2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)

A Water Level Prediction Model Based on ARIMA-RNN

[1]Guoyan Xu, [2]Yi Cheng, [3]Fan Liu, [4]Ping Ping, [5]Jie Sun
College of computer&Information Engineering
Hohai University
NanJing, China
[1]gy_xu@126.com, [2]1411681130@qq.com, [3]fanliu@hhu.edu.cn, [4]pingpingnjust@163.com, [5]820329829@qq.com

***Abstract*—Accurate water level prediction has important guiding value for scientific decision-making and planning. In order to fully excavate the information in the water level and improve the accuracy of water level prediction, a new water level prediction method is proposed. On the basis of ARIMA and RNN model, a new scheme based on ARIMA-RNN combined model for water level prediction is proposed. This method solves the problem that a single forecasting model can't take into account both linear and nonlinear components in data, and also solves the problem of precision reduction caused by simple addition of linear and nonlinear components in traditional combination schemes. In this scheme, the linear correlation components in the data are predicted by the Autoregressive Integrated Moving Average Model (ARIMA), and the nonlinear components are predicted by the Recurrent Neural Network (RNN), and the relationship between the two components is constructed by RNN. The model uses the data of daily water level and environmental factors related to water level as input vectors, and the water level in the next 30 days as output vectors. Experiment in Taihu Lake proves the validity of the model. ARIMA and RNN models were established to predict water level, and the results were compared with the results of the model proposed in this paper. It was found that the RMSE of ARIMA-RNN model was the smallest. The experimental results proved that the prediction model proposed in this paper could achieve better results.**

Keywords- water level, ARIMA, RNN, prediction

I. INTRODUCTION

Accurate prediction of water level is an important prerequisite for scientific decision-making and planning. The purpose of water level prediction is to establish accurate prediction model and reveal the changing law of rivers and lakes. Water level data are usually constructed into hydrological time series, and the prediction model of hydrological time series is used to predict water level. By predicting the water level, the information hidden in the data can be analyzed, which is of great significance to the allocation and management of water resources and the decision-making of flood control and disaster reduction.

At present, researchers at home and abroad have done some research on water level prediction and put forward many classical prediction algorithms. Box and Jenkins [1] proposed autoregressive integrated moving average model (ARIMA) in the 1970s, and systematically introduced the principles and methods of ARIMA applied in time series prediction. Birylo and Rzepecka et al. [2], combined with precipitation, surface runoff and evapotranspiration at the test site, established the ARIMA model to predict the groundwater level, analyzed the 12-month cycle, and obtained good results. Mirzavand [3] applied autologous regression (AR), moving average (MA), autoregressive moving average model (ARMA) and ARIMA to model and predict the fluctuation of groundwater level, proving that ARIMA model has advantages in the prediction of groundwater level. Zhen Yu and Guoping Lei [4] proposed the ARIMA model for daily water level prediction of three stations in the middle reaches of the Yangtze river, but the results showed that the accuracy of the model decreases with the extension of the prediction period. The ARIMA model only provides an effective short-term water level prediction method, but it does not play a good role in the prediction of long-term water level.

In recent years, artificial neural network (ANN) has been continuously developed and gradually applied to water level prediction. Neural network can capture and represent the nonlinear relationship between data input and output, which makes up for the limitation that traditional prediction models can only obtain linear relationship. Therefore, it is of great significance in water level prediction. Many researchers also use artificial neural network to predict water level [5,6]. Recurrent neural network (RNN) can identify the historical law between the nonlinear input and output of water level, and provide an effective method for the correlation between the nonlinear input and output of water level. Huang [7] established the RNN model and tested it in the south coast of Changdao. It provides a wonderful long-term prediction for the tidal and non-tidal water levels at the regional coastal entrance. In order to predict the hydrometeorological changes and the future lake level during human activities in Lake Euguedil, Veysel [8] constructed a RNN with multiple input structures, and established classical stochastic model, autoregressive model and autoregressive moving average model, which proved that RNN model has high accuracy and reliability in the prediction of water level change.

Due to the influence of various external factors, hydrological time series contains both linear and nonlinear components, and a single prediction model is often unable to accurately predict the water level. Therefore, some researchers try to combine different types of prediction

978-1-7281-0059-3/19/$31.00 ©2019 IEEE
DOI 10.1109/BigDataService.2019.00038

models to improve the prediction performance [9,10,11,12]. This paper proposes an ARIMA-RNN model for water level prediction of Taihu Lake. At the same time, the traditional model combination method is improved in this paper, which can not only mine the hidden information in the data comprehensively, but also improve the prediction effect.

II. Related Work

A. ARIMA Model

Autoregressive integrated moving average model (ARIMA) is built on the basis of historical data, the quantity of data determines the accuracy of the model. Its basic principle is to approximate the time series with a certain mathematical model, and to predict the future values with past values and current values. In this model, three parameters need to be set, that is, autoregressive term order, sequence difference order and moving average term order, the model can be expressed as

$$x_t = \mu + \varphi_1 x_{t-1} + \cdots + \varphi_p x_{t-p} - \theta_1 \varepsilon_{t-1} \cdots - \theta_q \varepsilon_{t-q} \quad (1)$$

where p represents autoregressive term order; d represents sequence difference order; q represents moving average term order; x_t refers the value of water level observed at time period t; $\varphi_p x_{t-p}$ represents the lagged values of x; $\theta_q \varepsilon_{t-q}$ represents the lagged errors.

Akaike information criteria (AIC) is used to evaluate the performances of models and select the degree p and q of an ARIMA model. AIC can be expressed as

$$AIC\,(k) = n \ln{(MSE)} + 2k \quad (2)$$

where n is the number of data point; k represents the number of free parameters; MSE presents mean square error.

B. RNN Model

Recurrent Neural Network introduces the concept of time series into the structure of the network, which has excellent adaptability in processing time series of water level. In an RNN network, the current input of a sequence is also related to the previous input. RNN will remember the previous information and apply it to the current calculation, that is, the input of the hidden layer will not only have the output of the input layer, but also include the output of the hidden layer at the previous time. The RNN consists of an input layer, a hidden layer, and an output layer. Its basic network structure is shown in Figure 1.

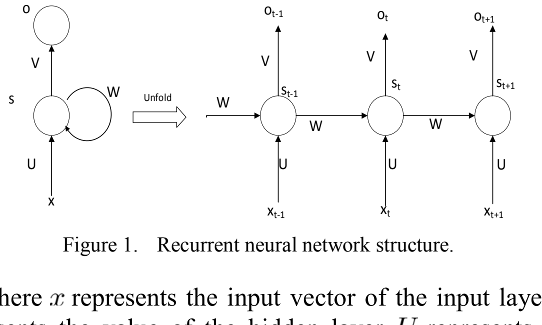

Figure 1. Recurrent neural network structure.

where x represents the input vector of the input layer, s represents the value of the hidden layer, U represents the weight matrix of the input layer to the hidden layer, o represents the output vector of the output layer, V represents the weight matrix of the hidden layer to the output layer, W represents the weight between the previous value of the hidden layer and this input. x_t is the input of time period t, $s_t$representing "memory" at time period t, o_t indicates the output at time period t. Obviously, we can see that the output of a certain time point not only relates to the input of the current time, but also to the previous state.

III. ARIMA-RNN Prediction Model

In this section, we first give an overview of our approach, which is the main step of establishing ARIMA-RNN. Then, the Principle of the model and all the steps are detailed in the following subsections.

A. Overview

Both ARIMA and RNN can predict water level, but their accuracy is not optimal. Since ARIMA model plays an important role in time series analysis, it is chosen to predict the linear components of water level data. Since the nonlinear activation function of RNN is able to analyze the nonlinear components in the water level, and RNN can capture long-term dependent information, it is considered to add RNN to the model. Therefore, the ARIMA-RNN model is proposed to improve the accuracy of prediction. Figure 2 shows the main steps of building ARIMA-RNN.

Step1: Data collection and extraction. Collect data from relevant departments, and divide the data into water level data and environmental data.

Step2: Build ARIMA model. Verify the stationarity of the data and get the parameters through AIC. Then train the model, the water level can be predicted and the residual can be calculated.

Step3: Use RNN model to predict residual. Normalize the environment data and residual data, then construct the training data, and we can get an RNN to predict residual sequence.

Step4: Predict future water level. The relationship between predicted value of residual, predicted value of water level and actual value can be determined through RNN network, and we can predict the water level of the next 30 days.

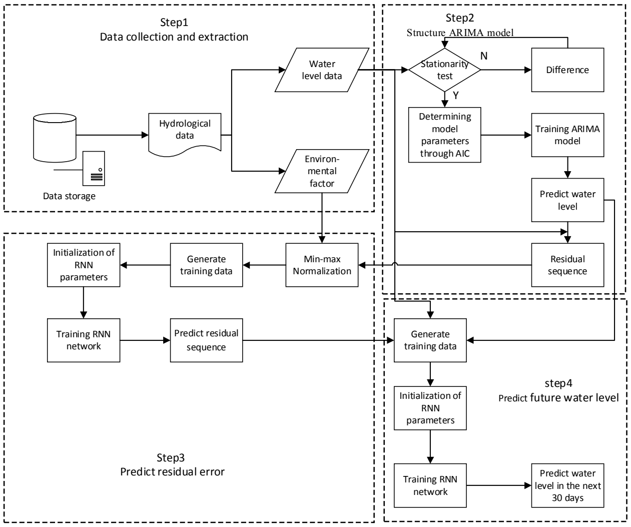

Figure 2. Process of establishing ARIMA-RNN model

B. Principle of ARIMA-RNN Model

A combined model is a way to deal with a problem while using two or more algorithms. In hydrological time series, there are both linear features and nonlinear features. A single predictive model cannot fully acquire the features. Therefore, the combined model should be used to capture linear features and nonlinear features to improve the prediction effect of the prediction model.

In the traditional way of combination, the combination can usually be expressed as follows:

$$x_t = L_t + N_t \qquad (3)$$

where x_t represents the hydrological time series; L_t represents the linear component in the hydrological time series; N_t represents the nonlinear component of the hydrological time series. However, in the traditional method, the model is constructed based on the addition of the linear component and the nonlinear component, but if the linear component and the nonlinear component have more complex relationships than simple plus, the accuracy of the model will be greatly reduced, and may be worse than single prediction model.

In this paper, a method is proposed to improve the traditional combinatorial method. This method uses a function to represent the relationship between linear components and nonlinear components in time series. The formula is as follows:

$$x_t = f\left(L_t, N_t\right) \qquad (4)$$

The main process of establishing the ARIMA-RNN model based on formula (4) is as follows:

1) Establish an Autoregressive Integrated Moving Average Model to process linear components in hydrological time series to obtain a sequence of predicted values. At the same time, the residual sequence of the nonlinear component can be obtained. The formula is as follows:

$$e_t = x_t - \widehat{L_t} \qquad (5)$$

where $\widehat{L_t}$ represents predicted values; e_t represents the residual sequence of the nonlinear component.

2) Use RNN to establish the prediction model to predict the nonlinear components and we can get the predicted values of the residual sequence. The formula is as follows:

$$N_t = f_1\left(e_{t-1}, e_{t-2} \cdots e_{t-n}\right) \qquad (6)$$

3) Reconstruct an RNN model to fit x_t, this model can get the relationship between linear components and nonlinear components. The relation between them is expressed as follows:

$$x_t = f_2\left(L_t, N_t\right) \qquad (7)$$

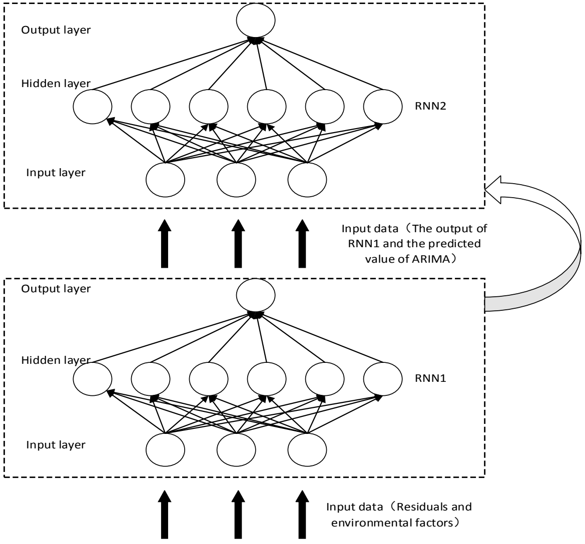
Output layer
Hidden layer
Input layer
RNN2
Input data（The output of RNN1 and the predicted value of ARIMA）
Output layer
Hidden layer
Input layer
RNN1
Input data（Residuals and environmental factors）

Figure 3. Network structure of the model

The combination of the two RNN networks is shown in Figure 3. This paper uses a new method to build hybrid model, which can better describe the relationship between linear components and nonlinear components, so it is better than traditional combination method in theory.

C. Algorithm and Analysis

The combined model in this paper is mainly composed of one ARIMA model and two RNN models. The combined structure of two RNNs is shown in figure 3. The ARIMA model mainly predicts the original time series, which can get the predicted value of water level and the residual value. The first RNN fits the residual sequence to obtain the predicted value of nonlinear components; the second RNN fits the linear component and the nonlinear components to get the relationship between them.

In the water level time series, the process of modeling ARIMA-RNN is as follows:

1) Use the unit root test method to judge whether the daily data of water level meets the requirement of stationarity. If the sequence is not stable, the data needs to be differentially operated, and then continue to judge until the sequence satisfies the stationarity requirement;

2) Use the AIC information criterion to get the autoregressive term order and moving average term order. Go through all the AIC values and get the minimum, and we can get the best parameters.

3) According to the optimal parameter set determined in the previous steps, build and train the ARIMA model, which

can predict the value of water level and can calculate the residual value.

4) The data used in the model should be normalized first, and the data include residual data, the extracted temperature and precipitation data. These data should be further processed before we can use them, they should be constructed into a two-dimensional matrix according to the format of the time window to get the training sets that can be used. The training set is in the following form:

$$trainX = \begin{bmatrix} e_1 & e_2 & \cdots & e_5 \\ e_2 & e_3 & \cdots & e_6 \\ \vdots & \vdots & \cdots & \vdots \\ e_{t-5} & e_{t-4} & \cdots & e_{t-1} \end{bmatrix} \tag{8}$$

$$trainY = \begin{bmatrix} e_6 \\ e_7 \\ \vdots \\ e_t \end{bmatrix} \tag{9}$$

5) Construct the first RNN network and initialize network parameters. Input the training data into the network for training, and when the number of iterations exceeds the set number, the training will end. After the anti-normalization of the output data, the predicted value of the residual sequence can be got.

6) Construct the second RNN network, normalize the input data and construct a training set of the data. The training set has the following form:

$$trainX = \begin{bmatrix} L_1 & N_1 \\ L_2 & N_2 \\ \vdots & \vdots \\ L_t & N_t \end{bmatrix} \tag{10}$$

$$trainY = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_t \end{bmatrix} \tag{11}$$

where L_t is the predicted value of ARIMA at time t, which represents the linear component of the original water level time series. N_t is the predicted value of the first RNN at time t, which represents the nonlinear component in the sequence. x_t represents the actual value at time t.

7) Initialize the parameters of the second RNN, and input the training data into the network for training.

8) Get the output of the model, reverse Normalization and get the predicted water level in the next 30 days.

The algorithm of the ARIMA-RNN model is as follows:

Algorithm1: Using ARIMA-RNN algorithm to predict water level of the next 30 days

```
Input: twl is training data set, H is environmental data
Output: predicted water level of the next 30 days
// Construct ARIMA model
adf=ADF (twl); // Stationary judgment
diff = 0
while adf [1]>=0.05 do
    diff=diff+1;
    adf=di(twl,diff)//Differential operation
end while
d=diff
for p=0 to pmax do
    for q=0 to qmax do
        m =ARIMA(twl,(p,d,q)) //Build ARIMA model
        AIC = aic(m) // Calculate the AIC
    end for
end for
p,q = idxmin(AIC)  // Get the optimal p and q
m = ARIMA(twl,(p,d,q))   //Determine the model
L,e= m.predict()      //Predict and calculate residuals
// Construct the first RNN network to predict residual
e_z,H_z =zscore(e,H)   // Normalization processing
trainX,trainY=create_dataset(e_z,H_z,look_back)      // Construct the training set
Create RNN1_model by Sequential()      // Construct RNN model
Initialize the parameters of RNN
RNN1= train（RNN1_model, train_X, train_Y）   // Train the RNN model
Get residual sequence predictions
//Build a second RNN model to determine the relationship between(L,N)and x
train=zscore(L,N,x)
train_X, train_Y=create_dataset(train)
Create RNN2_model by Sequential()
Initialize the parameters of RNN
RNN2= train（RNN2_model, train_X, train_Y）
//Predict future data
test = test_data()   //Construct data set
out = predict(test)   // Get output
Reverse normalization and predict the water level in the next 30 days
```

IV. Experiment and Analysis

A. Data Set and Indicator Selection

The experimental data set adopted in this paper is the daily average water level of Taihu Lake of Jiangsu Province in China from May 1, 2009 to September 2, 2018. The data set is divided according to the proportion of 9:1. The training data is sufficient to ensure that the model can achieve good results. Considering the influence of environmental factors, this paper adds the temperature and rainfall data to the RNN model to increase the accuracy of the model. The experimental software environment is Python 3.6.6, Windows 10 Ultimate operating system, and the hardware environment is 2.8GHz, 8G memory laptop. The result of ARIMA, RNN and ARIMA-RNN will be compared in the experiment to determine the accuracy of the prediction. Figure 4 is a line chart of the daily water level in Taihu Lake.

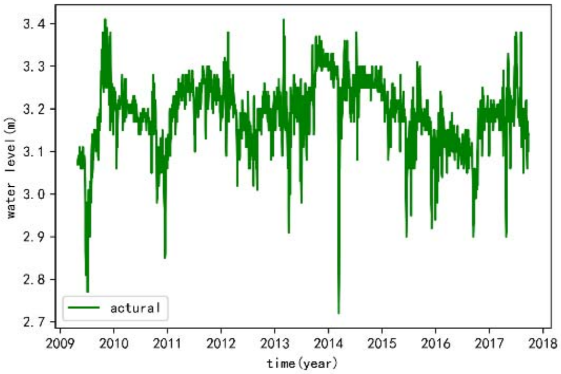

Figure 4. Water Level of Taihu Lake

In order to compare the experimental result, scientific evaluation indexes should be selected. In this paper, Root Mean Square Error (RMSE) and Forecast Accuracy (FA) are used as the evaluation criteria of the prediction model. These two evaluation indicators describe the prediction accuracy of the model from different angles. Smaller RMSE represents better predict effect. The formula of RMSE is described as follows:

$$RMSE = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}\left(x\_pred_i - x\_origin_i\right)\right)} \quad (12)$$

where x_pred_i represents the predicted value; x_origin_i represents the actual value; n is the number of the samples. FA indicates the proximity between predicted and actual values. The closer the value is to 1, the better the prediction model is. The formula of FA is described as follows:

$$FA = 1 - \frac{|x\_pred_i - x\_origin_i|}{x\_origin_i} \quad (13)$$

B. Experimental Results and Comparative Analysis

In this paper, the water level data and associated temperature and rainfall information are used as input of the model to predict the future water level. Figure 5 is the fitting result of predicted value and actual value, we can find that the model can fit the data well. Figure 6 is the predicted result of ARIMA-RNN model for the next 30 days.

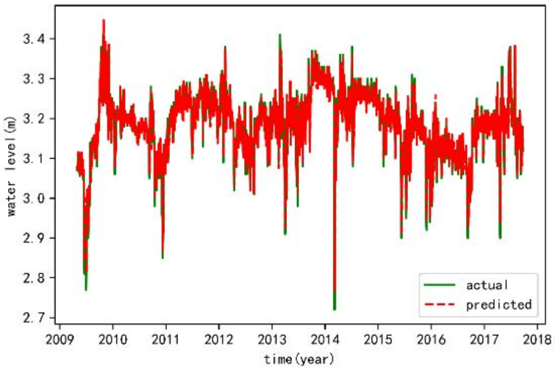

Figure 5. Combined model fit diagram. "actual" represents the actual water level and "predicted" represents the fitted water level.

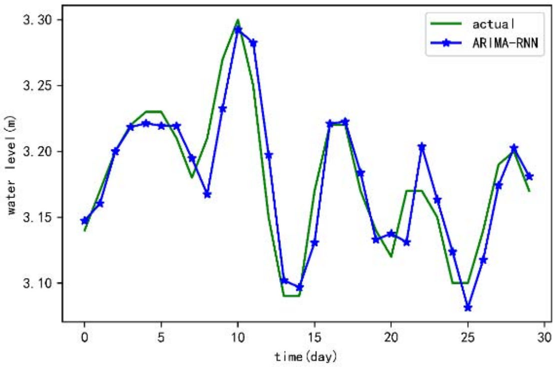

Figure 6. Prediction results of the ARIMA-RNN model. "actual" represents the actual water level for the next 30 days and "ARIMA-RNN" represents the predicted water level for the next 30 days.

In order to analyze the effect of the model easily, this experiment uses ARIMA, RNN and ARIMA models to predict and compares the predicted water level of different models in the next 30 days. Figure 7 is the predicted result of different models for the next 30 days.

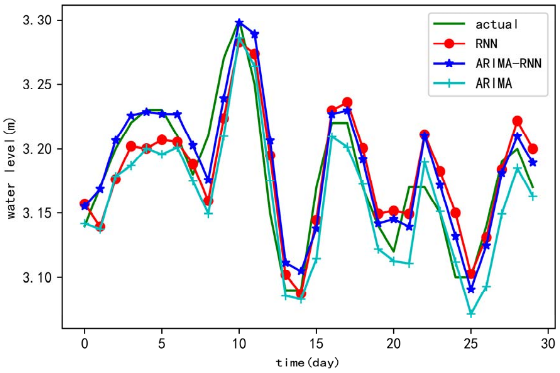

Figure 7. Comparison result for three models. "actual" represents the actual water level of the test data, "RNN" represents the predicted results of the RNN model, "ARIMA" represents the predicted results of the ARIMA model, "ARIMA-RNN" represents the predicted results of the ARIMA-RNN model.

It can be seen from Figure 7 that the ARIMA model has a good predictive effect on the water level at the beginning, but it can only depict the general trend a few days later, and the RNN and ARIMA-RNN models perform better. By observing the experimental result, it can be found that the prediction effect of the ARIMA-RNN model is better than RNN, and the predicted value is closer to the actual value. It can be proved that this new model does improve the accuracy of the prediction to a certain extent.

Using RMSE and FA to evaluate different models. The evaluation result of these models is shown in table 1.

TABLE I. PERFORMANCE EVALUATION FOR DIFFERENT MODELS

MODEL	RMSE	FA
ARIMA	0.047	97.24%
RNN	0.033	98.61%
ARIMA-RNN	0.021	99.17%

Based on the experimental result of the ARIMA, RNN and ARIMA-RNN model for water level prediction in the next 30 days, the RMSE of the ARIMA-RNN model is smaller, and the FA value is closest to 1, Therefore, it can be seen that the prediction result based on ARIMA-RNN proposed in this paper is optimal. From the results, we can see that compared with other models, ARIMA-RNN model has better prediction effect on future water level, and can grasp the overall trend and amplitude fluctuation more accurately. The experiment in this paper proves the advantage of the combined model in the prediction of water level, shows its scientific and rationality, and performs better in the prediction of actual model construction. Therefore, in the hydrological system, this model can achieve a good effect on water level prediction

V. Conclusion

In this paper, a prediction method based on ARIMA-RNN for water level information is proposed. The original water level sequence is predicted by ARIMA to obtain the predicted value and residual, and then RNN is used to fit the residual sequence. This paper also improves the combination of traditional models, using RNN to determine the relationship between linear and nonlinear components, instead of simple addition. From the experimental result, the combined model has obvious advantages in the time series of water level. This model can be applied to hydrological station, which can improve the effect of water level prediction. However, the water level data of one site is associated with other related sites in space, the model needs further optimization to improve the prediction accuracy.

Acknowledgment

This research is partially supported by the Water Resources Science and Technology Projects in Jiangsu Province under Grant No. 2017065 and the National Key R & D Program of China under Grant No. 2018YFC0407106.

References

[1] Box, G.E.P. and Jenkins, G.M. (1970) Time series analysis: Forecasting and control. Holden-Day, San Francisco.

[2] Birylo M , Rzepecka Z , Kuczynska-Siehien J , et al. Analysis of water budget prediction accuracy using ARIMA models[J]. Water Science and Technology: Water Supply, vol. 18, no.3 , pp. 819-830, June 2018.

[3] Mirzavand, Mohammad & Ghazavi, Reza. (2014). A Stochastic Modelling Technique for Groundwater Level Forecasting in an Arid Environment Using Time Series Methods. Water Resources Management, vol. 29, no .4, pp. 1315-1328, 2014.

[4] Z. Yu, G. Lei, Z. Jiang and F. Liu, "ARIMA modelling and forecasting of water level in the middle reach of the Yangtze River," 2017 4th International Conference on Transportation Information and Safety (ICTIS), Banff, AB, 2017, pp. 172-177. doi: 10.1109/ICTIS.2017.8047762.

[5] Sung Eun Kim,Il Won Seo. Artificial Neural Network ensemble modeling with conjunctive data clustering for water quality prediction in rivers[J]. Journal of Hydro-Environment Research, vol. 9, no. 3, p p. 325-339, September 1, 2015.

[6] K.S. Kasiviswanathan,Jianxun He,K.P. Sudheer,Joo-Hwa Tay. Potential application of wavelet neural network ensemble to forecast streamflow for flood management[J]. Journal of Hydrology, vol. 536, pp. 161-173, May 01, 2016.

[7] Huang W , Murray C , Kraus N , et al. Development of a regional neural network for coastal water level predictions[J]. Ocean Engineering, vol. 30, no. 17, pp. 2275-2295, December 2003.

[8] Veysel Güldal, Tongal H . Comparison of Recurrent Neural Network, Adaptive Neuro-Fuzzy Inference System and Stochastic Models in E?irdir Lake Level Forecasting[J]. Water Resources Management, vol. 24,pp. 105-128, 2010.

[9] M Khashei, M. Bijari, "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting[J]", Applied Soft Computing, vol. 11, no. 2, pp. 2664-2675, 2011.

[10] A H Nury, K Hasan, M J Alam, B. Comparative, "study of wavelet-ARIMA and wavelet-ANN models for temperature time series data in northeastern Bangladesh[J]", Journal of King Saud University – Science, vol. 29, pp.47-61, January 2017 .

[11] V Nourani, A H Baghanam, J Adamowski et al., "Applications of hybrid wavelet-Artificial Intelligence models in hydrology: A review[J]", Journal of Hydrology, vol. 514, pp. 358-377, 2014.

[12] Y Seo, S Kim, O Kisi et al., "Daily water level forecasting using wavelet decomposition and artificial intelligence techniques[J]", Journal of Hydrology, vol. 520, no. 520, pp. 224-243, 2015.