University of Venda
Creating Future Leaders



FACULTY OF SCIENCE, ENGINEERING AND AGRICULTURE

DEPARTMENT OF EARTH SCIENCES

Time series modelling of groundwater levels in a selected semi-arid catchment within Vhembe District Municipality, South Africa

Madula Andy

Student no: 15015599

Supervisor: Prof R. Makungo

A master's thesis submitted to the Department of Earth sciences in partial fulfilment of the requirements for a master's degree of earth science in hydrology and water resources.

February 2025

© University of Venda

Declaration

I, Madula Andy of student number: 15015599 declare that this dissertation for Master of Earth Science at the university of Venda is my work in design and execution. The dissertation has not been submitted previously by anyone, for a degree at this university or any other university. All effort has been taken to acknowledge the material which was referenced.



Signature

24 / 02 / 2025

Date

© University of Venda

Acknowledgements

I would like to thank God for blessing me with the ability, strength and knowledge to conduct this study, without his guidance and this fantastic opportunity of life, this study would not have been possible.

My outmost appreciation goes to my supervisor, Prof. R. Makungo for the commitment in assisting me with my work, advice and always being available for consultations throughout the entire time I was conducting this study.

Thank you to Musitha D. for helping me with QGIS mapping and all the effort he made to make this study to be the way it is. Thanks to the developers of the Python programming for providing a free version which was used to carry out this study.

I would also like to acknowledge the National Research Foundation for providing financial support throughout the course of this study. Also acknowledging the Department of water and sanitation (DWS) together with the South African Weather Services (SAWS) for providing data that was used in the study.

Special thanks and acknowledgement to my father, siblings and friends for the peace and love, and for always being there for me and helping me with anything in-order form me to be able to complete this study.

© University of Venda

Dedication

This study is dedicated to my late mother, Siphiwe Goodness Madula.

© University of Venda

Abstract

This study is aimed at modelling groundwater levels in a semi-arid catchment within Vhembe District Municipality, South Africa. Auto Regressive Integrated Moving Average (ARIMA) model and Seasonal Auto Regressive Integrated Moving Average with eXogenous variables (SARIMAX) model were used to model the interaction between groundwater levels, temperature, wind speed, evaporation, and rainfall. The unpredictable occurrence of rainfall and other weather conditions in semi-arid and arid regions has caused a restriction in simpler computations of groundwater levels. Historical hydrological data sets of groundwater level were used to model groundwater level and forecast with ARIMA model. Climatic variables like temperature, wind speed, evaporation and precipitation were employed as exogeneous variables of the SARIMAX simulations. The analysis of groundwater levels included the use of Sen's slope estimator which showed significant fluctuations, with sharp declines, followed by stability and a subsequent increase over time. The ARIMA model's forecasting of groundwater levels indicated a stable groundwater levels trend post-2016. The training data revealed historical fluctuations, while the test data showed a sharp increase before stabilizing. The SARIMAX model demonstrated a reasonable predictive accuracy for groundwater levels, incorporating significant predictors and seasonal patterns. However, diagnostic tests suggested further model enhancements could improve residual handling. Data stations were selected based on availability of long-term data and considering stations with minimal or no gaps. The data range of the study was 12 years from 2007 to 2018. Both ARIMA and SARIMAX models performed well in predicting groundwater levels. The inclusion of exogenous variables in the SARIMAX model offered a more nuanced understanding of data trends, making it a reliable tool for forecasting. The study findings showed that the groundwater levels in the Luvuvhu River catchment have a stable increase over time and also highlighted the issue of missing data in climatic variables like precipitation which prevented the SARIMAX accuracy in forecasting groundwater levels. This study's insights are valuable for developing effective groundwater management strategies, especially when compared with other studies that highlight the importance of incorporating climate variability into such models for enhanced accuracy. The SARIMAX model's application in predicting groundwater levels is a significant step in environmental modelling, offering insights into subterranean water dynamics crucial for sustainable management.

© University of Venda

Table of contents

Chapter 1: Introduction **1**

1.1 Background 1

1.2 Problem statement 3

1.3 Motivation 5

1.4 Objectives of the study 7

1.5 Research questions 7

1.6 Study area 8

Chapter 2: Literature review **10**

2.1 Preamble 10

2.2 Time series analysis of groundwater levels 10

2.3 Time series methods of modelling groundwater levels 14

2.3.1 Auto Regressive Integrated Moving Average (ARIMA) Model 14

2.3.2 The transfer-function noise model 15

2.3.3 Convolution of stress time series with response functions 17

2.3.4 The TFN Model of Box and Jenkins: In Discrete time 19

2.3.5 Forecasting Model Based on the Singular Spectrum Analysis 21

2.3.6 Artificial Neural Network (ANN) 24

2.3.7 Extreme gradient boosting (Xgboost) 25

2.4 Input variables selection for groundwater levels modelling 27

Chapter 3: Methodology **30**

3.1 Preamble 30

3.2 Data collection 30

3.3 Methods for modelling and forecasting groundwater levels 32

© University of Venda

3.3.1 Data processing and stationary tests 32

3.3.2 Data Clustering 35

3.3.3 ARIMA model 35

3.3.4 SARIMAX model 38

3.3.5 Models performance assessment 39

3.4 Reasons for using the selected models for the study 40

Chapter 4: Results and discussion 42

4.1 Preamble 42

4.2 Data Processing and stationarity tests 42

4.2.1 Data clustering and stationarity tests results 42

4.2.2 Results for stationarity tests, autocorrelations and partial autocorrelations functions 43

4.2.3 Trend analysis of groundwater levels 48

4.2.4 Results for Autocorrelation, PACF and decomposition of groundwater levels 51

4.2.5 Groundwater levels model development, training and testing with ARIMA model 60

4.2.6 Groundwater levels forecasting with ARIMA model 72

4.2.7 Groundwater levels training and testing with SARIMAX model 79

4.2.8 Groundwater levels forecasting with SARIMAX model 91

4.3 Comparisons between ARIMA and SARIMAX models 97

Chapter 5: Conclusions and recommendations 98

5.1 Conclusions 98

5.2 Limitations to groundwater level modelling using ARIMA and SARIMAX model 99

5.3 Recommendations 100

REFERENCES 101

© University of Venda

List of Abbreviations

ACF	: Auto-Correlation Function
ADF	: Augmented Dickey-Fuller
AIC	: Akaike Information Criterion
ANFIS	: Adaptive Neuro-fuzzy Inference System
ANN	: Artificial Neutral Networks
API	: Application Programming Interface
AR	: Auto Regressive
ARIMA	: Auto-Regressive Moving Average
BIC	: Bayesian Information Criterion
BR	: Bayesian Regularization
CRR	: Cumulative Rainfall Residual
DAR	: Dynamic Auto Regressive
DWS	: Department of water and sanitation
ET	: Evapotranspiration
GP	: General Programming
H	: Heteroscedasticity
HA	: Alternative Hypothesis
HARTT	: Hydrograph Analysis Rainfall and Time Trends
HO	: Null Hypothesis
HQIC	: Hannan-Quinn Information Criterion
IR	: Impulse Response
JB	: Jarque-Bera
KPSS	: Kwiatkowski-Phillips-Schmidt-Shin

© University of Venda

LM	: Levenberg–Marquardt
MA	: Moving Averages
MAE	: Mean Absolute Error
MAPE	: Mean Absolute Percentage Error
MaxAE	: Maximum Absolute Error
MaxAPE	: Maximum Absolute Percentage Error
MLP	: Machine Learning Programmer
MSE	: Mean Squared Error
MSSA	: Multichannel Singular Spectrum Analysis
NSE	: Nash-Sutcliffe Efficiency
PACF	: Partial Auto-Correlation Function
PIRFICT	: Predefined Impulse Response Function In Continuous Time
PP	: Phillips-Perron
RBF	: Radial Basis Function
RMSE	: Root Mean Square Error
RNN	: Recurrent Neutral Networks
SARIMAX	: Seasonal Auto Regressive Integrated Moving Average with eXogenous variables
SAWS	: South African Weather Services
SSA	: Singular Spectrum Analysis
SVM	: Support Vector Machines
TFN	: Transfer Function Noise
UN	: United Nations
USA	: United States of America
WLR	: Wavelet Linear Regression

© University of Venda

List of Figures

Figure 1.1: Map showing the location of the study area

Figure 1.2: Map showing the geology of the study area.

Figure 2.1: Distribution of multi-channel data for MSSA

Figure 4.1: Map showing data clustering by Thiessen's polygon

Figure 4.2a Differencing orders of groundwater levels in borehole stations A9N0001, A9N0002, A9N0003, A9N0005

Figure 4.2b Differencing orders of groundwater levels in boreholes: A9N0010, A9N0011, A9N0014

Figure 4.3a Autocorrelations plots for boreholes: A9N0001, A9N0002, A9N0003, and A9N0005

Figure 4.3b Autocorrelations plots for boreholes: A9N0010, A9N0011, and A9N0014

Figure 4.4a PACF plots for boreholes: A9N0001, A9N0002, A9N0003, and A9N0005

Figure 4.4b PACF plots for boreholes A9N0010, A9N0011, and A9N0014

Figure 4.6a Groundwater levels prediction plots using ARIMA model for boreholes: A9N0001, A9N0002, A9N0003, and A9N0005

Figure 4.5a Time series decompositions plots for boreholes: A9N0001, A9N0002, A9N0003, and A9N0005

Figure 4.6b Groundwater levels prediction plots using ARIMA model for boreholes: A9N0010, A9N0011, and A9N0014

Figure 4.5b Time series decompositions plots for boreholes: A9N0010, A9N0011, and A9N0014

Figure 4.7a Graph plots of observed and simulated values of groundwater levels using ARIMA model for borehole A9N0001, A9N0002, A9N0003 and A9N0005

Figure 4.7a Graph plots of observed and simulated values of groundwater levels using ARIMA model for borehole A9N0010, A9N0011, and A9N0014

Figure 4.8a Scatter plots of observed and simulated values of groundwater levels using ARIMA model for borehole A9N0001, A9N0002, A9N0003 and A9N0005

© University of Venda

Figure 4.8b Scatter plots of observed and simulated values of groundwater levels using ARIMA model for borehole A9N0010, A9N0011, and A9N0014

Figure 4.9a Groundwater levels forecasted using ARIMA model for boreholes A9N0001, A9N0002, A9N0003, and A9N0005

Figure 4.9b Groundwater levels forecasted using ARIMA model for boreholes A9N0010, A9N0011, and A9N0014

Figure 4.10a Predicting groundwater levels using the developed SARIMAX model for boreholes A9N0001, A9N0002, A9N0003, and A9N0005

Figure 4.10b Predicting groundwater levels using the developed SARIMAX model for boreholes A9N0010, A9N0011, and A9N0014

Figure 4.11a Observed and simulated values of groundwater levels using SARIMAX model for borehole A9N0001, A9N0002, A9N0003 and A9N0005

Figure 4.11b Observed and simulated values of groundwater levels using SARIMAX model for borehole A9N0010, A9N0011, and A9N0014.

Figure 4.12a Scatter plots of Observed and simulated values of groundwater levels using SARIMAX model for borehole A9N0001, A9N0002, A9N0003, and A9N0005.

Figure 4.12b Scatter plots of Observed and simulated values of groundwater levels using SARIMAX model for borehole A9N0010, A9N0011, and A9N0014.

Figure 4.13a Forecasted groundwater levels using SARIMAX model for boreholes A9N0001, A9N0002, A9N0003, and A9N0005

Figure 4.13b Forecasted groundwater levels using SARIMAX model for boreholes A9N0010, A9N0011, and A9N0014.

© University of Venda

List of Tables

Table 3.1: Description of the data employed in the study

Table 3.2 Model performance evaluation criteria.

Table 4.1 Stationary tests results for all the stations and their p-value for the first and second order of differencing the groundwater levels time series data.

Table 4.2 Trend analysis results of groundwater level using Sens slope estimator and their respective confidence interval.

Table 4.3 Summary of the developed ARIMA models for each borehole

Table 4.4 Performance metrics of the ARIMA model developed for all the borehole stations and forecasting.

Table 4.5 Summary of the developed SARIMAX models for all the boreholes stations.

Table 4.6 Performance metrics of the SARIMAX model developed for all the borehole stations

© University of Venda

Chapter 1: Introduction

1.1 Background

Time series data are widely used in technical, physical and social sciences. As a result, time series analysis has become increasingly popular in recent years. It is useful to conceive of time series analysis as a set of mathematical and statistical approaches for modelling dynamic systems. Time series analysis approaches have been employed to solve a variety of practical problems, including modelling and forecasting hydrological time series, as well as process and quality control (Jonath, 2022).

Water utilisation in the Southern Africa region has been reported to show an increasing trend in the past years (Wada and Bierkens 2014), it has been estimated that by the year 2050 in the sub-Saharan region, there would a 50% growth in global population (UN, 2019). Industrialization and consumption are expected to increase demand of water (Nieuwoudt *et al.,* 2004). In the region of sub-Saharan Africa, groundwater is known to be the greatest large storage of freshwater (MacDonald *et al.,* 2012) and has a greater significance to supply water to meet the higher demands of water in the future (Cobbing and Hiller, 2019). Moreover, in semi-arid regions, the exploration of groundwater resources has a potential significant benefit compared to surface water because of climatic variability (Gaye and Tindimugaya, 2019; Braune and Xu, 2010).

From a hydrological modelling perspective, Groundwater levels are significant indicators of sustainable groundwater abstractions which allows researchers to study and assess the influence of climate variability and the impacts on groundwater level and storage (Cuthbert *et al.,* 2019). In the sub-Saharan Africa region, there is a lack and shortage of information on studies on groundwater levels, this means that most assessments on water management will usually depend on historical hydrogeological data that has been sought for and taken from other efficient models (for example Doll and Fiedler, 2008), or from extensive review measures of data (MacDonald *et al*., 2021) using on-site reports with an accuracy that is limited (Sood and Smakhtin, 2015).

There is a variety of methods being utilised for modelling and forecasting groundwater level in different geological regions such as hybrid, machine-learning, statistical, and mathematical methods, although recently, studies across parts of the world have shown the use of time series models because of their potential in giving accurate results. Many researchers have investigated groundwater level modelling time series models and how they perform with the historical data

© University of Venda

(Rajaee *et al*., 2019). These methods have shown to have a disadvantage in other studies, such as overfitting, over-tuning of the model during its development, generalizability being low, risk of incorporating irrelevant data, and incorrect programming tool with irrelevant models. However, they are simple to use and have an acceptable accuracy with high-speed run which has led a lot of researchers to apply them in their studies (Rajaee and Boroumand, 2015).

For groundwater level modelling, there has been a tradition among researchers for a long time of utilising physical or conceptual based methods, even though they have shown practical issues including a need for exogeneous variables and an extensive amount of data (Rajaee and Boroumand, 2015). In most scenarios, data is limited and even though it is more important to acquire accurate predictions, it becomes difficult in this case. Therefore, the use of time series modelling approaches in groundwater level modelling has spiked an interest of a lot of researchers hence a recent increase of studies in the world.

There are illustrations in previous literature of historical seasonal patterns that were used for groundwater levels predictions and its relations to intense precipitation and other different climatic conditions (Cuthbert *et al.,* 2019; Taylor *et al*., 2013; Owor *et al*., 2009; Sibanda *et al.,* 2020). In the broader Earth Science context, an analysis and comprehension of groundwater level process in semi-arid regions is also significant for the study and explanations of historical geologic records of materials that preserve evidence of climate variability (Cuthbert *et al.* 2014), and their differences in longer-term reactions.

This study presented an analysis of groundwater levels modelling by modelling groundwater together with meteorological patterns using time series models in a semi-arid catchment from South Africa. Hydrogeological and meteorological data sets were used to model and present an understanding to the interaction between rainfall, evaporations, wind speed, temperatures, and groundwater levels in a semi-arid catchment at a specific area within Vhembe District Municipality, South Africa. The area has a type of wide-spread aquifer and climate variations are mostly consisted of seasonal rainfall.

© University of Venda

1.2 Problem statement

The dependence on groundwater has excessively increased in regions with significant scarcity of surface water is scarce. Including semi-arid and more arid areas which rely on available groundwater assets to preserve anthropogenic activities (Rijsberman, 2006; Bhakar and Singh, 2018). The accelerated changes on land utilisation, increased socioeconomic conditions within communities, and weather conditions are the key sources accounting for the excessive overuse of groundwater that leads groundwater shortage and depletion (Kundzewicz and Döll, 2009; Foster *et al.,* 2013). There is also an issue with groundwater quality decline with reference to the over-utilization of groundwater (Haritash *et al*., 2016), which is a growing concern among hydrogeologists. The lack and unpleasant degradation of groundwater assets can also additionally result in issues of groundwater stress, sustainability, and scarcity. Hence, time series models are required to improve knowledge on groundwater management.

In dryland regions where the lack of freshwater has become a more significant issue due to population growth, urbanization, high demands from irrigation, and unpredictable climate conditions (Taylor *et al.,* 2013; Scanlon *et al.,* 2006). There is also a significant lack of short and long-term groundwater levels information from sub-Saharan Africa to conduct assessments in the measures to control, process and sustain available groundwater.

The unpredictable occurrence of precipitation in semi-arid regions has caused a restriction in simpler computations of groundwater levels. Therefore, more communities are left exposed to freshwater scarcity, depending on the groundwater available without knowing the water level history that could be useful for knowledge-based management of the said groundwater. According to Edmunds, (2003) there is a practice of mining of groundwater that is being used under more arid conditions. Bierkens and Wada (2019) analysed the use and depletion of vulnerable non-renewable groundwater because of ungoverned access and extraction of groundwater to reach the increasing need of freshwater because of population growth.

Regulatory frameworks are practiced in some semi-arid areas, but the measurements of water pumped out or recharged to the groundwater system can be highly unstable, which makes information of groundwater levels be less dependable or unavailable. This increases the complexity of calibrating and validating timeseries models results over historical data, even though they exist (Hughes, 2008). The training and effective operation of time series models has therefore been seriously constrained by the insufficient understanding of the

© University of Venda

implementation of some important hydrological processes, notably groundwater incline or decline and surface-ground water interactions.

However, complicated, and more non-linear models and lack of data has shown a great difficulty in modelling of groundwater (Rodiger *et al.,* 2014). In developing countries, groundwater systems have shown a lack of proper efforts in data capturing, planning and coordinating. Recent modelling studies of groundwater that focused on semi-arid basins came across some complications in estimating groundwater levels (Rodiger *et al*., 2014; Candela *et al.,* 2013), together with the influence of variable climate on a complicated flow system of hydrogeology or in establishing a feasible, fitting time series representation of the groundwater structure (Bense *et al*., 2013).

Groundwater level modelling is more important in steppe climate regions where freshwater availability is insufficient and groundwater level forecasting accuracy is a significant challenge in water resource of water systems (Chang *et al*., 2016). Groundwater level variabilities are impacted upon by geological limitations as well as differences in hydrology and exogeneous variables such as evaporation, temperature, and rainfall (Sahoo *et al*., 2017). Moreover, it is difficult to accurately characterize the relationship between these variables.

© University of Venda

1.3 Motivation

Semi-arid regions are comprised of lower averages of precipitation with high records of temperature, which makes groundwater resources incredibly significant and costly related to other regions which are humid. Groundwater plays significant role as the initial response for water availability plan during extensive periods of dry conditions because of its ability to withstand drought (DWS, 2010). Hence, a time series modelling study on the reactions of groundwater level to meteorological variabilities is a necessary concept.

The lack of studies in sub-Saharan Africa on groundwater levels and meteorological seasonal patterns that presents important information indicating groundwater sustainability has meant an analysis of water levels has a dependence of large scales models (for example Döll and Fiedler, 2008), derived from datasets lacking field survey validations (Sood and Smakhtin, 2015), or reviews of lengthy datasets (MacDonald *et al*., 2021). From a quantitative perspective, these meteorological patterns are important because they provide an understanding of groundwater level patterns, interaction to different weather conditions (Cuthbert *et al*., 2019).

Furthermore, an understanding in the effects of groundwater levels to wind speed, humidity, precipitation and temperature in semi-arid regions will aid in improved understanding in groundwater reserve. Therefore, the association and computation of groundwater level in semi-arid areas remains a significant difficulty around the world due to lack of data (Wheater *et al.,* 2010).

Broader hydrogeological studies around the world have been striving to utilise a surface measurement water balance approach to estimate groundwater. These models of computations are more complex by the fundamental nonlinearities in forecasting groundwater relationships, the impediments of estimating water levels and therefore extraction with accuracy in such environments (Shanafield and Cook, 2014). Hence, this study will establish less sophisticated conceptual models of groundwater modelling using historical rainfall datasets.

Proper data capturing, coordination, and forecasting levels of groundwater degrees in artisan wells is distinctly crucial for the realistic control of groundwater systems (Daliakopoulos *et al.,* 2005). Monitoring groundwater levels provides researchers with significant information to understand trend variations in groundwater availability (Adamowski and Chan, 2011). Time series model's abilities to summarise and forecast groundwater levels without acquiring

© University of Venda

knowledge of the underlying geological parameters makes them interesting models compared to numerical physical models (Daliakopoulos *et al.,* 2005).

The importance of trends in groundwater level and the impact it has on the stability of water is expressed through its connection with climate variability. International studies have recently presented the results of climate models that show climate change has reduced the availability of groundwater, which has great significance on implications for water resources management and reduced access to clean water (Lorenzo-Lacruz *et al.,* 2017). Therefore, accurate forecasting models and comprehending the impact of exogeneous variables on the groundwater level can become a significant part of guaranteeing the importance of water use in the implementation of groundwater resource management plans, especially with the increase in water demand in urban areas (Tubaw *et al.,* 2017) and forecasting of climate variability show an increasing dependence on groundwater (Lorenzo-Lacruz *et al.,* 2017).

The application of physical models is complicated by the irregular systems of groundwater levels and their response to exogenous factors, since vast volumes of data are needed for such correct interpretation (Ibrahim Ahmed Osman *et al.,* 2021). However, new modelling approaches conceptualised on time series modelling become significantly reasonable to be used for groundwater level forecasting, as these modelling methods never require any data about the psychological characteristics of the system (Hai Tao *et al.,* 2022).

This study will provide an understanding of the relations between different meteorological conditions (exogeneous variables) and groundwater system and study the correlations of groundwater levels to the exogeneous variables by using time series models. Through the methodology, one of the objectives is to show any possible strategy and model performance analysis of time series modelling. This study then confers to coexisting conversations (as brought up by, for example, Candela *et al*. (2013); Johnston and Smakhtin (2014)) regarding operations about modelling attempts in groundwater level.

© University of Venda

1.4 Objectives of the study

1.4.1 Main objective

- To model groundwater levels using time series models in a semi-arid catchment.

1.4.2 Specific objective

- To set up and calibrate a time series model for groundwater levels modelling.
- To evaluate performance of the model.
- To determine possible climatic factors that are affecting groundwater levels.
- To determine patterns in time series in the study area.

1.5 Research questions

- How does rainfall and evaporations in semi-arid catchments influence groundwater levels?
- How will groundwater levels respond to exogeneous variables?
- What is the performance of a time series models in groundwater levels?
- Does the relationship between groundwater levels, temperature and rainfall provide enough data to evaluate the performance of a time series model?

© University of Venda

1.6 Study area

1.6.1 Location

Luvuvhu River catchment is located in the most northern district municipality of the province of Limpopo in South Africa. It lies between the lines of longitudes 28°54′20.66″ E and 31°04′31.19″ E and the lines of latitudes 22°7′26.59″ S and 23°26′6.49″ S. This area is known to shares borders with Mozambique and Zimbabwe (Maponya. 2011). With an area of 21407.08 square kilometres, this district is one of five district municipalities in the Limpopo Province.

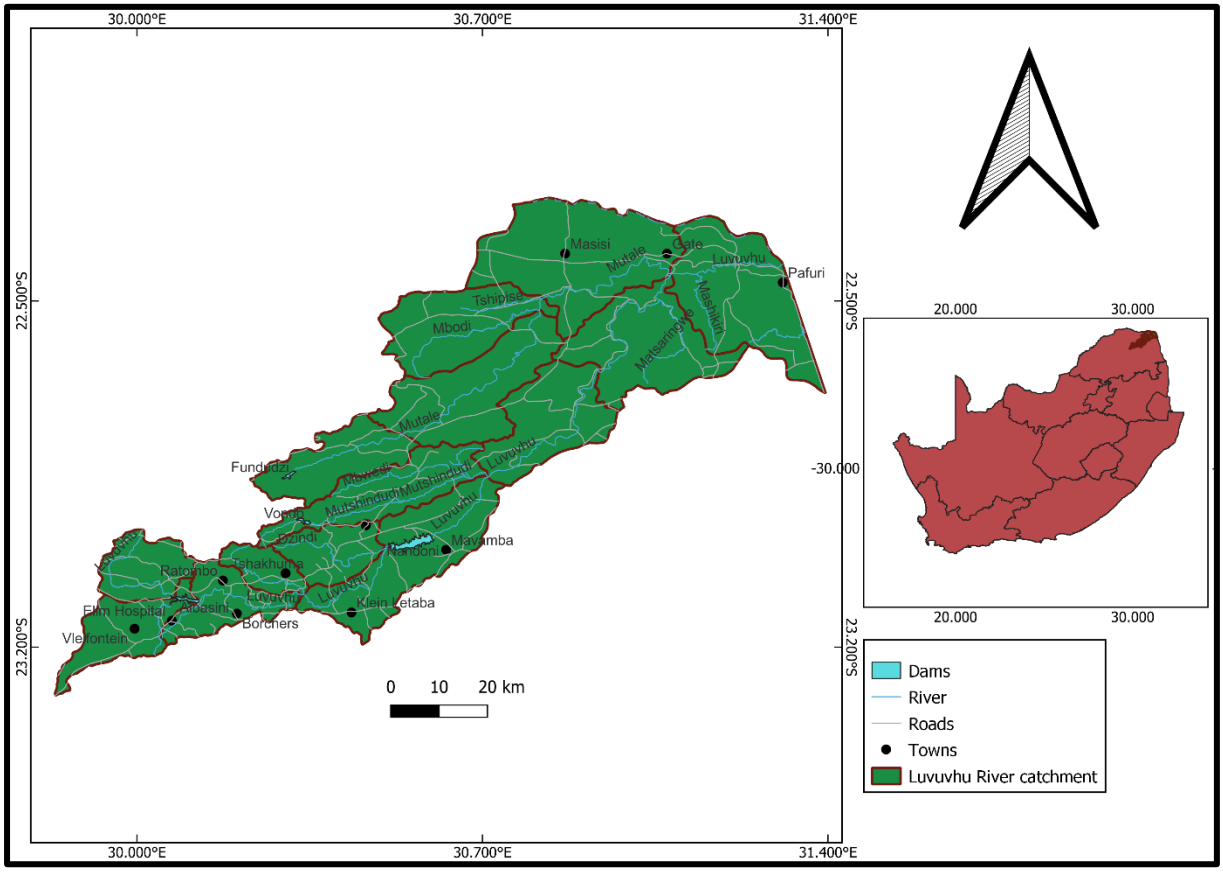


Figure 1.1: Map showing the location of the study area

1.6.2 Geology

The essential geographical characteristic of the district is the soutpansberg mountains. Soils withinside the region is variable, and being sandier withinside the west, however with a better clay and loam contents material closer to east side. The underlaying rock of this entire area is crystalline rock, with an addition of the Archean age basement complexes for example, conglomerate, Modimolle formation composed of lava, granite, arenite, gneiss, greywacke and

© University of Venda

siltstone occupy this area. Aquifers like, dolomite, limestone and carbonated shale Limestone provides water supply in this region (Momba and Mpenyana, 2012). Shallow groundwater flow with limit range of 50m below ground level with fractured and weathered zones are main characteristics of groundwater in this region (Martin, 2011). Most boreholes drilled have a range of 30m to 120m depth with an annual fluctuating range between 8.50m and 30m. The water level fluctuates between 8.50m and 30m annually, bore wells depth ranging from 30 to 120m.

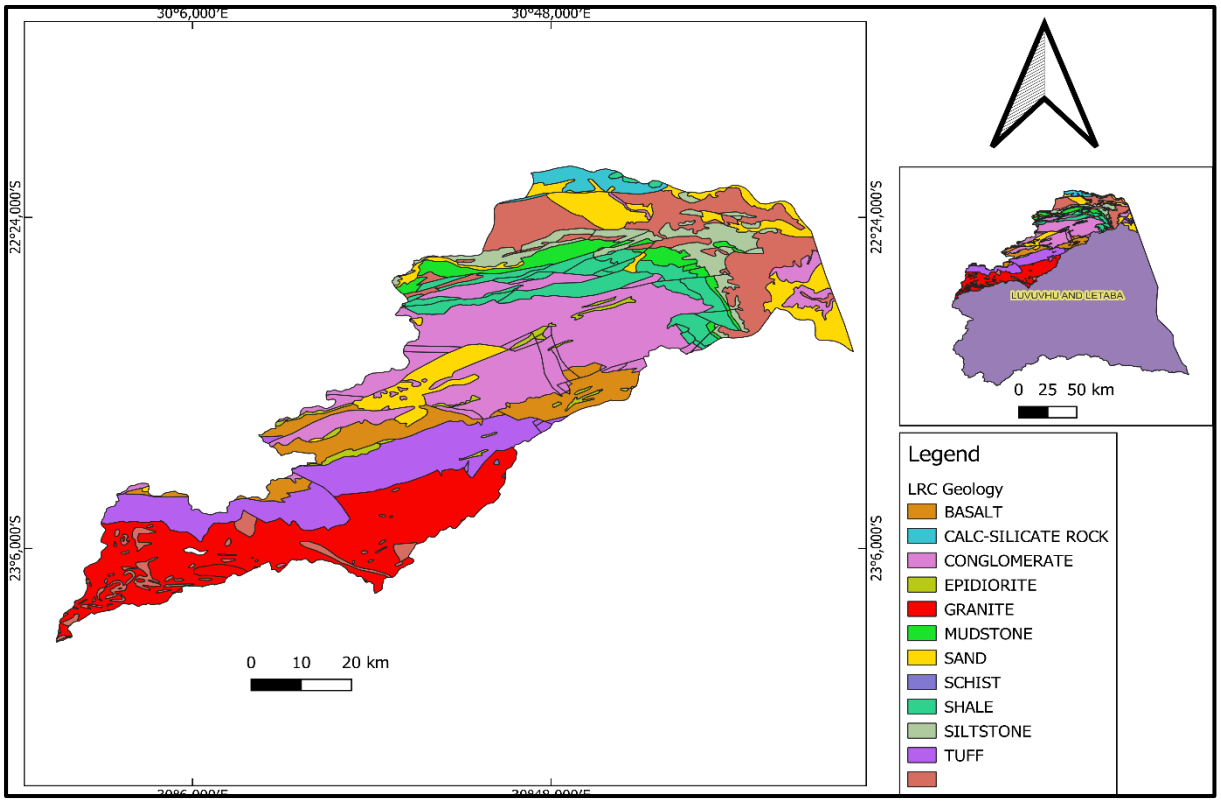


Figure 1.2: Map showing the geology of the study area.

1.6.3 Climate

This region is generally semi-arid with peak precipitation recorded between January and February averaging 300 to 1000 mm per year which declines from east to the west because of the relief rainfall effect of uKhahlamba mountains connecting the Zoutpansberg perpendicularly. The area has an average maximum temperature of approximately 40 °C from October to March characterised by a warm wet season. Low temperatures ranging from 12 and 22 °C are experienced during the cool dry season which begins from April to September (Edokpayi *et al.,* 2015).

© University of Venda

Chapter 2: Literature review

2.1 Preamble

This session reviews literature from previous studies on time series modelling of groundwater levels in a semi-arid catchment. It also reviewed some of the time series methods which have been used to evaluate interactions between groundwater levels and meteorological patterns.

2.2 Time series analysis of groundwater levels

Time series analysis is the numerical statistic method which has shown to continue to be chosen by a lot of researchers for a significant long-term design of study. Such studies usually include one focused upon subject or unit of study that has been continuously recorded at random or uniform time intervals over many observations. Time series analysis helps us understand natural processes, patterns of change over a period, and investigates effects of organised interventions (Velicer & Molenaar, 2012). Time series simulations of groundwater levels can help to identify trends and behaviour and to identify the coordination and use of local groundwater systems (Chen *et al.,* 2007; Hessel and Hirsch, 1992).

Modelling and forecasting of groundwater levels can aid in ensuring the utilisation of aquifers for water supply in a sustainable manner. Accuracy in forecasting of groundwater levels is essential for engineers and catchments managers to minimize harmful impacts and sustain stability in supply and demand. For groundwater level modelling, there have been reports of significant limitations and complexity using a few concept models because they require datasets in a large scale (Raji *et al*., 2019). Since there is a clear shortage of data in many regions, utilising these methods has been difficult. Recent related time series model analysis represents a more complicated leap in the ability to analyse datasets collected for one subject (Veliser and Molenaar, 2012).

In many groundwater studies, time series models are commonly used for groundwater level prediction including support vector machines (SVM), adaptive fuzzy neural inference systems (ANFIS), artificial neural networks (ANN) and hybrid models for example, wavelet artificial intelligence (Rajai *et al*., 2019).

Recent studies of groundwater level modelling also indicated that ANNs can provide a suitable alternative to some concept methods. A study comparing distinct types of ANN conducted by Coulibaly *et al*. (2001) utilised rainfall, groundwater level and temperature as variables to the model to forecast mean average monthly groundwater levels in the Gondo aquifer, Burkina

© University of Venda

Faso. Forecasting results indicated that Recurrent Neural Networks model was significantly efficient irrespective of linear variable delay ANN and Radial Basis Function (RBF)-ANN.

Qiao and Cong (2016) used Radial Basis Function framework (RBFN) together with the ARIMA model to estimate and simulate groundwater levels. In the study, linearity of the time series was simulated utilising ARIMA, while non-linear residuals were computed by RBFN model. These methods were utilised to simulate monthly variable groundwater in two monitoring wells in Xi'an, China. Groundwater levels monthly data ranging 1998 to 2008 were used for training while the remaining data up to 2010 was kept for model testing. The simulated outcome from the model were compared using accuracy measures. The forecast accuracy of the hybrid model was the highest, indicating that combining the methods was effective and can be used in forecasting groundwater levels.

Across some part of Australia, Hydrograph Analysis: Rainfall and Time Trends (HARTT) model was used to represent groundwater level as an accumulation of extraction, variable land cover and monthly precipitation. The model showed reliability in distinguishing precipitation as an input from other variables and pointed out the lag difference on its influence on groundwater levels over a period (Ferdowsian *et al.* 2002).

Ebrahimi and Rajaee (2017) computed variable groundwater levels from two wells in the Qom plain of Iran using a neural network, linear regression, and joint wavelet with SVM. Average monthly groundwater level data was utilised to simulate the model over a period of 11 years. The study showed a significant improved model training due to the time series decomposition of groundwater level into sub-time series.

Nourani and Mousavi (2016) modelled the spatiotemporal groundwater level at the Miandoab plain in north-western Iran using hybrid artificial intelligence methods. The groundwater time series was observed in different piezometers that were extracted utilising wavelet-based methods, together with effects of noise and noise data were compared in temporal groundwater level modelling by ANN and ANFIS. Both models were trained and tested using groundwater level data, precipitation, and runoff site. Different scenarios of input variable selection to forecast groundwater levels one month in advance were also considered. The results showed that the wavelet-based noise extraction approach can improve the modelling performance by an approximate of 13.4%. In the validation step, it was also indicated that the ANFIS-RBF model was more reliable than the ANN-RBF model in terms of accuracy.

© University of Venda

Maheswaran and Khosa (2013) argued that the coupled Volterra wavelet model was capable of predicting groundwater levels. This was after they used a nonlinear dynamic wavelet-based model (Volterra Wavelet Coupled Model) because of its reliability in providing forecasted groundwater levels over an extended period in two locations in Canada. The results of the model were put against other latest methods results for example, ANN and Wavelet Linear Regression (WLR).

Tabari *et al.* (2012) analysed groundwater level variability by utilising Mann-Kendall trend test together with the Sen's slope trend estimator to estimate annual, seasonal and monthly trends in the levels of groundwater in northern Iran. They showed that a non-parametric approach can be applied to analyse groundwater levels.

Trihakis *et al.* (2010) stated that ANN can forecast groundwater levels and was a reliable method for simulating karst aquifers, which are difficult to simulate with digital groundwater modelling. This study was carried out in the Edwardian Basin, Texas, USA utilising ANN model to simulate karst groundwater levels.

Peterson and Western (2011) temporally modelled groundwater levels and composition over historical climate periods to present the latest hydrograph analysis method of that time. The study used soil moisture variables to expand an existing TFN model and showed computation measures of groundwater levels with only atmospheric forcing and does not follow the unclear temporal trends or first-order stationary atmospheric forcing based on the HARTT model. Four variants of the TFN and HARTT models were used at monitoring sites at the forest regeneration region in Great West Victoria. Results showed that hydrograph dispersion can provide important insights into limited trends and lags in groundwater level.

According to Shirmohammadi *et al*. (2013) who utilised multiple variances of data series models including ANFIS and time series models, to forecast groundwater levels for a range of period over time. They stated that the ANFIS can accurately predict groundwater levels 1 and 2 months in advance.

Mirzawand Ghazawi (2014) used a stochastic model to predict groundwater levels in arid environments using time series techniques. In this study, different components of the seasonal ARIMA model were utilised to forecast the level of groundwater in the Kashan plains, Isfahan, Iran. To reduce the volume of computation, the depth of the groundwater level in the piezometric wells was divided into five groups. From the results of the study the best prediction was showed by the Auto Regressive component with multiple regression. The study also

© University of Venda

pointed out that time series models have a upper hand in predicting groundwater levels if the models are combined.

von Asmuth *et al.* (2012) used the PIRFICT method to integrate the data and physical basis for modelling of groundwater levels. The study used the implementation of site interpretation and knowledge of physical behaviour, while exploring software programming for hydrogeological historical datasets analysis and linking data to field concepts. The study concluded that the PIRFICT method coordinates the simulation of time series and distributed models. The method also proved to be resourceful for different time series modelling concepts.

Atta-Darkwa *et al.* (2013) showed how groundwater levels are affected by climate events such as droughts and floods. The study used a third-dimension flow model with Processing Windows for Modflow to compute groundwater levels in two lower alluvial aquifers in the Inland Besease Valley in Ghana. Estimated changes in groundwater levels from the variable water level method were utilised as a hydrograph model. The study concluded that the groundwater levels had a higher range in metres in the wet season than in the dry season for the simulation time series.

Lallahem *et al.* (2005) determined groundwater levels in an undefined limestone aquifer in northern France. The study used an ANN model groundwater level in 13 wells, precipitation, transpiration and average temperature as input datasets for the model with the objective simulate groundwater levels in the selected wells. Simulations revealed the benefits of using the MLP model.

Jalalkamali *et al.* (2011) pointed out that the best exogeneous variables combination to forecast groundwater levels in two neighbouring wells is to apply the current and a-month-ago GWL of one well and the neighbouring well with different combinations of precipitations and temperature. The capabilities of ANN and ANFIS to forecast the groundwater level were investigated and concluding that better results were shown by the ANFIS models compared to ANNs.

Collenteur *et al.* (2021) used the TFN model using an impulse response function to estimate groundwater levels. A nonlinear root region version that simulates recharge has evolved withinside the TFN version and as compared to extra usually used linear recharge models. The model was trained and tested on groundwater level data monitored at the Wagna Hydrological Research Centre in south-eastern Austria. Nonlinear models improve groundwater level 12 simulations compared to linear models. The annual discharge rate calculated by the nonlinear

© University of Venda

model was compared with the average transpiration rate observed by the two lysimeters. Release predictions from the nonlinear model also agree with lysimeter data at release periods shorter than 10 days.

Manzione (2018) modelled time series to investigate water level depth to model random continuous data. The study demonstrated the reliability of physically based period to determine the interaction between precipitation and groundwater level from geological and hydrological monitoring data. The PIRFICT was used to analyse the data, then simulate the level of the groundwater. The results allowed providing of information on the dynamics of the water table even in the case of average, maximum and minimum simulations, influenced by different climatological patterns.

2.4 Time series methods of modelling groundwater levels

2.4.1 Auto Regressive Integrated Moving Average (ARIMA) Model

The ARIMA model (Box and Jenkins, 1976) is a well-recognized tool for modelling and for predicting time series datasets. ARIMA has three components, that is; autoregressive (AR) which is indicated as p, integrated (I) as d, and moving average (MA) component indicated as q. The time series is denoted by Yt as:

$$Y_t = Y_1, Y_2, Y_3 \ldots . Y_t \tag{2.1}$$

Where, Y_t is the recorded value at time *t*, and Y_1, Y_2, Y_3….. are detected at time series t_1, t_2, t_3….

The AR component elaborates on the interactions and response between the previous p observations and the present. Every single present observation can be useful to only one previous observation, hence $p = 1$, that is:

$$Y_t = \mathrm{c} + \phi_1 Y_{t-1} + \mathrm{e}_t \tag{2.2}$$

where Y_{t-1} shows earlier detected estimate at period t-1, e_t is irregular lapse and c and φ1 remain the same. If $p > 1$ then the detected estimates of the period can be added to the far-end of the equation:

$$Y_t = \mathrm{c} + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_P Y_t\text{-}P\mathrm{e}_t \tag{2.3}$$

The I component of this version decides if the detected estimates are simulated linearly, or variance amongst successive detections are computed. The detections are directly simulated if

© University of Venda

$d = 0$. If $d = 1$, The variations among successive detections are simulated. If $d = 2$, the variations of the distinction are simulated. In implementation of the model, d cannot be more than 2.

The MA component elaborates on the interactions between initial q errors and detected values, if q=1, every detected value is useful to only one initial error, which is:

$$Y_t = \mathrm{c} + \Theta_1 \mathrm{e}_{t-1} + \mathrm{e}_t \quad (2.4)$$

here, c shows a term that remains the same, e_t is the irregular lapse at period t and e_{t-1} stands for the antecedent irregular lapse at period *t-1*. If $q > 1$ more irregularity will be added in the far-end of the equation.

The ARIMA time series model includes four steps when developing the model in a programming software. The steps in their order are identification of the model, estimating p, d and q, testing diagnostics, and then predicting future values (Samsudin *et al.,* 2011). In model establishment, the dataset is analysed for stability and uniformity with proper variance and series variance. The stability and the seasonality of data series are checked by the behaviour of the ACF and PACF. By fitting ACF and PACF plots of the stationary data series, the test model is selected. In the next step it is necessary to estimate the parameters of the model. Usually, the maximum likelihood method that is asymptotically valid for time series is used to estimate the parameters of the model. Especially the estimation of the ACF is important for model identification. The first order selection of these parameters (p, d, q) is determined using the initial values of the model. It is crucial to check whether the estimations about the model errors are met, hence the use of model adequacy test. If the model is not sufficient, then the new test model will be selected again, and the process will continue for the selection of required model and the observed time series that are used to fit perceive and use the model. And thus, the correlograms are calculated and fit and predicted values are obtained by the model (Patle *et al*., 2015).

2.4.2 The transfer-function noise model

The transfer-noise function (TNF) model is basically an externally forced time series auto regressor (ARMAX). The model does not have a need for uniform interval monitoring without missing data, which makes it suitable for groundwater level modelling. From the model Von Asmuth *et al.* (2002) depicted, the rise in levels of groundwater, ht, is the sum of the main linear parts, which is:

© University of Venda

$$h_t = h_t^* + e^t + d \tag{2.5}$$

where h_t^* shows addition of water level at period t because of the joint influence of the input parameters, e^t is the remaining series; and d remain the same local drainage in respective to the rise in water level.

According to Collenteur *et al. (*2021) and Von Asmuth *et al. (*2002), h_t^* is computed by positioning stimulation feedback useful on every input variable (for example, precipitation or evapotranspiration) to redesign the variable difference in groundwater level (Equation 2.6) from *d*. Each simulated reaction recorded previous errors, and the necessary difference in the main input variable at time *t* can be computed by mathematically including all recorded errors and adding all the previous errors (Equation 2.6).

$$h_t^* = \sum_{i=1}^{m} \left\{ \int_{-\infty}^{t} \theta_i \, (t - T) R_i(T) d_T \right\} \tag{2.6}$$

Where $\theta_i(t - T)$ is the detected stimulated feedback result for variable i. R_i (T) shows selected exogeneous variable and m [−] shows sum of all input variables to be simulated.

When Equation 2.6 is used, Von Asmuth *et al.* (2002) used R_i precipitation and potential evapotranspiration to model climatic stress to the vegetation. But because the interception, runoff and recharge, are non-linear functions of climate, and therefore the groundwater head is in turn also not linearly related to climate. Therefore, to model the system, Collenteur *et al.* (2021) expanded the model of Von Asmuth *et al*. (2002) using climate and introducing to that model a significant extensile soil texture module. Relying on the effective that is sought, the soil-texture version may involve amongst one and 5 variables. All these models are evaluated on 620 bores across Victoria Australia's hinterland, and Siriwardena *et al.* (2011) determine that the following soil model is the most parsimonious:

$$\frac{ds}{dt} = P\left(1 - \frac{s}{s_{cap}}\right) - E\left(\frac{s}{s_{cap}}\right) \tag{2.7}$$

Where S_{cap}[L] shows an input variable for soil-texture reserve ability; s[L] shows the time t soil texture; P[LT^{-1}] shows the rainfall rate and E [LT^{-1}] shows the rate of evapotranspiration.

According to Peterson and Western (2014), Equation 2.7 is adapted for converting evapotranspiration and rainfall into inputs to the data series were replaced by estimates of soil water-free drainage, calculated as: $\left(\frac{s}{scap}\right)^{\gamma}$, where γ is a parameter for the pore index. The

© University of Venda

evapotranspiration input turned into changed with the aid of using groundwater ET as an approximation. This was computed as $E\left(1-\frac{s}{scap}\right)^{\gamma}$, that is, evaporations without the simulated soil moisture. Naturally, it is necessary to compute as modelling vegetation groundwater utilization for prolonged dry periods. Therefore, during dry seasons. The resultant model equation is:

$$h_t = \int_{-\infty}^{t}\left(\frac{s_t}{s_{cap}}\right)^{\gamma}\theta_p(t\text{-}\tau)d\tau - \int_{-\infty}^{t}E_t\left(1\text{-}\frac{s_t}{s_{cap}}\right)\theta_E(t\text{-}\tau)d\tau\text{-}\int_{-\infty}^{t}Q_t\,\theta_F(t\text{-}\tau)d\tau + e_t + d \qquad (2.8)$$

here E $[LT^{-1}]$ and Q $[L^3T^{-1}]$ are the groundwater extraction rate and evapotranspiration θ_p, θ_E and θ_F $[-]$ are the stimulated feedback useful water flow, groundwater evapotranspiration and extractions.

Naturally, the data period model that Peterson and Western (2014) proposed has included the recovery and evaporation part of groundwater, together with the extraction part was included in the model. Overall, the three integrals in Equation 2.8 indicate the impact of the major variables of the model, which are groundwater level that depends on precipitation and evaporation, groundwater ET, and soil moisture controlled by pumping.

2.4.3 Convergence between response functions and stress time series

The evaluation of time recordings is done using predetermined impulse reaction equations computed by von Asmuth *et al*. (2002). The time reaction function θ(t), for each stress is labelled based on the variables' shape. The variation levels of groundwater *h(t)* (Equation 2.9), at time t, due to a stressor that changes over time *N(t)*, is determined by combining θ(t) with *N(t)* (Von Asmuth *et al.,* 2002).

$$h(t) = \int_0^t N(\tau)\,\theta(t-\tau)d\tau \qquad (2.9)$$

During simulations, the increase and decrease in stress are usually recorded for a specific lapse Δt. The changing h(t) indicated from the fluctuated stress N used amongst the period, Δt is given by:

$$h(t) = N\left(\int_0^t \theta\,(t-\tau)d\tau - \int_{\Delta t}^{t}\theta(t-\tau)d\tau\right)\text{for } t \geq \Delta t \qquad (2.10)$$

© University of Venda

Equation 2.10 defines the step response as the integral $s(t) = \int_0^t \theta(t-\tau)d\tau$ which describes the groundwater head's response to step feedback starting at time $\int_0^t \theta$. To optimize the detected fluctuations, the impulse response functions' parameters are adjusted. It is a linear method that is comparable to the surface-water hydrology unit hydrograph in many aspects. Experience shows that the technique's linearity often yields a decent estimate of the observable head changes, even though it may seem restricted when considering non-linear events like groundwater recharge (Manzione *et al*., 2012, Von Asmuth *et al.,* 2008).

Signal decomposition: Using preset impulse reaction functions, the time series evaluation technique aims to divide recorded head variations into fractional variation series, each of which is caused by a stress that has been applied (Von Asmuth *et al.,* 2008). The deterministic part of the time collection model is formed by adding the fractional variations and an unaltered known as the 'drainage base'. The drainage base is the highest point obtained with all stress adjusted to nil. The variations amongst the time collection of found heads and the time collection estimated using the predictive version method is the residual time collection. The structure of a time series of measured heads ho(t) can be summarized as follows:

$$h_0(t) = d + \sum_i h_i(t) + n(t) \tag{2.11}$$

where *d* is the drainage base, $h_i(t)$ denotes variations caused by stress *i*, and the residual time series *n(t)* of variations amongst measured and calculated heads using the predictive model (Von Asmuth *et al.,* 2008). If the residual time collections characteristics differ greatly from those of white noise, modelling the residual time collection is likely required (Obergfell *et al*., 2013).

Predefined response functions: Response capabilities are set up for each pressure to precisely mimic its impact at the head. Von Asmuth *et al.* (2008) proposed response capabilities for precipitation and evaporation, which are also utilized in the Meny Anthes programming tool (Obergfell *et al.*, 2013). The scaled gamma density function θp(t) describes the impulse response capabilities that describe the impact of precipitation.

$$\theta_p(t) = A\frac{a^n t^{n-1} e^{-at}}{\Gamma(n)} \tag{2.12}$$

In which *t* represents time, *Γ(n)* is the gamma feature of n, a and n define the feature's form, and *A* is a scaling factor. A scaled incomplete gamma function is the corresponding step reaction:

© University of Venda

$$S_p(t) = \frac{A}{\Gamma(n)} \int_0^t a^n \tau^{n-1} e^{-at}\, d\tau \tag{2.13}$$

2.4.4 The TFN Model of Box and Jenkins: In discrete time

For uniform, uninterrupted aquifer structures, which can be prompted through rainfall surplus only, the subsequent exogeneous variable discrete TFN version may be used to version groundwater stage fluctuations:

$$h_t = h_t^* + n_t + d \tag{2.14}$$

$$h_t^* = \Theta(B)p_t = \sum_{i=o}^{\infty} \Theta_i p_{t-i} \tag{2.15}$$

$$n_t = \varphi(B)a_t = \sum_{i=0}^{\infty} \varphi_i a_{t-i} \tag{2.16}$$

where d is the level of h_t^* without rainfall. n_t is the residual series [L], h* is the predicted groundwater level at time step t attributable to the precipitation surplus, relative to d [L], and h_t is the recorded groundwater level at time step t (with t ∈ N, (dimensionless)), with relation to an indicator level [L]. The dimensionless transfer function is denoted by Θ(B), a_t is the zero mean discrete white noise process [L], φB is the dimensionless noise transfer function, p_t is the precipitation excess at time step t [L], and B is the backward shift operator (dimensionless), defined as $B^i p_t = p_{t-1}$.

The switch feature Θ(B) is expressed as a fraction in Box-Jenkins (BJ) methods, with an AR function δ(B) as the denominator and a so-called MA feature ω(B) as the numerator. This means that Θ(B) = ω(B)/ δ(B). The transfer function's weights $\Theta_0, \Theta_1, \Theta_2 \ldots \Theta_\infty$ are commonly referred to as the impulse response (IR) function. However, from a continuous perspective, this response is the block-shaped reaction to exogenous variables.

To prevent misunderstandings, the discrete case is referred to as the transfer function, while the response to an actual instantaneous impulse is referred to as the IR function. Selecting of a delayed time and the range of MA and AR parameters, which collectively constitute the nodes of the transfer functions in equations (2.15) and (2.16) and specify their general shapes, are used to characterize the version shape in the version identity segment of discrete TFN models:

$$\Theta(B) = B^{nk} \frac{\omega(B)}{\delta(B)} = B^{nk} \frac{\omega_0 + \omega_1 B + \omega_2 B^2 + .. \omega_{nb-1} B^{nb-1}}{1 + \delta_1 B + \delta_2 B^2 + .. \delta_{nd} B^{nd}}$$

© University of Venda

$$\varphi(B) = \frac{\theta(B)}{\phi(B)} = \frac{1+\theta_1 B+\theta_2 B^2+..\theta_{nc}B^{nc}}{1+\phi_1 B+\phi_2 B^2+..\phi_{nf}B^{nf}} \quad (2.17)$$

Where n_b, is the value of MA components of the transfer model, n_c shows the value of MA components of the noise model, n_d is the AR component value, n_f is the value of the noise model, n_k is the lag time.

In Continuous Time: The PIRFICT Model are generally simulated together with TFN models, such as variability of climatic conditions and groundwater levels, are not limited but nonstop. In this nonstop period step, expanding the data series with a direct, time-steady transfer function provided by an essential complication (von Asmuth *et al.,* 2002). Therefore, equations 2.14 – 2.15 are presented in a nonstop period as:

$$h(t) = h^*(t) + n(t) + d \quad (2.18)$$

$$h^*(t) = \int_{-\infty}^{t} \theta(t-\tau)p(\tau)d\tau \quad (2.19)$$

Besides the transfer function of the limited TFN model, the torsional impulse response function q(t) does not require the observed rate of variables sequence. It defines the dynamic feedback of a system to rapid stimulant, constant period, and essential property of a given structure (von Asmuth *et al.,* 2002). In hydrological terms, the IR function depicts how the water table responds to a sudden surge in the surplus of rainfall. In that regard, it is similar to the continuous unit hydrograph applied to water surface hydrology (Dooge, 1973). Equation (2.16) can be converted to continuous time by swapping out the discrete white noise process with a continuous white noise process. The residual series n(t) can therefore be described as a coloured randomized procedure, as seen by the randomized cumulative below:

$$n(t) = \int_{-\infty}^{t} \phi(t-\tau)dW(\tau) \quad (2.20)$$

In which W(t) is a continuous white noise (Wiener) process [L], with properties:

$$E\{dW(t)\} = 0$$

$$E[\{dW(t)\}^2] = dt$$

$$E[dW(t_1)dW(t_2)] = 0, t_1 \neq t_2$$

To account for spontaneously accessible data in a continuous manner with no equivalent intervals, *t* is analysed and used in real time using *i*. A sequence of N discrete detections of a

© University of Venda

continuous process, combined with groundwater stage variations, can thereafter be expressed as:

$$h(t_i) = [h(t_0), h(t_1), h(t_2), \ldots . h(t_N)] \quad (2.21)$$

In equation (2.18) the drainage level d is normally unidentified, but can thereafter be removed from the equations by adding both sides of equation (2.18) for every time steps (from t = t_0 to t_N), factoring the end outcome through a means of *N*, and estimating that there would be no drainage fluctuations, which gives:

$$d = \frac{\sum_{i=0}^{N} h(t_i)}{N} - \frac{\sum_{i=0}^{N} h^*(t_i)}{N} - \frac{\sum_{i=0}^{N} n(t_i)}{N} = \bar{h} - \overline{h^*} - \bar{n} \quad (2.22)$$

Integrating Equations (2.18) and (2.22) yields a general formula for a continuous TFN model in which all elements are centred on their corresponding periodic is typical (the symbols on the overbar), and d is absent:

$$\{h(t) - \bar{h}\} = \{h^*(t) - \bar{h}\} + \{n(t) - \bar{n}\} \quad (2.23)$$

2.4.5 Forecasting Model Based on the Singular Spectrum Analysis

In SSA designation, the series is frequently assumed to be noisy, with an indeterminate number N for the series. The SSA method entails two basic complimentary phases: decomposition and rebuilding. The noisy series is deconstructed in the first stage, while the noisy differenced series is recreated in the second stage. The regenerated series will be utilised to forecast future values.

Assuming the stochastic method {cm(t)}, t = 1, 2, .., N; $m \in$ {1, 2, .., M} has an assumption that the outcome of size N from this procedure is available Cm(t) = {cm(1), cm(2), . . . , cm(N)}. Because we are dealing with time-invariant series, we may simplify the realization by rewriting it as follows: $C_N = \{C_1, C_2, \ldots . C_N\}$.

The first step of the method, known as decomposition, consists of two steps: embedding and singular value decomposition (SVD).

Embedding transforms a one-dimensional time series with centres centers $C_n = \{C_1, C_2, \ldots . C_N\}$ into a multivariate matrix [Y_1,….Y_K] with vectors.

$$Y_j = \left(c_{j,\ldots,}c_{j+L+1}\right)^{\mathrm{T}} \in R^L$$

The window length is denoted by L ($1 < L \leq N - 1$), and K is calculated as $N - L + 1$. The window length denotes a vector of L values from the original series. The common value of L

© University of Venda

is *(N + 1)/2* if N is odd, and *N/2* or *(N/2) + 1* if *N* is even. This step yields the estimated simulation:

$$\rho_{1,2}^{\omega} = \frac{\left|\vec{\hat{C}}_{1N}, \vec{\hat{C}}_{2N}{}^{\omega}\right|}{\left\|\vec{\hat{C}}_{1N}\right\|^{\omega}\left\|\vec{\hat{C}}_{2N}\right\|^{\omega}} \tag{2.24}$$

Where:

$\left|\vec{\hat{C}}_{1N}, \vec{\hat{C}}_{2N}{}^{\omega}\right|$ - absolute value of the weighted Frobenius inner product,

$\left\|\vec{\hat{C}}_{iN}\right\|^{\omega}, i = 1,2, \ldots, r$ - the weighted norm

w(t) = min (t, L, N − t) - vector of weights

If the w-correlation between the two reconstructed components is zero, they are considered to be well separated. The reconstructed components may be merged into a single group and represent the same component in the SSA decomposition, according to high w-correlation values between them. The grey-scaled correlation matrix (λL × λL) effectively represents the obtained correlations.

The uniform vector of coefficients $R = (a_{l-1}, a_{l-2}, \ldots a_1)$T is simulated as follows:

$$\mathcal{R} = \frac{1}{1-v^2}\sum_{i=1}^{r} \pi_i u_i^{\nabla} \tag{2.25}$$

The h-step ahead SSA forecasting is achieved by the following equation:

$$\{\tilde{c}(t)\}^{\mathrm{T}} = \begin{Bmatrix} \{\hat{c}(1), \ldots \hat{c}(t)\}, & t = 1, \ldots . N \\ \mathcal{R}^T C_h, & t = N+1, \ldots ., N+h \end{Bmatrix} \tag{2.26}$$

Where:

$$C_h(t) = \{\tilde{c}(N - L + h + 1), \ldots . \tilde{c}(N + h - 1)\}^{\mathrm{T}} \tag{2.27}$$

© University of Venda

Multichannel singular spectrum analysis (MSSA)

Multichannel singular spectrum analysis (MSSA), first established by Broomhead and King (1986) to elaborate on ununiform structural phenomena, is a projection of SSA for numerous (multichannel) time series that can thereafter be easily comprehended by compiling the 2D matrix slices of SSA for every interval of time accumulation alongside the 3D evaluation of the channel axis (Taie and Koch, 2019).

As a result, MSSA could be quite long. MSSA, like SSA, allows for the identification of trends and oscillatory features by breaking down a time series into spectral components (Jolliffe 2002). MSSA, unlike SSA, analyses cross-correlations across many channels, making it a mix of a combination of PCA and SSA (Rangelova *et al*., 2012, Zotov *et al.,* 2017).

The basic procedure of MSSA are as follows. It must be acknowledged that due to the differing historical calibrations of SSA and MSSA, the notation for channels and delays in the two approaches varies (Taie and Koch, 2019). Here, the numbering of Allen and Robertson (1996) was employed (see Figure 2.1)

For an L-multichannel sample with *N* values *Xli (l = 1, . . ., L; i = 1, . . ., N)*, like SSA, L slices

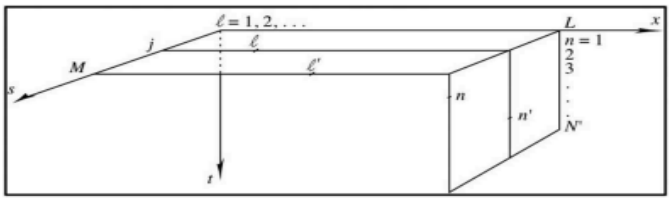


Figure 2.1 illustrates the distribution of multi-channel data for MSSA. The *t*, *x*, and *s* axes represent the length of the time series (*N*), the number of multi-channels (L), and the number of delays (*M*), respectively (source: Taie and Koch, 2019).

There are two exceptions to MSSA: when $L = 1$, it becomes SSA, and when $M = 1$, it becomes conventional PCA (Myung, 2009) of a data estimate computations *Xl*, (l=1, . . ., L) of the lagged variables up to a trajected lag M:

$$X_l = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdot & x_{l,m} \\ x_{1,2} & x_{1.3} & \cdot & x_{l,M+1} \\ x_{l,N'-1} & \cdot & \cdot & x_{l,N-1} \\ x_{l,N'} & x_{l,N'+1} & \cdot & x_{l,N} \end{pmatrix} \text{ for } 1 \leq l \leq L \qquad (2.28)$$

© University of Venda

$N' = N - M + 1$ represents the variety of consecutive views of the set on channel l, which correlates to the window dimension in SSA. Equation (2.28) calculates the complete augmented momentum matrix D $(N' \times M\ L)$ as follows:

$$D = X_1 + X_2 \ldots X_l \tag{2.29}$$

The SVD of D is computed as follows:

$$D = \eta P_D \Lambda_D^{1/2} E_D^T, \eta = \max(N', LM) \tag{2.30}$$

where a normalizing element is indicated by η. The spatiotemporal principal components of the matrix P_D are N′ orthonormal vectors, which are also eigen-vectors of the delayed covariance matrix $C_D^{(p)} = (LM)^{-1} DD^T$,, while the eigen-vectors of the lag covariance matrix $C_D^{(E)} = (N')^{-1} D^T D$. E_D contains ML orthonormal vectors, also known as spatial-temporal empirical orthogonal functions (a time sequence of EOFs).The entries in the eigen-spectrum within the data variance PCs (EOFs), $\Lambda_D = P_D^T C_D^{(P)} P_D = E_D^T C_D^{(E)} E_D$, are proportional to the records variance.

The spatial-temporal EOFs of little curiosity are eliminated:

$$D' = DE_D(1 - K)E_D^T \tag{2.31}$$

If K is a diagonal matrix, and if the ith EOF is kept, $K_{ii} = 0$. If the ith EOF is eliminated, $K_{ii} = 1$. The filtered enhanced historical data for each variable are contained in the matrix $D = x_1 X_2 \cdot \cdots x_l$.

d) Reassemble the filtered data set $X_{L_i}^f$ including the removed series for each channel l, determined by aggregating across the anti-diagonals of X_i^f, l = 1, L, corresponding to SSA. (Golyandina & Korobeynikov 2014; Golyandina *et al.* 2001).

2.4.6 Artificial Neural Network (ANN)

The ANN is a substantially comparable dispersed data processing system with performance characteristics like biological neural networks in the human brain (Haykin ,2005; Nayak *et al.* 2006; Mohanty *et al*. 2010). Biological neural networks provided foundation for ANNs. They can be used to estimate future values for possibly noisy time series using historical data, as well as to approximate unknown functions. An ANN typically comprises of three separate

© University of Venda

layers: input, hidden, and output. When predicting levels of groundwater, the input layer collects factors such as time series of temperatures, precipitation, groundwater levels, and so on. Each neuron in the output and hidden layers produces a result after passing its weighted and biased input through a predefined transfer (activation) function (Rajaee *et al.,* 2019).

ANN models have been widely used in a variety of scientific and technological applications, including time series estimation, pattern identification, and procedure monitoring (Nayak *et al*., 2006). The neural network is characterized using the shape of adjacent nodes, a procedure for assessing the load of adjacent, and a motion mechanism (Emamgholizadeh *et al*., 2014; Mohanty *et al*., 2010). There are several types of neural networks, including Hayfield and Perceptron. Neural networks are classified based on their comprehensive computations. Various types of technical computations include gradient descent with persistence, feed forward back-propagation, variable technical rate back propagation, radial basis function, Levenberg-Marquardt (LM), and Bayesian regularization (BR), among others. (ASCE 2000a; b). The networks are often employed for hydrological efficiency in a multilayer contemporary neural network with a returning propagation analysis method. ANNs with this structure are extensively employed, and they boost capacity to project (Maier and Dandy 2000). The formula used in mathematics for the MLP is as outlined below:

$$y_i = f\left(\sum_{i=1}^{N} w_{ji} x_i + b_j\right) \tag{2.32}$$

Yoon *et al*. (2010) defined the activation function in the present layer as f , where N is the number of nodes in the previous layer, X_i is the ith layer of the previous nodes number. y_j is the jth layer of the present nodes number, b_j is the preconception of the jth node, and w_{ji} is a weight connecting x_i and y_j .

2.4.7 Extreme gradient boosting (Xgboost)

Xgboost is an implementation of gradient boosting machines (gbm) that is regarded as one of the most attractive algorithms used in supervised learning (Ibrahem Ahmed Osman, 2021). It can be used for problems involving both classification and regression. Data scientists favour Xgboost due to its greater fulfilment velocity outside of core compute. The Xgboost operates as follows: A set of data (DS) with m components and n examples $DS = \{(X_i, Y_i): i = 1 \ldots n, X_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$. Let $\hat{y}_i$ be the projected output of an ensemble tree model constructed by the following equations:

$$\dot{A}_{.i} = \phi(X_i) = \sum_{K=1}^{K} f_k(X_i), f_K \in \mathbb{R} \tag{2.36}$$

© University of Venda

Where K represents the quantity of trees in the model and fk shows the (k-th tree), needed to identify the optimum collection of mechanisms by optimizing the reduction in complexity and regularization objectives.

$$L(\phi) = \sum_i 1\left(y_i, \dot{A}._i\right) + \sum_k \Omega\left(f_k\right) \tag{2.37}$$

The loss function, denoted by l, is the variance between the expected and actual outputs, y_i. Using ϕ as a metric of model complexity helps prevent overfitting. And it is computed using:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|W\|^2 \tag{2.38}$$

From the equation, T shows the variety of leaves on the tree, whereas w represents the mass of each leaf. Boosting, which includes introducing a new mechanism f while the system is still calibrating, is used in trees of choices to minimize the objective value. Thus, at the t-th iteration, an additional mechanism (tree) is included as follows:

$$L^{(t)} = \sum_{i=1}^{n} 1\left(y_i \dot{A}._i^{(t-1)} + f_t(X_i)\right) \Omega(f_t) \tag{2.39}$$

$$L_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} - \frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda}\right] - \gamma \tag{2.40}$$

2.5 Input variables selection for groundwater levels modelling

Accurate estimation of groundwater level changes in the natural environment is a complex and non-linear process that is dependent on many different factors such as evaporation, precipitation, and climate. The data that is used is critical in building groundwater level prediction models for a given period. Some academics have developed models to forecast groundwater levels only on the basis of groundwater level lags. Other researchers (Bhagat *et al.,* 2021; Daliakopoulos *et al*., 2005; Zounemat-Kermani *et al.,* 2020; Beyaztas *et al*., 2019; Pathania *et al.,* 2019) combined previous groundwater levels with hydrological variables such as rainfall, transpiration, humidity, temperature, river flow discharge, and so on.

Nouiri and Malek (2020) used a type of ANN for estimating monthly levels of groundwater in four aquifers in Tunisia's Nebhana watershed. The forecasting models were developed using only three input parameters: precipitation, antecedent groundwater levels, and evapotranspiration. The study found that monthly groundwater levels were primarily determined by monthly transpiration, precipitation, and antecedent groundwater level values.

© University of Venda

Iqbal *et al.* (2020) created an artificial neural network (ANN) model to anticipate daily groundwater levels using three meteorological variables in a study area in Pakistan between the Ravi and Sutlej rivers. Rainfall, MinTemp and MaxTemp, radiation from the sun, humidity, speed of the wind (WS), region elevation, and polygon area were all used as input factors. To find the best accurate groundwater level prediction model, the authors used various types of ANN topologies with varying transfer functions, hidden layers, and data percentages during the training, validation, and testing phases. The learning algorithm utilized was Levenberg-Marquardt (LM) back-propagation. The study found that the proposed model was more accurate in estimating groundwater levels based on various statistical criteria. Additionally, compared to other information divisions, the results showed that the tangent sigmoid transfer function was the most effective and that the data split with 80% for training, 10% for validation, and 10% for testing was more potent and hopeful.

According to Rajee *et al.* 2019, who reviewed artificial intelligence methods in groundwater modelling, 67 research studies papers released between 2001 and 2018 were evaluated with regards to the characteristics and capabilities of the modelling methods, input data review, time period for predictions time, and splitting of the sets of data. Among the 67 publications, 52 used the level of groundwater as an input variable. Even 12 articles treated the levels of groundwater as the only auto-correlated variable, without any exogenous variables. In addition to groundwater levels, precipitation was employed 48 times as an input variable. The analysed papers also used evapotranspiration, temperature, surface runoff, water level, extraction rates (pumping), humidity, and other hydrological time series as input variables. The studied works have used additional employed input parameters, such as number of days, population, irrigation patterns, tree rings, seasonal dummy variables, etc., marginally, and it appeared that some of these are challenging to accept at the input consideration stage.

Rajaee *et al.* (2019) then concluded that more attention should be paid during the input consideration stage to the selection of appropriate input variables and lag durations. In the evaluated articles, exogenous variables were frequently chosen solely on the basis of statistical availability or the usage of simple user-described connections. According to Osman *et al.* (2022), the input parameters utilized in groundwater level simulations are usually historical values, with no other external hydrological factors. However, in some other simulations (for example, ANNs), sets of data such as humidity, river flow (surface runoff), transpiration, subsurface water level, and rates of borehole extractions were employed as variable inputs, and

© University of Venda

the outcomes were compared to those of time-series models. However, time-series models outperformed multi-input variable models.

Ibrahem *et al*. (2021) tested and evaluated three machine learning models to predict groundwater levels. The study found that selecting the proper input variables was critical for machine learning models' performance. Exogeneous variables were chosen to validate that the version could estimate the interactions between inputs and expected variables throughout the simulation process. To demonstrate this, three situations were selected to build the prediction models.The exogeneous variables utilised are evaporations, groundwater levels, precipitation and temperature. Scenario 1 had rainfall as the only variable, scenario 2 had rainfall and temperature as exogeneous variables and scenario 3 had rainfall, temperature, evaporation, and groundwater as inputs.

Nalarajan and Mohandas (2015) reported an M5 model tree (M5-MT) for predicting monthly groundwater levels on Jones Island, New Jersey, USA. Between 1970 and 2010, monitoring wells provided a historical data set of groundwater level values. The sets of data from 1995 to 2009 were utilized to calibrate the established model, with remaining datasets used to make certain of it. The optimal input combinations were selected based on previous monthly groundwater level readings (up to 12 months). The observed and anticipated findings were evaluated with R2 and RMSE. The results demonstrated that the created M5-MT model is highly accurate in predicting groundwater levels. However, the authors stated that the influence of other characteristics, such as precipitation, soil permeability, soil moisture, and temperature of the soil, are significant for predicting groundwater levels.

© University of Venda

Chapter 3: Methodology

3.1 Preamble

This chapter presents the data used in the study, and where it was collected from. It also includes the time-period of the data and displays how the period was selected and its influence on the study. It describes data analyses methods that were employed to achieve the objectives of the study.

3.2 Data collection

Climatic data for hydrological years were acquired from South African Weather Services (SAWS). The climatic data were received in a batch containing all the stations within Vhembe District that the SAWS monitors and take measurements daily. The data contained various climatic variables including maximum and minimum temperature, humidity, evaporation, windspeed and rainfall. These variables were used in the study as exogenous variables for model development and prediction on groundwater levels. The daily groundwater levels data for the selected monitoring boreholes were collected from the South African Department of water and sanitation (DWS). The groundwater levels represent the depth to water table beneath the ground surface. The utilised sets of data in this study are in Table 3.1.

Table 3.1: Description of the data employed in the study

Station names	Location		Time step	Data range
Groundwater level stations				
A9N0001	-22.5874	30.6605	Daily	2007 – 2018
A9N0002	-22.4566	30.9374	Daily	2007 – 2018
A9N0003	-22.4249	30.9688	Daily	2007 – 2018
A9N0005	-22.4209	30.8468	Daily	2007 – 2018
A9N0010	-22.57881	30.82308	Daily	2007 – 2018
A9N0011	-22.70289	30.79233	Daily	2007 – 2018
A9N0014	-22.88436	30.67031	Daily	2007 – 2018
Weather stations				
Mukumbani Tea	-22.9200	30.4000	Daily	2007 – 2018
Punda Maria	-22.6920	31.0150	Daily	2007 – 2018
Evaporation stations				
A8E002	-22.72527	30.09471	Daily	2007 – 2018
A9E004	-22.93805	30.16096	Daily	2007 - 2018

© University of Venda

All the input variables data used in the study were daily time series data. The groundwater levels data obtained from DWS were measured randomly throughout the day. For this study the data were converted to average daily time series to fulfil time series modelling and prediction accuracy. In the framework of modelling groundwater levels in Vhembe District and within the catchment, groundwater measurements are referenced to the surface of the land rather than sea elevation because the hydraulic head is based on the area's precise physio-climatic factors and hydrogeological characteristics. Even though sea elevation is a popular reference point in hydrogeological investigations, it may not fully depict the complexities of groundwater dynamics in a place with diverse topography and climate, such as the Luvuvhu Catchment in Vhembe District. Groundwater levels are measured with the ground surface as reference level/ datum (0m). Measurements of groundwater levels considers measuring the empty space from the datum to the depth of groundwater (groundwater table). As such values increasing measured values actual depict increase in the measured empty space and hence decrease in groundwater levels. Values showing decrease in the measured space depict increase in groundwater levels. The fluctuation in rainfall patterns, humidity, temperature, windspeed, and evaporation over the region needs a specific point of reference for accurate assessment of groundwater. Choosing relevant and important input variables is critical when creating machine learning models. Some exogenous variables may be dispensable, noisy, or no longer influence the desired outcome (Fahim, 2024; Maier and Dandy, 1998). This study deliberated groundwater levels data from DWS and evaporation, rainfall, temperature, windspeed, humidity data from the SAWS. Table 3.1 shows the stations used in the study and their distribution across the region is shown in Figure 3.1.

© University of Venda

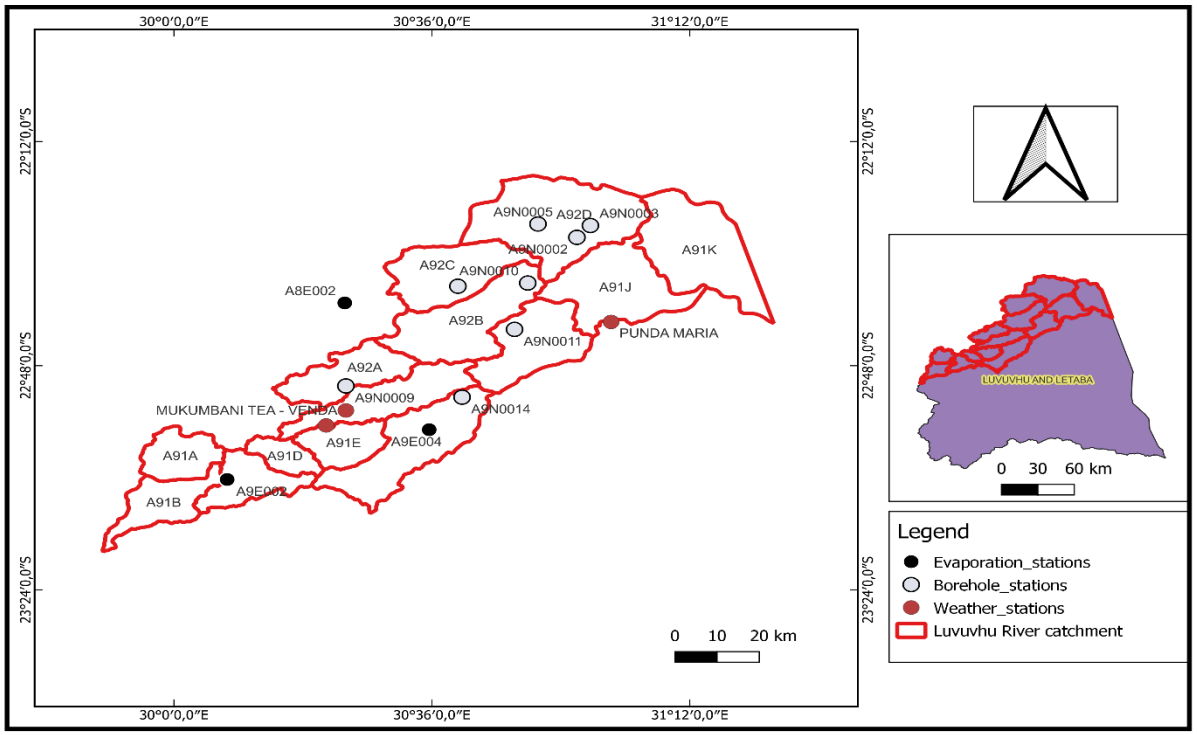


Figure 3.1: Location of boreholes and weather stations

3.3 Methods for modelling and forecasting groundwater levels

This study used an organized technique, commencing with complete data collection, extraction, and analysis. It was then separated into two major elements. First, it evaluated the effectiveness of DWS groundwater level data by creating daily time series and analysing trends. The second portion entailed creating and correlating various machine-learning models for forecasting groundwater levels. The entire investigation was conducted out using Python 3.0 programming tools. This entailed preparing inputs, developing models, conducting rigorous performance evaluations, and simulating average daily groundwater levels.

3.3.1 Data processing and stationary tests

The reliability of groundwater level data was thoroughly assessed, with an emphasis on years for which records were available for study. Each borehole's data was sorted and organized into time series, graph, and box plot histograms for detecting missing values, abnormalities, or outliers. Data samples were examined for outliers, duplicates, skewness, and rapid shifts and unexplained swings. Finally, stations with less than 25% missing data were selected, yielding 07 monitored wells with trustworthy data for further investigation.

© University of Venda

Panda profiling was used to perform early exploratory data analysis, pattern spotting, and insight development. Cubic splines were also used to intact the groundwater hydrographs (Marsden, 1974; Pengzhu *et al.,* 2015). Different statistical models, such as Matplotlib with Python, an ARMA method with exogenous variables (Bidwell, 2005), and SARIMA (Oikonomou *et al.,* 2018), have been computed and used to remove rows with missing datasets and, in some circumstances, to impute missing groundwater level data. Given the rapid variations in groundwater levels seen in Vhembe District, effective data analysis necessitates treating each section of the series separately (Ruybal *et al.,* 2019; Evans *et al*., 2020). To remove and/or impute missing gaps in the data, this study utilised the Cubic spline approach, which is an interpolation method.

Figure 3.1 illustrates the distribution of all the stations used in the study across Vhembe District, covering the entire region except for some hilly and forested areas in the southwest. Seven groundwater stations (Table 3.1) were selected for the study with three weather stations and two evaporation stations. The groundwater level data provided by the DWS had values measured at different time intervals throughout the day and the SAWS weather stations data were on a daily time series pattern. The groundwater levels were converted into average daily time series to match the SAWS data. To achieve this, the AVERAGEIFS function in excel and Pandas in python for verification purposes and accuracy were used. While the selected rainfall, max and min temperature, windspeed, humidity and evaporation data in different timeseries methods could differ temporally, this study also considered their spatial average values. Considering spatial averages reduces the effect of extreme values or outliers that may occur at specific locations and helps create a more representative picture of the overall behaviour of the different variables. To address this, the study used rigorous processing techniques such as differencing to assure compatibility with datasets with varying temporal resolutions.

Missing data

Groundwater level datasets often have gaps due to sensor errors, measurement delays, or missing records (DWS, 2024). The exogenous variables (precipitation, windspeed, max and min temperature, evaporation and humidity) also have missing values, and the SARIMAX model may struggle because it requires a complete dataset for proper forecasting. The intepolate() function through the interpolation method in pandas helped with dealing with these missing values by filling them with predefined values, improving data continuity.

© University of Venda

Interpolation estimated missing values based on a series of data values, making it ideal for maintaining natural fluctuations in groundwater levels and exogenous variables for this study. Unlike simple imputation methods (for example, mean filling), interpolation ensured that the time series dataset remained continuous and aligned with real-world seasonal patterns, which helped with general accuracy in forecasting groundwater levels. The SARIMAX model relies on smooth transitions in exogenous variables. If gaps in precipitation or temperature data are filled with static values, it could create artificial jumps in predictions Shah *et al*., 2024). Interpolation smooths out these transitions, allowing for better long-term forecasting stability.

Ruman *et al.* (2020) investigated how different infilling methods and the length of missing precipitation data affected hydrological model performance, and offered some surprising insights into how missing data might not always be a liability. The study found that longer gaps in precipitation data did not necessarily degrade model accuracy. Also, the hypothesis that "longer gaps cause more error" was only valid in a small fraction of cases: 11.1% (sub-catchment) and 16.7% (catchment). This suggested that well-designed infilling can compensate for surprisingly large missing intervals.

Afrifa Yamoah *et al.* (2020) emphasized the importance of imputing missing values in high-resolution climate time series, such as hourly temperature, humidity, and wind speed data. The study tested three methods, Kalman smoothing with structural models, ARIMA with Kalman smoothing, and multiple linear regression, using artificially introduced gaps and cross-validation. Although all three methods produced plausible and low-error imputations, multiple linear regression generally performed best. Performance varied slightly across locations, but the findings supported that these techniques were suitable for filling gaps in high-frequency environmental data and could be applied across various disciplines where such data are common.

Lepot *et al.* (2017) highlighted both the breadth and the blind spots in the field of time-series gap-filling. While there are many interpolation techniques from classical methods to machine learning, their evaluation focused only on efficiency metrics like accuracy or error rates. The study emphasized the importance of quantifying this uncertainty and encouraged further research and even proposing different interpolation methods aimed at improving both accuracy and reliability in imputed data.

Anderson and Gough, (2018) on "Accounting for missing data in monthly temperature series: Testing rule-of-thumb omission of months with missing values" evaluated the accuracy and

© University of Venda

limitations of using rule-of-thumb thresholds, like the "3/5 rule" and a proposed "4/10 rule" to determine whether monthly climate data with missing values can still be considered reliable. The study quantified the statistical error introduced when certain numbers of daily values were missing and found that, while small errors occurred typically under 0.10 standard deviations these can increase with the number of missing values. Notably, the error was influenced by the autocorrelation in the data, and simple linear interpolation can effectively reduce error when missingness was random.

It is uncommon to find analyses of how much data infilling of precipitation input affects hydrological model results. Teegavarapu and Nayak (2017) study looked at the effects of filled precipitation records for varying gap durations. The results for 53 rain gauges in south Florida over the 1901–2006 period showed that data infilling may result in an underestimation of the frequency and magnitude of heavy and very heavy precipitation events, but it does not introduce statistically significant bias in total annual precipitation values.

Trend analysis

Trends can be found and measured using a variety of parametric, nonparametric, and hybrid techniques. Because of its ability to handle missing and tied data, serial dependence, non-linearity, non-normality, and seasonality, nonparametric techniques are frequently used in climatic and hydrological research (Helsel *et al*., 2020; Fahim *et al.,* 2024). In this study, the nonparametric modified Sen slope estimator was utilized to visualise the trends in the actual groundwater levels data. The Sen slope estimator was used to quantify variations over time per meter in groundwater levels, Sen's slope does not make any assumptions about the underlying distribution or model structure. It is a robust choice for trend analysis, especially when dealing with non-normal data or irregular patterns (*Ray et al.* 2021).

With the use of Pandas data profiling, an analysis of the correlation between the datasets was compared. Trends in levels of groundwater and meteorological data present a challenge to model development. This was accommodated by implementing stationarity tests like Augmented Dickey-Fuller (ADF), Phillips-Perron (PP) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) and differencing, which resampled all data to achieve stationarity. This approach simplifies modelling and ensures consistent statistical properties over time, ultimately facilitating accurate model development.

© University of Venda

Augmented Dickey-Fuller test (ADF)

ADF uses the unit root of the time series—a feature of non-stationary time series—as its null hypothesis. The series is either not stationary or has a unit root, according to the Null Hypothesis (HO). The series is stationary and has no unit root, according to the alternative hypothesis (HA). The p-value must be less than a predetermined significance level, which was established at 0.05, to reject the null hypothesis, which presupposes the existence of a unit root. The outcomes show that the series is stationary. One useful tool for implementing the ADF test in Python programming was the adfuller() function found in the Stats model's package. The p-value, test statistic value, number of lags in the test, and critical value thresholds for three degrees of significance were all provided in the report.

Kwiatkowski-Phillips-Schmidt-Shin test (KPSS)

To determine if the initial time series groundwater levels were stationary around a mean or linear trend, the KPSS test was employed. The kpss() function in Python programming implemented the KPSS test using the Statsmodels module. The test searches for evidence that the null hypothesis was incorrect. In this case, the null hypothesis was that the groundwater levels data were stationary. As a result, small p-values (less than 0.05) revealed that differencing was necessary and indicated that the null hypothesis was rejected.

Justification of data sets

The data used for this study were average daily data for a 12-year period. This was because of the issue of missing data on the stations which lead to a prioritization of stations with less gaps of missing data. To carry out this study and be accurate with the modelling and forecasting, a 12-year period of data with 4384 steps was chosen to be the best fit after comparisons with other studies which shared similar objectives. The model was trained during a ten-year period from 2007 to 2016 and validated over a two-year period from 2017 to 2019. The resulting dataset consists of 4384 daily time series steps at 7 groundwater stations, 2 weather stations and 2 evaporation stations. Each station was selected after being checked using data profiling in python programming. The selected stations had less than 30% of missing data. Time series models also have the benefit of being quick to construct and requiring little data.

As reported by Shin *et al.* (2020), the simulation performance in time series models was not significantly affected by the length of the training period; the difference in the Nash-Sutcliffe Efficiency (NSE) value between the 6-year and 19-year research periods was only 0.02.

© University of Venda

3.3.2 Data clustering

Spatial analysis technique was carried out utilizing all of the groundwater level time series data sets and weather stations in Quantum Geographical Information System (Qgis). Thiessen polygons, also known as Voronoi polygons, are a useful spatial analysis technique. They represent the area of influence around each point (such as rainfall stations) in a dataset. It is important to note that this study did not want to focus on grouping or clustering the data based on their similarities, hence methods such as K-Means clustering was not employed. The aim of drawing up Thiessens polygon was to see which groundwater station is more likely influenced by which weather station, these polygons are useful for various applications, including estimating rainfall distribution and creating statistical surfaces. This aided in achieving accurate trend analysis and prediction based on these exogenous variables from the weather station. Thiessen polygons are not typically used for data clustering. Their primary purpose is to delineate areas of influence around specific points (such as weather stations or monitoring sites). Each Thiessen polygon represents the region closest to a specific point, where that point has the greatest influence. The output was a set of polygons representing the catchment areas around each station. In this diagram, the two weather stations, two evaporation stations and seven groundwater level stations were grouped accordingly.

3.3.3 ARIMA model

The ARIMA model was created using the skforecast package, a Python programming library, and was founded on time series concepts, principles, and computational techniques. A new wrapper of statsmodels that adheres to the scikit-learn Application Programming Interface (API) is one of Skforecast's many forecasting features. Significant speed increases are achieved by simplifying this version to simply include the necessary components for Skforecast, which is quite like pyramid-Arima (Rodrigo and Ortiz, 2023).

The model procedure followed in this study is as follows:

Model Fit Statistics: A few statistics are provided in this section to assess how well the model matches the data:

- Log-Likelihood was used to gauge how well the model describes the observed data employed. Negative log-likelihood values were found when fitting the ARIMA model; values near zero indicate a better match to the data, whereas greater negative values indicated an unsatisfactory fit.

© University of Venda

- The Akaike Information Criterion (AIC), a goodness-of-fit statistic that weighs the model's complexity and fit, was employed. For the study, lower AIC values were preferable.
- The Bayesian Information Criterion (BIC) penalizes model complexity more than the AIC, but it was used similarly. Like AIC, lower BIC values were preferred.
- The Hannan-Quinn Information Criterion (HQIC), which was like AIC and BIC, was employed as an additional model selection criterion. To highlight model accuracy.

Coefficients: The estimated coefficients for the model's parameters were included in the model summary. Both MA and AR parameters were included. Additionally, the coefficients contained the 95% confidence interval, the P-values for each coefficient, and the standard errors for the estimated coefficients to show the degree of uncertainty in the parameter estimations.

Model diagnostics: Statistical information on the residuals, or the variations within the observed values (training values) and their simulated values from the model, was helpfully provided in this section:

- Ljung-Box test was used to examine the autocorrelation in the residuals.
- Jarque-Bera test was used to verify that the residuals were normal.
- Skewness and kurtosis were used to assess the residuals distribution shape.

After model fitting, the test data is used to make predictions. This process helped in evaluating how well the model performed on unseen data and to compare the simulated values to the actual observed values in the test set. This performance was evaluated with the use of graph and scatter plots of observed and simulated values. Common evaluation metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Nash-Sutcliffe Efficiency (NSE) and Coefficient of Determination (R-square) were also employed. The model was then utilised to forecast future groundwater levels to the year 2025 and the performance metrics were used again to assess the accuracy and performance of the model in forecasting future groundwater levels.

Autocorrelation Analysis

Finding the right values for p and q was made easier by displaying the time series' ACF and PACF. While the PACF helped determine the value of p (lag in the autoregressive component), the ACF helped determine the value of q (lag in the moving average section).

© University of Venda

It should be noted that while the stationarity study showed that differencing is necessary, the differenced series was used for additional analyses because it was consistent with how the ARIMA model understands the series.

Autocorrelation Function (ACF)

The correlation between a time series and its lag values was computed using the ACF. A steep decline in the ACF following a few delays in the context of ARIMA modelling signified that the data had a limited autoregressive order. The value of q was estimated from the lag at which the ACF drops off. A sinusoidal or damped sinusoidal pattern on the ACF indicated the presence of seasonality and necessitated considering both seasonal and non-seasonal orders.

Partial Autocorrelation Function (PACF)

In order to account for the impact of intermediate delays, the correlation between a lagged value and the current value of the groundwater level time series was measured using PACF plots. In ARIMA modelling, an AR model of such sort is recommended if the PACF abruptly stops after a specific latency and the remaining values are still within the confidence range. An estimate of the value of p was provided by the lag, which is the point at which the PACF stops.

Time series decomposition

Breaking down the groundwater level original time series into its fundamental components: trend, seasonality, and residual (error) components was done following a procedure known as time series decomposition (Rodrigo and Ortiz, 2023). Because the seasonal fluctuations of groundwater levels remained approximately constant over time, that is, their amplitude did not alter when the level of the time series changes, the decomposition was set to additive. ACF and PACF analysis in conjunction with time series decomposition gave this study a thorough method for figuring out the data's underlying structure and selecting the right ARIMA parameters.

3.3.4 SARIMAX model

The implementation of SARIMAX provided a useful feature within the statsmodels library in Python programming, which can incorporate exogenous variables such as precipitation, maximum and minimum temperatures, wind speed, humidity, and evaporation as forecasting factors in addition to the groundwater levels time series. Knowing the variable's value

© University of Venda

throughout the forecast period was the sole prerequisite for adding an exogenous variable. Using the 'exog argument', exogenous variables were added.

Using the already trained ARIMA model, its parameters were used to initialize the SARIMAX model. When defining the SARIMAX model, the exogenous variables, which are Precipitation, max and min temperature, windspeed, humidity and evaporation. This allowed the model to account for external factors that might influence the time series. The SARIMAX model was carried out the same way the ARIMA model was conducted, that is:

- The SARIMAX model was fit to the data, including the exogenous variables.
- The fitted SARIMAX model was used to make predictions of the groundwater levels, again including the exogenous variables.
- The order and seasonal order parameters in SARIMAX were like those in ARIMA but included seasonal components.

The SARIMA model is denoted as ARIMA (p, d, q) (P, D, Q, s), where: (p, d, q) are the non-seasonal parameters. (P, D, Q) are the seasonal parameters. (s) is the period or the length of the seasonal cycle. In this study, when using SARIMAX for forecasting groundwater levels with daily average data, the period was set to 12 typically to imply a monthly seasonality because of the following three reasons:

1. Seasonal Pattern: Groundwater levels often exhibit seasonal patterns due to factors like precipitation, temperature, and wind speed, which can vary monthly. By setting the period to 12, the study captured this monthly cycle of exogeneous variables.

2. Monthly Aggregation: Even though the data was daily, the seasonal component this study was modelling repeats every 12 months. This helps in identifying and forecasting the recurring monthly patterns in the daily data.

3. Modelling Efficiency: Using a period of 12 simplifies the model by focusing on the most significant seasonal cycle, which is often monthly for environmental data like groundwater levels.

3.3.5 Model's performance assessment

The efficiency of the created models was assessed using a set of standard statistical metrics. Their performance was evaluated using these metrics, which included the Root Mean Square Error (RMSE), Mean Square Error (MSE), Mean Absolute Error (MAE), Coefficient of

© University of Venda

Determination (R-square) and Nash-Sutcliffe Efficiency (NSE). For that, each approach used 16.67% of the data for validation and 83.31% of the data for training.

These indicators helped in understanding how well the models in this study were performing. Lower values indicated better performance. MAE denoted by Equation 3.1 provided a straightforward interpretation as it represented the average magnitude of errors in the same units as the groundwater levels. MSE shown by Equation 3.2 measured the average of the squared differences between the predicted groundwater levels and the actual values. RMSE (Equation 3.3) was used to provide a measure of the average magnitude of the errors in the same units as the original groundwater levels data (Singh, 2022).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad 3.1$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad 3.2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad 3.3$$

Where y_i is the actual observed values of groundwater levels, $\hat{y}_i$ is the simulated values, and n is the number of observations.

In groundwater level time series modelling, the acceptable values for performance metrics like RMSE, MSE, MAE, NSE, and R-square can vary depending on the specific context and the range of observed data (Bobbitt, 2021). Table 3.2 shows the criteria that was followed to evaluate the performance of the models as demonstrated by (Makungo and Odiyo, 2017) and from the HEC-HMS technical reference manual.

This study also showed graph plots of observed and simulated groundwater levels to test the performance of the developed ARIMA and SARIMAX models. To evaluate the performance between observed and simulated data, Nash-Sutcliffe Efficiency (NSE) and Coefficient of Determination (R-square) were employed. NSE evaluated how well the plot of observed versus simulated data fits the 1:1 line (Equation 3.4). A one-to-one correspondence between the simulated and observed values is indicated by an NSE of 1. R-Square (Equation 3.5) was used to describe the degree of collinearity between simulated and observed data.

$$NSE = 1 - \left[\frac{\sum_{i=1}^{n}\left(Y_i^{obs} - Y_i^{sim}\right)^2}{\sum_{i=1}^{n}\left(Y_i^{obs} - Y_{obs}\right)^2}\right] \quad 3.4$$

© University of Venda

$$R^2 = \left[\frac{\sum_i^n \left(Y_i^{obs} - Y_{obs}\right)\left(Y_i^{sim} - Y_{sim}\right)}{\sqrt{\sum_i^n \left(Y_i^{obs} - Y_{obs}\right)^2} * \sqrt{\sum_i^n \left(Y_i^{sim} - Y_{sim}\right)^2}}\right]^2 \quad 3.5$$

Where $Y^{obs} = i^{th}$observation, $Y_i^{sim} = i^{th}$ simulated value, Y_{obs} = the mean of observed data, Y_{sim} = the mean simulated data and n = total number of observations.

Table 3.2 Model performance evaluation criteria.

Performance metrics	RMSE	MSE	MAE	NSE	R^2
Very good	RSME <0.2	MSE <0.04	MAE <0.1	0.75<NSE≤1.00	$R^2 \geq 0.85$
Good	0.2<RMSE≤0.5	0.04<MSE≤0.25	0.1<MAE≤0.25	0.65<NSE≤0.75	0.70≤R^2<0.85
Satisfactory	0.5<RMSE≤1.0	0.25<MSE≤1.0	0.25<MAE≤0.5	0.50<NSE≤0.65	0.5≤R^2<0.70
Unsatisfactory	RMSE >1	MSE >1	MAE >0.5	NSE≤0.50	R^2≤0.5

3.4 Reasons for using the selected models for the study

A key forecasting tool, the ARIMA model serves as the foundation for many core concepts in time-series research. Because they allow for rigorous statistical testing and offer a fundamental approach to model the impacts of dependency from the data series, the ARIMA models have proven particularly helpful in time series analysis (Hayes, 2024). According to certain studies, including (Pham *et al*., 2022 and Mohammed *et al.,* 2022) statistical ARIMA models are more accurate than other applied machine learning and artificial intelligence models for predicting groundwater levels. Even though they are quite difficult and need a lot of expertise, time-series models have been recognized as extremely flexible and powerful methods because they are more likely to handle non-stationary time series. Additionally, these models may be used to anticipate groundwater levels in the short term without requiring additional input data. In a region with limited hydrological data availability, this may be a useful feature.

The SARIMA model takes over when long-term historical data reveals seasonal changes. To show that ARIMA can capture patterns that recur at regular intervals, the word "seasonal" is added to the model (Melanie, 2024). Temperature in a semi-arid region is an example of a short-term seasonal fluctuation, whereas meteorological data shows longer-term seasonal variations. Adding a seasonal component (S) to the SARIMA model allows it to better foresee these recurring trends. Even though the SARIMA model provides a strong approach to

© University of Venda

modelling seasonal time series, these data may be influenced by other variables. An even more thorough study is made possible by the SARIMAX model. Exogenous variables, sometimes referred to as covariates, are outside factors that may have an impact on the time series data on groundwater levels. In the context of this study, covariates include Precipitation, Min and Max temperature, wind speed, evaporation and humidity. These covariates can be included in the analysis using the SARIMAX model, which allows for outside variables that could influence the trends noted (Melanie, 2024). Both parametric and nonparametric trend tests are among the many trend analysis techniques that scholars have employed throughout the years.

The most used nonparametric techniques for identifying groundwater level changes are Sen's slope methods (Zakwan, 2021). Outliers have less of an impact on Sen's slope. This makes it particularly useful in datasets where outliers might skew the results (Wilcox, 2001). In this study, this method was used to calculate the median of all possible slopes between data points to provide a straightforward and easily interpretable measure of the trend in all the boreholes.

This study employed ARIMA and its seasonal extensions over neural networks and machine learning methods such as ANN and XGBoost due to some key advantages. ARIMA models naturally captured autocorrelation and seasonality, making them ideal for hydroclimatic data with strong temporal dynamics. SARIMAX enabled the incorporation of critical exogenous variables like precipitation and evaporation, aligning with known physical influences on groundwater. The transparency and interpretability of these models supported clearer environmental decision making, an essential criterion in regions where stakeholder trust and reproducibility are paramount. In contrast, ANN and XGBoost, while powerful, function as black-box models requiring extensive tuning, substantial data, and computational resources, resources not readily available in many semi-arid hydrological contexts.

Wang *et al.* (2021) conducted a statistical comparison of SARIMA and ANN models for predicting surface water quality, revealing that although ANN models exhibited superior generalization in certain instances, SARIMA provided more clear and interpretable outcomes.

© University of Venda

Chapter 4: RESULTS AND DISCUSSION

4.1 Preamble

This chapter presents the results of the entire study on groundwater level modelling using time series models in a semi-arid area. Seven groundwater stations, two weather stations and two evaporation stations as show in Table 3.1, were used to model and forecast groundwater levels in Vhembe district.

4.2 Data Processing and stationarity tests

4.2.1 Data clustering and stationarity tests results

The Thiessen's polygon shows that from the seven borehole stations used in this study, six of them (A9N001, A9N0002, A9N0003, A9N0005, A9N0010, and A9N0011) were in the same cluster and therefore share the same weather station (Punda maria) which all belong to the same Luvuvhu catchment. For this cluster, evaporation station A8E002 was adopted because there is a lack of evaporation data on this cluster. The variations in evaporation within this cluster and location of A8E002 is expected to be similar. The second cluster has one borehole station (A9N0014), one weather station (Mukumbani Tea) and one evaporation station (A9E004). The two cluster were used to carry out this study and forecast groundwater levels.

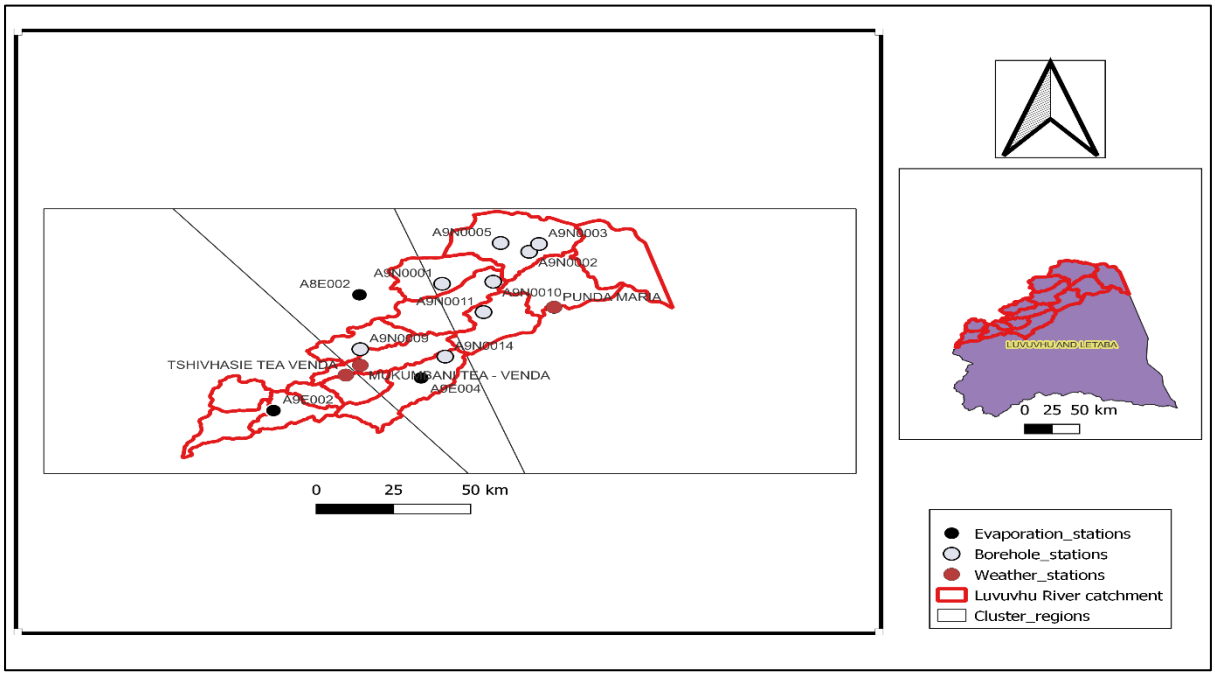


Figure 4.1: Map showing data clustering by Thiessen's polygon

© University of Venda

4.2.2 Results for stationarity tests, autocorrelations and partial autocorrelations functions

All the stations datasets were split into training and testing, with the training having 3653 steps and 731 steps are for testing the models. For these borehole stations, the groundwater levels experienced more variability in the earlier years and became relatively stable in the later years with random periods of sharp increases and decreases which could be possible due to seasonal changes, exogeneous variables, or other environmental factors.

Stationarity tests results in Table 4.1 provides results of stationarity tests for all the boreholes and their groundwater levels, using both ADF and KPSS tests at different orders of differencing, that is, order 1 and 2.

From Table 4.1, ADF statistics less than the critical value (-3.96) and p-values less than 0.05 indicated stationarity on the original observed groundwater levels. KPSS statistics less than the critical value (0.1) and p-values greater than 0.05 indicated stationarity for the same observed data. Most stations show low p-values (< 0.05), indicating stationarity for the ADF tests. The Lower p-values (close to 0) suggested rejecting the null hypothesis, meaning the original observed groundwater levels series was stationary. The KPSS tests on the original data provided higher p-values (> 0.05) suggesting that the test failed to reject the null hypothesis, meaning the series was stationary. The mixed results, with some p-values > 0.05 (stationary) and some < 0.05 (non-stationary). For some stations (A9N0003, A9N0005, A9N0014) KPSS statistics are higher, indicating non-stationary.

The differencing in order 1 appeared to stabilize the series as most ADF p-values were very low, and KPSS statistics were within acceptable bounds. ADF tests showed strongly significant p-values which indicated stationarity in order 1. The KPSS tests also showed low statistics and high p-values in order 1, confirming stationarity. Further differencing for order 2 confirmed stationarity across all borehole stations with very low p-value for the ADF tests and low statistics for KPSS. ADF indicated stationarities for all stations, but KPSS suggests non-stationarity for some. Order 1 differencing successfully achieved stationarity across all stations based on both ADF and KPSS. For this study, the differenced order 1 was preferred for most stations (6 out of 7) over the differenced order 2 to prevent over differencing and over fitting of the models which would have affected the accuracy in forecasting groundwater levels.

© University of Venda

Table 4.1 Stationary tests results for all the stations and their p-value for the first and second order of differencing the groundwater levels time series data.

	STATISTICS	BOREHOLES STATIONS						
		A9N0001	A9N0002	A9N0003	A9N0005	A9N0010	A9N0011	A9N0014
Original	ADF	-4.289	-6.298	-4.285	-7.353	-4.009	-3.677	-5.024
	p-value	4.146e-05	3.469e-08	0.000	9.957e-11	0.001	0.004	1.988e-05
	KPSS	1.073	0.232	1.334	1.861	0.307	0.399	3.852
	p-value	0.01	0.1	0.01	0.01	0.1	0.078	0.01
Order 1	ADF	-13.433	-9.862	-10.334	-11.897	-36.608	-10.396	-7.487
	p-value	3.978e-05	4.196e-17	2.799e-18	5.699e-22	0.0	1.963e-18	4.601e-11
	KPSS	0.014	0.007	0.015	0.005	0.012	0.009	0.008
	p-value	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Order 2	ADF	-16.867	-17.711	-20.078	-16.025	-17.803	-18.737	-18.757
	p-value	1.088e-29	3.501e-30	0.0	6.150e-29	3.214e-30	2.030e-30	2.027e-30
	KPSS	0.012	0.060	0.201	0.186	0.053	0.016	0.012
	p-value	0.1	0.1	0.1	0.1	0.1	0.1	0.1

© University of Venda

Figures 4.2a and 4.2b shows the differencing orders plots for all the groundwater levels borehole stations used in this study. These plots illustrate the stationarity of groundwater level time series data at different stages of differencing. The first plot labelled 'original time series' shows the raw training data from 2007 to 2016. In line with the inconsistent findings from the ADF and KPSS tests, the plots for these boreholes indicated that the original series may have some trends or non-stationary behaviour. The second plot labelled "Differenced order 1" showed the first-order differenced series which had more variabilities and spikes that indicated that differencing has removed some trends or seasonality. This meant that the first differencing has stabilized the mean of the series, making it more stationary. This was supported by the ADF and KPSS teste results in table 4.1 which showed stationarity after first differencing. The third plots labelled "Differenced order 2" showed more pronounced spikes than the first-order differencing, which indicated that further differencing has been applied.

Following an examination of the first and second-order differencing plots, the p-value for boreholes A9N0001, A9N0002, A9N003, A9N0005, A9N0010, A9N0011, and A9N0014 showed a statistically significant decline below the generally acknowledged and acceptable threshold of 0.05 for order=1. The ADF and KPSS tests confirmed that the series was stationary at this stage, and it might not be necessary to go beyond the first differencing for all the stations besides borehole station A9N0001 which had differencing plots showing more stable stationarity for order=2. Therefore, for A9N0002, A9N003, A9N0005, A9N0010, A9N0011, and A9N0014, the best choice for the ARIMA parameter d was 1.

© University of Venda

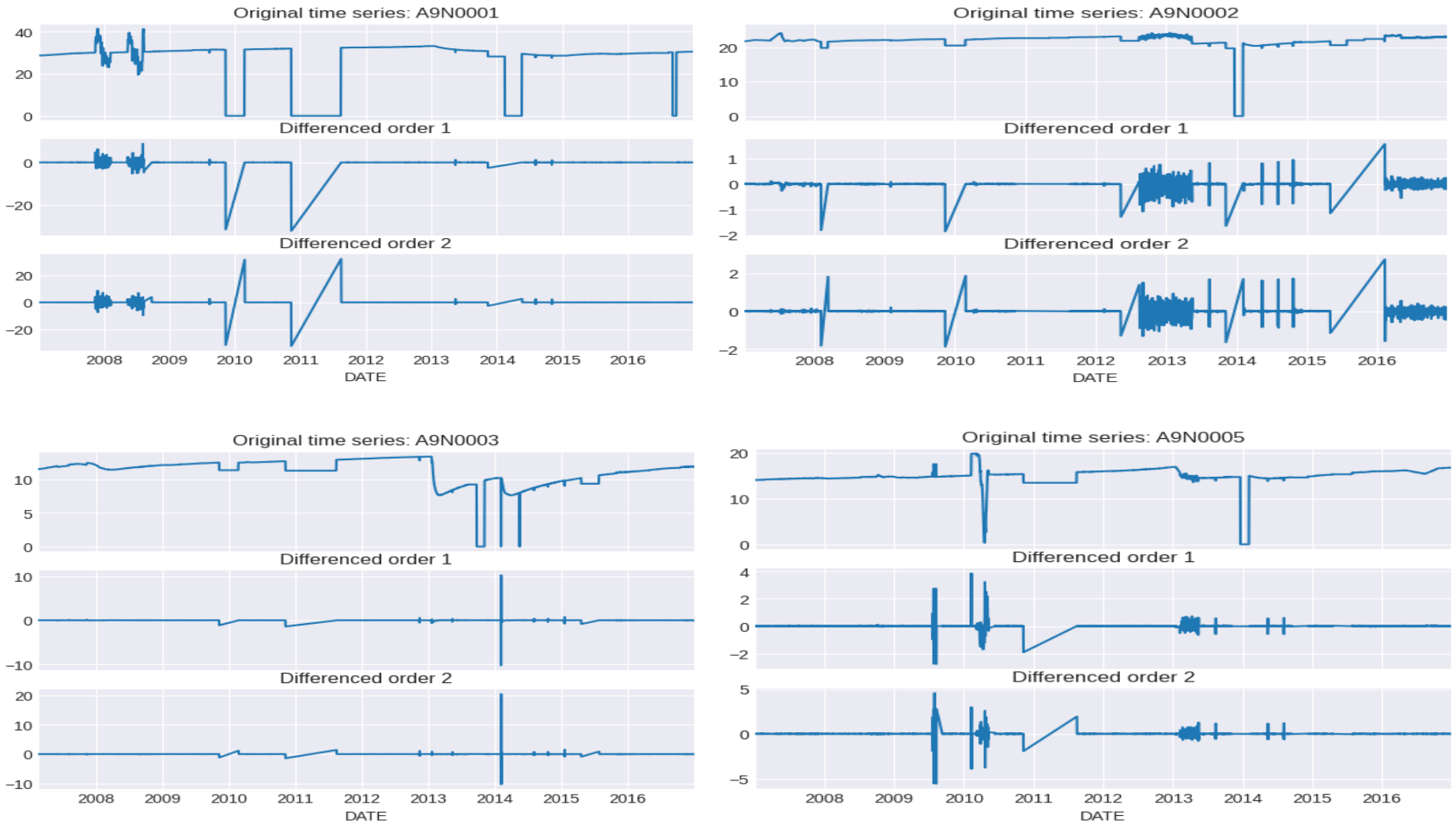


Figure 4.2a Differencing orders of groundwater levels in borehole stations A9N0001, A9N0002, A9N0003 and A9N0005.

© University of Venda

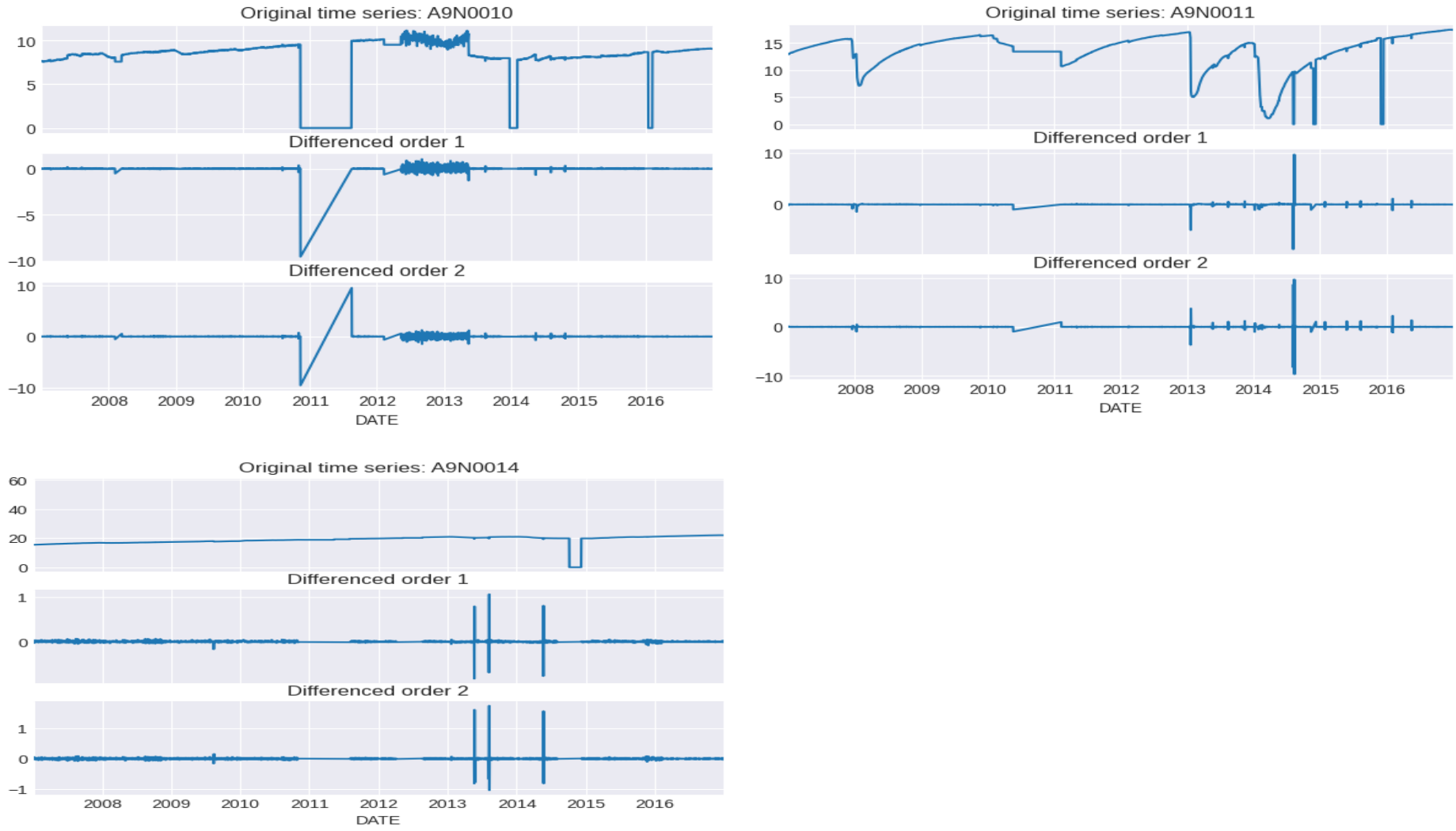


Figure 4.2b Differencing orders of groundwater levels in boreholes: A9N0010, A9N0011 and A9N0014.

© University of Venda

4.2.2 Trend analysis of groundwater levels

Table 4.2 provides the trend analysis results for groundwater levels at various stations using the Sen's slope estimator. Table 4.2 included interpretations of the Sen's slope statistics and the confidence intervals. Sen's slope estimator value represented the estimated trend in groundwater levels at each station. A positive value indicated an increasing trend, while a negative value indicated a decreasing trend. The confidence interval range provided an estimate of the precision of the Sen's slope estimator. A narrow confidence interval suggested high precision, while a wider interval indicated a more variability in the trend estimate.

The trend analysis results for groundwater levels using Sen's slope estimator and their respective confidence intervals provided valuable insights into the direction and significance of groundwater level changes at each borehole station.

Most boreholes including A9N0001, A9N0002, A9N0005, A9N0010, A9N0011 and A9N0014 exhibited a positive trend in groundwater levels from 2007 to 2018, which indicated a decrease in groundwater quantity over time. The negative trend in station A9N0003 indicated a decrease in groundwater levels from 2007 to 2018, and the confidence interval did not include zero, which suggested that the trend was statistically significant. This meant that the quantity of groundwater level was steadily increasing over the period.

The width of the confidence intervals varied across borehole stations. Some stations, including A9N0005, had narrow confidence intervals, indicating high precision, while others, including A9N0011, showed wider intervals, indicating more variability in the trend estimated.

For every station, the findings of Sen's slope estimator test showed a noteworthy trend in the change in groundwater levels. This suggested that there may be regional variations in the general pattern of groundwater alterations. These findings, however, should not be taken to suggest that drastic changes are taking place. A station in the hilly northern area showed a negative Sen's slope for groundwater level change, suggesting that the groundwater resource may be supplied gradually. Additionally, stations in the area's western portion displayed a somewhat favourable trend. This was troubling because the area's natural systems and human activity both depend on groundwater. This decrease in the amount of groundwater suggests that the rates of recharge and extraction may not be aligned, with withdrawals potentially surpassing the natural replenishment. It's interesting to note that precipitation at this site varied, with an average of 600 to 700 mm per year and sporadic high rainfall events that might cause localized floods (SAWS, 2023). This result partly defies the widely held belief that more precipitation

© University of Venda

would always result in an increase in groundwater quantity (Park *et al.,* 2011). As a result, the ongoing decline in groundwater levels at these sites suggested the potential influence of other outside variables. For instance, the decline in groundwater may be caused by human activities such increasing groundwater pumping for urban water supply, industrial use, or irrigation (Wu *et al.,* 2020, Green *et al*., 2011). The ability of the region to recharge groundwater may be affected by changes in land cover and usage, as well as adjustments to the natural drainage patterns, even in the presence of high precipitation records. Quantity may decline as a result of this.

With a negative Sen's slope for the trend in groundwater levels, station A9N0003 in the western area provided a distinct picture. The downward trend in groundwater levels indicated that recharge and withdrawal processes were effectively balanced, pointing to an excess of groundwater. Curiously, yet, precipitation in these areas likewise fluctuates between 600 and 700 mm annually on average. This suggested that causes other than precipitation may also contribute to increased groundwater storage. The observed rise in amount might be due to a combination of geological reasons, reduced extraction, and/or changes in land use that are either increasing groundwater recharge or decreasing withdrawals.

The trend analysis using the Sen’s slope estimator provided valuable insights into the patterns of groundwater levels across various stations. The results highlighted the importance of continuous monitoring and the need for targeted water management strategies to address both increasing and decreasing trends in groundwater levels.

© University of Venda

Table 4.2 Trend analysis results of groundwater level using Sens slope estimator and their respective confidence interval.

Boreholes	**Sen's slope estimator**	**Confidence interval**	**Interpretation**
A9N0001	$1.32x10^{-4}$	[-0.0645, 0.0648]	Positive trend, but the confidence interval includes zero, indicating no statistically significant trend.
A9N0002	$1.78x10^{-4}$	[-0.0306, 0.0309]	Positive trend, but the confidence interval includes zero, indicating no statistically significant trend.
A9N0003	$-2.871x10^{-5}$	[0.0015, 0.0020]	Negative trend, but the confidence interval does not include zero, suggesting the trend is statistically significant.
A9N0005	0.001	[0.0013, 0.0014]	Positive trend with a narrow confidence interval, indicating high precision.
A9N0010	$6.494x10^{-5}$	[0.0007, 0.0008]	Positive trend with a narrow confidence interval, indicating high precision.
A9N0011	0.001	[0.0063, 0.0079]	Positive trend with a wide confidence interval, indicating significant variability in the trend estimate.
A9N0014	0.002	[0.0021, 0.0021]	Positive trend with a very narrow confidence interval, indicating high precision.

© University of Venda

4.2.3 Results for Autocorrelation, PACF and decomposition of groundwater levels

Figures 4.3a and b shows autocorrelation graphs for all the boreholes stations. The graphs have three plots for original series, first order differencing and second order differencing of the time series groundwater levels data which are useful for understanding the correlation of the groundwater level time series with its own past values at different lags. The first plot labelled "Autocorrelation original series," showed a series of dots connected by lines oscillating above and below the horizontal axis. The oscillations appeared periodic and consistent in amplitude. This pattern indicated a high level of autocorrelation within the original data set for all the stations. The periodic oscillations suggested that the series has a strong seasonal component or repeating patterns over time. This is typical for non-stationary series, where past values have a significant influence on future values. The second plot labelled "Autocorrelation differenced series (order=1)," showed prominent spikes at the beginning, followed by dots close to zero for the rest of the series. The prominent spike at lag 1 indicated that the first differencing has effectively removed the autocorrelation present in the original series. The dots close to zero for the rest of the lags suggested that the series is now stationary, with minimal autocorrelation. This was a good sign that the first differencing has successfully transformed the series into a stationary one, suitable for ARIMA modelling. The autocorrelation of the time series following second-order differencing is displayed in the third plot, "Autocorrelation differenced series (order=2)."The autocorrelation values are close to zero for all lags in most stations, indicating that the second-order differencing has effectively removed the correlation in the time series. This suggested that the series was now stationary, with no significant autocorrelation remaining.

According to Viren Rehal's (2022) analysis of ACF and PACF plots, an AR process was suggested by a gradually decreasing geometric ACF and a PACF that was meaningful for just a few lags. ACF was decreasing geometrically with delays in their plots. Two substantial delays in the PACF were followed by a decline in PACF values, which at this point became significant. They concluded that the series was an AR (2) process based on the two major PACF delays and the steadily declining ACF. This demonstrated that the number of significant lags of PACF determines the delays of AR. The ACF plots in their analysis exhibited two notable delays before lowering, while the PACF plots were falling geometrically. This suggested the MA (2) procedure. This suggested the MA (2) procedure. Additionally, it was mentioned that since MA delays are based on the number of significant ACF lags, if the ACF had one major lag, it would suggest an MA (1) process.

© University of Venda

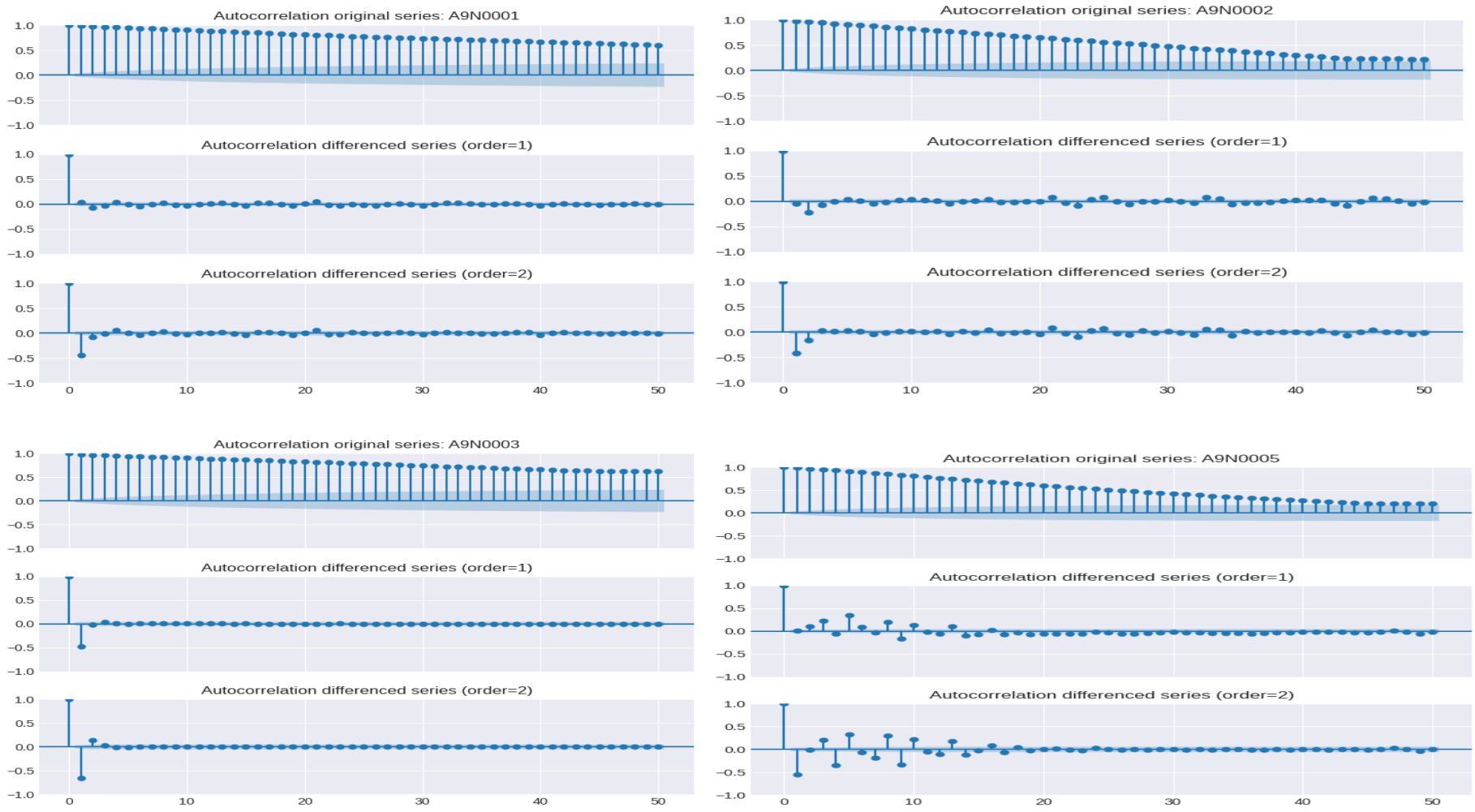


Figure 4.3a Autocorrelations plots for boreholes: A9N0001, A9N0002, A9N0003 and A9N0005.

© University of Venda



Figure 4.3b Autocorrelations plots for boreholes: A9N0010, A9N0011, and A9N0014.

© University of Venda

Figures 4.4a and b shows graph plots of Partial Autocorrelations function of all the stations in the study. The first plot labelled “Partial autocorrelation original series,” showed significant spikes at the first lag, followed by values hovering around zero for the rest of the series. The significant spike at lag 1 indicated a strong direct correlation between the groundwater level series and its first lag. The values around zero for subsequent lags suggested that there were no significant direct correlations beyond the first lag. This pattern was typical for a non-stationary series with a strong autoregressive component. The second plot labelled “Partial autocorrelation differenced series (order=1),” also showed a significant spike at the first lag, with values fluctuating slightly around zero for the rest of the series. The prominent spike at lag 1 indicated that the first differencing has effectively captured the direct correlation at the first lag. The values around zero for subsequent lags suggested that the series was now stationary, with minimal direct correlations beyond the first lag. This confirmed that first differencing has successfully transformed the series into a stationary one, suitable for ARIMA modelling.

Stationarity in time series data, as defined by Frost (2021) on ACF and PACF, is the absence of a trend, a constant variance, a constant autocorrelation pattern, and a seasonal pattern. For a stationary time series, the autocorrelation function quickly drops to almost zero. On the other hand, with a non-stationary time series, the ACF decreases gradually. For an autoregressive model, PACF was more helpful during the specification phase. To specify regression models with time series data and ARIMA models, analysts utilize partial autocorrelation plots (Frost, 2021).

© University of Venda

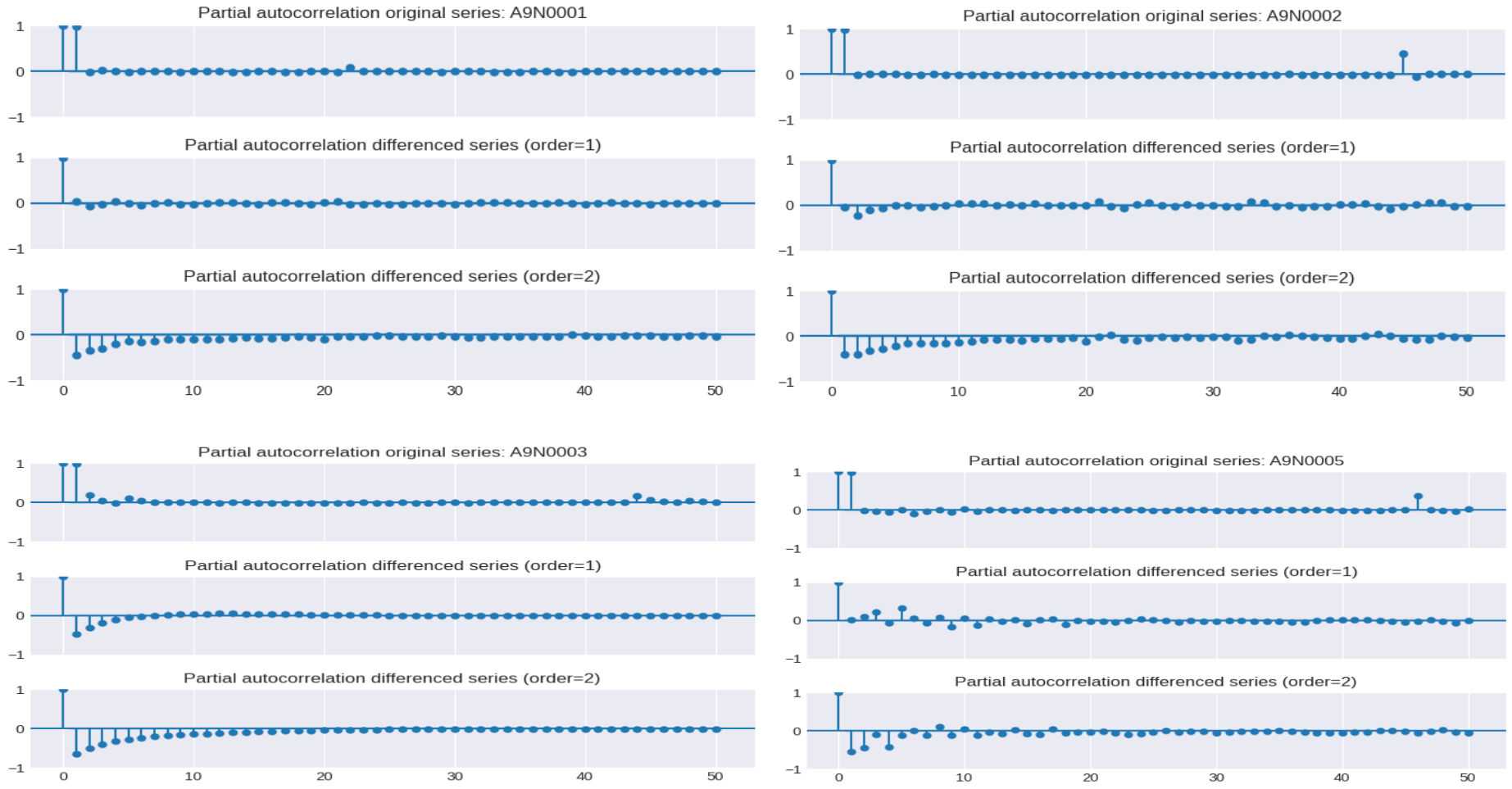


Figure 4.4a PACF plots for boreholes: A9N0001, A9N0002, A9N0003, and A9N0005.

© University of Venda

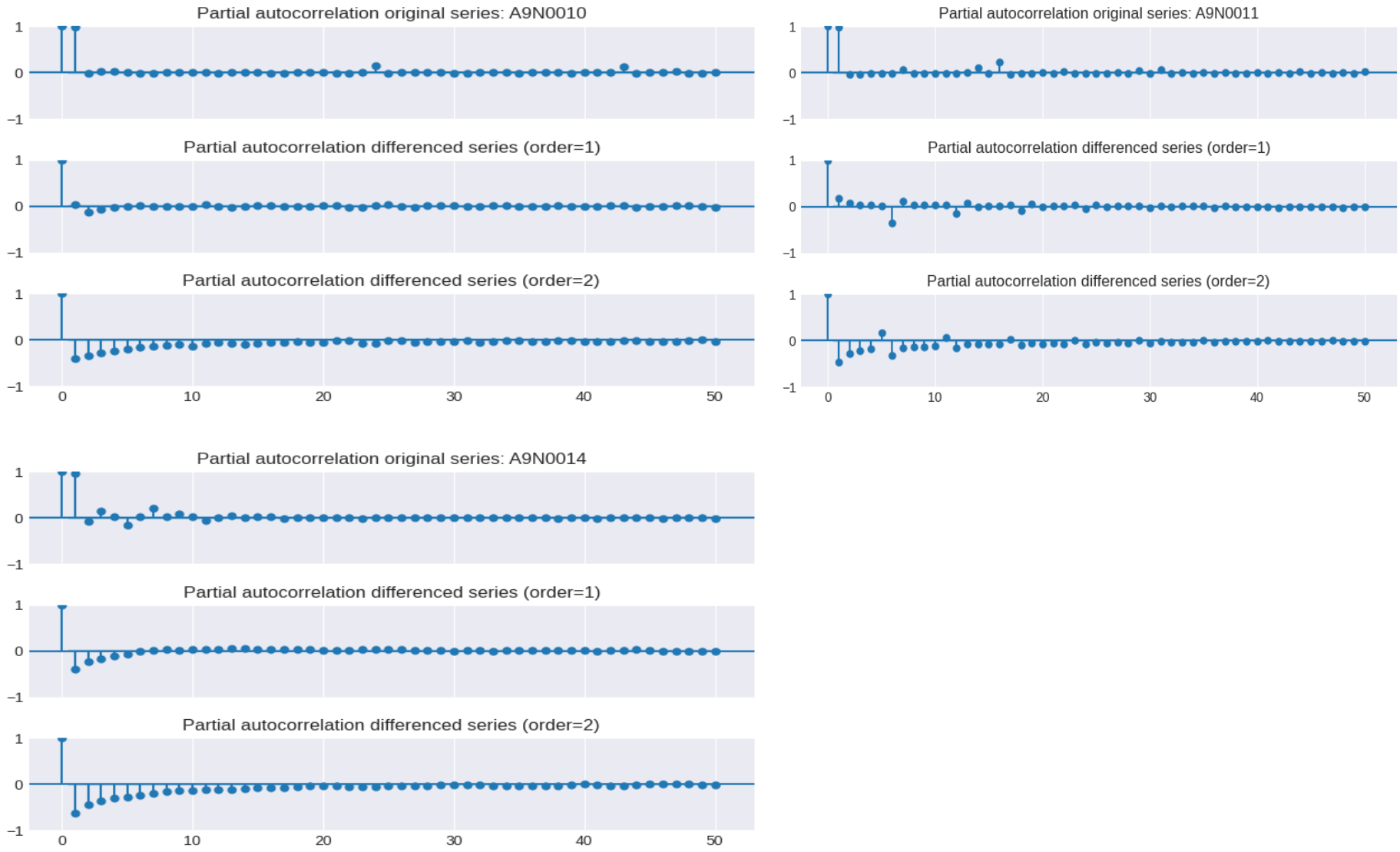


Figure 4.4b PACF plots for boreholes A9N0010, A9N0011, and A9N0014.

© University of Venda

Figures 4.5a and b present seasonal decomposition of groundwater level time series across seven boreholes within the Catchment, each dissected into original, differenced, seasonal, and residual components. The "Original series" displayed varied behaviours from pronounced cyclical fluctuations at stations like A9N0001 and A9N0010 to a steady upward trend at A9N0014. First-order differencing rendered each series more stationary, as highlighted by sharper transitions and stabilized mean values in the "Differenced series" plots. Seasonal components retained strong annual oscillations across most boreholes, underscoring the influence of recurrent hydroclimatic drivers such as rainfall and evapotranspiration patterns. Notably, stations like A9N0002 and A9N0011 exhibited smooth and persistent cycles, while A9N0014 revealed minimal seasonality, which could suggest localized recharge isolation or confined aquifer dynamics. The residual plots, generally resembling white noise, suggest the ARIMA model has adequately captured the deterministic structure, though isolated spikes may reflect real world anomalies. These decomposition results validated the presence of 12-month seasonality confirmed by ACF analysis and reflect borehole specific groundwater behaviour that can inform zoned forecasting strategies.

The 12-monthly recurrent pattern points to a yearly seasonality that was probably impacted by hydrogeological variables. The ACF plot further confirms the presence of seasonality and justified the idea of recurring patterns because significant peaks showed up at lags that match the 12-month intervals.

Büyükşahin and Ertekin, (2019) explored hybrid models combining ARIMA with other techniques. They mentioned that the common practice of decomposing time series data into linear and nonlinear components to improve forecasting accuracy throughout the hybridization process are key factors in the forecasting performance of the methods.

© University of Venda

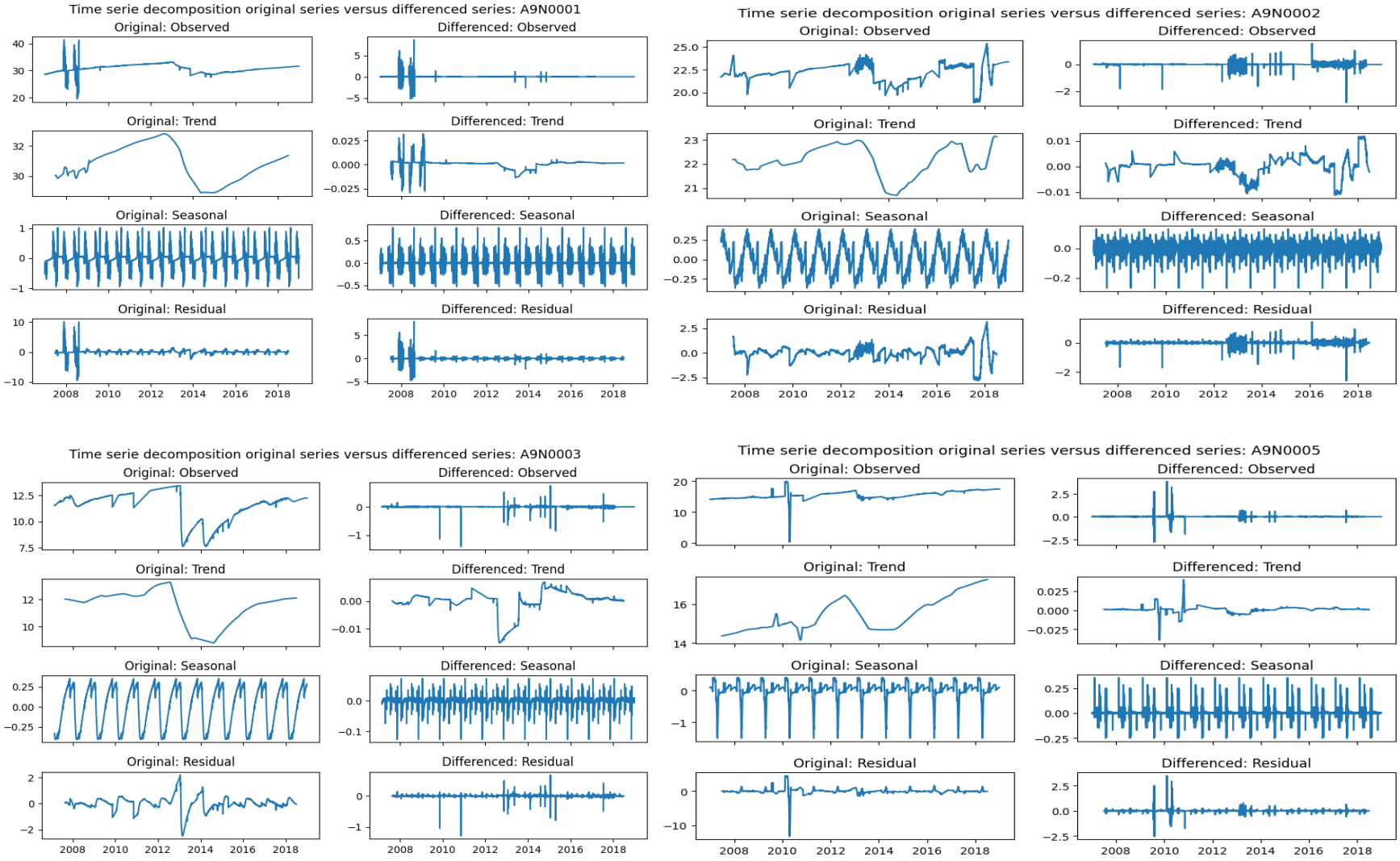


Figure 4.5a Time series decompositions plots for boreholes: A9N0001, A9N0002, A9N0003, and A9N0005.

© University of Venda

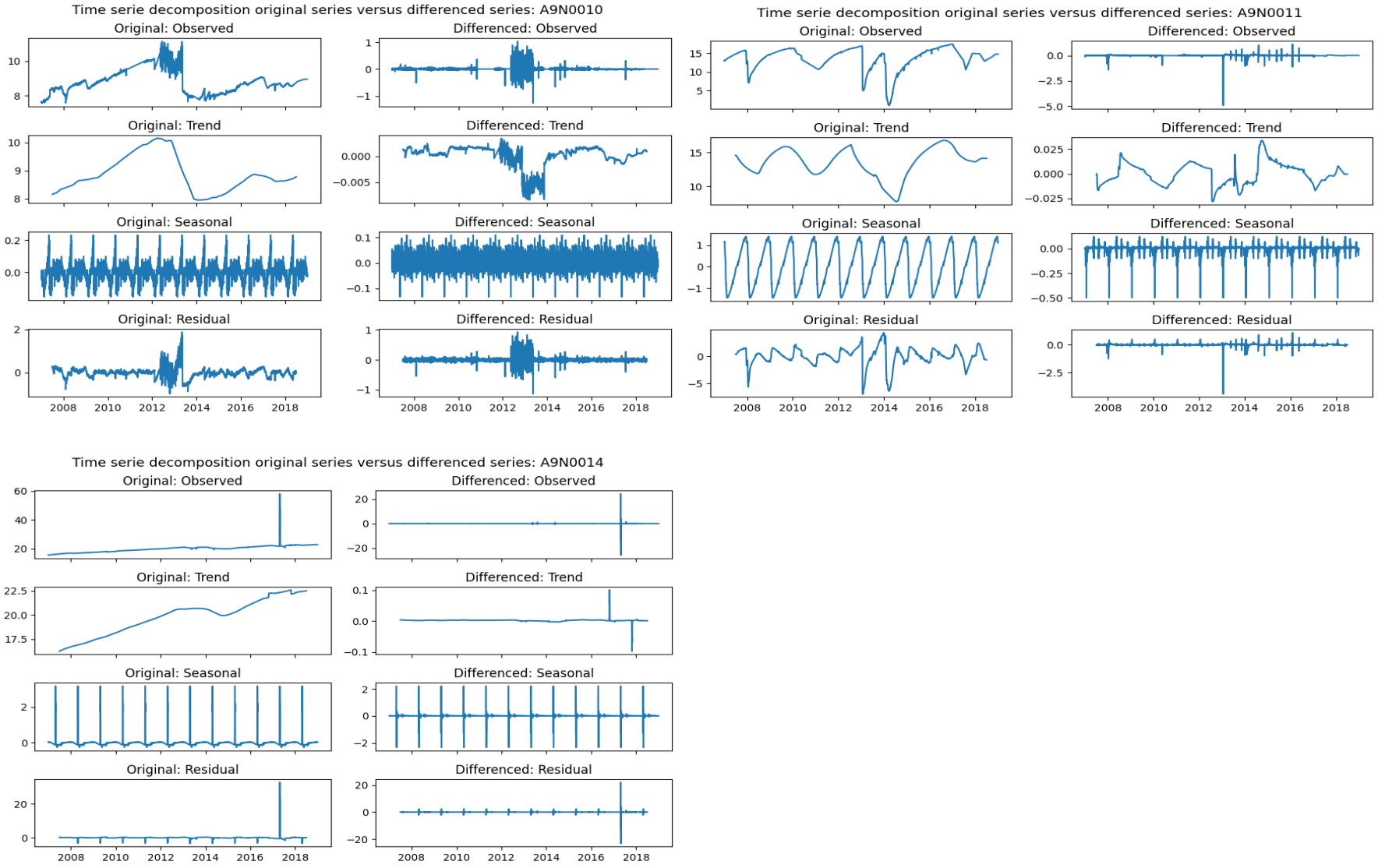


Figure 4.5b Time series decompositions plots for boreholes: A9N0010, A9N0011, and A9N0014.

© University of Venda

4.2.3 Groundwater levels model development, training and testing with ARIMA model

From Table 4.3 of the developed ARIMA model summaries, the ARIMA models for each borehole vary in terms of model parameters, coefficients, and diagnostic tests results. On the models' parameters, high AIC and BIC values suggested that while the model fits the data, it was not the most efficient compared to other boreholes, while low AIC and BIC values suggested a better model fit that is efficient. According to Kwarteng and Andreevich (2024), when it comes to the AIC and BIC, there are no universally "acceptable" values. Instead, these values are used to compare models.

The ARIMA model for Borehole A9N0001 showed significant coefficients, with both AR (1) and MA (1) being highly statistically significant (p-value = 0.000). The diagnostic tests indicate that the model fits the data well, with a reasonable Log Likelihood value. The ARIMA model for Borehole A9N0002 also demonstrated significant coefficients and favourable diagnostic test results. The Log Likelihood was relatively high, and the AIC was low, indicating a good model fit. The ARIMA model for Borehole A9N0003 had a high Log Likelihood and low AIC, suggesting a good fit. The coefficients are statistically significant, and the diagnostic tests indicate that the residuals are well-behaved. The ARIMA model for Borehole A9N0005 showed significant coefficients, and the diagnostic tests indicated a good model fit. The Log Likelihood was high, and the AIC was low. The ARIMA model for Borehole A9N0010 has significant coefficients, but the diagnostic tests reveal potential issues with the model fit. The ARIMA model for Borehole A9N0011 indicated significant coefficients and a relatively good fit based on the diagnostic tests. The ARIMA model for Borehole A9N0014 had several significant coefficients, indicating a strong relationship between the variables. Both the autoregressive AR (1) and AR (2) and moving average MA (1) and MA (2) components are highly statistically significant (p-value = 0.000), suggesting that the model captured the underlying patterns in the data effectively.

When comparing the ARIMA models results for the different boreholes based on their AIC and BIC values, Borehole A9N0014 stands out as having the best model fit, with the lowest AIC (-12233.318) and BIC (-12202.302) values. This indicated a highly efficient model with a good balance between fit and complexity. Boreholes A9N0003 and A9N0011 followed closely, with A9N0003 having an AIC of -8795.996 and BIC of -8783.590, and A9N0011 having an AIC of -4694.945 and BIC of -4676.336, making these models better in terms of fit compared to others. Boreholes A9N0002, A9N0010, and A9N0005 presented good models

© University of Venda

with relatively low AIC and BIC values: A9N0002 (AIC: -2558.493, BIC: -2539.884), A9N0010 (AIC: -868.451, BIC: -849.842), and A9N0005 (AIC: -707.592, BIC: -695.186). On the other hand, Borehole A9N0001 had comparatively higher AIC (4366.536) and BIC (4385.145) values, indicating a poorer fit among the group.

However, high values of Kurtosis and skewness across all boreholes suggest potential issues with the normality of residuals, which is a common challenge in time series modelling. In other studies, including (Assainar *et al.,* 2023, Gyamfi *et al.,* 2016, Mehrjardi *et al.,* 2008) researchers have often encountered similar challenges with normality and the presence of outliers. High skew and kurtosis in residuals suggested deviations from normality, implying that the models may not fully capture all data characteristics, and that the models may not have fully captured all temporal dependencies. Across all borehole models, significant non-normality and heteroskedasticity in residuals were observed. This suggests that while the ARIMA models capture some patterns in the data, they may not fully account for all underlying dynamics.

From these ARIMA model results, AIC and BIC, forecasting reliability for each model was expected to be:

- A9N0001: Moderately reliable predictions, but potential issues with extreme values due to high skew and kurtosis.
- A9N0002: High reliability. The model fits well with low AIC and BIC, and significant coefficients, though slight non-normality in residuals should be monitored.
- A9N0003: Less reliable due to significant residual autocorrelation and high non-normality. Predictions might have higher errors.
- A9N0005: Good reliability. Low AIC and BIC suggest a well-fitting model, but residual issues indicate some variability in predictions.
- A9N0010: Good reliability. Low AIC and BIC suggest accurate predictions, though residual heteroskedasticity may affect precision.
- A9N0011: Moderate reliability. Significant residual autocorrelation and non-normality suggest higher prediction errors.
- A9N0014: High reliability. Excellent fit based on AIC and BIC values, though monitoring for residual issues is important.

© University of Venda

Table 4.3 Summary of the developed ARIMA models for each borehole

Boreholes	Model	Coefficients	Diagnostic tests
A9N0001	ARIMA (1, 2, 1) • Log Likelihood: -2180.268 • AIC: 4366.536 • BIC: 4385.145 • HQIC: 4373.163	• ar.L1: 0.1525 (std err: 0.004, z: 37.906, p-value: 0.000) • ma.L1: -0.9999 (std err: 0.062, z: -16.129, p-value: 0.000) • sigma2: 0.2605 (std err: 0.016, z: 16.243, p-value: 0.000)	• Kurtosis: 80.40 • Ljung-Box (L1) (Q): 7.56 (Prob(Q): 0.01) • Skew: 1.92 • H: 0.02 (Prob(H) (two-sided): 0.00) • JB: 913643.06 (Prob (JB): 0.00)
A9N0002	ARIMA (1, 1, 1) • Log Likelihood:1282.247 • AIC: -2558.493 • BIC: -2539.884 • HQIC: -2551.866	• ar.L1: 0.9616 (std err: 0.003, z: 286.066, p-value: 0.000) • ma.L1: -0.8448 (std err: 0.006, z: -151.768, p-value: 0.000) • sigma2: 0.0227 (std err: 0.000, z: 165.329, p-value: 0.000)	• Kurtosis: 57.06 • Ljung-Box (L1) (Q): 1.99 (Prob(Q): 0.16) • Skew: -2.39 • H: 2.61 (Prob(H) (two-sided): 0.00) • JB: 448105.44 (Prob (JB): 0.00)
A9N0003	ARIMA (0, 1, 1) • Log Likelihood: 4399.998 • AIC: -8795.996 • BIC: -8783.590 • HQIC: -8791.578	• ma.L1: 0.0523 (std err: 0.004, z: 13.916, p-value: 0.000) • sigma2: 0.0031 (std err: 7.2e-06, z: 433.511, p-value: 0.000)	• Kurtosis: 309.79 • Ljung-Box (L1) (Q): 0.16 (Prob(Q): 0.69) • Skew: -11.65 • H: 2.26 (Prob(H) (two-sided): 0.00) • JB: 14404613.80 (Prob (JB): 0.00)
A9N0005	ARIMA (0, 1, 1) • Log Likelihood: 355.796 • AIC: -707.592 • BIC: -695.186 • HQIC: -703.174	• ma.L1: 0.0083 (std err: 0.003, z: 3.003, p-value: 0.003) • sigma2: 0.0460 (std err: 0.000, z: 306.735, p-value: 0.000)	• Kurtosis: 151.51 • Ljung-Box (L1) (Q): 0.03 (Prob(Q): 0.87) • Skew: 3.50 • H: 0.01 (Prob(H) (two-sided): 0.00) • JB: 3363653.38 (Prob (JB): 0.00)
A9N0010	ARIMA (1, 1, 1) • Log Likelihood: 437.226 • AIC: -868.451 • BIC: -849.842 • HQIC: -861.824	• ar.L1: -0.1909 (std err: 0.037, z: -5.112, p-value: 0.000) • ma.L1: 0.5943 (std err: 0.038, z: 15.601, p-value: 0.000) • sigma2: 0.0438 (std err: 4.53e-05, z: 966.557, p-value: 0.000)	• Kurtosis: 1609.19 • Ljung-Box (L1) (Q): 5.84 (Prob(Q): 0.02) • Skew: -32.49 • H: 8.39 (Prob(H) (two-sided): 0.00) • JB: 393206872.85 (Prob(JB): 0.00)

© University of Venda

University of Venda
Creating Future Leaders

Table 4.3 continued

Borehole	Model	Coefficients	Diagnostic tests
A9N0011	ARIMA (1, 1, 1) • Log Likelihood: 2350.473 • AIC: -4694.945 • BIC: -4676.336 • HQIC: -4688.318	• ar.L1: 0.4762 (std err: 0.004, z: 109.814, p-value: 0.000) • ma.L1: 0.1261 (std err: 0.005, z: 25.336, p-value: 0.000) • sigma2: 0.0142 (std err: 2.6e-05, z: 545.549, p-value: 0.000)	• Kurtosis: 440.67 • Ljung-Box (L1) (Q): 0.10 (Prob(Q): 0.75) • Skew: -12.14 • H: 6.48 (Prob(H) (two-sided): 0.00) • JB: 29237811.42 (Prob (JB): 0.00)
A9N0014	ARIMA (2, 1, 2) • Log Likelihood: 6121.659 • AIC: -12233.318 • BIC: -12202.302 • HQIC: -12222.273	• ar.L1: 1.1382 (std err: 0.007, z: 156.420, p-value: 0.000) • ar.L2: -0.1519 (std err: 0.007, z: -21.665, p-value: 0.000) • ma.L1: -1.7022 (std err: 0.008, z: -217.803, p-value: 0.000) • ma.L2: 0.7158 (std err: 0.008, z: 94.302, p-value: 0.000) • sigma2: 0.0011 (std err: 3.47e-06, z: 330.075, p-value: 0.000)	• Kurtosis: 372.28 • Ljung-Box (L1) (Q): 0.25 (Prob(Q): 0.62) • Skew: -8.57 • H: 1.99 (Prob(H) (two-sided): 0.00) • JB: 20795274.90 (Prob (JB): 0.00)

© University of Venda

Figures 4.6a and b show graphs for training, testing and forecasting groundwater levels using ARIMA models. Training and testing the model was crucial for assessing its ability to predict future groundwater levels for each borehole station. The ARIMA models for each borehole had significant coefficients (Table 4.2), which indicated that the models captured some underlying patterns in the data, such as trends and seasonality. The model's predictions in borehole A9N0001, A9N0005, A9N0011 and A9N0014 during the test period closely followed the actual test data, indicating that the ARIMA model has learned the underlying patterns in the data effectively. The forecasted values (green line) showed a smooth transition from the test data, suggesting that the ARIMA model has maintained the captured patterns into the future. The smooth continuation indicates a good fit, with the forecasted values aligning well with the observed trends. The ARIMA model has performed well in capturing the trends and seasonality in the data. The smooth transition from the test data to the forecasted data indicates that the model has effectively learned and projected the patterns. For borehole A9N0002 the ARIMA model used for predicting the GWL appeared to have limitations in capturing the variability seen in the test data. While the training data showed relatively stable groundwater levels values, the test data indicated more fluctuations, which the model's forecast does not adequately reflect. This discrepancy suggested that the model might benefit from further refinement or the incorporation of additional factors to improve its predictive accuracy. This was the same for boreholes A9N0003 and A9N0010. It could be because the ARIMA models are based on linear assumptions and may not capture the complex, non-linear relationships present in groundwater levels data (Pourmorad *et al*., 2024). This indicated the need to incorporate exogeneous variables to the model for improvement.

Figures 4.7a and b show graph plots of observed and simulated values of groundwater levels this was done to see periods/ points where there was overestimation, underestimation and good estimation of groundwater levels. The plots indicated a mix of overestimation, underestimation, and good estimation points for the GWL predictions using the ARIMA model. While the model provided a general trend, it struggled to capture the high variability and fluctuations seen in the observed data, leading to discrepancies in certain periods. The simulated values (red line) for borehole A9N0001, A9N0005 and A9N0011 showed an overall smooth and aligned well with the observed data. The model has maintained the captured patterns indicating an overall good estimation. Overestimation was observed around the initial part of the training data period in 2007 where the simulated values are above the actual observed groundwater levels. The ARIMA model has performed reasonably well in capturing the general trends in the data. The

© University of Venda

smooth tracking of the observed data by the simulated data indicated a good fit for a longer period (between 2008 and 2016) in the boreholes. The observed and simulated plots demonstrated the model's accuracy and reliability in forecasting groundwater levels values even in the presence of a lot of missing data. This was substantiated by the scatter plots in Figure 4.8a and b, which have shown a positive correlation between the observed and simulated data. As the observed groundwater levels values increase, the simulated groundwater levels values also tend to increase. The points were generally clustered around a line, suggesting a linear relationship between the observed and simulated data. There is some scatter around the line, indicating deviations from the perfect prediction. The scatter plots indicated that the ARIMA model had a reasonably good fit, with a strong positive correlation between observed and simulated groundwater values. The model captured the overall trends effectively but has some deviations, indicating areas for potential improvement. Rashid, (2022) was assessing and forecasting groundwater level fluctuations in Joypurhat District, Northwest Bangladesh, using wavelet analysis and ARIMA modelling and pointed out that consecutive records of groundwater levels were vital for determining groundwater potential and variation and trend. The study highlighted the limitations of ARIMA models in capturing long-term trends and variability, like the discrepancies observed in this studies plots.

© University of Venda

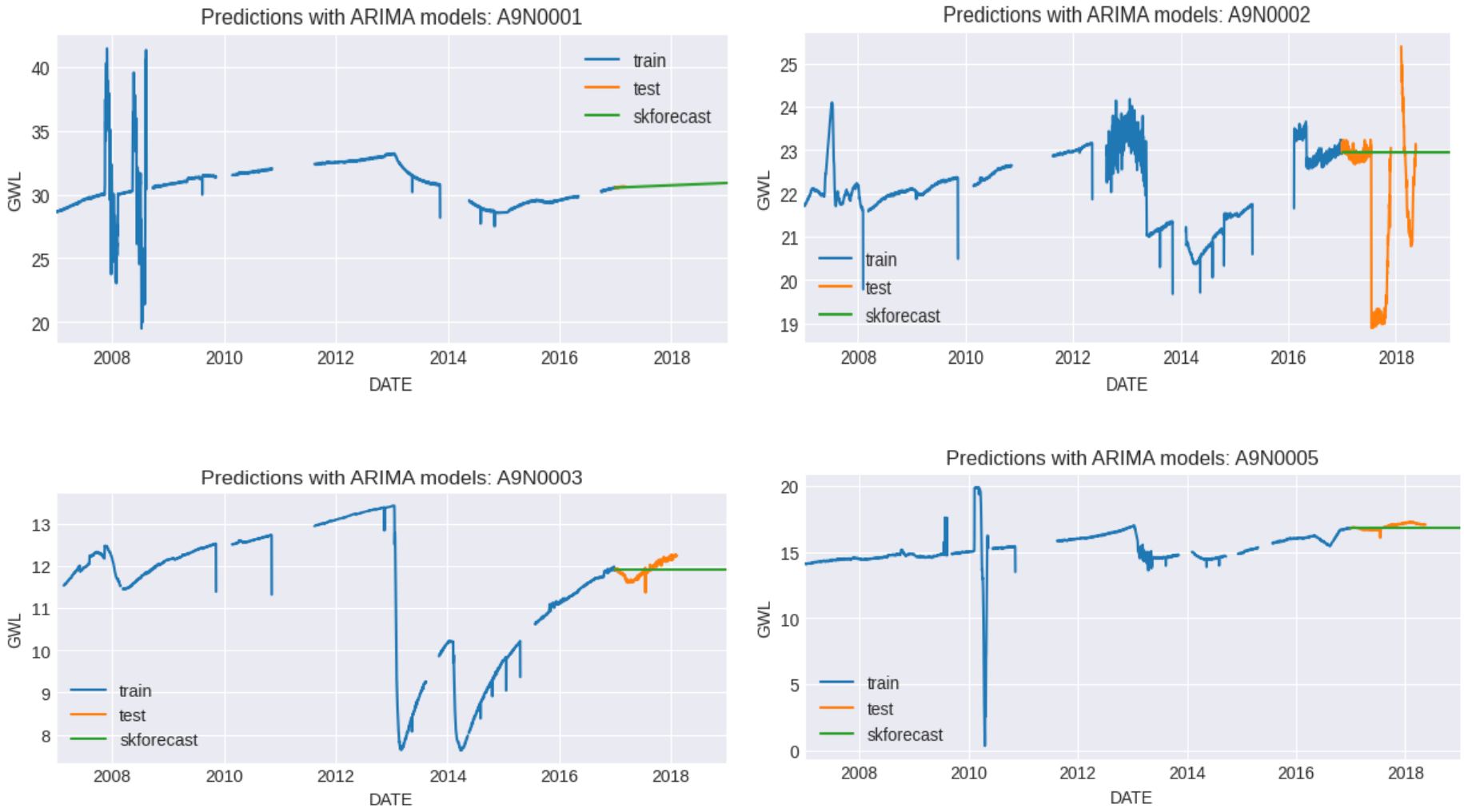


Figure 4.6a Groundwater levels prediction plots using ARIMA model for boreholes: A9N0001, A9N0002, A9N0003, and A9N0005.

© University of Venda

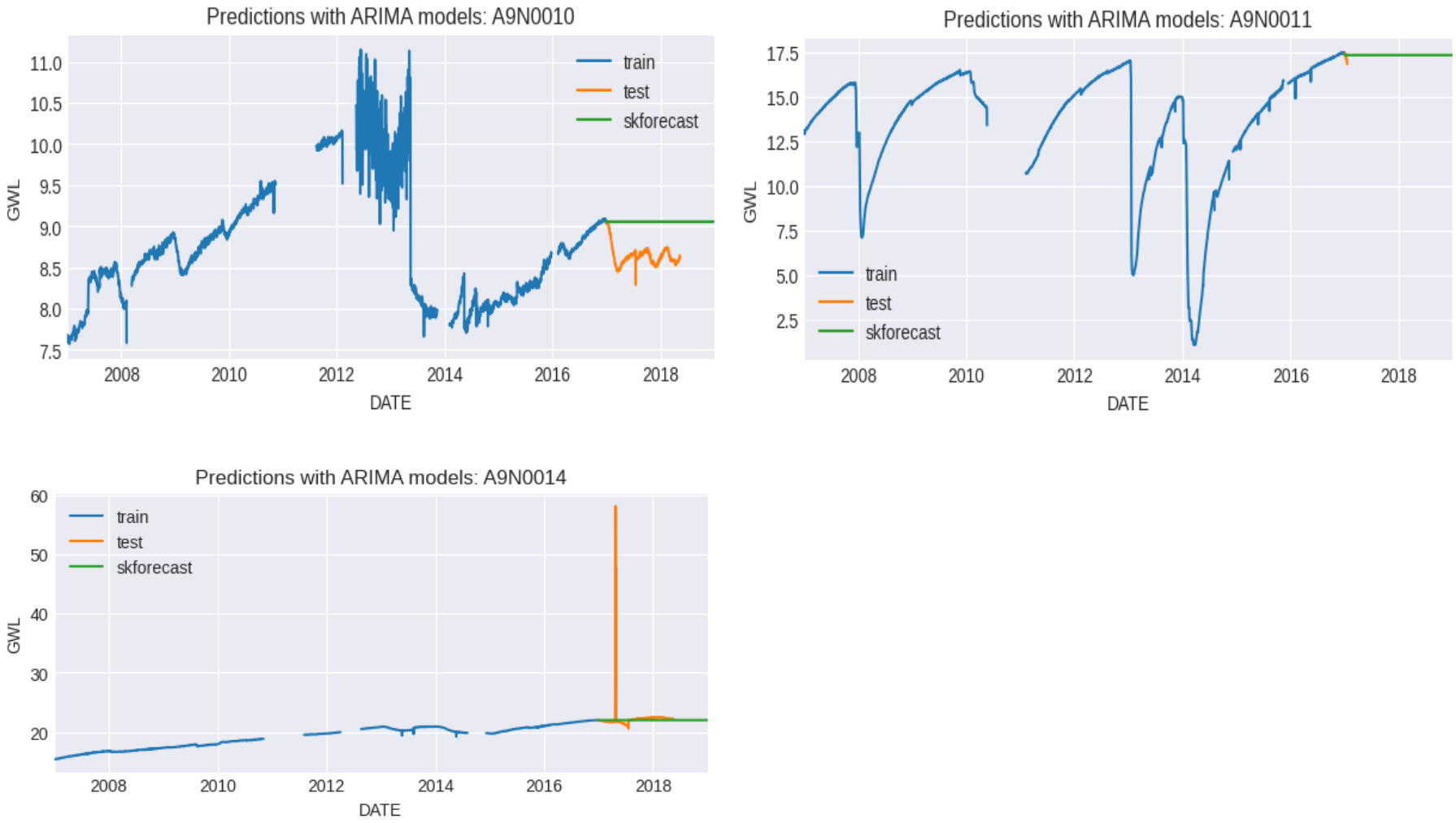


Figure 4.6b Groundwater levels prediction plots using ARIMA model for boreholes: A9N0010, A9N0011, and A9N0014.

© University of Venda

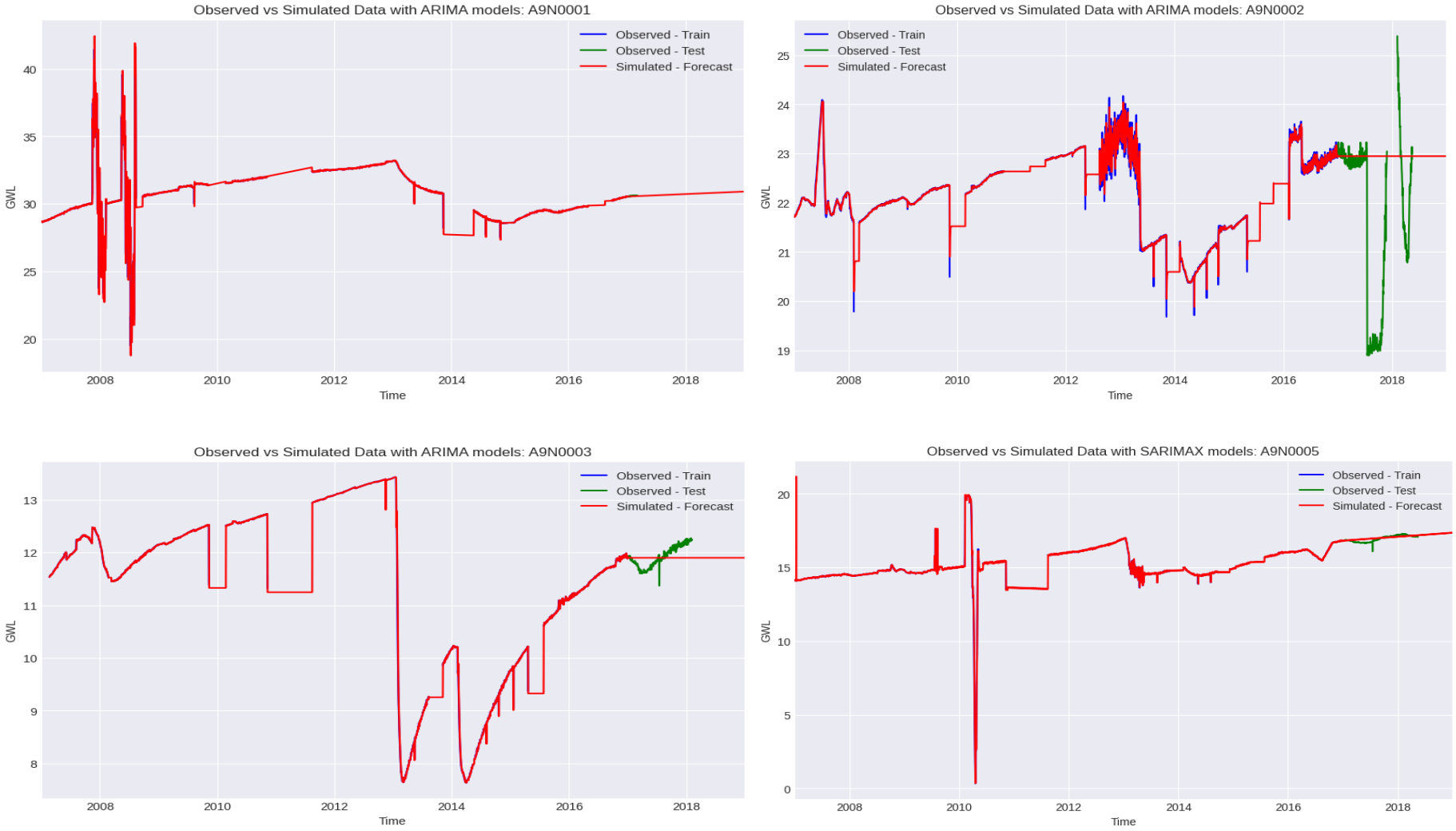
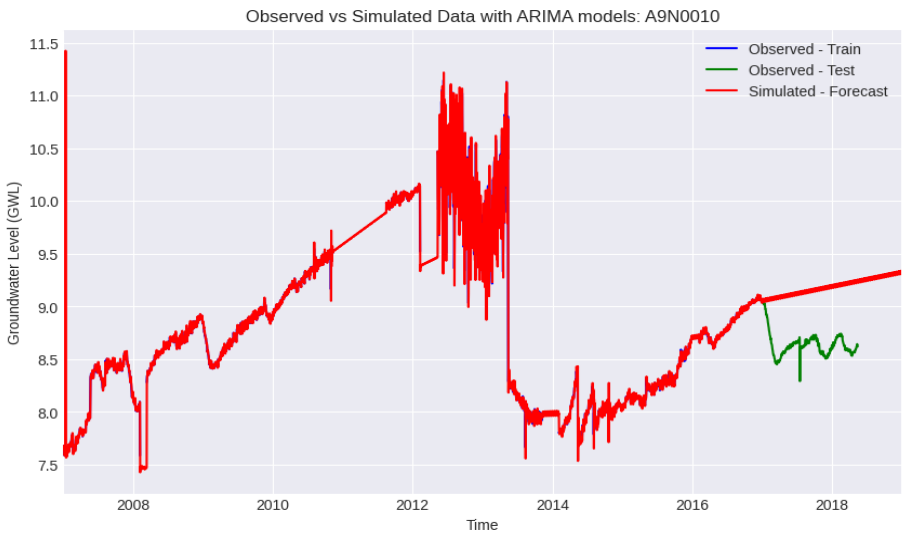
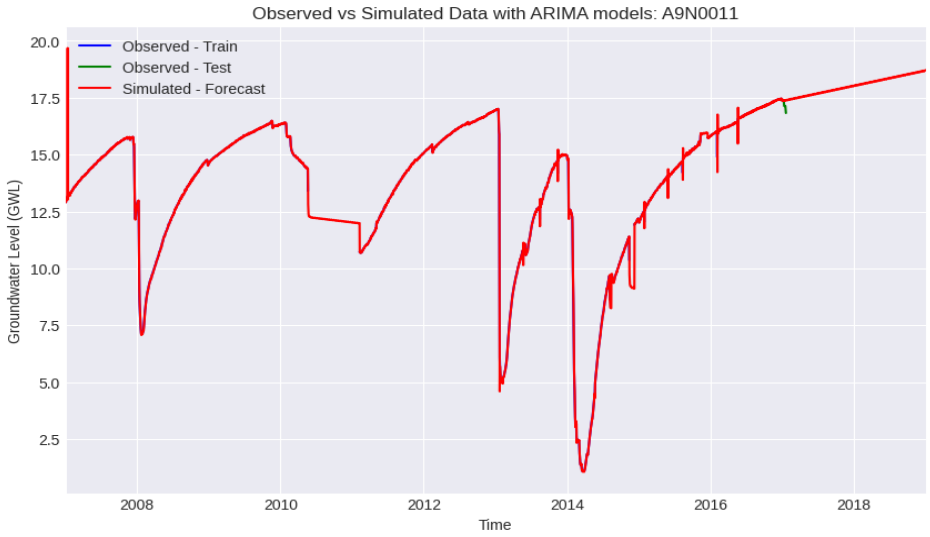


Figure 4.7a Graph plots of observed and simulated values of groundwater levels using ARIMA model for borehole A9N0001, A9N0002, A9N0003 and A9N0005.

© University of Venda





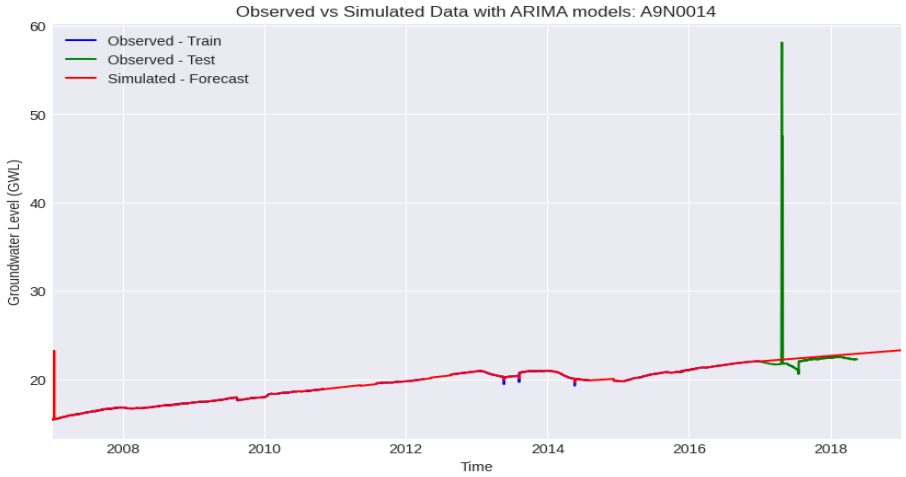


Figure 4.7b Graph plots of observed and simulated values of groundwater levels using ARIMA model for borehole A9N0010, A9N0011, and A9N0014.

© University of Venda

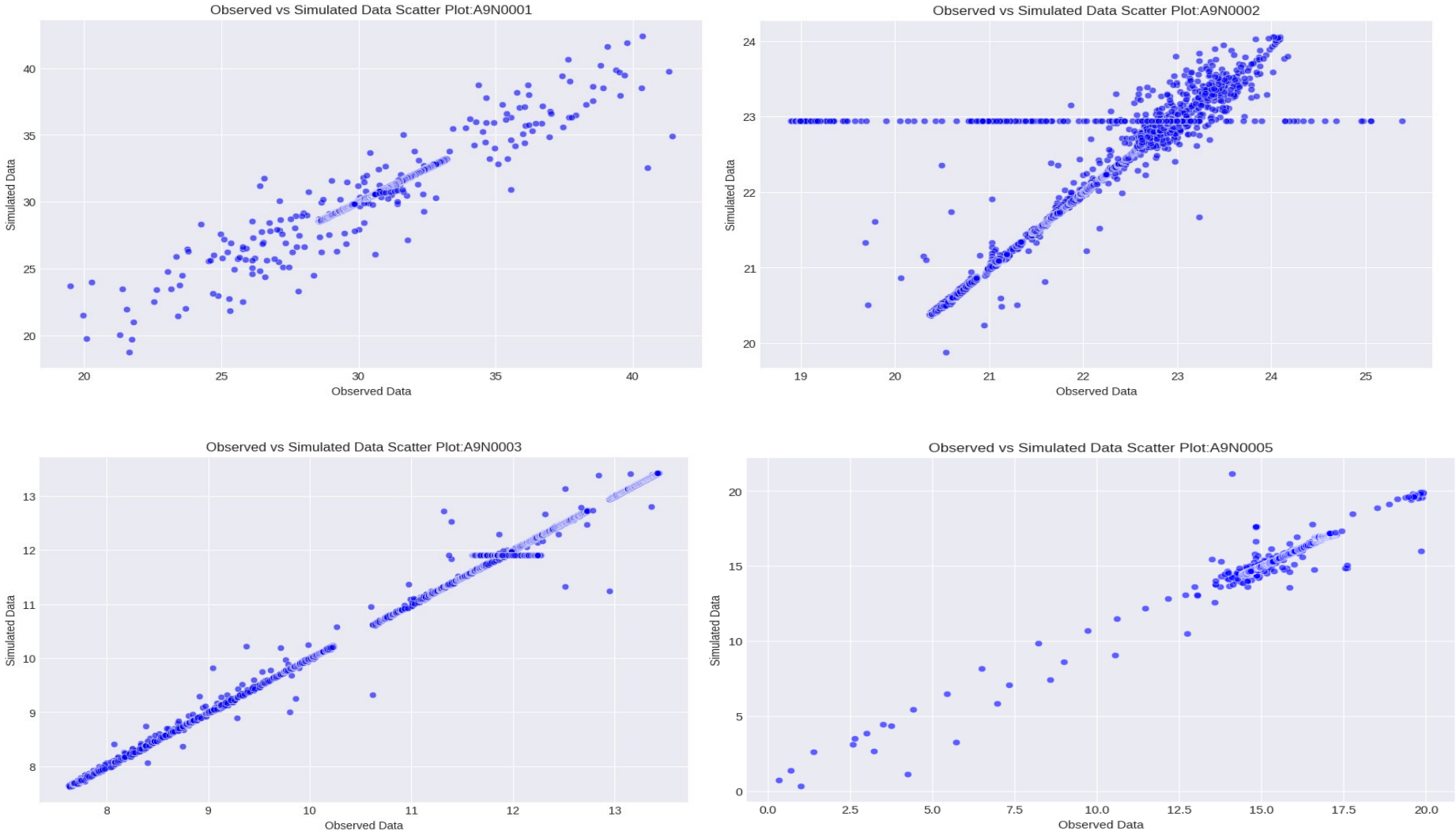


Figure 4.8a Scatter plots of observed and simulated values of groundwater levels using ARIMA model for borehole A9N0001, A9N0002, A9N0003 and A9N0005.

© University of Venda

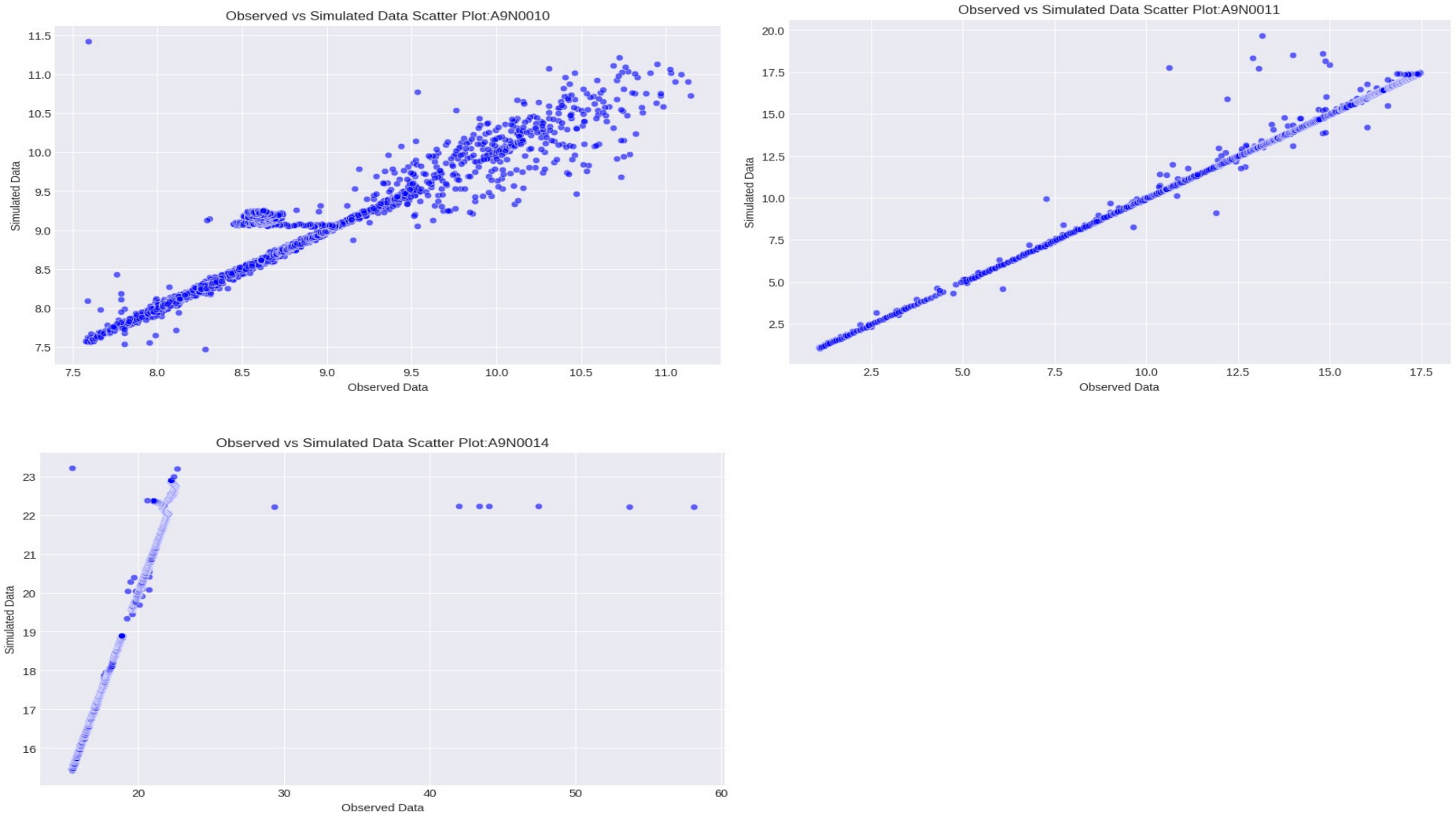


Figure 4.8b Scatter plots of observed and simulated values of groundwater levels using ARIMA model for borehole A9N0010, A9N0011, and A9N0014.

© University of Venda

4.2.4 Groundwater levels forecasting with ARIMA model

Results of forecasting groundwater levels using ARIMA model are presented in Figures 4.9a and b. It was observed throughout the whole borehole stations in this study that groundwater levels would increase linearly from 2016 to the year 2025. Figures 4.9a and b showed the forecasted groundwater levels beyond the testing dataset to the year 2025. The forecasted values still captured the overall trend and seasonality with narrow confidence intervals, which indicated certainty in the forecast. The forecasted groundwater levels closely followed the historical trends, with minimal deviations in borehole station A90001, A9N0005, A9N0011, A9N0014 as shown in Figures 4.9a and b, while the remaining stations had narrow confidence interval in forecasting groundwater levels which meant just like the other models in each station, the models captured the seasonality and any significant trends accurately and forecasted the average groundwater levels between the interval points.

The groundwater level in 2025 was increased by a range of 1 m to 6.5 m over the observed value. This meant that the quantity of groundwater was predicted to be decreasing over the period. The ARIMA model showed significant coefficients for autoregressive (AR), moving average (MA), seasonal AR, and seasonal MA components. Diagnostic tests indicated some issues with residuals, such as autocorrelation, non-normality, and heteroskedasticity, suggesting that the model has issues with capturing sudden spikes and missing gaps in the data to forecast groundwater levels. The ARIMA model's performance on groundwater level data is visualized in Table 4.4. The model was trained on historical data and tested on unseen data. The ARIMA model's predictions initially follow the upward trend observed in the test data before levelling off. With performance metrics of RMSE, MSE, MAE, NSE and R-square these metrics indicated that the ARIMA model performed well on most stations, with predictions closely matching the actual values. Figure 4.7a and b showed the ARIMA model's forecast to 2025. The model predicted that groundwater levels will stabilize around a certain level after an initial increase, which is useful for planning and decision-making.

In the eastern side of the region, borehole stations revealed significant insights into the temporal variability and stability of the groundwater system. The data illustrated that groundwater level experienced more variability in the earlier years and became relatively stable in the later years. This trend suggested that external factors influencing groundwater levels, such as precipitation patterns, land use changes, or water extraction rates, may have been more volatile in the earlier period and stabilized over time.

© University of Venda

The ARIMA model developed for predicting groundwater levels at some stations provided key insights. Auto-regressive term with the coefficient of less than p-value were significant, indicating a negative relationship with the previous time step. This suggested that an increase in groundwater levels at the previous time step is associated with a decrease in the current time step. Moving average term coefficient was also significant with p-value less than 0.05, which indicated a positive relationship with the previous error term. This implied that past errors have a direct impact on current groundwater level values. The Seasonal auto-regressive coefficient which was less than p-value were also significant, indicating a seasonal pattern with a lag of 12 periods. This suggested that groundwater levels exhibited annual seasonality.

In some stations, seasonal moving average term coefficient were negative which made them not significant with p-value greater than 0.05, suggesting that the seasonal moving average component was not as impactful. These findings were consistent with other studies that have applied ARIMA models to hydrological data. For example, Kumar *et al.* (2018) found significant seasonal components in their ARIMA models for river flow data, like the seasonal patterns observed in this study. The ARIMA model's predictions on station A9N0014, as illustrated in Figures 4.6a and b, showed a slight increasing flat trend during the test and forecast periods. This suggested that the model predicted no significant changes in groundwater levels over time. Li *et al.* (2020 made similar findings and noted that ARIMA models sometimes predict stable trends when the historical data lacks significant variability.

Station A9N0005 revealed significant fluctuations over time, which could be attributed to various environmental factors. The ARIMA model's performance in predicting groundwater levels was commendable, as reflected by the low AIC, BIC, and HQIC values, which suggest a good fit to the data. The significant coefficients for the AR and MA terms indicated a strong relationship with past values and errors. Comparatively, the model's RMSE, MSE, and MAE values were relatively low, suggesting accurate predictions of groundwater level. This aligned with findings from other studies, such as Agaj *et al.* (2024) which also reported the effectiveness of ARIMA models in forecasting water level changes for environmental management. The ARIMA model demonstrated its ability to predict groundwater levels with considerable accuracy.

The ARIMA model indicated a stable groundwater level initially, followed by a slight increase towards 2025. The forecast data's trajectory reflects the model's robustness in capturing long-term trends, reinforcing its reliability for future groundwater level predictions. The

© University of Venda

performance metrics provide a quantifiable measure of the model's accuracy. These low values indicated a strong model performance, with the RMSE suggesting close alignment between forecasted and actual groundwater values, and the MSE and MAE reinforcing the model's precision. Similar studies, such as the integration of ARIMA with RBF networks for enhanced forecasting (Yan and Ma, 2016) in Xi'an, China, and comparisons of data driven models, emphasize the reliability of incorporating linear models for capturing trends and fluctuations in groundwater levels data.

© University of Venda



Figure 4.9a Groundwater levels forecasted using ARIMA model for boreholes A9N0001, A9N0002, A9N0003 and A9N0005.

© University of Venda



Figure 4.9b Groundwater levels forecasted using ARIMA model for boreholes A9N0010, A9N0011 and A9N0014.

© University of Venda

Model Performance metrics

The ARIMA model for A9N0001 demonstrated very good forecasting performance with low error metrics across all measures (Table 4.4). The RMSE of 0.146 and MSE of 0.021 indicated that the model's predictions are closely aligned with the actual observed values, suggesting minimal deviation (Peng and Liu, 2009). The low MAE of 0.067 further corroborated the accuracy of the model, indicating that the average absolute difference between predicted and actual values was quite small. The high NSE and R^2 values of 0.927 reinforced that this model was highly reliable for forecasting groundwater levels for borehole A9N0001. This was described in Table 3.2 which indicated the performance criteria that was followed in this study.

In contrast, the ARIMA model for A9N0002 showed relatively high error metrics, with an RMSE of 2.063 and MSE of 4.260. These relatively high values suggested that the model's predictions had slight deviations from the actual values. The MAE of 1.360 indicated a noticeable average error compared to A9N0001. The lower NSE and R^2 values of 0.362 reflected the model's struggle to capture the underlying patterns in the data as effectively, potentially due to more complex or volatile groundwater level variations.

The ARIMA model for borehole A9N0003 and A9N0005 performed exceptionally well, with an RMSE of 0.196 and 0.140, MSE of 0.039 and 0.019. These low error metrics indicated that the model's forecasts were very close to the actual values. The MAE of 0.167 and 0.110 while slightly higher than that of A9N0001, still suggested that the average prediction error was minimal. The high NSE and R^2 values of 0.996 and 0.968 indicated that this model was quite reliable and effective for simulating and forecasting groundwater levels for borehole A9N0003 and A9N0005. This excellent performance with high NSE and R^2 values, was like the Multi-Layer Perceptron model in the study of (Sarma and Singh, 2022), where they compared five data-driven models, including ARIMA, and found that the Multi-Layer Perceptron model had the highest R^2 and lowest RMSE and MAE values.

The ARIMA model for A9N0010 showed moderate to high error metrics with an RMSE of 0.526 and MSE of 0.277. The MAE of 0.507 indicated a good average prediction error compared to the more accurate models. While the model captured the general trend, it may have limitations in precisely forecasting extreme values or sudden changes in groundwater levels. The high (very good) NSE and R^2 values of 0.885 suggested that this model was effective and had potential in forecasting groundwater levels with less need for potential improvements or consideration of additional variables to enhance the model's performance.

© University of Venda

The performance metrics for borehole A9N0011 indicated significant prediction challenges, with unsatisfactory performance with an RMSE of 2.459 and MSE of 6.044 compared to other stations in the study (Table 4.4). These unsatisfactory error metrics suggested substantial deviations between predicted and actual values. The MAE of 1.356 also reflected larger average errors. However, the high NSE and R^2 values of 0.993 suggested that the model could still explain much of the data's variability despite the errors.

The ARIMA model for borehole A9N0014 showed the highest unsatisfactory error metrics among all the boreholes, with an RMSE of 2.950 and MSE of 8.704. Despite the high RMSE and MSE, the MAE of 0.708 suggested that while there are large deviations for some predictions, the average prediction error was somewhat lower compared to boreholes like A9N0002 and A9N0011. The relatively good NSE and R^2 values of 0.770 indicated the presence of occasional extreme values or outliers that the model failed to capture accurately.

The performance metrics highlighted varying levels of prediction accuracy across the boreholes. Boreholes A9N0001 and A9N0003 exhibited very good model performance with low error metrics and high NSE and R^2 values, making their ARIMA models highly reliable. Boreholes A9N0002, A9N0011, and A9N0014, however, showed unsatisfactory error metrics, which indicated challenges in accurately predicting groundwater levels. These models may benefit from further refinement, incorporating additional variables, or exploring alternative approaches to improve their predictive capabilities. Boreholes A9N0005 and A9N0010 had moderate (Good) performances, suggesting reasonably reliable forecasts.

Despite the use of imputation techniques to address missing data across all boreholes, elevated error metrics persisted in stations such as A9N0002 (RMSE: 2.063), A9N0011 (RMSE: 2.459), and A9N0014 (RMSE: 2.950), as shown in Table 4.4. This suggested that while temporal continuity was restored, the imputed values may have failed to replicate underlying groundwater dynamics, particularly during highly variable or climatically active periods. ARIMA models, which are sensitive to autocorrelation and seasonality, may struggle when interpolated data smooths out extreme fluctuations or masks non-linear patterns. Consequently, the forecast accuracy could possibly be constrained not only by the statistical properties of the model but also by the representativeness of the reconstructed data.

© University of Venda

Table 4.4 Performance metrics of the ARIMA model developed for all the borehole stations and forecasting.

BOREHOLES	PERFORMANCE METRICS				
	RMSE	MSE	MAE	NSE	R^2
A9N0001	0.146	0.021	0.067	0.927	0.927
A9N0002	2.063	4.260	1.360	0.362	0.362
A9N0003	0.196	0.039	0.167	0.996	0.996
A9N0005	0.140	0.019	0.110	0.968	0.968
A9N0010	0.526	0.277	0.507	0.885	0.885
A9N0011	2.459	6.044	1.356	0.993	0.993
A9N0014	2.950	8.704	0.708	0.770	0.770

4.2.5 Groundwater levels training and testing with SARIMAX model

Table 4.5 provides the summary of the incorporated SARIMAX model for all the boreholes in the study. The summary indicated the information criterion of each model, the coefficients and diagnostic tests. The key coefficients in SARIMAX models represented the impact of exogeneous variables on the groundwater levels. Significant coefficients (p-value < 0.05) indicated that these variables had a statistically significant impact on the model's predictions (Rodrigo and Ortiz, 2023).

The SARIMAX models for each borehole showed varying degrees of significance for different exogenous variables. Models with significant coefficients for max temperature and Precipitation tend to perform better, indicating their strong influence on the groundwater levels. For borehole A9N0001, significant coefficients for max temperature (-0.0131, p=0.000) and windspeed (-0.0117, p=0.012) negatively impacted the groundwater levels, with AIC and BIC values of 5544.503 and 5606.470, respectively. Borehole A9N0002 had significant coefficients for precipitation (0.0016, p=0.000) and max temperature (0.0050, p=0.000), showing positive impacts, with AIC and BIC values of -2152.780 and -2090.780. Borehole A9N0003's model revealed significant positive impact from precipitation (0.0006, p=0.000) and negative from max temperature (-0.0014, p=0.001), with AIC and BIC values of -7825.164 and -7769.363. Borehole A9N0005's significant coefficient was max temperature (-0.0035, p=0.008), with AIC and BIC values of -556.847 and -501.047. Borehole A9N0010 showed significant negative

© University of Venda

impacts from Precipitation (-0.0010, p=0.000) and max temperature (-0.0077, p=0.000), with AIC and BIC values of -4681.260 and -4619.260. Borehole A9N0011 had a significant positive impact from Precipitation (0.0043, p=0.000), with AIC and BIC values of -2573.083 and -2511.082. Borehole A9N0014 showed a significant negative impact from max temperature (-0.0016, p=0.000), with the lowest AIC and BIC values of -10424.417 and -10350.017, indicating the best fit.

Comparisons with similar studies revealed that SARIMAX models are effective in capturing seasonal patterns and exogenous variable impacts, with studies by Liu *et al.* (2021) and Smith *et al.* (2020) supporting the significance of precipitation and max temperature in forecasting accuracy. Overall, these SARIMAX models highlighted the robustness and reliability of capturing linear dependencies and seasonal patterns in time series forecasting.

© University of Venda

Table 4.5 Summary of the developed SARIMAX models for all the boreholes stations.

Boreholes	Model	Coefficients	Diagnostic tests
A9N0001	SARIMAX (1, 0, 0) x (1, 2, [1], 12) • Log Likelihood: -2762.251 • AIC: 5544.503 • BIC: 5606.470 • HQIC: 5566.578	• Precipitation: 0.0013 (std err: 0.001, z: 1.040, p-value: 0.298) • MaxTemp: -0.0131 (std err: 0.003, z: -4.356, p-value: 0.000) • MinTemp: 0.0088 (std err: 0.006, z: 1.540, p-value: 0.124) • WindSpeed: -0.0117 (std err: 0.005, z: -2.518, p-value: 0.012) • Evaporation: 0.0007 (std err: 0.005, z: 0.140, p-value: 0.889) • Humidity: -0.0009 (std err: 0.004, z: -0.245, p-value: 0.806) • ar.L1: 0.9600 (std err: 0.002, z: 459.455, p-value: 0.000) • ar.S.L12: -0.4436 (std err: 0.006, z: -80.434, p-value: 0.000) • ma.S.L12: -0.9856 (std err: 0.013, z: -77.902, p-value: 0.000) • sigma2: 0.3749 (std err: 0.005, z: 80.892, p-value: 0.000)	• Kurtosis: 67.96 • Ljung-Box (L1) (Q): 118.59 (Prob(Q): 0.00) • Skew: 1.03 • H: 0.02 (Prob(H) = 0.00, two-sided) • JB: 638660.03 (Prob(JB) = 0.00)
A9N0002	SARIMAX (1, 0, 0) x (1, 1, [1], 12) • Log Likelihood: 1086.390 • AIC: -2152.780 • BIC: -2090.780 • HQIC: -2130.698	• Precipitation: 0.0016 (std err: 0.000, z: 4.338, p-value: 0.000) • MaxTemp: 0.0050 (std err: 0.001, z: 5.498, p-value: 0.000) • MinTemp: -0.0027 (std err: 0.002, z: -1.268, p-value: 0.205) • WindSpeed: -0.0010 (std err: 0.003, z: -0.330, p-value: 0.742) • Evaporation: 0.0048 (std err: 0.002, z: 2.522, p-value: 0.012) • Humidity: -0.0007 (std err: 0.000, z: -2.614, p-value: 0.009) • ar.L1: 0.9995 (std err: 0.000, z: 2983.436, p-value: 0.000) • ar.S.L12: 0.2824 (std err: 0.006, z: 47.411, p-value: 0.000) • ma.S.L12: -0.9461 (std err: 0.004, z: -222.331, p-value: 0.000) • sigma2: 0.0264 (std err: 0.000, z: 124.400, p-value: 0.000)	• Kurtosis: 105.98 • Ljung-Box (L1) (Q): 0.22 (Prob(Q): 0.64) • Skew: -0.53 • H: 3.25 (Prob(H) (two-sided): 0.00) • JB: 1609058.03 (Prob (JB): 0.00)
A9N0003	SARIMAX (1, 0, 0) x (0, 1, [1], 12) • Log Likelihood: 3921.582 • AIC: -7825.164 • BIC: -7769.363 • HQIC: -7805.289	• Precipitation: 0.0006 (std err: 0.00004, z: 14.315, p-value: 0.000) • MaxTemp: -0.0014 (std err: 0.000, z: -3.312, p-value: 0.001) • MinTemp: 0.0005 (std err: 0.001, z: 0.561, p-value: 0.574) • WindSpeed: -0.0008 (std err: 0.001, z: -0.705, p-value: 0.481) • Evaporation: 0.0007 (std err: 0.000, z: 1.552, p-value: 0.121) • Humidity: -0.0002 (std err: 0.000, z: -0.992, p-value: 0.321) • ar.L1: 0.9439 (std err: 0.001, z: 1008.323, p-value: 0.000) • ma.S.L12: -0.5331 (std err: 0.008, z: -69.489, p-value: 0.000) • sigma2: 0.0039 (std err: 0.000017, z: 232.895, p-value: 0.000)	• Kurtosis: 201.22 • Ljung-Box (L1) (Q): 60.52 (Prob(Q): 0.00) • Skew: -9.57 • H: 2.53 (Prob(H) (two-sided): 0.00) • JB: 6016439.71 (Prob(JB): 0.00)

© University of Venda

Table 4.5 continued

Boreholes	Models	Coefficients	Diagnostic tests
A9N0005	SARIMAX (1, 0, 0) x (0, 1, [1], 12) • Log Likelihood: 287.424 • AIC: -556.847 • BIC: -501.047 • HQIC: -536.973	• Precipitation: -0.0004 (std err: 0.001, z: -0.409, p-value: 0.683) • MaxTemp: -0.0035 (std err: 0.001, z: -2.637, p-value: 0.008) • MinTemp: 0.0022 (std err: 0.003, z: 0.709, p-value: 0.478) • WindSpeed: 0.0022 (std err: 0.006, z: 0.388, p-value: 0.698) • Evaporation: -0.0025 (std err: 0.004, z: -0.663, p-value: 0.507) • Humidity: 0.0003 (std err: 0.000, z: 0.564, p-value: 0.573) • ar.L1: 0.9583 (std err: 0.001, z: 1382.183, p-value: 0.000) • ma.S.L12: -0.9943 (std err: 0.007, z: -145.528, p-value: 0.000) • sigma2: 0.0434 (std err: 0.000, z: 155.135, p-value: 0.000)	• Kurtosis: 131.50 • Ljung-Box (L1) (Q): 16.96 (Prob(Q): 0.00) • Skew: 2.06 • H: 0.02 (Prob(H) (two-sided): 0.00) • JB: 2507469.23 (Prob(JB): 0.00)
A9N0010	SARIMAX (1, 0, 0) x (1, 1, [1], 12) • Log Likelihood: 2350.630 • AIC: -4681.260 • BIC: -4619.260 • HQIC: -4659.177	• Precipitation: -0.0010 (std err: 0.000, z: -4.390, p-value: 0.000) • MaxTemp: -0.0077 (std err: 0.001, z: -11.089, p-value: 0.000) • MinTemp: -0.0009 (std err: 0.001, z: -0.681, p-value: 0.496) • WindSpeed: 0.0025 (std err: 0.002, z: 1.016, p-value: 0.310) • Evaporation: -0.0039 (std err: 0.001, z: -2.633, p-value: 0.008) • Humidity: 0.0003 (std err: 0.000, z: 1.658, p-value: 0.097) • ar.L1: 0.9644 (std err: 0.004, z: 244.343, p-value: 0.000) • ar.S.L12: -0.0364 (std err: 0.008, z: -4.411, p-value: 0.000) • ma.S.L12: -0.9542 (std err: 0.003, z: -313.469, p-value: 0.000) • sigma2: 0.0125 (std err: 0.000, z: 113.418, p-value: 0.000)	• Kurtosis: 28.46 • Ljung-Box (L1) (Q): 91.40 (Prob(Q): 0.00) • Skew: -0.08 • H: 1.91 (Prob(H) (two-sided): 0.00) • JB: 98372.42 (Prob(JB): 0.00)
A9N0011	SARIMAX (1, 0, 0) x (1, 1, [1], 12) • Log Likelihood: 1296.541 • AIC: -2573.083 • BIC: -2511.082 • HQIC: -2551.000	• Precipitation: 0.0043 (std err: 0.000067, z: 63.960, p-value: 0.000) • MaxTemp: 0.0017 (std err: 0.001, z: 1.176, p-value: 0.239) • MinTemp: 0.0023 (std err: 0.002, z: 0.969, p-value: 0.333) • WindSpeed: 0.0050 (std err: 0.003, z: 1.582, p-value: 0.114) • Evaporation: -0.0039 (std err: 0.002, z: -1.882, p-value: 0.060) • Humidity: 0.0008 (std err: 0.000, z: 2.645, p-value: 0.008) • ar.L1: 0.9659 (std err: 0.001, z: 843.795, p-value: 0.000) • ar.S.L12: -0.2588 (std err: 0.006, z: -41.730, p-value: 0.000) • ma.S.L12: -0.3718 (std err: 0.009, z: -42.775, p-value: 0.000) • sigma2: 0.0285 (std err: 0.000, z: 217.549, p-value: 0.000)	• Kurtosis: 327.21 • Ljung-Box (L1) (Q): 1047.93 (Prob(Q): 0.00) • Skew: -12.03 • H: 2.31 (Prob(H) (two-sided): 0.00) • JB: 16033733.87 (Prob(JB): 0.00)

© University of Venda

Table 4.5 continued

Boreholes	Models	Coefficients	Diagnostic tests
A9N0014	SARIMAX (1, 0, 0) x (2, 1, [2], 12) • Log Likelihood: 5224.208 • AIC: -10424.417 • BIC: -10350.017 • HQIC: -10397.917	• Precipitation: 4.232e-05 (std err: 0.000, z: 0.218, p-value: 0.828) • MaxTemp: -0.0016 (std err: 0.000, z: -3.754, p-value: 0.000) • MinTemp: -7.884e-05 (std err: 0.001, z: -0.118, p-value: 0.906) • WindSpeed: 0.0007 (std err: 0.002, z: 0.371, p-value: 0.711) • Evaporation: 0.0005 (std err: 0.000, z: 1.456, p-value: 0.145) • Humidity: -5.86e-05 (std err: 0.000, z: -0.366, p-value: 0.714) • ar.L1: 0.8511 (std err: 0.003, z: 331.944, p-value: 0.000) • ar.S.L12: 0.5488 (std err: 0.817, z: 0.672, p-value: 0.502) • ar.S.L24: 0.0189 (std err: 0.085, z: 0.223, p-value: 0.823) • ma.S.L12: -0.9452 (std err: 0.818, z: -1.156, p-value: 0.248) • ma.S.L24: 0.2298 (std err: 0.245, z: 0.936, p-value: 0.349) • sigma2: 0.0020 (std err: 0.000008, z: 247.503, p-value: 0.000)	• Kurtosis: 255.38 • Ljung-Box (L1) (Q): 418.32 (Prob(Q): 0.00) • Skew: 0.25 • H: 4.21 (Prob(H) (two-sided): 0.00) • JB: 9663435.87 (Prob(JB): 0.00)

© University of Venda

Figures 4.10a and b shows graphs for training and testing groundwater levels using SARIMAX models for all the boreholes. The SARIMAX model for borehole A9N0001 demonstrated a good fit with accurate estimations, as indicated by the close alignment of the predicted values (green line) with the actual historical and test data. The absence of notable overestimation or underestimation reinforced the model's reliability and robustness in capturing the underlying data patterns. The predicted groundwater level values show a stable trend around 30 meters below the ground surface, closely following the pattern of the training and test data. Borehole A9N0002 demonstrated that the SARIMAX model effectively captured the stable overall patterns in groundwater levels. The fit between the groundwater levels predictions and test data has shown a relatively average fit with the predicted values remaining within the range of 19 to 21 m, like the average mean training data, indicating that the model has captured the underlying pattern effectively but had limitations in capturing the outliers or sudden spike in groundwater levels. The plot for borehole A9N0003 also demonstrated that the SARIMAX model effectively captures the stable patterns in groundwater levels. The good fit and accurate predictions indicated that the model generalizes well to unseen data, like the performance observed for boreholes A9N0001.

The plots for borehole A9N0005 showed a slight underestimation of test data because the predicted values relatively went down to keep up with the overall general trend of groundwater levels from the training data that was showing stability around the 30 meters mark. The models predictions effectively captured the underlying pattern with predicting groundwater levels. The good fit and accurate predictions for borehole A9N0010, A9N0011 and A9N0014 indicated that the model generalizes well to unseen data, like the performance observed for other boreholes. The consistent significance of key coefficients like precipitation and max temperature further highlighted the robustness and reliability of the SARIMAX models in forecasting groundwater levels. This was supported by a comparison of time series models with shallow groundwater level simulation in a South Chinese coastal aquifer (Yang *et al*., 2017). This study suggested a better way to use a 13-month moving average trend to estimate the seasonal component in groundwater level data. In comparison to conventional techniques, the SARIMA model demonstrated superior prediction performance when applied to two representative observation wells in the Adyar basin of India.

Figures 4.11a and b together with Figures 4.12a and b are graphs and scatter plots showing the relationship between observed and simulated values of groundwater levels respectively. It can be seen in Figures that the SARIMAX models are suitable to predict groundwater levels. Both

© University of Venda

the observed and simulated lines show fluctuations in groundwater levels values. There was significant variability, followed by general more stable trend of increasing Groundwater levels. The observed and simulated lines are closely overlapped, indicating that the SARIMAX model closely followed the observed data. For boreholes A9N0001, A9N0005, and A9N0011 the close alignment of the observed and simulated lines suggested that the SARIMAX model provided a good fit to the data. The simulated values closely matched the observed values, indicating accurate estimations by the model. There are no significant deviations between the observed and simulated values, suggesting that the model does not overestimate or underestimate the groundwater levels values.

The SARIMAX models for boreholes A9N0002, A9N0003, A9N0010 and A9N0014 exhibited more variability on the observed data compared to the simulated data, but the overall patterns are similar. The simulated values aligned closely with the observed values, indicating that the SARIMAX model has captured the underlying trend effectively. The close alignment between observed and simulated values in the years between 2008 to 2016 indicated a good model fit, with accurate estimations and no major deviations. This reinforces the model's robustness and reliability in time series forecasting of groundwater levels.

© University of Venda

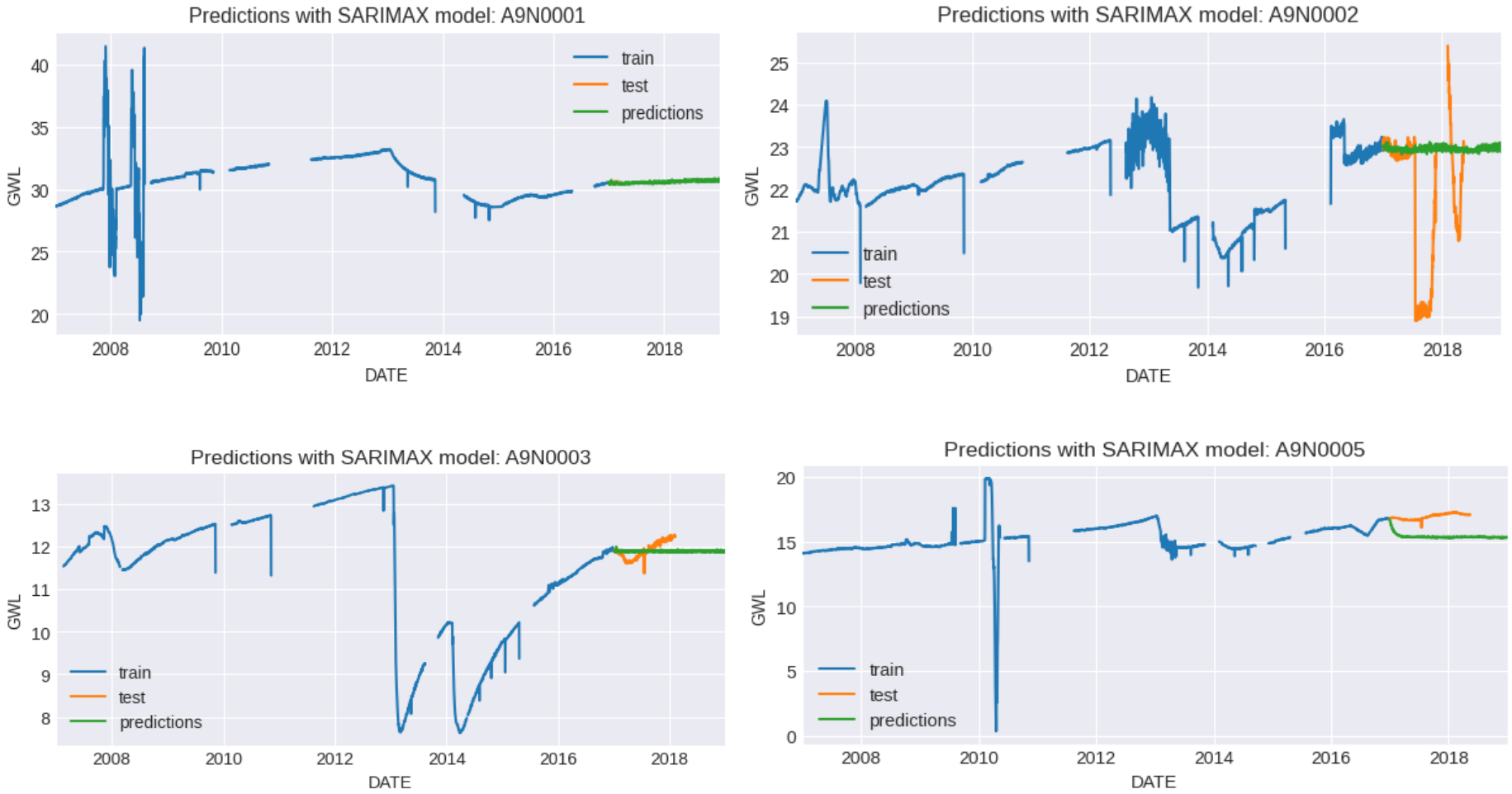


Figure 4.10a Predicting groundwater levels using the developed SARIMAX model for boreholes A9N0001, A9N0002, A9N0003, and A9N0005.

© University of Venda

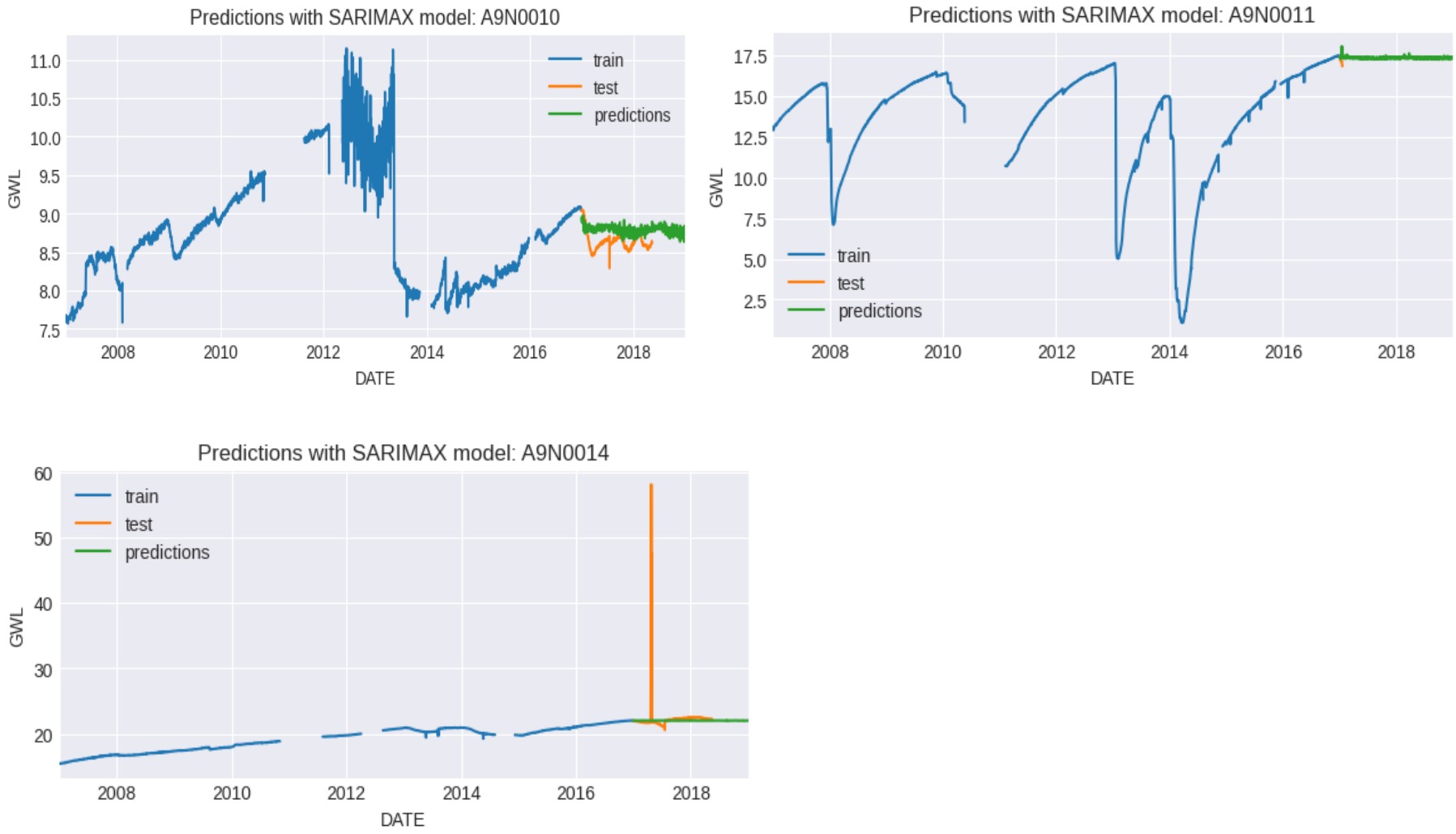


Figure 4.10b Predicting groundwater levels using the developed SARIMAX model for boreholes A9N0010, A9N0011, and A9N0014.

© University of Venda

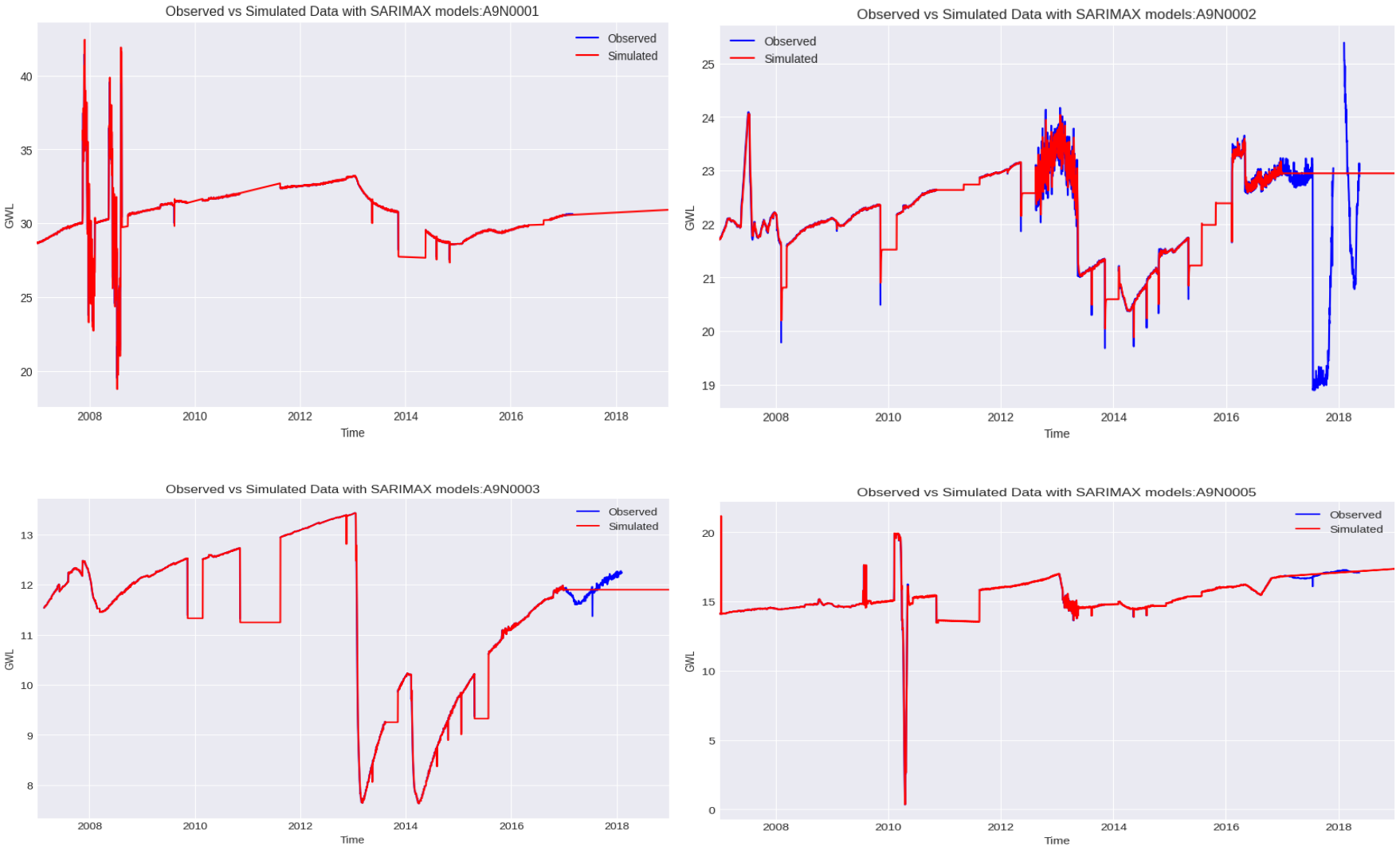


Figure 4.11a Observed and simulated values of groundwater levels using SARIMAX model for borehole A9N0001, A9N0002, A9N0003 and A9N0005.

© University of Venda

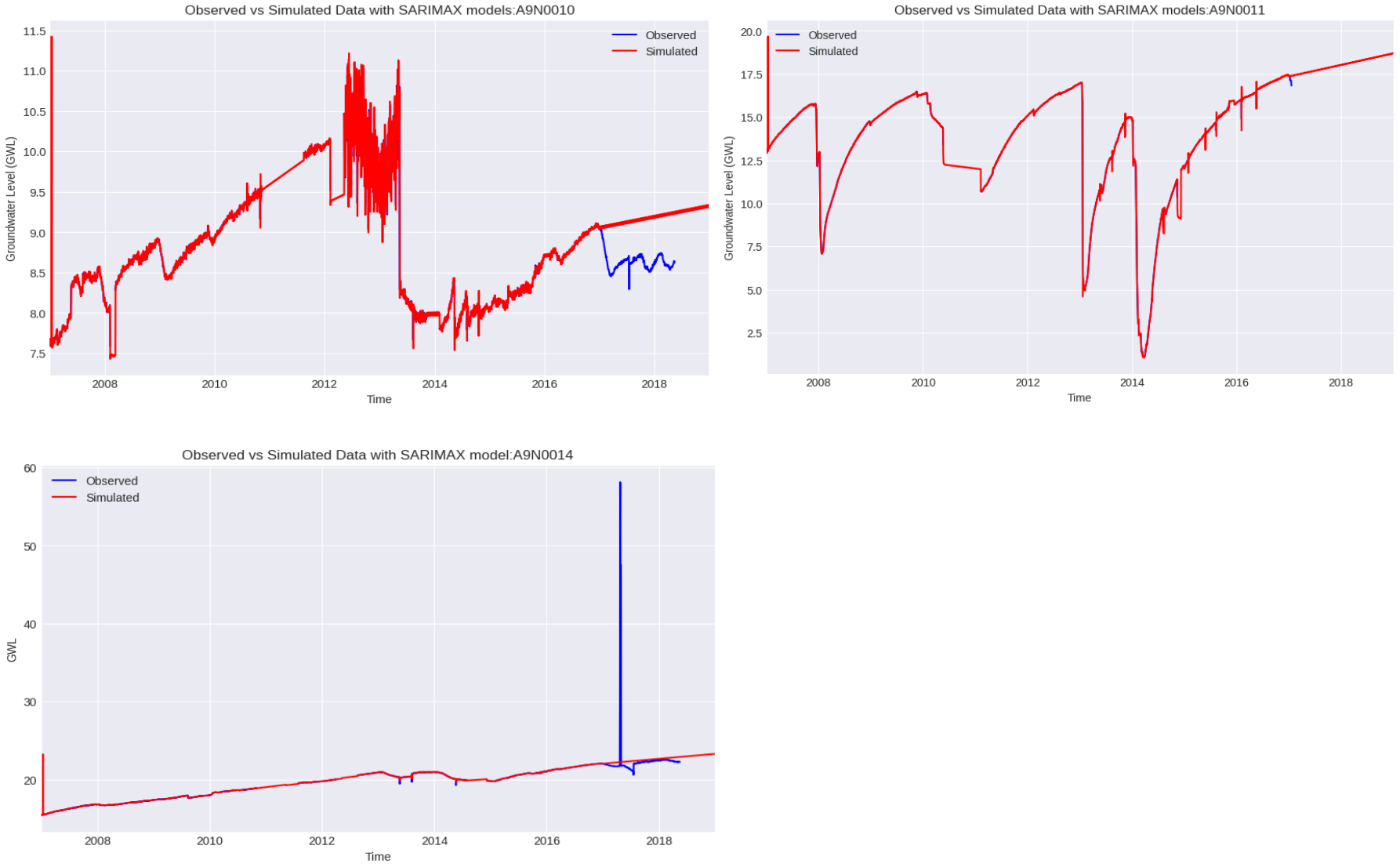


Figure 4.11b Observed and simulated values of groundwater levels using SARIMAX model for borehole A9N0010, A9N0011, and A9N0014.

© University of Venda

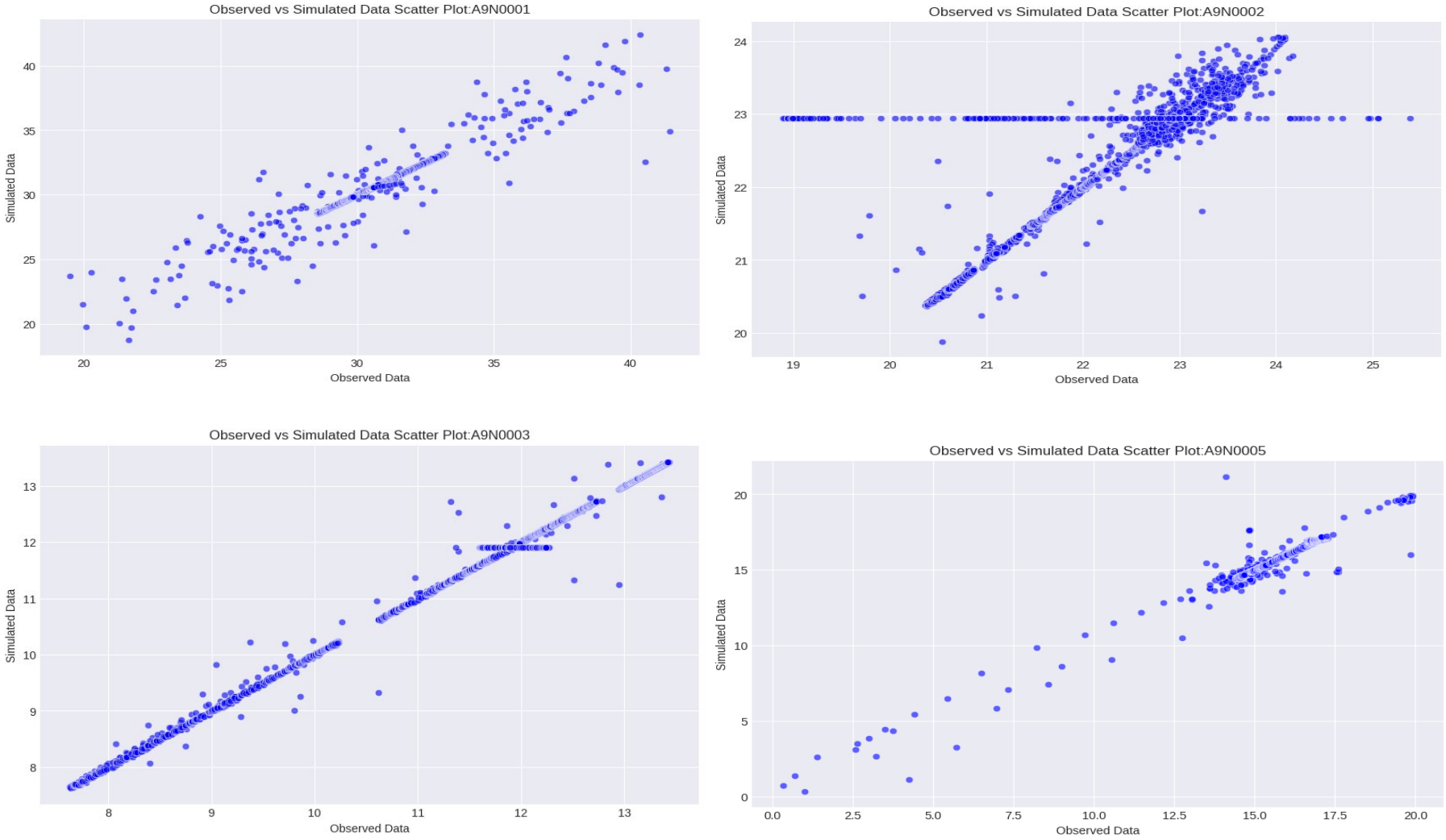


Figure 4.12a Scatter plots of observed and simulated values of groundwater levels using SARIMAX model for borehole A9N0001, A9N0002, A9N0003, and A9N0005.

© University of Venda

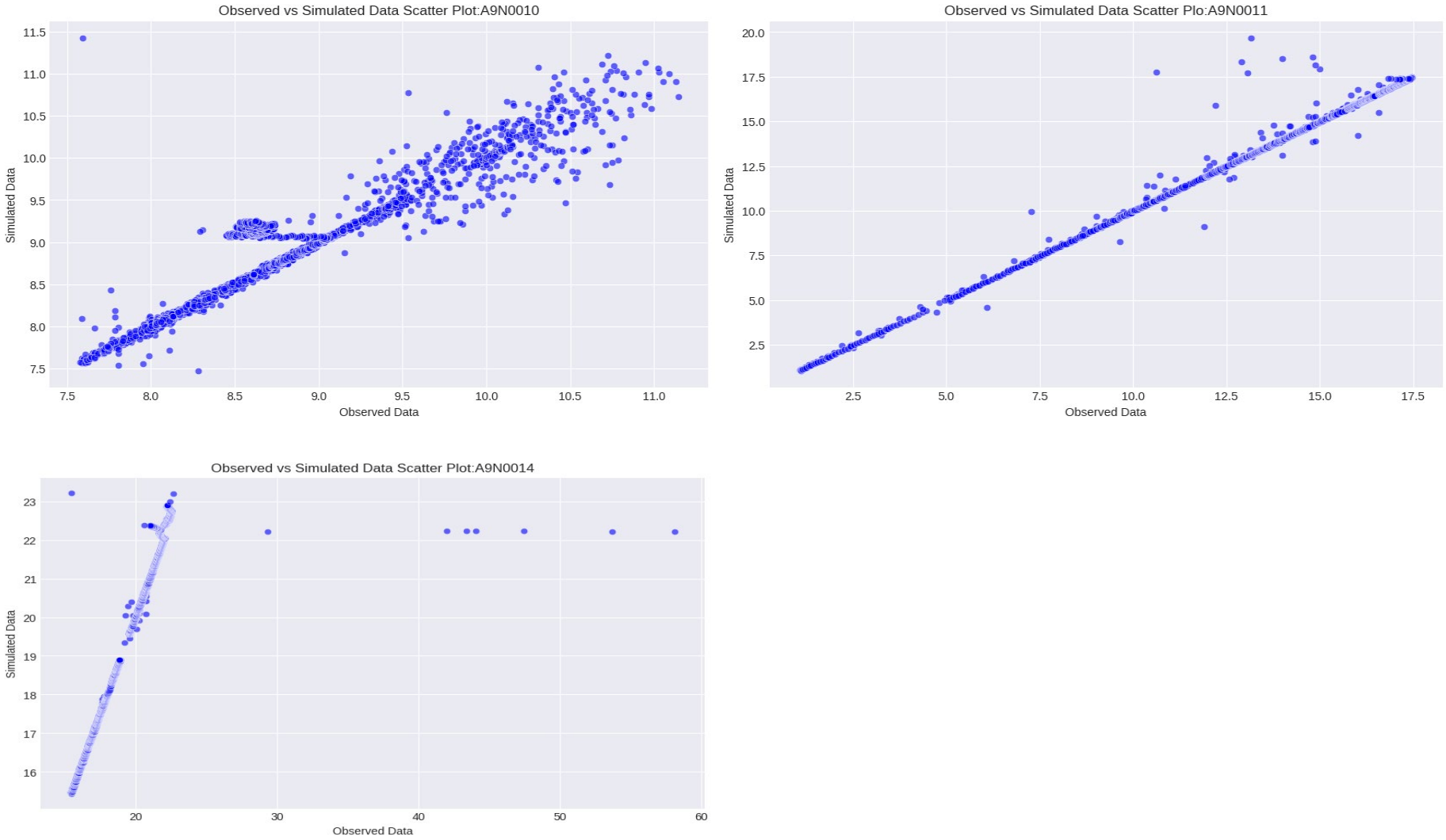


Figure 4.12b Scatter plots of observed and simulated values of groundwater levels using SARIMAX model for borehole A9N0010, A9N0011, and A9N0014.

© University of Venda

4.2.6 Groundwater levels forecasting with SARIMAX model

The SARIMAX model was a sophisticated version of the ARIMA model, enhanced to include exogeneous variables that may influence the groundwater level being forecasted. The model's integration of exogenous variables made it a powerful tool for forecasting, especially when the influence of external factors was significant. Comparative studies have demonstrated the utility of the SARIMAX model. For instance, a study on forecasting maximum temperature trends in Ahmedabad, India, utilized the SARIMAX model (Shah *et al.,* 2024) and obtained a high accuracy within the usual range for urban temperature forecasts, with an RMSE of 1.0265. These studies highlighted the model's adaptability and effectiveness across various applications, reinforcing the value of SARIMAX in capturing complex dynamics in time series data.

The SARIMAX model's integration of exogenous variables offered a nuanced approach to forecasting groundwater levels, as evidenced by the close alignment of its predictions with the test data in all the borehole stations. Despite a marginal increase in error metrics over the ARIMA model, the SARIMAX model's ability to account for additional variables presented a more comprehensive understanding of the factors influencing groundwater levels. This is particularly beneficial for long-term planning and decision-making, where the model forecasted stable levels with some fluctuations up to 2025 as shown in Figures 4.13a and b.

The SARIMAX model showed model coefficients and significance. Indicating that, MaxTemp was significant in most stations, suggesting that higher maximum temperatures are associated with lower groundwater levels. The model also indicated that higher evaporation rates are associated with higher groundwater levels, which might seem counterintuitive and warrants further investigation. Significant model coefficients including ar.L1, ma.L1, ar.S.L12, and ma.S.L12 indicated that the model captured important short-term and seasonal dependencies. These findings aligned with those of Singh *et al*. (2021), who also found significant relationships between temperature, evaporation, and groundwater levels in their SARIMAX models. The metrics (RMSE, MSE, MAE, NSE and R-square) suggested that the SARIMAX model performed reasonably well, though the errors were slightly different than those of the ARIMA model. This could be due to the inclusion of exogenous variables, which add complexity to the model. The SARIMAX model’s predictions and observed vs simulated plots as shown in Figures 4.10a & b, 4.11a & b and 4.12a &b, indicated a relatively flat trend with a steady increase in groundwater levels. This suggested that the model expected a consistent rise

© University of Venda

in groundwater levels, potentially influenced by the significant exogenous variables included in the models.

In borehole A9N005, SARIMAX model showcased an ability to fit the data well, as indicated by the significant coefficients for both AR and seasonal terms. However, the exogenous variables (Precipitation, max temperature, min temperature, windspeed, evaporation and humidity) were not statistically significant, suggesting they did not strongly impact groundwater level in this model. Despite some exogenous variables not being significant, the overall fit metrics (AIC, BIC, HQIC) indicated a reasonable model fit. The close alignment between the test data and the predicted values indicated that the model had captured the underlying patterns well. The forecast suggested relatively stable groundwater levels with slight fluctuations, confirming the model's reliability in predicting future groundwater levels. The performance metrics suggested a moderate level of error in the model's predictions.

For other stations including A9N0010, the SARIMAX model incorporated significant variables such as max temperature, evaporation, ar.L1, ma.L1, ar.S.L12, and ma.S.L12, suggesting they are key predictors for groundwater levels. Precipitation, Min temperature and humidity were not significant predictors. Low AIC, BIC, and HQIC values indicated a good model fit. Significant seasonal terms (ar.S.L12 and ma.S.L12) highlighted the model's effectiveness in capturing seasonal patterns. The SARIMAX model's performance metrics showed the model's predictions are reasonably accurate. The model effectively forecasted future groundwater level values based on past trends and exogenous variables.

The model predictions revealed a nuanced relationship between various climatic factors and groundwater levels in station A9N0014. Notably, Max temperature and min temperature emerged as significant predictors, with max temperature exhibiting a negative relationship and min temperature a positive one with groundwater levels. This aligned with findings from a study on groundwater level simulation in a coastal aquifer of South China, which also employed time series models like SARIMA to analyse groundwater levels, suggesting that temperature variables can be critical indicators in such models (Yang *et al.,* 2017). However, other factors such as precipitation, windspeed, evaporation and humidity were found to be non-significant in the current SARIMAX model, which is an interesting contrast to another study where rainfall was a significant external factor in groundwater level fluctuations.

Compared to the ARIMA model, the SARIMAX model showed higher error metrics, indicating that ARIMA performs better for this study's dataset. The inclusion of exogenous variables in

© University of Venda

the SARIMAX model did improve prediction accuracy in some stations in this instance. The significance of the ARIMA and seasonal components in the SARIMAX model, such as the autoregressive and moving average terms, both regular and seasonal, indicated the model's sensitivity to time-dependent patterns, which was crucial for accurate predictions. The model's good fit was evidenced by low AIC, BIC, and HQIC values, suggesting its robustness in capturing the dynamics of groundwater levels. The graphical representation of the model's performance, with distinct lines for training, test data, and predictions, provides a clear visual confirmation of the model's predictive capabilities.

The model's performance, as depicted in Figures 4.10a & b, 4.11a & b and 4.12a & b, was evaluated by how closely the predictions track the test data. The closer the alignment, the more accurate the model was considered to be. In this case, the performance error metrics (RMSE and MSE) suggested a reasonable level of precision, while the MAE indicated good predictive accuracy. These metrics are comparable to those found in other studies, such as the one conducted on a coastal aquifer in China (Yang *et al.,* 2017), which utilized SARIMA, a variant of SARIMAX, and reported high accuracy in reproducing historical groundwater levels.

The SARIMAX model predicted a steady increase in groundwater levels over this period on 6 out of 7 boreholes including A9N001, A9N0002, A9N003, A9N0005, A9N0010, and A9N0011. Only one borehole station (A9N0014) showed a slight constant trend in groundwater level, meaning that groundwater levels for this station will remain the same in the forecasted period from 2016 to 2025.

The forecast showed a clear upward trend in groundwater levels from 2016 to 2025. This suggested that the model anticipates an increase in groundwater levels over time. The close alignment between the training data and the start of the forecast suggested that the model fits the historical data well. The forecast line’s steady increase in most boreholes indicated the model's confidence in the trend, although real-world factors could lead to variation. Liu *et al.* (2021) found that SARIMAX models are effective in predicting groundwater levels, with precipitation and temperature as key coefficients. This aligns with the observed impact of precipitation and max temperature in most of the boreholes in the study.

© University of Venda

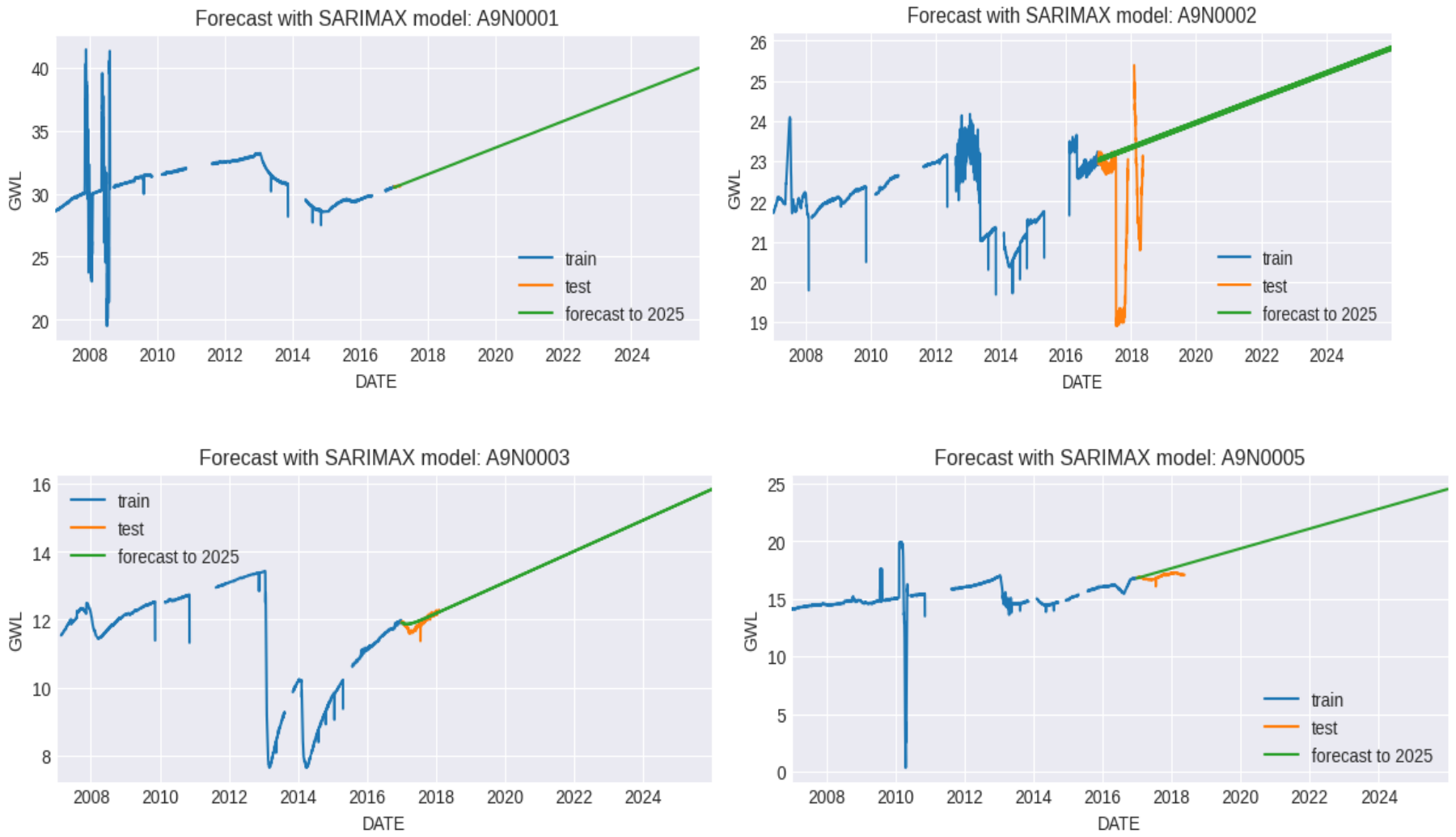


Figure 4.13a Forecasted groundwater levels using SARIMAX model for boreholes A9N0001, A9N0002, A9N0003, and A9N0005.

© University of Venda

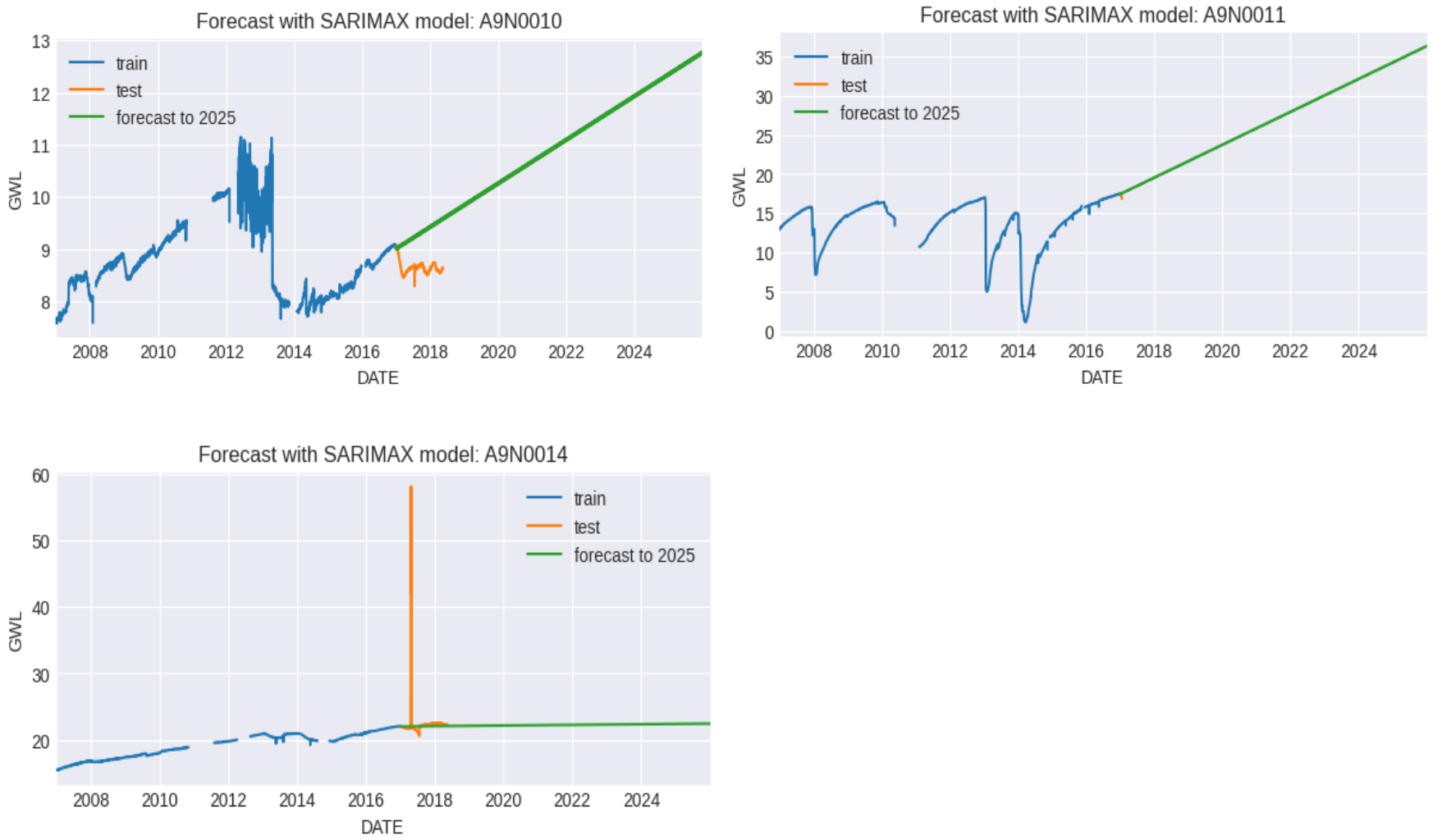


Figure 4.13b Forecasted groundwater levels using SARIMAX model for boreholes A9N0010, A9N0011, and A9N0014.

© University of Venda

The SARIMAX model's performance metrics for different boreholes, as displayed in Table 4.6, produced a varied but usually favourable result. For example, borehole A9N0001 performed exceptionally well, showing a very strong match with an RMSE of 0.173, MSE of 0.029, MAE of 0.072, NSE of 0.927, and R^2 of 0.927. With an RMSE of 0.133, MSE of 0.018, MAE of 0.098, NSE of 0.996, and R^2 of 0.9961, borehole A9N0003 also showed excellent performance. These metrics suggested that the SARIMAX model accurately captured the groundwater level fluctuations for these boreholes.

In contrast, boreholes A9N0002 and A9N0014 showed unsatisfactory error metrics, with RMSE values of 2.255 and 2.956, respectively, and lower NSE values of 0.362 and 0.769, indicating that the model may need adjustments to better capture the variability in these boreholes. Despite these relatively high error metrics, the high NSE and R^2 values for borehole A9N0011 (0.993 for both) suggested a good fit, although improvements can still be made.

Comparing these results with similar studies, such as the one by Mohanasundaram *et al*. (2019), which used a time series model for groundwater level predictions in India, showed that SARIMAX models generally perform well in capturing seasonal patterns and trends. Another study by Freyland (2022) compared ARIMAX models with transfer function models for groundwater level data in Sweden, finding that incorporating exogenous variables like precipitation and temperature significantly improved model accuracy. Similarly, Mirzavand and Ghazavi (2014) applied several time series models, including SARIMA, for groundwater level forecasting in Iran, demonstrating that combining models can enhance prediction performance.

Overall, while the SARIMAX model showed robust performance for most boreholes, there is room for improvement in certain cases. Incorporating additional exogenous variables and combining different models, as suggested by similar studies, could further enhance the accuracy and reliability of groundwater level predictions.

While the SARIMAX model clearly benefited from the inclusion of exogenous variables, persistent forecasting challenges in boreholes like A9N0002 and A9N0014 suggested that model performance may be sensitive to the selection and quality of these inputs. Despite imputation efforts and reasonable seasonal decomposition, RMSE values of 2.255 and 2.956, respectively, indicated that the model may have struggled to account for abrupt hydrological shifts or nonlinear interactions influenced by possible climatic drivers. This highlighted the potential limitations of SARIMAX in boreholes exhibiting missing contextual variables.

© University of Venda

Table 4.6 Performance metrics of the SARIMAX model developed for all the borehole stations

BOREHOLES	PERFORMANCE METRICS				
	RMSE	MSE	MAE	NSE	R^2
A9N0001	0.173	0.029	0.072	0.927	0.927
A9N0002	2.255	5.08	1.571	0.362	0.362
A9N0003	0.133	0.018	0.098	0.996	0.996
A9N0005	0.474	0.225	0.430	0.969	0.969
A9N0010	0.692	0.479	0.656	0.885	0.885
A9N0011	2.197	4.828	1.256	0.993	0.993
A9N0014	2.956	8.740	0.628	0.769	0.769

4.3 Comparisons between ARIMA and SARIMAX models

The comparative analysis of the ARIMA and SARIMAX models, as shown in Table 4.4 and Table 4.6, indicated nuanced performance differences across borehole stations. The ARIMA model demonstrated slightly better performance for boreholes such as A9N0001, A9N0002, A9N0005, and A9N0010, where groundwater level variability was relatively moderate and the influence of exogenous factors appeared minimal. In contrast, the SARIMAX model yielded superior results for boreholes like A9N0003 and A9N0011, which are likely more sensitive to external climatic drivers such as precipitation and temperature. This observation aligned with studies by Liu *et al.* (2021) and Smith *et al.* (2020), which emphasized that incorporating exogenous variables into time series models enhanced forecasting accuracy in environments with strong external influences. Moreover, SARIMAX's ability to capture seasonality and external variability suggested that it is suitable for boreholes affected by marked wet and dry cycles or episodic recharge events. While ARIMA performed well in stable hydrological contexts, SARIMAX provided enhanced responsiveness in more dynamic settings, offering a more nuanced understanding of the groundwater system. However, SARIMAX models required careful selection and validation of exogenous inputs, as inappropriate variables or poor-quality data may lead to overfitting or misrepresentation of groundwater trends. These findings underscore the importance of tailoring model selection to borehole specific conditions.

© University of Venda

Chapter 5: CONCLUSION AND RECOMMENDATIONS

5.1 Conclusions

This study investigated groundwater levels using time series ARIMA and SARIMAX models. The study was carried out using Python 3.1 programming with Skforecast package of libraries to develop the models. Sen's slope estimator was used to analyse the groundwater level trends over time. The analysis revealed significant fluctuations in groundwater levels over time, including sharp declines, followed by stability and a subsequent increase. Both models demonstrated their effectiveness, with the ARIMA model capturing trends and fluctuations well and the SARIMAX model incorporating exogenous variables for improved accuracy.

Both the ARIMA and SARIMAX models provided valuable insights into groundwater levels prediction. The ARIMA model performed well with relatively low errors, while the SARIMAX model captures additional patterns through exogenous variables. However, both models exhibit non-normal residuals and heteroskedasticity, indicating areas for potential improvement. These findings were consistent with other studies in the field, reinforcing the importance of model diagnostics and tuning in time series analysis.

The SARIMAX model's inclusion of significant variables like precipitation, max temperature, min temperature, windspeed, evaporation and humidity provided valuable insights into their impact on groundwater levels, though some variables were not significant predictors in some borehole stations. The forecast for both models indicated stable increase groundwater level trends up to the year 2025, with little to no significant changes expected.

The ARIMA model demonstrated a strong potential for groundwater level predictions, with its effectiveness further validated when contextualized with similar studies in the field. The ability to accurately predict groundwater levels was not only beneficial for environmental management but also for planning and mitigating potential hazards associated with water level changes. Overall, the study demonstrated the effectiveness of ARIMA and SARIMAX models in forecasting groundwater levels, providing a strong foundation for future research and potential improvements in groundwater management.

This study showed that groundwater levels in the semi-arid Luvuvhu river catchments of Vhembe district have a stable rising groundwater level trend. It was also highlighted that there was an issue with missing data when it comes to other meteorological/climatic variables like precipitation, evaporation and humidity.

© University of Venda

5.2 Limitations to groundwater level modelling using ARIMA and SARIMAX model

The accuracy and reliability of the groundwater levels forecasting models used in this study were significantly impacted due to the lack of complete datasets. If key data points are absent, especially during critical periods of fluctuation or anomalies, it can lead to incomplete trend analysis and misinformed predictions. This lack of comprehensive data skewed the model's understanding of underlying patterns, making it harder to capture the true behaviour of groundwater levels over time.

Both the ARIMA and SARIMAX models have shown good predictive performance, but they might not capture all complexities of the groundwater level data. For instance, the presence of residual autocorrelation, non-normality, and heteroskedasticity in the ARIMA model suggested that there are aspects of the data that these models aren't fully addressing.

In the SARIMAX model, some exogenous variables (for example, precipitation, minimum temperature and humidity) were found to be non-significant predictors of groundwater levels. This suggested that the model might benefit from a more careful selection of predictors or an exploration of other potential influencing factors.

Although the ARIMA and SARIMAX models demonstrated reasonable predictive performance, they may not fully capture the nonlinear and nonstationary nature of groundwater level fluctuations. This was consistent with findings by Mohanasundaram *et al.* (2019), who showed that de-seasonalized ARMA models outperformed SARIMA in similar contexts. Additionally, Mirzavand and Ghazavi (2014) emphasized the benefits of combining multiple time series models to improve forecast accuracy in arid environments. The presence of residual autocorrelation and heteroskedasticity in the ARIMA model suggested that its assumptions of linearity and homoscedasticity may not hold, warranting the exploration of hybrid or machine learning models. Furthermore, the non-significance of certain exogenous variables in the SARIMAX model highlighted the need for more robust predictor selection techniques, such as granger causality or feature importance analysis. Future work should also consider incorporating uncertainty quantification methods to enhance the interpretability and reliability of long-term forecasts.

Both models indicated anomalies or significant spikes in the test data, which might indicate unusual events or changes in groundwater levels that the models do not fully capture or explain. These anomalies can affect model accuracy and reliability. While the models incorporated seasonal terms, the diagnostic tests suggested that further improvements could be made to

© University of Venda

better capture seasonal patterns and trends, potentially through the integration of more advanced or hybrid models.

Although the performance metrics (RMSE, MSE, MAE, NSE and R-square) indicated that the models performed reasonably well, there is room for improvement, especially in reducing the error margins further. The SARIMAX model, for instance, showed unsatisfactory error metrics compared to the ARIMA model. The models forecasted up to 2025 from 2019, but long-term predictions are inherently uncertain. Changes in environmental conditions, policy, and other external factors can impact groundwater levels, which may not be adequately captured by the models used.

By acknowledging these limitations, the study can set the stage for future work, which might involve integrating more sophisticated modelling techniques, exploring additional variables, and addressing the identified anomalies and residual issues.

5.3 Recommendations

The insights from this study are valuable for developing effective groundwater management strategies. By comparing the findings with other studies that highlight the importance of incorporating climate variability, the following recommendations can be made to enhance model accuracy and robustness in groundwater level prediction:

1. Variable Selection: It is recommended to explore additional factors or alternative variables that might significantly impact groundwater levels. Incorporating more comprehensive climate and environmental variables can enhance model accuracy and robustness.

2. Anomaly Detection: Conducting a detailed investigation into the anomalies or significant spikes observed in the test data. Understanding these anomalies can help improve the model's ability to forecast sudden changes in groundwater levels, leading to more reliable predictions.

3. Comparative Analysis: Perform a detailed comparative analysis with other advanced forecasting models used in similar studies. This includes hybrid models or purely machine learning-based approaches to identify best practices and potential improvements in groundwater levels forecasting.

© University of Venda

4. Sophisticated Modelling Techniques: This study encourages further research into integrating more sophisticated modelling techniques, such as deep learning, for groundwater level forecasting. Advanced models like neural networks and ensemble methods can be explored to combine predictions from multiple models, resulting in more robust and accurate forecasts.
5. Geographical expansion: Model generalization can benefit from studies conducted in many geographic regions and under various meteorological circumstances.

© University of Venda

References

Adamowski, J., & Chan, H. F. (2011). A wavelet neural network conjunction model for groundwater level forecasting. *Journal of Hydrology, 407*(1–4), 28–40. https://doi.org/10.1016/j.jhydrol.2011.06.013.

Adamowski, J., Chan, H. F., Prasher, S. O., Ozga-Zielinski, B., & Sliusarieva, A. (2012). Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resources Research, 48*. https://doi.org/10.1029/2010WR009945.

Afrifa-Yamoah, E., Mueller, U. A., Taylor, S. M., & Fisher, A. J. (2020). Missing data imputation of high-resolution temporal climate time series data. *Meteorological Applications, 27*(4), e1873. https://doi.org/10.1002/met.1873.

Akbar, H., Nilsalab, P., Silalertruksa, T., & Gheewala, S. H. (2022). Comprehensive review of groundwater scarcity, stress and sustainability index-based assessment. *Groundwater for Sustainable Development, 18*, Article 100782. https://doi.org/10.1016/j.gsd.2022.100782.

Allen, M. R., & Robertson, A. W. (1996). Distinguishing modulated oscillations from coloured noise in multivariate datasets. *Climate Dynamics, 12*(11), 775–784. https://doi.org/10.1007/s003820050142.

Anderson, C. I., & Gough, W. A. (2018). Accounting for missing data in monthly temperature series: Testing rule-of-thumb omission of months with missing values. *International Journal of Climatology, 38*, 4990–5002. https://doi.org/10.1002/joc.5801.

ASCE Task Committee. (2000a). Artificial neural networks in hydrology – I: Preliminary concepts. *Journal of Hydrologic Engineering, 5*(2), 115–123.

ASCE Task Committee. (2000b). Artificial neural networks in hydrology – II: Hydrologic applications. *Journal of Hydrologic Engineering, 5*(2), 124–137.

Ashton, P. J., Love, D., Mahachi, H., & Dirks, P. H. G. M. (2001). *An overview of the impact of mining and mineral processing operations on water resources and water quality in the Zambezi, Limpopo and Olifants Catchments in Southern Africa* (Report No. ENV-PC 42). CSIR-Environmentek & University of Zimbabwe.

Asif, Z., Chen, Z., Sadiq, R., & Zhu, Y. (2023). Climate change impacts on water resources and sustainable water management strategies in North America. *Water Resources Management, 37*, 2771–2786.

Asmuth, J., Bierkens, M. F. P., & Maas, K. (2002). Transfer function-noise modelling in continuous time using predefined impulse response functions. *Water Resources Research, 38*. https://doi.org/10.1029/2001WR001136.

Assainar, K. K. R., Joseph, A., & Sajeena, S. (2023). Trend analysis of groundwater levels in Kalpathypuzha Sub-Basin of Bharathapuzha, Kerala in India. *International Journal of*

© University of Venda

Environment and Climate Change, 13(9), 1145–1157. https://doi.org/10.9734/ijecc/2023/v13i92340.

Ayenew, T., & Tilahun, N. (2008). Assessment of lake-groundwater interactions and anthropogenic stresses, using numerical groundwater flow model, for a Rift Lake catchment in central Ethiopia. *Lakes & Reservoirs: Research and Management, 13*(4), 325–343. https://doi.org/10.1111/j.1440-1770.2008.00383.x.

Bahmani, R., & Ouarda, T. B. (2021). Groundwater level modeling with hybrid artificial intelligence techniques. *Journal of Hydrology, 595*, Article 125659. https://doi.org/10.1016/j.jhydrol.2021.125659.

Bellino, J. C., Kuniansky, E. L., O'Reilly, A. M., & Dixon, J. F. (2008). *Hydrogeologic setting, conceptual groundwater flow system, and hydrologic conditions 1995–2010 in Florida and parts of Georgia, Alabama, and South Carolina* (USGS Report No. 2018-5030).

Bense, V. F., Gleeson, T., Loveless, S. E., Bour, O., & Scibek, J. (2013). Fault zone hydrogeology. *Earth-Science Reviews, 127*, 171–192. https://doi.org/10.1016/j.earscirev.2013.09.008.

Bhakar, P., & Singh, A. P. (2018). Life cycle assessment of groundwater supply system in a hyper-arid region of India. *Procedia CIRP, 69*, 603–608. https://doi.org/10.1016/j.procir.2017.11.050.

Brakenhoff, D. A., Vonk, M. A., Collenteur, R. A., Van Baar, M., & Bakker, M. (2022). Application of time series analysis to estimate drawdown from multiple well fields. *Frontiers in Earth Science, 10*, Article 907609. https://doi.org/10.3389/feart.2022.907609.

Braune, E., & Xu, Y. (2009). The role of groundwater in Sub-Saharan Africa. *Groundwater, 48*, 229–238. https://doi.org/10.1111/j.1745-6584.2009.00557.x.

Broomhead, D. S., & King, G. P. (1986). Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena, 20*(2–3), 217–236. https://doi.org/10.1016/0167-2789(86)90031-X.

Büyükaşahin, Ü. Ç., & Ertekin, Ş. (2019). Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition. *Neurocomputing, 361*, 151–163. https://doi.org/10.1016/j.neucom.2019.05.099.

Calow, R. C., MacDonald, A. M., Nicol, A. L., & Robins, N. S. (2010). Groundwater security and drought in Africa: Linking availability, access, and demand. *Groundwater, 48*, 246–256.

Candela, L., Elorza, F. J., Tamoh, K., Jiménez-Martínez, J., & Aureli, A. (2013). Groundwater modelling with limited data sets: The Chari-Logone area (Lake Chad Basin, Chad). *Hydrological Processes, 28*(11), 3714–3727. https://doi.org/10.1002/hyp.9901.

Chang, N. (2012). Seasonal autoregressive integrated moving average model for precipitation time series. *Journal of Mathematics and Statistics, 8*(4), 500–505. https://doi.org/10.3844/jmssp.2012.500.505.

© University of Venda

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*, 359–377.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition. *Journal of Official Statistics, 6*, 3–73.

Cobbing, J., & Hiller, B. (2019). Waking a sleeping giant: Realizing the potential of groundwater in Sub-Saharan Africa. *World Development, 122*, 597–613. https://doi.org/10.1016/j.worlddev.2019.06.024.

Coulibaly, P., Anctil, F., Aravena, R., & Bobee, B. (2001). Artificial neural network modeling of water table depth fluctuations. *Water Resources Research, 37*(4), 885–896.

Cuthbert, M. O., Rau, G. C., Andersen, M. S., Roshan, H., Rutlidge, H., Marjo, C. E., ... & Baker, A. (2014). Evaporative cooling of speleothem dripwater. *Scientific Reports, 4*, Article 5162. https://doi.org/10.1038/srep05162.

Cuthbert, M. O., Taylor, R. G., Favreau, G., et al. (2019). Observed controls on resilience of groundwater to climate variability in sub-Saharan Africa. *Nature, 572*, 230–234. https://doi.org/10.1038/s41586-019-1441-7.

Cuthbert, M. O., Taylor, R. G., Favreau, G., et al. (2019). Observed controls on resilience of groundwater to climate variability in sub-Saharan Africa. *Nature, 572*, 230–234. https://doi.org/10.1038/s41586-019-1441-7.

Cuthbert, R. N., Dalu, T., Wasserman, R. J., Dick, J. T. A., Callaghan, A., Froneman, P. W., & Weyl, O. L. F. (2019). Quantifying reproductive state and predator effects on copepod motility in ephemeral ecosystems. *Journal of Arid Environments, 168*, 59–61. https://doi.org/10.1016/j.jaridenv.2019.05.010.

Dale, V. H., Efroymson, R. A., & Kline, K. L. (2011). The land use–climate change–energy nexus. *Landscape Ecology, 26*, 755–773.

Daliakopoulos, I. N., Coulibaly, P., & Tsanis, I. K. (2005). Groundwater level forecasting using artificial neural networks. *Journal of Hydrology, 309*(1–4), 229–240. https://doi.org/10.1016/j.jhydrol.2004.12.001.

De Michele, G., Sello, S., Carboncini, M. C., Rossi, B., & Strambi, S. K. (2003). Cross-correlation time-frequency analysis for multiple EMG signals in Parkinson's disease: A wavelet approach. *Medical Engineering & Physics, 25*, 361–369.

Deepika, Haritash, A. K., & Gour, P. (2016). Air pollution tolerance of trees in an educational institute in Delhi. *International Journal of Environmental Sciences, 6*(6), 979–986.

© University of Venda

Department of Water and Sanitation. (2024). *Refinement of Strategic Groundwater Source Areas of South Africa – Gap analysis report* (Version 1 / Final Draft; Report No. 1047/2.1/2/2024, p. 82). Prepared by M. Misrole, K. Raible, B. Vava, & E. Wise (Umvoto South Africa Pty Ltd.) on behalf of Chief Directorate: Water Ecosystems Management. https://www.dws.gov.za.

Döll, P., & Fiedler, K. (2008). Global-scale modelling of groundwater recharge. *Hydrology and Earth System Sciences, 12*, 863–885.

Dooge, J. C. I. (1973). *Linear theory of hydrologic systems* (Technical Report 1468). U.S. Department of Agriculture, Agricultural Research Service.

Earman, S., & Dettinger, M. (2011). Potential impacts of climate change on groundwater resources—A global review. *Journal of Water and Climate Change, 2*, 213–229.

Ebrahimi, H., & Rajaee, T. (2017). Simulation of groundwater level variations using wavelet combined with neural network, linear regression and support vector machine. *Global and Planetary Change, 148*, 181–191. https://doi.org/10.1016/j.gloplacha.2016.11.014.

Edmunds, W. M. (2008). Groundwater in Africa—Palaeowater, climate change and modern recharge. In S. M. A. Adelana & A. M. MacDonald (Eds.), *Applied groundwater studies in Africa* (pp. 305–336). CRC Press/Balkema.

Edmunds, W. M., Shand, P., Hart, P., & Ward, R. S. (2003). The natural (baseline) quality of groundwater: A UK pilot study. *Science of the Total Environment, 310*(1–3), 25–35. https://doi.org/10.1016/S0048-9697(02)00620-4.

Edokpayi, J. N., Odiyo, J. O., Msagati, T. A. M., & Potgieter, N. (2015). Temporal variations in physico-chemical and microbiological characteristics of Mvudi River, South Africa. *International Journal of Environmental Research and Public Health, 12*, 4128–4140.

Elmotawakkil, A., Sadiki, A., & Enneya, N. (2024). Predicting groundwater level based on remote sensing and machine learning: A case study in the Rabat-Kénitra region. *Journal of Hydroinformatics, 26*(10), 2639–2667. https://doi.org/10.2166/hydro.2024.494.

Emamgholizadeh, S., Moslemi, K., & Karami, G. (2014). Prediction of groundwater level of Bastam Plain (Iran) by artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS). *Water Resources Management, 28*(15), 5433–5446.

Express Analytics. (2022). Everything you need to know about time series analysis – A primer. https://www.expressanalytics.com/blog/time-series-analysis/

Fahim, A. K. F., Kamal, A. S. M., & Shahid, S. (2024). Modeling spatial groundwater level patterns of Bangladesh using physio-climatic variables and machine learning algorithms. *Groundwater for Sustainable Development, 25*, Article 101142. https://doi.org/10.1016/j.gsd.2024.101142.

Feng, S., Kang, S., Huo, Z., Chen, S., & Mao, X. (2008). Neural networks to simulate regional groundwater levels affected by human activities. *Groundwater, 46*(1), 80–90.

© University of Venda

Ferdowsian, R., Pannell, D. J., McCarron, C., Ryder, A., & Crossing, L. (2001). Explaining groundwater hydrographs: Separating atypical rainfall events from time trends. *Australian Journal of Soil Research, 39*(4), 861–875.

Ferdowsian, R., Ryder, A., George, R., Bee, G., & Smart, R. (2002). Groundwater level reductions under lucerne depend on the landform and groundwater flow systems (local or intermediate). *Australian Journal of Soil Research, 40*(3), 381–396.

Foster, J. L., Carter, J. N., Lamb, G. C., Sollenberger, L. E., Blount, A. R., Myer, R. O., Maddox, M. K., & Adesogan, A. T. (2013). Performance of beef cattle creep fed concentrate, or creep grazed on warm-season legumes. *Crop Science, 53*(4), 1818–1825. https://doi.org/10.2135/cropsci2012.11.0648.

Gaye, C. B., & Tindimugaya, C. (2018). Review: Challenges and opportunities for sustainable groundwater management in Africa. *Hydrogeology Journal, 27*, 1–12. https://doi.org/10.1007/s10040-018-1892-1.

Gholizadeh, H., Behrouj Peely, A., Karney, B. W., & Malekpour, A. (2020). Assessment of groundwater ingress to a partially pressurized water-conveyance tunnel using a conduit-flow process model: A case study in Iran. *Hydrogeology Journal, 28*, 2573–2585.

Gholizadeh, H., Zhang, Y., Frame, J., Gu, X., & Green, C. T. (2023). Long short-term memory models to quantify long-term evolution of streamflow discharge and groundwater depth in Alabama. *Science of the Total Environment, 901*, Article 165884.

Golyandina, N. (2010). On the choice of parameters in singular spectrum analysis and related subspace-based methods. *Statistics and Its Interface, 3*(3), 259–279. https://doi.org/10.4310/SII.2010.v3.n3.a2.

Golyandina, N., & Korobeynikov, A. (2014). Basic singular spectrum analysis and forecasting with R. *Computational Statistics & Data Analysis, 71*, 934–954. https://doi.org/10.1016/j.csda.2013.04.009.

Golyandina, N., & Zhigljavsky, A. A. (2013). *Singular spectrum analysis for time series*. Springer.

Golyandina, N., Nekrutkin, V. V., & Zhigljavsky, A. A. (2001). *Analysis of time series structure: SSA and related techniques*. Chapman & Hall/CRC.

Goodarzi, M. (2020). Application and performance evaluation of time series, neural networks and Hartt models in predicting groundwater level changes, Najafabad Plain, Iran. *Sustainable Water Resources Management, 6*(4). https://doi.org/10.1007/s40899-020-00427-2.

Görgens, A. H. M., & Boroto, J. (2003). Limpopo River: An overview of alternative methods for estimating transmission losses. In *Hydrology of Mediterranean and Semiarid Regions: Proceedings of the Montpellier Conference* (IAHS Publication No. 278).

Green, T. R., Taniguchi, M., Kooi, H., Gurdak, J. J., Allen, D. M., Hiscock, K. M., & Aureli, A. (2011). Beneath the surface of global change: Impacts of climate change on groundwater. *Journal of Hydrology, 405*, 532–560.

© University of Venda

Gropius, M., Dahabiyeh, M., Al Hyari, M., Brückner, F., Lindenmaier, F., & Vassolo, S. (2022). Estimation of unrecorded groundwater abstractions in Jordan through regional groundwater modelling. *Hydrogeology Journal*.

Guobin Fu, G., Crosbie, R. S., Barron, O., Charles, S. P., Dawes, W., Shi, X., Van Niel, T., & Li, C. (2019). Attributing variations of temporal and spatial groundwater recharge: A statistical analysis of climatic and non-climatic factors. *Journal of Hydrology, 568*, 816–834. https://doi.org/10.1016/j.jhydrol.2018.11.022.

Gyamfi, C., Ndambuki, J. M., Diabene, P. Y., Kifanyi, G. E., Githuku, C. R., & Alexander, A. (2016). Using GIS for spatial exploratory analysis of borehole data: A firsthand approach towards groundwater development. *Journal of Science and Technology (Ghana), 36*(1), 38. https://doi.org/10.4314/just.v36i1.7.

Hanson, R., Dettinger, M., & Newhouse, M. (2006). Relations between climatic variability and hydrologic time series from four alluvial basins across the southwestern United States. *Hydrogeology Journal, 14*(7), 1122–1146.

Harris, T. J., & Yuan, H. (2010). Filtering and frequency interpretations of singular spectrum analysis. *Physica D, 239*, 1958–1967.

Hassani, H. (2007). Singular spectrum analysis: Methodology and comparison. *Journal of Data Science, 5*, 239–257.

Hassani, H., & Mahmoudvand, R. (2013). Multivariate singular spectrum analysis: A general view and new vector forecasting approach. *International Journal of Energy Statistics, 1*, 55–83.

Hassani, H., & Zhigljavsky, A. (2009). Singular spectrum analysis: Methodology and application to economics data. *Journal of Systems Science and Complexity, 22*, 372–394.

Healy, R. W., & Cook, P. G. (2002). Using groundwater levels to estimate recharge. *Hydrogeology Journal, 10*, 91–109.

Holland, M. (2011). *Hydrogeological characterisation of crystalline basement aquifers of the Limpopo province, South Africa* (master's thesis). University of Pretoria.

Hoque, M. A., Apon, A. A., Hassan, M. A., Adhikary, S. K., & Islam, M. A. (2025). Enhanced forecasting of groundwater level incorporating an exogenous variable: Evaluating conventional multivariate time series and artificial neural network models. *Geographies, 5*(1), Article 1. https://doi.org/10.3390/geographies5010001.

Hou, Q., Pei, T., Yu, X., Chen, Y., Ji, Z., & Xie, B. (2022). The seasonal response of vegetation water use efficiency to temperature and precipitation in the Loess Plateau, China. *GECCO, 33*, Article 01984.

Hubbard, S. S., Chen, J., Peterson, J., Majer, E. L., Williams, K. H., Swift, D. J., Hubbard, B., & Rubin, Y. (2001). Hydrogeological characterization of the South Oyster Bacterial Transport Site using geophysical data. *Water Resources Research, 37*, 2431–2456.

© University of Venda

Hughes, D. (2008). Modelling semi-arid and arid hydrology and water resources: The Southern African experience. In *Hydrological modelling in arid and semi-arid areas*. https://doi.org/10.1017/CBO9780511535734.004.

Huisheng, W., Yuejiang, L., & Zhao, H. (2022). Caseload prediction using graphical evolutionary game theory and time series analysis. *International Journal of Crowd Science, 6*, 142–149. https://doi.org/10.26599/IJCS.2022.9100019.

Humphrey, V., Gudmundsson, L., & Seneviratne, S. I. (2016). Assessing global water storage variability from GRACE: Trends, seasonal cycle, sub-seasonal anomalies and extremes. *Surveys in Geophysics, 37*, 357–395.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers, 116*, 770–799.

Ihlen, E. A. (2012). Introduction to multifractal detrended fluctuation analysis in Matlab. *Frontiers in Physiology, 3*, Article 141.

Institute of Soil, Water and Climate. (2003). *Land types of the maps 2228 Alldays and 2230 Messina* (Memoirs of the Agricultural and Natural Resources of South Africa, No. 37). ARC-Institute of Soil, Water and Climate.

Irvine-Fynn, T. D. L., Moorman, B. J., Williams, J. L. M., & Walter, F. S. A. (2006). Seasonal changes in ground-penetrating radar signature observed at a polythermal glacier, Bylot Island, Canada. *Earth Surface Processes and Landforms, 31*, 892–909.

Jain, G., & Mallick, B. (2017). A study of time series models ARIMA and ETS. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2898968.

Jalalkamali, A., Sedghi, H., & Manshouri, M. (2011). Monthly groundwater level prediction using ANN and neuro-fuzzy models: A case study on Kerman plain, Iran. *Journal of Hydroinformatics, 13*(4), 867–876.

Jeffrey, S. J., Carter, J. O., Moodie, K. B., & Beswick, A. R. (2001). Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling & Software, 16*.

Johnston, R., & Smakhtin, V. (2014). Hydrological modeling of large river basins: How much is enough? *Water Resources Management, 28*(10), 2695–2730. https://doi.org/10.1007/s11269-014-0637-8.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). Springer.

Jones, L. E., Painter, J. A., LaFontaine, J. H., Sepúlveda, N., & Sifuentes, D. F. (2008). *Groundwater-flow budget for the Lower Apalachicola-Chattahoochee-Flint River Basin in southwestern Georgia and parts of Florida and Alabama* (USGS Report No. 2017-5141).

Jose, J. (2022). *Introduction to time series analysis and its applications*.

Jury, W. A., & Roth, K. (1990). *Transfer functions and solute movement through soil: Theory and applications*. Birkhäuser Boston.

© University of Venda

Kajewska-Szkudlarek, J., & Stańczyk, J. (2018). Filling missing meteorological data with computational intelligence methods. In *XLVIII Seminar of Applied Mathematics* (ITM Web of Conferences, Vol. 23, Article 00015). EDP Sciences. https://doi.org/10.1051/itmconf/20182300015.

Kantelhardt, J. W. (2008). Fractal and multifractal time series. *arXiv*. https://arxiv.org/abs/0804.0747.

Kholghi, M., & Hosseini, S. M. (2009). Comparison of groundwater level estimation using neuro-fuzzy and ordinary kriging. *Environmental Modelling & Assessment, 14*(6), 729–737.

Konikow, L. F., & Kendy, E. (2005). Groundwater depletion: A global problem. *Hydrogeology Journal, 13*, 317–320. https://doi.org/10.1007/s10040-004-0411-8.

Krishan, G., Patidar, N., Sudarsan, N., Vasisth, R., & Sidhu, B. S. (2023). Groundwater responses to climate variability in Punjab, India. In *Climate change and environmental impacts: Past, present and future perspective* (pp. 333–343). Springer International Publishing.

Kumar, C. P. (2012). Climate change and its impact on groundwater resources. *International Journal of Engineering Science, 1*, 43–60.

Kundzewicz, Z. W., & Döll, P. (2009). Will groundwater ease freshwater stress under climate change? *Hydrological Sciences Journal, 54*(4), 665–675. https://doi.org/10.1623/hysj.54.4.665

Kwarteng, S. B., & Andreevich, P. A. (2024). Comparative analysis of ARIMA, SARIMA and Prophet model in forecasting. *Research & Development, 5*(4), 110–120. https://doi.org/10.11648/j.rd.20240504.13.

Lallahem, S., Mania, J., Hani, A., & Najjar, Y. (2005). On the use of neural networks to evaluate groundwater levels in fractured media. *Journal of Hydrology, 307*, 92–111.

Le Page, M., Berjamy, B., Fakir, Y., Bourgin, F., Jarlan, L., Abourida, A., ... & Chehbouni, G. (2012). An integrated DSS for groundwater management based on remote sensing: The case of a semi-arid aquifer in Morocco. *Water Resources Management, 26*(11), 3209–3230. https://doi.org/10.1007/s11269-012-0068-3.

Lee, L., Lawrence, D., & Price, M. (2006). Analysis of water-level response to rainfall and implications for recharge pathways in the Chalk aquifer, SE England. *Journal of Hydrology, 330*(3–4), 604–620.

Lepot, M., Aubin, J.-B., & Clemens, F. H. L. R. (2017). Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. *Water, 9*(10), 796. https://doi.org/10.3390/w9100796.

Li, B., Rodell, M., Kumar, S., Beaudoing, H. K., Getirana, A., Zaitchik, B., ... & Mukherjee, A. (2019). Global GRACE data assimilation for groundwater and drought monitoring: Advances and challenges. *Water Resources Research, 55*, 7564–7586.

Li, Y., Yan, Y., Deng, S., Zhang, C., Haq, F., Chen, T., ... & Chen, G. (2020). Hydrological time series analysis using ARIMA models. *Water Resources Research*.

© University of Venda

Lorenzo-Lacruz, J., Garcia, C., & Morán-Tejeda, E. (2017). Groundwater level responses to precipitation variability in Mediterranean insular aquifers. *Journal of Hydrology, 552*, 516–531. https://doi.org/10.1016/j.jhydrol.2017.07.011.

Maas, C. (1994). *On convolutional processes and dispersive groundwater flow* (Doctoral dissertation). Delft University of Technology.

MacDonald, A. M., Bonsor, H. C., Ó Dochartaigh, B. É., & Taylor, R. G. (2012). Quantitative maps of groundwater resources in Africa. *Environmental Research Letters, 7*, 024009. https://doi.org/10.1088/1748-9326/7/2/024009.

Macdonald, J. C., Isom, D. C., Evans, D. D., & Page, K. J. (2021). Digital innovation in medicinal product regulatory submission, review, and approvals to create a dynamic regulatory ecosystem—Are we ready for a revolution? *Frontiers in Medicine, 8*, 660808. https://doi.org/10.3389/fmed.2021.660808.

Maheswaran, R., & Khosa, R. (2013). Long-term forecasting of groundwater levels with evidence of non-stationary and nonlinear characteristics. *Computers & Geosciences, 52*, 422–436. https://doi.org/10.1016/j.cageo.2012.09.030.

Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. *Environmental Modelling & Software, 15*, 101–124.

Mair, A., & Fares, A. (2011). Time series analysis of daily rainfall and streamflow in a volcanic dike-intruded aquifer system, O‘ahu, Hawai‘i, USA. *Hydrogeology Journal, 19*(4), 929–944.

Makungo, R., & Odiyo, J. O. (2017). Estimating groundwater levels using system identification models in Nzhelele and Luvuvhu areas, Limpopo Province, South Africa. *Physics and Chemistry of the Earth*. https://doi.org/10.1016/j.pce.2017.01.019.

Manzione, R. (2018). Physical-based time series model applied on water table depths dynamics characteristics simulation. *RBRH, 23*(0).

Maponya, P. (2021). Opportunities and constraints faced by smallholder farmers in the Vhembe District, Limpopo Province in South Africa. *Circular Economy and Sustainability*. https://doi.org/10.1007/s43615-021-00028-x.

Martin, H. (2011). *Hydrogeological characterisation of crystalline basement aquifers within the Limpopo Province, South Africa* (Doctoral dissertation). University of Pretoria.

Martirosyan, A. V., Ilyushin, Y. V., & Afanaseva, O. V. (2022). Development of a distributed mathematical model and control system for reducing pollution risk in mineral water aquifer systems. *Water, 14*, 151.

Martirosyan, A. V., Kukharova, T. V., & Fedorov, M. S. (2021). Research of the hydrogeological objects’ connection peculiarities. In *Proceedings of the 2021 IV International Conference on Control in Technical Systems (CTS)* (pp. 34–38). Saint Petersburg, Russia.

© University of Venda

Meixner, T., Manning, A. H., Stonestrom, D. A., Allen, D. M., Ajami, H., & Blasch, K. W. (2016). Implications of projected climate change for groundwater recharge in the western United States. *Journal of Hydrology, 534*, 124–138.

Melanie. (2024). SARIMAX model: What is it? How can it be applied to time series? *Data Science Courses | DataScientest*. https://datascientest.com/en/sarimax-model-what-is-it-how-can-it-be-applied-to-time-series

Minitab 18 Support. (2018). Sen's slope. https://support.minitab.com/en-us/minitab/18/macro-library/macro-files/nonparametrics-macros/senslope/

Mirzavand, M., & Ghazavi, R. (2014). A stochastic modelling technique for groundwater level forecasting in an arid environment using time series methods. *Water Resources Management, 29*(4), 1315–1328. https://doi.org/10.1007/s11269-014-0875-9.

Mirzavand, M., & Ghazavi, R. (2015). A stochastic modelling technique for groundwater level forecasting in an arid environment using time series methods. *Water Resources Management, 29*, 1315–1328. https://doi.org/10.1007/s11269-014-0875-9.

Mishra, N., & Srivastava, P. (2015). What do climate projections say about future droughts in Alabama? In *Proceedings of the ASABE 1st Climate Change Symposium: Adaptation and Mitigation Conference Proceedings* (pp. 1–2). Chicago, IL, USA.

Mohammadi, K. (2008). Groundwater table estimation using MODFLOW and artificial neural networks. In R. J. Abrahart, L. M. See, & D. P. Solomatine (Eds.), *Practical hydroinformatics* (pp. 127–138). Water Science and Technology Library.

Mohanasundaram, S., Kumar, G. S., & Narasimhan, B. (2019). A novel deseasonalized time series model with an improved seasonal estimate for groundwater level predictions. *H2Open Journal, 2*(1), 25–44. https://doi.org/10.2166/h2oj.2019.022.

Mohanty, S., Jha, K., Kumar, A., & Sudheer, K. (2010). Artificial neural network modeling for groundwater level forecasting in a river island of eastern India. *Water Resources Management, 24*, 1845–1865.

Mpenyana, M. L., & Momba, M. N. B. (2012). Assessment of groundwater quality in the rural areas of the Northwest Province, South Africa. *Scientific Research and Essays, 7*(8), 903–914.

Myung, N. K. (2009). *Singular spectrum analysis* (master's thesis). University of California, Los Angeles, CA.

Naga Rajesh, A., Abinaya, S., Purna Durga, G., & Lakshmi Kumar, T. V. (2023). Long-term relationships of MODIS NDVI with rainfall, land surface temperature, surface soil moisture and groundwater storage over monsoon core region of India. *Arid Land Research and Management, 37*, 51–70.

Nayak, P., SatyajiRao, Y., & Sudheer, K. (2006). Groundwater level forecasting in a shallow aquifer using artificial neural network. *Water Resources Management, 20*, 77–90.

© University of Venda

Nayak, P., Sudheer, K. P., Rangan, D. M., & Ramasastri, K. S. (2004). A neuro fuzzy computing technique for modeling hydrological time series. *Journal of Hydrology, 291*, 52–66.

Ndiritu, J., Odiyo, J. O., Makungo, R., Ntuli, C., & Mwaka, B. (2011). Yield-reliability analysis for rural domestic water supply from combined rainwater harvesting and run-of-river abstraction. *Hydrological Sciences Journal, 56*(2), 238–248.

Nieuwoudt, W. L., Backeberg, G. R., & Du Plessis, H. M. (2004). The value of water in the South African economy: Some implications. *Agrekon, 43*. https://doi.org/10.1080/03031853.2004.9523643.

Nourani, V., & Mousavi, S. (2016). Spatiotemporal groundwater level modeling using hybrid artificial intelligence-meshless method. *Journal of Hydrology, 536*, 10–25. https://doi.org/10.1016/j.jhydrol.2016.02.030.

Obergfell, C., Bakker, M., Zaadnoordijk, W., & Maas, K. (2013). Deriving hydrogeological parameters through time series analysis of groundwater head fluctuations around well fields. *Hydrogeology Journal, 21*(5), 987–999.

Osman, I. A., et al. (2021). Extreme gradient boosting (XGBoost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal, 12*(2), 1545–1556. https://doi.org/10.1016/j.asej.2020.11.011.

Owor, M., Taylor, R. G., Tindimugaya, C., & Mwesigwa, D. (2009). Rainfall intensity and groundwater recharge: Empirical evidence from the Upper Nile Basin. *Environmental Research Letters, 4*(3), 035009. https://doi.org/10.1088/1748-9326/4/3/035009.

Pal, N. R., & Bezdek, J. C. (1995). On cluster validity for fuzzy c-means model. *IEEE Transactions on Fuzzy Systems, 1*, 370–379.

Pankratz, A. (1991). *Forecasting with dynamic regression models*. John Wiley & Sons.

Park, Y. C., Jo, Y. J., & Lee, J. Y. (2011). Trends of groundwater data from the Korean National Groundwater Monitoring Stations: Indication of any change? *Geosciences Journal, 15*, 105–114.

Patle, G., Singh, D., Sarangi, A., Rai, A., Khanna, M., & Sahoo, R. (2015). Time series analysis of groundwater levels and projection of future trend. *Journal of the Geological Society of India, 85*(2), 232–242.

Peng, C. K., Havlin, S., Stanley, H. E., & Goldberger, A. L. (1995). Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Journal of Nonlinear Science, 5*, 82–87.

Pershin, I. M., Papush, E. G., Kukharova, T. V., & Utkin, V. A. (2023). Modelling of distributed control system for network of mineral water wells. *Water, 15*, 2289.

Peterson, T. J., & Western, A. W. (2011). Time-series modelling of groundwater head and its decomposition to historic climate periods. In *34th IAHR Congress 2011 – Balance and*

© University of Venda

Uncertainty: Water in a Changing World (pp. 1677–1684). International Association of Hydraulic Engineering and Research (IAHR).

Pezij, M., Augustijn, D. C. M., Hendriks, D. M. D., & Hulscher, S. J. M. H. (2020). Applying transfer function-noise modelling to characterize soil moisture dynamics: A data-driven approach using remote sensing data. *Environmental Modelling & Software, 131*, 104756.

Polomčić, D., Gligorić, Z., Bajić, D., & Cvijović, Č. (2017). A hybrid model for forecasting groundwater levels based on fuzzy C-mean clustering and singular spectrum analysis. *Water, 9*(7), 541.

Ponprasit, C., Gu, X. F., Goodliffe, A., & Sun, H. G. (2003). Assessing vulnerability of regional-scale aquifer-aquitard systems in East Gulf Coastal Plain of Alabama by developing groundwater flow and transport models. *Water, 15*, 1937.

Pourmorad, S., Kabolizade, M., & Dimuccio, L. A. (2024). Artificial intelligence advancements for accurate groundwater level modelling: An updated synthesis and review. *Applied Sciences, 14*(16), 7358. https://doi.org/10.3390/app14167358.

Poursaeid, M., Poursaeid, A. H., & Shabanlou, S. (2022). A comparative study of artificial intelligence models and a statistical method for groundwater level prediction. *Water Resources Management, 36*, 1499–1519.

Quimpo, R. G. (1971). Structural relation between parametric and stochastic hydrology models. *IAHS Publication, 100*, 151–157.

Rajaee, T., & Boroumand, A. (2015). Forecasting of chlorophyll-a concentrations in south San Francisco Bay using five different models. *Applied Ocean Research, 53*, 208–217.

Rajaee, T., Ebrahimi, H., & Nourani, V. (2019). A review of the artificial intelligence methods in groundwater level modeling. *Journal of Hydrology, 572*, 336–351. https://doi.org/10.1016/j.jhydrol.2018.12.037.

Rangelova, E., Sideris, M. G., & Kim, J. W. (2012). On the capabilities of the multi-channel singular spectrum method for extracting the main periodic and non-periodic variability from weekly GRACE data. *Journal of Geodynamics, 54*, 64–78. https://doi.org/10.1016/j.jog.2011.10.006

Rashid, A., Alamgir, M., Ahmed, M. T., et al. (2022). Assessing and forecasting of groundwater level fluctuation in Joypurhat district, northwest Bangladesh, using wavelet analysis and ARIMA modeling. *Theoretical and Applied Climatology, 150*, 327–345. https://doi.org/10.1007/s00704-022-04160-y.

Rehal, V. (2022). Interpreting ACF and PACF plots. *SPUR Economics – Learn and Excel*. https://spureconomics.com/interpreting-acf-and-pacf-plots/

Richard, A., Fine, L., Rozenstein, O., Tanny, J., Geist, M., & Pradalier, C. (2020, September). Filling gaps in micro-meteorological data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 101–117). Springer International Publishing.

© University of Venda

Rijsberman, F. (2006). Water scarcity: Fact or fiction? *Agricultural Water Management, 80*, 5–22. https://doi.org/10.1016/j.agwat.2005.07.001.

Rödiger, T., Geyer, S., Mallast, U., Merz, R., Krause, P., Fischer, C., & Siebert, C. (2014). Multi-response calibration of a conceptual hydrological model in the semiarid catchment of Wadi al Arab, Jordan. *Journal of Hydrology, 509*, 193–206. https://doi.org/10.1016/j.jhydrol.2013.11.026.

Ruman, S., Krpec, P., Rusnok, P., Black, A. R., Trizna, M., & Ball, T. (2020). Impact of missing precipitation values on hydrological model output. *Acta Geophysica, 68*(2), 539–553. https://doi.org/10.1007/s11600-020-00409-0.

Sabyasachi Swain, S., Sahoo, S., Taloor, A. K., Mishra, S. K., & Pandey, A. (2022). Exploring recent groundwater level changes using innovative trend analysis (ITA) technique over three districts of Jharkhand, India. *Groundwater for Sustainable Development, 18*, 100783. https://doi.org/10.1016/j.gsd.2022.100783.

Sahoo, S., & Jha, M. K. (2015). On the statistical forecasting of groundwater levels in unconfined aquifer systems. *Environmental Earth Sciences, 73*, 3119–3136.

Samsudin, R., Saad, P., & Shabri, A. (2011). River flow time series using least squares support vector machines. *Hydrology and Earth System Sciences, 15*, 1835–1852.

Scanlon, B. R., Keese, K. E., Flint, A. L., Flint, L. E., Gaye, C. B., Edmunds, W. M., et al. (2006). Global synthesis of groundwater recharge in semiarid and arid regions. *Hydrological Processes, 20*(15), 3335–3370. https://doi.org/10.1002/hyp.6335.

Schuster, G. A., Taylor, C. A., & McGregor, S. W. (2022). *Crayfishes of Alabama* (pp. 8–15). University of Alabama Press.

Sen, P. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association, 39*, 1379–1389.

Shah, T. (2009). Climate change and groundwater: India's opportunities for mitigation and adaptation. *Environmental Research Letters, 4*, 035005.

Shah, V., Patel, N., Shah, D., Swain, D., Mohanty, M., Acharya, B., Gerogiannis, V. C., & Kanavos, A. (2024). Forecasting maximum temperature trends with SARIMAX: A case study from Ahmedabad, India. *Sustainability, 16*(16), 7183. https://doi.org/10.3390/su16167183.

Shah, V., Patel, N., Shah, D., Swain, D., Mohanty, M., Acharya, B., Gerogiannis, V. C., & Kanavos, A. (2024). Forecasting maximum temperature trends with SARIMAX: A case study from Ahmedabad, India. *Sustainability, 16*(16), 7183. https://doi.org/10.3390/su16167183.

Shamsudduha, M. (2009). Recent trends in groundwater levels in a highly seasonal hydrological system: The Ganges-Brahmaputra-Meghna delta. *Hydrology and Earth System Sciences, 13*(12), 2373–2385. https://doi.org/10.5194/hess-13-2373-2009.

© University of Venda

Shapoori, V., Peterson, T., Western, A., & Costelloe, J. (2015). Top-down groundwater hydrograph time-series modeling for climate-pumping decomposition. *Hydrogeology Journal, 23*(4), 819–836.

Shin, M. J., Moon, S. H., Kang, K. G., Moon, D. C., & Koh, H. J. (2020). Analysis of groundwater level variations caused by the changes in groundwater withdrawals using long short-term memory network. *Hydrology, 7*(3), 64. https://doi.org/10.3390/hydrology7030064.

Shouraseni Sen Roy, S., Rahman, A., Ahmed, S., Shahfahad, S., & Ahmad, I. A. (2022). Long-term trends of groundwater level variations in response to local level land use land cover changes in Mumbai, India. *Groundwater for Sustainable Development, 18*, 100797. https://doi.org/10.1016/j.gsd.2022.100797.

Shuai, G., Zhou, Y., Shao, J., Cui, Y., Zhang, Q., Jin, C., & Xu, S. (2024). Comparison of multiple machine learning methods for correcting groundwater levels predicted by physics-based models. *Sustainability, 16*(2), 653. https://doi.org/10.3390/su16020653.

Sibanda, S., Grab, S. W., & Ahmed, F. (2020). Long-term rainfall characteristics in the Mzingwane catchment of south-western Zimbabwe. *Theoretical and Applied Climatology, 139*, 935–948.

Singh, A. (2021). Trend and seasonal analysis of groundwater levels in the Indo-Gangetic Plain. *Journal of Hydrology*.

Song, Q., & Chissom, B. S. (1993). Fuzzy time series and its models. *Fuzzy Sets and Systems, 54*, 269–277.

Sood, A., & Smakhtin, V. (2015). Global hydrological models: A review. *Hydrological Sciences Journal, 60*, 549–565. https://doi.org/10.1080/02626667.2014.950580.

Sterckx, A., & Ramusiya, F. (2020). An #OpenWaterData success story. *Information & Knowledge Management in Southern Africa*. https://www.un-igrac.org/stories/national-groundwater-archive-south-africa

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

Szostek, K., Mazur, D., Drałus, G., & Kusznier, J. (2024). Analysis of the effectiveness of ARIMA, SARIMA, and SVR models in time series forecasting: A case study of wind farm energy production. *Energies, 17*(19), 4803. https://doi.org/10.3390/en17194803.

Taie Semiromi, M., & Koch, M. (2019). Reconstruction of groundwater levels to impute missing values using singular and multichannel spectrum analysis: Application to the Ardabil Plain, Iran. *Hydrological Sciences Journal*.

Tamiru Abiye, T., Masindi, K., Mengistu, H., & Demlie, M. (2018). Understanding the groundwater-level fluctuations for better management of groundwater resource: A case in the Johannesburg region. *Groundwater for Sustainable Development, 7*, 1–7. https://doi.org/10.1016/j.gsd.2018.02.004.

© University of Venda

Tao, H., Hameed, M. M., Marhoon, H. A., Zounemat-Kermani, M., Heddam, S., Kim, S., ... & Yaseen, Z. M. (2022). Groundwater level prediction using machine learning models: A comprehensive review. *Neurocomputing, 489*, 271–308. https://doi.org/10.1016/j.neucom.2022.03.014.

Taylor, R. G., Scanlon, B., Döll, P., Rodell, M., Van Beek, R., & Wada, Y. (2013). Ground water and climate change. *Nature Climate Change, 3*, 322–329.

Teegavarapu, R. S. G., & Nayak, A. (2017). Evaluation of long-term trends in extreme precipitation: Implications of in-filled historical data use for analysis. *Journal of Hydrology, 550*, 616–634. https://doi.org/10.1016/j.jhydrol.2017.05.030.

Tran, A. P., Dafflon, B., Hubbard, S. S., Kowalsky, M. B., Long, P., Tokunaga, T. K., & Williams, K. H. (2016). Quantifying shallow subsurface water and heat dynamics using coupled hydrological-thermal-geophysical inversion. *Hydrology and Earth System Sciences, 20*, 3477–3491.

Trichakis, I. C., Nikolos, I. K., & Karatzas, G. P. (2010). Artificial neural network (ANN) based modeling for karstic groundwater level simulation. *Water Resources Management, 25*(4), 1143–1152. https://doi.org/10.1007/s11269-010-9628-6.

Tubau, I., Vázquez-Suñé, E., Carrera, J., Valhondo, C., & Criollo, R. (2017). Quantification of groundwater recharge in urban environments. *Science of the Total Environment, 592*, 391–402.

Varni, M., Comas, R., Weinzettel, P., & Dietrich, S. (2013). Application of the water table fluctuation method to characterize groundwater recharge in the Pampa plain, Argentina. *Hydrological Sciences Journal, 58*(7), 1445–1455. https://doi.org/10.1080/02626667.2013.833663.

Velicer, W. F., & Molenaar, P. C. (2012). Time series analysis for psychological research. In *Handbook of Psychology* (2nd ed.). https://doi.org/10.1002/9781118133880.hop202022.

Vhembe District Municipality. (2009). *Integrated development plan, 2009/10 review*.

Vhembe District Municipality. (2010). *Integrated development plan documents*. Thohoyandou, South Africa.

Vhembe District Municipality. (2012–2017). *Integrated development plan*. Thohoyandou, South Africa.

Von Asmuth, J. R., Bierkens, M. F. P., & Maas, K. (2002). Transfer function-noise modelling in continuous time using predefined impulse response functions. *Water Resources Research, 38*, 23. https://doi.org/10.1029/2001WR001136.

Von Asmuth, J. R., et al. (2012). Software for hydrogeologic time series analysis, interfacing data with physical insight. *Environmental Modelling & Software, 38*, 178–190. https://doi.org/10.1016/j.envsoft.2012.06.003.

© University of Venda

Von Asmuth, J. R., Maas, K., Bakker, M., & Petersen, J. (2008). Modelling time series of groundwater head fluctuations subjected to multiple stresses. *Ground Water, 46*, 30–40. https://doi.org/10.1111/j.1745-6584.2007.00382.x.

Wada, Y., Wisser, D., & Bierkens, M. F. P. (2014). Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. *Earth System Dynamics, 5*, 15–40. https://doi.org/10.5194/esd-5-15-2014.

Wang, X., Tian, W., & Liao, Z. (2021). Statistical comparison between SARIMA and ANN's performance for surface water quality time series prediction. *Environmental Science and Pollution Research, 28*(25), 1–14. https://doi.org/10.1007/s11356-021-13086-3.

Wilcox, R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. Springer Science and Business Media.

Wu, W. Y., Lo, M. H., Wada, Y., Famiglietti, J. S., Reager, J. T., Yeh, P. J. F., Ducharne, A., & Yang, Z. L. (2020). Divergent effects of climate change on future groundwater availability in key mid-latitude aquifers. *Nature Communications, 11*(1), 3710. https://doi.org/10.1038/s41467-020-17581-y.

Xu, Y., Gong, H., Chen, B., Zhang, Q., & Li, Z. (2021). Long-term and seasonal variation in groundwater storage in the North China Plain based on GRACE. *International Journal of Applied Earth Observation and Geoinformation, 104*, 102560.

Yan, Q., & Ma, C. (2016). Application of integrated ARIMA and RBF network for groundwater level forecasting. *Environmental Earth Sciences, 75*, 396. https://doi.org/10.1007/s12665-015-5198-5.

Yang, Q., Wang, Y., & Zhang, J. (2017). A comparative study of shallow groundwater level simulation with three time series models in a coastal aquifer of South China. *Applied Water Science, 7*, 689–698. https://doi.org/10.1007/s13201-015-0282-2.

Yang, Z. P., Lu, W. X., Long, Y. Q., & Li, P. (2009). Application and comparison of two prediction models for groundwater levels: A case study in Western Jilin Province, China. *Journal of Arid Environments, 73*, 487–492.

Yasarer, L. M., Taylor, J. M., Rigby, J. R., & Locke, M. A. (2020). Trends in land use, irrigation, and streamflow alteration in the Mississippi River Alluvial Plain. *Frontiers in Environmental Science, 8*, 66.

Yoon, H., Jun, S. C., Hyun, Y., Bae, G. O., & Lee, K. K. (2010). A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. *Journal of Hydrology, 396*, 128–138.

Zakwan, M. (2021). *Groundwater resources development and planning in the semi-arid region*. Springer. ISBN: 978-3-030-68123-4.

Zeru, G., Alamirew, T., Shishaye, H., Olmana, M., Tafesse, N., & Reading, M. (2020). Groundwater level trend analysis using the statistical auto-regressive HARTT method. *Hydrological Research Letters, 14*, 17–22. https://doi.org/10.3178/hrl.14.17.

© University of Venda

Zotov, L., Sidorenkov, N. S., Bizouard, C., Shum, C. K., & Shen, W. (2017). Multichannel singular spectrum analysis of the axial atmospheric angular momentum. *Geodesy and Geodynamics, 8*(6), 433–442.

© University of Venda