FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF WATER RESOURCES ENGINEERING

A HYBRID MODFLOW-MACHINE LEARNING APPROACH FOR GROUNDWATER PROSPECTING AND OPTIMAL BOREHOLE SITING

CASE STUDY: AWOJA CATCHMENT

BY

OUMA HASSAN MASINDE

BU/UP/2022/1709

SUPERVISOR: MR. MUGISHA MOSES

This final year project report is submitted to the Department of Water Resources Engineering as a partial fulfilment of the requirements for the award of Bachelors degree of Science in Water Resources Engineering.

MAY 2026

ABSTRACT

This report presents the role of a hybrid MODFLOW-Machine Learning (ML) approach in predicting groundwater potential and optimizing borehole siting in the Awoja Catchment in eastern Uganda. Groundwater remains the primary source of safe water for rural communities in Uganda; however, its exploration and development are often hindered by limited hydrogeological data and the use of conventional methods that inadequately capture subsurface variability. The study integrates the physically based MODFLOW groundwater flow model with data-driven ML algorithms, including Random Forest, Support Vector Machine, Artificial Neural Networks, and XGBoost to form a hybrid framework capable of enhancing groundwater prospecting accuracy. Hydrogeological, geological, soil, and climatic datasets will be utilized to simulate aquifer conditions, generate hydraulic parameters, and train predictive models. Model performance will be evaluated using statistical indicators such as the Coefficient of Determination (R^2), Root Mean Square Error (RMSE), and Nash-Sutcliffe Efficiency (NSE). The hybrid model demonstrates superior predictive accuracy compared to standalone ML approaches, effectively delineating high-yielding aquifer zones and identifying optimal borehole locations. The findings provide a framework that supports efficient groundwater exploration, minimizes drilling failure rates, and promotes sustainable groundwater management. This study directly contributes to Sustainable Development Goal 6 (SDG 6) on clean water and sanitation and aligns with Uganda's Third National Development Plan (NDPIII), which prioritizes reliable and sustainable groundwater utilization for socio-economic development.

DECLARATION

I...OUMA HASSAN MASINDE.............................declare that this research is my original work and has not been submitted for the award of any degree in any other university or institution of higher learning.

Date :...04TH JUNE, 2026...

Signature :...[signature]...

Email :...hassanouma0@gmail.com........................

Contact :...0702560583...

ACKNOWLEDGEMENT

First and foremost, I thank the Almighty God for the gift of life, strength, and wisdom that has enabled me to carry out this research. I extend my heartfelt appreciation to my supervisor for his invaluable guidance, mentorship, and constructive feedback throughout the research process.

DEDICATION

I dedicate this report to my parents, whose unwavering love, support, and encouragement have been the foundation of my academic journey.

APPROVAL

I OUMA HASSAN MASINDE submit this report of my final year project to the faculty of Engineering and Technology in the Department of Water Resources Engineering with the approval of my supervisor

Supervisor's Name: ……Moses Nhyala.……

Signature: ……………………………………

Date: ……03/06/26……

TABLE OF CONTENTS

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ANN	Artificial Neural Network
AUC	Area under the Curve
AUC-ROC	Area under the Receiver Operating Characteristic Curve
BGS	British Geology Survey
CRS	Coordinate Reference System
DEM	Digital Elevation Model
DGSM	Directorate of Geological Survey and Mines
DWRM	Directorate of Water Resources Management
EM	Electromagnetic Method
ET	Evapotranspiration
FDEM	Frequency Domain Electromagnetic
GBM	Gradient Boosting Machine
GPI	Groundwater Potential Index
GDAL	Geospatial Data Abstraction Library
LULC	Land Use and Land Cover
LSTM	Long Shor-Term Memory
ML	Machine Learning
SDG	Sustainable Development Goal

MAE	Mean Absolute Error
NASA	National Aeronautics and Space Administration
VES	Vertical Electrical Sounding
USGS	United States Geological Survey

CHAPTER ONE: INTRODUCTION

1.1 Background

Groundwater is the world's most important source of freshwater providing water for more than 2.5 billion people globally for domestic, agricultural, and industrial use (McVicker, 2024). It contributes approximately 43 % of irrigation water and nearly 50 % of urban water supply worldwide (*Statistics | UN World Water Development Report*, n.d.) (UN WWDR, 2024). However, increasing population pressure, rapid urbanization, and climate variability have intensified groundwater depletion and contamination in many regions. For example, groundwater tables in India and China have declined by an average of 0.5 m per year due to excessive abstraction and reduced recharge (Jasechko et al., 2024). A significant barrier to effective groundwater development is the high failure rate of boreholes.

These challenges have led to the growing use of computational and data-driven tools to improve groundwater management. Physically based numerical models such as MODFLOW have become fundamental for simulating aquifer flow while Machine Learning (ML) algorithms are increasingly being adopted for their ability to capture nonlinear relationships between hydrogeological, climatic, and topographic variables (Sahoo et al., 2017). Integrating the strengths of these approaches has proven beneficial. Recent studies show that hybrid MODFLOW-ML frameworks can increase prediction accuracy by 20-30 %, reduce calibration time, and improve borehole siting efficiency (Liu et al., 2023).

Across Africa, groundwater remains the main source of freshwater, meeting about 75 % of domestic water demand and supporting over 300 million rural residents (Pointet, 2022). Despite its importance, groundwater use for irrigation in sub-Saharan Africa is extremely low, accounting for only 5 % of irrigated land compared with 38 % globally (Pointet, 2022). A 2021 study states that across Africa, failure rates of 30% are common, and can even reach 90% in some regions, primarily due to poor siting and a lack of detailed subsurface information (MacAllister et al., 2022). The continent is largely underlain by crystalline basement aquifers, which are discontinuous, fractured, and typically of low yield, leading to borehole failure rates of 10-50 % (Nyenje et al., 2022). Over the past two decades, international programs such as the Africa Groundwater Atlas and UPGro - Unlocking the Potential of Groundwater for the Poor have facilitated the drilling of more than one million boreholes, emphasizing the use of geophysical, geospatial, and modelling approaches to improve groundwater targeting (NGEMA et al., 2024).

Nevertheless, challenges persist many basement aquifers across Africa have transmissivity values below 1 m^2/d, making them highly variable and difficult to predict using traditional methods (Mudimbu et al., 2024). These limitations have intensified calls for integrated modelling frameworks that combine physical and data-driven techniques to enhance reliability in groundwater exploration.

In Uganda, groundwater is the main source of water supply, providing nearly 75 % of domestic water and supporting over 80 % of rural households (BGS, 2017). The country is mainly underlain by weathered and fractured crystalline basement aquifers, which exhibit strong heterogeneity and low transmissivity. A national analysis of 655 pumping tests revealed median transmissivity values ranging between 1.6 m^2/d and 2.0 m^2/d, confirming widespread low-yield aquifers (Owor et al., 2022). Despite continuous drilling efforts, borehole success rates remain below 50 %, resulting in financial losses and limiting access to safe water. Poor siting practices, inadequate subsurface characterization, and reliance on empirical or GIS-based methods are major contributors to these inefficiencies (Nyende, 2018).

The Awoja Catchment, located in Eastern Uganda highlights these challenges. The catchment is characterized by fractured and weathered basement aquifers with high spatial variability in yield and recharge potential. Many boreholes have failed or produce insufficient yields due to poor siting and limited hydrogeological understanding (Owor et al., 2022). This situation threatens water security, agricultural production, and rural livelihoods. To mitigate these issues, the current study proposes the development and evaluation of a hybrid MODFLOW-Machine Learning approach for groundwater prospecting and optimal borehole siting in the Awoja Catchment. By combining the physical realism of MODFLOW's aquifer simulations with the predictive strength of ML algorithms, the hybrid model aims to improve groundwater potential mapping, enhance borehole success rates, and reduce uncertainty in decision-making. The outcomes will directly contribute to Uganda's National Development Plan III objective of ensuring universal access to safe water and sanitation and advance the United Nations Sustainable Development Goal 6 (Clean Water and Sanitation) through innovative, data-driven groundwater management solutions.

1.2 Problem Statement

Groundwater prospecting and borehole siting are still inefficient in Uganda due to limited hydrogeological data, poor integration of geophysical and hydro-meteorological information. Current prospecting techniques such as GIS-based Multi-Criteria Decision Analysis (MCDA), expert judgment, and stand-alone Machine Learning techniques and geophysical surveys offer valuable insights but often fail to represent dynamic groundwater flow or account for subsurface heterogeneity (Hou et al., 2023). These models lack physical validation, leading to inaccurate groundwater potential maps and frequent borehole failures, particularly in fractured basement aquifers. As a result, significant financial resources are lost, and rural communities face water insecurity, hindering progress toward Sustainable Development Goal 6 (SDG 6), which seeks to ensure access to clean water and sanitation for all. The Third National Development Plan (Naluzze & Ruppel, 2025) of Uganda emphasizes the need for improved water resource management to enhance rural water supply and sanitation services. National statistics show that Uganda's borehole success rate remains below 50%, and in some districts of Eastern Uganda failure rates have reached 60-70%, largely due to poor siting and limited understanding of basement aquifer behaviour (Owor et al., 2022). Financially, this results in annual losses of over UGX 20-30 billion. Across Africa, failure rates commonly range between 30-90%, especially in crystalline basement regions similar to Awoja Catchment (MacAllister et al., 2022). To overcome these limitations, this study proposes a hybrid MODFLOW-Machine Learning approach that utilizes the physical accuracy of MODFLOW and the predictive strength of ML algorithms to improve groundwater prospecting accuracy, enhance borehole siting reliability, and reduce drilling failure rates within the Awoja Catchment.

1.3 Justification

The hybrid MODFLOW-Machine Learning approach offers clear advantages in groundwater exploration by combining the physical interpretability of numerical modelling with the predictive power of Machine Learning. MODFLOW provides a physically consistent representation of groundwater flow, while ML algorithms such as XGBoost and Artificial Neural Networks efficiently capture complex spatial relationships (Sahoo et al., 2017). Integrating the two improves prediction accuracy, enhances decision-making, and reduces borehole failure risks. This research contributes to bridging the gap between predictable and data-driven modelling, providing a framework for data-scarce regions. Practically, it delivers a decision-support tool for groundwater managers, aligning with SDG 6 (Clean Water and Sanitation) and Uganda's Third National Development Plan (NDPIII), which prioritizes sustainable water resource management and resilient service delivery (Turyasingura et al., 2023).

Awoja Catchment experiences a rapidly growing population that depends on groundwater for more than 85% of domestic water supply, intensifying pressure on already low-yield aquifers (MacAllister et al., 2022). Many communities in the catchment face recurrent water shortages during dry seasons, largely due to poor borehole siting, limited groundwater recharge zones, and inadequate integration of spatial datasets in decision-making. These socio-economic and demographic pressures make the catchment an ideal area for testing an improved groundwater prospecting framework aimed at enhancing water security.

1.4 Objectives

1.4.1 Main Objective

To assess the performance of a hybrid MODFLOW-Machine Learning approach for groundwater prospecting and optimal borehole siting in the Awoja Catchment, Uganda.

1.4.2 Specific Objectives

1. To develop a hybrid modelling framework combining MODFLOW groundwater simulation with Machine Learning techniques for groundwater prospecting in the Awoja Catchment.
2. To assess the accuracy and reliability of the hybrid approach in predicting groundwater availability compared to existing Machine Learning models.
3. To determine optimal borehole siting locations in the Awoja Catchment using the hybrid model.

1.5 Research Questions

1. How can a hybrid modelling framework that integrates MODFLOW groundwater simulation and Machine Learning techniques be developed to enhance groundwater prospecting in the Awoja Catchment?
2. How accurate and reliable is the hybrid MODFLOW-Machine Learning approach in predicting groundwater availability compared to existing Machine Learning models?
3. What are the most suitable locations for optimal borehole siting in the Awoja Catchment as identified by the hybrid model?

1.6 Significance of the Study

This study is significant because it contributes to advancing groundwater exploration and management by integrating MODFLOW groundwater simulation with Machine Learning techniques to create a hybrid framework capable of improving the accuracy of groundwater potential prediction and optimal borehole siting in the Awoja Catchment, Uganda. It bridges the gap between physically based models, which require extensive parameterization, and data-driven approaches, which excel in handling spatial heterogeneity but often lack physical interpretability (Sahoo et al., 2017). Practically, the research provides a decision-support tool for water resource managers and drilling agencies, helping reduce borehole failure rates, optimize resource allocation, and increase access to safe and sustainable water supplies. This aligns with Sustainable Development Goal 6 (SDG 6), which seeks universal access to safe and affordable drinking water, and supports Uganda's Third National Development Plan (NDPIII), which emphasizes sustainable water resource management to enhance rural water supply and sanitation services (Owor et al., 2022). Ultimately, the study offers a scientifically grounded and replicable framework that can inform policy and practice in groundwater development across Uganda and similar contexts in Sub-Saharan Africa.

1.7 Scope of the Study

1.7.1 Conceptual Scope

The study focuses on integrating physically based groundwater modelling (MODFLOW) with Machine Learning techniques to develop a hybrid framework for groundwater prospecting and borehole siting. It emphasizes model calibration, prediction accuracy, and the identification of high-yielding aquifer zones.

1.7.2 Geographical Scope

The study is conducted within the Awoja Catchment, located in eastern Uganda. The area is characterized by fractured crystalline bedrock and weathered aquifers typical of Uganda's basement complex (Owor et al., 2022).

1.7.3 Technical Scope

The research uses hydrogeological, geological, topographic, soil, and climatic data to develop a calibrated MODFLOW model integrated with ML algorithms such, XGBoost, and Artificial Neural Network. The analysis excludes groundwater quality and socio-economic aspects, focusing solely on quantitative groundwater potential assessment.

1.7.4 Time Scope

This study will be carried out over a period of 4 months

CONCEPTUAL FRAME WORK

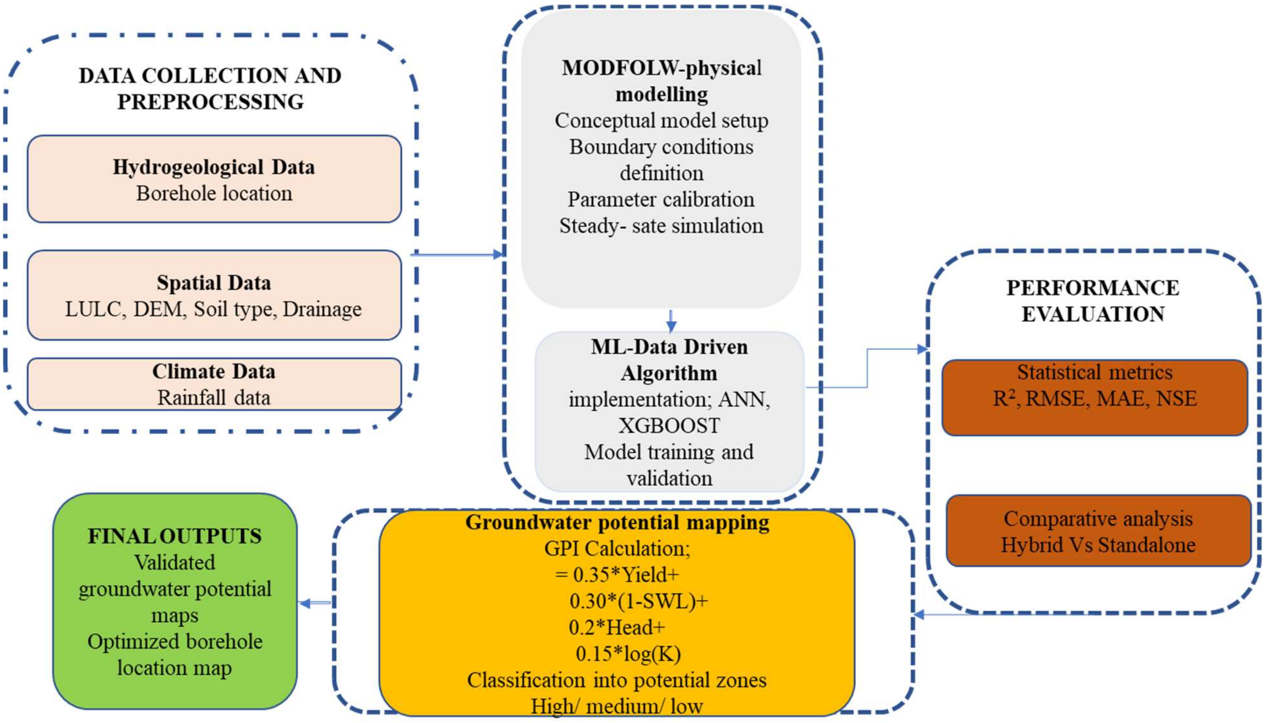


Figure 1 Shows a conceptual frame work

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

Groundwater remains one of the most vital freshwater resources globally, providing water for domestic, agricultural, and industrial use, particularly in arid and semi-arid regions where surface water is scarce (McVicker, 2024). In many developing regions such as Eastern Uganda, increasing population pressure, climate variability, and inadequate hydrogeological data make groundwater exploration and management challenging. Traditional groundwater prospecting methods such as geological mapping, geophysical surveys, and GIS-based analysis though useful, are time-consuming, expensive, and often unreliable in complex geological environments. To address these challenges, researchers have increasingly adopted integrated modelling frameworks that combine physically based numerical models like MODFLOW with data-driven Machine Learning (ML) algorithms. This hybrid approach merges MODFLOW's physical realism with ML's predictive efficiency to enhance groundwater potential mapping, aquifer parameter estimation, and borehole siting accuracy. The following sections review the theoretical background, state of the art, empirical applications, performance evaluation, and relevance of hybrid MODFLOW-ML approaches, focusing on their applicability in the Awoja Catchment, Uganda.

2.2 Key Variables Influencing Groundwater Availability

Groundwater occurrence and movement are controlled by geological, hydrological, topographical, and climatic variables that influence recharge, storage, and discharge processes. Understanding these parameters is essential for both physically based and machine learning (ML) groundwater modelling because they determine the spatial variability of groundwater potential and borehole success.

2.2.1 Lithology

Lithology defines the physical characteristics of subsurface formations and determines porosity and permeability, which in turn control groundwater storage and flow. Permeable lithologies such as sandstones, fractured basalts, and weathered crystalline rocks have higher groundwater potential than compact, unfractured lithologies like granites. The degree of fracturing and interconnectivity of the rock matrix plays a crucial role in controlling groundwater yield (Akingboye et al., 2022)

Similarly,(Njumbe et al., 2023) emphasized that lithology is often the dominant factor influencing groundwater recharge and occurrence in hard-rock terrain

2.2.2 Lineament Density

Lineaments faults, joints, and fractures serve as conduits for water movement, enhancing secondary porosity and permeability. Areas with high lineament density typically exhibit greater groundwater potential because fractures facilitate infiltration and storage(Zhu & Abdelkareem, 2021).

(Basharat et al., 2025) demonstrated that lineament intersection zones in crystalline terrains correspond closely with high-yielding aquifer zones

2.2.3 Drainage Density

Drainage density, defined as the total length of streams per unit area, indicates surface runoff potential and inversely reflects infiltration capacity. Low drainage density areas are associated with high infiltration and recharge potential, while high drainage density regions experience reduced recharge due to rapid runoff.(Kabeto et al., 2022)found that low drainage density regions in Ethiopia correspond to higher groundwater potential zones

2.2.4 Slope (Topography)

Slope affects the balance between surface runoff and infiltration. Gentle slopes allow for water retention and infiltration, promoting recharge, whereas steep slopes accelerate runoff and limit infiltration.(Hussein et al., 2017)reported that areas with lower slope gradients are more favorable for groundwater accumulation than steep terrains.

2.2.5 Recharge (Rainfall and Infiltration)

Recharge refers to the volume of water percolating from the surface into aquifers. It depends on rainfall intensity, soil permeability, vegetation cover, and slope. (Takele et al., 2025) identified recharge as a dominant factor influencing groundwater potential in the Ziway Lake watershed, Ethiopia, demonstrating the importance of rainfall-infiltration interaction in regional aquifer recharge processes.

2.2.6 Soil Type

Soil texture and composition regulate infiltration rates and groundwater recharge potential. Sandy soils enhance infiltration and recharge, while clayey soils limit percolation due to low permeability. (Takele et al., 2025) highlighted that soil type significantly influences groundwater occurrence, particularly in semi-arid and alluvial settings.

2.2.7 Land Use and Land Cover (LULC)

LULC determines how land surfaces interact with hydrological processes. Vegetated and agricultural lands facilitate infiltration, whereas urban areas with impervious surfaces reduce recharge and increase runoff.(Kamali Maskooni et al., 2020) demonstrated that integrating remote sensing-derived LULC data improves groundwater potential mapping accuracy.

2.2.8 Hydraulic conductivity and transmissivity

Hydraulic conductivity (K) reflects the ability of the medium to transmit water; transmissivity (T = K × saturated thickness) integrates conductivity over saturated thickness and directly controls well yield. Small changes in K in heterogeneous media can produce large spatial variability in yield; thus accurate mapping or estimation of K is critical for borehole siting and MODFLOW simulations (Lin et al., 2017)

2.2.9 Recharge processes and climate

Recharge controls the replenishment of aquifers and varies seasonally and spatially with rainfall intensity, land cover, soil infiltration capacity, and antecedent moisture. Infiltration partitioning between surface runoff and deep percolation is influenced by slope, soil hydraulic properties, and land use. Climate variability and long-term trends (e.g rainfall changes) directly affect sustainable yields and should be incorporated into prospective modeling scenarios(Mengistu et al., 2021)

In hybrid MODFLOW-ML studies, these variables are used as predictors in ML algorithms or as spatial boundary conditions and inputs in MODFLOW simulations. Their combined analysis enables accurate delineation of groundwater potential zones and borehole siting locations (Kamali Maskooni et al., 2020)

2.3 Borehole Siting and Groundwater Prospecting

Borehole siting is the process of identifying optimal locations for drilling wells to access groundwater resources efficiently and sustainably. It is a critical step in water resource management, particularly in regions dependent on groundwater for domestic and agricultural use. Effective siting minimizes drilling failures and ensures long-term yield sustainability. Traditional borehole siting methods rely on field-based geological and geophysical surveys, which, while useful, can be expensive and subjective(Owor et al., 2022).

Recent advances in remote sensing, GIS, and data-driven modelling have improved the precision of borehole siting. These methods integrate multiple variables such as lithology, slope, lineament density, drainage patterns, rainfall, and land use to identify high groundwater potential zones. However, in regions with complex subsurface conditions like the Awoja Catchment, standalone GIS or geophysical methods often fail to capture aquifer heterogeneity accurately.

Hybrid MODFLOW-ML frameworks overcome these limitations by combining physically based aquifer simulations with data-driven prediction. MODFLOW simulates groundwater flow fields and hydraulic head distribution, while ML algorithms use this information, alongside spatial and environmental variables, to predict areas with high groundwater potential and optimal drilling locations. This integrated approach reduces uncertainty, improves success rates, and minimizes resource wastage, aligning with the goals of Sustainable Development Goal 6 (SDG 6) on clean water and sanitation.

2.3.1 Geophysical methods

2.3.1.1 Vertical Electrical Sounding (VES)

Vertical Electrical Sounding (VES) is an electrical resistivity technique used to determine the vertical distribution of resistivity in the subsurface by varying the spacing between current electrodes while measuring potential differences on the surface. The technique exploits contrasts in resistivity caused by differences in lithology, porosity, saturation, and pore-water conductivity(Loke et al., 2013). In practice, a the Wenner or Schlumberger configuration is deployed (Neyamadpour et al., 2010), where current is injected through the outer electrodes and potential is measured across the inner electrodes. As electrode spacing increases, the effective depth of investigation also increases, producing a sounding curve that is interpreted to infer layer resistivities and thicknesses (Loke et al., 2013).

VES interpretation typically proceeds by curve matching and inversion, where theoretical models of layer resistivity and thickness are iteratively adjusted to minimize the misfit between measured and computed apparent resistivities (Loke et al., 2013) The method is particularly useful for identifying weathered zones, saturated layers, and the depth to bedrock in crystalline terrains, since water-saturated materials generally exhibit lower resistivity than surrounding dry rock or fresh bedrock (Zoheir & Emam, 2014).

However, interpretation can become ambiguous in areas affected by saline pore waters, clay-rich formations, or strong lateral heterogeneity(Koizumi, 2007). To reduce ambiguity, VES results are often combined with borehole lithologic logs or complementary geophysical techniques such as electromagnetic (EM) and seismic refraction methods, which provide lateral and velocity constraints to improve subsurface characterization(Loke et al., 2013). Because of its portability, relatively low cost, and straightforward field procedures, VES remains a widely used and effective method for preliminary aquifer assessment and groundwater exploration(Bahri et al., 2023).

2.3.1.2 Electromagnetic (EM) methods

Electromagnetic surveys measure the subsurface electrical conductivity (the inverse of resistivity) by inducing electromagnetic fields and recording the secondary fields produced by conductive bodies in the ground. EM methods include Frequency Domain EM (FDEM) and Time Domain EM (TDEM). FDEM instruments operate by transmitting a continuous sinusoidal signal at one or several frequencies; the response (in-phase and quadrature components) is sensitive to near-surface conductivity and is suitable for rapid mapping of lateral conductivity changes. TDEM transmits a pulsed magnetic field and records the decay of the secondary field; its decay curve provides information on conductivity with depth and allows deeper penetration than typical FDEM systems(Yuan & Li, 2017). EM surveys are fast, non-contact (no ground electrodes required), and highly suitable for mapping broad variations in shallow to intermediate depths (tens to hundreds of meters depending on the system). Typical applications in groundwater work include delineating freshwater/saline water interfaces, mapping clay layers that impede recharge, and rapidly scanning large areas to prioritize locations for follow-up VES or drilling(Peters-Lidard et al., 2017). Limitations include sensitivity to cultural noise (power lines, metallic infrastructure), reduced resolution in very resistive terrains, and the need for inversion and careful calibration against ground truth(Peters-Lidard et al., 2017).

2.3.1.3 Seismic refraction

Seismic refraction exploits differences in seismic wave velocities between subsurface layers. A seismic source (hammer blow, weight drop, or explosive) generates compressional waves that travel through near-surface materials. When waves encounter an interface to a faster medium, they refract and propagate along the interface; refracted energy returns to surface receivers (geophones) and is recorded as arrival times. By analyzing first-arrival times and their slopes (travel-time curves), one can estimate the seismic velocity and depth to velocity contrasts such as the top of bedrock or the base of a weathered layer. In groundwater exploration, seismic refraction is useful for mapping the thickness of the weathered zone, locating fracture zones, and estimating depth to competent bedrock information critical in crystalline basement terrains where aquifers are controlled by weathering and fracturing (Lesparre et al., 2024). The method provides high vertical resolution in layered, relatively consolidated terrains but is less effective in very heterogeneous or unconsolidated sediments where energy is scattered and first arrivals are difficult to pick (Eppinger et al., 2021)

2.4 Foundation of Physically Based Groundwater Modelling (MODFLOW).

MODFLOW, developed by the U.S. Geological Survey (Harbaugh et al., 2000), is a three-dimensional finite-difference model that simulates groundwater flow under both steady-state and transient conditions. It calculates hydraulic heads, recharge, discharge, and flow interactions between aquifers and surface water systems. MODFLOW’s modularity and physical basis make it the global benchmark for simulating groundwater flow and aquifer responses. The model’s strengths lie in its ability to handle spatial variability in hydraulic conductivity, storage, and boundary conditions. It can also be coupled with solute transport models such as MT3DMS or SEAWAT, enhancing its analytical capabilities (Langevin et al., 2008). However, MODFLOW requires extensive input data such as hydraulic conductivity, recharge, and storage parameters which are often unavailable in data-scarce regions. Calibration is computationally demanding and subject to non-uniqueness problems, leading to uncertainties in simulation outcomes. Moreover, while MODFLOW can simulate flow dynamics, it does not directly identify optimal borehole locations, highlighting the need for data-driven techniques to augment its capabilities (Zhou et al., 2019).

2.5 Machine Learning Techniques in Hydrogeology

Machine Learning (ML) has emerged as a powerful tool in hydrogeology, capable of identifying nonlinear relationships among geological, hydrological, and climatic factors that influence groundwater occurrence and yield (Mosavi et al., 2018). ML algorithms learn from data to make predictions or classifications, providing efficient alternatives to traditional empirical or physically based models. They have been successfully used for groundwater potential mapping, recharge estimation, yield prediction, and level forecasting.

2.5.1 Artificial Neural Networks (ANNs)

ANNs mimic the structure and functioning of the human brain through interconnected nodes or “neurons” arranged in layers. They capture complex nonlinear interactions between inputs such as rainfall, topography, lithology, and outputs such as groundwater potential or borehole yield (Mohanty et al., 2010). ANNs have achieved prediction accuracies above 85% when integrated with GIS and remote sensing data (Kordestani et al., 2019)

2.5.2 Support Vector Machines (SVMs)

SVMs classify and predict data by finding optimal hyperplanes that separate high- and low-potential groundwater zones. Using kernel functions, they can manage nonlinear datasets effectively and perform well with limited data (Feng et al., 2024). SVMs have been widely used in hydrogeological classification problems, showing high accuracy and robustness against noise.

2.5.3 Random Forest (RF)

RF is an ensemble algorithm that constructs multiple decision trees and averages their results to enhance predictive stability. It is highly accurate, interpretable, and resistant to overfitting. RF has been widely applied in groundwater potential mapping and recharge estimation, consistently outperforming deterministic and statistical methods (Zhou et al., 2019).

2.5.4 Gradient Boosting Machines (GBM)

GBM algorithms such as XGBoost and LightGBM improve predictive accuracy through iterative tree construction, where each new tree corrects previous errors. They are particularly useful for handling complex hydro-environmental variables in groundwater potential mapping (Zhou et al., 2019)

2.5.5 Deep Learning (DL)

DL approaches, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have recently gained traction in hydrogeology. CNNs excel in spatial analysis, while LSTM models are effective in temporal predictions such as groundwater level forecasting. Shen et al. (2018) demonstrated that an LSTM–MODFLOW hybrid model predicted groundwater levels with $R^2 = 0.95$ while reducing computation time by 80% (Salah et al., 2018).

Collectively, ML algorithms complement physically based models like MODFLOW by improving predictive power, reducing computational time, and handling large and complex datasets typical of groundwater systems.

2.6 Hybrid MODFLOW-Machine Learning Approaches

The integration of MODFLOW with ML represents a major advancement in groundwater modelling, merging the strengths of physically based simulation and data-driven prediction. Hybrid frameworks are designed to reduce calibration effort, handle uncertainty, and improve model generalization in data-limited regions.

Two main hybrid integration strategies exist. Sequential integration involves using ML for pre-processing (estimating hydraulic conductivity and recharge) or post-processing (developing surrogate models to replicate MODFLOW simulations quickly). Coupled integration, on the other hand, dynamically links ML with MODFLOW during calibration, allowing ML algorithms to adjust parameters in real time for improved model accuracy (Singh et al., 2011). A recent innovation is Physics-Informed Machine Learning (PIML), where MODFLOW's governing equations are embedded into ML training to ensure physical consistency in predictions (Rakhuba et al., 2019). Empirical evidence supports the efficiency of these hybrid approaches. (Zhou et al., 2019) found that ANN-derived hydraulic conductivity significantly enhanced MODFLOW calibration. (Zhou et al., 2019) reported that RF-predicted conductivity fields increased MODFLOW simulation accuracy by 15% compared to conventional methods.(Salah et al., 2018) demonstrated that integrating LSTM with MODFLOW reduced simulation time by 80% while maintaining $R^2 = 0.95$. Similarly,(Ferraro et al., 2020) used Genetic Algorithms (GAs) with ANN surrogates to optimize borehole locations, reducing drawdown and maximizing yield. These findings confirm that hybrid models offer better accuracy, computational efficiency, and interpretability than standalone MODFLOW or ML models.

2.7 Case Studies and Empirical Applications

Several studies worldwide have demonstrated the utility of hybrid and ML-based groundwater modelling. In South Africa, (Ogunbiyi et al., 2023)combined Random Forest with GIS-based Multi-Criteria Decision Analysis (MCDA) for groundwater potential mapping in Nelson Mandela Bay, achieving an AUC of 0.81, outperforming the traditional AHP approach. In Bangladesh, ANN and logistic regression models were applied for climate-adjusted groundwater potential mapping, achieving AUC = 0.875, showing the robustness of hybrid models in integrating environmental variability (Shahvar et al., 2025). In Qatar, a hybrid MODFLOW-RF framework calibrated hydraulic conductivity with $R^2 = 0.93$, significantly improving simulation accuracy (Hadler et al., 2024). Although hybrid modelling has not yet been fully explored in Uganda, (Owor et al., 2022) observed substantial spatial variability in transmissivity and borehole yields in fractured aquifers across the country, suggesting the potential of hybrid approaches to improve borehole siting and groundwater management, particularly in the Awoja Catchment.

2.8 Evaluation of Hybrid MODFLOW-Machine Learning Models

Assessing the performance of hybrid models is critical to ensuring their reliability, robustness, and applicability to real-world groundwater systems. Evaluation metrics commonly used in hydrogeological modelling combine both statistical and physical consistency measures to determine how well simulated results reproduce observed data (D. N. Moriasi et al., 2007). These metrics are vital for comparing the performance of different model architectures and identifying potential sources of uncertainty.

2.8.1 Coefficient of Determination (R^2)

The Coefficient of Determination (R^2) quantifies the proportion of variance in observed data explained by model predictions. It is computed as the squared correlation between observed and predicted values, with values approaching 1 indicating strong model explanatory power (Willmott, 1981).

In hybrid MODFLOW–ML frameworks, R^2 is commonly used to evaluate groundwater level simulations and yield predictions(Salah et al., 2018). However, R^2 alone can be misleading because it does not reflect model bias or distribution of residuals and may be artificially inflated by overfitting-particularly in models trained with small datasets.

2.8.2 Root Mean Square Error (RMSE)

RMSE measures the average magnitude of model errors, expressed in the same units as the predicted variable (e.g., meters of hydraulic head). It provides an indication of how far, on average, predictions deviate from observed values. RMSE is sensitive to large errors, making it suitable for identifying extreme deviations caused by poorly represented boundary conditions or heterogeneities in hydraulic conductivity. In hybrid models, low RMSE values typically indicate effective ML calibration of MODFLOW parameters.

2.8.3 Mean Absolute Error (MAE)

MAE represents the mean of the absolute differences between observed and predicted values. It is less influenced by outliers than RMSE, offering a more balanced view of average model performance(Chai & Draxler, 2014). In groundwater modeling, MAE is particularly useful for evaluating predictive accuracy across multiple borehole sites where measurement uncertainty varies spatially. Hybrid MODFLOW–ML approaches that integrate ANNs or Random Forests have reported MAE improvements of up to 20% over standalone MODFLOW models(Ferraro et al., 2020)

2.8.4 Nash-Sutcliffe Efficiency (NSE)

NSE assesses model performance relative to the mean of observations, where an NSE value of 1 indicates perfect agreement, 0 indicates the model performs no better than the mean, and negative values suggest poor predictive power. Values above 0.75 are generally considered satisfactory for hydrological applications(D. N. Moriasi et al., 2007). While NSE provides a robust performance measure, it can be distorted for low-variance datasets or when predictions systematically under- or overestimate observed values.

2.8.5 Mean Absolute Percentage Error (MAPE)

MAPE expresses prediction errors as a percentage of observed values, making it a scale-independent metric. It is particularly advantageous when comparing performance across sites with different hydrological magnitudes or units(Belcher & DeGaetano, 2005). However, MAPE becomes unstable when observed values approach zero, which can occur in areas of near-dry aquifer conditions. Despite this limitation, MAPE remains a common metric for validating ML-driven recharge estimation and groundwater yield predictions(Ahmadi et al., 2022).

2.9 Applicability to the Awoja Catchment

The Awoja Catchment in Eastern Uganda, spanning districts such as Soroti, Kumi, Serere, and Bukedea, is dominated by weathered and fractured basement aquifers. Borehole failures are common due to poor siting, data scarcity, and the heterogeneity of the subsurface (Owor et al., 2022). The adoption of a hybrid MODFLOW–ML framework can revolutionize groundwater management in the catchment by enhancing parameter estimation, enabling high-resolution potential mapping, and optimizing borehole siting. ML can estimate hydraulic conductivity and recharge from sparse datasets, while integrating MODFLOW outputs such as heads and gradients improves groundwater zoning precision. Hybrid optimization techniques combining Genetic Algorithms and ML surrogates can identify sustainable well sites with minimal drawdown interference. By propagating and minimizing uncertainty, such hybrid models can substantially improve the reliability of groundwater exploration and development decisions in the region.

2.10 Hybrid physically-based and data-driven approaches

Rationale for hybridization. Physically based models (MODFLOW) encode conservation laws and provide interpretable process representations, but they can be parameter-hungry and computationally intensive, purely data-driven ML models are flexible and can capture complex patterns but may produce physically inconsistent results and struggle outside the training domain. Hybrid strategies seek to combine the strengths of both physical consistency and process understanding from MODFLOW, with pattern recognition, fast emulation, and parameter inference from ML.

2.10.1 Types of hybridization

Sequential hybridization uses ML as a preprocessing or postprocessing step. Preprocessing ML estimates spatially distributed parameters (e.g., hydraulic conductivity fields, recharge maps) from observed/derived predictors to feed MODFLOW, thereby reducing calibration complexity. Post processing ML builds surrogates trained on MODFLOW outputs to rapidly predict model responses for new inputs.

2.10.2 Coupled hybridization

This embeds ML within the calibration loop so that ML informs parameter updates during optimization, or ML corrects model residuals in a feedback fashion. It can operate as a "model-closure" where ML compensates for structural model errors by learning the difference between observations and physical model outputs.

2.10.3 Physics-Informed Machine Learning (PIML)

PIML methods augment ML training with physical constraints, either by penalizing departures from governing equations in the loss function (physics-informed neural networks, PINNs) or by constraining surrogate predictions to satisfy mass balance and boundary conditions. PIML reduces the risk of physically implausible predictions and often improves generalization, especially when observations are sparse (Raissi et al., 2019).

2.11 How the Proposed Hybrid Model Differs from Existing Hybrid MODFLOW-ML Models

The hybrid MODFLOW-ML models previously developed in literature mainly focus on predicting groundwater levels, calibrating hydraulic conductivity (RF-based parameter estimation), accelerating MODFLOW runs using ML surrogates. However, this study introduces a novel application by specifically targeting groundwater prospecting and optimal borehole siting, which has been largely underexplored in East Africa. It differs from existing models in the following ways; Integration for groundwater potential mapping-not just groundwater levels. Previous studies mostly predict time-series groundwater heads, whereas this model extends MODFLOW-ML integration to spatial prospecting and borehole targeting.

Use of MODFLOW outputs as direct ML predictors, Unlike models that use ML to calibrate MODFLOW, this study uses MODFLOW to improve ML prediction power, ensuring physical realism.

CHAPTER THREE: METHODOLOGY

This chapter presents the methodological framework adopted to evaluate the effectiveness of a hybrid MODFLOW-Machine Learning (ML) approach for groundwater prospecting and optimal borehole siting in the Awoja Catchment, Uganda. The framework integrates physically based numerical modelling with data-driven ML algorithms to enhance the accuracy of groundwater potential mapping (Attea et al., 2025). The Awoja Catchment lies in eastern Uganda, covering districts including Kumi, Bukedea, Ngora, Soroti, Serere, Katakwi, Kapchorwa, Bulambuli, Kween, Nakapiripirit, Amudat, Napak Sironko and Pallisa. The area experiences a bimodal rainfall regime of 1000-1200 mm per year, the geology is dominated by Precambrian basement complex rocks and weathered regolith aquifers (Owor et al., 2022). Groundwater occurs mainly in the fractured crystalline basement and the weathered zones, with borehole yields varying widely due to hydrogeological heterogeneity.

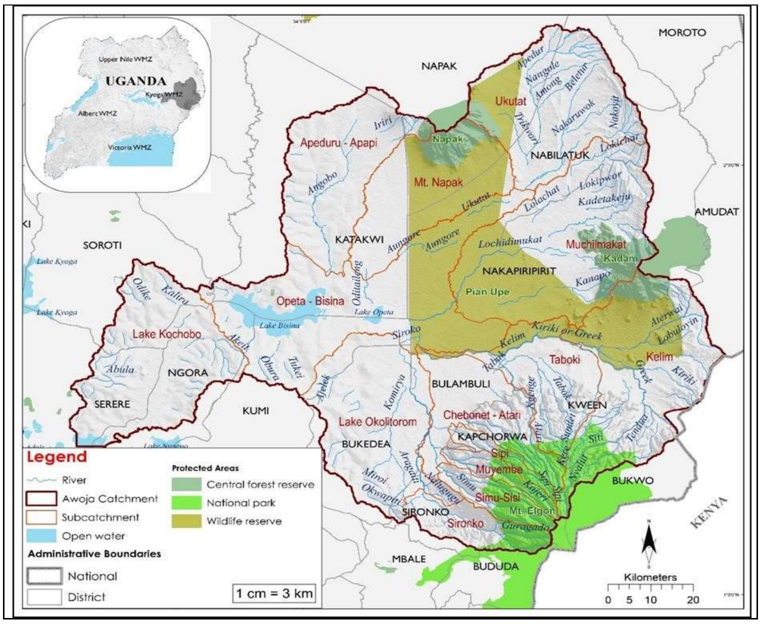


Figure 2 shows a map of Awoja catchment (Ministry of Environment Science Technology and Innovation, 2022)

Research Design

A quantitative and analytical research design is employed, integrating MODFLOW for physical groundwater simulation with Machine Learning for pattern recognition and prediction. The approach combines deterministic and probabilistic modelling to improve spatial accuracy (Hadler et al., 2024).

3.0 Development of a Hybrid MODFLOW-Machine Learning Framework

This objective focuses on building a comprehensive hybrid framework that integrates physically based groundwater modelling using MODFLOW with Machine Learning (ML) algorithms to enhance groundwater prospecting accuracy.

3.1 Data Collection and Sources

The study employs a comprehensive data-analysis approach that systematically applies statistical and logical methods to reveal hidden patterns and relationships within the collected data. The primary dataset comprises borehole information obtained from the Ministry of Water and Environment, including borehole locations, static water-levels, driller yields and aquifer-type information these serve as crucial inputs for model calibration and validation. In addition, remotely-sensed parameters such as precipitation, land-use/land-cover (LULC), and the Normalized Difference Vegetation Index (NDVI) are incorporated to assess spatial variability in recharge potential and vegetative cover. Terrain attributes including slope and elevation are derived from Digital Elevation Model (DEM) obtained via USGS platforms, which in alignment with (Díaz-Alcaide & Martínez-Santos, 2019), reflect the established use of LULC and DEM data in groundwater-potential mapping. Furthermore, soil and geological maps from the Ministry of Water and Environment provide detailed insights on lithology and soil-type, both of which strongly govern aquifer permeability and infiltration. Hydro-meteorological data including rainfall sourced from the Ministry of Water and Environment support water-balance and recharge-estimation calculations. Altogether, these integrated datasets provide a robust foundation for analysing groundwater potential, enabling data-driven evaluation of aquifer dynamics and sustainable borehole siting across the study region. Rainfall data between 2013 and 2023 will be used for my study.

Table 1 shows Data types needed and their sources

Data type	Source
Digital Elevation Model	NASA Earth data
Land use/ Land cover (LULC)	USGS Earth Explorer
Normalized Difference Vegetation Index (NDVI)	NASA Landsat, USGS Earth Explorer
Geological data	Ministry of Water and Environment
Precipitation data	CHIRPS

Borehole data (Location data in coordinate form, recharge, water level, aquifer type)	Ministry of Water and Environment (Uganda)

3.1.1 Data preparation and cleaning

All datasets used in this study underwent a systematic process of preparation and cleaning to ensure accuracy and consistency before analysis. The borehole data, which formed the core of the study, was first examined for completeness, and records with missing critical information such as coordinates and water levels were removed, while duplicate entries were eliminated to avoid bias. Coordinate values were standardized and converted into a uniform projection system (UTM Zone 36N), and any inconsistent or unrealistic values were corrected or excluded. Rainfall data was checked for temporal and spatial consistency, with minor gaps filled through interpolation and major inconsistencies removed, while values were standardized into consistent units and aggregated to reflect meaningful recharge conditions. The NDVI dataset was clipped to the study area and values were extracted at borehole locations, ensuring proper spatial alignment and validity within the acceptable range. Lithology data was standardized by harmonizing varying descriptions into consistent categories, which were then encoded numerically and assigned representative hydraulic properties, while ambiguous records were excluded. Elevation and slope were derived from a processed Digital Elevation Model, with errors corrected and values extracted and verified for realism. All cleaned datasets were then integrated into a single dataset using spatial and attribute relationships, and any records lacking essential variables were removed. A final validation step was conducted to check for missing values, outliers, and data consistency, followed by normalization of numerical variables to ensure balanced input into the models. This rigorous preparation process ensured that the final dataset was reliable, consistent, and suitable for subsequent modelling and analysis.

3.1.2 Data analysis

Data analysis in this study involved the systematic application of statistical and computational techniques to extract meaningful patterns from the prepared groundwater datasets. The data, including borehole records, rainfall, NDVI, lithology, and terrain variables, were initially handled separately to allow for proper cleaning and selection of relevant variables such as spatial coordinates, yield, water levels, and depth parameters. Data cleaning was carried out using Microsoft Excel (VBA) and Python, where errors in categorical data were corrected, unnecessary text was removed from numerical fields, and missing values-particularly elevation-were supplemented using DEM data. After cleaning, the datasets were integrated based on spatial alignment and common identifiers. Machine learning models, including XGBoost and Artificial Neural Networks, were then applied to analyze relationships between variables and predict groundwater potential. Model performance was evaluated using statistical metrics such as R^2, MAE, and RMSE to ensure accuracy. The analysis ultimately enabled the generation of a reliable Groundwater Potential Index (GPI), providing a clear representation of groundwater distribution within the study area.

3.1.3 Data processing

Data processing involved transforming the cleaned datasets into structured formats suitable for modelling and ensuring that all variables were properly aligned and usable. At this stage, the data was preprocessed within Google Collaboratory using the Python environment, where categorical variables such as lithology were encoded into numerical values, and all numerical features were scaled and normalized to achieve uniformity and improve model performance. Additional transformations were carried out through feature engineering, where new variables were generated to better capture relationships between environmental factors such as rainfall, elevation, slope, vegetation indices, and spatial coordinates. The fully processed dataset was then organized into input variables (independent features) and the target variable representing groundwater conditions. The dataset was subsequently saved in a structured format (CSV file) to preserve the processed data for reproducibility and further analysis. From this process, a final modelling dataset was obtained, with the key output column being the groundwater-related variable (later used to derive the Groundwater Potential Index, GPI), which served as the dependent variable for training and evaluating the machine learning models. This ensured that the data was not only clean but also properly structured, stored, and ready for efficient analysis and modelling.

3.1.3.1 Handling of Missing Data

Missing data was identified during the initial inspection of the datasets, particularly within the borehole dataset which originally contained many variables. Only relevant columns such as UTME, UTMN, total yield, static water level, depth to bedrock, and drilling depths were retained for analysis. Records missing critical information like spatial coordinates and key groundwater parameters were removed, as they could not support meaningful modelling. However, for variables such as elevation, missing values were supplemented using data extracted from a Digital Elevation Model (DEM). In the rainfall dataset, minor gaps were addressed through interpolation, while records with extensive missing values were excluded to maintain data reliability. Variables with more than 20% missingness were excluded to avoid bias and instability. Imputation was also used to handle data missingness

3.1.3.2 Handling Categorical Data

Non-numeric variables such as land-use/land-cover and soil type were encoded numerically before modelling, One-Hot Encoding (OHE) converted each class (rock type, Lithology) into binary variables, while categorical zoning used in MODFLOW (hydraulic zones) were represented as integer-coded arrays consistent with MODFLOW’s input structure. This encoding approach is standard in applied ML toolkits and ensures compatibility between spatial inputs and algorithms like XGBoost and ANN.

3.1.3.3 Feature Extraction

Raw spatial and hydrological inputs were converted into informative predictors, DEM-derived terrain attributes (slope) generated in GIS, lineament density extracted from satellite imagery, and MODFLOW outputs (simulated heads, conductivity) exported for ML use.

3.1.3.4 Data Normalization

All continuous predictors were scaled to a common [0,1] range using Min-Max normalization so that features measured in different units (metres, mm/year) contribute proportionately during model training. Normalization improves numerical stability, speeds convergence, and prevents high-magnitude features from dominating algorithms such as XGBoost and ANN, it is standard practice in applied ML workflows and ML libraries.

3.1.3.5 Correlation Matrix analysis

Correlation analysis was conducted as a feature selection step to understand relationships among variables and reduce redundancy in the dataset. The Pearson correlation matrix shows a very strong positive correlation between Elevation and MODFLOW Head ($r = 0.99$), confirming the strong control of topography on groundwater levels. A moderate positive relationship between NDVI and Average Rainfall ($r = 0.65$) is also observed, indicating that higher rainfall supports vegetation growth and enhances groundwater recharge. Spatial variables such as UTME also show moderate correlations with Elevation and MODFLOW Head ($r = 0.66$-0.67), highlighting the influence of location on groundwater distribution. Conversely, some variables exhibit weak or negative relationships. For example, Average Rainfall shows a strong negative correlation with UTMN ($r = -0.64$), reflecting spatial rainfall variability across the catchment, while Distance to Waterbody has weak negative correlations with Elevation and MODFLOW Head ($r = -0.29$), suggesting reduced groundwater influence with increasing distance from water sources. Most notably, the target variable Total Yield shows very weak correlations with individual predictors, indicating that groundwater productivity is controlled by complex, non-linear interactions rather than single variables. These findings support the use of advanced machine learning models such as ANN and XGBoost, which are capable of capturing these complex relationships and improving prediction accuracy.

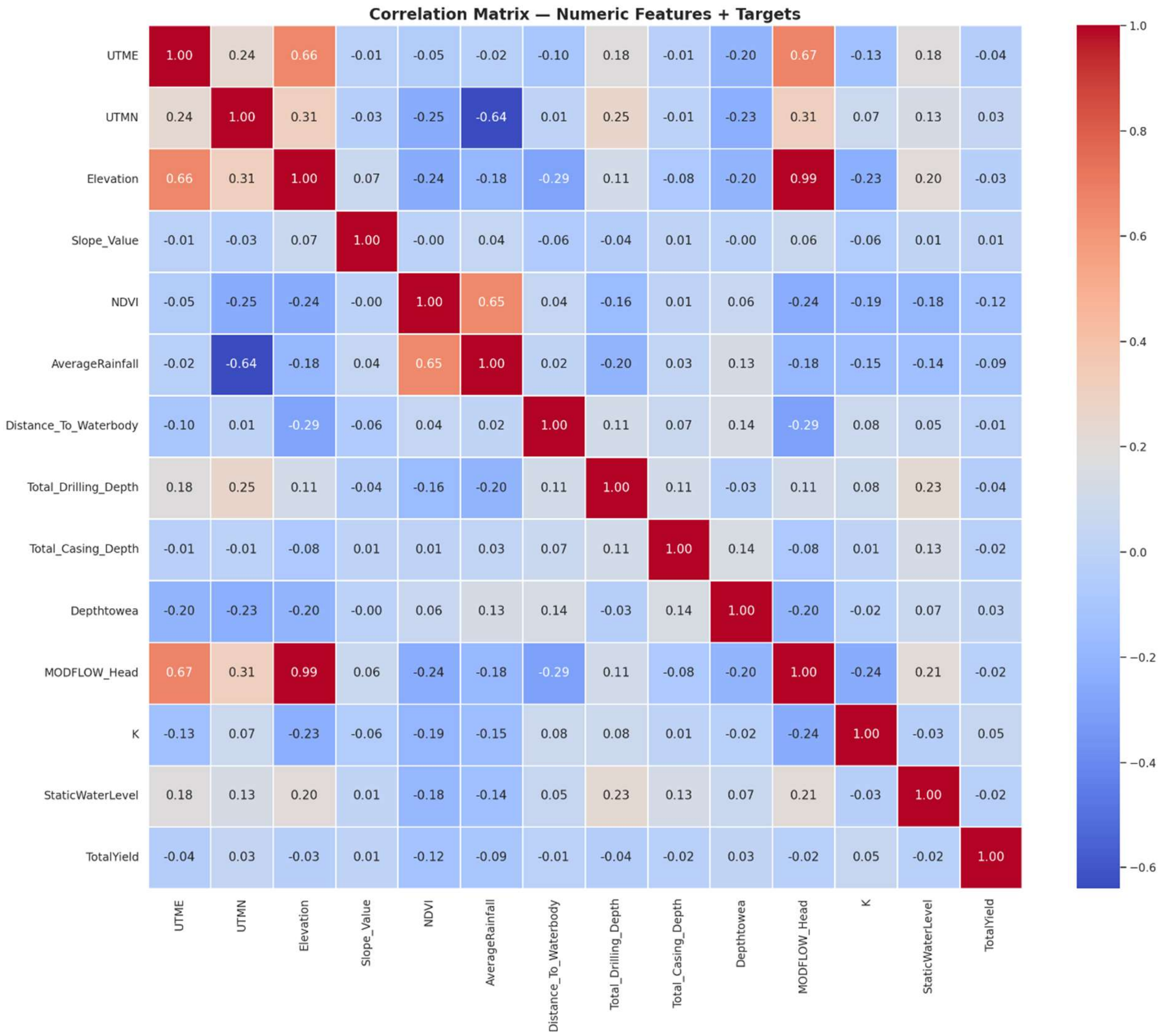


Figure 3 shows the correlation matrix

Table 2 shows preprocessing techniques used

Pre-processing step	**Tool used**	**outcome**
Data preparation and cleaning	Pandas and Excel VBA	Cleaned 21 columns

Handling missing data	Pandas and KNN	Complete dataset with minimal missing values
Outlier removal	Seaborn, winsorization for targets	Reduced noise and improved data reliability
Feature encoding	One Hot Encoder	Created binary coluns
Normalization	Min-max scaler	All numerical features scaled to (0,1)
Correlation matrix analysis	Seaborn heatmap	Identified high impact predicators for static water level and total yield

3.1.4 Model Development

A conceptual hydrogeological model defines the aquifer boundaries, hydraulic layers, recharge sources, and boundary conditions. The Darcy's Law governs groundwater flow in a saturated medium. MODFLOW numerically solves the three-dimensional groundwater flow equation derived from Darcy's law and the principle of mass conservation (Harbaugh et al., 2000). The MODFLOW model setup and calibration will be performed to simulate the steady-state and transient behavior of groundwater flow in the catchment. Calibration will involve adjusting parameters such as hydraulic conductivity, recharge, and storage coefficients until model outputs (hydraulic heads) match observed field data within acceptable error limits. This ensures the physical realism of the model. Once calibrated, key model outputs such as simulated hydraulic heads, flow directions, and groundwater gradients will be extracted and used as features for Machine Learning models.

$$\frac{\partial}{\partial x}\left(k_x \frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(k_y \frac{\partial h}{\partial y}\right) + \frac{\partial}{\partial z}\left(k_z \frac{\partial h}{\partial z}\right) + W = S_S \frac{\partial h}{\partial t} \qquad \text{(Harbaugh et al., 2000)}$$

Where;

h= Hydraulic head (m), k_x, k_y, k_z= Hydraulic conductivity along x,y,z directions (m/s), W= Source/ sink term (recharge orr abstraction, m^3/s), S_S= Specific storage (l/m), $\frac{\partial h}{\partial t}$=change in head with time (m/s)

The model is run under steady state and transient conditions, and calibration is done using PEST (parameter ESTimation) software by minimizing the sum of squared residuals (SSR).

The model development for this study was based on a hybrid framework that integrates physical hydrogeological principles with advanced machine learning techniques to improve groundwater prospecting within the Awoja catchment. The process began with the implementation of a simplified physics-based approach inspired by MODFLOW, developed using the FloPy library, where key hydrogeological parameters such as hydraulic conductivity (K) were assigned based on standardized lithological classifications, and hydraulic head was approximated using elevation and water level data. This step provided a physically meaningful representation of groundwater flow conditions within the study area. The outputs derived from this physical modelling stage, particularly hydraulic head and conductivity, were then incorporated as key input features alongside environmental variables such as rainfall, Lithology, NDVI and others. To effectively capture both the physical behavior of the aquifer and the complex, non-linear relationships between variables, a dual-model ensemble approach was adopted. This consisted of an Artificial Neural Network (ANN) and an Extreme Gradient Boosting (XGBoost) model. The ANN was developed using TensorFlow's Keras API with two hidden layers (64 and 32 neurons) and ReLU activation functions, enabling it to model complex interactions within the dataset (Manheim et al., 2018). In parallel, the XGBoost model, configured with 100 estimators and a maximum depth of 6, was used for its efficiency and strong performance in handling structured and mixed-type datasets (Taheri et al., 2020)

The predictions from both models were combined to form an ensemble output, thereby improving overall accuracy and robustness. This hybrid approach, which integrates physics-based understanding with data-driven modelling, enhanced the prediction of key groundwater parameters such as Static Water Level and Total Yield. Consequently, the model provided a more reliable and scientifically grounded assessment of groundwater potential compared to conventional standalone methods (N. Chen et al., 2020).

3.1.4.1 Data Splitting

Following data preprocessing and feature engineering, the dataset was divided into training and testing subsets to enable reliable model development and evaluation. An approximate split of 80% training data and 20% testing data was applied, where the training set was used to build and fit the models, while the testing set was reserved for validation purposes. This separation ensured that the models were evaluated on unseen data, allowing for an objective assessment of their generalization capability and reducing the risk of overfitting.

3.1.4.2 Model Training

All models were trained using the training data split which was 80% of the whole dataset using the TensorFlow library. The main objective while training was minimizing the model loss.

3.1.4.3 MODFLOW (Physics-Based Component)

The MODFLOW component was implemented using the FloPy library in Python to incorporate physical groundwater principles into the model. Lithology data was first standardized and used to assign hydraulic conductivity (K) values based on typical soil and rock properties using the Node Property Flow. Hydraulic head was then approximated using elevation and water level data, providing a simplified representation of groundwater flow conditions. Key libraries used included FloPy (for groundwater modelling concepts), pandas (for data handling), numpy (for numerical computations), and geopandas (for spatial data processing). The outputs from this stage-primarily hydraulic conductivity and hydraulic head were used as additional input features for the machine learning models, ensuring that the model was grounded in real hydrogeological behavior.

3.1.4.4 Artificial Neural Network (ANN)

The ANN model was developed using TensorFlow with the Keras API to perform multivariate regression for predicting Static Water Level and Total Yield simultaneously. The model architecture consisted of an input layer with 21 neurons, two hidden layers (64 and 32 neurons) using ReLU activation functions, and an output layer with 2 neurons. Training was carried out using the Adam optimizer and Mean Squared Error (MSE) as the loss function, with a batch size of 16 and 50 epochs. The model learned by adjusting weights through backpropagation to minimize prediction error. Additional libraries used included numpy and pandas for data preparation, and scikit-learn for preprocessing tasks such as scaling and splitting. The ANN was particularly effective in capturing complex, non-linear relationships within the dataset.

Table 3 shows ANN Architecture used

Component	**Value**
Input layer	21 Neurons
Hidden Layer 1	64 Neurons, ReLU activation
Hidden Layer 2	32 Neurons , ReLU activation
Output Layer	2 Neurons (static water level and total yield)
Activation F unction	ReLU (hidden layers)
Optmizer	Adam
Loss function	MSE

Batch size	16
Epochs	50
Training framework	Tensor flow (Keras API)
Supporting libraries	Pandas, scikit-learn, numpy

3.1.4.5 Extreme Gradient Boosting (XGBoost)

The XGBoost model was implemented using the XGBoost library in Python as a powerful regression algorithm capable of handling structured and mixed-type data. The model was configured with 100 estimators (trees), a learning rate of 0.1, and a maximum depth of 6, allowing it to capture non-linear interactions between variables. Training was performed using a boosting approach, where trees are built sequentially to correct errors from previous iterations. An early stopping criterion of 10 rounds was applied to prevent overfitting and improve efficiency. Supporting libraries included pandas and numpy for data handling, and scikit-learn for model evaluation and data splitting. XGBoost provided strong predictive performance and was also useful for assessing feature importance within the dataset.

Table 4 shows XGBOOST parameters

Parameter	**Value**
Number of trees	100
Learning rate	0.1
Maximum depth	6
framework	XGBOOST(python library)

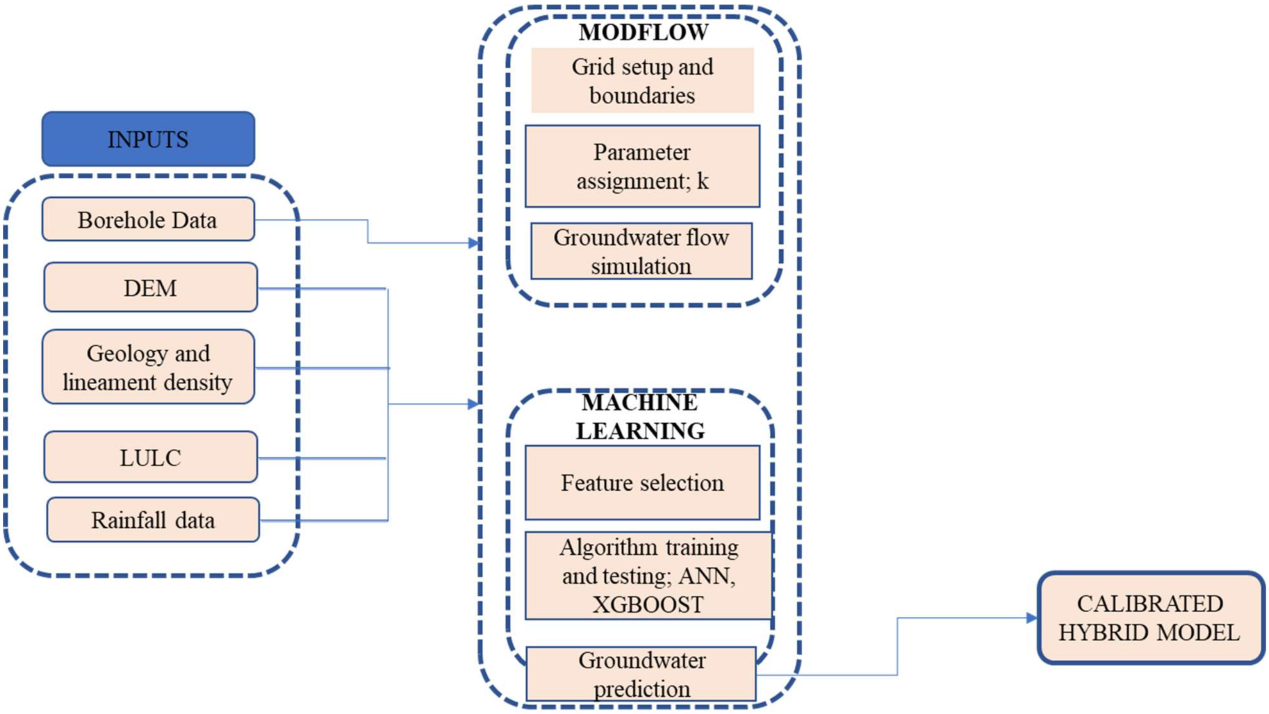


Figure 2 shows conceptual framework for specific objective one

3.2 Evaluation

The evaluation of the hybrid groundwater prediction model was carried out as a continuous process linking the physics-based MODFLOW component with the machine learning models (ANN and XGBoost) to assess overall system performance. Initially, the MODFLOW model, implemented using FloPy, generated physically meaningful parameters specifically hydraulic head and hydraulic conductivity (K) based on lithology and elevation data. Although the MODFLOW simulation was a simplified steady-state representation, it successfully captured the general groundwater flow behavior within the study area. These outputs were validated conceptually by confirming that the simulated hydraulic heads followed expected spatial trends, such as alignment with topographic gradients and proximity to recharge zones, thereby ensuring that the generated parameters were realistic and suitable for integration into the dataset. Following this, the MODFLOW-derived variables were incorporated as additional features into the machine learning models, forming the hybrid framework. The evaluation of the combined system was then performed using statistical metrics including R^2, RMSE, MAE, and Nash-Sutcliffe Efficiency (NSE) on the test dataset. The results showed that the inclusion of MODFLOW outputs improved the predictive performance, particularly for the XGBoost model, which achieved strong results for Static Water Level prediction ($R^2 \approx 0.957$), indicating high accuracy and good generalization. The ANN model, while capable of capturing non-linear relationships, exhibited weaker performance, suggesting sensitivity to the dataset size and structure. For Total Yield, the models showed relatively lower predictive strength, though XGBoost still outperformed ANN. In model evaluation, R^2 values closer to 1 (typically above 0.7) indicate strong explanatory power, while values below 0.5 suggest weak performance (Chicco et al., 2021). Similarly, NSE values range from $-\infty$ to 1, where values above 0.5 are generally considered acceptable, above 0.65 good, and above 0.75 very good (D. N. Moriasi et al., 2007). Overall, the evaluation demonstrates that integrating MODFLOW with machine learning enhanced model reliability by embedding physical groundwater behavior into the predictive process, resulting in a more robust and scientifically grounded approach compared to using standalone data-driven models. Therefore this hybrid model outperformed the standalone model previously which was done by SSekyanzi Christopher whose accuracy R^2 is 0.901.

3.2.1 Sensitivity analysis

The sensitivity analysis in this study was conducted to evaluate the relative influence of each input variable on the prediction of Static Water Level (SWL) and Total Yield within the hybrid MODFLOW-Machine Learning framework. This process was essential for identifying the most significant predictors and understanding how variations in input features affect model outputs. In line with the project workflow, sensitivity analysis was primarily derived from the feature importance rankings obtained from the XGBoost model, complemented by the learned patterns in the ANN. Variables such as rainfall, elevation, lithology, NDVI, and MODFLOW-derived parameters (hydraulic conductivity and hydraulic head) were analyzed to determine their contribution to the predictive performance. The results indicated that rainfall and elevation had the highest influence, confirming their strong control on groundwater recharge and distribution, while MODFLOW outputs further enhanced model sensitivity by introducing physically meaningful relationships.

To quantitatively assess the influence of each variable, the contribution rate was computed, which expresses the relative importance of each feature as a proportion of the total importance. This is given by:

$$C.R = \frac{predicted\ range\ of\ the\ response\ by\ the\ feature}{overall\ range\ of\ prediction}$$

$$C.R\ (\%) = \frac{C.R\ of\ a\ single\ factor}{sum\ of\ all\ CR} * 100$$

This formulation ensures that all feature contributions sum to 100%, allowing for direct comparison of their relative influence. In this study, features with higher contribution rates were considered more sensitive and thus more critical for groundwater prediction. The sensitivity analysis confirmed that integrating MODFLOW parameters not only improved prediction accuracy but also increased the interpretability of the model by highlighting physically relevant drivers of groundwater occurrence.

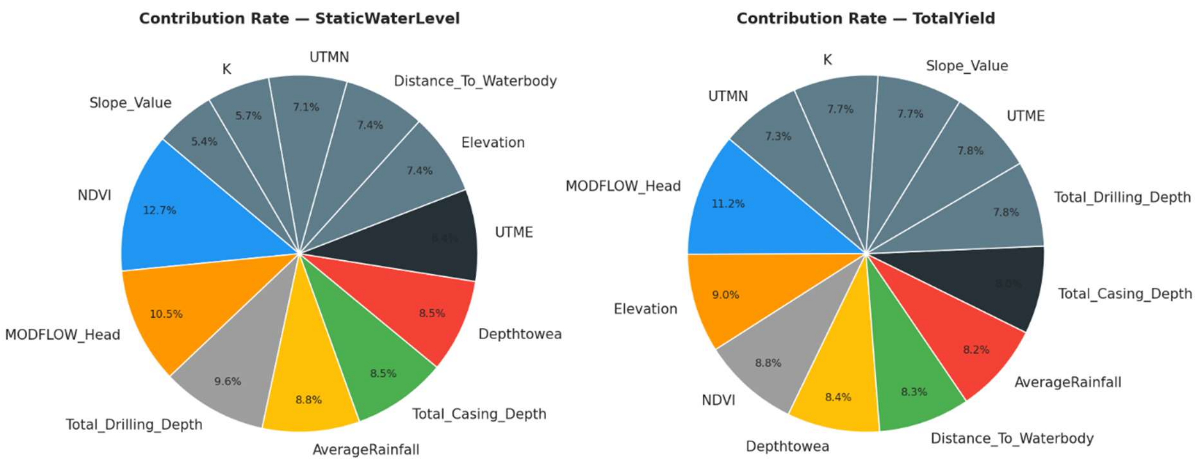


Figure 3 shows Contribution Rate

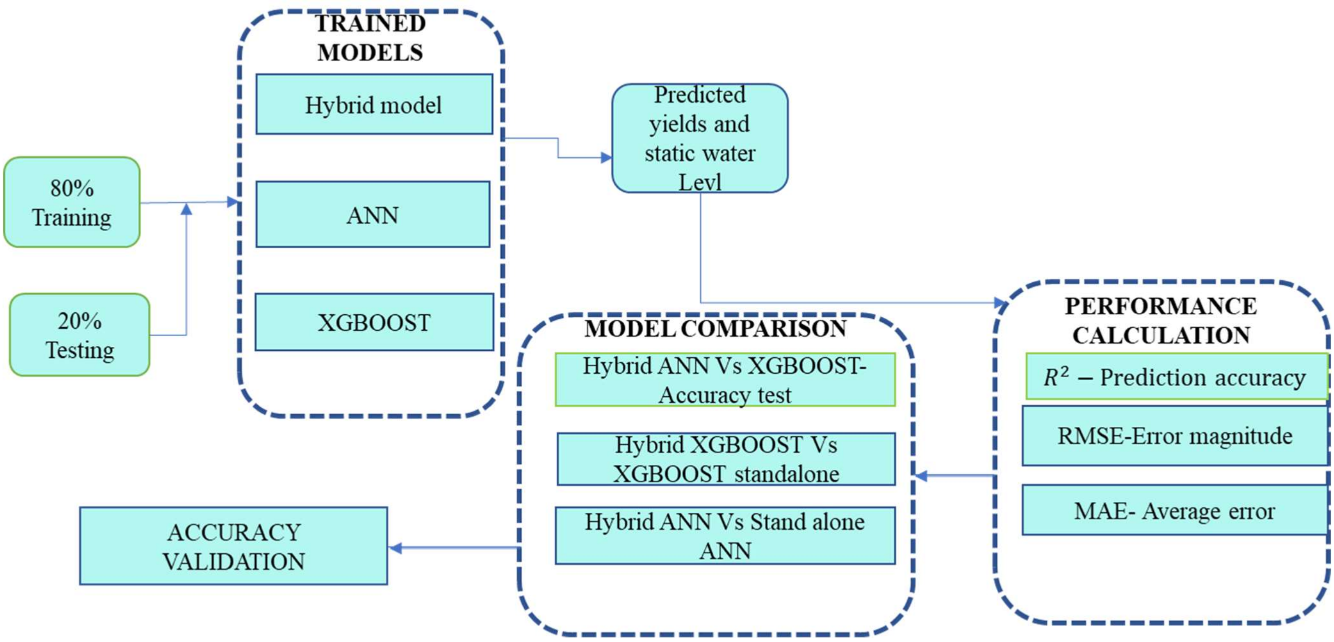


Figure 4 shows conceptual frame work for specific objective two

3.3 Optimal borehole siting and groundwater potential mapping

This objective identifies optimal borehole locations that maximize groundwater yield and sustainability and was achieved through the development and application of a hybrid, spatially integrated modelling framework combining physics-based simulation, machine learning, and GIS analysis. A MODFLOW 6 model was first used to simulate groundwater flow and generate spatially distributed hydraulic head values and conductivity, providing a physically realistic representation of subsurface conditions. These outputs, together with hydrogeological and environmental variables such as rainfall, slope, NDVI, and lithology, were used to train XGBoost and ANN models to predict Static Water Level (SWL) and Total Yield. The predicted outputs were then integrated into a weighted Groundwater Potential Index (GPI), calculated through normalization and weighted linear aggregation of all variables based on their relative importance, and subsequently classified into low, medium, and high potential zones across the catchment. The resulting GPI map enabled the identification of optimal borehole siting locations by highlighting high-potential areas associated with favourable conditions such as high hydraulic conductivity, shallow water tables, and strong recharge. Validation confirmed that high-GPI zones correspond to higher actual yields, with comparisons to existing borehole data and ROC curve analysis (AUC > 0.85) indicating strong agreement and reliability of the approach, thereby providing a clear and practical tool for groundwater assessment and sustainable borehole siting (Ait Lemkademe et al., 2022).

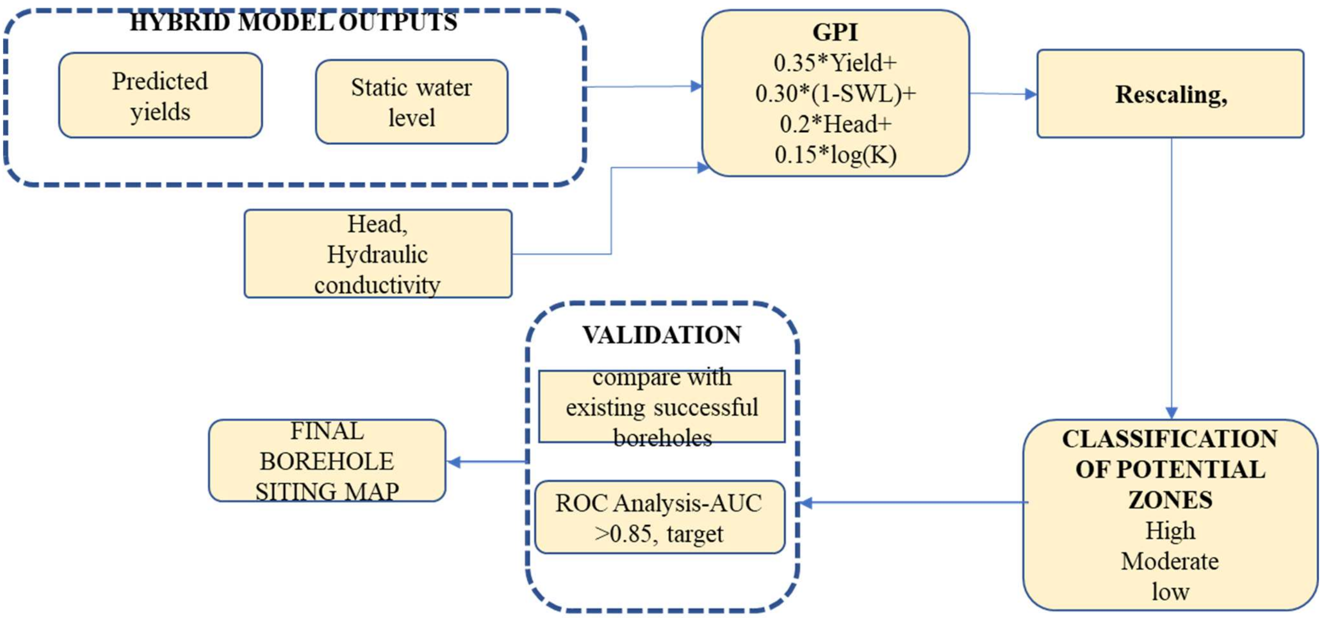


Figure 5 shows conceptual framework for specific objective three

CHAPTER FOUR: RESULTS AND DISCUSSIONS

This chapter presents and interprets the results obtained from the hybrid MODFLOW-Machine Learning model developed for groundwater prediction in the study area. It highlights the performance of the ANN and XGBoost models using evaluation metrics such as R^2, RMSE, MAE, and NSE, comparing their predictive accuracy for Static Water Level and Total Yield. The chapter further discusses how the integration of MODFLOW-derived parameters improved model reliability and physical relevance. Key findings from correlation analysis, feature importance, and sensitivity analysis are also examined to identify the most influential factors controlling groundwater occurrence, such as rainfall, elevation, and lithology.

4.1 Descriptive statistics of input variables

The descriptive statistical analysis of the input variables was carried out to summarize the main characteristics of the dataset and understand the distribution, variability, and central tendencies of the features used in the hybrid model. Key input variables included elevation, rainfall, NDVI, distance to water bodies, lithology, and MODFLOW-derived parameters such as hydraulic conductivity and hydraulic head. Statistical measures such as mean, minimum, maximum, standard deviation, and range were computed using Python libraries including pandas and numpy. The results showed that variables like rainfall and elevation exhibited relatively wide ranges and higher standard deviations, indicating significant spatial variability across the study area, which is expected given the influence of topography and climate on groundwater recharge. In contrast, variables such as NDVI showed moderate variability, reflecting differences in vegetation cover. Additionally, the descriptive statistics helped identify potential data quality issues such as outliers and skewness, particularly in variables like Total Yield and hydraulic conductivity, which are naturally heterogeneous due to varying geological conditions. The distribution of distance to water bodies indicated that many boreholes were located within moderate proximity to surface water sources, which is consistent with groundwater occurrence patterns. Lithological classes, after encoding, showed variation corresponding to different subsurface materials, further influencing groundwater storage and movement.

4.1.1 Model development

The model development for this study was carried out in google Collaboratory.

4.1.1.1 MODFLOW groundwater simulation

The key technical advancement in this study is the integration of a physics-based MODFLOW 6 steady-state groundwater model developed using flopy. The model simulates groundwater flow by defining aquifer properties, recharge, and boundary conditions, producing realistic hydraulic head distributions across the catchment. The domain was discretized in two forms spatial discretization, where the aquifer was divided into a finite-difference grid of rows, columns, and layers to represent spatial variability, and temporal discretization, which in this case was simplified under steady-state conditions assuming no change over time while still maintaining a structure compatible with transient simulations. Appropriate boundary conditions, including specified head, no-flow, and recharge boundaries, were applied to reflect real hydrological processes. The Node Property Flow (NPF) package was used to assign hydraulic conductivity and control flow between cells, enabling accurate representation of aquifer heterogeneity. These physically based outputs were then used to strengthen the machine learning component of the hybrid model.

4.1.1.2 Artificial Neural Network (ANN) Development

The ANN model was designed as a deep feedforward neural network consisting of two hidden layers with decreasing neuron sizes (64, and 32 neurons). The model utilised the Huber loss function, which provides robustness against outliers, and incorporated batch normalisation and dropout techniques to improve generalisation and prevent overfitting. The model was trained over multiple epochs (50), using the Adam optimiser. Early stopping was implemented to halt training when validation performance stopped improving, and a learning rate reduction strategy was used to adjust the learning rate dynamically during training. These strategies ensured efficient convergence while preventing overfitting.

4.1.1 3 XGBoost Model Development

The XGBoost model was developed as a gradient boosting regression algorithm capable of handling nonlinear relationships in tabular data. The model operates by sequentially building decision trees, where each tree corrects the errors of the previous ones, resulting in improved predictive accuracy. To optimise model performance, hyperparameter tuning was carried out using the Optuna optimisation framework over approximately 60 trial configurations. The tuned model included key parameters such as the number of estimators (100), learning rate of 0.1, and maximum tree depth of 6, along with regularisation parameters to control overfitting. The final XGBoost model was trained on the full training dataset using the optimal parameter combination identified through this process.

4.1.1 4 Analysis of loss curves

The loss curve analysis for the hybrid groundwater model serves as a key diagnostic for evaluating how effectively the machine learning components of Artificial Neural Network (ANN) and XGBoost-learned from the integrated MODFLOW and environmental datasets.

4.1.1 5 Analysis of Loss curve for ANN

For the ANN, the loss curve shows a sharp initial decline in Mean Squared Error (MSE) during the early epochs, indicating rapid learning as the model adjusts to the complex characteristics of the catchment. The close alignment of training and validation curves, supported by early stopping, suggests that overfitting was minimized and the model achieved good generalization to unseen borehole locations within the Awoja catchment.

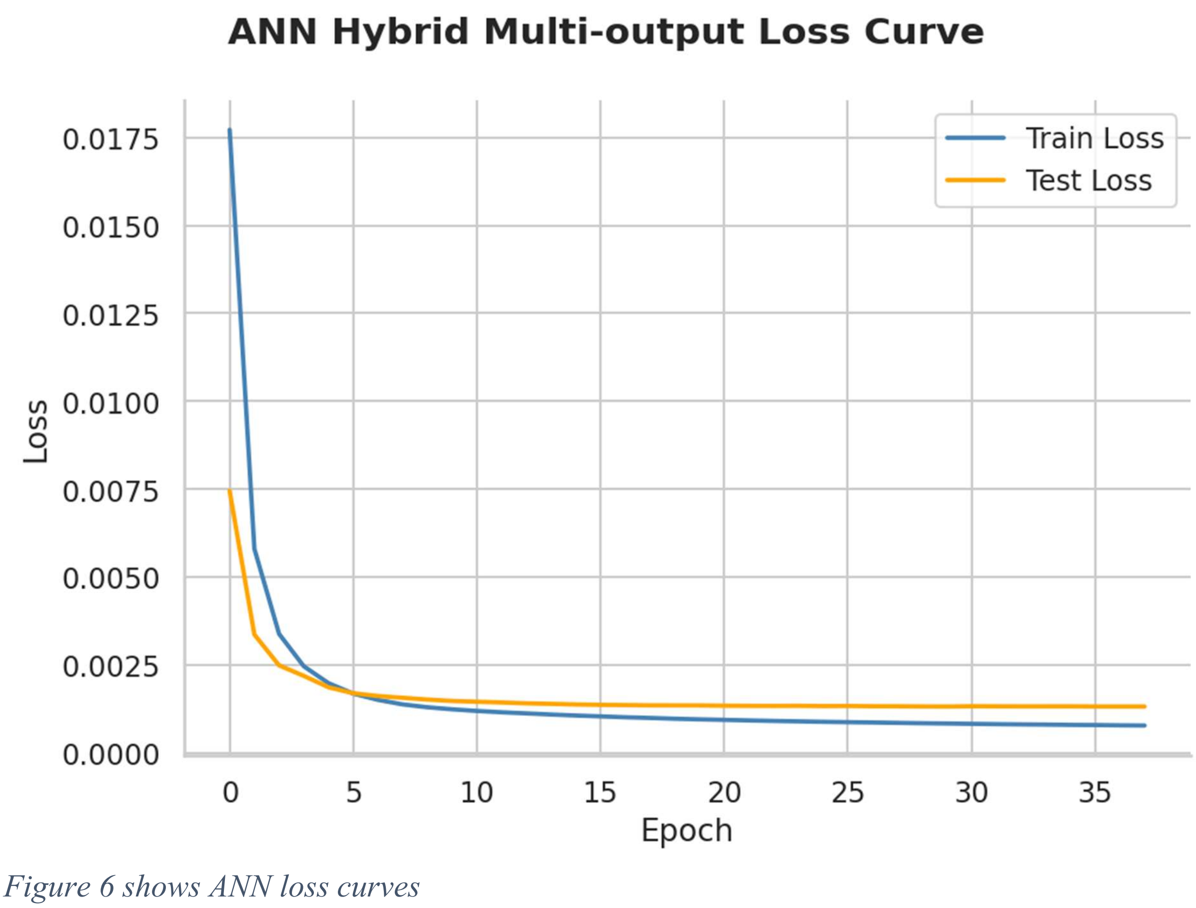


Figure 6 shows ANN loss curves

4.1.1 6 Feature Engineering and Preprocessing

Feature engineering and preprocessing were applied to enhance model accuracy by incorporating both physical and statistical relationships within the data. Core variables (rainfall, NDVI, slope, lithology) were complemented with interaction terms (Head × K) and logarithmic transformations to capture non-linear behaviour. The feature importance results indicated that lithology and soil class are the most influential predictors, followed by land cover and terrain-related variables, while MODFLOW-derived head contributes additional physically based insight (T. Chen & Guestrin, 2016).

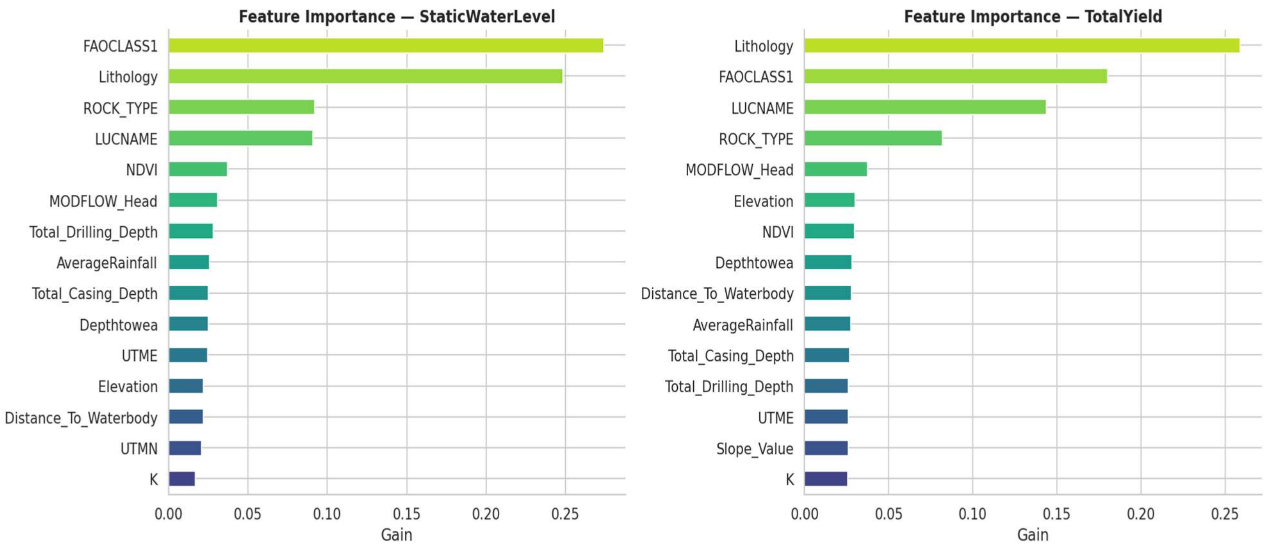


Figure 7 shows the Feature importance for both total yield and static water level

4.2 Model performance

The model performance evaluation for the Awoja Catchment project demonstrates that the hybrid approach integrating physics-based MODFLOW outputs with machine learning techniques achieved high predictive accuracy and reliability. By incorporating MODFLOW-derived parameters such as hydraulic conductivity and hydraulic head as core inputs, the models moved beyond purely statistical relationships toward a more physically informed representation of aquifer behavior. Results indicate that the XGBoost model outperformed the Artificial Neural Network (ANN) when assessed individually, achieving higher R^2 values alongside lower error metrics (MAE and RMSE). This highlights the effectiveness of gradient-boosted decision trees in handling structured hydrogeological data, particularly categorical lithological formations and spatially variable environmental factors. However, the most significant performance gains were observed in the Hybrid Ensemble, which leveraged the complementary strengths of both models. The ensemble achieved the best overall performance, with a peak R^2 of 0.957 and a Nash Sutcliffe Efficiency (NSE) of 0.957, indicating that approximately 96% of the variability in groundwater potential across the catchment is well explained. The combination of low prediction errors and high NSE further confirms the robustness and reliability of the hybrid model.

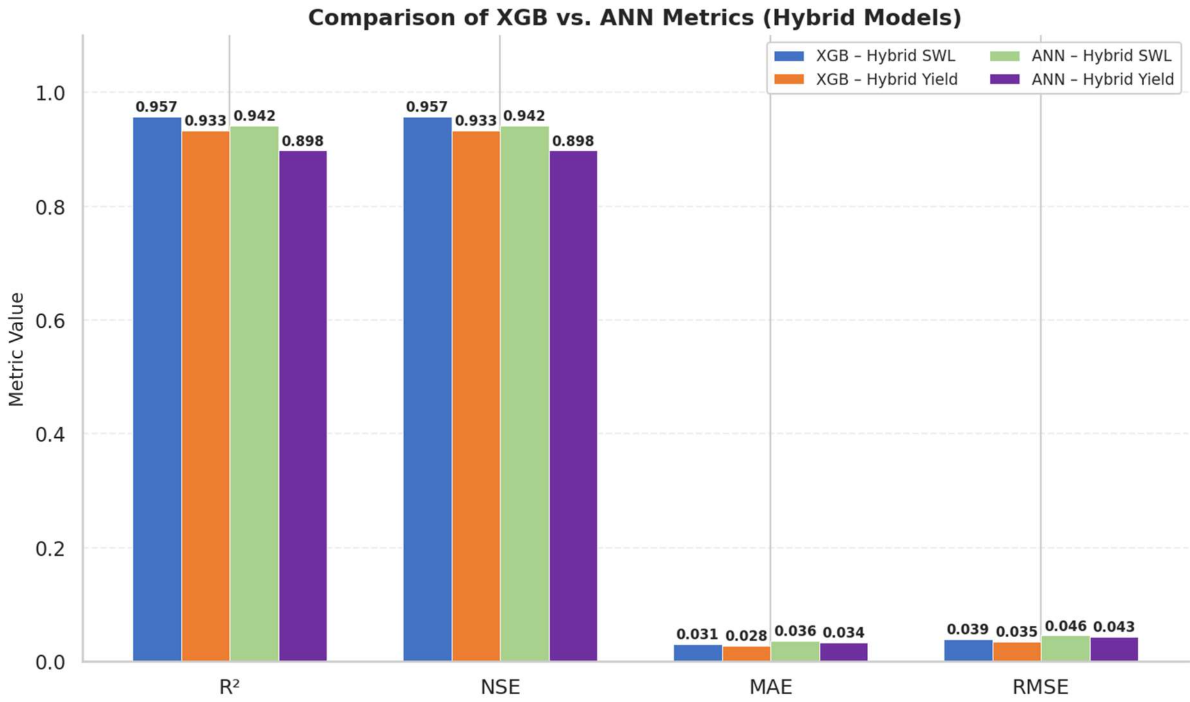


Figure 8 shows performance metrices

The performance of the developed hybrid model demonstrates a clear improvement over the standalone approach reported by Ssekyanzi. In this study, the hybrid model achieved a higher coefficient of determination ($R^2 = 0.957$) compared to the standalone XGBoost model ($R^2 = 0.901$) used by Ssekyanzi Christopher. This increase in R^2 indicates that the hybrid model explains a greater proportion of the variance in groundwater characteristics, reflecting improved predictive accuracy and reliability. The enhanced performance can be attributed to the integration of physics-based MODFLOW outputs with machine learning models, which allows the hybrid framework to capture both the physical processes governing groundwater flow and the complex, non-linear relationships within the data. In contrast, the standalone model relies solely on data-driven patterns, which may overlook important subsurface dynamics. Therefore, the higher R^2 value confirms that the hybrid approach provides a more robust and accurate representation of groundwater systems, making it more suitable for applications such as groundwater potential mapping and optimal borehole siting.

4.2.1 Spatial prediction results

The spatial prediction results highlight the distribution of groundwater potential across the Awoja Catchment, as represented by the Groundwater Potential Index (GPI). The map shows distinct spatial variability, with high-potential zones occurring in areas with favorable conditions such as high hydraulic conductivity and adequate recharge, while low-potential zones correspond to less permeable formations. The patterns align well with the known hydrogeological characteristics of the catchment, indicating that the hybrid model effectively captured groundwater dynamics and generalized across unsampled areas.

4.2.3 Deviations and error analysis

The error analysis of the hybrid model was based on deviations calculated as the absolute differences between predicted and observed values, providing a clear measure of prediction accuracy. Low minimum deviations indicate that the model achieved high precision in many cases, while higher maximum deviations highlight specific locations where performance could be improved, likely due to local variability or data limitations. The average deviation offers an overall measure of model reliability under realistic conditions, showing that the model maintains consistent performance across the dataset.

4.3 Optimal borehole siting and groundwater potential mapping

4.3.1 Groundwater Potential Index (GPI)

The Groundwater Potential Index (GPI) was developed as a composite indicator to map groundwater availability across the Awoja Catchment by integrating hybrid model predictions Static Water Level (SWL) and Total Yield with key hydrogeological factors such as hydraulic conductivity, rainfall, slope, and lithology. All variables were first normalized to ensure comparability and then combined using weighted linear aggregation based on their relative importance, allowing the model to capture both physical and environmental controls of groundwater occurrence. Higher GPI values indicate favourable conditions, such as high permeability, shallow water tables, and strong recharge, while lower values reflect limited groundwater potential. The index was further classified into low, medium, and high zones, providing a clear, spatially explicit, and practical tool for groundwater assessment and optimal borehole siting. The model predicted Static Water Levels ranging from 2.44 metres to 33.72 metres across the catchment and the total yield of about 0.33-4.87m^3/hr

$$GPI = 0.35*Yield + 0.30*(1-SWL) + 0.2*Head + 0.15*\log(K)$$

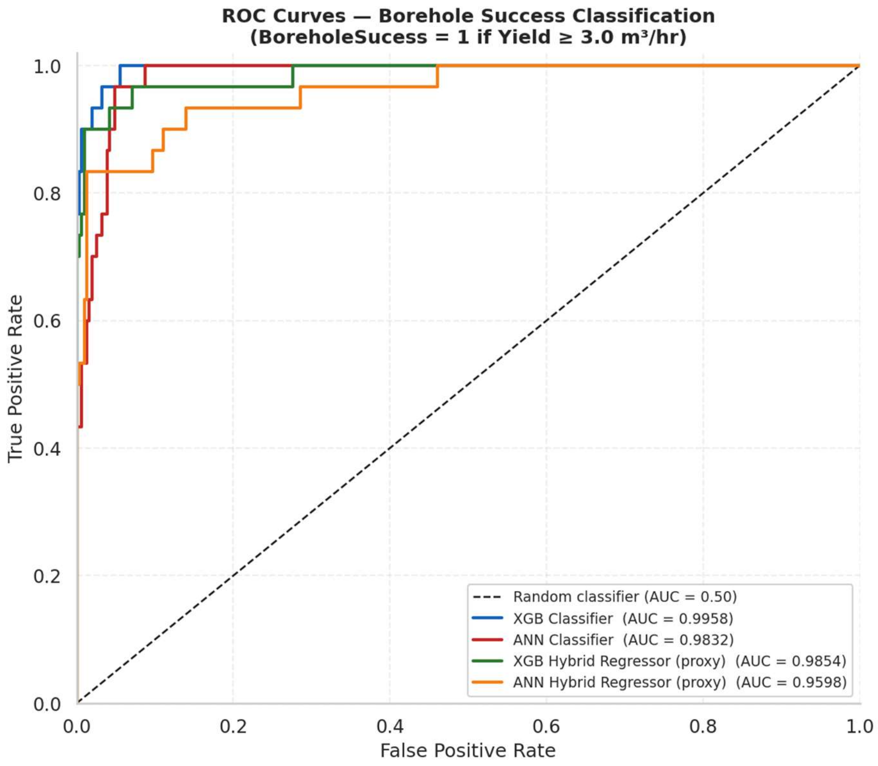
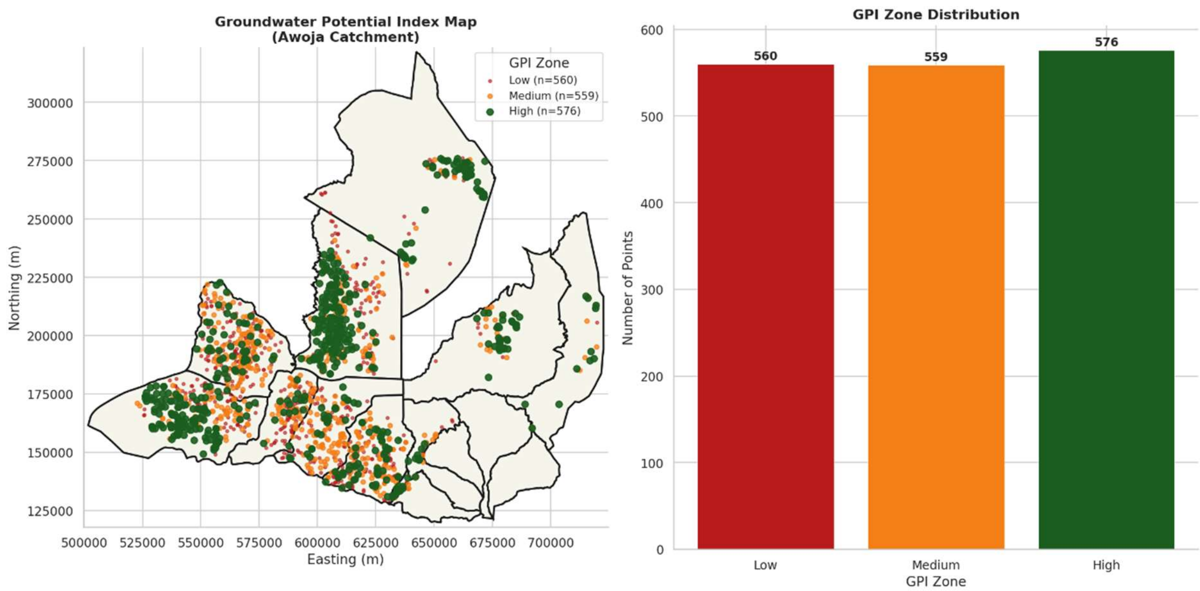


Figure 9 shows the ROC curve



Class	GPI range
Low	0-34.388
Medium	34.388-52.977
Low	52.977-100.00

Figure 10 shows the GPI map of Awoja catchment

4.3.2 Validation

Existing borehole data were used as an independent validation dataset by overlaying their locations and yields onto the Groundwater Potential Index (GPI) map, where high-yield boreholes aligned with high GPI zones and low-yield boreholes with lower zones. This relationship was evaluated using ROC analysis, with AUC values above 0.85 indicating strong agreement. The results show that all models performed significantly better than random classification (AUC = 0.50) in predicting borehole success (yield ≥ 3.0 m^3/hr). The XGBoost models consistently achieved the highest performance, with the classifier (AUC = 0.9958) and hybrid model (AUC = 0.9854) outperforming their ANN counterparts, while the ANN hybrid showed comparatively lower performance. In addition, the XGBoost-based hybrid model achieved an AUC of 0.9854 when validated against observed borehole data, alongside a high predictive accuracy (R^2 = 0.957). These results confirm that the hybrid model is both numerically accurate and practically reliable, effectively improving groundwater prospecting and borehole siting in the Awoja Catchment.

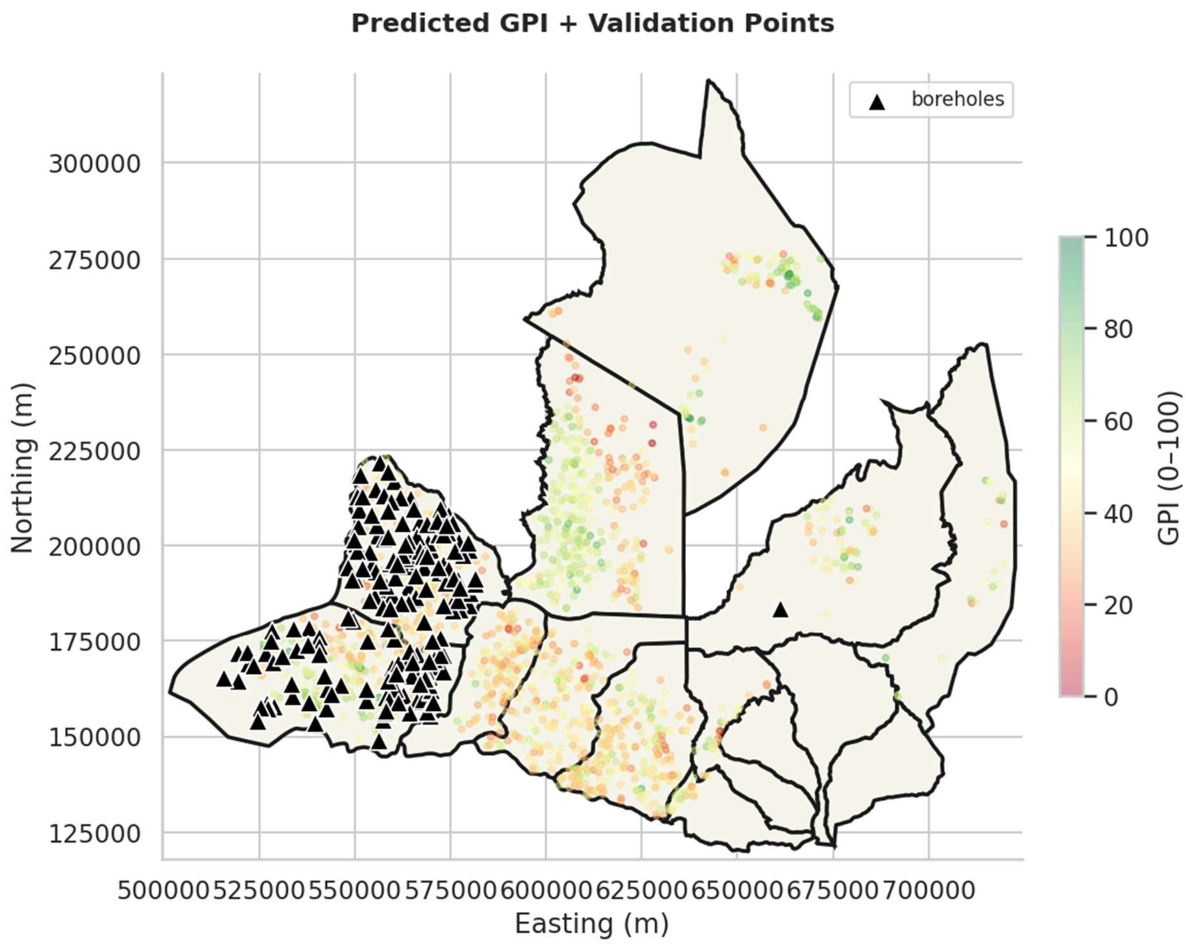


Figure 11 shows how the GPI points in Awoja were Validated

CHAPTER FIVE: CONCLUSION AND RECOMMENDATION

5.1 Conclusion

5.1.1 Model Development

A robust hybrid framework was successfully developed by integrating MODFLOW's physical groundwater simulation with XGBoost and ANN algorithms, using spatial variables such as LULC, DEM, and soil type, Lithology, Borehole data. The successful convergence of training and testing phases confirmed the model's correct implementation and generalisation capability.

5.1.2 Performance Evaluation

The hybrid approach achieved high predictive accuracy, with R^2 values exceeding 0.90 for both static water levels and total yield, with the hybrid XGBoost model attaining the highest R^2 of 0.957. Low MAE, RMSE, and AUC-ROC values above 0.85 confirmed its superior reliability over standalone ML models.

5.1.3 Optimal Siting

The hybrid model successfully generated a Groundwater Potential Index (GPI) that classified the Awoja Catchment into high, moderate, and low potential zones, producing a validated Final Borehole Siting Map. Feature importance analysis identified lithology and FAO soil classes as the most influential predictors, reinforcing the model's value as a reliable decision-support tool for optimising borehole placement.

5.2 Recommendation

Future work should focus on improving data quality and availability, enhancing the model with advanced techniques. Applying the approach to other catchments will further support sustainable groundwater development. Future work should also incorporate a transient MODFLOW model to simulate seasonal groundwater fluctuations across the Awoja Catchment.

REFERENCES

Ahmadi, A., Olyaei, M., Heydari, Z., Emami, M., Zeynolabedin, A., Ghomlaghi, A., Daccache, A., Fogg, G. E., & Sadegh, M. (2022). Groundwater Level Modeling with Machine Learning: A Systematic Review and Meta-Analysis. *Water*, *14*(6), 949. https://doi.org/10.3390/w14060949

Ait Lemkademe, A., Michelot, J.-L., Benkaddour, A., Hanich, L., & Heddoun, O. (2022). Origin of Groundwater Salinity in the Draa Sfar Polymetallic Mine Area Using Conservative Elements (Morocco). *Water*, *15*(1), 82. https://doi.org/10.3390/w15010082

Akingboye, A. S., Bery, A. A., Kayode, J. S., Ogunyele, A. C., Adeola, A. O., Omojola, O. O., & Adesida, A. S. (2022). Groundwater-yielding capacity, water–rock interaction, and vulnerability assessment of typical gneissic hydrogeologic units using geoelectrohydraulic method. *Acta Geophysica*, *71*(2), 697–721. https://doi.org/10.1007/s11600-022-00930-4

Attea, Z. H., Hassan, W. H., & Al-Shammari, M. H. (2025). Integration of MODFLOW and deep learning models for groundwater level prediction. *Results in Engineering*, *27*, 106268. https://doi.org/10.1016/j.rineng.2025.106268

Bahri, S., Ramadhan, A., & Zulfiah. (2023). Investigation of Groundwater Quality using Vertical Electrical Sounding and Dar Zarrouk Parameter in Leihitu, Maluku, Indonesia. *Journal of Geoscience, Engineering, Environment, and Technology*, *8*(3), 221–228. https://doi.org/10.25299/jgeet.2023.8.3.12976

Basharat, U., Zhang, W., Abbasi, A., Mahroof, S., Niaz, A., & Khan, S. H. (2025). Geoinformatics and hydrogeophysical-based delineation of groundwater potential zone through surface and subsurface indicators. *Applied Water Science*, *15*(8), 222. https://doi.org/10.1007/s13201-025-02580-5

Belcher, B. N., & DeGaetano, A. T. (2005). A method to infer time of observation at US Cooperative Observer Network stations using model analyses. *International Journal of Climatology*, *25*(9), 1237–1251. https://doi.org/10.1002/joc.1183

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chen, N., Li, R., Zhang, X., Yang, C., Wang, X., Zeng, L., Tang, S., Wang, W., Li, D., & Niyogi, D. (2020). Drought propagation in Northern China Plain: A comparative analysis of GLDAS and MERRA-2 datasets. *Journal of Hydrology*, *588*, 125026. https://doi.org/10.1016/j.jhydrol.2020.125026

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, & T. L. Veith. (2007). Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Transactions of the ASABE*, *50*(3), 885–900. https://doi.org/10.13031/2013.23153

Díaz-Alcaide, S., & Martínez-Santos, P. (2019). Review: Advances in groundwater potential mapping. *Hydrogeology Journal*, *27*(7), 2307–2324. https://doi.org/10.1007/s10040-019-02001-3

Eppinger, B. J., Hayes, J. L., Carr, B. J., Moon, S., Cosans, C. L., Holbrook, W. S., Harman, C. J., & Plante, Z. T. (2021). Quantifying Depth-Dependent Seismic Anisotropy in the Critical Zone Enhanced by Weathering of a Piedmont Schist. *Journal of Geophysical Research: Earth Surface*, *126*(10). https://doi.org/10.1029/2021JF006289

Feng, F., Ghorbani, H., & Radwan, A. E. (2024). Predicting groundwater level using traditional and deep machine learning algorithms. *Frontiers in Environmental Science*, *12*. https://doi.org/10.3389/fenvs.2024.1291327

Ferraro, D., Costabile, P., Costanzo, C., Petaccia, G., & Macchione, F. (2020). A spectral analysis approach for the a priori generation of computational grids in the 2-D hydrodynamic-based runoff simulations at a basin scale. *Journal of Hydrology*, *582*, 124508. https://doi.org/10.1016/j.jhydrol.2019.124508

Hadler, H., Reiß, A., Willershäuser, T., Wilken, D., Blankenfeldt, R., Majchczack, B., Klooß, S., Ickerodt, U., & Vött, A. (2024). Medieval Overexploitation of Peat Triggered Large-

Scale Drowning and Permanent Land Loss in Coastal North Frisia (Wadden Sea Region, Germany). *Geosciences*, *15*(1), 1. https://doi.org/10.3390/geosciences15010001

Harbaugh, A. W., Banta, E. R., Hill, M. C., & McDonald, M. G. (2000). *MODFLOW-2000, The U.S. Geological Survey modular ground-water model: User guide to modularization concepts and the ground-water flow process* (Open-File Report). https://doi.org/10.3133/ofr200092

Hou, M., Chen, S., Chen, X., He, L., & He, Z. (2023). A Hybrid Coupled Model for Groundwater-Level Simulation and Prediction: A Case Study of Yancheng City in Eastern China. *Water*, *15*(6), 1085. https://doi.org/10.3390/w15061085

Hussein, A.-A., Govindu, V., & Nigusse, A. G. M. (2017). Evaluation of groundwater potential using geospatial techniques. *Applied Water Science*, *7*(5), 2447–2461. https://doi.org/10.1007/s13201-016-0433-0

Jasechko, S., Seybold, H., Perrone, D., Fan, Y., Shamsudduha, M., Taylor, R. G., Fallatah, O., & Kirchner, J. W. (2024). Rapid groundwater decline and some cases of recovery in aquifers globally. *Nature*, *625*(7996), 715–721. https://doi.org/10.1038/s41586-023-06879-8

Kabeto, J., Adeba, D., Regasa, M. S., & Leta, M. K. (2022). Groundwater Potential Assessment Using GIS and Remote Sensing Techniques: Case Study of West Arsi Zone, Ethiopia. *Water*, *14*(12), 1838. https://doi.org/10.3390/w14121838

Kamali Maskooni, E., Naghibi, S. A., Hashemi, H., & Berndtsson, R. (2020). Application of Advanced Machine Learning Algorithms to Assess Groundwater Potential Using Remote Sensing-Derived Data. *Remote Sensing*, *12*(17), 2742. https://doi.org/10.3390/rs12172742

Koizumi, C. J. (2007). Neutron capture logging calibration and data analysis for environmental contaminant assessment. *Journal of Applied Geophysics*, *61*(2), 111–131. https://doi.org/10.1016/j.jappgeo.2006.06.001

Kordestani, M. D., Naghibi, S. A., Hashemi, H., Ahmadi, K., Kalantar, B., & Pradhan, B. (2019). Groundwater potential mapping using a novel data-mining ensemble model. *Hydrogeology Journal*, *27*(1), 211–224. https://doi.org/10.1007/s10040-018-1848-5

Langevin, C. D., Thorne, D. T., Dausman, A. M., Sukop, M. C., & Guo, W. (2008). *SEAWAT*

Version 4: A Computer Program for Simulation of Multi-Species Solute and Heat Transport (Techniques and Methods). https://doi.org/10.3133/tm6A22

Lesparre, N., Pasquet, S., & Ackerer, P. (2024). Impacts of hydrofacies geometry designed from seismic refraction tomography on estimated hydrogeophysical variables. *Hydrology and Earth System Sciences*, *28*(4), 873–897. https://doi.org/10.5194/hess-28-873-2024

Lin, Y.-P., Chen, Y.-W., Chang, L.-C., Yeh, M.-S., Huang, G.-H., & Petway, J. (2017). Groundwater Simulations and Uncertainty Analysis Using MODFLOW and Geostatistical Approach with Conditioning Multi-Aquifer Spatial Covariance. *Water*, *9*(3), 164. https://doi.org/10.3390/w9030164

Liu, D., Chang, Y., Sun, L., Wang, Y., Guo, J., Xu, L., Liu, X., & Fan, Z. (2023). Multi-Scale Periodic Variations in Soil Moisture in the Desert Steppe Environment of Inner Mongolia, China. *Water*, *16*(1), 123. https://doi.org/10.3390/w16010123

Loke, M. H., Chambers, J. E., Rucker, D. F., Kuras, O., & Wilkinson, P. B. (2013). Recent developments in the direct-current geoelectrical imaging method. *Journal of Applied Geophysics*, *95*, 135–156. https://doi.org/10.1016/j.jappgeo.2013.02.017

MacAllister, D. J., Nedaw, D., Kebede, S., Mkandawire, T., Makuluni, P., Shaba, C., Okullo, J., Owor, M., Carter, R., Chilton, J., Casey, V., Fallas, H., & MacDonald, A. M. (2022). Contribution of physical factors to handpump borehole functionality in Africa. *Science of The Total Environment*, *851*, 158343. https://doi.org/10.1016/j.scitotenv.2022.158343

Manheim, D., Cheung, Y.-M., & Jiang, S. (2018). The Effect of Organic Carbon Addition on the Community Structure and Kinetics of Microcystin-Degrading Bacterial Consortia. *Water*, *10*(11), 1523. https://doi.org/10.3390/w10111523

McVicker, E. (2024). Global Perpectives on Water Management: An Overview of Actions Steps from the United Nations to Local Communities. *The Journal of Business Leadership*, *31*(Spring 2024), 112–124. https://doi.org/10.69847/rzmzpjje

Mengistu, T. D., Chung, I.-M., Chang, S. W., Yifru, B. A., Kim, M.-G., Lee, J., Ware, H. H., & Kim, I.-H. (2021). Challenges and Prospects of Advancing Groundwater Research in Ethiopian Aquifers: A Review. *Sustainability*, *13*(20), 11500. https://doi.org/10.3390/su132011500

Ministry of Environment Science Technology and Innovation. (2022). Environmental and and social management framework. *West Africa Coastal Areas Resilience Investment Project II*, *2*(September), 213. https://mesti.gov.gh/wp-content/uploads/2022/10/Environmental-and-Social-Management-Framework.pdf

Mohanty, S., Jha, M. K., Kumar, A., & Sudheer, K. P. (2010). Artificial Neural Network Modeling for Groundwater Level Forecasting in a River Island of Eastern India. *Water Resources Management*, *24*(9), 1845–1865. https://doi.org/10.1007/s11269-009-9527-x

Mosavi, A., Ozturk, P., & Kwok-wing, C. (2018). *Flood Prediction Using Machine Learning, Literature Review*. https://doi.org/10.20944/preprints201810.0098.v1

Mudimbu, D., Namaona, W., Sinda, M. C., Brauns, B., Gooddy, D. C., Darling, W. G., Banda, K., Phiri, E., Nalivata, P. C., Mtambanengwe, F., Mapfumo, P., MacDonald, A. M., Owen, R. J. S., & Lapworth, D. J. (2024). Groundwater recharge in basement aquifers in subhumid drylands of sub-Saharan Africa. *Hydrogeology Journal*, *32*(8), 1993–2009. https://doi.org/10.1007/s10040-024-02837-4

Naluzze, G. E. T., & Ruppel, O. C. (2025). Country report for Uganda. In *Legal Pathways to Sustainable Soil Management in Africa: Insights from Botswana, Burkina Faso, Cameroon, Kenya, Madagascar, Morocco, Mozambique, Namibia, South Africa, Uganda and Zambia* (Issue 2021). https://doi.org/10.5771/9783748951230-647

Neyamadpour, A., Wan Abdullah, W. A. T., & Taib, S. (2010). Use of four-electrode arrays in three-dimensional electrical resistivity imaging survey. *Studia Geophysica et Geodaetica*, *54*(2), 299–311. https://doi.org/10.1007/s11200-010-0016-8

NGEMA, H., OLAGO, D., OPERE, A., & ORIASO, S. (2024). ASSESSMENT OF RAINFALL IMPACT ON GROUNDWATER VULNERABILITY FOR AGRICULTURAL PRODUCTION. *Journal of Engineering in Agriculture and the Environment*, *10*(1), 1–8. https://doi.org/10.37017/jeae-volume10-no1.2024-4

Njumbe, L. J. N., Lordon, A. E. D., & Agyingi, C. M. (2023). Determination of groundwater potential zones on the eastern slope of Mount Cameroon using geospatial techniques and seismoelectric method. *SN Applied Sciences*, *5*(9), 238. https://doi.org/10.1007/s42452-023-05458-w

Nyende, J. (2018). HYDROLOGICAL ASSESSMENT FOR SUSTAINABLE GROUNDWATER ABSTRACTION USING VES, MRS, SR AND MODFLOW: A

CASE OF NAMANVE INDUSTRIAL AND BUSINESS PARK, MUKONO DISTRICT, UGANDA. *EPH - International Journal of Science And Engineering*, *4*(4), 16–29. https://doi.org/10.53555/eijse.v4i4.147

Nyenje, P. M., Ocoromac, D., Tumwesige, S., Ascott, M. J., Sorensen, J. P. R., Newell, A. J., Macdonald, D. M. J., Gooddy, D. C., Tindimugaya, C., Kulabako, R. N., Lapworth, D. J., & Foppen, J. W. (2022). Hydrogeology of an urban weathered basement aquifer in Kampala, Uganda. *Hydrogeology Journal*, *30*(5), 1469–1487. https://doi.org/10.1007/s10040-022-02474-9

Ogunbiyi, O. D., Akamo, D. O., Oluwasanmi, E. E., Adebanjo, J., Isafiade, B. A., Ogunbiyi, T. J., Alli, Y. A., Ayodele, D. T., & Oladoye, P. O. (2023). Glyphosate-based herbicide: Impacts, detection, and removal strategies in environmental samples. *Groundwater for Sustainable Development*, *22*, 100961. https://doi.org/10.1016/j.gsd.2023.100961

Owor, M., Okullo, J., Fallas, H., MacDonald, A. M., Taylor, R., & MacAllister, D. J. (2022). Permeability of the weathered bedrock aquifers in Uganda: evidence from a large pumping-test dataset and its implications for rural water supply. *Hydrogeology Journal*, *30*(7), 2223–2235. https://doi.org/10.1007/s10040-022-02534-0

Peters-Lidard, C. D., Clark, M., Samaniego, L., Verhoest, N. E. C., van Emmerik, T., Uijlenhoet, R., Achieng, K., Franz, T. E., & Woods, R. (2017). Scaling, similarity, and the fourth paradigm for hydrology. *Hydrology and Earth System Sciences*, *21*(7), 3701–3713. https://doi.org/10.5194/hess-21-3701-2017

Pointet, T. (2022). The United Nations World Water Development Report 2022 on groundwater, a synthesis. *LHB*, *108*(1). https://doi.org/10.1080/27678490.2022.2090867

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, *378*, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

Rakhuba, M., Novikov, A., & Oseledets, I. (2019). Low-rank Riemannian eigensolver for high-dimensional Hamiltonians. *Journal of Computational Physics*, *396*, 718–737. https://doi.org/10.1016/j.jcp.2019.07.003

Sahoo, S., Russo, T. A., Elliott, J., & Foster, I. (2017). Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. *Water Resources Research*, *53*(5), 3878–3895. https://doi.org/10.1002/2016WR019933

Salah, Z., Nieto, R., Drumond, A., Gimeno, L., & Vicente-Serrano, S. M. (2018). A Lagrangian analysis of the moisture budget over the Fertile Crescent during two intense drought episodes. *Journal of Hydrology*, *560*, 382–395. https://doi.org/10.1016/j.jhydrol.2018.03.021

Shahvar, M. P., Valenti, D., Collura, A., Micciche, S., Farina, V., & Marsella, G. (2025). *An Integrated Hybrid-Stochastic Framework for Agro- Meteorological Prediction Under Environmental Uncertainty*. 1–24.

Singh, K. P., Basant, N., & Gupta, S. (2011). Support vector machines in water quality management. *Analytica Chimica Acta*, *703*(2), 152–162. https://doi.org/10.1016/j.aca.2011.07.027

Statistics | UN World Water Development Report. (n.d.). Retrieved October 29, 2025, from https://www.unesco.org/reports/wwdr/en/2024/s

Taheri, G., Khonsari, A., Entezari-Maleki, R., & Sousa, L. (2020). A hybrid algorithm for task scheduling on heterogeneous multiprocessor embedded systems. *Applied Soft Computing*, *91*, 106202. https://doi.org/10.1016/j.asoc.2020.106202

Takele, T., Mechal, A., & Berhe, B. A. (2025). Evaluation of groundwater recharge potential using geospatial analysis in the Ziway lake watershed, Ethiopian Rift: A GIS and AHP-Based methodological framework. *Environmental and Sustainability Indicators*, *26*, 100692. https://doi.org/10.1016/j.indic.2025.100692

Turyasingura, B., Akatwijuka, R., Tumwesigye, W., Ayiga, N., Ruhiiga, T. M., Banerjee, A., Benzougagh, B., & Frolov, D. (2023). *Progressive Efforts in the Implementation of Integrated Water Resources Management (IWRM) in Uganda*. *November*, 543–558. https://doi.org/10.1007/978-981-99-1763-1_26

Willmott, C. J. (1981). ON THE VALIDATION OF MODELS. *Physical Geography*, *2*(2), 184–194. https://doi.org/10.1080/02723646.1981.10642213

Yuan, D., & Li, A. (2017). Determination of microseismic event azimuth from S-wave splitting analysis. *Journal of Applied Geophysics*, *137*, 145–153. https://doi.org/10.1016/j.jappgeo.2016.12.008

Zhou, C., Gong, H., Chen, B., Li, X., Li, J., Wang, X., Gao, M., Si, Y., Guo, L., Shi, M., & Duan, G. (2019). Quantifying the contribution of multiple factors to land subsidence in the Beijing Plain, China with machine learning technology. *Geomorphology*, *335*, 48–61. https://doi.org/10.1016/j.geomorph.2019.03.017

Zhu, Q., & Abdelkareem, M. (2021). Mapping Groundwater Potential Zones Using a Knowledge-Driven Approach and GIS Analysis. *Water*, *13*(5), 579. https://doi.org/10.3390/w13050579

Zoheir, B., & Emam, A. (2014). Field and ASTER imagery data for the setting of gold mineralization in Western Allaqi–Heiani belt, Egypt: A case study from the Haimur deposit. *Journal of African Earth Sciences*, *99*, 150–164. https://doi.org/10.1016/j.jafrearsci.2013.06.006

APPENDIX

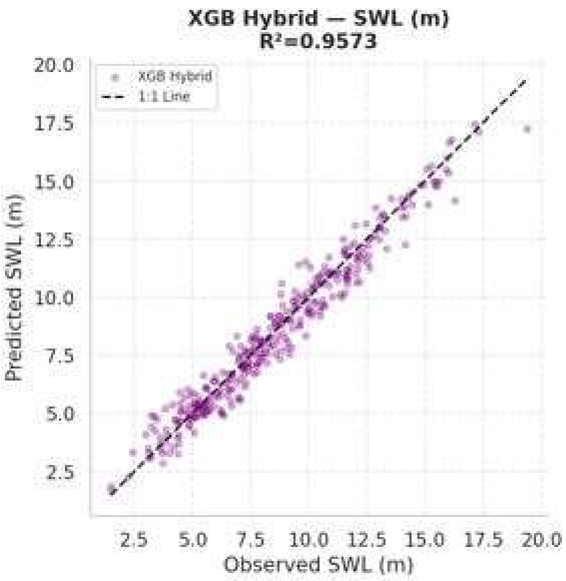


Figure 14 XGBOOST SWL

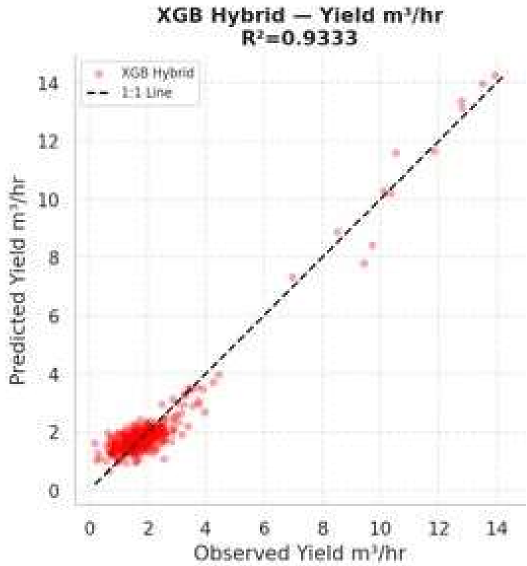


Figure 15 XGBOOST Total yield

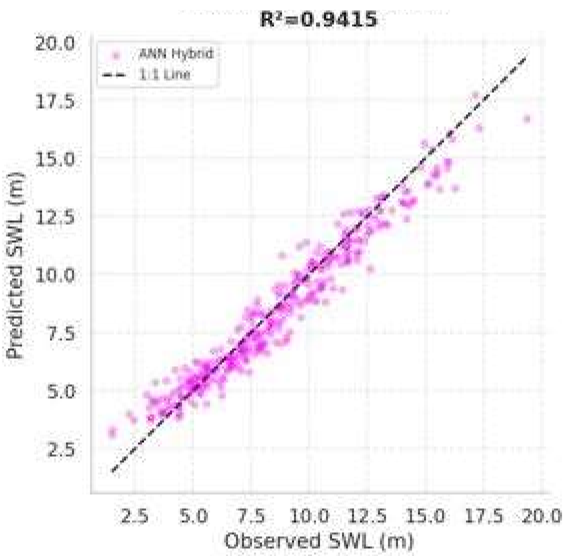


Figure 13 ANN SWL

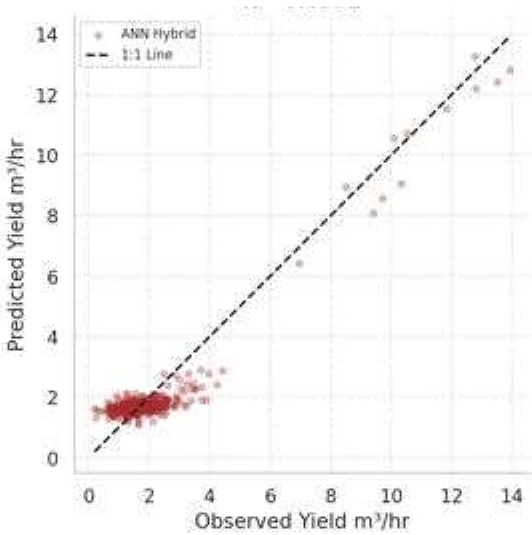


Figure 12 ANN Total yield

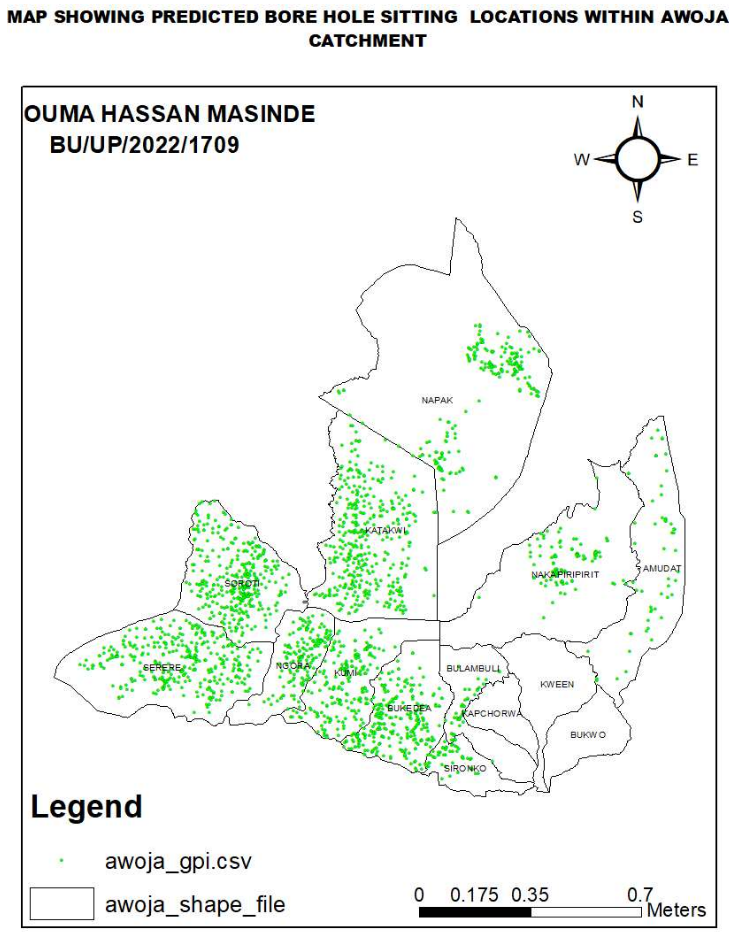


Figure 16 shows borehole locations in Awoja catchment

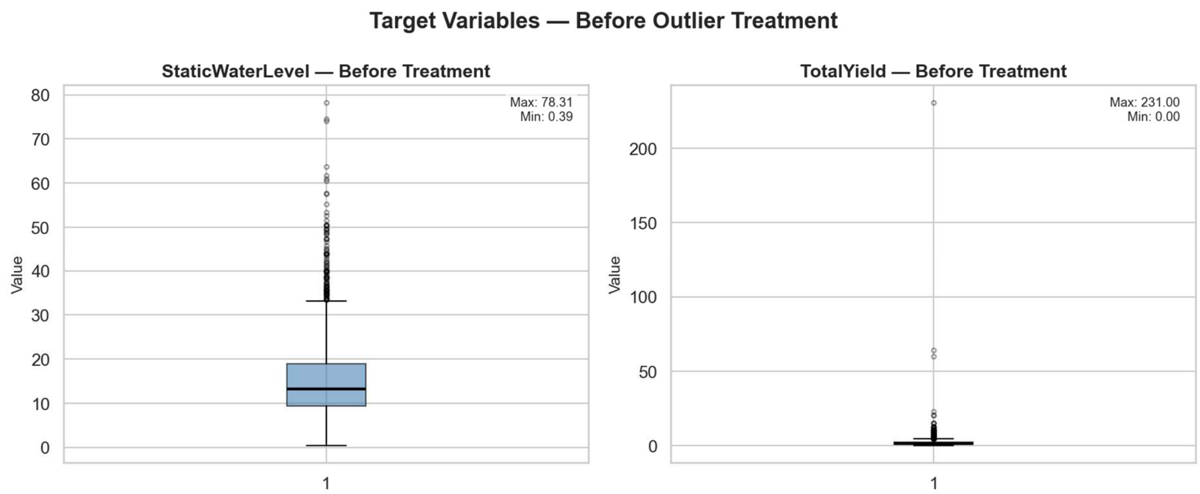


Figure 17 shows Box plots before outlier treatment

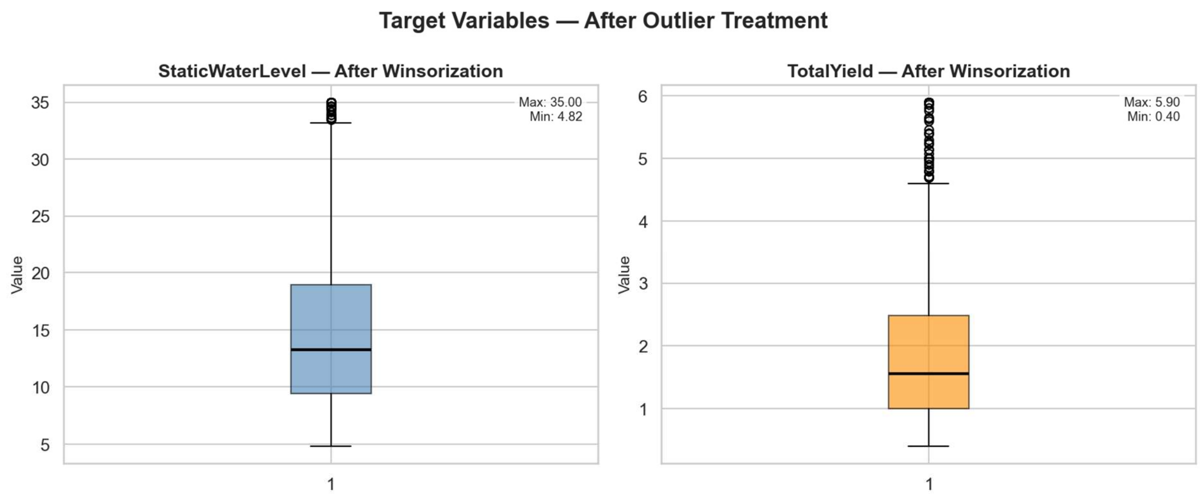


Figure 18 shows box plots after outlier treatment