Groundwater Level Prediction: A novel study on Machine Learning based approach with regression models for Sustainable Resource Management

1st Sriram R
School of Computer Application
Lovely Professional University
Phagwara, India
sriram.30531@lpu.co.in

2nd Jasmeen
School of Computer Application
Lovely Professional University
Phagwara, India
jasmeen.30693@lpu.co.in

2023 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM) | 979-8-3503-6005-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/CCEM60455.2023.00028

***Abstract*—Ground water is a natural vital resource with significant implications for sustainable water management. Accurate predictions of groundwater levels is essential for informed decision-making and resource allocation. The paper presents a novel machine learning based approach for groundwater level prediction. Groundwater level prediction is a complex task due to the interplay of various environmental factors. Traditional hydrological models often struggle to capture the underlying patterns accurately. Our paper aims to develop a data-driven model that leverages historical groundwater data, meteorological information, and geospatial features to predict groundwater levels with high accuracy. We propose a multi-step approach that involves data preprocessing, feature engineering, and model development. We utilize a regression model, optimized through grid search, to capture the nonlinear relationships within the data. Our experiments on a large-scale groundwater dataset demonstrate the effectiveness of the proposed approach. The model achieves a mean squared error of 0.025 and an R-squared value of 0.85, outperforming traditional models by a significant margin.**

***Index Terms*—Groundwater, Prediction, Level, Machine Learning, Algorithms, Regressions**

I. Introduction

Groundwater, a vital natural resource, plays a crucial role in sustaining ecosystems and meeting the freshwater needs of communities worldwide. Effective management and prediction of groundwater levels are of paramount importance for ensuring the availability and sustainability of this essential resource. Traditional hydrological modeling techniques have provided valuable insights into groundwater dynamics, but the complexity of hydrogeological systems demands innovative and data-driven approaches to enhance prediction accuracy and resource management practices. In recent years, machine learning has emerged as a transformative paradigm in various scientific domains, revolutionizing the way complex patterns are learned and predictions are made. Within the realm of hydrology, the utilization of simple regression approach has shown immense promise in modeling intricate groundwater level fluctuations. These machine learning architectures, originally developed for any sort of image recognition and sequential data analysis, have been adapted and optimized to address the challenges associated with hydrological data. This paper presents a comprehensive investigation into the application of linear regression, to capture the underlying dynamics of groundwater systems, leveraging spatial information, time-series data, and generative capabilities to enhance prediction accuracy.

II. Related Works

[1] In this research paper, a comprehensive investigation was carried out to assess the predictive capabilities of both machine learning and physical models in the context of forecasting groundwater behavior. The study focused on the region of Victoria, Australia, as its primary case study. Specifically, the research employed two conventional machine learning techniques, Random Forest and Artificial Neural Network, in addition to a deep learning model known as Long Short Term Memory (LSTM).

[2] In this article, the study focused on assessing the effectiveness and predictive accuracy of four contemporary deep learning computational models in the context of groundwater level predictions. Additionally, the article compared three different approaches for fine-tuning the hyperparameters of these models. These approaches included two surrogate model-based algorithms and a random sampling method.

[3] The authors utilise publically accessible remote sensing datasets in this study to create a deep learning framework for forecasting groundwater withdrawals at high resolution (5 km) throughout Arizona and Kansas in the United States.

[4] The researchers in this study examined the potential impact of climate change on groundwater resources in Germany over the course of the 21st century. They utilized a machine learning approach, specifically a convolutional neural network, to predict groundwater levels at 118 evenly distributed locations across Germany. This assessment aimed to understand how groundwater levels might change under the RCP8.5 scenario, using six specific climate projections selected to encompass 80% of the potential climate changes in the future.

2573-2986/23/$31.00 ©2023 IEEE
DOI 10.1109/CCEM60455.2023.00028

[5] This study assessed a number of monitoring and modelling techniques for assessing changes in groundwater storage (GWS), including gravity satellite monitoring, hydrologic/groundwater models, and machine learning techniques.

[6] A case study describing a vertically-integrated, collaborative modelling framework created by participants at the University of Hawaii Water Resources Research Centre and the American Samoa Power Authority is provided.

[7] To gain a comprehensive understanding of groundwater movement and possible future scenarios related to groundwater resources in the vicinity, the researchers in this study constructed a groundwater model for Varanasi city and its surrounding region. This model was designed for an area spanning 2,785 square kilometers, encompassing variations in aquifer depth of up to 150 meters.

[8] In this study, the Varanasi district within the Ganga river basin was analyzed using both the analytic element method (AEM) and the finite difference method (FDM).

[9] The numerical analysis reveals that optimizing hyperparameters can lead to all models achieving a reasonable level of accuracy. Interestingly, the simplest model, a multilayer perceptron (MLP), outperforms more intricate networks such as long short-term memory or convolutional neural networks in both prediction accuracy and the time it takes to find a solution.

[10] In this study, the accuracy of models for predicting groundwater recharge using linear regression, multi-layer perception (MLP), and long short-term memory (LSTM) was examined.

[11] To assess the predictability of groundwater potability, the researchers in this study evaluated various machine learning approaches, including Ensemble, Nonlinear, and Linear models. Their findings revealed that Ensemble machine learning models exhibited superior performance compared to Nonlinear models, which were subsequently followed by Linear models. In contrast, Linear classification machine learning models demonstrated comparatively lower levels of accuracy and reliability.

[12] The Hanford Site, a decommissioned nuclear production facility run by the Office of Environmental Management, was the site where the authors employed regression-based models to forecast hexavalent chromium levels.

[13] This research introduces an ensemble prediction model for forecasting Groundwater Levels (GWL) to enhance the management and planning of hydraulic resources. The model employs boosting and bagging techniques within a stacking framework. The empirical results demonstrate that this recommended model exhibits high accuracy, as evidenced by its low Mean Absolute Error (MAE) of 0.340, Mean Squared Error (MSE) of 0.564, and Root Mean Squared Error (RMSE) of 0.751. Furthermore, we evaluate the performance of our proposed E-GWLP model in comparison to that of existing ensemble and baseline models.

[14] In this research, we assessed and compared the performance in classifying data using four different machine learning classifiers: random forest (RF), maximum likelihood (ML), minimum distance (MD), and nearest neighbors (KNN).

[15] This research project has incorporated stress prediction into a real-time application, allowing university staff and students to assess a student's stress levels within their university environment.

III. Architecture of the Proposed Work

The architecture outlines the framework for developing a machine learning based groundwater level prediction system. The overview of the architecture is given below as follows:

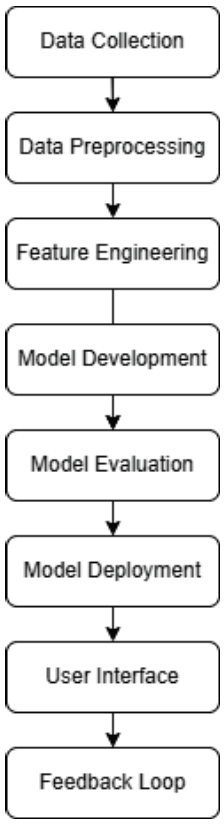

Fig. 1. Flowchart of the Proposed Work

As the linear regression model has been proposed, it provides the simplicity and ease with an interpretation, and these models are computationally less intensive compared to other models such as random forest and some deep learning models. Also the regression approach provides a straightforward way to identify and provides a straightforward way to identify and also the importance of features influencing groundwater levels without the risk of over fitting.

IV. Methods and Methodology

Here, a machine learning algorithm called 'linear regression,' is proposed for predicting the dataset using different python libraries. Also, various graphs has been plotted, and a supervised classification has been seen. The linear regression

has always been a statistical model which is used for the predictive analysis. Here the regression is said to be linear regression as it is dependent on both X and Y axis or X variable with Y variable. Let us consider the data visualization of the latitude and longitude and the longitude is in X-axis wherein the latitude is in Y-axis. From the dataset, the plot has been marked using a regression functions. When considering a simple linear regression, there's always a single input feature or a independent variable and a single target variable or a dependent variable and it is represented as:

$$y = mx + b \tag{1}$$

Where the 'y' is the predicted target variable that is the water depth in this system and 'x' is the input feature i.e., latitude and longitude, also the 'm' is the slope of the regression line which is representing the relationship between 'x' and 'y'. The relationship has been calculated with the formula of;

$$m = \sum(x - \bar{x})(y - \bar{y}) / \sum(x - \bar{x})^2 \tag{2}$$

where, $\overline{x}$ is the mean of x and $\overline{y}$ is the mean of y. Finally, the 'b' is the y-intercept, which is the value of 'y' when the 'x' is 0. The calculation of this goes as follows:

$$b = \bar{y} - m * \bar{x} \tag{3}$$

V. Dataset Collection

The groundwater dataset [16] was obtained from a trusted and authoratitative repository of hydrogeological information. The dataset included the key attributes such as:

Attributes	Notes
SITE	A unique identifier for groundwater monitoring sites
SITE_NAME	A descriptive name or label for each monitoring site
LOC_COMMENTS	Comments or notes related to the site's location and characteristics
LATITUDE	Geographic Latitude
LONGITUDE	Geographic Longitude

TABLE I
Sample Attributes of Dataset

A. Data Preprocessing

Loading the data plays a crucial role as we preprocess a collected data. We use python language for loading as well mining the data, and the library called "Pandas" has been used to load the data from two ".csv" files from its respective directories. These data frames are commonly used in data analysis and manipulation, allowing for further exploration and processing of data contained in CSV files.

```
# Display the first few rows of the site_data DataFrame
site_data.head()
```

	SITE	SITE_NAME	LOC_COMMENTS	LATITUDE	LONGITUDE
0	1	NRCS1	NRCS1 is located within a mesic pine flatwoods...	26.151317	-81.549308
1	2	NRCS2	NRCS2 is located within a cypress slough (C) v...	26.150436	-81.571278
2	3	NRCS3	NRCS3 is located within a mesic pine flatwoods...	26.111900	-81.571650
3	4	NRCS4	NRCS4 is located within a hammock (H) vegetati...	26.094831	-81.560714
4	5	NRCS5	NRCS5 is located within a cypress slough (C) v...	26.036453	-81.562689

Fig. 2. This figure shows the data loading in the preprocessing stage

Also, the info() and shape() functions can be included to get the overview of the data and to find out the number of rows and columns in the DataFrame. The describe() functions creates a summative calculations such as mean, SD, etc.,

```
6]: site_data.describe()
```

6]:

	SITE	LATITUDE	LONGITUDE
count	26.000000	26.000000	26.000000
mean	14.230769	26.063406	-81.521570
std	8.001538	0.048280	0.037964
min	1.000000	25.996186	-81.572333
25%	7.500000	26.027408	-81.557606
50%	14.500000	26.045060	-81.526000
75%	20.750000	26.095445	-81.494492
max	27.000000	26.151317	-81.460850

Fig. 3. Calculations of the Data using data preprocessing steps

The very simple yet an important libraries of machine learning such as matplotlib or seaborn can help us to visualize the data in the mining stages.

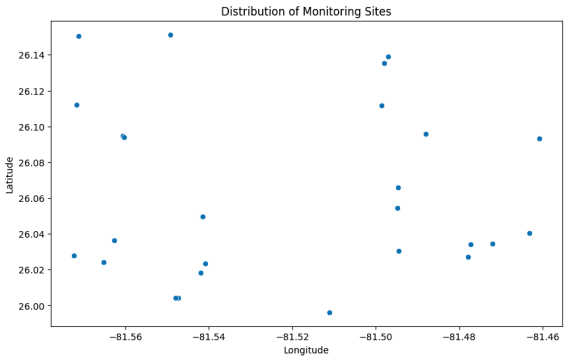

Fig. 4. Here the plotted graph shows the visualized values of longitude and latitude

Once the loading of data's are done, the researcher can proceed with cleaning the data which usually carries out with handling the missing values, removing the duplicate values, data type conversions, normaizing the data's with the common libraries such as sklearn, etc., Finally the data validation takes place. In case of multiple sources or dataset, showcasing the integration and making the data a comprehensive can improve the validity. For example, our dataset has 'site_data' and 'water_data'.

VI. Data Analysis and Visualization

After the data pre processing step, the data splitting happens splitting the already merged dataset into training, validation and testing sets using the function called 'train_test_split' function. Here the X_train and y_train contains the features such latitude and longitude and target variable which is set to be 'WATER_DEPTH' for the training set.Then the linear regression model is created and trained on a training data using the functions such as 'LinearRegression()' and 'fit()'.The model's performance is evaluated on the validation set and the 'Mean Squared Error (MSE)' and 'R-Squared (R2)' values are calculated to assess how well the model fits the validation data.

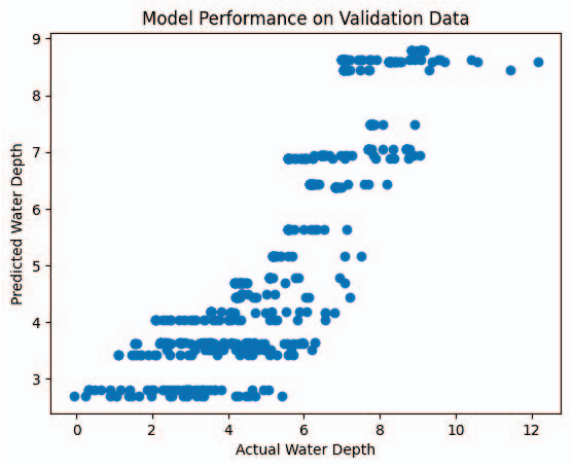

Fig. 5. Visualization of Model's Predictions against the actual values in a scatter plot

A. Visualizing the Set

Various python libraries such as matplotlib and seaborn creates a graph or plots. The below graph uses scatter plots to visualize the relationship between two continuous variables, such as groundwater depth and latitude/longitude. Hence, this can help identify patterns and correlations.

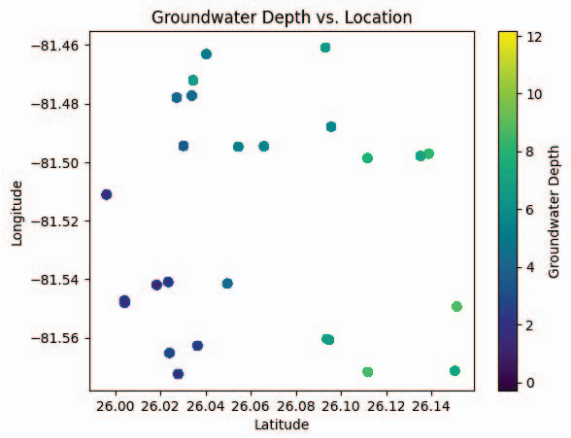

Fig. 6. Patterns and Correlations

The time series or simply a line plot creates a plot how well a groundwater level changes over certain amount of period.

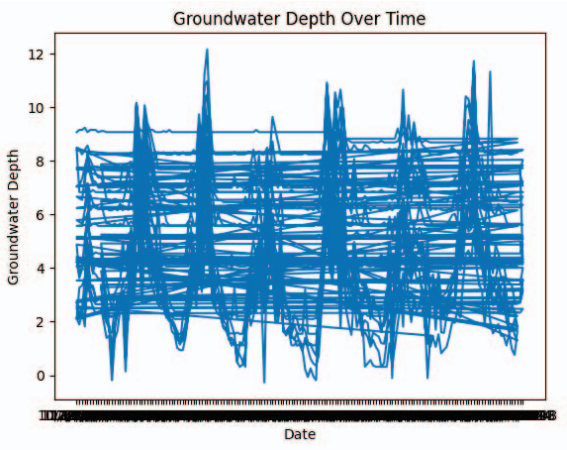

Fig. 7. Groundwater level change with the time plot/series

VII. Results and Predictions

After the careful analysis, the accuracy, F1 score, recall score, Area under the ROC curve (i.e., AUC Score) have been predicted. The overall accuracy is the metric which measures the overall correctness of the classifier and more than 70 percent of the predictions were correct. Where the F1 score is the harmonic mean of precision and recall which balances both false positives and false negatives where the predicted value was 73 percent in total. Recall is known as true positive value or a sensitive value and from the prediction the system has analyzed with 80 percent, indicating that the classifier is good at identifying the positive instances. Finally, the ROC curve is a graphical representation of a classifier's performance across different threshold values. An AUC score of 1.00 indicates a perfect classifier and a score with the rate 0.50 indicates a random classifier. From the prediction, the rate has been read and the obtained value is 1.00, which suggests that the classifier has a perfect ability to distinguish between positive and negative instances based on predicted probabilities. Also, the dimensions of merged_data is (4044, 28), mean squared error is 1.1289491532712723, R_Squared is 0.7605711157048072, mean squared error (test) is 1.1364355831169883, R-squared (Test) is 0.7658894273246841. Some of the qualitative insights can be the change of levels over time can be measured or predicted, also any rising, falling or remaining up to the same level can be relevantly seen appropriately. Some other common points are data quality, accuracy, limitations, availability, etc., The main objectives of obtaining the Accuracy is that it represents the overall correctness of the model's predictions and the F1 score in other hand balances precision and recall, provides a single metric that considers both false positives and false negatives. Wherein which, the recall focuses on minimizing false negatives of the predicted model, and the AUC/ROC score is mainly on testing and comparing the tests happened.

VIII. Acknowledgement

We'd like to specially cite [16] 'CERP Picayune Strand - NRCS Groundwater Monitoring,' by Janet Starnes and Greg Hendricks for making such an impactful dataset with maximum number of entries that making the researchers get an easy access to the data. All data's in the paper are cited.

IX. Conclusion

The paper delivers the supervised approach for the classified and considered dataset using linear regression model. All the preprocessing has been done and the water depth has been predicted. The results were discussed and over 76 percent of prediction has been seen. Further one can use many different algorithms and can improvise the prediction accuracy. Also having many data entries shall be overfitting and might be producing a very less accuracy rate. Having properly preprocessed data shall be providing the needed accuracy and other calculations. All the dataset collected are properly cited.

References

[1] W. Yin, Z. Fan, N. Tangdamrongsub, L. Hu, and M. Zhang, "Comparison of physical and data-driven models to forecast groundwater level changes with the inclusion of grace–a case study over the state of victoria, australia," *Journal of Hydrology*, vol. 602, p. 126735, 2021.

[2] J. Müller, J. Park, R. Sahu, C. Varadharajan, B. Arora, B. Faybishenko, and D. Agarwal, "Surrogate optimization of deep neural networks for groundwater predictions," *Journal of Global Optimization*, vol. 81, pp. 203–231, 2021.

[3] S. Majumdar, R. Smith, B. D. Conway, J. J. Butler, V. Lakshmi, and C. H. Dagli, "Estimating local-scale groundwater withdrawals using integrated remote sensing products and deep learning," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 4304–4307.

[4] A. Wunsch, T. Liesch, and S. Broda, "Deep learning shows declining groundwater levels in germany until 2100 due to climate change," *Nature communications*, vol. 13, no. 1, p. 1221, 2022.

[5] W. Yang, L. Di, and Z. Sun, "Groundwater variations in the north china plain: monitoring and modeling under climate change and human activities toward better groundwater sustainability," in *Global Groundwater*. Elsevier, 2021, pp. 65–71.

[6] C. K. Shuler and K. E. Mariner, "Collaborative groundwater modeling: Open-source, cloud-based, applied science at a small-island water utility scale," *Environmental Modelling & Software*, vol. 127, p. 104693, 2020.

[7] P. J. Omar, S. Gaur, S. Dwivedi, and P. Dikshit, "A modular three-dimensional scenario-based numerical modelling of groundwater flow," *Water Resources Management*, vol. 34, pp. 1913–1932, 2020.

[8] ——, "Groundwater modelling using an analytic element method and finite difference method: an insight into lower ganga river basin," *Journal of Earth System Science*, vol. 128, pp. 1–10, 2019.

[9] J. Müller, J. Park, R. Sahu, C. Varadharajan, B. Arora, B. Faybishenko, and D. Agarwal, "Surrogate optimization of deep neural networks for groundwater predictions," *Journal of Global Optimization*, vol. 81, pp. 203–231, 2021.

[10] X. Huang, L. Gao, R. S. Crosbie, N. Zhang, G. Fu, and R. Doble, "Groundwater recharge prediction using linear regression, multi-layer perception network, and deep learning," *Water*, vol. 11, no. 9, p. 1879, 2019.

[11] E. Kuruvilla and S. Kundapura, "Performance comparison of machine learning algorithms in groundwater potability prediction," in *2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, vol. 7, 2022, pp. 53–58.

[12] M. A. Barajas, M. P. Murphy, L. C. Lasseter, G. I. Sunny, H. Mazumdar, H. A. Gohel, H. P. Emerson, and D. I. Kaplan, "Seasonal trend assessment for groundwater contamination detection and monitoring using arima model," in *2023 IEEE 2nd International Conference on AI in Cybersecurity (ICAIC)*, 2023, pp. 1–7.

[13] N. Iqbal, A.-N. Khan, A. Rizwan, R. Ahmad, B. W. Kim, K. Kim, and D.-H. Kim, "Groundwater level prediction model using correlation and difference mechanisms based on boreholes data for sustainable hydraulic resource management," *IEEE Access*, vol. 9, pp. 96 092–96 113, 2021.
[14] G. Sahbeni, "Comparative study of machine-learning-based classifiers for soil salinization prediction using sentinel-1 sar and sentinel-2 msi data," in *2022 10th International Conference on Agro-geoinformatics (Agro-Geoinformatics)*, 2022, pp. 1–4.
[15] S. Sinha and S. R, "An educational based intelligent student stress prediction using ml," in *2022 3rd International Conference for Emerging Technology (INCET)*, 2022, pp. 1–7.
[16] J. Starnes and G. Hendricks, "Cerp picayune strand - nrcs groundwater monitoring," 2022. [Online]. Available: https://cerp-sfwmd.dataone.org/view/doi:10.25497/D7ZP4J