BIG DATA ANALYTICS AND MODELLING LOCALISING TRANSBOUNDARY DATA SETS IN SOUTHERN AFRICA: A CASE STUDY APPROACH

Z Gaffoor, K Pietersen, A Bagula, N Jovanovic, T Kanyerere and G Wanangwa

USAID FROM THE AMERICAN PEOPLE
science & innovation
Department: Science and Innovation
REPUBLIC OF SOUTH AFRICA
GROUNDWATER MANAGEMENT INSTITUTE
IBM Research | Africa
SWP SUSTAINABLE WATER PARTNERSHIP
USGS science for a changing world
WATER RESEARCH COMMISSION
TT 843/20

Big Data Analytics and Modelling

Localising transboundary data sets in Southern Africa: A case study approach

Report to the
WATER RESEARCH COMMISSION

by

Z Gaffoor[1], K Pietersen[1, 3], A Bagula[1], N Jovanovic[1,2], T Kanyerere[1] and G Wanangwa[4]
[1]University of the Western Cape
[2]CSIR, Stellenbosch
[3]L2K2 Consultants (Pty) Ltd.
[4]Regional Irrigation and Water Development Office – South, Ministry,
Department of Water Resources (Ministry of Agriculture, Irrigation and Water Development),
Blantyre, Malawi

WRC Report No. TT 843/20

February 2021

Obtainable from
Water Research Commission
Private Bag X03
Gezina
Pretoria, 0031

orders@wrc.org.a or download from www.wrc.org.za

The publication of this report emanates from a project entitled *Localizing Transboundary Data Sets in Southern Africa: A case study approach* (WRC Project No. K5/2878).

This report forms part of a series of four reports. The other reports are:

- *Imagining Solutions for Extracting Further Value from Existing Datasets on Surface and Groundwater Resources in Southern Africa* (WRC Report no. TT 842/20)
- *Data Analytics and Transboundary Water Collaboration. Theme 1: Consolidation of Data and Application of Big Data Tools to Enhance National and Transboundary Data Sets in Southern Africa that Support Decision-Making for Security of Water Resources* (WRC Report no. TT 844/20)
- *Machine Learning Models for Groundwater Availability – Incorporating a Framework for a Sustainable Groundwater Strategy* (WRC Report no. TT 845/20)

ISBN 978-0-6392-0236-5
Printed in the Republic of South Africa

Executive summary

Big Data analytics is a novel and innovative package of tools and methods used to analyse and transform large volumes of heterogeneous data into information. Big Data is the term used to describe these vast collections of heterogeneous, multidimensional datasets generated by various sources, such as sensors, connected devices, computer simulations, satellite observations, research missions, ground-based monitoring, social media, and business transactions to name a few. In the groundwater discipline Big Data can provide new methods of information discovery, that can support efforts of sustainable groundwater management. However, the application of Big Data analytics is particularly nascent in the groundwater discipline. The data in the groundwater discipline is challenging to integrate and analyses. This is largely due to differences in spatial and temporal scale, the multi-dimensionality of various datasets, and the complexity of the natural systems. Big Data analytics can play a role in addressing these complexities, by transforming large groundwater datasets into actionable information to improve groundwater management.

Thus, the purpose of this research is to investigate the use of Big Data analytics to integrate, match and model groundwater data, especially at a local scale, to improve sustainable groundwater management. A case study application is undertaken, with a focus on transboundary aquifers in the Southern African Development Community (SADC). These aquifers are vital water resources for region, but their sustainable management is hindered, in part due to a lack of relevant groundwater data. The unique challenges experienced in SADC provide an opportunity to test Big Data analytics in data scarce regions, by augmenting the data gaps with new sources of data.

A transboundary aquifer (TBA) analytics framework was designed to provide a methodical approach to the application of Big Data analytics to support groundwater management. The main components of the framework include case study area selection, establishment and selection of various groundwater management scenarios and a set of sustainability indicators, collection and ingestion of relevant data to support the groundwater management scenario, the integration and modelling of sustainability indicators using Big Data analytics to better inform local groundwater management. In addition, downscaling and merging data of different spatial resolutions are included in the framework.

In modern Big Data environment, the computing resources required to store and analyse vast amounts of groundwater data, is generally beyond the capabilities of standard computing machines. Therefore, this research presented an architecture for a novel Big

Data platform to facilitate the collection, storage, processing, analyses, and information delivery of groundwater Big Data. The architecture proposes a multi-layered approach, with a “data sources” layers which constitute the physical layer where all raw data are located, a data collection/identification layer which identifies and acquires data, a middleware layer which is dedicated to the storage and analysis of the data, and the application layer where applications related to different analysis and groundwater management scenarios are implemented. This type of architecture was designed to be deployed either as a stand-alone system, that is housed at an organization’s premises, or a federated cloud computing infrastructure, where resources are shared amongst federation members. The latter is preferred due to the shared resources, fault tolerance handling (federated systems limit total system failure), and data sharing capabilities.

Two case study areas were chosen to explore the applications of Big Data analytics to groundwater management: The Zeerust/Lobatse/Ramotswa Dolomite aquifers of Botswana and South Africa, and the Shire Valley Alluvial Transboundary aquifer in Malawi and Mozambique. The status quo of the TBAs was discussed within the context of undesirable results, based on a set of groundwater management scenarios adapted from the California Department of Water Resources. From this, the scenario of chronic lowering of groundwater levels was investigated using Big Data analytics.

Relevant data considered important to the understanding of groundwater levels in the aquifer were collected from various remote sensing and land surface model sources. This includes hydroclimatic variables such as precipitation and evapotranspiration, as well as aquifer properties, such as aquifer type and groundwater storage changes. In total 9 predictor variables were chosen to model a single predictant variable, namely groundwater levels, using Big Data analytics. The data was pre-processed into a set of comparable variables with consistent temporal and spatial resolution. 30-day groundwater level changes were calculated (predictant). The regional scale data were integrated with the local data, where raster values were converted to tabular data and appended to the groundwater level change records.

In this research, we applied a machine learning approach to model and predict groundwater level changes across the study areas. Specifically, we relied on a Gradient Boosting Decision Tree (GBDT) to develop a generalised machine learning model to predict 30-day groundwater level changes at ~5 x 5 km resolution across the study areas. Through a process of model training, validation, and testing, a GBDT model was developed for the Ramotswa aquifer that predicted 30-day groundwater level changes with a mean absolute error of ~17.8 cm. Unfortunately, for the Shire Valley TBA, sufficient data series was only available for a single borehole, where monthly

groundwater level changes were predicted with a mean absolute error of 34.6 cm. The model for the Ramotswa aquifer was then used to predict monthly groundwater level changes at ~5 x 5 km. The results revealed significant groundwater level declines across the study area between 2002 and 2019.

The results illustrate the potential for Big Data analytics to provide information on chronic lowering of groundwater levels in the study area. However, concerns regarding model performance in terms of predicting extreme values are still present, as well as issues regarding effective model training. This is a consequence of aquifer processes such as abstraction and episodic events that could not be accounted for. Nonetheless the results suggest that additional data and the inclusion of features such as abstraction may improve model performance.

Overall, the use of Big Data analytics to support groundwater management in the SADC region has potential uses. However, the data need to be readily available and at a sufficient level to develop machine learning models. In many cases, various Big Data sources can be augmented and integrated to improve data availability, but in situ observation are still required to validate model performance. Beyond this, challenges associated with the collection and processing of large datasets can be expected, and they need to be overcome. Finally, the data, tools and information need to be packaged into forms that facilitate decision making on the ground.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AAAS	Archiving as A Service
AGDC	Australian Geoscience data cube
AI	Artificial Intelligence
API	Application programming interface
BDA	Big data analytics
CoP	Community of Practice
CSIR	Council for Scientific and Industrial Research
CSR	Centre For Space Research
CW	Canopy Water Storage
ECMWF	European Centre for Medium Weather Forecasts
DSI	Department Science and Innovation
EFB	Exclusive Feature Bundling
ESA	European Space Agency
ESGF	Earth System Grid Federation
GBDT	Gradient Boosting Decision Tree
GCM	Global circulation models
GFZ	GeofurschungsZentrum
GIS	Geographic information system
GLDAS	Global Land Data Assimilation System
GMAs	Groundwater Management Area
GMUs	Groundwater Management Units
GOSS	Gradient Boosting Decision Tree
GRACE	Gravity Recovery and Climate Experiment
GRU	Groundwater Resource Unit
GUA	Groundwater Unit of Assessment
GWL	Groundwater Level
HPC	High Performance Computing
IAAS	Infrastructure as a service
IoT	Internet of things
MAE	Mean Absolute Error
MODIS	Moderate Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration
NCI	National Computational Infrastructure
NOAA	National Oceanic and Atmospheric Administration
NW	North West
OGC	Open Geospatial Consortium

PAAS	Platform as a Service
PAIRS	Physical Analytics Integrated Data Repository and Services
SAAS	Software as a Service
SADC	Southern African Development Community
SADC-GMI	SADC-Groundwater Management Institute
SQL queries	Structured Query Language queries
SWE	Snow Water Equivalent Thickness
SWP	Sustainable Water Partnership
TBAs	Transboundary aquifers
TRMM	Tropical Rainfall Measuring Mission
USAID	United States Agency for International Development
USGS	United States Geological Survey
UWC	University of the Western Cape
WRC	Water Research Commission

Mapping convention

All maps are projected in WGS 84 (EPSG:4326) co-ordinate reference system. The units used are degrees. Maps are displayed at the same scale.

Acknowledgements

This research was funded by the Big Data Analytics and Transboundary Water Collaboration for Southern Africa, a multi-agency coalition with the Department of Science and Innovation South Africa, the USAID, the Southern African Development Community-Groundwater Management Institute (SADC-GMI), the South African Water Research Commission (WRC), the Sustainable Water Partnership and the United States Geological Survey (USGS).

1. Introduction and background

1.1. The Big Data Analytics and Transboundary Water Collaboration for Southern Africa

This research project, managed by the Water Research Commission (WRC) of South Africa, is part of a series of four projects under the Big Data Analytics and Transboundary Water Collaboration for Southern Africa, bringing together key stakeholders in Water and Big Data sectors.

The *Collaboration* was first conceptualised in 2014 during the African Leaders Forum in Washington D.C., between United States Agency for International Development (USAID) Global Development Lab and IBM Africa Research, which had opened its first hub in Nairobi (Kenya) in 2013, followed by the Johannesburg Lab in 2015. Since the early 2000s, the regional USAID mission for Southern Africa had been intensifying its regional support for transboundary water systems with both the Ramotswa Aquifer Project, involving Botswana and South Africa and the Resilience in the Limpopo River Basin Program (currently in its second phase with the Resilient Waters Programme, covering the entire Southern Africa region, with a focus on the Limpopo and Okavango River Systems). As part of this process, USAID had also been engaging with the Southern African Development Community (SADC): Groundwater Management Institute (GMI) and the Department of Science and Innovation of South Africa to support knowledge and technological advancement in the region. The focus of this multi-agency collaboration was agreed as Big Data Analytics and Transboundary Water. On April 3, 2017, the partners met with a multi-stakeholder regional group in a dynamic "Idea Jam" hosted by the IBM Africa Research Lab in Johannesburg. The objective was twofold:

- To answer the broad question "how best can big data analytics be used to enhance transboundary water management", and
- To identity the research questions, which would have guided the projects.

Requiring the collaboration of at least five high profile government agencies and private institutions, it took over one year to move from the Idea Jam to the launch of the Call for Proposals in August 2018, and the awarding of the four research projects in January 2019.

1.1.1. The Collaboration: its partners and objectives

Currently, the Collaboration has seven partners, with a joint function for USAID Global Development Lab, Water Office, and Southern African Mission. The partners each contributed to the development of the research projects based on own technical and funding capacity (Figure 1). The total funds provided by the Funding Partners to research directly amount to USD $ 500,000 (40%, 40%, 20%). IBM Africa contributed with the provision of the venue in Johannesburg, ad hoc, but more importantly, by sponsoring the internship programme to the five candidates from the research projects.

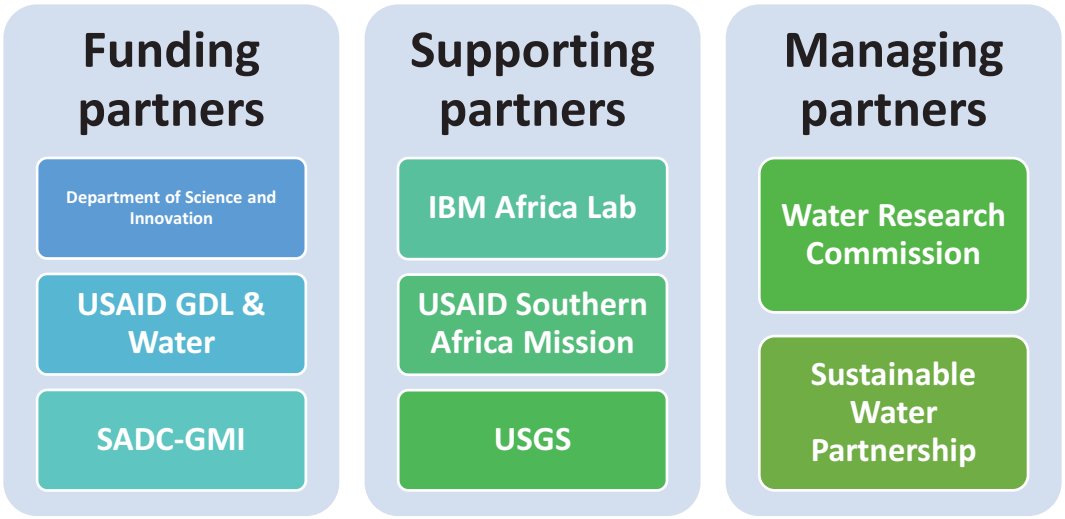


Figure 1: Collaboration partners & functions

The WRC is primarily tasked to oversee the financial and implementation management of the four research projects, as well as final reporting. The Sustainable Water Partnership (SWP) was called in by USAID in 2018 to act as the overarching Programme Coordinator, tasked with providing relation management, overall objective achievement, direction, and positioning for the Collaboration in the region, and the fostering of a Community of Practice.

The United States Geological Survey (USGS), IBM Research and SWP provided three sets of online training on issues pertaining to the focal topics of the Collaboration, which are now available on the Collaboration YouTube channel.

The Collaboration partners defined the objectives for this first phase of action (Table 1). However, the long-term vision is to create a Community of Practice (CoP) for research

and innovation on Big Data for Water Security, building on the multi-donor environment which has proven successful.

Table 1: Collaboration goals and objectives

Goals	Objectives
Enhance current understanding of shared groundwater resources	Improve transboundary groundwater management and collaboration
Provide big data skills development, capacity building and networking opportunities for Southern African researchers and their students	To foster multi-agency collaborative funding opportunities
To promote innovative thinking and application of Big Data Analytics to the Transboundary Water sector for integrated decision-making	To plant the seed for a growing community of pioneers in the use of Big Data Analytics for the study and management of Transboundary Water Aquifers

1.1.2. Research projects: funding and training

The four projects were awarded between December 2018 and January 2019, with a focus on a secondary river basin in the region: the Ramotswa, part of the Limpopo River Basin, spanning Botswana and South Africa. All the lead institutions of the project teams have partnered, see Figure 2, with Botswana government and private institutions, as well as other leaders in previous water programmes in the area, such as UN-IGRAC[1] (partner of Team 1) and IWMI[2], implementers of the Ramotswa 2 USAID Project.

[1] International Groundwater Resources Assessment Centre of the United Nations
[2] International Water Management Institute

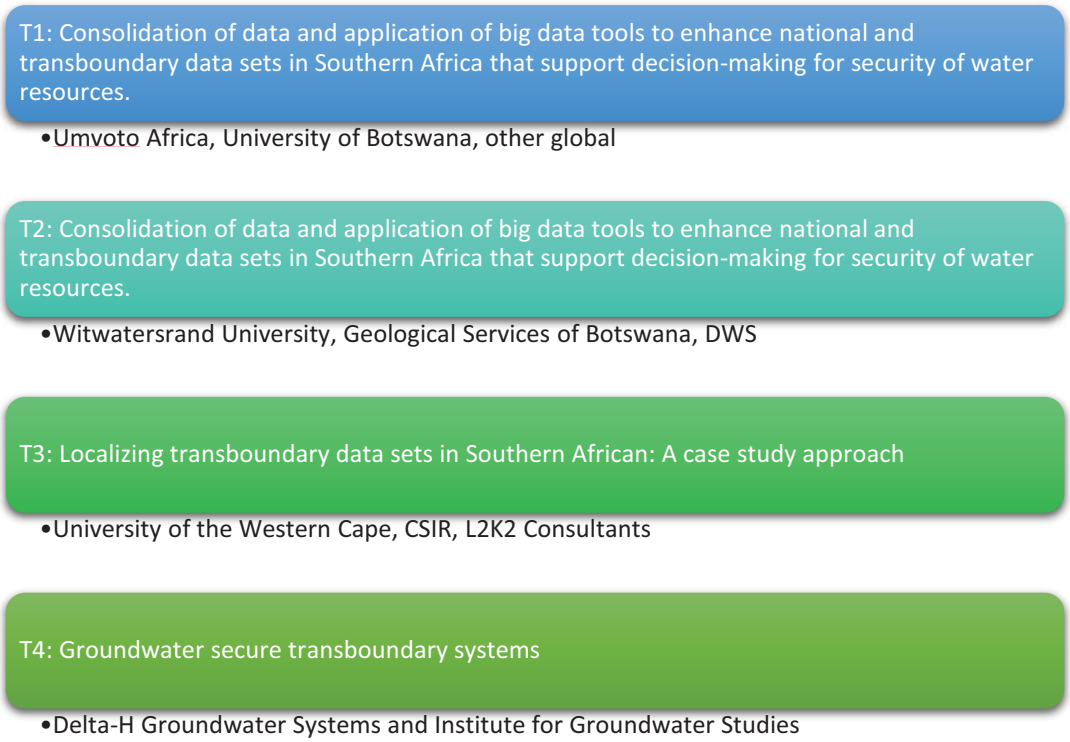


Figure 2: Title of the four thematic areas and projects

Despite working independently to address own project topics, the four research teams have progressively worked together to provide better integration for their outcomes. This process was led by the SWP in respect of providing a communication forum for the team leaders but was enhanced by the Internship Programme. The IBM mentors created a dedicated team and engaged the interns as individuals, as well as a group to help each other resolve new questions in coding and Machine Learning.

1.1.3. The future prospects

As the current phase is coming to an end with the closing of the four research projects, the Collaboration partners are already identifying new opportunities to build on the lessons learnt and address the gaps recognised in this preliminary work, enhance the partnership to include national and regional government stakeholders, as well as new funding partners.

The focus of the Collaboration will remain the nexus between Big Data Analytics and (Transboundary) Water Security, recognising the inter-relatedness of successful water management in both national and shared aquifers to both human development and environmental goals.

1.2. Theme interaction and integration

A CoP approach towards theme interaction and integration was adopted. Through the practice, the theme leaders shared information and developed knowledge resulting from joint activities and discussions. This included:

- **Meetings**: Biweekly meetings, Reference Group meetings
- **Data and information sharing**:
  - **Theme 1**: Provided data pertaining to the Ramotswa TBA, which included depth to groundwater levels, GIS vector and raster layers, and other attribute data
  - **Theme 4**: Provided depth to groundwater level data for monitoring stations on the South African side of the Ramotswa TBA
- **Workshops**: Closing Workshop of the RAMOTSWA 2 Project, Data Storage Solution Workshop Online series
- **Webinars and training**: USGS Technical Webinar Series; Big Data for Water Security: Scalable Geospatiotemporal Data Integration & Systems; SWP & IUCN Webinar Series on Transboundary Water Governance for Water Security
- **Conferences**: 2nd SADC-GMI; Conference; 2019 Groundwater Division Conference: Conservation, Demand & Surety; 3rd SADC-GMI Conference
- **Internship**: A handful of young scientists from various research institutes, of which Zaheed Gaffoor (PhD candidate) was one, were chosen to participate in an internship under the mentorship of IBM Research Africa. IBM Research Africa represents a technical and training partner in the Transboundary Water Collaboration for Southern Africa and is a world class data science and artificial intelligence research institute with a proven track record in Big Data analytics (BDAs). The skills transfer learning during the internship included:
  - Python coding and scripting skills
  - Machine learning implementation
  - Application programming interface (API) interfacing
  - General data science skills (ingestion, pre-processing, exploration, etc.)
  - Various analysis outputs (e.g. maps, datasets, and figures)

1.3. Motivation

Groundwater of sufficient quantity and good quality is vital to the socio-economic development aims of Member States of the Southern African Development Community (SADC). In the SADC there are groundwater resources that transcend national boundaries, known as transboundary aquifer systems (TBAs). The TBAs are associated with important aquifers making them crucial for the strengthening of water and international cooperation among the Member States of SADC.

Conventional groundwater data such as from in-situ monitoring programs, historical reports, and computer simulations, have long been the source of information to support groundwater decision-making. However, recent assessment of the state of groundwater data in the SADC-region by the SADC-GMI revealed many constraints including limited human resources, equipment and financial capacity for collection, analysis, management, retrieval, and sharing of data; inconsistencies in data collection and routine quality control; data storage in different formats and difficulty in data access, use or interpretation (SADC-GMI et al., 2019a, b). Advances in remote sensing missions, atmospheric and land surface models, social media, and other internet-related platforms provide new sources of data for groundwater (Figure 3).

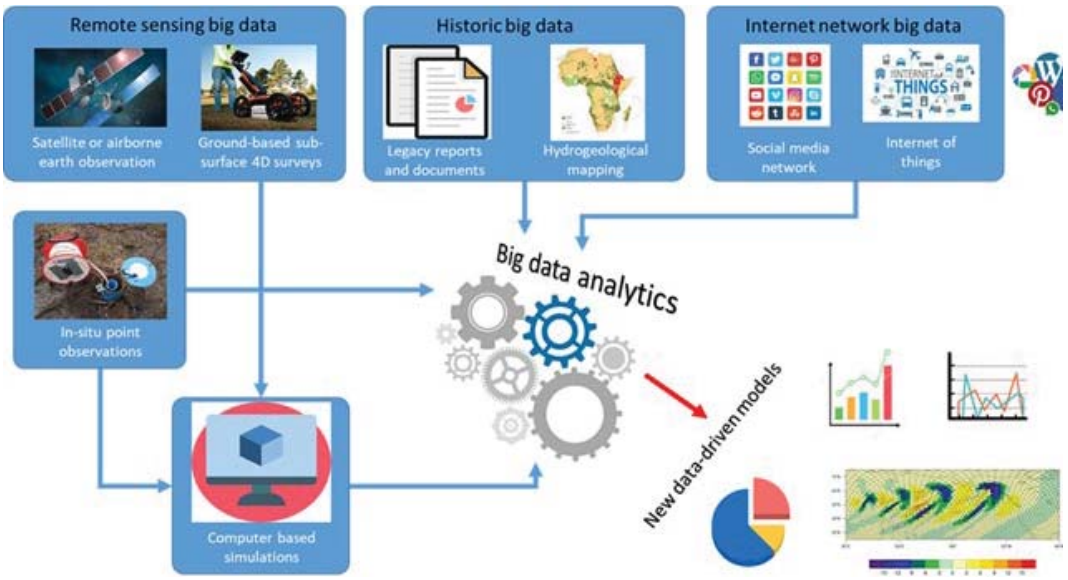


Figure 3: Sources of big data

The issue with remote sensing data and computer-based models is that the spatial resolution is more suited for regional or global studies whereas groundwater-related investigations are mostly needed at the local level. For local groundwater management

in the SADC, the data must be downscaled to a finer spatial resolution. To this end, this project explores the use of BDAs to harmonise the data sources and link models and data of different scales to support groundwater decision-making in the TBAs of the SADC at the local level.

1.4. Aims and objectives

The research was funded by the BDAs and Transboundary Water Collaboration for Southern Africa (section 1.1) and carried out by the University of the Western Cape (UWC) in partnership with the CSIR entitled "Localising transboundary data sets in Southern Africa: A case study approach." The specific aims of the study were to:

- Select case study area(s) where there are active transboundary data sets in addition to reliable local data sets and identify which data elements are of high value for joint decision-making
- Identify and prioritize data parameters critical for local decision-making by developing a matrix for various management scenarios (exploitation, protection, forecasting)
- Match, integrate and model local data with regional data (data from official and nonofficial sources)
- Integrate and downscale datasets for finer scale decision making using big data tools
- Extrapolate to similar settings (e.g. geology, climate, aquifers)
- Evaluate and test whether big data tools can model local data results at the transboundary level
- Identify and assess important problems/issues regarding transboundary water systems and their management

1.5. Project team

To ensure a multi-disciplinary project team for the full scope of the research, UWC (through its Institute for Water Studies and Department of Computing Sciences), CSIR and L2K2 Consultants (Pty) Ltd have joined forces to form the research team led by UWC. Capacity building is a vital component to fulfil the project objectives and the students that contributed to the project included:

- Zaheed Gaffoor (PhD): Assessing the feasibility of Big Data analytics to support local groundwater decision making
- Gift Wanangwa (PhD): A validation study on using BDA in decision-making for managing transboundary alluvial aquifer: A case study of the Shire Valley aquifer in Malawi
- Lutho Ntantiso (MSc): Federated Machine Learning for Modelling Transboundary Datasets in the Southern African Region

1.6. Report structure

Chapter 1 provides the introduction and background to the study. This chapter provides a summary of the transboundary collaboration and an overview of theme interaction and integration. The aims and objectives of the project are discussed, and the project team affiliations introduced with a list of students that contributed to the project.

BDAs and its applicability to support local groundwater management in the Southern African Development Community is discussed in Chapter 2. The chapter reflects on the current knowledge regarding the application of BDAs in the groundwater sciences. BDAs contribution to groundwater management was found to be two-fold. Firstly, BDAs can address issues of data scarcity by consolidating data available from different sources, both traditional and unconventional. Secondly, BDAs can transform data into usable information that can support groundwater management, especially at a local scale.

The methodology for the research included case study area selection, scenario development, federated cloud storage infrastructure development and downscaling regional to local scale in TBAs using BDAs tools. This is discussed in Chapter 3 which introduces the Transboundary Analytics Framework. Using the Transboundary Aquifer Analytics Framework, a holistic approach is applied that provides integration of the various objectives of this research.

Chapter 4 introduces a multi-layered architecture for big data processing aiming at collecting different types of data in order and applying different BDA to the data to get useful insights which can be used by water researchers and decision makers for the understanding, management, and enhancement of water resources.

To model local data and regional data in an integrated manner, a machine learning approach was applied. This is discussed in Chapter 5. The machine learning model relies on a set of predictor variables (regional scale hydroclimatic variables, e.g. groundwater

storage, precipitation, run-off), to predict a predictant (local scale variable such as depth to water level). This required a significant amount of data to be collected and pre-processed as discussed in Chapter 5.

Chapter 6 describes the set-up and execution of a machine learning algorithm that was used to model groundwater level changes in the case study areas. A gradient boosting decision tree machine learning model was developed and implemented in the case study areas. The model was designed, trained, and calibrated to predict groundwater level changes based on a set of hydroclimatic, land surface and hydrogeological variables.

Chapter 7 provide reflections on the learning and profiling opportunities linked to the project.

2. Literature review

2.1. Introduction

This chapter is based on a paper that was published in Water (Gaffoor et al., 2020). BDA describes the use of advanced and traditional analytical techniques to leverage vast quantities of heterogeneous data, in order to provide valuable insights that can be used to propel optimization, development and knowledge discovery (Kitchin and McArdle, 2016; Adamala, 2017). To date, the surge of data from online social media activities, internet activities, business transactions, scientific missions, digitization, and sensor technologies, among many others, benefit many industries in understanding their operational environment. Collectively these data are referred to as big data.

The earth sciences discipline, like many other scientific disciplines, has been driven into the big data era with the advancement of sensor technologies, such as remote sensing, that continually collect new data (Guo, 2017). This has paved the way for the introduction of data-driven approaches in the earth science discipline. It is not a surprise that in recent times the potential for big data to support knowledge discovery in the hydrogeological discipline has become apparent (Adamala, 2017). For example:

- Chalh et al. (2015) showcased the use of the big data open platform to support water resource management in the Foum Tillich watershed, Morocco. The platform utilizes several tools such as stochastic models, simulations, hydraulic and hydrological models, high performance computing, grid computing, decision support tools, big data analysis systems, communication and diffusion systems, database management, geographic information system (GIS) and knowledge-based expert systems to extract information from a variety of heterogeneous datasets. Through decision support tools such as hypsometrical approach, users can understand the impacts of various future management scenarios
- Lee et al. (2019) demonstrated the potential of BDAs to mapping groundwater potential in Goyang-si, South Korea, by combining data from borehole-pumping activities and satellite-based earth observation data

In fact, recent interest in BDAs has spurred a special section in Water Resources Research focusing entirely on the application of BDAs in hydrological research (Water Resources Research, 2020). Nonetheless, applications of big data are still very incipient

in the discipline of hydrogeology. As such, the range of applicability of big data in the hydrogeological field has not been fully explored. The aim of this chapter is to highlight the potential role BDAs and big data can play in supporting groundwater management in SADC.

2.2. Big data: concepts and role in groundwater science

2.2.1. Defining big data

Big data are referred to as collections of very huge datasets with a great diversity of types that makes it difficult to be collected, stored and analysed by conventional tools and techniques (Chen et al., 2014; Ylijoki and Porras, 2016). Big data have a few characteristics that separate them from generally large datasets. These characteristics are recognized as the Vs of big data (Gandomi and Haider, 2015):

- ***Volume*** – big data consist of enormous quantities of data, generally beyond a threshold of one terabyte, however this change with time, sector, data types and use case
- ***Velocity*** – big data are generated at an exceptionally high rate, such that the volume of big data increases rapidly over time
- ***Variety*** – big data are composed of a variety of different data types from a variety of sources

The three Vs (volume, velocity and variety) are the commonly defined features of big data, which were first coined by (Laney, 2001). Since then, industry experts have added additional Vs to define big data:

- IBM added ***veracity*** – which describes the inherent inaccuracy and uncertainty present in most large datasets and complex datasets (Zikopoulos et al., 2012)
- SAS introduced ***variability & complexity*** – which describe the ever changing nature of big data over time with respect to velocity and variety (Gandomi and Haider, 2015; Lee, 2017)
- Oracle introduced ***value*** as an additional V – which stipulates that big data must contain new knowledge or improve operational efficiency for them to have any meaning in terms of financial investment (Gandomi and Haider, 2015; Lee, 2017). This value is usually achieved using analytics which transforms the raw data into useful information

For SADC groundwater to realize the value of big data, thought must be given to understanding the Vs in the context of groundwater big data in Southern Africa, as well as the analytics required to turn these data into useful information for groundwater management.

Big data types play a role in how big data are managed from data to information. They can be broadly categorized into structured and unstructured data (Gandomi and Haider, 2015). Structured data are any type of data that can easily be stored, categorized, and referenced in tabular form. The main tool to store, access and query this type of data is through relational databases, making them easily readable by machines (Lee, 2017). For example, conventional hydrological data generated through in situ monitoring commonly constitute point information that can easily be captured in relational databases and conventional spreadsheets. This is typical of structured data.

On the other hand, text, video, audio, and images are examples of unstructured data. These lack higher structural organization and are not easily stored in relational databases (Lee, 2017). For example, videos of a flooding events or social media posts related to various aspects of water and groundwater, constitute unstructured data relevant to groundwater. In addition, remote-sensing images constitute unstructured data, but the meta-data attached to the image is structured (Guo, 2017; Wang et al., 2019). Unstructured data are particularly difficult for machine programs to extract information from, at least with traditional techniques. Semi-structured data have some form of structure; however, these tend to be very irregular and often heterogeneous, which makes categorization challenging. Emails and XML files fall into the semi-structured data type (Gandomi and Haider, 2015; Lee, 2017; Lin et al., 2018).

2.2.2. Sources and nature of big data in groundwater sciences

A common awareness among data scientists is that not all big data are the same and that the structure and nature of big data and how we analyse them depend on the domain (Guo, 2017). For example, geospatial data differ from text data (such as from social media posts) and the techniques and tools used to collect, store and analyse each of these types of data will be different (Chen et al., 2014). The result is that one needs to fully understand the specificities of the relevant data sources and what information is required from these data before appropriate big data tools, techniques and analytics can be applied.

Data in the groundwater domain has not been static. Over the years, groundwater scientists have explored various sources to collect groundwater data. Table 2 illustrates these sources of data relevant to groundwater. Table 2 includes the traditional sources of groundwater data such as in situ observations or hydrogeological maps, as well as modern data sources such as remote sensing, social media, or Internet of things (IoT). Individually, some of these sources may not have the characteristics of big data, but when harnessed together they provide some substantial opportunities for knowledge discovery. Large scale data assimilation models are one example of such systems that incorporate data from different sources, such as field activities, remote sensing, and computer simulations. However, at the moment they do not ingest data from unconventional big data sources, such as social media (Zhang and Moore, 2014).

Table 2: Sources of data in the groundwater domain from a big data context.

Source	Description	Characteristics
Field activities	Data generated from field activities such as monitoring, drilling, and pumping activities	Structured data format Limited coverage (spatially and temporally) Local
Historical	Legacy reports, maps, and documents	Unstructured Local or regional Text or images
Remote sensing	Satellite, airborne or groundwater-based earth observation	Unstructured and structured Multidimensional Voluminous Regional
Computer simulation	Data generated through computer-based models	Unstructured and structured Multidimensional Voluminous Regional
Social media and the web	Data available on webpages and social media post	Unstructured Textual, images, videos, or audio Multidimensional Heterogeneous Voluminous Local
Internet of Things	Data available from connected devices	Unstructured and structured Heterogeneous Multidimensional Local

2.3. Methods in Big Data Analytics

The value of big data is truly realized when it is transformed into useful information. BDA covers a comprehensive package of advanced analytical, statistical, mathematical and graphic methods that can be used to transform the data into useful information (Russom, 2011).

According to (Russom, 2011), BDAs is advanced analytics operating on big data. Many of the tools and techniques employed in BDAs, such as machine learning, have been available for many years (Watson, 2014). It is only recently, with the surge in big data, that the value of these advanced analytical techniques has been realized. Compared to traditional analytics approaches, advanced analytical techniques perform well when dealing with very large, heterogeneous datasets, requiring less data pre-processing, as shown in Table 3 (Tsai et al., 2015). For example, machine learning can work on both structured and unstructured data, while traditional analytics works well only on structured data. One of the major differences between traditional analytics and BDAs is the processing platforms required. Big data generally require parallel processing methods to effectively analyse these large datasets. BDAs methods are designed to operate over multiple distributed processors, whereas traditional analytics methods are generally designed to operate on single machines (Tsai et al., 2015). Traditional analytical methods are only efficient when significant sampling and dimensional reduction methods (e.g. principal component analysis, genetic algorithm) are used to reduce data size. In addition, traditional analytics are not suited for parallel processing frameworks. BDAs together with traditional analytics may allow us to leverage various sources and types of groundwater big data, turning them into useful information for a groundwater manager to use.

Table 3: Traditional analytics vs BDAs

	Traditional Analytics	**Big Data Analytics**
Focus	Descriptive analytics and diagnosis analytics	Predictive analysis and prescriptive analytics
Datasets	Limited datasets with structured data. Adoption of simple data models	Large scale datasets with more types of data. Adoption of complex data models
Analysis	Looks to what happened and why?	Provides new insights and forecasts
Processing	Generally capable of being run on a single machine (centralized processing)	Requires parallel processing across multiple machines (distributed processing)

Source: adapted from (Tsai et al., 2015; Almeida, 2018)

Generally, BDAs include traditional analytics such as data mining, statistical analysis, SQL queries (Structured Query Language queries) and data visualization, which work well on structured data. Advanced analytical techniques such as natural language processing, text analytics, video analytics, audio analytics, artificial intelligence and machine learning work well with heterogeneous unstructured data (Russom, 2011; Gandomi and Haider, 2015). An assemblage of these techniques is usually used to turn raw big data into information. For example, in shale analytics, a combination of data mining, machine learning, artificial intelligence, correlation analysis and pattern recognition is used to extract information from text reports, sensor data and geophysical surveys from thousands of existing well operations. This information is then used to predict the success of new well operations (Mohaghegh et al., 2017). In this case, the combination of analytics is uniquely designed to extract value from the types of data present in shale gas operations. To leverage big data in groundwater in SADC, a similar set of unique analytical operations is needed to extract information from the types of data expected. It is also important to note that the type of analytics required should address the problem being investigated.

The spectrum of BDA techniques is vast and an explanation of all these techniques is beyond the scope of this study. However, understanding the role various BDAs play in deriving information from data are key to derive the knowledge required to improve decision-making. For example, Table 4 presents a summary of common BDA techniques and the typical methods they include. These techniques can be used for a myriad of tasks such as extracting information from text data (text analytics), video files (video analytics) and audio data (audio analytics) and even geospatial data (Gandomi and Haider, 2015). Hence, data collected from citizen science initiatives, remote-sensing data, social media

data and conventional hydrological data can be turned into useful information for advancing understanding in groundwater management.

Generally, the role of BDAs is to understand historical events or observations (descriptive analytics), what will occur based on historical observation (predictive analytics) and what is the best solution under uncertainty (prescriptive analytics) (Sun and Huo, 2019). Translating this to a groundwater context allows us to understand what the fundamental interrelation and operation of various hydrogeological processes are based on current data (descriptive analytics), using this knowledge to predict future groundwater scenarios (predictive analytics) and then understanding what the best actions are going forward (prescriptive analytics). This is where the paradigm shifts towards emphasis on data-driven solutions, allowing our analysis to be prescribed by trends in the data rather than theory.

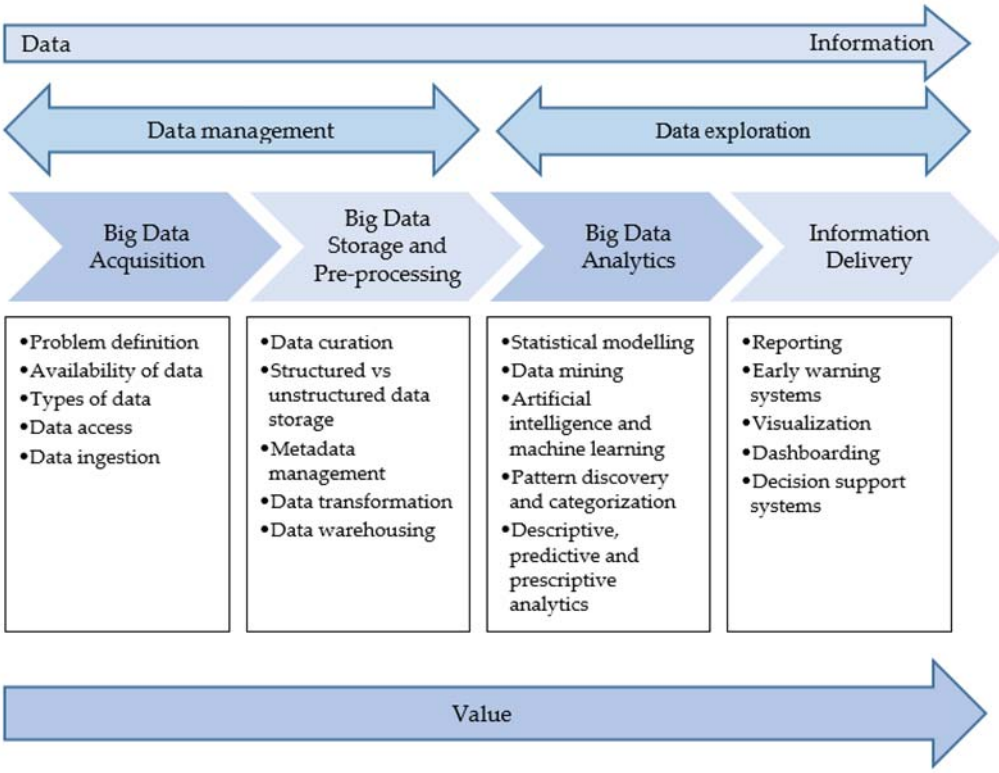
Table 4: Summary of BDA techniques

Techniques	Description	Examples of Computational Methods
Statistics	Collection, organization, and interpretation data	Descriptive statistics, regression, correlation, factor analysis, clustering, hypothesis testing, probabilistic statistics
Data mining	The process of extracting new information, such as patterns, from large datasets	SQL queries, machine-learning, statistics, feature selection
Artificial intelligence (AI)	The role of developing computer systems that imitate, amplify, and automate intelligent behaviour of human beings	Statistical learning, optimization methods, deep learning
Machine learning	Subset of AI, concerned with using self-learning computer algorithms to recognize features in empirical data	Artificial neural networks, support vector machine, random forest, k-means clustering, natural language processing
Uncertainty analysis	Techniques used to quantify and handle uncertainty in big data	Data cleaning, probability theory, Bayesian theory, Shannon's entropy, rough set theory, fuzzy set theory
Visualization	The use of graphic means to represent large datasets	Tables, graphs, images, feature extraction, geometric modelling

Source: (Gandomi and Haider, 2015; Ali et al., 2016; Hariri et al., 2019)

2.4. Big Data Analytics frameworks and platforms

Big data platforms are enterprise scale solutions used to facilitate the use of big data to meet a specific industry need. They are generally a collection of hardware and software layers, built upon a specific big data processing framework (Chapter 4). The function of modern big data platforms is to leverage big data. This is achieved through a process of data acquisition, data storage and pre-processing, data transformation through analytics and information dissemination (Becker et al., 2016). Figure 4 illustrates a general reference framework for big data, which includes the typical features or components required for any big data platform.



Source: adapted from Jony et al. (2016); Faroukhi et al. (2020)

Figure 4: BDAs value chain

Data acquisition revolves around connecting to relevant data sources, determine individual data products and ingestion mechanisms. Here, one must consider the type of data being collected (e.g. structured versus unstructured), access and usage protocols for the various sources, the volumes of data required (which influences how data will be transmitted from the source to the processing location) and meta-data generation (Nasser and Tariq, 2015). For example, the size of some data products makes it

impractical to retrieve data from the sources repeatedly for analytical queries. In this case, it may be more advantages to ingest entire datasets and store on local systems. The complexities associated with data collection make the data itself an important component of any big data platform.

Data pre-processing focuses on addressing the quality and uncertainty in the data, as well as the conversion of unstructured data to structured data. The purpose of this component is to create analysis-ready datasets. In this step, one must consider the type of data required for analytical operations, data cleaning protocols that are necessary, the uncertainty of the data and the post-processing algorithms that can be applied to improve accuracy in the raw data. The caveats (i.e. limitations and inaccuracies) of individual datasets will be important in this step (Padgavankar and Gupta, 2014). Once the data have been pre-processed, then data storage can take place. This requires knowledge on how data are to be curated, the type of data being stored (i.e. structured, or unstructured), the processing environment required, meta-data and the indexing paradigm. For example, in the Earth Science domain data will most certainly be geospatial in nature, indexing the data along temporal and spatial dimension would support faster and more versatile analytical operations (Alarabi et al., 2018).

Figure 4 also illustrates how the value of big data increases across the value chain. BDA plays an important role in the value chain, leveraging big data in driving the knowledge discovery process, as we move from raw data to useful information. In this component, many of the analytical methods described in Section 2.3 will be useful. However, developing data-driven modelling through machine learning and artificial intelligence is perhaps the current status quo in terms of extracting value from the data. Descriptive, predictive, and prescriptive analytical models, if feasible, can provide additional tools to support groundwater management. For example, descriptive and predictive models may allow simulation of current and future groundwater conditions, while prescriptive models may allow determination of the impact of various management decisions. Finally, usable information must be disseminated in the form of maps, figures, and tables (etc.). This information can be usable as it is or it can be incorporated into decision support systems, early warning systems or dashboards to facilitate decisions (Figure 4).

Addressing some of the challenges facing groundwater management in SADC may require a holistic solution such as a big data platform. For example, the disparate nature of groundwater big data could be centralized, the application of analytics could be simplified with built-in methods and functions, and the information could easily be accessed through web-based services. Hence, big data frameworks and platforms that

can be used to implement a big data approach in support of sustainable groundwater management in SADC are reviewed below.

Many of the data sources described in section 2.2.2 house their data in large data warehouses or centres, which are distributed across the globe. These data centres can be accessed through various web-based platforms, such as Earth Explorer[3], EarthData[4], ESA Earth Online[5]. For example, most data generated by NASA missions get stored in distributed active archive centres across the United States, which can be accessed through various web-based platforms and software (Blumenfeld, 2018). However, navigating, extracting and processing vast amounts of remote-sensing data from various data sources to apply to a specific objective, such as to support groundwater management in SADC region, can be technically challenging (Cui et al., 2018). Often, specialist skills and tools are required to properly integrate and use the vast volumes of groundwater big data available.

To address some of these challenges, many agencies have developed special platforms that can be used to leverage these big data. The Australian Geoscience Data Cube (AGDC) is an example of a purpose-built big data platform that focuses on leveraging remote-sensing big data, particularly Landsat, for Australian geoscience applications (Lewis et al., 2017). Hence, the platforms, data collection, storage and analysis features are tailored toward managing geo-spatial remote-sensing data. For example, data ingestion and pre-processing components focus largely on refining incoming raw data into analysis-ready products before data storage, using standard techniques. Data storage follows a multidimensional data array format with geospatial indexing (Data Cube). The architecture for this system is supported by the National Computational Infrastructure (NCI) Facility and their high-performance computing framework.

EarthServer is a geospatial big data platform that is more generalized and interoperable, by focusing development on open geospatial data standards, such as those provided by the Open Geospatial Consortium (OGC) (Baumann et al., 2016). The platform is supported by the Rasdaman framework, which is an array-based, fully implemented parallel storage and processing platform. The platform allows various front-end applications to be attached for specific use cases.

[3] https://earthexplorer.usgs.gov/

[4] https://earthdata.nasa.gov/

[5] https://earth.esa.int/web/guest/data-access

IBM's physical analytics integrated data repository and services (PAIRS) is another geospatial big data platform (Klein et al., 2015; Lu et al., 2016). Its focus is largely on facilitating and simplifying the collection, integration, pre-processing, storage, retrieval, and analysis of heterogeneous spatial data. Data are collected and pre-processed into analysis-ready products, indexed, and stored along a common geo-spatial grid. Frameworks such as Hadoop and HBase support the storage and processing. Unlike the other platforms that focus on raster data, PAIRS provide facility for unstructured data types such as from IoT and social media. The unstructured data are transformed and stored alongside the raster data.

The Earth System Grid Federation (ESGF) is a multi-agency, international collaboration focusing on the sharing of climate-related data (Cinquini et al., 2014). The design of the ESGF is based on geographical independent data nodes that are built on common infrastructure. The nodes adopt common federation protocols and APIs (Application Programming Interfaces) that facilitate peer-peer communication and transfer of data. Now the ESGF is not an analytics platform, instead focusing on data indexing and data access.

Besides the aforementioned big data platforms, there are a number of big data geospatial frameworks that can be implemented as geospatial big data processing solutions. These include ST-Hadoop (Alarabi et al., 2018), SpatialHadoop (Eldawy and Mokbel, 2015), Hadoop-GIS (Aji et al., 2013), GeoWave (Whitby et al., 2017) and GeoSpark (Yu et al., 2015), among others. These frameworks facilitate the distributed or parallel processing of geospatial big data.

2.5. Challenges in applying Big Data to groundwater management

According to (Sivarajah et al., 2017), there are numerous challenges that are faced by experts when trying to implement BDAs, but these can be divided into three broad categories:

- Data challenges relate to the nature of big data itself (e.g. volume, velocity and variety, etc.)
- Process challenges relate to how to capture, integrate and transform data, how to select the right model for analysis and how to provide the results
- Management challenges cover issues such as privacy, governance, institutionalization, security, among others.

These challenges are further exacerbated by the technological limitations of current information systems (Fan et al., 2014).

Like all other domains, big data in groundwater within SADC region are expected to have considerable volume, velocity, and variety. For example, the data for a 10° × 10° tile from MODIS Evapotranspiration dataset for the SADC region can be as large as 20 GB. Multiplying by additional variables and additional tiles needed to model a groundwater management scenario across the entire SADC region would result in the dataset growing rapidly. The technological requirements to store and process such large heterogeneous volumes of data often require dedicated systems beyond the capabilities of conventional desktop systems (Fan et al., 2014). In this instance, technologies such as parallel processing infrastructure and clustered computing systems have come to the fore (Fan et al., 2014). However, the computational capabilities of many SADC member states may not be advanced enough to facilitate big data approaches. Furthermore, an obvious bottleneck when ingesting huge volumes of data are the high network speed required to move and process big data (Bonner et al., 2017). This requirement is often lacking in less developed African regions and may even be non-existent in rural regions.

Lastly, big data management challenges are experienced within a SADC context, especially when dealing with transboundary aquifers. The transparency of data sharing across international boundaries is not always welcomed by individual states. Data ownership and data access is often restricted to certain individual or institutions and come with many caveats for their use (Pietersen and Beekman, 2016). This is certainly the case when security issues are present with sharing or use of data. The institutional barriers may become a roadblock. Furthermore, management practices employed by member states are not always aligned with each other (Pietersen and Beekman, 2016). The consequence is that the decisions taken based on the data may be contradicting within transboundary aquifers, ultimately affecting the sustainable management of groundwater.

2.6. Conclusion

Groundwater science is generating increasing amounts of data from scientific experiments, sensor arrays, monitoring programs, remote sensing – even social media. Increasing attention is being paid to leveraging these vast volumes of data for new knowledge discovery in groundwater. Improving sustainable groundwater management in SADC is one use case where big data and BDAs may be useful. BDAs contribution to groundwater management can be two-fold. Firstly, BDAs can address issues of data

scarcity by consolidating data available from different sources, both traditional and unconventional. Secondly, BDAs can transform data into usable information that can support groundwater management, especially at a local scale. The consensus in the literature is that BDAs techniques and methods provide benefits beyond traditional analytics, when dealing with large heterogeneous datasets and are particularly useful when performing data-driven modelling. Advanced analytics such as machine learning have shown a promising insight when modelling groundwater processes. However, the choice of data and the choice of analytical techniques to achieve the analysis goal is critical to ensure data integrity and accuracy along the life cycle of the data. Proper management of data and analytical processes is imperative in this case.

3. Methodology

3.1. Introduction

The methodology for our research includes case study area selection, scenario development, federated cloud storage infrastructure development and downscaling regional to local scale in TBAs using BDA tools (Figure 5). The application of BDA tools will facilitate understanding of complex systems that traditional methods cannot reveal or provide insights leading to improved understanding of groundwater resources.

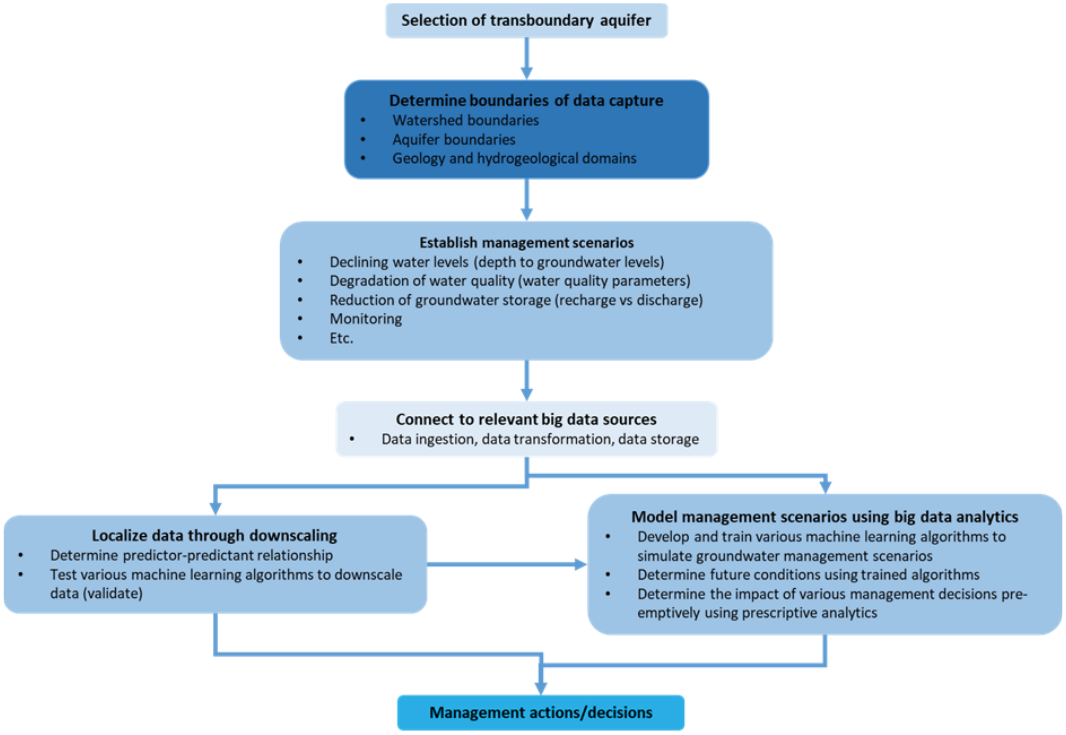


Figure 5: Transboundary aquifer analytics framework

3.1.1. Case study area selection

The Zeerust/Lobatse/Ramotswa Dolomite[6] Basin Aquifer and the Shire Valley Alluvial Aquifer have been selected as case study areas. The karstic dolomite aquifer straddles

[6] During the proposal and scoping phase of the study it was decided to select the Zeerust / Ramotswa /Lobatse dolomite basin aquifer, as the case study area for the application of TBA analytics framework. This choice as influenced by the availability of data compared to other TBAs and the extent of previous works in the aquifer. Following the initial data exploration phase, it was realised that the Ramotswa aquifer was not of great enough areal extent to allow sufficient coverage of remote sensing regional data. This is particularly true for GRACE data, which has a resolution of 110 kilometre (km) (or 1° x 1°). In addition, the inadequate temporal and spatial coverage of the in-situ data record limits the applications of various BDA techniques. In order to address the abovementioned issues, the case study area was expanded to include the dolomite aquifers extending into the North-West and Gauteng provinces of South Africa.

the international border between Botswana and South Africa and extends into the Northwest and Gauteng Provinces of South Africa (Figure 6). Issues related to the dolomite aquifer include decreasing groundwater levels, reduction in groundwater storage, nitrate and faecal pollution, sinkhole formation due to dewatering and disappearance of spring flow (CSIR, 2003; DWAF, 2006; Altchenko et al., 2017; Baqa, 2017; Modisha, 2017; Cobbing, 2018; Cobbing and de Wit, 2018; Pietersen et al., 2018; Nijsten et al., 2018). The Shire Valley Alluvial Aquifer covers parts of Malawi and Mozambique. The alluvial aquifer has high salinity, fluoride and nitrate concentration levels which constitute a significant risk to the health of the consumer (van Weert et al., 2009; Monjerezi et al., 2011; Pavelic et al. ,2012; Grimason et al., 2013a).

3.1.2. Groundwater management scenarios

The status quo of the TBAs was discussed within the context of undesirable results, which include one or more of the following effects (CDWR, 2017; Kiparsky et al., 2017; Niles and Hammond Wagner, 2019):

- Chronic lowering of groundwater levels indicating a significant and unreasonable depletion of supply if continued over the planning and implementation horizon.
- Significant and unreasonable reduction of groundwater storage.
- Significant and unreasonable seawater intrusion.
- Significant and unreasonable degraded water quality, including the migration of contaminant plumes that impair water supplies.
- Significant and unreasonable land subsidence that substantially interferes with surface land uses.
- Depletions of interconnected surface water that have significant and unreasonable adverse impacts on beneficial uses of the surface water.

These are the common groundwater management scenarios, that can be expected when ensuring sustainable management of an aquifer. From this set of indicators, chronic lowering of groundwater levels was selected as a use case scenario to apply the downscaling and machine learning modelling.

3.1.3. Federated cloud storage infrastructure development

The federated cloud storage infrastructure links or federates independent local cloud storages together into a digital pool (Kurze et al., 2011) that enables sharing of computational and storage resources amongst federation members. Through cloud federation, a decentralised network of resources is distributed across multiple platforms, which increases fault tolerance over the network infrastructure. This allows continual operation of the system in the event of failure in some components of the system. A federated cloud storage infrastructure housed within a Big Data architecture has benefits in collecting, storing, transforming, analysing, and disseminating large, complex, multisource, and heterogeneous data sets (Figure 14).

3.1.4. Downscaling and modelling

This downscaling involved selection of potential predictors and predictants based on local groundwater management scenario (chronic lowering of groundwater levels), selection of datasets pertaining to potential predictors, development of machine learning models by training and validation, prediction of relevant data parameters using the best identified and calibrated model.

Figure 6: *Map of the case study areas in brown (The Ramotswa study area includes the dolomites of the Malmani subgroup extending into South Africa)*

3.2. Data Providers

There are two main challenges affecting data collection and storage of groundwater data. The first challenge is that collecting data from field activities is generally sporadic. For example, field monitoring data collection has been limited by the number and distribution of sampling sites having generally decreased over the years. This has manifested into a generally sparsely populated (both temporally and spatially) data records. Secondly, data storage is disparate, and in various formats. For example, some countries store data in centralised databases, while others only store data on spreadsheets or in hardcopy form. These challenges affect data retrieval, sharing and

analysis ultimately affecting groundwater management. At present, the surge of data from online social media activity, internet activity, business transactions, scientific missions, digitization, and sensor technologies, amongst many others, can benefit groundwater management (Figure 3).

The following is a list of ground-based observation datasets included in the analysis:

- **South African Department of Water and Sanitation Hydstra database** is an archive for data collected via the hydrological monitoring networks. It includes data related to surface water gauging stations, rainfall gauging stations and groundwater monitoring stations. Hence it includes both groundwater level and groundwater chemistry data. For the purpose of analysis, the depths to groundwater level data were extracted from the Hydstra database.
- **South African Department of Water and Sanitation National Groundwater Archive (NGA)** is a database for various groundwater-related data and information. This database includes data from drilling activities (e.g. borehole construction details, lithology intersections, water strikes), data related to aquifer properties (e.g. yield, depth to water level, hydrogeochemistry) and various metadata related to boreholes and other groundwater sites (e.g. spring chemistry, geosites name and location), amongst much more. The following data were extracted from NGA database: depth to groundwater level and names and location for geosites in the study area.
- **Ramotswa Information Management System** which is a database and information management platform with hydrological data and research related to the Zeerust/Ramotswa/Lobatse dolomite basin aquifer. It includes data collected from groundwater monitoring stations, surface water data, groundwater geophysical exploration, population demographics, administrative data, and socio-economic data, amongst many others. The following is a list of data used from this database: depth to groundwater level and borehole identifiers.

Remote sensing data have global coverage, higher temporal resolution, and covers a wide set of hydrological parameters, including groundwater (Elbeih, 2015). This makes the use of remote sensing data applications particularly useful in data scarce regions such as the SADC.

Earth observation missions using remote sensing applications, particularly through satellites started in the late 1950s with the launch of the Sputnik 1 satellite

(Tatem et al., 2009). Since then, many more missions have taken place, leading to an array of satellites, which now collect information relevant to studying various components of the hydrological cycle (Table 5). For example, the Gravity Recovery and Climate Experiment (GRACE) has been collecting terrestrial water storage data since 2002 (NASA, 2002). GRACE can indirectly be used to infer groundwater storage changes, making it a useful data source for groundwater investigations in the SADC Region (van der Gun, 2012; Chen and Wang, 2018).

Table 5: Various remote sensing missions collecting hydrological earth observation data

Mission	**Agency**	**Hydrological application**	**Spatial resolution (km)**	**Temporal resolution (day)**
Soil moisture and ocean salinity (SMOS)	European Space Agency (ESA)	Soil moisture	36	3
Global precipitation measurement (GPM)	National Aeronautics and Space Administration (NASA)/ Japan Aerospace Exploration Agency (JAXA)	Precipitation	5	0.125
Soil moisture active and passive (SMAP)	NASA	Soil moisture	36	3
Gravity recovery and climate experiment (GRACE)	NASA	Gravity field (groundwater)	110	30
GRACE-FO	NASA	Gravity field (groundwater)	180	30
Sentinel-1a	ESA	Soil moisture	0.1-0.005	12
Sentinel-1b	ESA	Soil moisture	0.1-0.005	12
Sentinel-2A	ESA	Vegetation/ Land Cover/ Irrigated Area	0.02	10
Sentinel-2B	ESA	Vegetation/ Land Cover/ Irrigated Area	0.02	10
Sentinel-3A	ESA	Vegetation/ Land Cover/ Irrigated Area	0.3	2
Proba-V	ESA	Vegetation/ Land Cover/ Irrigated Area	0.35	2

Mission	Agency	Hydrological application	Spatial resolution (km)	Temporal resolution (day)
Fengyun	China	Land surface Temp/ NDVI/ Soil moisture	25-1000	Daily to monthly
Haiyang serial satellites	China			
Terra/MODIS (moderate-resolution imaging spectroradiometer)	NASA	Evapotranspiration/ Vegetation/ Land Cover/ Irrigated Area	0.250-1	1
Aqua/MODIS (moderate-resolution imaging spectroradiometer)	NASA	Evapotranspiration/ Vegetation/ Land Cover/ Irrigated Area	0.250-1	1
LANDSAT 8	United States Geological Survey (USGS)	Evapotranspiration/ Vegetation/ Land Cover/ Irrigated Area	0.5	1
Tropical Rainfall Measuring Mission (TRMM)	NASA/JAXA	Precipitation (tropics and subtropics)		
ICESat	NASA	Vegetation and topography	various	various
ICESat 2	NASA	Vegetation and topography	various	various
Suomi/VIIRS	NASA/ National Oceanic and Atmospheric Administration (NOAA)	Evapotranspiration	0.5	1
GCOM-W/AMSR2	ESA/JAXA	Soil moisture/ Precipitation	15-50	

These missions generate large volumes of data. For instance, the SMAP missions can generate 458 Gigabytes of data daily (Chen and Wang, 2018). Most of these data is open source and can easily be accessed by the public.

3.3. Case Study overview – Dolomite Aquifer

The Dolomite Aquifer forms a geologically connected and regionally extensive formation of predominantly carbonate and chert sequence of rocks (Figure 7). The carbonate rocks are the main source of groundwater for local populations and irrigation activities in the region.

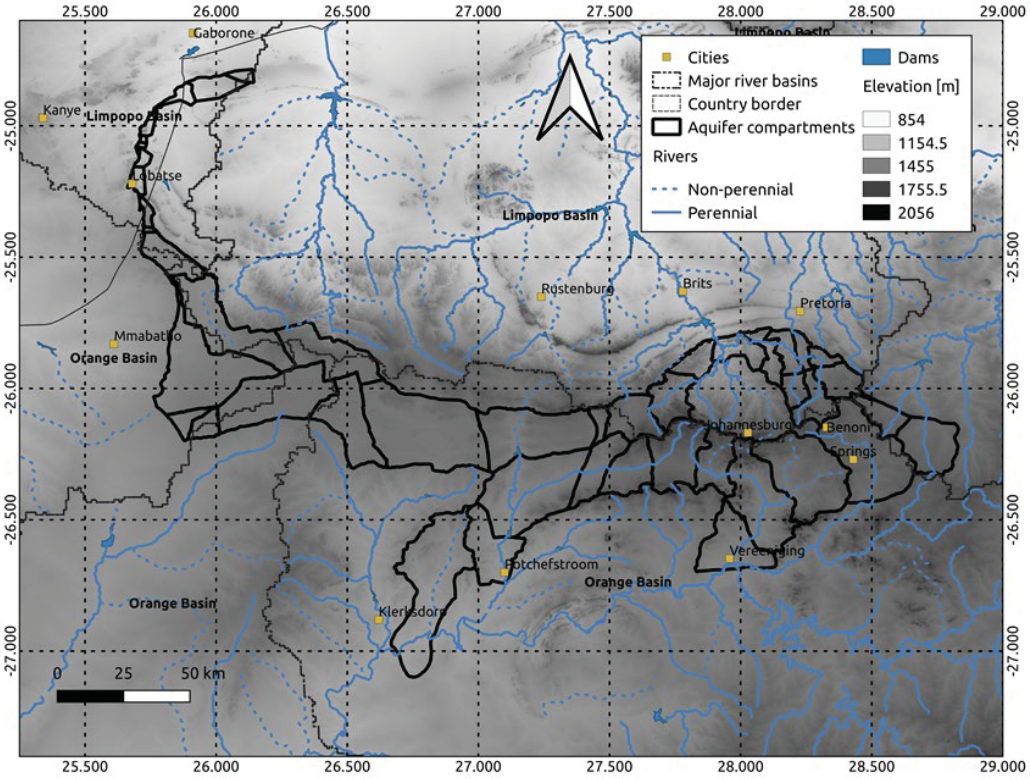


Figure 7: Topographical map of the study area

3.3.1. Hydrology and Topography

The Dolomite Aquifer underlies parts of the Limpopo River Basin in the north sections, and parts of the Orange River Basin to the south (Figure 7). Several perennial and non-perennial surface water drainage systems exist, some of which are important sources of irrigation water (Cobbing et al., 2016). Groundwater drainage emanates as springs, that feed surface water flows, in topographical lows (Wiegmans et al., 2013). There are no regional scale water bodies. The topography can be described as slightly undulating plain, with slight hills and ravines. In some places the land surface is almost flat. There are no major escarpments in the study area.

3.3.2. Geology and Hydrogeology

The study area is underlain by several geological formations (Figure 8). The oldest rocks are composed of basement granitic rocks, possibly of the Kaapvaal Craton, and other Achaean granitic intrusions (CGS, 2008). Following the development of the Kaapvaal Craton, a period of andesitic flows, coupled with erosion and deposition, formed the Dominion Group (Bumby et al., 2012). Overlying this, chronologically, are the thick sedimentary marine and fluvial deposits of the Witwatersrand Supergroup. These

deposits were formed circa 2.98 to 2.78 Ga, in what is believed to be first a passive margin basin, followed by an evolution into a foreland basin, due to orogenic activity (Koglin et al., 2010). Directly overlying the Witwatersrand Supergroup are the thick deposits of volcanic rocks of the Ventersdorp Supergroup (Bumby et al., 2012). Following this are the deposits of the Transvaal Supergroup. The Transvaal Supergroup is composed predominantly of carbonates rocks, in the lower sequences, and sedimentary layers in the upper sequences. It is believed to have formed circa 2,7 Ga, in a shallow marine environment (Manzi et al., 2013; Cobbing et al., 2016). In the study area, these rocks are overlain by the deposits of the Karoo Supergroup. The Karoo supergroups deposits are predominately argillaceous deposits, that formed circa 300 Ma in a retro-arc foreland basin (Cairncross, 2001; Catuneanu et al., 2005). The presence of numerous igneous intrusions, such as the Post-Transvaal Diabase and Karoo Dolerite, as well as large scale tectonics events have significantly deformed the region. There are a number of Quaternary Alluvial and Chert layers that exist as surficial deposits in the study area.

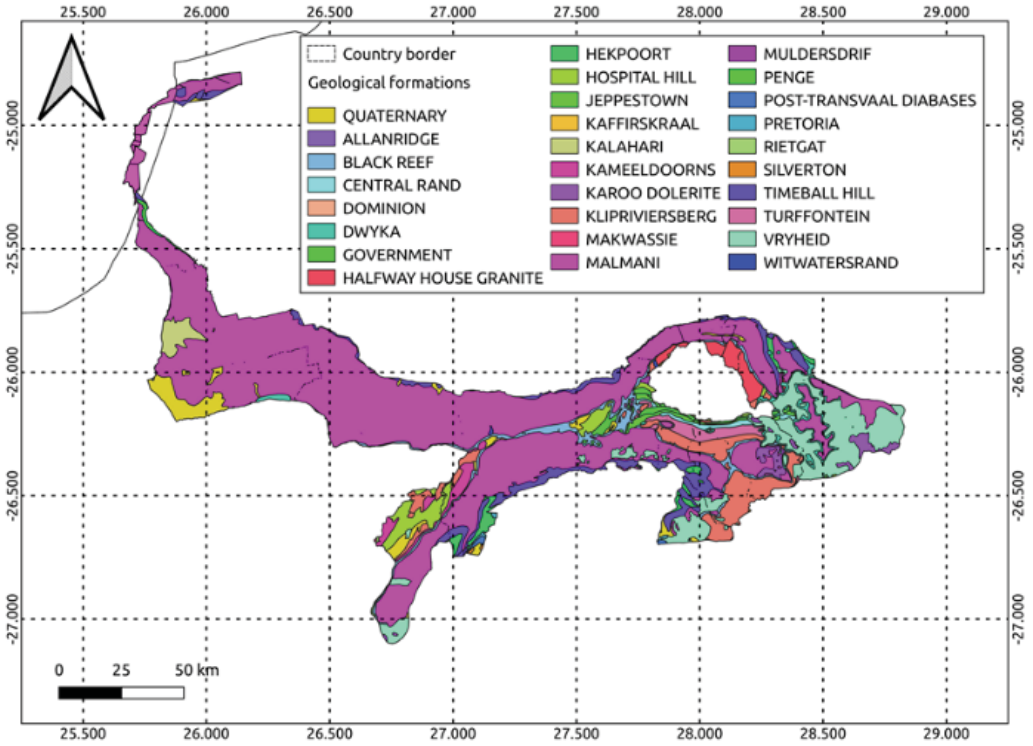


Figure 8: Simplified geology of the study area

The major geological unit outcropping in the study area are the rocks of the Transvaal Supergroup. More specifically, those of the Malmani Subgroup. The Malmani Subgroup presents the most favourable aquifers in the study area. While the rocks of the Witwatersrand and Karoo Supergroups are classified as fractured and low permeability

aquifers, the Malmani subgroup has undergone extensive karstic weathering to expose joints and conduits in the rock as highly permeable sources of groundwater (Cobbing et al., 2016).

However, the presence of dolerite dykes, faults, and geologic contacts, partition the study area into almost distinct aquifer units or “compartments” (Figure 9). According to (Cobbing et al., 2016) a number of studies have identified and classified compartments into resource assessment units, while in the Ramotswa aquifer section, work by (Altchenko et al., 2017), identified a 13 compartments:

- Groundwater Management Area (GMAs): which are based on hydrological drainage boundaries, such as quaternary catchments. These form the larger divisions in the study area.
- Groundwater Management Units (GMUs: which are based on hydrological drainage boundaries, hydrogeological features, such as aquifer boundaries and water levels, and other hydrological features, to identify hydrogeologically connected zones.
- GUAs (or GRUs): which are based on hydrogeological, hydrological and ecological criteria, for the purpose of defining units of analysis (Wiegmans et al., 2013).

These dykes are considered impermeable, however, it is known that groundwater connection through surface and near surface drainage between compartments can occur (Cobbing et al., 2016). This feature can sometimes be expressed as springs that form near compartment boundaries, and that can decant into neighbouring compartments.

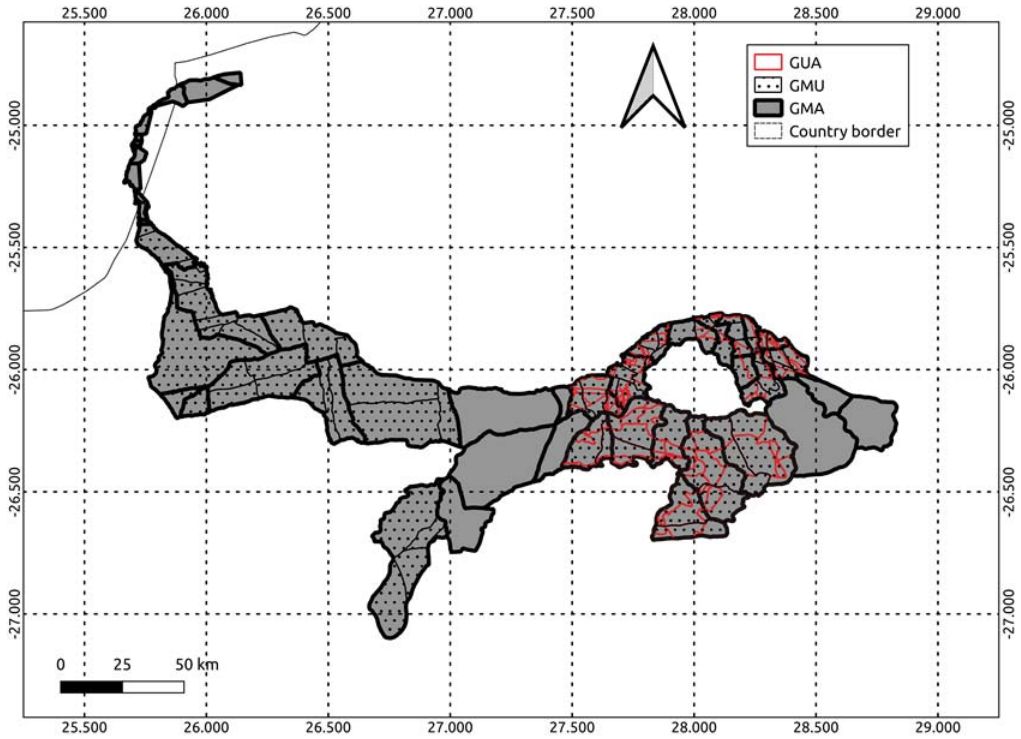


Figure 9: Hydrological compartments of the study area

3.3.3. Groundwater Levels

The study area has extensive groundwater level data records, compared to other regions in Southern Africa. While an extensive analysis of groundwater levels was not undertaken in the study area, there are a number of reports that have highlighted features of the groundwater level (Wiegmans et al., 2013; Cobbing et al., 2016). One of the consequences of the compartmentalization is the large disparity between groundwater levels from adjacent compartments. While the groundwater levels within a compartment are said to be more uniform, and flat. In most cases the groundwater level follows the topography (Cobbing et al., 2016).

3.3.4. Groundwater Use

On the South African side, the dolomitic aquifers of the North West are one of the most utilized aquifers in the country. They provide water for the region's agricultural activities, as well as providing fresh water to several towns in the region. In addition, The Molopo Eye, which is a high yielding spring, is part of large scale domestic water supply scheme for the town of Mafikeng (Cobbing et al., 2016). While, on the Botswana

side there are a number of wellfields that are used for both domestic and industrial fresh water supply.

3.3.5. Groundwater Quality

Limited information is available regarding the regional aspects of groundwater quality in the aquifer. However, (Cobbing, 2018) reported that the dolomites in the North West and Gauteng region are generally classified as having a good water quality, with some localities even being described as having a pristine groundwater quality. However, on the Botswana side, Altchenko et al. (2017) reported serious issues of nitrate contamination related to pit latrines and waste water treatment works.

3.4. Case Study overview – Shire Aquifer

The Shire Valley Alluvial Aquifer (Shire Valley TBA) is a transboundary alluvial aquifer situated on the southern border of Malawi and central Mozambique (Altchenko and Villholth, 2013). The aquifer is situated in the Shire River sub-basin, which is part of the Zambezi River Basin, and has an aerial extent of roughly 5,454 km^2. The Shire Valley Alluvial Aquifer is an important fresh water resource in the region (Chairuca et al., 2019).

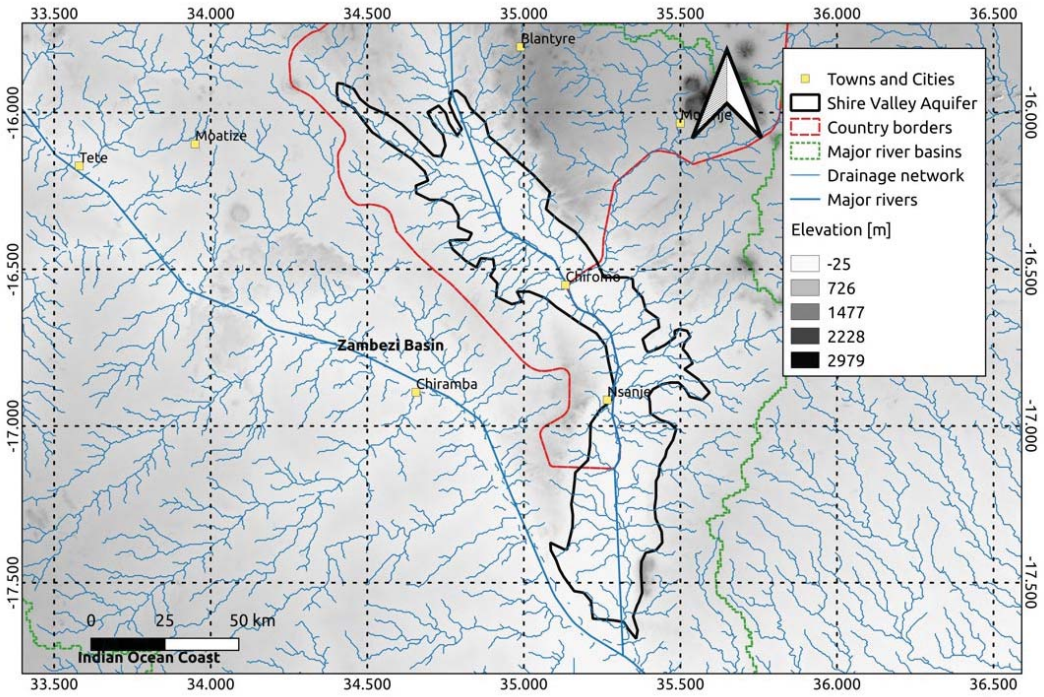


Figure 10: Topographical map of the Shire Valley TBA

3.4.1. Hydrology and Topography

The Shire Valley TBA is situated in a topographical low, relative to surrounding areas. As the name suggest, this can be described as a valley system, bordered on the east and west by prominent escarpments, such as the Thyolo escarpments (Grimason et al., 2013b). In the north-east of Figure 10, the topography increases in elevation, forming distinctive peaks that are almost 3,000 m high. As mentioned, the Shire Valley Alluvial aquifer is part of the Shire River Sub-basin, which is one of the most important contributors of surface run-off to the Zambezi River (Chairuca et al., 2019). The main hydrological feature in the basin includes the Shire River, which originates as the main outflow of Lake Malawi to the north. The Shire River flows south through the Shire Valley, before discharging into the Zambezi River.

3.4.2. Geology and Hydrogeology

Pre-Cambrian geology of the region is primarily concerned with a number of crustal development events, that perhaps are part of the formation of the Gondwana supercontinent (Chairuca et al., 2019). Post-Cambrian can be characterized by the break of Gondwana, which created deep troughs during rifting. These troughs where then filled with Karoo age sediments, which outcrop today (Figure 11) (Chairuca et al., 2019). The main geological influence on the region is the East African Rift System, which is the major north-south trending continental rift system formed during the Miocene (Monjerezi et al., 2011). The continental extension, and crustal faulting caused by the rift system has created a number rift valleys, which are horst and graben style tectonic basins, bordered by uplifted crustal blocks (Wood and Guth, 2020). These features are the major control on Post-Karoo sedimentation in region.

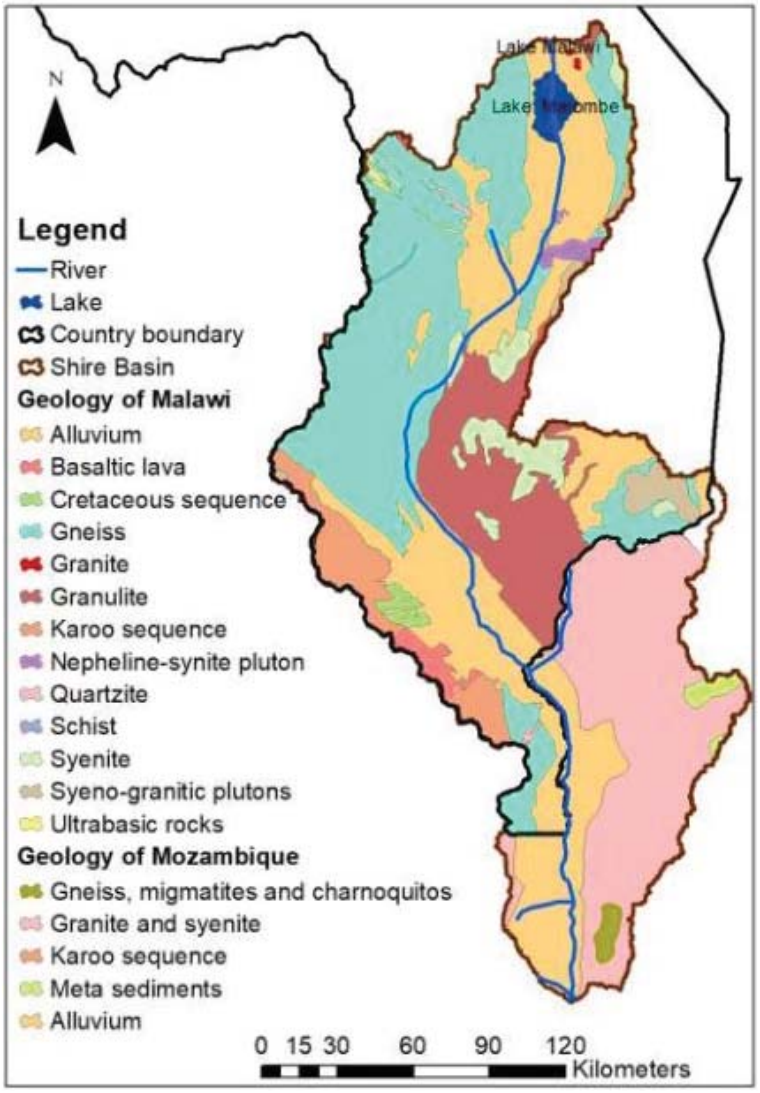


Figure 11: Simplified geology of the Shire River Basin (Chairuca et al., 2019)

The Shire Valley and the Shire Valley Alluvial Aquifer occur at the southernmost tip of the western branch of the East African Rift System. The Shire Valley TBA is composed of a thick sequence of Quaternary aged deposits of unconsolidated, clays, silts, sands and gravels (Chairuca et al., 2019). This can be seen in Figure 11 as the southern outcropping of alluvium. This aquifer can be described as a highly yielding, unconfined aquifer. The thick sediments provide large storage capacity for groundwater. Although rainfall in the Shire Valley TBA is low relative to surrounding regions, the high recharge capacity of the aquifer, make the aquifer a favourable target for groundwater exploration (Chairuca et al., 2019)

3.4.3. Groundwater Levels

Groundwater levels in the aquifer are generally close to the surface. In areas close to surface water bodies, the groundwater level is shallow (5-10 m). However, in general the depth to groundwater ranges from 10-30 m. The limited time-series datasets available for the aquifer prohibits an extensive review of the groundwater levels. However, there are situations of significant groundwater level decline in parts of the aquifer (Chairuca et al., 2019).

3.4.4. Groundwater Use

Groundwater is the main fresh water source for majority of the rural population in the Shire Valley Basin. Groundwater is also extensively used for agricultural watering (Chairuca et al., 2019). Groundwater extraction through shallow boreholes and open wells occurs extensively throughout the valley, while deeper boreholes are becoming more prevalent in recent times. Although handpumps are the main mechanism for groundwater extraction, motorised pumps are employed by farmers.

3.4.5. Groundwater Quality

In the Shire Valley TBA, the groundwater quality suffers from a large degree of salinization (Monjerezi and Ngongondo, 2012; Grimason et al., 2013b). In some parts of the aquifer the salinity has rendered the groundwater unusable. While this is true for parts of the aquifer on the Malawi side, on the Mozambique side much fresher groundwater quality has been reported (Chairuca et al., 2019). The exact extent of the salinization is not fully known, however it is believed to be caused by evaporites forming near the surface, as well as water rock interaction (Monjerezi and Ngongondo, 2012). Besides salinization there are issues for nitrate contamination as a result of informal sanitation practices. In addition, there are further concerns of excessive fluoride, arsenic, iron, and manganese concentration in the groundwater in the aquifer.

3.5. Conclusion

The methodology outlined in this section reflects the current knowledge regarding the application of BDAs in the groundwater sciences. Using the Transboundary Aquifer Analytics Framework, a holistic approach is applied that provides integration of the various objectives on this research. The Framework provides a view of a typical

application from problem identification, through data collection, data analysis and finally improved management decisions. At its essence, the framework is a conceptual understanding of the various components required to match, integrate and model local data with regional data, for the specific purpose of improving groundwater management decisions.

To demonstrate the use of the framework, it is applied to a SADC setting, where two case study areas were chosen. The Dolomitic aquifers of the Ramotswa and NW Gauteng region, provide an opportunity to test the methodology in a karstic aquifer, complicated by geological and hydrogeological discontinuities. While with the Shire Valley TBA, the methodology can be tested in an unconfined unconsolidated aquifer, with limited data coverage on a local scale. Together they can provide valuable insights into the applicable of Big Data and BDAs in groundwater.

4. Big data processing architecture

4.1. Introduction

Transboundary aquifers information systems are assumed to use data from different sources (Figure 3). The processing of the aquifers data will benefit from high performance computing (HPC) infrastructures which are able to store the massive datasets collected from the different groundwater data sources and meet the diverse processing requirements demanded from groundwater management applications. At its onset, HPC was commonly associated with scientific computing for scientific research using supercomputers and computer clusters. Nowadays, HPC has evolved toward the relatively more recent cloud computing model that builds on decades of research in virtualization, distributed computing, utility computing, networking, web and software service to provide mainly three services to users: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). Traditional IT services are delivered on premises and managed by the organization. They include networking, storage, servers, virtualization and operating system, Middleware, Runtime, Data and Applications. As presented in Figure 13, the services which are moved from the organization to be delivered by a third party cloud provider include i) networking, storage, servers, virtualization and operating system for the IaaS model ii) networking, storage, servers, virtualization and operating system, Middleware and Runtime for the PaaS model and iii) all the services including networking, storage, servers, virtualization and operating system, Middleware, Runtime, Data and Applications SaaS model.

Cloud technologies have become the technologies of choice for solving large scale data/compute intensive problems as they present undeniable advantages over traditional HPC. These include through the concept of moving computation to data, a more data-centred approach to parallel computing, and better quality of services provided by these technologies. This evolution led to the emergence of new ecosystems suitable to accommodate Big Data applications' requirements such as i) Cloudera with the Hadoop and Spark systems which are suitable for batch and streaming processing respectively ii) Amazon Ec2, iii) Microsoft Azure, iv) Google App Engine and v) many other emerging systems from niche areas. Some of the advantages of these emerging ecosystems include on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. Many of these ecosystems are equipped with technologies of the fourth industrial revolution, including BDA tools in order to provide decision support to geo-physicists, hydrologists and other stakeholders in water

research projects while meeting i) the economic constraints related to the acquisition of equipment and software and affordability of human expertise, ii) the engineering constraints related to the feasibility of the data processing solutions with existing/future technologies and iii) transboundary constraints associated to country ownership of both data and water resources. However, only a few African countries which are involved in transboundary aquifers research can afford world-class HPC processing infrastructures and building such facilities is often time consuming thus discouraging for multiple deployments.

4.2. A multi-layered architecture

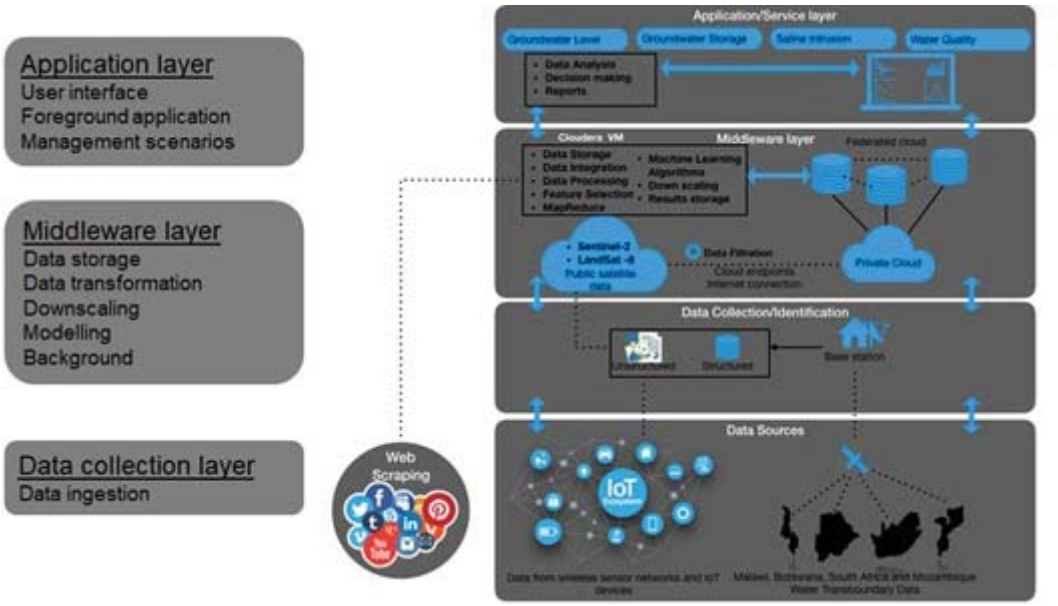


Figure 12: Big data architecture

The different processes which are involved in TBA analytics applications range from information collection through data acquisition, transport of the information close to its processing place and its processing to get useful insights from the information collected, to visualization of the results and utilization of these results in different water-related applications. The implementation of these different processes will follow a multi-layer architecture including four main layers:

– A “data sources” which constitutes the physical layer where all data sources are located. These include data collected from wireless sensor and IoT devices, Web-scraped data and transboundary data collected from satellites, geophysics and other sources.

- A Data collection/Identification layer that consists of a layer that captures all information related to the data sources and processes involved with the acquisition and identification of data sources.
- A middleware layer that consists of an interface between users/consumers of the data and the lower layers that capture, identify and move data close to the user in the middleware layer. It is a layer often referred to as adaptation layer where the information collected from the lower layers is stored and processed to meet the demands of the users' applications for the system. It is at this layer that BDA is performed to get insight into the massive datasets collected from the data sources.
- The application layer is where different applications related to different research questions are defined and implemented users' demands. In groundwater research, some of these applications may involve ground water level reduction, ground water storage reduction, saline water intrusion, water quality degradation and many other groundwater management applications.

The big data infrastructure proposed in this work is an implementation of the big data architecture depicted by Figure 12 where a complete workflow and the different components and processes of such a workflow are revealed.

Two different but complementary models have been considered for this work:

- **Standalone model** (Figure 13) where each organization handles its storage and processing tasks in isolation on its premise if it is endowed with the necessary resources but with possibility of exchanging data and results with other organizations
- **Federated platform** (Figure 14) that involves communication between organization using a cooperative model where data can be stored anywhere and processes can be run anyway in the federation to achieve a common goal.

Two deployment models are proposed for the federated platform: a centralized deployment with CSIR hosting the workflows and data of the different organizations by meeting their storage and processing demands and a hybrid deployment where software containers are used to move the processing to the data residing in different organizations' data repositories. Thus, meeting transboundary collaboration requirements. While for non-constrained data requirements, CSIR plays the role of data storage and processing unit and archiving. While the centralized model with CSIR being the storage and processing infrastructure for all organizations has the advantage of

rapid deployment, the hybrid model implementation is a more complex and time-consuming distributed process that should be phased in different deployment steps to enable a clean and robust implementation with the right security support for both data and groundwater management resources.

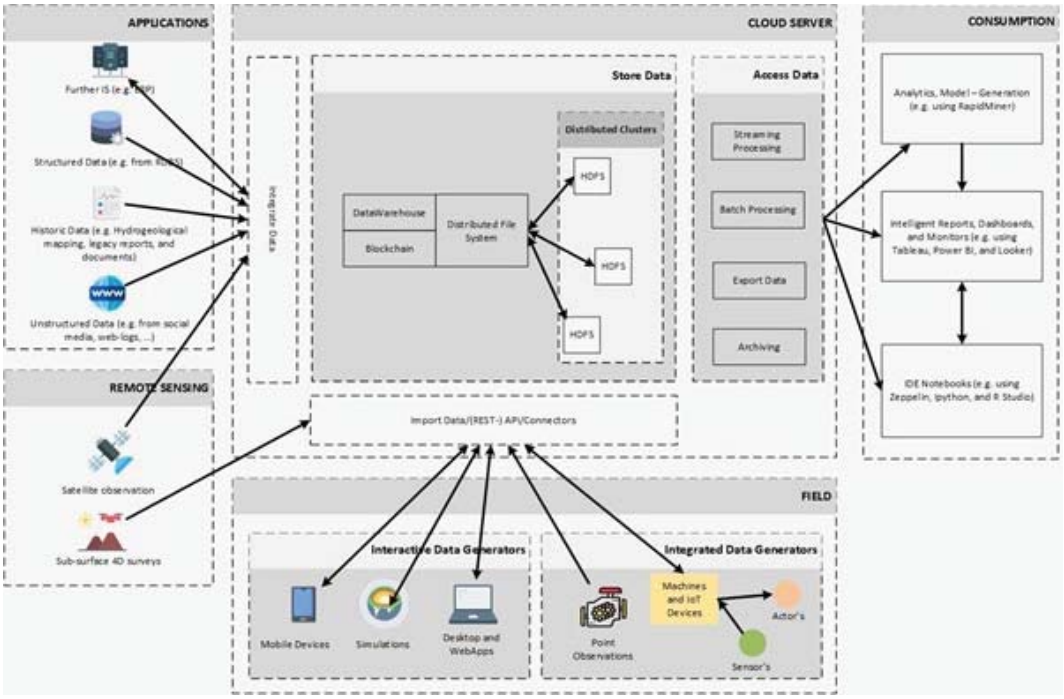


Figure 13: Standalone Model

The architecture depicted by Figure 13 reveals different key components that need to be present in the big data infrastructure:

- Integrated data sources emerging from interactive data generators and integrated data, remote sensing data and data resulting from applications are used as input to the big data infrastructure
- A Datawarehouse consisting of a distributed file system with possibly of using both SQL and NoSQL databases and storing data on a blockchain from where it is read for processing
- A cluster for the distributed file system
- A data processing referred to as data access comprising stream processing, batch processing, archiving and exportation of data
- A data consumption module where different analytics are used to analyse and visualise the data received from the processing module.

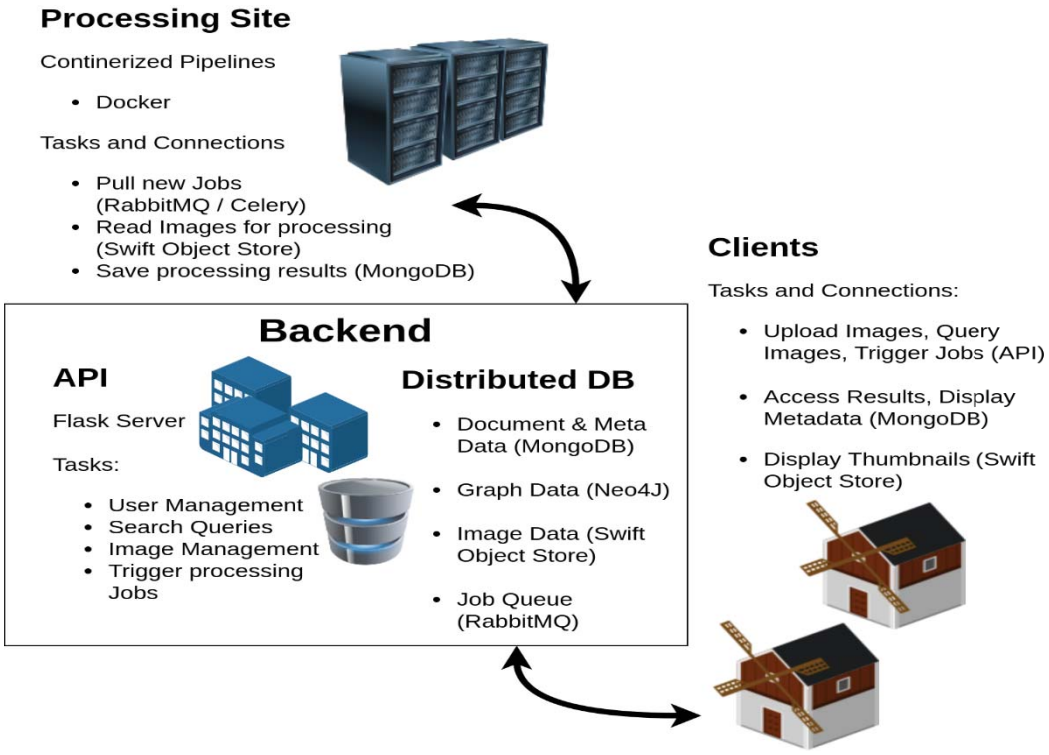


Figure 14: Federated Cloud Infrastructure

Figure 14 depicts the main building blocks of a federated cloud infrastructure called Water-Flow designed for groundwater research data processing. Water-Flow is a built around docker-based microservices architecture where i) each pipeline and element is self-contained and ii) scheduling is performed by routed job queues for more than a million Jobs/sec. it can be deployed as a federated and modular cloud infrastructure with

- Data consisting of satellite images or from other sources are stored on your company server, at home, at one of the big cloud providers or a local data centre you trust.
- Modularity revealed by host components hosted wherever needed and even separately.
- Standardization and expandability expressed by the linkage of any computing task.
- Openness revealed by a system which is 100% Open Source and community focused.

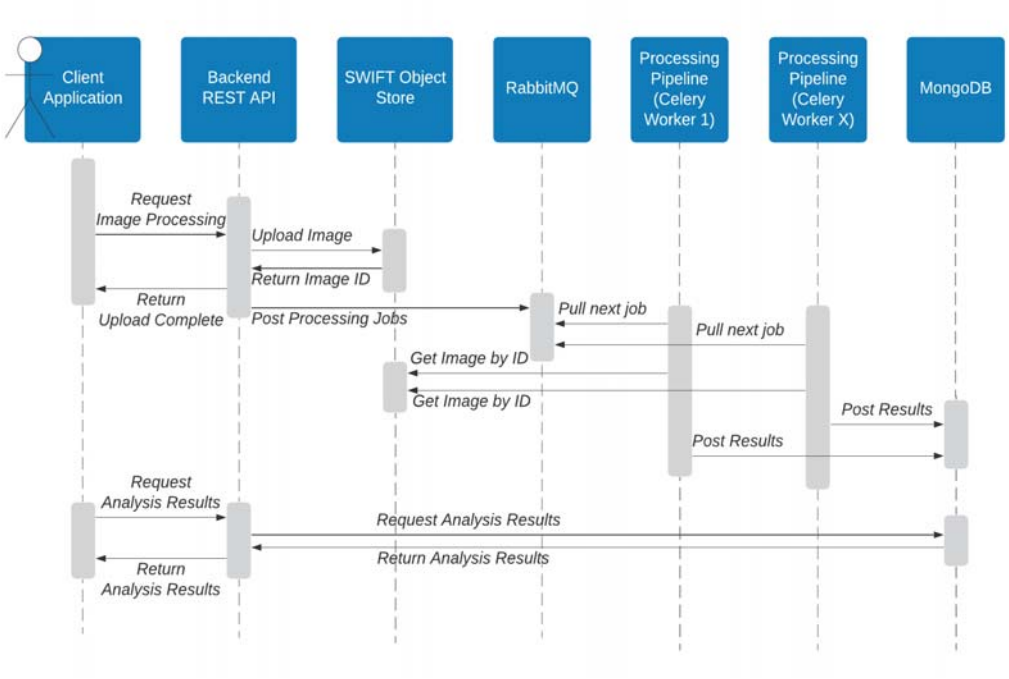


Figure 15: Processing pipeline

Figure 15 depicts Water-Flow processing pipeline that reveals how requests for processing data (here represented by satellite images but can be expanded to other datasets) submitted by client applications are processed in different stations/steps and how the results are sent back to client applications.

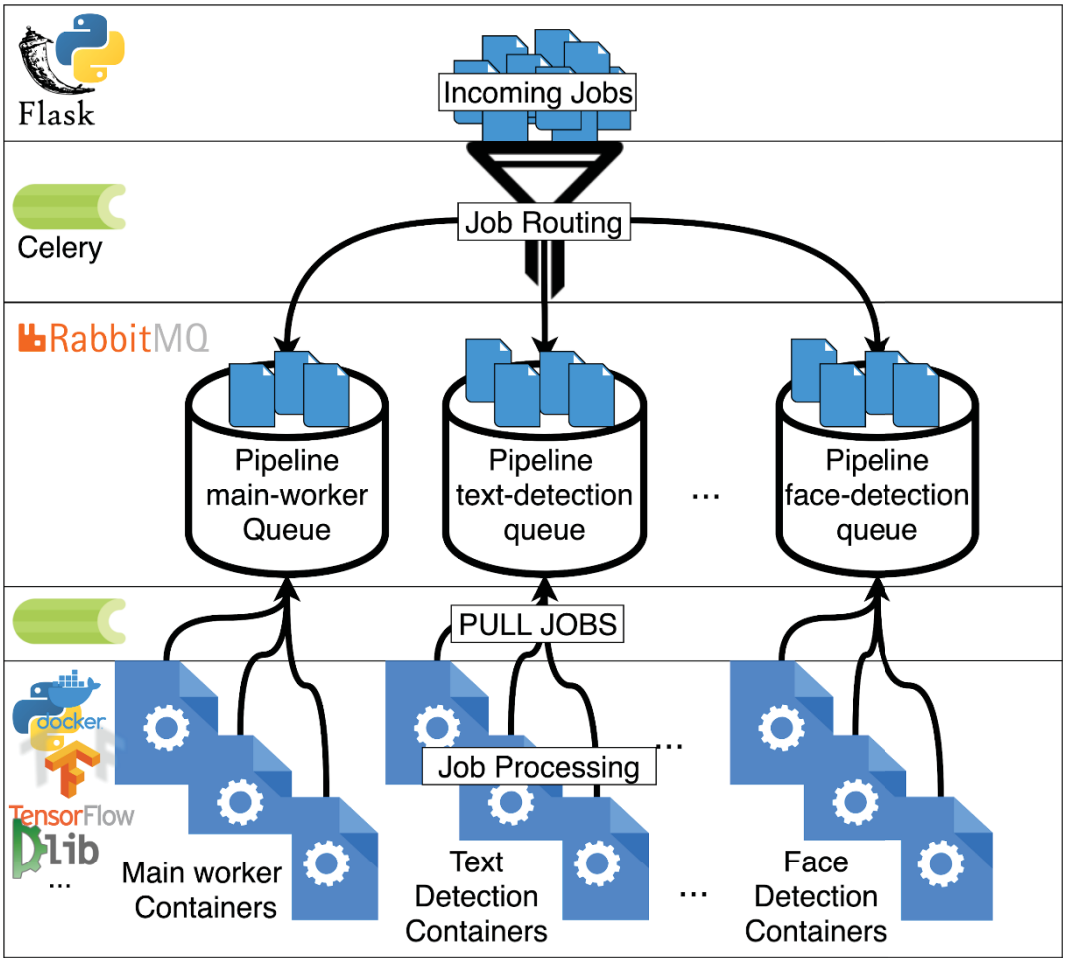


Figure 16: Scheduling capability

Figure 16 reveals the scheduling capability of the proposed Water-Flow platform where incoming jobs are routed to one of the pipelines where they are queued before being pulled by different docker containers for processing. Some of the key features of our processing pipeline include:

- Implementation of a control (management plane) plane which is used to find and allocate processing resources for the client tasks and route these tasks to where they should be processed in the cloud federation
- Implementation of a data plane (processing plane) which is used to manage data and execute tasks on the data.
- Separation of functionalities between the control and data plane by having the control plane managed by the Celery task queue system with the RabbitMQ lightweight communication protocol while the data plane is managed by a combination of a container-based docker environment with the TensorFlow and SciKit-Learn libraries to provide distributed capabilities in the processing of machine learning algorithms

Note that “Celery” is an open-source asynchronous task queue (or job queue) system, which is used in production systems, for instance Instagram, to process millions of tasks every day. It is built around a distributed message passing model used to support task scheduling with focus on real time operations. It is based on execution units, called tasks, which are executed concurrently on one or more worker nodes using multiprocessing, the concurrent networking library for python called eventlet or the coroutine-based python networking library gevent which was inspired by eventlet but features a more consistent API, a simpler implementation and better performance. These libraries enable tasks to be executed asynchronously (in the background) or synchronously (wait until ready). While Celery is written in Python, the protocol can be implemented in any language and operate with other languages using webhooks. Different Celery clients exist for different programming languages/environments including RCelery for Ruby, a PHP client, a Go client, and a Node.js client. While different message brokers are recommended for Celery and different databases are supported, we have adopted for this work RabbitMQ as message broker and MongoDB and CouchDB as database systems for Water-Flow.

4.3. Integration into ilifu and the national big data infrastructure

Ilifu is an infrastructure for data-intensive research partly funded by the Department of Science and Innovation (DSI) with the expectation of enabling South African researchers to be world leaders in the strategic science domains of astronomy and bioinformatics. The project is operated by a consortium of universities and research organisations in the Western and Northern Cape. It is a regional node in the national data infrastructure created to support the National Integrated Cyberinfrastructure System of South Africa (NICIS). The Ilifu partners have further developed scalable systems for cloud-based provisioning of data-centric resources, and have prototyped a tiered, federated cloud infrastructure with consortium partners and external collaborators. The goals of the Ilifu project are to i) provide a new model for provisioning of data-intensive research infrastructure to researchers ii) federate cloud systems to create a common eResearch cyberinfrastructure system and iii) demonstrate cloud-based solutions for strategic projects in astronomy and bioinformatics.

Hydrological research involves massive datasets collected from different sources and requires data intensive processes which can be handled by the Ilifu project’s infrastructure and hence extend the initial mission and goals of the project initially intended only to astronomy and bioinformatics to data intensive hydrological research. The project is an opportunity to pilot such extension by implementing a phased

integration of the models and data produced by different organizations into the Ilifu infrastructure. While some of the organizations might have their own on-premises infrastructure capable of storing and processing all data related to the project, some other research groups might need to migrate all or part of their work to the Ilifu infrastructure following one of the four service migration models:

- "Archiving-as-a-Service (AaaS)" where all IT services which are traditionally managed on-premises are still kept on-premises but data is replicated on the Ilifu infrastructure for archiving and sharing/exchange with other interested groundwater research and management organizations. The services which are traditionally managed on-premises include:
  - Applications
  - Data
  - Runtime
  - Middleware
  - Operating System
  - Virtualization
  - Servers
  - Storage
  - Networking
- SaaS where all the services traditionally managed on-premises are migrated to the Ilifu infrastructure to be delivered as a service to the groundwater research and management organizations. These include:
  - Applications
  - Data
  - Runtime
  - Middleware
  - Operating System
  - Virtualization
  - Servers
  - Storage
  - Networking
- IaaS where some services such as Applications, Data, Runtime and Middleware are still managed by the groundwater research and management organizations while the Operating system, Virtualization Servers, Storage and Networking are migrated to the Ilifu infrastructure to be delivered as a service to the various stakeholders.

- PaaS where many of the services including Runtime, Middleware, Operating System, Virtualization, Servers and Networking are migrated to the Ilifu infrastructure to be delivered as a service to various organizations while only Applications and Data are managed by relevant stakeholders.

The integration proposed above is in line with the main design principles behind the Ilifu project as it is based on i) a similar federation model and ii) data intensive processing principles. To the best of our knowledge, Ilifu is also based on a hybrid container/virtualization HPC principle which are compatible with the proposed big data infrastructure above. This will speed up the integration of the Big Data and Transboundary Water Collaboration research into Ilifu.

The current specifications of the ilifu platform include:

- 110 x compute nodes, 32 CPUs, 256 GB RAM
- 2 x compute nodes, 32 CPUs, 512 GB RAM
- 4 x GPU nodes, 32 CPUs, 256 GB RAM, 2 x Tesla P100 16 GB GPU
- 400 TB BeeGFS storage
- 2.9 PB CephFS storage

Two user interaction and processing environments can be used to access the ilifu resources:

- SLURM batch scheduler
- JupyterLab

SLURM is a batch scheduler and job manager that allows multiple data processing tasks to run using script files. JupyterLab is an interactive coding and processing environment. It operates through the use of notebooks, which can display data, code, as well as information output. The analytical work conducted for this research was conducted using the JupyterLab environment in ilifu. A unique user workspace was provided for our research team, under the pseudonym “Transboundary Water Management”. In addition, a specialised container/kernel (container for software packages) was setup to provide the software requirements for our data analysis. This container is available to all users. Currently, the ilifu platform is only configured to operate via one of the two processing environments. As such additional work will be required to implement a Big Data decision support system that is envisaged by this research.

4.4. Conclusion

The big data infrastructure proposed in this research is based on a multi-layer architecture aiming at collecting different types of data in order and applying different big data analytics to the data to get useful insights which will be used by water researchers and decision makers for the understanding, management and enhancement of water resources. Such an infrastructure could be implemented using either (a) a standalone model with data storage and processing localised in a centralized location or (b) a federated model with data storage and processing distributed in many locations. The ilifu project, which is the current Big Data processing environment hosted in NICIS provides access to a tiered and federated HPC cloud infrastructure. The ilifu platform was utilized to perform the data analysis and modelling in this research, allowing us resources beyond the capabilities of desktop machines. It is expected that with time, the project will migrate to a federated model which provides specific applications in the context of sustainable groundwater management. This will allow more flexibility upon data growth and fits better for a transboundary, multi-organization context.

5. Data processing to match, integrate and model local data with regional data

5.1. Introduction

To model local data and regional data in an integrated manner, a machine learning approach is applied. The machine learning model relies on a set of predictor variables (regional scale hydroclimatic variables, e.g. groundwater storage, precipitation, run-off), to predict a predictant (local scale variable such as depth to water level). The approach used in this case can be used to generate high resolution maps of regional groundwater parameters, such as groundwater level changes, from remote sensed hydro-climatic variables (Seyoum et al., 2019).

All data were acquired from publicly available satellites, modelled, and in-situ based observations. In addition, predictor variable importance was also analysed in this study. The following section describes the data used as well as methods employed to produce high resolution groundwater level maps. These maps provide the information to inspect possible issues of chronic lowering of groundwater levels. In each section details of any pre-processing of data that has been performed is included.

5.2. Data

A set of 9 hydroclimatic parameters are chosen as predictors variables (groundwater storage anomaly, soil moisture, evapotranspiration, precipitation, run-off, land surface temperature, land cover, aquifer type and the aquifer compartments), to predict a single predictant variable (groundwater level changes). These variables are considered major hydrological and hydrogeological components affecting the terrestrial water system in the study area. The variables are aggregated into a number of features used in the machine learning model (section 5.4). In the case of the Shire Valley Aquifer, land cover, aquifer type and aquifer compartments are not included, due to a lack of data, and or being statistically irrelevant features.

5.2.1. GRACE derived terrestrial water storage anomaly

The GRACE Tellus mission represents a breakthrough in our ability to measure and monitor changes in Earth's cryosphere, hydrosphere, and oceanographic components[7]. The GRACE Tellus mission consists of twin satellites measuring changes in Earth's gravity field. GRACE and GRACE-FO level-1 instrument data is fed to three processing centres NASA's Jet Propulsion Laboratory (JPL), GeoforschungsZentrum Potsdam (GFZ), and Center for Space Research at University of Texas (CSR). These three processing centres are responsible for delivering GRACE level-2 and level-3 data products. This includes monthly changes in terrestrial water storage (ΔTWS), based on gravitation field anomalies from land mass changes (Wahr et al., 1998)[8]. Each centre relies on various post-processing algorithms to derive monthly gravity field changes (Level-2) and monthly terrestrial water storage changes (Level-3). The results are three different solutions for GRACE derived terrestrial water storage changes (Sakumura et al., 2014).

Currently there are two major level-1 post-processing derivatives representing terrestrial water storage changes:

- Spherical harmonics-based solutions
- Mass concentration blocks (mascon) based solution

The spherical harmonics version relies on resolving earth gravity field using a set spherical harmonic (Stokes) coefficient at approximately monthly intervals, complete to degree and order 120[8] (Swenson and Wahr, 2006; Swenson et al., 2008). However, the spherical harmonic versions are beset by issues of signal/noise ratio at short wavelengths, and correlated errors, amongst others. This requires additional post-processing adjustment such as applying spatial smoothing filters, as well as the application of de-striping filters. The mascon version on the other hand relies on surface spherical cap mascons to directly estimate mass variations from the inter satellite range-rate measurements (Watkins et al., 2015). The signal loss for the mascon version is considered negligible, hence requiring no post-processing. For this reason, the mascon version is preferred.

[7] https://grace.jpl.nasa.gov/

[8] Spherical harmonics are used to solved geodesic functions on spherical surface. In this case the stokes coefficients and their required settings are used to resolve the anisotropy from the GRACE satellite data.

The native resolution of GRACE data is close to 3°x3°, However, Level-3 GRACE derived terrestrial water storage anomalies are provided as latitude and longitude gridded products at various resolutions. The spherical harmonics version has a resolution of 1°x1°. The mascon version is provided on a global 0.25°x0.25° grid, with ocean signals masked. No optional gain factors need to be applied to these data. In this application we extract latest mascon version, Release 06 version 02 of GRACE and GRACE-FO level-3 monthly terrestrial water storage anomalies, for April 2002-March 2020, from the Center for Space Research at University of Texas[9] (Save et al., 2016).

5.2.2. GLDAS NOAH derived terrestrial water storage anomaly

The GRACE ΔTWS data detailed above include water storage in the entire terrestrial water column – water stored as groundwater, soil moisture, canopy water storage, snow-water storage and surface water bodies (Rodell et al., 2007). To extract the groundwater signal, the various terrestrial water storage components must be removed from the GRACE signal (Rodell et al., 2007). The Global Land Data Assimilation System (GLDAS) in combination with land surface modelling is designed to provide optimal fields of land surface fluxes, through using remote sensing and ground-based observations (Rodell et al., 2004). The GLDAS provide data on various land surface states such as evapotranspiration, soil moisture, land surface energy fluxes to name a few. There are 5 land-surface model derivatives of GLDAS, Noah, CLM, VIC, Mosaic, and Catchment land surface models. Together the various models provide data on land surface states as 1° or 0.25° gridded data products at 3 hourly, daily or monthly intervals from 1948-present.

In this application GLDAS TWS variables: soil moisture (SM), canopy water storage (CW), and snow water equivalent thickness (SWE), were extracted from the model. The data represent monthly averages from April 2002-March 2020 for on 0.25°x0.25° grid[10].

5.2.3. ECMWF ERA5-Land soil moisture data

ERA5-Land is the latest generation reanalysis dataset developed through advanced land surface modelling, by the European Centre for Medium Weather Forecasts (ECMWF). ERA5-Land typically provides data for various land surface states, including single level atmospheric variables such as, precipitation and skin temperature. The ERA5-Land is

[9]http://download.csr.utexas.edu/outgoing/grace/RL06_mascons/CSR_GRACE_GRACEFO_RL06_Mascons_all-corrections_v02.nc

[10] https://disc.gsfc.nasa.gov/datasets/GLDAS_NOAH025_M_2.1/summary?keywords=GLDAS

decoupled from the atmosphere and is run as a single simulation using atmospheric forcing.

The ERA-land component provides data at a 9 km resolution over land surfaces, as global gridded data product. Data extend from 1981 to presents, at an hourly temporal resolution. Soil moisture plays a key role in the terrestrial water budget. Here we extract soil moisture data for a total of 4 layers (0-7 cm, 7-28 cm, 28-100 cm, and 100-289 cm), extending from January 2000-December 2019 at an hourly interval[11]. The units of measure for this parameter are stored as m^3/m^3.

5.2.4. ECMWF ERA5-Land run-off data

Water from falling precipitation, snow melt, or from surface water-groundwater interactions either enters the ground to be stored in soil moisture and groundwater or otherwise drains away to become run-off. Run-off drains away either over the surface (surface-runoff) or as shallow sub-surface flow (sub-surface run-off). The ECMWF ERA5-Land reanalysis dataset provides global run-off data as gridded 9 km data products. Here we extract an aggregated total run-off data (sum of surface and sub-surface run-off), for the study area, from January 2000-December 2019 at an hourly interval[12]. The units of measure for this parameter are stored as m.

5.2.5. ECMWF ERA 5 precipitation data

Precipitation either as liquid water or frozen, generally provide the main flux in-terms of recharge to many aquifers. The ECMWF ERA5-Land reanalysis dataset provides global total precipitation as gridded 9 km data products. Here we extract total precipitation data for the study area, from January 2000-December 2019 at an hourly interval. The units of measure for this parameter are stored as m.

[11] Data downloaded from https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview

[12] Data downloaded from https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview

5.2.6. ECMWF ERA 5 evapotranspiration data

Many hydrological systems lose water through processes of evaporation, or transpiration. The ECMWF ERA5-Land reanalysis dataset provides global gridded 9 km data products on various evaporation and transpiration components of the terrestrial water budget including, evaporation from bare soil, evaporation from open water surfaces, evaporation from the top of canopy, evaporation from vegetation transpiration. Here we extract total evaporation product (evapotranspiration) for the study area, from January 2000-December 2019 at an hourly interval. The units of measure for this parameter are stored as m.

5.2.7. In-situ groundwater level measurements

5.2.7.1. Dolomite Aquifer

In the absence of relatively abundant local or in-situ groundwater storage measurements, in-situ (well-based) groundwater level observations are used as the target (predictant) variable. Groundwater monitoring efforts in the study area have generated groundwater level observations extending back towards 1938 till current. Here were extract a regional set of depth to groundwater level data from the South African National Department of Water and Sanitation's Hydstra database, as well as the Ramotswa Information Management System for the study area. The data include observation from 1938-2019, from a total of 1480 boreholes. Figure 17 displays the distribution of boreholes in the Ramotswa TBA. As can be seen this region benefits with relatively abundant distribution of groundwater level monitoring points compared to other SADC aquifers.

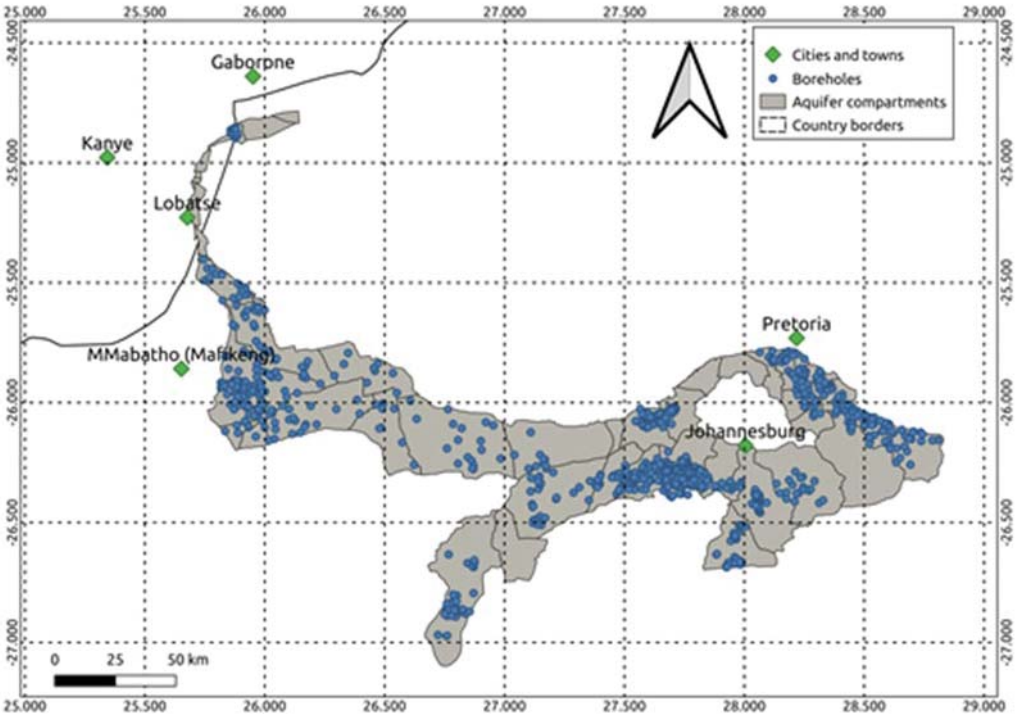


Figure 17: Groundwater level monitoring points within the Dolomite Aquifer

5.2.7.2. Shire Alluvial Aquifer

Figure 18 displays the distribution of water level monitoring points in the Shire Alluvial Valley TBA. In contrast to the Ramotswa there are only 3 boreholes which have sufficient time series data. Additionally, these boreholes are all based on the Malawi side of the aquifer. The time series includes a limited number of observations, starting from 2009 and ending in 2013. Only a single borehole has data that extends until 2019. The data was acquired from the Malawi Ministry of Forestry and Natural Resources.

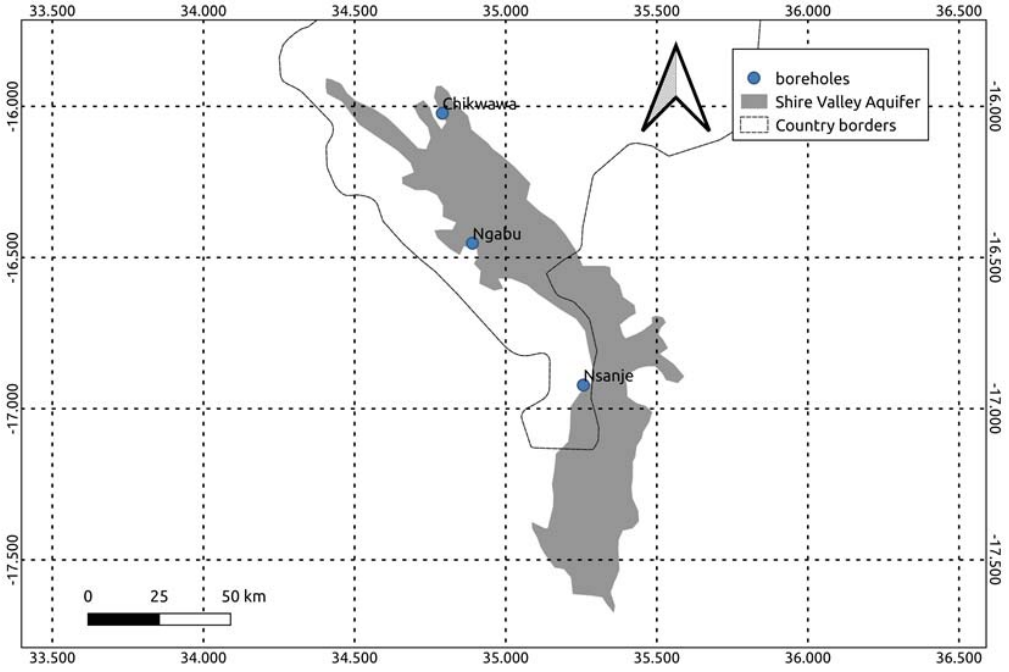


Figure 18: Groundwater level monitoring points within the Shire Valley TBA

5.2.8. Aquifer compartments

One important hydrogeological characteristic of the study area is the partitioning of the aquifer into hydrological and hydrogeological discontinuous compartment, as explained in section 3.3.2. These features play an important part in the hydrodynamics of the aquifer, and so their influence is included in the model. In the case of the Dolomite Aquifers, the GRU compartment feature is removed from further analysis, as spatial representation of this data across the study is limited. It must also be stated that this feature is not relevant to the Shire Valley TBA.

5.2.9. Aquifer type

In addition to the above aquifer compartments, aquifer type is also considered an important feature in understanding the hydrodynamics of the aquifer. The aquifer type data is extracted from the SADC hydrogeology map (SADC, 2010). In the case of the Shire Valley TBA, this feature has not been included. The applicability of isotropic data will have low importance on the prediction in this case.

5.2.10. Land cover

Land cover plays an influential role in the recharge attributes of various aquifers. In this application, land cover data is extracted from the European Space Agency (ESA) Climate Change Initiative (CCI) Land Cover datasets [13]. The data are provided as 300 m resolution, yearly gridded data products. In this case the latest land cover grid (2018), is used to represent the land cover for the study area. In the case of the Shire Valley TBA, this feature has not been included. The applicability of isotropic data will have low importance on the prediction in this case.

5.3. Pre-processing

Several pre-processing steps are necessary to convert the data described above into comparative features that can be related to one another, and that can be used to develop the machine learning model. The following section describes in detail the pre-processing of the raw data. Table 6 summarises the pre-processing outcomes, showing the original raw data and the final processed form. It must be stated that all data were converted to units of centimetres for convenient comparison, where applicable.

Table 6: Parameters and pre-processing results

Data	Original temporal and spatial resolution	Original units	Final temporal and spatial resolution	Final units
Soil moisture	1 hourly, 0,1°x0.1°	m^3/m^3	Daily, 0,1°x0.1°	Centimetres, mean daily
Precipitation	1 hourly, 0,1°x0.1°	metres	Daily, 0,1°x0.1°	Centimetres, cumulative
Run-off	1 hourly, 0,1°x0.1°	metres	Daily, 0,1°x0.1°	Centimetres, cumulative
Evapotranspiration	1 hourly, 0,1°x0.1°	metres	Daily, 0,1°x0.1°	Centimetres, cumulative
Land surface temperature	1 hourly, 0,1°x0.1°	Kelvin	Daily, 0,1°x0.1°	Kelvin, mean daily
GRACE TWS	Monthly, 0,25°x0,25°	Centimetres, terrestrial water storage anomaly	Monthly, 0,1°x0,1°	Centimetres, groundwater storage anomaly

[13] https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-land-cover?tab=overview

Data	Original temporal and spatial resolution	Original units	Final temporal and spatial resolution	Final units
GLDAS TWS	Monthly, 0,25°x0,25°	kg/m^2, various terrestrial water storage variables	Monthly, 0,25°x0,25°	Centimetres, total terrestrial water storage
Groundwater levels	Various, N/A	metres, depth to groundwater	30-day, N/A	Centimetre, groundwater level anomaly
Compartments (Ramotswa only)	N/A	N/A	N/A	N/A
Aquifer type (Ramotswa only)	N/A	N/A	N/A	N/A
Land cover (Ramotswa only)	Yearly, 300x300 m	N/A	N/A	2018, 300x300 m

5.3.1. GRACE data

The following is a description of the pre-processing steps conducted to extract the GRACE ΔTWS into GRACE groundwater storage anomaly (ΔGWS). Due to inconsistencies in satellite data collection, GRACE data typically have a number of missing observations in the time series. There are 216 months in the observation period (2002/04-2020/03), while data exist for only 184 months. Gaps in the data were filled by substituting the monthly mean. Firstly, the observations were grouped according to calendar month, and the mean for each group (calendar month) was calculated. These values were substituted for the corresponding missing months in the time series. Figure 19 displays the net change (i.e. cumulative change) in the GRACE-derived terrestrial water storage anomaly for the Ramotswa TBA. These data suggest a maximum decrease in total water storage for the study area of 92,28 cm, compared to the baseline period (2004-2009). Figure 20 displays the net change (i.e. cumulative change) in the GRACE-derived terrestrial water storage anomaly for the Shire Valley TBA. In this case there appears to be an increase in terrestrial water storage in the southern parts of the aquifer while the northern parts indicate an overall decrease in terrestrial water storage.

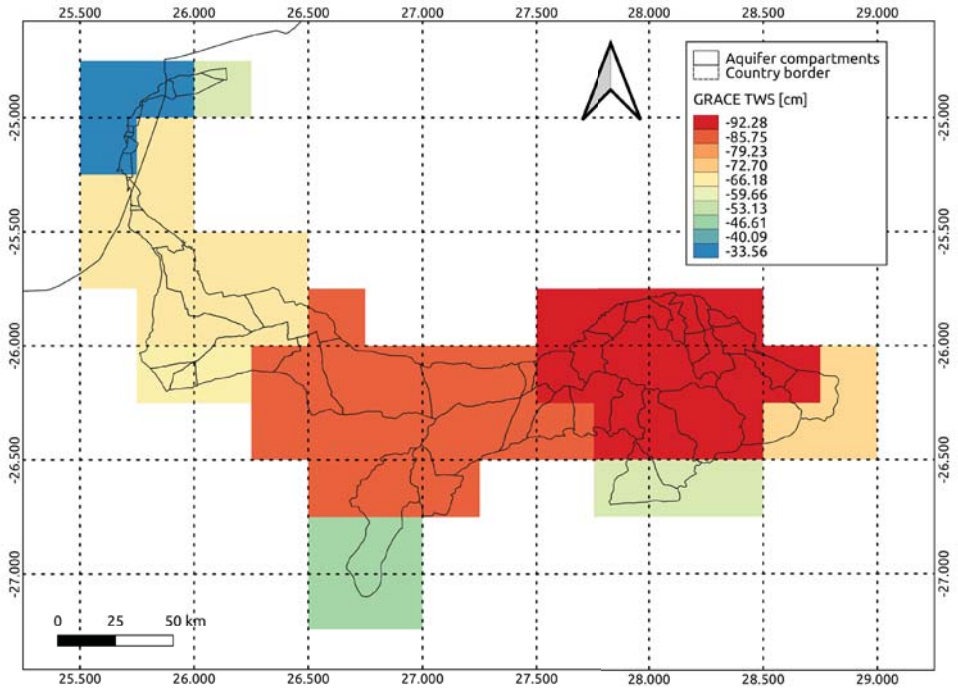


Figure 19: Net GRACE-derived terrestrial water storage anomaly 2002-2020 for the Dolomite aquifer

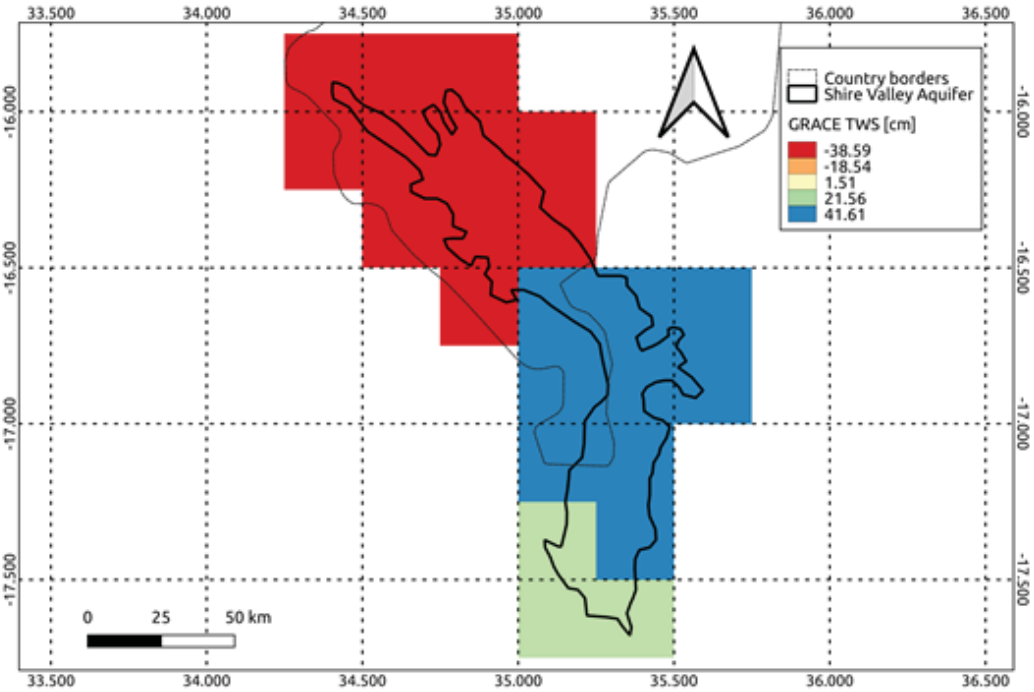


Figure 20: Net GRACE-derived terrestrial water storage anomaly 2002-2020 for the Shire Valley TBA

For the GLDAS TWS, the data are presented in units of kg/m^2. All the units were converted to cm. This is to ensure compatibility to the GRACE ΔTWS mascon units, which is in cm. Thereafter the individual components (SM, SWE, CW) were aggregated, by summation. This value reflects the land surface component of the total terrestrial water

budget. However, GRACE data reflect anomalies relative to a mean baseline period (2004-2009). For the GLDAS TWS data to be compatible to the GRACE data, anomalies must be calculated relative to this same baseline period. Firstly, the mean GLDAS TWS value was calculated for the months between 2004 and 2009. This mean value is then subtracted from each monthly time-step in the GLDAS TWS timeseries. This new value reflects GLDAS ΔTWS relative to the baseline period. Figure 21 displays the net change in the GLDAS-based terrestrial water storage anomaly for the Dolomites aquifer. This suggests a large maximum decrease of 1056.89 cm in surface and near surface water content, compared to the baseline (2004-2009).

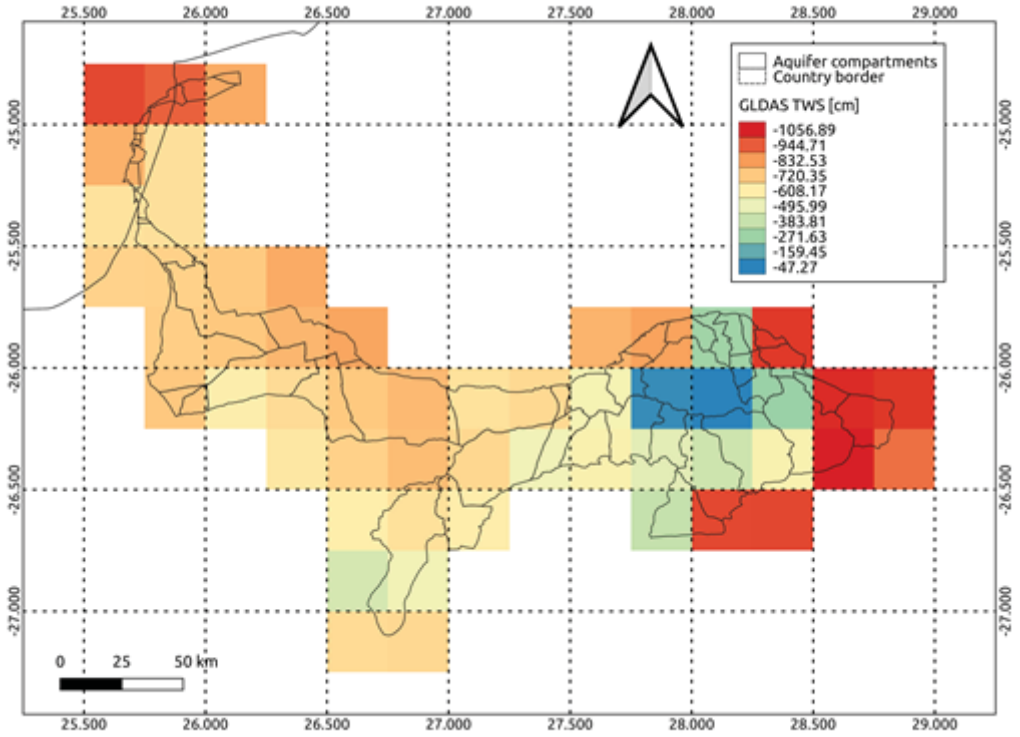


Figure 21: Net GLDAS-based terrestrial water storage anomaly 2002-2020 for the Dolomite aquifer

Figure 22 displays the net change in the GLDAS-based terrestrial water storage anomaly for the Shire Valley TBA. The data indicate an overall decrease in surface and near surface water content, compared to the baseline.

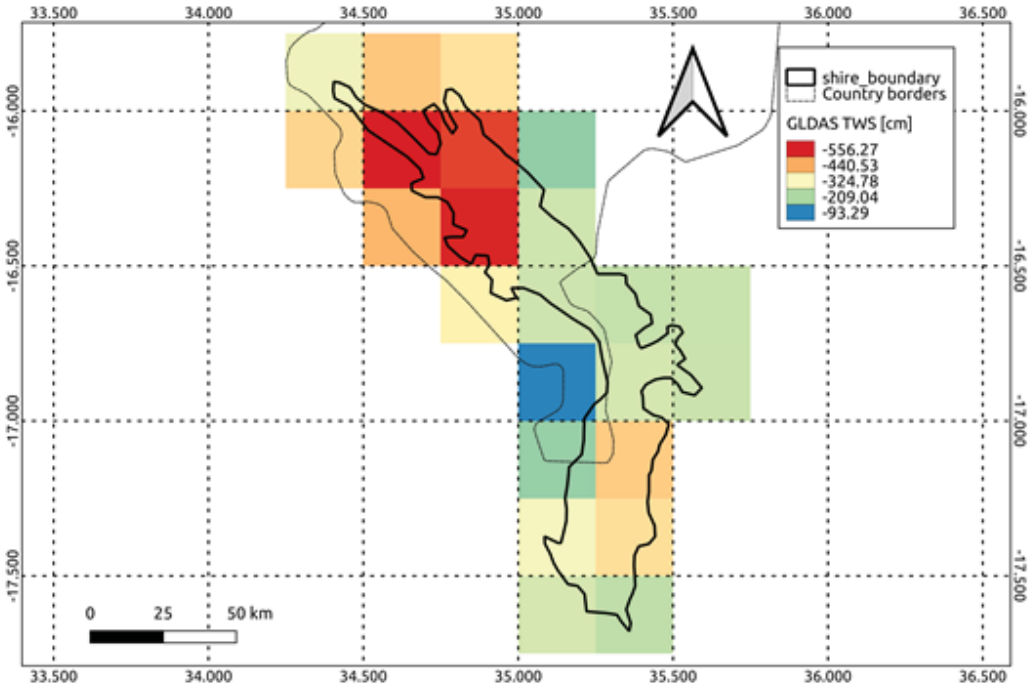


Figure 22: Net GLDAS-based terrestrial water storage anomaly 2002-2020 for the Shire Valley TBA

In order to determine the ΔGWS signal within the GRACE ΔTWS mascon data, the various terrestrial water components must be removed from the model. In this case a water mass balance approach was used (Rodell et al., 2007): ΔGWS = ΔTWS – Δ (SM + SWE + CW). For every time step the corresponding GLDAS ΔTWS is subtracted from the GRACE ΔTWS mascon data. It is important to note that although it is quite reasonable to assume the terrestrial water constitutes most of the GRACE ΔTWS signal, changes in surface water storage and biomass can have an effect (Rodell et al., 2007). However, these components are not included in the model. Figure 23 displays the net GRACE-derived groundwater storage anomaly for the Ramotswa/NW Dolomites. This suggests an overall gain in groundwater storage, compared to the baseline period (2004-2009). While Figure 24 displays the net GRACE-derived groundwater storage anomaly of the Shire Valley TBA. Here there is an overall increase in groundwater storage compared to the baseline period.

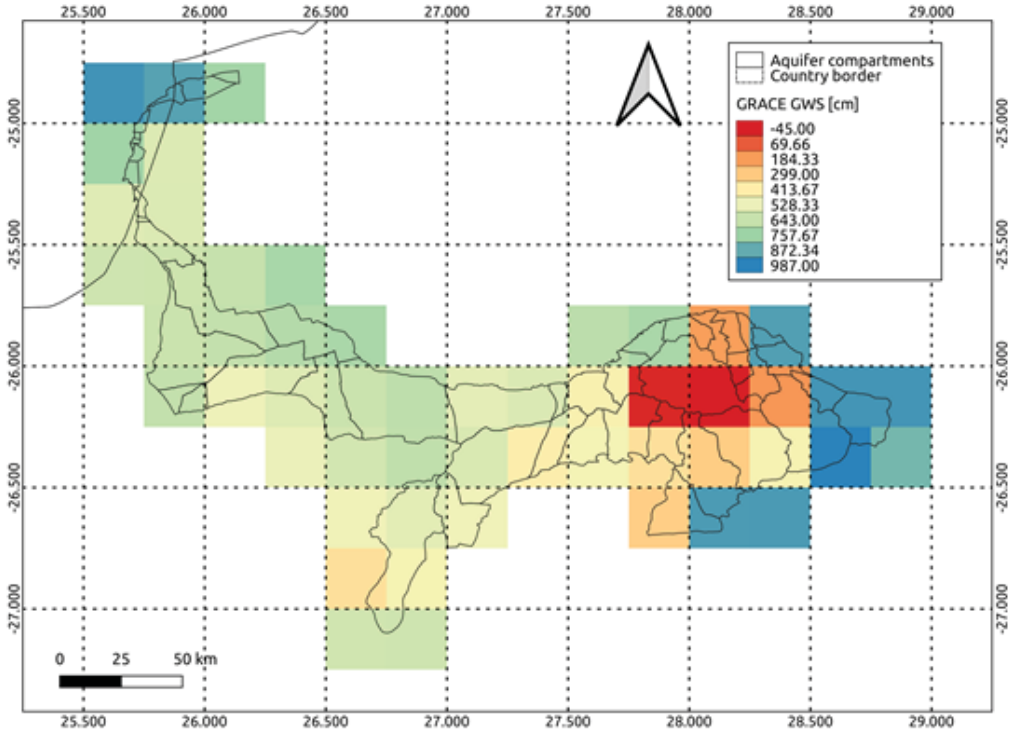


Figure 23: Net GRACE-derived groundwater storage anomaly 2002-2020 for the Dolomite Aquifer

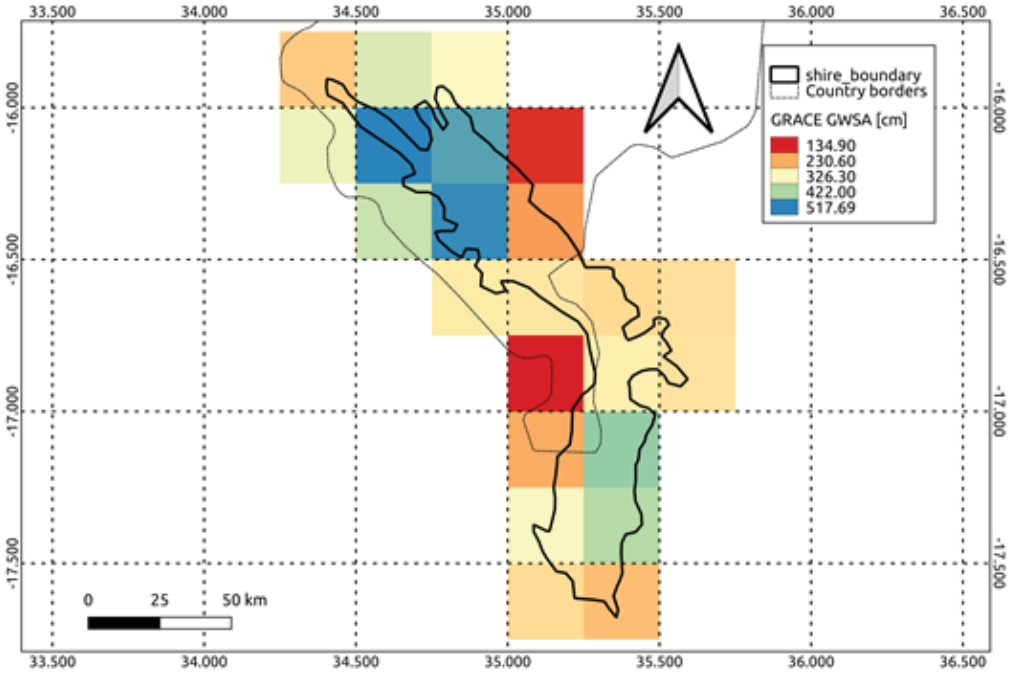


Figure 24: Net GRACE-derived groundwater storage anomaly 2002-2020 for the Shire Valley TBA

Finally, the GRACE ΔGWS data is then re-gridded from a resolution of 0,25° x 0,25° to a resolution of 0,1° x 0.1°. This is done to match the resolution of the land surface variables

provided by ERA5. This task is accomplished using a bilinear interpolation method (Miro and Famiglietti, 2018).

5.3.2. Groundwater level data

5.3.2.1. Dolomite Aquifer

The depth to groundwater level (GWL) data were aggregated into daily values using the mean. Next, GWL data were cleaned using a z-score function to remove outliers beyond 2 standard deviations away from the mean. In addition, any borehole that only had less than two records was removed from the dataset. Thereafter boreholes outside the study area boundary were removed. Thereafter the 28-32-day groundwater level changes were calculated for every observation, where possible. The average groundwater level change between the 28-32-day series is set as the 30-day groundwater level change for each observation in the record. The final dataset contained approximately 35 000 30-day groundwater level changes.

5.3.2.2. Shire Alluvial Aquifer TBA

Due to the lack of available depth to groundwater level data in the Shire Valley TBA, the pre-processing applied to the Dolomite Aquifer could not be applied to the Shire Valley TBA. Instead, the data were first aggregated into monthly mean depth to groundwater levels. Gaps in the time-series were filled in using a linear interpolation. Thereafter the monthly change in groundwater levels was calculated.

5.4. Data aggregation and integration

In order to match and integrate local data with regional data, the above 30-day groundwater level change data are used to extract data from the regional datasets into a table that represents a set of features for machine learning application. For example, for every groundwater level change record, the corresponding soil moisture anomaly, cumulative evapotranspiration, GRACE groundwater storage anomaly, cumulative precipitation, and cumulative run-off for the preceding 30, 60 and 90 days is calculated and appended to each groundwater level change record. Here the borehole locations are used to extract data from the underlying pixel value. Although it is challenging to account for probable lag times regarding the recharge response in groundwater levels, the inclusion of an extended aggregation period up to 90 days may allow for this. Table

7 provides a breakdown of the features aggregated during this process. While Figures 23-37 display examples of each predictor used in the machine learning model.

Table 7: Breakdown of the features generated through the data aggregation algorithm

Features	Description
tprecip_30	Total precipitation 30 days before record
tprecip_60	Total precipitation 60 days before record
tprecip_90	Total precipitation 90 days before record
tevap_30	Total evapotranspiration 30 days before record
tevap_60	Total evapotranspiration 60 days before record
tevap_90	Total evapotranspiration 90 days before record
tro_30	Total runoff 30 days before record
tro_60	Total runoff 60 days before record
tro_90	Total runoff 90 days before record
sma_30	30-day average soil moisture
sma_60	60-day average soil moisture
sma_90	90-day average soil moisture
lst_30	30-day average land surface temperature
lst_60	60-day average land surface temperature
lst_90	90-day average land surface temperature
ggwsa	Mean GRACE groundwater storage changes 30 days before record
aqtype_code	Aquifer type code
gma_code	GMA compartment code
gmu_code	GMU compartment code
lc_code	Land cover code

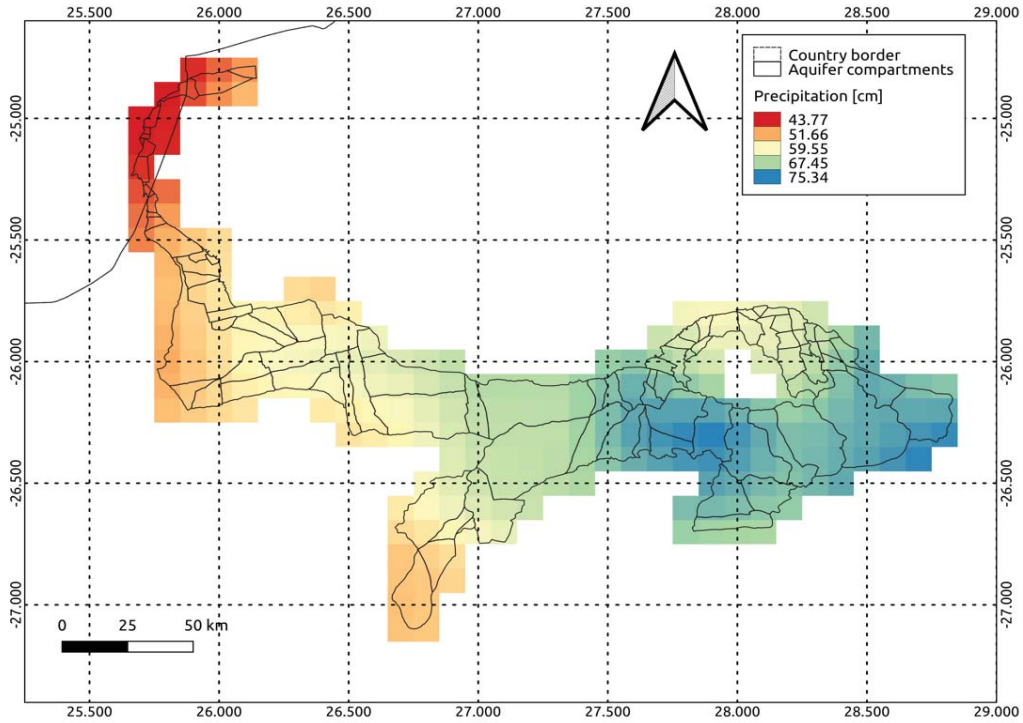


Figure 25: Mean annual total precipitation for the Dolomite aquifers14

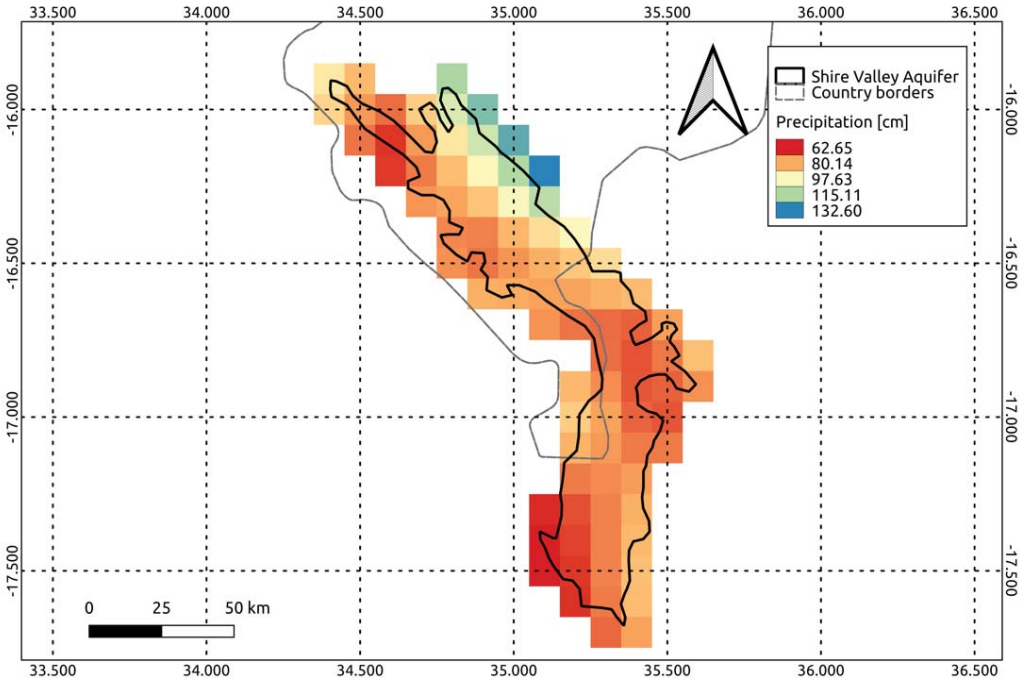


Figure 26: Mean annual total precipitation for the Shire Valley TBA

14 The units for figure 23-34 are in centimetres. Care should be taken when comparing model variables with observations, because observations are often local to a particular point in space and time, rather than representing averages over a model grid box and model time step.

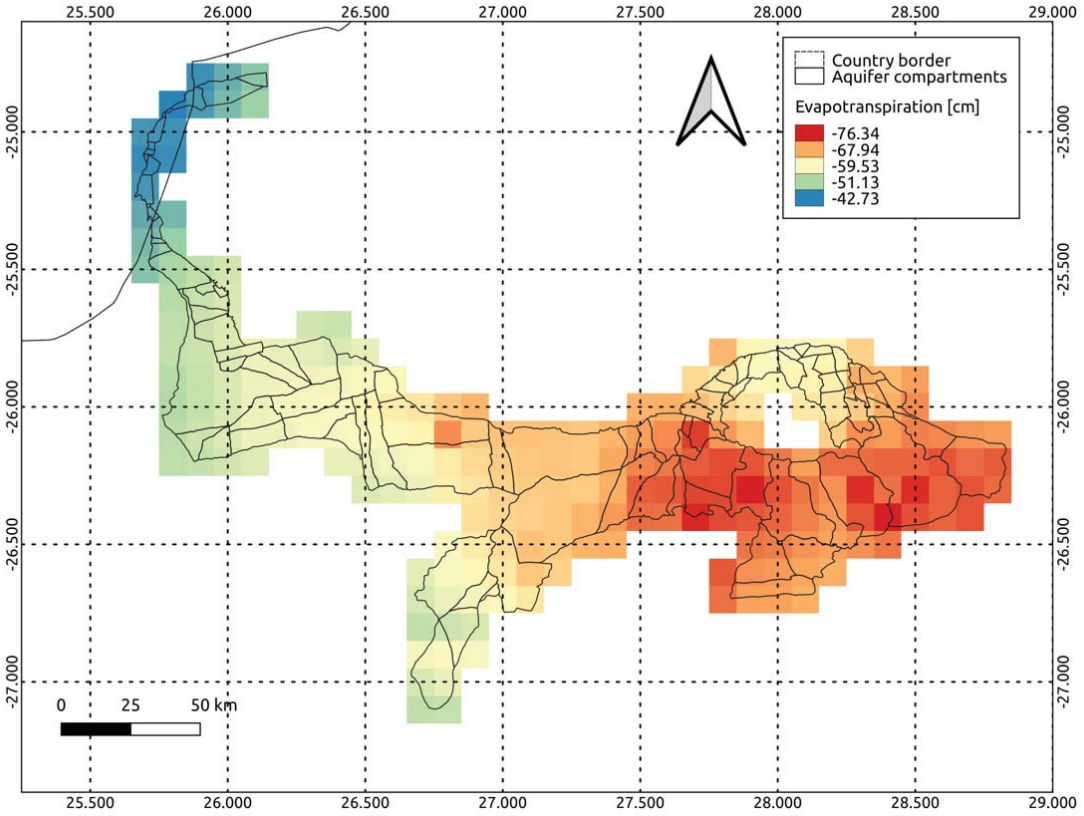


Figure 27: Mean annual total evapotranspiration for the Dolomite aquifers (negative sign indicates upward flux)

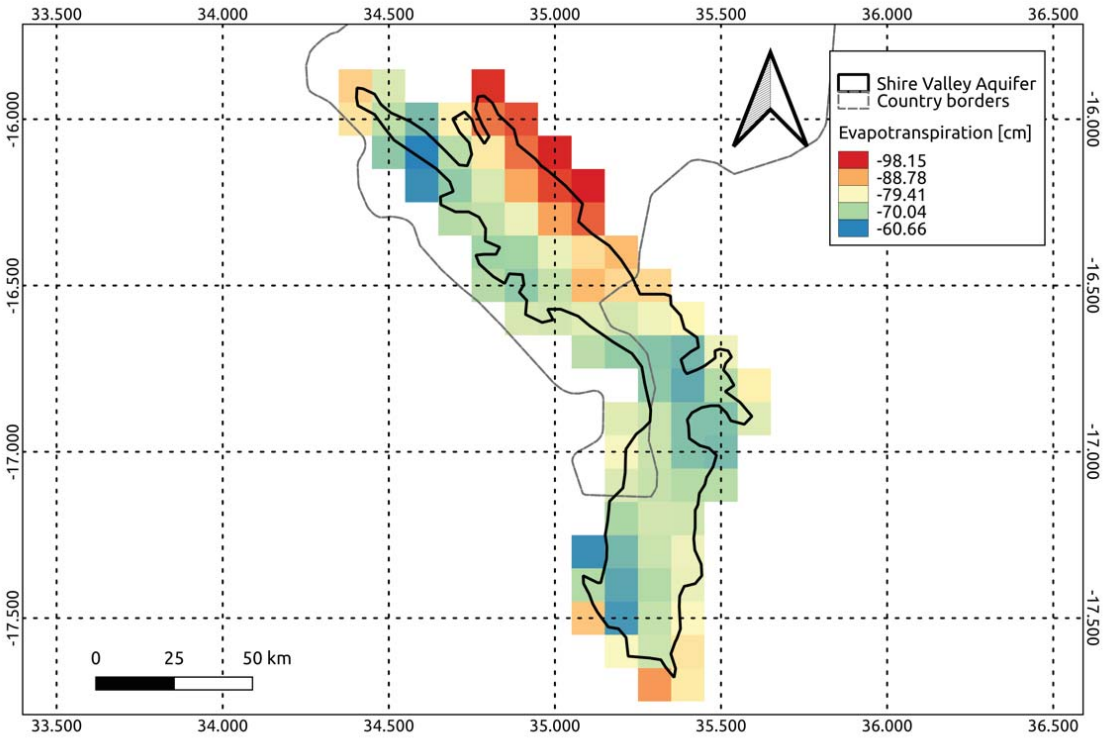


Figure 28: Mean annual total evapotranspiration for the Shire Valley TBA (negative sign indicates upward flux)

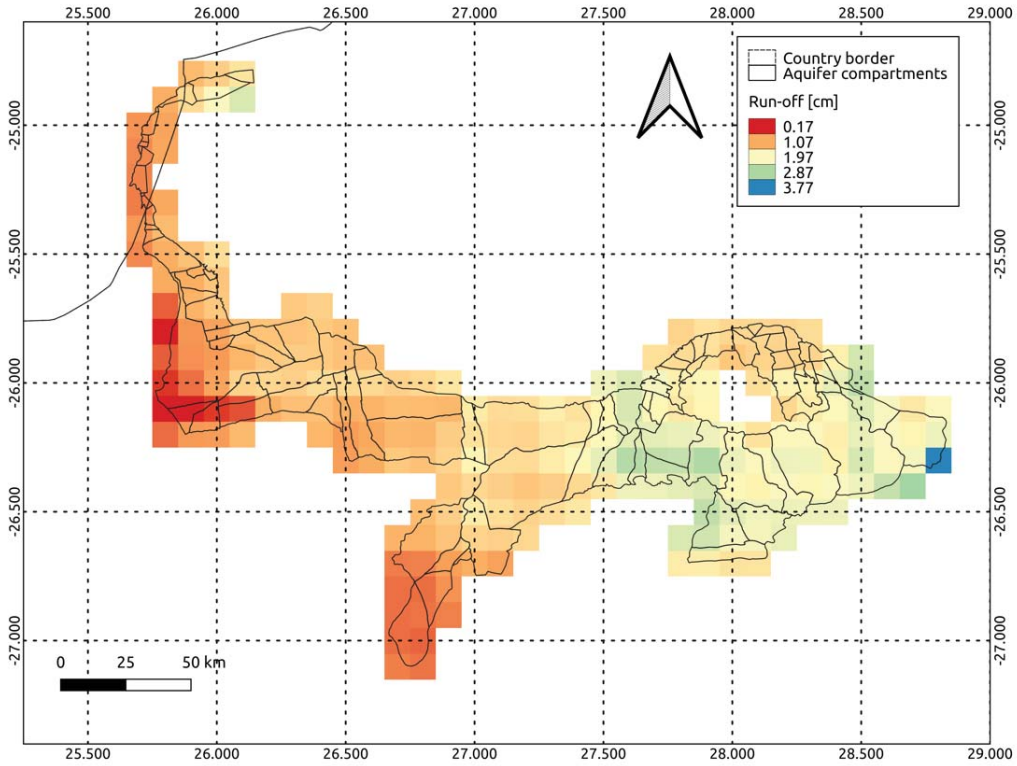


Figure 29: Mean annual total run-off for the Dolomite aquifers

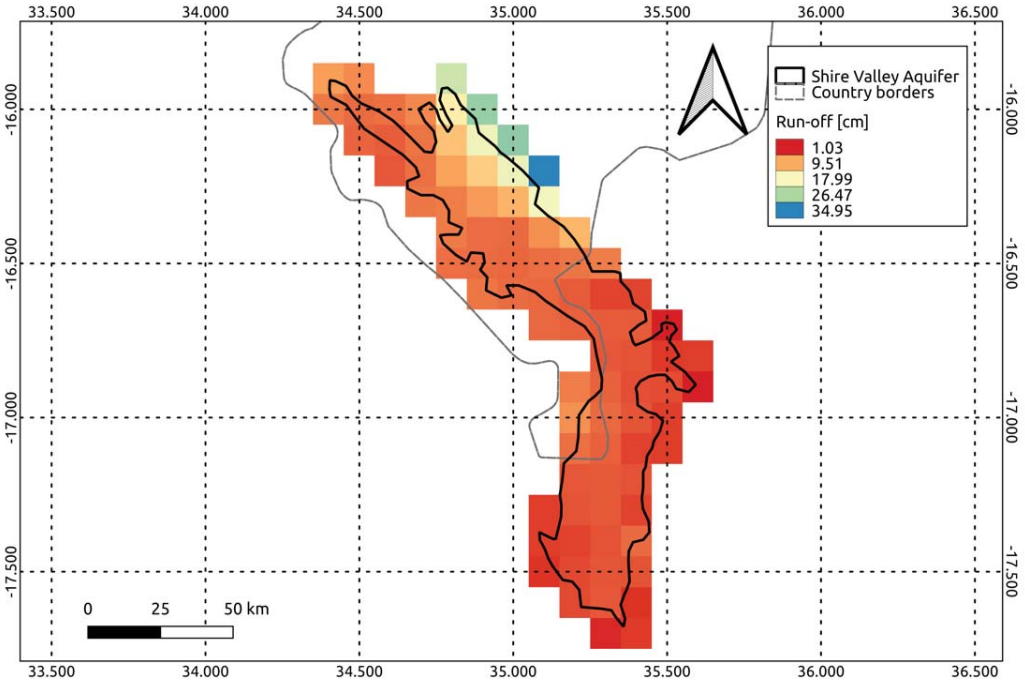


Figure 30: Mean annual total run-off for the Shire Valley TBA

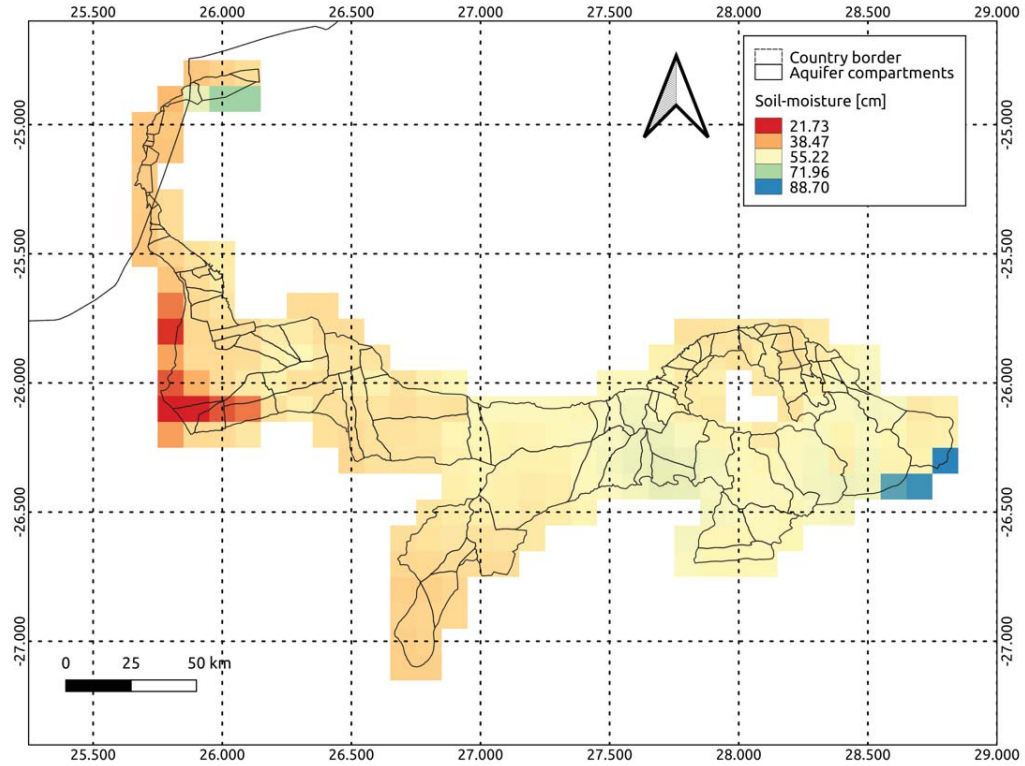


Figure 31: Mean soil moisture content up to soil depth of 298 cm for the Dolomite aquifers

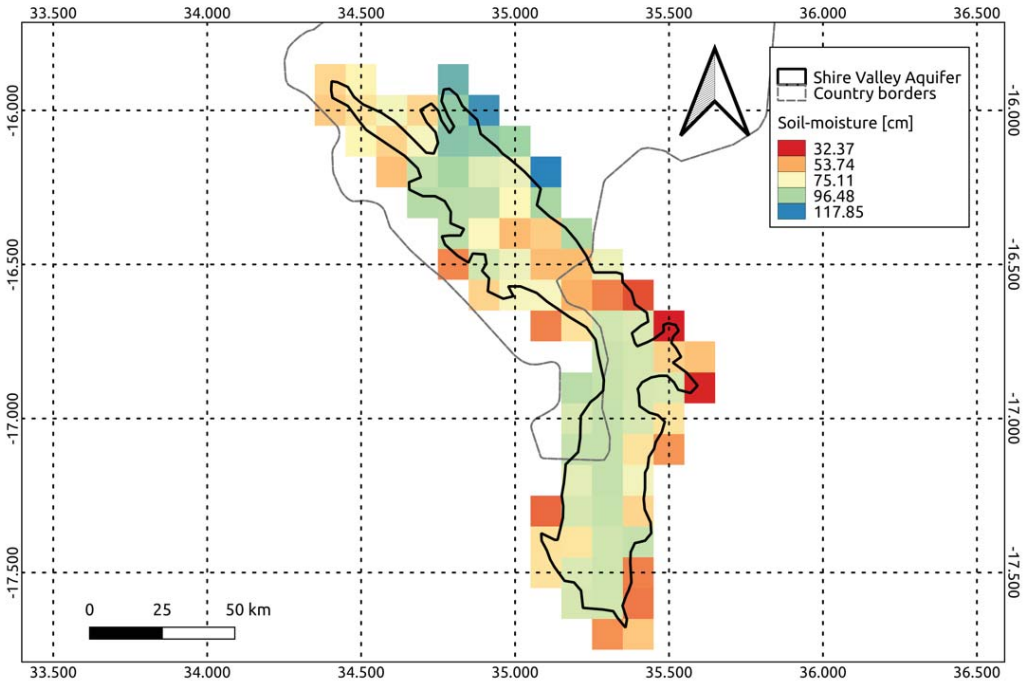


Figure 32: Mean soil moisture content up to soil depth of 298 cm for the Shire Valley TBA

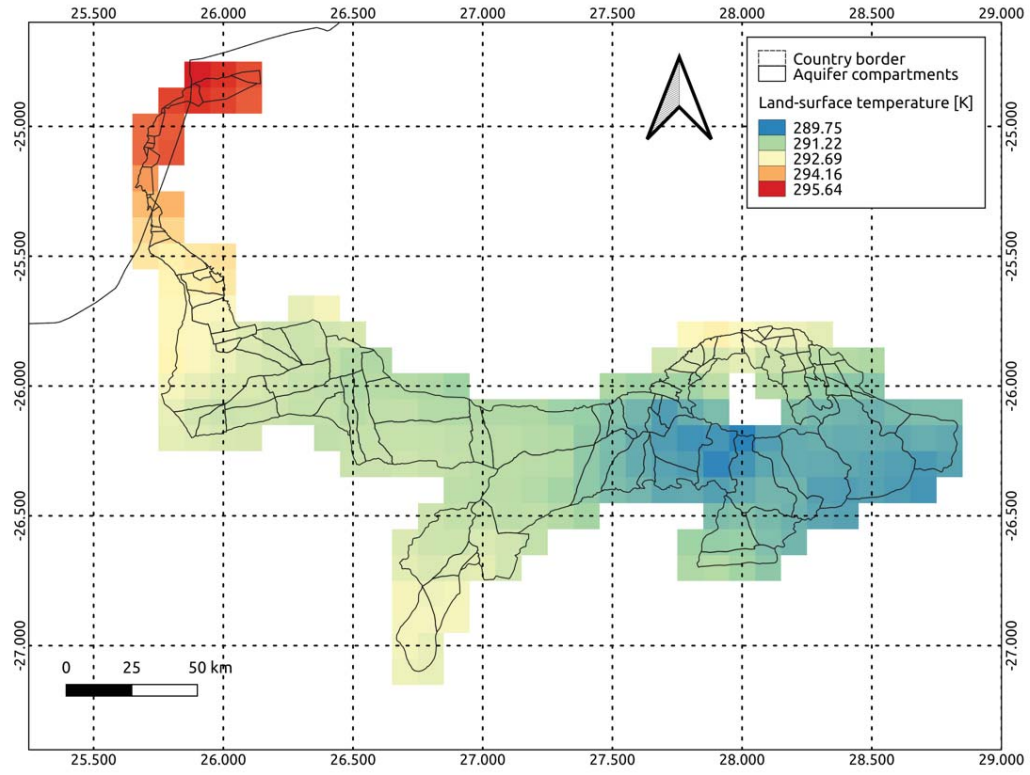


Figure 33: Mean land surface temperature for the Dolomite aquifers

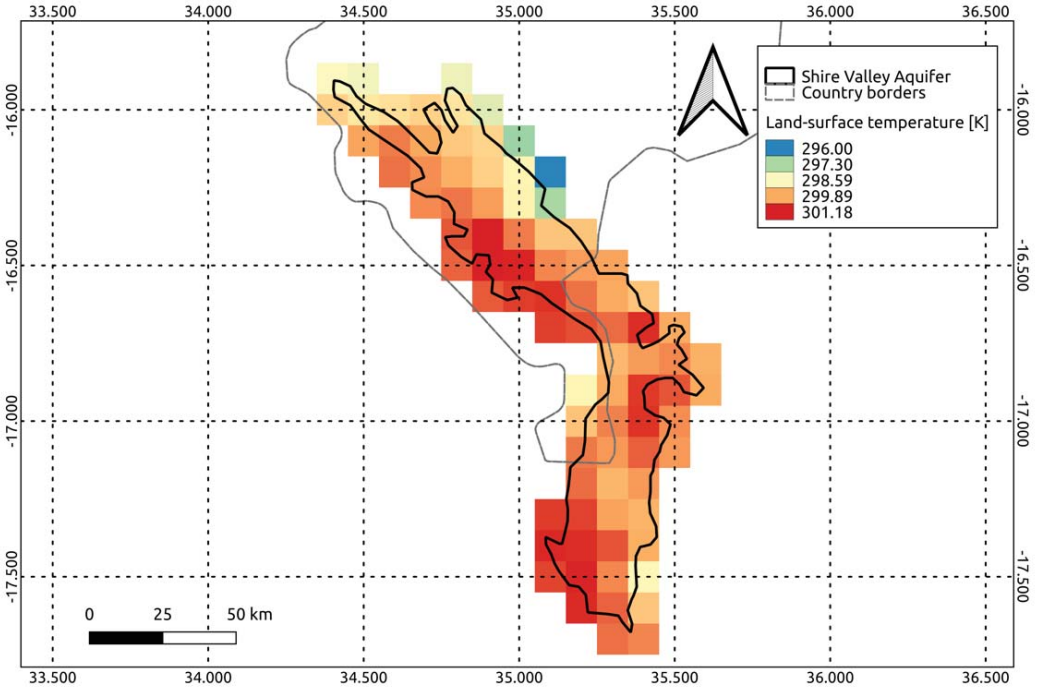


Figure 34: Mean land surface temperature for the Shire Valley TBA

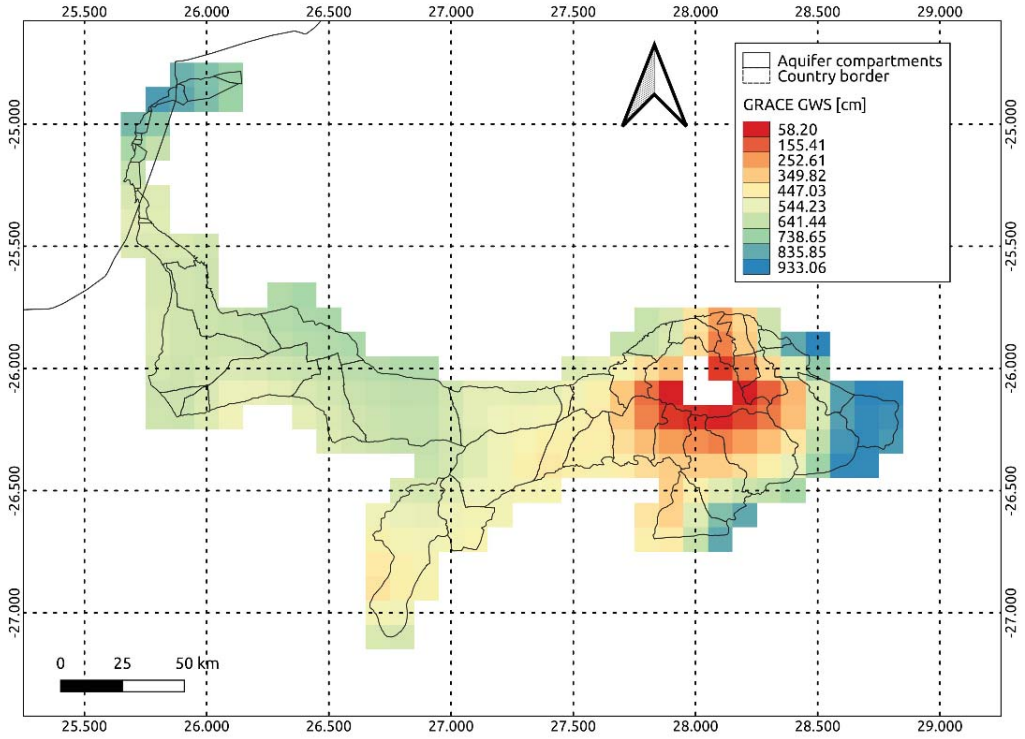


Figure 35: Net re-gridded GRACE-derived groundwater storage anomaly 2002-2019 for the Dolomites aquifers

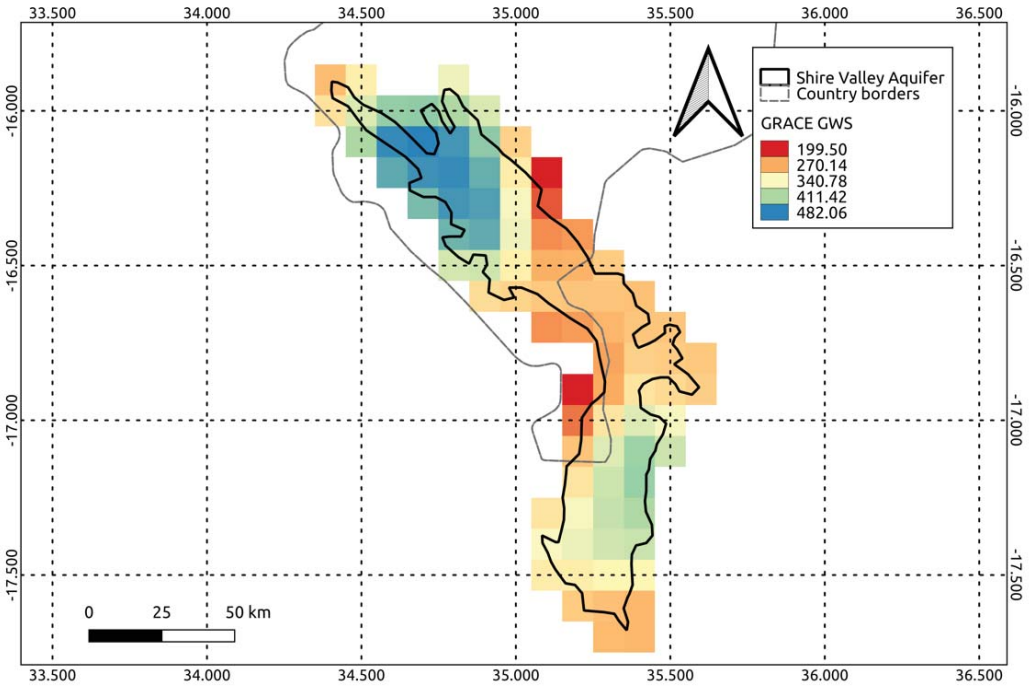


Figure 36: Net re-gridded GRACE-derived groundwater storage anomaly 2002-2019 for the Shire Valley TBA

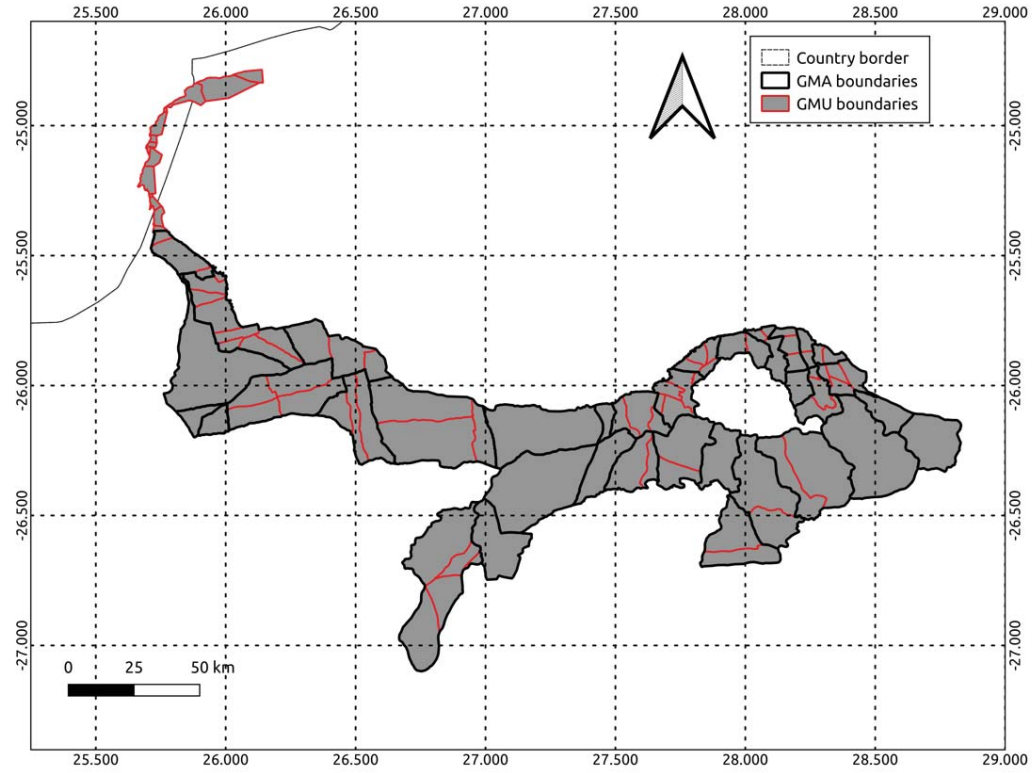


Figure 37: Aquifer compartments classified as GMAs (Cobbing et al., 2016)

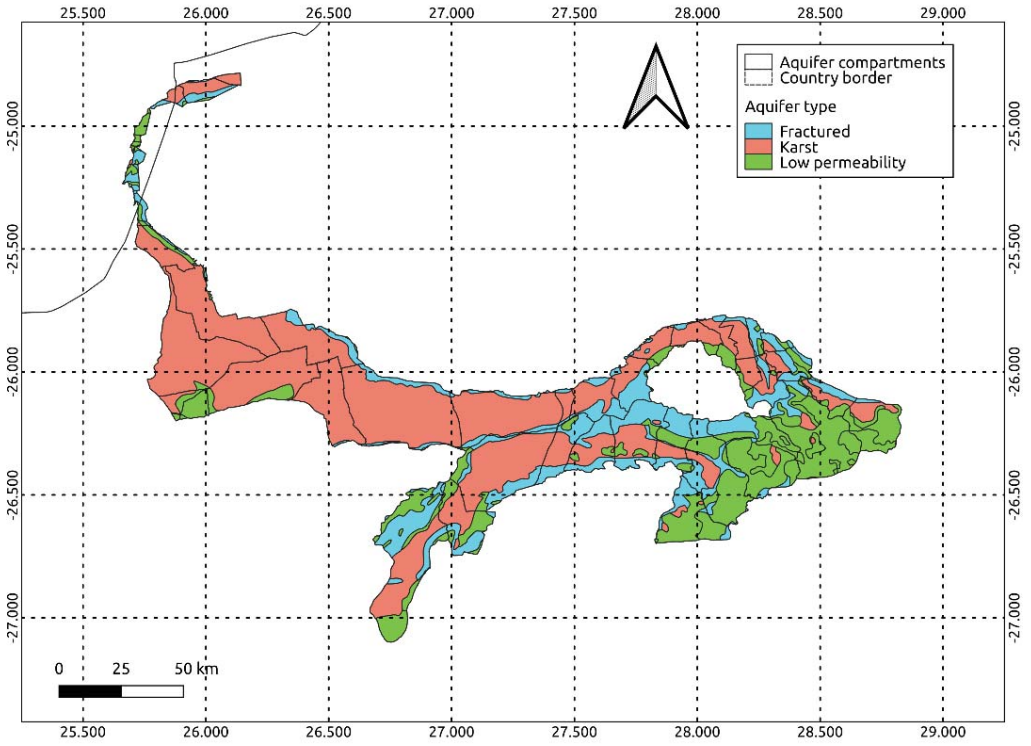


Figure 38: Aquifer types (SADC, 2010)

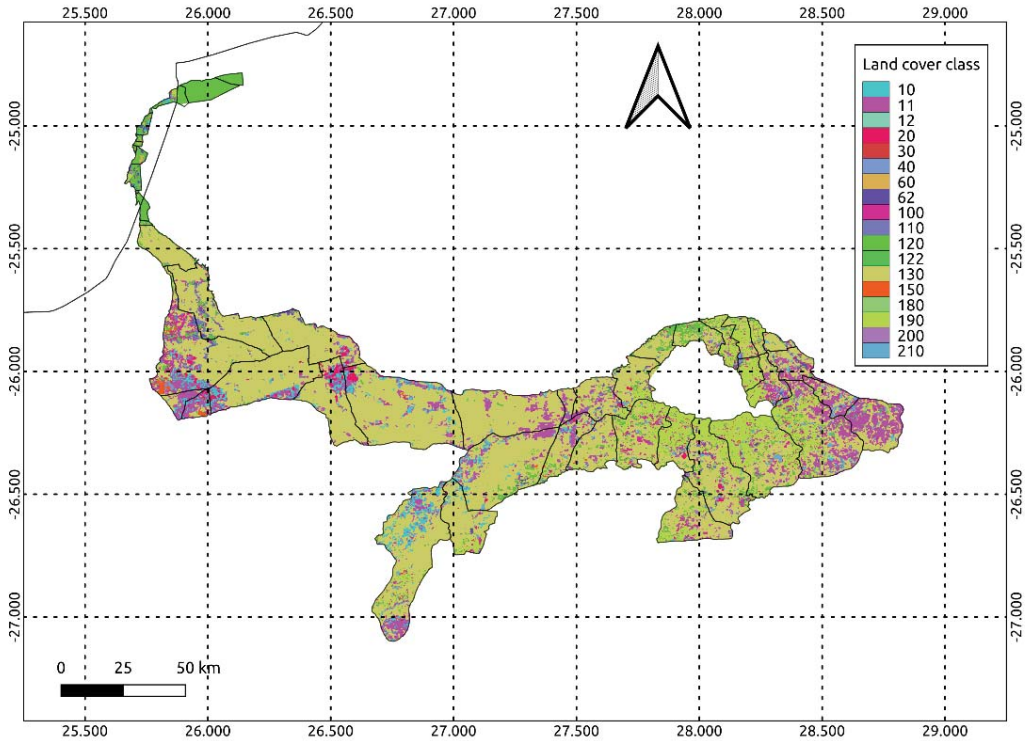


Figure 39: Land cover for the study area (legend see Appendix 1)

5.5. Conclusion

During this research, a significant amount of data was collected and pre-processed. The data originate from various sources and are stored in a number of different formats. This is even more true for local scale data, such as in-situ groundwater observations. The regional scale data on the other hand are more well organized and are encoded and stored according to internationally recognized standards. The regional scale data are also more easily acquired, due to being publicly available via web-based platforms. Compared to in-situ groundwater data, which in many cases requires communication with relevant authorities and data providers, before being approved for sharing.

Pre-processing of data into analysis ready datasets require a significant amount of time. This is largely due to the inconsistencies in data storage formats. In particular, the way in-situ groundwater observations are stored and shared are significantly different from one data provider to the next. For example, the typical storage mechanism for in-situ groundwater observations is in spreadsheets and table. However, the format of these tables tends to be completely different from one data provider to the next. Tables thus require editing before data can be collated together. In some cases, the editing must be done manually, one table at a time.

Beyond the initially editing, the data goes through a transformation processes that is designed to convert the data into a set of analysis ready features. This component is specific to the use case and will differ depending on the application. In terms of this application, critical thought had to be given to the feature engineering phase, the purpose of which is to provide a set of features for machine learning. It is imperative that the features be based on hydrogeological understanding of the aquifer. For example, the groundwater level changes were used to understand the response to various hydroclimatic features. Also, the use of 30-, 60- and 90-day changes were included to account for the possible lag times between recharge and groundwater level responses.

In its entirety the data collections and pre-processing are perhaps the most time-consuming portion of any BDAs applications.

6. Application of machine learning algorithm to case study areas

6.1. Introduction

In the previous section the sourcing and pre-processing of the relevant data used in this study was described. The following sections describe the set-up and execution of a machine learning algorithm that is used to model groundwater level changes in the case study areas. The information generated from the model will allow inspection of the selected groundwater management scenario. In this case, chronic lowering of groundwater levels.

6.2. Machine learning algorithm

In this study we used a decision tree-based learning algorithm, implemented within a gradient boosting framework (Ke et al., 2017). Gradient Boosting Decision Trees (GBDT) are popular machine learning algorithms, that can be used for regression, classification, and ranking. Decision trees have advantages of being easy to interpret, handle missing values, not influenced by outliers, do not need a priori information, and can handle irrelevant features (Seyoum et al., 2019). In addition, these models are highly efficient and accurate (Ke et al., 2017). Modern decision trees are simple supervised tree-like models that rely on a set of decisions (branches) based on conditions (nodes) that end up qualifying/predicting an outcome (leaf) (Gupta, 2017). Figure 40 illustrates a simple decision tree, in which we determine whether an individual will survive or not on board a sinking Titanic. Here three features or attributes are used to reach an outcome: gender, age, and number of spouses or child on-board. Technically speaking decision tree models partition a predictor space (dataset space) into rectangular subspaces, based on conditional probability of the outcomes of decisions from given conditions, using the training data (Elith et al., 2008; Leonard, 2017). In regression tree models each subspace or region is fit with constant, such as the mean response for observations in that region (Elith et al., 2008). The tree grows (new leaves forming) as new conditions are met and decisions are made to reach an eventual prediction or rather a terminal node. The outcome in most cases is generally the predictant data (i.e. label data).

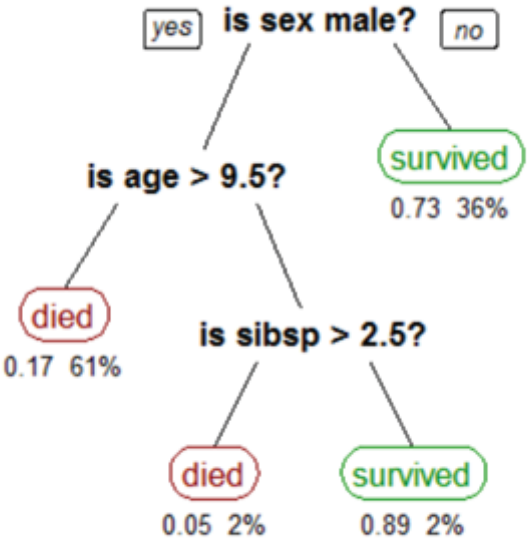


Figure 40: Example of a decision tree used to determine if a passenger onboard the Titanic will live or die based on probabilities within the covariant space (Gupta, 2017)

Instead of a single decision tree model, a gradient boosting relies on fitting many decision trees to the data, and then combining them in an ensemble fashion to produce a robust model. Typically, successive models are built using the residual errors of the preceding model, adjusting to further reduce the residual error as boosting progresses (Elith et al., 2008; Seyoum et al.. 2019). Unlike other ensemble methods (bagging, stacking and model averaging), boosting relies on successive iterations, where the overall accuracy is improved by reducing the gradient of a loss function (measure of predictive performance) at successive iterations (Schapire, 2003). In our implementation of GBDT we rely on a model put forward by (Ke et al., 2017), that uses histogram-based approach to find node points in the predictor space, as well as using a Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) method to reduce data instances. In addition, trees are grown leaf-wise, instead of depth-wise. The implementation is executed using the LightGBM module[15] in a python environment.

[15] https://lightgbm.readthedocs.io/en/latest/#

6.3. Model design

6.3.1. Model design Dolomite aquifer

In this study we implement a GBDT framework, to develop a generalised decision tree model that can be used to compute the 30-day groundwater level change at any locality in the study area based on a set of input variables (Table 7). The process to achieve this involves reducing the boreholes that have over 1000 records by random sampling of 1000 records. Thereafter the data is randomly split into training, validation and test sets along the borehole identity axis. Essentially this ensures that the model is set-up to train on the entire time-series from a set of boreholes and validate on another set of boreholes. In addition, the data is anonymised by removing the borehole identity attribute. These processes ensure that the training data is more generalized and not specific to an individual borehole.

Repeated iterations of model training are done based on different random splitting and machine learning trails, to ensure repeated model runs during training are robust, and consistent. This means that the model should train the same regardless of which boreholes are involved in training and which are involved in validation. The aim of this is to ensure consistent predictive powers across all boreholes, akin to a more generalized model of the aquifer. We then test the results on a test set initially removed from the training phase.

Model hyperparameter tuning was also incorporated in the model design. This involved the use of Scikit-Learns GridSearchCV function to perform a grid search of the optimal parameter settings for the GBDT model. In this case, a fourfold cross-validation for each of the 16128 possible combinations of parameters were run to ascertain the optimal parameters. Finally, the model with lowest prediction score is selected, and used to predict the groundwater level change at the centre node of a predefined prediction grid (Figure 41). The grid has a resolution of 0.5^{o}x0.5^{o} (~5km). Firstly, a monthly time-series (2002-2019) is populated with data from the predictor variables as set out in Table 7. Thereafter the corresponding groundwater level change is predicted using the features and model with the lowest mean absolute error. Finally, the results are validated using change in observed depth to groundwater level data.

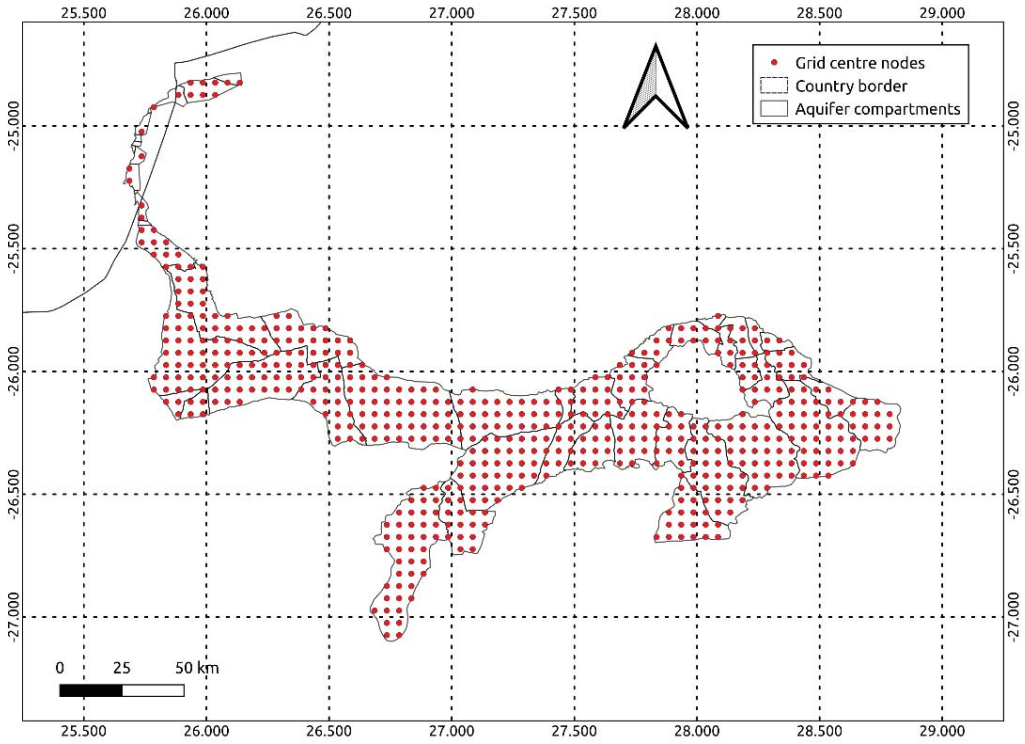


Figure 41: Grid centre nodes used as prediction points for application of the trained model

6.3.2. Model design Shire Alluvial Aquifer

Due to the lack of available data, there are a limited number of model setups that can be applied in the case of the Shire Valley alluvial aquifer. Only a single borehole, Ngabu (Figure 18), provided a long enough time-series. Unlike with the Dolomites Aquifer, here the groundwater levels were aggregated into monthly values, and the difference in groundwater levels between the months calculated. The GBDT is set up to predict monthly groundwater level changes, using the features presented in Table 6. The dataset is split into a training, validation and test set along the time-series (i.e. the first 80 months are used as the training set, the next 30 months are used as the validation set, while the remaining 22 months are used as a test set).

6.4. Results

6.4.1. Model results Dolomite Aquifer

A total of 100 model runs were conducted, with each run representing a different random splitting of the dataset in training and validation sets (testing sets removed

beforehand) (Appendix 2). Table 7 indicates the results of the ten best iterations of this process. The mean absolute errors for the training and test phase are shown. In this case the training score is used to rank the model runs. As can be seen model run 16 produces the best results, at least in terms of lowest mean absolute error. For this model the MAE is roughly 18 cm. In general, this means that on average the model is predicting the groundwater level change with an error of 18 cm. (Seyoum et al., 2019) report similar MAE for their experiment. All the runs display a large difference in mean absolute error from training to testing evaluations. As well as displaying large differences between subsequent model runs. This is not ideal, as it suggests that the randomization during splitting of training and testing is not generating consistent covariance amongst feature during subsequent runs. This would result in different covariant subspaces being created during training on each successive model. However, it must also be noted that there is a marked improvement compared to previous reported results (Deliverable 4: Lessons learnt from case studies). This can be seen in the mean score for all 100 models, which is significantly lower than in previous reports.

Table 8: Score for the top ten model runs (Score=Mean Absolute Error, units=cm)

Model No.	Training score	Testing score
16	17,848946	26,740144
41	19,069898	25,916309
21	21,944976	25,505946
89	22,018323	26,304860
47	22,405129	26,273599
2	22,464920	29,883113
17	22,533392	25,710877
40	23,624674	25,152207
90	23,669164	28,685849
32	24,494933	25,539539
Mean score (all 100 models)	33,206389	27,552611

In-addition to the above table the Figures 40-42 display the predicted vs true values, for the training, validation, and test dataset, respectively. In this case the results are based on the top ranked model (model 16). Here it can be seen that the model struggles to effectively predict outliers or rather extreme values in the dataset. In-addition the plots illustrate that the model is making repeatedly similar predictions (linear artifacts of the data points). This may in-turn be affecting the overall MAE. This may also be a consequence of the categorization of the data into aquifer compartments, or the

possibility that the model is not training effectively. These results are consistent across the training, validation and testing sets.

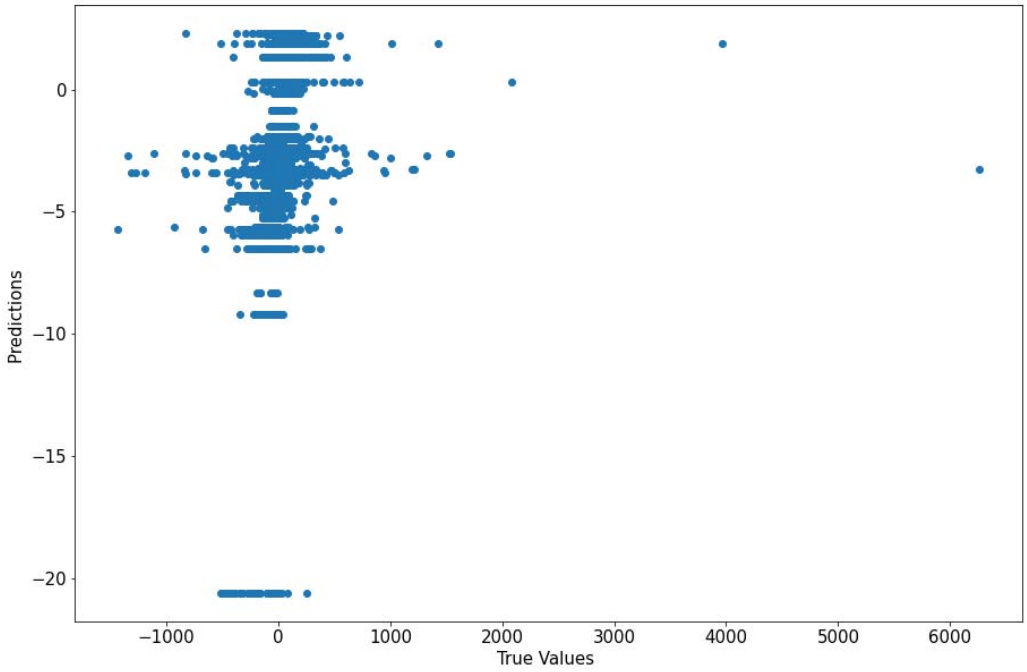


Figure 42: Scatter plot of predicted vs true values for the training set

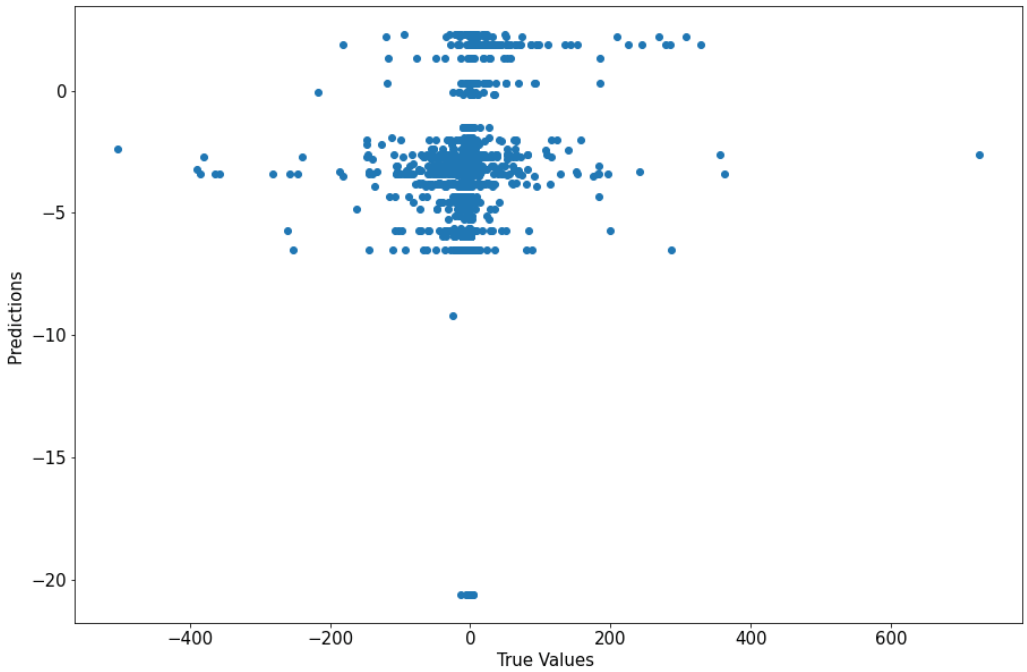


Figure 43: Scatter plot of predicted vs true values for the valid set

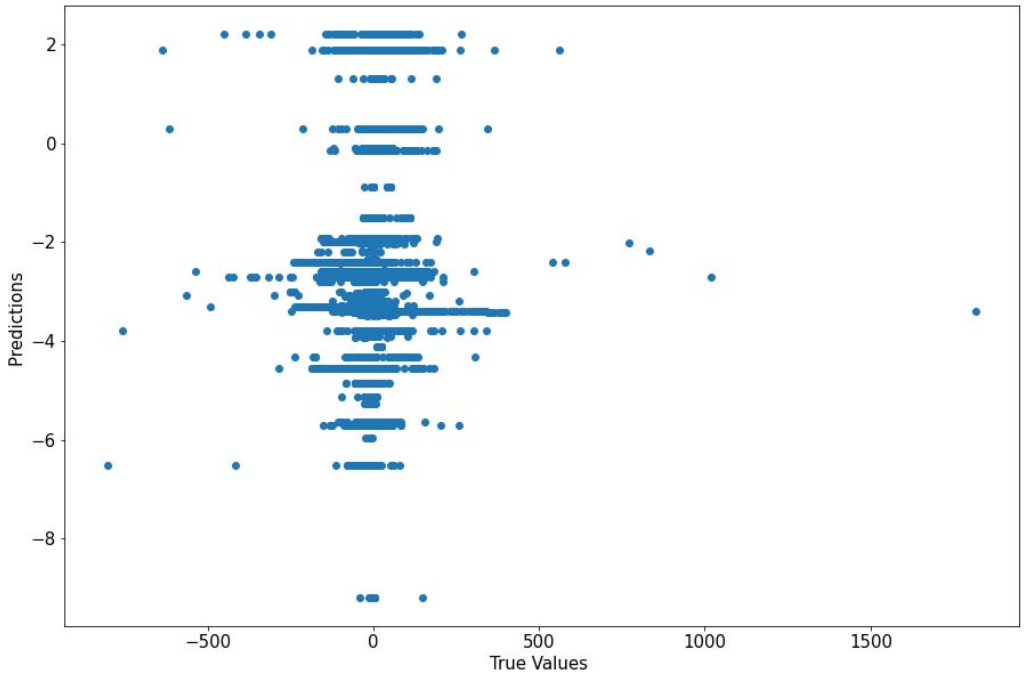


Figure 44: Scatter plot of predicted vs true values for the test set

6.4.2. Model results Shire Alluvial Aquifer

For the single Ngabu borehole, the model training and validation produced a mean absolute error of 34,5799 cm. During the training process the model fails to learn effectively, instead perhaps memorizing the training data. This is reflected in the training where the model fails to significantly improve the MAE through successive iterations. Figure 45 illustrates the predicted vs true values for the training dataset. Figure 46 displays the predicted vs true values for the validation dataset. In this case the predicted values have a higher degree of error. The model also does not accurately predict extreme values.

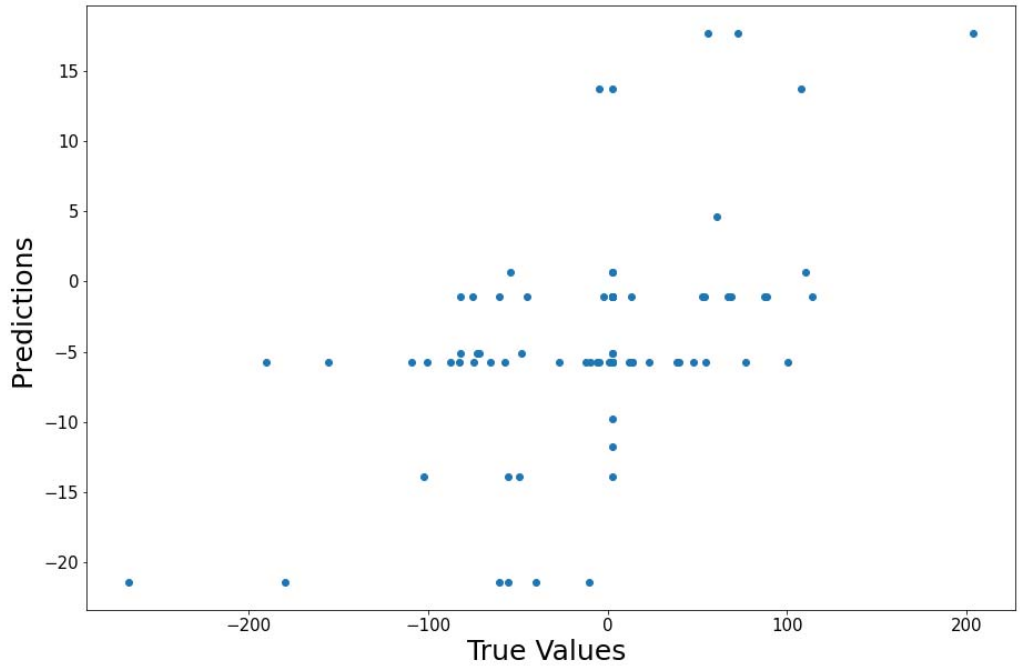


Figure 45: Scatter plot of predicted vs true values for the training dataset

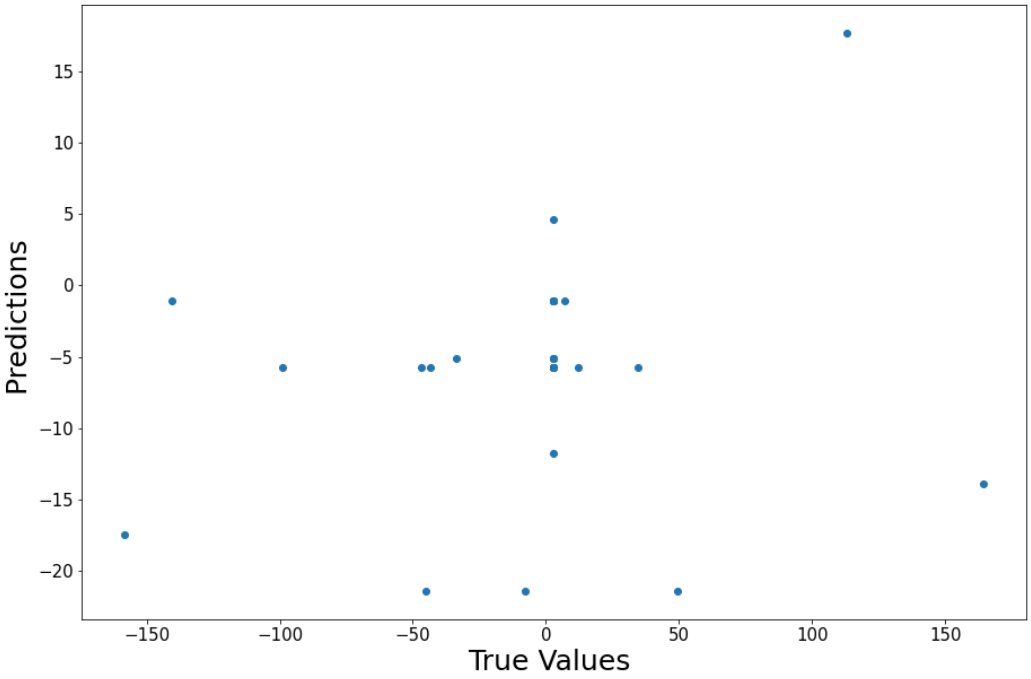


Figure 46: Scatter plot of predicted vs true values for the validation dataset

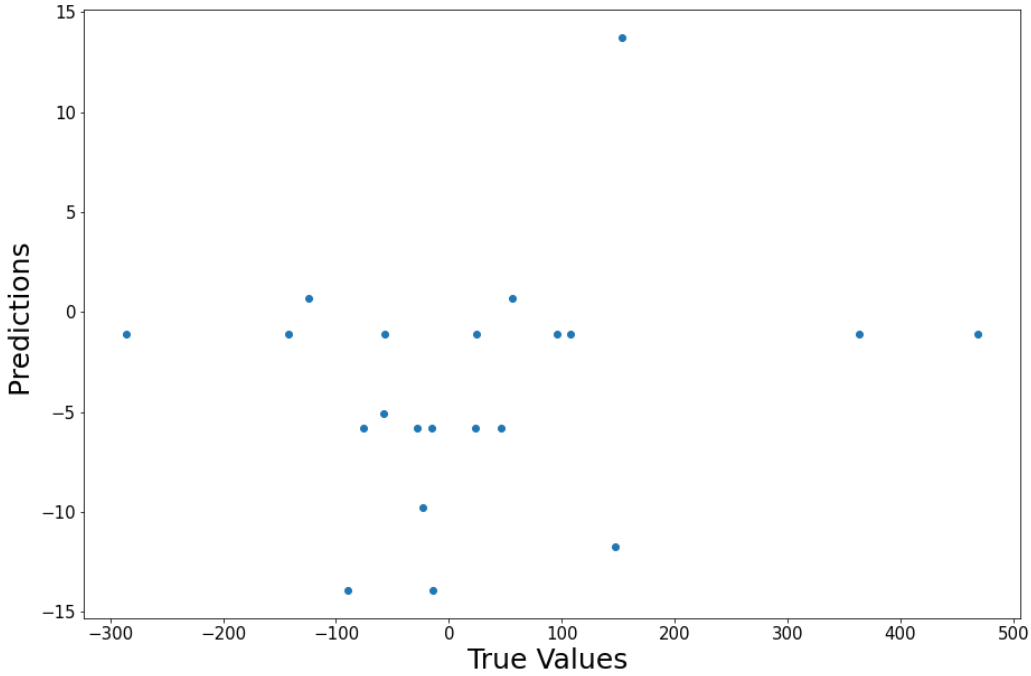


Figure 47: Scatter plot of predicted vs true values for the test dataset

6.4.3. Feature importance Dolomite Aquifer

As stated, the feature importance during model runs was also investigated. Appendix 3 displays the feature importance for the top ten model runs. Figure 48 displays the feature important with the top model iteration (model 16). Here it can be seen that the 30-day average land surface temperature has the highest influence on the prediction of groundwater level change. While the GRACE derived groundwater storage is considered the 2nd most influential feature. Theoretically groundwater storage change is proportional to groundwater level changes. Depending on the model run, either GRACE GWS or Land surface temperature are considered the most influential feature. In general, it also appears that the aquifer compartments play a large role in the prediction, as suggested previously. In most cases the hydroclimatic variables such as precipitation and evapotranspiration have a lower influence on the prediction. In general, also the 30- and 90-day features appear to have greater influence than the 60-day features.

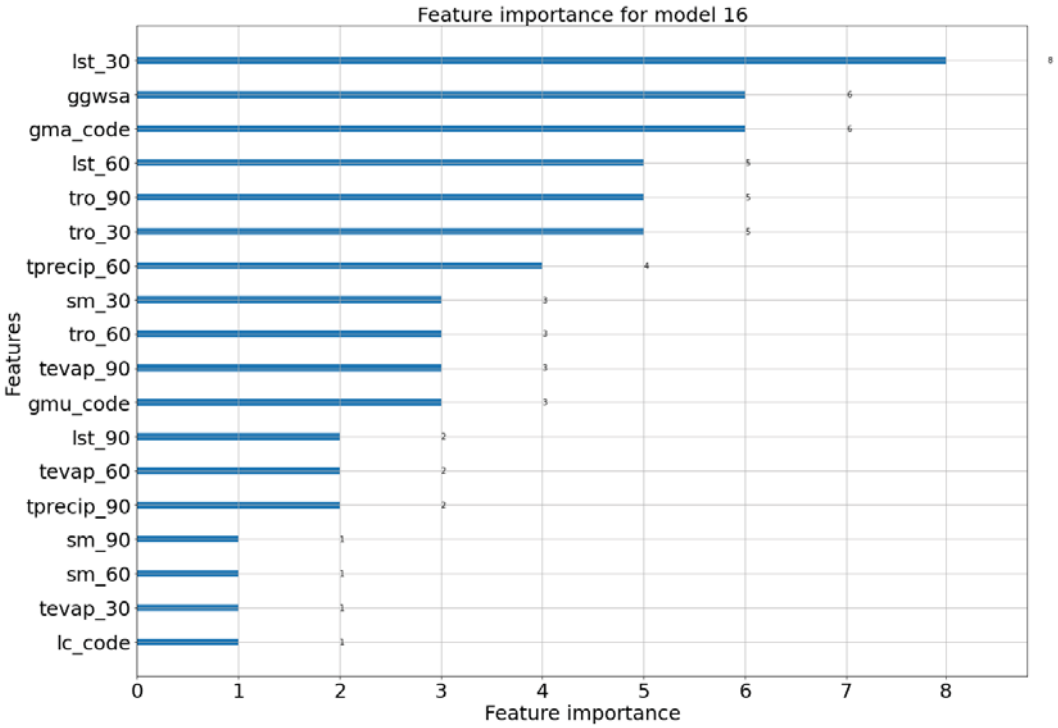


Figure 48: Feature importance for model #16

6.4.4. Feature importance Shire Alluvial Aquifer

Figure 49 shows the feature importance of the model. For this model in the Shire Valley TBA, the most influential predictor is the 30-day cumulative evapotranspiration, with the precipitation and soil moisture also influencing the prediction. The size of the training data limits the iterative training process, with is reflected in the fact that not all the features are considered. This may be a consequence of the model hyper parameter settings, where the trees are set up with a small number or leaves containing a limited number of data points.

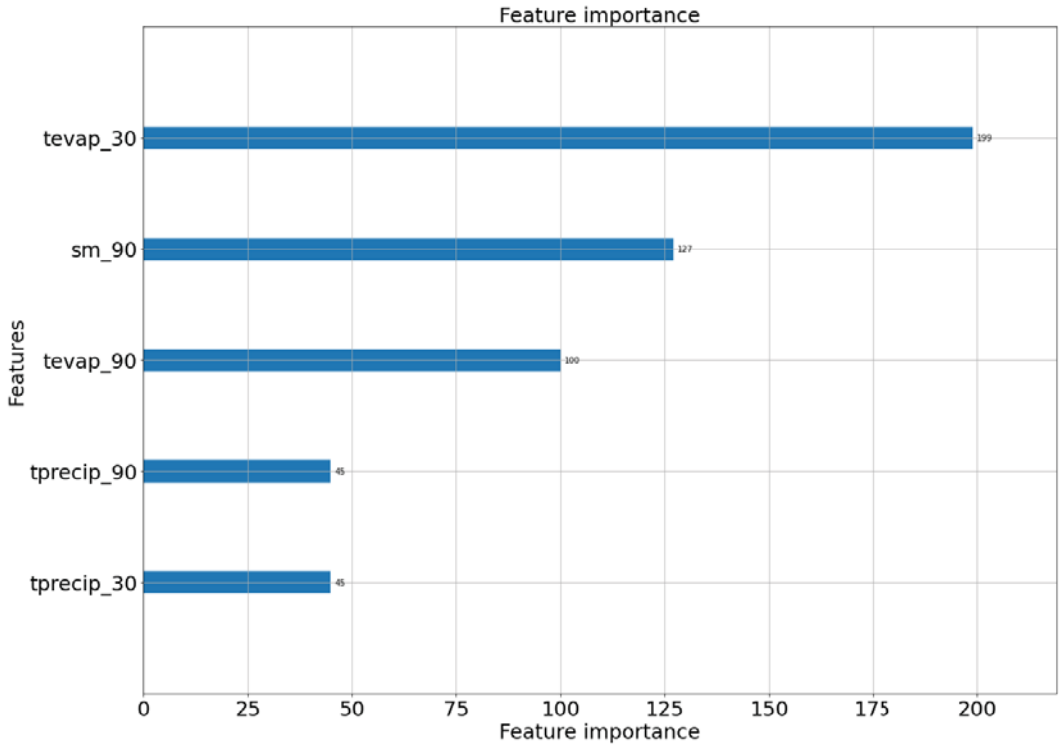


Figure 49: Feature importance for the GBDT model in the Shire Valley TBA

6.4.5. Downscaling grid results Dolomite Aquifer

Here we use the top ranked model (model 16) to generate a set of monthly groundwater level change maps from April 2002-November 2019. Appendix 4 contains a set of triennial net groundwater level change as predicted by the model. While a rasterized version of the net modelled groundwater level change across the entire time-series is displayed in Figure 50. Based on the results, it can be suggested that groundwater level is significantly lower compared to the start of the study period (2002). In-fact when grouped into 3-year cumulative values, there has been a consistent negative change in groundwater levels across the study area. Significant decreases are predicted in the Ramotswa sections, as well as in the middle and southern outcroppings of the dolomites.

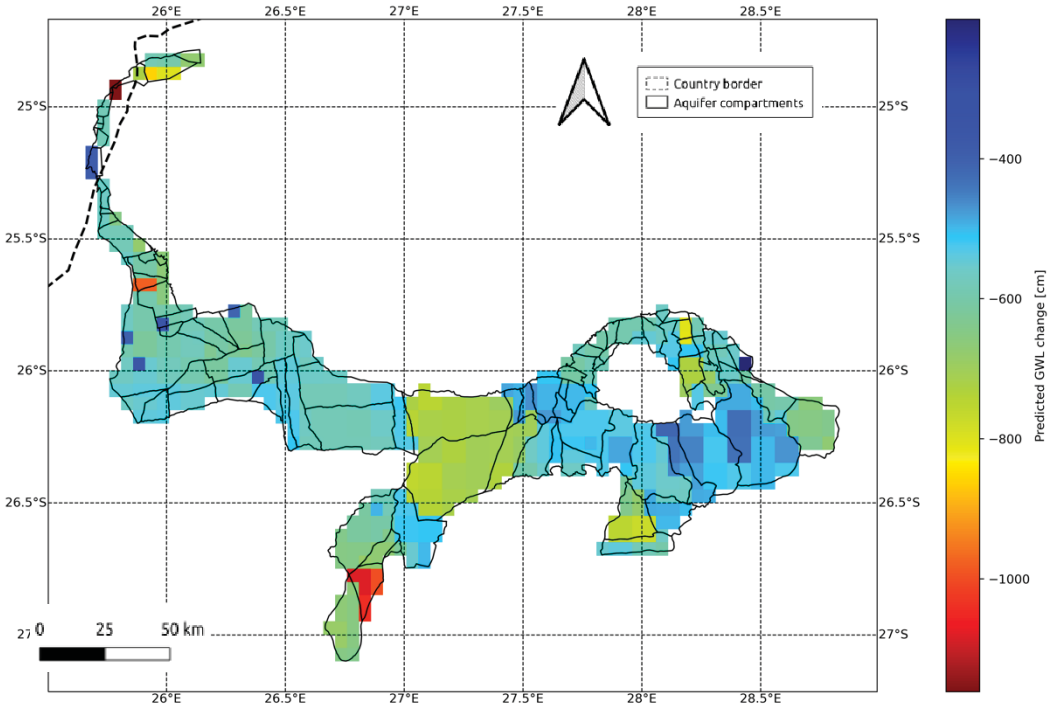


Figure 50: Net modelled groundwater level anomaly for the study area 2002-2019 (cm)

6.4.6. Validation of Dolomite Aquifer

To validate the modelled groundwater level changes, a sample of boreholes are chosen to compare observed groundwater level changes against predicted changes. Firstly, the time series of boreholes are aggregated into mean monthly groundwater level values. Thereafter only boreholes that have more than 150 records were retained for further analysis. Data gaps for months without data were filled using a linear interpolation approach. The difference between monthly groundwater levels was then calculated. Lastly, outliers are removed from the monthly validation dataset, as was done with pre-processing of water level data for model development.

The mean absolute error between the predicted groundwater level change and the observed groundwater level change is calculated at 17,181 cm. Even while extreme values are removed from the validation dataset, the model fails to accurately predict extreme values that still remain in monthly groundwater level changes (Figures 49-51). However, it must be pointed out that comparing a gridded product to point data is not expected to yield exact matches. Perhaps a more representative approach is to calculate the average observed groundwater level change per grid cell, to avoid the effects of outliers.

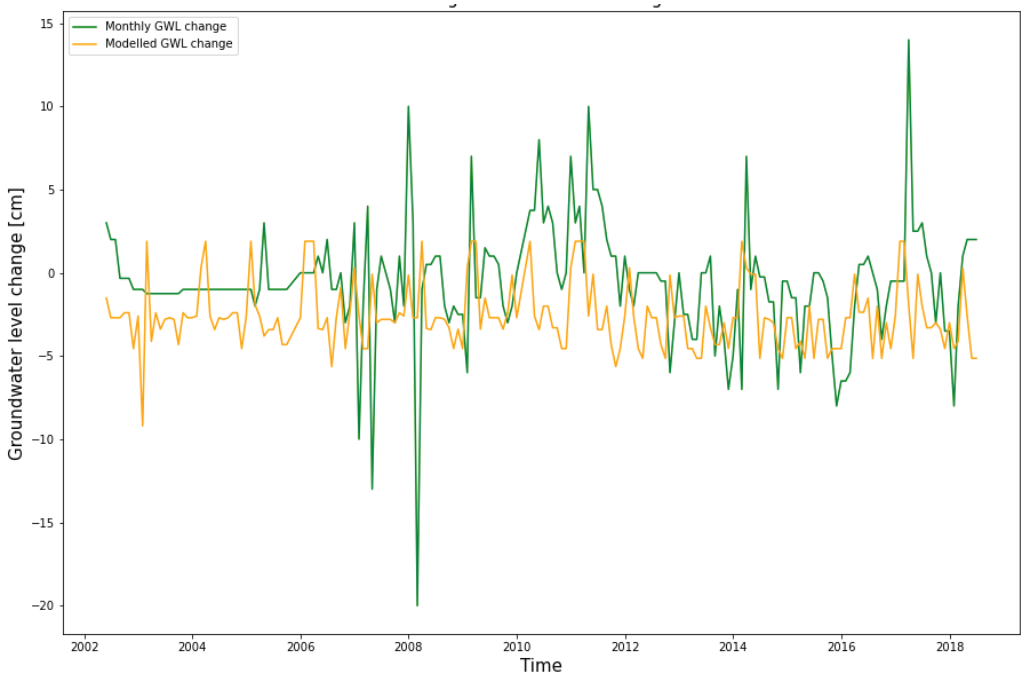


Figure 51: Borehole with low mean absolute error

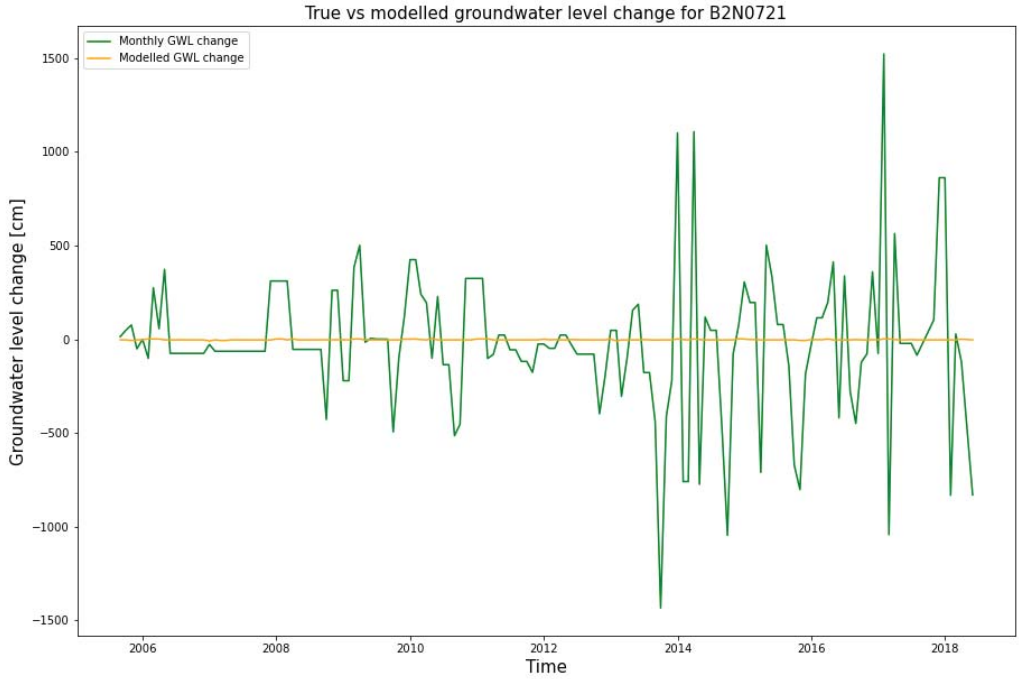


Figure 52: Borehole with the highest mean absolute error

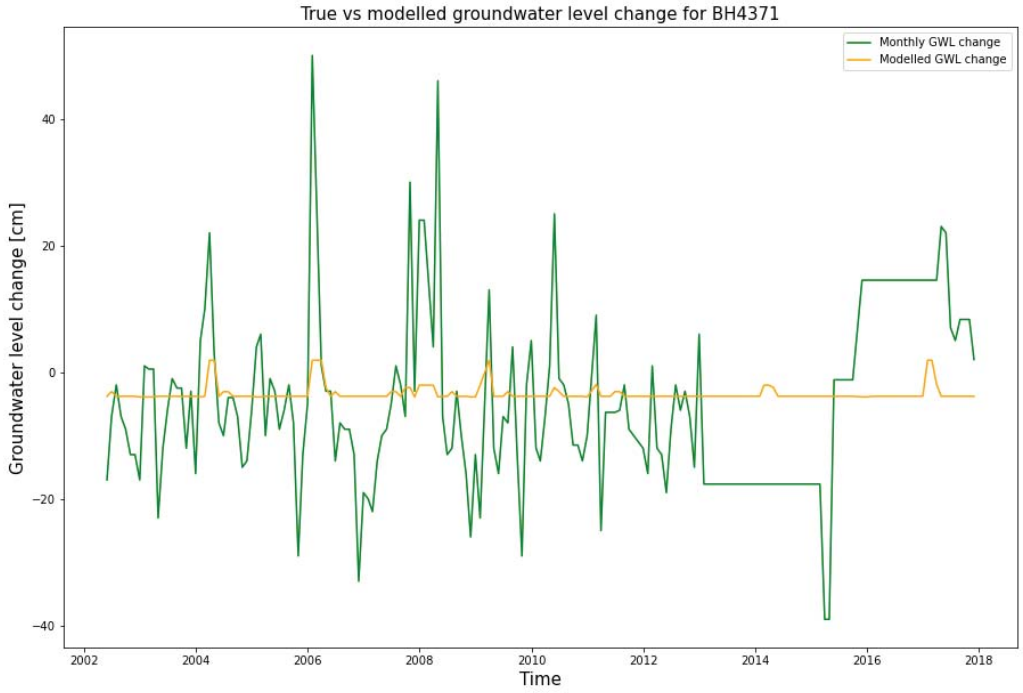


Figure 53: Borehole in the Ramotswa section of the aquifer

Figure 54 displays the rasterized mean absolute error across the study area. There is an overall, low mean absolute error across the study area, with only regions in the east and small portion in the centre of the study area displaying large mean absolute errors. In addition, the regions that display large decreases in groundwater levels coincide with regions with the low overall mean absolute error. This suggests that indeed significant groundwater level declines have occurred in parts of the aquifer. It further highlights the need for investigations into boreholes in the east of the study area indicating rather extreme changes in groundwater level month on month.

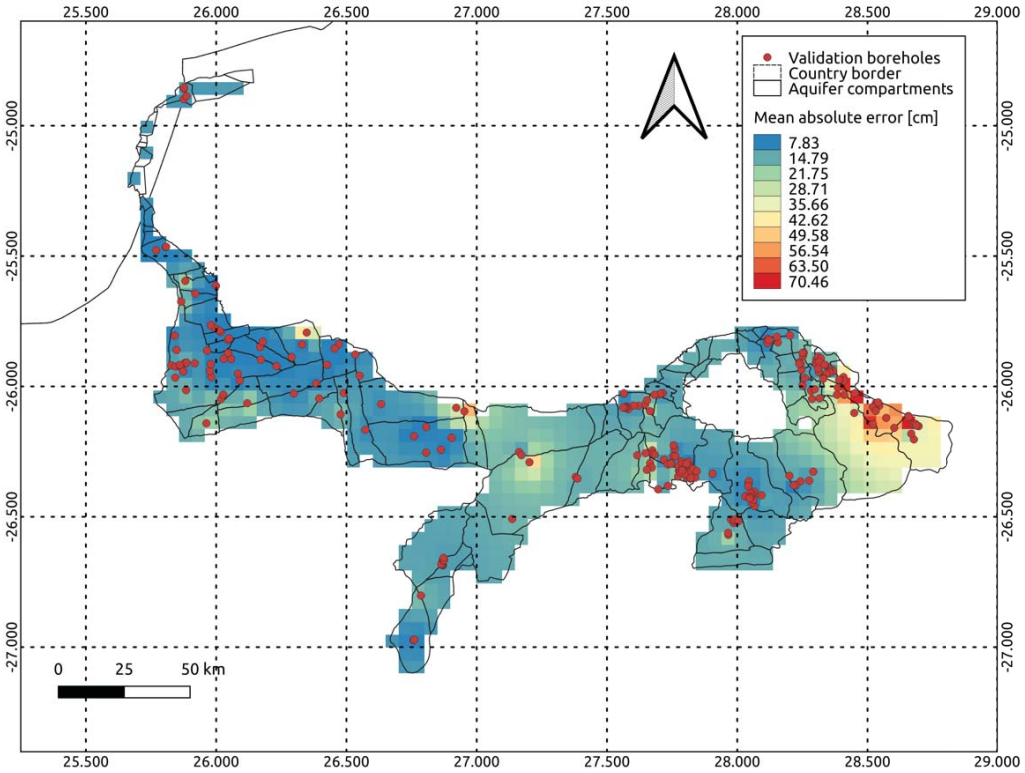


Figure 54: Map of the mean absolute error across the study area. Boreholes used in the validation of the mode are also shown.

6.5. Conclusion

A gradient boosting decision tree machine learning model was developed and implemented in the case study areas. The model was designed, trained and calibrated to predict groundwater level changes based on a set of hydroclimatic, land surface and hydrogeological variables. In addition, by training and validating on different sets of boreholes, the model is robust in terms of spatial distribution. Hence, it can be applied to various localities within the aquifer. The model, once calibrated and tested, produces a MAE of ~18 cm. When compared to results of other studies such as (Seyoum et al., 2019), the results appear to be in-line. However, when comparing the true values vs predicted values, there are issues with predicting extreme values, as the model tends to under-predict extreme values. This may be a result of the non-linearity associated with extreme episodic events, as well as the impacts associated with abstraction, which are not included in the model. In addition, the inherent limitations associated with the temporal and spatial distribution of the groundwater level datasets in capturing the aquifer system, reduces the effectiveness of the model in real world scenarios (i.e. the model can only perform as well as the data allow). None the less, in the current state the model has provided information on the lowering of groundwater levels, suggesting that indeed there is potential to its application in groundwater management scenarios.

In this regard, further experimentation and analysis is required, with additional refinement of the model to ensure robustness and accuracy in real world deployment.

Feature importance was also evaluated in this application. Out of all the features used, the GRACE-derived groundwater storage anomaly, the soil moisture and the land surface temperature are the most important features in terms of predicting groundwater levels. This implies a greater connection between these variables' response to groundwater levels. This highlights the importance of GRACE data in understanding regional groundwater levels, and that GRACE also can play a role in modelling local groundwater level changes.

Finally, the calibrated and validated model was applied to a typical real-world scenario, which in this case is the production of high-resolution groundwater level change maps in the Dolomites Aquifer. The validation of these maps produced a MAE of ~ 17 cm. In some cases, the model incorrectly predicted the direction of change (i.e. an increase vs a decrease in groundwater levels). While in general the MAE for majority of the boreholes is relatively low, there are a number of boreholes with large prediction errors that are skewing the overall MAE. It must however be noted that the validation of the maps was done with point data compared to the gridded maps. It may be more appropriate to validate against average groundwater level changes per grid cell.

For the Shire Valley TBA, a machine learning model was developed for a single borehole (Ngabu). The model was setup to predict the monthly groundwater level changes based on the selected features. In this case the limited spatial and temporal distribution prevents a similar application as in the Ramotswa/NW Dolomites. The model can thus be only used to predict groundwater level change for the Ngabu borehole. However, the limited data prevents the model from being tested. In conclusion it is difficult to assess the validity and potential usefulness of this model, and the regional scale data that is used to develop the model.

7. Reflections on the learning opportunities linked to the project

7.1. Challenges encountered

The following is a list of lessons learnt during the processing and application of this work. This lessons not only focus on the specific methods, but also within the greater context of the project.

- Large datasets are challenging to manipulate on standard desktop machines. For example, the data used in this application (just covering the study area) totals approximately 17-18 GB. This requires an equivalently sized machine memory allotment to process. Techniques such as chucking, parallel processing, and blocked algorithms can be used to overcome this challenge. However, this is just a fraction of data available for SADC. Drawing value from these data will require resources beyond single machines.
- Data cleaning and validation is an important step during pre-processing. The data products available are not always in a "ready-to-use" format. The presence of outliers, duplicate values, inappropriate fill values in raster and meta-data affect the analysis results if not properly accounted for.
- The data from remote sensing and simulated data from models are not always robust across time and space. Many caveats exist for specific datasets that must be accounted for. GRACE data are plagued by errors during post-processing of Level-1 data. This includes instrument issues, satellite technical issues and analysis errors. In addition, resolving the Earth's mass changes is not a perfect science. For example, higher latitudes, the effects of coastlines and regions of extensive de-glaciation must be analysed with care. In this context GRACE data do not perfectly represent the parameter being observed. The same can be said of modelled data, like GLDAS, which have issues of misrepresenting inter-annual changes in land surface states.
- The resolution of GRACE data influences the validity of small-scale investigations. The native resolution of GRACE is close to 300 km. The finer resolution products mentioned in this report are produced through resampling, provided by the data creators. Hence, attempting to analyse small scale signals in GRACE groundwater storage introduces additional uncertainty into the model.

- Machine learning models are only as good as the data that are fed into them. Data pre-processing and validation are fundamental steps, to ensure model representativeness and accuracy. This process often requires more time than the design and development of the machine learning model itself.
  - This could be the cause of modelling errors in the above analysis. Specifically, the data aggregation step may be flawed considering archetypal lag times in recharge responses. That is not adjusted for during this step.
  - It is thus apparent that more care and effort is needed to improve the quality and quantity of data both on the ground and from other big data sources.
- The machine learning models developed in this research are unambiguous. Meaning that they are designed to provide predictions on a single target variable or predictant. In this case, they can only be used to predict 30-day groundwater level changes in the respective study area. They cannot be generalized to model other properties of the aquifer.

7.2. Project Recommendations

7.2.1. Policy and transboundary aquifer management scenarios

Groundwater depletion has increased considerably in major aquifer systems of the world. Continuing groundwater over-exploitation in such regions is unsustainable over multiple generations. In addition, climate variability and change influence groundwater systems both directly through changes in replenishment by recharge and indirectly through changes in groundwater use. There is an expectation of increased occurrence of extreme hydrological events requiring the need to develop improved adaptive capacity to flooding and drought.

The TBAs of the SADC are associated with 'high-importance' aquifers making them crucial to improve water security and international cooperation among SADC Member States. Most investigations, however, have focussed on regional understanding of TBAs, but as far as groundwater is concerned, management and operational decisions are made at national and local levels which are more useful for sustainable management of the resource. Using BDA, the various sources of data and information can be integrated to produce new knowledge on local groundwater conditions. This would allow more informed decision-making.

Firstly, it is essential to conceptualize the purpose of BDAs. This will facilitate deciding on areal domain, time periods, what data to use as well as the methodological approach. The conceptualization involves the identification of the common problems occurring in the transboundary aquifer and the definition of sustainability goals. The following recommendations are made:

- For sustainable groundwater management, sustainability indicators are recommended based on the California Department of Water Resources (2017) – namely the effects caused by groundwater conditions occurring throughout the transboundary aquifer that, when significant and unreasonable, become undesirable results. The undesirable effects are listed in Table 9.
- Response to undesirable results implies adopting quantitative metrics and setting minimum thresholds. Minimum thresholds need to be quantified to represent the groundwater conditions causing an undesirable result(s) in the basin, similar to the Resource Quality Objectives as defined in the National Water Act (1998). Metrics and thresholds are site-specific, and they must be established through local studies and stakeholder engagement.
- Establishing minimum thresholds for groundwater levels at a given representative monitoring site needs to take into consideration historical groundwater conditions in the basin, the average, minimum, and maximum borehole depths of municipal, agricultural, and domestic boreholes, and the potential impacts of changing groundwater levels on groundwater dependent ecosystems.

Table 9: Undesirable effects in transboundary aquifers and associated metrics.

Undesirable impacts	**Sustainability and threshold metrics**
Chronic lowering of groundwater levels indicating a significant and unreasonable depletion of supply	Groundwater level Volume and rate of abstraction
Significant and unreasonable reduction of groundwater storage	Total volume of groundwater storage
Significant and unreasonable seawater intrusion (or saline intrusion)	Chloride concentration isocontour Groundwater level
Significant and unreasonable degraded water quality, including the migration of contaminant plumes that impair water supplies	Migration of plumes Isocontour of contaminants
Significant and unreasonable land subsidence that substantially interferes with surface land uses	Rate and extent of land subsidence

Undesirable impacts	Sustainability and threshold metrics
Depletions of interconnected surface water that have significant and unreasonable adverse impacts on beneficial uses of the surface water	Volume and rate of surface water depletion

7.2.2. Recommendations on Big Data infrastructure

To extract usable information from the variety of sources of information to the benefit of researchers, water managers and decision-makers, these data need to be processed through capable Big Data platforms, transformed into useful information, and packaged into tools that facilitate interpretation and decision-making on the ground. The following are proposed recommendations that can be implemented to develop a Big Data platform that can support data driven applications of sustainable groundwater management in SADC. Recommendations are summarized related to the BDAs infrastructure, the hardware and software requirements as well database and storage requirements.

- The TBA analytics architecture presented in Figure 12 proposes the integration of many structured and unstructured data and sources by designing and implementing local clouds for storing local data sets and interconnecting these local clouds through a federated cloud infrastructure designed to present to the users a unified view of a transboundary database management scheme. Tools that are required for this operation include: i) data ingestion operators for ingesting data in multiple formats and curating the data in a data lake; ii) data transformation operators to validate data and apply complex transformation; iii) feature extractors or query tools to facilitate the retrieval of data.
- A federated cloud storage infrastructure linking independent local cloud storages and databases is recommended because it can provide access to data providers through publish-subscriptions services that by-pass the need for local data storage and overcome the limitations associated with client-server service models
- Two deployment models are proposed for the federated platform: a centralized deployment with CSIR hosting the workflows of the different research groups and meeting their storage and processing demands, and a hybrid deployment where i) software containers are used to move the processing to the data residing in different research groups' data repositories, thus meeting

transboundary requirements while ii) for non-constrained data requirements, CSIR plays the role of data storage and processing unit and archiving.

- The volume of data and information produced by novel technologies makes it essential to develop database and storage facilities. This is particularly true for Earth observation data where users are facing challenges of exponentially growing data archives due to the increasing number of sensors and products that are making available layers of information at increasingly fine space resolution and at higher frequency (Lewis et al., 2017).
- The nature of data, often unstructured, includes different sources and formats, such as web-portals, web-based geographic information system (GIS) tools, password protected portals, cloud storage, portable storage devices, hardcopy maps, theses, reports, newsletters, documents, videos and podcasts (Hu et al., 2014). This makes it also imperative to develop database and storage facilities that can manage heterogeneous datasets. It should also allow easy access and provide user-friendly meta-data content.

The hardware and software requirements for a TBA project may be associated with providing two different services models or a combination of both: i) infrastructure as a service model of service delivery (Table 10) ii) a platform as a service (Table 11).

Table 10: Infrastructure As A Service – Compute Modules

SN	Type	Description	Specification	Quantity
1	General Purpose	General day-to-day use, low end	Small scale ARM like Processors: 1 to 16 vCPU cores, 2 to 32GiB Memory, 10Gbps Network, EBS storage.	15
		General day-to-day use but high end	Intel like Processors: 2-64 vCPU, 8-258GiB Memory, up to 10Gbps Network, dedicated EBS storage	10
2	Compute Optimized	High performance computing	Intel Xeon E5 v3 like Processors, 2-36 vCPU, 3 to 60GiB Memory, EBS storage	10
3	Accelerated Computing	Machine Learning & Gaming optimized	1-8 Nvidia GPUs, 8-96 vCPU, 61-768GiB Memory, 16-256 GPU Memory, up to 100Gbps Network, dedicated EBS storage	5
4	Storage Optimized	NoSQL db, data warehousing	Intel Xeon E5 v4 like Processors: 2-72 vCPU, 15-512 GiB Memory, 1x 475GB to 8x 2TB SSD storage, up to 25Gbps network	10
				50

Table 11: Platform As A Service (set of Open-source tools) – Operating System: Debian-based (Free)

SN	Description	Installed / Available Tools
1	IoT	RabbitMQ, Mosquitto, Arduino, Node-RED, M2MLabs, ThingsBoard
2	Machine Learning	Weka, Scikit, TensorFlow,
3	Big Data	Apache Kafka,
4	Database	MongoDB, MySQL, MariaDB, PostgreSQL, Apache Hive
5	Analytics	Tableau, PowerBI, Apache Spark
6	NLP	Semantria, Trackur
7	Programming	Eclipse IDE, VSCode, Python, Java, R, MatLab/Octave
8	Web Development	Joomla, WordPress, CSS, HTML, JavaScript, PHP, Apache Cordova
9	App Servers	XAMP/WAMP, Apache Hadoop, JDK
10	Others	Docker, Cloudera, Kubernetes, OpenStack Suite

The processing of such large volumes of aquifers' data can benefit from existing high-performance computing (HPC) infrastructures which are able to store massive datasets collected from different water data sources and meet the diverse processing requirements demanded from water applications. At its onset, HPC was commonly associated with scientific computing for scientific research using super-computers and computer clusters. Nowadays, HPC has evolved toward the relatively more recent cloud computing model that builds on decades of research in virtualization, distributed computing, utility computing, networking, web and software service.

Cloud technologies have become the technologies of choice for solving large scale data/compute intensive problems as they present undeniable advantages over traditional HPC. However, only a few African countries which are involved in TBAs research can afford world-class processing infrastructures and building such facilities is often time consuming thus discouraging for multiple deployments. It is therefore necessary to strengthen the capabilities in SADC countries to make use of these technological opportunities.

7.2.3. Recommendations on applications of Big Data Analytics

During this project, research was conducted to test and demonstrate the application of BDAs at two case study sites, namely the Zeerust/Lobatse/Ramotswa Dolomite Basin Aquifer and the Shire Valley Alluvial Aquifer. The main focus of the project was to apply BDAs primarily for the purpose of downscaling big data (spatial and time series) to the level where local decisions can be taken on water resource assessment, planning and management. Several recommendations emanated from the case studies are summarized below for the data pre-processing, machine learning development and application, and decision-making phases.

7.2.3.1. Data and processing

- Site selection and area domain are crucial in order to make use of sources and data in a consistent manner. Satellite-derived information (e.g. Gravity Recovery and Climate Experiment GRACE data) can be expected to be at much coarser resolution than data collected on the ground for the purpose of localized groundwater management. Pre-processing is therefore required.
- The size of the investigated aquifers must be consistent in relation to the regional data used. For example, in the Ramotswa case study, it was realized that the aquifer was not of great enough areal extent to allow sufficient coverage of remote sensing regional data (GRACE data at a resolution of 110 km, or 1^{o} x 1^{o}). In order to overcome this limitation, the case study area was expanded to include the dolomite aquifers extending into the North-West and Gauteng provinces of South Africa.
- A wealth of large-scale data and information has been recently generated thanks to novel technologies such as satellite remote sensing and Global Circulation Models (GCM). This information is generally available in spatial format and with time series that span up to a few decades. Time series are often available for near-real time analysis and decision-making. This is highly valuable especially for areas were *in situ* information is scarce, however the spatial resolution is often too coarse for applications to localized water management (e.g. wellfields and individual boreholes). It is therefore recommended that BDAs methodologies and approaches make use of the opportunity to add value to these novel technologies and available large-scale datasets.
- Large scale data are stored in commonly used NetCDF (.nc extension) or HDF5 formats. These are multidimensional data storage containers, capable of storing different types of data. Data need to be extracted from these storage

containers to be plotted in a GIS environment. The process can be automated by writing scripts using common programming languages, e.g. Python.

- Processing large volumes of data may require appropriate machine memory allotments. Techniques such as chucking, parallel processing, and blocked algorithms can be used to overcome this challenge.
- One of the most important concepts when collecting and analysing Big Data is dealing with the uncertainty. Machine learning models are only as good as the data that are fed into it. Possible sources of uncertainty are:
  - Missing data, poor data capturing, measurement errors
  - Inherent errors to raw data (e.g. sensor, satellite, instrumental errors)
  - Large, highly heterogeneous, multi-dimensional datasets (unstructured, inconsistent, incomplete, and noisy data)
  - The scaling factors used to reduce signal loss in GRACE data in the post-processing phase
  - Extraction and re-sampling to render consistency between datasets at different spatial resolution
  - Data aggregation may be flawed considering archetypal lag times in recharge responses
- Possible means of reducing uncertainties are:
  - Selection of proper techniques when dealing with Big Data
  - Data pre-processing, cleaning and validation are fundamental steps, often requiring more time than the design of the model itself
  - Traditional methods are outlier detection, removal of duplicates, missing data detection, handling and unifying datasets, ranges check
- More complex techniques that can be employed are Probability theory, Bayesian theory, Shannon's entropy, Rough set theory, Fuzzy set theory.
- A pre-processing stage is required for ground data (local scale). This consists in formatting data, combining data from different sources (e.g. in spreadsheets), removing duplicates and missing records, and data quality control. Proper care must be taken to ensure the quality of data is adequate.
- Data cleaning and quality control are an important step during pre-processing of regional scale datasets as well. The data products available are not always in a "ready-to-use" format. The presence of outliers, duplicate values, inappropriate fill values in raster files affect the analysis results if not properly accounted for.
- Substantial pre-processing is required in order to render consistency to the different sources of data:
  - Data need to be consistent in terms of spatial and temporal resolution

- Missing data are likely to occur and appropriate methodologies need to be applied for data patching or removal of missing records
- Realistic limits and constraints need to be set in order make optimal use of existing data and information (e.g. omit monitoring points that have too little data; set realistic periods of groundwater level response to rainfall events, etc.)
- Consistency in units needs to be checked.

– Sufficient ground data (both spatial and temporal) are required in order to train models and algorithms as predictive tools for the specific application and problems to be solved. Efforts must be made to ensure the amount of data on groundwater is increased spatially and temporally. To this end, modern smart sensor technologies, and IoT devices are encouraged.

7.2.3.2. Application of Big Data analytics

A large basket of BDAs techniques is available that can be used to solve specific problems. These are presented in Table 11.

Table 12: Summary classification of BDAs techniques

Techniques	**Examples of computational methods**
Statistics	Descriptive statistics, Regression, Correlation, Factor analysis, Clustering, Hypothesis testing, Probabilistic statistics
Data mining	SQL queries, Machine Learning, Statistics, Feature selection
Artificial intelligence (AI)	Statistical learning, Optimization methods, Deep learning
Machine learning (subset of AI)	Artificial neural networks, Support vector machine, Random forest, K mean clustering, Natural language processing
Uncertainty analysis	Data cleaning, Probability theory, Bayesian theory, Shannon's entropy, Rough set theory, Fuzzy set theory
Visualization	Tables, Graphs, Images, Feature extraction, Geometric modelling

– BDAs can be specifically applied to address downscaling of regional data for catchment scale management. A large number of approaches, tools and techniques were reviewed in the literature. Amongst the approaches, dynamic and statistical approaches were considered:
  - Statistical approaches were traditionally preferred as they are less complex to develop and because of low computational costs

- Dynamic approaches provide more mechanistic and physically-sound results and they are better suited to describe non-linear relationships between variables in complex systems (Deliverable 2b of the project), but are complex to develop, requiring high computational resources
- BDAs can be categorised as statistical approach with the advantage of being able to model non-linear relationships and handle large streams of heterogeneous data

- Unsupervised machine learning algorithms are better suited for grouping and classifying, whereas supervised machine learning algorithms are an excellent predictive (regression) tool. Supervised machine learning algorithms are therefore recommended for downscaling applications.
- Literature has shown that many machine learning algorithms can be applied to downscaling of regional datasets. The most appropriate downscaling techniques were selected based on the literature review and discussions with IBM Africa. In the particular case study of this project, the machine learning algorithm chosen for downscaling regional groundwater data was the Gradient Boosting Decision Trees (GBDT).
- For downscaling regional groundwater data, machine learning methods such as the Gradient Boosting Decision Trees (GBDT) can be used. GBDT are versatile algorithms that can be used for regression, classification and ranking. Decision trees have advantages of being easy to interpret, they handle missing values, they are not influenced by outliers, they do not need *a priori* information, and they can handle irrelevant features. In addition, these models are highly efficient and accurate.
- The implementation of GBDT can be executed using the LightGBM module[16] in a Python environment (other languages are also available).
- Feature engineer is a critical step in the development of a machine learning model. Feature engineering is a formulation of a set of input and output features (predictors and predictants, respectively). Data need to be converted to sets of features that can accurately predict the output feature (predictant).
- ***Groundwater levels*** are expected to be the most common predictant in relation to sustainable groundwater management. However, depending on the problem-solving application, predictants can also be ***groundwater quality, land subsidence or other variables***. National databases such as the National Groundwater Archive are common sources of this information.

[16] https://lightgbm.readthedocs.io/en/latest/#

- In this case study application, the recommended predictors in downscaling regional data for groundwater level modelling are summarized in Table 12.
- It is useful to develop a baseline model, e.g. relationship between precipitation (predictor) and groundwater level (predictant), and subsequently add incrementally the various input variables to assess the sensitivity of groundwater level to each predictor.
- Proper care must be taken during the interpretation of machine learning model development and application. It is not always as easy as selecting based on training or validation score. Additional scrutiny of the training process and model performance results are required to ensure models are rigid and applicable.
- When datasets contain a limited number of samples, it is recommended to use a pixel-based rather than a sample-based approach, and consider several pixels as an input to predict the future value of a groundwater variable pixel

Table 13: Summary of recommended predictors for case study area used in this project (independent variables, model inputs), and sources of information.

Independent variable (predictor)	**Source**
Precipitation	ECMWF ERA5-Land https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview
Groundwater storage anomaly	Release 06 version 03 of GRACE Tellus mission and GRACE-FO Level-3 monthly land surface mass changes based on spherical harmonics, https://podaac.jpl.nasa.gov/GRACE?sections=about%2Bdata GLDAS NOAH, https://podaac-tools.jpl.nasa.gov/drive/files/allData/tellus/L3/gldas_monthly/netcdfRelease 5 version 4 NCAR CLM scaling factors for the spherical harmonics to reduce the signal loss during pre-processing caused by filtering data to remove noise and correlated errors.
Evapotranspiration	ECMWF ERA5-Land, https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview
Soil moisture	ECMWF ERA5-Land, https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview
Run-off	ECMWF ERA5-Land, https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview

Independent variable (predictor)	Source
Land surface temperature	ECMWF ERA5-Land, https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview
Aquifer compartments	Based on hydrogeological boundaries
Aquifer type	SADC Hydrogeology map (2010)
Landcover	ESA Climate Change Institute Land Cover product, https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-land-cover?tab=overview
Dependant variable (predictant)	**Source**
Groundwater level anomalies	Groundwater level data, RIMS, HYDSTRA

7.2.4. Decision-making

The main purpose of BDAs is to transform large volumes of heterogeneous data into usable information and knowledge to support decision-making by water managers on the ground. Decision-making tools need to be tailored according to users' needs, especially at local scale, in the form of dashboards, early warning systems, maps, graphs, tables, web- and other user-friendly applications. The huge amount of data does not make it viable to analyse data manually. This is a limitation in terms of near-real time decision-making. It is therefore recommended to automate, where possible, the results relevant to some decisions in order to promote and facilitate near-real time decision-making.

7.3. Further research recommendations and conclusion

Some requirements for further research are:

– Groundwater level simulation using BDAs and remote sensing has been documented. However, groundwater quality analysis using BDAs and remote sensing has not been explored. It is recommended that further research addresses this gap. For example, groundwater quality is known to be dependent on various land and sub-surface activities (e.g. land use, mining and geology). BDAs and remote sensing could help in better understanding and

managing these linkages. In addition, predictive analytics can be used to map and to predict groundwater quality in 3 dimensions, using *in situ* data and machine learning algorithms (Ransom et al., 2017).

- Many of the uncertainties in the analytical results are driven by inconsistencies in spatial and temporal resolution of ancillary data as well as errors in the various pre-processing and post-processing steps to render data spatially consistent. Future investments in transboundary contexts should take into consideration spatial and temporal resolutions required in data collection to solve specific problems.
- The technological requirements to store and process large heterogeneous volumes of data often require dedicated systems beyond the capabilities of conventional desktop systems. This is particularly a problem in many SADC Member States that do not have the computational capabilities to facilitate Big Data approaches. Furthermore, ingesting huge volumes of data has implications with the network speed required to move and process Big Data. High performance computing (HPC) infrastructures and the design of federated platforms is therefore fundamental in extracting the best value from BDAs.
- The transparency of data sharing across international boundaries is not always common amongst individual states. Data ownership and data access is often restricted to certain individuals or institutions. The institutional barriers and management practices employed by member states are not always aligned with each other. The consequence is that this can ultimately affect the sustainable management of groundwater. Management challenges cover issues such as privacy, governance, institutionalization, security, among others, and this need to be addressed *a priori.*
- In South Africa, comprehensive and integrated groundwater-related databases at national level (that encompass different scales with additional added-value of Internet-based tools for data processing, interpretation and groundwater management support) do not exist. The development of a centralized data repository would be highly beneficial based on the recommendation of the National Water Security Framework (2019) and the Data Storage Solution Online Workshops that took place on 20-24 April 2020. The benefits of such repository would be two-fold:
  - To collate existing groundwater-related information into a centralized repository that would be easy to access, generate and extract valuable information in support of a quick response in groundwater management

- To develop tailor-made computerized tools and software to add value to the existing information and support decisions in groundwater management and planning

– Such data repository would require a strong involvement of data scientists, geohydrologists and potential users of the information since its inception. Much of the design should be based on the users' needs. In particular, the development of value-adding functionalities such as dashboards, early warning systems and visualization tools tailor-made to specific problems, users and transboundary aquifers would be highly advantageous.

References

Adamala S (2017) An Overview of Big Data Applications in Water Resources Engineering. Machine Learning Research 2:10-18. https://doi.org/10.11648/j.mlr.20170201.12

Aji A, Wang F, Vo H, et al. (2013) Hadoop gis: A high performance spatial data warehousing system over mapreduce. Proceedings of the VLDB Endowment 6:1009-1020. https://doi.org/10.14778/2536222.2536227

Alarabi L, Mokbel MF, Musleh M (2018) ST-Hadoop: a MapReduce framework for spatio-temporal data. GeoInformatica 22:785-813. https://doi.org/10.1007/s10707-018-0325-6

Ali A, Qadir J, Rasool R ur, et al. (2016) Big data for development: applications and techniques. Big Data Analytics 1:2. https://doi.org/10.1186/s41044-016-0002-4

Almeida F (2018) Big Data: Concept, Potentialities and Vulnerabilities. Emerging Science Journal 2:1-10. https://doi.org/10.28991/esj-2018-01123

Altchenko Y, Genco A, Pierce K, et al. (2017) Resilience in the Limpopo Basin : The potential role of the transboundary Ramotswa aquifer. United States Agency for International Development, Pretoria

Altchenko Y, Villholth KG (2013) Transboundary aquifer mapping and management in Africa: a harmonised approach. Hydrogeology Journal 21:1497-1517. https://doi.org/10.1007/s10040-013-1002-3

Baqa SS (2017) Groundwater Recharge Assessment in the Upper Limpopo River Basin: A Case Study in Ramotswa Dolomitic Aquifer. University of the Witwatersrand

Baumann P, Mazzetti P, Ungar J, et al. (2016) Big Data Analytics for Earth Sciences: the EarthServer approach. International Journal of Digital Earth 9:3-29. https://doi.org/10.1080/17538947.2014.1003106

Becker T, Curry E, Jentzsch A, Palmetshofer W (2016) New horizons for a data-driven economy: Roadmaps and action plans for technology, businesses, policy, and society. In: New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe. Springer International Publishing, pp 277-291

Blumenfeld J (2018) EOSDIS DAACs Celebrate Milestones of Service to Global Data Users. https://earthdata.nasa.gov/learn/articles/tools-and-technology-articles/daac-overview-and-milestones. Accessed 1 Dec 2020

Bonner S, Kureshi I, Brennan J, Theodoropoulos G (2017) Chapter 14 – Exploring the Evolution of Big Data Technologies. In: Mistrik I, Bahsoon R, Ali N, et al. (eds) Software Architecture for Big Data and the Cloud. Morgan Kaufmann, Boston, pp 253-283

Bumby AJ, Eriksson PG, Catuneanu O, et al. (2012) Meso-Archaean and Palaeo-Proterozoic sedimentary sequence stratigraphy of the Kaapvaal Craton. Marine and Petroleum Geology 33:92-116. https://doi.org/10.1016/j.marpetgeo.2011.09.010

Cairncross B (2001) An overview of the Permian (Karoo) coal deposits of southern Africa. Journal of African Earth Sciences 33:529-562. https://doi.org/10.1016/S0899-5362(01)00088-4

Catuneanu O, Wopfner H, Eriksson PG, et al. (2005) The Karoo basins of south-central Africa. Journal of African Earth Sciences 43:211-253. https://doi.org/10.1016/j.jafrearsci.2005.07.007

CDWR (2017) DRAFT Sustainable Management Criteria

CGS (2008) 1:2000000 chronostratigraphic map of South Africa. Pretoria, South Africa

Chairuca L, Chintengo P, Ebrahim G, et al. (2019) Transboundary diagnostic analysis of the Shire River Aquifer System. Southern African Development Community Groundwater Management Institute, Bloemfontein

Chalh R, Bakkoury Z, Ouazar D, Hasnaoui MD (2015) Big data open platform for water resources management. In: 2015 International Conference on Cloud Technologies and Applications (CloudTech). pp 1-8

Chen CLP, Zhang CY, Philip Chen CL, Zhang CY (2014) Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences 275:314-347. https://doi.org/10.1016/j.ins.2014.01.015

Chen L, Wang L (2018) Recent advance in earth observation big data for hydrology. Big Earth Data 2:86-107. https://doi.org/10.1080/20964471.2018.1435072

Cinquini L, Crichton D, Mattmann C, et al. (2014) The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data. Future Generation Computer Systems 36:400-417. https://doi.org/10.1016/j.future.2013.07.002

Cobbing J (2018) An updated water balance for the Grootfontein aquifer near Mahikeng. Water SA 44:. https://doi.org/10.4314/wsa.v44i1.07

Cobbing J, Eales K, Rossouw T (2016) The path to successful water user associations in the north west dolomite aquifers. Pretoria

Cobbing JE, de Wit M (2018) The Grootfontein aquifer: Governance of a hydro social system at Nash equilibrium. South African Journal of Science 114:1-7. https://doi.org/10.17159/sajs.2018/20170230

CSIR (2003) Protection and Strategic Uses of Groundwater Resources in the Transboundary Limpopo Basin and Drought Prone Areas of the SADC Region-Groundwater Situation Analysis in the Limpopo River Basin

Cui Y, Chen X, Gao J, et al. (2018) Global water cycle and remote sensing big data: overview, challenge, and opportunities. Big Earth Data 2:282-297. https://doi.org/10.1080/20964471.2018.1548052

DWAF (2006) A Strategy for Water Allocation Reform in South Africa. Pretoria, South Africa

Elbeih SF (2015) An overview of integrated remote sensing and GIS for groundwater mapping in Egypt. Ain Shams Engineering Journal 6:1-15. https://doi.org/10.1016/j.asej.2014.08.008

Eldawy A, Mokbel MF (2015) SpatialHadoop: A MapReduce framework for spatial data. In: 2015 IEEE 31st International Conference on Data Engineering. pp 1352-1363

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. Journal of Animal Ecology 77:802-813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Fan J, Han F, Liu H (2014) Challenges of Big Data analysis. National Science Review 1:293-314. https://doi.org/10.1093/nsr/nwt032

Faroukhi AZ, El Alaoui I, Gahi Y, Amine A (2020) Big data monetization throughout Big Data Value Chain: a comprehensive review. Journal of Big Data 7:1-22. https://doi.org/10.1186/s40537-019-0281-5

Gaffoor Z, Pietersen K, Jovanovic N, et al. (2020) Big Data Analytics and Its Role to Support Groundwater Management in the Southern African Development Community. Water 2020, Vol 12, Page 2796 12:2796. https://doi.org/10.3390/W12102796

Gandomi A, Haider M (2015) Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35:137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Grimason AM, Beattie TK, Morse TD, et al. (2013a) Classification and quality of groundwater supplies in the Lower Shire Valley, Malawi – Part 2: Classification of borehole water supplies in Chikhwawa, Malawi. Water SA 39:573-582. https://doi.org/10.4314/wsa.v39i4.17

Grimason AM, Morse TD, Beattie TK, et al. (2013b) Classification and quality of groundwater supplies in the Lower Shire Valley, Malawi – Part 1: Physico-chemical quality of borehole water supplies in Chikhwawa, Malawi. Water SA 39:563-572. https://doi.org/10.4314/wsa.v39i4.16

Guo H (2017) Big Earth data : A new frontier in Earth and information sciences Big Earth data : A new frontier in Earth and information. Big Earth Data 1:4-20. https://doi.org/10.1080/20964471.2017.1403062

Gupta P (2017) Decision Trees in Machine Learning. https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052. Accessed 20 Oct 2020

Hariri RH, Fredericks EM, Bowers KM (2019) Uncertainty in big data analytics: survey, opportunities, and challenges. Journal of Big Data 6:44. https://doi.org/10.1186/s40537-019-0206-3

Jony RI, Rony RI, Rahat A, Rahman M (2016) Big Data Characteristics, Value Chain and Challenges. In: 1st International Conference on Advanced Information and Communication Technology. Chittagong, Bangladesh

Ke G, Meng Q, Finley T, et al. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Guyon I, Luxburg U V, Bengio S, et al. (eds) Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp 3146-3154

Kiparsky M, Milman A, Owen D, Fisher A (2017) The Importance of Institutional Design for Distributed Local-Level Governance of Groundwater: The Case of California's Sustainable Groundwater Management Act. Water 9:755. https://doi.org/10.3390/w9100755

Kitchin R, McArdle G (2016) What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society 3:205395171663113. https://doi.org/10.1177/2053951716631130

Klein LJ, Marianno FJ, Albrecht CM, et al. (2015) PAIRS: A scalable geo-spatial data analytics platform. In: 2015 IEEE International Conference on Big Data (Big Data). pp 1290-1298

Koglin N, Zeh A, Frimmel HE, Gerdes A (2010) New constraints on the auriferous Witwatersrand sediment provenance from combined detrital zircon U-Pb and Lu-Hf isotope data for the Eldorado Reef (Central Rand Group, South Africa). Precambrian Research 183:817-824. https://doi.org/10.1016/j.precamres.2010.09.009

Kurze T, Klems M, Bermbach D, et al. (2011) Cloud Federation. In: The 2nd International Conference on Cloud Computing, GRIDs and Virtualisation. Rome, Italy

Laney D (2001) 3D Data Management: Controlling Data Volume, Velocity, and Variety | BibSonomy. https://www.bibsonomy.org/bibtex/742811cb00b303261f79a98e9b80bf49. Accessed 28 Sep 2020

Lee I (2017) Big data: Dimensions, evolution, impacts, and challenges. Business Horizons 60:293-303. https://doi.org/10.1016/j.bushor.2017.01.004

Lee S, Hyun Y, Lee M-J (2019) Groundwater Potential Mapping Using Data Mining Models of Big Data Analysis in Goyang-si, South Korea. Sustainability 11:1678. https://doi.org/10.3390/su11061678

Leonard LC (2017) Chapter One – Web-Based Behavioral Modeling for Continuous User Authentication (CUA). In: Memon AMBT-A in C (ed). Elsevier, pp 1-44

Lewis A, Oliver S, Lymburner L, et al. (2017) The Australian Geoscience Data Cube – Foundations and lessons learned. Remote Sensing of Environment 202:276-292. https://doi.org/10.1016/j.rse.2017.03.015

Lin Y, Jun Z, Hongyan M, et al. (2018) A Method of Extracting The Semi-structured Data Implication Rules. Procedia Computer Science 131:706-716. https://doi.org/10.1016/j.procs.2018.04.315

Lu S, Shao X, Freitag M, et al. (2016) IBM PAIRS curated big data service for accelerated geospatial data analytics and discovery. In: Proceedings – 2016 IEEE International Conference on Big Data, Big Data 2016. Institute of Electrical and Electronics Engineers Inc., pp 2672-2675

Manzi M, Hein K, King N, Durrheim R (2013) Neoarchaean tectonic history of the Witwatersrand Basin and Ventersdorp Supergroup: New constraints from high-resolution 3D seismic reflection data. Tectonophysics 590:94-105. https://doi.org/10.1016/j.tecto.2013.01.014

Miro ME, Famiglietti JS (2018) Downscaling GRACE remote sensing datasets to high-resolution groundwater storage change maps of California's Central Valley. Remote Sensing 10:. https://doi.org/10.3390/rs10010143

Modisha RCO (2017) Investigation of the Ramotswa Trandboundary Aquifer Area, groundwater flow and pollution. University of the Witwatersrand

Mohaghegh SD, Gaskari R, Maysami M (2017) Shale Analytics: Making Production and Operational Decisions Based on Facts: A Case Study in Marcellus Shale. Society of Petroleum Engineers Hydraulic Fracturing Technology Conference and Exhibition 23. https://doi.org/10.2118/184822-MS

Monjerezi M, Ngongondo C (2012) Quality of Groundwater Resources in Chikhwawa, Lower Shire Valley, Malawi. Water Quality, Exposure and Health 4:39-53. https://doi.org/10.1007/s12403-012-0064-0

Monjerezi M, Vogt RD, Aagaard P, Saka JDK (2011) Hydro-geochemical processes in an area with saline groundwater in lower Shire River valley, Malawi: An integrated application of hierarchical cluster and principal component analyses. Applied Geochemistry 26:1399-1413. https://doi.org/10.1016/j.apgeochem.2011.05.013

NASA (2002) GRACE Launch

Nasser T, Tariq R (2015) Big Data Challenges. Journal of Computer Engineering & Information Technology 4:4-11

Nijsten G-JJ, Christelis G, Villholth KG, et al. (2018) Transboundary aquifers of Africa: Review of the current state of knowledge and progress towards sustainable development and management. Journal of Hydrology: Regional Studies 20:1-14. https://doi.org/10.1016/j.ejrh.2018.03.004

Niles MT, Hammond Wagner CR (2019) The carrot or the stick? Drivers of California farmer support for varying groundwater management policies. Environmental Research Communications 1:045001. https://doi.org/10.1088/2515-7620/ab1778

Padgavankar MH, Gupta SR (2014) Big Data Storage and Challenges. Journal of Computer Science and Information Technologies 5:2218-2223

Pavelic P, Keraita B, Giordana M (2012) Ground water Availability and Use in Sub-Saharan Africa: A review of 15 countries

Pietersen K, Beekman H (2016) Groundwater Management in the Southern African Development Community. Southern Development Community Groundwater Management Institute, Bloemfontein, South Africa

Pietersen K, Beekman H, Cobbing J, Kanyerere T (2018) Consultancy for capacity needs assessment to determine priority challenges for capacity development initiatives in Member States. Southern African Development Community Groundwater Management Institute, Bloemfontein, South Africa

Rodell M, Chen J, Kato H, et al. (2007) Estimating groundwater storage changes in the Mississippi River basin (USA) using GRACE. Hydrogeology Journal 15:159-166. https://doi.org/10.1007/s10040-006-0103-7

Rodell M, Houser PR, Jambor U, et al. (2004) The Global Land Data Assimilation System. Bulletin of the American Meteorological Society 85:381-394. https://doi.org/10.1175/BAMS-85-3-381

Russom P (2011) Big data analytics. Renton, Washington

SADC (2010) Explanatory Brochure for the South African Development Community (SADC) Hydrogeological Map & Atlas. Gaborone, Botswana

SADC-GMI, IGRAC, IGS (2019a) SADC Framework for Groundwater Data Collection and Data Management. Bloemfontein, South Africa

SADC-GMI, IGRAC, IGS (2019b) State of Groundwater Data Collection and Data Management in SADC Member States. Bloemfontein, South Africa

Sakumura C, Bettadpur S, Bruinsma S (2014) Ensemble prediction and intercomparison analysis of GRACE time-variable gravity field models. Geophysical Research Letters 41:1389-1397. https://doi.org/10.1002/2013GL058632

Save H, Bettadpur S, Tapley BD (2016) High-resolution CSR GRACE RL05 mascons. Journal of Geophysical Research: Solid Earth 121:7547-7569. https://doi.org/10.1002/2016JB013007

Schapire RE (2003) The Boosting Approach to Machine Learning: An Overview. In: Denison DD, Hansen MH, Holmes CC, et al. (eds) Nonlinear Estimation and Classification. Springer New York, New York, NY, pp 149-171

Seyoum WM, Kwon D, Milewski AM (2019) Downscaling GRACE TWSA data into high-resolution groundwater level anomaly using machine learning-based models in a glacial aquifer system. Remote Sensing 11:. https://doi.org/10.3390/rs11070824

Sivarajah U, Kamal MM, Irani Z, Weerakkody V (2017) Critical analysis of Big Data challenges and analytical methods. Journal of Business Research 70:263-286. https://doi.org/10.1016/j.jbusres.2016.08.001

Sun Z, Huo Y (2019) The Spectrum of Big Data Analytics. Journal of Computer Information Systems 1-9. https://doi.org/10.1080/08874417.2019.1571456

Swenson S, Chambers D, Wahr J (2008) Estimating geocenter variations from a combination of GRACE and ocean model output. Journal of Geophysical Research: Solid Earth 113:. https://doi.org/10.1029/2007JB005338

Swenson S, Wahr J (2006) Post-processing removal of correlated errors in GRACE data. Geophysical Research Letters 33:. https://doi.org/10.1029/2005GL025285

Tatem AJ, Goetz SJ, Hay SI (2009) UKPMC Funders Group Fifty Years of Earth Observation Satellites : Earth 96:1-7. https://doi.org/10.1511/2008.74.390.Fifty

Tsai C-WW, Lai C-FF, Chao H-CC, Vasilakos A V. (2015) Big data analytics: a survey. Journal of Big Data U6 – ctx_ver=Z3988-2004&ctx_enc=info%3Aofi%2Fenc%3AUTF-8&rfr_id=info:sid/summon.serialssolutions.com&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&rft.genre=article&rft.atitle=Big+data+analytics%3A+a+survey&rft.jtitle=Journal+of+Big+Data& 2:1-32. https://doi.org/10.1186/s40537-015-0030-3

Van der Gun J (2012) Groundwater and Global Change: Trends, Opportunities and Challenges

Van Weert F, van der Gun J, Reckman J (2009) Global overview on saline groundwater occurence and genesis. Report Nr. GP-2009.1. International Groundwater Resources Assessment Centre 107

Wahr J, Molenaar M, Bryan F (1998) Time variability of the Earth's gravity field: Hydrological and oceanic effects and their possible detection using GRACE. Journal of Geophysical Research: Solid Earth 103:30205-30229. https://doi.org/10.1029/98jb02844

Wang S, Li G, Yao X, et al. (2019) A Distributed Storage and Access Approach for Massive Remote Sensing Data in MongoDB. ISPRS International Journal of Geo-Information 8:533. https://doi.org/10.3390/ijgi8120533

Water Resources Research (2020) Big Data and Machine Learning in Water Sciences: Recent Progress and Their Use in Advancing Science: Water Resources Research. https://agupubs.onlinelibrary.wiley.com/doi/toc/10.1002/(ISSN)1944-7973.MACHINELEARN. Accessed 28 Sep 2020

Watkins MM, Wiese DN, Yuan DN, et al. (2015) Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons. Journal of Geophysical Research: Solid Earth 120:2648-2671. https://doi.org/10.1002/2014JB011547

Watson HJ (2014) Tutorial: Big Data Analytics: Concepts, Technologies, and Applications. Communications of the Association for Information Systems 34:. https://doi.org/10.17705/1CAIS.03462

Whitby MA, Fecher R, Bennight C (2017) GeoWave: Utilizing distributed key-value stores for multidimensional data. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, pp 105-122

Wiegmans FE, Holland M, Janse van Rensburg H (2013) Groundwater Resource Directed Measures for Maloney's Eye Catchment. Water Research Commission, Pretoria

Wood J, Guth A (2020) East Africa's Great Rift Valley: A Complex Rift System

Ylijoki O, Porras J (2016) Perspectives to Definition of Big Data: A Mapping Study and Discussion. Journal of Innovation Management 4:69-91. https://doi.org/10.24840/2183-0606_004.001_0006

Yu J, Wu J, Sarwat M (2015) GeoSpark: A Cluster Computing Framework for Processing Large-Scale Spatial Data. In: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. Association for Computing Machinery, New York, NY, USA

Zhang Z, Moore J (2014) Data assimilation. In: Zhang Z, Moore J (eds) Mathematical and Physical Fundamentals of Climate Change – 1st Edition, 1st edn. Elsevier, pp 291-3011

Zikopoulos P, deRoos D, Parasuraman K, et al. (2012) Harness the Power of Big Data The IBM Big Data Platform. McGraw-Hill Osborne Media

Appendix 1: Legend for land cover map

Label		Value		Color	RGB
Global	Regional	Global	Regional		
No Data		0			0, 0, 0
Cropland, rainfed		10			255, 255, 100
	Cropland, rainfed, herbaceous cover		11		255, 255, 100
	Cropland, rainfed, tree or shrub cover		12		255, 255, 0
Cropland, irrigated or post-flooding		20			170, 240, 240
Mosaic cropland (>50%) / natural vegetation (tree, shrub, herbaceous cover) (<50%)		30			220, 240, 100
Mosaic natural vegetation (tree, shrub, herbaceous cover) (>50%) / cropland (<50%)		40			200, 200, 100
Tree cover, broadleaved, evergreen, closed to open (>15%)		50			0, 100, 0
Tree cover, broadleaved, deciduous, closed to open (>15%)		60			0, 160, 0
	Tree cover, broadleaved, deciduous, closed (>40%)		61		0, 160, 0
	Tree cover, broadleaved, deciduous, open (15-40%)		62		170, 200, 0
Tree cover, needleleaved, evergreen, closed to open (>15%)		70			0, 60, 0
	Tree cover, needleleaved, evergreen, closed (>40%)		71		0, 60, 0
	Tree cover, needleleaved, evergreen, open (15-40%)		72		0, 80, 0
Tree cover, needleleaved, deciduous, closed to open (>15%)		80			40, 80, 0
	Tree cover, needleleaved, deciduous, closed (>40%)		81		40, 80, 0

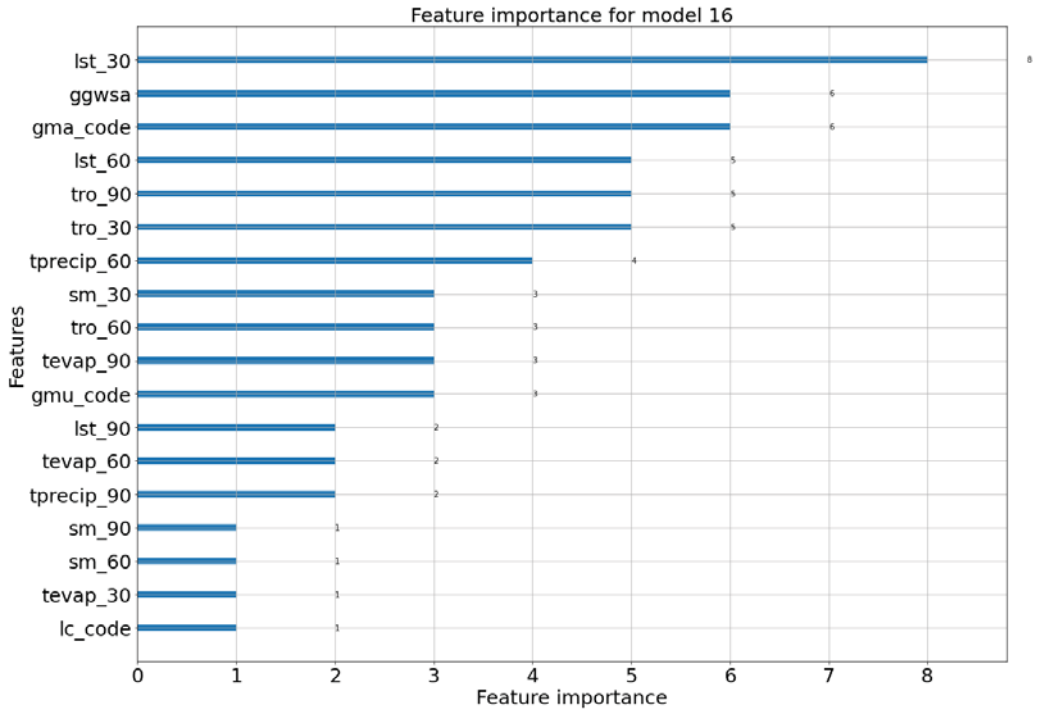
Label		Value		Color	RGB
Global	Regional	Global	Regional		
	Tree cover, needleleaved, deciduous, open (15-40%)		82		40, 100, 0
Tree cover, mixed leaf type (broadleaved and needleleaved)		90			120, 130, 0
Mosaic tree and shrub (>50%) / herbaceous cover (<50%)		100			140, 160, 0
Mosaic herbaceous cover (>50%) / tree and shrub (<50%)		110			190, 150, 0
Shrubland		120			150, 100, 0
	Evergreen shrubland		121		150, 100, 0
	Deciduous shrubland		122		150, 100, 0
Grassland		130			255, 180, 50
Lichens and mosses		140			255, 220, 210
Sparse vegetation (tree, shrub, herbaceous cover) (<15%)		150			255, 235, 175
	Sparse tree (<15%)		151		255, 200, 100
	Sparse shrub (<15%)		152		255, 210, 120
	Sparse herbaceous cover (<15%)		153		255, 235, 175
Tree cover, flooded, fresh or brackish water		160			0, 120, 90
Tree cover, flooded, saline water		170			0, 150, 120
Shrub or herbaceous cover, flooded, fresh/saline/brackish water		180			0, 220, 130
Urban areas		190			195, 20, 0
Bare areas		200			255, 245, 215
	Consolidated bare areas	201			220, 220, 220
	Unconsolidated bare areas	202			255, 245, 215
Water bodies		210			0, 70, 200
Permanent snow and ice		220			255, 255, 255

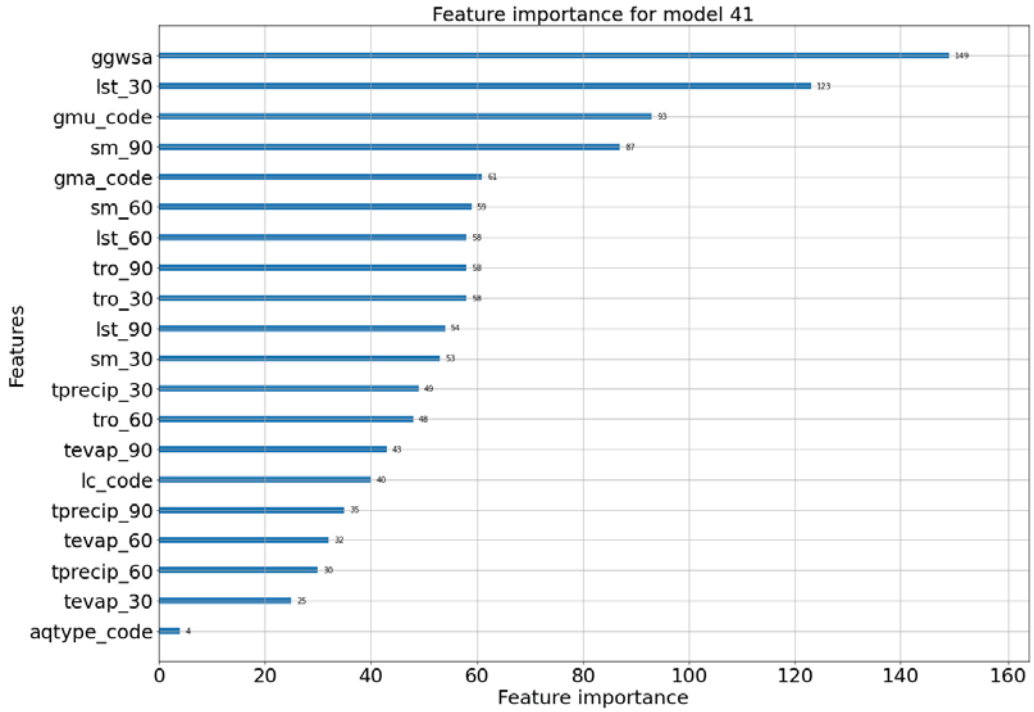
Appendix 2: Table of mean absolute errors for all model runs

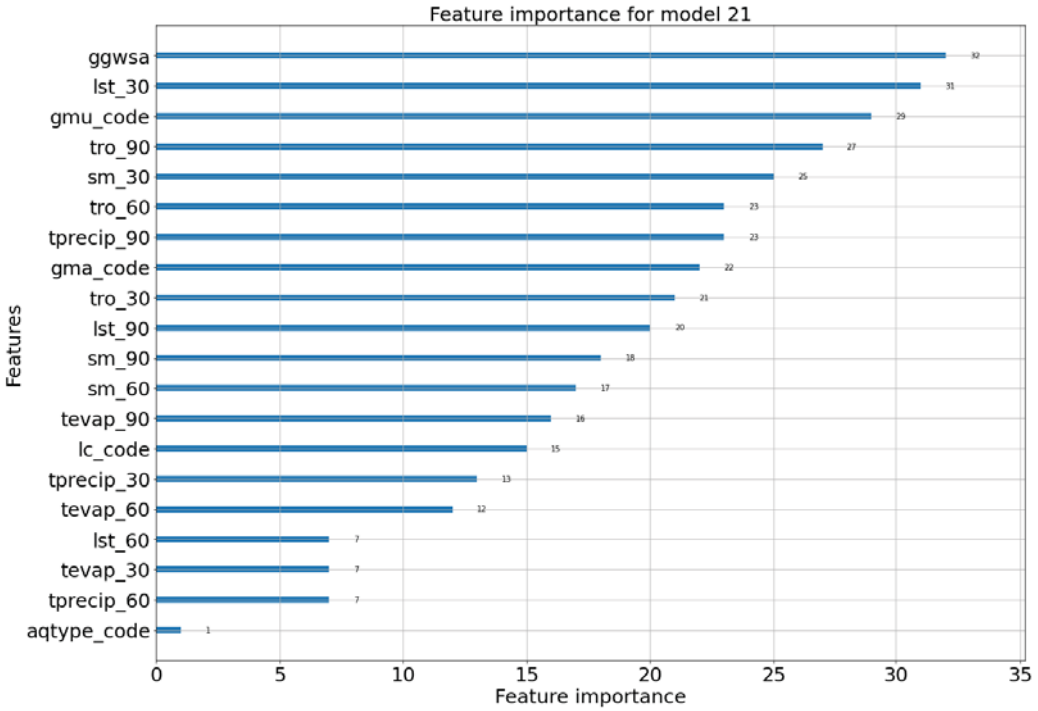
Model No.	Training score	Testing score
16	17,848946	26,740144
41	19,069898	25,916309
21	21,944976	25,505946
89	22,018323	26,304860
47	22,405129	26,273599
2	22,464920	29,883113
17	22,533392	25,710877
40	23,624674	25,152207
90	23,669164	28,685849
32	24,494933	25,539539
56	24,565496	26,250806
59	24,713275	26,474276
75	25,135341	25,413766
30	25,202213	26,673129
63	25,277708	28,552989
0	25,826891	30,254467
65	25,916797	29,236090
37	26,130676	25,655081
24	26,525331	27,707772
77	26,619471	26,711580
29	26,773775	27,910805
39	26,991959	25,797266
15	27,030893	26,496871
43	27,118175	29,696530
100	27,501247	28,236069
86	27,848208	25,696593
4	27,889247	26,711192
14	27,942828	29,141449
87	28,021943	26,610208
88	28,234004	26,316298
31	28,324220	25,502928
85	28,352378	26,445886
27	28,468128	29,666193
5	28,644791	27,888238
95	28,831406	26,017404
62	29,124826	26,411477
44	29,145013	29,937435
84	29,374263	27,624600
23	30,092430	28,519709
76	30,215790	25,676533
69	30,555003	28,462353
51	30,914281	27,804763
19	31,312477	27,083782
54	31,557019	25,998050
61	31,915258	29,101971
57	31,984579	25,373199
79	32,011243	29,603712
82	32,165766	30,977532
11	32,315520	28,321670
94	32,656596	28,670207
18	32,795368	26,218912

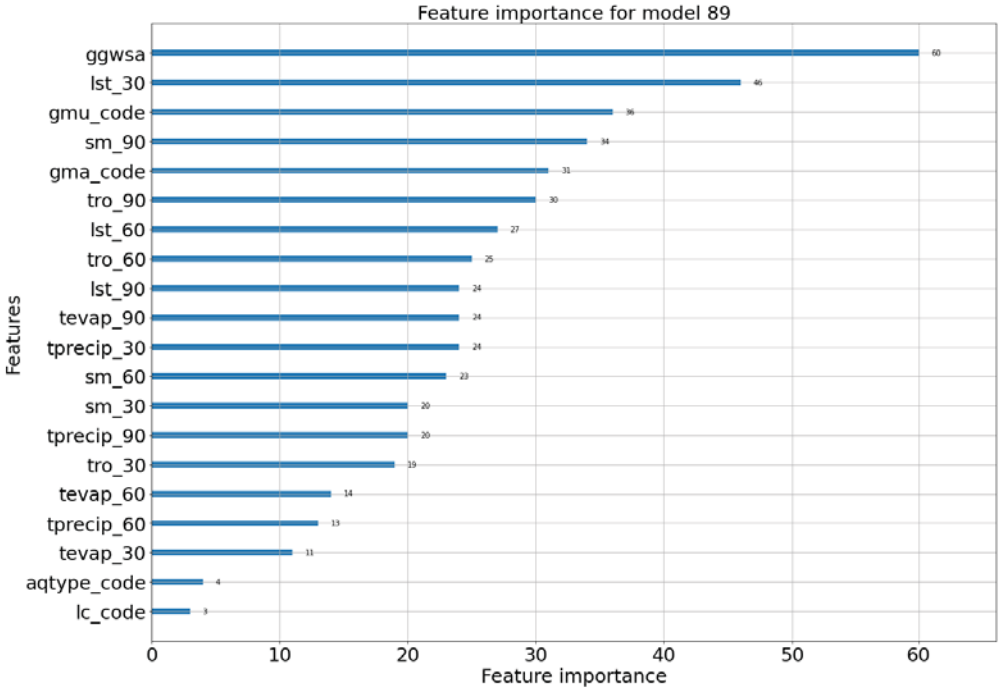
Model No.	Training score	Testing score
28	33,320437	33,355326
25	33,505135	26,719099
71	33,695243	28,036191
9	34,315373	27,997933
91	34,334866	28,842793
72	34,513080	25,353502
33	34,625607	28,925912
60	34,744977	31,913824
67	34,958223	28,993225
8	35,139224	29,064702
6	35,261152	27,464458
1	35,343800	29,389803
92	36,123219	28,647168
96	36,292218	28,986263
64	36,306192	28,251349
26	36,619845	26,322898
50	36,853506	29,259121
45	36,968243	27,287602
70	36,973009	28,231565
97	37,030195	27,930289
98	37,213935	27,919437
7	37,328470	28,036410
36	37,761582	26,026738
12	38,707984	30,271429
13	39,013492	32,319562
34	39,661979	28,642646
81	40,343670	25,345343
38	40,933082	27,517918
55	41,193410	25,788522
80	42,001618	26,347467
74	42,419811	29,044834
46	42,493109	26,207590
48	43,005858	27,159635
68	43,608749	26,520068
78	43,753281	29,660488
73	43,916673	29,054615
20	44,396021	25,521243
3	45,721230	28,370919
99	46,976839	25,578834
22	47,360974	28,006448
83	48,504386	25,181396
49	48,788778	25,439512
42	49,757235	26,091757
35	49,808832	29,342935
10	50,596474	26,156514
93	53,138890	28,951393
52	54,778764	27,789299
66	55,245309	25,046291
58	56,263872	28,561853
53	60,911358	28,648473
Mean error	**33,20638932**	**27,552611**

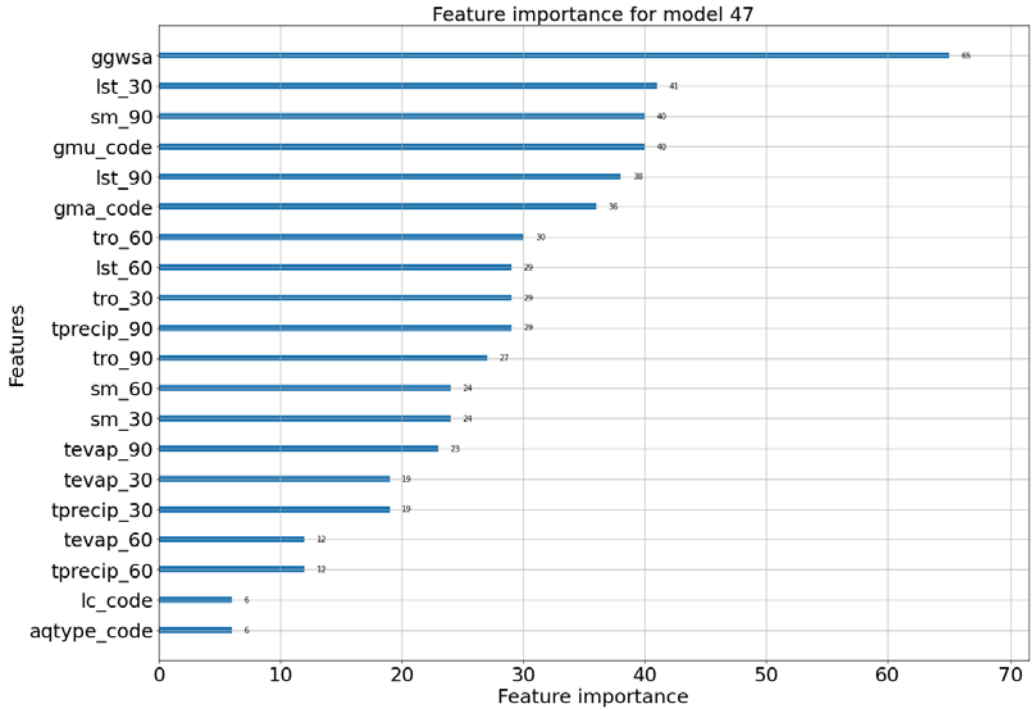
Appendix 3: Model feature importance (top ten models only)

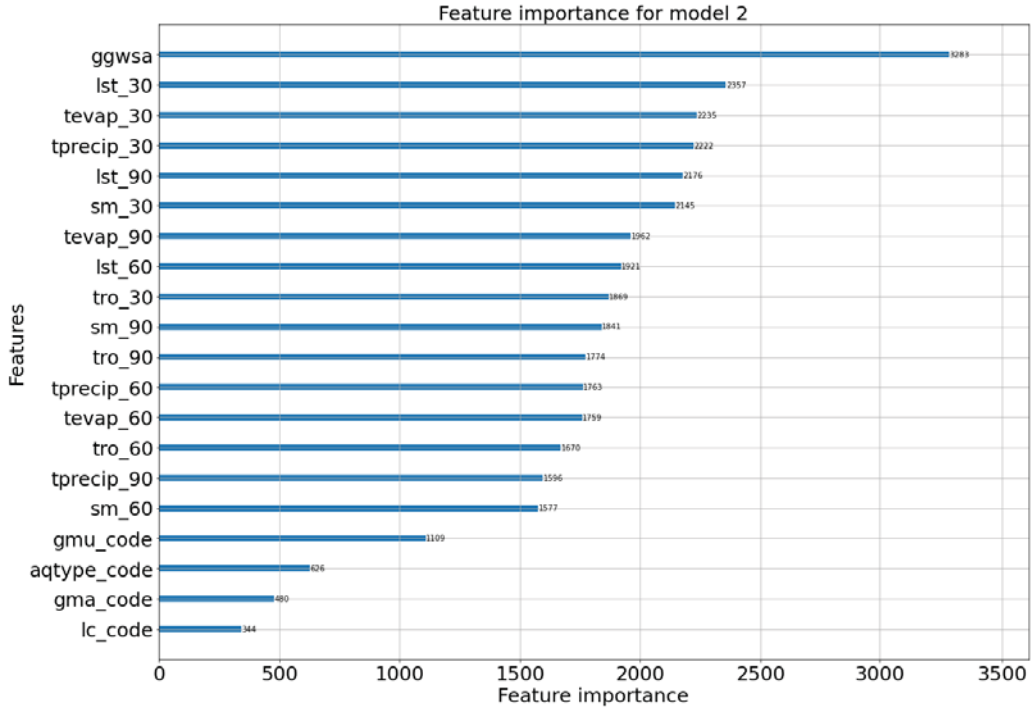


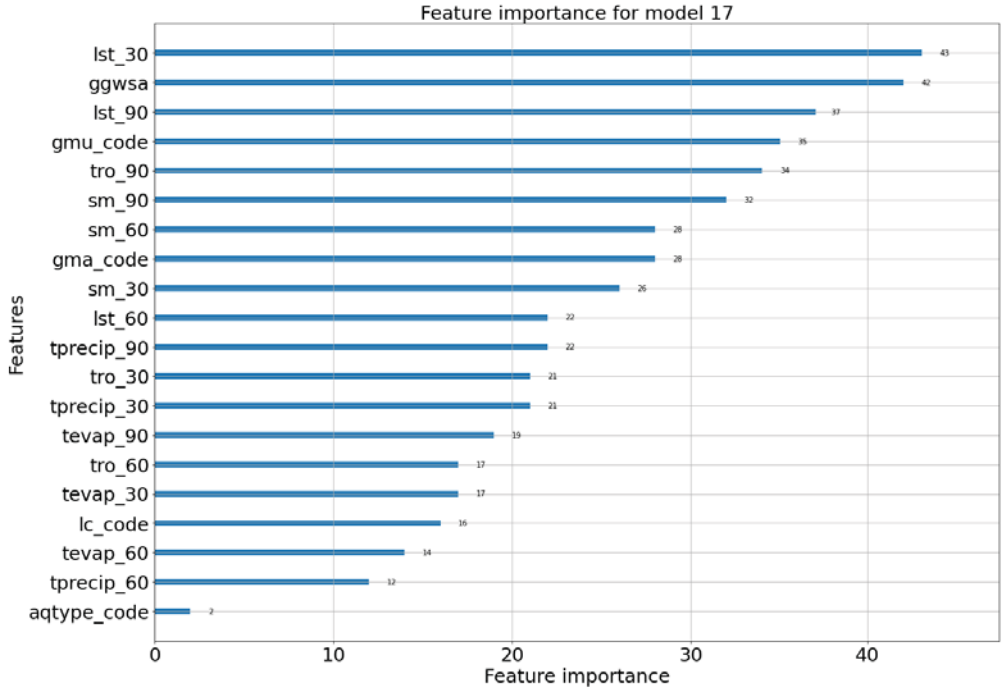


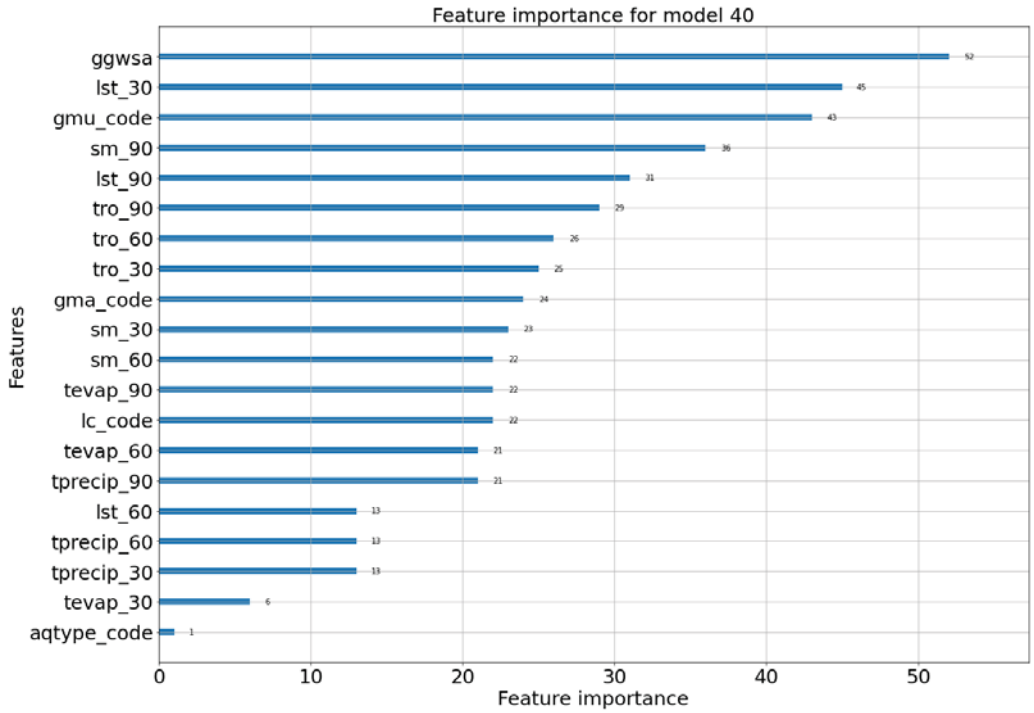

Feature importance for model 21
ggwsa
lst_30
gmu_code
tro_90
sm_30
tro_60
tprecip_90
gma_code
tro_30
lst_90
sm_90
sm_60
tevap_90
lc_code
tprecip_30
tevap_60
lst_60
tevap_30
tprecip_60
aqtype_code
Features
0
5
10
15
20
25
30
35
Feature importance

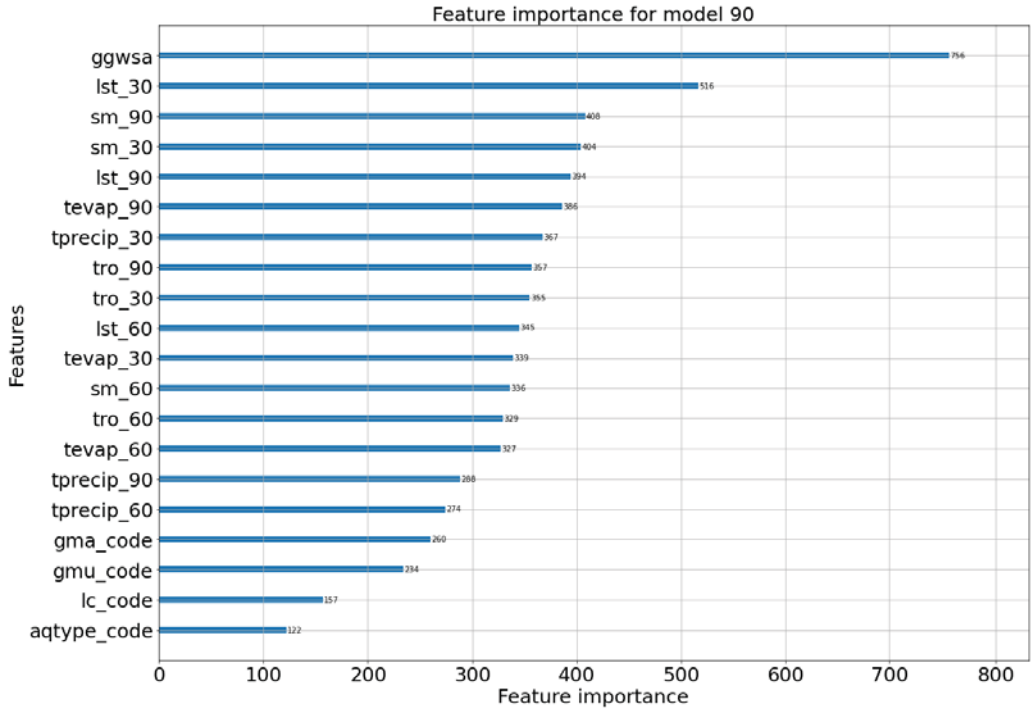

Feature importance for model 89
ggwsa
lst_30
gmu_code
sm_90
gma_code
tro_90
lst_60
tro_60
lst_90
tevap_90
tprecip_30
sm_60
sm_30
tprecip_90
tro_30
tevap_60
tprecip_60
tevap_30
aqtype_code
lc_code
Features
0
10
20
30
40
50
60
Feature importance

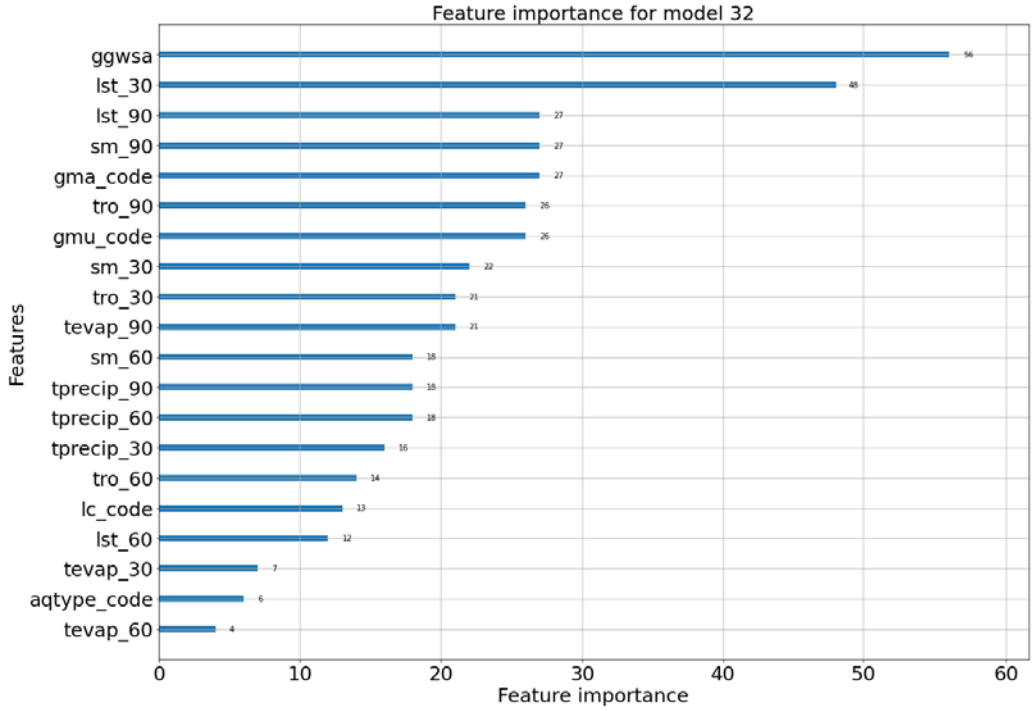

Feature importance for model 47
ggwsa
lst_30
sm_90
gmu_code
lst_90
gma_code
tro_60
lst_60
tro_30
tprecip_90
tro_90
sm_60
sm_30
tevap_90
tevap_30
tprecip_30
tevap_60
tprecip_60
lc_code
aqtype_code
Features
0
10
20
30
40
50
60
70
Feature importance


Feature importance for model 2
ggwsa
lst_30
tevap_30
tprecip_30
lst_90
sm_30
tevap_90
lst_60
tro_30
sm_90
tro_90
tprecip_60
tevap_60
tro_60
tprecip_90
sm_60
gmu_code
aqtype_code
gma_code
lc_code
Features
0
500
1000
1500
2000
2500
3000
3500
Feature importance

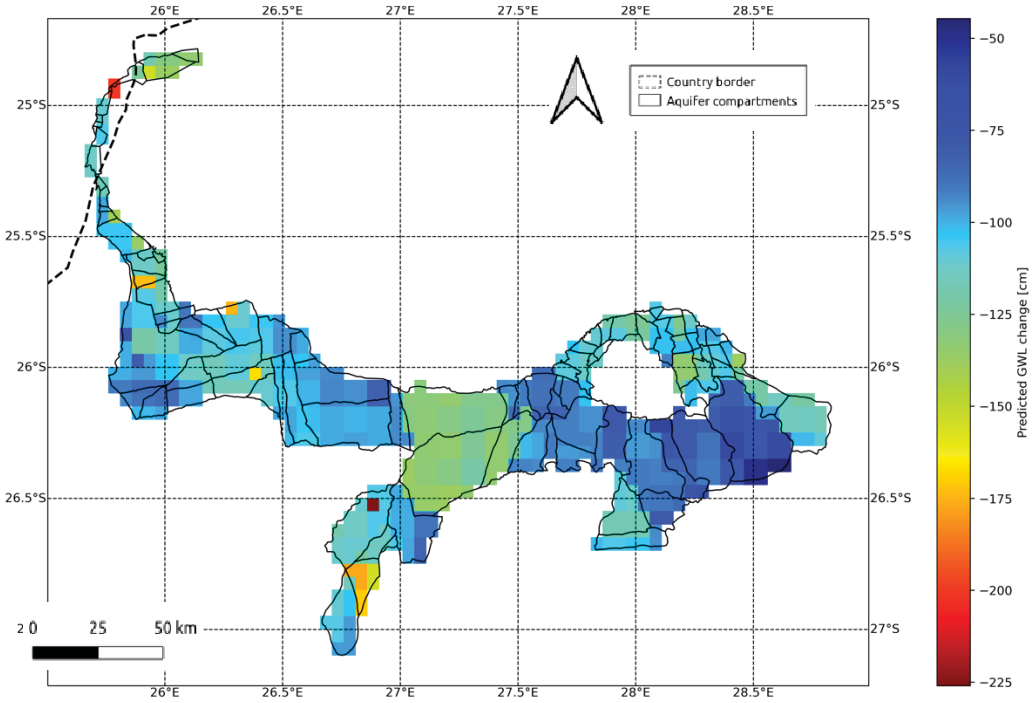

Feature importance for model 17
lst_30
ggwsa
lst_90
gmu_code
tro_90
sm_90
sm_60
gma_code
sm_30
lst_60
tprecip_90
tro_30
tprecip_30
tevap_90
tro_60
tevap_30
lc_code
tevap_60
tprecip_60
aqtype_code
43
42
37
35
34
32
28
28
26
22
22
21
21
19
17
17
16
14
12
2
Features
0
10
20
30
40
Feature importance


Feature importance for model 40
ggwsa
lst_30
gmu_code
sm_90
lst_90
tro_90
tro_60
tro_30
gma_code
sm_30
sm_60
tevap_90
lc_code
tevap_60
tprecip_90
lst_60
tprecip_60
tprecip_30
tevap_30
aqtype_code
52
45
43
36
31
29
26
25
24
23
22
22
22
21
21
13
13
13
6
1
Features
0
10
20
30
40
50
Feature importance

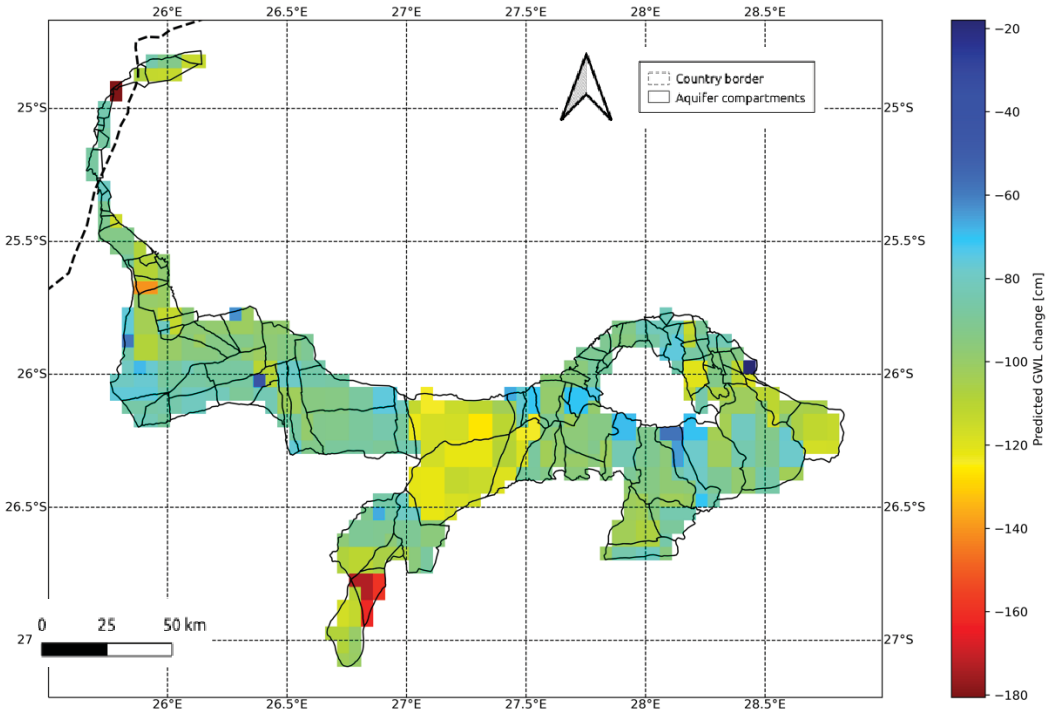

Feature importance for model 90
ggwsa
lst_30
sm_90
sm_30
lst_90
tevap_90
tprecip_30
tro_90
tro_30
lst_60
tevap_30
sm_60
tro_60
tevap_60
tprecip_90
tprecip_60
gma_code
gmu_code
lc_code
aqtype_code
Features
0
100
200
300
400
500
600
700
800
Feature importance


Feature importance for model 32
ggwsa
lst_30
lst_90
sm_90
gma_code
tro_90
gmu_code
sm_30
tro_30
tevap_90
sm_60
tprecip_90
tprecip_60
tprecip_30
tro_60
lc_code
lst_60
tevap_30
aqtype_code
tevap_60
Features
0
10
20
30
40
50
60
Feature importance

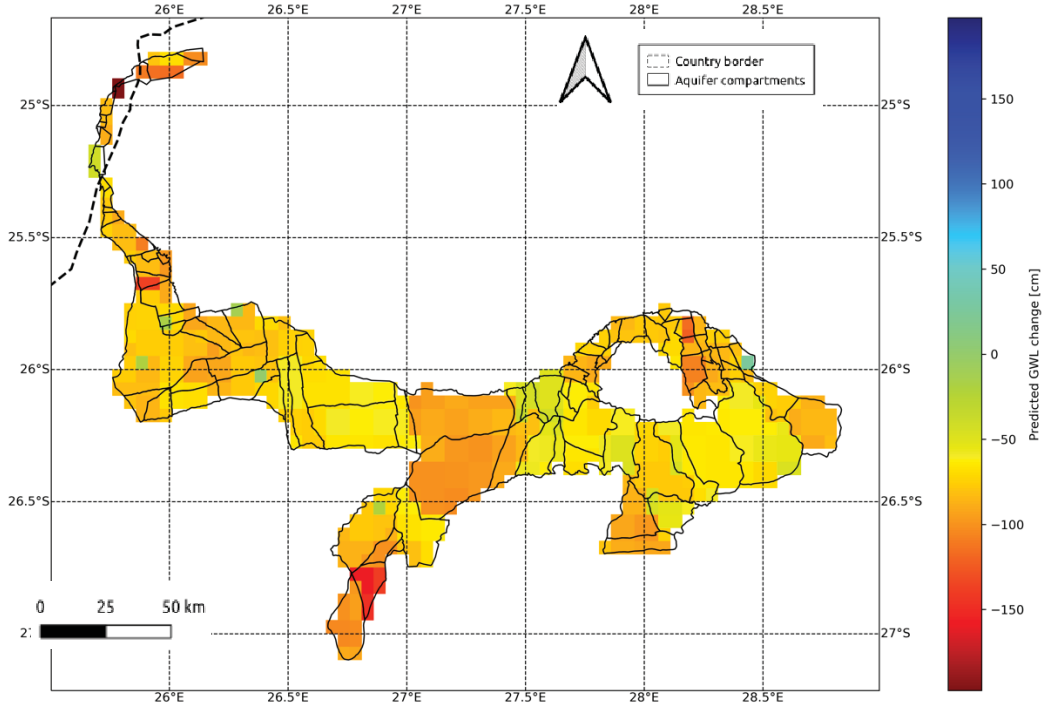
Appendix 4: Triennial net groundwater level change maps (2003-2019)



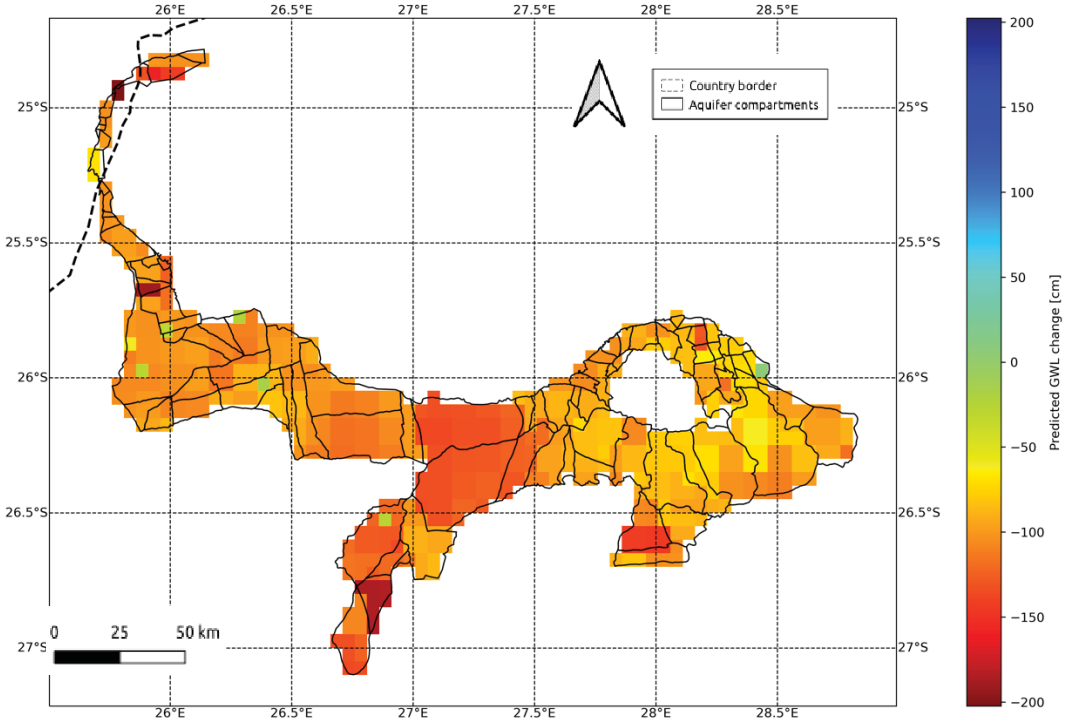
Change in groundwater level 2003-2005



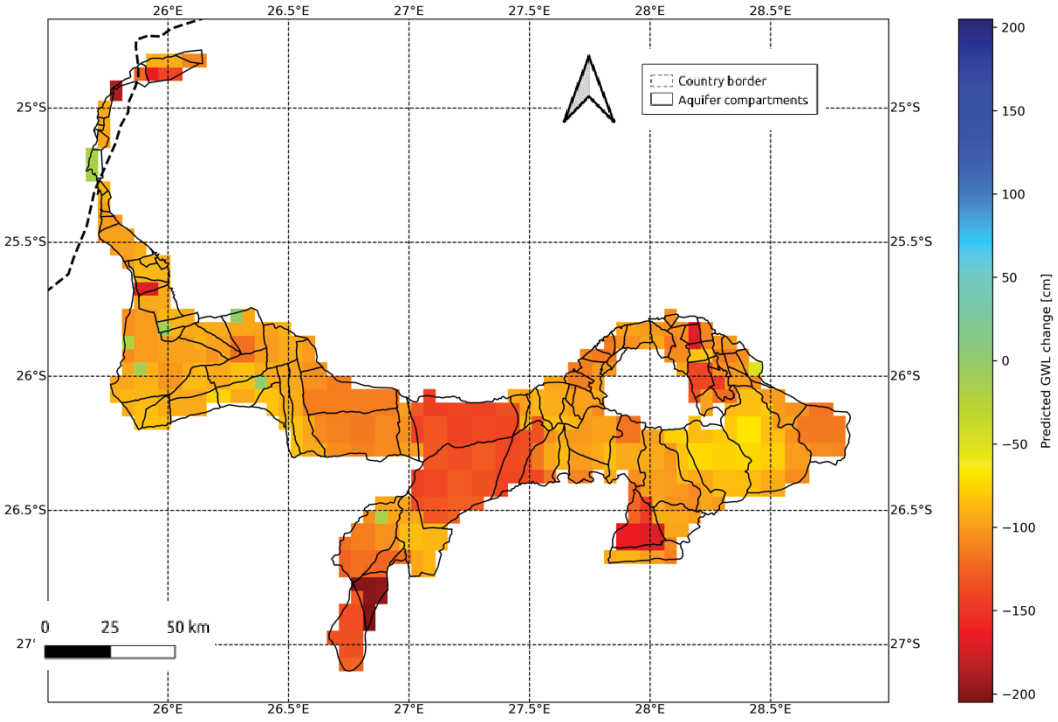
Change in groundwater level 2006-2008



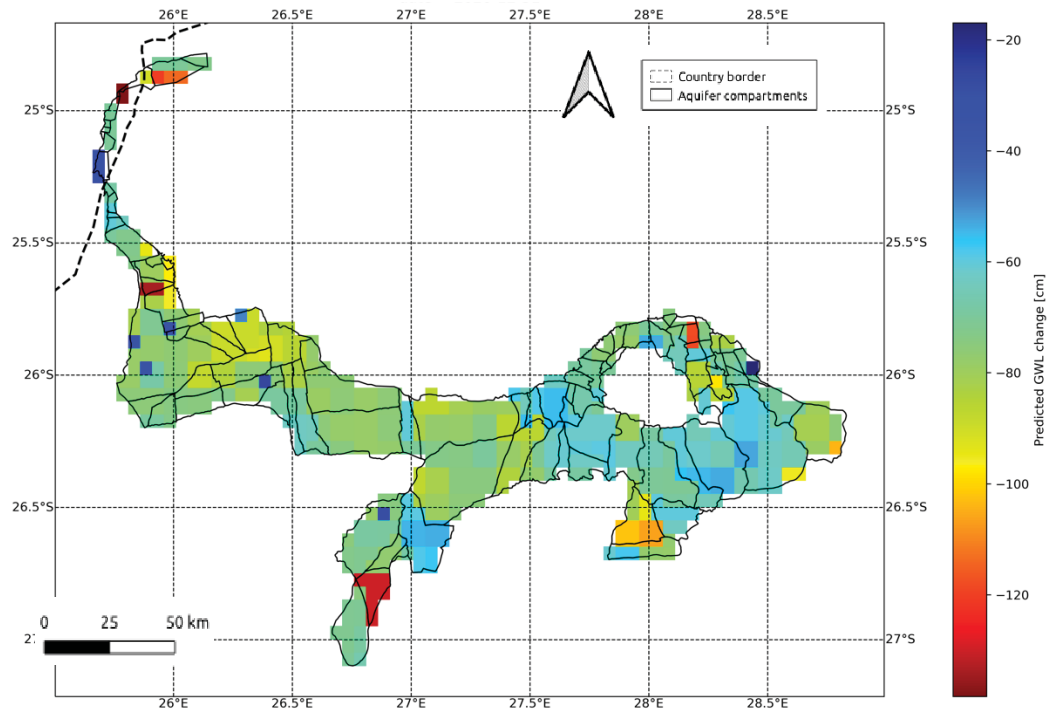
Change in groundwater level 2009-2011



Change in groundwater level 2012-2014



Change in groundwater level 2015-2017



Change in groundwater level 2018-2019