The International Journal of Engineering and Information Technology

journal homepage:www.ijeit.misuratau.edu.ly

# Groundwater Level Forecasting Using Random Forest and Linear Regression Neural Network Models

**Amna Elhawil** University of Tripoli
**Alarabi Naji** University of Tripoli
**Malak Nuesry** University of Tripoli
**Almabruk Sanossi** University of Tripoli

*Abstract*—**Predicting the groundwater level has recently become very important research topic especially with the rise of population density and consequently increasing the water demand. This paper uses the Random Forest and linear regression neural network models to predict the groundwater level of Wadi-Alshaty district in the South West part of Libya. The results are compared with that obtained using the hydrologic long-term forecasting graphical approach. One of the most important findings of this study is the effectiveness of the neural network models to investigate the fluctuation of the groundwater levels over time (20 years).**

*Index Terms:* **Groundwater level forecasting, Random Forest, Linear regression, Wadi-Alshaty.**

## I. INTRODUCTION

Libya is located in the north part of Africa. About 70% of its area is located in the Great Desert. In 1953 the search for oil in the deserts of southern Libya led to the discovery not only of significant oil reserves, but also vast quantities of freshwater trapped in aquifers under the Libyan desert [1]. Since that, the groundwater in the south of Libya has become a very important source of water. In 2019, the United Nations Children's Fund (UNICEF) reported that Libya is considered as one of the most water scarce countries in the world. The annual amount of renewal water per person is 108 cubic meters, whereas the minimum international limit is about 1000 cubic meters per year. The groundwater represents 97% of the water supply in Libya [2]. The prediction of groundwater level has been tackled using time series forecasting techniques. The groundwater level and the flood are both predicted using machine-learning techniques including Multivariate Linear Regression (MLR), Multilayer Perceptron (MLP), Random Forest (RF), eXtreme Gradient Boosting (XGB) and Support Vector Machine (SVR) [3] and [4]. On the other hand, the artificial neural network (ANN) is adapted for regional flood frequency analysis in [5] and for rainfall intensity-duration-frequency in [6].

In this paper, the most widely used techniques of time series prediction are used to estimate the groundwater levels and depths of wells at different cities in the South West part of Libya. These techniques are: Linear Regression (LR) and Random Forest (RF) machine learning techniques. The results are compared with that computed using frequency analysis techniques.

## II. TIME SERIES FORECASTING TECHNIQUES

Time series forecasting means making predictions based on historical time data. The technique builds models through scientific historical analysis and uses them to drive future predictions. In general, there are two basic classes of models: traditional and machine learning models, as shown in Figure. 1. Traditional time series techniques include, but not limited to, Linear Regression, Autoregressive Integrated Moving Average (ARIMA) [7], Prophet [8], neural Prophet and vector autoregression (VAR). Each technique has its limitation. For example, ARIMA and Prophet are univariate models, both deal with single time series. The future values are predicted using the previous time series. On the other hand, VAR is multivariate model. It can deal with multiple time series. Actually, it uses the predicted values to forecast the future. The main disadvantage of VAR is it requires a lot more quality data to come up with reasonable predictions [8]. However, machine learning models have shown good learning capability even with complicated time series. Machine learning is divided into two subcategories: supervised and unsupervised machine learning. In supervised learning, a training set, contains the correct answer, is used to teach the models to yield the desired output. Whereas, unsupervised learning uses unlabeled data. The model discovers patterns and features to solve problems. Supervised learning techniques can handle two types of problems: classification and regression. Classification is a kind of problem wherein the results are categorical in nature like 'Yes' or 'No', 'True' or 'False' etc. On the other hand, regression problem predicts a continuous quantity outputs or values. Random forest is

Received 22 Oct , 2021; revised 7 Nov, 2021; accepted 4 Nov, 2021.

Available online 18 Aug, 2021.

one of the main supervised learning algorithms in addition to neural networks, naive Bayes, linear regression, logistic regression, support vector machine (SVM) and K-nearest neighbor algorithms. Random forest can be used for both classification and regression purposes. For time series prediction, random forests and linear regression models are mostly used.

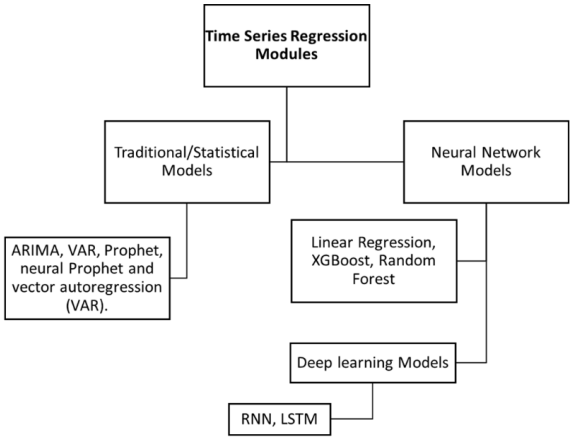


Figure 1. Classification of time series prediction techniques

### A. *Hydrological long-term forecasting (HLTF)*

Water demand of future population is estimated based on future population growth. Prediction of future population is divided into two types: short-term estimates for 1 to 10 years, and long-term forecasting for 10 to 50 or more years [9]. The techniques are implemented graphically or mathematically. The graphical method is based on plotting the population of past census years against time, sketching a curve that fits the data, and extending this curve into the future to obtain the projected population. On the other hand, the mathematical approach assumes three forms of population growth: geometric growth, arithmetic growth, and declining rate of growth. As shown in Figure. 2 each segment has a separate relation. The historic population data of the study area may be plotted on a regular graph. Depending on whether the shape is similar to *ab*, *bc*, or *cd*, the relationship of that segment should be used for population projection [9].

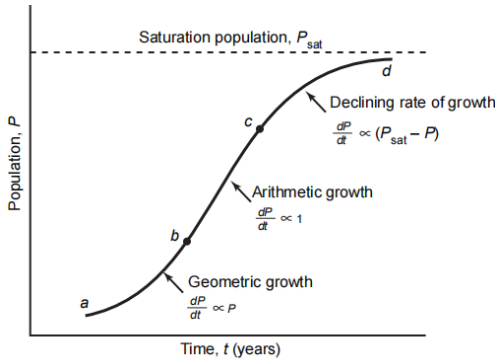


Figure 2. Population growth curve.
Note: Adapted from [9], pp. 6

### B. *Random forests regression model (RF)*

Random forests are made out of a collection of decision trees generated using random bootstrapped samples of the dataset. Random forests technique was first proposed by L. Breiman in 1984 [10]. The "forest" references a collection of uncorrelated decision trees, which are then merged together to reduce variance and create more accurate data predictions. This is called bagging [11]. The bagging technique was first introduced by Breiman in 1996 [12]. Random forest can maintain high accuracy predictions. As shown in Figure. 3, this technique works as follows:

1. The model starts by randomly selecting samples, called bootstrapped samples, from the training dataset.
2. The model then constructs many decision trees from bootstrapped samples. The number of the trees is a hyperparameter depending on the size of the training set.
3. The model is trained by running all the trees in parallel. The output prediction is given by

$$y = \frac{1}{n}\sum_{i=1}^{n} p_i. \qquad (1)$$

where $n$ is the number of trees, $p_i$ is the prediction of each tree. The size of the trees is controlled by setting the number of samples required to be at a leaf node and maximum depth of the tree (the largest possible length between the root to a leaf).

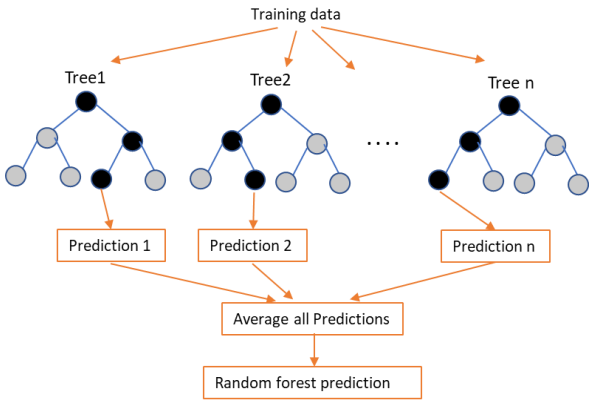


Figure 3. The structure of the random forest trees

### C. *Ordinary Least Squares Linear regression model (OLSLR)*

Regression models estimate the target output by finding out the relationship between time series data. As shown in Figure. 4, this relationship in the ordinary least squares (or simple) linear regression models is linear, that means the explanatory variable ($y$) is predicted based on a given independent variable ($x$) [13].

$$y = \beta_0 + \beta_1 x + e \qquad (2)$$

The coefficients $\beta_0$ and $\beta_1$ denote the intercept and the slope of the line respectively. The intercept $\beta_0$ represents the predicted value of y when x = 0 . The

slope $\beta_1$ represents the average predicted change in $y$ resulting from a one unit increase in x. $e$ is the random error component.

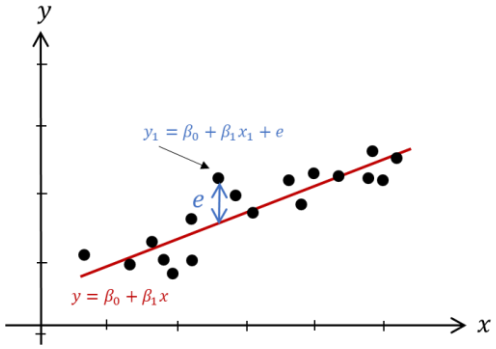


Figure 4. Linear regression prediction

The predictions are based on minimizing the sum of the least square errors for the training data. That is accomplished by finding the best fit regression line. The loss function is given as

$$Loss = \frac{1}{2}\sum_{i=1}^{n}(f(x_i) - y_i)^2 \quad (3)$$

Or

$$Loss = \frac{1}{2}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - y_i)^2 \quad (4)$$

## III. STUDY AREA AND DATASET DESCRIPTION

The study area lies in Wadi-Alshaty district, in the central-west of Libya, approximately 640 kilometers south of Tripoli. It occupies a land area of 15,330 km$^2$. Most of this area is desert as shown in Figure. 6. The estimated population, at mid-year 2021, is 95,294 [14]. The groundwater is considered to be the major source of water for this region. This study includes four main wells located in the following Oases: Burgan, Umm-Aljdawal, Wanzarik and Ashkda. These Oases, as shown in Figures. 5 and 6, are close to Sabha city, which is one of the biggest Cities in Libya with a population of 153,454 [14]. The studies show that the hydraulic transmissibility coefficient of these wells is about 2.59 x $10^{-2}$ m$^2$/day and the storage coefficient is about 5.5 x $10^{-3}$ [15]. The groundwater reservoirs of Wanzarik, Umm-Aljdawal and Ashkda are confined artesian aquifer due to the presence of existing clay layers while the Burgan wells are in non-artesian condition thus the water does not flow self-contained to the surface [15].


Figure 5 Wadi-Alshaty district in Libya's map and the oases included in this study


Figure 6. Satellite image of Wadi-Alshaty district in Libya

In this work, the input data is the annual groundwater level in meters, listed in Table 1. It is measured by The General Company of Water and Sanitation- Sabha in the period from 1984 to 2013 [16]. Figure. 7 illustrates the measured groundwater level. It can be seen that, the groundwater level in general decreases over the years. In fact, this area is suffering from excessive extraction of groundwater which produce a continuous drop in water levels [17]. The depletion of Umm-Aljdawal Oasis is over the expectation by 19.42 meters in the 30 years (1984 to 2013) as shown in Table 2. That is due to its population growth. In addition to that this Oasis is an agricultural land. The frequent pumping of water has been increased leading to a dangerous reduction in the groundwater supply. Furthermore, Wanzarik Oasis, which has smaller populations and quite far from Sabha city as shown in Figure. 8, has a depletion of only 2.73 over the same period.

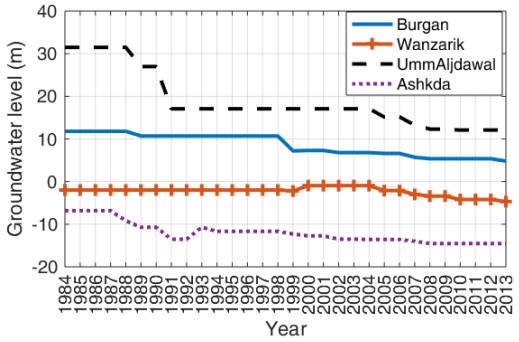


Figure 7. Groundwater Hydrograph of the oases


Figure 8. Satellite image of Wanzarik Oasis (top left) and Sabha city (bottom right)

Table 1. The measured groundwater levels of the oasis from 1984 to 2013

| Year | Burgan | Wanzarik | UmmAljdawal | Ashkda |
|---|---|---|---|---|
| 1984 | 11.8 | -1.97 | 31.5 | -6.84 |
| 1985 | 11.8 | -1.97 | 31.5 | -6.84 |
| 1986 | 11.8 | -1.97 | 31.5 | -6.84 |
| 1987 | 11.8 | -1.97 | 31.5 | -6.84 |
| 1988 | 11.8 | -1.97 | 31.5 | -9.18 |
| 1989 | 10.7 | -1.97 | 27 | -10.73 |
| 1990 | 10.7 | -1.97 | 27 | -10.73 |
| 1991 | 10.7 | -1.97 | 17.08 | -13.54 |
| 1992 | 10.7 | -1.97 | 17.08 | -13.54 |
| 1993 | 10.7 | -1.97 | 17.08 | -10.72 |
| 1994 | 10.7 | -1.97 | 17.08 | -11.68 |
| 1995 | 10.7 | -1.97 | 17.08 | -11.68 |
| 1996 | 10.7 | -1.97 | 17.08 | -11.68 |
| 1997 | 10.7 | -1.97 | 17.08 | -11.68 |
| 1998 | 10.7 | -1.97 | 17.08 | -11.68 |
| 1999 | 7.2 | -2.21 | 17.08 | -12.3 |
| 2000 | 7.3 | -0.95 | 17.08 | -12.75 |
| 2001 | 7.3 | -0.95 | 17.08 | -12.75 |
| 2002 | 6.8 | -0.95 | 17.08 | -13.51 |
| 2003 | 6.8 | -0.95 | 17.08 | -13.51 |
| 2004 | 6.8 | -0.95 | 17.08 | -13.6 |
| 2005 | 6.6 | -2.1 | 15.17 | -13.6 |
| 2006 | 6.6 | -2.1 | 15.17 | -13.6 |
| 2007 | 5.7 | -3.01 | 13.27 | -14 |
| 2008 | 5.36 | -3.41 | 12.31 | -14.53 |
| 2009 | 5.36 | -3.41 | 12.31 | -14.53 |
| 2010 | 5.36 | -4.2 | 12.08 | -14.53 |
| 2011 | 5.36 | -4.2 | 12.08 | -14.53 |
| 2012 | 5.36 | -4.2 | 12.08 | -14.53 |
| 2013 | 3.8 | -4.7 | 12.08 | -14.53 |

Table 2. The depletion of groundwater in the oasis from 1984 to 2013

| # | Oasis | Decline *of* Groundwater level from 1984 to 2013 *(meters)* |
|---|---|---|
| 1 | Burgan | -8 |
| 2 | Wanzarik | -2.73 |
| 3 | Umm-Aljdawal | -19.42 |
| 4 | Ashkda | -6.84 |

## IV. MODEL'S SETTING

As described in [15], the long-term graphical curve-fitting approach is applied. It is a simple linear regression of the form $y = a + bT$. where $T$ is the time and $y$ is the groundwater level. $a$ and $b$ are constants. The procedure works by finding the values of the coefficients $a$ and $b$ for the straight line. The slope of the fitted line represents the general trend of water level. Negative slope (or trend) means the water level declining over the period whereas the positive trend represents water level rising over years.

For neural network models, the data is preprocessed by restructuring it as a time series supervised learning problem. As illustrated in section II, RF has hyperparameters that are tuned to improve its performance. The optimal chosen values are as following:

1. The number of trees is 300.
2. Number of variables in each split (sliding window) is 6.
3. The depth of each tree is 10.
4. Minimum number of data points placed in a node before the node is split is 2.
5. Minimum number of data points allowed in a leaf node is 1.

Furthermore, the intercept of OLSLR model is included. This produces unbiased model even if the true value of the intercept is approximately zero. The main procedure of both models has the following steps:

1. Build the model with the appropriate specifications
2. Train the model using the training dataset. Then the models are run to predict the groundwater levels. These predictions are used to evaluate the performance of the model.
3. Calculate the accuracy by comparing the difference between the actual and predicted values as it will be explained in the next section.

## V. MODEL'S EVALUATION

The performance of used models is measured using two metrics: Mean Absolute Error (MAE) and Mean Squared Error (MSE). They are defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_{ti} - y_{pi}| \quad (5)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{ti} - y_{pi})^2 \quad (6)$$

where $y_{ti}$ and $y_{pi}$ are the true and predicted values respectively. High MAE or MSE values mean that the predictions are far away from the true values.

## VI. RESULTS

In this paper, the results of HLTF technique are considered from the study in the reference [15]. These data are compared with the results of the neural network models. In this section we will describe our results and compare them with that of HLTF. Machine learning models must be trained and tested using different datasets to avoid overfitting. For this reason, the dataset, which is the measured data, is divided into three sets: training, testing, and evaluating dataset. The split percentages are

60%, 20% and 20% respectively. The training dataset, which includes both the input and the output, is used to train the models to find the correct pattern that maps the input data to the target output. Once the models are successively trained, they are tested by applying another input data, called *test* dataset. The performance of the model is measured using the test and evaluate datasets. The training and testing steps are repeated until the satisfying accuracy is achieved. At this stage, the model is ready to work. The obtained performance characteristics are listed in Table 3 and Figurers. 9 and 10. The UmmAljdawal oasis has the large error as shown in Figures. 9 and 10. It is clear that the accuracy of the OLSLR model is better than that of RF model. This result agrees with that achieved by [18]. The problem we deal with is basically linear problem. For this reason, OLSLR could successfully extrapolate good predictions. On the other hand, the nature of RF is nonlinear model. It sometimes may not efficiently predict linear predictions. That is clearly shown in the results of the groundwater levels of Wanzarik oasis in the period 2014 and 2015.

Table 3. List of MSE and MAE errors of RF and OLSLR models

| Oasis | RF | | OLSLR | |
|---|---|---|---|---|
| | MSE | MAE | MSE | MAE |
| Burgan | 0.20 | 0.45 | 0.01 | 0.08 |
| Wanzarik | 0.03 | 0.13 | 0.10 | 0.17 |
| Umm-Aljdawal | 1.77 | 1.22 | 0.53 | 0.41 |
| Ashkda | 0.23 | 0.44 | 0.08 | 0.16 |

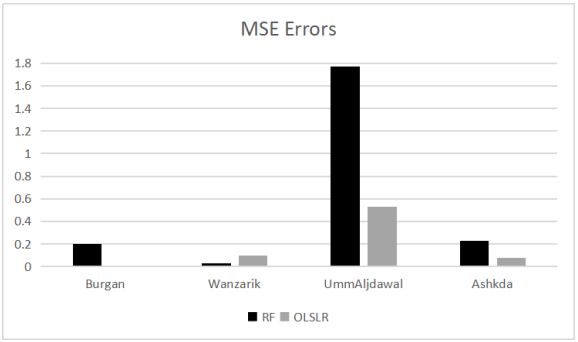


Figure 9. The Mean Square Errors (MSE) of RF and OLSLR models

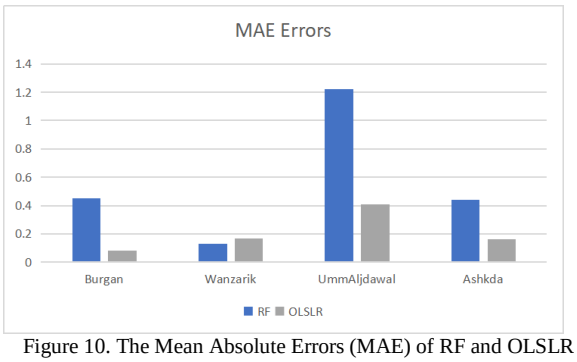


Figure 10. The Mean Absolute Errors (MAE) of RF and OLSLR models

Finally, the predictions are compared with that obtained from the traditional HLTF method. The results are depicted in Figure. 11. The OLSLR model produces better agreement with the HLTF method than the RF model.

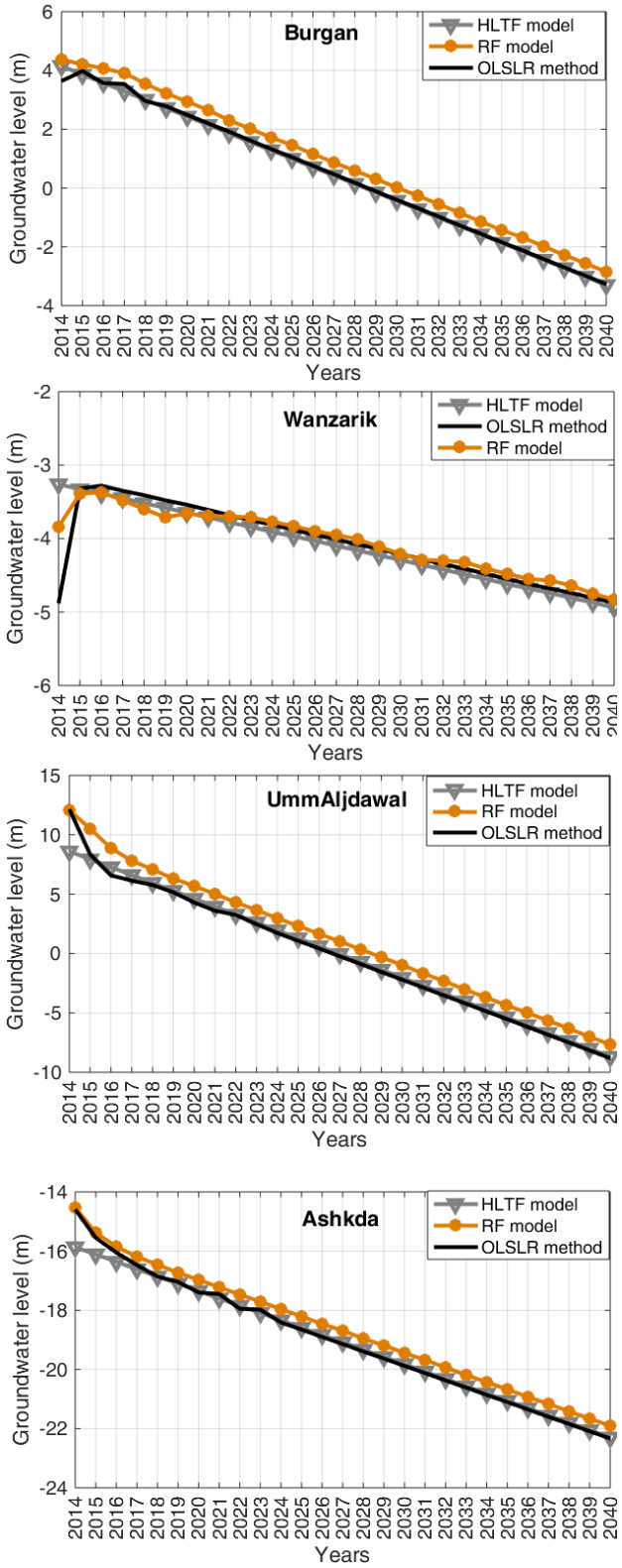


Figure 11. The predictions of the groundwater depth of each oasis

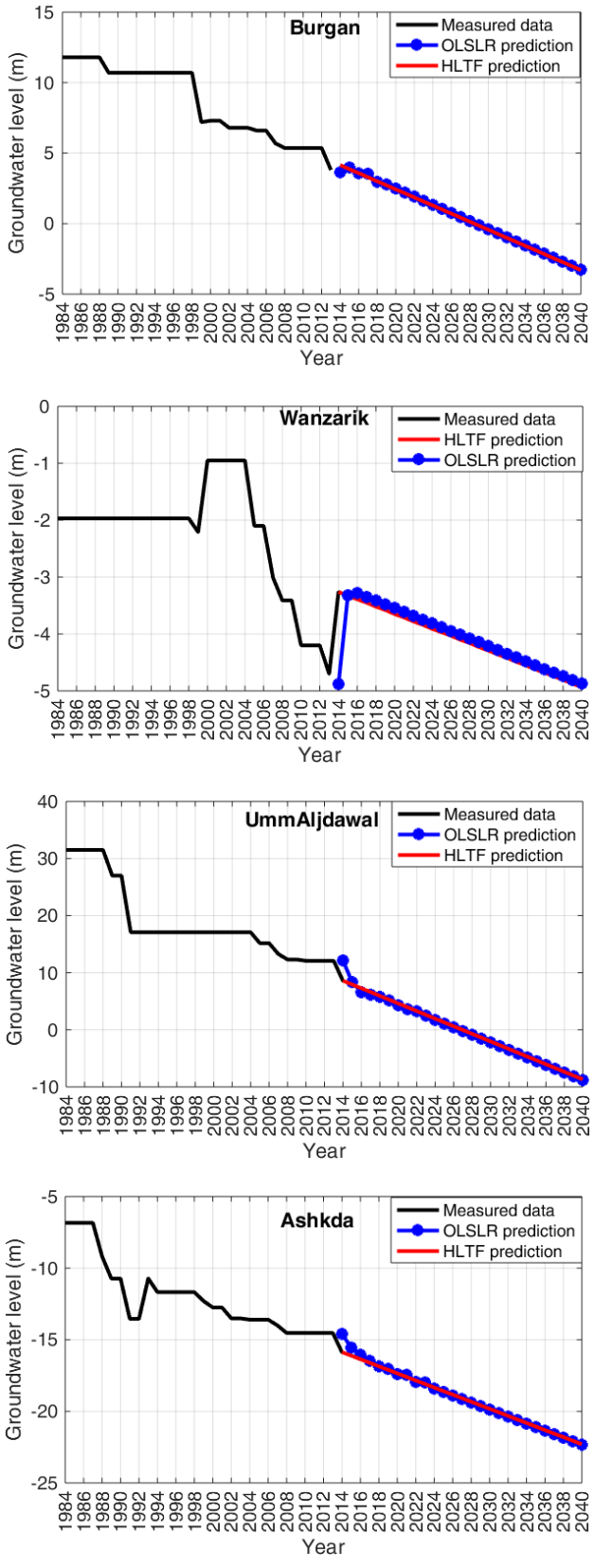


Figure 12. Combination of the measured and predictions of the groundwater depth

The overall predictions of the groundwater levels are shown in Figure. 12. It is obvious from the predictions that the aquifers are continuously depleted over the years reflecting limited and gradually degraded water resource. That is related to the population growth which increases the water demand. However, there is an urgent need to develop alternative water resources such as treated wastewater, rainwater harvesting and desalination of seawater.

Furthermore, it is obvious that from Figure. 12 the groundwater levels of Wanzarik oasis rises for the years 2000 – 2004. That is due to two reasons, first the damage of the well's engines which prevents water withdrawal. During this period the wells are fed by rainwater in the winter. The second reason might be changing the measuring team or their measurement equipment.

## VII. CONCLISION

This paper provides a comparison between RF and OLSLR models with the HLTF method for predicting the time series of groundwater level of some wells in four Oases in the west of Libya. In this study, the behavior of groundwater levels of the wells is mostly linear, Thus the results of the OLSLR agree well with that of the HLTF method. The basic information extracted from the predictions is that there is an increase in the groundwater withdrawal in the area of the study. This problem leads to deepening the wells which causes an increase in the cost of the water production and as a consequence an environment destruction. It is recommended to develop alternative water resources.

## REFERENCES

[1] Libya's water supply, "*Plumbing the Sahara",* https://www.economist.com/graphic-detail/2011/03/11/plumbing-the-sahara, May 2011,

[2] تقييم قدرات مؤسسات إمداد المياه في ليبيا-الملخص "الهيئة العامة للموارد المائية التنفيذي", 2019, pp. 5.

[3] E. A. Hussein, C. Thron, M. Ghaziasgar, A. Bagula and M. Vaccari, "*Groundwater Prediction Using Machine-Learning Tools*", Scholarly Journal, vol. 13, no. 11, 2020, DOI:10.3390/ a13110300.

[4] A. Mosavi, P. Ozturk and K. Chau, "*Flood Prediction Using Machine Learning Models: Literature Review*", Water 2018, 10, 1536. https://doi.org/10.3390/w10111 536.

[5] S. Kordrostami, M. A Alim, F. Karim and A. Rahman, "*Regional Flood Frequency Analysis Using an Artificial Neural Network Model*", Geosciences, vol. 10, no. 4, 2020, 127; https://doi.org/ 10.3390/ geosciences10040127.

[6] R. Acar, S. Çelik and S. Senocak, "*Rainfall intensity-duration-frequency (IDF) model using an artificial neural network approach*", Journal of Scientific and Industrial Research, vol. 67, no. 3, March 2008.

[7] W. Palma, "*Time Series Analysis*", Wiley, 2016, pp. 298.

[8] O. Valenzuela, F. Rojas, L.Javier Herrera, H. Pomares and I. Rojas, "*Theory and Applications of Time Series Analysis: Selected Contributions from ITISE 2019*", Springer International Publishing, Nov 21, 2020.

[9] R. S. Gupta , *Hydrology & Hydraulic Systems*, 2ed edition, 2017, pp. 5-10.

[10] L. Breiman, "*Random forests. Machine Learning*", 45, pp. 5–32, 2001. https://doi.org/10.1023/A:1010933404324.

[11] IBM Cloud Education, "*Supervised Learning*", August 2020 https://www.ibm.com/cloud/learn/supervised-learning

[12] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton, 1984.

[13] M. Gupta_OMG, "*ML: Linear Regression*", Sep. 2018, https://www.geeksforgeeks.org/ml-linear-regression/

[14] 24أخبـار ليبيـا, "*أنفوغرافيك / تقرير سكان ليبيا حسب المناطق لسنة 2020*", https://akhbarlibya24.net/, April 2021.

[15] A. A. Naji, "*Environmental impact of high obstruction of ground water in WadiAlshati region*", M.S. dissertation, Dept. Civil ENg., University of Sabha, 2019.

[16] "*General Water Desalination Company*", https://www.libyaobserver.ly/general-water-desalination-company

[17] United Nations, "*Country Intervention on Drought- Libya*", https://sdgs.un.org/documents/country-intervention-drought-libya-19926, Department of Economic and Social Affairs Sustainable Development, retrieved Sep. 2021.

[18] K. Lin, Q. Lin, C. Zhou and J. Yao, "*Time Series Prediction Based on Linear Regression and SVR"*, Third International Conference on Natural Computation (ICNC 2007), 2007, pp. 688-691, doi: 10.1109/ICNC.2007.780.