# Water Table Analysis  using Machine Learning

Aishwarya Kulkarni[1] ,Shivangi Negi[1],Sumedha Raghu[1] and Dr. Vijaya Shetty[2]

[1] Department of Computer Science and Engineering ,Nitte Meenakshi Institute of Technology

Bangalore 50064,India

[2]Department of Computer Science and Engineering ,Associate Professor,

Nitte Meenakshi Institute of Technology Bangalore 50064,India

email:vijayashetty.s@nmit.ac.in; skay4kulkarni@gmail.com;

negishivangi3@gmail.com;sumedharaghu@gmail.com

**Abstract.** Groundwater is the primary water resource in arid and semi-arid areas. Monitoring water table fluctuations is essential for predicting the groundwater levels to outline the future needs. In this study, a thorough analysis is carried out on the prediction of groundwater levels in Karnataka. 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 variations. The prediction ability of these models was probed and evaluated using yearly groundwater level data scraped together 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 evaluate the performance of these models. These evaluation measures highlight the capability of  models to keep up with the shift in groundwater levels.

**Keywords:** groundwater levels; machine learning; data-driven; random forest; gradient boosting; long short-term memory

## 1 Introduction

Groundwater is characterized as the water present in the sections between the soil pore spaces and within the cleavages of rock formations. The depth at which soil pore fractures and gaps in rock become saturated with water is determined as the water table [1]. In several arid and semi-arid areas, groundwater has emerged as an important source of water required for domestic, irrigation, urban, and industrial activities. India is a substantial consumer of groundwater in the world using 230 cubic kilometres of groundwater per year, which is equivalent to the quarter of the global total [2]. Therefore, sustainable development of groundwater resources is essential for precise quantitative analysis, which is necessary for India due to its prevalent semi-arid and arid climate. Constant monitoring of groundwater levels is important to prevent the misuse of groundwater resources that can usher to local water rationing, limiting in agricultural yields, wells going dried-up or generating unpredictable groundwater quality changes, variations in flow patterns of groundwater emerging in the inflow of meagre quality water and seawater intervention in coastal areas [3]. The water levels, if forecasted prior, might facilitate the executives to plan better, the groundwater usage. In this research, we focus on observation 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 suitable in several arid and semi-arid areas as a consequence of insufficient data. Meanwhile, in data-driven modelling with machine learning methodology, our model attempts to establish a direct relationship among the inputs and outputs of the system without having any knowledge about the interior structure of the physical process [4]. The focus here is to use temporal data inputs (historic groundwater level, weather, and rainfall data) to learn the best approximation of the groundwater level values.

Recurrent neural networks (RNNs), a technique of deep learning, are a prominent choice for designing groundwater time series data due to their ability to retain a memory of previous network conditions, but they face challenges in acquiring long term dependencies within variables as weights associated with the network reaches to zero or become exceedingly large during model training [5]. LSTM, a class of RNN can avoid these training problems by eliminating unnecessary information being passed to future model states while retaining a memory of relevant past events [5]. LSTM networks have recently incorporated to model the groundwater table in an inland agricultural area of China on a monthly time step basis [6].

In our study, we are training models based on a set of related attributes to generate optimal predictions for the missing groundwater level values in our dataset using binary classification and later applying data-driven techniques to evaluate the performance of our models. Despite the growing applications of data-driven approaches in surface water problems, there are hardly any studies related to groundwater in arid and semi-arid areas [4]. Therefore, the emphasis of this study is on the implementation of data-driven models with machine learning (i.e., RF AND GB) and deep learning (i.e., LSTM) and comparison of two ensemble methods (i.e., RF and GB) for forecasting groundwater levels in Karnataka, India.

## 2 Literature Survey

Various works are being implemented in the field of hydrology and groundwater study which also includes prediction analysis. The right form of data extraction is highly necessary for an accurate model [13], A complex web of factors that determines groundwater levels, which are: Rainfall, aquifers, precipitation levels, seasonal changes, groundwater storage patterns, water extraction. As highlighted in [7], [8], [9], Artificial Neural Networks (ANN) is ideal for forecasting based on the implementation of data from wells and shallow aquifers, 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 wavelet integration with the support vector machine (SVM),which 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 a study of groundwater level data in three districts of Maharashtra - Thane, Latur, and Sangli. Analysis of data of above 100 observation wells in each of these districts and developed seasonal models to represent the groundwater behaviour. Three different types of models were developed-periodic, polynomial, and rainfall models. In [13], a detailed survey was conducted concerning the prediction of groundwater levels of Ljubljana polje aquifer. Three different datasets from Ljubljana (Slovenia) and Skiathos (Greece): weather pump sensor data from Skiathos and, information. 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. Gradient boosting was also implemented. From the literature survey, neural networks were found to be widely used for predictive analysis. Hence for this Gradient Boosting and Random Forest are the suitable algorithms to be used for predicting 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 future values.

## 3 System Architecture

The presented system as shown in Fig 1, is developed using machine learning and deep learning application using which the dataset has been generated 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 of groundwater conditions is split into testing and training data. Firstly, cleaning of data is carried out by alternating the empty cells with the mean set of values. To compare the difference between a dataset with NULL values and that without one, the data is fit into the two algorithms that are Random forest and Gradient boosting. The loss functions MSE (Mean Square Error) is presented on the models. Also, the output obtained is plotted which highlights the most important parameter among all the existing parameters. Further, the LSTM (Long short-term memory network) is applied which is an RNN (Recurrent neural network) model to predict future water level values These results are intended to be given to the respective authority for necessary actions.

Data pre-processing is the approach wherein data is gathered from the specific account using the XLS sheet. Since the gathered data is unstructured and not well defined, the pre-processing techniques are used which include normalization, data cleaning, dimension reduction. This pre-processed data is reviewed and analysed several times by considering all of its parameters before fitting it into the model. Next, the data is subjected to the prediction analyser using the algorithms that include gradient boosting, random forest, and LSTM namely. Finally, the outcome is compared to know how accurate the prediction is, also the result of the data cleaning and the replaced values is represented as graphs.

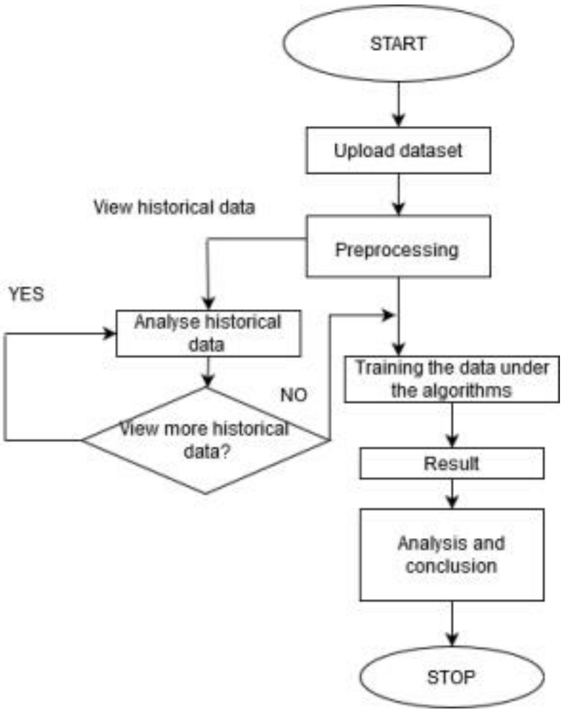

**Fig. 1**:System Architecture of Groundwater level prediction

## 4 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 season as well as the Post Rabi and Post Kharif crop seasons. It also consists of location data including the well-code, its district, state, site name, and site type. Fig 2 shows the dataset that has been used.

| **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 | 3.72 |
| 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 |

**Fig. 2**:Dataset

### 4.1 Pre-Processing

The data gathered is compiled from various government and water board 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 importance. The dataset compiled contained a large amount 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 alternate the 0s with the mean value of each column. Certain parameters like well code and site name were then dropped as it was not required for any of the methods employed in this project.

| **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(administrative area) where the observational well is located |
| BLOCK_NAME | Name of the sub-regions in the district where the observational well is located |
| LAT | Latitude of the observational well location |
| LON | Longitude of the observational 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 observational well |
| YEAR_OBS | Year during which the observation has been recorded |
| MONSOON | Groundwater levels during the monsoon |
| POMRB | Groundwater levels post monsoon during the rabi crop season(October to November) |
| POMKH | Groundwater levels post monsoon during the kharif crop season(June to October) |
| PREMON | Groundwater levels before the monsoon season |

**Table 1:** Dataset Description

Then a feature importance was calculated for the four features YEAR_OBS, MONSOON, POMKH AND POMRB that contain groundwater level values. The plot shows POMKH(post monsoon kharif ) (Fig 3) has the highest importance. It is the top feature contributing to the model, while (year of observation) has the lowest importance. When training trees one can compute the quantity, the feature contributes to decreasing the weighted impurity, which in case of regression trees is variance.

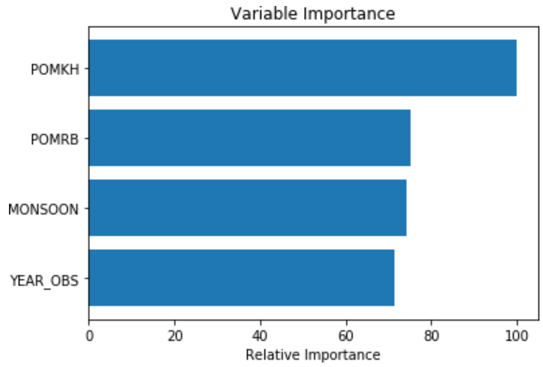

**Fig. 3:** Feature importance after pre-processing

## 5 Algorithms

**Random Forest :**

Random forest is one of the ensemble learning methodology that combines the concepts of classification and regression tasks with the help of multiple decision trees and a technique named bagging (Bootstrap Aggregation) with some additional degree of randomization.

Given a training set Z = z1, ..., zn with responses Y = y1, ..., yn, bagging repeatedly (M times) chooses a random sample with replacement and fits trees to generate new training sets:

For m = 1, ..., M :

- Sample, with replacement, n training examples from Z, Y called $Z_m$ , $Y_m$
- Train a classification or regression tree $f_m$ on $Z_m$ , $Y_m$

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

$$\hat{f} = \frac{1}{M}\sum_{m=1}^{M} f_m(z') \qquad (1)$$

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

An individual decision tree is considered as a weak predictor (low bias, high variance) but is fast enough to build. More trees give you a more robust model and prevent overfitting. When modelling, the data is resampled with replacement and for each sampling, a new classifier is trained. A new object type is classified based on new attributes and each tree votes a classification. The forest selects the classifications having the most votes of all the other trees in the forest.

In general, different classifiers overfit the data differently, and via voting, those disparities are averaged out.

**Gradient Boosting :**

Gradient boosting is a power technique for building predictive models. Gradient Boosting is about taking a model that by itself is a weak predictive model and combining that model with other models of the same type to produce a more accurate model. The idea is to compute a sequence of simple decisions trees, where each successive tree is built for the prediction residuals of the preceding tree.

In gradient boosting the weak model is called a weak learner. The term used when combining various machine learning models is an ensemble. The weak learner in XGBoost is a decision tree. The ensemble works in a forward stage-wise manner by adding classifiers one at a time such that the upcoming classifier is trained to improvise the previously trained ensemble, instituting a weak learner to improve the faults of the already present weak learners.

**LSTM :**

Long short-term memory (LSTM) units are units of a recurrent neural network (RNN)[5]. An RNN composed of LSTM units is often called an LSTM network. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series.

The forward pass of LSTM network is in the below equations:

$$fo_t = \sigma_g(W_{fo}\, x_t + C_{fo}\, h_{t-1} + b_{fo}) \qquad (2)$$

$$ip_t = \sigma_g(W_{ip}\, x_t + C_{ip}\, h_{t-1} + b_{ip}) \qquad (3)$$

$$op_t = \sigma_g(W_{op}\, x_t + C_{op}\, h_{t-1} + b_{op}) \qquad (4)$$

Here $fo_t$ , $ip_t$ and $op_t$ can be described as forget gate, input gate and output gate, respectively. The matrices Wq contains the weight of the input and Cq contains recurrent connection and $\sigma_g$ 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.

## 6 Methodology

The groundwater dataset initially contained 14 parameters [Table 1], and these parameters had a lot of NaN values. During pre-processing, we replaced the NULL values 0. Using an imputer function, we replaced the 0 valued data with the mean value calculated for each column. The feature importance of the parameters containing numerical values was checked. We applied the random forest and gradient boosting algorithms 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 the replacement of the null values were also calculated. We divided the dataset as a 70 : 30 train-test split and applied the two techniques to determine accuracies, Mean squared errors, Confusion matrices, and classification reports.

For the random forest algorithm, we used the threshold value of 8.19, deliberated 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 identify the data relationship between the true-positive, true-negative, false positive 4 and false negative, as shown in Table 2. In the test split that contained 3519 values, the following was observed. We then calculated the accuracy and MSE using the random forest regression algorithm. For gradient boosting algorithm, we made use of gradient boosting regressor function from the python libraries to calculate the accuracy

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

**Table 2** : Confusion matrix from Random Forest Regression

and MSE of both before and after replacing the null values. These algorithms were also used to predict the missing data in the dataset. We then implemented an RNN network, LSTM i.e. Long Short Term Memory, to predict future groundwater levels, and to check the MSE values for training and test values, where we also check the accuracy of the predicted values compared to its test and train set values. By implementing the algorithms, we yield useful and necessary results.

## 7 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 [Table 2.] is provided. Let Truly Positive be TF, Truly Negative-TN ,Falsely Positive-FP ,Falsely Negative -FN, Actual results-(TP+FP) and Predicted results- (TP+FN); Based on the same it is found that the Recall(TP / Predicted results), Precision (TP/Actual results) ,f1-score (2*(Precision*Recall)/Precision+Recall) and Accuracy(TP+TN/TP+TN+FN+FP) 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. Thus, we understand that all the measures of the confusion matrix have been in increased in their values which asserts the replacement.

On implementing gradient boosting we again have two conditions i.e. with the replacement of null values with mean and without replacement and we find that the MSE from 31.93 to 19.58 when replaced. In gradient boosting, it is found that in each iteration, we fit a base learner to the negative gradient of the loss function and later multiply the prediction with a constant and add it to the value from the previous iteration. From implementing the two algorithms, we observe that the MSE and Accuracy using the gradient boosting yields the leading results for the data set. Table 3. shows the Comparison between Random Forest Regression and Gradient Boosting. In the Gradient Boosting algorithm, we also checked for the training and test set deviance, where deviance is the goodness-of-fit statistic for a statistical model and we found that the deviance from the actual value is very minimal as seen in Fig 4.

| **Algorithm** | **MSE before replacement (out of 100)** | **MSE after replacement (out of 100)** | **Increase in accuracy (in percentage)** |
|---|---|---|---|
| **Random Forest** | 32.1665 | 20.3613 | 5% |
| **Gradient Boosting** | 31.9301 | 19.5847 | 8% |

**Table 3:** Comparison between Random Forest and Gradient Boosting

We have also implemented an RNN for prediction of future groundwater levels using the LSTM algorithm, which shows the improvement in the Root Mean Squared Error (RMSE). We can compare the predicted values with the existing one is in the data set which is learnt by the Machine Learning model that we have used.

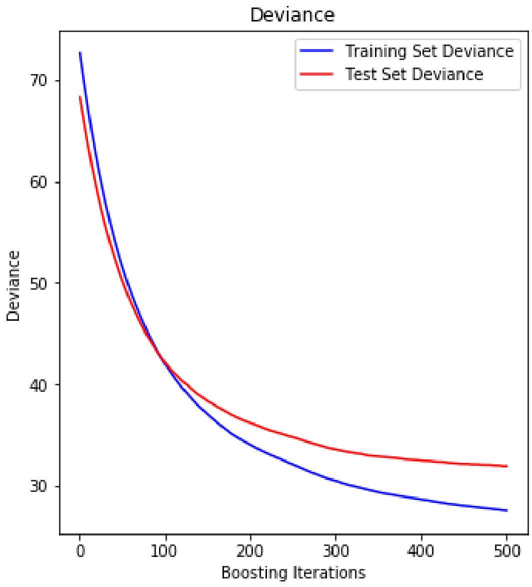

**Fig. 4:** Data deviation as seen in Gradient Boosting

Out of every 100 data values, we find that the RMSE score of the LSTM model for trained data compared to predicted values scored a 7.87, whereas for the tested data compared 5 to predicted values has scored a 13.36, which indicates how close the observed data points are to the model's predicted values. The overall performance of the models is found to be acceptable based on the high correlation efficiency.

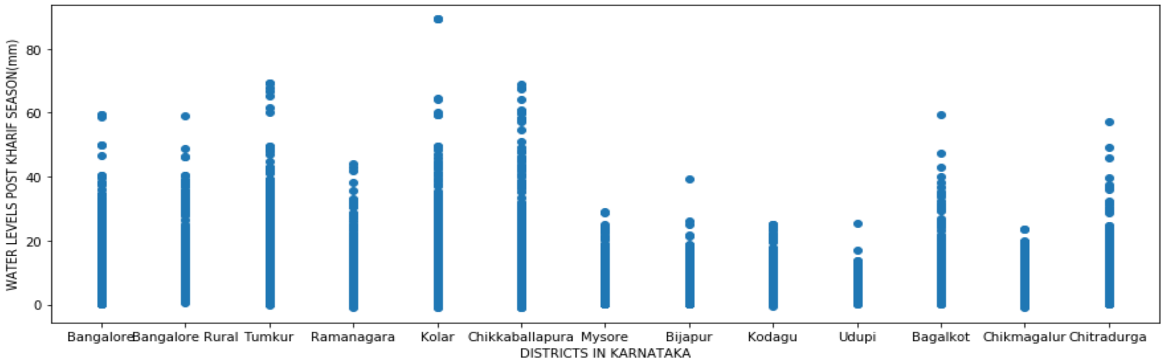

**Fig. 5:** Groundwater level distribution in various districts

The Fig. 5. is a scatterplot that depicts the groundwater level distribution in the 13 districts : Bangalore, Bangalore rural, Tumkur, Ramanagara, Kolar, Chikkaballapura, Mysore, Bijapur, Kodagu, Udupi, Bagalkot, Chikmagalur and Chitradurga. We can observe that regions such as Bijapur and Mysore receive lesser rainfall and therefore have lower groundwater levels than when compared to districts such as Kolar and Chikkaballapura. We can therefore conclude that these trends are able to predict when a district is in critical condition with respect to groundwater levels.

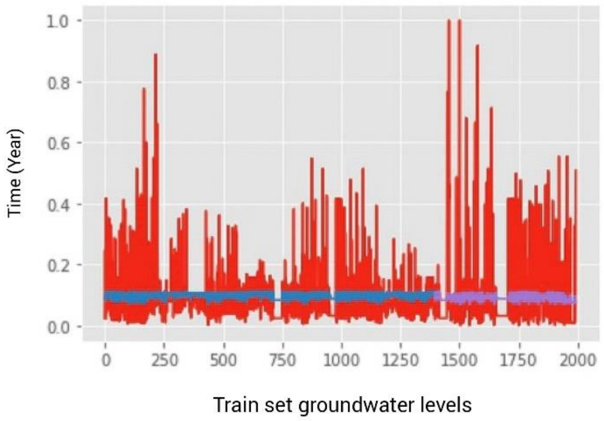

**Fig. 6:** LSTM prediction using train set

The Fig 6 is a graph comparing the train set values of the dataset (depicted in red) to the future values predicted by the LSTM network (depicted in blue) over the course of a year.

Similarly, The Fig 7 is a graph comparing the test set values of the dataset (depicted in red) to the future values predicted by the LSTM network (depicted in blue) over the course of a year. We can observe that the performance of the LSTM network is better with respect to test set values than with respect to the train set values. The overall performance of the models is found acceptable built on high correlation efficiency.

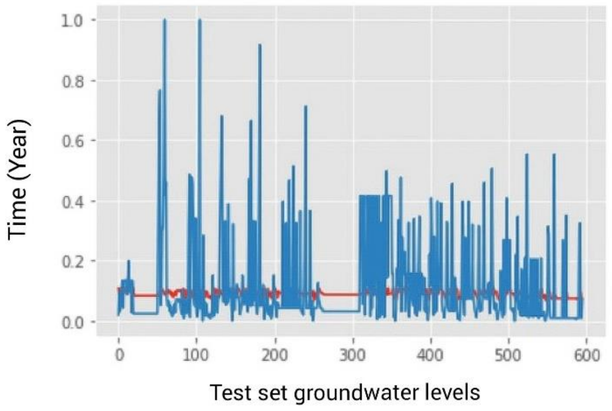

**Fig. 7:** LSTM prediction using test set

## 8 Conclusion

In this paper, we propose a groundwater level forecasting system. Three data-driven methodologies are tested based on various Machine Learning algorithms; namely, random forests, gradient boosting, and LSTM. The system is regarding 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 designed 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 paper are useful to the research community of our nation involved in groundwater management and protection.

# 9 References

[1] Richard Greenburg (2005). *The Ocean Moon: Search for an Alien Biosphere*. Springer Praxis Books.

[2] *The World Bank,* India Groundwater: a Valuable but Diminishing Resource, *india-groundwater-critical-diminishing,* 6 March 2012

[3] " *Groundwater Governance.* "Thematic Papers on Groundwater". April 2016. PDF File.

[4] Mutao Huang and Yong Tian, "Prediction of Groundwater Level for Sustainable Water Management in an Arid Basin Using Data-driven Models", Proceedings of the 2015 International Conference on Sustainable Energy and Environmental Engineering, 2015, pp. 134-137, 10.2991/seee-15.2015.33

[5] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735

[6] Zhang, Jianfeng & Zhu, Yan & Zhang, Xiaoping & Ye, Ming & Yang, Jinzhong. (2018). Developing a Long Short-Term Memory (LSTM) based Model for Predicting Water Table Depth in Agricultural Areas. Journal of Hydrology. 561. 10.1016/j.jhydrol.2018.04.065.

[7] X. Mao, S. Shang and X. Liu, "Groundwater level predictions using artificial neural networks," in Tsinghua Science and Technology, vol. 7, no. 6, pp. 574-579, Dec. 2002.

[8] Nayak, P.C., Rao, Y.R.S. & Sudheer, K.P. Groundwater Level Forecasting in a Shallow Aquifer Using Artificial Neural Network Approach. *Water Resour Manage* 20, 77–90 (2006)

[9] Z. Yang, W. Lu, Y. Long and P. Li, "Application of Back-Propagation Artificial Neural Network Models for Prediction of Groundwater Levels: Case study in Western Jilin Province, China," 2008 2nd International Conference on Bioinformatics and Biomedical Engineering, Shanghai, 2008, pp. 3203-3206, doi: 10.1109/ICBBE.2008.1130.

[10] J. Liu, J. Chang and W. Zhang, "Groundwater Level Dynamic Prediction Based on Chaos Optimization and Support Vector Machine," 2009 Third International Conference on Genetic and Evolutionary Computing, Guilin, 2009, pp. 39-43, doi: 10.1109/WGEC.2009.25.

[11] Ch. Suryanarayana, Ch. Sudheer, Vazeer Mahammood, B.K. Panigrahi, An integrated wavelet-support vector machine for groundwater level prediction in Visakhapatnam, India, Neurocomputing, Volume 145, 2014, Pages 324-335, ISSN 0925-2312

[12] "Temporal Models for Groundwater Level Prediction in Regions of Maharashtra"

Dissertation Report - Lalit Kumar(June 2012)

Ch. Suryanarayana a,n, Ch.Sudheer b, VazeerMahammood c, B.K.Panigrahi d

[13] Kenda, Klemen & Senozetnik, Matej & Klemen, Kristina & Mladenić, Dunja. (2018). Groundwater Modeling with Machine Learning Techniques: Ljubljana polje Aquifer. Proceedings. 2. 10.3390/proceedings2110697.