TUNISIAN REPUBLIC
***

esprit
Ecole Supérieure Privée
d'Ingénierie et de Technologies

Engineering Internship Report

Artificial Intelligence Decision Support System for Groundwater Management under Climate Change: Application to Mornag Plain in Tunisia

Prepared By

Youssef TFIFHA

Supervised By

Dr. Nehla DEBBABI

And

Manel ENNAHEDH

Abstract

This paper aims to study the impact of climate change on groundwater levels in the Mornag plain in Tunisia. Indeed, in the last few decades, aquifers all over the world have experienced notable water level variability due to the spatiotemporal variability of rainfall and temperature. Therefore, for reliable groundwater management in a climate change context, it is mandatory to analyze and estimate its level variability. In this study, we focus on the plain of Mornag, located in the southeast of Tunisia, since it represents 33% of the national agricultural production. From this point, we have collected historical piezometric and pluviometric data covering the period 2005–2017. Knowing the pluviometric data, our goal is to predict the piezometric one. This issue has already been studied using classical numerical groundwater modeling such as Modflow and Feflow. Despite the fact that the results are unsatisfactory, these techniques are data and time-consuming. To address all of these shortcomings, we propose using two artificial intelligence (AI) approaches: the Extreme Gradient Boosting (XGBoost) approach, which has demonstrated excellent performance in the literature, and the well-known in our context, which involves the use of a Long-Short Term Memory (LSTM) Neural Network. For better results, we have added supplementary features to our dataset, such as the cluster zone (zones with the same characteristics) and the Standardized Precipitation Index (SPI), which can identify drought at different time scales. Both approaches have been executed entirely on the GPU for time acceleration. Compared with traditional existing methods, they have both shown a high level of accuracy, which confirms their adequacy for groundwater level forecasting. The proposed prediction models will be used for evaluating the repercussions of climate change on groundwater levels under the different scenarios RCP 4.5 and RCP 8.5 for the period of 2017–2090. It will be evaluated for three future periods: 2017–2040 (short term), 2041–2065 (medium term), and 2066–2090 (long term). The analysis of the future results using AI will be considered as a new decision support system used to optimize the management of our limited resources in order to satisfy the needs of the population in terms of drinking water and agricultural production.

Keywords

Artificial Intelligence, Climate Change, Decision Support System, Forecasting, Groundwater Level, Representative Concentration Pathway

Résumé

Cet article a pour objectif d'étudier l'impact du changement climatique sur les niveaux des eaux souterraines dans la plaine du Mornag en Tunisie. En effet, au cours des dernières décennies, les aquifères du monde entier ont connu une variabilité notable de leur niveau en raison de la variabilité spatio-temporelle des précipitations et de la température. Par conséquent, pour une gestion fiable des eaux souterraines dans un contexte de changement climatique, il est obligatoire d'analyser et d'estimer la variabilité de leur niveau. Dans cette étude, nous nous focalisons sur la plaine de Mornag, située dans le sud-est de la Tunisie, puisqu'elle représente 33% de la production agricole nationale. A partir de là, nous avons collecté des données historiques piézométriques et pluviométriques couvrant la période 2005-2017. Connaissant les données pluviométriques, notre objectif est de prédire les données piézométriques. Cette question a déjà été étudiée en utilisant la modélisation numérique classique des eaux souterraines, comme Modflow et Feflow. Malgré le fait que les résultats ne soient pas satisfaisants, ces techniques sont gourmands en données et prennent du temps. Pour remédier à tous ces défauts, nous proposons d'utiliser deux approches d'intelligence artificielle (IA) : l'approche Extreme Gradient Boosting (XGBoost), qui a démontré d'excellentes performances dans la littérature, et l'approche bien connue dans notre contexte, qui implique l'utilisation d'un réseau de neurones à mémoire à long et court terme (LSTM). Pour obtenir de meilleurs résultats, nous avons ajouté des caractéristiques supplémentaires à notre jeu de données, telles que la zone de regroupement (zones ayant les mêmes caractéristiques) et l'indice de précipitation standardisé (SPI), qui permet d'identifier la sécheresse à différentes échelles de temps. Les deux approches ont été exécutées entièrement sur le GPU pour l'accélération du temps. Comparées aux méthodes traditionnelles existantes, elles ont toutes deux montré un haut niveau de précision, ce qui confirme leur adéquation pour la prévision du niveau des eaux souterraines. Les modèles de prévision proposés seront utilisés pour évaluer les répercussions du changement climatique sur les niveaux des eaux souterraines dans les différents scénarios RCP 4.5 et RCP 8.5 pour la période 2017-2090. Il sera évalué pour trois périodes futures : 2017-2040 (court terme), 2041-2065 (moyen terme), et 2066-2090 (long terme). L'analyse des résultats futurs à l'aide de l'IA sera considérée comme un nouveau système d'aide à la décision utilisé pour optimiser la gestion de nos ressources limitées afin de satisfaire les besoins de la population en termes d'eau potable et de production agricole.

Mots clés

Intelligence artificielle, changement climatique, système d'aide à la décision, prévision, niveau des eaux souterraines, voie de concentration représentative.

Table of contents

List of Tables

List of figures

General Introduction

Any new information, fact, or skill obtained by an individual or group of persons via theoretical or practical comprehension of a subject is characterized as knowledge. Nonetheless, no dictionary or encyclopedia emphasizes that the most essential aspect of accumulating that information is to advance, teach others, and provide insight into what is to come. With everything going on in the world right now, from economic shortfalls to natural disasters, the only knowledge we require is how to ensure our survival, and understanding water resources is the key to doing so.

GroundWater (GW) is one of the most important and significant sources of water in the world, as it affects many facets of human life, such as industrial growth, agricultural production, drinking water provision, etc [1].

Unfortunately, over recent decades, aquifers all over the world have experienced significant GroundWater Level (GWL) volatility that makes water resource management challenging [2]. Two key factors are behind this volatility; the increase in water consumption and climate change. Indeed, due to the fast urban growth, the consumption of water resources has been dramatically increased. As a consequence, the balance between human and ecosystem needs has become difficult to maintain. On the other hand, climate change has played a crucial role in GWL variability, with mainly the spatiotemporal unpredictable changes of RainFall (RF) and Temperature [3]. As a matter of fact, according to UNESCO, GW is the sole way for 2.5 billion people throughout the world to meet their daily water demands [4], however, climate change endangers these resources constantly.

With full knowledge of the facts, a complete study of historical, current, and future GWL variability is required for policymakers and practitioners to develop water resource planning and management strategies, as well as prevent drought for the upcoming years. Both research and public interest in the projected climatic consequences on GW have intensified in recent years. Indeed, GWL has long been anticipated using a variety of numerically based conceptual models, including Feflow and ModFlow [5].

These models, developed by the United States Geological Survey (USGS), are widely considered the worldwide benchmark for predicting and forecasting GW. Though the non-linearity of these water systems and their response to climate conditions make modeling using numerical models difficult. In addition, these classical models are mostly data and time-consuming, hard to set up and maintain, and therefore expensive.

In recent years, the limitations of traditional numerical models have been widely addressed through the application of Artificial Intelligence (AI) models since they offer high accuracy with relatively less parametriza- tion. Among these models, we find the Long Short-Term Memory (LSTM) neural network model [6] that has shown great performance in literature when handling recurrent data, as for instance, for sign language translation [7], video prediction [8], or weather forecasting [9], to name a few.

In addition to this latter DL algorithm, Extreme Gradient Boosting (XGBoost), an efficient and scalable implementation of the gradient boosting framework, will be developed as well. Indeed, this Ensemble Learning (EL) model has gained a lot of notoriety in forecasting various climate related issues and has shown excellent results.

Since LSTM and XGBoost are well adapted to deal with sequential data, such as time series, we propose in this paper to use them to model the GWL of the Mornag plain in Tunisia. Corresponding to our region of interest, the Mornag plain is one of the most important plains in Tunisia since it contributes 33% of the population's needs in terms of drinking water and agricultural production, as well as to prevent upcoming drought.

We will utilize the CRoss-Industry Standard Process for Data Mining (CRISP-DM) methodology in this project because of its tremendous capability, adaptability, and utility when employing analysis to tackle a variety of hard situations. It's the golden thread that connects practically every client interaction. This model represents an idealized timeline of occurrences. Many jobs may be executed in a different order in practice, and it is frequently essential to return to earlier tasks and repeat some procedures. The model does not try to capture every potential path through the data mining process.

This report is divided into seven chapters: Chapter 1 will cover the overall presentation of the project and business understanding. The second chapter is devoted to data understanding and preparation . The third portion focuses on data modeling. The fourth section will cover the deployment of our models and their projection with the RCP data. The fifth section includes an examination of our findings. Finally, conclusions are drawn .

Chapter I : Business Understanding

1. Introduction

First, in this introductory chapter, we will clarify the basic concept and the broad context of the project through the presentation of the project, an analysis of the existing solution, the issue it poses, and the development of an alternative solution, as well as to identify key aspects that may impact the project's result . The methodology of the project will also be presented.

2. General Context

GW is water that exists underground in cracks and crevices in soil, sand, and rock. It is stored in and slowly travels through aquifers, which are geologic formations of soil, sand, and rocks. It is found practically everywhere. As a matter of fact, the water table can be deep or shallow, rising or falling based on a variety of conditions. Extensive RF or melting snow can raise the water table, whereas heavy pumping of GW supplies might lower it. Rain and snow melt that penetrates down into the cracks and fissures beneath the land's surface replenishes or recharges these water resources. Fig. 1 illustrates the GW circulation flow principals, which will be a key factor in the understanding of the data.

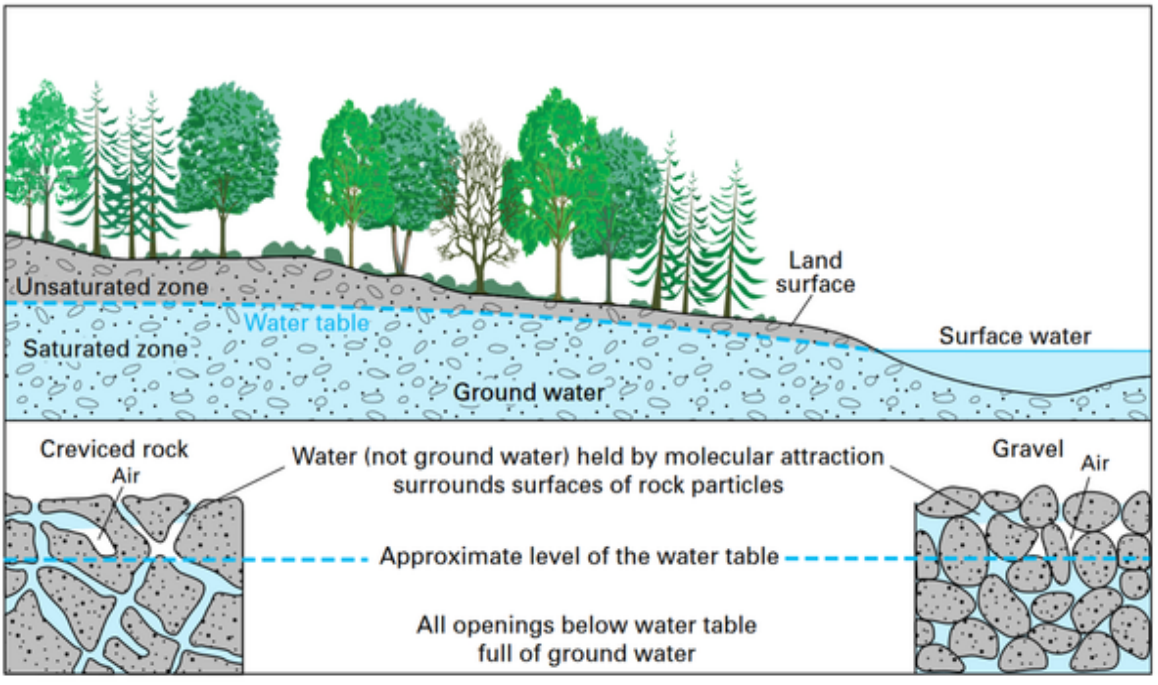


Fig. 1 Groundwater Circulation Principles in General

Indeed, GW is the world's greatest source of usable, fresh water. Domestic, agricultural, and industrial water demands in many regions of the world, particularly when surface water sources are unavailable, can only be satisfied by utilizing GW.

The USGS compares GW to money in a bank account. If money is withdrawn quicker than it is deposited, there will ultimately be account-supply issues. Consequently, pumping water out of the earth quicker than it is recharged generates comparable issues in the long run.

People in certain parts of the world experience severe water scarcity because groundwater is depleted quicker than it is restored naturally. Human activities contaminate groundwater in other regions. So our main goal is to forecast the level of subsurface water under changing climate conditions using automatic learning models.

Continuous GW pumping is the primary cause of groundwater depletion. Some of the negative consequences of GW depletion include:

- Water table subsidence: excessive pumping can decrease the GW table and prevent wells from reaching GW
- Cost increases: as the water table drops, water must be pushed farther to reach the surface, which requires more energy. In severe instances, utilizing such a well can be prohibitively expensive
- Surface water supply shortages: GW and surface water are inextricably linked. When groundwater is overused, the supply of lakes, streams, and rivers that are related to groundwater is reduced
- Subsidence of land: when there is a lack of support below the earth, land subsidence develops. When the soil collapses, compacts, and descends, it is most typically caused by human activity, most notably the misuse of GW
- Concerns about water quality: excessive pumping at the shore might cause saltwater to travel inland and upward, contaminating the water supply

As a result, for policymakers and practitioners to design water resource planning and management strategies in the coming years, a thorough understanding of groundwater levels (GWLs) in the past, present, and future variability is essential.

3. Goals and Objectives

As the context of the project becomes more clear, we need to set the goals and objectives of this research.

- Geostatistical analysis of climate and groundwater level data
- Understanding the relationship between climate data and groundwater levels based on historical data
- Developing models for groundwater level prediction based on AI methods
- Simulate groundwater level change under climate change scenarios

4. Region of interest

This study focuses on the Mornag plain which is located in northern Tunisia, 20 kilometers southeast of the capital Tunis. The study area climate is considered arid to semi-arid with moderate temperature. The annual RF is approximately 526mm [11]. As illustrated in Fig. 2 this plain is drained by two major rivers (Meliane and El Hma).

The surface area of the Mornag aquifer is about 200km². It stretches over 14 km from Tunis Gulf (Mediterranean Sea) in the North to the Khledia hills in the South. It is limited to the West by the Rades hills and its surroundings and to the East by the J. Rourouf mountain and its surroundings. Its hydraulic system consists of unconfined (GW lodged in recent Quaternary series) and confined aquifers (a deep aquifer which groups a series of 4 systems that occurred in ancient Quaternary, Oli- gocene, Miocene, and Eocene sediments). The aquifer system of the Mornag plain is characterized by the presence of the most dense observation system of GWL in Tunisia [12]. In this study, we focus on the unconfined aquifer. Thus, it is more and more exploited: in 2015, the exploitation rate reached 195 % with a deficit equal to -6.62 mm3/year [13]. This massive exploitation created an important piezometric depression and, consequently, an increase in the salinity of the studied groundwater [11].

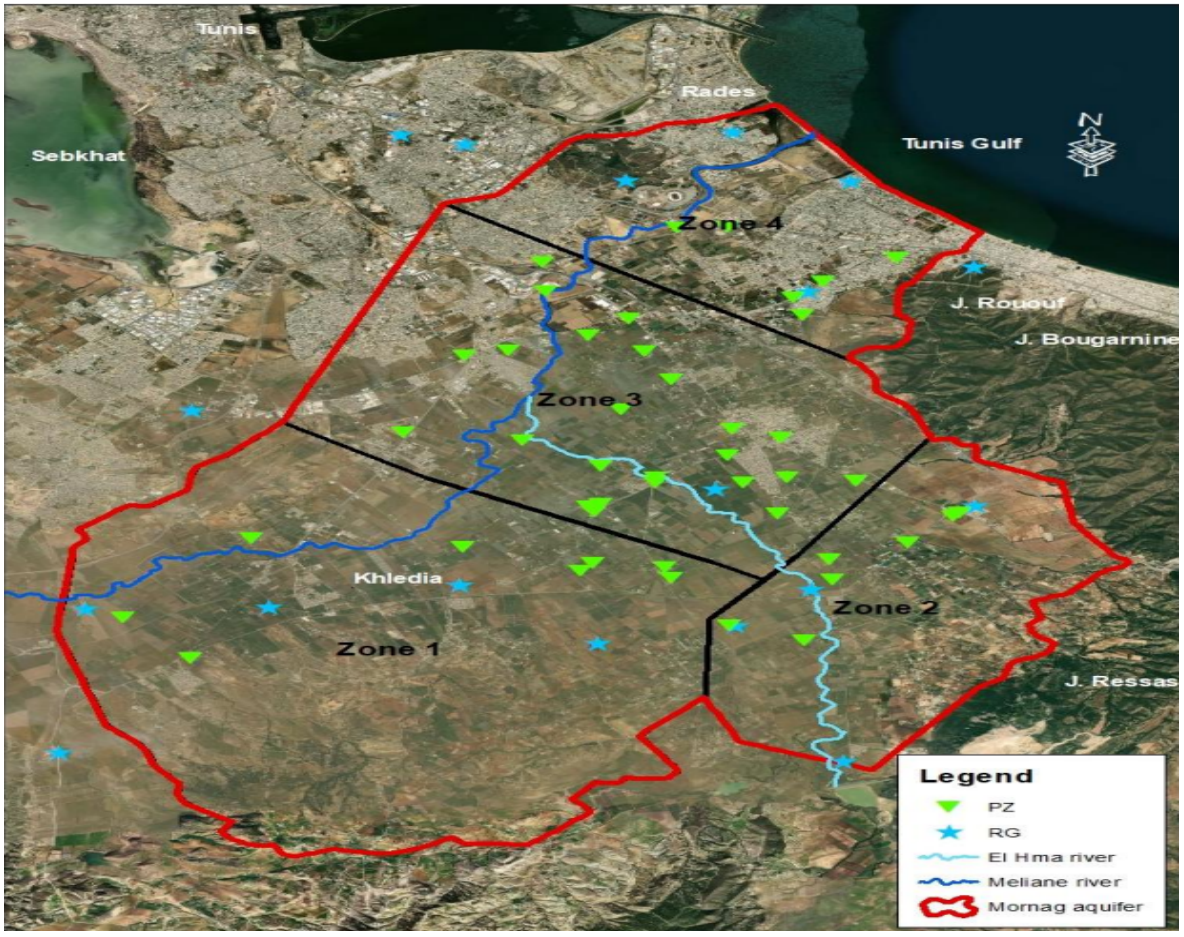


Fig. 2 Groundwater Circulation Principles in General

This aquifer is monitored by 44 piezometric stations and 18 pluviometric observation points, recording the GWL and RF, respectively, which allow hydrologists and researchers to better understand and investigate the Mornag aquifer system.

5. Methodology: CRoss-Industry Standard Process for Data Mining(CRISP-DM)

A hierarchical process model is used to describe the CRISP-DM technique. It contains the project stages, their associated tasks, and their outcomes. A data driven project's life cycle is divided into six parts, as shown in Fig 3 below. The stages are not in any particular order. The arrows represent just the most essential and common relationships between phases, but in a specific project, the conclusion of each phase determines which phase, or which specific job within a phase, must be completed next. The outer circle in Figure 3 represents the cyclical nature of the project.

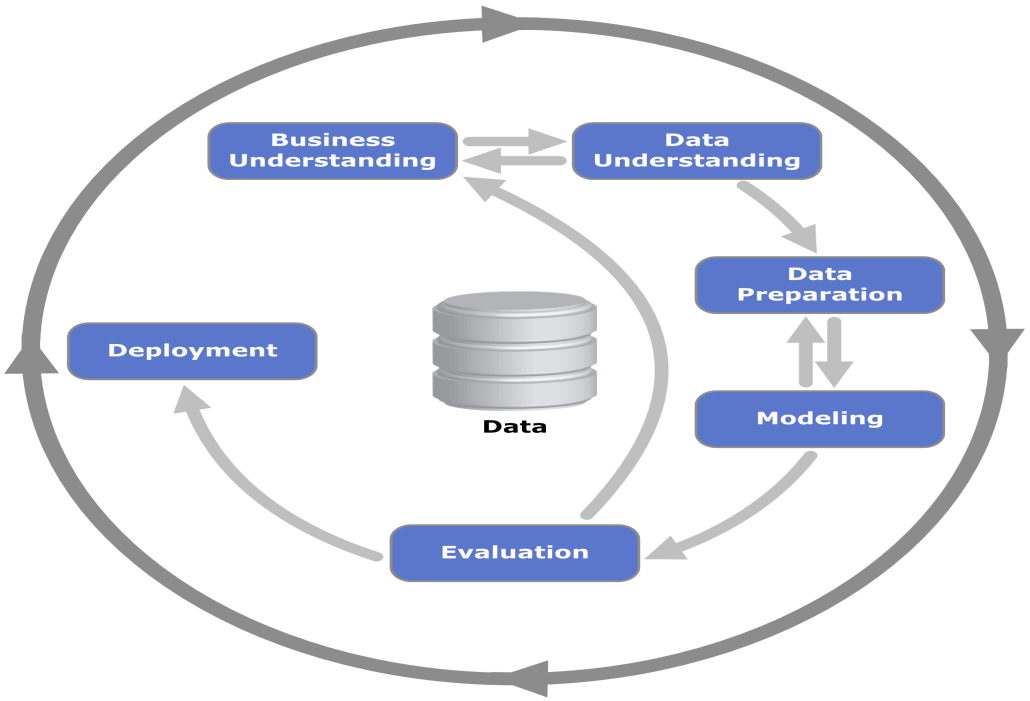


Fig. 3 Phases of the CRISP-DM Process Model

In the following, we outline each phase briefly:

- **Business Understanding**

This first phase focuses on understanding the project objectives and needs from a business standpoint and then transforming this information into a data mining issue definition and a preliminary project plan tailored to meet the goals.

- **Data Understanding**

The data understanding phase begins with data collection and continues with actions to become acquainted with the data, find data quality issues, uncover early insights into the data, or detect intriguing subsets to create hypotheses about hidden information. Business understanding and data understanding are inextricably linked. The formulation of the data mining challenge and the project strategy both need some knowledge of the available data.

- **Data Preparation**

The data preparation phase encompasses all operations that result in the final dataset, which will be supplied to the modeling tools being constructed from the original raw data.

Data preparation procedures are likely to be repeated several times, in no particular order. Attribute selection; data cleansing; attribute generation; and data transformation are all tasks and procedures that are generally present in every project .

- **Modeling**

Various modeling approaches are chosen and employed in this phase, and their parameters are calibrated to ideal levels. Typically, numerous strategies exist for the same data mining issue type. Some approaches necessitate the use of specialized data formats. Data preparation and modeling are inextricably linked. Often, when modeling, one discovers data difficulties or receives ideas for creating new data.

- **Evaluation**

Based on data analysis, at this point in the project, we have created one or more models that appear to be of high quality. Before proceeding with the final deployment of the model, it is critical to thoroughly examine the model and review the procedures used to develop the model to ensure that it meets the business objectives. One significant goal is to identify whether any critical business issues have been overlooked. A choice on how to use the data mining results should be made at the end of this step.

- **Deployment**

In most cases, the creation of the model is not the conclusion of the project. Typically, the acquired knowledge must be structured and presented in a way that the client can use.The deployment step might be as easy as creating a report or as sophisticated as establishing a repeatable data mining process, depending on the needs. In many circumstances, the user, not the data analyst, will perform the deployment processes. In any event, it is critical to understand what activities must be undertaken ahead of time in order to employ the produced models.

6. Business Success Criteria

A success criterion seeks to demonstrate that a project deliverable fulfills the needs of the company. These deliverables are subsequently used by project stakeholders to provide the benefits. Prior to or at the end of the project, success criteria might be set and measured.

- Map information from many study papers and obtain information on the relationships between entities, such as the seasonal effect of RF and the characteristics of the ground on GWL.
- Use revolutionary AI methodologies, ML and DL techniques to get a more accurate depiction of the GWL under climate change scenarios in the next few years than the standard numerical approach.

7. Data Mining Success Criteria

In this part, the project success criteria will be defined in technical terms. As with corporate success criteria, they may need to be described in subjective terms, in which case the person or people making the subjective assessment must be recognized.

- a successful use and understanding of the AI technology for GWL forecasting
- outperforming the numerical models in terms of accuracy
- Simulating GWL under different climate change scenarios

8. Tools

8.1. Python

For computer scientists, Python is the most extensively used open source programming language. This language has risen to prominence in infrastructure management, data analysis, and software development. Python, among other things, allows developers to concentrate on what they do rather than how they do it. It has liberated developers from the form limitations that plagued them in previous languages. As a result, creating code in Python is faster than in other languages.

8.2. Anaconda

Anaconda is a free and open source distribution of the Python and R programming languages for the creation of data science and machine learning applications (large-scale data processing, predictive analysis, scientific computing), with the goal of simplifying package management and deployment. The conda package management system manages package versions.

8.3. JupyterLab

JupyterLab is the most recent interactive development environment for notebooks, code, and data on the web. Users may create and arrange workflows in data science, scientific computing, computational journalism, and machine learning using its versatile interface. A modular design encourages additions to enhance and increase functionality.

9. Conclusion

This chapter explains the goal of the project, the success criteria, and the resources. To recap, it is the cornerstone of this project's next moves.

Chapter II : Data Understanding and Preparation

1. Introduction

This section aims to describe the Mornag plain dataset as well as the added features that will be used to enhance the GWL forecast.

2. Original Dataset Explorations

Originally, our dataset is composed of 285480 rows with 4 features; daily RF (RFd), Rain Gauge (RG), GWL, and PieZometer (PZ). Indeed, since 1971, using the 44 PZs, GWL data has been gathered twice a year, and from 2005 up to the present, RF has been measured daily using RGs. The collected dataset contains some missing values (8,8% for RFd and 4% for GWL), to manage them we have used a KNN imputation. As the GWL of the Mornag aquifer changes both geographically (by location and depth) and temporally (hourly, daily, seasonally, inter-annually, and at longer time scales), it will be helpful to use all of these factors to add new features for a better understanding of the GWL behavior. Fig 4 describes the process of GWL measurements.

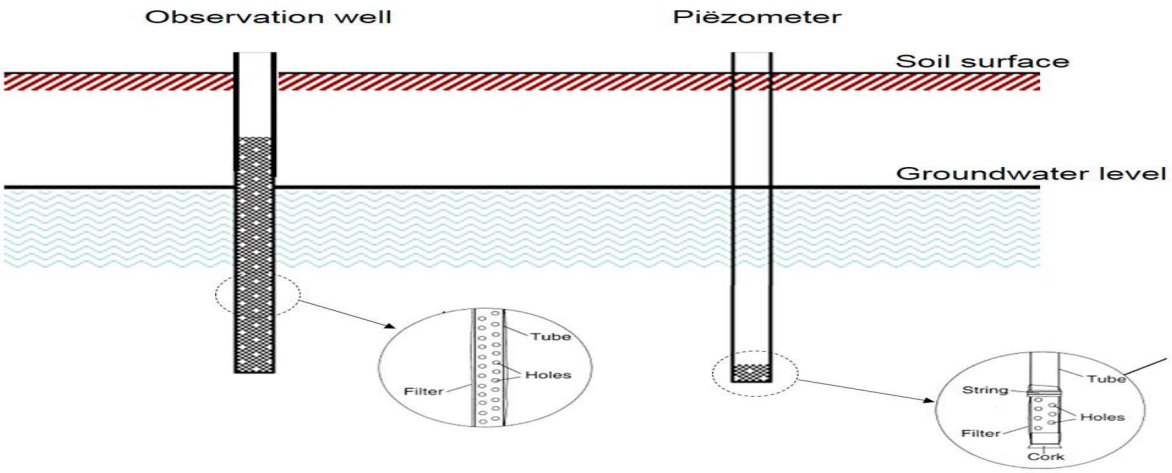


Fig. 4 GWL measurement process

Using querying, data visualization, and reporting approaches, we will answer data mining problems at this level.This step may also add to or improve the data description and quality reports, as well as feed into the transformation and other data preparation procedures required for future research.

Table 1. Table summarizing preliminary collected data.

	RG	RFd	Piezometer	GWL
Missing values	-	8.8%	-	4%
Frequency	-	daily	-	trimestrial
Type	Categorical	Continuous	Categorical	continuous

3. Data Preparation

One of the most critical and time-consuming components of data mining is data preparation. In fact, data preparation is projected to account for 50 to 70% of project time and effort. Devoting enough work to the earlier phases of knowing the company and understanding the data can help to shorten this step, but preparing and integrating the data for data mining still takes time. Indeed, our main issue is dealing with Missing Completely at Random (MCAR). When the missing variable is fully unsystematic, MCAR happens. When missing values occur at random in our dataset, the chance of missing data is unrelated to any other variable and unrelated to the variable with missing values itself. In order to handle missing data, we have used The KNN Imptuer, which is a distance-based imputation approach that necessitates data normalization. Otherwise, the KNN Imputer will create biased substitutes for missing values due to the varied scales of our data.

4. Feature Engineering and Extraction

This is the stage of the project when we must determine the data to use for the analysis. The relevance of the data to our data mining aims, the quality of the data, and technological restrictions such as data volume or data type limits can all be utilized to make this decision. It is important to note that data selection encompasses both the selection of characteristics (columns) and the selection of records (rows) in a table.

4.1 Named Zones

As the GWL of the Mornag aquifer changes both geographically (by location and depth) and temporally (hourly, daily, seasonally, inter-annually, and at longer time scales), it will be helpful to use all of these factors to add new features for better understanding of the GWL behavior. As a matter of fact, observing the GWL over the time, we have distinguished 4 different delimited zones (shown in Fig. 1.); each one clusters PZs with similar variability. Hence, the first added feature is named Zone which indicates the piezometric site. Fig.5 below describes the shape of Mornag aquifer while each color corresponds to a specific zone.

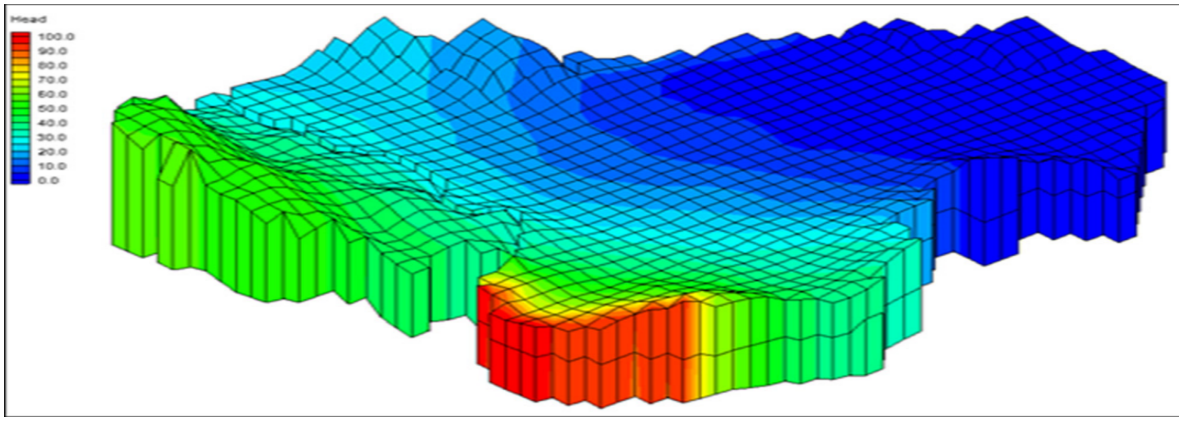


Fig. 5 3D shape of the Mornag aquifer model calibrated (via GMS).

4.2. Long and short term RF dependencies

Moreover, because of the strong association between GWL and both long-term and short-term RF levels [14], we have included the data seasonal influences such as; monthly RF (RFm), trimestrial RF (RFt), semestrial RF (RFm) and yearly RF (RFy). While seasonal climatic fluctuations are the most important driver of GWL variations, nearby surface water bodies, such as rivers, can also have an impact on groundwater levels.

Table 2. Summary of additional features.

	RF_d	RF_m	RF_t	RF_s	RF_y
Freq	daily	monthly	trimestrial	semestriel	yearly

Above in Table 2 an overview of the dataset with descriptions of seasonal characteristics can be found .

4.3. Standardized Precipitation Index (SPI)

Therefore, investigating the Standardized Precipitation Index (SPI) will yield important information about the wetness and dryness category described by the feature SPI-C [15]. Table 3 below summarizes each SPI category.

$$Let\ (RF_i)\ \ be\ the\ Daily\ rainfall\ of\ the\ day\ i,$$

$$RF_m\ \ be\ the\ Average\ rainfall\ of\ the\ series\ on\ the\ timescale\ considered$$

$$and\ S\ \ be\ the\ Standard\ deviation\ of\ the\ series\ on\ the\ timescale\ considered.$$

$$SPI = \frac{RF_i - RF_m}{S}$$

Table 3. SPI categories

	SPI > 2	1 < SPI < 2	0 < SPI < 1	- 1 < SPI < 0	-2 < SPI < -1	SPI < -2
Class Degree	extremely wet	very wet	moderately wet	moderately dry	severely dry	extremely dry

5. Conclusion

Finally, we consider as well Month and Year as two features since the chronological dependencies. As a summary, the new dataset is composed of 13 features with 4 categorical features; RG, PZ, SPI-C, Zone and 9 continuous features; RFd, RFm, RFt, RFs, RFy, SPI, Month, Year, GWL.

Chapter III : Data Modeling

1. Introduction

In the initial stage of modeling, we must choose the actual modeling approach to be used. Although a tool was chosen at the business knowledge step, the specific modeling approach will be chosen at this stage. Because numerous approaches will be used, we will complete this work independently for each methodology.

2. Artificial intelligence for groundwater level modeling

2.1 Long-Short Term Memory:LSTM

Long Short-Term Memory (LSTM) networks are an improved version of the traditional Recurrent Neural Networks (RNNs) [16] that are frequently used to handle sequential data, such as time series.

As stated in the literature, RNNs suffer from the vanishing gradient problem during backpropagation, where the gradient gets less and less with every layer, during the network training, until it is too small to reach the deepest levels. This drawback makes basic RNN memory unable to learn past information [6].

Alternatively, LSTMs came to recall long-term dependencies since they were specifically developed to address this issue. With LSTM, training errors retain their values, which overcomes the vanishing gradient problem and allows learning from sequences hundreds of timesteps long [6]. The initial LSTM model consists of a single hidden LSTM layer that could be composed of several Memory Blocks (MBs). A MB like depicted in Fig. 2. (B), which is composed of 3 sigmoid (σ) separate network layers and one hyperbolic tangent (Tanh) one, corresponds to the key contribution of the LSTM neural network, since the decision to consider or throw away information is taken inside. Indeed, the LSTM MB has three gates to govern the information flow; for a given time t, coupling the input x and the t output of the previous hidden state h , the forget gate regulates which and how much cell information is $t-1$ f t forgotten, the input gate i controls which inputs are used to update the old cell state (convoying important t information) c , into the new cell state and as for the output gate it defines which cell memory elements $t-1$ c t o t are used to update the hidden state h of the LSTM cell [16].

Using a unique LSTM model as initially developed, the past information stored inside a sequence is barely captured [6]. Alternatively, one could use the stacked LSTM which is a model expansion that involves several hidden LSTM layers, with many MBs for each of them, enhancing the ability to capture more complicated associations in the dataset. Based on experimentation, in this paper, we propose the stacked LSTM architecture illustrated in Fig. 2. (A) for which two LSTM layers are used, each one consists of 50 MBs.

As input, our stacked LSTM model uses 30 timesteps of the encoded sequence X1 to X12 corresponding to our 12 features. The stacked LSTM is followed by a dropout layer in order to reduce overfitting while improving model performance. Finally, we add a Fully-Connected (FC) layer giving rise to the output layer, denoted by Y and corresponding to the GWL in our study.

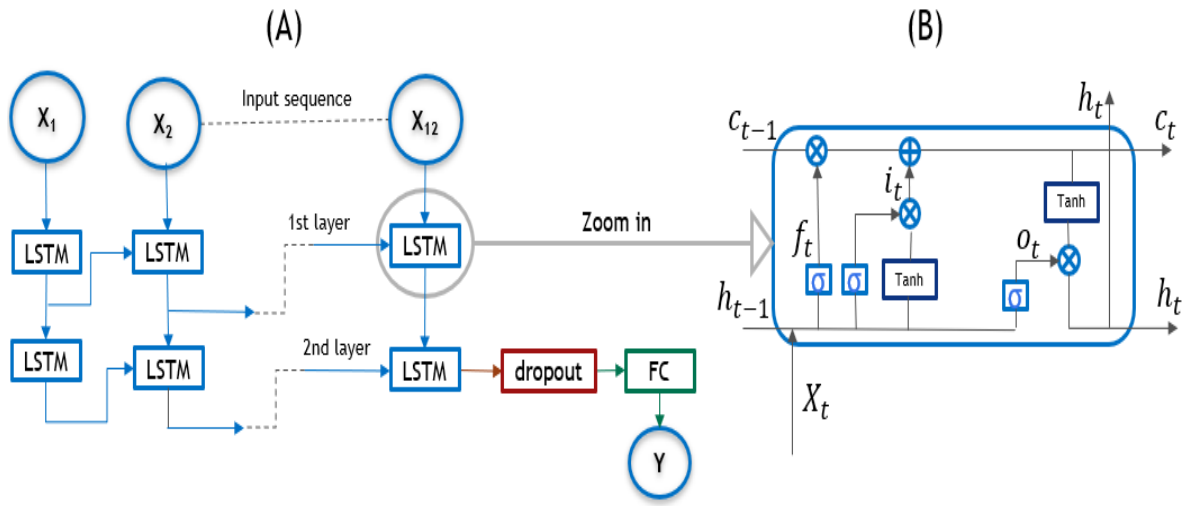


Fig. 6 Illustration of the proposed LSTM Network architecture for GWL forecasting

2.2. GPU Accelerated eXtreme Gradient Boosting (XGBoost)

It is an ensemble ML approach based on decision trees. XGBoost is a Gradient Boosted Decision Trees implementation in which we construct new models and integrate them into an ensemble. Unlike random forests, we construct trees one at a time, with each new tree contributing to the correction of faults committed by previously trained trees. By merging weak learners, the XGBoost reduces model residuals and boosts predictive power. The XGBoost algorithm is highly parallelizable by requiring scans across gradient values and using these partial sums to evaluate the quality of splits at every possible split in the training set. By utilizing fast parallel prefix operations to scan through all possible splits as well as parallel radix sorting to repartition data, the GPU accelerated version builds a decision tree for a given boosting iteration one level at a time, processing the entire dataset concurrently on the GPU.

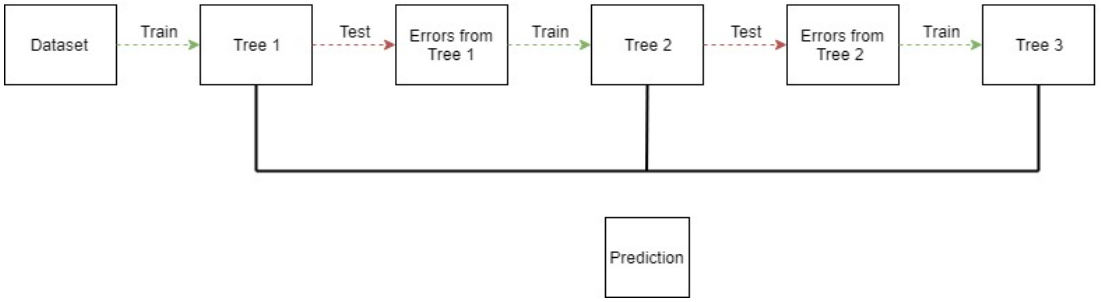


Fig. 7 XGBoost Architecture

To summarize, tree boosting is particularly effective because it uses adaptively determined neighborhoods to fit additive tree models with rich representational capabilities. Adaptive neighborhoods have the ability to apply varying degrees of flexibility in various parts of the input space. As a result, it will be capable of performing autonomous feature selection and capturing high-order interactions without crashing. As a result, it appears to be resistant to the curse of dimensionality. Individual trees are cleverly penalized by XGBoost. As a result, the trees might contain a variable number of terminal nodes. XGBoost can also use penalization to reduce leaf weights. The advantage of this approach is that not all leaf weights are lowered by the same factor, but leaf weights calculated using less evidence in the data are shrunk more significantly. Once again, the bias-variance tradeoff is considered during model fitting. Furthermore, XGBoost uses Newton boosting rather than gradient boosting. This increases the likelihood that XGBoost will learn better tree architectures. Because the tree structure determines the neighborhoods, XGBoost should aim for better neighborhoods. Finally, XGBoost can be demonstrated to learn better than any other model by employing a higher-order approximation of the optimization problem at each iteration.

3. Model Evaluation

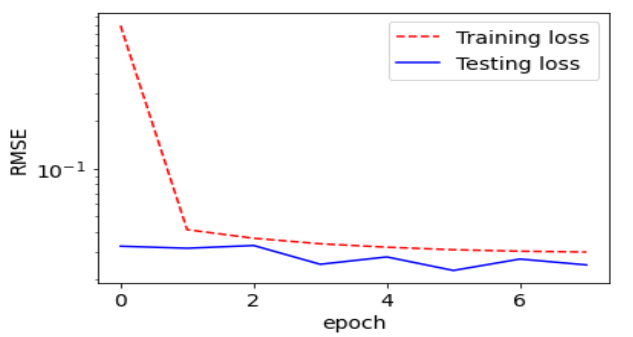
Both approaches using LSTM Neural Networks and GPU Accelerated XGBoost models have used the following hardware resources: GPU NVIDIA™ GeForce RTX 3050 TI and CPU AMD Ryzen 5 5600H with Radeon Graphics 3.30 GHz.

3.1 LSTM Evaluation

To train the proposed stacked LSTM model, we have used 70% of our dataset corresponding to the period from 04-2005 to 04-2011 and have kept 30% of data for the test phase, corresponding to the period from 05-2011 to 08-2015.

As a result of hyperparameter tuning, we have used the "Adam" optimizer which is a stochastic gradient descent technique that has shown great efficacy and resilience in modeling hyperparameters. We have set the corresponding batch size to 512 and have considered 8 epochs. For the dropout layer, we have considered a rate of 0.2 and as a loss function we have used the RMSE. Once trained, we have used the proposed stacked LSTM model to predict GWL using test data. In Fig. 8, we illustrate training and testing losses (in a log-scale) using RMSE. Both loss curves are decreasing and reaching a low point with a small gap of order 10 which −2 underlines a good fit quality.

Fig. 8 Validation and train losses.



Moreover, for each PZ, we represent the obtained RMSEs in Fig. 9.; results depict very acceptable values. Based on identical conditions and data, and over all PZs, we have in addition compared forecasting results using the proposed stacked LSTM and the Modflow one; with an RMSE of 0.85m, LSTM highly outperforms the Modflow, which has an RMSE of 6.9m [17]. The findings of this study highlights the use of AI techniques for Mornag GWL forecasting, which makes it a good decision support system for water resources management.

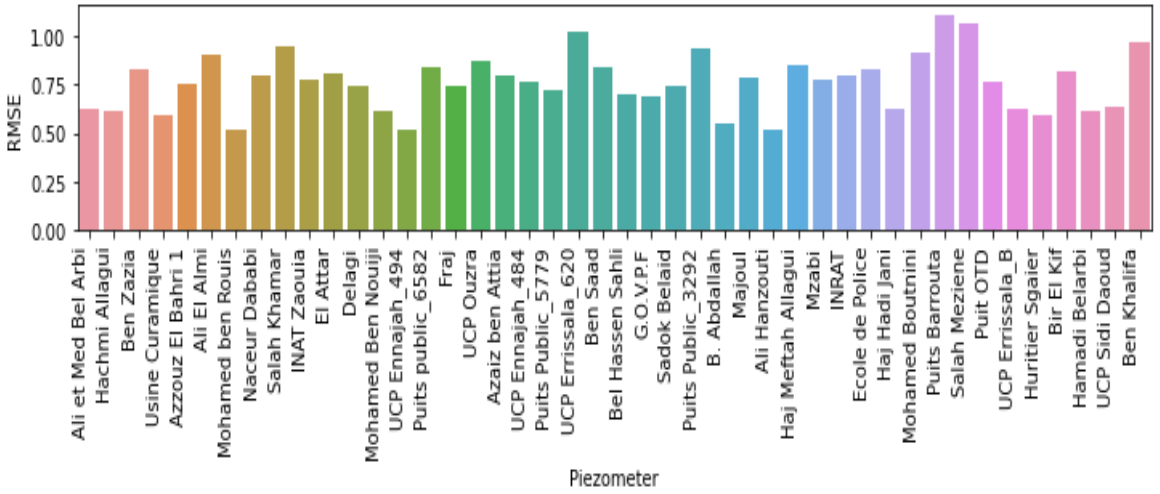


Fig. 9 RMSE of all piezometric stations using data from 2013 to 2015 (LSTM).

3.2 XGBoost Evaluation

XGBoostThe results of modeling the GWL based on the identical conditions and data demonstrate that XGBoost performed the best, with an RMSE of 0.13 compared to the LSTM, with an RMSE of 0.78m . The findings of this study also imply that IA techniques and algorithms may be utilized to estimate and forecast the GWL with great accuracy. This XGBoost approach was using the following parameters:

'objective': 'reg:squarederror' , 'eval_metric': 'rmse', 'booster' :'gbtree', 'tree_method': 'gpu_hist', "min_child_weight":4,'eta': 0.01, 'max_depth': 10, 'subsample': 0.7, 'colsample_bytree': 0.8, 'alpha':0.001, 'random_state': 42

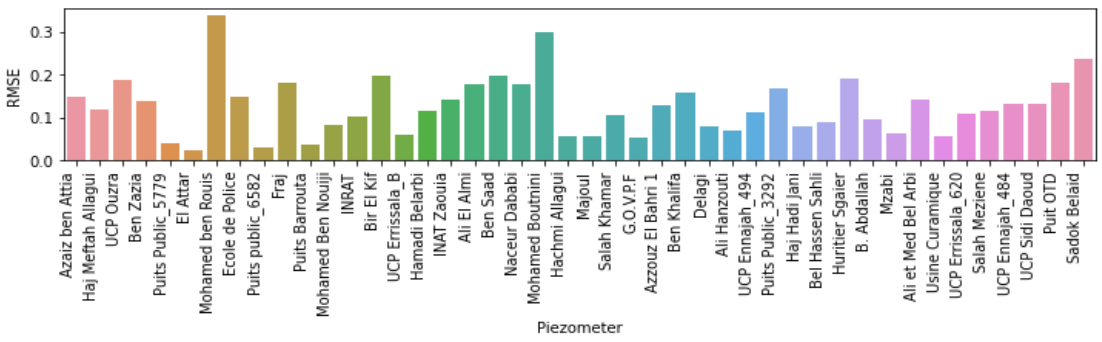


Fig.10 RMSE of all piezometric stations using data from 2013 to 2015 (XGBoost).

In Fig 10. above, we show the performance of the LSTM model by emphasizing the derived RMSE for different piezometric stations on the y-axis.Data-driven models and associated workflow models provide a fresh look at groundwater modeling that is valuable for scientists, policymakers, and water users. According to this viewpoint, this work investigates and compares the productivity of a data-driven modeling technique established with a DL architecture on a real-world case study as well as an operational workflow, to a typical process-driven methodology.

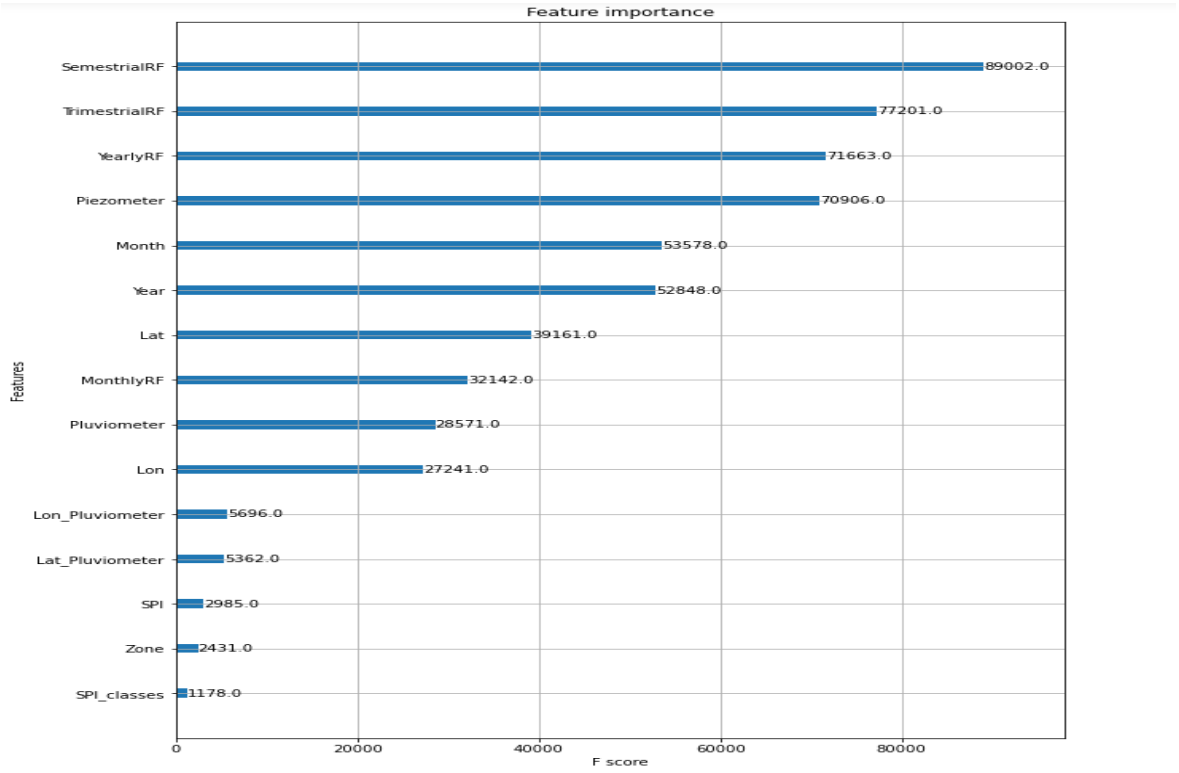


Fig. 11 Feature Importance

As discussed in the previous sections and according to fig.10 above, we can confirm the hypothesis that the present value of GWL is highly dependent on long term and short term RF values.

4. Conclusion

At this stage, we have assessed how well the model meets your business objectives and determined whether there is a commercial reason why the model is inadequate. Another alternative is to test the theories or the models on test applications in real-world applications, if time and money constraints allow. The evaluation step entails evaluating all of the other data exploration results that you have generated. The results of data exploration include models that are required to be related to the original commercial objectives, as well as other conclusions that are not required to be related to the original commercial objectives but may reveal challenges, information, or additional advice for future direction.

Chapter IV : Model Deployment

1. Introduction

Deployment is the process of integrating a machine learning model into an existing production environment in order to make real data-driven business choices. It is one of the final steps of the machine learning life cycle and one of the most time-consuming.

2. Representative Concentration Pathway

The forecasting will be applied using both the Representative Concentration Pathway (RCP) 4.5 and 8.5, which are two of the scenarios of radiative forcing trajectories to the year 2100 that were established by the Intergovernmental Panel on Climate Change (IPCC) . These scenarios are not only unpredictable because of restrictions such as the selection of future virtual scenarios, an imperfect physical knowledge of numerous self-connections, and computing capabilities, but also because of the geospatial variability. Different RCP models have been established and validated, but we can distinguish that some of those models are highly optimistic and others are very pessimistic. In previous research and applications, they were corrected using the mean values of the different models. The Fig. 11 illustrates a sample of the RCP 4.5 data.

date	2090-04-01	2059-12-01	2045-09-01	2093-11-01	2063-10-01
Piezometer	UCP Errissala_620	Delagi	Majoul	Bel Hassen Sahli	INRAT
Pluviometer	MORNEG FERME ESSADIR	KHELIDIA CTV	BOUMHEL BASSATINE MU	FOUCHANA FERME GAMOU	MORNAG SIDI ZEYED
YearlyRF	351.766364	346.14	357.683636	442.995455	330.992727
SemestrialRF	147.888182	146.031818	160.05	170.434545	174.249091
TrimestrialRF	63.953636	96.971818	48.320909	128.039091	130.078182
MonthlyRF	4.958537	25.906645	6.638738	43.973034	48.012949
Mean	7.746364	36.991818	13.871818	45.128182	40.567273
Zone	3	1	4	3	2
SPI	-1.40697	0.289634	-1.051617	0.761646	0.497056
SPI_classes	Severely dry	Moderately Wet	Severely dry	Moderately Wet	Moderately Wet
Lat	36.667605	36.631954	36.736674	36.681987	36.628555
Lon	10.306164	10.166459	10.273134	10.282895	10.281398
Lat_Pluviometer	36.66806	36.63694	36.72167	36.69056	36.63111
Lon_Pluviometer	-10.27917	-10.19417	-10.29778	-10.18028	-10.2825
Month	4	12	9	11	10
Year	2090	2059	2045	2093	2063

Fig. 12 Sample of RCP 4.5 CC scenario

Given that the models above were moderated for a single piezometric well (Tunis-Carthage Station), which is 20 kilometers from the nearest Morang station, we decided to define our

own model using the linear regression algorithm on the various RCP models and compare the results to the actual data we have. Indeed, the Rades Ouafa station, utilizing a mix of RCP models, has been shown to be the closest station to the projected values of CC scenarios, yielding a considerably more accurate result than the mean technique.

Dep. Variable:	RADES OUAFA	R-squared (uncentered):	0.808
Model:	OLS	Adj. R-squared (uncentered):	0.768
Method:	Least Squares	F-statistic:	20.19
Date:	Mon, 08 Aug 2022	Prob (F-statistic):	5.74e-14
Time:	16:10:24	Log-Likelihood:	-246.65
No. Observations:	58	AIC:	513.3
Df Residuals:	48	BIC:	533.9
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
MIROC-ESM	0.2422	0.115	2.100	0.041	0.010	0.474
CNRM-CM5	0.1303	0.101	1.289	0.204	-0.073	0.333
CanESM2	-0.1064	0.120	-0.887	0.379	-0.347	0.135
FGOALS-s2	-0.0688	0.117	-0.588	0.559	-0.304	0.166
BNU-ESM	0.1350	0.163	0.829	0.411	-0.193	0.463
MIROC5	0.0378	0.136	0.278	0.782	-0.235	0.311
GFDL-ESM2G	-0.0436	0.112	-0.389	0.699	-0.269	0.182
MIROC-ESM-CHEM	0.0478	0.121	0.395	0.694	-0.196	0.291
GFDL-ESM2M	0.2823	0.091	3.112	0.003	0.100	0.465
MRI-CGCM3	0.0971	0.055	1.753	0.086	-0.014	0.208
MRI-CGCM3.1	0.0971	0.055	1.753	0.086	-0.014	0.208

Fig. 13 Linear Regression results

$$Let\ R^2\ be\ the\ coefficient\ of\ determination,$$
$$RSS\ be\ the\ sum\ of\ squared\ residuals\ and$$
$$TSS\ be\ the\ total\ sum\ of\ squares$$
$$We\ have\ R^2 = 1 - \frac{RSS}{TSS}$$
$$Let\ N\ be\ the\ total\ sample\ size\ and$$
$$P\ be\ the\ number\ of\ independent\ variables$$
$$P\ be\ the\ number\ of\ independent\ variables$$

$Then\ we\ have\ Adjusted.\ R^2 = 1 - \frac{(1-R^2)(N-1)}{N-P-1}$ $an\ adjusted\ coefficient\ of\ determination\ that\ consi$
$independent\ variables\ which\ actually\ have\ an\ effect\ on\ the\ performance\ of\ the\ model.$

As demonstrated above in Fig. 12, this station has the lowest Akaike information criterion (AIC) and the greatest adjusted R-Squared value . We can also see that the p values of several of the models are much lower than 0.05, indicating that they are very significant. The next phase is to anticipate the new CC scenarios using the new model in order to gain a more precise understanding and give decision assistance to policymakers.

3. Simulating GWL using RCP 4.5 and 8.5

For our simulated results up to 2100, GW resources in the Mornag aquifer were affected by climate change due to a decline in natural recharge from reduced precipitation (the mean will be 19.42% less at the end of the century for RCP 4.5;and 44.86% less for RCP 8.5). As Fig. 13. shows, the absolute variations in GWL under RCP 4.5 and RCP 8.5 may seem small (between 1 and 5 m), but the fact that we are investigating a shallow aquifer (thickness between 30 and 50m) reinforces the importance of results in terms of water availability for vegetation and agriculture. A drop of tens of centimeters (depending on the thickness of the aquifer) can be vital for plants during hot and dry periods, if therefore GW is no longer accessible [18].

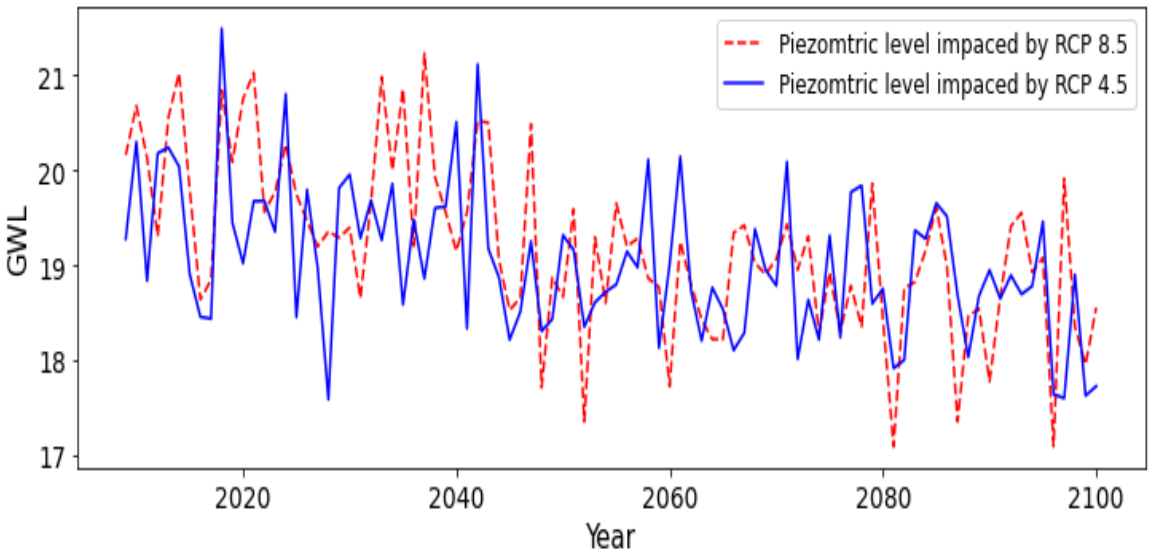


Fig. 13 Linear Regression results

This decrease in GWL presents a great concern for the future of irrigated agriculture in the study area as some farms would be abandoned due to GW unavailability. Nevertheless, there is an urgent need for adaptation measures that take into account these impacts of climate change on GW resources, in particular to improve productivity in the agriculture sector, such as an extensive reconversion of gravity irrigation to drip irrigation and adapted crops that are water efficient and more resilient to climate change.

4. Findings of the analysis

The following are the findings of the analysis:

- Among the published state-of-the-art research and critical evaluation of AI vs numerical models, LSTM has emerged as one of the most promising and popular techniques for modeling and simulating complicated hydrogeological systems.
- Considerations such as correlation analysis, statistical analysis, and research area characteristics must be included in the selection of input parameters.
- Prior to the GWL investigation under CC situations, precipitation rose relatively to

historical levels in all scenarios and assessment periods.

- Extreme precipitation is anticipated to increase globally as temperatures rise, in contrast to normal yearly precipitation.
- GWL declined gradually in all four cluster zones owing to the increase in evapotranspiration induced by the temperature rise and GW usage increase.
- The predicting findings for RCP 4.5 and 8.5 demonstrate that the GWL trend is diminishing over time due to the huge fall in RF.
- As a further step in our work, we will simulate each RCP model independently to emphasize the influence of climate change on GWL.

5. Conclusion

Finally, if the results of data mining become part of everyday business and its surroundings, monitoring and maintenance become critical challenges. A well-planned maintenance approach can assist in avoiding unduly extended periods of inappropriate usage of data mining findings. A specific monitoring process plan is required for the project to monitor the deployment of the data mining findings. This strategy considers the unique type of deployment.

General Conclusion

The Mornag plain has already faced water scarcity due to anthropogenic activities such as over-pumping and climate change conditions as, during the last few decades, an increasing number of drought years has been observed. Therefore, the present study focuses on the assessment of the pressure that climate change will impose on GWL in the future. However, an Artificial Intelligence Decision Support System was conducted, while the Standardized Precipitation Index SPI was adopted, for the first time in Mornag plain, in order to identify climate change impacts on groundwater resources, at the same time, explore the use of SPI as an indicator to predict GW responses to climate conditions and the use of the AI modeling techniques in designing and planning GW management. The elaborated intelligent model LSTM has shown significant results in GWL prediction with an RMSE below 1m. In addition, the main finding of the assessment of climate change impacts indicates that the predicted hydrological drought events will affect the water table fluctuation in the medium and long term with a drawdown up to 5m. Thus, these results are of great importance as key information for decision-makers regarding the future of the sustainable exploitation of groundwater resources in the aquifer.

References

[1] Vaux, H.. "Groundwater under stress: the importance of management." Environmental Earth Sciences 62.1 (2011): 19-23.

[2] Anand, B., et al. "Long-term trend detection and spatiotemporal analysis of groundwater levels using GIS techniques in Lower Bhavani River basin, Tamil Nadu, India." Environment, Development and Sustainability 22.4 (2020): 2779-2800.

[3] Jeppesen, E., et al. "Ecological impacts of global warming and water abstraction on lakes and reservoirs due to changes in water level and related changes in salinity." Hydrobiologia 750.1 (2015): 201-227.

[4] UNESCO. "World's groundwater resources are suffering from poor governance." UNESCO Natural Sciences Sector News (2012).

[5] Alloisio, S. A. R. A. H., et al. "Groundwater modeling for large-scale mine dewatering in Chile: MODFLOW or FEFLOW." Water Management Consultants. Chile (2004).

[6] Zhang, J., Zeng, Y., & Starly, B. (2021). Recurrent neural networks with long term temporal dependencies in machine tool wear diagnosis and prognosis. SN Applied Sciences, 3(4), 1-13.

[7] Guo, D., Zhou, W., Li, H., & Wang, M. (2018, April). Hierarchical lstm for sign language translation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).
[8] Wang, Y., Jiang, L., Yang, M. H., Li, L. J., Long, M., & Fei-Fei, L. (2018, September). Eidetic 3D LSTM: A model for video prediction and beyond. In International conference on learning representations.

[9] Karevan, Z., & Suykens, J. A. (2020). Transductive LSTM for time-series prediction: An application to weather forecasting. Neural Networks, 125, 1-9.

[10] Van Vuuren, D. P., et al. "The representative concentration pathways: an overview." Climatic change 109.1 (2011): 5-31.

[11] Zaafouri, A. (2020). Impact de l'urbanisation sur la zone humide de sebkha Ariana et dynamique sédimentaire du littoral (Golfe de Tunis) (Doctoral dissertation, Université de Sfax. École Nationale d'Ingénieurs de Sfax-ENIS).

[12] Horriche, F. (2004). Contribution a l'analyse et a la rationalisation des reseaux piezometriques. These de Doctorat, Universite Tunis El Manar, ENIT.

[13] Menani, M (2015). Evaluation du risque de conflit autour des eaux trasnfrontalieres du systeme aquifere du sahara septentrional (SASS). Larhyss Jouranl P-ISSN 1112-3680/E-ISSN 2521-9782, (22),59-59

[14] Ahmadi, S. H., & Sedghamiz, A. (2007). Geostatistical analysis of spatial and temporal variations of groundwater level. Environmental monitoring and assessment, 129(1), 277-294.

[15] Cancelliere, A., Mauro, G. D., Bonaccorso, B., & Rossi, G. (2007). Drought forecasting using the standardized precipitation index. Water resources management, 21(5), 801-819.
[16] Bianchi, F. M., Maiorino, E., Kampffmeyer, M. C., Rizzi, A., & Jenssen, R. (2017). Recurrent neural networks for short-term load forecasting: an overview and comparative analysis.

[17] Ennahedh, M., Hariga-Tlatli, N., Tarhouni, J., Hydrogeological modeling for the aquifer system of the Mornag plain (Tunisia) for future real-time management.3rd conference of the Arabian Journal of Geosciences

[18] Wunsch, A., Liesch, T. & Broda, S. Deep learning shows declining groundwater levels in Germany until 2100 due to climate change. Nat Commun 13, 1221 (2022).

Glossary

Data Science DS : a multidimensional, multidisciplinary field of research that employs a variety of scientific approaches, sophisticated analytics techniques, and predictive modeling algorithms to extract relevant insights from data in order to assist in answering key business or scientific problems in a variety of fields. It combines a wide variety of technical and non-technical abilities and typically necessitates extensive domain knowledge in the sector in which it is used in order to appropriately analyze the given data and the resulting conclusions.

Artificial Intelligence AI : a subfield of computer science that uses machine learning, programming, and data science approaches to teach computers to behave intelligently. AI systems are diverse, with differing degrees of sophistication. They can range from rule-based systems to machine learning-based systems and can identify fraud, recognize objects, translate languages, predict stock prices, and much more.

Deep Learning DL : a subset of machine learning methods based on multilayered artificial neural networks (ANN) influenced by brain structure ANNs are extremely adaptable and can learn from massive quantities of data to produce highly accurate results. They are frequently at the heart of several data science and machine learning use cases, such as picture or sound recognition, language translation, and other complex issues.

Machine Learning ML : a subfield of artificial intelligence (AI) that offers a collection of algorithms for learning patterns and trends from historical data. Without being explicitly programmed, the goal of ML is to predict future events and generalize beyond the data points in the training set. There are two types of machine learning algorithms: supervised and unsupervised, each with a variety of strategies suited to various use situations.

Time Series : A time series is a collection of observations of a variable recorded at various times and sorted in time order. Time series measurements are typically made at equal intervals in time. Stock market prices or temperature over a specific time period are examples of time series.

Regression : Regression is a supervised learning problem in which continuous outcomes must be predicted based on input characteristics. A regression model learns the connection between one or more independent characteristics and the target variable, and then utilizes the developed function to predict data that has not yet been observed. Linear regression and ridge regression are two examples of regression algorithms. Price prediction is a common regression problem.

Feature : a non-linear variable that is used as an input in a machine learning model.

Dataframe : a tabular data structure with designated axes (rows and columns) that can be of many forms.

Feature Engineering : Feature engineering is the process of transforming raw characteristics into features that better reflect the underlying problem and are more suitable for machine learning algorithms by utilizing domain knowledge and subject matter expertise.

Gradient descent : an iterative optimization approach used in machine learning to minimize the cost function by determining the best values for the function's parameters.

Imputation : The technique of filling in missing values in a dataset is known as imputation. Imputation methods might be statistical (mean/mode imputation) or machine learning-based (KNN imputation).

Root Mean Squared Error (RMSE) : The square root of the mean squared error is the root mean squared error (RMSE). This assessment metric is more intuitive than MSE since the result is easier to understand when measured in the same units as the original data.

Model tuning : the practice of modifying hyperparameters to enhance model accuracy while avoiding overfitting.

Hyperparamaters : properties of a machine learning model that are explicitly specified before the training process begins. Hyperparameters, unlike other parameters, cannot be calculated or learnt directly from data. We can identify the ideal values for hyperparameters by tuning them and assessing the resulting model performance. Setting a hyperparameter intuitively is similar to tuning a radio knob to achieve a perfect signal.

Scikit-Learn : Python's most useful and stable machine learning library. It offers a set of fast tools for machine learning and statistical modeling, such as classification, regression, clustering, and dimensionality reduction, via a Python interface.

Comma Separated Values CSV : a phrase used to describe a computer file that contains tabular data shown in plain text. It's set up such that the data may be imported into tables later on, such as in Excel. As such, it is frequently seen as a simple type of spreadsheet.