Prediction of Groundwater Level in an Unconfined Aquifer with Machine Learning

Xiao Chang
College of Business and Information Science, Tuskegee University, Tuskegee, AL 36088, USA

Eniola Webster-Esho
College of Business and Information Science, Tuskegee University, Tuskegee, AL 36088, USA

Corresponding author:
Xiao Chang
xchang@tuskegee.edu

Acknowledgement
This work is supported by the National Science Foundation (NSF) EPSCoR Track-2 grant (Award # 2019561) and NSF HBCU-UP Targeted Infusion Project (TIP) program (Award # 2306141). Any opinions, findings, conclusions, and recommendations expressed are those of the authors and do not necessarily reflect the views of NSF.

Abstract

Water is one of the most important resources for life on Earth. Groundwater is a critical natural resource that sustains human and ecological systems, providing essential water supplies for domestic, agricultural, and industrial use. Groundwater level (GWL) prediction is critical for planning drinking water supply and agriculture activities. This study investigated the approach to predicting GWL based on the data of GWL and the environmental factors of previous days with machine learning methods, including linear regression, decision tree, random forest, and artificial neural networks. All the machine learning methods achieved a MAPE of 0.09 or less in the experiment except decision tree. The experimental results show the models learned from the historical data of GWL and the environmental factors can predict GWL effectively.

Keywords: Groundwater level, machine learning, time series, regression

1. Introduction

One of the most important resources for life on Earth is water. Water exists in a variety of forms, including surface water, groundwater, and atmospheric water. Each form has its own set of properties and characteristics. The term "surface water" refers to water found in lakes, rivers, streams, and other bodies of water visible on the earth's surface. Groundwater is beneath the earth's surface and contained in aquifers, which are soil and rock formations. In an unconfined aquifer, groundwater is in direct contact with the atmosphere through the open pore spaces of the overlying soil or rock. In an unconfined aquifer, the groundwater level in a well is the same as the groundwater level outside the well.

Groundwater is a critical natural resource that supports human and ecological systems, providing essential water supplies for domestic, agricultural, and industrial use. It provides a consistent and long-term source of water that usually contains fewer chemical pollutants and other contaminants than surface water. It provides a consistent and long-term source of water that is frequently of higher quality than surface water sources.

Groundwater recharges surface water sources like rivers and lakes, keeps them flowing during dry periods, and supports a diverse range of plant and animal life. Groundwater also helps to keep our natural environment in balance by regulating the earth's temperature and acting as a natural filter for pollutants. The prediction of groundwater levels is helpful for effective water management and sustainable use in the support of life and environment.

Traditionally, physical models have been used for groundwater level prediction. However they are often computationally intensive and require significant data inputs (Nourani et al., 2011). Calibrations of these models are very difficult, since many parameters need to be controlled, particularly in chalky media. Additionally, these models need an enormous amount of good data and a complete realization of the essential physical processes in the system (Chen et al., 2009).

In the recent years, machine learning (ML) has emerged as a promising alternative for groundwater level (GWL) prediction, as it can effectively model complex relationships between groundwater level and environmental variables using data-driven approaches (Khedri, et al., 2020; Sahoo et al., 2017; Cho et al.; 2011; Sahoo et al., 2005). Artificial neural network (ANN) has been applied to groundwater level prediction with rainfall and temperature (Adamowski and Chan, 2011; Adiat et al., 2020; Coulibaly et al., 2001; Daliakopoulos et al., 2005; Juan et al., 2015). Dash et al. (2009) studied a hybrid neural model that is combination of (ANN-GA) employing an ANN model and genetic algorithms (GA) for accurate forecasts of groundwater levels in basin of Orissa State, India. Jalalkamali et al. (2011) studied the neuro-fuzzy (NF) and ANN methods to forecast the groundwater levels in Kerman plain of Iran. Shiri and Kisi (2011) evaluated the implementation of genetic programming (GP) and an adaptive neuro-fuzzy inference system (ANFIS) to predict groundwater level fluctuations using several benchmarks. The results of their findings showed the performance of GP was relatively better than that of the ANFIS model. Safieh et. al. (2020) evaluated a multilayer perceptron neural network (MLPNN) and an M5 model tree (M5-MT) in modelling groundwater level fluctuation in an Indian coastal aquifer. The evaluation results showed that the M5-MT outperformed the MLPNN model in estimating the GWL in the aquifer case study.

In this study, we investigated the approach to predicting groundwater level in an unconfined aquifer in North Carolina, the United States, with the observations of GWL and environmental factors, including precipitation, temperature, evapotranspiration, and surface pressure, of the previous days using machine learning methods. The multiple machine learning models were employed to construct GWL prediction models. The performance of the machine learning methods were compared in the experiment.

2. Method

2.1 Data set

The area under study is Haywood County located in North Carolina, the United States. According to the U.S. Census Bureau, Haywood county has a total area of 555 square miles (1,440 km2), of which 554 square miles (1,430 km2) is land and 0.9 square miles (2.3 km2) (0.2%) is water. The daily GWL data collected in the observation well located in located in an unconfined aquifer and in Haywood County in North Carolina, the United States, was downloaded from the USGS website (USGS 2023), which includes the GWL data collected from January 1, 2000 to December 31, 2019. The daily data of the other four factors was also downloaded and included in the dataset, including daily precipitation, temperature, evapotranspiration, and surface pressure. The historical data of daily GWL and the other four factors are be used to construct GWL forecasting models.

2.2 **Machine Learning Methods**

GWL prediction is a problem of time series prediction. We convert the time series prediction to regression by splitting the long time series into multiple short time series using a time window. The time window is slide along the time by one time step at each shift from the oldest time to the latest time in the data set or from the latest time to the oldest time. The GWL values and the values of other factors within the time window form a short time series. GWL of the last time step within a time window is treated as a target variable. GWL and environmental factors are considered as variables may have a dependent relationship with the target variable. A new data set with short time series can be generated from original time series. Any regression methods can be applied to construct GWL prediction models.

2.2.1 Linear Regression

Linear regression is a statistical technique for estimating the relationship between two variables by fitting a linear equation to the observed data. Linear regression can be used to identify a linear relationship between one dependent variable and one or more independent variables. The assumptions of multivariate analysis are normal distribution, linearity, freedom from extreme values and having no multiple ties between independent variables. (Gulden et al., 2013)

2.2.2 Decision Tree Regression

The structure of a decision tree (DT) is used to create regression or classification models. A DT is developed incrementally while a dataset is broken down into smaller and smaller subsets. A DT contains a root node, interior nodes, and leaf nodes. All the nodes of a decision tree are connected by branches. DT regressor predicts a continuous numeric value as an output based on a set of input features. DT learning algorithm employs a recursive binary splitting technique in which, at each split, it selects the input feature with the greatest information gain in terms of reducing the variance of the output values. A cost function, such as the mean squared error (MSE), is minimized at each split to reduce variance in training associated with each node.

2.2.3 Random Forest Regression

An ensemble of decision trees is used in the Random Forest (RF) regression algorithm to make prediction. RF regression is an extension of the DT regression, where multiple decision trees are trained on the subsets of training data and their predictions are averaged to improve the model's performance and avoid overfitting. Randomization is used to select the best node to split on when the individual trees in the RF are constructed. Breiman (2001) introduced additional randomness during the process of building decision trees using the classification and regression trees (CART). The Gini index heuristics are used to evaluate the subset of features chosen for each interior node using this method. In each interior node, the split feature is selected based on the feature's Gini index.

2.2.4 Artificial Neural Network (ANN)

An ANN is designed to mimic the structure and function of the human brain. It consists of interconnected nodes that work together to process information. The input layer is the first layer. It houses the input neurons that send data to the hidden layer. The hidden layer computes on the input data and sends the results to the output layer. The inputs from the input layer are multiplied by the weights that are associated with the connections between nodes. The multiplied values are added together to create the weighted sum. Then, an appropriate activation function is applied to weighted sum of inputs for generating output.

3. Experiment and Results

3.1. Data set

The data set contains the daily GWL data and surface pressure measured in the observation well, and the precipitation, temperature, evapotranspiration of Haywood County in North Carolina, the United States, from January 1, 2000 through December 31, 2019. After the rows with null values were removed, the data set contains a total of 7280 records of daily GWL, precipitation, temperature, evapotranspiration, and surface pressure. The daily GWL and other environmental factors are numeric variables.

3.2. Data preparation

The min-max normalization was performed to map the values of each numeric variable to a range [0, 1]. This was used to even out the weight of the one variable with other variables in the dataset. The training data was split into training and test sets. Training data set consists of the values of daily GWL and other factors from January 1, 2000 through December 31, 2016 with 6187 records. The test data set contains the data from January 1, 2017 to December 31, 2019 with 1093 records.

3.3. Evaluation Metrics

The evaluation metrics used to evaluate the performance of the models are mean absolute percentage error (MAPE) and mean squared error (MSE). MAPE is the average or mean of absolute percentage errors of forecast. Error is defined as the difference between actual value and predicted value. MAPE is computed by adding percentage errors without regard to sign. It provides the error in terms of percentages, the smaller the MAPE the better the prediction.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

where MAPE is mean absolute percentage error, n is number of times the summation iteration happens, A_t is actual value, and F_t is predicted value.

Mean Squared Error (MSE) is defined as mean or average of the square of the difference between actual and predicted values. This metric indicates how close a predicted value is to the actual value, the closer to zero the better the prediction.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where MSE is mean squared error, n is number of data points, y_i is observed values, and $\hat{y}_i$ is predicted values.

3.4. Evaluation results

The machine learning models, including linear regression, decision tree, random forest and ANN, were trained on the GWL data included in the training set. The trained models were applied to predict the GWL values of each day from January 1, 2017 to December 31, 2019 based on the data included in the test set. The evaluation results of the machine learning models with MAPE and MSE are given in Table 1 by year. The linear regression model achieved the MAPEs of 0.05, 0.08, and 0.05 in the prediction of daily GWLs in 2017, 2018 and 2019, which are lower than the MAPEs of decision tree, random forest and ANN in the GWL prediction. The learned linear regression model achieved the MSEs of 0.13, 0.18 and 0.11 respectively in the prediction of daily GWL values in 2017, 2018 and 2019, which are also lower than the MSEs of decision tree, random forest and ANN in the GWL prediction. The evaluation results show that the linear regression models outperformed the other three machine learning models tested in the GWL prediction task.

The daily GWL values from January 1, 2017 to December 31, 2019 predicted by the learned linear regression model are plotted in red in Figure 1. The actual daily GWL values from January 1, 2017 to December 31, 2019 are plotted in blue in Figure 1. We can see the predicted daily GWL values are close to the actual daily GWL values on most of the days, which demonstrates the good performance of linear regression in the GWL prediction.

Table 1. Evaluation results of the machine learning models in the GWL prediction

	Year	Linear Regression	Decision Tree	Random Forest	Artificial Neural Network
MAPE	2017	0.05	0.14	0.08	0.07
	2018	0.08	0.18	0.11	0.13
	2019	0.05	0.14	0.07	0.08
	Average	0.06	0.15	0.09	0.09
MSE	2017	0.13	1.07	0.33	0.19
	2018	0.18	0.73	0.29	0.39
	2019	0.11	0.93	0.2	0.2
	Average	0.14	0.91	0.27	0.26

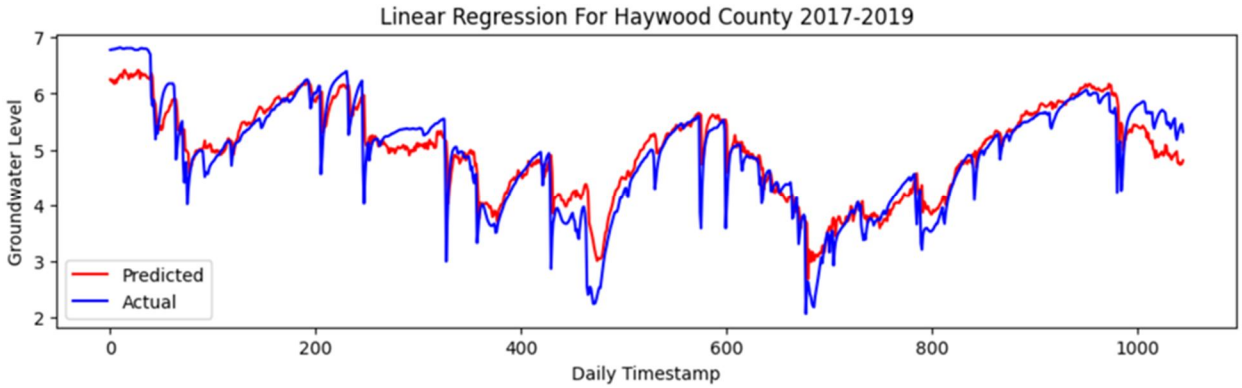

Figure 1. Actual daily GWLs from 2017 to 2019 and the daily GWLs predicted by the linear regression model

4. Discussion

The dramatic weather changes in some seasons may make the GWL prediction to be more challenging and lead to larger GWL prediction errors. The prediction results of the linear regression model were summarized by averaging the daily MAPEs of each month in 2017, 2018 and 2019. The evaluation results by month are shown with the bar plot in Figure 2. From the results, we can see that the average MAPEs in March of 2017, April of 2018, and November of 2018 are much higher than the average MAPEs in other months. The average MAPEs in March of 2017, April of 2018, and November of 2018 are 0.17, 0.38 and 0.28, respectively. The high average MAPEs indicate that the insufficient training data or more environmental parameters related to the GWL fluctuations in Haywood County should be incorporated into the GWL prediction model.

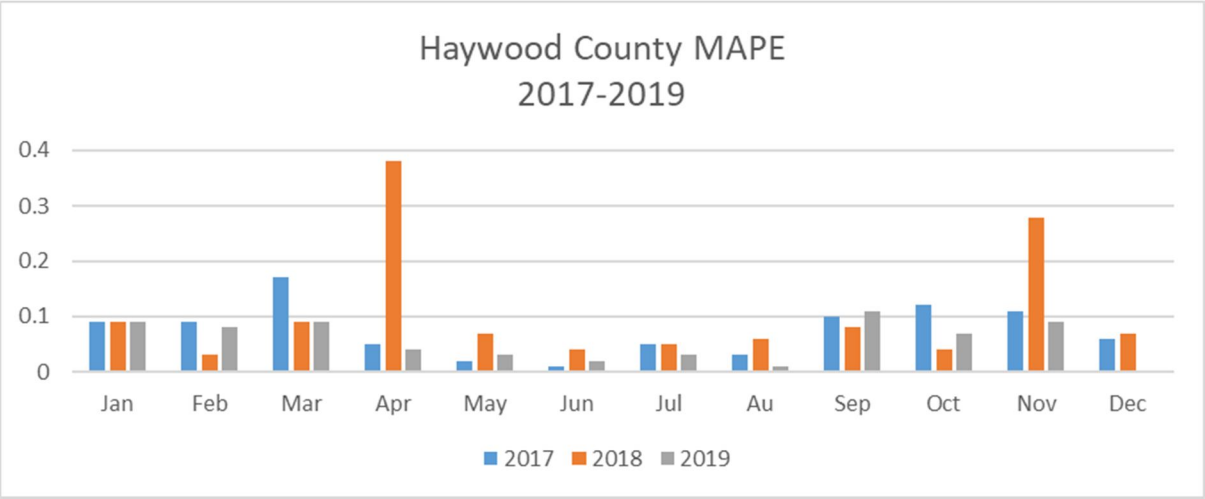

Figure 2. Monthly MAPE for Haywood County from 2017 to 2019 using Random Forest Regression

We conducted an experiment to test the impact of the training size on the performance of the GWL prediction model constructed with linear regression. The initial training data set includes the data of 10 years from 2000-2010. The GWL prediction model was trained on the training data set. The performance of the trained model was evaluated on the on the daily GWLs in 2019. Then the data of the next subsequent year was added to the training data set. The next GWL prediction model was trained on the incremental training data set. The performance of the new GWL prediction model was also evaluated on the daily GWLs in 2019. The training and evaluation processes were conducted repeatedly by adding the data of the next subsequent year to the training data set. The performance evaluation of the new GWL prediction model was always performed on the daily GWLs in 2019. In the end of the process, the last training data set consists of the data of 18 years from 2000-2018. The pairs of the size of the training data set by number of years covered and the MAPE achieved by the model trained on the training data set are plotted and given in Figure 3. The learning curve given in Figure 3 indicates that the MAPE dropped dramatically after about 17 years historical data of GWL and environmental factors were included in the training data. In addition to adding more data to the training data set, incorporating more hydrological and meteorological factors into the GWL prediction model may be helpful for improving the accuracy of the GWL prediction models.

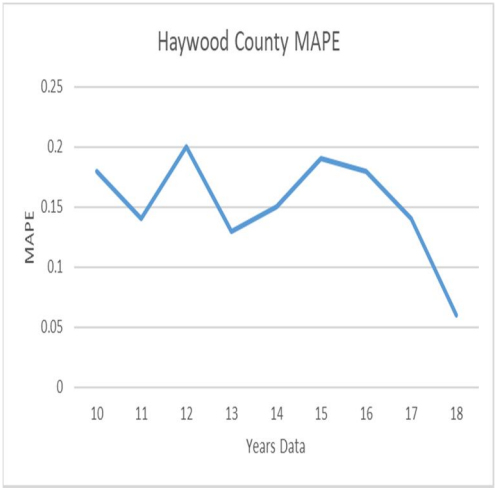

Figure 3. The learning curve of the GWL prediction model with the linear regression. Years Data n means the historical data of the n+1 years from 2000 to 2010. For example, Years Data 10 means the historical data of the 11 years from 2000 to 2010.

5. Conclusion

In this study, we investigated the approach to the prediction of groundwater level in the observation well in an unconfined aquifer located in Haywood County in North Carolina, United States, with machine learning. In addition to GWL, four environmental factors were incorporated into the prediction models. Linear regression, decision tree regression, random forest regression, and ANN regression were employed to construct the GWL prediction models. The experimental results show that the machine learning models learned from the historical data of GWL and the environmental factors can predict groundwater level with good accuracy. The GWL prediction using machine learning would be useful for monitoring groundwater conditions and informing future planning of drinking water supply and agricultural activities.

References

Adamowski, J., Chan, F.H., 2011. A wavelet neural network conjunction model for groundwater level forecasting. J. Hydrol. 407, 28–40. https://doi.org/10.1016/j. jhydrol.2011.06.013.

Adiat, K.A.N., Ajayi, O.F., Akinlalu, A.A., Tijani, I.B., 2020. Prediction of groundwater level in basement complex terrain using artificial neural network: a case of ljebu-Jesa, southwestern, Nigeria. Appl. Water Sci. 10 (8).

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. doi: 10.1023/A:1010933404324 [Crossref], [Web of Science ®], [Google Scholar]

Chen Lh, Chen Ct and Pan Yg (2009) Groundwater level prediction using SOM-RBFN multisite model. J. Hydrol. Eng. 15 (8) 624–631. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000218

Cho, K.H., Sthiannopkao, S., Pachepsky, Y.A., Kim, K.W., Kim, J.H., 2011. Prediction of contamination potential of groundwater arsenic in Cambodia, Laos, and Thailand using artificial neural network. Water Res. 45 (17), 5535–5544. https://doi.org/10.1016/j.watres.2011.08.010.

Coulibaly, P., Anctil, F., Aravena, R., Bob´ee, B., 2001. Artificial neural network modeling of water table depth fluctuations. Water Resour. Res. 37 (4), 885–896.

Daliakopoulos, I.N., Coulibaly, P., Tsanis, I.K., 2005. Groundwater level forecasting using artificial neural networks. J. Hydrol. 309, 229–240. https://doi.org/10.1016/j. jhydrol.2004.12.001.

Dash, N.B., Panda, S.N., Remesan, R., Sahoo, N., 2009. Hybrid neural modeling for groundwater level prediction. Neural Comput. Appl. 19 (8), 1251–1263.

Gulden K. U. and Nese G. 2013. A study on multiple linear regression analysis. Procedia - Social and Behavioral Sciences 106 (2013) 234 – 240

Jalalkamali, A., Sedghi, H., Manshouri, M., 2011. Monthly groundwater level prediction using ANN and neuro-fuzzy models: a case study on Kerman plain, Iran. J. Hydroinf. 13 (4), 867–876.

Juan, C., Genxu, W., Tianxu, M., 2015. Simulation and prediction of suprapermafrost groundwater level variation in response to climate change using a neural network model. J. Hydrol. 529, 1211–1220. https://doi.org/10.1016/j.jhydrol.2015.09.038.

Khedri, A., Kalantari, N., Vadiati, M., 2020. Comparison study of artificial intelligence method for short term groundwater level prediction in the northeast Gachsaran unconfined aquifer. Water Supply 20 (3), 909–921.

Nourani V, Kisi Ö and Komasi M (2011) Two hybrid artificial intelligence approaches for modeling rainfall–runoff process. J.Hydrol. 402 (1–2). 41–59. https://doi.org/10.1016/j.jhydrol.2011.03.002

Safieh J, Rebwar D and Forough J 2020. Modelling groundwater level fluctuation in an Indian coastal aquifer. Water SA 46(4) 665–671 / Oct 2020 https://doi.org/10.17159/wsa/2020.v46.i4.9081

Sahoo, G.B., Ray, C., Wade, H.F., 2005. Pesticide prediction in ground water in North Carolina domestic wells using artificial neural networks. Ecol. Model. 183 (1), 29–46. https://doi.org/10.1016/j.ecolmodel.2004.07.021

Sahoo, S., Russo, T.A., Elliott, J., Foster, I., 2017. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the US. Water Resour. Res. 53 (5), 3878–3895.

Shiri J and Kisi Ö (2011) Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. Comput. Geosci. 37 (10) 1692–1701. https://doi.org/10.1016/j.cageo.2010.11.010

USGS 2023 https://waterdata.usgs.gov/monitoring-location/352315082484401