
**UNIVERSITY OF RWANDA**

**COLLEGE OF SCIENCE AND TECHNOLOGY**

**AFRICAN CENTRE OF EXCELLENCE IN INTERNET OF THINGS (ACEIoT)**

# LOW-COST, LOW-POWER IoT-BASED SYSTEM AND MACHINE LEARNING MODEL FOR MONITORING AND PREDICTION OF GROUNDWATER QUANTITY

**PhD. Thesis submitted in the fulfilment of requirements of award of PhD Degree in Internet of Things –Wireless Intelligent Sensor Networking.**

**Omar Haji Kombo**

**AUGUST, 2022**


**UNIVERSITY OF RWANDA**

**COLLEGE OF SCIENCE AND TECHNOLOGY**

**AFRICAN CENTRE OF EXCELLENCE IN INTERNET OF THINGS (ACEIoT)**

# LOW-COST, LOW-POWER IoT-BASED SYSTEM AND MACHINE LEARNING MODEL FOR MONITORING AND PREDICTION OF GROUNDWATER QUANTITY

**PhD. Thesis submitted in the fulfilment of requirements of award of PhD Degree in Internet of Things –Wireless Intelligent Sensor Networking.**

**Omar Haji Kombo**
**Registration Number: 218014374**

**Thesis Supervisor: Prof. Santhi Kumaran, PhD.**

**Thesis Resident Supervisor: Dr. Emmanuel Ndashimye, PhD.**

**AUGUST, 2022**

**DECLARATION**

I Omar Haji Kombo do hereby declare that the dissertation entitled *"Low-Cost, Low-Power, IoT Based System and Machine Learning Model for Monitoring and Prediction of Groundwater Quantity"* to be submitted for the Degree of Doctor of Philosophy is my original work and the dissertation has not formed the basis for the award of any degree, diploma, associateship or fellowship of similar other titles. It has not been submitted to any other University or Institution for the award of any degree or diploma.

Signature: 

OMAR HAJI KOMBO

Omar Haji Kombo, a Ph.D. student of UR-ACEIoT student ID 218014374, successfully defended the thesis entitled "**LOW-COST, LOW-POWER IOT-BASED SYSTEM AND MACHINE LEARNING MODEL FOR MONITORING AND PREDICTION OF GROUNDWATER QUANTITY**", which he prepared after fulfilling the requirements specified in the associated legislations, before the thesis examination members whose signatures are below.

**Thesis Supervisor: Prof. Santhi Kumaran**

Copperbelt University, Zambia.

**Co-Supervisor: Dr. Emmanuel Ndashimye**

University of Rwanda, Rwanda.

**Viva Voce Members: Prof. Prof. Denis Ndanguza**

University of Rwanda, Rwanda.

**Dr. Uduak A. Umoh**

University of Uyo, Nigeria.

**Dr. Abdi T. Abdalla**

University of Dar es Salaam, Tanzania.

**Date of Defense : 29/8/2022**

***To my family,***

## FOREWORD

First and foremost, I want to express my gratitude to the Almighty God for his grace, support, and guidance throughout my PhD journey. I thank him for guiding me through all of the different journeys I have had to take over the last four years. Throughout my studies, I have been fortunate to receive a great deal of assistance, without which my research and thesis would not have been possible.

I deeply appreciate Prof. Santhi Kumaran, my principal supervisor, and my co-supervisors, Dr. Emmanuel Ndashimye and Engineer Alastair Bovim, who took on the challenge and accepted me as their PhD student. They supported me and contributed significantly to the intellectual content of this doctoral project at every stage; without their help, I doubt I would have been able to complete this research.

It would be difficult for me to pursue my doctoral studies if my employer, the State University of Zanzibar, did not grant me study leave and financial support. I am eternally grateful for this opportunity. I would also like to express my heartfelt appreciation and gratitude to Rwanda Water and Forestry (RWFA), Zanzibar Water Authority (ZAWA), Rwanda Meteorological Agency (MeteoRwanda), and Tanzania Meteorological Agency for their assistance and technical advice. I am very grateful to all my co-authors and the anonymous reviewers who provided very constructive comments for the improvement of my research articles.

I wish to thank my wonderful spouses, Kazija Khamis and Wasila Abdalla, for their patience, encouragement, and moral support during my PhD journey, without forgetting my children who were thousands of kilometers away from their father. In addition, I would like to express my profound gratitude to all my colleagues at the ACEIoT and the State University of Zanzibar, whose encouragement, seminar feedback, and comments influenced my research.

This research project acknowledges the financial support of the African Centre of Excellence in Internet of Things (ACEIoT) and the World Bank. Opinions expressed and conclusions reached are those of the author and should not be attributed to the ACEIoT or World Bank.

AUGUST, 2022 Omar Haji Kombo

**TABLE OF CONTENTS**

FOREWORD ..... xi

TABLE OF CONTENTS ..... xiii

LIST OF ABBREVIATIONS AND ACRONYMS ..... xix

LIST OF SYMBOLS ..... xxiii

LIST OF TABLES ..... xxv

LIST OF FIGURES ..... xxvii

ABSTRACT ..... xxxi

CHAPTER 1 ..... 1

INTRODUCTION ..... 1

1.1 General Introduction ..... 1

1.2 Research Background and Motivation ..... 2

1.2.1 Groundwater Situation in Rwanda's Eastern Province ..... 3

1.2.2 Rwanda's Groundwater Management Strategic Plan ..... 3

1.2.3 The Current Groundwater Monitoring Practice in Rwanda ..... 4

1.3 Statement of the Problem ..... 4

1.4 Research Aim and Objectives ..... 5

1.5 Methodology ..... 6

1.6 Rational for and Significance of the Research ..... 7

1.7 Context of the Research ..... 8

1.8 Thesis Approach and Outline ..... 8

CHAPTER 2 ..... 11

LITERATURE REVIEW ..... 11

2.1 Introduction ..... 11

2.2 Groundwater and its Significance in Life ..... 11

2.3 Climatic and Human Influence on Groundwater Variability ..... 12

2.4 Aquifer Management, Practices, Issues, and Constraints ..... 14
2.4.1 Monitoring Changes in Groundwater Table Depths ..... 15
2.4.2 The Importance of Long-term Groundwater Level Data and Information Sharing ..... 17
2.5 Wireless Sensor Networks (WSNs) and the Internet of Things (IoT) ..... 18
2.5.1 Primer on LoRa Radio Technology ..... 22
2.5.2 Low-power, Low-cost WSNs and the Open Designs ..... 23
2.5.3 Powering and Sustaining the WSNs ..... 24
2.5.4 Use of the Internet of Things for Groundwater Table Monitoring ..... 25
2.6 Artificial Intelligence and Its Role in Hydrology ..... 27
2.6.1 Machine Learning and Appropriate Frameworks and Tools ..... 27
2.6.2 Machine Learning as Water Management Tool ..... 28
2.6.3 Machine Learning Based Predictive Modeling and its Role in Groundwater Analysis ..... 29
2.6.3.1 K-Nearest Neighbor ..... 30
2.6.3.2 Artificial Neural Network ..... 31
2.6.3.3 Support Vector Machine ..... 31
2.6.3.4 Random Forest ..... 33
2.6.3.5 Ensemble Mode Decomposition ..... 35
2.7 The Use of Machine Learning and IoT for Groundwater Management ..... 36
2.8 Research Methodological Approach ..... 37
CHAPTER 3 ..... 39
DEVELOPMENT AND VALIDATION OF A LOW-COST, LOW-POWER ENERGY-HARVESTING, LoRa-GSM, IoT SYSTEM FOR MONITORING GROUNDWATER LEVELS ..... 39
3.1 Introduction ..... 39
3.2. Contributions ..... 40
3.3. Materials and Methods ..... 40
3.3.1. Study Site Description ..... 41

3.3.2 Description of the System Design and Realization ..... 42
3.3.3 Preparation of Water Table Depth Probe ..... 47
3.3.4 Data Correction and Storage Description ..... 50
3.3.5 Cloud Server and Data Visualization ..... 51
3.3.6 Energy Autonomy of the LWNGM ..... 51
3.4. Experiment Evaluation and Test Results ..... 53
3.4.1 LWNGM System Deployment ..... 54
3.4.2 Configuration, Key Parameters, and Network Performance ..... 56
3.4.3 System Testing and Data ..... 56
3.4.4 Cost and Simplicity of the Systems ..... 59
3.4.5 Energy and Lifetime of the End Node ..... 60
3.5 Discussion ..... 62
3.6 Conclusion ..... 65
CHAPTER 4 ..... 67
LONG-TERM GROUNDWATER LEVEL PREDICTION MODEL BASED ON HYBRID KNN-RF TECHNIQUE ..... 67
4.1 Introduction ..... 67
4.2. Contributions ..... 68
4.3 Case Study and Data Processing ..... 68
4.3.1. Study Area and Data ..... 68
4.3.2 Data Preparation ..... 71
4.3.3 Model Performance and Evaluation Measures ..... 75
4.4 Realization and Evaluation of the Ensemble KNN-RF Method ..... 76
4.4.1 Development of the KNN-RF Ensemble ..... 78
4.4.2 Tuning Parameter and Input Selection ..... 78
4.4.3 Training and Testing of the Model ..... 80
4.4.4. Prediction of Seasonal Changes in Groundwater Depths ..... 81

4.5 Experimental Results and Discussion ..... 81
4.6 Conclusions ..... 86
CHAPTER 5 ..... 89
AN ENSEMBLE MODE DECOMPOSITION COMBINED WITH SVR-RF MODEL FOR PREDICTION OF GROUNDWATER LEVEL: THE CASE OF EASTERN RWANDAN AQUIFERS ..... 89
5.1 Introduction ..... 89
5.2 Contributions ..... 90
5.3 Materials, Tools and Methodology ..... 90
5.3.1 Case Study and Available Data ..... 90
5.3.2 Initial Data Preparation and Tools ..... 93
5.3.3 Methods ..... 94
5.4 Results and discussion ..... 98
5.5 Conclusions ..... 103
CHAPTER 6 ..... 105
RESEARCH OUTCOMES, CONCLUSIONS, AND RECOMMENDATIONS ..... 105
6.1 Introduction ..... 105
6.2 Summary of the Outcomes ..... 105
6.2.1 Development of an Affordable IoT System for Groundwater Monitoring ..... 105
6.2.2 Field Validation and Evaluation of the Affordable IoT Groundwater Monitoring System ..... 106
6.2.3 Development of an Efficient Machine Learning Method for Seasonal Groundwater Quantity Prediction ..... 107
6.2.4 Evaluation of the ML techniques for Seasonal Water Table Forecasting ..... 108
6.3 Limitations of the Research ..... 108
6.4 The Implications of Affordable IoT and ML technologies Outcomes ..... 108
6.5 Recommendations for LCSN and ML in Groundwater Hydrology and Research ..... 109
6.5.1 Adoption of Affordable and IoT-based Groundwater Management Tools ..... 109

6.5.2 Recommendation for Further IoT and ML Research in Aquifer Management.....109
6.6 Conclusion.....110
REFERENCES.....111
APPENDIX A ..... 131
PUBLICATIONS..... 131
A.1 Published Papers .....131

## LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| **6LoWPAN** | : IPv6 over Low Power Wireless Personal Area Network |
| **AES** | : Advanced Encryption Standard |
| **AI** | : Artificial Intelligence |
| **ANN** | : Artificial Neural Network |
| **ASCE** | : American Society of Civil Engineers |
| **BW** | : Bandwidth |
| **Cr** | : Code Rate |
| **CSS** | : Chirp Spread Spectrum |
| **dB** | : Decibel |
| **DOAJ** | : Directory of open access journals |
| **DR** | : Data Rate |
| **DSR** | : Design Science Research |
| **ELM** | : Extreme Learning Machine |
| **EMA** | : Exponential Moving Average |
| **FEC** | : Forward Error Correction |
| **FSK** | : Frequency Shift Keying |
| **FUOTA** | : Firm-ware Update Over –The-Air |
| **GE** | : Genetical Programming |
| **GFSK** | : Gaussian Frequency Shift Keying |
| **GMSK** | : Gaussian Minimum Shift Keying |
| **GoR** | : Government of Rwanda |
| **GPRS** | : Global Packet Radio Service |
| **GSM** | : Global System for Mobile |
| **I2C** | : Integrated-Integrated Communication |
| **ICT** | : Information Technology |
| **IEEE** | : Institute of Electrical and Electronic Engineers, Inc. |
| **IoT** | : Internet of Things |

**IoT4GRW** : Internet of Things for Groundwater Resources

**IP** : Internet Protocol

**ISM** : Industrial Scientific and Medical Band

**IT** : Information System

**JSON** : JavaScript Object Notation

**KNN** : K-Nearest Neighbor

**KNN-RF** : K-Nearest Neighbor Random Forest

**LAN** : Local Area Network

**LCS** : Low Cost Sensors

**LCSN** : Low-cost Sensor Network

**LGMWN** : Low cost, low power Groundwater

**LiPo** : Lithium Polymer battery

**LoRa™** : Long-range Communication

**LoRaWAN** : Long-range Wide Area Network

**LPWAN** : Low Power Wide Area Network

**LTE** : Long Term Evolution

**MA** : Moving Average

**MAC** : Media Access Control

**MAE** : Mean Absolute Error

**MCU** : Microcontroller

**MDPI** : Multidisciplinary Digital Publishing Institute

**MeteoRwanda** : Rwanda Meteorological Agency

**ML** : Machine Learning

**MWSN** : Monitoring Wireless Sensor Network

**MQTT** : Message Queuing Telemetry Transport

**MQTT** : Message Queuing Telemetry Transport

**MSE** : Mean Absolute Error

**MSK** : Minimum Shift Keying

**NSE** : Nash-Sutcliffe Efficiency

**OOK** : On-off Keying

**PVC** : Polyvinyl Chloride

**$R^2$** : Coefficient of determination

**RF** : Random Forest

**RMSE** : Root Mean Squared Error

**RSS** : Received Signal Strength Indicator

**RWFA** : Rwanda Water Forestry Authority

**SF** : Spread Factor

**SGP** : Smart Groundwater Portal

**SNR** : Signal to Noise Ratio

**SPI** : Serial Peripheral Interface

**Station ID** : Station Identity Number

**SVM** : Support Vector Machine

**TMA** : Tanzania Meteorological Agency

**TP** : Transmission Power

**USB** : Universal Serial Bus

**UTP** : Untwisted Pairs

**WiFi** : Wireless Fidelity

**WLAN** : Wireless Local Area Network

**WSN** : Wireless Sensor Network

**ZAWA** : Zanzibar Water Authority

## LIST OF SYMBOLS

| | |
|---|---|
| **A** | **:** Ampere |
| **C** | **:** Capacitance |
| $\widehat{\boldsymbol{D}}$ | **:** Distance |
| **E** | **:** Energy |
| $\boldsymbol{\epsilon}$ | : Error value |
| $\boldsymbol{\alpha}$ | : Decay value |
| **J** | **:** Joule |
| **M** | : Matrix |
| **T, t** | **:** Lag time in days |
| **V** | **:** Volt |
| **W** | **:** Watt |
| $\boldsymbol{w}$ | **:** Weight vector |

## LIST OF TABLES

Table 2.1 Comparisons of LPWAN technologies for large-scale IoT implementation ..... 21
Table 2.2 RFM95W modem specification ..... 23
Table 2.3 Analysis of the selected closely related studies for comparison ..... 26
Table 3.1 Selected components for the development of the LGMWN ..... 42
Table 3.2 Configured LoRa transmission parameters ..... 56
Table 3.3 Performance metrics for the LGWMN LoRa network ..... 56
Table 3.4 Summary of the quantity and pricing of the LWNGM components ..... 59
Table 4.1 Summary of the selected monitoring well and its main features ..... 69
Table 4.2 Performance evaluation results for 15, 30, 60, and 90 days lead-time groundwater level variations using the KNN-RF model. ..... 82
Table 4.3 NSE performance evaluation results for 15, 30, 60, and 90 day-ahead groundwater level predictions using the KNN-RF model. ..... 83
Table 4.4 Summary of the selected parameters during training of the SVR, ANN, and KNN-RF models. ..... 86
Table 5.1 Characteristics of Rugarama and Mukarange aquifers**.** ..... 92
Table 5.2 Training and test data portions for Rugarama and Mukarange datasets ..... 97
Table 5.3 Performance of EEMD-SVR-RF on Mukaange and Rugarama dataset ..... 99

**LIST OF FIGURES**

Figure 2.1 Multiple water pumps installed in a community well. ..... 12
Figure 2.2 The layout of a four-stage IoT solutions architecture..... 20
Figure 2.3 The comparison of data rate and range capacity of LPWAN radio technologies based on positioning..... 21
Figure 2.4 An illustration design consideration for an energy harvesting system ..... 25
Figure 2.5 A general schematic representation of the design science research approach [164]. ..... 38
Figure 3.1 Location of bandamaji monitoring well in (a) Africa (b) Tanzania and (c) Zanzibar (Unguja) - Donge, Mnyimbi ..... 42
Figure 3.2 The functional scheme of data gathering and storage. ..... 44
Figure 3.3 Block scheme of the LWNGM's field node architecture. ..... 46
Figure 3.4 The LWNGM field node (a) the potted MS5803-14A and protected in aluminum cylinder (b) the field-node circuitry in a water-proof enclosure (top removed) (c) the MS5803-14A potted in a plastic container, housed in a PVC cylinder (d) the damaged MS5803-14A sensor in a potted plastic container. ..... 49
Figure 3.5 The I2C differential interfacing connectors attached to each end of the Cat 6 UTP sensor connection cable (25 m)..... 49
Figure 3.6 Solar devices for powering the nodes (a) 3.7V LiPo battery (b) solar panel (c) charger unit for batteries. ..... 51
Figure 3.7 LWNGM platform: the well, devices, sensors, power source, and communication protocols deployment. ..... 53
Figure 3.8 Field deployment of LWNGM (a) LG01-P long-range gateway with 4G LTE USB dongle (b) field node in a waterproof enclosure with mini solar panel deployed mounted to a pole at Bandamaji monitoring well. ..... 55
Figure 3.9 Comparison of LoRa radio performance statistics in terms of RSSI, PDR, and airtime for various studies. ..... 57
Figure 3.10 (a) A plot of daily average water depths at bandamaji monitoring well over a two-week deployment period. ..... 58
Figure 3.10 (b) Web dashboard for LWNGM showing a plot of two weeks data from Bandamaji station at 6-hour intervals. ..... 58

Figure 4.1 Map of Rwanda shows the location of the groundwater monitoring wells and weather stations in the eastern province..... 70
Figure 4.2 Eastern Province, Rwanda. Map shows hydrogeological features of the study area [76]. ..... 71
Figure 4.3 Time series plot of precipitation and groundwater level collected from the Mukarange monitoring borehole. ..... 73
Figure 4.4 Time series plot of temperature and groundwater level collected from the Mukarange monitoring borehole ..... 74
Figure 4.5 Time series plot of solar radiation and groundwater level collected from the Mukarange monitoring borehole ..... 74
Figure 4.6 Flow chart of the KNN-RF method ..... 77
Figure 4.7 Diagram of the time series 4-sliding window validation method. Adapted from [208]. ..... 80
Figure 4.8 Comparison of the performance obtained by the SVR, ANN, RF, KNN, and KNN-RF models on groundwater level prediction for: (a) 15 day lead-time, (b) 30 day lead-time (c) 60 day lead-time, and (d) 90 day lead-time..... 82
Figure 4.9 Assessment of the actual and the estimated groundwater levels of the optimal KNN-RF model for 15, 30, 60, and 90 days (corresponding to a–d, respectively) lead time in the testing phase. ..... 84
Figure 4.10 Comparison of the 15, 30, 60, and 90 days (corresponding to a–d, respectively) estimated and the actual levels yielded by the optimal KNN-RF technique..... 85
Figure 5.1 Study area and location of groundwater stations (Rugarama and Mukarange) and Kawangire weather station in Eastern Province- Rwanda. ..... 91
Figure 5.2 The relationships between Rugarama groundwater level and the weather parameters (a) air humidity (b) solar radiation (c) precipitation (d) air temperature ..... 93
Figure 5.3 The relationships between Mukarange groundwater level and the weather parameters (a) air humidity (b) solar radiation (c) precipitation (d) air temperature..... 93
Figure 5.4 Schematic layout of EEMD –SVR-RF method. ..... 95
Figure 5.5 Decomposed levels for (a) Rugarama and (b) Mukarange groundwater data. ..... 96
Figure 5.6 Evaluation outcome for the EEMD-SVR, EEMD-ANN, EEMD-SVM and EEMD-SVR-RF models for the nighty-day lead-time prediction of groundwater level at Rugarama observatory station: (a) RMSE metric (b) Rsquared metric (c).MAE metric (d) NSE metric. ..... 100

Figure 5.7 Mukarange regression (a) at 15- day horizon (b) 30-day horizon (c) 60-day horizon (d) 90-day horizon. ..... 101
Figure 5.8 Rugarama regression (a) at 15-day horizon (b) 30-day horizon (c) 60-day horizon (d) 90-day horizon ..... 102
Figure 5.9 Observed and estimated water table depth at (a) Rugarama observatory (b) Mukarange observatory station ..... 103

**ABSTRACT**

Groundwater is the most dependable source of freshwater supply on the planet. Meanwhile, the prohibitively expensive cost of groundwater data loggers, telemetry, and data analysis tools makes responsive groundwater management difficult. The situation is exacerbated in Sub-Saharan African countries with a severe lack of groundwater data. In Rwanda, for example, groundwater level data is collected using standalone sensors via field patrols, and there are no effective tools for analyzing this data. This research aimed to develop a low-cost, low-power Internet of Things (IoT)-based system and Machine Learning (ML) model for monitoring and predicting groundwater quantity in order to provide managers and other stakeholders with inexpensive tools for groundwater management in Eastern Rwanda. Historical hydro-climatic data were obtained from the Rwanda Meteorological Agency (MeteoRwanda) and the Rwanda Water and Forestry Authority (RWFA), and system specifications were obtained from managers and other stakeholders. A data logger with a submersible water table depth probe was built using redesigned low-cost MS5803-14BA and MBE280 sensors, an improved I2C interface, a real-time clock, a microSD module, mini solar charger, and an ATmega328P-based framework. A low-power, long-range telemetry system was developed using an open-source Dragino LoRa transceiver and a 4G LTE dongle. The system was deployed at the Bandamaji groundwater station for two weeks, allowing for near real-time data collection, analysis, and validation of its power consumption, cost, and network efficiency. The findings show that the system has a relatively low cost of around USD 310.168, a promising efficacy with a daily energy consumption of about 12% of the battery's capacity of 66,600J. The network performance is 84.46% for PDR, -83 for RSSI, and each send takes about 37.13 seconds. Predictive analytics tools were developed by combining RF with SVR and KNN methods in order to improve prediction efficiency and accuracy. These ensemble machine learning techniques were calibrated and tested using well-prepared datasets. Multiple hyper-parameters and lagged inputs were also tested iteratively until the best results were obtained. The EEMD-SVR-RF technique improves prediction accuracy ($R^2$) at 90-day lead time by 5.1832%, 49.8543%, and 2.5083%, respectively, when compared to SVR, ANN, and RF methods. In addition, when compared to other models, this model also has the smallest errors of 0.0038 m for MAE, and 0.0011 m for RMSE. Moreover, the SVR-RF with EEMD preprocessing outperforms the EMA-KNN-RF with an $R^2$ of 0.9608, RMSE of 0.0011, MAE of 0.0382, and NSE of 0.9586. These results are comparatively better and more

insightful for groundwater management and the advancement of IoT and AI-based hydrology solutions.

# CHAPTER 1

# INTRODUCTION

## 1.1 General Introduction

Groundwater is the most important source of freshwater for a variety of uses. This resource is critical for combating poverty, ensuring water and food security, creating decent jobs, socio-economic development, and adapting to climate change. Because of its underground nature, it is frequently undervalued, mismanaged, and abused. In some areas, the rate at which groundwater is abstracted is alarming threatening human life and the ecosystem [1]. Groundwater must be preserved for current and future generations. This goal is attainable if there is reliable information to guide policy and resource governance. Such information can be obtained through continuous and effective monitoring of groundwater utilization and changes.

Many countries, particularly developing countries, lack access to tools for observing and modeling aquifers. Cost and technical skills are major factors preventing these tools from reaching low-income countries [2], [3] The Internet of Things (IoT) and Artificial Intelligence (AI) technologies that are emerging have the potential to revolutionize groundwater management. The thesis argues that inexpensive automatic IoT-enabled groundwater monitoring improves accuracy, efficiency, and reliability of groundwater levels time series data. It also argues that open source predictive machine learning analysis produces reliable seasonal groundwater forecasts.

In relation to the current research topic, this chapter aims to provide the motivations for the study, describe its background, and outline the statement of research aim and objectives. First, it introduces the background and motivation of the research, as well as the potential of IoT and ML technologies for groundwater management. The research problem statement is followed by a discussion of the research objectives in relation to the overall purpose of the study. Finally, it presents a summary of the contributions of the thesis as well as its overall structure.

## 1.2 Research Background and Motivation

Groundwater is the world's largest (more than 97%) and most important unfrozen fresh water reserve, serving as a major source of water supply in many countries, particularly in arid and semi-arid regions, for drinking, irrigation, and animal husbandry [1]. This resource has a high potential for helping to achieve the United Nations (UN) Sustainable Development Goals (SDGs), particularly SDG6 [4], which states that water should be accessible to all and managed sustainably. Groundwater is critical to the fight against poverty, food and water security, the creation of decent jobs, socioeconomic development, and society and economy resilience to climate change [5]. In a broader sense, groundwater is linked to more than 50% of the SDG targets [6]. Meanwhile, while groundwater is the most exploited raw material on the planet, it is under-monitored and under-managed [6]. Mismanagement of groundwater, as well as its frequent abuse through over-exploitation, endangers the entire water cycle, and thus endangers human well-being and the survival of all life [7]. Nearly half of the world's population relies on groundwater as their primary source of drinking water, and groundwater accounts for approximately 43% of the water used in agriculture and food production [4]. It also serves as a major supply for the global ecology's preservation and sustainability. Global water demand is steadily increasing, resulting in groundwater scarcity. Almost half the world's major aquifers are at risk of depletion and drying out [4][5]. One of the driving forces behind the lowered water tables is unsustainable aquifer abstraction.

In Africa, while more than 75% of the population is dependent on groundwater for basic water supply, there is growing interest and excitement about the potential of groundwater for irrigation [7][8]. Africa's groundwater demands are expected to rise substantially as the continent's population and economy expand at a rapid pace, but knowledge about this resource is limited [8]. With recurring draughts causing human and animal deaths as well as economic consequences, Sub-Saharan Africa's population reliant on groundwater grows during dry seasons. The East African drought of 2011 resulted in food shortages for over ten million people and over 260,000 deaths [8]. Groundwater supplies approximately 80% of Rwanda's fresh water, with Eastern Province being the most reliant on this resource [9]. To meet rising demands and natural influences, improved and adaptive groundwater management strategies are required.

### 1.2.1 Groundwater Situation in Rwanda's Eastern Province

Eastern province is typically composed of mountains, hills, and wide deep valleys, the floors of which are 200 meters below the surrounding hills and mountains. Historically, this region has been plagued by recurring droughts and water scarcity [8]–[12], as it receives approximately 25% less annual rainfall than the rest of the country [11][13]. As a result of this situation, the government developed and made available all of the region's springs. The water from these springs is of acceptable quality, but yields are low, especially during the long dry season, when some of the springs may run dry. As a result, the use of springs in this region is extremely difficult, leaving groundwater wells as the most viable solution to water scarcity [14]. This province has the most drilled wells, with at least 400 boreholes and wells as of 2009 [10]. The test wells were drilled for the first time between May and October 1985 in order to facilitate a study aimed at the development of groundwater aquifers in this zone [14]. Since 1994, the eastern province has seen extensive borehole drilling and shallow well construction [10]. With a large number of wells, as this part of the country receives less than 1000 mm of rain on average, and the average evaporation rate is estimated to be between 1,000 and 1,500 mm [6], production is limited, forcing resource managers to devise new methods of managing water resources equitably and sustainably. Nonetheless, there has been a rapid depletion of aquifers over the last three decades due to higher abstraction rates and the greatest climate change impact on the Eastern Province [4][12]. These effects of climate variability are likely to worsen, resulting in longer drought periods. Aquifers will undoubtedly continue to be a major source of fresh water at demand points in this region.

### 1.2.2 Rwanda's Groundwater Management Strategic Plan

Rwanda's water management policy (enacted in 2010 and reviewed on 04/2011), aims to ensure that the country's groundwater resources are managed in a holistic and long-term manner, utilizing an Integrated Water Resources Management (IWRM) approach. The policy aims to achieve three major goals: (i) to protect, conserve, manage, and develop Rwanda's groundwater resources in an integrated and sustainable manner; (ii) to ensure that groundwater resources are available in sufficient quantity and quality for current and future generations' socio-economic and ecological needs; and (iii) to ensure that decisions affecting groundwater resource management are made in a coordinated manner and with the participation of all stakeholders at local, national, and international levels.

Among other things, the policy directs the use of cost-effective groundwater resource assessment and monitoring systems, recognizing that rational decisions cannot be made without reliable information and the ability to apply knowledge appropriately [10]. On this basis, the strategic plan [10], was created to provide translation and implementation of the water management policy.

### 1.2.3 The Current Groundwater Monitoring Practice in Rwanda

Even though groundwater accounts for the majority of managed water resources in many low-and middle-income countries, hydrogeological capacity is lacking. This frequently includes both technical and institutional capacity [7]. The current groundwater monitoring practice in Rwanda employs standalone pressure transducer sensors, and seasonal field patrols are conducted to collect data from those sensors. This practice necessitates significant amount of human, material, and financial resources and may result in inaccuracies and missing data.

The Internet of Things (IoT) and Machine Learning (ML) are potential technologies for improved groundwater resources management. The availability of inexpensive, low-power wireless sensors and telemetry opens up new avenues for groundwater hydrology research and development. IoT offers ground-based, continuous, and remote monitoring, as well as data analytics that can be utilized to inform groundwater decisions.

## 1.3 Statement of the Problem

The conventional groundwater data logging instruments have high power consumption, are expensive and difficult to deploy logistically [2], [3]. In addition to that, most groundwater data analysis and decision support tools have prohibitively high licensing costs and are only useful for short-term groundwater predictions [15]. These factors contribute to unsustainable aquifer management by causing irregular and inefficient monitoring, as well as ineffective analysis of groundwater responses. Although, groundwater monitoring and forecasting have been the focus of extensive hydrology research, but the long-term hydrological data are scarce in most parts of the world [9]– [11], owing in part to a lack of affordable groundwater monitoring tools, making groundwater evaluation and groundwater decision-making difficult [13], [14], [16], [17]. Currently, groundwater monitoring in Rwanda is done with stand-alone water level sensors, and there is no effective seasonal forecast of groundwater variability [3]. These factors limit our ability to mitigate the effects of droughts and water scarcity. Innovative solutions for efficient and cost-effective

groundwater management are critical in order to provide managers and policymakers with easily accessible information [16].

This research focused on the development of an energy-efficient, affordable IoT system for monitoring of groundwater tables, as well as an efficient machine learning model for seasonal prediction of groundwater availability, in order to provide groundwater managers and other stakeholders with relevant and reliable information.

## 1.4 Research Aim and Objectives

The research aimed to develop reliable IoT and ML-based tools for monitoring and predictive-analysis of groundwater levels in order to aid in sensible decision-making in sustainable groundwater resource management.

Specifically, the research will attain the following objectives:

i. To obtain requirements for the Internet of Things-enabled groundwater table depths monitoring system.
ii. To design and develop a low-cost, low-power, energy-harvesting Wireless Sensor Network (WSN) for remote and near real-time groundwater level monitoring.
iii. To validate the developed WSN's affordability, energy efficiency, and network efficacy through field deployment.
iv. To develop an efficient machine learning model for seasonal prediction of groundwater levels.
v. To evaluate the developed machine learning model using appropriate performance metrics.

The present research sought to leverage the potential of emerging open source (IoT and ML) technologies to create and evaluate two tools: an IoT-enabled groundwater level monitoring system and machine learning for seasonal groundwater level prediction, as well as empirical evaluation of these tools. This solution is referred to as the Internet of Things for groundwater (IoT4GRW). Hence, the primary contribution of this research is to provide novel, affordable, automated, and efficient groundwater management tools to support the findings that groundwater is undermanaged [6], in many parts of the world, particularly Rwanda [18], due to the high cost of instrumentation and a lack of appropriate technical skills[3], [18]. It also contributes to a better understanding of methods for IoT-enabled environmental monitoring and forecasting solutions.

## 1.5 Methodology

The research was carried out in four systematic phases. It began with a review of existing tools and methods for monitoring and analyzing groundwater responses to natural and anthropogenic effects in Africa and elsewhere. Groundwater variables as well as the potential influencing factors were identified. Different types of groundwater monitoring practices were considered, as well as their efficiency, cost and energy implications, simplicity, and limitations. In addition to that, the effectiveness, accuracy, interpretability, and resource requirements of the existing groundwater predictive analysis techniques were also evaluated. The preliminary literature review was followed by requirements gathering from groundwater engineers, managers, and other stakeholders. Furthermore, weather and water table data were collected, exploratory analyses were performed, and environmental data treatment methods were applied.

In the second phase, machine learning predictive algorithm for groundwater was conceptualized, designed, developed and evaluated. This phase began with data preparation to create the dataset (combination of real-world weather and water level data) for each groundwater station, where different data treatment methods were employed. The dataset comprises temperature, precipitation, solar radiation, humidity, and groundwater level data from three groundwater stations in Rwanda's Easter Province. Pre-processing was also performed to ensure that the data is suitable for the selected models. Then, the best performing machine learning algorithms (KNN, RF, SVR, and ANN) were selected, iteratively trained, and compared. Following that, a baseline in model performance was established to provide a point of comparison when comparing these models to their hybrid versions. A hybrid and ensemble predictive modeling approaches were used to improve the efficiency and accuracy by reducing the models' generalization errors on training set. Thereafter, all hybrid models were iteratively calibrated and tested with various hyper-parameters and lagged inputs until the best results are obtained. The models were evaluated using the MAE, NSE, RMSE, and $R^2$ performance metrics. All model development was done in an open source python environment, and all testing and comparisons were done with a python-based hydrology framework. Based on the comparison of these models, the model with the highest accuracy in seasonal (3 moths) groundwater level prediction was determined, chosen and applied to prediction.

The third phase involved the design, development, and application of a low-cost, low-power IoT-enabled system for continuous observation of groundwater levels. The literature and requirements

gathered from stakeholders influenced the system's design. The designed solution was realized using an open source development board and software tools. Stakeholders were also involved in the prototype's development and testing. The high resolution low-cost pressure sensors were redesigned, and the I2C protocol's transmission distance was improved to create an inexpensive submersible groundwater table depths probe. The sensors chosen are low-power, and the solar energy was used to provide the nodes with energy autonomy. The monitoring system includes a GIS-enabled web-visualization portal for data sharing. The prototype was then deployed at the groundwater observatory station for field testing.

The fourth phase included an evaluation of the efficacy, cost, power, and energy expenditure, as well as their implications for the applicability of the developed groundwater monitoring prototype.

## 1.6 Rational for and Significance of the Research

The present research is claimed to be significant in that it sought not only to address the lack of continuous, affordable, and effective groundwater management tools but it also offers in depths insights into the design, development, empirical analysis of WSN and ML models, as well as the adoption of LoRa technology in groundwater observation and improvement of I2C WSN protocol. The research is also consistent with Rwanda's water management policy (Enacted in 2010) and the United Nations Sustainable Development Goals (SDGs-2030). The outcomes could inform the actions that Rwanda's Ministry of natural resources may need to take to improve groundwater resource management. Moreover, the current research focuses on the development of both an affordable IoT-enabled groundwater level monitoring system and an open-source machine learning groundwater level predictive model in Rwanda (whereas many studies, for example, the studies cited above do not), so this study is expected to contribute to the literature on methods for efficiently developing, evaluating, and utilizing IoT and ML technologies in groundwater management at an affordable cost.

Lastly, the Rwandan government intends to increase groundwater usage through irrigation expansion [11], while the Ministry of natural resources is working to increase data collection and interpretation to improve understanding of the available quantity of water resources [16]. Hence, the current research is both timely and important in terms of providing managers with appropriate tools that facilitate the provision of appropriate information to policymakers.

## 1.7 Context of the Research

The current research was carried out in Rwanda's Eastern Province. This province's primary source of freshwater is groundwater, and it is prone to recurring droughts [12]. While it receives 600-700 mm of annual rainfall, which is 25% less than the rest of the country [13], it also has a high rate of evaporation ranging from 1000 to 15000 mm. In this province, there are three groundwater observatory stations: Ruhuha, Mukarange, and Rugarama, as well as three weather stations: Nyagatare, Kawangire, and Ngoma. RWFA (which is under Rwanda's Ministry of Natural Resources) manages the groundwater stations, while MeteoRwanda manages the weather stations. RWFA installed stand-alone diver sensors and barometric sensors in December 2016, which are used to sample groundwater water column pressure and atmospheric pressure at 12-hour intervals. This information is gathered through seasonal (3-4 month) patrols in the field. MeteoRwanda provides weather data relevant to this study (air temperature, humidity, solar radiation, and precipitation). In current practice, groundwater data are not linked to climatic information, and only simple analysis is performed on groundwater data.

Although our case study is in Rwanda's Easter Province, the WSN prototype was deployed in Tanzania (at the Bandamaji groundwater station) due to the COVID-19 quarantine, which prevented the researcher from accessing the original case study in Eastern Rwanda. Bandamaji station was chosen for two reasons: first, it is located in the North region of Zanzibar-Tanzania, which has higher groundwater consumption, particularly for rice cultivation [19], and second, it was the only properly secured area among the North region's groundwater stations during the research period. The station's round-the-clock security ensures the safety of the WSN's devices.

## 1.8 Thesis Approach and Outline

This thesis adapted its author’s published research papers. The first chapter provided the background and motivation for the research on IoT and ML technologies for groundwater management in Rwanda. It also describes the groundwater situation in Rwanda's Eastern Province, Rwanda's groundwater management strategic plan, the current groundwater monitoring practice in Rwanda, a statement of the problem, the objectives of the study, and the rationale for and significance of the study. The context of the study is also provided, and the chapter concludes with a thesis outline.

Chapter 2 examines the relevant literature on the status of groundwater resources in eastern Rwanda, groundwater management practices, emerging technologies, and research methodology approaches. Chapter 3 describes the design and implementation of a low-cost, low-power, energy-harvesting, Internet of Things-enabled groundwater monitoring system (achieved research objectives 1, 2 and 3). Chapter 4 describes the machine learning hybrid KNN-RF approach for long-term forecasting of groundwater levels (achieved objectives 4 and 5). Chapter 5 focuses on an advanced machine learning prediction method that combines EEMD and a hybrid SVR-RF model. It adds two new groundwater monitoring stations to the groundwater prediction network and improved forecasting efficacy and accuracy (also achieved objectives 4 and 5).

Chapter 6 concludes the thesis by providing a summary of the research outcomes and conclusions drawn from those findings discussed in Chapters 3, 4, and 5. This chapter discusses the limitations of the current study, the implications of the research findings, and recommendations and suggestions for future research in this field.

Along the lines of the thesis outline, the following chapter provides a review of the literature that informed the current research.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Introduction

This chapter reviews relevant literature from African and global contexts that inform groundwater management research through monitoring and forecasting using IoT and ML technologies. The following subheadings will be used to organize this review of related literature.

## 2.2 Groundwater and its Significance in Life

Groundwater is a common pool resource that is stored beneath the earth's surface. This limited resource is stored in aquifers, which are areas of permeable rocks. The boundary between water-saturated and unsaturated ground is defined by the water table. Water is abundant beneath the water table in the rocks and soil [20]. Groundwater is critical to sustaining ecosystems and enabling human adaptation to climate variability and change because it is the world's largest distributed store of fresh water. Aquifers are the primary source of freshwater supplies on a global scale. Groundwater serves nearly half of the world's population and provides half of the water used in agriculture and food production. This resource is critical to development agendas such as the UN Sustainable Development Goals (SDGs) 2 and 6 to end hunger, achieve food security, and promote sustainable agriculture by 2030. In Africa, where rainfall and river discharge are among the most variable on the planet, groundwater is a critical source of freshwater [21]. Groundwater, as opposed to surface waters, provides a distributed, relatively low-cost, and climate-resilient source of freshwater to meet rapidly growing demand in Sub-Saharan Africa (SSA) associated with expanding access to safe water and improving food security through irrigation. The rate at which aquifers are depleted is alarming; Figure 2.1 depicts multiple water pumps installed in a single well, demonstrating the possibility of a water table overdraft in our case study.

Figure 2.1 Multiple water pumps installed in a community well.

Groundwater, for example, supplies more than 75 % of Africa's freshwater supplies and 80% of Rwanda's water sources. To ensure the long-term viability of this resource in SSA, and particularly in Rwanda, frameworks for institutional coordination and stakeholder participation, information management programs, and capacity building are required [22], [23]. Groundwater resource is likely to become more strategic for global water and food security as more frequent and intense climate extremes (droughts and floods) increase variability in precipitation, soil moisture, and surface water [24]. However, over the last century, the use of this resource has also increased dramatically [15]. This implies that improved livelihoods and climate change adaptation will be much more difficult to achieve if available groundwater resources are not used effectively.

The subheadings that follow provide in-depth reviews of the factors influencing groundwater hydrology as well as the methods used to address the aforementioned challenges.

## 2.3 Climatic and Human Influence on Groundwater Variability

It is widely acknowledged that there is a strong relationship between climate variability and groundwater quantity. According to UN-Water [23] and the Intergovernmental Panel on Climate Change (IPCC) [24], climate change has a direct impact on the groundwater cycle and, as a result, on the quantity and quality of groundwater resources available to meet human and environmental

demands. Over time, the wider scientific community has come to realize that single discipline studies cannot adequately answer management questions [25], prompting scientists to incorporate climatic factors into water table modeling. These scientific and societal efforts are primarily intended to characterize and comprehend the response of aquifers to the effects of weather and climate uncertainty as a result of global warming [26]–[28]. Effective water table assessment and projection are impossible without taking weather influences into account. Following that, the future risks to groundwater resources from global change are typically assessed by driving hydrological models with climate model outputs [22].

More research [29][23][24] is confirming the noticeable effects of climate variability on droughts and aquifers. Studies indicate that there are strong spatial correlations among observation sites (in both data and model calculations), indicating the existence of a distinct relationship between large-scale atmospheric circulation patterns and groundwater dynamics. Hora et al. [25] in India, Bowes et al. [30] in the United States, and Taylor et al. [21] in Sub-Saharan aquifers highlight significant links between climate stress and groundwater development in these continents. Hence, the availability of groundwater resources for drinking water may be influenced by climate change and other future developments.

The relationship between hydrology and climatic conditions is referred to as hydro-climatology, which is a broad discipline concerned with understanding the workings of the hydrological cycle in the context of climate change. Significant progress in obtaining a satisfactory hydro-climatological understanding of groundwater variability remained to be made, particularly if hydro-climatological knowledge is to be fully integrated into water resource management and planning [27]. There are numerous potential links between climate change, extreme events, and disasters on one hand and groundwater fluctuations on the other [26]. Climate change's consequences are a growing source of concern for water managers. These resources are distributed globally due to the erratic influences of climate and physiographic structures [25]. Given the harsh realities of climate change, the hydro-climatology information is needed in greater quantities and for longer periods of time than ever before [29]. A combination of in-situ groundwater level and meteorological observations can be used to provide early warning of agricultural and hydrological droughts [30]. This is especially important because climate change is expected to cause higher temperatures, more intense rainfall, and longer dry seasons in Rwanda [31].

Human activities, on the other hand, have a wide range of effects on groundwater systems and availability. Numerous studies [32]–[35], have found significant impacts of human activities, primarily irrigation and groundwater exploitation. The study in [36], suggests that human-caused impacts on groundwater availability have long-term consequences. Groundwater extraction can have an effect on groundwater levels, potentially affecting other water use, such as agriculture and groundwater-dependent ecosystems [28]. With that said, quantitative estimates of climate change's hydrologic effects are critical for understanding and resolving potential water resource problems associated with domestic water use, industry, power generation, agriculture, transportation, future water resource system planning and management, and environmental protection [24].

Because of the relationships between surface and groundwater systems, activities that alter surface stages have been shown to influence aquifers and groundwater levels. A recent study [37] discovered that groundwater behavior was influenced by river stages that were transferred to the aquifer. However, Doll et al. [35], revealed that human influence occurs on a global scale, resulting in altered aquifer yields, particularly in arid and semi-arid regions with intensive irrigation practices.

## 2.4 Aquifer Management, Practices, Issues, and Constraints

Groundwater management entails balancing resource exploitation (in terms of quantity, quality, and relevant links to other natural resources) with rising demand for water for broad economic development and livelihoods [38][1]. The UNESCO [7], defines groundwater management as the control of groundwater abstraction and quality while also addressing the effects of groundwater abstraction on ecosystems, surface waters, land subsidence, and other factors. In support of these definitions, Foster et al. [39], state that the primary goal of groundwater management is to develop groundwater resources on the basis of a policy plan, as well as to monitor and control the effects of abstraction on the aquifers. To provide further clarification on groundwater management, [7] asserts that controlling the location and quantity of aquifer water withdrawals is one of the most important aspects of groundwater management. Thus, effective groundwater management must prevent or at least reduce aquifer overdraft.

The evaluation of groundwater issues and the implementation of management solutions necessitate the collection of hydrogeological data that are both 'baseline' and 'time-variant' in nature. As we saw in the subsection 2.3, groundwater levels (hydraulic heads) fluctuate over time as a result of human

influence and changing climatic conditions. Monitoring groundwater abstraction and aquifer water levels provides critical information for groundwater resource management [39]. One of the most important tools for obtaining information needed for adequate decision-making about sound and sustainable development and protection of groundwater resources is well-organized and implemented groundwater monitoring. This viewpoint is shared in [40], who believes that monitoring and assessment are critical tools for long-term management of groundwater resources. Monitoring is the collecting of time-varying data in its broadest sense, while groundwater monitoring is the collection, analysis, and storage of a variety of data on a regular basis in accordance with specific circumstances and aims.

Long-term water-level monitoring during periods of significant land-use change is specifically important for aquifer protection. Therefore, accurate ground-water level data must be collected over sufficient time periods to allow for proper development, management, and protection of the Nation's ground-water resources. Meteorological data, such as precipitation, aid in the interpretation of changes in water levels in observation wells. Meteorological and stream-flow data are frequently available from other sources; however, if not, some monitoring of variables such as stream-flow and precipitation may be required to supplement water-level data, particularly in remote areas or small watersheds. The frequency with which water levels are measured is one of the most important aspects of a water level monitoring program. Water-level fluctuations caused by changes in climatic conditions and water-level trends caused by changes in land-use or water-management practices are more likely to be "sampled" when data is collected in a systematic, long-term manner. The availability of long-term water-level records improves the ability to forecast future water levels significantly [41]. Water-level hydrographs, which are graphical plots that show changes in water levels over time, are an especially useful type of data reporting. Such hydrographs show the range of water-level fluctuations, seasonal water-level variations, and the cumulative effects of short-term and long-term hydrologic stresses [40]. This kind of information has the potential to assist water managers in meeting the recommendations outlined in the Brandtland commission's report [42], which focuses on environmental and effective natural resource utilization and preservation.

### 2.4.1 Monitoring Changes in Groundwater Table Depths

Groundwater resources are becoming increasingly important in supplying freshwater. This resource is an important source of freshwater throughout the tropics, providing access to safe water for

domestic, agricultural, and industrial purposes near the point of demand [43]. Water table monitoring is critical for guiding evidence-based decision making required for long-term water resource management and governance [44]. Levett et al. [45], argue that, environmental water monitoring should be regarded as a fundamental component of environmental science and policy. The continuous observations of groundwater tables under growing demands and climatic stressors has received the interest of scientists all over the world due to its importance to human, the environment, and the economy. According to Graham [46], groundwater level time series can also be thought of as the sum of the net groundwater recharge-discharge. This implies that time series data represents the net amount of available water in the aquifer.

Satellite has been one of the most used methods for groundwater observation over the last decades. However, according to a number of researchers, most satellite products are still too coarse for water management purposes, making precipitation downscaling a high-risk activity [47]. The recent advancement in ground-based water resource observation using wireless sensor networks has revolutionized the way this resource is managed. It reduces the cost of monitoring while improving the spatial and temporal coverage of monitoring networks, as well as the overall efficiency of the process. Although the Internet of Things has seen extensive applications in surface water hydrology, the literature only reports on a small number of IoT applications for groundwater management. To the best of the author's knowledge, there are two studies that have used wireless sensors for groundwater monitoring. One of these applications is discussed in [48], where a water table probe was created using low-cost sensor. Authors suggested that data logger has high accuracy and can be applied to monitoring of changes in water table depths in the well. However, a short I2C channel was used and no telemetries for data transfer are reported, only data saving on SD memory card is described. A study reported in [49], have also used low-cost sensors to data logger that records water flow in a flooded cave and drip rate of water in the cave. Their solution is compact and well designed, but it only saves data on memory card. None of these studies have used long-range communication to relay data to the remote receiving node or server. The paucity of deployment of IoT-enabled aquifer monitoring networks is primarily due to the high cost of underwater data loggers and their telemetry [3]. As a result, there has been a scarcity of long-term observations of changes in groundwater storage in low-income regions, widening the knowledge gap in groundwater hydrology.

As stated in the Subsection (2.2), groundwater, as opposed to surface waters, provides a distributed source of freshwater to meet the rapidly increasing demand in Sub-Saharan Africa (SSA) associated with expanding access to safe water and improving food security through irrigation [21], there is an urgent need to improve funding and technical skills in order to provide long-term monitoring of these resources [50], [51]. Automatic groundwater level monitoring using under-water pressure transducer sensors will provide critical data for drought monitoring and assessment.

### 2.4.2 The Importance of Long-term Groundwater Level Data and Information Sharing

Hydrology as a geoscience is built on observations and data from long-term experimental watersheds. They allow us to benchmark process understanding, observe trends and natural cycles, and are required for testing predictive models. Long-term experimental watersheds are also places where new measurement technologies are developed. These studies provide critical evidence for understanding and managing the provision of clean water supplies, predicting and mitigating the effects of floods, and protecting ecosystem services provided by rivers and wetlands. They also demonstrate how to manage land and water in an integrated, sustainable manner that reduces environmental and economic costs [52]**.** As the time series lengthens, this provides unparalleled insights into the sustainability of groundwater. Hence, the long-term groundwater data continue to have a fundamental value that grows over time.

The importance of persistent groundwater data collection and preservation has been emphasized for a long time. Nearly fifty-five years ago, two legendary figures in hydrology, Hewlett et al. [53] and Leopold [54], highlighted the significance of long-term data in managing current and future water resources. According to Hewlett et al. [53], the observational data collected from long-term experimental watersheds distributed around the world are the foundations of scientific hydrology and the evidence base for sustainable water management. Many basic data acquisition and research programs were established in response to the first International hydrological decade (1965–1974), which aimed to expand quantitative process-based understanding of the hydrological cycle [55].

Data streams at long-term sites enable us to identify and quantify the relationships between rainfall, soil moisture, groundwater, and runoff, facilitating understanding of flood and drought risk in various types of landscapes [52], [56]. However, as time series lengthen, the value and scientific leverage of these sites become equally invaluable in providing critical long-term data that provides

context for more focused research driven by hypothesis testing [57]. Long-term monitoring resulting from curiosity-driven research is critical for policymakers and society, serving as the foundation for rational decision making [45]. In most cases, the lack of a mechanisms to commit long-term funding makes long-term aquifers monitoring difficult for policymakers [52].

Several researchers have indicated that long-term data are more important than ever because they allow us to see trends and study the effects of various types of human and environmental induced groundwater changes, especially in light of the uncertain effects of projected climate change and increasing development activities [22], [52]. These long-term records will, however, play an important role in developing a better understanding of the hydrogeological and climatic conditions that control access and sustain well yields, informing where, when, and how groundwater withdrawals can contribute sustainably to building resilience and alleviating poverty, as represented by the United Nations SDGs [21]. They are also crucial for calibrating groundwater models, and they serve as the foundation for accurate simulations of future abstraction scenarios [39]. In general, these data are more important than ever before, as we work to preserve limited resources for current and future generations [52]. Despite the benefits and importance of long-term data for scientists and water managers, there is a lack of data and information sharing, particularly in low-income countries [7]. A large body of literature demonstrates that, even in places where groundwater data is continuously collected, it is rarely interpreted and shared with relevant stakeholders.

A thorough review on the role of IoT in continuous data collection from remote and difficult-to-access locations is presented in the subsection 2.5.

## 2.5 Wireless Sensor Networks (WSNs) and the Internet of Things (IoT)

Local networks, which belong to the network layer and typically include nodes equipped with sensors and/or actuators, are generally organized as Wireless Sensor Networks (WSNs), which are networks made up of spatially distributed sensing nodes that are frequently used to monitor and record environmental conditions. Data is typically collected via a (possibly multi-hop) route to a concentrator, also known as a "sink" or "gateway," which stores and/or forwards it outside the WSN. To be thorough, the sink is in charge of connecting the WSN to the outside world via a communication protocol that, in general, can differ from the protocol used to connect sensor nodes. In fact, the devices used in WSNs typically operate with limited energy for sensing, processing, and

communication tasks [58]–[60], igniting a cross-layer design approach that typically necessitates the joint consideration of distributed signal or data processing, medium access control, and communication protocols. These WSNs constitute the majority of IoT platforms. According to researchers [61]–[63], the IoT is a novel paradigm that is rapidly gaining traction in the context of modern wireless telecommunications. In broad terms, the IoT is a constantly growing collection of linked devices that capture and share data. The International Telecommunication Union (ITU) [64], defines IoT as *"a global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving interoperable information and communication technologies".* These definitions indicate that IoT is a system of interconnected, internet-connected objects capable of collecting and transmitting data over a wireless network with little or no human intervention. Wireless sensors make it possible for these objects to communicate with one another. Indeed, information exchange between sensor nodes and the rest of the IoT platform components is frequently accomplished by combining long-range (e.g., cellular LTE, SigFox, NB-IoT, and LoRaWAN) and low/medium-range (e.g., Z-Wave, ZigBee, Wi-Fi, NFC, RFID, Bluetooth Low Energy (BLE) or IEEE 802.11) communication technologies. Because the IoT is heavily reliant on WSN, it is necessary to have technical knowledge of WSN and their protocols, as well as a thorough understanding of the IoT ecosystem, when designing a reliable IoT solution [65]. While there are several models for IoT ecosystems [66]–[69], some of them overlook critical aspects of this technology in terms of environmental and hydrology monitoring. Figure 2.2 depicts an example of an ecosystem model that provides a high-level overview of IoT solutions [70].

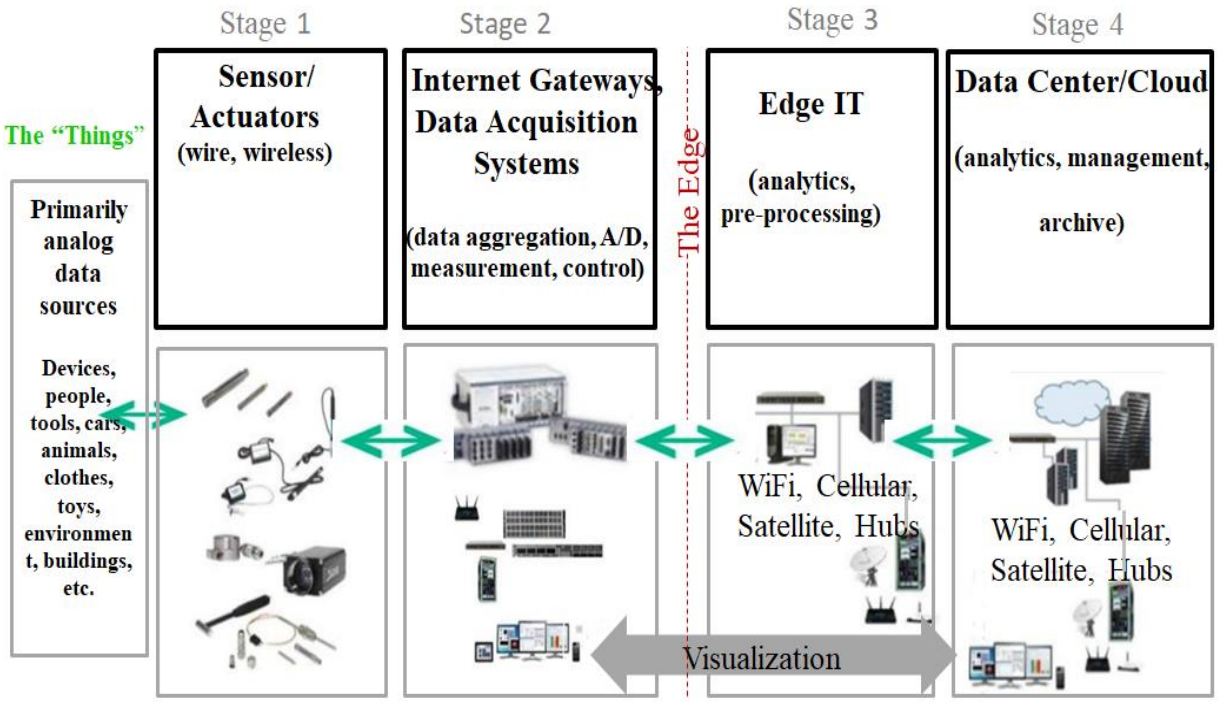



Figure 2.2 The layout of a four-stage IoT solutions architecture [70].

Among the IoT protocols listed above, are those specifically designed to support low-bandwidth, battery-powered devices. These are referred to as a set of low-power wireless wide area network (LPWAN) technologies. Many scientists are interested in an open source, long-range (LoRa) protocol within this family of protocols. LoRa is a wireless modulation scheme derived from Chirp Spread Spectrum (CSS) technology that operates at the physical network layer and encodes information on radio waves using chirp pulses. This scheme is best suited for applications that require small chunks of data to be transmitted at low bit rates [71]–[75]. Working at up to 2.4 GHz on license-free sub-gigahertz bands, it also supports industrial, scientific, and medical (ISM) frequencies (e.g., 868MHz, 433MHz, and 915MHz). The Low Power Wide Area Network (LoRaWAN) is built on top of LoRa modulation at a media access control (MAC) layer. The LoRa Alliance [76], released the specification, which depicts the LoRaWAN network protocol, on June 16, 2015. This standard enables seamless interoperability among smart things without the need for complex local installations, allowing users, developers, and businesses to implement the IoT with greater freedom. A 6LoWPAN is an improved version of the LPWAN designed by the Internet Engineering Task Force (IETF) to support IPv6 addressing. However, because of its complexity and high overhead, it was not applicable to constrained devices [77].

When choosing the right LPWAN technology for an IoT application, several factors must be considered, including quality of service, battery life, latency, scalability, payload length, coverage, range, deployment, and cost. The comparisons of short, medium, and long-range, wireless protocols with their respective bandwidth capacities are depicted in the Figure 2.3.

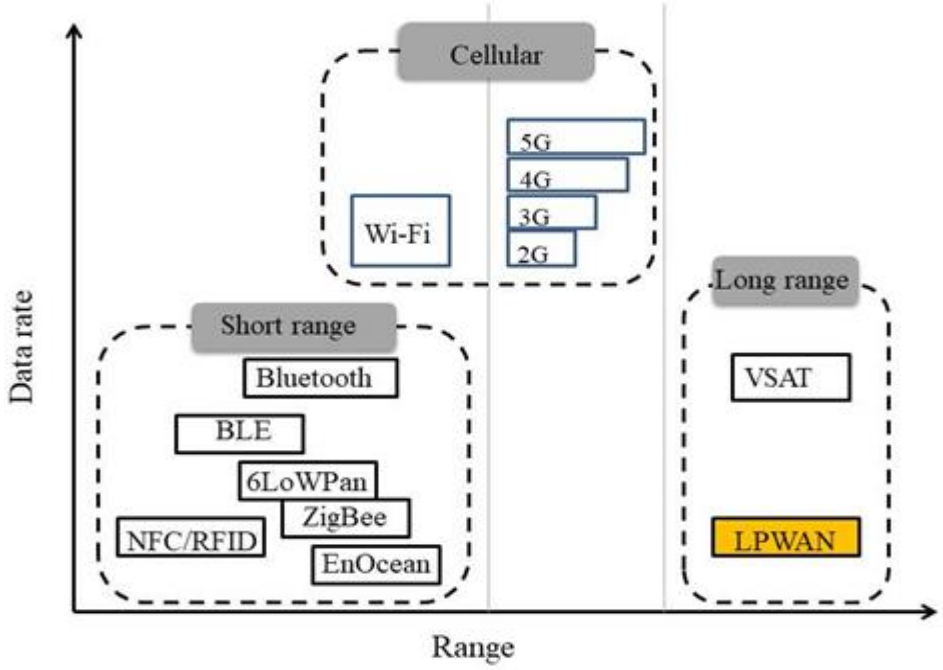



Figure 2.3 The comparison of data rate and range capacity of LPWAN radio technologies based on positioning [72].

Similarly, when considering the application of large-scale IoT deployment, it is also critical to understand the differences between the various schemes in an LPWAN family in order to make an informed choice of the appropriate technology. Table 2.1 summarizes the key technical features of various low-power wireless communication schemes.

Table 2.1 Comparisons of LPWAN technologies for large-scale IoT implementation

| | **NB-IoT** | **LoRaWAN** | **Sigfox** | **LTE-M** |
|---|---|---|---|---|
| **End-device cost** | $12 | $2.000-$5.210 | $2.080 | $20-30 |
| **Maximum payload** | 1600 bytes | 242 bytes | 12 bytes | 256 bytes |
| **Range** | 10 km (rural), 1 km (urban) | 5 km (rural), 20 km (urban) | 10 km (rural), 40 km (urban) | 1 km (urban), 10 km (rural) |

| | | | | |
|---|---|---|---|---|
| **Bandwidth** | 200 kHz | 125 kHz and 500 kHz | 100 Hz | 1.400 MHz-500 MHz |
| **Maximum data rate** | 200 kbps | 50 kbps | 100 bps | 1 Mbps both in uplink & downlink |
| **Latency** | 1.600 s - 10 s | <400 ms [78] | 30 s | 10 ms -15 ms |
| **Transmission power** | 46 dBm | 2/20 dBm | 14.500 dBm | 20/23 dBm |
| **Allows private network** | No | Yes | No | No |
| **Battery life** | Good | Long | Long [79] | Low |

According to the data in Figure 2.2 and Table 2.1, each of these schemes has advantages and disadvantages over the others. Sigfox and LoRaWAN are similar in many ways, but LoRaWAN has many advantages over other techniques. The ones that stand out are its ability to support millions of messages per gateway, multitenant operability, and security embedded end-to-end AES-128 for both network and application [72], [76]. While the NB-IoT has advantages in terms of latency and quality of service, it is difficult to implement firmware-update-over-the-air (FUOTA) or file transfers, has limited support for node mobility, its downlink is very limited, and its services are prohibitively expensive [66], [67]. Subsection 2.5.1 provides an overview of LoRa technology, including cost and energy implications.

### 2.5.1 Primer on LoRa Radio Technology

As described in Section 2.5, LoRa operates at the physical layer and consumes less energy. This scheme also uses end-to-end encryption for trusted network security. The basic LoRaWAN network architecture is star-of-stars, in which end-devices communicate with gateways via LoRaWAN. A network server is normally connected to a higher throughput network (WiFi, Ethernet, 3G or 4G) at the far end of the LoRaWAN and receives raw LoRaWAN frames from the end devices via the gateway. The FEC, bandwidth (BW), and spread factor (SP) are the three most important LoRa parameters to configure [80]. These parameters have a significant impact on network performance, with BW having the most sway. Normally, one or more radio transceivers connect the two sides to the field nodes and the gateway. RFM95W is a widely used, low-power LoRa enabling transceiver. It consumes little current and is resistant to interference [81]. Table 2.2 presents the typical RFM95W LoRa modem specifications. One or more radio transceivers are typically used to connect

the two sides of the field nodes and the gateway. RFM95W based LoRa transceivers are not only recommended for its low power, but also for its impervious to interference.

Table 2.2 RFM95W modem specification

| Parameter | Description |
|---|---|
| Modulation type | LoRa TM, GMSK, FSK, OOK, GFSK, and MSK |
| Link Budget | 168 dB (Maximum) |
| Low RX Current | 10.300 mA, 200 nA for register retention |
| RF output vs Input Power | +20 dB -100mW (Constant RF) |
| Packet appliance | 256 bytes (Maximum) with CRC |
| Synthesizer | 61Hz |
| Dynamic Range RSSI | 127dB |

According to Table 2.2, the RFM95 modem provides a consistent amount of power to the receiving node and can tolerate temporary attenuation.

Subsection 2.5.2 discussed open-source, low-power, and low-cost IoT technologies.

### 2.5.2 Low-power, Low-cost WSNs and the Open Designs

WSN dates back to the 1950s, when the United States (USA) first introduced this technology for military purposes. Soon after the 1980s, governments and universities began to use WSNs in a variety of applications, including forest fire detection, weather stations, and many others [58],[60]. Sensor cost has been an impediment to rapid adoption in many applications over the years [82]. Recent advances in electronics miniaturization and communication models have allowed for the creation of low-cost, low-power micro-scale sensors, microcontroller boards, and telemetries. The sharing of open designs and software has also contributed to the growth of low-cost sensor networks (LCSNs) [83], [84]. In this regard, there are various open-source IoT prototyping and development boards available that support various wireless connectivity, processing, and memory capabilities. Common boards include the ESP32, ESP8266, Arduino family (Uno, Nano, Mini, and Mega),

Raspberry Pi, Pycom Fipy, and others. This advancement has resulted in a high availability of spatial WSN sensing capacity at a fraction of the cost of conventional instruments. Most importantly, this technology also allows researchers and end-users to customize the LCSNs based on their needs [3]. Studies [3], [49], [83], [84] on the applicability of affordable sensors have shown that LCSNs have large potential for improving in situ environmental and hydrology observation. Despite the growth of LCSNs, there have been few studies using this technology on water resources [76], particularly in developing regions, and even fewer on groundwater hydrology globally [3], [84]. Therefore, there is a need to expand the use of LCSN to allow for a broader collection of groundwater data, particularly in Sub-Saharan Africa aquifers.

### 2.5.3 Powering and Sustaining the WSNs

To enable long-term operation, WSN nodes must be self-sufficient. These nodes are typically deployed in remote and difficult-to-access locations that require a reliable, long-lasting power source. Using replaceable batteries when they are depleted is not only expensive, but also difficult. Therefore, relying solely on battery power may not be able to sustain the network for an extended period of time [85]. This is where alternative energy generation methods are required. The available ambient energy can be harvested and stored in rechargeable batteries (e.g., lithium polymer) depending on the environment where the nodes have been deployed and the type of energy harvester(s). These renewable energy sources include thermal, light, chemical, kinetic, and electromagnetic energy [86]. These energies are converted to electricity using harvesting electronic circuitry. Among other sources, light in the form of solar energy harvested by photovoltaic cells is the most efficient [87]. The components for ambient energy harvesting must be chosen carefully in this quest [88]. In fact, the capacity of harvesters and reachable battery is chosen based on the dynamics of the load as influenced by WSN operations (see Figure 2.4).

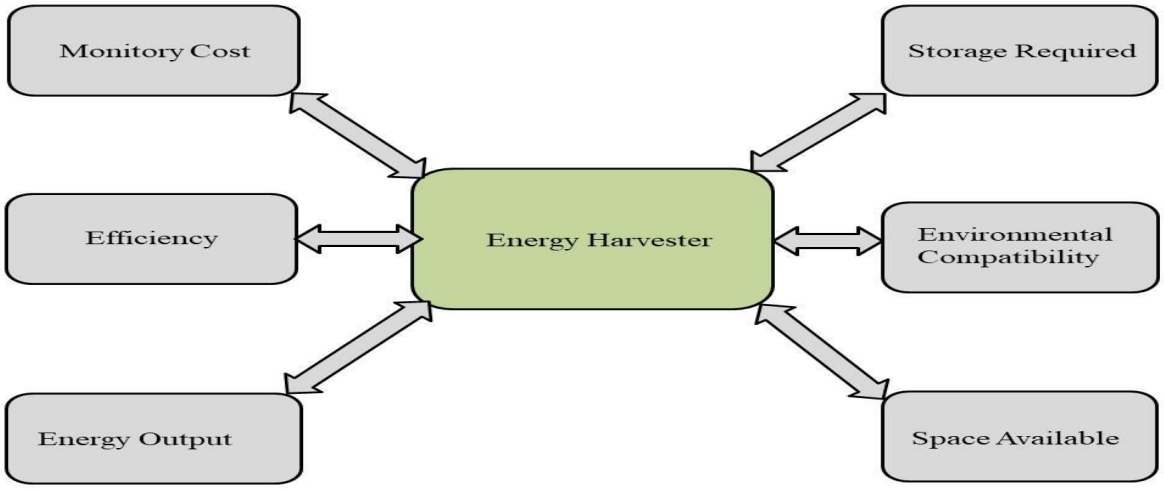



Figure 2.4 An illustration design consideration for an energy harvesting system [86].

Energy harvesters must be small enough to fit into the available space while also providing adequate ventilation. A major aspect for an energy-efficient system is the availability of a sufficient amount of energy source at the deployment region [87]. Additionally, and perhaps most importantly, the production should be energy and cost efficient.

### 2.5.4 Use of the Internet of Things for Groundwater Table Monitoring

The Internet of Things (IoT) has gained traction in environmental research in recent years, and ground-based aquifer monitoring is far more recommended than remote sensing**.** Hence, the adoption of low-cost instrumentation in wireless sensor technologies has greater potential in this context.

As previously stated, the application of WSN and IoT to surface water management have received a lot of attention [89]–[92] [93] but less attention has been paid to its application in groundwater resources [94] [55]. Previous research has revealed a scarcity of low-cost groundwater sensing deployments in both developed and developing countries [34], and we review these findings here. Anumalla et al. [95] presented one of the earliest studies in this field, in which pressure sensors and field programmable arrays were used to monitor the levels of ground-water tables in Western Nebraska aquifers in the United States. To relay data from a remote field to the data processing unit, the authors used a 2.4 GPRS/GSM communication scheme. The information gathered was disseminated through a web application and text messages. The authors made a significant breakthrough in water table monitoring, but their system relies on expensive, energy-intensive

devices and network protocols. Calderwood et al. [84], used cellular network telemetry to develop an open-source, low-cost WSN for real-time groundwater management. A proprietary data collection unit (Solinst kit) and an open-source data handling and visualization web application comprise the system. Their findings appear promising; however, the proprietary logger and telemetry used may limit the system's application in low-income areas due to its higher cost.

In addition, Chan et al. [48], proposed a groundwater observatory model that includes a low-cost probe built with low-cost pressure sensors (NXP MPX5010DP and MS5803-02ba) to measure water table depths and an SD memory card for data storage. The submersible sensor is housed in a water-resistant (aluminum tube) enclosure, while a common logger in a separate open-to-air enclosure samples atmospheric pressure. The authors established in that study that the redesigned sensor performed better in terms of accuracy, which was closely related to a commercial version. Their approach is sound, but the lack of telemetry and the use of primary batteries limit the logger's life span, making it unsuitable for water table monitoring in remote areas. Furthermore, Beddows et al. [49], propose a data logger for measuring cave water flow and drip rate. The logger is powered by three alkaline AA batteries and features a real-time clock, SD memory card low-cost sensors, and is built on the Arduino platform. This system, however, only stores data on SD memory cards and depends on primary batteries, necessitating field visits to collect data and replace batteries, making it unsuitable for aquifer management.

Rahman et al. [2], recently presented an architecture for water resource management that incorporates artificial intelligence and the internet of things. Their smart system monitors water usage using a Raspberry Pi-based platform and low-cost sensors. However, the system is only intended for water management at the end-user level, not aquifer management, and it employs the ZigBee protocol, which covers a relatively short distance (up to 400 meters) outdoors. Moreover, no information on energy expenditure, overall cost, or system deployment approach is provided.

The analysis of the works on the subject under discussion is summarized in Table 2.3.

Table 2.3 Analysis of the selected closely related studies for comparison

| | [84] | [49] | [95] | [96] |
|---|---|---|---|---|
| Sensor type | Solinst Leveloggers | MS5803-02BA and NXP | Unidata pressure sensor | Not specified |

| | | | | |
|---|---|---|---|---|
| | | MPX5010DP | | |
| Processing board | Solinst in-built board | Arduino Pro Mini or Nano | PIC12F675 microcontroller on Altera Nios FPGA board | Not specified |
| Network Technologies | GPRS/GSM* | Not specified | 802.11 based WLAN | GRPRS/GSM* |
| Power consumption analysis | Not specified | Details not Provided | Details not provided | Not reported |
| Energy harvesting | Not reported | Not reported | Not reported | Not included |
| Cost analysis | Detailed reported | Reported | Reported | Reported |

GSM= Global System for Mobile Communications, GPRS= General Packets Radio Services.

According to Table 2.3, most of the analyzed studies used commercial pressure transducers and did not provide information on the overall cost, energy, and power analysis of the solutions reported in these studies. It is also evident that low-cost, long-distance communication technologies (e.g., LoRa) were not used. Instead, high-power, high-cost networks were used.

## 2.6 Artificial Intelligence and Its Role in Hydrology

According to the field's founder, McCarthy [97], artificial intelligence (AI) is the art and science of creating intelligent machines, particularly clever computer programs. It is similar to using computers to study human intelligence, but AI does not have to be restricted to physiologically observable ways. AI is divided into several sub-fields, including Machine Learning (ML), Robotics (RBT), Computer Vision (CV), Natural Language Processing (NLP), and Artificial Neural Networks (ANN). These sub-fields enable AI to be used in a wide range of disciplines, including health, industry, entertainment, hydrology, and many others. The present research focuses primarily on the ML aspect of AI. Subsection 2.6.1 describes ML and related frameworks as part of AI technology.

### 2.6.1 Machine Learning and Appropriate Frameworks and Tools

Machine learning and Artificial Intelligence complement each other, and the next breakthrough will come not only from pushing each of them further, but also from combining them [98]. In recent years, AI, specifically machine learning (ML), has grown rapidly in the context of data analysis and

computing, allowing applications to function intelligently [99]. ML is widely regarded as one of the most popular latest technologies in the fourth industrial revolution because it allows systems to learn and improve based on experience without being explicitly programmed (4IR or Industry 4.0). Thus, ML algorithms are critical for intelligently analyzing hydrology data and developing the corresponding real-world applications. These algorithms are typically classified into four categories: supervised, unsupervised, semi-supervised, and reinforcement learning [99]. Supervised learning employs labeled training data, unsupervised learning employs unlabeled training data, semi-supervised learning employs both unlabeled and labeled data, and reinforcement learning employs an environmental-driven approach. Overall, the effectiveness and efficiency of a machine learning solution are determined by the nature and characteristics of the data, as well as the performance of the learning algorithms [100].

There are numerous ML tools and frameworks that support various programming languages such as C/C++, Python, R, Java, JavaScript, and others, with Python, JavaScript, and R being the most popular open-source languages. In this list, Python has become the most widely used for ML development and applications. TensorFlow, Scikit-Learn, Theano, Keras, and WEKA are examples of common frameworks. Scikit-Learn is a well-known Python-based framework that has an ensemble feature for combining predictions from multiple supervised models, works without special hardware support, and has a constantly updated and comprehensive set of algorithms and implementations. Furthermore, it is a part of many ecosystems; it is closely related to statistical and scientific Python packages [101]. A large number of libraries and visualization tools for dealing with hydrology issues, such as CFM, OpenHydrology, PyQGIS, and Hydrostat, are available for free and open sharing.

### 2.6.2 Machine Learning as Water Management Tool

Given the widespread use of ML in hydrological modeling over the last several decades, it is currently frequently employed to describe a variety of hydrological processes [102]. The hydrology community may use this technology to fully leverage the power of massive volumes of data in diverse hydrology subdomains using machine learning [47]. In the subject of hydrology, machine learning has been utilized to better grasp hydrological complexities [102], such as runoff modeling and water quality forecasts. This interpreted data is critical for hydrological management and will be

useful to water managers and policymakers. Several works from various application scenarios have demonstrated the use of ML to support water decisions [47], [103]–[105].

### 2.6.3 Machine Learning Based Predictive Modeling and its Role in Groundwater Analysis

The scientific community agrees that machine learning algorithms, when supplemented with data from in-field monitoring nodes, can improve models and forecasts of water availability, droughts, and other water-related events. Predictive modeling, according to Shmueli [106], is the practice of using a statistical model or data mining method to forecast new or future observations on data. As indicated by Cranmer et al. [107], predictive models try to provide a probabilistic model that fits testing data that was not utilized to estimate the model's parameters well. Predictive analytics is a science that can produce future insights with high accuracy. This analytics makes use of a wide range of methods and technologies, including big data, data mining, statistical modeling, machine learning, and various mathematical processes. Organizations use predictive analytics to sift through current and historical data to detect trends and forecast events and conditions that should occur at a specific time based on parameters supplied. Models can be developed, for example, to discover relationships between climatic parameters and changes in water table depths. Hydrologic process modeling is a critical tool for the efficient management of water resource systems [108]. Although studies have shown that extensive training of models with large amounts of prior data is the reason for higher predictive accuracy [109], it is strongly encouraged to broaden ML application in predicting the hydrologic cycle even in areas where data is scarce [47]. This has not been the case in Sub-Saharan Africa, where the use of ML in hydrology is still in its early stages [110][21]. The frequently used ML based in groundwater hydrology are random forest (RF), support vector regression (SVR), artificial neural network (ANN), and k-nearest neighbor (KNN). In addition to these methods, some hybrid approaches have been devised in pursuit of enhancing the efficiency of predictive accuracy. In groundwater hydrology, the most commonly used ML-based models are random forest (RF) [111]–[113], support vector machine (SVR) [114]–[117], artificial neural network (ANN) [103], [118]–[120], and k-nearest neighbor (KNN) [121]. In addition to these methods, some hybrid approaches have been developed in the pursuit of improving predictive accuracy efficiency [122]–[125]. These ML methods are calibrated using appropriately prepared historical samples before being applied to prediction, and then tested using unseen sub-samples from the same distribution as the training sample. Data preparation entails processes such as removing

unrealistic data points, scaling data to the appropriate range, and dividing the sample into training, validation, and testing sets. Most of the time, the calibration set is greater than twice the sum of the other subsets. These common ML techniques are represented mathematically and described in the following subheadings.

### 2.6.3.1 K-Nearest Neighbor

K-Nearest Neighbor is a simple and robust method for regression and classification. The KNN method can estimate unknown data points based on their distance from known data points (training sample) [126][127]. While attained observations are incomplete and noisy, the KNN technique is one of the best methods for ML-based forecasting [128], [129]. Moreover, using this method, the most significant data points can be identified from noisy samples. For continuous data, voting is done based on the distance measured between data points using either Minkowski, Euclidean, Chebyshev, or Manhattan metrics. Suppose, we have two sets $X$, and $Y$, each of them has $t$ number of items, such that $X = (x_1, x_2, \ldots, x_t)$, and $Y = (y_1, y_2, \ldots, y_t)$ so long as $(j = 1,2,3, \ldots, t)$. Then, distance between the desired data point and nearby points can be calculated using the aforementioned metrics. Distance between the desired point to the closest points is then defined as:

$$\widehat{D}(X, Y) = \sum_{j=1}^{t} [(|x_j - y_j|^r)]^{1/r}, \quad (2.1)$$

where, $r$ is a positive real number, and $\widehat{D}$ is the calculated distance.

To anticipate the variations in water table depths, we perform the following steps:

1. Use Equation 2.1 to calculate the distance between a new sample and each of the adjacent points.
2. Sort all values calculated in step 1, by increasing order.
3. Utilize the greedy search technique to determine the optimal value of K, based on RMSE.
4. Enumerate an inverse distance weighted mean using K neighboring examples.
5. Return average as the approximated value.

In the above scheme, K is a user-configurable parameter that represents the number of contiguous features to be included in the calculation of average votes. Then, prediction of variations in groundwater levels is obtained as the average weighted distance between samples.

### 2.6.3.2 Artificial Neural Network

The artificial neural network is a combination of multiple interconnected neurons that learn cardinal relationships in a set of data in the same way as the human brain operates [120], [130], [131] . Interconnected neurons make an input layer, one or more hidden layers, and an output layer. As input data are fed through the input layer, neurons in the hidden layer(s) compute the output using connection weights and bias. One of the two stages in which ANNs are used is in the training phase. In this phase, a training algorithm such as conjugate gradient momentum, Levenberg–Marqardt, backpropagation, Adam, gradient-descent, or Bayesian regularization is selected and the suitable connection weights are determined. The feed-forward network was trained using the back-propagation method to avoid an over-fitting issue. Another stage is the real application of the trained neural network. The estimated value(s) is then obtained as:

$$Y_e = f\sum_{j=1}^{k}(x_i w_{ij} + b), \tag{2.2}$$

where $x_i$ is the input examples, $Y_e$ is the approximated output, $b$ is bias term, $f$ is an activation function, and $w_i$ is the weight of the vertices. In a three-layered multilayer perceptron network, the transformation of the weighted inputs is accomplished using a rectifier linear unit (ReLU), which is defined as:

$$x_r = \max(0, x_i), \tag{2.3}$$

$x_r$ symbolizes the transformed input passing in the hidden layer, $x_i$ is the raw input. All $x_i$ values greater than zero are mapped to their respective y-values, while all $x_i$ smaller than zero are assigned to zero. This makes ReLU computational modest and be able to efficiently handle negative inputs, and also offers smoother optimization [132].

### 2.6.3.3 Support Vector Machine

The support vector machine for regression problems is termed as a support vector regression (SVR). SVRs are supervised learning techniques introduced by Cortes and Vapnik [133]. These are powerful methods that utilize structural risk minimization to obtain optimal solutions [133], [134]. SVR accomplishes risk minimization measures using a set of several input vectors while conducting an estimation of non-linear targets through regression processes [134][135]. Based on the assumption that, there is a relationship between the dependent variable *y* and independent variables

($X_1$, $X_2$, $X_3$,…,$X_t$), the SVR model estimates a function $f(x)$, which determines the target values $y_i$ plus the admissible error $\epsilon$. In an SVR model, data processing is conducted in a hyperplane and it starts as a linear transformation of the time-series. The linear representation of an SVR algebraic function is given by:

$$F = \langle w, x \rangle + b, \tag{2.4}$$

$$Y = f(x) + \epsilon, \tag{2.5}$$

where $b \epsilon \mathbf{R}$, $x$, $w \epsilon x$, and $\langle w.x \rangle$ are the bias, inputs signals, weight vector, and the dot product between $x$ and $w$ respectively. Then to minimize the norm $\| w \|^2 = \langle w.x \rangle$ we need to find smallest possible value of $w$ as:

$$minimize \frac{1}{2} \| w \|^2, \tag{2.6}$$

subject to:

$$\langle w, x \rangle + b - Y_i \leq \epsilon; Y_i - \langle w, x \rangle - b \leq \epsilon; \epsilon \geq 0; i = 1,2,\ldots,t \tag{2.7}$$

As the primary goal is to get a function *f(x)* that can be used to calculate a set of observed $x_i$ and estimated $y_i$ values with the level of accuracy bounded by $\epsilon$, this is realized by minimizing a regularized risk function in (Equation (2.6)) with constraints stipulated in inequalities given below (Equation (2.7)). Two relaxed variables ($\xi$ , $\xi^*$) are incorporated in (Equation (2.8)) to allow for some error tolerance.

$$\frac{1}{2} \| w \|^2 + C(\sum_{i=1}^{t} (\xi + \xi^*)), \tag{2.8}$$

subject to:

$$\langle w, x \rangle + b - Y_i \leq \epsilon + \xi^*; Y_i - \langle w, x \rangle - b \leq \epsilon + \xi; \xi^* + \xi \geq 0; i = 1,2,\ldots,t. \tag{2.9}$$

where $t$ is the total number of model input features, and $C$ is a user-configurable parameter that manages the influence of each supporting vector in generalization and stability of the SVR model. Ultimately, Eq.2.9 can be reformulated to:

$$F(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{t} (\alpha_i - \alpha_i^*) k\langle x, x_i \rangle + b \tag{2.10}$$

where $k\langle x, x_i \rangle$ is the kernel, and $\alpha_i$,$\alpha_i^*$ are Lagrangian multipliers. Before applying SVR in data processing, the appropriate kernel and support vectors should be determined. The SVR kernel

provides mapping of the non-linear features in a high dimensional space while converting it into a normal linear format. Thus, SVR suits well for complex interrelationship among features in environmental modeling [136]. There are several types of kernels, such as polynomial, multi-layer perceptron, exponential, and Gaussian radial basis function. Radial basis essence has shown a commended performance in hydrologic studies [135][137]. The exponential radial bias function is given by:

$$k(x_i, x) = \exp(-\| (x_i - x_t) \|^2) \div 2\sigma^2, \quad (2.11)$$

### 2.6.3.4 Random Forest

Based on decision trees with the application of bootstrap aggregation, random forests (RF) for regression problems and classification were introduced by Breiman [138]. A forest of diverse trees is developed using randomly chosen features selected from random subsets of the original training data. As a large number of trees is produced, classification results are obtained from the popular class, while in regression problems, the result is computed as the average value obtained from all the individual regression trees [139][140]. In the current study, we focus on the regression type of RF. A forest may contain several trees as specified by the user. Suppose the number of trees in the forest is denoted by *M*. The random forest method works in the following manner:

1. Randomly fetch different subsets *xi* from a given dataset *X*
2. Use sampled data to create *M* decision trees.
3. Enumerate average of the votes from the decision trees.
4. Return the average as the final approximated value.

Randomness is applied at two levels of the random forests: during data selection and in attribute selection. Since the regression trees are created from random vectors selected from training dataset *X*, each leaf-node contains a constant estimate of *Y*. As an example, the data points $\{(x_1, y_1), (x_2, y_2), \dots, (x_M, y_M)\}$are selected as samples for the leaf-nodes, and the anticipated data can be modeled as the averaged predictions from all the individual regression trees as:

$$\hat{y}(x_j) = \frac{1}{M} \sum_{m=1}^{M} y_m(x_j), \quad (2.12)$$

Such that, $j = 1,2,\ldots,M$, and where $\hat{y}(x_j)$ is the estimated result. In RF, tuning-parameters have a great effect on the ability of the model [141]. The most important tuning parameters for RF in Scikit-Learn are n_estimators, random_state, n_jobs, min  sample_leaf, max_features [141].

The hydrology forecast can be made with a short (few days or weeks) or long (seasons or years) lead time. Short-term and long-term forecast outcomes may be relevant depending on the type of information required to support groundwater management decisions. The further into the future you go in such modeling; the more difficult it becomes to predict how groundwater changes will occur. Seasonal forecasts, in most cases, have a greater potential to inform groundwater decisions [142]. The majority of the research in the literature has focused on short-term water table forecasting, has used a large amount of input, and has not been conducted under temperate climate conditions [143]. For example, Zhou et al. [144] present a comparison of ANN and SVM for water table depth modeling. The authors used discrete wavelets in data preparation, and their results show that SVM outperforms the ANN model in terms of accuracy. Similarly, the SVM method is said to outperform the adaptive neuro-fuzzy inference system and the ANN [145], [146]. Furthermore, Natarajan et al. [147], evaluate the accuracy of SVM, extreme learning (ELM), genetic programming (GP), and ANN in groundwater level simulation. Their findings suggest that ELM outperforms GP, SVM, and ANN in terms of precision. Additionally, the studies  [148], [136] provide a thorough explanation of the application of SVM in hydrology. Guzman et al. [124] present one of the few studies focused on seasonal forecast of water tables, in which a Recurrent Neural Network is trained using a sufficient amount of historical precipitation and groundwater level time series data and applied to predict water level up to three months ahead with promising accuracy. These findings are promising for irrigation water management, but they may not be applicable to Sub-Saharan Africa, which has a variety of aquifers and a temperate climate.In addition, recent research has investigated the effectiveness of a non-linear auto-regressive network with exogenous input (NARX) in modeling variations in water table heights using precipitation and temperature data [136], [149], [150]. NARX performed well in seasonal predictions of groundwater depths for various aquifer types, according to Wunsch et al. [150] and Guzman et al. [136], [149]. The authors' results were obtained with a large training sample, which is not common in Sub-Saharan Africa. Furthermore, the climate is different, with precipitation not being the only significant recharge of aquifers in humid areas. For forecasting daily

levels, [151] prefers SVM to NARX. Some studies attempted to combine multiple techniques in order to improve the accuracy of machine learning models [152], [124]. For instance, the authors of [152] describe a hybrid method for simulating groundwater levels that combines Wavelet, ANN, SVM, and empirical mode decomposition.

### 2.6.3.5 Ensemble Mode Decomposition

The EMD method, first proposed by [153], is an adaptive and empirical approach for data analysis. This data driven method is designed for the data having nonlinear and nonstationary characteristics. It decomposes original time series into several simpler components "mono component functions" known as intrinsic mode functions (IMFs). IMFs should meet two principle conditions: the number of extreme values and that of zero crossings must be equal or at most differ by one in the whole signal segment, and the second condition is that at any point, the mean of envelope defined by the local minima and that of the local maxima must be zero [153]. For a given series $x(t)$ the EMD can decompose this input signal into narrow band IMFs using sifting process as flows:

(i) Identify and extract all the minima and maxima

(ii) Connect all the local minima and maxima by a cubic spline to form the lower and upper envelope.

(iii) Calculate the mean value $m_1$ of these envelops and subtract it from the signal $h_1 = x(t) - m_1$.

(iv) Check if $h_1$ fulfills two criteria for IMF. Otherwise, repeat steps (i)-(iv) until satisfies the IMF criteria.

Supposing, the conditions are satisfied after $j$ times of iteration, then:

$$h_{1j} = h_{1(j-1)} - m_{1j} \tag{2.13}$$

then

$d_1 = h_{1j}$ is the first IMF.

The residual value is obtained as presented in Equation (2.14).

$$r_1 = x(t) - d_1 \tag{2.14}$$

The residue is considered as the input signal to be used in another sifting process to get next IMF $d_2$, since it contains all the remaining frequency information of the original data.

$$r_3 = r_1 - d_2 \tag{2.15}$$

$$r_n = r_{n-1} - d_n$$

The algorithm stops when $r_n$ is a trend component such that no more IMF can be extracted.
With $j^{th}$ degree of freedom, the function can be presented in Equation 2.16.

$$x(t) = \sum_{j=1}^{I} (d_j(t) + r_1(t)) \tag{2.16}$$

where I is the total number of IMFs, $r_I(t)$ is the residual of the decomposed signal, and $d_j(t)$ is the $j^{th}$ IMF.

Although the EMD has high locality and adaptability, but suffers from instability, end interpolation and mode mixing challenges [154]. The ensemble empirical mode decomposition (EEMD) is an improved version of the EMD in attempt to overcome the end effects and mode mixing issues experienced in EMD computation. This improvement avoids the pattern aliasing issue in EMD since it contains sifting mechanism that injects white–noise in the signals and treats the mean values as the final result. EEMD captures well the dynamics of hydrology[155][156]. The EEMD works in the following manner:

(i) Inject white noise $wn(t)$ to the original series $x(t)$

$$X(t) = x(t) + wn(t) \tag{2.17}$$

(ii) Extract IMFs from the data with added white noise

(iii) Iterate step 1 and step 2 adding different white noise to the series in each iteration.

(iv) Get the ensemble means of the extracted IMFs as the final result

## 2.7 The Use of Machine Learning and IoT for Groundwater Management

In summary, the majority of the literature on the use of IoT in hydrology has focused on surface water, with few discussions on low-power, long-range, and energy harvesting solutions for groundwater management. It is also demonstrated that there is a paucity of groundwater data and little research on the use of appropriate technologies such as IoT in Sub-Saharan Africa. Even the scant information available on aquifers is almost never shared with multiple stakeholders. Furthermore, there are a few studies in the literature that report on machine learning-based seasonal water table forecasts, particularly in temperate regions. This information is almost never shared with

multiple stakeholders. Literature encourages more aquifer exploration through the use of sustainable and cost-effective solutions to assist groundwater managers and policymakers. On the other hand, the body of literature discusses a variety of approaches to developing or improving technological solutions in novel ways while generating new knowledge about the process and application. However, the design science methodological approach is best suited to the current research due to its steps that are appropriate for the desired outcomes.

## 2.8 Research Methodological Approach

Literature presents several approaches to scientific studies that focused on devising solutions to identified problems and develop new knowledge iteratively. The present study adopts a design science research (DSR) methodological approach which was first ideated in 1969 by Simon [157]. Over the years, DSR has been assessed, improved, and recognized as an acceptable approach to research in information systems and other fields [158]. While natural and social sciences seek to explore, explain, describe, and predict, design science seeks to create something that does not yet exist or to improve on existing solutions.

According to Johannesson et al. [159], the goal of design science in information systems and technology is to create novel artifacts such as models, methods, and systems that help people develop, use, and maintain IT solutions. The authors went on to say that DSR is typically used to create artifacts that address practical issues that people face in a variety of contexts. Humans create artifacts to solve or address a specific problem. These can include physical objects, methods, guidelines, products, services, and processes, among other things. In support of the above-mentioned researchers' viewpoint, Schallmo et al. [160], elaborate that, in addition to creating an artifact, DSR also generates new knowledge about those artifacts, as well as their use and environment. Furthermore, Holmstrom et al. [161], stated that this methodological approach is human-centered and can be used to address ill-defined or unknown problems. With regard to the preceding viewpoint, and that of Ben Mahmoud-Jouini et al. [162], [163], DSR is best suited to decision contexts with high uncertainty and ambiguity because it involves abduction and experimentation with multiple alternative solutions that actively mediate a variety of tensions between possibilities and constraints. The DSR approach is implemented in five major steps (shown in Figure 2.5), namely, problem explanation, requirement definition, design and development, demonstration, and artifact evaluation.

The author of the present research believes that DSR is the best fit for the current research, which aimed at designing and developing an IoT-based solution as well as a machine learning method to provide a practical solution to address groundwater management problems in eastern Rwanda. The proposed solution consists of an autonomous monitoring network and decision support tools that were developed iteratively and through experimentation. The prototype was also validated in the field and its reliability and efficacy were assessed. The outcomes are made available to a wide range of stakeholders via web and mobile applications.

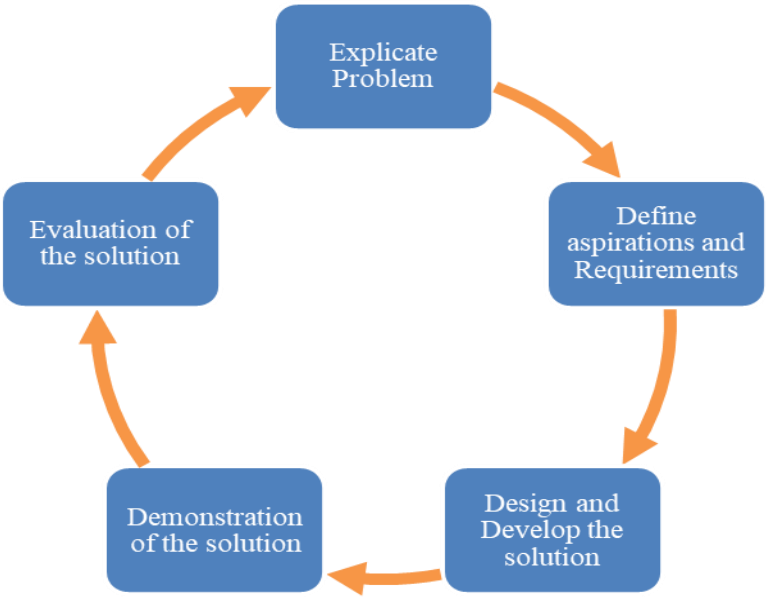



Figure 2.5 A general schematic representation of the design science research approach [164].

# CHAPTER 3

# DEVELOPMENT AND VALIDATION OF A LOW-COST, LOW-POWER ENERGY-HARVESTING, LoRa-GSM, IoT SYSTEM FOR MONITORING GROUNDWATER LEVELS

## 3.1 Introduction

This chapter presents the design, development, deployment, and validation of a low-cot, low-power IoT-based system (LWNGM) that makes use of LoRa and GSM communication technologies as well as energy-harvesting technique, to achieve the objectives one, two, and three of this research. The objectives are namely: *To obtain requirements for the Internet of Things-enabled groundwater table depths monitoring system, To design and develop a low-cost, low-power, energy-harvesting Wireless Sensor Network (WSN) for remote and near real-time groundwater level monitoring- and To validate the developed WSN's affordability, energy efficiency, and network efficacy through field deployment.*

The proposed low-cost, low-power, wireless sensor network for groundwater monitoring (LWNGM) was developed to provide near real-time groundwater level data to support prudent decision making in groundwater resource management in Eastern Province in Rwanda but the LWNGM was deployed in Zanzibar, Tanzania because of the reasons explained in the Section 1.7 in Chapter 1. The system is based on the ATmega328P microcontroller platform and incorporates MS5803-14BA and MB280 sensors. The I2C communication channels between the sensors and the microcontroller were extended using 25-meter PVC cables. The electronics were potted and protected in a waterproof aluminum cylinder. The Arduino UNO wakes up in six-hour intervals for measurements and data-logging to the SD card, and at twelve-hour intervals for relaying data (in batches) to the LoRa gateway, before it goes back into a deep-sleep mode for the rest of the time (duty cycle < 1%). The average power consumption for the end node was 104.081 mW. The power autonomy of all nodes is provided by a 3.7 V, 5000 mAh rechargeable LiPo battery, and a 9 V, 600 mAh rechargeable Li-ion battery, respectively, which are supported by 6 V and, 3 W solar chargers. The data processing and storage components, as well as the data visualization dashboard, were created

using free and open-source software. The LWNGM was reasonably priced, ranging between $350 and $400. Practical evaluation determined that the system is reliable and transferable, particularly in areas with a limited budget for hydrologic management.

Section 3.2 outlines the research's major contributions. The materials and methods used in this work are explained in Section 3.3. Section 3.4 describes the experimental design and evaluation of the LWNGM system. Section 3.5 delves into the results of the experiment described in Section 3.4. Section 3.6 concludes the chapter and provides an outlook.

In this study, we created and tested a groundwater monitoring platform based on low-power, low-cost sensors, open-source tools, and low-power solar energy harvesters. The system periodically logs the water table depth data to an SD card on site, and then relays the recorded data to the LG01 LoRa gateway in batch, twice a day. This study will broaden the regional and global understanding of simple and low-cost WSN technology for hydrology management. It will also improve evidence-based consultation to assist decision-makers in making better decisions for sensible water resource management.

## 3.2. Contributions

The research work presented in this chapter has made the following contributions: First, developed a conceptual design of low-cost, low-power, autonomous WSN for groundwater monitoring based on non- proprietary software. Second, it has developing an energy-harvesting wireless sensor network for continuous and near real-time monitoring of groundwater. Third, has developing an integrated system that combines sensor-based remote monitoring with downstream units for open data sharing with policymakers, scientists, and the general public. Fourth, has conducted a practical analysis and evaluation of the cost and energy expenditure for the water table monitoring model in order to better understand the cost and energy implications of affordable WSN technology.

## 3.3. Materials and Methods

In this section, we provide a detailed elucidation of the field sites, materials, and methods utilized for this investigation.

### 3.3.1. Study Site Description

Bandamaji station is found in Donge Mnyimbi at −5.968399, 39.250488, with an elevation of approximately 37 m in Zanzibar, Tanzania (Figure 3.1). This area is approximately 25 km from the Zanzibar stone town. The groundwater observatory station is under the Zanzibar Water Authority (ZAWA) [165], [166], a government organization responsible for the management and distribution of water supply in the Isles. The station is situated approximately 87 m from one of the huge ponds in Zanzibar. The majority of the people in Mnyimbi and nearby villages get their water from drilled wells.

Groundwater is the primary source of freshwater for more than a quarter (25%) of Tanzanians [167], and is the primary source of water (more than 70%) in the Zanzibar Islands [168][169]. However, according to some studies, groundwater extraction in Zanzibar is unsustainable, and some boreholes are no longer operational [165], [170], [171]. At the end of each month, ZAWA monitors patrols across the country to collect water quality and quantity measurements from bored observational wells. Normally, the only available water table depth-monitoring tool is the beeper tape. These patrols are time-consumin.,mg, human-resource- intensive, and material-resource-intensive. Current monitoring practices and frequencies have a significant impact on the quality and availability of continuous water table data. As a result, there is a need for affordable and continuous groundwater monitoring in Zanzibar [165], [170].

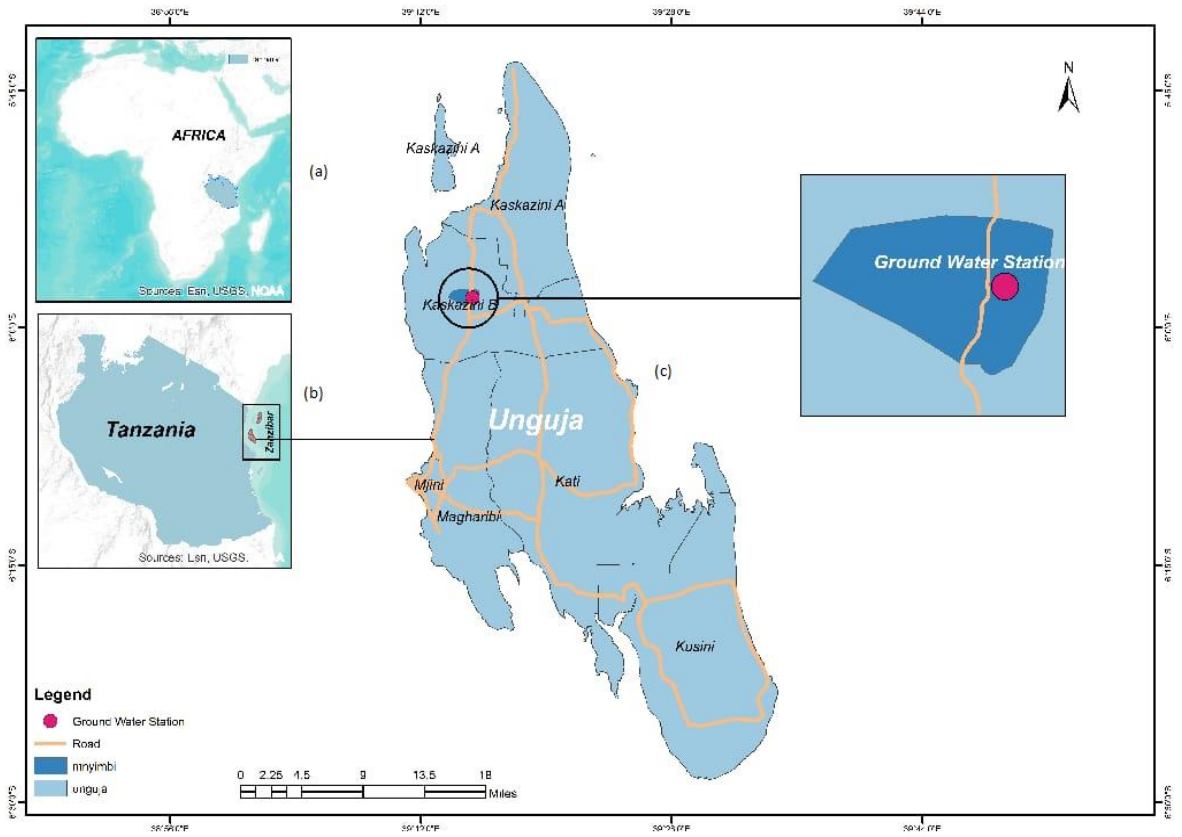



Figure 3.1 Location of bandamaji monitoring well in (a) Africa (b) Tanzania and (c) Zanzibar (Unguja) - Donge, Mnyimbi

### 3.3.2 Description of the System Design and Realization

One of the driving forces behind this study's attempt to create a WSN-based platform for monitoring variations in water table depths is technological advancement in low-power sensors and telemetry. The system is intended to have low power and low cost in nature. To achieve this goal, the system comprises four low-cost, low-power main components: the terminal unit (end node), gateway, network server, and cloud application server. The low-cost and low-power components (shown in Table 3.1) are redesigned and used to realize the intended system.

Table 3.1 Selected components for the development of the LGMWN.

| Component | Function | Version |
|---|---|---|
| Arduino board | Microcontroller to host sensors, memory and clock. | UNO R3 |
| MicroSD card shield | To host with the MicroSD card | Generic/Robotlyn SD card breakout |

| | | |
|---|---|---|
| | | shield |
| LoRa gateway | Linking between LoRa Nodes and the network Server | LG01-P,868 MHz |
| MicroSD card | Field data storage | Class 4, 16 GB |
| Solderboard | Solderboard for prototyping | |
| Solar harvester | Solar energy collector | 9 V/12 V, 3 W/6 W |
| Real-time clock | Accurate time keeping | DS3231 |
| Sensor connection cable | To link sensors to microcontroller | UTP Cat 6 |
| I2C differential interface | To extend the length of the I2C connections | PCA9615 |
| LoRa breakout shield | LoRa shield to connect to LoRa gateway | V1.4 |
| 4G LTE USB modem | 4G LTE modem to provide backhaul for LoRa | EC25-EU mini PCIe |
| Adjustable DC to DC converter | DC to DC converter to provide 9V and 12V for nodes and gateway | 5 V/8 V/9 V/12 V |

### 3.2.2.1 GATEWAY HARDWARE AND PROGRAM

The gateway (GW) is the connection point through which end nodes send data to the server. The Dragino LG01-P gateway (Dragino, Shenzhen, Guangdong, China) embeds a Semtech SX1276 LoRa module with an RMF95 chipset (Semtech Corporation, USA), and connects the LoRa wireless network to the Internet protocol (IP). This gateway supports LoRaWAN protocol on a single channel and controlled by the customizable OpenWrt Linux-based platform. It has a 100 mA current rating and a 12 V voltage rating. The gateway operates at 868 MHz with a 3 dBi gain antenna, receives sensor data via LoRa radio with an average sensitivity of -148 dBm, and relays it to the local server via a GPRR/GSM backhaul. The Quectel 4G LTE EC25-EU USB dongle (Quectel, Shanghai, China) was connected to the cellular network at 50 Mbps and delivered downlink data at 150 Mbps. A 9 V to 12 V DC voltage booster module (Shenzhen iSmart, Guangdong, China) was connected to a 9 V, 600 mAh Li-ion rechargeable battery and two 9 V, 3 W solar panels connected in series to power the LG01 gateway at 12 V. Li-ion batteries have several advantages over other types of

batteries, but the most important reason they were chosen to power the gateway is their high voltage capacity and longer life. The software that controls the gateway is written in the C programming language in an open source Arduino Integrated Environment (IDE), downloaded from the Arduino website (Arduino Somerville, MA, USA). The gateway relays data to the local server using the message queuing telemetry transport (MQTT) protocol. The protocol is well suited to the size and format of the message sent to the server, as well as the processing devices [172], [173]. The messages are formatted in JSON format for ease of reading and to reduce server load. Figure 3.2 depicts the scheme used to collect and relay data from the field to the network server. The configurations for the gateway and the 4G LTE USB dongle were completed on a Linux-based console using a secure shell connection (SSH), web user interface, and AT command (via the default IP address 10.130.1.1).

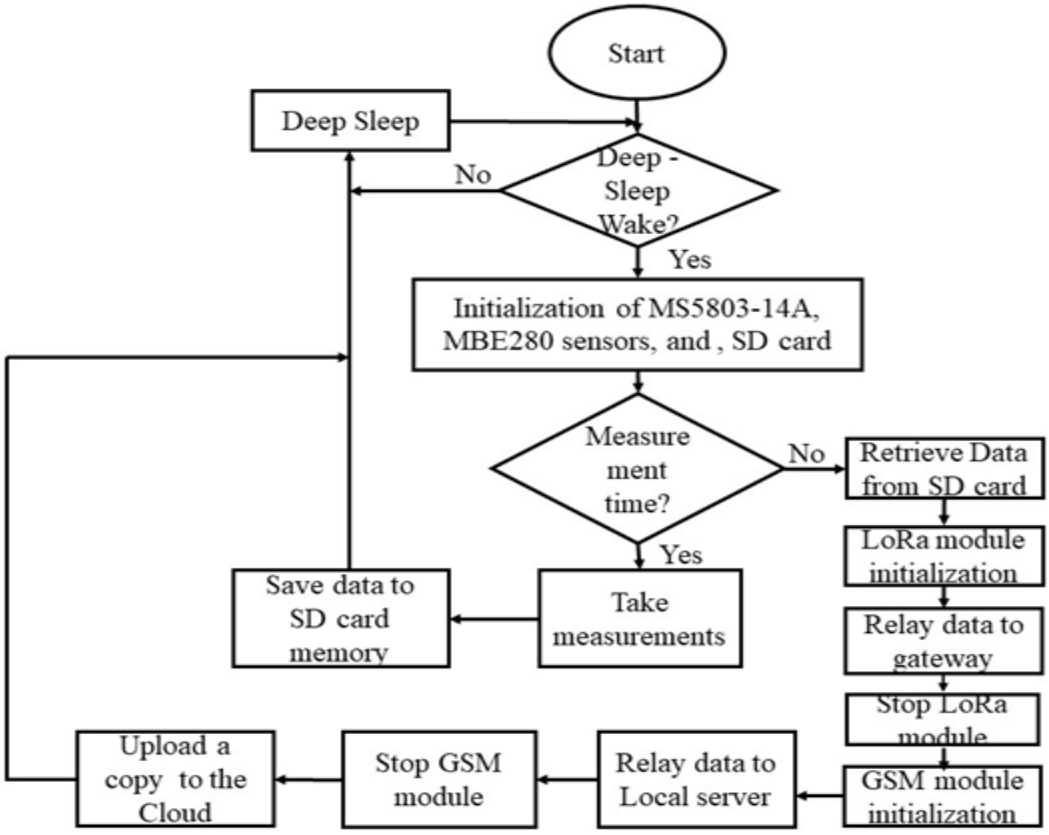



Figure 3.2 The functional scheme of data gathering and storage.

### 3.2.2.2 FIELD SENSOR MODULE HARDWARE AND PROGRAM

An Arduino Uno R3 microcontroller (MCU) served as the foundation for the field-sensing node (Figure 3.3). The MCU board communicates with the LoRa transceiver version 1.4 (Dragino, Shenzhen, Guangdong, China) and the micro SD card unit (Robotdyn, Zhuhai, GD, China) via the serial peripheral interface (SPI) and input-output (I/O) lines.

The Arduino Uno also communicates with the two pressure sensors and the external DS3231 real-time clock (RTC) module (Maxim Integrated, San Jose, CA, USA) via the I2C bus lines. The solar energy harvester module (Heltec, Chengdu, China) is linked to the 3.7 V, 5000 mAh rechargeable LiPo battery, which supplies voltage to the MCU. The first sensor of the field node is a low power, high-resolution pressure sensor breakout MS5803-14BA (SparkFun, Colorado, USA) that measures the pressure exerted by the water above it. The second sensor is a low power, humidity, barometric pressure, and temperature MBE280 sensor breakout (Adafruit, New York, USA), which linked to the 3.7 V, 5000 mAh rechargeable LiPo battery, which supplies voltage to the MCU which is used to compensate for atmospheric pressure. The MS5803-14BA (MS) has a maximum voltage of 3.3 V, whereas MBE280 (MB) has an operating voltage range of 3.3 V to 5 V. The RTC clock generates a high-precision and reliable date and time using a separate power source from a long-lasting 3 V lithium coin cell battery [174], [175], whereas the micro SD card (SD) module is powered through a 3.3 V pin of the MCU. The ATmega328P microcontroller was embedded in an Arduino UNO (R3). R3 has a voltage rating of 3.3 V to 5 V and a maximum current of 50 mA, according to the datasheet.

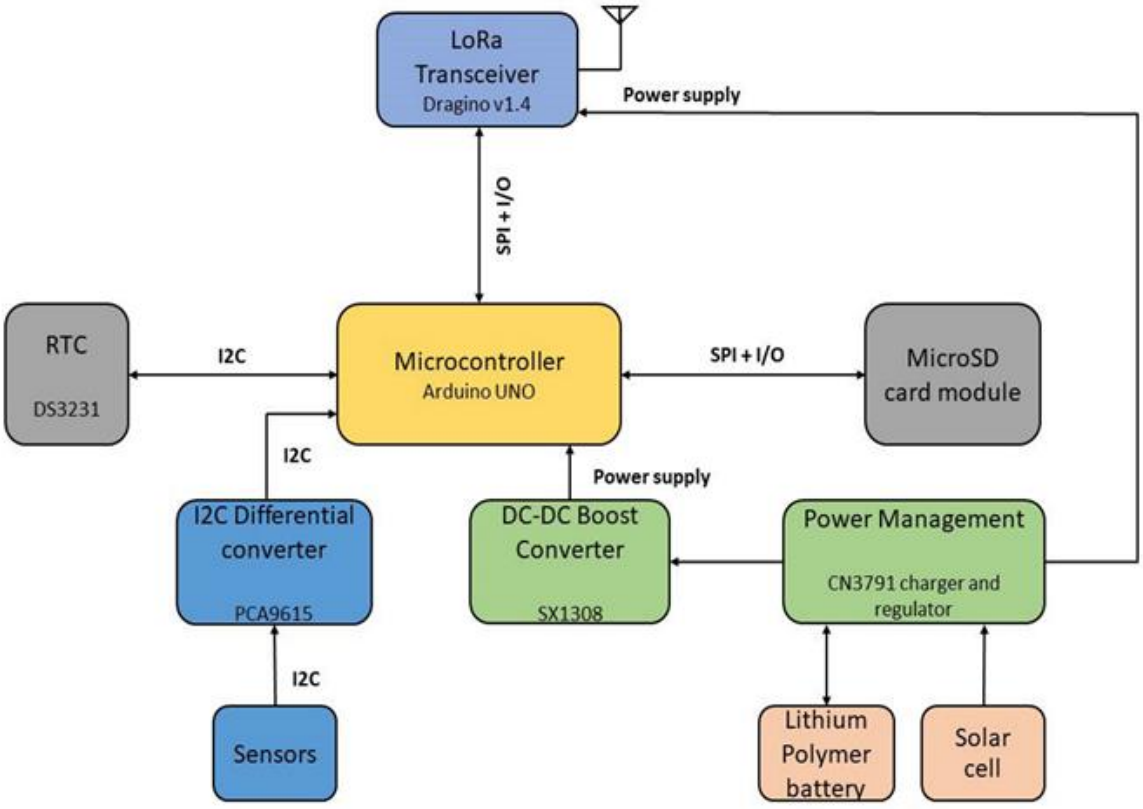



Figure 3.3 Block scheme of the LWNGM's field node architecture.

The RFM95W module on-board the Dragino transceiver (version 1.4) is specified to accept 3.3 V or 5 V input volt- age. In this case, a 3.3 V supply is used to meet the low-power scheme. The entire field node is powered by a rechargeable 3.7 V, 5000mAh LiPo battery via a 3.7 V to 5 V DC power boost converter (Shenzhen iSmart, Guangdong, China), which is supplemented by a 6 V, 3 W solar energy harvester. Because the applied power level (less than 12 V) is safe for the MCU, it is fed to the Arduino via a 5 V pin, avoiding any potential voltage regulator losses. The program that controls the end-node was also developed using the C language in Arduino IDE, Arduino, Somerville, MA, USA. The Arduino UNO microcontroller is programmed to spend most of the time sleeping to save power and meet low-power operational constraints. Every six hours (6 h), the RTC sends triggers (programmable interrupts) to wake up the MCU, allowing the water table depth sensor and atmospheric pressure sensor to perform measurements (see Figure 3.2). Before the data was temporarily logged into the SD card (16 GB ScanDisk) as a CSV file, each record was time stamped

and the measuring node identification number was attached. The end node returns to sleep. The RTC wakes up the MCU every twelve hours (12 h) to initialize and read data from the SD card. The LoRa shield is then activated to send values to the gateway. After successfully sending the data, the LoRa radio is turned off. Then the end-node returns to sleep. In this scheme, the end node spends the majority of its time in the low-power mode.

### 3.3.3 Preparation of Water Table Depth Probe

The following procedure was used to adapt the MS 5803-14B sensor (TE connectivity) for the desired task. To make the sensor and the I2C differential connector (PCA9615) waterproof, they were potted in epoxy resin. Following the wired connection of the PCA9615 and the MS5803 sensor, both devices were enclosed in a watertight metallic tube measuring 9.5 cm (3.74 in) length and 4.5 cm (1.772 in) diameter for waterproofing reinforcement. A 6 mm hole was drilled at one end of the metallic tube to expose the pressure measuring diaphragm to the water in the well. Another 6 mm hole was drilled on the opposite end of the tube to allow us to draw in the UTP cable that connects to the differential I2C breakout that uses the PCA9615 integrated circuit, as shown in Figure 3.4(a). This cable connects the sensor to the microcontroller board via two I2C extenders (PCA9615 converters). The tube was also potted to prevent water from leaking through the joints and drilled holes. The MBE280 climatic pressure sensor was epoxy-potted and hung 3 m from the top of the well casing in the bored well. As shown in Figure 3.4(b), the Arduino UNO board, LoRa shield, and all connected electronics were enclosed in a waterproof PVC enclosure. The unit was mounted on a pole 1.5 m above the ground, corresponding to the height of the end node. This position allowed the 868 MHz LoRa board's 3 dBi gain antenna to establish a line of sight with the 868 MHz GL01-P LoRa gateway's 3d Bi antenna, which was positioned 3 m above the ground. The three common bus protocols for communications between digital sensors and microcontrollers are the serial peripheral interface (SPI), one-wire, and integrated-integrated communication (I2C). All of these protocols are supported by both hardware and software libraries. The SPI and one-wire protocols can transfer large amounts of data at higher data rates, but they have drawbacks such as the SPI requiring dedicated communication pins and a one-wire with complex data synchronization at the receiver side. Moreover, the one-wire scheme is unpopular among hardware manufacturers and is susceptible to cable capacitance and noise. Despite the fact that the I2C has slow data and a shorter data transfer distance (<1 m) [49], we chose it for prototyping because of its ability to share and save communication pins. The most difficult task in the redesign and preparation of the end

device was to pot the sensor and the I2C differential interface with epoxy for underwater applications. Although the diaphragm of the MS5803 is contained and protected by a stainless steel cap, it may be easily damaged if it comes into contact with glossy and sticky materials such as the potting epoxy resin during the potting process. Several devices were damaged during the potting and testing phases before we came up with a viable solution. For example, when we potted and enclosed the MS sensor in a plastic container and housed the container in a PVC cylinder (see Figures 3.4(c) and 3.4(d)), the attempt was unsuccessful, allowing water to enter and damage it. The I2C converters (Sparkfun, Electronics, Colorado, USA) were used on Cat 6, UTP cable to overcome the limitation of the I2C protocol (capacitance effects on the signals increase with cable length). To extend the wire up to 25 m, one differential convertor was attached at each end of the cable (see Figure 3.5). Unlike most of the studies that used the MS5803 sensor family with cable length <10 m [176], [49], [177], this study successfully applied MS5803-14BA with a cable length of 25 m and overcome the limitations of the I2C protocol. As a result, a high-pressure sensor (MS5803) could be installed in the borehole at a depth of 15 m below ground. RJ45 connectors were used to connect the PCA9615 devices to both ends of the UTP cables. The serial data line (SDA), serial clock line (SCL), 3.3 V input line, and ground terminal point (GND) on the PCA9615 device correspond to the four connection pins on the MS5803 sensor. Hence, only four of the eight cores of the UTP cables were used to transfer data between the sensor and the UNO board via PCA9615 devices using the I2C protocol. Likewise, MBE280 used the same connection arrangements and protocols.

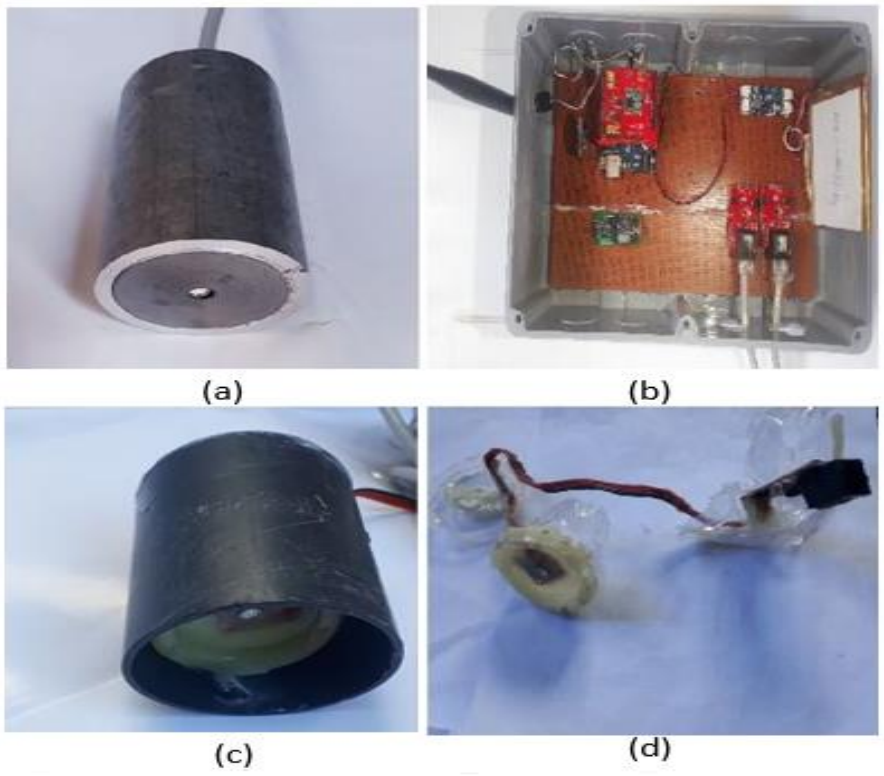



Figure 3.4 The LWNGM field node (a) the potted MS5803-14A and protected in aluminum cylinder (b) the field-node circuitry in a water-proof enclosure (top removed) (c) the MS5803-14A potted in a plastic container, housed in a PVC cylinder (d) the damaged MS5803-14A sensor in a potted plastic container.

Figure 3.5 The I2C differential interfacing connectors attached to each end of the Cat 6 UTP sensor connection cable (25 m).

Potting was done to protect the sensor from water and moisture during the construction of the barometric pressure probe using a BME280 sensor (MBE), but the length of the cable was set to 10 m, which is significantly shorter than the length of the underwater MS5803 cable. This is primarily due to the fact that the BME280 for atmospheric pressure measurement is hung inside the borehole three meters (3 m) from the borehole casting rather than being submerged in water.

### 3.3.4 Data Correction and Storage Description

The local server (LS) receives sensor data from the LG01 gateway twice a day via a wired connection. When data are received from the gateway, the Python script written in Python 3.6.6 Integrated Development Environment (IDE) runs to perform data processing by computing daily averages from raw data and then correcting the data for the influence of cable length and barometric pressure. The calculations of the measured water table depths were carried out in the following manner. Each averaged value of the water column pressure was first compensated for the average value of the atmospheric pressure using (3.1).

$$W_d(t) = P_w(t) - P_b(t) \tag{3.1}$$

Where, $W_d(t)$ refers to the depth of the water at time *t*, $P_w(t)$ the water column pressure on the submersible sensor at time *t*, and $P_b(t)$ ambient pressure at time *t*. We also need to calculate the actual length of the cable attached to the sensor. The ($C_l$) is computed in (3.2).

$$C_l(0) = P_w(0) + D_w(0) \tag{3.2}$$

Where, $C_l(0)$ is refers to the length of the cable at time $t = 0$, $P_w(0)$ pressure of water column above the submersible sensor at time $t = 0$, and $D_w(0)$ the depth to water at time $t = 0$. From equation 3.1 and 3.2, the actual water table height is subsequently obtained as follows:

$$A_h(t) = S_d - C_l(0) + W_d(t) \tag{3.3}$$

Where, $A_h(t)$ refers to the actual depth of the water table at time t, $S_d$ the average sea level datum in the study area.

The processed groundwater data are saved in the MySQL database. MySQL is a well-known, open-source, high- performance database that can be used for both on-premises and cloud-based IoT

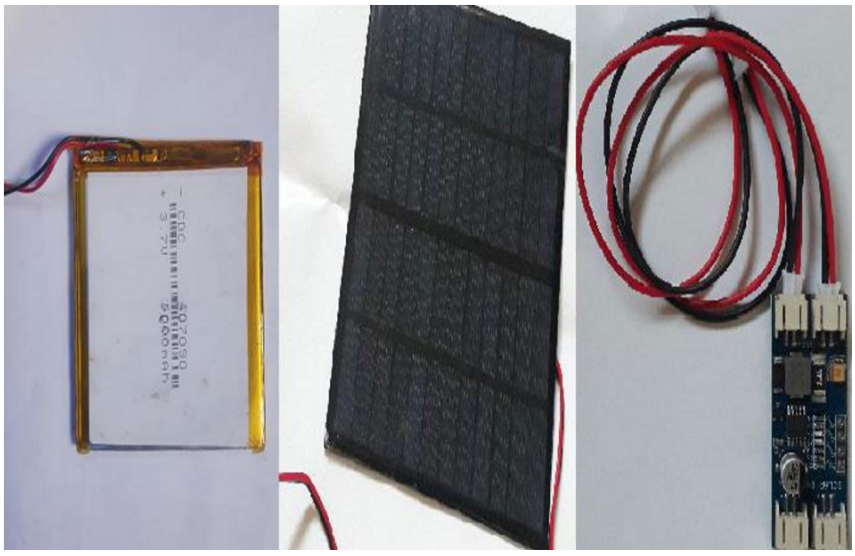
applications [178]. The data are then copied and uploaded to cloud storage for backup, sharing, and visualization (as explained in the next subsection). The database design allows for scalability, allowing for the addition of data from new observational stations. It can handle data from multiple stations while requiring only minor changes to the database configurations.

### 3.3.5 Cloud Server and Data Visualization

Sharing information among stakeholders and end users is a significant step toward resource management coherence [176]. The LGWMN cloud-based web portal was created to allow data sharing with potential stakeholders. The MySQL-powered website was created using the PHP and Java scripts. The monitoring dashboard has four primary functions: A map that allows the user to navigate the location of the borehole well, charts that show the trends and patterns of the water table depth variations over a specified period, and downloadable data in a CSV file format. The database also includes a configuration with information about the location and data rendering.

### 3.3.6 Energy Autonomy of the LWNGM

Energy autonomy is a critical requirement for IoT systems, particularly those deployed in remote and difficult-to-reach locations [179]. The chosen solar panels have a 3 W, 6 V solar charger that serves as a backup to a 3.7 V, 5000 mAh LiPo battery (Figure 3.6) that powers the end node. The gateway is powered by a DC-to-DC power booster with a constant output voltage of 12 V and a current rating of 1 A. This booster is linked to a 9 V, 600 mAh Li-ion battery, which is recharged by two mini solar panels (6 V, 3 W each) connected in series to produce a total output voltage of 12 V.



(a) (b) (c)

Figure 3.6 Solar devices for powering the nodes (a) 3.7V LiPo battery (b) solar panel (c) charger unit for batteries.

Despite the fact that we have two solar energy harvesters, we now only consider the estimation of charge and discharge statistics for the end node solar panel and battery. The CN3791 charger module charges the LiPo battery via a connected solar panel. The module also protects the battery from overcharging. The open circuit voltage ($V_{cc}$) of the solar panel is 6 V, the specific load voltage ($V_l$) is 4.2 V, the short circuit current ($I_{cc}$) is 2000 mA, the typical current at load (Il) is 500 mA, and the maximum power is 3 W. When fully charged, the LiPo outputs 4.2 V, which is connected to the Arduino board's 5 V pin. This pin was selected as the power input pin for the microcontroller. Weather conditions are one of the most important factors influencing the performance of solar panels [85]. According to the data we collected on the power efficiency of the solar panel connected to the end-node, the panel operates at 5.5 V/338 mA, 4.3 V/205 mA, and 4.0 V/123 mA for sunny, cloudy, and rainy hours, respectively. Tanzania has a high level of solar energy, with 2800-3500 hours of sunlight per year and global horizontal radiation ranging from 4 to 7 kWh per $m^2$ per day [85]. Based on this information, we assume that the study site has an average of eight sunny, one cloudy, and one rainy hours per day. The average energy ($C_{av}$) produced by the solar panel is then calculated using our solar performance data, as given in (3.4).

$$\boldsymbol{C_{av}} = [(5.5\text{ V} \times 338\text{ mA} \times 10^{-3} \times 7) + (4.3\text{ V} \times 205\text{ mA} \times 10^{-3} \times 2) + (4.0\text{ V} \times 123\text{ mA} \times 10^{-3} \times 1)] \times 3600 \quad (3.4)$$

$$= 54{,}964.8\text{ J}$$

The results computed in (3.4) can be used to estimate the time required to charge the battery ($t_{ch}$); however, we must first calculate the amount of energy produced by the battery. Given a battery with a charge capacity ($Batt_{cap}$) of 5000 mAh and a voltage rating ($Batt_{vt}$) of 3.7 V, the energy of the battery (Batten) as a function of Battcap and battery voltage $Batt_{vt}$ is computed as given in (3.5).

$$Batt_{en} = (Batt_{cap} \times Batt_{vt} \times 60\text{s} \times 60\text{s}) \quad (3.5)$$

$$= 66{,}600\text{ J}$$

The $t_{ch}$ is then calculated by dividing the battery energy by the energy produced by the solar panel, as shown in (3.6).

$$t_{ch} = \frac{Batt_{cap}}{C_{av}} \quad (3.6)$$

$$= \frac{(5000\text{mAh} \times 3.7 \times 60 \times 60)}{54964.8\ x1000}$$

= 1.212 days

According to the calculation in (3.6), it takes approximately 1.212 days (29.080 h) for the solar energy harvester to fully charge the battery. As a result, the LGWMN will be powered on a daily basis by this power source (solar charger). The LWNGM architecture is depicted in Figure 3.7, with the sensor devices connected to the microcontroller board via I2C differential interfacing and the RTC directly connected to the microcontroller's I2C connections. The SPI bus connects the micro SD card component and the LoRa breakout to the MCU. Underwater and ambient pressure sensors were installed in the well. A LoRa-based connection is provided between the LoRa breakout and the gateway. The GW is outfitted with a 4G LTE dongle that connects it to the network server, which is linked to the cloud-based application server. Rechargeable solar-powered batteries power the nodes. The local server handles data processing, whereas the cloud server handles data backup, sharing, and visualization.

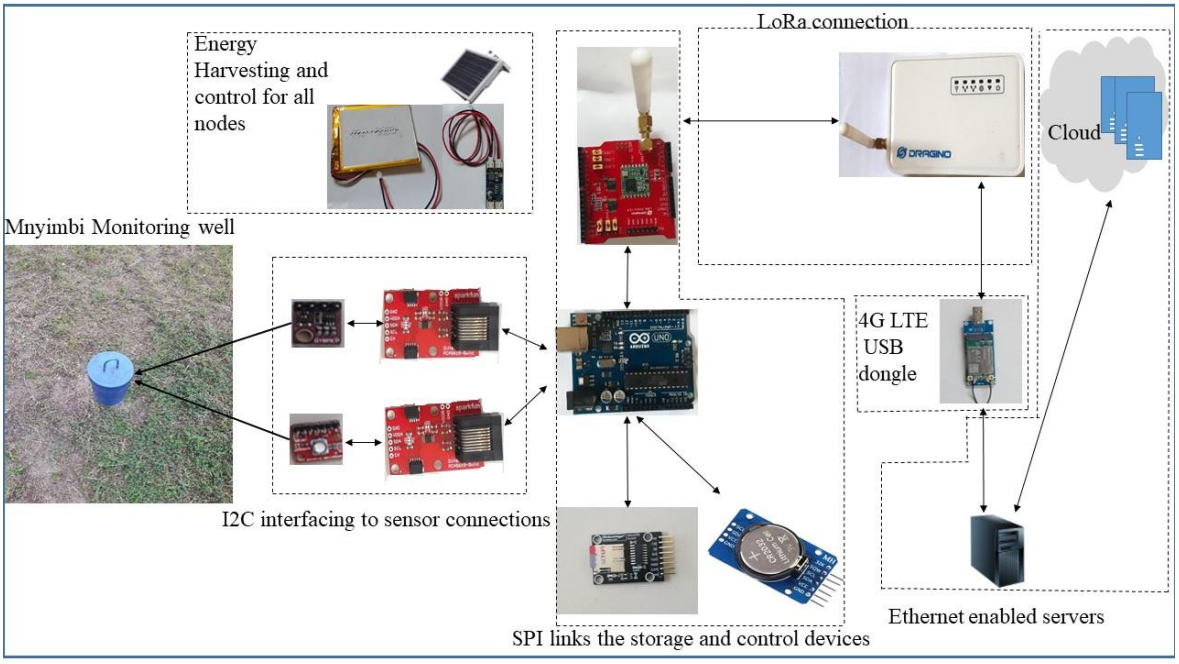



Figure 3.7 LWNGM platform: the well, devices, sensors, power source, and communication protocols deployment.

## 3.4. Experiment Evaluation and Test Results

This section describes the process and results of an experiment related to this case study in the adoption of low-cost, low-power WSN-based equipment. The end node, gateway, and web dashboard were designed, built, programmed, and tested to realize the proposed LWNGM prototype. After the

system was deployed in the field, it was evaluated. At the time of our prototype's deployment, there were no sensors in the case study, and monthly measurements were taken with a beeper tap. The tap is normally lowered into the borehole and makes a sound when it comes into contact with water. We used a beeper tap to measure the depth to water prior to installing the electronic sensor in the Bandamaji observation well to establish a reference point for automatic measurements. Each water table depth and air pressure measurement, as well as the identification number of the associated measuring sensor and the timestamp generated by the DS3231 clock, are temporarily logged into the SD card memory by the node. The recorded values are then relayed to the gateway at the scheduled time before the data are transmitted from the gateway to the network server.

### 3.4.1 LWNGM System Deployment

The LoRa-GSM-based logger was deployed between May 2 and 15, 2021. The system continuously sampled and transmitted data to the local server over a two-week period. The locations of the end-node and the gateway were carefully chosen to provide the best possible line of sight through the area with minimal vegetation at the study site. At the distance of 125 m from the end-node, the gateway is located in the doors near an open window (shown in Figure 3.8(a)). The vegetation and tall trees slightly shaded the line of sight between the gateway and end nodes. The line of sight is clear in most of the time. Figure 3.8(b) depicts the field node of the deployed system. The deployment also considered the possibility of scaling up the network while retaining the single-hop topology. This allows for the addition of more end nodes to share the gateway while ensuring network longevity. To maximize the energy of the system, the solar energy harvesters were placed in areas where there was no shade or materials such as tree leaves that could block the surface of the solar panels.


(a)

(b)

Figure 3.8 Field deployment of LWNGM (a) LG01-P long-range gateway with 4G LTE USB dongle (b) field node in a waterproof enclosure with mini solar panel deployed mounted to a pole at Bandamaji monitoring well.

### 3.4.2 Configuration, Key Parameters, and Network Performance

The optimal efficiency for the LoRa link was achieved with key parameter settings of 14 dB, 4/5, 868 MHz, 125 kHz, and 7, respectively, for transmission power, coding rate, transmission frequency, bandwidth, and spread factor (see Table 3.2).

Table 3.2 Configured LoRa transmission parameters

| Parameter | Value |
|---|---|
| Transmission power | 14 dBm |
| Bandwidth | 125 kHz |
| Frequency | 868 MHz |
| Spread factor | SF 7 |
| Coding rate | 4/5 |

To evaluate the performance of our network, we sent a series of packets from the end node to the gateway and computed the packet delivery ratio (PDR) for those packets. The PDR is calculated as the ratio of the total number of packets that successfully arrive at the receiver (gateway) to the total number of packets that leave the source (end node). We also measured the received signal strength (RSSI) and airtime for packet transmission in addition to the PDR. Table 3.3 shows the average of the LoRa radio performance values.

Table 3.3 Performance metrics for the LGWMN LoRa network

| PDR (%) | RSSI (dB) | Airtime (ms) |
|---|---|---|
| 84.4600 | -83 | 37.1300 |

### 3.4.3 System Testing and Data

The deployment in the field allowed the LWNGM to be tested. The mean water level was calculated in six-hour (6 h) intervals of data sampling cycle and twelve-hour (12 h) intervals of data forwarding cycle to provide daily average levels. Each send had an average airtime of 19 ms. The obtained test results were compared to those of previous studies. Figure 3.9 compares the performance of LoRa

radio links demonstrated by our proposed system (LWNGM) to studies conducted by Codeluppi et al. [180], Yousuf et al. [82], Cattani et al. [181], and Augustin et al. [75]. During the field deployment of the LWNGM system, sampled data were recorded and visualized using a web dashboard. Figure 3.10 (a) depicts the daily average water table depths recorded over a two-week period, and Figure 3.10 (b) depicts a web dashboard that updates groundwater depth data every six hours.

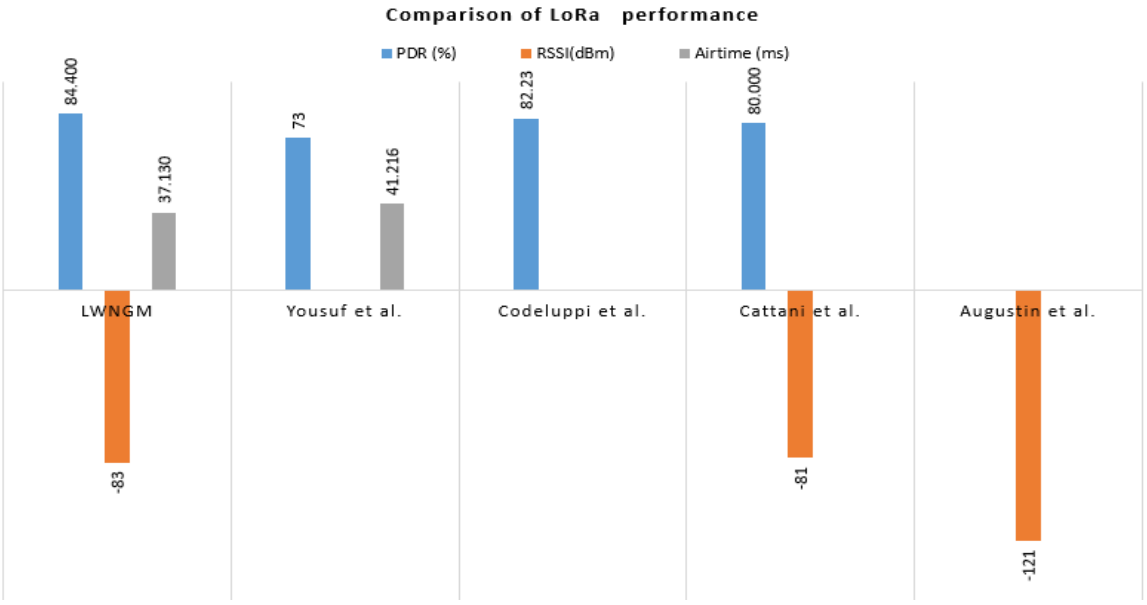



Figure 3.9 Comparison of LoRa radio performance statistics in terms of RSSI, PDR, and airtime for various studies.

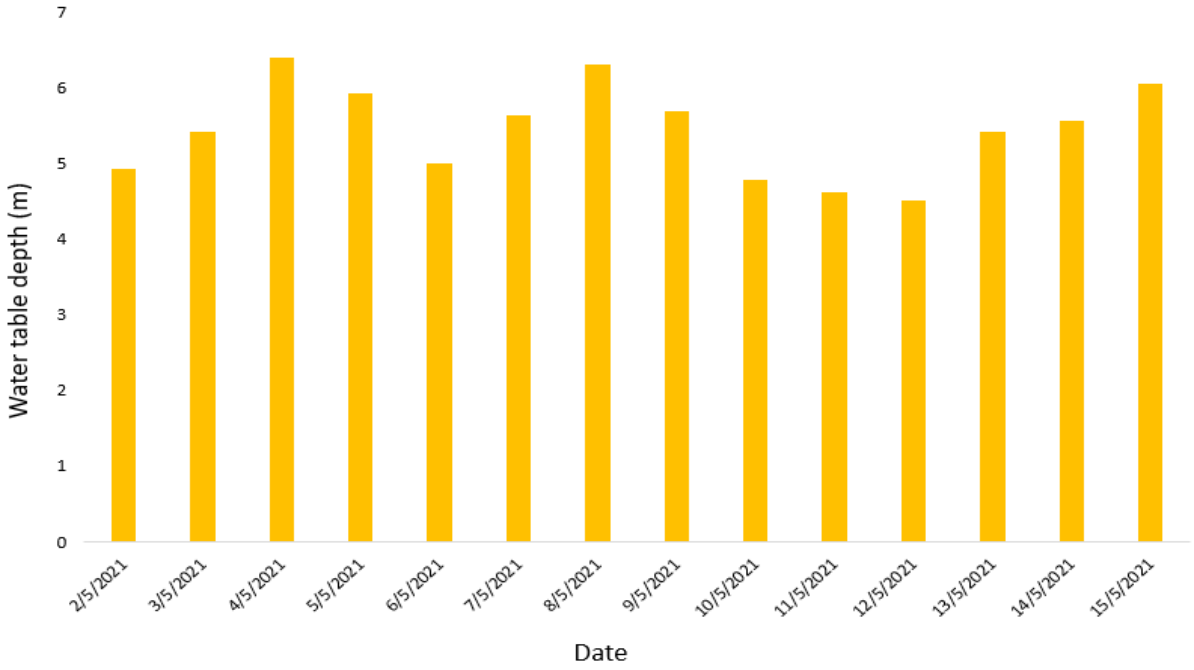



Figure 3.10 (a) A plot of daily average water depths at bandamaji monitoring well over a two-week deployment period.

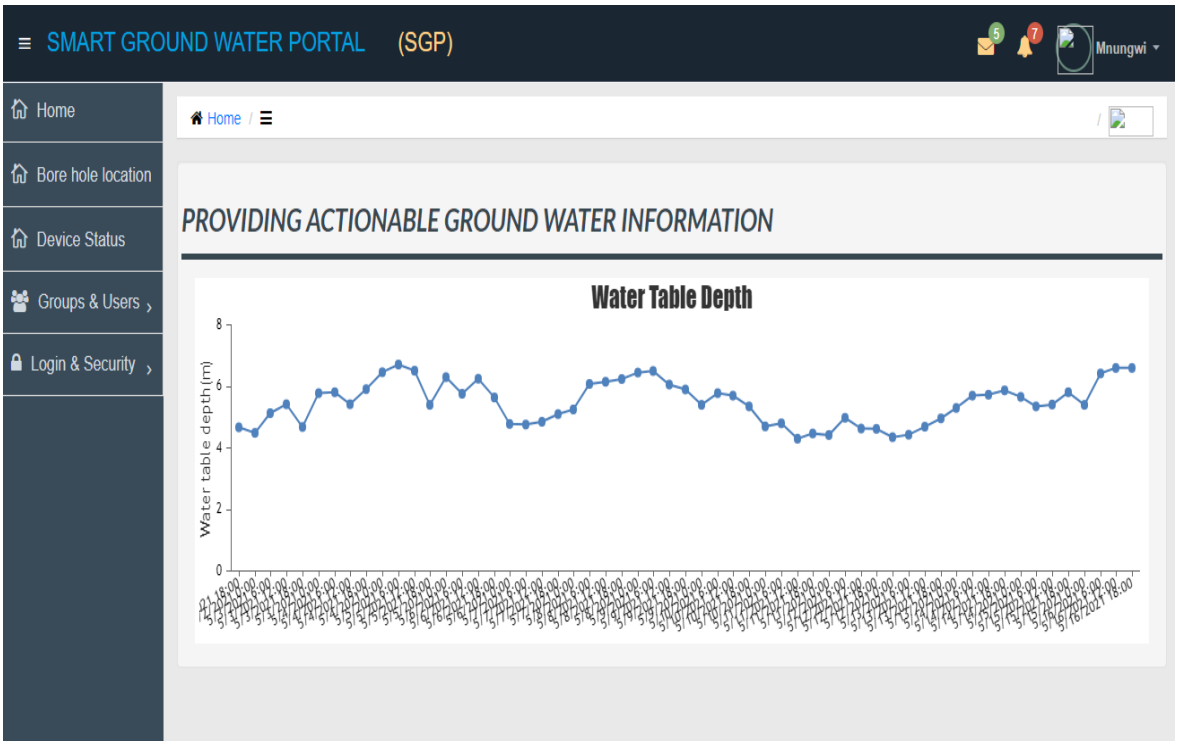



Figure 3.10 (b) Web dashboard for LWNGM showing a plot of two weeks data from Bandamaji station at 6-hour intervals.

The front and back ends of the LWNGM web portal were designed to be flexible enough to accommodate data from additional observational stations. The user interactive map is being developed, which will allow the navigation and visualization of data from multiple stations.

### 3.4.4 Cost and Simplicity of the Systems

The costs of the components used to construct the LGWMN are summarized in Table 3.4. It is worth noting that the prices of these devices vary depending on the supplier chosen. These prices include shipping costs. The system is simple and easy to replicate. Although the project consists of multiple domains, a moderate wireless network and electronics skills and tools can accomplish the development of the system presented here. The total cost of the components used to construct the LWNGM was USD 310.168. This means that this system can be built for less than 400 USD.

Table 3.4 Summary of the quantity and pricing of the LWNGM components

| Component | Number | Unit Price ($) | Total Price ($) |
|---|---|---|---|
| Arduino board | 1 | 26.6200 | 26.6200 |
| MicroSD card shield | 1 | 2.7900 | 2.7900 |
| MicroSD card | 1 | 4.5950 | 4.5950 |
| Pressure sensor MS5803-14BA | 1 | 22.0400 | 22.0400 |
| LiPo | 1 | 6.7000 | 6.7000 |
| LoRa gateway | 1 | 96.8300 | 96.8300 |
| Solar energy harvester | 1 | 53.8500 | 53.8500 |
| Real-time clock | 1 | 4.7000 | 4.7000 |
| Sensor connection cable | 35 m | 0.4330 | 15.1500 |

| | | | |
|---|---|---|---|
| I2C differential interface | 2 | 9.5290 | 19.0580 |
| LoRa breakout shield | 1 | 38.1250 | 38.1250 |
| 4G LTE USB modem | 1 | 9.8900 | 9.8900 |
| DC to DC converter | 1 | 3.1300 | 3.1300 |

### 3.4.5 Energy and Lifetime of the End Node

The experimental evaluation of energy dissipation by the sensors and the general network is explained below. In all scenarios, we estimated the average power expense (data sampling state, data transmission state, and deep sleep state). We also estimated the lifespan of the battery that powers the end node. A battery that powers the end-node has voltage rating of 3.7 V at power capacity of 5000 mAh. To calculate the energy expended by the end node, we consider the current drawn by each of the individual components at the various aforementioned states. According to the current ratings of the devices, the standby current of the pressure sensors is less than 0.15 µA for MS5803 and 0.5 µA for MBE280. The MCU's standby current was 28 mA. The RTC was powered by a backup coin cell at 3 V, 0.19 A. The sleep currents for the UNO board, MBE280, MS5803-14BA, and LoRa transceiver breakouts are 27.9 mA, 0.1 µA, 0.1 µA, and 1 mA, respectively.

The MBE280 sensors have an average data sampling duration of 1 s at 1.8 µA and the MS5803-14BA sensors have an average data sampling duration of 1.1 ms at 1 µA. Unlike writing to an SD card, the UNO takes an average of 7 ms to complete at a current of 31 mA. In contrast, the SD card draws an average of 0.11 mA while sleeping. The sleep state consumed a total current of 28.0112 mA in this case. The node remained in deep sleep ($t_{sleep}$) for 3598.653 s. We can calculate the amount of charge consumed as the product of sleep duration and the current drawn while sleeping. As a result, the charge expended ($C_{sleep}$, in $m_{As}$) was 100,801.869 mAs. Similarly, the end node enters a data sampling state for a sampling time ($t_{samp}$) of 1.007 s and the consumption ($C_{samp}$) of 31.282 mAs. Here, we define the data-sampling state as the process of measuring data and logging it to an SD card. Furthermore, relaying the sampled data to the gateway consumes an average current of 28 mA at a transmission time ($t_{tx}$) of 0.34 s, with a total charge consumption ($C_{tx}$ ) of 9.52 mAs. The complete cycle of the operation includes both the active and sleep times of the end node. During

duty cycle *d*, the amount of energy consumed by this node reaches a maximum value, which is given as.

$$d = \frac{t_{act}}{T} \tag{3.7}$$

where *T* refers to the total time spent in the entire cycle, $t_{act}$ denotes the time at which the device is in the wake state. The LGWMS records two (2) measurements ($N_{samp}$) in each data-sampling period, with an average sampling time ($t_{avsamp}$) of 0.34 s. The total charge consumption over the cycle ($O_{Ccycle}$) is expressed in (3.8) and the battery cycle in (3.9).

$$OC_{cycle} = N_{samp}.C_{samp}(t_{samp}) + N_{samp}.C_{sleep}(t_{sleep}) \tag{3.8}$$

The number of battery cycles that would be performed on this battery with a given battery capacity (Battcap) is.

$$Batt_{cycle} = \frac{Batt_{cap}}{OC_{cycle}} \tag{3.9}$$

Because we need to estimate the battery's lifespan ($L_{batt}$), the number of battery cycles ($B_{attcycle}$) must be multiplied by the battery cycle time ($t_{cycle}$). The cycle time is defined as the sum of the sleep time ($t_{sleep}$), measurement time ($t_{samp}$), and data transmission time ($t_{tx}$) in seconds (3.10).

$$t_{cycle} = N_{samp}.t_{sleep} + N_{samp}.(t_{samp} + t_{avsamp}) + t_{tx} \tag{3.10}$$

$$L_{batt} = OC_{cycle}.t_{cycle} = \frac{Batt_{cap}}{OC_{cycle}}.(N_{samp}.(t_{sleep} + t_{samp} + t_{avsamp}) + t_{tx}) \tag{3.11}$$

By substituting (3.8), (3.9), (3.10), and (3.11), the life span of the battery ($L_{batt}$) that provides energy to the end node is calculated as follows.

$$L_{batt} = \frac{Batt_{cap}.(N_{samp}.(t_{sleep} + t_{samp} + t_{avsamp}) + t_{tx})}{N_{samp}.(C_{samp} + C_{sleep}) + C_{tx}} \tag{3.12}$$

Using the capacity of the battery used in the experiment and previous numerical values, (3.12) produces 177.948 h, which is approximately equivalent to 8 days for the estimation of the battery's lifetime in hours. The total energy expended by the end node ($E_T$) is the sum of the energy expended

in data sampling ($E_{samp}$), data transmission ($E_{tx}$), and sleep mode ($E_{sleep}$). Given that the end-node is powered by a 3.3 V supply voltage and that the average sampling time is 1.007 s, the energy expended in data sampling will be.

$$E_{samp} = 31.0028 \text{ mA x } 1.0028 \text{ s x } 3.3 \text{ V x } 24$$

$$= 24.623 \text{ J}$$

For the transmission duration of 0.34 s, the energy consumed for relaying data to the gateway is calculated as given below.

$$E_{tx} = 28 \text{ mA x } 0.34 \text{ s x } 3.3 \text{ V x } 24$$

$$= 0.754 \text{ J}$$

Because the node sleeps for 3598.653 s, the energy spent while sleeping is. $E_{sleep}$

$$E_{sleep} = 27.9 \text{ mA x } 3598.653 \text{ s x } 3.3 \text{ V x } 24$$

$$= 7951.872 \text{ J}$$

Consequently, the total energy spent by the end node (ET) per day is.

$$E_T = E_{tx} + E_{samp} + E_{sleep} \tag{3.13}$$

$$= 24.623 \text{ J} + 0.754 \text{ J} + 79518.72 \text{ J}$$

$$= 7977.249 \text{ J}$$

Based on the energy calculations, the end node's total daily energy consumption in both active and sleep modes was 7977.249 J.

## 3.5 Discussion

Groundwater monitoring on a regular and affordable cost serves as the foundation for estimating, assessing, and forecasting the quantity of this resource. The high cost of commercial instrumentation is one of the major impediments to the rapid adoption of WSNs in wider hydrologic applications, particularly in developing countries [182], [183], [184]. Ordinary water depth electronic sensors are relatively expensive. For example, the Van Essen diver sensor costs approximately $830, and the HOBO water depth probe costs approximately $495 per unit. Our redesigned water table depth probe, on the other hand, costs approximately $55. The use of potting of high precision electronic

sensors and I2C extenders is essential for the success of the construction of water-depth probes for the LWNGM system. Moreover, the application of free and open source software has also reduced the establishment and operational cost for and may improve sustainability aspect of the LWNGM. This finding is consistent with the one reported in [84]. Furthermore, the LWNGM development procedure is relatively simple and does not require advanced technical skills.

Incorporating user-redesigned LoRa tools in the monitoring of environmental phenomena, on the other hand, lowers the overall cost of the system while improving efficiency and reliability. The game changer that has provided a wider community with LoRAWAN tools is an open-source instrument and software [82]. The redesign of off-the-shelf instruments and the deployment of low-cost WSNs allow communities and organizations with limited resources to adopt this technology more efficiently. The LWNGM offers another opportunity for wider adoption in hydrology and other fields at a cost less than ($500) comparable commercial solutions (i.e., monitoring systems that use the Van Essen water depth probe, which costs approximately ($830). Sending the sensor node into a sleep state lowers the duty cycle and increases battery life, which is especially important in low-cost and low-power nodes. Ideally, with a sleep mode that brings low duty cycle (<1%), our system used approximately 1.343 percent of the battery energy. This is consistent with the findings in [85]. Although it takes approximately 1.212 day to fully charge the chosen battery using a mini solar cell, it can power the end node for approximately 8 days before it needs to be recharged. Connecting two 6 V panels in series to charge the 9 V battery that powers the gateway via a 9 V-to-12 V booster, on the other hand, produced a promising resource optimization result. This not only saved money, but also allowed the gateway to be deployed in off-grid areas. This dependable source of power also allows for the prevention of data loss due to power outages.

Fine-tuning and configuration of network parameters can result in the effective use of LoRa technology for efficient communication. The tuning parameters, among other things, allow for the optimization of power consumption, signal distance, and data rate. The greater the SF, the greater the PDR, and the greater the observed airtime. Similarly, as the airtime lengthens, so does the power budget [82].

Furthermore, the topography and local environment at the site influence signal reception quality. Despite its strong modulation technique, LoRa is sensitive to the presence of reflections and obstacles. Mnyimbi is a rural area with little vegetation and fewer signal obstacles and reflectors.

These favorable conditions allowed for the majority of the time to be spent installing LWNGM nodes in positions that maintain a clear line of sight between LWNGM nodes. The line of least resistance to tree leaves was established at 1.5 m for the end node and 3 m for the LoRa-4G-based gateway. Despite the slight signal attenuation (due to the tree branches moving with the wind), the LoRa signal was found to be stable and reliable. No significant impact on the received signals was observed due to the moderate weather conditions, with the mean air temperature and onboard temperature both below 40◦C and rain rates below 100 mm/h. The results of our experiments show that signals from the LWNGM's end node arrive at the gateway node correctly (RSSI=83 percent) with a low number of retransmissions. The vast majority of packets arrive at the gateway correctly (PDR > 80%) and with a reasonable airtime (40 ms). This was accomplished with SF7, CR4/5, a TX of 14 dB, a BW of 125, and the CRC enabled. As shown in Figure 3.9, the system reliability level demonstrated by the proposed LWNGM is more promising than that presented in [75], [82], [180], [181].

Two possible explanations for these results are the distance (125 m) and the slightly clear line of sight (light vegetation) between the gateway and the end node. The accuracy of the samples collected by the sensors has a direct impact on the measurement quality. According to the results, the MS5803-14BA sensor data deviated by 1.1 % from the standard measurement performed with a beeper. Furthermore, the MBE280 data deviated from standard atmospheric pressure by 0.37 %. The linear calibration produced the most accurate results in filling the gaps between the standard measurements and the Arduino-based measurements in this case. As a result, data collected by the LWNGM system has become more reliable. This is consistent with the results of the study in [84], which confirms the applicability of linear calibration in sensor data correction. Furthermore, in order to maintain the accuracy of the measurements over time, the effective cable length must be recalculated every six months [185].

Unlike manual data collection, where the cost of labor and field visits is a significant barrier to informed groundwater management, low-cost, automated monitoring with energy harvesting creates the potential for continuous observations, especially in low- and middle-income countries. Similarly, the quality and reliability of data collected by LWNGM are far superior to data collected using traditional methods. This is due to the fact that in traditional data collection practices, errors are easily introduced, and missing data points are a common occurrence.

## 3.6 Conclusion

This chapter has described the design, development, application, and evaluation of an autonomous low-power, low-cost IoT-based system with energy harvesting to provide a proof-of-concept for practical monitoring of water table depths. The developed LWNGM system consists of four parts: data acquisition, data management, energy harvesting and management, and data storage and visualization. The developed low-cost solution is built on an open platform. The LWNGM generates critical information for more efficient assessment and management of groundwater tables. Furthermore, the system's information down streaming capability allows for additional research in the fields of hydrology and sensor networks.

We potted the electronic sensor and extended the I2C-enabled communication channel up to 25 m via a PVC cable, for underwater application in the borehole. The system runs on batteries supported by the reliable tiny solar cells. The outstanding efficiency and low cost of redesigned sensor nodes and energy harvesters have been evaluated to be the promising alternatives to conventional instruments. The prototype system was used to monitor groundwater wells at the Bandamaji station in Zanzibar, Tanzania. The system is easily transferable, even to least developed countries, because it is built with low-cost components and does not require advanced technical skills. The system performed admirably and allowed for near real-time monitoring of changes in water table depth. Additionally, the groundwater monitoring network can be expanded to a larger area or even the entire country thanks to the inexpensive and simple-to-deploy system that is being proposed. This can be done by setting up multichannel outdoor LoRa gateways to link numerous end nodes spread throughout the nation. In order to improve the monitoring and evaluation of the aquifers, the system also offers the possibility of integrating low-cost water quality sensors.

# CHAPTER 4

# LONG-TERM GROUNDWATER LEVEL PREDICTION MODEL BASED ON HYBRID KNN-RF TECHNIQUE

## 4.1 Introduction

This chapter describes the development and evaluation of a hybrid KNN-RF machine learning technique that implements research objectives four and five, which state: *To develop an efficient machine learning model for seasonal prediction of groundwater levels, and To evaluate the developed machine learning model using appropriate performance metrics.*

In order to achieve the aforementioned objectives, an ensemble KNN-RF with time series preprocessing was developed to predict seasonal variations in groundwater table levels in a data-scarce environment. Provided with limited and noisy examples, we harness the merits of both techniques by combining them in a hybrid manner. Reliable seasonal prediction of groundwater levels is not always possible when the quality and the amount of available on-site groundwater data are limited. In this chapter, a hybrid K-Nearest Neighbor-Random Forest (KNN-RF) is used for the prediction of variations in groundwater levels (L) of an aquifer with the groundwater relatively close to the surface (<10 m) is presented. First, the time-series smoothing methods are applied to improve the quality of groundwater data. Then, the ensemble K-Nearest Neighbor-Random Forest (KNN-RF) model is treated using hydro-climatic data for the prediction of variations in the levels of the groundwater tables up to three months ahead. Climatic and groundwater data collected from eastern Rwanda were used for validation of the model on a rolling window basis. Potential predictors were: the observed daily mean temperature (T), precipitation (P), and daily maximum solar radiation (S). Previous day's precipitation P (reference day(t) − 1), solar radiation S (t), temperature T (t), and groundwater level L (t) showed the highest variation in the fluctuations of the groundwater tables. The KNN-RF model presents its results in an intelligible manner. Experimental results have confirmed the high performance of the proposed model in terms of root mean square error (RMSE = 0.0031), mean absolute error (MAE = 0.0022), Nash–Sutcliffe (NSE = 0.9346), and coefficient of determination ($R^2$ = 0.9387) at 90-day lead-time.

The chapter also describes the context of the research, preparation of the data, and the metrics used to assess the performance of the hybrid KNN-RF model. It also explains the development, tuning and testing of this model as well as the obtained outcomes.

## 4.2. Contributions

The work described in this chapter has made the following contributions: first, it devised and tested the performance of the KNN-RF ensemble scheme with limited data; second, it characterized the seasonal response of the permeable fractured aquifer in a temperate region with limited groundwater studies; and third, it compares the proposed KNN-RF model to conventional groundwater modeling techniques (SVM, RF, KNN, and ANN). In doing so, the chapter provides Rwanda's first ML-based seasonal approximation of groundwater level. It also sheds light on the novel KNN-RF approach's suitability for sub-Saharan semi-arid conditions.

## 4.3 Case Study and Data Processing

This section discusses the characteristics of the area under investigation, the sources, the nature of the research data, the preparation of the data, and the evaluation metrics used for the current exploration.

### 4.3.1. Study Area and Data

The investigated well is found in eastern Rwanda, which lies between 29.86875E–29.90625E and 2.30625S–2.26875S with a total area of 9813 $km^2$ (3789 sqmi). This region is relatively flat with the altitude ranging between 1000 and 1500 m [186]. During the study period between December 2016 and December 2018, the majority of the rain showered in the wet season between March and May (90%), with rainfall ranges between 450 mm and 500 mm. This is less than when compared to other parts of the country, that receive on average between 600 mm and 800 mm annually [187]. The eastern province is characterized by the highest evapotranspiration rate in Rwanda. The average annual temperature varies between 15.70 °C and 24.20 °C. The average minimum temperatures (13 °C–16.65 °C) are recorded in May and June, while the maximum average temperatures (24.3 °C–30.3 °C) are recorded in July and September [187]. The eastern province is the most populated area in Rwanda [188], and is heavily reliant on groundwater as a source of fresh water.

The groundwater abstraction rate is 742 $m^3$/h, while the demand is estimated to be between 3069 and 7672 $m^3$/h [189]. This area has highly heterogeneous types of aquifers. Those aquifers range from low permeable fractured (schist), which is located in Rugarama; permeable fractured (quartzite) located in Mukarange; and fractured (granite), which is found in Ruhuha. The Rwanda Water and Forestry Authority (RWFA) has groundwater stations in each of those three aforementioned areas. A summary of the monitoring well and its features are shown in Table 4.1.

Table 4.1 Summary of the selected monitoring well and its main features

| **Station ID** | **Name** | **Latitude** | **Longitude** | **Aquifer** | **Availability of data** | **Data time resolution** |
|---|---|---|---|---|---|---|
| F6 | Kayonza-Mukarange | 1.89874154 | 30.5065299 | Permeable Fractured | Dec 3, 2016-Dec 30, 2018 | Daily |

Groundwater data for 2016, 2017, and 2018 were gathered from the Rwanda Water and Forestry Authority (RWFA). Weather records (precipitation, solar radiation, and temperature) are available for a longer period, and researchers decided to use only data matching to the observational period of the groundwater level. Weather data (precipitation, solar radiation, and temperature) for the period of two years, from 3 December 2016, to 30 December 2018, were obtained from nearby weather stations (Kawangire, Kibungo, and Nyagatare) that are operated by the Rwanda meteorological agency (Meteo-Rwanda). Temperature is a daily minimum and maximum observed metric measured in Celsius, daily precipitation is recorded in millimeters, while groundwater level is measured in centimeters obtained from two measurements of groundwater depth per day. For consistency reasons, the groundwater unit was converted into meters. Solar radiation is measured in watt per meter square ($W/m$) and is included in the study because it influences the evaporation and evapotranspiration [190]. The locations of the groundwater and weather stations in the case study are depicted in Figure 4.1.

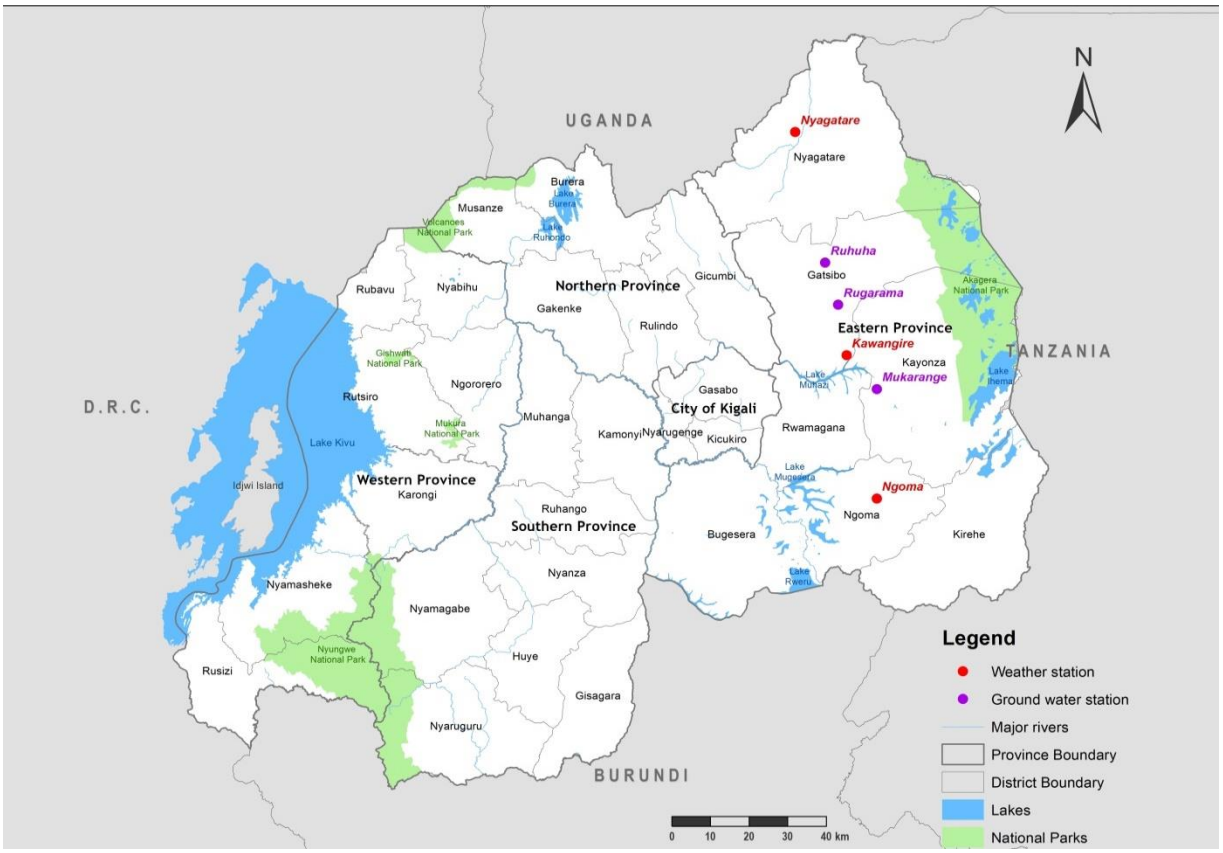



Figure 4.1 Map of Rwanda shows the location of the groundwater monitoring wells and weather stations in the eastern province.

The eastern province has the highest number of boreholes and shallow wells in Rwanda. Generally, the eastern part has high-localized fractured aquifers with moderate groundwater yields, as represented in Figure 4.2. Alluvium based aquifers are mostly connected to fast-flowing rivers. These aquifers exhibit high groundwater potential in the eastern province [189]. The studied aquifer (Mukarange) is made of quartzite rocks fused on a schist base with a relatively high yield.

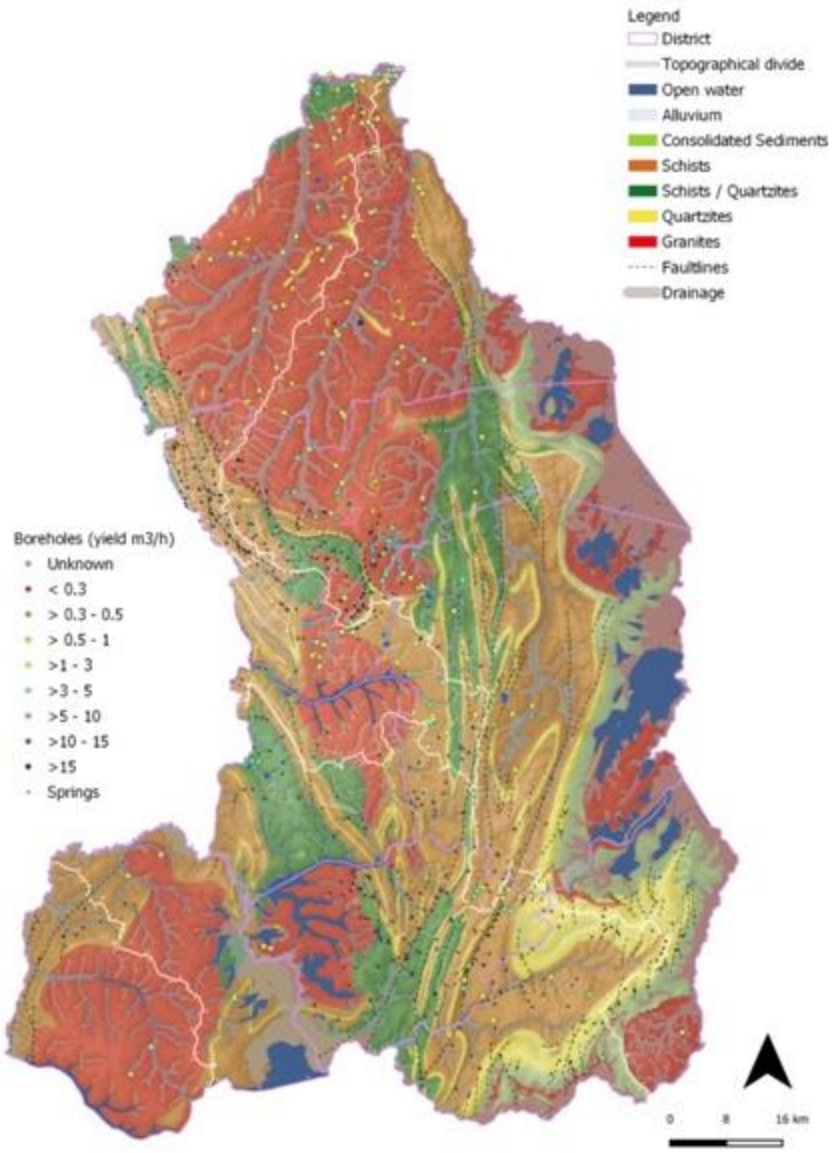



Figure 4.2 Eastern Province, Rwanda. Map shows hydrogeological features of the study area [76].

### 4.3.2 Data Preparation

The state of the input data is one of the key factors that determines the level of accuracy of ML-based predictions. Preprocessing and rectification of the variables ensure that all features receive equal attention throughout the training process [117][136][141]. A total of 759 daily observations from each of the above-mentioned stations were analyzed and prepared before application to the designated task [191]. Water level data were available on 12 h intervals and precipitation and temperature data were on a 24 h basis. In order to set common time intervals, we converted temperature and levels to daily averages. Data preparation was carried out with Python (3.6.6)

programming language using Pandas, Numpy, Matplotlib, and SciPy data analysis libraries [192]. The RF, SVR, ANN, KNN, and KNN-RF models for prediction of groundwater levels were also realized in Python using the Scikit-Learn machine learning library (version 0.20) [193].

Water level data were available between 3 December 2016, and 30 December 2018. Weather data, including precipitation, temperature, and solar radiation data between 2010 and 2018 were acquired, and weather data between December 2016 and December 2018 that correlated to the water level data were utilized in the experiment. Evapotranspiration is one of the key factors that influence groundwater level oscillations [117]. Since evapotranspiration data were not available, solar radiation data were successfully substituted instead, as suggested in [194], [195]. Temperature and groundwater level (GWL) data were converted to mean values to reduce variance among data points as recommended by [196], [197]. During the analysis of the data, it was discovered that groundwater data had irregular patterns. Then, the time-series data filtering method (exponential moving average) was also used to improve the quality of that data. The exponential weighted moving average (rolling mean) produced a superior output of the groundwater level samples. This not only filtered the data, but also revealed long term trends from the data. The exponential rolling mean of a sequence $S = (x_1, x_2, \ldots, x_k)$, is:

$$S_k = \alpha \sum_{j=1}^{k} (1-\alpha)^j x[k-j], \tag{4.1}$$

$$\alpha = 1/(1+idc), idc \geq 0, \tag{4.2}$$

where $S_k$ is the filtered data, $k$ is the size of $S$, $\alpha$ is the decay in the interval [0,1], $idc$ is the initial value of the decay and $x$ is the input data. As the exponential weighted average is calculated, the decay $\alpha$ value decreases exponentially in such a way the most recent observations are assigned higher weights than the old ones. For proper scaling of the time-series data, for each stage the features($X_j$) were converted in the range between -1 and 1 with the formula:

$$X_s = \frac{X_j - X_{mean}}{2X_{max}}, \tag{4.3}$$

where, $X_s$ stands for standardized value, $X_{max}$, and $X_{min}$ is the maximum value and minimum value of the features to be scaled, respectively. Despite the effort made to improve the samples, data from two of the observation boreholes (Ruhuha and Rugarama) found to be incurable. Therefore, only one borehole (Mukarange) with 759 observations is considered for the current investigation. Then this groundwater level data was matched with weather data recorded in the same time period (2016-2018) from the nearby station located between -1.81 latitude and 30.43 longitude in Kawangire. The useful preprocessed dataset is shown in the pair-wise plots in Figures 4.3, 4.4, and 4.5. The time lagged water table predictors have great positive effects on the estimated levels [198]. Thus, the smoothed Mukarange data was then converted into four days time lags ($t-1$, $t-2$, $t-3$, and $t-4$) for better comparison of all the models. More details about these lags is given in the methodology section.

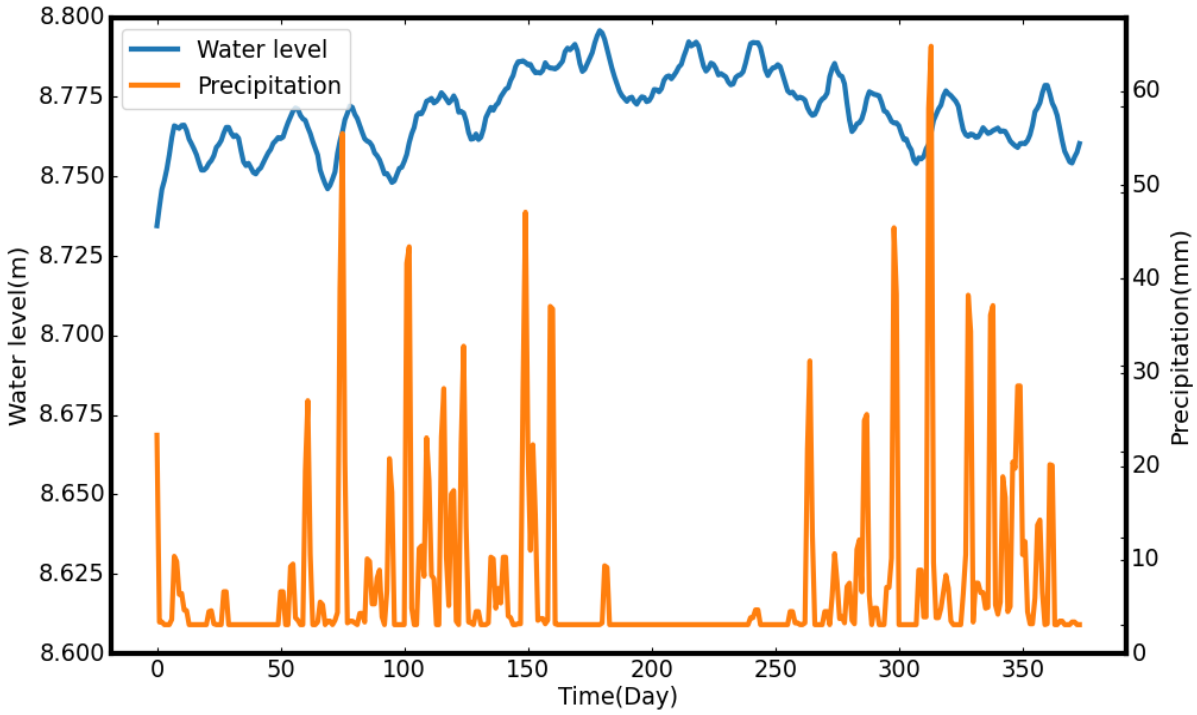



Figure 4.3 Time series plot of precipitation and groundwater level collected from the Mukarange monitoring borehole.

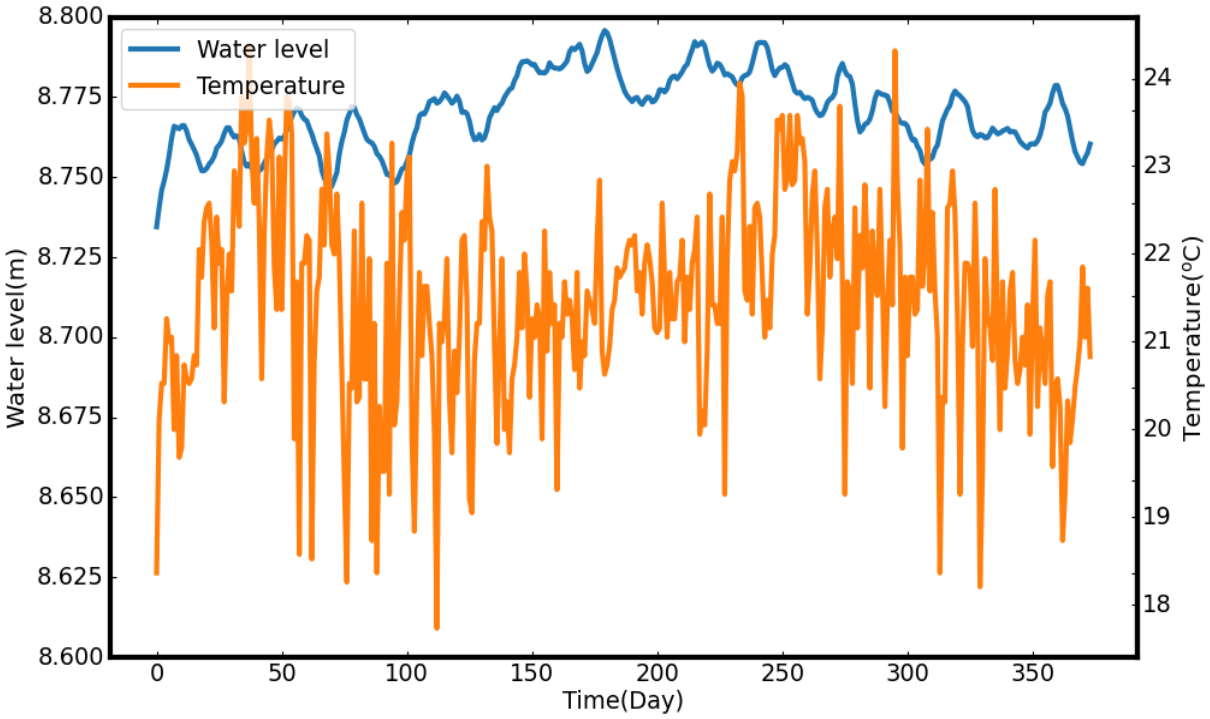



Figure 4.4 Time series plot of temperature and groundwater level collected from the Mukarange monitoring borehole

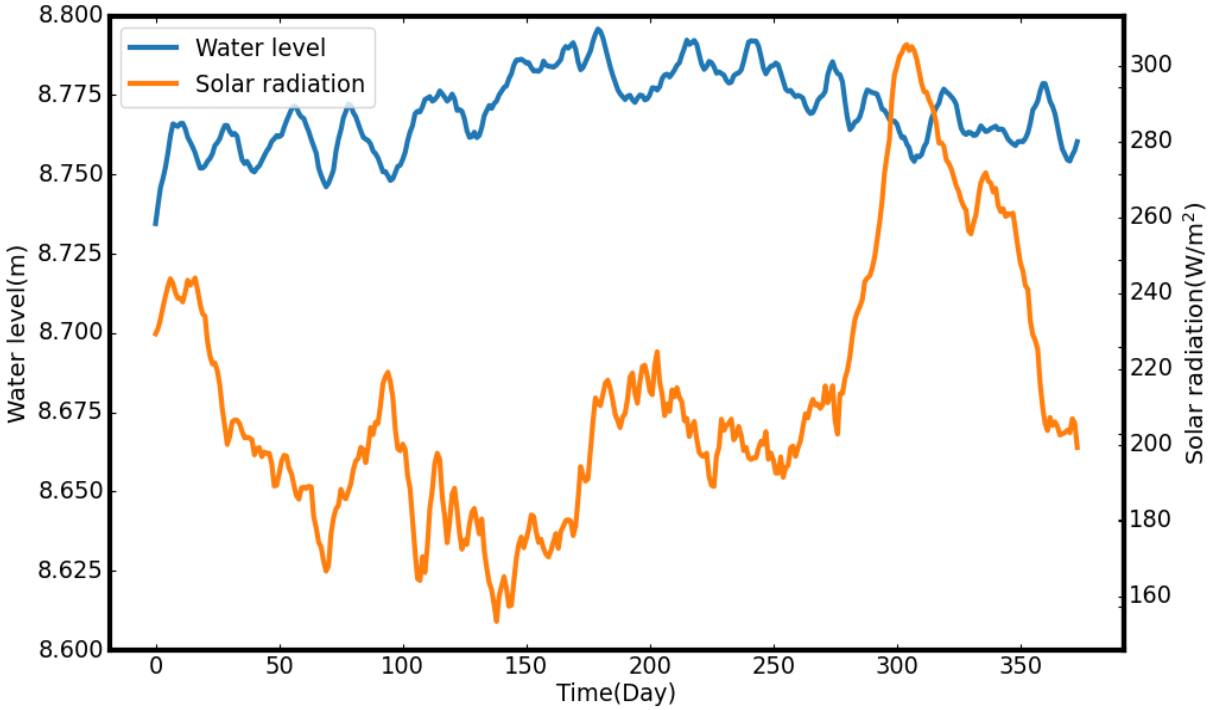



Figure 4.5 Time series plot of solar radiation and groundwater level collected from the Mukarange monitoring borehole

From Figure 4.3, Figure 4.4 and Figure 4.5, it is quite clear that in dry periods (June–August and January–February) there are noticeable declines in groundwater level (GWL), during the study interval (December 2016 to December 2018). This could be connected to the higher evaporation rate, reduced replenishment, and increased groundwater withdrawals due to excessive temperatures in the studied area. Conversely, higher GWLs are observed during wet periods (March–May and September–December), which can be attributed to increased groundwater restoration and reduced abstractions.

### 4.3.3 Model Performance and Evaluation Measures

It is important that the prediction model is properly evaluated to assess its performance [199], [200]. We estimated the predictive ability of the KNN-RF ensemble model on groundwater levels and evaluated it against SVM, KNN, RF, and ANN models based on mean absolute error (MAE), root mean square error (RMSE), the Nash–Sutcliffe efficiency coefficient (NSE), and the coefficient of determination ($R^2$). MAE, RMSE, and $R^2$ were selected because they limit the bias of models against acute events [200]. In addition, MAE and RMSE provide a finer comparison between models, especially in data-scarce situations [200]. NSE is another efficient coefficient used to gauge the relative magnitude of residual variance against the variance of observational data [201][200][202]. MAE, RMSE, and $R^2$ range between 0 and 1, while NSE is between $-\alpha$ and 1. The highest agreement between the estimated and observed values is reached when MAE = 0, RMSE = 0, NSE = 1, and $R^2$ =1. All measurements of the performance of the models were conducted using the hydrostats library (a Python package designed distinctively for hydrology studies) [203]. MAE, RMSE, NSE, and $R^2$ are defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |Q_i^E - Q_i^A|, \tag{4.4}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} (Q_i^E - Q_i^A)^2}, \tag{4.5}$$

$$R^2 = \frac{\sum_{i=1}^{n} [(Q_i^E - \overline{Q}_i^E)(Q_i^A - \overline{Q}_i^A)]^2}{\sum_{i=1}^{n} (Q_i^E - \overline{Q}_i^E)^2 \sum_{i=1}^{n} (Q_i^A - \overline{Q}_i^A)^2}, \tag{4.6}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}[(Q_i^E - Q_i^A)]^2}{\sum_{i=1}^{n}(Q_i^A - \overline{Q}_i^A)^2}, \quad (4.7)$$

where $Q^E$ is the estimated change in groundwater level, $Q^A$ is the actual or observed groundwater level, and *n* is the total number of input data points.

## 4.4 Realization and Evaluation of the Ensemble KNN-RF Method

The principal ambition of this study was to propose a decisive model to characterize the seasonal response of the fractured aquifer in eastern Rwanda, through quantification of seasonal deviations in water table depths in data-scarce situations. There is a requirement [204][205][12] for accessible and simple tools that offer actionable insights for the adaptive management of groundwater resources on a seasonal basis. With that requirement in mind, we propose to estimate seasonal groundwater levels using an innovative ensemble KNN-RF model with an exponentially weighted average preprocess of three predictors (solar radiation, precipitation, and temperature). The workflow of predictive modeling and validation setup is illustrated in Figure 4.6.

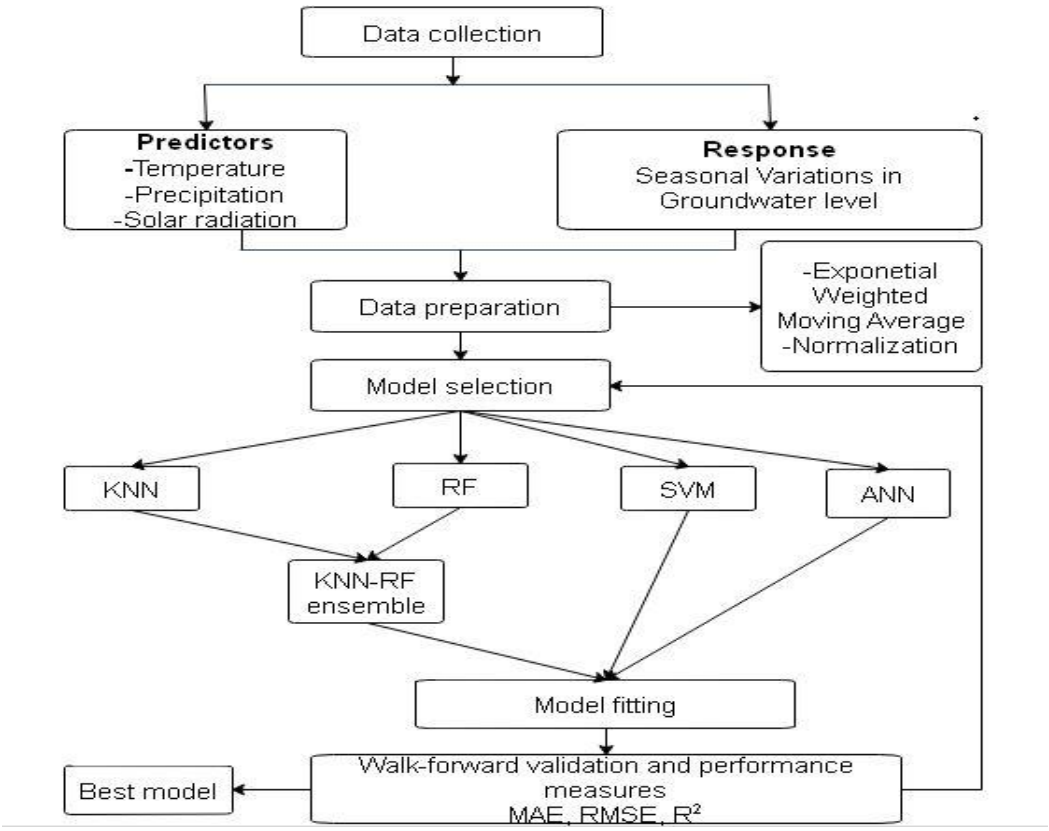



Figure 4.6 Flow chart of the KNN-RF method

Potential features are selected from the collected data. Solar radiation, precipitation, temperature, and GWL data are refined and scaled for proper format and, subsequently, the candidate machine learning models are chosen. The two models (RF and KNN), both of which are non-complex and capable of working with both small and big datasets [206][207][126], are combined to overcome the disadvantages of small datasets as well as enhance predictive accuracy. In the final step, the models are tested and compared against NSE, MAE, RMSE, and $R^2$ using both estimated and observed groundwater data. A rolling window testing and validation method is employed and the most effective and useful model is determined based on the performance contrasts. This is most suited for the temporal nature of time series data[208], since the size of training and validation sets are sampled with respect to the desired forecast length (corresponding to 15, 30, 60, and 90 days) at the end of the series. A model that can be easily adopted using the available information, particularly in resource scarce areas, will be most feasible and practical for sensible decision making on groundwater resources.

### 4.4.1 Development of the KNN-RF Ensemble

Focusing on the improvement of seasonal predictions, a hybrid KNN-RF technique is developed and validated. As indicated in the Introduction section, the RF and KNN methods have good data-representation ability; however, these methods do not perform optimally when fed with tiny datasets. To overcome this limitation, we merge the above models in a hybrid manner. The two base regressors, KNN, and RF are fitted on the whole training set and, using the test set, the models yield predictions individually. The results are then averaged to produce the final result. The final result of the KNN-RF ensemble is given by:

$$\mu(x) = \frac{1}{N}\sum_{n=1}^{N} \omega_n p_n(x), \tag{4.20}$$

where $\mu(x)$ is the final weighted average result of the ensemble model, $\omega_n$ is the weight allocated to the $n^{th}$ regressor, which is based on the *MSE* performance; $p_n$ is the prediction from $n^{th}$ model; $x$ is the sample data points. The ensemble based on the KNN-RF method enhances predictive performance in the following aspects [209].

1. Supports using fewer samples to adequately represent data distribution.
2. Limits the generalization error.
3. Controls variance in a small dataset.
4. Relieves the processing burden for model selection.

Uniform weights are assigned to all estimators in the KNN-RF model. The base RF model is set to perform bootstrapping on the training subset, which reduces similarities in the trees. This, therefore, benefits the performance of the model provided that a small number of training examples are accessible. The KNN base model is set to use the distance between data points as the proximity criterion. Tuning parameters for KNN-RF are leaf_size, metric, random_ state, n_jobs, n_neighbors, $p$, and weights.

### 4.4.2 Tuning Parameter and Input Selection

During the training phase, the appropriate tuning parameters are chosen based on the model performance gains to establish the proper architecture for each model. The following describes how the optimal model's hyper-parameters and input feature combinations were identified and tested in the Scikit-Learn framework. For KNN, the Chebyshev, Minkowski, Euclidean, and Manhattan

length metrics were tested. The Minkowski measure emerged as the best choice [210]. The ideal value of K was found using a grid-search procedure [211], and it varied with the sample size, which was determined by the prediction range (more details about the portions of the sample used for the adjustments and testing of the models are given in the next subsection). For ANN, the trial-and-error technique was used to determine the finest number of hidden layer neurons based on the least RMSE [147], [212], [213]. Fourteen hidden neurons produce the best output. The adaptive learning (Adam) optimization scheme was found to be the most suitable for the dataset used in the current investigation [214]. The ReLU, linear, and tanh activation functions were tested and resulted in ReLU enumerating the most precise results [132][215].

Considering the SVR, the most appropriate values of the epsilon, cost, gamma, and the kernel (rbf, poly, sigmoid) were verified using the trial-and-error technique [147] producing the best factors of 0.01, 1.0, scale, and RBF, respectively [216]. For RF and KNN-RF, the ideal number of estimators is calculated using the grid search procedure [217]. The number of learners affects the processing speed of the model. Whilst a large number of learners improves the reliability of the model, it also slows down processing speed [141]. The ideal number of estimators is 200, max_depth of the trees is 15, leaf_size is 30, max_feature is n _feature, min_sample_leaf is 1, random_state is none, n_jobs is 1, and min_split is 2. Similarly, the n_neighbors is 3 when the prediction period is 15 or 30 days, and 2 when the prediction period is 60 or 90 days, *p* is 2, metric is Minkowski, random _ state is 0, and weights is distance.

For proper and comparable evaluation of the models, the input series of the time lagged precipitation, groundwater level, temperature, and solar radiation were arranged in twelve combinations: P (t − 1) L (t) T (t) S (t), P (t − 2) L (t) T (t) S (t), P (t − 3) L (t) T (t) S (t), P (t − 4) L (t) T (t) S (t), P (t) L (t) T (t) S (t − 1), P (t) L (t) T (t) S (t − 2), P (t) L (t) T (t) S (t − 3), P (t) L (t) T (t) S (t − 4), P (t) L (t) T (t − 1) S (t), P (t) L (t) T (t − 2) S (t), P (t) L (t) T (t − 3) S (t), and P (t) L (t) T (t − 4) S (t). Once the best parameters and input arrangements were established, the KNN − RF, SVM, ANN, RF, and KNN models were trained as described in the next subsection.

### 4.4.3 Training and Testing of the Model

When it comes to the training and evaluation of the models, time series predictive modeling has numerous distinctive traits and peculiarities that need a different approach to supervised learning problems [208]. There are intrinsic interrelationships between the data points measured across time, and during the training and testing of the models, the temporal structure of the series needs to be maintained. Typically, the arbitrary splitting of the time series dataset from different points in time is irrelevant to time-based data because it causes inherent biases [208]. The rolling window or walk-forward validation is best suited for time series-based forecasting as it facilitates updating of the predictions as new data come in. In this approach, the holdout values are sampled at the end of the dataset temporally. Figure 4.7 shows a graphic demonstration of the rolling windows validation procedure.

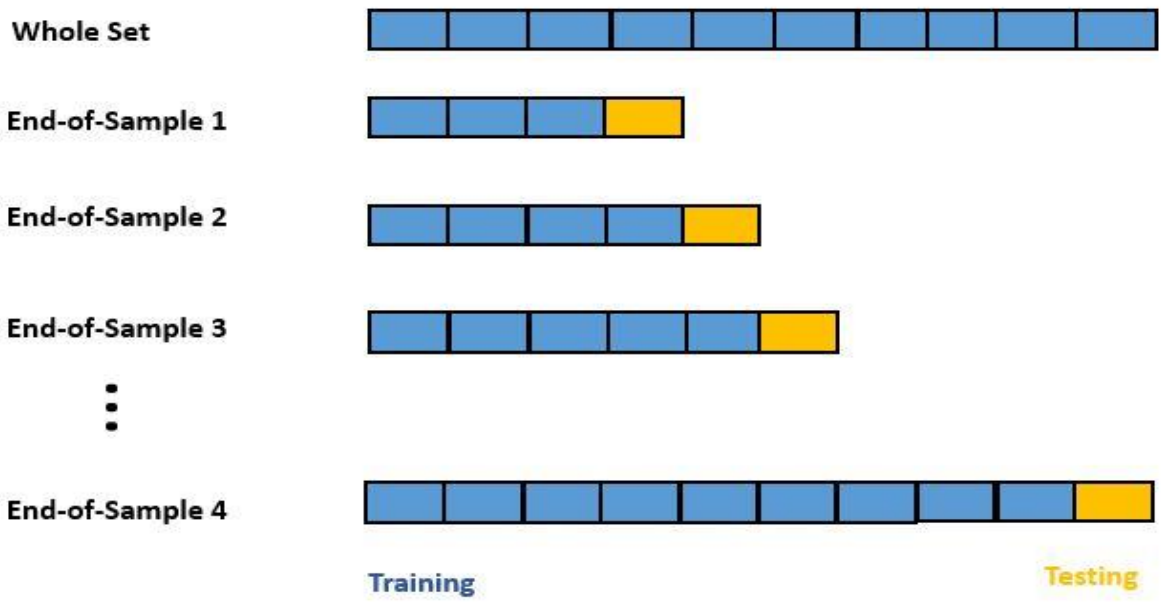



Figure 4.7 Diagram of the time series 4-sliding window validation method. Adapted from [208].

The size of the holdout sample is determined by the prediction scope and, therefore, the width of the rolling window is equal to the desired forecast length. The training and validation of the KNN-RF, ANN, SVM, RF, and KNN models was conducted with the rolling window technique. It was completed using four different portions of training and holdout values corresponding to the prediction of 15-day-ahead (t + 15), 30-day-ahead (t + 30), 60-day-ahead (t + 60), and 90-day-ahead (t + 90). The training and validation percentages for these prediction horizons were

88.1423−11.8577, 92.0949−7.9051, 96.0473−3.9527, and 98.0237−1.9763, respectively. The trained models that were used for prediction are explained in the next subsection.

#### 4.4.4. Prediction of Seasonal Changes in Groundwater Depths

In this work, seasonal forecasting is the forecast of 90 days lead-time groundwater level variations. For comparison, other prediction periods of 15, 30, and 60-day were also evaluated. As previously explained, the 15, 30, 60, and 90 days predictions were implemented by changing the size of the hold-out sample.

## 4.5 Experimental Results and Discussion

The predictive capacity of the KNN-RF technique was investigated and the results were compared to the four general models. The results of the 15, 30, 60, and 90 days lead-time groundwater level predictions at the Mukarange borehole using the KNN-RF, RF, SVM, ANN, and KNN models are presented in Figure 4.8. According to Figure 4.8, at all horizons the KNN-RF model achieved the best performance with respect to NSE,MAE,RMSE and $R^2$ values. For this model, the RMSE values range between 0.0030 and 0.0035, while NSE values were between 0.913 and 0.9741 during the validation stage. The KNN model obtained the best results for the short term (15–30 day-ahead), while RF obtained improved accuracy for long-range (60–90 day-ahead) estimations. The SVR model tried to catch up with the long changes of the levels and outperformed the ANN model. Similar findings were reported in the studies that compared the above methods for the modeling of groundwater tables [218][144][219]. At all lead times, the ANN method overpredicted the observed values. The low performance of the ANN method in training and testing phases on small-sized samples could be attributed to the data requirements of this model [220]. Compared with the RF, ANN, and SVM models, the KNN model had higher performance scores. This is in contrast with the outcomes reported by Rahmati et al. [221]. Meanwhile, RF is found to be superior to the SVM model, which is consistent with the conclusion made by Naghibi et al. [222].

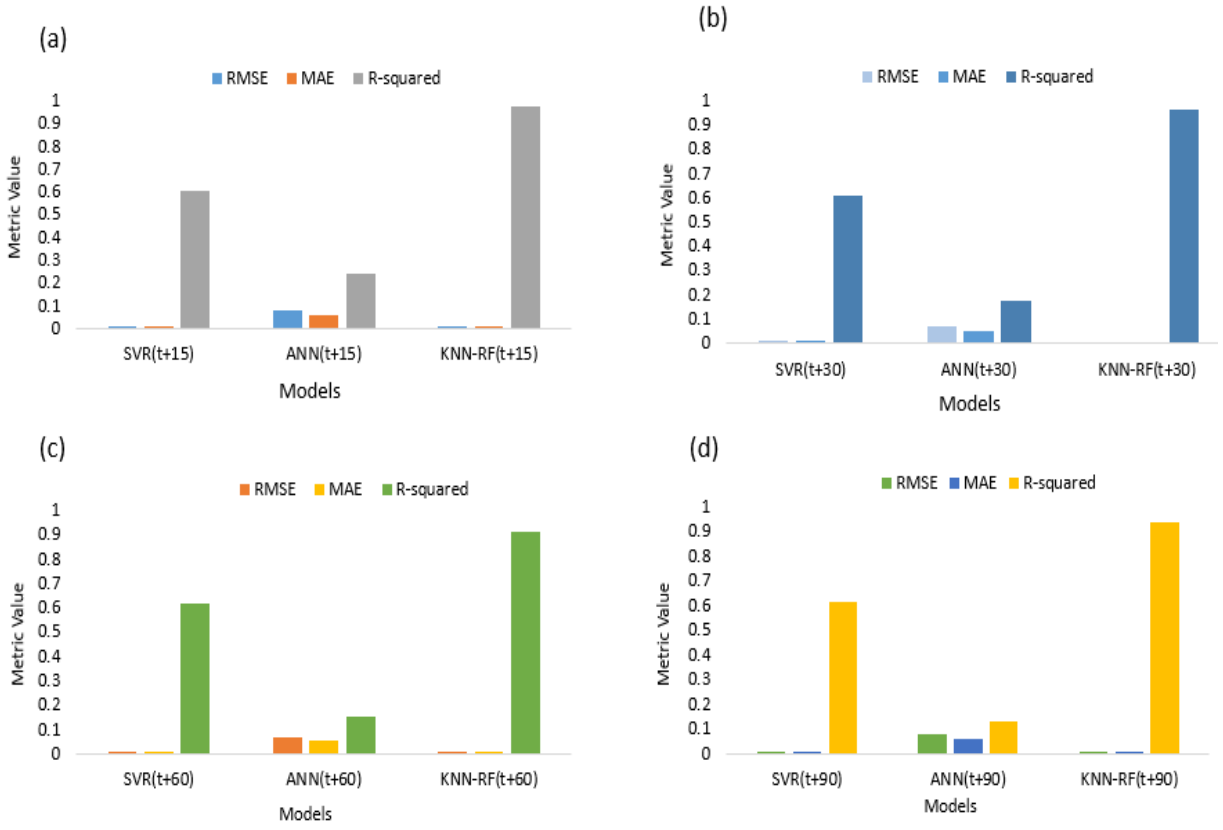



Figure 4.8 Comparison of the performance obtained by the SVR, ANN, RF, KNN, and KNN-RF models on groundwater level prediction for: (a) 15 day lead-time, (b) 30 day lead-time (c) 60 day lead-time, and (d) 90 day lead-time.

From Table 4.2 and Table 4.3, the highest accuracy of the KNN-RF for different horizons was achieved with precipitation (P−1), solar radiation S (t), temperature T (t), and groundwater level L (t) time-lags in the testing phase. The most accurate outcomes are shown for the prediction at 15-days ahead. It is also seen that the accuracy of the forecasted results declines with the length of the prediction. These results are in corroboration with the studies in [112], [149]. Conversely, the 90-day prediction obtained better results than 60-day prediction. The largest difference between the MAE and RMSE values is perceived for the 60-day predictions. The $R^2$ criterion values showed higher importance for longer lead-times, while NSE provided the overall description of the predictive power of the models.

Table 4.2 Performance evaluation results for 15, 30, 60, and 90 days lead-time groundwater level variations using the KNN-RF model.

| | L(t+15) | | | L(t+30) | | | L(t+60) | | | L(t+90) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Input arrangement | **RMSE** | **MAE** | $R^2$ | **RMSE** | **MAE** | $R^2$ | **RMSE** | **MAE** | $R^2$ | **RMSE** | **MAE** | $R^2$ |
| **P(t-1)S(t)T(t)** | **0.0022** | **0.0019** | **0.9791** | **0.0026** | **0.0020** | **0.9619** | **0.0036** | **0.0025** | **0.9185** | **0.0031** | **0.0022** | **0.9387** |
| P(t-2)S(t)T(t) | 0.0061 | 0.0050 | 0.8397 | 0.0069 | 0.0050 | 0.7114 | 0.0068 | 0.0051 | 0.7049 | 0.0065 | 0.0048 | 0.7143 |
| P(t-3)S(t)T(t) | 0.0054 | 0.0043 | 0.8909 | 0.0064 | 0.0047 | 0.7527 | 0.0059 | 0.0044 | 0.7948 | 0.0059 | 0.0044 | 0.7807 |
| P(t-4)S(t)T(t) | 0.0049 | 0.0042 | 0.9179 | 0.0069 | 0.0051 | 0.7739 | 0.0056 | 0.0042 | 0.7989 | 0.0057 | 0.0043 | 0.7882 |
| P(t)S(t-1)T(t) | 0.0060 | 0.0051 | 0.8401 | 0.0067 | 0.0051 | 0.7308 | 0.0064 | 0.0048 | 0.7454 | 0.0061 | 0.0046 | 0.7660 |
| P(t)S(t-2)T(t) | 0.0060 | 0.0051 | 0.8749 | 0.0061 | 0.0048 | 0.7872 | 0.0059 | 0.0047 | 0.7993 | 0.0061 | 0.0046 | 0.7706 |
| P(t)S(t-3)T(t) | 0.0058 | 0.0048 | 0.8639 | 0.0065 | 0.0048 | 0.7438 | 0.0059 | 0.0044 | 0.7948 | 0.0059 | 0.0044 | 0.7783 |
| P(t)S(t-4)T(t) | 0.0058 | 0.0050 | 0.8635 | 0.0062 | 0.0049 | 0.7739 | 0.0061 | 0.0049 | 0.7537 | 0.0057 | 0.0044 | 0.7840 |
| P(t)S(t)T(t-1) | 0.0061 | 0.0050 | 0.8391 | 0.0065 | 0.0049 | 0.7577 | 0.0065 | 0.0049 | 0.7408 | 0.0065 | 0.0048 | 0.7277 |
| P(t)S(t)T(t-2) | 0.0061 | 0.0050 | 0.8411 | 0.0065 | 0.0049 | 0.7601 | 0.0065 | 0.0049 | 0.7518 | 0.0062 | 0.0046 | 0.7553 |
| P(t)S(t)T(t-3) | 0.0058 | 0.0048 | 0.8539 | 0.0064 | 0.0048 | 0.7545 | 0.0062 | 0.0048 | 0.7646 | 0.0059 | 0.0046 | 0.7711 |
| P(t)S(t)T(t-4) | 0.0057 | 0.0047 | 0.8663 | 0.0066 | 0.0049 | 0.7312 | 0.0062 | 0.0048 | 0.7578 | 0.0059 | 0.0045 | 0.7679 |

T represents temperature, L represents groundwater level, S represents solar radiation, and P represents precipitation. The highest $R^2$ and lowest MAE and RMSE are in bold.

Table 4.3 NSE performance evaluation results for 15, 30, 60, and 90 day-ahead groundwater level predictions using the KNN-RF model.

| | L(t+15) | L(t+30) | L(t+60) | L(t+90) |
|---|---|---|---|---|
| Input arrangement | | | | |
| P(t-1)S(t)T(t) | 0.9741 | 0.9540 | 0.9130 | 0.9346 |
| P(t-2)S(t)T(t) | 0.7957 | 0.6792 | 0.6657 | 0.6898 |
| P(t-3)S(t)T(t) | 0.8385 | 0.7239 | 0.7532 | 0.7483 |
| P(t-4)S(t)T(t) | 0.8697 | 0.6798 | 0.7720 | 0.7595 |
| P(t)S(t-1)T(t) | 0.7987 | 0.6977 | 0.7083 | 0.7267 |
| P(t)S(t-2)T(t) | 0.8014 | 0.7458 | 0.7488 | 0.7290 |
| P(t)S(t-3)T(t) | 0.8105 | 0.7136 | 0.7532 | 0.7431 |
| P(t)S(t-4)T(t) | 0.8160 | 0.7373 | 0.7339 | 0.7659 |
| P(t)S(t)T(t-1) | 0.7934 | 0.7168 | 0.6968 | 0.6951 |
| P(t)S(t)T(t-2) | 0.7933 | 0.7173 | 0.7023 | 0.7212 |
| P(t)S(t)T(t-3) | 0.8133 | 0.7256 | 0.7274 | 0.7456 |
| P(t)S(t)T(t-4) | 0.8214 | 0.7055 | 0.7260 | 0.7430 |

T represents temperature, L represents groundwater level, S solar radiation, and P represents precipitation. The highest $NSE$ are in bold.

Figure 4.9 delineates the results of the comparison of the relationships between the actual and KNN-RF estimated groundwater levels for different horizons. These results show that there is high association between the actual and estimated levels for all four-time horizons. The 15-day range exhibits the largest value of $R^2$, since most of the predictions are closer to the straight line. It was also found that the 60-day prediction range showed the relatively lowest value of $R^2$ compared to

other ranges. This also supports the results presented in Figure 4.8, which showed higher error values for the 60-day prediction stage than those of the other predictions. Similarly, the hydrographs in Figure 4.10 show that the KNN-RF technique reproduced and fairly represented the seasonal oscillations of the depths of the groundwater tables. However, it is quite obvious that the 60-day prediction is not as accurate as the predictions for other time-horizons.

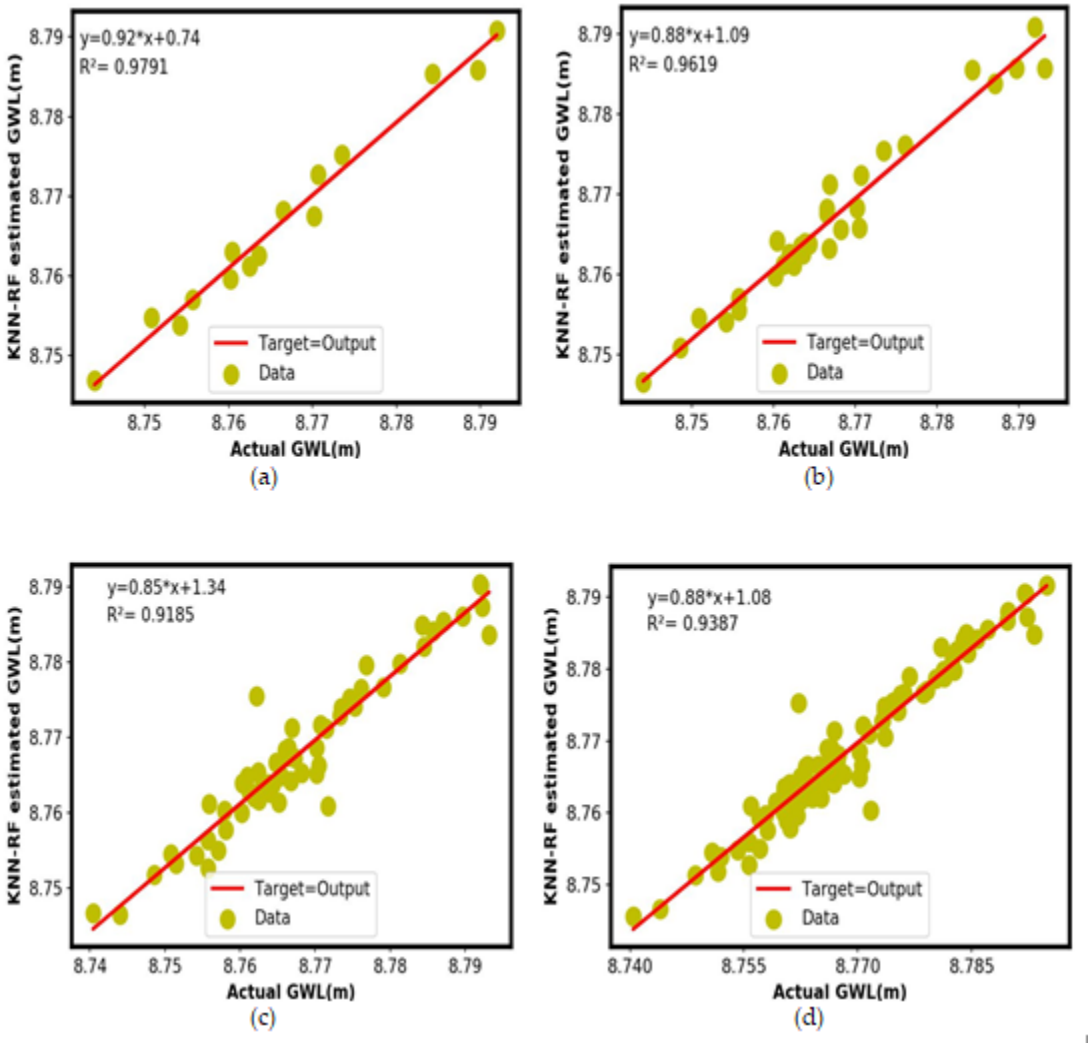



Figure 4.9 Assessment of the actual and the estimated groundwater levels of the optimal KNN-RF model for 15, 30, 60, and 90 days (corresponding to a–d, respectively) lead time in the testing phase.

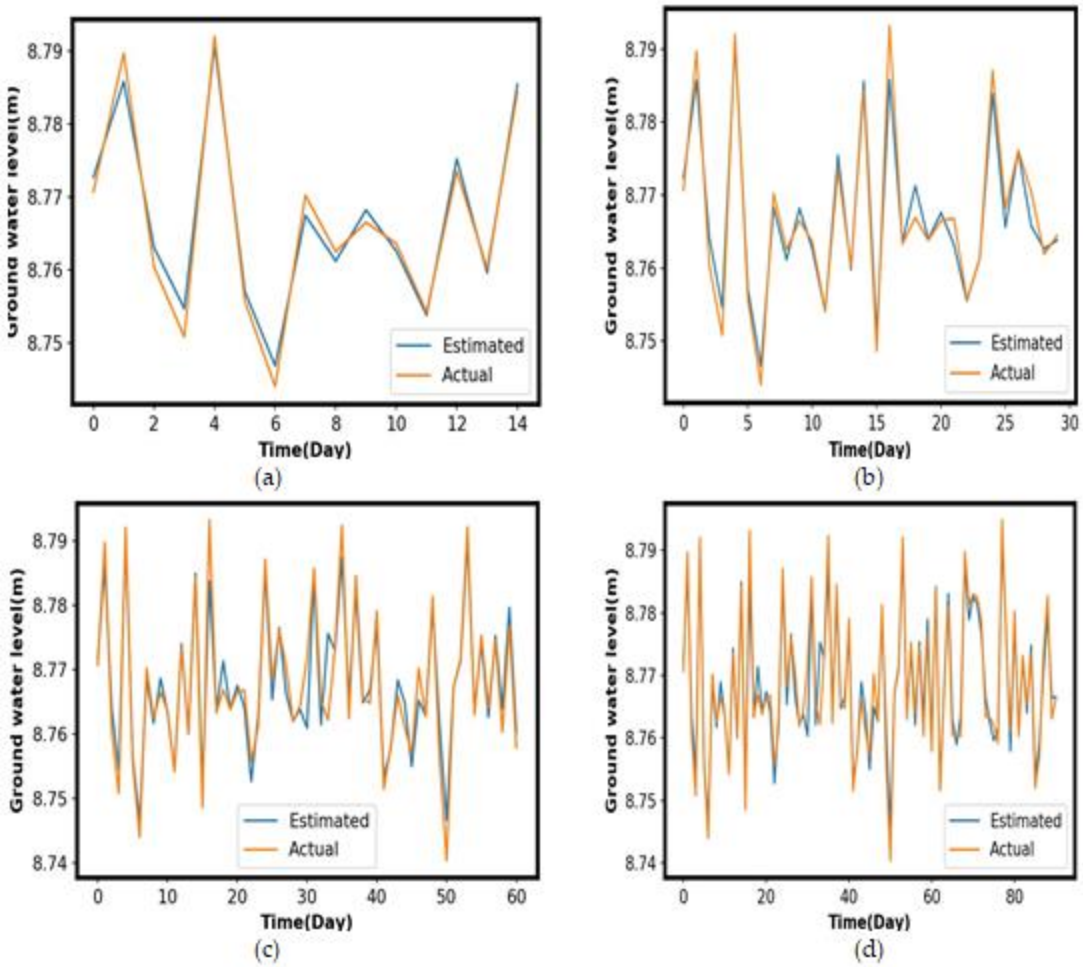



Figure 4.10 Comparison of the 15, 30, 60, and 90 days (corresponding to a–d, respectively) estimated and the actual levels yielded by the optimal KNN-RF technique

.

The results obtained have also demonstrated the significant role of input and tuning parameter selection. The combinations of solar radiation, temperature, precipitation, and previous groundwater level with an appropriate time-lag improved the seasonal estimation of the groundwater heights. It was found that for the KNN-RF technique, the lagged precipitation improved NSE, MAE, RMSE, and $R^2$ scores, for 32.6%, 53.19%, 51.57%, and 27.38%, respectively. This suggests that there is a huge potential for the infiltrated and percolated rain water to raise the groundwater table for the Mukarange aquifer. The hydrographs in Figure 4.10 show high agreement between observed and predicted water levels which implies that the KNN-RF has higher level of multi-step prediction accuracy.

For feature selection, Lindsey et al. [194] and Kelly et al. [195] showed that the use of solar radiation as a substitute for evapotranspiration as a suitable option for capturing the dynamics of groundwater depth. Our results confirmed the high influence of solar radiation on the long-term variability of groundwater levels in the semi-arid area.

Results also confirmed that tuning parameters have a great influence on the model's final results. These parameters led to an improved generalization capability for all models. Considering the SVR, it was found that RBF yielded the best performance, while epsilon and gamma are the most influential parameters for the determination of the appropriate architecture of the SVR model. This is in corroboration with the findings in [136]. With six (6) input features, the finest number of nodes in the hidden layer of the ANN was found to be 14, which is consistent with the conclusion reached by Kayzoglu et al. [223]. Whilst the adaptive learning scheme and the ReLU function are commonly used with large datasets, our results suggest that these parameters can also work well with a limited dataset. One of the possible reasons for this outcome is the sparsity of the available samples. We found that limiting the number of learners to 200 and the depth of the trees to 15 had positive effects on the generalization ability of the RF model and overcame the overfitting issue on the Mukarange dataset. The best results for the KNN-RF, SVM, and ANN methods were achieved with the structures presented in Table 4.4.

Table 4.4 Summary of the selected parameters during training of the SVR, ANN, and KNN-RF models.

| **SVR** | **ANN** | **KNN-RF** |
|---|---|---|
| Epsilon: 0.0100 | Epsilon: 1e-04 | Epsilon: 0.0100 |
| Kernel: RBF | Hidden layer: 1 | Number of trees: 200 |
| Gamma: Scale | Hidden layer neurons: 14 | Maximum depth: 15 |
| Soft-margin(C): 1.0000 | Activation function: ReLU | Weights: distance |
| Support vectors: 137 | Learning mode: Adaptive | Metric: Minkowski |
| Degree: 3 | Max-iterations: 500 | Number of estimators: 50 |
| | Training algorithm: Adam | Algorithm: Auto |

## 4.6 Conclusions

This chapter has described the development and validation of performance and capacity of an ensemble KNN-RF regression approach in predicting seasonal groundwater levels for the fractured aquifer with limited data. Groundwater level data and its significant meteorological drivers (solar radiation, temperature, and precipitation) collected from Mukarange in eastern Rwanda were used

for the analysis. From the experimental analysis, it was found that the KNN-RF ensemble approach is stable with enhanced generalization competence and prediction accuracy. The results also indicated that, by using the sliding window validation procedure, the KNN-RF model captured slightly well with the time-based changes in the depths of the groundwater tables. Inclusion of the solar radiation as a substitute for evapotranspiration resulted in an improved prediction accuracy. The results of the study suggest that KNN-RF is well-suited for the forecasting of seasonal variations in groundwater depths with limited samples. The values of the analytical measures showed that, in all prediction ranges, the KNN-RF technique achieved the most promising results compared to those obtained by the ANN, KNN, SVM, and RF models. The NSE and $R^2$ values of the ensemble KNN-RF technique were higher than those yielded by the above methods, and the values of RMSE and MAE of the ensemble KNN-RF were smaller than those produced by the other methods. The research used data from one groundwater observation station over a short duration, and it has been concluded that more data could improve the predictive accuracy of the model. This can be achieved simply and effectively by updating the model as data become available, since a sliding window method has been used. In addition, the KNN-RF model was shown to be an advanced alternative to the SVR, KNN, RF, and ANN models. The results from this study would be useful for the planning and management of groundwater resources. Our proposed model could be readily transferable or adapted to other areas, specifically those with similar aquifers where the availability and quantity of data is challenging.

# CHAPTER 5

# AN ENSEMBLE MODE DECOMPOSITION COMBINED WITH SVR-RF MODEL FOR PREDICTION OF GROUNDWATER LEVEL: THE CASE OF EASTERN RWANDAN AQUIFERS

## 5.1 Introduction

This chapter presents a machine learning predictive technique that combines SVR and RF methods with the use of an EEMD preprocessing approach to provide an improved prediction model for achieving objectives five and six. The chapter also compares the seasonal prediction performance with that of the conventional methods.

Modelling and prediction of groundwater table is one of the strategic practices for more effective utilization, planning and adaptation of this scarce resource. Machine learning (ML) methods are gaining prominence with advance in modelling of water resources. However, there is limited practical usage of ML owing to some limitations such as a deficiency of hydrologic data and the complexity of groundwater system. This chapter's predictive model used groundwater tables and associated weather data from Eastern Rwanda to predict variations in water table depths using a hybrid SVR-RF with ensemble empirical mode decomposition (EEMD) preprocessing (EEMD-SVR-RF). Historical time-series air humidity, temperature, solar radiation, precipitation, and groundwater levels from Rugarama and Mukarange observatory stations were utilized to calibrate and evaluate the methods. These datasets show high relevance in predicting water table depths up to three-month lead time using the proposed methodology. For both sites and all prediction time steps, the EEMD-SVM-RF method demonstrated relatively higher accuracy and performance. At 90-day lead-time, the enhanced model obtained 0.9608 of the coefficient of determination ($R^2$) compared to 0.911, 0.9367, and 0.4818 obtained by the EEMD-SVM, EEMD-RF, and EEMD-ANN, respectively. Likewise, the proposed method obtained slightly higher Nasch-Schimitt efficiency (NSE) of 0.9586 compared to 0.8903, 0.4598, and 0.9092 for the EEMD-SVM, EEMD-ANN, and EEMD-RF at nighty-day prediction horizon, respectively. These results affirm the applicability of the EEMD-SVR-RF method to groundwater forecasting even with limited hydrologic data.

## 5.2 Contributions

The contribution of the work presented in this chapter is two-fold. First, it presents the only attempt implementing a hybrid SVR-RF with EEMD preprocess for groundwater prediction. Second, it shares the practical application of the ML-based to seasonal prediction in temperate sub-Saharan African aquifers.

## 5.3 Materials, Tools and Methodology

### 5.3.1 Case Study and Available Data

The study area, Eastern province in Rwanda, lies between 29.86875E – 29.90625E and 2.30625S - 2.26875S and covers 9,813 $km^2$. The weather in the study area between the months of July to September, is sunny and dry, with average maximum temperatures ranging between 24.3$^oC$ and 30.3$^oC$, whilst the between May and June the minimum average temperature ranges between 13$^oC$ and 16.65$^oC$. The period between March and May is the wet season with rainfall ranging between 450 mm and 500 mm.

Two observational wells were considered for this study. One of the wells (Mukarange) is in a low permeable aquifer in Kayonza, while the other (Rugarama) is found in a permeable aquifer in Gatsibo (Figure 5.1). Rwanda Water and Forestry Authority (RWFA) is in charge of these boreholes. RWFA provided the water table depths time series from these boreholes for the period from December 3, 2016 to October 30, 2020 (see Table 5.1). The RWFA groundwater measurements were taken at 12-hour intervals. Therefore, we transformed these measurements to daily average values in order to calculate mean levels.

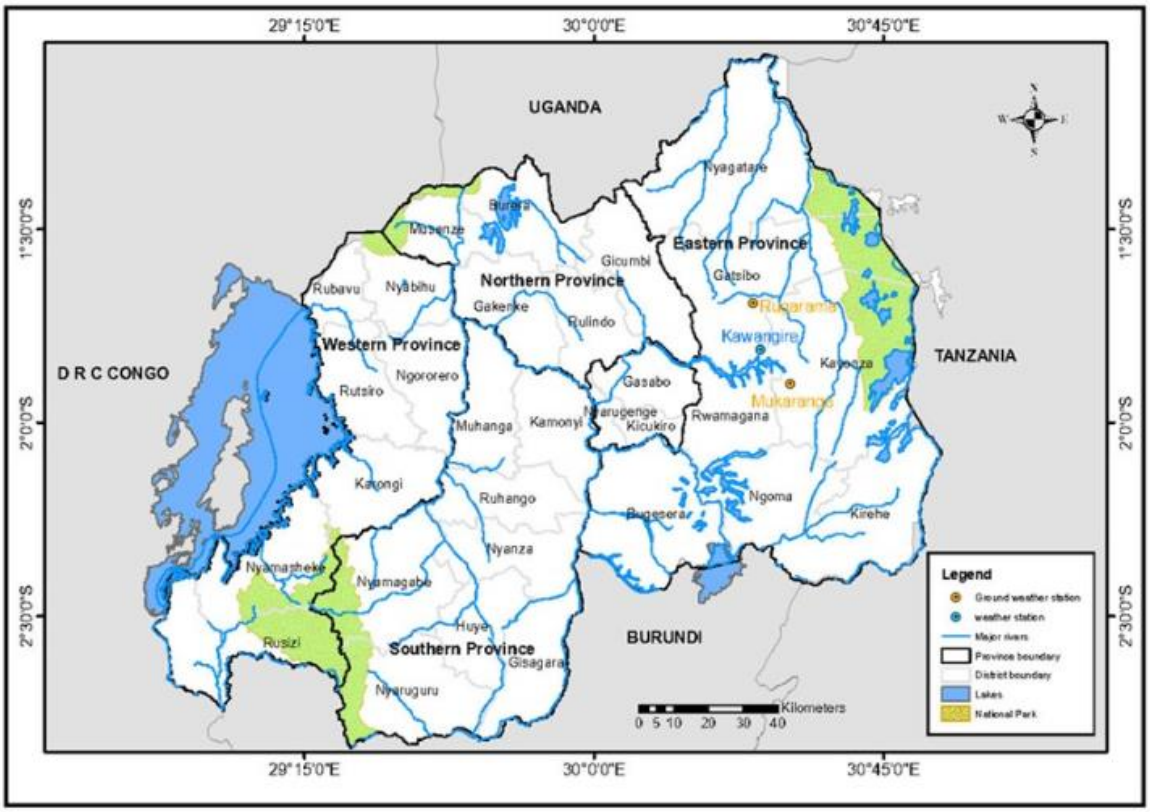



Figure 5.1 Study area and location of groundwater stations (Rugarama and Mukarange) and Kawangire weather station in Eastern Province- Rwanda.

Table 5.1 Characteristics of Rugarama and Mukarange aquifers.

| Identifier | Station name | Latitude | Longitude | Data resolution | Availability of data | Aquifer |
|---|---|---|---|---|---|---|
| F6 | Mukarange | 1.89874154 | 30.5065299 | Daily | Dec 3,2016-Oct 30, 2020 | Permeable fracture |
| F7 | Rugarama | 1.69129110 | 30.4121294 | Daily | Dec 3,2016-Oct 30, 2020 | Low permeable |

Weather data for the study period of 2016 to 2020 was acquired from the Kawangire weather station (managed by Rwanda Meteorological Agency - MeteoRwanda).This station is common to both Mukarange and Rugarama groundwater observatory stations. The data are the minimum and maximum air temperature ($^{o}$C), average daily air humidity (in percentage g/m$^3$), daily precipitation (mm) and daily solar radiation (W/m$^2$). Temperature data were converted to mean values to provide all data in the same time-interval. Figures 5.2 and 5.3 represent the hydro-climatic data gathered from the above mentioned organizations.

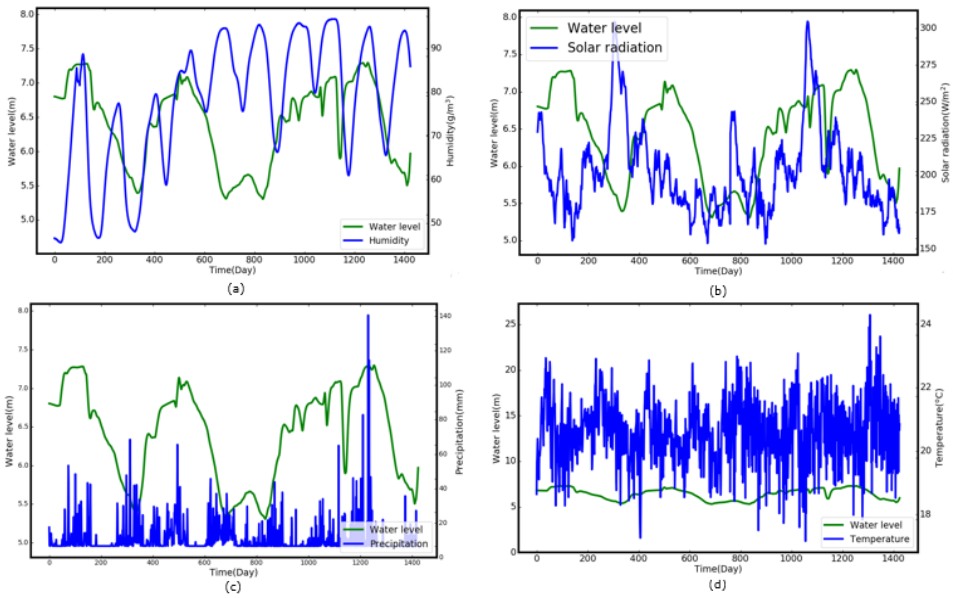



Figure 5.2 The relationships between Rugarama groundwater level and the weather parameters (a) air humidity (b) solar radiation (c) precipitation (d) air temperature

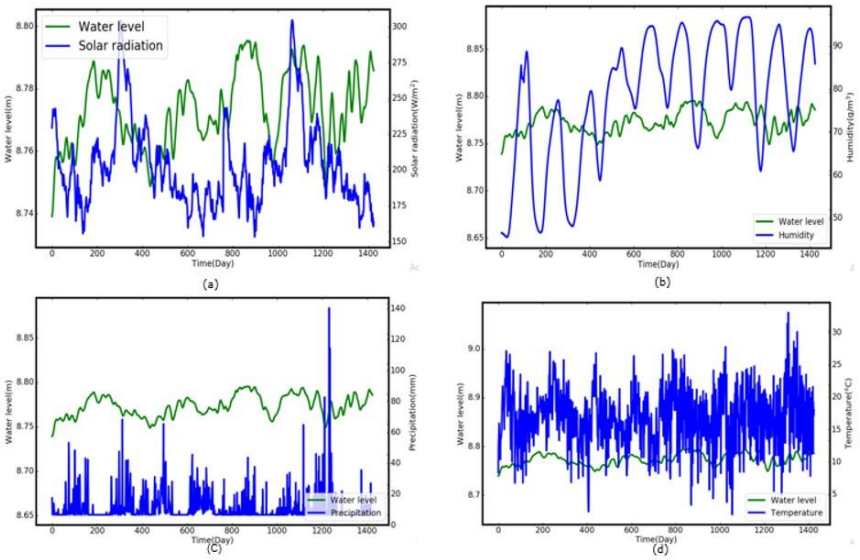



Figure 5.3 The relationships between Mukarange groundwater level and the weather parameters (a) air humidity (b) solar radiation (c) precipitation (d) air temperature

## 5.3.2 Initial Data Preparation and Tools

All experiments were carried out in the Python 3.6 environment, and the Numpy, Scikit-learn, Hydrostats, Matplotlib, CSV, and Pandas libraries were used. To reveal the underlying features for

the prediction model, the raw data are cleaned and scaled. The features are scaled between -1 and 1, using Equation 5.1. After that, the temperature and groundwater measurements were averaged on daily basis, using Equation 5.2.

$$X_c = \frac{X_r - X_{min}}{2X_{max}} \tag{5.1}$$

$$X_m = \frac{\sum_{i=1}^{K} x_i}{K} \tag{5.2}$$

where, $X_m$ is the calculated mean value of the data points, $x_i$ is the individual data points, and $k$ is the total number of the data points in the observation.

### 5.3.3 Methods

The development

#### 5.3.3.1 EEMD –SVR-RF Approach

The preprocessing of the normalized dataset is performed using EEMD, which extracts suitable information from the samples. Thereafter, the transformed and well formatted data with higher signal to noise ratio is then fed to the SVR-RF hybrid model as well as the SVR, ANN, and RF models for calibration and testing of these models, as depicted in Figure 5.4. The SVR-RF model uses the data to predict possible changes in water table depth using Equation 5.3. The outcomes from these models are compared to determine the best performing method.

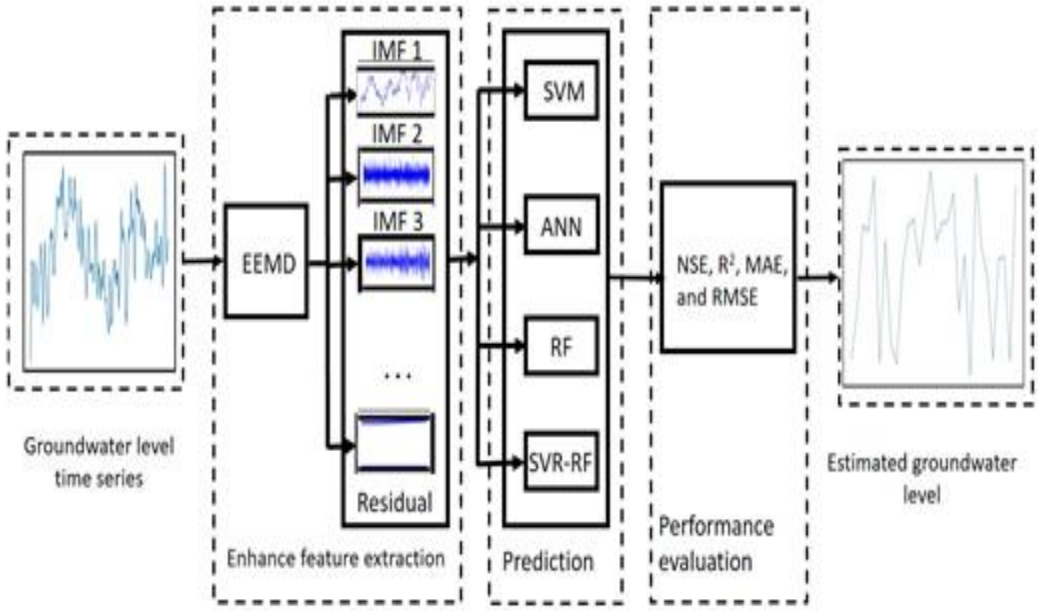



Figure 5.4 Schematic layout of EEMD –SVR-RF method.

$$\mu(x) = \frac{1}{K}\sum_{k=1}^{K} \omega_k p_k(x) \tag{5.3}$$

where, $\mu(x)$ is the final weighted average result of the SVR-RF model, $p_k$ is the prediction from $k^{th}$ model, $\omega_k$ is the weight allocated to $k^{th}$ regressor, and $\boldsymbol{x}$ is the sample data points.

The IMFs of different frequencies extracted from Rugarama and Mularange datasets are depicted in Figure 5.5.

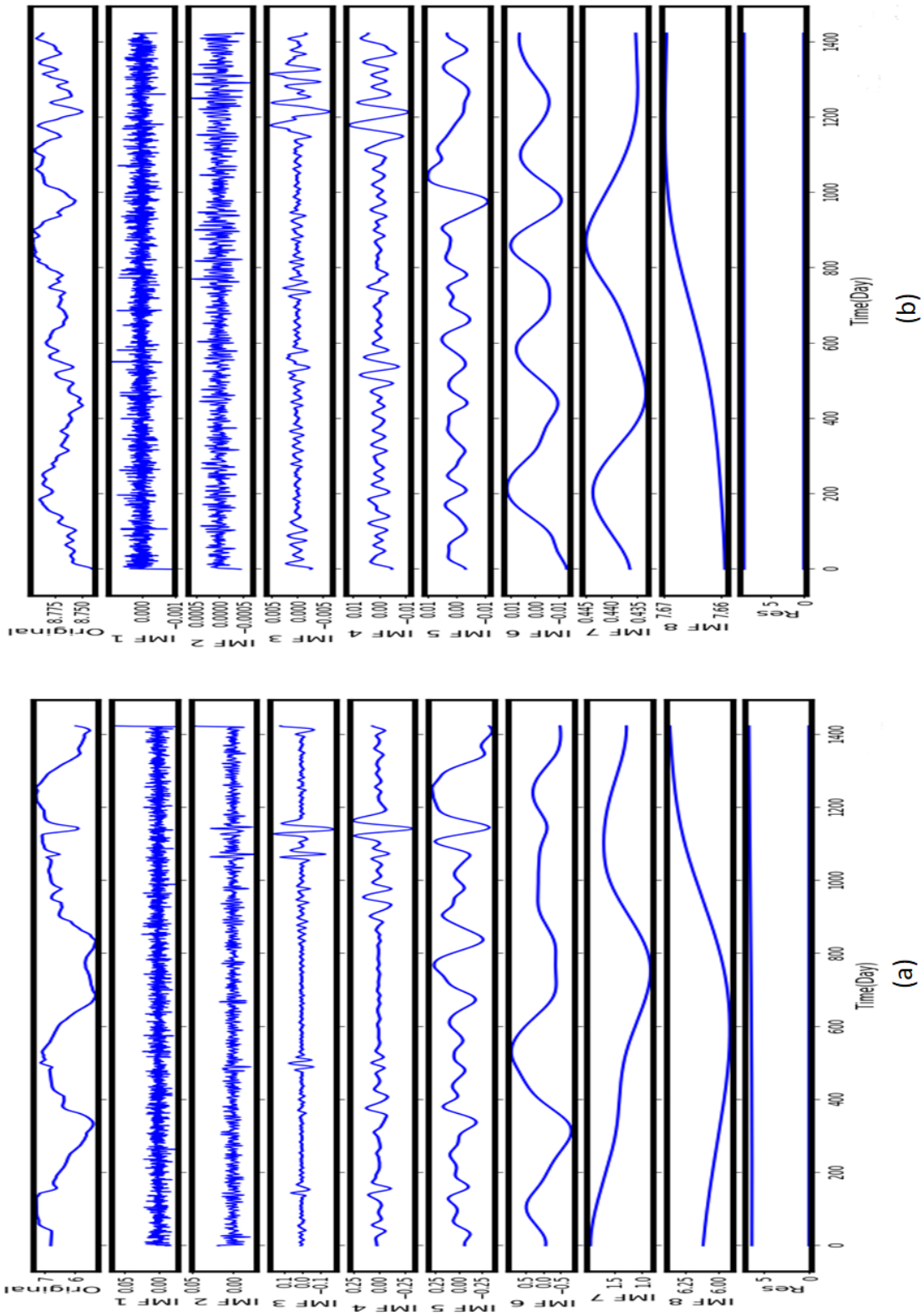



Figure 5.5 Decomposed levels for (a) Rugarama and (b) Mukarange groundwater data.

### 5.3.3.5 Data Split, Training and Evaluation of the Models

The EEMD-SVR-RF models with Rugarama and Mukarange water table depth predictors were fitted on rolling windows mode (15, 30,60, and 90-day prediction horizons) in turn, and evaluated based on Root Mean Squared Error (MSE), Nash-Schmitt efficiency (NSE), Coefficient of determination ($R^2$), and Mean Absolute Error (MAE) which are presented in equations 5.4, 5.5, 5.6, and 5.7. The hydrostats python library was used to calculate these statistical performance measures. The subsets of the training examples and test data are outlined in Table 5.2.

$$NSE = 1 - \frac{\sum_{k=1}^{K} [(Q_k^E - Q_k^M)]^2}{\sum_{k=1}^{K} (Q_k^M - \overline{Q}_k^M)^2} \tag{5.4}$$

$$R^2 = \frac{\sum_{k=1}^{K} [(Q_k^E - \overline{Q}_k^E)(Q_k^M - \overline{Q}_k^M)]^2}{\sum_{k=1}^{K} (Q_k^E - \overline{Q}_k^E)^2 \sum_{k=1}^{K} (Q_k^M - \overline{Q}_k^M)^2} \tag{5.5}$$

$$MAE = \frac{1}{K} \sum_{k=1}^{K} \left| Q_k^E - Q_k^M \right| \tag{5.6}$$

$$RMSE = \sqrt{\sum_{k=1}^{K} (Q_k^E - Q_k^M)^2} \tag{5.7}$$

Where, $\boldsymbol{Q^M}$ is the item of the measured data, $\boldsymbol{Q^E}$is item of the predicted values, K is the total number of the measurements/observations.

Table 5.2 Training and test data portions for Rugarama and Mukarange datasets

| | Training(data points) | Test(data points) |
|---|---|---|
| T+15 | 1410 | 15 |
| T+30 | 1395 | 30 |
| T+60 | 1365 | 60 |
| T+90 | 1335 | 90 |

## 5.4 Results and discussion

In this section, the results of the experiment that created and evaluated a hybrid EEMD-SVR-RF method are provided and discussed. In that experiment the model was also compared against EEMD-SVR, EEMD-ANN, and EEMD-RF approaches in terms of MAE, NSE, $R^2$, and RMSE metrics. According to Figure 5.6, EEMD-SVR-RF approach achieved relatively superior performance, followed by EEMD-RF and EEMD-SVR. With lowest $R^2$ and NSE values, as well as highest MAE and RMSE values, ANN is the least effective method on both datasets. This is in corroboration with the findings reported in [224]. Moreover, for both aquifers, the results in Figures 5.7, 5.8, and 5.9 indicate significant correlation between observed and EEMD-SVR-RF estimated water table depths.

Furthermore, as can be seen in Figures 5.7 and 5.8, the majority of the data points are closer to the regression lines, signifying the EEMD-SVR-RF approach's exceptional skill on both datasets. Nevertheless, performance of this approach on Rugarama dataset is slightly better than that on the Mukarange dataset. This is also confirmed by the results shown in Table 5.3, which show that the model estimation has an RMSE of 0.0011 and $R^2$ value of 0.9608 on Rugarama dataset, compared to an RMSE of 0.0032 and $R^2$ value of 0.9375 on Mukarange dataset, at 90-day lead time D(t+90) prediction. The EEMD-SVR-RF is more efficient at longer lead times than at shorter lead times, as can be observed from Table 5.3, with the best performance at a 60-day lead time.

In general, the experimental test results demonstrate that the proposed hybrid EEMD-SVR-RF produces relatively better estimations of the changes in water table depths compared to conventional methods. For example, using Rugarama dataset, at 90-day lead time, the EEMD-SVR-RF outperformed EEMD-RF, EEMD-SVR and EEMD-ANN by 2.41%, 4.98%, and 47.9%, respectively. Moreover, it reduces the RMSE error by 0.7%, 0.54%, and 0.25% compared to EEMD-SVR, EEMD-ANN, and EEMD-RF, respectively.

Table 5.3 Performance of EEMD-SVR-RF on Mukaange and Rugarama dataset

| | D(t+15) | | | D(t+30) | | | D(t+60) | | | D(t+90) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | RMSE | MAE | $R^2$ | RMSE | MAE | $R^2$ | RMSE | MAE | $R^2$ | RMSE | MAE | $R^2$ |
| Rugarama | 0.0031 | 0.0624 | 0.9038 | 0.0015 | 0.0539 | 0.9423 | 0.0013 | 0.0344 | 0.9694 | 0.0011 | 0.0382 | 0.9608 |
| Mukarange | 0.0052 | 0.0022 | 0.8643 | 0.0037 | 0.0014 | 0.9079 | 0.0028 | 0.0011 | 0.9532 | 0.0032 | 0.0012 | 0.9375 |

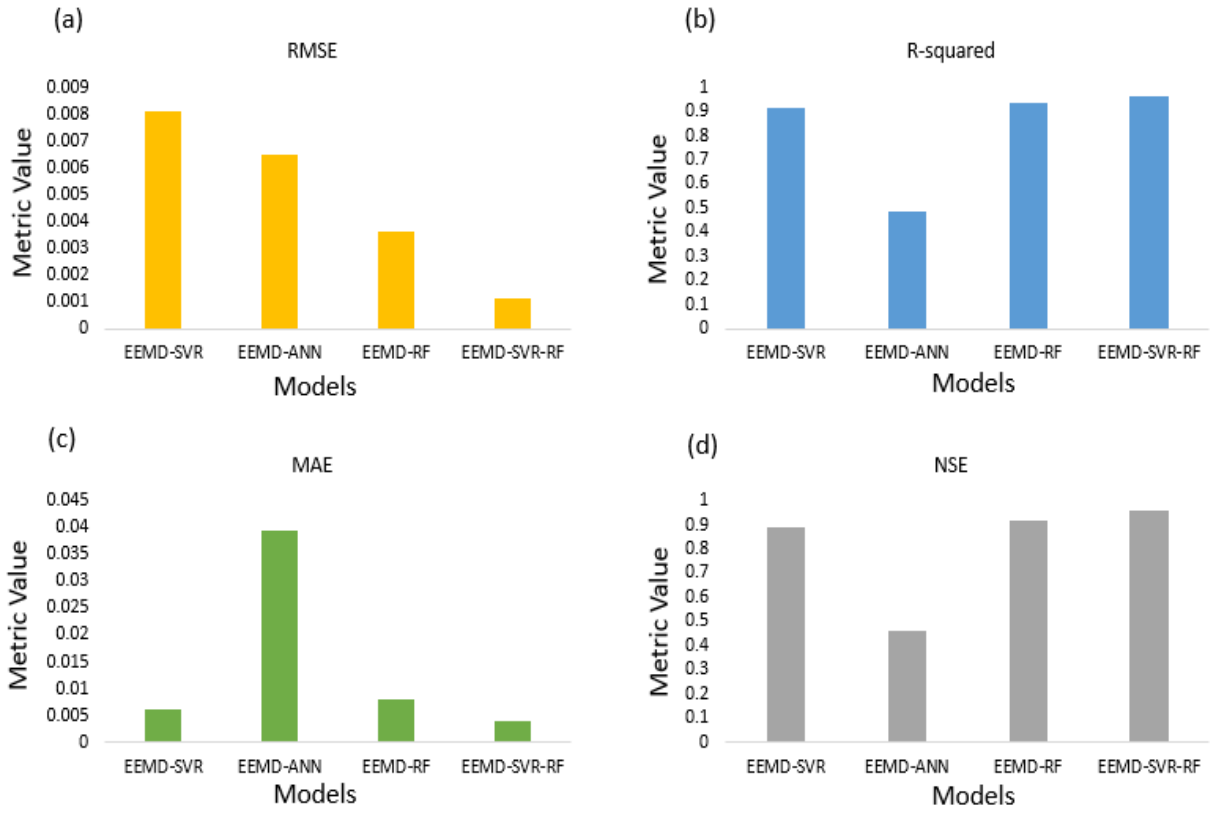



Figure 5.6 Evaluation outcome for the EEMD-SVR, EEMD-ANN, EEMD-SVM and EEMD-SVR-RF models for the nighty-day lead-time prediction of groundwater level at Rugarama observatory station: (a) RMSE metric (b) Rsquared metric (c).MAE metric (d) NSE metric.

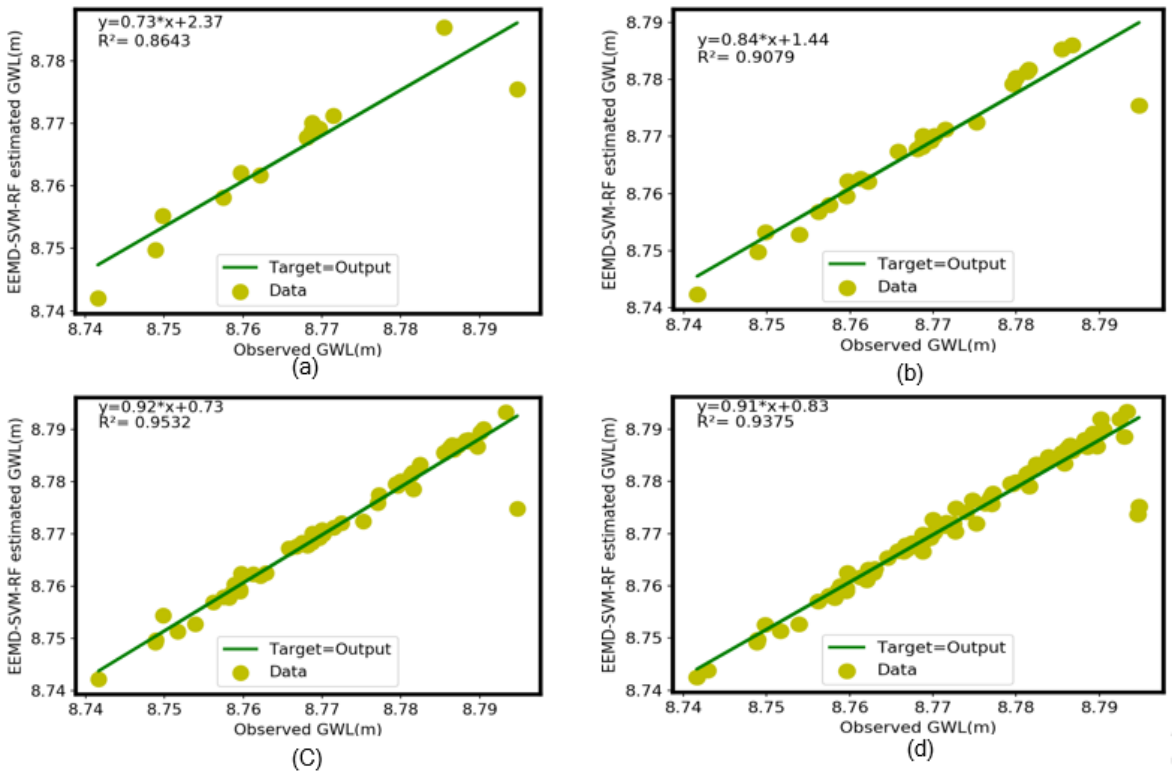



Figure 5.7 Mukarange regression (a) at 15- day horizon (b) 30-day horizon (c) 60-day horizon (d) 90-day horizon.

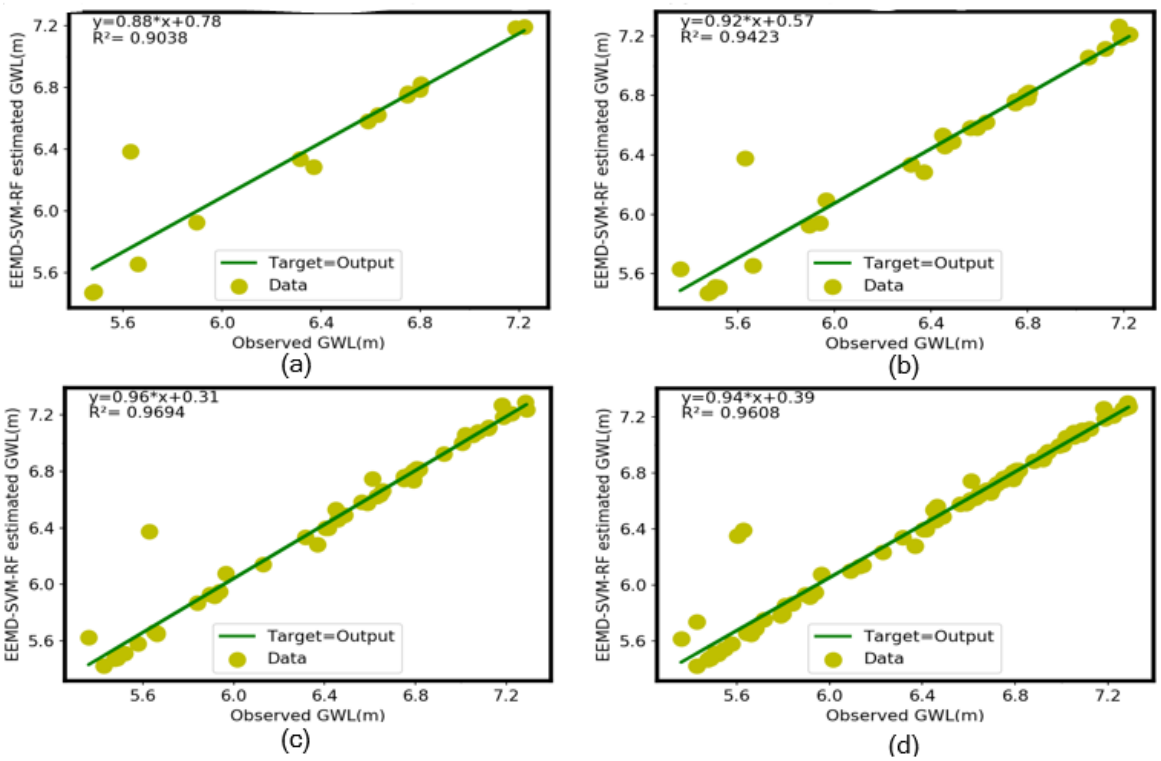
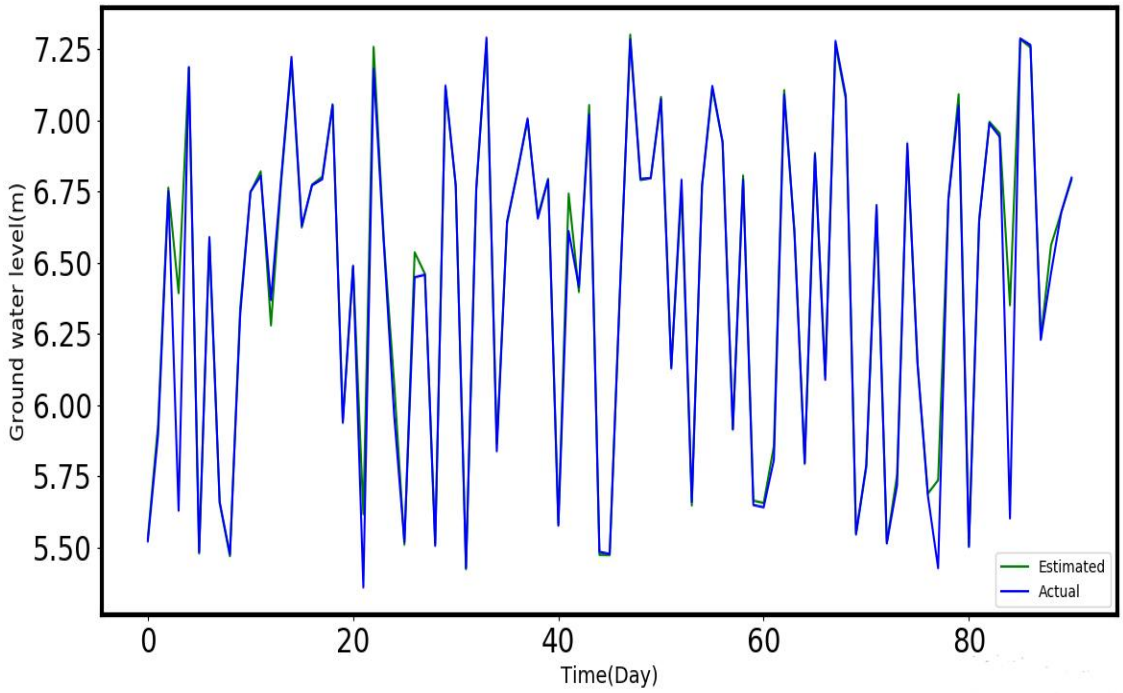
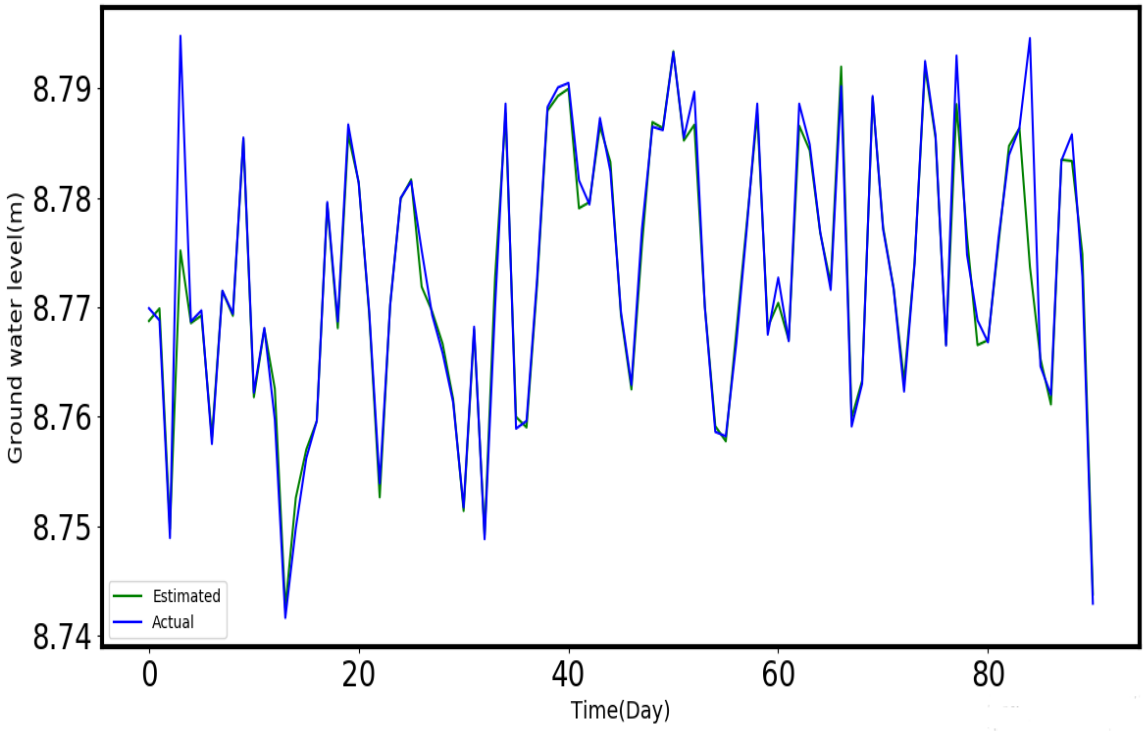



Figure 5.8 Rugarama regression (a) at 15-day horizon (b) 30-day horizon (c) 60-day horizon (d) 90-day horizon




(a)




(b)

Figure 5.9 Observed and estimated water table depth at (a) Rugarama observatory (b) Mukarange observatory station

## 5.5 Conclusions

This chapter has described the development, testing, and application of evaluated a hybrid SVR-RF method with EEMD preprocess using dataset from Rugarama and Mukarange groundwater observatory stations in Eastern Rwanda. The achieved results highlight the significance of employing weather data (temperature, solar radiation, humidity, and precipitation) to calibrate and evaluate hydrological models. The EEMD has evaluated to be a suitable data preprocessing strategy for improving performance of the hybrid approach. Most importantly, the EEMD-SVR-RF outperforms the EEMD-SVR, EEMD-ANN, and RF approaches across all prediction time steps and datasets. The SVR-RF with EEMD preprocess has demonstrated to be effective at capturing changes in water table depths. As a result, the model has the potential to be a useful tool for hydrological management and model selection.

# CHAPTER 6

# RESEARCH OUTCOMES, CONCLUSIONS, AND RECOMMENDATIONS

## 6.1 Introduction

This chapter describes the main contributions in relation to the research objectives in Section 6.2, some research gaps and blind spots in Section 6.3, suggestions for future directions in Section 6.4, and a general conclusion in Section 6.5.

## 6.2 Summary of the Outcomes

The present research sought to develop groundwater management tools based on IoT and ML technologies to aid in informed groundwater decisions in Rwanda. Chapter 3 of this thesis has presented and evaluated practically an inexpensive, low-power IoT system in relation to the first and second objectives of the research. Chapter 4 presents a hybrid ML technique for predicting seasonal water table depths using hydro-climatic data, while Chapter 5 presented an ensemble improved ML method for more accurate and reliable groundwater depths predictions in relation to the third research objective.

### 6.2.1 Development of an Affordable IoT System for Groundwater Monitoring

While Rwanda's current groundwater abstraction rate is estimated to be 80% [9] this rate will inevitably rise as water demand rises and the Ministry of Agriculture plans to expand irrigation [11]. This necessitates affordable, continuous, and reliable monitoring of this resource in order to improve current practice and inform related decisions. In this regard, Rwanda's water management policy (2010, revised in 2011) and other stakeholders call for innovative solutions to inform water management and plan for sustainable groundwater utilization [10], [16], [205].

An inexpensive IoT enabled system for remote and continuous monitoring of the quantity of the water tables was realized through the redesign of low-cost high resolution pressure sensors, I2C communication scheme, and adoption of non-proprietary software and low-power, long-distance communication method. The submersible water depth probe developed in this research costs around $55, as opposed to a conventional probe, which costs between $495 and $830. In addition to that, the

developed IoT system includes a downstream module that allows multiple stakeholders open access to and sharing of groundwater information. This tool was developed in response to the need to provide water managers, policymakers, and other interested stakeholders with easily accessible information to assist them in making sound groundwater decisions at the appropriate time. Furthermore, the tools described here, in terms of design, development, and field deployment, contribute to the body of knowledge on the design, development, and evaluation of IoT groundwater-based solutions.

### 6.2.2 Field Validation and Evaluation of the Affordable IoT Groundwater Monitoring System

Following the development of a low-cost, automatic IoT-enabled aquifer depth monitoring system, a field deployment was carried out. Field evaluation results over a two-week period show that the collected data are accurate and closely correlate (98.9%) with standard measurements, and consistent with the findings of Beddows et al. [49]. The outstanding accuracy of the developed water table depth probe suggests a higher level of reliability for the proposed system (LWNGM) for aquifer monitoring. These findings also indicated that the system is efficient, with a network reliability of around (80% packet delivery ratio and 83% signal strength) and a relatively low daily energy consumption (12%). This indicates that the system is has highly reliable network communication. The research presented here is the first to present a practical analysis of end-node and LoRa energy consumption in the context of groundwater monitoring. Energy consumption is 12% of total battery capacity of 66,600J, allowing the system to operate autonomously on solar energy. The analysis revealed, on the other hand, that the cellular internet connection between the gateway and the local server costs $1.3 per week and $5 per month. Taking into account our hardware cost ($310.168), these costs are relatively lower than those reported in other studies. For example, consider the low-cost solution presented in a recent study by Calderwood et al. [84], which has a weekly cost of $981. Additionally, the LWNGM runs on a rechargeable battery for an extended period of time, requiring a mini solar panel to be recharged after approximately 8 days of operation. Unlike the conventional systems, which are expensive and have difficult logistics, the LWNGM system’s low cost, consistency, and deployment logistics of the LWNGM system enable broader adoption of efficient and cost-effective aquifer monitoring, even in low-income areas.

### 6.2.3 Development of an Efficient Machine Learning Method for Seasonal Groundwater Quantity Prediction

As the data collection is scaled up and for a longer period of time (e.g. using the automated near real-time tool described in Chapter 3), the need for efficient, reliable, affordable, and simple tools for data analysis grows**.** Two different MLs were developed in order to meet objective three, which is to provide an efficient ML model for seasonal prediction on multiple and varying length of groundwater datasets.

There were some inconsistencies and gaps in the historical groundwater and associated weather data collected from multiple organizations (RWFA and Rwanda Meteorological Office). Appropriate data preparation procedures were thus implemented in accordance with standard procedure were also discovered to have resulted in improved model accuracy. These data preparation methods were used to transform the data and overcome the inconsistency. The square root, logarithmic, cube root, moving average, max transformation, and weighted moving average are among these methods. The weighted moving average produced smoothed data that was more realistic. To determine appropriate values for filling gaps in both weather and groundwater level data, a moving average based on neighboring (non-gap) values was used. The data was also scaled between -1 and 1 to make the features more visible to the prediction models. Additionally, groundwater level data from each monitoring station was temporally combined with weather data to form time-series datasets. Moreover, combining the RF method with SVR, ANN, and KNN to create KNN-RF and SVR-RF hybrid techniques resulted in more robust models with higher accuracy. A dictionary was created for each of these ensemble models, and a cross validation method was used to determine the appropriate number of base models for each ensemble model. The appropriate hyper-parameters for each base model as well as the ensemble models were then determined using a grid search. The final results of each hybrid model were obtained by averaging the output of each individual method. Additionally, to avoid bias in the predictive analysis, each training phase used a different size of training sample, while testing was done using separate portions of the datasets using the sliding windows approach.

Finally, and most importantly, the development process, as well as the ML methods proposed in this research, made use of open source tools, allowing for open and unlimited access to multiple hydrology stakeholders.

### 6.2.4 Evaluation of the ML techniques for Seasonal Water Table Forecasting

The two developed ML techniques were tested on real-world data from the case study of the present research. To preserve the time series nature of this data, all training and testing stages were completed using a rolling windows method. The first ML technique described in Chapter 4, showed promising seasonal forecast accuracy (93.87%), when trained with limited examples. The second ML method described in Chapter 5 produces the most accurate predictions (96.08%) on two relatively rich datasets. The use of ensemble empirical decomposition for data preparation is one of the key improvements to this model that accounts for the more accurate results.

## 6.3 Limitations of the Research

At the time of the machine learning predictive method's implementation, the temporal resolution of the collected historical water table data was limited. This data came from only three groundwater observation boreholes, two of which had many missing values beyond treatment. Ensemble models were created and applied to prediction using a limited dataset at first to overcome these constraints. Another limitation is that the author's developed IoT system only collected empirical data from a single observational well due to various constraints. By the time the groundwater records had grown a little, the second ML technique had been developed to overcome the first ML technique's limitations and work with a larger dataset.

On the other hand, the redesigned water table depth probe had a small amount of bias (1.1 %) from the standard value. A linear calibration was performed to compensate for this deviation in order to ensure more accurate measurements and greater reliability. Moreover, the solar energy harvester is normally deprived of an energy source during prolonged cloudy and rainy days. The discussed IoT-based system is powered by a battery that lasts about 8 days, ensuring that the system operates normally even when there is little solar radiation (see Chapter 3 for detailed explanation). This preserves the quality and quantity of groundwater data in both sunny and cloudy conditions.

## 6.4 The Implications of Affordable IoT and ML technologies Outcomes

The outcomes of the current study (described in Chapters 3, 4, and 5) are highly significant because they provide more affordable and efficient tools and information for the management of precious

groundwater resources, theoretical insights into the development and application of IoT and AI solutions in hydrology, and policymakers with information.

The developed low-cost, automated groundwater observation tools because they are important for improving groundwater management, particularly in low-income and data-scarce regions. Machine learning models are useful and practical tools for forecasting droughts and future changes in water availability.

The outcomes and methodological approaches are most applicable to future research in improving tools for monitoring and data collection in hydrology and hydrogeology, as well as predictive analysis.

## 6.5 Recommendations for LCSN and ML in Groundwater Hydrology and Research

The recommendations of the present study basically focus on the adoption of affordable tools to enable effective and continuous monitoring and control of the precious groundwater resources. The recommendations are presented in the subsequent subsections.

### 6.5.1 Adoption of Affordable and IoT-based Groundwater Management Tools

Due to the lack of near real-time, continuous and affordable aquifers monitoring tools and effective inexpensive groundwater predictive analysis tools data it is recommended that the use of the tools discussed in Chapters 3, 4, and 5 be accepted and used for groundwater management and policy directives.

### 6.5.2 Recommendation for Further IoT and ML Research in Aquifer Management

Based on some limitations discussed in the Section 6.3, we encourage further research into the application of IoT to provide more sophisticated groundwater management tools at affordable costs and facilitate knowledge transfer between countries, particularly in the developing world. In this regard, including a variety of experimental setups and user-centered design approaches may be more appropriate. We also recommend that more effective energy sources and storage be investigated in order to power wireless sensors installed in observational boreholes. Hence, more research is required to make IoT-based systems more autonomous and resilient in the face of natural disasters.

More research is required on developing physics-guided ML methods that improve learning by taking into account both physical and stochastic characteristics of aquifers to improve forecast accuracy.

## 6.6 Conclusion

This research proposed an ML and IoT enable system in order to put affordable and effective tools in the hands of managers, policymakers, and other interested parties in order to equip them with the right tools for the right decisions at the right time. These have been realized through the redesign of low-cost, low-power sensors as well as open-access tools. The proposed tools have repeatedly demonstrated high accuracy and efficacy in the field and real-world data evaluation. An autonomous water table monitoring system generates measurements that closely correlate with standard values, and the ML has demonstrated consistent performance in seasonal water table forecasting.

Under the current conditions, these groundwater decision support tools become especially relevant, providing instruments for uncertainty quantification and the natural account of fluctuating interfering factors influencing the variation in water table depths.

Finally, the findings of this study help to advance knowledge in both theoretical and practical aspects of WSN, IoT, hydrology, machine learning, and environmental management.

## REFERENCES


[1] R. Connor and M. Miletto, “The United Nations World Water Development Report 2022-GROUNDWATER: Making the Invisible Visible,” Perugia, Italy, 2022.

[2] S. A. Rahaman and H. A. Jamil, “An Efficient Low-Cost Smart Water Monitoring System Based on the Internet of Things and Artificial Intelligence Techniques,” in *Artificial Intelligence Applications for Smart Societies. In Studies in Distributed Intelligence book series* ., M. Elhoseny, M., Shankar, K., Abdel-Basset, Ed. Springer, Cham, 2021, pp. 33–47. doi: 10.1007/978-3-030-63068-3_3.

[3] F. Mao, K. Khamis, S. Krause, J. Clark, and D. M. Hannah, “Low-Cost Environmental Sensor Networks: Recent Advances and Future Directions,” *Front. Earth Sci.*, vol. 7, no. September, pp. 1–7, 2019, doi: 10.3389/feart.2019.00221.

[4] R. B. Johnston, “Arsenic and the 2030 Agenda for sustainable development,” 2016. doi: 10.1201/b20466-7.

[5] P. Thomson and J. Koehler, “Performance-oriented Monitoring for the Water SDG – Challenges, Tensions and Opportunities,” *Aquat. Procedia*, vol. 6, no. 0, pp. 87–95, 2016, doi: 10.1016/j.aqpro.2016.06.010.

[6] L. Guppy, P. Uyttendaele, K. G. Villholth, and V. Smakhtin, “Groundwater and Sustainable Development Goals : Analysis of Interlinkages,” Hamilton, 2018. [Online]. Available: http://inweh.unu.edu/publications/

[7] UNESCO, “GROUNDWATER - Making the invisible visible, The United Nations World Water Development Report 2022,” Paris, 2022.

[8] C. B. Gaye and C. Tindimugaya, “Review: Challenges and opportunities for sustainable groundwater management in Africa,” *Hydrogeol. J.*, vol. 27, no. 3, pp. 1099–1110, 2019, doi: 10.1007/s10040-018-1892-1.

[9] R. O. RWANDA, “CONSULTANCY SERVICES FOR DEVELOPMENT OF RWANDA NATIONAL WATER RESOURCES MASTER PLAN,” 2014.

[10] R. O. RWANDA, “WATER RESOURCES MANAGEMENT SUB-SECTOR STRATEGIC PLAN (2011 – 2015),” Kigali, Rwanda, 2015.

[11] M. M. Malesu *et al.*, “Rwanda Irrigation Master Plan,” Nairobi, 2010.

[12] R. O. RWANDA, “Strategic Programme for Climate Resilience ( SPCR ),” 2017.

[13] RWFA, “Terms of Reference: Groundwater and Recharge and storage Enhancement in

Eastern Province of Rwanda," Kigali, 2018. [Online]. Available: https://waterportal.rwb.rw/report/ground-water-recharge-and-storage-enhancement-eastern-province-rwanda

[14] U. Nations, "Grounewater in Eastern, Central and Southern Africa," New York, 1989.

[15] A. Alamanos and A. Rolston, "Development of a Decision Support System for Sustainable Environmental Management and Stakeholder Engagement," vol. 8, no. 40, pp. 1–20, 2021.

[16] A. TWAHIRWA, "Innovation to sustainable water management in Rwanda," in *Innovation to sustainable water management in Rwanda*, 2017, pp. 1–5.

[17] O. P. Abimbola, J. Wenninger, R. Venneker, and A. R. Mittelstet, "The assessment of water resources in ungauged catchments in Rwanda," *J. Hydrol. Reg. Stud.*, vol. 13, no. September 2017, pp. 274–289, 2017, doi: 10.1016/j.ejrh.2017.09.001.

[18] UN-WWAP, *The United Nations World Water Development Report 2015: Water for a Sustainable World*. UNESCO, 2015. doi: 978-92-3-100071-3.

[19] WWF GEF, "Food Systems, Land Use and Restoration in Tanzania's Forest Landscape," 2021. [Online]. Available: https://files.worldwildlife.org/wwfcmsprod/files/Publication/file/7vu2a568c8_10262_Tanzania_FOLUR_ProDoc_submission_7_track_chage_1_

[20] T. Gleeson, M. Cuthbert, G. Ferguson, and D. Perrone, "Global Groundwater Sustainability, Resources, and Systems in the Anthropocene," *Annu. Rev. Earth Planet. Sci.*, vol. 48, pp. 431–463, 2020, doi: 10.1146/annurev-earth-071719-055251.

[21] R. G. Taylor, G. Favreau, B. R. Scanlon, and K. G. Villholth, "Topical Collection: Determining groundwater sustainability from long-term piezometry in Sub-Saharan Africa," *Hydrogeol. J.*, vol. 27, no. 2, pp. 443–446, 2019, doi: 10.1007/s10040-019-01946-9.

[22] S. Stoll, H. J. Hendricks Franssen, R. Barthel, and W. Kinzelbach, "What can we learn from long-term groundwater data to improve climate change impact studies?," *Hydrol. Earth Syst. Sci.*, vol. 15, no. 12, pp. 3861–3875, 2011, doi: 10.5194/hess-15-3861-2011.

[23] A. Tuinhof, S. Foster, F. van Steenbergen, A. Talbi, and M. Wishart, "Appropriate Groundwater Management Policy for Sub-Saharan Africa," *Appropr. Groundw. Manag. Policy Sub-Saharan Africa*, 2011, doi: 10.1596/27363.

[24] I. Shiklomanov, H. Lins, and E. STAKhiv, "Hydrology and water resources," *Clim. Chang. IPCC Impacts Assess.*, 1990, [Online]. Available:

http://scholar.google.com/scholar?q=related:nA_1a254upcJ:scholar.google.com/&hl=en&num=20&as_sdt=0,5%5Cnpapers3://publication/uuid/6A5154FF-6921-4CB5-BD1E-9B08C3ADC119

[25] P. Fitch, B. Brodaric, M. Stenson, and N. Booth, *Integrated groundwater data management*. Springer Open, 2016. doi: 10.1007/978-3-319-23576-9_26.

[26] M. P. Clark *et al.*, "Characterizing Uncertainty of the Hydrologic Impacts of Climate Change," *Curr. Clim. Chang. Reports*, vol. 2, no. 2, pp. 55–64, 2016, doi: 10.1007/s40641-016-0034-x.

[27] G. Mcgregor, "Hydroclimatology , modes of climatic variability and stream flow , lake and groundwater level variability : A progress report," 2017, doi: 10.1177/0309133317726537.

[28] K. C. Urama and N. Ozor, "IMPACTS OF CLIMATE CHANGE ON WATER RESOURCES IN AFRICA : the Role of Adaptation," no. December, pp. 1–29, 2010.

[29] M. O. Cuthbert *et al.*, "Global patterns and dynamics of climate–groundwater interactions," *Nat. Clim. Chang.*, vol. 9, no. 2, pp. 137–141, 2019, doi: 10.1038/s41558-018-0386-4.

[30] B. D. Bowes, J. M. Sadler, M. M. Morsy, M. Behl, and J. L. Goodall, "Forecasting groundwater table in a flood prone coastal city with long short-term memory and recurrent neural networks," *Water (Switzerland)*, vol. 11, no. 5, pp. 1–38, 2019, doi: 10.3390/w11051098.

[31] K. Warner *et al.*, "Climate change profile: Rwanda," *Netherlands Minist. Foreign Aff.*, no. July, p. 16, 2015.

[32] W. M. Alley, "Flow and Storage in Groundwater Systems," vol. 1985, no. 2002, 2008, doi: 10.1126/science.1067123.

[33] S. feng Yan, S. en Yu, Y. bai Wu, D. feng Pan, and J. gen Dong, "Understanding groundwater table using a statistical model," *Water Sci. Eng.*, vol. 11, no. 1, pp. 1–7, 2018, doi: 10.1016/j.wse.2018.03.003.

[34] Q. Tang *et al.*, "Hydrological monitoring and seasonal forecasting: Progress and perspectives," *J. Geogr. Sci.*, vol. 26, no. 7, pp. 904–920, 2016, doi: 10.1007/s11442-016-1306-z.

[35] P. Döll *et al.*, "Impact of water withdrawals from groundwater and surface water on continental water storage variations," *J. Geodyn.*, vol. 59–60, pp. 143–156, 2012, doi: 10.1016/j.jog.2011.05.001.

[36] R. G. Taylor, "Ground water and climate change," *Nat. Clim. Chang.*, vol. 3, no. November 2012, pp. 1–13, 2013, doi: 10.1038/NCLIMATE1744.

[37] J. C. Haas and S. Birk, "Characterizing the spatiotemporal variability of groundwater levels

of alluvial aquifers in different settings using drought indices," *Hydrol. Earth Syst. Sci.*, vol. 21, no. 5, pp. 2421–2448, 2017, doi: 10.5194/hess-21-2421-2017.

[38] E. K. White, T. J. Peterson, J. Costelloe, A. W. Western, and E. Carrara, "Can we manage groundwater? A method to determine the quantitative testability of groundwater management plans," *Water Resour. Res.*, vol. 52, no. 6, pp. 4863–4882, 2016, doi: 10.1002/2015WR018474.

[39] S. Foster, A. Tuinhof, and H. Garduño, "Groundwater Development in Sub-Saharan Africa. A Strategic Overview of Key Issues and Major Needs," 2006. [Online]. Available: https://www.un-igrac.org/sites/default/files/resources/files/GWMATE case profile - Sub Saharan Africa.pdf

[40] C. J. Taylor and W. M. Alley, "Ground-water-level monitoring and the importance of long-term water-level data," 2001.

[41] M. Mazzoleni, L. Alfonso, and D. Solomatine, "Influence of spatial distribution of sensors and observation accuracy on the assimilation of distributed streamflow data in hydrological modelling," *Hydrol. Sci. J.*, vol. 62, no. 3, pp. 389–407, 2017, doi: 10.1080/02626667.2016.1247211.

[42] W. Visser and G. H. Brundtland, "Our Common Future," New York, Oxford University Press, 1987. doi: 10.9774/gleaf.978-1-907643-44-6_12.

[43] D. O. V. Kotchoni, J. M. Vouillamoz, F. M. A. Lawson, P. Adjomayi, M. Boukari, and R. G. Taylor, "Relationships between rainfall and groundwater recharge in seasonally humid Benin: a comparative analysis of long-term hydrographs in sedimentary and crystalline aquifers," *Hydrogeol. J.*, vol. 27, no. 2, pp. 447–457, 2019, doi: 10.1007/s10040-018-1806-2.

[44] N. Njue *et al.*, "Citizen science in hydrological monitoring and ecosystem services management: State of the art and future prospects," *Sci. Total Environ.*, vol. 693, no. 133531, 2019, doi: 10.1016/j.scitotenv.2019.07.337.

[45] G. M. Lovett *et al.*, "Who needs environmental monitoring?," *Front. Ecol. Environ.*, vol. 5, no. 5, pp. 253–260, 2007, doi: 10.1890/1540-9295(2007)5[253:WNEM]2.0.CO;2.

[46] A. G. Law, "Stochastic Analysis of Groundwater Level Time Series in the Western United States.," *Colo State Univ (Fort Collins), Hydrol Pap.*, no. 68, p. 33, 1974.

[47] C. Shen, X. Chen, and E. Laloy, "Editorial: Broadening the Use of Machine Learning in Hydrology," *Front. Water*, vol. 3, no. May, pp. 1–4, 2021, doi: 10.3389/frwa.2021.681023.

[48] K. Chan *et al.*, "Low-cost electronic sensors for environmental research : Pitfalls and opportunities," *Prog. Phys. Geogr.*, vol. 45, no. 3, pp. 305–338, 2021, doi:

10.1177/0309133320956567.

[49] P. A. Beddows and Edward K. Mallon, “Cave Pearl Data Logger : A Flexible Arduino-Based Harsh Environments,” *sesnors*, vol. 18, no. 530, pp. 1–26, 2018, doi: 10.3390/s18020530.

[50] Jousma. G., “Guideline on : Groundwater monitoring for general reference purposes,” *Solutions*, no. June, p. 165, 2008.

[51] J. T. . N. V. for W. and C. S. in A. Snow *et al.*, “A New Vision for Weather and Climate Services in Africa,” 2016.

[52] D. Tetzlaff, C. Soulsby, S. K. Carey, J. P. Mcnamara, and H. Laudon, “The essential value of long-term experimental data for hydrology and water management,” *Water Resour. Res.*, pp. 2598–2604, 2017, doi: 10.1002/2017WR020838.Received.

[53] J. D. Hewlett, H. W. Lull, and K. G. Reinhart, “In Defense of Experimental Watersheds,” *Water Resour. Res.*, vol. 5, no. 1, pp. 306–316, 1969, doi: 10.1029/WR005i001p00306.

[54] I. H. Decade, “Results ~ f research on representative and experimental basins Résultats de recherches sur les bassins représentatifs et,” *Actes Colloq.*, vol. 1, 1973.

[55] B. Docker and I. Robinson, “Environmental water management in Australia: Experience from the Murray-Darling Basin,” *Int. J. Water Resour. Dev.*, vol. 30, no. 1, pp. 164–177, 2014, doi: 10.1080/07900627.2013.792039.

[56] C. Huntingford *et al.*, “Potential influences on the United Kingdom’s floods of winter 2013/14,” *Nat. Clim. Chang.*, vol. 4, no. 9, pp. 769–777, 2014, doi: 10.1038/nclimate2314.

[57] R. J. McCutcheon, J. P. McNamara, M. J. Kohn, and S. L. Evans, “An evaluation of the ecohydrological separation hypothesis in a semiarid catchment,” *Hydrol. Process.*, vol. 31, no. 4, pp. 783–799, 2017, doi: 10.1002/hyp.11052.

[58] L. Mainetti, L. Patrono, and a. Vilei, “Evolution of wireless sensor networks towards the Internet of Things: A survey,” *SoftCOM*, pp. 1–6, 2011, [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6064380%5Cnhttp://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6064380

[59] C. Buratti, A. Conti, D. Dardari, and R. Verdone, “An overview on wireless sensor networks technology and evolution,” *Sensors*, vol. 9, no. 9, pp. 6869–6896, 2009, doi: 10.3390/s90906869.

[60] S. Labs, “The Evolution of Wireless Sensor Networks,” 2004. [Online]. Available: http://www.silabs.com/Support Documents/TechnicalDocs/evolution-of-wireless-sensor-networks.pdf

[61] I. Khan *et al.*, “Wsn Virtualization: Survey”.

[62] L. Atzorii, A. Lera, and G. Morabito, “The internet of things: a survey,” *Inf. Syst. Front.*, vol. 17, no. 2, pp. 1–20, 2010, doi: doi:10.1016/ j.comnet.2010.05.010.

[63] J. E. Siegel, S. Kumar, and S. E. Sarma, “The Future Internet of Things: Secure, Efficient, and Model-Based,” *IEEE Internet Things J.*, no. October, 2017, doi: 10.1109/JIOT.2017.2755620.

[64] P. Biggs, J. Garrity, and C. LaSalle, “Harnessing-IoT-Global-Development.pdf,” 2015. [Online]. Available: https://www.itu.int/en/action/broadband/Documents/Harnessing-IoT-Global-Development.pdf

[65] M. Ahmad, “Designing for the Internet of Things: A paradigm shift in reliability,” *Proc. - Electron. Components Technol. Conf.*, vol. 2015-July, pp. 1758–1766, 2015, doi: 10.1109/ECTC.2015.7159836.

[66] I. Gronbaek, “Architecture for the Internet of Things (IoT): API and interconnect,” *Proc. - 2nd Int. Conf. Sens. Technol. Appl., SENSORCOMM 2008, Incl. MESH 2008 Conf. Mesh Networks; ENOPT 2008 Energy Optim. Wirel. Sensors Networks, UNWAT 2008 Under Water Sensors Syst.*, pp. 802–807, 2008, doi: 10.1109/SENSORCOMM.2008.20.

[67] R. Khan, S. U. Khan, R. Zaheer, and S. Khan, “Future internet: The internet of things architecture, possible applications and key challenges,” *Proc. - 10th Int. Conf. Front. Inf. Technol. FIT 2012*, pp. 257–260, 2012, doi: 10.1109/FIT.2012.53.

[68] T. Robles *et al.*, “An IoT based reference architecture for smart water management processes,” *J. Wirel. Mob. Networks, Ubiquitous Comput. Dependable Appl.*, vol. 6, no. 1, pp. 4–23, 2015.

[69] K. Amemiya and A. Nakao, “Layer-Integrated Edge Distributed Data Store for Real-time and Stateful Services,” *Proc. IEEE/IFIP Netw. Oper. Manag. Symp. 2020 Manag. Age Softwarization Artif. Intell. NOMS 2020*, pp. 1–9, 2020, doi: 10.1109/NOMS47738.2020.9110436.

[70] T. Bradicich, “The Intelligent Edge: What it is, what it’s not, and why it’s useful,” 2018. https://www.hpe.com/us/en/insights/articles/the-intelligent-edge-what-it-is-what-its-not-and-why-its-useful-1711.html

[71] J. Haxhibeqiri, F. Van den Abeele, I. Moerman, and J. Hoebeke, “LoRa scalability: A simulation model based on interference measurements,” *Sensors (Switzerland)*, vol. 17, no. 6, 2017, doi: 10.3390/s17061193.

[72] K. Mekki, E. Bajic, F. Chaxel, and F. Meyer, “A comparative study of LPWAN technologies

for large-scale IoT deployment," *ICT Express*, vol. 5, no. 1, pp. 1–7, 2019, doi: 10.1016/j.icte.2017.12.005.

[73] A. Lavric and A. Loan Petrariu, "High-Density Low Power Wide Area Networks," *Proc. 10th Int. Conf. Electron. Comput. Artif. Intell. ECAI 2018*, 2019, doi: 10.1109/ECAI.2018.8678997.

[74] S. Bertoldo, L. Carosso, E. Marchetta, M. Paredes, and M. Allegretti, "Feasibility Analysis of a LoRa-Based WSN Using Public Transport," *Appl. Syst. Innov.*, vol. 1, no. 4, p. 49, 2018, doi: 10.3390/asi1040049.

[75] A. Augustin, J. Yi, T. Clausen, and W. M. Townsley, "A study of Lora: Long range & low power networks for the internet of things," *Sensors (Switzerland)*, vol. 16, no. 9, pp. 1–18, 2016, doi: 10.3390/s16091466.

[76] LoRa™ Alliance, "LoRaWAN™ Specification v1.0.2," 2016. [Online]. Available: https://lora-alliance.org/sites/default/files/2018-05/lorawan1_0_2-20161012_1398_1.pdf

[77] A. L. I. Nikoukar, S. Raza, A. Poole, M. Güneş, and B. Dezfouli, "Low-Power Wireless for the Internet of Things : Standards and Applications," *IEEE Access*, vol. 6, pp. 67893–67926, 2018, doi: 10.1109/ACCESS.2018.2879189.

[78] A. Potsch and F. Hammer, "Towards End-to-End Latency of LoRaWAN: Experimental Analysis and IIoT Applicability," *IEEE Int. Work. Fact. Commun. Syst. - Proceedings, WFCS*, vol. 2019-May, 2019, doi: 10.1109/WFCS.2019.8758033.

[79] C. Gomez, J. C. Veras, R. Vidal, L. Casals, and J. Paradells, "A sigfox energy consumption model," *Sensors (Switzerland)*, vol. 19, no. 3, 2019, doi: 10.3390/s19030681.

[80] Semtech, "AN1200.22 LoRa Modulation Basics," Flynn Road, Camarillo, California, 2015. [Online]. Available: http://www.semtech.com/images/datasheet/an1200.22.pdf

[81] R. Ghanaatian, O. Afisiadis, M. Cotting, and A. Burg, "Lora Digital Receiver Analysis and Implementation," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2019, vol. 2019-May, pp. 1498–1502. doi: 10.1109/ICASSP.2019.8683504.

[82] A. M. Yousuf, E. M. Rochester, and M. Ghaderi, "A low-cost LoRaWAN testbed for IoT: Implementation and measurements," *IEEE World Forum Internet Things, WF-IoT 2018 - Proc.*, vol. 2018-Janua, pp. 361–366, 2018, doi: 10.1109/WF-IoT.2018.8355180.

[83] P. Kumar *et al.*, "The rise of low-cost sensing for managing air pollution in cities," *Environ. Int.*, vol. 75, pp. 199–205, 2014, doi: 10.1016/j.envint.2014.11.019.

[84] A. J. Calderwood, R. A. Pauloo, A. M. Yoder, and G. E. Fogg, “Low-cost, open source wireless sensor network for real-time, scalable groundwater monitoring,” *Water (Switzerland)*, vol. 12, no. 4, pp. 1–17, 2020, doi: 10.3390/W12041066.

[85] C. J. García-Orellana, M. Macías-Macías, H. M. González-Velasco, A. García-Manso, and R. Gallardo-Caballero, “Low-power and low-cost environmental iot electronic nose using initial action period measurements,” *Sensors (Switzerland)*, vol. 19, no. 14, 2019, doi: 10.3390/s19143183.

[86] N. Harris, “Powering the Environmental Internet of Things,” 2019, doi: 10.3390/s19081940.

[87] N. Garg and R. Garg, “Energy harvesting in IoT devices: A survey,” *Proc. Int. Conf. Intell. Sustain. Syst. ICISS 2017*, no. Iciss, pp. 127–131, 2018, doi: 10.1109/ISS1.2017.8389371.

[88] T. Ruan, Z. J. Chew, and M. Zhu, “Energy-Aware Approaches for Energy Harvesting Powered Wireless Sensor Nodes,” *IEEE Sens. J.*, vol. 17, no. 7, pp. 2165–2173, 2017, doi: 10.1109/JSEN.2017.2665680.

[89] R. Martac, N. Milivojevic, V. Milivojevic, V. Cirovic, and D. Barac, “Using internet of things in monitoring and management of dams in Serbia,” *Facta Univ. - Ser. Electron. Energ.*, vol. 29, no. 3, pp. 419–435, 2016, doi: 10.2298/FUEE1603419M.

[90] P. M. Pujar, H. H. Kenchannavar, R. M. Kulkarni, and U. P. Kulkarni, “Real-time water quality monitoring through Internet of Things and ANOVA-based analysis: a case study on river Krishna,” *Appl. Water Sci.*, vol. 10, no. 1, pp. 1–16, 2020, doi: 10.1007/s13201-019-1111-9.

[91] M. A. Nasirudin, U. N. Za’bah, and O. Sidek, “Fresh water real-time monitoring system based on Wireless Sensor Network and GSM,” *Open Syst. (ICOS), 2011 IEEE Conf.*, pp. 354–357, 2011, doi: 10.1109/ICOS.2011.6079290.

[92] C. Encinas, E. Ruiz, J. Cortez, and A. Espinoza, “Design and implementation of a distributed IoT system for the monitoring of water quality in aquaculture,” 2017, doi: 10.1109/WTS.2017.7943540.

[93] Q. Zhou *et al.*, “Real-time management of groundwater resources based on wireless sensors networks,” *J. Sens. Actuator Networks*, vol. 7, no. 1, pp. 1–11, 2018, doi: 10.3390/jsan7010004.

[94] L. A. Méndez-Barroso, J. A. Rivas-Márquez, I. Sosa-Tinoco, and A. Robles-Morúa, “Design and implementation of a low-cost multiparameter probe to evaluate the temporal variations of water quality conditions on an estuarine lagoon system,” *Environ. Monit. Assess.*, vol. 192, no. 11, 2020, doi: 10.1007/s10661-020-08677-5.

[95] S. Anumalla, B. Ramamurthy, D. C. Gosselin, and M. Burbach, “Ground Water Monitoring

using Smart Sensors," *2005 IEEE Int. Conf. Electro Inf. Technol.*, pp. 1–6, 2005, doi: 10.1109/EIT.2005.1626962.

[96] X. Li, X. Cheng, P. Gong, and K. Yan, "Design and implementation of a wireless sensor network-based remote water-level monitoring system," *Sensors*, vol. 11, no. 2, pp. 1706–1720, 2011, doi: 10.3390/s110201706.

[97] J. McCarthy, "What Is Artificial Intelligence Anyway," 1985.

[98] K. Kersting, "Machine Learning and Artificial Intelligence: Two Fellow Travelers on the Quest for Intelligent Behavior in Machines," *Front. Big Data*, vol. 1, no. November, pp. 1–4, 2018, doi: 10.3389/fdata.2018.00006.

[99] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[100] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00318-5.

[101] G. Nguyen *et al.*, "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 77–124, 2019, doi: 10.1007/s10462-018-09679-z.

[102] B. Mohammadi, "A review on the applications of machine learning for runoff modeling," *Sustain. Water Resour. Manag.*, vol. 7, no. 6, pp. 1–11, 2021, doi: 10.1007/s40899-021-00584-y.

[103] T. Peng, J. Zhou, C. Zhang, and W. Fu, "Streamflow forecasting using empirical wavelet transform and artificial neural networks," *Water (Switzerland)*, vol. 9, no. 6, pp. 1–20, 2017, doi: 10.3390/w9060406.

[104] E. Rozos, "Machine learning, urbanwater resources management and operating policy," *Resources*, vol. 8, no. 4, pp. 2–13, 2019, doi: 10.3390/RESOURCES8040173.

[105] A. neuro approach to water management systems Kambli and S. Modi, "Fuzzy neuro approach to water management systems," *ACM Int. Conf. Proceeding Ser.*, pp. 215–220, 2019, doi: 10.1145/3310986.3311026.

[106] G. Shmueli, "To explain or to predict?," *Stat. Sci.*, vol. 25, no. 3, pp. 289–310, 2010, doi: 10.1214/10-STS330.

[107] S. J. Cranmer and B. A. Desmarais, "What can we learn from predictive modeling?," *Polit. Anal.*, vol. 25, no. 2, pp. 145–166, 2017, doi: 10.1017/pan.2017.3.

[108] P. Birbal, H. Azamathulla, L. Leon, J. Hosein, and V. Kumar, “Predictive modelling of the stage-discharge relationship using gene-expression programming,” *Water Supply*, vol. 21, no. 7, pp. 3503–3514, 2021, doi: 10.2166/ws.2021.111.

[109] G. S. Nearing *et al.*, “What Role Does Hydrological Science Play in the Age of Machine Learning?,” *Water Resour. Res.*, vol. 57, no. 3, 2021, doi: 10.1029/2020WR028091.

[110] H. Ji, Y. Chen, G. Fang, Z. Li, W. Duan, and Q. Zhang, “Adaptability of machine learning methods and hydrological models to discharge simulations in data-sparse glaciated watersheds,” *J. Arid Land*, vol. 13, no. 6, pp. 549–567, 2021, doi: 10.1007/s40333-021-0066-5.

[111] V. M. Herrera, T. M. Khoshgoftaar, F. Villanustre, and B. Furht, *Random forest implementation and optimization for Big Data analytics on LexisNexis’s high performance computing cluster platform*, vol. 6, no. 1. Springer International Publishing, 2019. doi: 10.1186/s40537-019-0232-1.

[112] X. Wang, T. Liu, X. Zheng, H. Peng, J. Xin, and B. Zhang, “Short - term prediction of groundwater level using improved random forest regression with a combination of random features,” *Appl. Water Sci.*, vol. 8, no. 5, pp. 1–12, 2018, doi: 10.1007/s13201-018-0742-6.

[113] S. A. Naghibi, H. R. Pourghasemi, and B. Dixon, “GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran,” *Environ. Monit. Assess.*, vol. 188, no. 1, pp. 1–27, 2016, doi: 10.1007/s10661-015-5049-6.

[114] S. Huang, J. Chang, Q. Huang, and Y. Chen, “Monthly streamflow prediction using modified EMD-based support vector machine,” *J. Hydrol.*, vol. 511, pp. 764–775, 2014, doi: 10.1016/j.jhydrol.2014.01.062.

[115] S. Raghavendra and P. C. Deka, “Support vector machine applications in the field of hydrology: A review,” *Appl. Soft Comput. J.*, vol. 19, pp. 372–386, 2014, doi: 10.1016/j.asoc.2014.02.002.

[116] O. Kisi and K. S. Parmar, “Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution,” *J. Hydrol.*, vol. 534, pp. 104–112, 2016, doi: 10.1016/j.jhydrol.2015.12.014.

[117] H. Yu, X. Wen, and Q. Feng, “Comparative Study of Hybrid-Wavelet Artificial Intelligence Models for Monthly Groundwater Depth Forecasting in Extreme Arid Regions , Northwest China,” vol. 32, pp. 301–321, 2017, doi: 10.1007/s11269-017-1811-6.

[118] S. Sahoo, T. A. Russo, J. Elliott, and I. Foster, “Machine Learning Algorithms for Modeling Groundwater Level Changes in Agricultural Regions of the U . S .,” *Water Resour. Res.*, vol. 53, no. 5, pp. 3878–3895, 2017.

[119] S. Mohanty, M. K. Jha, S. K. Raul, R. K. Panda, and K. P. Sudheer, “Using Artificial Neural Network Approach for Simultaneous Forecasting of Weekly Groundwater Levels at Multiple Sites,” *Water Resour. Manag.*, vol. 29, no. 15, pp. 5521–5532, 2015, doi: 10.1007/s11269-015-1132-6.

[120] ASCE Task Committee on Application of Artificial Neural Network in Hydrology, “Artificial Neural Network in Hydrology. I:Priliminary Concepts,” *J. Hydrol. Eng.*, vol. 5, no. 2, pp. 115–123, 2000.

[121] W. Sun and B. Trevor, “Combining k-nearest-neighbor models for annual peak breakup flow forecasting,” *Cold Reg. Sci. Technol.*, vol. 143, pp. 59–69, 2017, doi: 10.1016/j.coldregions.2017.08.009.

[122] T. Song, W. Ding, H. Liu, J. Wu, H. Zhou, and J. Chu, “Uncertainty Quantification in Machine Learning Modeling for Multi-Step Time Series Forecasting : Example of Recurrent Neural Networks in,” *Water (Switzerland)*, vol. 12, no. 912, pp. 2–19, 2020.

[123] A. D. Gorgij, O. Kisi, and A. A. Moghaddam, “Groundwater budget forecasting, using hybrid wavelet-ANN-GP modelling: A case study of Azarshahr Plain, East Azerbaijan, Iran,” *Hydrol. Res.*, vol. 48, no. 2, pp. 455–467, 2017, doi: 10.2166/nh.2016.202.

[124] R. Barzegar, E. Fijani, A. Asghari, and E. Tziritis, “Science of the Total Environment Forecasting of groundwater level fl uctuations using ensemble hybrid multi-wavelet neural network-based models,” *Sci. Total Environ.*, vol. 599–600, pp. 20–31, 2017, doi: 10.1016/j.scitotenv.2017.04.189.

[125] F. J. Chang, L. C. Chang, C. W. Huang, and I. F. Kao, “Prediction of monthly regional groundwater levels through hybrid soft-computing techniques,” *J. Hydrol.*, vol. 541, pp. 965–976, 2016, doi: 10.1016/j.jhydrol.2016.08.006.

[126] A. Navot, L. Shpigelman, N. Tishby, and E. Vaadia, “Nearest neighbor based feature selection for regression and its application to neural activity,” *Adv. Neural Inf. Process. Syst.*, no. January, pp. 995–1002, 2005.

[127] K. Fukunaga and L. D. Hostetler, “k-Nearest-Neighbor Bayes-Risk Estimation,” *IEEE Trans. Inf. Theory*, vol. 21, no. 3, pp. 285–293, 1975, doi: 10.1109/TIT.1975.1055373.

[128] D. Nguyen, S. Ouala, L. Drumetz, and R. Fablet, “EM-like Learning Chaotic Dynamics from

Noisy and Partial Observations," *arxiv:1903.10335v1*, pp. 1–14, 2019, [Online]. Available: http://arxiv.org/abs/1903.10335

[129] J. Brajard, A. Carrassi, M. Bocquet, and L. Bertino, "Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: a case study with the Lorenz 96 model," *Geosci. Model Dev. Discuss.*, no. May, pp. 1–21, 2019, doi: 10.5194/gmd-2019-136.

[130] B. Vaheddoost, "Application of hybrid ANN-whale optimization model in evaluation of the field capacity and the permanent wilting point of the soils," 2020.

[131] R. Moazenzadeh and B. Mohammadi, "Assessment of bio-inspired metaheuristic optimisation algorithms for estimating soil temperature," *Geoderma*, vol. 353, no. June, pp. 152–171, 2019, doi: 10.1016/j.geoderma.2019.06.028.

[132] A. N. Samatin Njikam and H. Zhao, "A novel activation function for multilayer feed-forward neural networks," *Appl. Intell.*, vol. 45, no. 1, pp. 75–82, 2016, doi: 10.1007/s10489-015-0744-0.

[133] C. Cortes and V. ; Vapnik, "Support-Vector Networks," *Mach. Leaming*, vol. 20, no. 5, pp. 273–297, 1995, doi: 10.1109/64.163674.

[134] P. Aghelpour, B. Mohammadi, and S. M. Biazar, "Long-term monthly average temperature forecasting in some climate types of Iran, using the models SARIMA, SVR, and SVR-FA," *Theor. Appl. Climatol.*, vol. 138, no. 3–4, pp. 1471–1480, 2019, doi: 10.1007/s00704-019-02905-w.

[135] G. F. Lin, G. R. Chen, M. C. Wu, and Y. C. Chou, "Effective forecasting of hourly typhoon rainfall using support vector machines," *Water Resour. Res.*, vol. 45, no. 8, pp. 1–11, 2009, doi: 10.1029/2009WR007911.

[136] S. M. Guzman, J. O. Paz, M. L. M. Tagert, and A. E. Mercer, "Evaluation of Seasonally Classified Inputs for the Prediction of Daily Groundwater Levels: NARX Networks Vs Support Vector Machines," *Environ. Model. Assess.*, vol. 24, no. 2, pp. 223–234, 2019, doi: 10.1007/s10666-018-9639-x.

[137] M. K. Goyal, B. Bharti, J. Quilty, J. Adamowski, and A. Pandey, "Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS," *Expert Syst. Appl.*, vol. 41, no. 11, pp. 5267–5276, 2014, doi: 10.1016/j.eswa.2014.02.047.

[138] L. Breiman, "RANDOM FORESTS," *Univ. Calif.*, pp. 1–33, 2001.

[139] L. Breiman, "Bagging predictors," in *In International Workshop on Multiple Classifier Systems*, 1996, vol. 24, pp. 123–140. doi: 10.3390/risks8030083.

[140] Y. Wang, S. T. Xia, Q. Tang, J. Wu, and X. Zhu, “A novel consistent random forest framework: Bernoulli random forests,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 8, pp. 3510–3523, 2018, doi: 10.1109/TNNLS.2017.2729778.

[141] B. Mohammadi, “Predicting total phosphorus levels as indicators for shallow lake management,” *Ecol. Indic.*, vol. 107, no. August, p. 105664, 2019, doi: 10.1016/j.ecolind.2019.105664.

[142] J. L. Neumann *et al.*, “Can seasonal hydrological forecasts inform local decisions and actions? A decision-making activity,” *Geosci. Commun.*, vol. 1, no. 1, pp. 35–57, 2018, doi: 10.5194/gc-1-35-2018.

[143] H. Tao *et al.*, “Groundwater level prediction using machine learning models: A comprehensive review,” *Neurocomputing*, vol. 489, no. C, pp. 271–308, 2022, doi: 10.1016/j.neucom.2022.03.014.

[144] H. Yoon, Y. Kim, K. Ha, S.-H. Lee, and G.-P. Kim, “Comparative Evaluation of ANN- and SVM-Time Series Models for Predicting Freshwater-Saltwater Interface Fluctuations,” *Water (Switzerland)*, vol. 323, no. 9, pp. 1–16, 2017, doi: 10.3390/w9050323.

[145] M. Mokhtarzad, F. Eskandari, N. Jamshidi Vanjani, and A. Arabasadi, “Drought forecasting by ANN, ANFIS, and SVM and comparison of the models,” *Environ. Earth Sci.*, vol. 76, no. 21, pp. 0–9, 2017, doi: 10.1007/s12665-017-7064-0.

[146] Y. Gong, Y. Zhang, S. Lan, and H. Wang, “A Comparative Study of Artificial Neural Networks, Support Vector Machines and Adaptive Neuro Fuzzy Inference System for Forecasting Groundwater Levels near Lake Okeechobee, Florida,” *Water Resour. Manag.*, vol. 30, no. 1, pp. 375–391, 2016, doi: 10.1007/s11269-015-1167-8.

[147] N. Natarajan and C. Sudheer, “Groundwater level forecasting using soft computing techniques,” *Neural Comput. Appl.*, vol. 5, 2019, doi: 10.1007/s00521-019-04234-5.

[148] C. Neurocomputing An integrated wavelet-support vector machine for groundwater level prediction in Visakhapatnam , IndiaSuryanarayana, C. Sudheer, V. Mahammood, and B. K. Panigrahi, “Neurocomputing An integrated wavelet-support vector machine for groundwater level prediction in Visakhapatnam , India,” *Neurocomputing*, vol. 145, pp. 324–335, 2014, doi: 10.1016/j.neucom.2014.05.026.

[149] S. M. Guzman, J. O. Paz, and M. L. M. Tagert, “The Use of NARX Neural Networks to Forecast Daily Groundwater Levels,” *Water Resour. Manag.*, vol. 31, no. 5, pp. 1591–1603, 2017,

doi: 10.1007/s11269-017-1598-5.

[150] A. Wunsch, T. Liesch, and S. Broda, “Forecasting Groundwater Levels using Nonlinear Autoregressive Networks with Exogenous Input ( NARX ),” *J. Hydrol.*, 2018, doi: 10.1016/j.jhydrol.2018.01.045.

[151] S. M. Guzmán, J. O. Paz, M. L. M. Tagert, A. E. Mercer, and J. W. Pote, “An integrated SVR and crop model to estimate the impacts of irrigation on daily groundwater levels,” *Agric. Syst.*, vol. 159, pp. 248–259, 2018, doi: 10.1016/j.agsy.2017.01.017.

[152] Y. Gong, Z. Wang, G. Xu, and Z. Zhang, “A comparative study of groundwater level forecasting using data-driven models based on ensemble empirical mode decomposition,” *Water (Switzerland)*, vol. 10, no. 6, 2018, doi: 10.3390/w10060730.

[153] I. Branch, P. Sciences, W. Island, S. Spring, N. Surface, and C. Division, “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis,” 1998.

[154] A. Stallone, A. Cicone, and M. Materassi, “New insights and best practices for the successful use of Empirical Mode Decomposition , Iterative Filtering and derived algorithms,” *Sci. Rep.*, pp. 1–15, 2020, doi: 10.1038/s41598-020-72193-2.

[155] T. Y. Chu and W. C. Huang, “Application of empirical mode decomposition method to synthesize flow data: A case study of Hushan Reservoir in Taiwan,” *Water (Switzerland)*, vol. 12, no. 4, 2020, doi: 10.3390/W12040927.

[156] W.-C. Huang, T.-Y. Chu, Y.-S. Jhang, and J.-L. Lee, “Data Synthesis Based on Empirical Mode Decomposition,” *J. Hydrol. Eng.*, vol. 25, no. 7, p. 04020028, 2020, doi: 10.1061/(asce)he.1943-5584.0001935.

[157] H. A. Simon, *The Sciences of the Artificial Third edition*. London, England: MIT Press, 1969.

[158] J. Venable, J. Pries-Heje, and R. Baskerville, “FEDS: A Framework for Evaluation in Design Science Research,” *Eur. J. Inf. Syst.*, vol. 25, no. 1, pp. 77–89, 2016, doi: 10.1057/ejis.2014.36.

[159] P. Johannesson and E. Perjons, *An introduction to design science*, vol. 9783319106. 2014. doi: 10.1007/978-3-319-10632-8.

[160] D. Schallmo, C. A. Williams, and K. Lang, “An Integrated Design Thinking Approach-Literature Review, Basic Principles and Roadmap for Design Thinking Digital Transformation of Business Models in the Israeli AgTech Landscape View project An Integrated Design Thinking

Approach-Literature Review, Bas," no. July, 2018, [Online]. Available: www.ispim.org.

[161] J. Holmström, M. Ketokivi, and A. P. Hameri, "Bridging practice and theory: A design science approach," *Decis. Sci.*, vol. 40, no. 1, pp. 65–87, 2009, doi: 10.1111/j.1540-5915.2008.00221.x.

[162] S. B. Mahmoud-Jouini, C. Midler, and P. Silberzahn, "Contributions of Design Thinking to Project Management in an Innovation Context," *Proj. Manag. J.*, vol. 47, no. 2, pp. 144–156, 2016, doi: 10.1002/pmj.

[163] C. Suero Montero and A. F. Kapinga, "Design science research strengthened: integrating co-creation and co-design," *IFIP Adv. Inf. Commun. Technol.*, vol. 551, pp. 486–495, 2019, doi: 10.1007/978-3-030-18400-1_40.

[164] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *J. Manag. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, 2007, doi: 10.2753/MIS0742-1222240302.

[165] ZAWA, "ZANZIBAR WATER AUTHORITY STRATEGIC BUSINESS PLAN ( 2013 - 2018 )," Zanzibar, 2018.

[166] Revolutionary Government of Zanzibar, "National Environmental Policy for Zanzibar," Zainzibar, 1992.

[167] E. Elisante and A. N. N. Muzuka, "Occurrence of nitrate in Tanzanian groundwater aquifers: A review," *Appl. Water Sci.*, vol. 7, no. 1, pp. 71–87, 2017, doi: 10.1007/s13201-015-0269-z.

[168] H. Mwevura, M. R. Haji, W. J. Othman, and C. J. Okafor, "Seasonal Assessment of Quality of Groundwater from Private Owned Wells in Unguja Island Zanzibar," *Int. J. Trop. Dis. Heal.*, no. April, pp. 30–45, 2021, doi: 10.9734/ijtdh/2021/v42i430449.

[169] E. Hansson, "Groundwater on Zanzibar - use and pollutants," 2010. [Online]. Available: bioenv.gu.se/digitalAssets/1322/1322530_erik-hansson.pdf

[170] R. A. M. M. Rubhera, "Groundwater quality degradation due to salt water intrusion in Zanzibar Municipality," *African J. Environ. Sci. Technol.*, vol. 9, no. 9, pp. 734–740, 2015, doi: 10.5897/ajest2015.1931.

[171] Z. P. Ali and M. J. Rwiza, "Assessment of the impact of groundwater pumpage on water supply sustainability in Zanzibar, Tanzania," *Environ. Earth Sci.*, vol. 79, no. 21, 2020, doi: 10.1007/s12665-020-09240-8.

[172] F. E. Colchester, H. G. Marais, P. Thomson, R. Hope, and D. A. Clifton, "Accidental

infrastructure for groundwater monitoring in Africa," *Environ. Model. Softw.*, vol. 91, pp. 241–250, 2017, doi: 10.1016/j.envsoft.2017.01.026.

[173] T. Jabeen, H. Ashraf, and A. Ullah, "A survey on healthcare data security in wireless body area networks," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 10, pp. 9841–9854, 2021, doi: 10.1007/s12652-020-02728-y.

[174] M. A. Al Imran, Y. Dalveren, B. Tavli, and A. Kara, "Optimal operation mode selection for energy-efficient light-weight multi-hop time synchronization in linear wireless sensor networks," *Eurasip J. Wirel. Commun. Netw.*, vol. 2020, no. 1, 2020, doi: 10.1186/s13638-020-01744-y.

[175] A. D. Wickert, C. T. Sandell, B. Schulz, and G. H. C. Ng, "Open-source Arduino-compatible data loggers designed for field research," *Hydrol. Earth Syst. Sci.*, vol. 23, no. 4, pp. 2065–2076, 2019, doi: 10.5194/hess-23-2065-2019.

[176] K. Chan *et al.*, *Low-cost electronic sensors for environmental research: Pitfalls and opportunities*, vol. 45, no. 3. 2021. doi: 10.1177/0309133320956567.

[177] T. P. Lyman, K. Elsmore, B. Gaylord, J. E. K. Byrnes, and L. P. Miller, "Open Wave Height Logger : An open source pressure sensor data logger for wave measurement," pp. 335–345, 2020, doi: 10.1002/lom3.10370.

[178] L. Jiang, L. Da Xu, H. Cai, Z. Jiang, F. Bu, and B. Xu, "An IoT-Oriented Data Storage Framework in Cloud Computing Platform," *Ind. Informatics, IEEE Trans.*, vol. 10, no. 2, pp. 1443–1451, 2014, doi: 10.1109/TII.2014.2306384.

[179] L. J. Chien, M. Drieberg, P. Sebastian, and L. H. Hiung, "A simple solar energy harvester for wireless sensor networks," *Int. Conf. Intell. Adv. Syst. ICIAS 2016*, 2017, doi: 10.1109/ICIAS.2016.7824104.

[180] G. Codeluppi, A. Cilfone, L. Davoli, and G. Ferrari, "LoraFarM: A LoRaWAN-based smart farming modular IoT architecture," *Sensors (Switzerland)*, vol. 20, no. 7, 2020, doi: 10.3390/s20072028.

[181] M. Cattani, C. A. Boano, and K. Römer, "An experimental evaluation of the reliability of lora long-range low-power wireless communication," *J. Sens. Actuator Networks*, vol. 6, no. 2, 2017, doi: 10.3390/jsan6020007.

[182] S. Oiro, J. C. Comte, C. Soulsby, A. MacDonald, and C. Mwakamba, "Depletion of groundwater resources under rapid urbanisation in Africa: recent and future trends in the Nairobi Aquifer System, Kenya," *Hydrogeol. J.*, vol. 28, no. 8, pp. 2635–2656, 2020, doi: 10.1007/s10040-

020-02236-5.

[183] R. . HEALY, “Groundwater resilience in sub-Saharan Africa,” *Nature*, vol. 572, pp. 185–186, 2019.

[184] A. van der Hoogen, B. Scholtz, and A. P. Calitz, “Using Theories to Design a Value Alignment Model for Smart City Initiatives,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12066 LNCS, pp. 55–66, 2020, doi: 10.1007/978-3-030-44999-5_5.

[185] F. Tauro *et al.*, “Measurements and observations in the XXI century (MOXXI): Innovation and multi-disciplinarity to sense the hydrological cycle,” *Hydrol. Sci. J.*, vol. 63, no. 2, pp. 169–196, 2018, doi: 10.1080/02626667.2017.1420191.

[186] Republic of Rwanda Ministry of Disaster Management and Refugee Affairs, “The National Disaster Management Policy Revision of the 2009 National Disaster Management Policy,” no. June, 2012.

[187] Rwanda Meteorological Agency, “Weather Data,” Kigali, Rwanda, 2018.

[188] National Institute of Statistics of Rwanda (NISR), “2012 Population and Housing Census (PHC) Brochure,” 2013. [Online]. Available: http://www.statistics.gov.rw/publication/rphc4-brochure

[189] G. Niyidufasha, “Groundwater Potential Eastern Province-Rwanda,” 2019.

[190] T. Kamai and S. Assouline, “Evaporation From Deep Aquifers in Arid Regions: Analytical Model for Combined Liquid and Vapor Water Fluxes,” *Water Resour. Res.*, vol. 54, no. 7, pp. 4805–4822, 2018, doi: 10.1029/2018WR023030.

[191] J. H. Zhao, Z. Dong, and Z. Xu, “Effective feature preprocessing for time series forecasting,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4093 LNAI, pp. 769–781, 2006, doi: 10.1007/11811305_84.

[192] “Python Release Python 3.6.6 | Python.org,” *Python366*, 2018. https://www.python.org/downloads/release/python-366/ (accessed Jul. 01, 2021).

[193] Scikit-Learn, “Scikit-Learn. Scikit-Learn 2.20,” *ScikitLearn*, 2018. https://www.scikit-learn.org/stable/install.html/ (accessed Feb. 19, 2019).

[194] S. D. Lindsey and R. K. Farnsworth, “Sources of solar radiation estimates and their effect on daily potential evaporation for use in streamflow modeling,” *J. Hydrol.*, vol. 201, no. 1–4, pp. 348–366, 1997, doi: 10.1016/S0022-1694(97)00046-2.

[195] M. Hocking and B. F. J. Kelly, “Groundwater recharge and time lag measurement through Vertosols using impulse response functions,” *J. Hydrol.*, vol. 535, pp. 22–35, 2016, doi: 10.1016/j.jhydrol.2016.01.042.

[196] C. Pappas, S. M. Papalexiou, and D. Koutsoyiannis, “A quick gap filling of missing hydrometeorological data,” *J. Geophys. Res. Atmos.*, vol. 119, pp. 9290–9300, 2014, doi: 10.1002/2014JD021633.Received.

[197] Y. Gao, C. Merz, G. Lischeid, and M. Schneider, “A review on missing hydrological data processing,” *Environ. Earth Sci.*, vol. 77, no. 47, 2018, doi: 10.1007/s12665-018-7228-6.

[198] H. Yoon, S. C. Jun, Y. Hyun, G. O. Bae, and K. K. Lee, “A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer,” *J. Hydrol.*, vol. 396, no. 1–2, pp. 128–138, 2011, doi: 10.1016/j.jhydrol.2010.11.002.

[199] D. N. Moriasi, M. W. Gitau, N. Pai, and P. Daggupati, “Hydrologic and water quality models: Performance measures and evaluation criteria,” *Trans. ASABE*, vol. 58, no. 6, pp. 1763–1785, 2015, doi: 10.13031/trans.58.10715.

[200] N. D. Bennett *et al.*, “Characterising performance of environmental models,” *Environ. Model. Softw.*, vol. 40, pp. 1–20, 2013, doi: 10.1016/j.envsoft.2012.09.011.

[201] X. Wen, J. Si, Z. He, J. Wu, H. Shao, and H. Yu, “Support-Vector-Machine-Based Models for Modeling Daily Reference Evapotranspiration With Limited Climatic Data in Extreme Arid Regions,” *Water Resour. Manag.*, vol. 29, no. 9, pp. 3195–3209, 2015, doi: 10.1007/s11269-015-0990-2.

[202] A. Ritter and R. Muñoz-Carpena, “Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments,” *J. Hydrol.*, vol. 480, pp. 33–45, 2013, doi: 10.1016/j.jhydrol.2012.12.004.

[203] W. Roberts, G. P. Williams, E. Jackson, E. J. Nelson, and D. P. Ames, “Hydrostats: A Python Package for Characterizing Errors between Observed and Predicted Time Series,” *Hydrology*, vol. 5, no. 4, p. 66, 2018, doi: 10.3390/hydrology5040066.

[204] W. Yang, “The Hydroclimate of East Africa: Seasonal cycle, Decadal Variability and Human-induced Climate Change,” Columbia University, 2015.

[205] J. Aboniyo, D. Umulisa, A. Bizimana, J. Marie, P. Kwisanga, and K. A. Mourad, “National Water Resources Management Authority for A Sustainable Water Use in Rwanda,” *Sustain. Resour. Manag. J.*, vol. 2, no. 3, pp. 1–15, 2017, doi: 10.5281/zenodo.801026.

[206] H. Tyralis and G. Papacharalampous, “A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources,” *Water*, vol. 11, no. 910, pp. 1–37, 2019.

[207] G. Biau, E. Scornet, and J. Welbl, “Neural Random Forests,” *Sankhya A*, no. 1984, pp. 1–40, 2018, doi: 10.1007/s13171-018-0133-y.

[208] A. Labarr, “How Good is That Forecast ? The Nuances of Prediction Evaluation Across Time,” pp. 1–7, 2018.

[209] T. G. Dietterich and Oregon, “Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Models,” *Oncogene*, vol. 12, no. 2, pp. 1–15, 2000.

[210] R. Cordeiro De Amorim and B. Mirkin, “Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering,” *Pattern Recognit.*, vol. 45, no. 3, pp. 1061–1075, 2012, doi: 10.1016/j.patcog.2011.08.012.

[211] P. Baronti, P. Pillai, V. W. C. Chook, S. Chessa, A. Gotta, and Y. F. Hu, “Wireless sensor networks: A survey on the state of the art and the 802.15.4 and ZigBee standards,” *Comput. Commun.*, vol. 30, no. 7, pp. 1655–1695, 2007, doi: 10.1016/j.comcom.2006.12.020.

[212] Y. Sun, D. Wendi, D. E. Kim, and S. Liong, “Technical note : Application of artificial neural networks in groundwater table forecasting – a case study in a Singapore swamp forest,” pp. 1405–1412, 2016, doi: 10.5194/hess-20-1405-2016.

[213] V. Moosavi, M. Vafakhah, and B. Shirmohammadi, “A Wavelet-ANFIS Hybrid Model for Groundwater Level Forecasting for Different Prediction Periods A Wavelet-ANFIS Hybrid Model for Groundwater Level Forecasting for Different Prediction Periods,” no. March, 2013, doi: 10.1007/s11269-012-0239-2.

[214] D. P. . Kingma and J. Ba, “Fixing weight decay regularization in Adam,” *arXiv 2014, arXiv:1412.6980*, pp. 1–4, 2018.

[215] H. Zhang, T. W. Weng, P. Y. Chen, C. J. Hsieh, and L. Daniel, “Efficient neural network robustness certification with general activation functions,” *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. NeurIPS, pp. 4939–4948, 2018.

[216] K. Y. Choy and C. W. Chan, “Modelling of river discharges and rainfall using radial basis function networks based on support vector regression,” *Int. J. Syst. Sci.*, vol. 34, no. 14–15, pp. 763–773, 2003, doi: 10.1080/00207720310001640241.

[217] P. Probst, M. N. Wright, and A. L. Boulesteix, “Hyperparameters and tuning strategies for

random forest," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 3, pp. 1–15, 2019, doi: 10.1002/widm.1301.

[218] T. Zhou, F. Wang, and Z. Yang, "Comparative analysis of ANN and SVM models combined with wavelet preprocess for groundwater depth prediction," *Water (Switzerland)*, vol. 9, no. 10, 2017, doi: 10.3390/w9100781.

[219] H. Yoon, Y. Hyun, K. Ha, K. Lee, and G. Kim, "A method to improve the stability and accuracy of ANN- and SVM-based time series models for long- term groundwater level predictions," *Comput. Geosci.*, vol. 90, pp. 144–155, 2016, doi: 10.1016/j.cageo.2016.03.002.

[220] P. Flach, "Machine Learning: The Art and Science of Algorithms That Make Sense of Data," *Cambridge Univ. Press*, 2012.

[221] O. Rahmati *et al.*, "Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods," *Sci. Total Environ.*, vol. 688, pp. 855–866, 2019, doi: 10.1016/j.scitotenv.2019.06.320.

[222] S. A. Naghibi, K. Ahmadi, and A. Daneshi, "Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping," *Water Resour. Manag.*, vol. 31, no. 9, pp. 2761–2775, 2017, doi: 10.1007/s11269-017-1660-3.

[223] T. Kavzoglu and P. M. Mather, "The use of backpropagating artificial neural networks in land cover classification," *Int. J. Remote Sens.*, vol. 24, no. 23, pp. 4907–4938, 2003, doi: 10.1080/0143116031000114851.

[224] C. Chen, W. He, H. Zhou, Y. Xue, and M. Zhu, "A comparative study among machine learning and numerical models for simulating groundwater dynamics in the Heihe River Basin, northwestern China," *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, 2020, doi: 10.1038/s41598-020-60698-9.

# APPENDIX A

## PUBLICATIONS

### A.1 Published Papers

The following is a list of published papers.

1. O. H. Kombo, S. Kumaran, and A. Bovim, E. Ndashimye, A. Bovim, “An Ensemble Mode Decomposition Combined with SVR-RF Model for Prediction of Groundwater Level: The Case of Eastern Rwandan Aquifers,” In: Silhavy, R. (eds) Cybernetics Perspectives in Systems. CSOC 2022. Lecture Notes in Networks and Systems, vol 503. Springer, Cham., 2022, https://doi.org/10.1007/978-3-031-09073-8_27

2. O. H. Kombo, S. Kumaran, and A. Bovim, “Design and Application of a Low-Cost, Low-Power, LoRa-GSM, IoT Enabled System for Monitoring of Groundwater Resources with Energy Harvesting Integration,” IEEE Access, vol. 9, pp. 128417–128433, 2021, doi: 10.1109/ACCESS.2021.3112519.

*3.* O. H. Kombo, S. Kumaran, Y. H. Sheikh, A. Bovim, and K. Jayavel, “Long-term groundwater level prediction model based on hybrid KNN-RF technique,” Hydrology, vol. 7, no. 3, pp. 1–24, 2020, doi: 10.3390/HYDROLOGY7030059.