Water table analysis using Machine Learning

Aishwarya Kulkarni, Shivangi Negi, Sumedha Raghu
Dr. Vijaya Shetty S.
Nitte Meenakshi Institute of Technology

Abstract

Groundwater is commonly the most important water resource in semi-arid areas. Monitoring water table fluctuations is essential for predicting the groundwater levels to plan for future needs. In this study, a thorough analysis is conducted concerning the prediction of groundwater levels in Karnataka. Two nonlinear data-driven models (i.e; Random Forest (RF) and gradient boosting (GB)) along with a variant of standard RNNs (Long Short-Term Memory - LSTM) were proposed to predict groundwater level fluctuations. The prediction capability of these models was investigated and evaluated using yearly groundwater level data collected from observation wells located in various districts of Karnataka in India. The statistical parameters Correlation coefficient (R), Mean Square Error (MSE), precision, recall, f1 score, support, and accuracy were used to assess the performance of these models. Different evaluation metrics highlight the capability of these models to catch the trend of groundwater level fluctuations.

Keywords: groundwater levels; machine learning; data-driven; random forest; gradient boosting

I INTRODUCTION

Groundwater is defined as the water present in the areas between the soil pore spaces and within the fractures of rock formations. The depth at which soil pore fractures and voids in rock become saturated with water is defined as the water table [1]. In many areas, groundwater has emerged as an important source of water required for domestic, irrigation, urban, and industrial activities, especially in arid and semi-arid areas. India being the largest user of groundwater in the world uses 230 cubic kilometers of groundwater per year, over a quarter of the global total [2]. For this reason, sustainable development of groundwater resources requires precise quantitative assessment and this is vital for India due to prevalent semi-arid and arid climate. Different weather conditions and usage rates are essential for the effective utilization and management of groundwater resources. Constant monitoring of the groundwater levels is important to prevent the misuse of groundwater resources that can lead to local water rationing, excessive reductions in agricultural yields, wells going dried up or producing erratic groundwater quality changes, changes in flow patterns of groundwater resulting in the inflow of poorer quality water and seawater intrusion in coastal areas [3]. The water levels, if forecasted well prior to, might facilitate the administrators to plan better, the groundwater utilization. In this research, we focus on observations wells from various districts of Karnataka.

Traditionally, process-based models are often employed to perform groundwater simulation and predications, which rely on spatial data of the observed system dynamics. However, they are not applicable in many arid and semi-arid regions due to data limitations. On the other hand, in data-driven modeling with machine learning techniques, our model attempts to identify a direct mapping between the inputs and outputs of the system without reaching an understanding of the internal structure of the physical process [4]. The goal here is to predict groundwater levels based on temporal data inputs (historic groundwater level, weather, and rainfall data) and outputs (groundwater level).

Recurrent neural networks (RNNs) are a popular choice for modeling groundwater time series data as they can retain a memory of past network conditions, but they face difficulty in capturing long term dependencies between variables due to exploding and vanishing gradient, where weights in the network go to zero or become extremely large during model training [5]. LSTM, a type of RNN is able to avoid these training problems by eliminating unnecessary information being passed to future model states while retaining a memory of important past events. LSTM networks have also recently been used to model the groundwater table on a monthly time step in an inland agricultural area of China [6]. This agriculture-focused

study provides valuable information on the advantages of LSTM for groundwater level prediction over a basic feed-forward neural network. Groundwater level prediction is a regression problem. Based on available data we are trying to generate the best possible predictions for the groundwater level for the missing values in our dataset using binary classification. Data-driven algorithms use historical data to learn the best approximation of an underlying process. Despite the growing applications of data-driven approaches in the surface water problems, there have been only a few studies related to groundwater in arid and semi-arid regions [4]. The focus of this study is the application and comparison of two data-driven models (i.e., RF and GB) for forecasting groundwater levels in Karnataka, India.

II LITERATURE SURVEY

Various works have been carried out in the field of groundwater which also include prediction analysis. The right form data extraction is highly necessary for an accurate model [4], A complex web of factors that determines groundwater levels, which are : Rainfall, aquifers, precipitation levels, seasonal changes, patterns of groundwater storage, water extraction. As highlighted in [7],[8],[9], Artificial Neural Networks (ANN) is ideal for forecasting based on implementation on data from wells and shallow aquifiers respectively. Further,different algorithms are analysed [10]. Least Squares Support Vector Machine (LSSVM) was used for dynamic forecasting of wells in Mongolia. The paper [11] helped understand the wavlet integration with support vector machine (SVM), and that it is a regression model that is being implemented to check the fluctuations in the water level. In [12], The focus is on quantitative estimates of groundwater temporally and spatially. They have performed analysis of groundwater level data in three districts of Maharashtra - Thane, Latur and Sangli. Analysis for data of more than 100 observation wells in each of these districts and developed seasonal models to represent the groundwater behavior. Three different type of models were developed-periodic, polynomial and rainfall models. In [13], a thorough analysis is conducted concerning the prediction of groundwater levels of Ljubljana polje aquifer. Three different datasets from Ljubljana (Slovenia) and Skiathos (Greece): information, pump sensor data from Skiathos and weather data. This paper processed the problem as a regression problem and hence implemented regression trees. They have highlighted the optimization when using ensembles by using random forests which is an ensemble of trees. Gradient boosting was also implemented. From the literature survey, neural networks were found to be widely used when it comes for predictive analysis. Hence for this Gradient Boosting and Random Forest are the suitable algorithms which will be used to predict the missing values and also compare the two algorithms to see which has better accuracy. LSTM (Long short term memory) which is RNN (Recurrent neural network) model is considered for predicting the future values.

III SYSTEM ARCHITECTURE

The proposed system developed using machine learning and deep learning application in which the dataset has been generated using different sources and later analysed. The groundwater dataset is gathered from various aquifers and observational wells of several districts across Karnataka. The dataset which comprises of the multiple parameters pertaining to groundwater conditions is split into training and testing data. Firstly cleaning of data is carried out by replacing the null values with the mean set of values. To compare and contrast the difference between a dataset with missing values and that with out one, the data is fit into the two algorithms that is Random forest and Gradient boosting. The loss functions MSE (Mean Square Error) presented based on the models. Also output obtained are plotted which highlight the most importance of a certain parameter among all the existing parameters.Further to predict future groundwater levels, the LSTM (Long short term memory network) is applied which is a RNN (Recurrent neural network) model. These results are intended to be given to the respective authority to carry out necessary actions.

Data preprocessing is the approach where in data is gathered from the specific account using XLS sheet. Since the gathered data is unstructured and not well-defined, the preprocessing techniques are used which include normalization, data cleaning, dimension reduction. This pre-processed data is reviewed and analysed several times by taking into consideration all of its parameters before fitting it into the model. Next the data is subjected to the prediction analyser in the form of the algorithms that includes gradient boosting, random forest and LSTM namely. Finally the outcome is compared to know how accurate is the prediction, also the result of the data cleaning and the replaced values is represnted in the form of a set of graphs. The system architecture of the project is depicted by Figure 1.

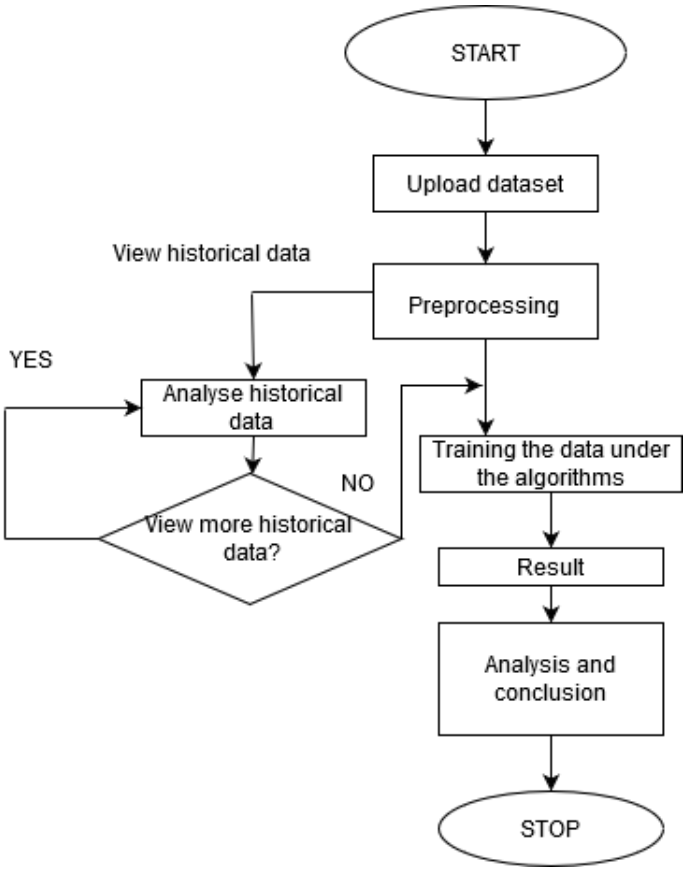

Figure 1: System Architecture

IV DATASET

Table 1 represents the attributes in a given dataset. The pre-processing is done on this dataset. The dataset consists of groundwater levels in the pre-monsoon and monsoon seasons as well as the Post Rabi and Post Kharif crop seasons. The dataset also consists of location data such as the well code, its district, state, site name and site type. The dataset we have chosen is shown in figure 2.

H	I	J	K	L	M	N
SITE_TYPE	WLCODE	YEAR_OBS	MONSOON	POMRB	POMKH	PREMON
Bore Well	W05243	2018	19.63	18.11	NA	NA
Dug Well	W24336	2018	NA	NA	4.13	3.7
Bore Well	W05497	2018	23	24.57	NA	NA
Bore Well	W06424	2018	NA	59.32	NA	NA
Bore Well	W21201	2018	NA	59.3	NA	NA
Bore Well	W05727	2018	NA	21.06	37.5	NA
Bore Well	W05731	2018	NA	7.64	9.1	NA
Bore Well	W06428	2018	NA	13.5	23.9	NA
Bore Well	W21202	2018	NA	5.46	6.36	NA
Dug Well	W18666	2018	NA	2.17	2.55	NA
Dug Well	W05726	2018	NA	4.3	3.5	NA
Bore Well	W21088	2018	NA	18.48	20.1	NA
Dug Well	W05245	2018	NA	NA	1.17	1.47
Dug Well	W05250	2018	NA	4.57	3.87	NA

Figure 2: Dataset

IV.I PRE-PROCESSING

The data collected is compiled from various government sources. The dataset initially required a significant amount of pre-processing. The techniques used on the groundwater dataset are data cleaning, data reduction and checking feature impotance. The dataset complied contained a large amout of NaN (Not a Number) values, due to which data was not usable. We first replaced the NaN values with 0s and then used an imputer function to replace the 0s with the mean value of each column. Certain parameters such as well code and site name were then dropped as it was not required for any of the methods employed in this project. Then a feature importance was calculated for the four features YEAR_OBS, MONSOON, POMKH AND POMRB that contain groundwater level values.

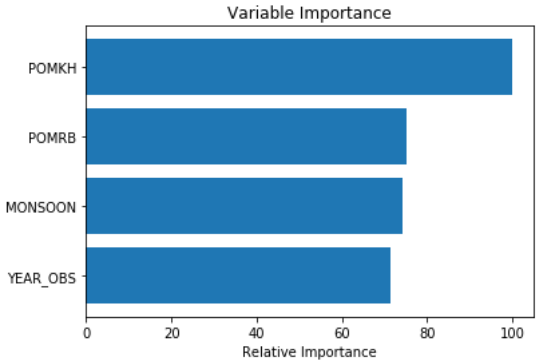

Figure 3: Feature importance after pre-processing

The plot as shown in Figure 3 depicts that POMKH (post monsoon kharif) has the highest importance. It is the top feature contributing to the predictions of the model and YEAR_OBS (year of observation) has the lowest importance. When training a tree, each feature contribution can be be computed while training, to decreasing the weighted impurity, which in case of regression trees is variance.

Attribute	Attribute Description
STATE	Name of the state where the observational well is located
DISTRICT	Name of the district where the observational well is located
TEH_NAME	Name of the tehsil (an administrative area) where the observational well is located
BLOCK_NAME	Name of the sub-region in the district where the observational well is located
LAT	Latitude of the observation well location
LON	Longitude of the observation well location
SITE_NAME	Name of the site of the observational well
SITE_TYPE	Type of the site i.e. bore well or dug well
WLCODE	A specific unique number assigned to each observation well
YEAR_OBS	Year during which the observation has been recorded
MONSOON	Ground water levels during the monsoon season
POMRB	Ground water level post monsoon during the rabi crop season (October to November)
POMKH	Ground water level post monsoon during the kharif crop season (June to October)
PREMON	Ground water levels before the monsoon season

Table 1.Data set description

V ALGORITHMS

Random Forest : Random forest is an ensemble learning method that combines the concepts of classification and regression tasks with the use of multiple decision trees and a technique called bagging (Bootstrap Aggregation) with some additional degree of randomization.

Given a training set $X = x1, ..., xn$ with responses $Y = y1, ..., yn$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to samples which are :

For $b = 1, ..., B$:

- Sample, with replacement, n training examples from X, Y called X_b, Y_b.
- Train a classification or regression tree f_b on X_b, Y_b.

Bootstrap aggregation uses the following formula to predict unseen samples by averaging prediction from individual regression trees :

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(\acute{x}) \tag{1}$$

Where $\hat{f}$ represents the prediction for unseen samples.

A single decision tree is a weak predictor (low bias, high variance) but is relatively fast to build. More trees give you a more robust model and prevent overfitting. When modeling, the data is resampled with replacement and for each sampling, a new classifier is trained. A new object is classified based on new attributes and each tree gives a classification. The forest chooses the classifications having the most votes of all the other trees in the forest and takes the average difference from the output of different trees. In general, different classifiers overfit the data in a different way, and through voting, those differences are averaged out.

Gradient Boosting : Gradient boosting is also based on decision trees and it builts one tree at a time. Boosting relies on weak learners (high bias, low variance) such as shallow trees, sometimes even as small as decision stumps (trees with two leaves). Here, model (ensemble) works in a forward stage-wise manner by adding one classifier at a time so that the next classifier is trained to improve the already trained ensemble, introducing a weak learner to improve the shortcomings of the existing weak learners. The gradient boosting method assumes a real-valued y and seeks an approximation $\hat{F}(x)$ in the form of a weighted sum of functions $h_i(x)$ from a class of weak learners γ.

$$\hat{F}(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + const \tag{2}$$

Long Short-term Memory Neural Networks : LSTM neural networks are a type of RNN that was developed to overcome the vanishing and exploding gradient obstacles of traditional RNNs [5]. The LSTM architecture minimizes gradient problems by enforcing constant error flow between hidden cell states, without passing through an activation function. This study uses LSTM cells with three gates (forget gate, input gate, and output gate). The forward pass of LSTM networks can be described by the following equations :

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{3}$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{4}$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{5}$$

Where f_t, i_t and o_t can be described as forget gate, input gate and output gate respectively. The matrices W_q contains the weight of the input and U_q contains recurrent connections and σ_g is the sigmoid activation function

used in the LSTM network. The network output was calculated by stacking a fully connected layer on top of the LSTM cell. The product of the output layer is the forecast of the groundwater level for the coming season.

VI METHODOLOGY

The groundwater dataset initially contained 14 parameters [Table 1], and these parameters had a lot of NaN values. During pre-processing, the NaN values were replaced with the value of $'0'$. Using an imputer function, 0 valued data has been replaced with the mean value calculated for each individual column. The feature importance of the parameters was also checked, which contained numerical values. Random Forest and Gradient Boosting algorithms were applied on the dataset both before and after replacing the 0 valued data with the mean imputation. The accuracies of the two algorithms before and after replacement of the null values was calculated. The dataset was divided into a 80:20 train-test split and the two techniques were applied to determine accuracies, Mean squared errors, Confusion matrices and classification reports.

For the random forest algorithm, a threshold value of 8.19 was used, calculated using the collective mean of all the groundwater values, to divide the data into 0 and 1 classes where any value below the threshold is classified as 0 and anything above the threshold is classified as 1. This division was created for the test and predicted classes to to identify the data relationship between the true-positive, true-negative, false positive and false negative. In the test split that contained 3519 values, the following was observed.

	TRUE POSITIVE	FALSE NEGATIVE
FALSE POSITIVE	2001	235
TRUE NEGATIVE	439	844

Table 2. Confusion Matrix from Random Forest Regression

The accuracy and mean squared error were then calculated using the random forest regression algorithm. For the gradient boosting algorithm, gradient boosting regressor function was used from the python libraries to calculate the accuracy and mean squared error both before and after replacing the null values. These algorithms were also used to predict the missing data in the dataset.

An RNN network, LSTM i.e. Long Short Term Memory was then implemented, to predict future groundwater levels, and to check the MSE values for training and test values, where the accuracy of the predicted values is also checked and is compared to its test and train set values. By implementing the algorithms, useful and necessary results have been yielded.

VII RESULTS

On extracting data from various sources, the data is cleaned and pre-processed, and is then fed to the random forest regressor model. To measure the effectiveness of the model the Confusion Matrix is provided. It is found that the recall, precision, f1-score and accuracy to have increased from 0.81, 0.81, 0.81, 0.809 to 0.85, 0.85, 0.85 and 0.51 respectively after replacing the null values with mean values.

On implementing gradient boosting there are again have two conditions i.e. with replacement of null values with mean and without replacement and it has been found that the MSE changes from 31.93 to 19.58 when replaced. In gradient boosting, it can be observed that at every iteration, a base learner is then fit to the negative gradient of the loss function and multiply our prediction with a constant and add it to the value from previous iteration. From implementing the two algorithms, it can be observed that the MSE and Accuracy using the gradient boosting yields the best results for the dataset.

ALGORITHM	**MSE BEFORE REPLACEMENT** (out of 100)	**MSE AFTER REPLACEMENT** (out of 100)	**INCREASE IN ACCURACY** (in percentage)
RANDOM FOREST REGRESSION	32.1665	20.3613	5%
GRADIENT BOOSTING	31.9301	19.5847	8%

Table 3. Comparison between Random Forest Regression and Gradient Boosting

In the Gradient Boosting algorithm, the training set and test set deviance is also checked, where deviance is defined as a goodness-of-fit statistic for a statistical model and it is found that the deviance from the actual value is very minimal as seen in Figure 4.

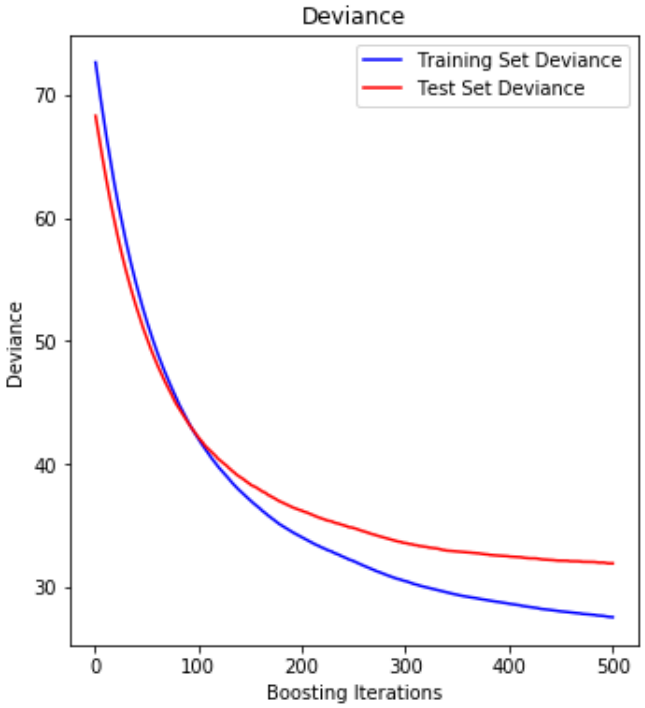

Figure 4: Data deviation as seen in Gradient Boosting

An RNN is also implemented for prediction of future groundwater levels using LSTM algorithm, which shows the improvement in the Root Mean Squared Error (RMSE). It is possible to compare the predicted values with the existing values in the dataset which is learnt by the Machine Learning model that has been used. Out of every 100 data values, it can be seen that the RMSE score of the LSTM model for train set compared to predicted values scored a 7.87, whereas for the test set compared to predicted values has scored a 13.36, which is low and beneficial. The overall performance of the models are found to be acceptable based on the high correlation efficiency.

VIII CONCLUSION

In this study, a groundwater level forecasting system is proposed. Three data-driven methodologies are tested based on various Machine Learning algorithms; namely, random forests, gradient boosting and LSTM. The system is based on using past groundwater levels data in the pre-monsoon and monsoon seasons as well as the Post Rabi and Post Kharif crop seasons to predict the groundwater levels for the missing values in the dataset, as well as for future usage. Analysis of the results indicated that the developed gradient boosting model provided a good prediction of groundwater levels, with considerably good accuracy and lower value MSE. The methodology and findings demonstrated in this study are useful to the research community of our nation involved in groundwater management and protection.

IX REFERENCES

[1] Groundwater https://en.wikipedia.org/wiki/Groundwater

[2] India Groundwater- World Bank Group https://www.worldbank.org/en/news/india-groundwater-critical-diminishing

[3] Thematic Papers on Groundwater-FAO http://www.fao.org/3/a-i6040e.pdf

[4] Prediction of Groundwater Level for Sustainable Water Management in an Arid Basin Using Data-driven Models Mutao Huang and Yong Tian

[5] Long short-term memory (1997) Hochreiter, S.; Schmidhuber, U.

[6] Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas.(2018) Zhang, J.; Zhu, Y.; Zhang, X.; Ye, M.; Yang, J.

[7] Groundwater Level Predictions Using Artificial Neural Networks (Dec 2002) MAO Xiaomin , SHANG Songhao, LIU Xiang

[8] Groundwater Level Forecasting in a Shallow Aquifer Using Artificial Neural Net- work Approach Purna C. Nayak1, Y. R. Satyaji Rao1 And K. P. Sudheer2(2006)

[9] Application of Back-Propagation Artificial Neural Network Models for Prediction of Groundwater Levels: Case study in Western Jilin Province, China Zhongping (2008)

[10] Groundwater level Dynamic prediction based on Chaos Optimization and Support Vector Machine (2009) Jin Liu, Jian-xia Chang Wen-ge Zhang

[11] An integrated wavelet-support vector machine for groundwater level prediction in Visakhapatnam,India(2014)

[12] Temporal Models for Groundwater Level Prediction in Regions of Maharashtra Dissertation Report - Lalit Kumar(2014) Ch. Suryanarayana a,n, Ch.Sudheer

b, VazeerMaham- mood c, B.K.Panigrahi d

[13] Groundwater Modeling with Machine Learning Techniques: Ljubljana polje Aquifer - Klemen Kenda , Matej Cerin , Mark Bogataj , Matej Senoetnik , Kristina Klemen , Petra Pergar , Chrysi Laspidou and Dunja Mladenic (2018)