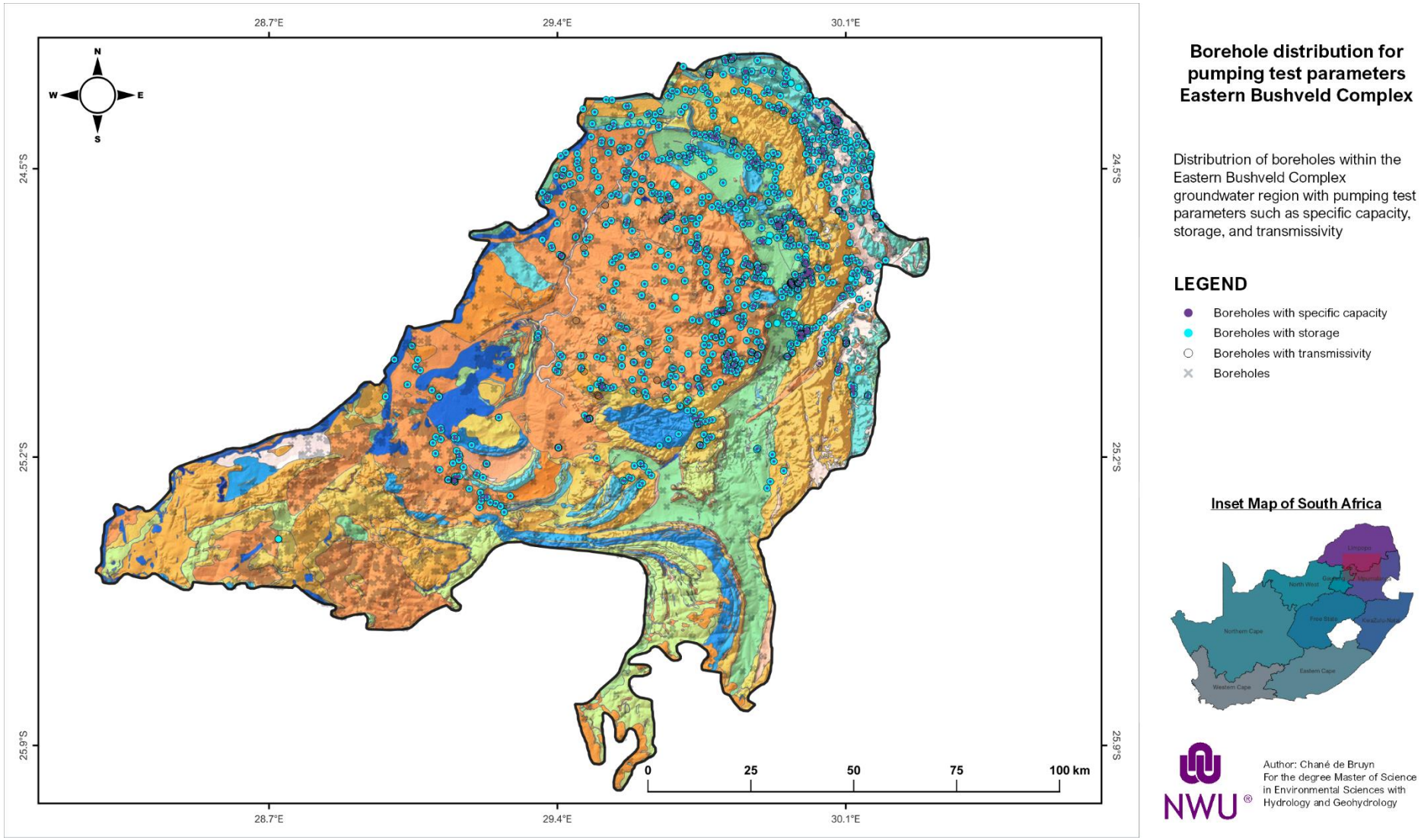


Figure 5-14: Borehole distribution in the Eastern Bushveld Complex region – pumping test parameters

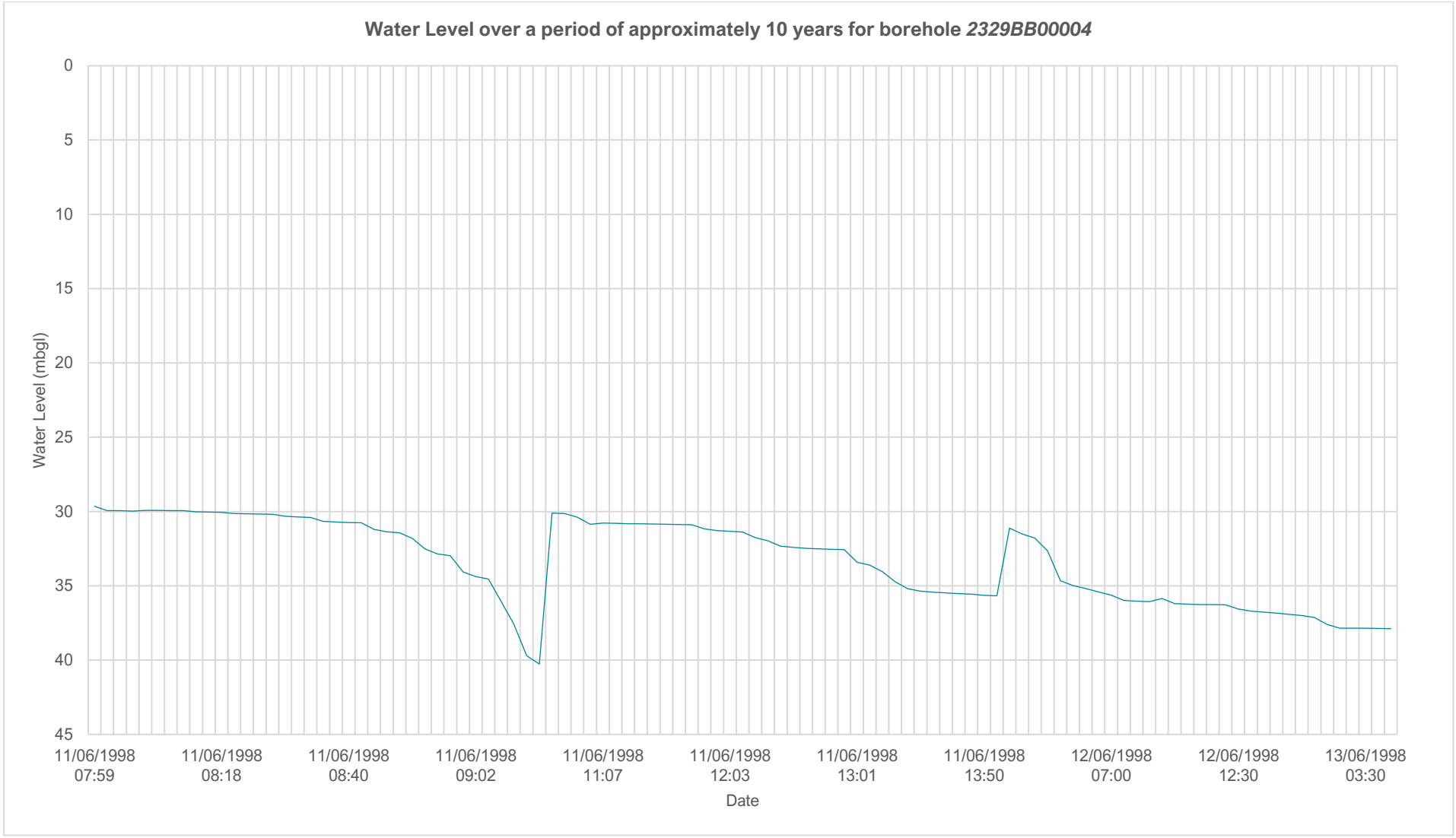


Figure 5-15: Time series water levels for borehole 2429BDC0001

5.2.2 Water level predictions

Groundwater levels strongly correlate with topography, as depicted in Figure 5-16. Various boreholes deviated in the lower to mid elevations. The elevation of the region ranges from approximately 600 mamsl in the lower areas in the north-east of the region to 2 099 mamsl in the higher mountainous areas towards the west. Figure 5-17 shows the elevation in tandem with rivers and quaternary catchments.

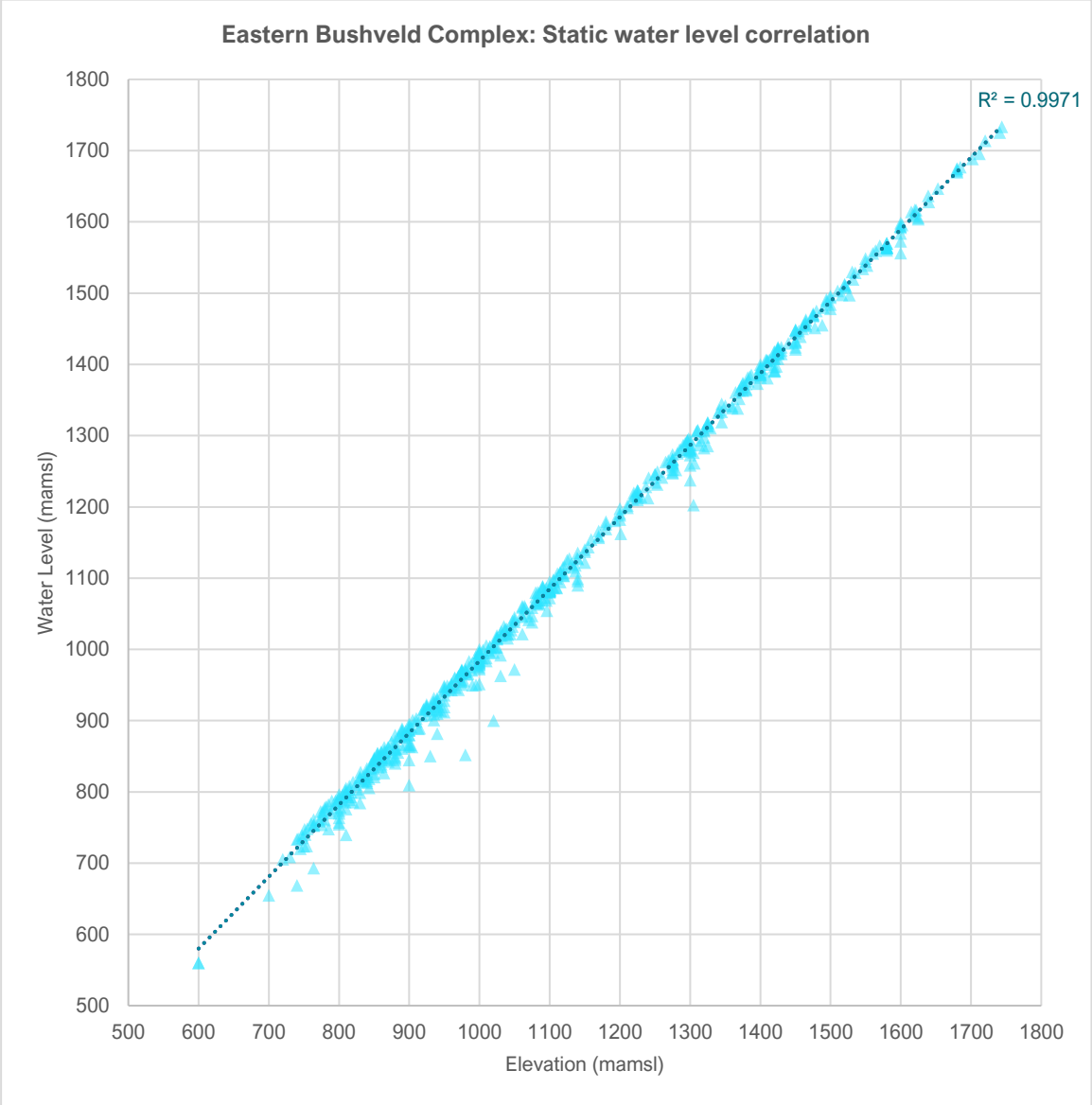
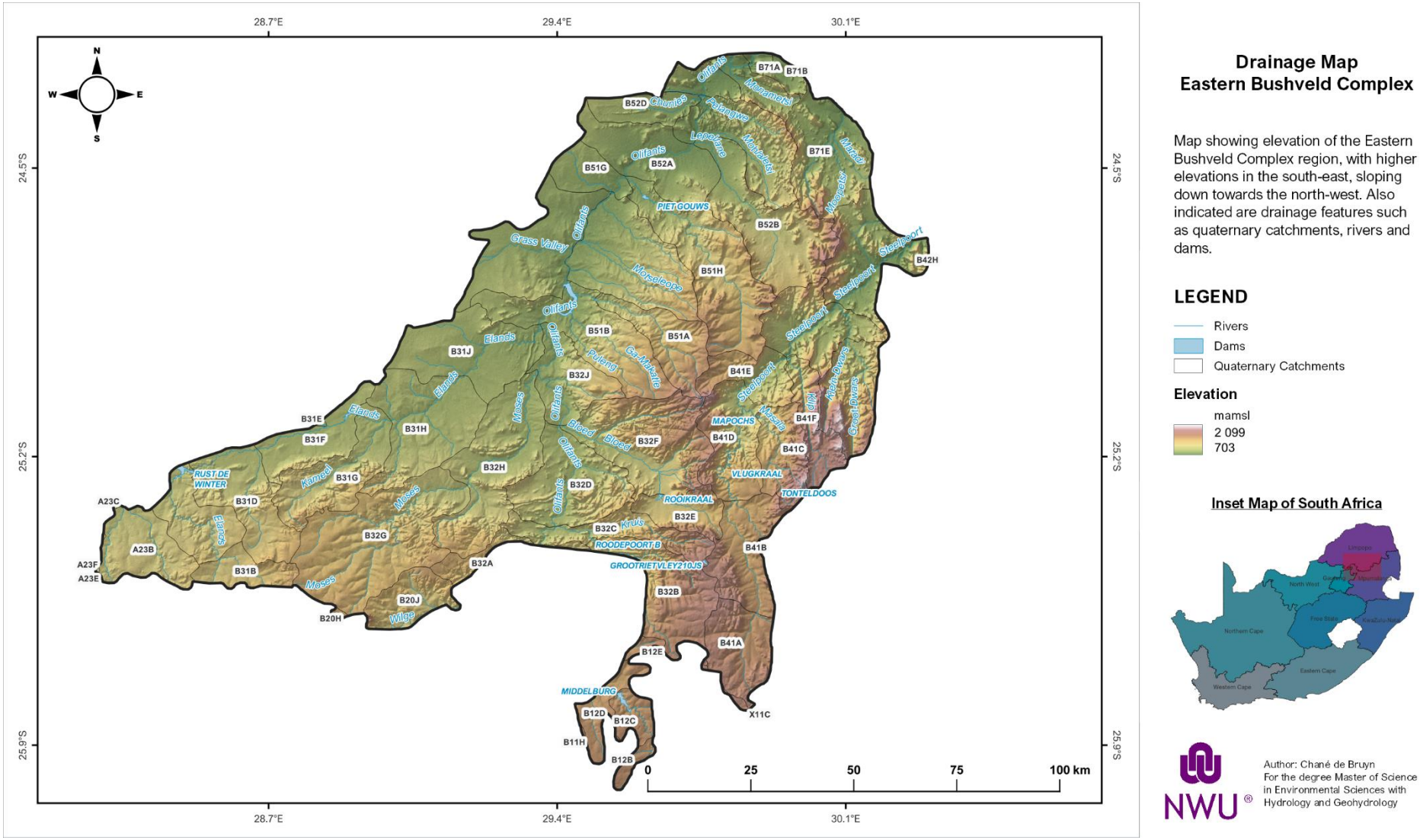


Figure 5-16: Eastern Bushveld Complex static water level correlation



Drainage Map
Eastern Bushveld Complex

Map showing elevation of the Eastern Bushveld Complex region, with higher elevations in the south-east, sloping down towards the north-west. Also indicated are drainage features such as quaternary catchments, rivers and dams.

LEGEND

- Rivers
- Dams
- Quaternary Catchments

Elevation

mamsl
2 099
703

Inset Map of South Africa

Author: Chané de Bruyn
For the degree Master of Science in Environmental Sciences with Hydrology and Geohydrology

Figure 5-17: Eastern Bushveld Complex elevation and drainage map

By means of the methodology employed, elevation was found to be the crucial driver behind water levels. Storativity, mean annual precipitation, and geology were also major drivers. By using these four parameters, static water levels for the Eastern Bushveld Complex were predicted at relatively high accuracy rates. Refer to Annexure D – Maps, which indicates the geospatial distribution of these parameters.

Figure 5-18 reflects the SVR algorithm and Bayesian interpolation predictions alongside the observed static water levels. Both SVR and the Bayesian interpolation predicted water levels were very close to those of the elevation, whereas the observed water levels deviated from the trend line to a greater extent. The SVR model and the Bayesian interpolation performed nearly identically: SVR performed marginally better.

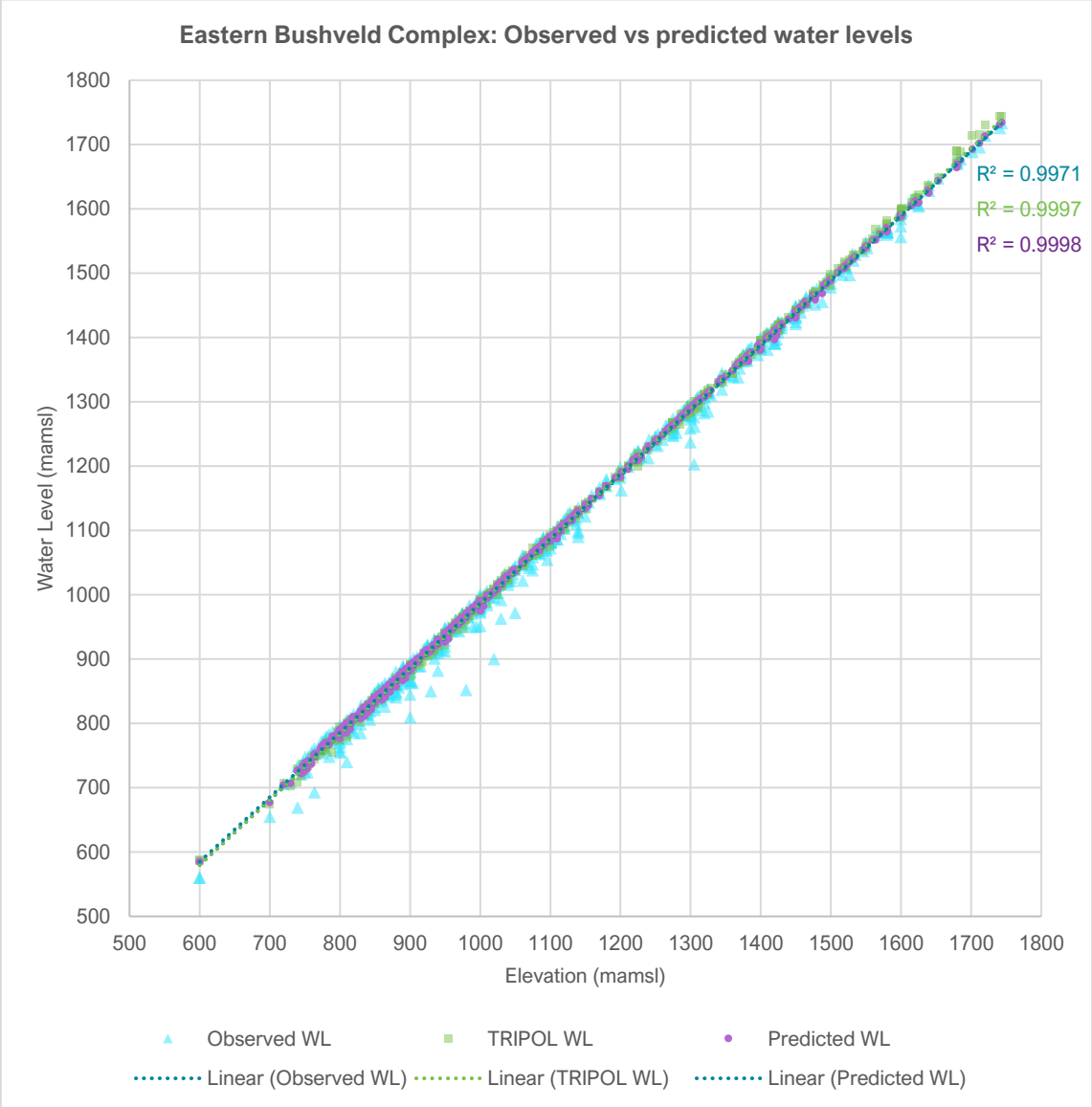


Figure 5-18: Eastern Bushveld Complex predicted water level correlation

The SVR model predicted numeric static water levels with a Pearson correlation of 99.86% and a RMSE of 13.96. Only a subtle difference occurred between predictions of the SVR model (Figure 5-19) and those of the Bayesian interpolation, which had a Pearson correlation of 99.85% and an RMSE of 14.64, therefore rendering the SVR model slightly better. This indicates a stronger correlation between water level and elevation than the other parameters, given that the Bayesian interpolation used only elevation as a parameter. Therefore, it could be presumed that elevation is the primary driver behind water levels in the Eastern Bushveld Complex.

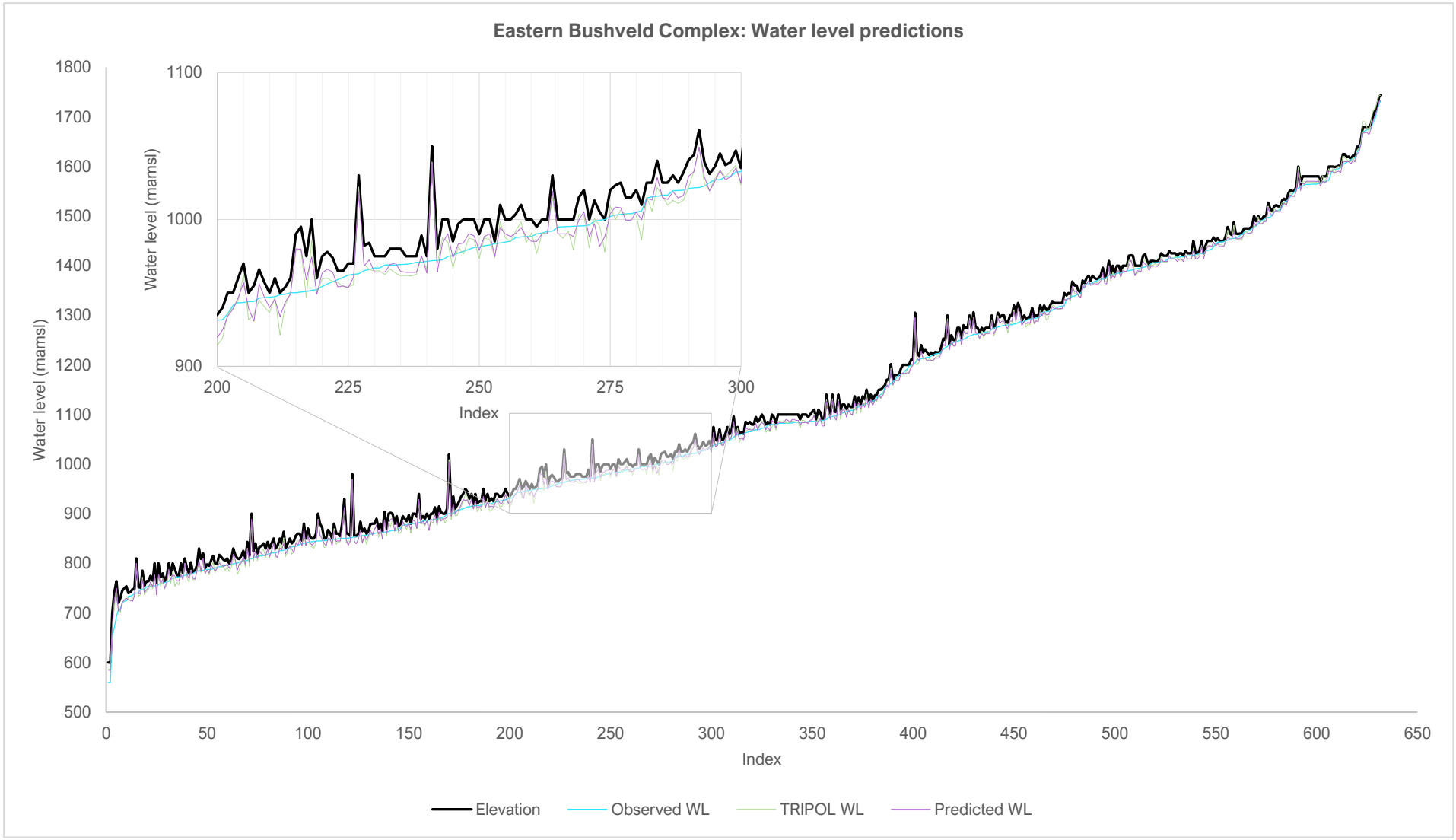


Figure 5-19: Eastern Bushveld Complex predicted water level prediction correlation

Random-forest classification and water level class intervals of 100 m showed an accuracy level of 90.48% and a Kappa value of 0.89, indicating a perfect prediction (Figure 5-20). As in the case of the Lowveld case study, a second run of the model was conducted and elevation was omitted. The performance of the model degraded markedly, lapsing to an accuracy level of only 56.98% and a Kappa value of 0.51, which corresponds to a moderate prediction. The model showed a 33.5% decline in accuracy, whereas the Lowveld study only showed a 17.55% decline in accuracy. This supports to the hypothesis that elevation is a major driver for the Eastern Bushveld Complex water levels.

630		OBSERVED CLASS																	
		A	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	Tot	CE	UA
PREDICTED CLASS	A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N/A	N/A
	E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N/A	N/A
	F	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	100%	0%
	G	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	3	100%	0%
	H	0	0	0	0	55	8	0	0	0	0	0	0	0	0	0	63	13%	87%
	I	0	0	0	0	3	96	3	0	0	0	0	0	0	0	0	102	6%	94%
	J	0	0	0	0	0	3	105	5	0	0	0	0	0	0	0	113	7%	93%
	K	0	0	0	0	0	0	4	71	4	0	0	0	0	0	0	79	10%	90%
	L	0	0	0	0	0	0	0	3	36	3	0	0	0	0	0	42	14%	86%
	M	0	0	0	0	0	0	0	0	0	56	4	0	0	0	0	60	7%	93%
	N	0	0	0	0	0	0	0	0	0	4	54	3	0	0	0	61	11%	89%
	O	0	0	0	0	0	0	0	0	0	0	0	53	2	0	0	55	4%	96%
	P	0	0	0	0	0	0	0	0	0	0	0	1	32	0	0	33	3%	97%
	Q	0	0	0	0	0	0	0	0	0	0	0	0	3	11	0	14	21%	79%
	R	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1	4	75%	25%
	Tot	0	0	0	0	62	107	112	79	40	63	58	57	37	14	1	570		
	OE	N/A	N/A	N/A	N/A	11%	10%	6%	10%	10%	11%	7%	7%	14%	21%	0%			
	PA	N/A	N/A	N/A	N/A	89%	90%	94%	90%	90%	89%	93%	93%	86%	79%	100%			
	OCA	90.48%																p_o	90%
	K	0.89		Perfect														p_e	0.12

Figure 5-20: Eastern Bushveld Complex water level classification confusion matrix

5.2.3 Yield predictions

As mentioned, the parameters for yield predictions were challenging to establish with the available data. The resulting parameter used included pump-test-established parameters such as transmissivity, storage coefficient, and specific capacity, recharge, mean annual precipitation, mean annual runoff, quaternary level baseflow, and geology.

The random-forest regression model predicted numerical yields with a Pearson correlation of 72.09% and an RMSE of 3.31. Figure 5-21 illustrates the predicted yields versus the observed yields.

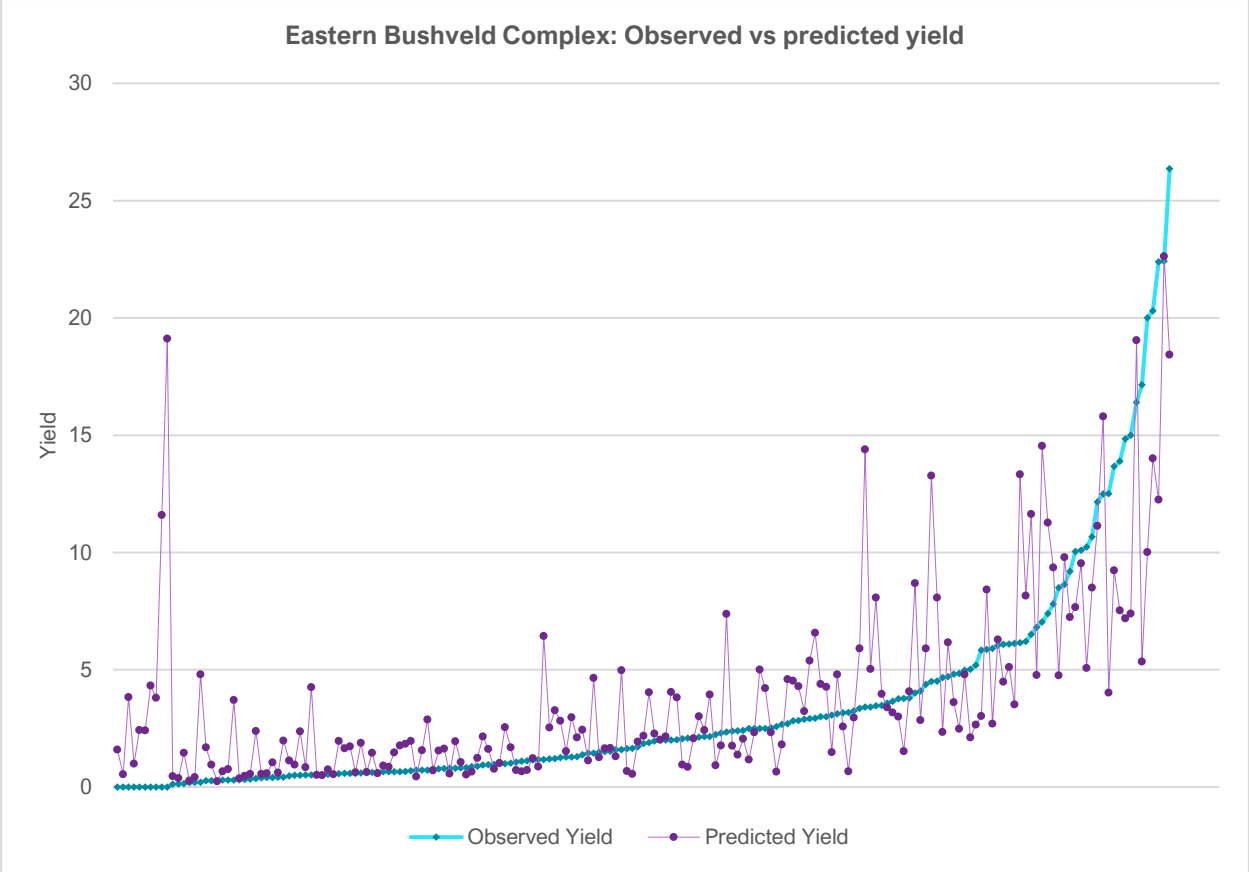


Figure 5-21: Eastern Bushveld Complex predicted yield

The random-forest classification model could predict yield classes at an accuracy of 59.69% and had a Kappa value of 0.44, which is considered to be only moderate, as depicted in Figure 5-22.

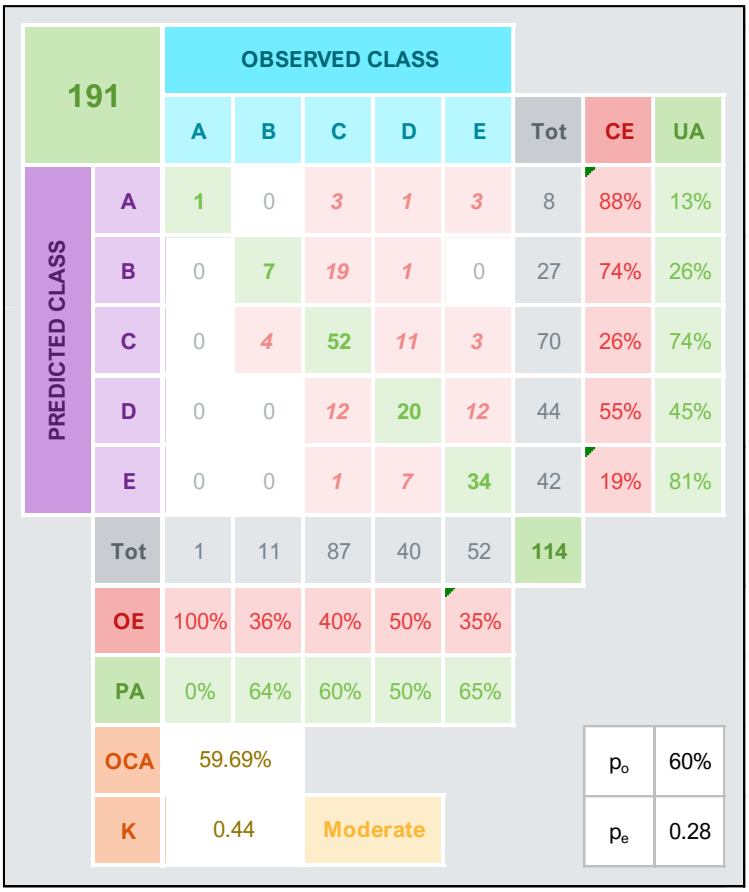


Figure 5-22: Eastern Bushveld Complex yield classification confusion matrix

5.3 Taung-Prieska Belt case study

The third case study was conducted in the Taung-Prieska Belt, also termed Dry Harts-Lower Vaal-Orange Lowland due to rivers that occur in the region. This area was chosen to contrast to the previous two areas, as it contains significantly less data and is located in a more arid region. The location of the region is depicted in Figure 5-23.

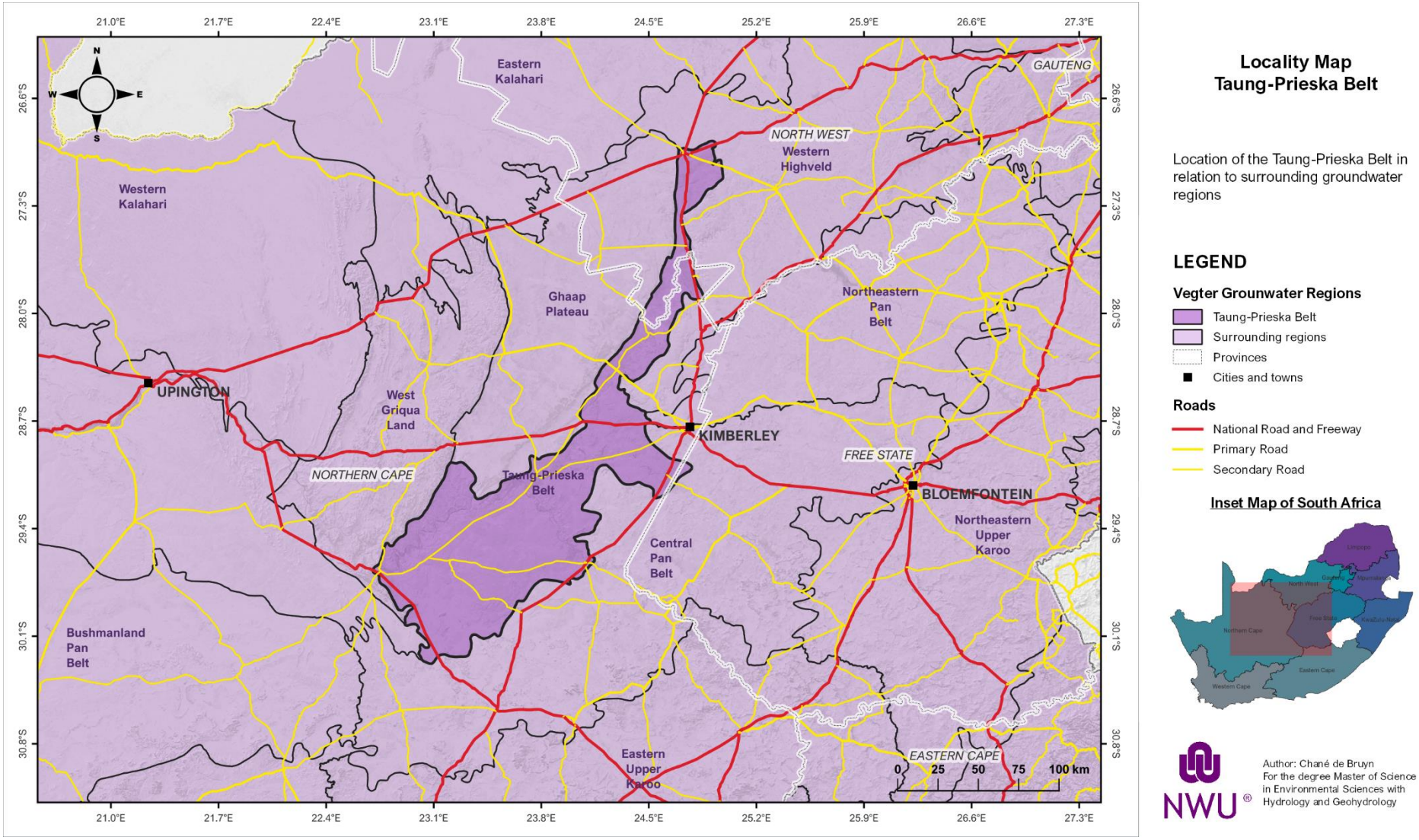


Figure 5-23: Locality map of the Taung-Prieska Belt groundwater region

5.3.1 Background

The larger part of the Taung-Prieska Belt is located in the Northern Cape with a small portion of the northern section situated in the North West Province. The region has an estimated area of 19 206 km^2 and contains 1 645 boreholes. Only 575 of these enjoy static water level data and 738 have yield data (Figure 5-24). There are no pumping test details available for any borehole within the region. Table 5-4 summarises the borehole distribution regarding specific data and the density.

Table 5-4: Borehole density for the Taung-Prieska Belt region

Borehole with specific data	Total boreholes	Density (boreholes/km^2)
All boreholes	1 645	0.09
Boreholes with water level	795	0.04
Boreholes with static water level	575	0.03
Boreholes with yield	738	0.04
Boreholes with transmissivity	0	0
Boreholes with storage	0	0
Boreholes with specific capacity	0	0

Only one water level is noted for the majority of boreholes (more than 80%). There are, however, boreholes with significant numbers of entry. Seven boreholes have entries that reach well above a thousand. The boreholes with the most entries is *2624DC00033* with 8 402 entries, spanning a time period of approximately 46 years. In total, 4 327 individual days of data is captured. The distribution and change in static water levels are depicted in Figure 5-25. Regarding all static water level entries for the region, the data spans a time period from 1913/11/05 to 2015/04/16, amounting to approximately 101 years.

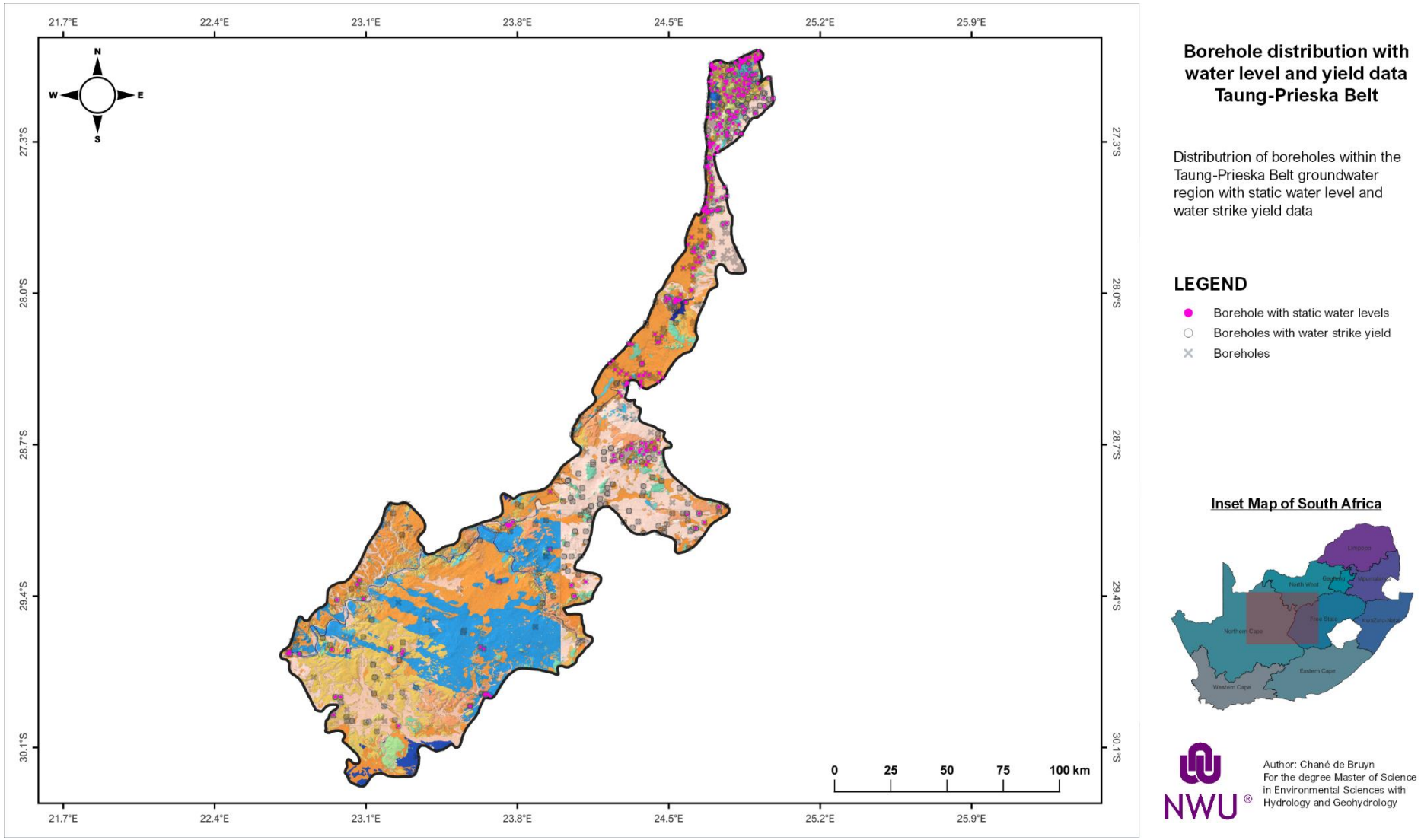


Figure 5-24: Borehole distribution in the Taung-Prieska Belt region – static water levels and yield

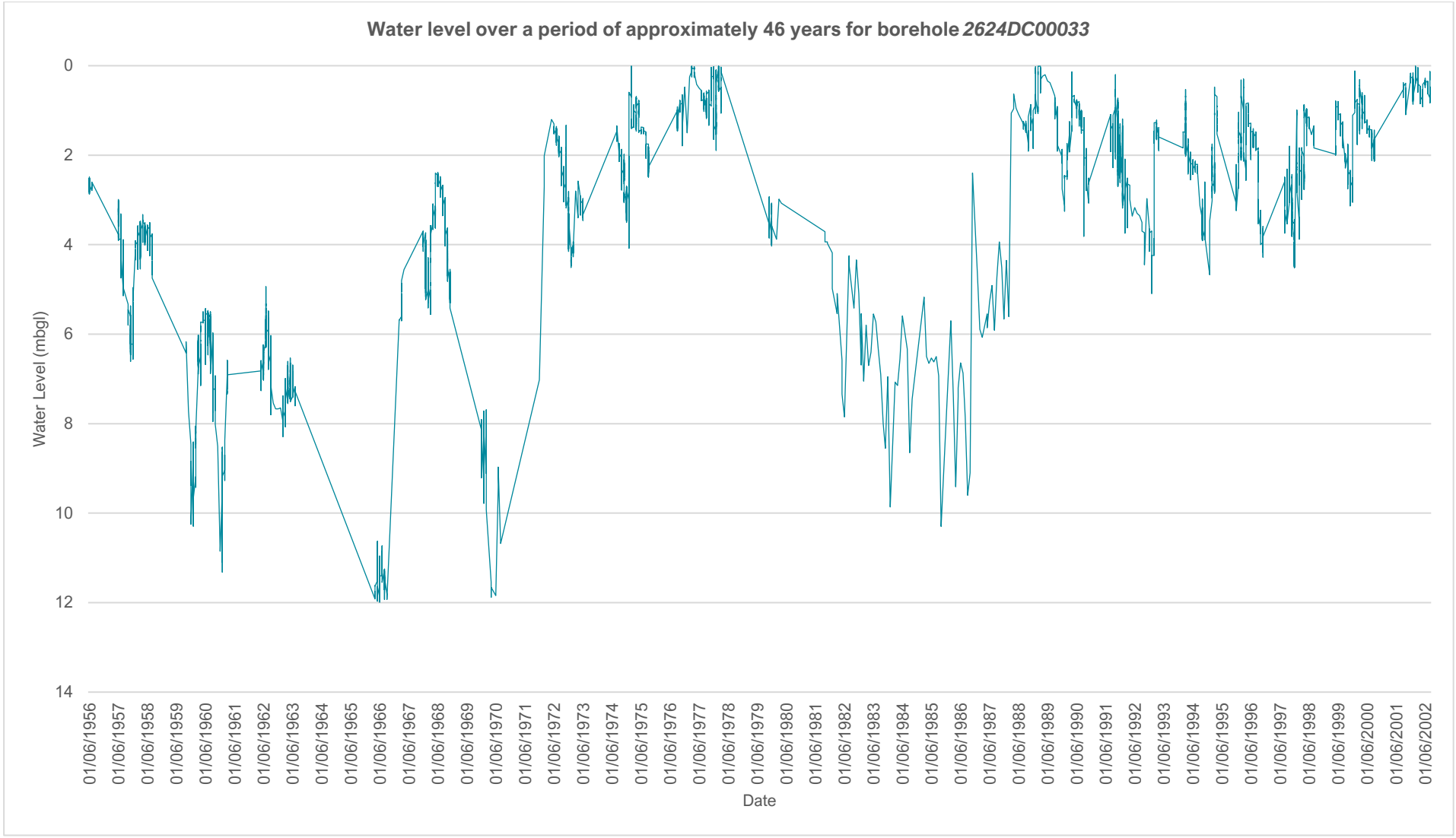


Figure 5-25: Time series water levels for borehole 2624DC00033

5.3.2 Water level predictions

Groundwater levels strongly correspond with the topography of the landscape, as illustrated in Figure 5-26. The elevation in the region ranges from 1 322 mamsl to 912 mamsl in the west where the Orange river flows out of the region. Figure 5-27 shows the elevation along with rivers and quaternary catchments.

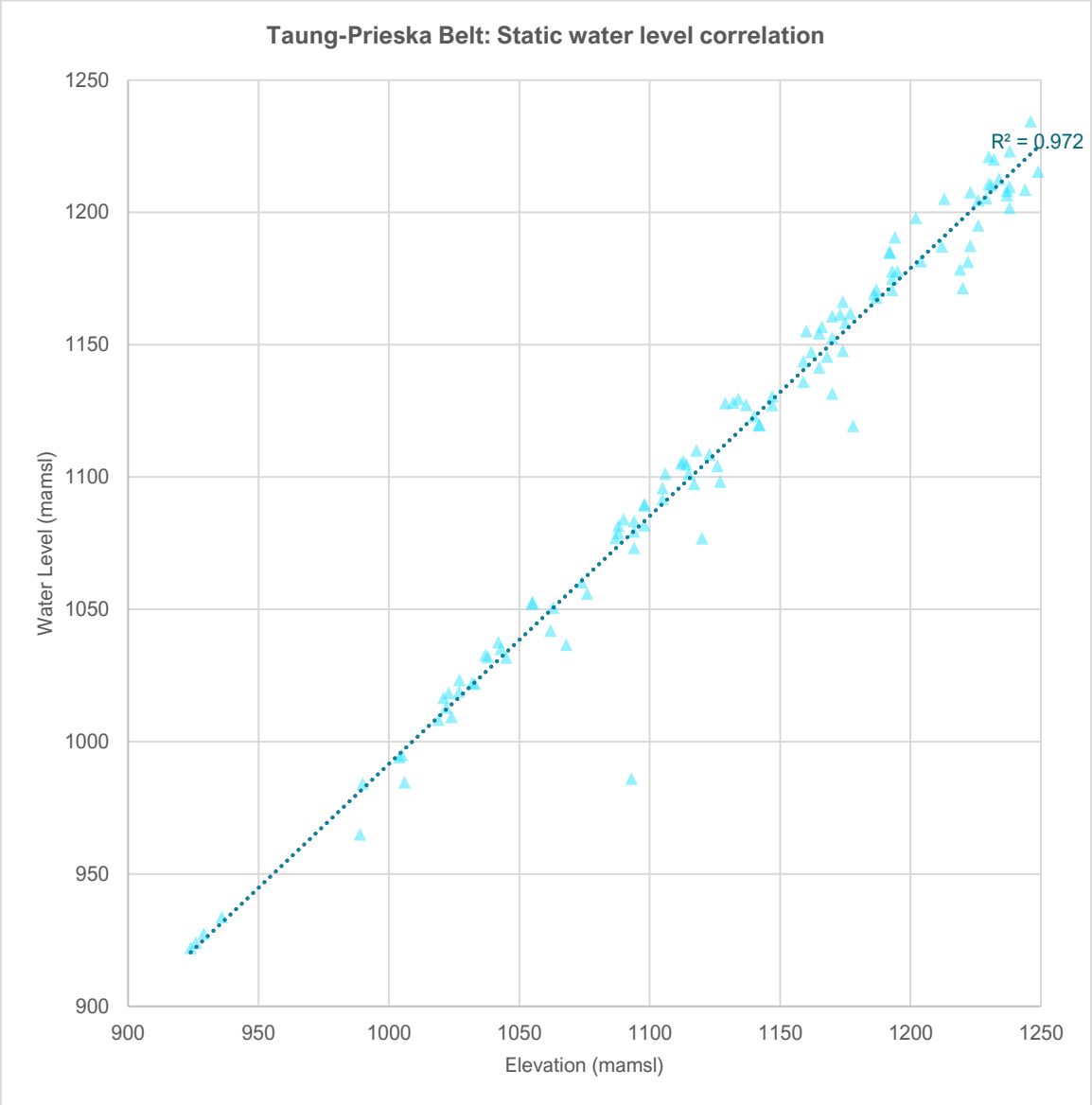
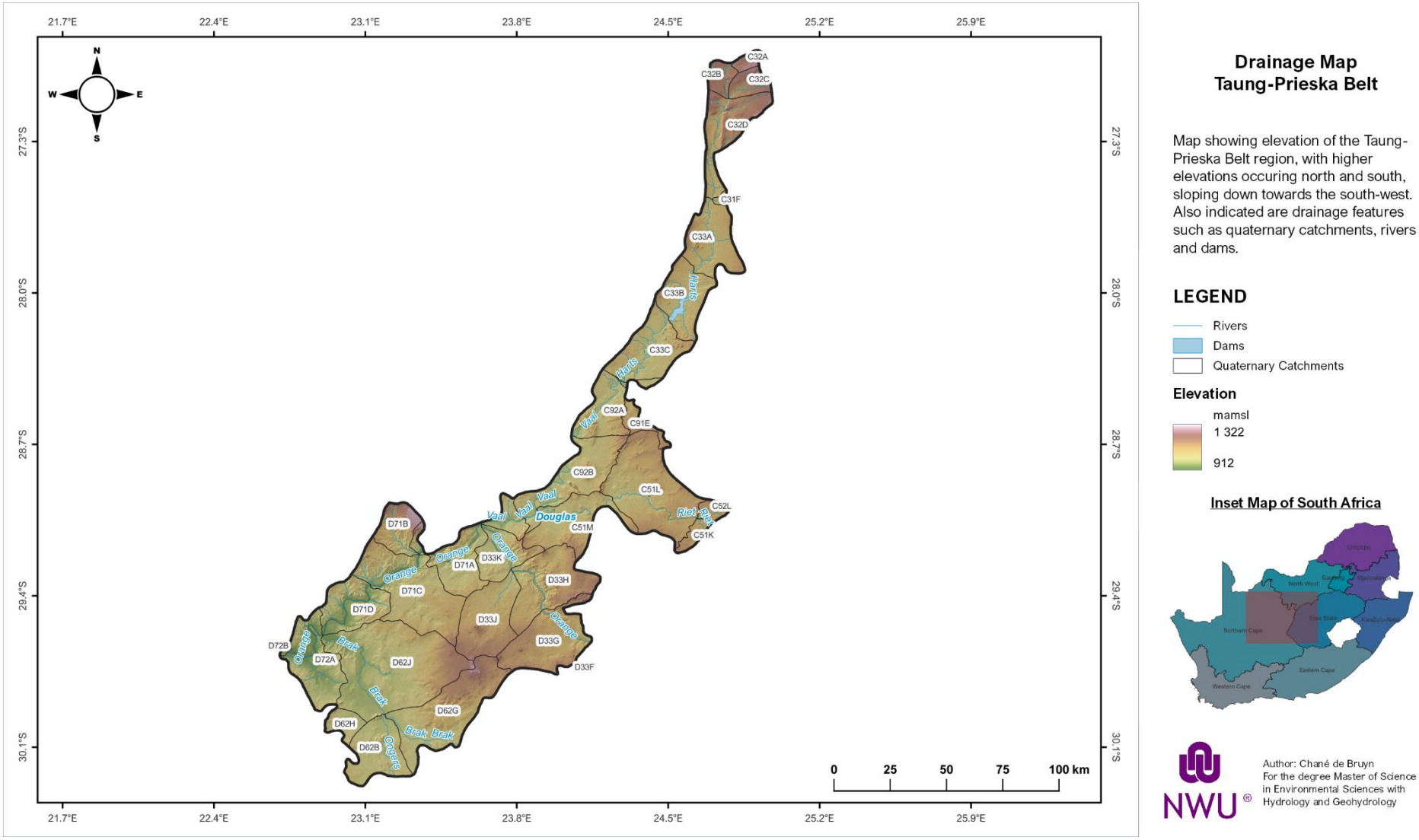


Figure 5-26: Taung-Prieska Belt static water level correlation



Drainage Map
Taung-Prieska Belt

Map showing elevation of the Taung-Prieska Belt region, with higher elevations occuring north and south, sloping down towards the south-west. Also indicated are drainage features such as quaternary catchments, rivers and dams.

LEGEND

Rivers
Dams
Quaternary Catchments

Elevation

mamsl
1 322
912

Inset Map of South Africa

Author: Chané de Bruyn
For the degree Master of Science in Environmental Sciences with Hydrology and Geohydrology

Figure 5-27: Taung-Prieska Belt elevation and drainage map

By implementing the present methodology and by using all four critical parameters, water levels could be predicted with fair accuracy. The SVR model and the Bayesian interpolation predicted relatively uniform results with some variation. The results are indicted in Figure 5-28. Annexure D – Maps indicate the geospatial distribution of the mentioned parameters.

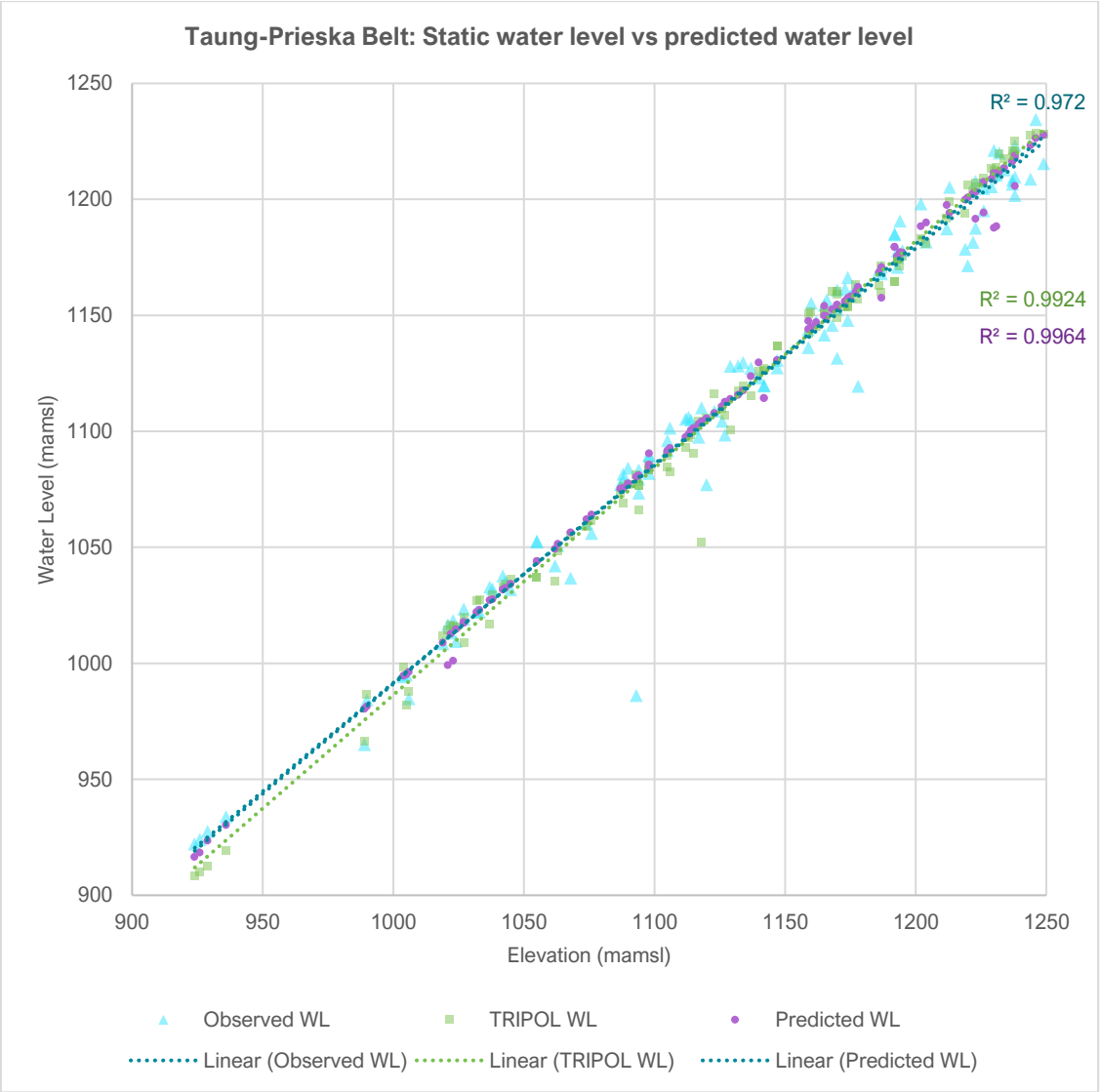


Figure 5-28: Taung-Prieska Belt predicted water level correlation

By using SVR, the regression model produced predictions with a Pearson correlation of 98.45% and an RMSE of 13.47. Only minor differences occurred between the results of the SVR model and those of the Bayesian interpolation (Figure 5-29). The Bayesian interpolation had a Pearson correlation of 98.10%, and an RMSE of 15.37. Given this small difference in results, it can be presumed that elevation is the primary parameter for predictions.

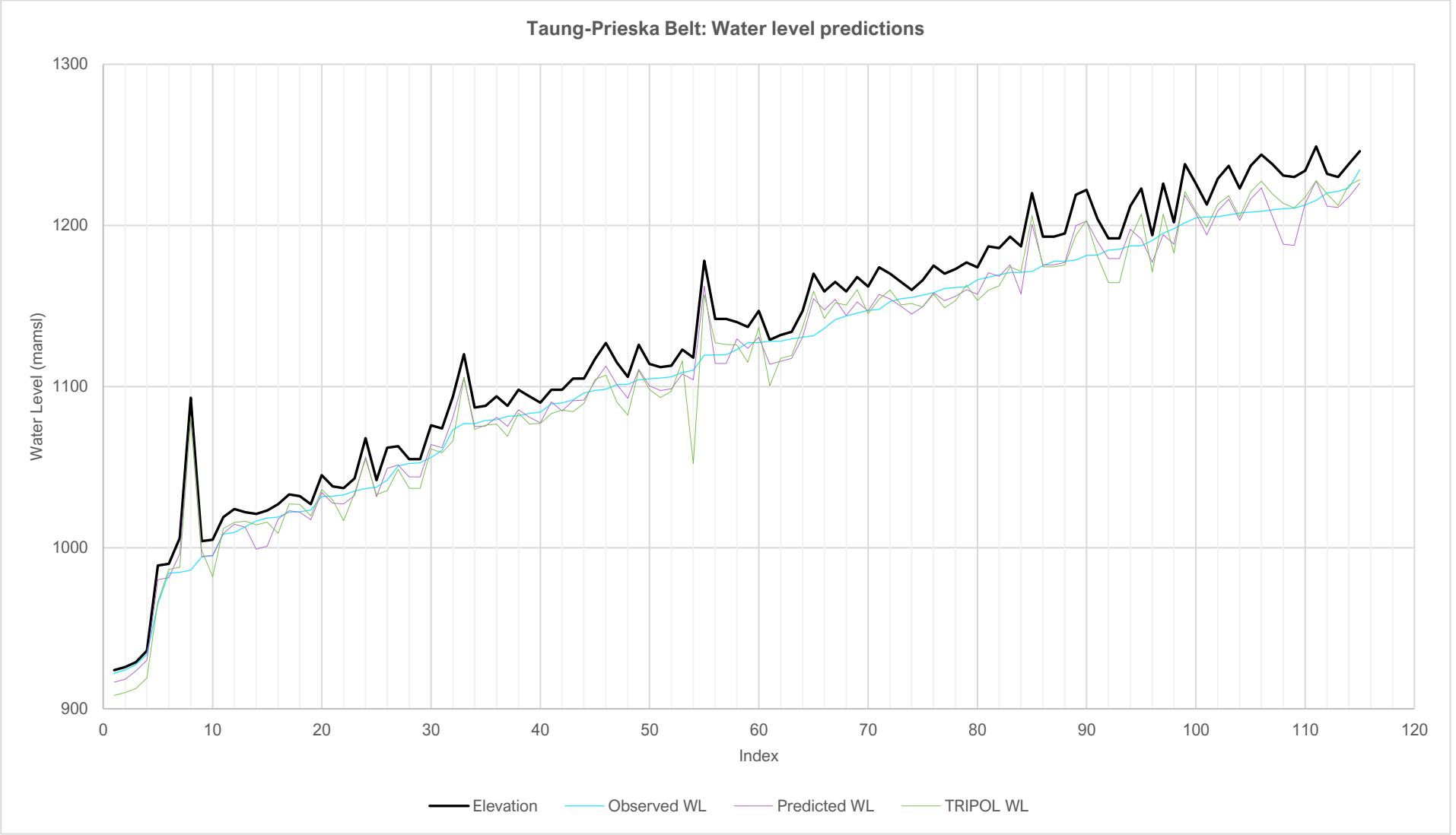


Figure 5-29: Taung-Prieska Belt water level prediction correlation

The random-forest classification model classified water levels on the basis of 100 m intervals with an accuracy of 87.83% and a Kappa value of 0.80, which is considered substantial (Figure 5-30). A second run was conducted to assess how well the model would perform with the critical parameter elevation omitted from the dataset. The model performed well, with an accuracy of 75.65%, which is regarded to be moderate, and a Kappa value of 0.61. This possibly established that the other parameters were of equal importance to elevation for a class-based water level.

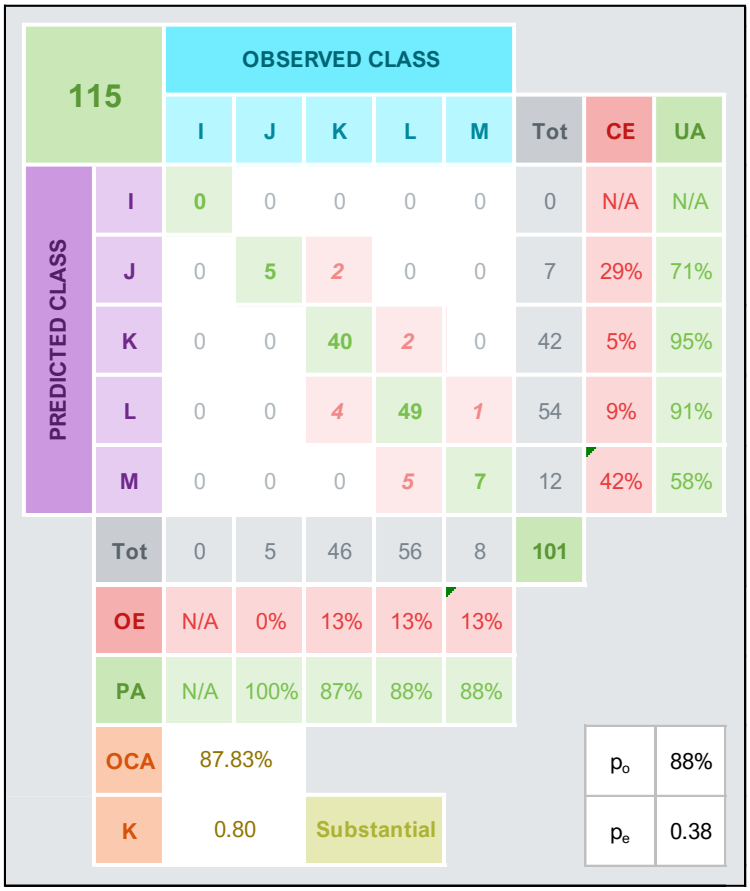


Figure 5-30: Taung-Prieska Belt water level classification confusion matrix

5.3.3 Yield predictions

The methodology established parameters which possibly influenced the yield, including pump-test parameters such as transmissivity, storage coefficient and specific capacity, recharge, mean annual precipitation, mean annual runoff, quaternary level baseflow, geology, and the number of water strikes present. No pump-test parameter data were available for the region and these had to be omitted from the model.

The results were unexpectedly good considering the lack of such critical parameters. The random-forest regression model predicted numerical yields with a Pearson correlation of 58.43% and an RMSE of 0.44. Figure 5-31 illustrates the yield predictions versus observed yields.

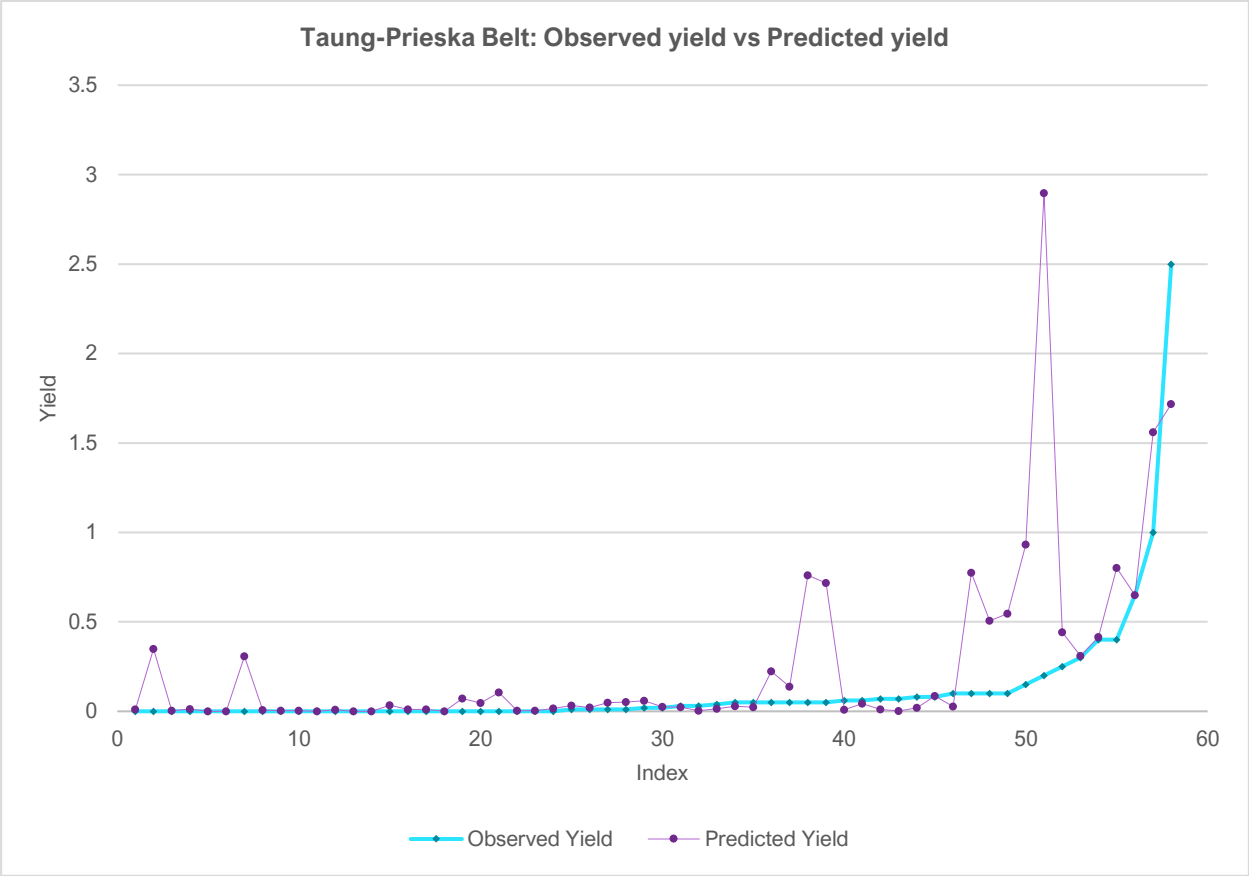


Figure 5-31: Taung-Prieska Belt predicted yield

The random-forest classification model predicted yield classes at an accuracy level of 77.61% and had a Kappa value of 0.57, which is considered moderate. The confusion matrix is displayed in Figure 5-32.

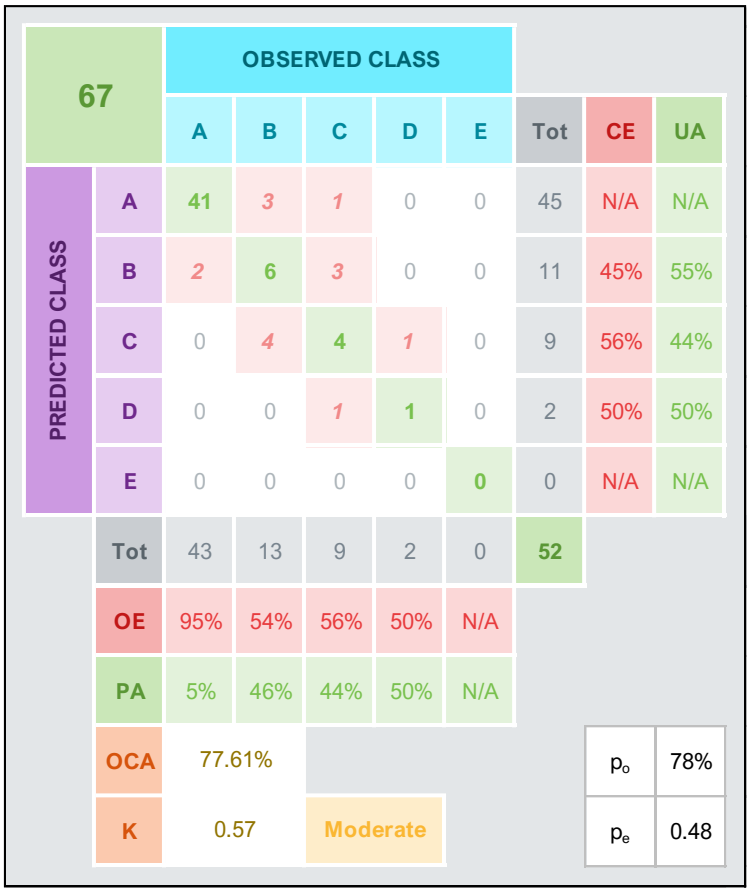


Figure 5-32: Taung-Prieska Belt yield classification confusion matrix

CHAPTER 6: RESULTS AND DISCUSSION

This chapter critically evaluates the results obtained in the three case studies conducted, the latter as discussed in the preceding chapter. The present chapter discusses whether the hypothesis has been proved.

6.1 Water level modelling

The methodology used here established that SVR was the best-performing algorithm to use for modelling static water levels whereas, for classification, the random-forest approach performed best. These were used to model the static water levels of three groundwater regions of which the results are presented in Table 6-1.

Table 6-1: Static water level model results obtained from case studies

Case Study Area	Regression		Classification	
	Pearson Correlation	RMSE	Classification Accuracy	Kappa Coefficient
Lowveld	99.69%	13.74	91.78%	0.90
Eastern Bushveld Complex	99.86%	13.96	90.48%	0.89
Taung-Prieska Belt	98.45%	13.47	87.83%	0.80

The accuracy metrics for the models in each study area all occur in approximately the same order, despite the fact that the Taung-Prieska Belt had far less data available than the others. This indicates that the parameters established during the methodology could be used as indicators in unexplored areas in order to characterise the water table. It was observed during the case studies that elevation is the primary driver behind water table occurrence. Geology also factors into the water level, which is a parameter that was established during the methodology. The other parameters are MAP, which is a driving force behind groundwater recharge, and the storativity coefficient.

Consider that water levels within groundwater databases are skewed. This engendered the high correlation with elevation, which holds true for unconfined aquifers, that is, those that are not confined by an impermeable layer. Uncased boreholes drilled in confined aquifers tend to mimic the behaviour of surrounding boreholes that are located in unconfined aquifers. Consequently, an equilibrium is reached in the system that skews the data. Ultimately, the data in the national

databases do not amount to a perfect representation of the system, because boreholes were not drilled and cased-off to ensure the integrity of the system and study the setting.

Another set of issues that arises from the national groundwater dataset, specifically the NGA, is known artefacts. Geographic positions captured within the NGA are not necessarily accurately taken in the field. Many of the older boreholes were captured before the advent of GPS, and the centroid of the farm portion was used as the geographic location. Therefore, any data from the spatial datasets do not necessarily represent the setting in which the borehole is situated.

Despite these issues, water levels are still modelled with high rates of accuracy. Nonetheless, these matters should be kept in mind in future.

6.2 Yield modelling

Modelling the yield and establishing primary drivers of this parameter proved to be more challenging than in the case of the water levels. The random-forest algorithm performed best with regard to numerical yield value modelling and classification of yield. The results are indicated in Table 6-2.

Table 6-2: Yield model results obtained from case studies

Case Study Area	Regression		Classification	
	Pearson Correlation	RMSE	Classification Accuracy	Kappa Coefficient
Lowveld	56.23%	2.93	57.36%	0.40
Eastern Bushveld Complex	72.09%	3.31	59.69%	0.44
Taung-Prieska Belt	58.43%	0.44	77.61%	0.57

The results of yield modelling are not as uniform as those of the static water levels. The Lowveld gave the most yield data, but does not have higher accuracy rates than those of the other two case studies. In fact, the Taung-Prieska Belt yield classification predicted fairly well for an area with exponentially less data compared to the other areas. Therefore, it might be assumed that, in contrast to static water level modelling, yield modelling does not suffer a 'one size fits all' model. The model should be adjusted to the regional setting. Although parameters have been established that could be important drivers, such as transmissivity, storage coefficient, specific capacity, geology, and recharge and baseflow, these cannot be considered as the primary drivers for any and all regions. Yield was limited by the fact that it is strongly dependent on geology, while the available lithological logs could not be used in the analysis, as described. Furthermore, borehole yields were averaged for the purpose of analysis, which could also have skewed the data.

Although isolated model results seem to model yield with a considerable degree of accuracy, the overall result suggests that further investigation is needed to establish the primary drivers behind yield.

CHAPTER 7: CONCLUSIONS AND RECOMMENDATIONS

The aim of this study was to mine the groundwater databases of South Africa and make use of data-driven modelling through machine learning, so as to classify relationships between borehole parameters such as static water level and yield and the surrounding settings. The model was compiled with the aim of utilising it for different regions across South Africa, and not to limit it to one area. Three case studies were conducted across South Africa to test the validity of the methodology proposed.

A single dataset was created by using three different groundwater information sources, namely the NGA, the GRIP, and geospatial data. This dataset was then used to model static water levels and borehole yield by using five different machine-learning algorithms. During the modelling phase, certain parameters were identified that could influence borehole parameters. The parameter with the most influence was elevation, which was to be expected due to the distinct correlation between groundwater level and topographic elevation. Second most important were mean annual precipitation (MAP), storativity (as gridded values from GIS datasets), and five sequential lithologies.

On evaluation of the model results and considering the relevant extant literature that has been reviewed here, the following conclusions are made:

- The foundations of a well-performing model are input data. The quality of the data will considerably impact the results of a model. Even though large quantities of data were available during this study, the quality of the data remains unknown.
- Borehole-specific geology could not be used in the modelling phase as it was too complex to gain a representative value which could be used by the algorithms. This was a critical dataset that could have influenced the results of the models.
- Not all algorithms are equally suited to a single task. Naive Bayes and K-Nearest neighbour classification do not manage high-dimension data well. This was established during the study, as these, especially naive Bayes, consistently underperformed.
- Although data-driven models are able to model the dependant variable of the system with a reasonable degree of accuracy, the drivers of the system are not immediately apparent. A trial-and-error approach must be taken in order to establish the drivers of a system, and even then, the interrelationships between the independent variables are not known.
- Static water levels could be modelled with significant accuracy, but borehole yield drivers are still uncertain and require further research.

The machine learning models created for this study, have been adjusted for the selected parameters and was designed with the aim of generalisation. Although the three case studies represent different groundwater regions and tested the applicability of the models for different scenarios, South Africa has many different and diverse regions. The true generalisation of the models is still to be tested in other regions of the country. For example, the most prominent parameter used in the model is elevation and has been proved to be a determining factor in the regions tested. Yet, the dolomitic aquifers of the North West Province is notorious for their complexity in simulating understood processes. Groundwater levels are very different in these dolomite or karst aquifers and do not necessarily follow the topography.

As per the literature review, there exists a hesitancy in the water resource sector to make use of data-driven models as opposed to the widely accepted process-based models. This hesitance may be in part due to the lack in competence of computer science skills. A similar challenge was faced during this study; therefore, it is recommended that members of the water resource community be exposed to these topics conventionally considered outside their field.

Data quality and quantity has been a central topic in this study. Section 3 discussed the available data for South African geohydrological databases and the quality of their data. It may be concluded that certain regions, such as the Limpopo Province, enjoys extensive and relatively good quality data, where regions in, for example, the Northern Cape, may not have publicly available data on which to conduct through studies.

On the basis of these, the following recommendations can be made:

- Data quality is more important than data quantity. Good quality data should be used to refine the models.
- The models created used the algorithms in their most basic form. Arguments for each algorithm can be tweaked to improve calibrate models.
- Consideration should be given as to how complex data such as lithological logs can be represented in the dataset to be usable for the model.

Finally, the hypothesis that applying data mining and machine-learning techniques on borehole data will improve a geohydrological characterisation of unexplored areas has tested true.

BIBLIOGRAPHY

Aggarwal, C.C. 2015. *Data mining: The textbook*. Cham: Springer International Publishing.

Akossou, A.Y.J. & Palm, R. 2013. Impact of data structure on the estimation R-square and adjusted R-square in linear regression. *International Journal of Mathematics and Computation* 20(3):84-93.

Alaliyat, S. 2008. *Video–based fall detection in elderly's houses*. Gjøvik: Gjøvik University College. (Thesis – MSc).

Allwright, A., Witthueser, K., Cobbing, J., Mallory, S. & Sawunyama, T. 2013. Development of a groundwater resource assessment methodology for South Africa: Towards a holistic approach. *Water Research Commission Report* No. 2048/1/13.

Arabameri, A., Roy, J., Saha, S., Blaschke, T., Ghorbanzadeh, O. & Tien Bui, D. 2019. Application of probabilistic and machine learning models for groundwater potentiality mapping in Damghan sedimentary plain, Iran. *Remote sensing* 11(24):3015-3049.

Aranibar, L.A.Q. 1994. *Learning fuzzy logic from examples*. Athens, OH: Ohio University. (Thesis – MSc).

Babovic, V. 2005. Data mining in hydrology. *Hydrological Processes* 19:1511-1515.

Bagaria, J. 2019. Set theory. In: *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/spr2020/entries/set-theory/ Date of access: 26 Aug. 2020.

Batini, C. & Scannapieco, M. 2016. Data and information quality – dimensions, principles and techniques. In: Carey, M.J. & Ceri, S., eds. *Data-centric systems and applications*. Cham: Springer International. pp. 5-7.

Bougher, B.B. 2009. *Machine learning applications to geophysical data analysis*. Vancouver: The University of British Columbia. (Thesis – MSc).

Bramer, M. 2016. *Principles of data mining*. 3rd ed. London: Springer.

Brownlee, J. 2021. *Regression metrics for machine learning.* https://machinelearningmastery.com/regression-metrics-for-machine-learning/ Date of access: 20 Feb. 2023.

Caté, A., Perozzi, L., Gloaguen, E. & Blouin, M. 2017. Machine learning as a tool for geologists. *The leading edge* 36(3):215-219.

Cichosz, P. 2015. *Data mining algorithms: Explained using R*. Hoboken, NJ: Wiley.

Dennis, R. & Dennis, I. 2020. Geo-statistical analysis and sub-delineation of all vegter regions. *North-West University*, K5/2745 Deliverable 6.

Devi, G.K., Ganasri, B.P. & Dwarakish, G.S. 2015. A review on hydrologicl models. *Aquatic Procedia* 4:1001-1007.

DFFE. 2022. *GIS data downloads* [Datasets]. DFFE. https://egis.environment.gov.za/data_egis/data_download/current

DFFE. 2021. South African national land-cover 2020 accuracy report. *Department of Forestry, Fisheries and Environment Public Release Report* version 1.0.4.

Diez, P. 2018. *Smart wheelchairs and brain-computer interfaces: Mobile assistive technologies*. Cambridge, MA: Academic Press.

DWA. 2009. Review of GRA1, GRA2 and international assessment methodologies. *Department of Water Affairs Report* No. P RSA 000/00/11609/6.

DWS (Department of Water and Sanitation). 2020 *National groundwater archive (NGA) stored borehole distribution*. https://www.dws.gov.za/Groundwater/data/boreholedist.aspx Date of access: 9 Nov. 2020.

Elefteriadou, L. 2014. Mathematical and empirical models. In: *An Introduction to traffic flow theory*. New York, NY: Springer. pp. 129-135.

Freeze, R.A. & Cherry, J.A. 1979. *Groundwater*. Englewood Cliffs, NJ: Prentice-Hall.

Gaaloul, N., Eslamian, S. & Ostad-Ali-Askari, K. 2018. Boreholes. In: Bobrowsky, P.T. & Marker, B., eds. *Encyclopedia of engineering geology*. Cham: Springer. pp. 68-73.

García, S., Luengo, J. & Herrera, F. 2015. *Data preprocessing in data mining*. Vol. 72. Cham, Switzerland: Springer International Publishing.

Gardner, S.A. 1992. Spelling errors in online databased: What the technical communicator should know. *Technica Communication* 39(1):50-53.

Gaur, P. 2012. Neural networks in data mining. *International Journal of Electronics and Computer Science Engineering* 1(3):1449-1453.

Goltz, M. & Huang, J. 2017. *Analytical modeling of solute transport in groundwater: Using models to understand the effect of natural processes on containment fate and transport*. Hoboken, NJ: Wiley.

GRIP (Groundwater Resource Information Project). s.a. *About the GRIP project.* http://griplimpopo.co.za/about/ Date of access: 11 Nov. 2020.

Grus, J. 2015. *Data science from scratch*. Sebastopol, CA: O'Rielly Media.

Hand, D.J. 2013. Data mining. In: El-Shaarawi, A.H. & Piegorsch, W.W., eds. *Encyclopedia of Environmetrics*. Hoboken, NJ: John Wiley & Sons, Ltd. pp. 1-4.

Hand, D.J., Mannila, H. & Smyth, P. 2001. *Principles of data mining*. Cambridge, MA: MIT Press.

Jing, H., He, X., Tian, Y., Lancia, M., Cao, G., Crivellari, A., … Zheng, C. 2022. Comparison and interpretation of data-driven models for simulating site-specific human-impacted groundwater dynamics in the North China plain. *Journal of Hydrology* 616:128751.

Joyce, J. 2019. Bayes' theorem. In: *Standford encyclopedia of philosophy*. https://plato.stanford.edu/archives/spr2019/entries/bayes-theorem/ Date of access: 3 Mar. 2020.

Kapitanova, K, Son, S.H., Kang, K.D. 2012. Using fuzzy logic for robust event detection in wireless sensor networks. *Ad Hoc Networks* 10(4):709-722.

Kenda, K., Čerin, M., Bogataj, M., Senožetnik, M., Klemen, K., Pergar, P., Laspidou, C. & Mladenić, D. 2018. Groundwater modeling with machine learning techniques: Ljubljana Polje aquifer. *Proceedings* 2(11):697-704.

Khan, A.N., Kim, B.W., Rizwan, A., Ahmad, R., Iqbal, N., Kim, K. & Kim, D.H. 2023. A new method for determination of optimal borehole drilling location considering drillist cost minimization and sustaible groundwater management. *ACS Omega* 2023(8):10806-10821.

Kim, J.H. & Jackson, R.B. 2012. A global analysis of groundwater recharge for vegetation, climate, and soils. *Vadose Zone Journal* 11(1):159-174.

Kohavi, R. & Becker, B. 1996. Dataset for: *Adult data set* [Dataset]. UCI machine learning repository. https://archive.ics.uci.edu/ml/datasets/adult

Kumar, C.P. 2019. An overview of commonly used groundwater modelling software. *International Journal of Advanced Research in Science, Engineering and Technology* 6(1):7854-7865.

Landis, J.R. & Koch, G.G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1):159-174.

Larose, D.T. 2005. *Discovering knowledge in data: An introduction to data mining*. Hoboken, NJ: Wiley.

Larose, C.D. & Larose, D.T. 2019. *Data science using python and R*. Hoboken, NJ: Wiley.

Lerner, D.N. & Harris, B. 2009. The relationship between land use and groundwater resources and quality. *Land Use Policy* 26(1):s265-s273.

Maliva, R.G. 2016. *Aquifer characterization techniques*. Cham: Springer.

Meinzer, O.E. 1934. The history and development of ground-water hydrology. *Journal of the Washington Academy of Sciences* 24(1):6-32.

Melville, P. & Sindhwani, V. 2017. Recommender systems. In: Sammut, C. & Webb, G.I., eds. *Encyclopedia of machine learning and data mining*. New York: Springer. pp. 1047.

Mijwel, M.M. 2018. *Artificial neural networks – advantages and disadvantages*. https://www.linkedin.com/pulse/artificial-neural-networks-advantages-disadvantages-maad-m-mijwel/ Date of access: 18 Mar. 2023.

Mitchell-Guthrie, P. 2014. *Looking backwards, looking forwards: SAS, data mining, and machine learning* [Blog post]. https://blogs.sas.com/content/subconsciousmusings/2014/08/22/looking-backwards-looking-forwards-sas-data-mining-and-machine-learning/ Date of access: 19 Oct. 2021.

Müller, B., Reinhardt, J. & Strickland, M.T. 1995. *Neural networks: An introduction*. 2nd ed. New York, NY: Springer.

Montgomery, D.C., Peck, E.A. & Vining, G.G. 2012. *Introduction to linear regression analysis*. 5th ed. Hoboken, NJ: Wiley.

Neelamegam, S. & Ramaraj, E. 2013. Classification algorithm in data mining: An overview. *International Journal of P2P Network Trends and Technology* 3(5):1-5.

Nell, J.P. & van Huyssteen, C.W. 2014. Geology and groundwater regions to quantify primary salinity, sodicity and alkalinity in South Africa. *South African Journal of Plant and Soil* 31(3):127-135.

NGA (National Groundwater Archive). s.a.(a) *Data disclaimer*. https://www.dws.gov.za/NGANet/Resources/Docs/Disclaimer.htm Date of access: 9 Nov. 2020.

NGA (National Groundwater Archive). s.a.(b) *Glossary*. https://www.dws.gov.za/NGANet/Resources/Docs/Glossary.htm Date of access: 9 Nov. 2020.

NGA (National Groundwater Archive). s.a.(c) *Site map*. https://www.dws.gov.za/NGANet/Resources/docs/SiteMap.htm Date of access: 9 Nov. 2020.

NGA (National Groundwater Archive). s.a.(d) *About us.* https://www.dws.gov.za/NGANet/Resources/Docs/About%20Us.htm Date of access: 9 Nov. 2020.

Noble, W.S. 2006. What is a support vector machine? *Nature Biotechnology* 24(12):1565-1567.

Oliveira, P., Rodrigues, F. & Henriques, P.R. 2005. A formal definition of data quality problems. In: *International Conference on Information Quality (MIT ICIQ Conference)*, Cambridge. https://dblp.org/rec/conf/iq/OliveiraRH05 Date of access: 31 Aug. 2020.

Oyebode, O.K., Adeyemo, J.A. & Otieno, F.A.O. 2015. Comparison of two data-driven modelling techniques for long-term streamflow prediction using limited datasets. *Journal of the South African Institution of Civil Engineering* 57(3):9-17.

Oyebode, O., Otieno, F. & Adeyemo, J. 2014. Review of three data-driven modelling techniques for hydrological modelling and forecasting. *Fresenius Environmental Bulletin* 23(7):1443-1454.

Pipino, L.L., Lee, Y.W. & Wang, R.Y. 2002. Data quality assessment. *Communications of the ACM* 45(4):211-218.

QGIS Development Team. 2021. QGIS Desktop (Version 3.20.3) [Software]. Available at: https://www.qgis.org/en/site/forusers/download.html

Rojas, R. 1996. *Neural networks: A systematic introduction*. New York, NY: Springer.

Rosli, M.M., Tempero, E. & Luxton-Reilly, A. 2016. What is in our datasets? Describing a structure of datasets. In: Gedeon, T., eds. *Conference proceedings*. ACSW '16: Australasian Computer Science Week Multiconference, Canberra, Australia. New York: Association for Computing Machinery. pp. 1-10.

Rosli, M.M., Tempero, E. & Luxton-Reilly, A. 2018. Evaluating the quality of datasets in software engineering. *Advanced Science Letters* 24(10):7232-7239.

RSA. 1998. National Water Act. In: Government Gazette. Pretoria, Republic of South Africa.

Russell, S.T. & Norvig, P. 2010. *Artificial intelligence: A modern approach*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Saha, S. 2018. *What is the c4.5 algorithm and how does it work?* https://towardsdatascience.com/what-is-the-c4-5-algorithm-and-how-does-it-work-2b971a9e7db0 Date of access: 18 Mar. 2023.

Sahoo, S., Russo, T.A., Elliott, J. & Foster, I. 2017. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. *Water Resource Research* 53:3878-3895.

Sammut, C. & Webb, G.I., *eds*. 2017. *Encyclopedia of machine learning and data mining*. Boston, MA: Springer.

Santner, T.J. & Duffy, D.E. 1989. *The statistical analysis of discrete data*. New York, NY: Springer Science+Business Media.

Shirmohammadi, B., Vafakhah, M., Moosavi, V. & Moghaddamnia, A. 2013. Application of several data-driven techniques for predicting groundwater level. *Water Resources Management* 27:419-432.

Sirsat, M. 2019. *What is confusion matrix and advanced classification metrics?* https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html Date of access: 1 Apr. 2020.

Sivakumar, B. & Berndtsson, R. 2010. *Advances in data-based approaches for hydrologic modeling and forecasting*. Singapore: World Scientific Publishing.

Solomatine, D.P. & Ostfeld, A. 2008. Data-driven modellling: Some past experiences and new approaches. *Journal of Hydroinformatics* 10(1):3-22.

Sun, J., Hu, L., Li, D., Sun, K. & Yang, Z. 2022. Data-driven models for accurate groundwater level prediction and their practical significance in groundwater management. *Journal of Hydrology* 608:127630.

SuperDataScience. 2020. Machine learning A-Z: Download codes and datasets [scripts]. https://www.superdatascience.com/pages/machine-learning

Taylor, C.J. & Alley, W.M. 2001. Ground-water-level monitoring and the importance of long-term water-level data. *U.S. Geological Survey Circular* No. 1217.

Taylor, R.G., Koussis, A.T. & Tindimugaya, C. 2009. Groundwater and climate in Africa – a review. *Hydrological Sciences Journal* 54(4):655-664.

Tehrany, M.S., Pradhan, B. & Jebur M.N. 2013. Spatial prediction of flood susceptible areas using rule based decision tree (DT) and novel ensemble bivariate and multivariate statistical models in GIS. *Journal of Hydrology* 504:69-79.

Teng, X. & Gong, Y. 2018. Research on application of machine learning in data mining. In *IOP conference series: Materials science and engineering*, 392(6). Bristol: IOP Publishing.

Provost, A.M., Reilly, T.E., Harbaugh, A.W. & Pollock, D.W. 2009. U.S. geological survey groundwater modeling software: Making sense of a complex natural resource. *U.S. Geological Fact Sheet* No. 2009-3105.

Villholth, K.G. & Giordano, M. 2007. Groundwater use in a global perspective – can it be managed? In: Villholth, K.G. & Giordano, M., eds. *The agricultural groundwater revolution: Opportunities and threats to development*. Oxfordshire: Wallingford. pp. 393-401.

WRC (Water Research Commission). 2012. GIS Maps [Geospatial dataset]. Water resources of South Africa, 2012 study. https://www.waterresourceswr2012.co.za/resource-centre/

Wheater, H., Sorooshian, S. & Sharma, K.D., eds. 2007. *Hydrological modelling in arid and semi-arid areas*. New York, NY: Cambridge University Press.

Wu, S. 2020. *3 Best metrics to evaluate regression model?* https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b Date of access: 20 Feb. 2023.

Xu, Y. & Beekman, H.E. 2019. Review: Groundwater recharge estimation in arid and semi-arid Southern Africa. *Hydrogeology Journal* 27:929-943.

Yan, X. & Su, X.G. 2009. *Linear regression analysis: Theory and computing*. Singapore: World Scientific.

Zadeh, L.A. 1988. Fuzzy Logic. *Computer*, 21(4):83-93.

Zaki, M.J. & Meira Jr, W. 2014. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge, UK: Cambridge University Press.

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., … Ye, L. 2022. A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health* 1(2):107-116.

ANNEXURES

8.1 Annexure A – NGA database

Table 8-1: NGA available features for export

Geosite Information	Construction Completion	Water Levels	Abstraction	Discharge Rate	Field Measurements	Reference	Other Numbers	Depth & Diameter	Casing	Openings & Screens	Fill Materials	Developments	Piezometer	Equipment Installed	Geosites Linked to Bulk Meter	Water Strike	Downhole Geophysics	Pumping Test Details	Yield Test	Operational Recommendations	Site Visits	Owners	Lithology
Data Owner	Construction Completion Date	Reference Point	Meter Type	Measurement Date & Time	Piezometer Number	Reference Type	Other Number Type	Measurement Date	Casing Column Number	Casing Column Number	Casing Column Number	Development Methods	Piezometer Number	Monitoring Type	Data Owner Of Linked Geosite	Construction Completion Date	Logging Date	Pumping Test Start Date	~~Yield Test Start Date~~	Pumping Test Start Date	Visit Date	Owner Name	Descriptor Designator
Identifier	Starting Construction Completion Date	Reference Height	Measurement Date & Time	Starting Measurement Date & Time	Measurement Date & Time	Library Report Number	Other Number	Data Source	Casing Collar Height	Depth To Top	Depth To Top	Chemical Types Used	Piezometer Purpose	Installation Date	Identifier Of Linked Geosite	Yield Measurement Method	Logging Unit	Pumping Test Start Time	~~Yield Test Start Time~~	Pumping Test Start Time	Visit Reason	Address	Descriptor Name
Geosite Type	Ending Construction Completion Date	Measuring Method	Starting Measurement Date & Time	Ending Measurement Date & Time	Starting Measurement Date & Time	Report Name	Reporting Institution	Reporting Institution	Observed Casing	Depth To Bottom	Depth To Bottom	Development Date	Depth To Top	Decommissioned Date	Start Date	Groundwater Occurrence	Logging Company	Pumping Test End Date	Test Type	Recommendation Date	Reporting Institution	Contact Details	Event Date
Mine Type	Construction Cost	Water Level Status	Ending Measurement Date & Time	Discharge Rate	Ending Measurement Date & Time	Report Date	Assignor	Depth To Bottom	Depth To Top	Number Of Openings	Fill Type	Duration Of Development	Depth To Bottom	Data Source	End Date	Water Strike Type	Logging Contractor	Pumping Test End Time	Static Water Level	Source Of Recommendation	Reporting Institution Address	Reporting Institution	Depth To Top
Confidential	Construction Method	Drawdown Period	Data Source	Discharge Type	Measurement Depth	Consultant's Report Number		Penetration Information Available	Depth To Bottom	Opening Method	Gravel Pack	Finish Type	Piezometer Height	Reporting Institution		Seepage Value	Logging Company Address	Test Method	Constant Yield Test Total Duration Hours	Purpose Indicator	Reporting Institution Contact Number	Visit Date	Depth To Bottom
Reference Datum	Construction Company	Recovery Period	Reporting Institution	Discharge Method	Electrical Conductivity	Located At		Depth Qualifier	Casing Material	Opening Width			Depth From Casing / Lining Collar	Pump Type		Depth To Top	Logging Company Contact Number	Pump Type	Constant Yield Test Total Duration Minutes	Protection Zone Up Gradient	Site Visitor		Lithology Name
Coordinate Method	Construction Contractor	Piezometer Number	Hour Meter Reading	Data Source	Ph Class	Geology Indicator		Diameter	Other Casing Material	Opening Length			Piezometer Length	Depth To Pump Intake		Depth To Bottom	Paper Trace Reference Number	Depth To Pump Intake	Analysis Method	Protection Zone Down Gradient	Site Visitor Address		Primary Colour
Coordinate GPS Accuracy	Construction Company Contact Details	Measurement Date & Time	Conversion Factor / Constant Value	Reporting Institution	Ph Value			Opencast Mine Length	Inner Diameter	Opening Diameter			Installed Date	Pump Power Source		Total Blow Yield Value	Data Cassette / Video Number	Reporting Institution	Other Analysis Methods	Water Quality Class	Site Visitor Contact Number		Colour Qualifier
Elevation	Construction Company Contact Number	Starting Measurement Date & Time	Quantity		Temperature			Opencast Mine Width	Outer Diameter	Screen Manufacturer			Decommissioned Date	Pump Power Rating		Contribution Value	Measurement Type	Testing Company	Specific Capacity	Chemical Suitability			Secondary Colour
Elevation Method	Drilling Fluid	Ending Measurement Date & Time	Power Meter Reading		HCO3			Opencast Mine Depth	Deepest Interval Closed	Screen Specification			Inner Diameter	Pump Manufacturer		Seepage Indicator For Contribution	Special Measurements	Testing Contractor	Transmissivity	Pump Type			Composition Qualifier
Elevation GPS Accuracy	Additives	Water Level	Water Meter Collective Reading(Measurement)		Data Source			Underground Mine Shaft Depth	Diameter				Outer Diameter	Pump Serial Number				Testing Company Address	Storativity	Depth To Pump Intake			Fabric Qualifier
Elevation Reference Point	Additional Additives	Data Source	Water Meter Reading		Reporting Institution			Tunnel Shape	Lining Collar Height				Piezometer Material	Pump Riser Main Material				Testing Company Contact Number	Specific Yield	Duty Cycle			Fabric Attribute
Farm Name	Other Construction Method	Reporting Institution	Water Meter Measurement Reason					Tunnel Cross-Section	Lining Material				Other Piezometer Material	Pump Riser Diameter				Analysis Method	Leakage / Drainage Factor	Recommended Abstraction Yield			Material Type
Farm Number	Protection Method	Latest Water Level Measurement Only	Measurement Method					Tunnel Length	Other Lining Material				Reporting Institution	Meter Type				Other Analysis Methods	Hydraulic Resistance	Recovery Period			Formation Type
Town	Other Protection Method	Water Levels (with or without pumping tests)	Measurement Status						Drain Casing Length					Meter Serial Number				Specific Capacity	Hydraulic Conductivity	Operational Period			Loss Percentage
Portion			Abstraction (with or without pumping tests)						Collector Diameter					Supplying Company				Transmissivity	Water Quality Class	Piezometer Number			Loss Reason
Village									Arm Diameter					Supplying Contractor				Storativity	Step Number	Water Level Drawdown			Other Loss Reason
Geosite Status									Arm Number					Supplying Company Address				Specific Yield	Step Duration	Water Level Measurement Frequency			Texture Qualifier
Date When Status Was Observed									Arm Length					Supplying Company Contact Number				Leakage / Drainage Factor	Average Discharge Rate	Monitoring Period			Hardness Qualifier
Starting Geosite Status Date														Bulk Meter Indicator				Hydraulic Resistance	Measurement Date & Time Type				Particle Shape

Ending Geosite Status Date														Water Meter Maximum Measurement Value				Hydraulic Conductivity	~~Actual Measurement Date~~				Sorting
Reporting Institution														Rainfall Monitoring Equipment Type				Water Quality Class	~~Actual Measurement Time~~				Boulder Percentage
Map Number														Rainfall Other Equipment Type					Elapsed Time				Cobble Percentage
Province														Electronic Data Logger					~~Water Level Type~~				Pebble Percentage
Registration District														Other Electronic Data Logger Type					Water Level				Granule Percentage
Quaternary Drainage Region														Electronic Data Logger Manufacturer					Water Level Status				Sand Percentage
Site Selector														Other Electronic Data Logger Manufacturer					Discharge Rate				Silt Percentage
Water Consumer														Electronic Data Logger Serial Number					Quantity				Clay Percentage
Comment														Measurement Method									Degree Of Weathering
Latitude														Other Measurement Method									Degree Of Fracturing
Longitude														Water Level Monitoring Equipment Type									
Water Management Area																							
Municipal District Old																							
Municipal District New																							
Surface Geology																							
Hydrogeological Region																							
Water User Association																							
Geomorphology																							
Land Cover																							
Taste Of Water																							
Intended Geosite Purpose																							
Observed/Actual Water Uses																							
Drain Type																							
Mine Groundwater Extraction Method																							
Mine Status																							
Mined Commodity																							
Sinkhole Type																							
Sinkhole Cause																							
Sinkhole Probable Trigger																							
Sinkhole Classification																							
Spring Seasonality Regime																							
Spring Geomorphological/Geological Type																							
Spring Classification Through Discharge Rate																							
Tunnel Purpose																							
Tunnel Usage																							

Table 8-2: Column completeness results for the NGA

Attribute	Column Completeness
Identifier	1
Geosite Type	1
Latitude	1
Longitude	1
Data Owner	1
Borehole Depth	0.6851
Borehole Diameter	0.6851
Lithology Name	0.3839
Water Strike Type	0.3528
Depth to Top (Water Strike)	0.3522
Water Level	0.3241
Discharge Rate	0.2174
Seepage Value (Water Strike)	0.2058
Weathering Degree (Lithology)	0.1184
Electrical Conductivity	0.0467
Depth To Bottom (Water Strike)	0.0220
Fracturing Degree (Lithology)	0.0193
Abstraction Quantity	0.0183
pH Value	0.0126
pH Class	0.0105
Temperature	0.0095
Groundwater Occurrence	0.0057
HCO_3	0.0001

Table 8-3: Column completeness results for the GRIP

Attribute	Column Completeness
GRIP site ID number	1
GRIP borehole number	1
H Area	1
Quaternary catchment area	1
Farm name	1
Farm number	1
Province	1
District municipality	1
Local municipality	1
Settlement name	1
Settlement ID	1
Longitude	1
Latitude	1
Power	0.9997
Comment	0.9995
Equipment	0.9977
Borehole depth	0.5239
Water level	0.4056
Water level date taken	0.4056
Depth to pump intake	0.3274
Discharge rate	0.3273
Duty cycle	0.3273
Daily abstractions	0.3273
Quality	0.3164
Alternative borehole number 2	0.1376
Regional borehole number	0.0554
Alternative borehole number 1	0.0479
Alternative settlement name	0.0104

8.2 Annexure B – GRIP database example

GRIP Site ID number	GRIP borehole Number	H Area	Quaternary Catchment area	Regional borehole number	Alternative Borehole Number 1	Alternative Borehole Number 2	Farm Name	Farm Number	Province	District Municipality	Local Municipality	Settlement Name	Settlement ID
H01-2429BDV0003	H01-2125	H01	B52G	-	G45030	-	KOPPIEKRAAL	LPKS475	Limpopo	Capricorn	Lepele-Nkumpi	Serobaneng	1851
H01-2429BCV0014	H01-2318	H01	B52G	-	-	-	MOLSGAT	LPKS439	Limpopo	Capricorn	Lepele-Nkumpi	Masite	1147
H02-2429DDV0015	H02-0907	H02	B52B	-	-	-	GROBLERSVREDE	LPKS844	Limpopo	“Greater Sekhukhune”	Makhuduthamaga	Ga-Ratau	438
H02-2429DBN0060	H02-1150	H02	B52B	-	-	-	“GELUKS LOCATION”	LPKS000	Limpopo	“Greater Sekhukhune”	Makhuduthamaga	Manganeng	1022

Alternative Settlement name	Longitude [WGS84]	Latitude [WGS84]	Borehole depth [m]	Waterlevel [mbgl]	Water level date taken	Depth to pump intake [m]	Discharge rate [l/s]	Duty cycle [hours]	Daily Abstraction [m3/day]	Equipment	Power	Quality	Comment
-	29.60052	-24.23661	125	17.95	2001-10-02	66	3	24	259.2	“No equipment”	“No power”	-	TESTED
-	29.67855	-24.31084	37	12.62	2003-02-01	24	0.7	24	60.48	“Submersible pump”	“Electric motor”	“CLASS 0”	TESTED
-	29.95428	-24.78826	64.2	9.51	2005-06-23	24	0.5	24	43.2	“Momo-type pump”	Diesel	“CLASS 2”	TESTED
-	29.97279	-24.6857	47.72	19.66	2006-07-04	36	0.5	24	43.2	Handpump	Hand	“CLASS 3”	TESTED

8.3 Annexure C – Model Scripts

Scripts presented are those used for water level prediction. Those of yield prediction are exactly the same, only the input dataset and dependent variable differ.

8.3.1 Regression

```
# Multiple Linear Regression

# Importing the dataset
dataset = read.csv('Method_WL_mamsl.csv')
str(dataset)

# Removing columns from imported dataset
dataset = subset(dataset, select = -c(WL_100m))
# dataset = subset(dataset, select = -c(Elevation))
str(dataset)

# Numerical to factor
dataset$LITHO_1 = factor(dataset$LITHO_1)
dataset$LITHO_2 = factor(dataset$LITHO_2)
dataset$LITHO_3 = factor(dataset$LITHO_3)
dataset$LITHO_4 = factor(dataset$LITHO_4)
dataset$LITHO_5 = factor(dataset$LITHO_5)
str(dataset)

# Splitting the dataset into the Training set and Test set
# install.packages('caTools')
library(caTools)
set.seed(1234)
split = sample.split(dataset$WL_mamsl, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)

# Fitting Multiple Linear Regression to the Training set
regressor = lm(formula = WL_mamsl ~ .,
               data = training_set)

# Predicting the Test set results
y_pred = predict(regressor, newdata = test_set)

# Plot results
plot(test_set$WL_mamsl, type = "l", col = "red")
lines(y_pred, col = "grey")
lines(moving_average, col = "blue")

# Correlation
Pearson = cor(test_set$WL_mamsl, y_pred, method = c("pearson"))
RMSE = sqrt(mean((test_set$WL_mamsl - y_pred)^2))
MAPE = MAPE(y_pred,test_set$WL_mamsl)
MAE = MAE(test_set$WL_mamsl,y_pred)

print(Pearson*100)
print(RMSE)
print(MAPE)
print(MAE)

# Significance value of features
summary(regressor)
```

```
# Support Vector Regression

# Importing the dataset
dataset = read.csv('Method_WL_mamsl.csv')
str(dataset)

# Removing columns from imported dataset
dataset = subset(dataset, select = -c(WL_100m))
# dataset = subset(dataset, select = -c(Elevation))
str(dataset)

# Numerical to factor
dataset$LITHO_1 = factor(dataset$LITHO_1)
dataset$LITHO_2 = factor(dataset$LITHO_2)
dataset$LITHO_3 = factor(dataset$LITHO_3)
dataset$LITHO_4 = factor(dataset$LITHO_4)
dataset$LITHO_5 = factor(dataset$LITHO_5)
str(dataset)


# Splitting the dataset into the Training set and Test set
library(caTools)
set.seed(123)
split = sample.split(dataset$WL_mamsl, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)

# Fitting Support Vector Regression to the Training set
library(e1071)
regressor_L_nu = svm(formula = WL_mamsl ~ .,
                 data = training_set,
                 type = 'nu-regression',
                 kernel = 'linear')
y_pred_L_nu = predict(regressor_L_nu, newdata = test_set)

# Plot results
plot(test_set$WL_mamsl, type = "l", col = "red")
lines(y_pred_L_nu, col = "grey")
lines(moving_average, col = "blue")

# Correlation
Pearson_L_nu = cor(test_set$WL_mamsl, y_pred_L_nu, method = c("pearson"))
RMSE_L_nu = sqrt(mean((test_set$WL_mamsl - y_pred_L_nu)^2))
MAPE = MAPE(y_pred_L_nu,test_set$WL_mamsl)
MAE = MAE(test_set$WL_mamsl,y_pred_L_nu)

print(Pearson_L_nu*100)
print(RMSE_L_nu)
print(MAPE*100)
print(MAE)

# Significance value of features
summary(regressor_L_nu)
```

```
# Decision Tree Regression

# Importing the dataset
dataset = read.csv('Method_WL_mamsl.csv')
str(dataset)

# Removing columns from imported dataset
dataset = subset(dataset, select = -c(WL_100m))
dataset = subset(dataset, select = -c(Elevation))
str(dataset)

# Numerical to factor
dataset$LITHO_1 = factor(dataset$LITHO_1)
dataset$LITHO_2 = factor(dataset$LITHO_2)
dataset$LITHO_3 = factor(dataset$LITHO_3)
dataset$LITHO_4 = factor(dataset$LITHO_4)
dataset$LITHO_5 = factor(dataset$LITHO_5)
str(dataset)

# Splitting the dataset into the Training set and Test set
# install.packages('caTools')
library(caTools)
set.seed(123)
split = sample.split(dataset$WL_mamsl, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)

# Fitting Decision Tree Regression to the Training set
# install.packages('rpart')
library(rpart)
regressor = rpart(formula = WL_mamsl ~ .,
                  data =    training_set,
                  control = rpart.control(minsplit = 2))

# Predicting a new result with Decision Tree Regression
y_pred = predict(regressor, newdata = test_set)

# Plot results
plot(test_set$WL_mamsl, type = "l", col = "red")
lines(y_pred, col = "grey")
lines(moving_average, col = "blue")

# Correlation
Pearson = cor(test_set$WL_mamsl, y_pred, method = c("pearson"))
RMSE = sqrt(mean((test_set$WL_mamsl - y_pred)^2))
MAPE = MAPE(y_pred,test_set$WL_mamsl)
MAE = MAE(test_set$WL_mamsl,y_pred)

print(Pearson*100)
print(RMSE)
print(MAPE*100)
print(MAE)

# Significance value of features
summary(regressor)
```

```
# Random Forest Regression

# Importing the dataset
dataset = read.csv('Method_WL_mamsl.csv')
str(dataset)

# Removing columns from imported dataset
dataset = subset(dataset, select = -c(WL_100m))
dataset = subset(dataset, select = -c(Elevation))
str(dataset)

# Numerical to factor
dataset$LITHO_1 = factor(dataset$LITHO_1)
dataset$LITHO_2 = factor(dataset$LITHO_2)
dataset$LITHO_3 = factor(dataset$LITHO_3)
dataset$LITHO_4 = factor(dataset$LITHO_4)
dataset$LITHO_5 = factor(dataset$LITHO_5)
str(dataset)

# Splitting the dataset into the Training set and Test set
# install.packages('caTools')
library(caTools)
set.seed(123)
split = sample.split(dataset$WL_mamsl, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)

# Fitting Random Forest Regression to the Training set
# install.packages('randomForest')
library(randomForest)
set.seed(123)
regressor = randomForest(formula = WL_mamsl ~ .,
                         data = training_set,
                         ntree = 100,
                         keep.forest = TRUE,
                         importance = TRUE)

# Predicting a new result with Random Forest Regression
y_pred = predict(regressor, newdata = test_set)


# Get variable importance from the model fit
print(regressor)
importance(regressor)
varImpPlot(regressor)

# Plot results
plot(test_set$WL_mamsl, type = "l", col = "red")
lines(y_pred, col = "grey")
lines(moving_average, col = "blue")

# Correlation
Pearson = cor(test_set$WL_mamsl, y_pred, method = c("pearson"))
RMSE = sqrt(mean((test_set$WL_mamsl - y_pred)^2))
MAPE = MAPE(y_pred,test_set$WL_mamsl)
MAE = MAE(test_set$WL_mamsl,y_pred)

print(Pearson*100)
print(RMSE)
print(MAPE*100)
print(MAE)
```

8.3.2 Classification

```
# K-Nearest Neighbors (K-NN)

# Importing the dataset
dataset = read.csv('Method_WL_mamsl.csv')
str(dataset)

# Removing columns from imported dataset
dataset = subset(dataset, select = -c(WL_mamsl))
# dataset = subset(dataset, select = -c(Elevation))
str(dataset)

# Numerical to factor
dataset$LITHO_1 = factor(dataset$LITHO_1)
dataset$LITHO_2 = factor(dataset$LITHO_2)
dataset$LITHO_3 = factor(dataset$LITHO_3)
dataset$LITHO_4 = factor(dataset$LITHO_4)
dataset$LITHO_5 = factor(dataset$LITHO_5)
dataset$WL_100m = factor(dataset$WL_100m)
str(dataset)

# Splitting the dataset into the Training set and Test set
# install.packages('caTools')
library(caTools)
set.seed(123)
split = sample.split(dataset$WL_100m, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)

# Fitting K-NN to the Training set and Predicting the Test set results
library(class)
y_pred = knn(train = training_set[, -9],
             test = test_set[, -9],
             cl = training_set[, 9],
             k = 1)

# Making the Confusion Matrix
cm = table(test_set[, 9], y_pred)
print(cm)
Misclassification = 1 - sum(diag(cm))/sum(cm)
Correctly_Classified = 1 - Misclassification
print(Misclassification*100)
print(Correctly_Classified*100)

# Calculating Kappa value
diagonal.counts = diag(cm)
N = sum(cm)
row.marginal.props = rowSums(cm)/N
col.marginal.props = colSums(cm)/N
Po = sum(diagonal.counts)/N
Pe = sum(row.marginal.props*col.marginal.props)
k = (Po - Pe)/(1 - Pe)
k
```

```
# Support Vector Machine (SVM)

# Importing the dataset
dataset = read.csv('Method_WL_mamsl.csv')
str(dataset)

# Removing columns from imported dataset
dataset = subset(dataset, select = -c(WL_mamsl))
dataset = subset(dataset, select = -c(Elevation))
str(dataset)

# Numerical to factor
dataset$LITHO_1 = factor(dataset$LITHO_1)
dataset$LITHO_2 = factor(dataset$LITHO_2)
dataset$LITHO_3 = factor(dataset$LITHO_3)
dataset$LITHO_4 = factor(dataset$LITHO_4)
dataset$LITHO_5 = factor(dataset$LITHO_5)
dataset$WL_100m = factor(dataset$WL_100m)
str(dataset)

# Splitting the dataset into the Training set and Test set
# install.packages('caTools')
library(caTools)
set.seed(123)
split = sample.split(dataset$WL_100m, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)

# Fitting SVM to the Training set
# install.packages('e1071')
library(e1071)
classifier = svm(formula = WL_100m ~ .,
                 data = training_set,
                 type = 'C-classification',
                 kernel = 'linear')

# Predicting the Test set results
y_pred = predict(classifier, newdata = test_set[-8])

# Making the Confusion Matrix
cm = table(test_set[, 8], y_pred)
print(cm)
Misclassification = 1 - sum(diag(cm))/sum(cm)
Correctly_Classified = 1 - Misclassification
print(Misclassification*100)
print(Correctly_Classified*100)

# Calculating Kappa value
diagonal.counts = diag(cm)
N = sum(cm)
row.marginal.props = rowSums(cm)/N
col.marginal.props = colSums(cm)/N
Po = sum(diagonal.counts)/N
Pe = sum(row.marginal.props*col.marginal.props)
k = (Po - Pe)/(1 - Pe)
k
```

```
# Naive Bayes

# Importing the dataset
dataset = read.csv('Method_WL_mamsl.csv')
str(dataset)

# Removing columns from imported dataset
dataset = subset(dataset, select = -c(WL_mamsl))
dataset = subset(dataset, select = -c(Elevation))
str(dataset)

# Numerical to factor
dataset$LITHO_1 = factor(dataset$LITHO_1)
dataset$LITHO_2 = factor(dataset$LITHO_2)
dataset$LITHO_3 = factor(dataset$LITHO_3)
dataset$LITHO_4 = factor(dataset$LITHO_4)
dataset$LITHO_5 = factor(dataset$LITHO_5)
dataset$WL_100m = factor(dataset$WL_100m)
str(dataset)

# Splitting the dataset into the Training set and Test set
# install.packages('caTools')
library(caTools)
set.seed(123)
split = sample.split(dataset$WL_100m, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)

# Fitting naiveBayes to the Training set
# install.packages('e1071')
library(e1071)
classifier = naiveBayes(x = training_set[, -8],
                        y = training_set$WL_100m)

# Predicting the Test set results
y_pred = predict(classifier, newdata = test_set[-8])

# Making the Confusion Matrix
cm = table(test_set[, 8], y_pred)
print(cm)
Misclassification = 1 - sum(diag(cm))/sum(cm)
Correctly_Classified = 1 - Misclassification
print(Misclassification*100)
print(Correctly_Classified*100)

# Calculating Kappa value
diagonal.counts = diag(cm)
N = sum(cm)
row.marginal.props = rowSums(cm)/N
col.marginal.props = colSums(cm)/N
Po = sum(diagonal.counts)/N
Pe = sum(row.marginal.props*col.marginal.props)
k = (Po - Pe)/(1 - Pe)
k
```

```
# Decision Tree Classification

# Importing the dataset
dataset = read.csv('Method_WL_mamsl.csv')
str(dataset)

# Removing columns from imported dataset
dataset = subset(dataset, select = -c(WL_mamsl))
dataset = subset(dataset, select = -c(Elevation))
str(dataset)

# Numerical to factor
dataset$LITHO_1 = factor(dataset$LITHO_1)
dataset$LITHO_2 = factor(dataset$LITHO_2)
dataset$LITHO_3 = factor(dataset$LITHO_3)
dataset$LITHO_4 = factor(dataset$LITHO_4)
dataset$LITHO_5 = factor(dataset$LITHO_5)
dataset$WL_100m = factor(dataset$WL_100m)
str(dataset)

# Splitting the dataset into the Training set and Test set
# install.packages('caTools')
library(caTools)
set.seed(123)
split = sample.split(dataset$WL_100m, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)

# Fitting DT to the Training set
library(RWeka)
classifier = J48(formula = WL_100m ~ .,
                 data = training_set)
summary(classifier)

# Predicting the Test set results
y_pred = predict(classifier, newdata = test_set[-8])

# Making the Confusion Matrix
cm = table(test_set[, 8], y_pred)
print(cm)
summary(cm)
Misclassification = 1 - sum(diag(cm))/sum(cm)
Correctly_Classified = 1 - Misclassification
print(Misclassification*100)
print(Correctly_Classified*100)

# Calculating Kappa value
diagonal.counts = diag(cm)
N = sum(cm)
row.marginal.props = rowSums(cm)/N
col.marginal.props = colSums(cm)/N
Po = sum(diagonal.counts)/N
Pe = sum(row.marginal.props*col.marginal.props)
k = (Po - Pe)/(1 - Pe)
k
```

```
# Random Forest Classification

# Importing the dataset
dataset = read.csv('Method_WL_mamsl.csv')
str(dataset)

# Removing columns from imported dataset
dataset = subset(dataset, select = -c(WL_mamsl))
# dataset = subset(dataset, select = -c(Elevation))
str(dataset)

# Numerical to factor
dataset$LITHO_1 = factor(dataset$LITHO_1)
dataset$LITHO_2 = factor(dataset$LITHO_2)
dataset$LITHO_3 = factor(dataset$LITHO_3)
dataset$LITHO_4 = factor(dataset$LITHO_4)
dataset$LITHO_5 = factor(dataset$LITHO_5)
dataset$WL_100m = factor(dataset$WL_100m)
str(dataset)

# Splitting the dataset into the Training set and Test set
# install.packages('caTools')
library(caTools)
set.seed(123) #12345
split = sample.split(dataset$WL_100m, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)

# Fitting Random Forest Classification to the Training set
# install.packages('randomForest')
library(randomForest)
classifier = randomForest(x = training_set[, -9], y = training_set$WL_100m, ntree = 100, keep.forest = TRUE,
importance = TRUE, nodesize = 5)

# Predicting the Test set results
y_pred = predict(classifier, newdata = test_set[-9])

# Making the Confusion Matrix
cm = table(test_set[, 9], y_pred)
print(cm)
Misclassification = 1 - sum(diag(cm))/sum(cm)
Correctly_Classified = 1 - Misclassification
print(Misclassification*100)
print(Correctly_Classified*100)

# Calculating Kappa value
diagonal.counts = diag(cm)
N = sum(cm)
row.marginal.props = rowSums(cm)/N
col.marginal.props = colSums(cm)/N
Po = sum(diagonal.counts)/N
Pe = sum(row.marginal.props*col.marginal.props)
k = (Po - Pe)/(1 - Pe)
k
```

8.4 Annexure D – Maps

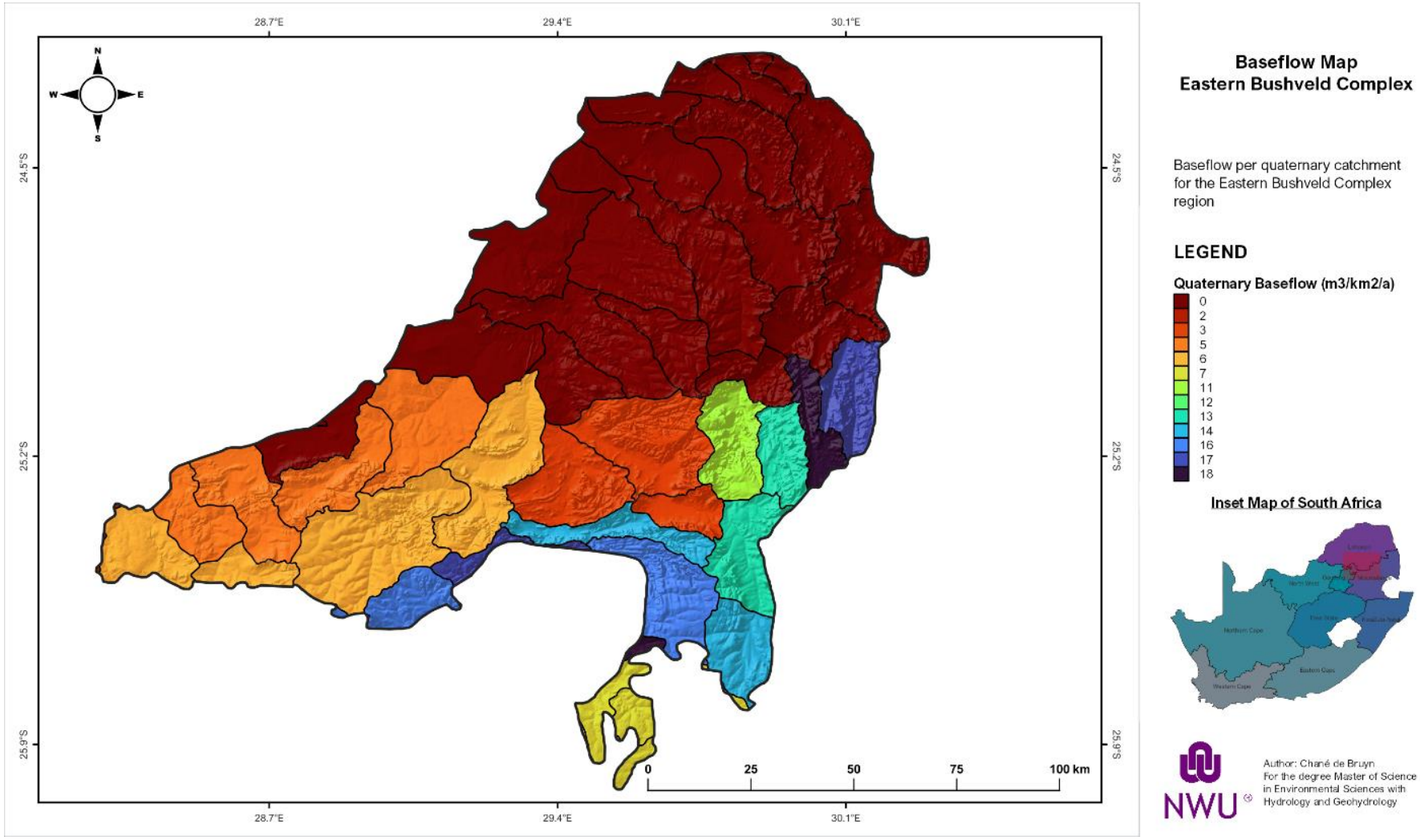


Figure 8-1: Eastern Bushveld Complex – Baseflow

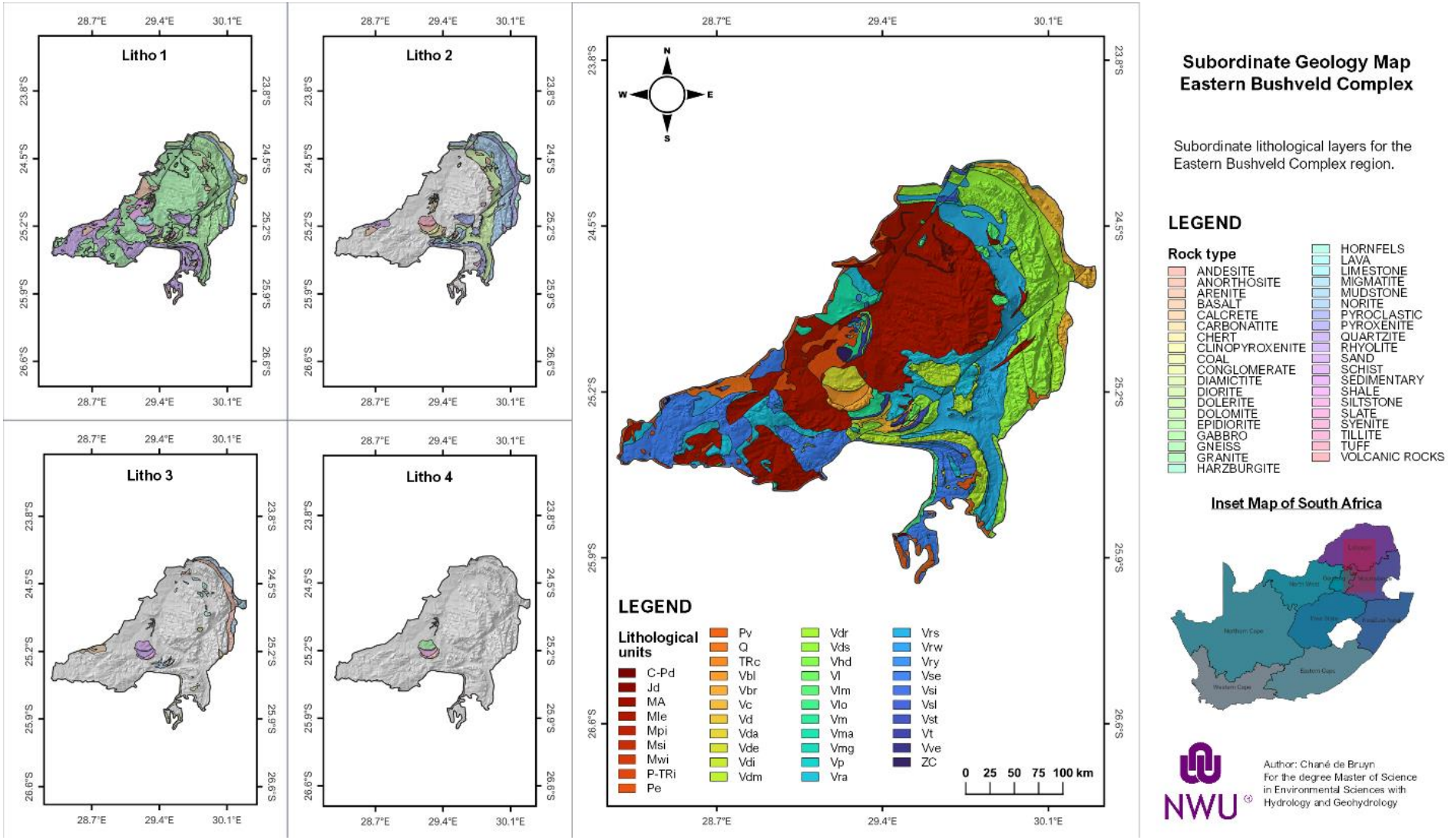


Figure 8-2: Eastern Bushveld Complex - Lithology

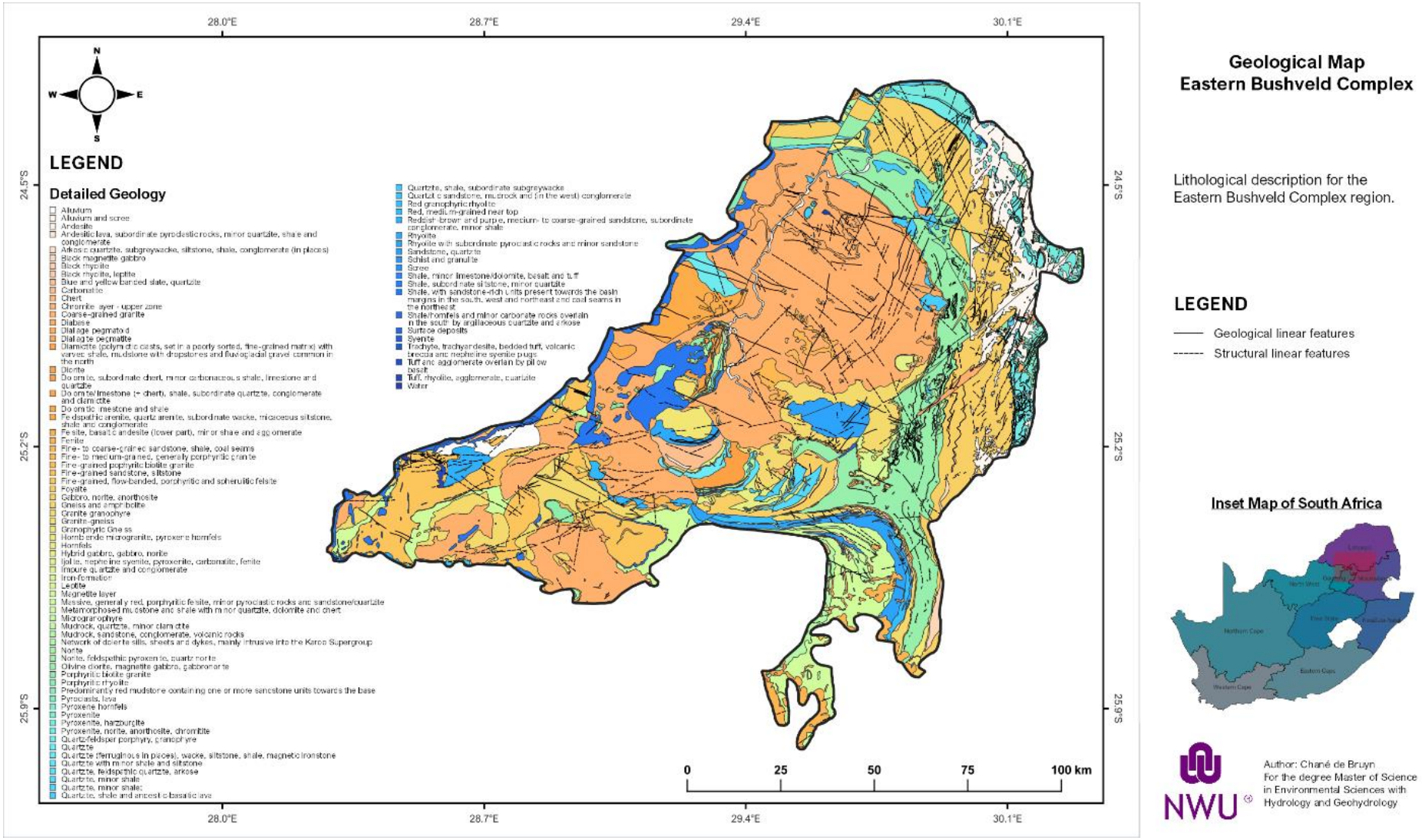


Figure 8-3: Eastern Bushveld Complex – Geology

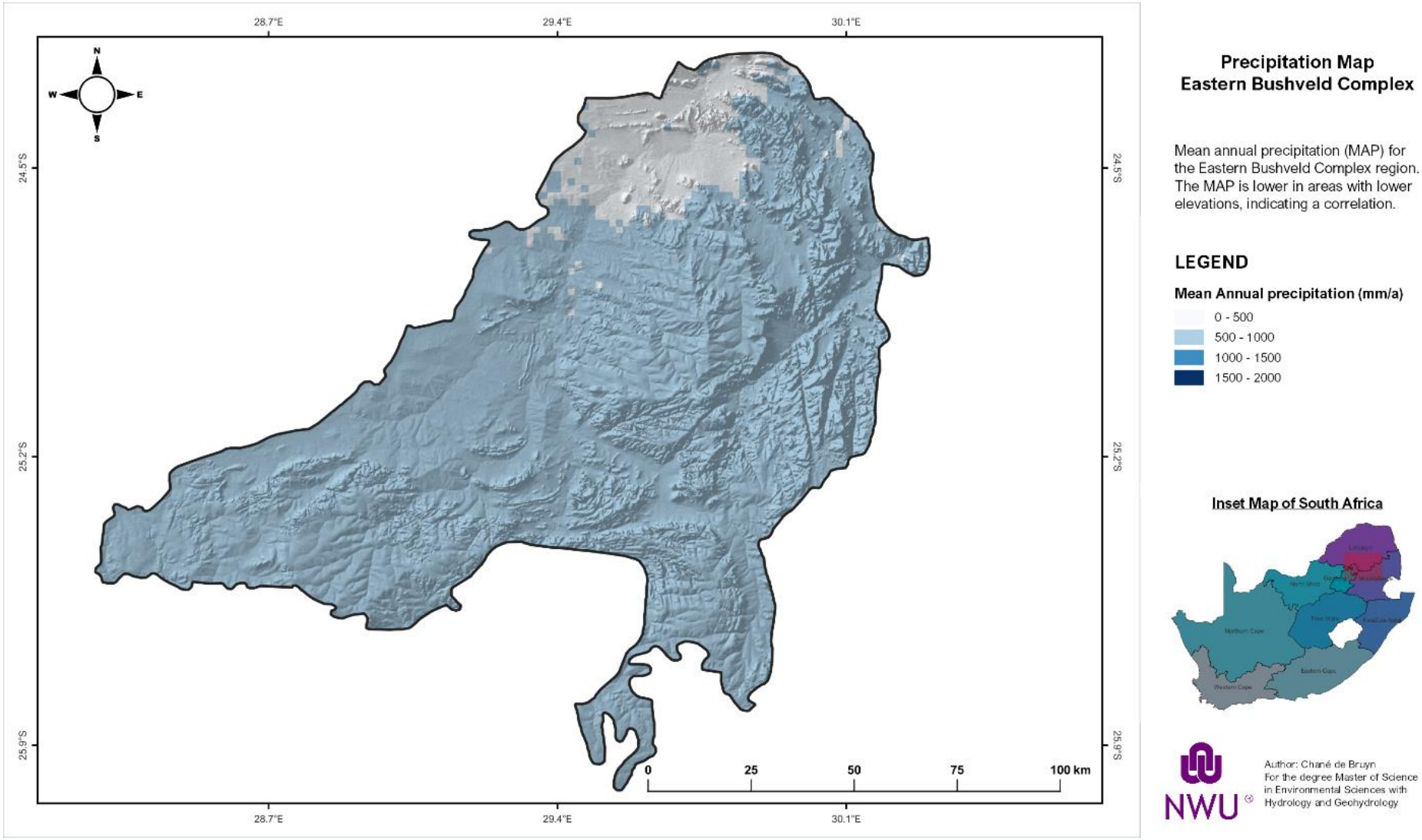


Figure 8-4: Eastern Bushveld Complex - Precipitation

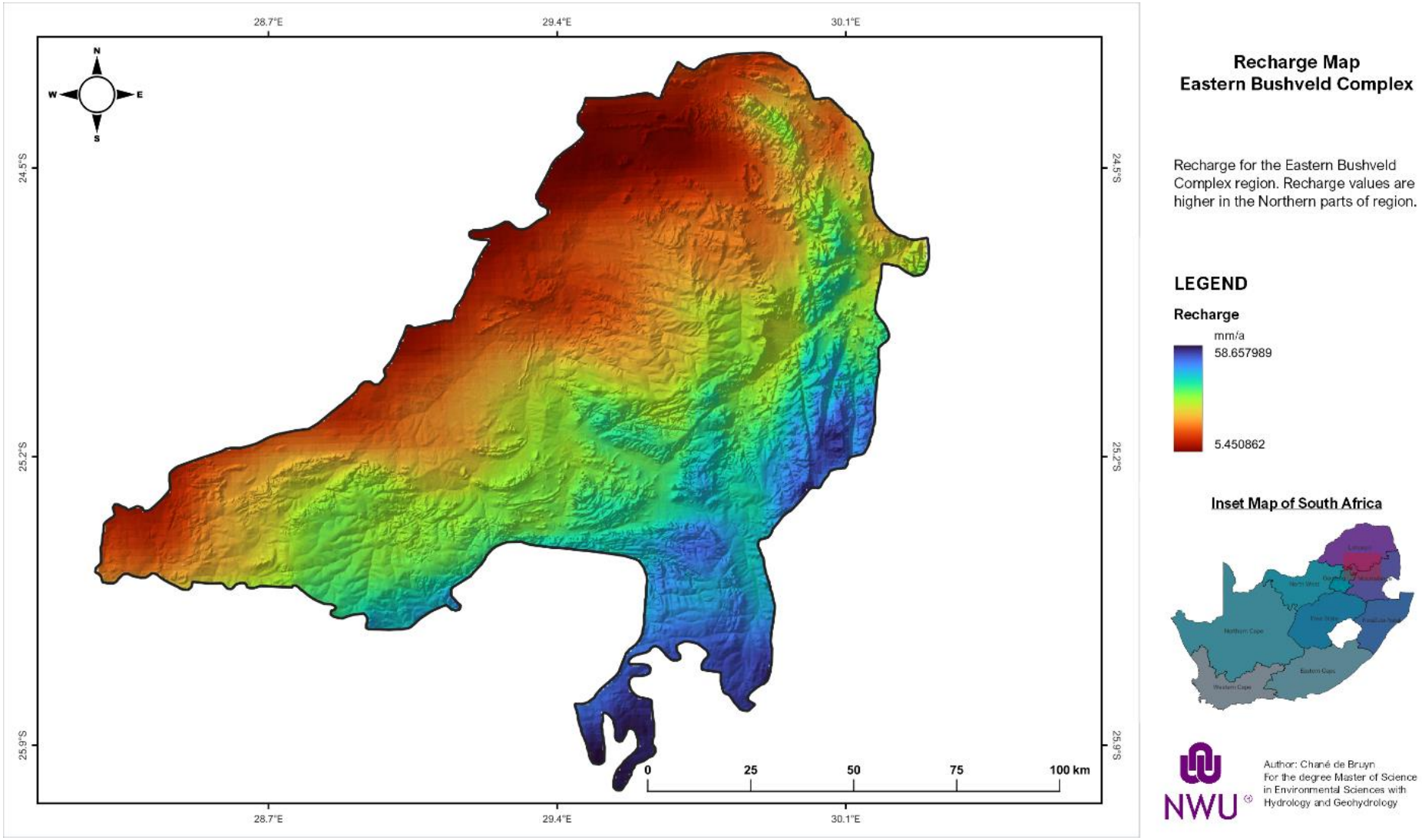


Figure 8-5: Eastern Bushveld Complex - Recharge

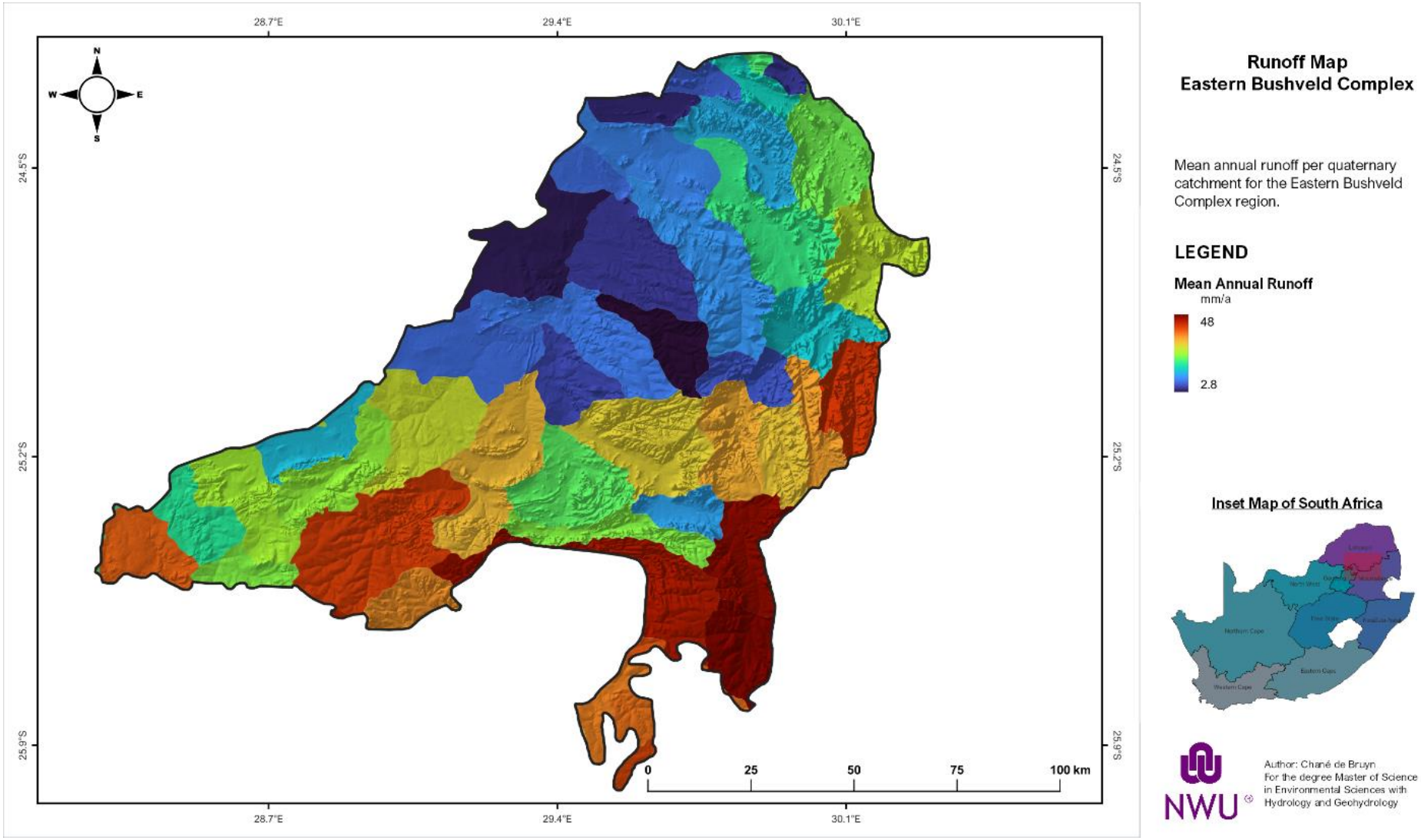


Figure 8-6: Eastern Bushveld Complex - Runoff

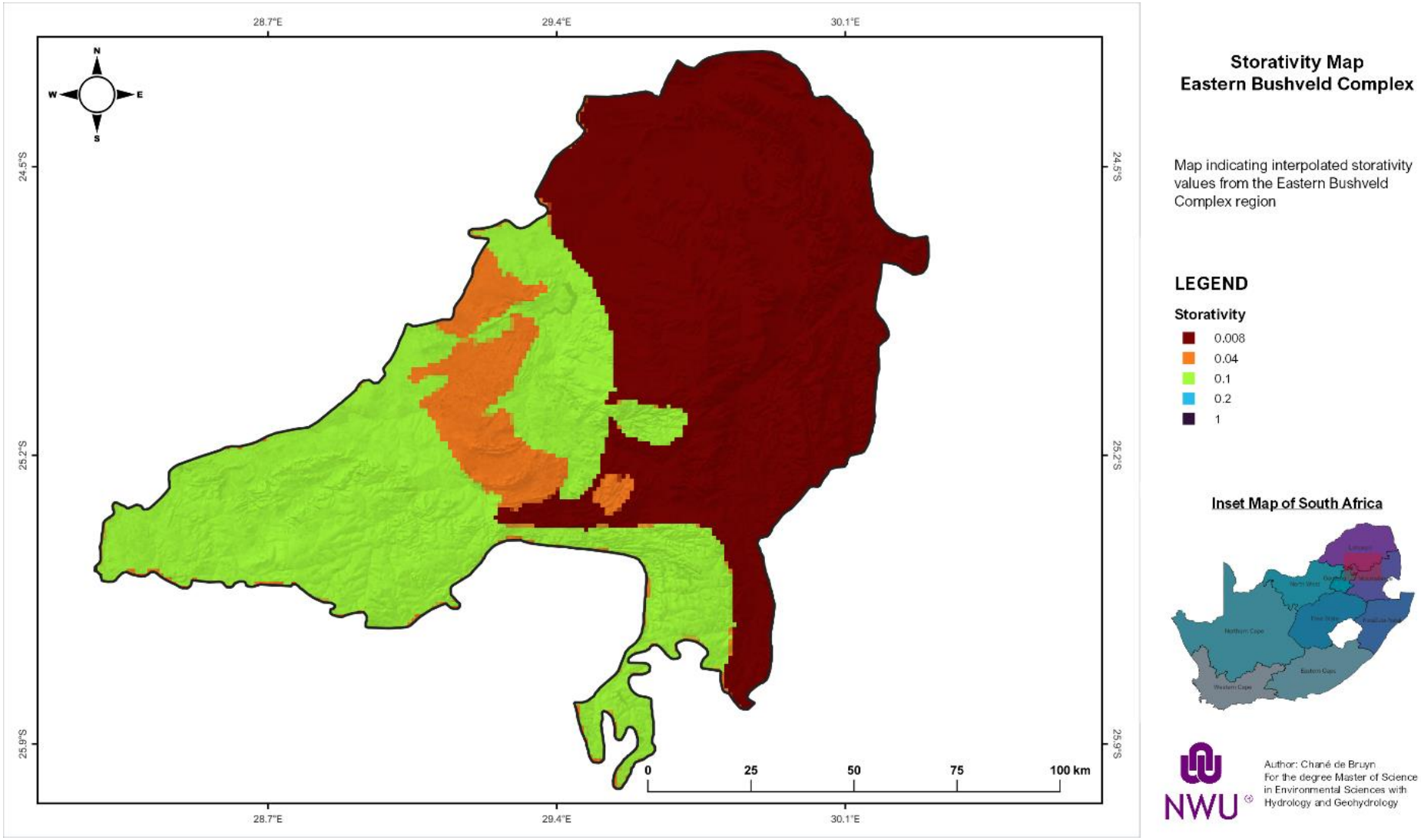


Figure 8-7: Eastern Bushveld Complex - Storativity

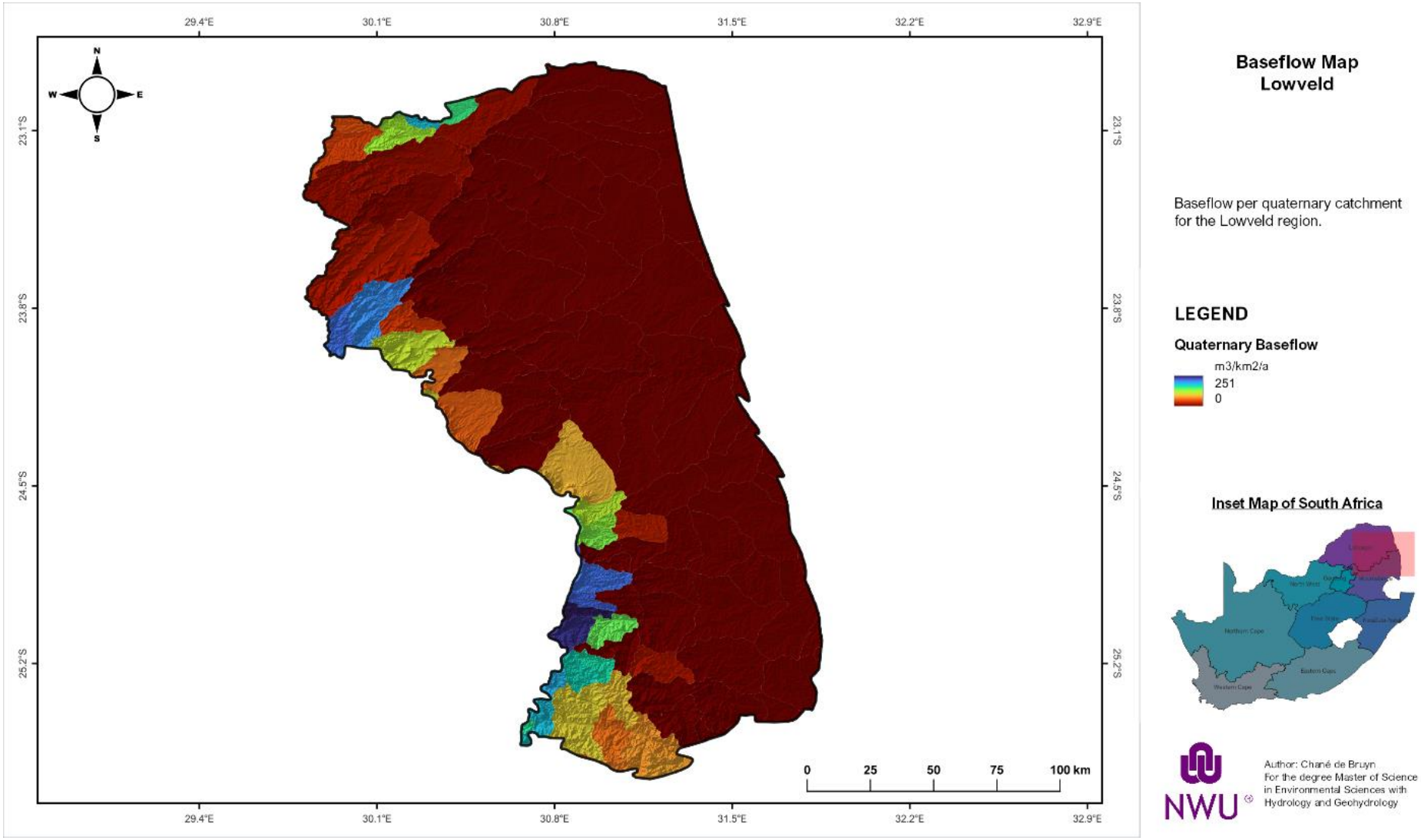


Figure 8-8: Lowveld - Baseflow

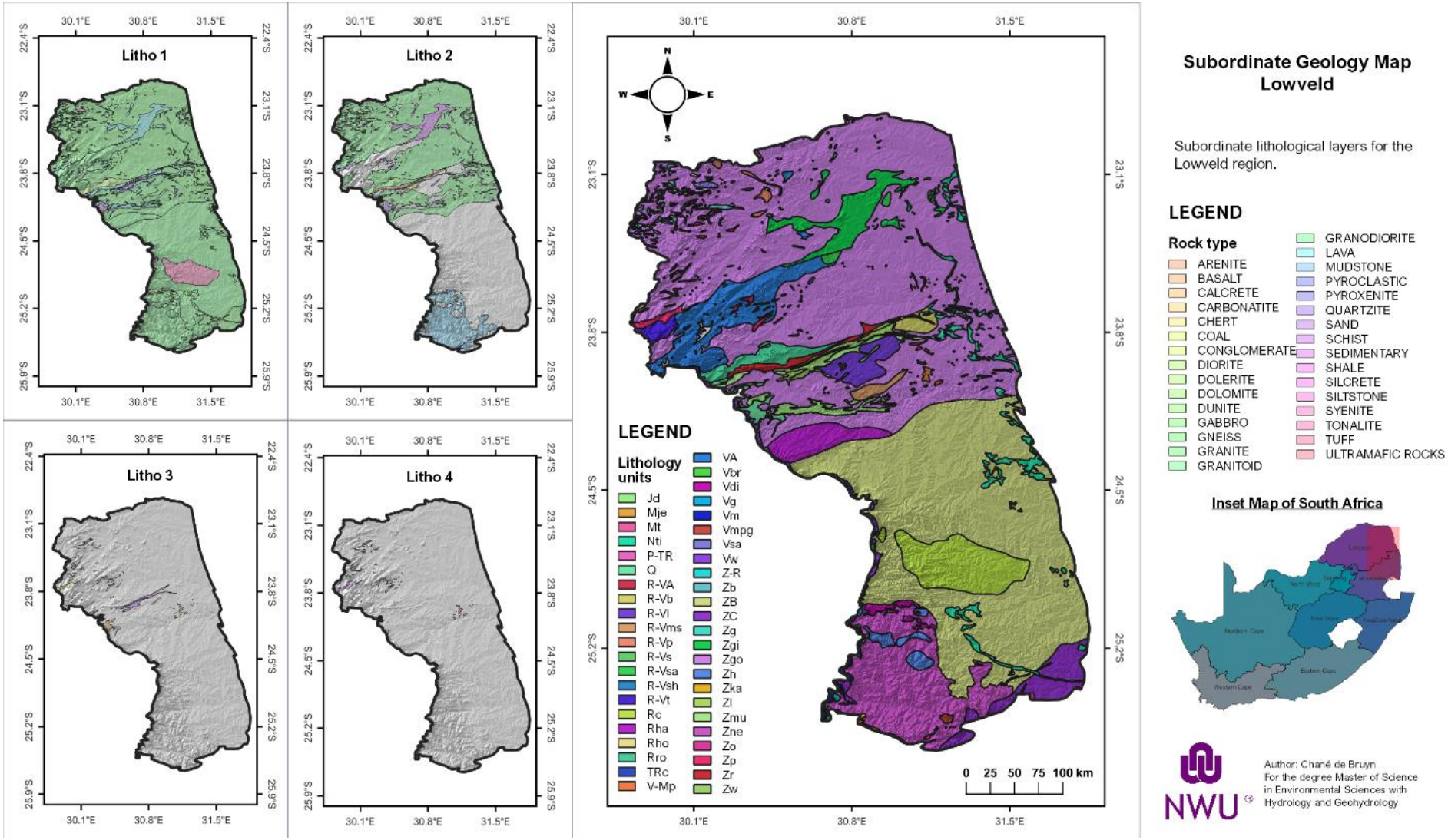


Figure 8-9: Lowveld - Lithology

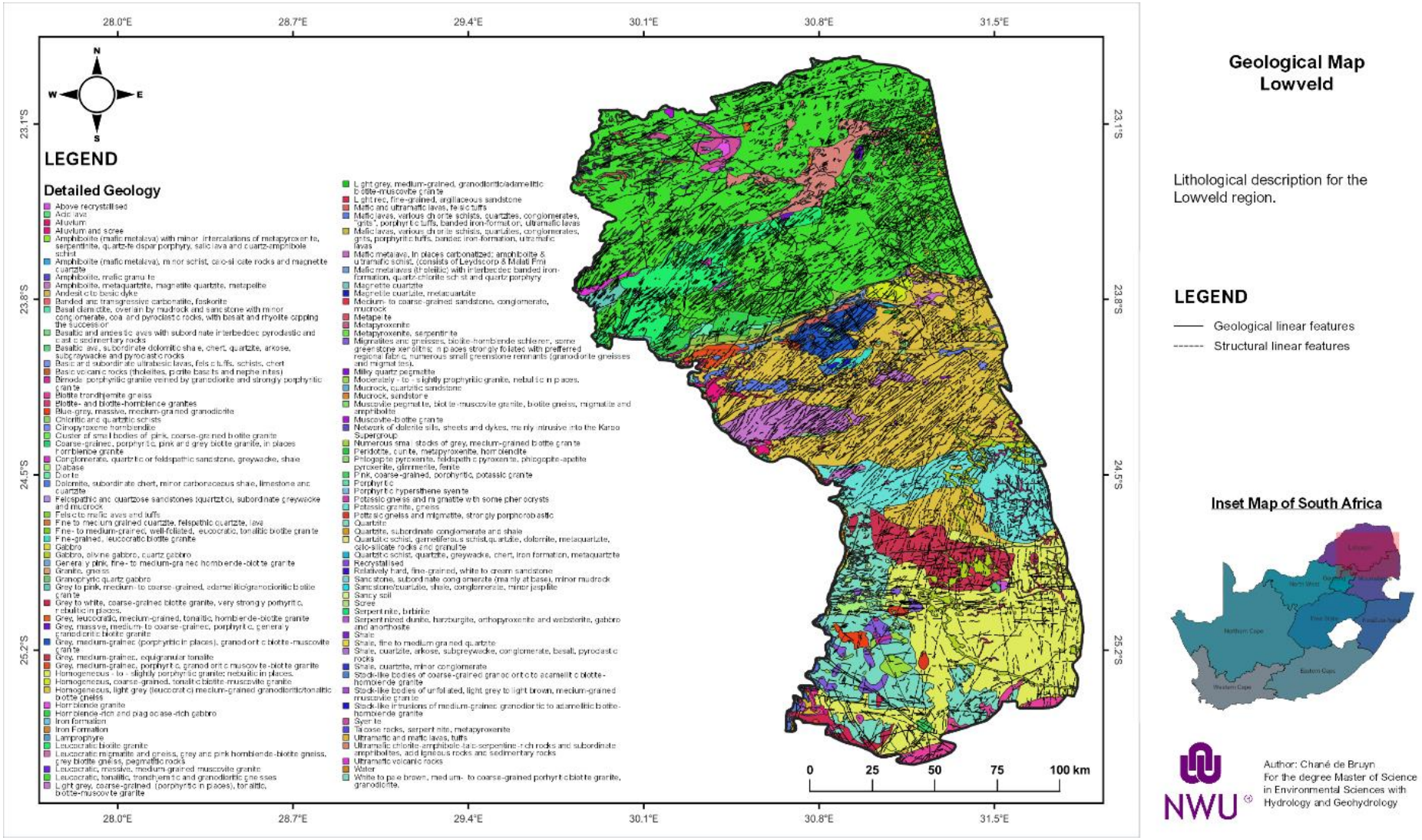


Figure 8-10: Lowveld - Geology

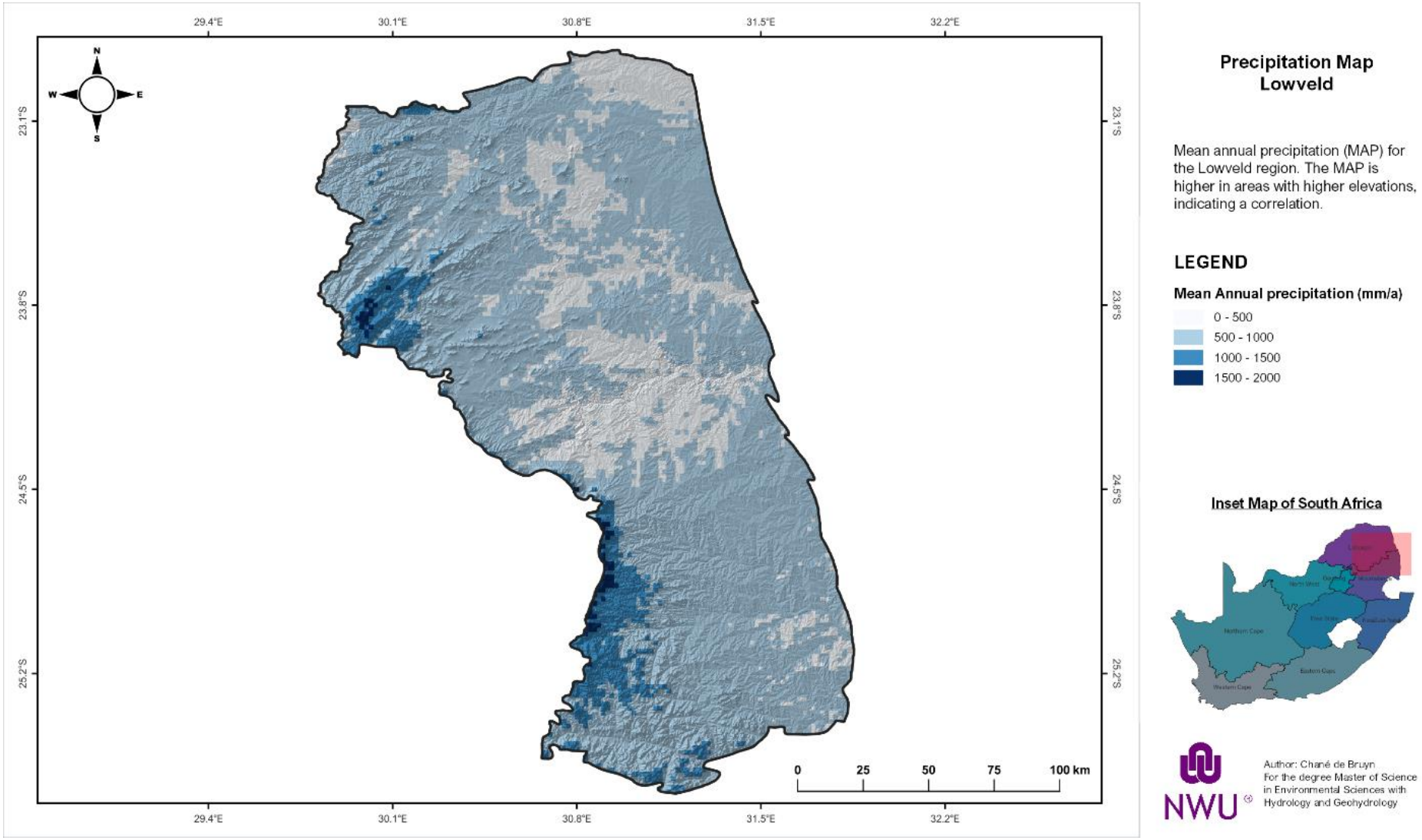


Figure 8-11: Lowveld - Precipitation

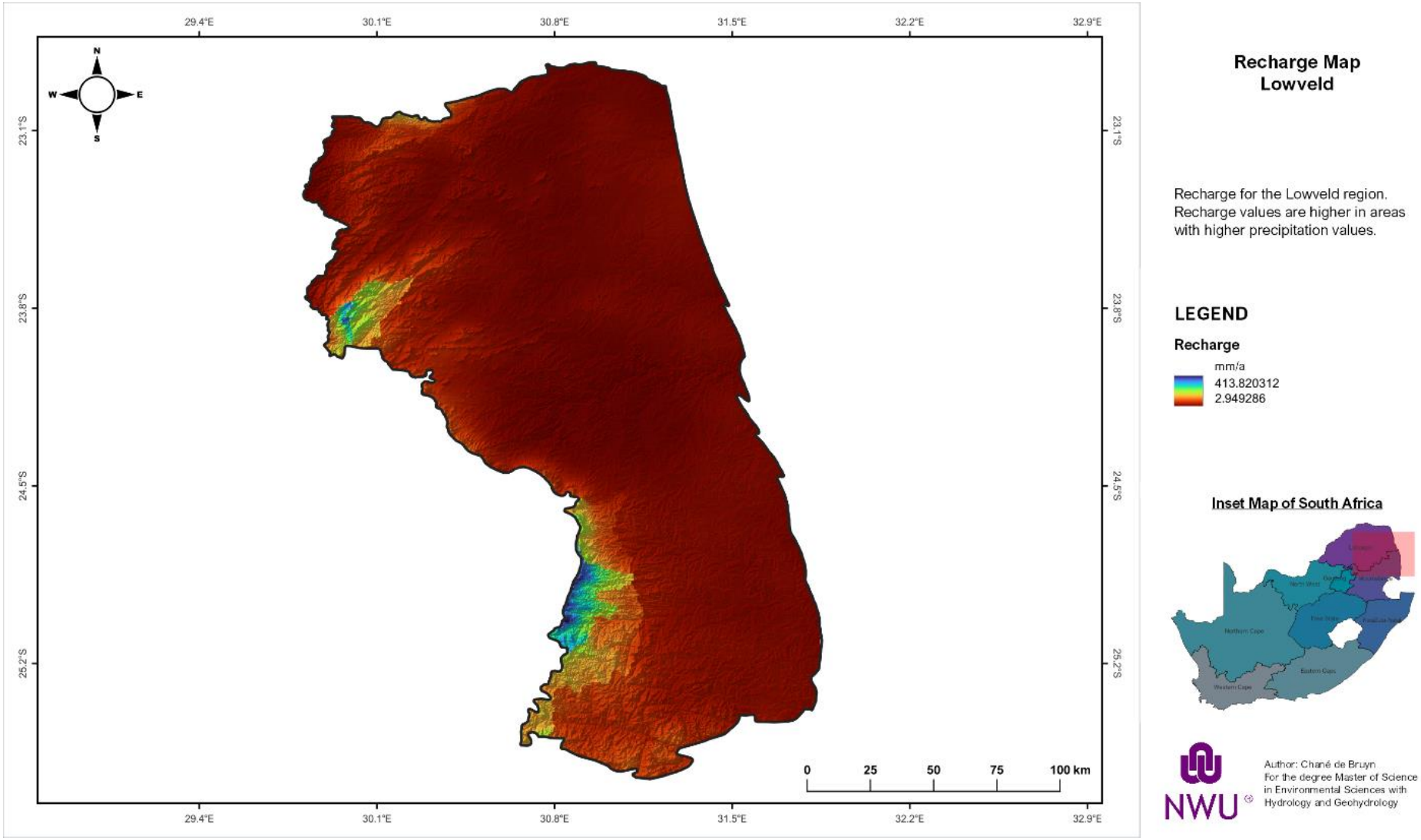


Figure 8-12: Lowveld - Recharge

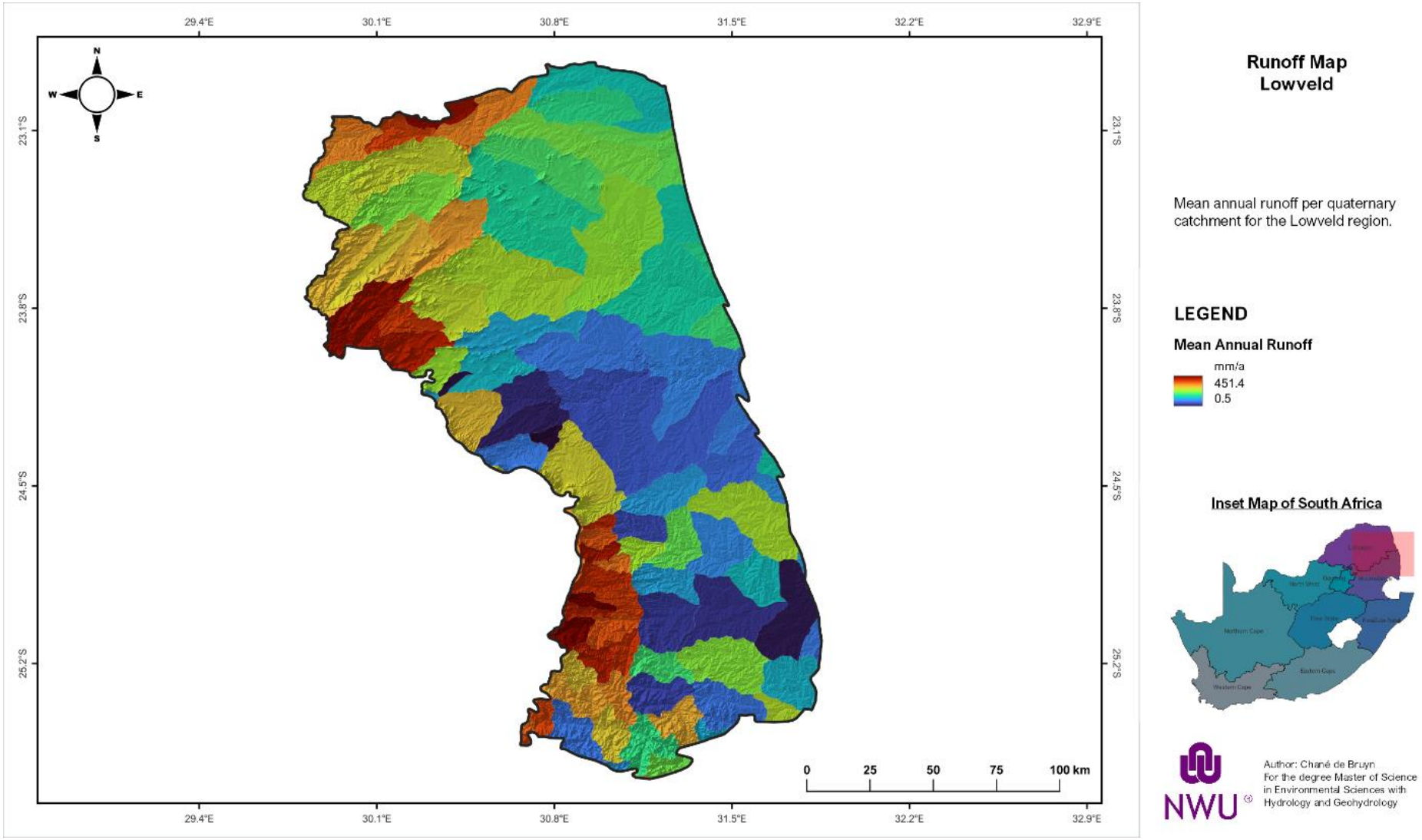


Figure 8-13: Lowveld - Runoff

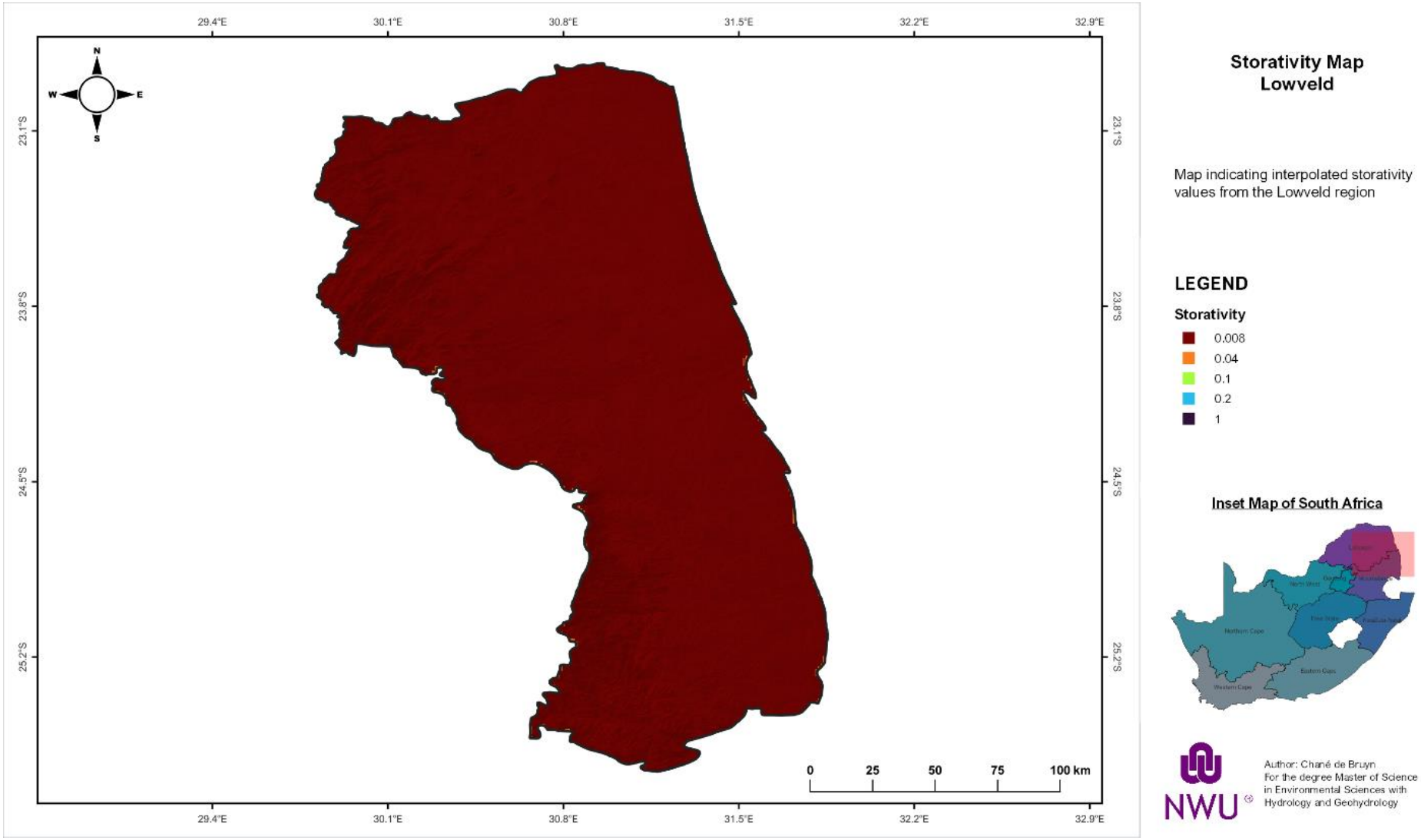


Figure 8-14: Lowveld - Storativity

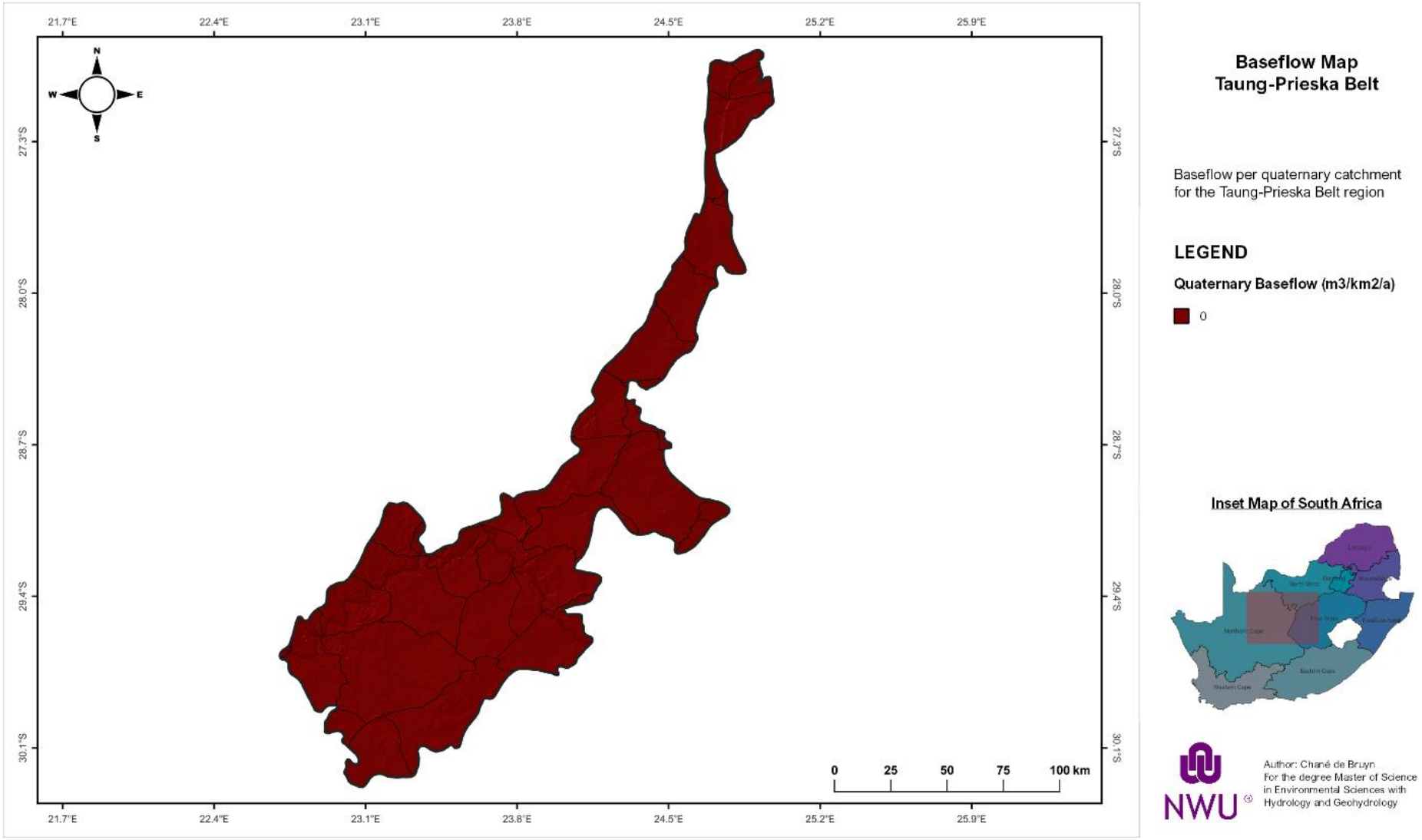


Figure 8-15: Taung-Prieska Belt - Baseflow

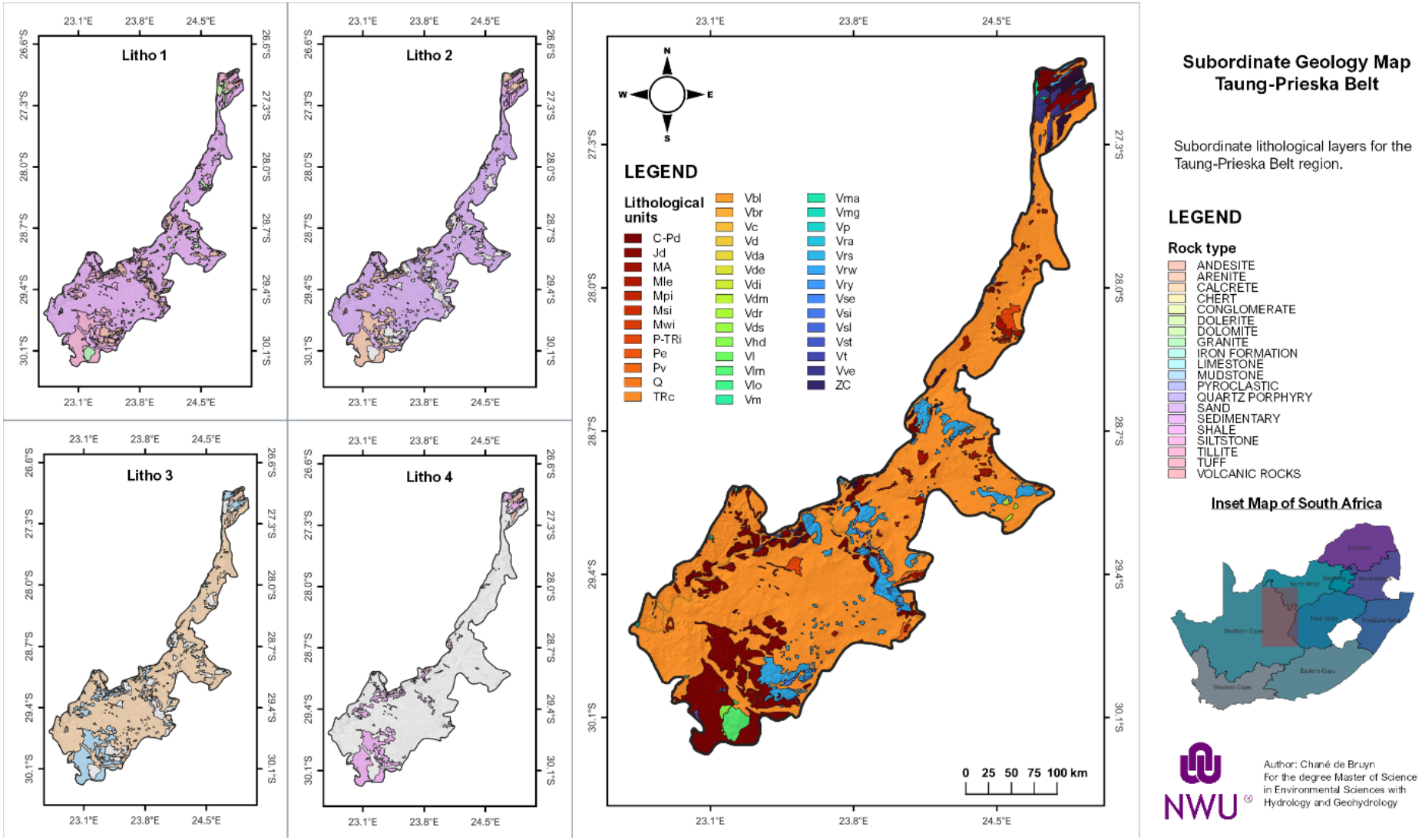
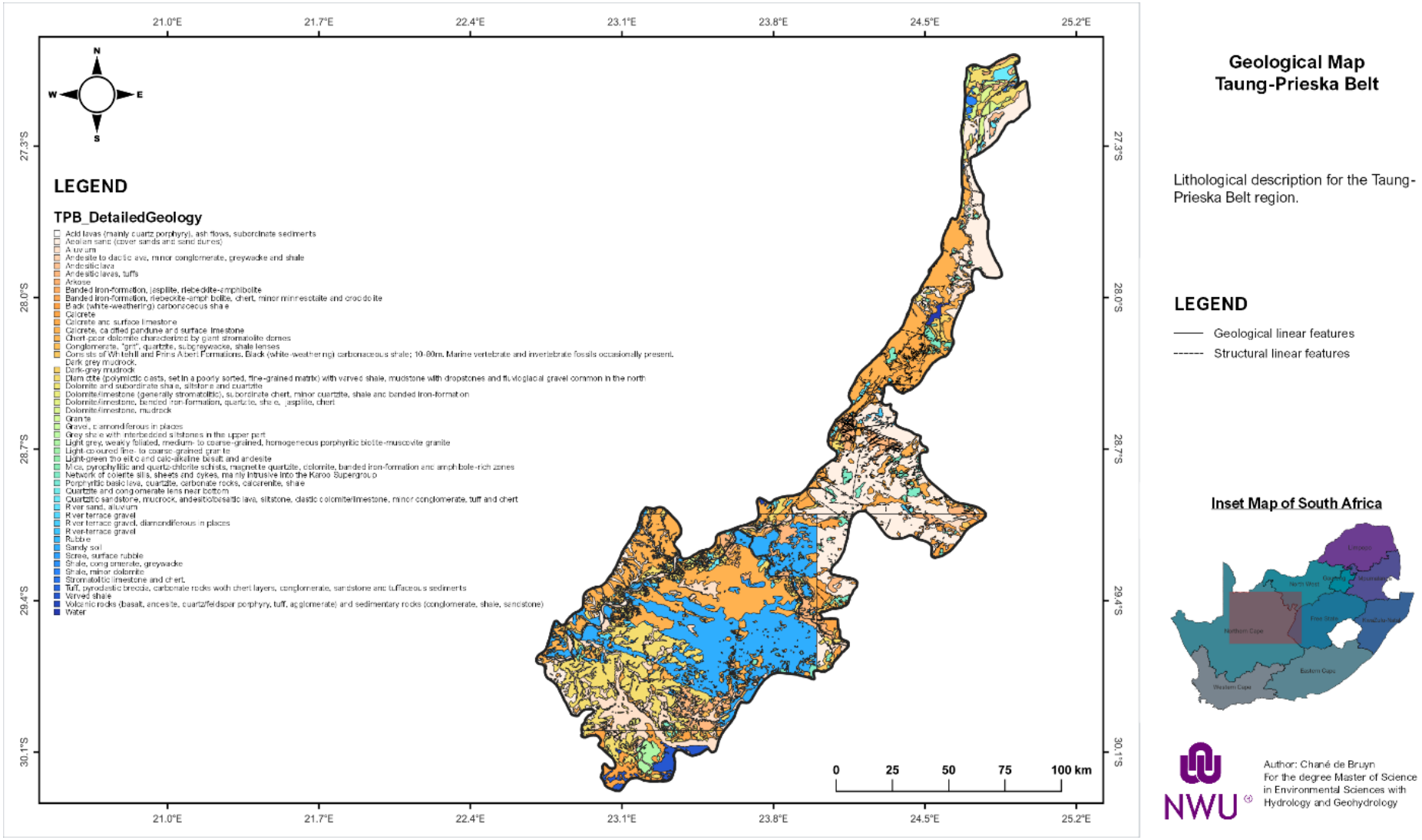


Figure 8-16: Taung-Prieska Belt - Lithology



Geological Map
Taung-Prieska Belt

Lithological description for the Taung-Prieska Belt region.

LEGEND

—— Geological linear features

------ Structural linear features

Inset Map of South Africa

NWU

Author: Chané de Bruyn
For the degree Master of Science in Environmental Sciences with Hydrology and Geohydrology

Figure 8-17: Taung-Prieska Belt - Geology

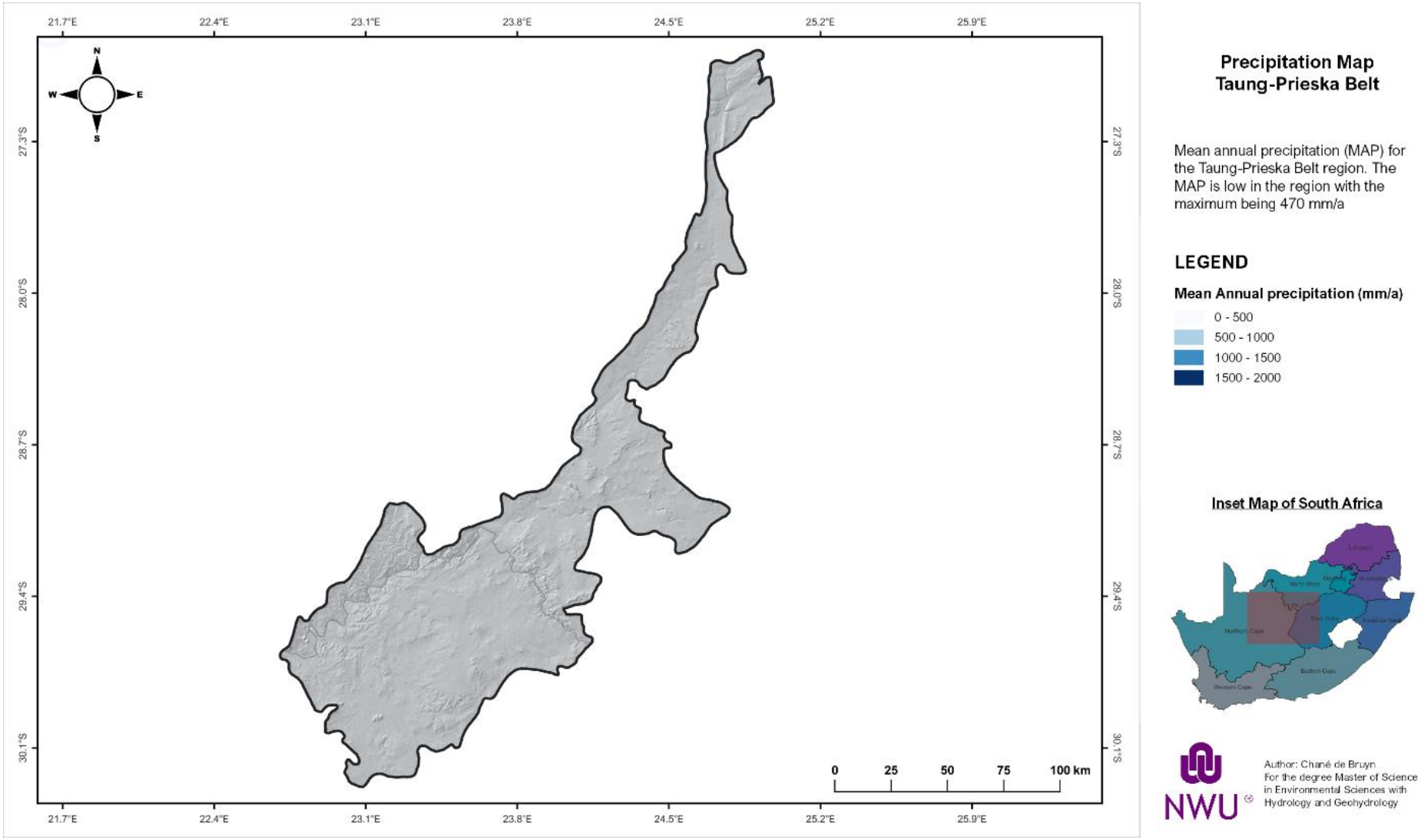


Figure 8-18: Taung-Prieska Belt - Precipitation

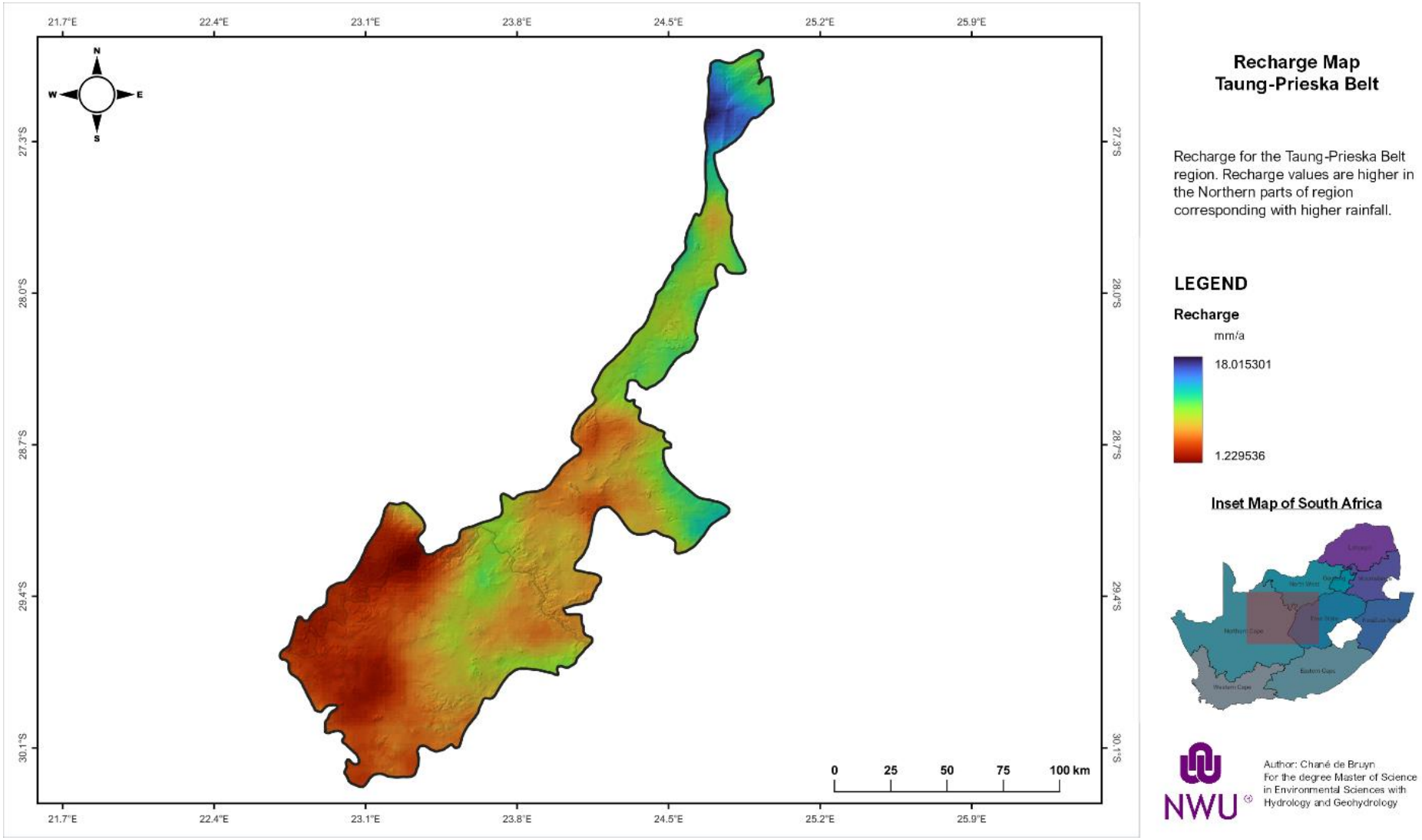


Figure 8-19: Taung-Prieska Belt - Recharge

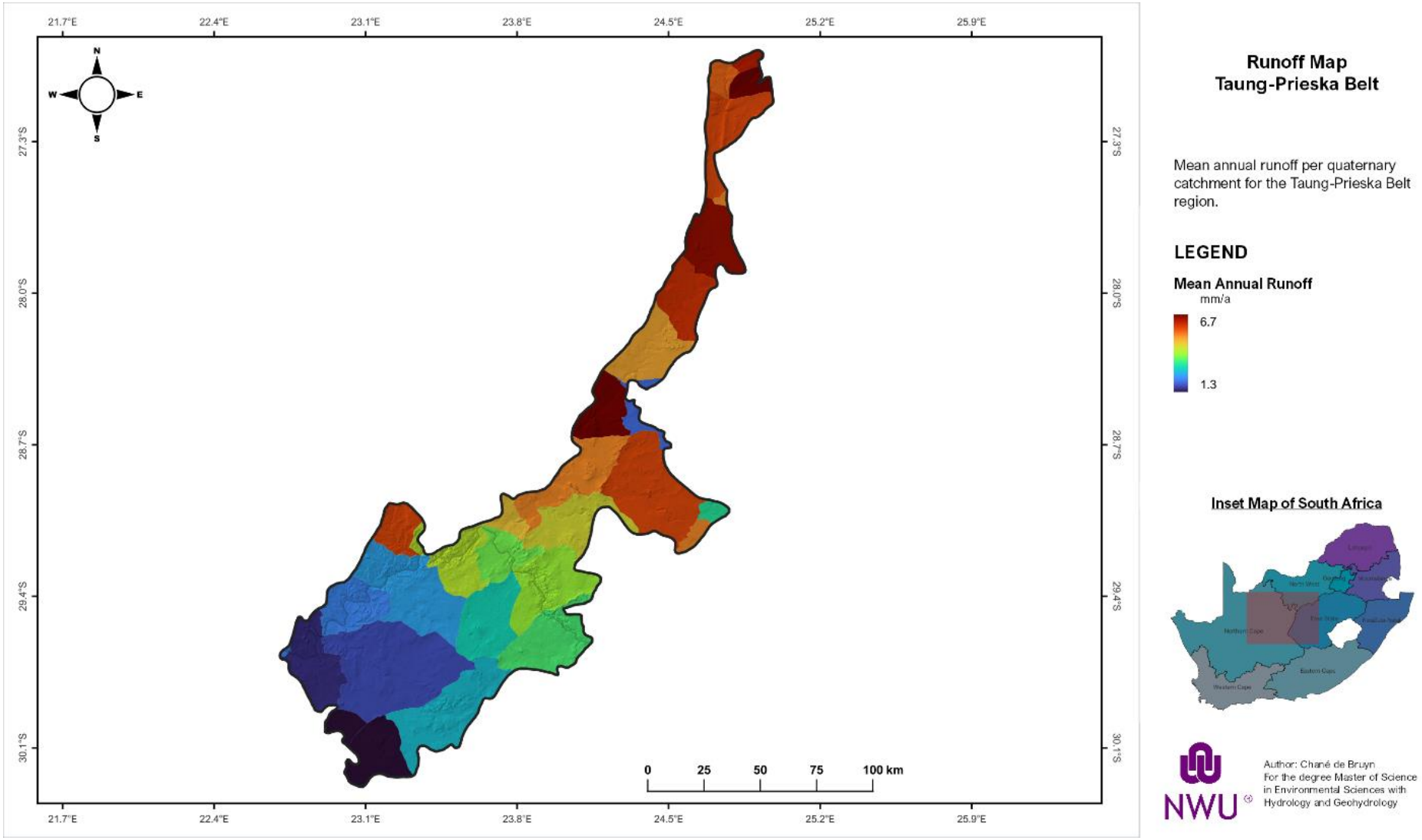


Figure 8-20: Taung-Prieska Belt - Runoff

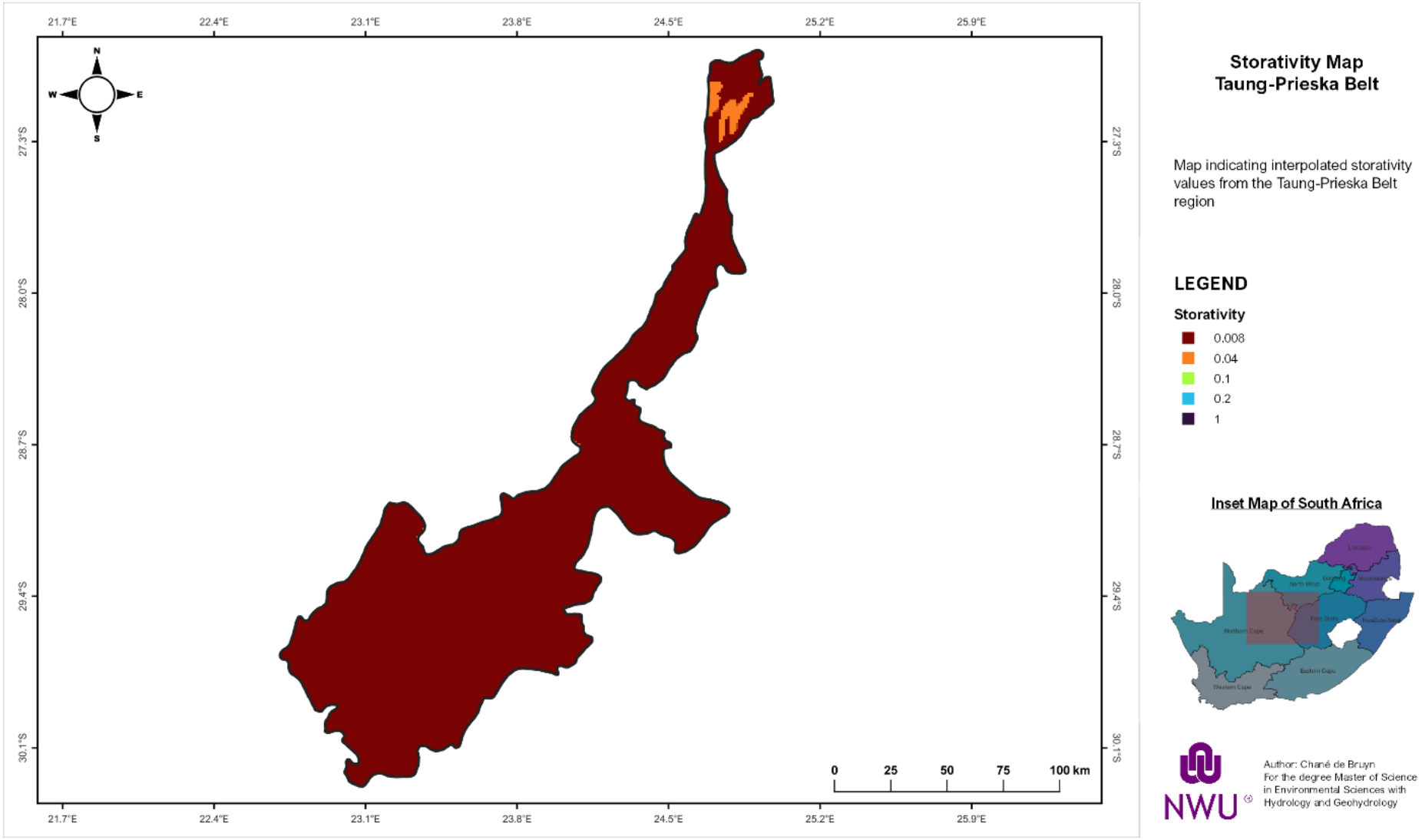


Figure 8-21: Taung-Prieska Belt - Storativity