

Forecasting of Groundwater Level using Artificial Neural Network by incorporating river recharge and river bank infiltration

Mohd Khairul Nizar Shamsuddin^{1,*}, *Faradiella Mohd Kusin*², *Wan Nor Azmin Sulaiman*², *Mohammad Firuz Ramli*², *Mohamad Faizal Tajul Baharuddin*³, *Mohd Shalahuddin Adnan*³

¹Hydrogeology Research Centre, National Hydraulic Research Institute of Malaysia, Lot 5377, Jalan Putra Permai, 43300 Seri Kembangan, Selangor, Malaysia

²Faculty of Environmental Studies, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

³Department of Environmental and Water Resources, Faculty of Civil and Environmental Engineering, Universiti Tun Hussein Onn Malaysia, Johor

Abstract. Groundwater tables forecasting during implemented river bank infiltration (RBI) method is important to identify adequate storage of groundwater aquifer for water supply purposes. This study illustrates the development and application of artificial neural networks (ANNs) to predict groundwater tables in two vertical wells located in confined aquifer adjacent to the Langat River. ANN model was used in this study is based on the long period forecasting of daily groundwater tables. ANN models were carried out to predict groundwater tables for 1 day ahead at two different geological materials. The input to the ANN models consider of daily rainfall, river stage, water level, stream flow rate, temperature and groundwater level. Two different type of ANNs structure were used to predict the fluctuation of groundwater tables and compared the best forecasting values. The performance of different models structure of the ANN is used to identify the fluctuation of the groundwater table and provide acceptable predictions. Dynamics prediction and time series of the system can be implemented in two possible ways of modelling. The coefficient correlation (R), Mean Square Error (MSE), Root Mean Square Error (RMSE) and coefficient determination (R^2) were chosen as the selection criteria of the best model. The statistical values for DW1 are 0.8649, 0.0356, 0.01, and 0.748 respectively. While for DW2 the statistical values are 0.7392, 0.0781, 0.0139, and 0.546 respectively. Based on these results, it clearly shows that accurate predictions can be achieved with time series 1-day ahead of forecasting groundwater table and the interaction between river and aquifer can be examine. The findings of the study can be used to assist policy marker to manage groundwater resources by using RBI method.

*Corresponding author: nizar@nahrim.gov.my

1 Introduction

River Bank Infiltration (RBI) is a natural filter process to improve the drinking water quality obtained from surface water flow through aquifer and mixture with groundwater. Surface waters are infiltrated through the aquifer media in the pumping wells during pumping activities and subsequently influence river and groundwater level [1-7] stated that river and aquifer interactions are exist when waters from vertical and horizontal wells that are located in alluvial adjacent to rivers/lakes are pumped. It must be managed in an integrated way to provide efficient water supplies and also concerns over the conservation of the natural environment. Changing of river water levels and groundwater level will effect flow and contaminant transport through infiltration in aquifer media [1]. The dynamic fluctuation of the river water levels during pumping activities does not only influenced by soil sediments clogging of the riverbed but also other factors such as climate, hydrogeological properties and pumping rates [5-6]. Therefore, groundwater tables monitoring is essential to be conducted for understanding river-aquifer interactions. Groundwater aquifer located closed to river have direct influence on river water level. At distances more than a few hundred meters from the river, the groundwater table depth is generally several meters higher than the river water level and subsequently does not effect to river water levels. In addition, groundwater tables are also not affected immediately by river water depth increase due to very low groundwater velocity. On the other hand, the other factors such as precipitation and stream flow can influence quickly the river water depth.

Network design and monitoring of groundwater tables depends on hydrogeological conditions and available logistical resources. Although mathematical and conceptual models are the main tools for representing hydrological variables and understanding the physical processes in a system but it has practical limitations [8]. When the data is not sufficient and accurate prediction is required, artificial neural network (ANN) models can be a good option [8]. The ability of ANNs to identify the relationship of hydrological variables patterns makes ANNs sufficiently to solve complex hydrologic problems [9]. Recently, ANNs have been successfully applied for identifying the temporal data to calculate groundwater level [10].[11-15] has developed an ANN model for hydrology and hydrogeology.

The aim of this study is to look of the applicability of ANNs in forecasting groundwater tables in a site at Jenderam Hilir, Selangor, Malaysia. The Jenderam Hilir is where a pilot project to develop a better understanding of sustainable water resources, and introduce RBI in Malaysia is located. This site was chosen due to the high water demand in the area and groundwater is seen as one of the source with very high potential to be developed as supplementary sources to meet the high public water supply demand. The applicability of the methodology is demonstrated when using ANN training algorithms namely Lavenberg-Marquardt (LM) algorithm for predicting groundwater tables in a Jenderam Hilir located in the tropical humid countries. The major focus of the current study is to investigate the potential of ANN approach in groundwater modeling focus on water level fluctuation and to predict the relationship between river and aquifer of the study area.

2 Materials and methods

2.1 Study area

The study area is located in a flat area of Langat River Basin at the confluence of Semenyih Rivers, Langat Rivers and a small Jenderam Hilir River flow to Langat River in the area.

These two rivers are the main source of raw water for Selangor state. This area is a former tin mining area in which there are three major ponds (ponds A, B and C) which are interconnected (Fig. 1). These three ponds function as storage to increase the capacity of raw water at the water intake. The study area is bounded by the hilly areas in the north and east, and by the sea in the southwest. The maximum elevation is 297 m and the minimum is 7.981 m, above mean sea level. The exact location is between latitude 2° 53' 28.56" N and 2° 53' 39.75" N, and longitude 101° 42' 03.78" E and 101° 44' 14.58" E, covering an area of 10km². The daily average temperatures vary from 27-30°C. This area is experiencing two monsoon periods; Northeast monsoon from October to January (wet season) and the Southwest monsoon from May to September (dry season). The annual precipitation in 2015 is 2,013 mm and 70-80% of annual precipitation is concentrated in November and December. The mean stream flow is 47.13 m³/s and the highest river water level is about 5.64 m. The daily average of rainfall, stream flow, temperature, river water level were obtained from the Drainage Irrigation Department, Malaysia (DID) and groundwater levels were measured from the two pumping test wells, DW1 and DW2. The data (rainfall, stream flow, temperature, river water level, groundwater level) used for this study are from February 2015 to March 2016.

2.2 Design of ANN, Training Algorithms and Feed forward neural network (FNN)

A ANN model is characterised by its architecture that represents the pattern of connection between nodes, its method of determining the connection weights and the activation function [16] and will be used to reproduce groundwater fluctuations for long periods using the observed time series data within twelve months of monitoring period (February 2015 to March 2016). In order to perform the system identification, the neural model is first trained to perform 1-day a-head predictions of the groundwater table using previously observed groundwater level. Once this autoregressive model has been developed, simulations were carried out by feeding back its output as the simulation times increases. In this study feed forward neural network architecture has been used and coupled with ANN training algorithms and Levenberg–Marquardt (LM) algorithm. These algorithms have been used to identify a suitable procedure which performs to predict daily groundwater levels over the study area. Feed forward neural networks architecture and the corresponding learning algorithm can be viewed as a generalisation of the popular least-mean-square (LMS) algorithm [17]. A multilayer perceptron network consists of an input layer, one or more hidden layers of computation nodes, and an output layer. Fig. 2 shows a typical feed forward network with one hidden layer consisting of three nodes, four input neurons and one output. The input signal propagates through the network in a forward direction, layer by layer. The various parameters, which affect the performance of network, are number of hidden nodes, number of hidden layers, learning rate and momentum factor. The ANN architecture consist of an input layer with three nodes, a hidden layer with varying nodes and one output layer with single node, thus leading to a multi input single output network. In the present study, the 406 input nodes represent initial groundwater levels at the two sites, daily rainfall, average daily river stage, average daily stream flow and daily temperature (Fig. 3). The two output nodes represent groundwater levels at the two sites in the next time step (i.e., 1 day a head). The input data is the groundwater level during time 1 day ahead, 't-4', 't-3', 't-2', 't-1' and 't' and output data is groundwater level during time 't'. The structure of the neural network was determined by trial and error. The optimal number of nodes in the hidden layer and the stopping criteria were optimised by trial and error for obtaining accurate output. The activation function of the hidden layer and output layer was set as log-sigmoid transfer function as this proved by trial and error to be the best

among a set of other options. In this study, supervised type of learning with a batch mode of data feeding was used in ANN modelling by available data. The data length of 406 data sets has been dividing into 70% for training and 30% for testing. The entire ANN modelling was performed by using Neural Works Predict software and EXCEL.

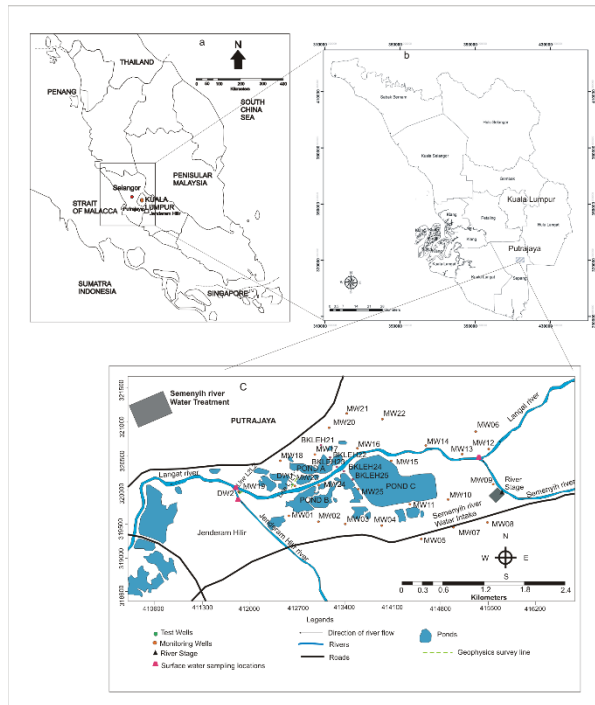


Fig. 1. Location of study area. a. Location of Selangor (box) within the Peninsular Malaysia. b. Location of Jenderam Hilir (shaded) in Selangor. c. Details of the study area including the locations of the Langkat River, Semenyih River, monitoring wells and pumping wells.

2.2.1 Dynamic and Time series prediction

The temperature and groundwater level at time t at the DW1 and DW2, is a function of the past groundwater level at real time, the rainfall, stream flow and river water level from the Dengkil station. The lag is still at one day. In the time series prediction, time series are a sequence of number where in this case they are a sequence of groundwater level form the DW1 and DW2 in daily series or at time, (t) day). The groundwater level at time t , is a function of the past groundwater level at time $t-1$, $t-2, t-3$ and $t-4$ as shown below;

$$Y = f[y(t-1), y(t-2), y(t-3), (y(t-4))] \tag{1}$$

Where $y(t)$ is the models prediction of groundwater level form in the BI site at time t , $y(t-1)$, $y-2)$, $y-3)$ and $y(t-4)$.

2.2.2 Evaluation of ANN efficiency

The predictive performance of ANN are measured by four efficiency terms; the correlation coefficient (R); the mean error (ME), i.e the systematic difference between the predicted and measured values; the mean square error (MSE); and the root mean square error (RMSE). The ANN responses are more precise if R, MSE, RMSE, and ME are found to be close to 1, 0, 0 and 0, respectively.

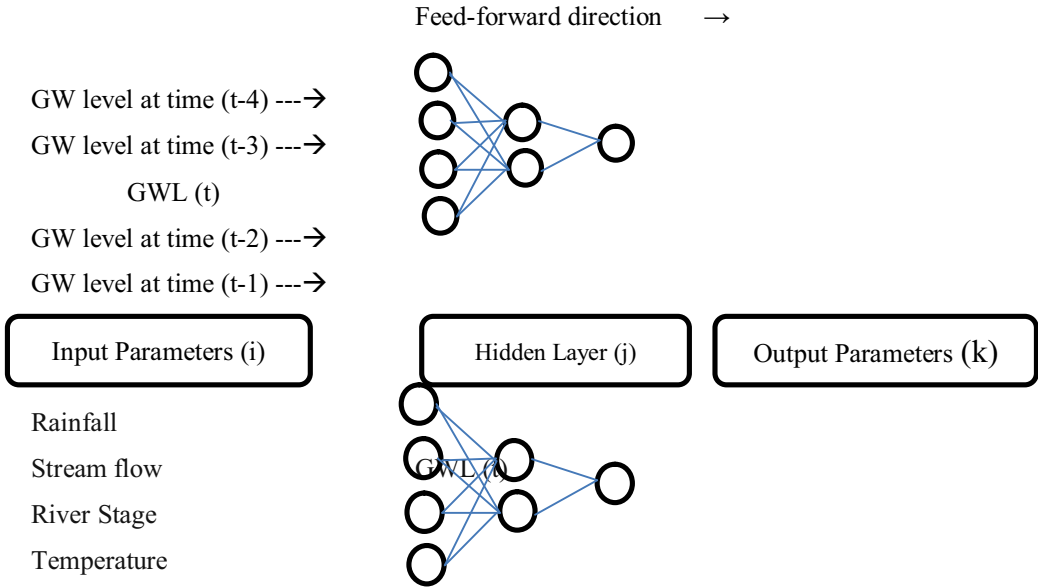


Fig. 2. Typical feed forward neural network of the study

2.2.3 Correlation Coefficient (R)

The correlation coefficient is a commonly used statistical parameter and provides information on the strength of linear relationship between the observed and predicted by the compute value. The value of R close to 1.0 indicates good model performance and can be calculated using equation below:

$$R = \frac{\sum_{i=1}^n (x_{obs} - \hat{x}_{obs})(x_{pre} - \hat{x}_{pre})}{\sqrt{\sum_{i=1}^n (x_{obs} - \hat{x}_{obs})^2 \sum_{i=1}^n (x_{pre} - \hat{x}_{pre})^2}} \quad (2)$$

Where, X_{obs} = observed groundwater \hat{x}_{obs} observed levels=mean of X_{obs} , X_{pre} =predicted groundwater level, \hat{x}_{pre} = mean of X_{pre} and n =the number of data set used for evaluation.

2.2.4 Mean square Error (MSE) and Root Mean Square Error (RMSE)

For every data point, take the difference of the corresponding estimated values, and square the values. Then add up all those values for all data points, and divide by the number of points. The squaring is done so negative values do not cancel positive values. Smaller MSE indicates better prediction of the data. The MSE has the units squared of the parameter estimated.

$$MSE = \frac{1}{n} \sum_{i=1}^n (x_{obs} - x_{pre}) \frac{1}{n} \sum_{i=1}^n (x_{abs} - x_{pre}) \quad (3)$$

It is probably the most easily interpreted statistic, since it has the same units as the parameter estimated. The RMSE is thus the difference, on average, of an observed data and the estimated data.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (x_{abs} - x_{pre})^2}{n}} \quad (4)$$

Different neural networks architectures were developed in order to establish a relationship between the input and output. All the networks were of the feed forward type. The network architectures were trained by varying the number of hidden layer and then by varying number of neurons in each hidden layer.

2.3 Coefficient Determination (R^2) and Residual error (RE)

The coefficient of determination R^2 (or sometimes r^2) is another measure of how well the least squares equation performs as a predictor of y .

$$\hat{y} = b_0 + b_1 x \quad (5)$$

The R^2 is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a graph. The higher the R^2 , the more useful of the model, R^2 takes on values between 0 and 1. The residual error in the results is given by:

$$E_i = y_{obs} - \hat{y}_{pred} \quad (6)$$

Where y_{obs} is the observed and \hat{y}_{pre} the predicted groundwater level. The percentage error of a variable is given by:

$$Ei = \frac{y_{obs} - \hat{y}_{pre}}{y_{abs}} \times 100\% \quad (7)$$

3 Results and discussion

In the present study, the ANN model was designed to predict groundwater levels in two test wells with 1 day a-head time using a set of suitable input parameters. The input parameters for the ANN model were decided by considering the parameters potentially to affecting the groundwater level. A cross correlation analysis between the water levels in the test wells at various lags suggested that Lag 1 correlation is highly significant in the water level time series in all the two test wells. To examine the effect of rainfall, river stage, surface flow and temperature on groundwater, daily groundwater levels were plotted (Fig. 3). It shows that groundwater levels are generally higher on rainy days which indicate that rainfall is a paramount parameter that influences groundwater levels compared to river stage (Fig. 3) and other temperature parameters (Fig.4). In a semi-confined aquifer, apart from rainfall,

surface water flow is another parameter that can influence the recharge to groundwater in BI sites (see Fig. 3). Therefore, daily surface water flow was also considered as one of the input parameters for the ANN model. The ANN models were trained with four input nodes for dynamic series model (Model 1) and five input nodes for time-series model (Model 2). The data were divided into two sets namely training and testing data. The calculated performance statistics for both models are shown in Table 1. Based on Table 1, the coefficient determination, R^2 for the DW1 and DW2 for Model 1 is about 0.748 and 0.546 and for Model 2 is 0.640 and 0.602 respectively (Fig. 4 and 5). For a perfect predictor, the coefficient determination should be +1 or -1. In general the definition of R tells us that 100 R^2 is the percentage of the total variation of the predicted values which is explained by, or is due to their relationship with actual values. This is an important measure of the relationship between two variables; beyond this it permits valid comparisons of the strength of several relationships. Based on this value, indicating that the Model 1 and 2 is good even though the R^2 value is not really reaching 1. The correlation coefficient was also high on Model 1 for the DW1 and DW2 which is 0.8649 and 0.7392 and also model 2 which is 0.8006 and 0.7765 respectively. Fig. 5a, positive (measure of direction) correlation or direct relationship indicates that a high score on the one variable is associated with a low score on the second variable. A negative correlation or inverse relationship indicates that a high score on one variable is associated with a low score on the second variable. The magnitude of the correlation coefficient indicates the strength of the relationship between the two variables. This magnitude can vary from 0.00 to 1.00. The closer the correlation coefficient is to either -1.00 or +1.00 the stronger the relationship. The stronger relationship between two variables is related produced the better prediction. This phenomenon can be seen in Fig. 4 and 5 for Model 1 and Model 2 respectively. Model 1 and 2 gives a clear view that the predicted and the actual process output are related which mean that the prediction output is quite near to the actual data which is the model 1 and 2 and it is a good correlation. As shown in Fig. 7 and 8, that the actual and predicted output for Model 1 and 2 is reasonably good and the groundwater level is always being recharged by river during the high flow period but during the low flow period the groundwater is recharging to the river. The prediction model is reasonably good in showing the relationship between river and groundwater. The residue analysis was also carried out in this study. The residual analysis is very useful in helping us to identify the performance of both models. The analysis for the residual is presented in Fig. 5. It can be seen that the residual for Model 1 and 2 is more consistent where output is near to zero even though in some samples the residual is high (Fig. 6 and 7). It shows that model output or prediction always near to the actual data. From the residual analysis, it determines that for the negative values, river gains from the aquifer and if river losses it shows the positive value. The statistical adequacies of the developed models for 1-day ahead forecasts for DW1 and DW2 test wells are summarized in Table 1. It is observed from Table 1, that the model performance is good and the models have forecasted the water levels with reasonable accuracy in terms of all statistical indices during calibration and validation. The correlation statistics that evaluates the linear correlation between the observed and the computed water table is consistent. The RMSE statistic, which is a measure of residual variance that shows the global goodness of fit between the computed and observed water levels, is very good as is evidenced by a low value (< 0.4 m) during both training and testing. While a 1-day ahead forecasts at higher lead time are required for efficient planning of RBI method or conjunctive use. A further analysis was performed by using the forecasted water levels at the DW1 and DW2 test wells as input to the models. An analysis that evaluates the input sensitivity to the model predictions was carried out by developing two models; one using rainfall, surface water level, temperature, stream flow and rainfall as input model. The results were found to be not very accurate and may be because the recharge time for water to reach the groundwater is quite high

compared to the lag period considered in the model. The results presented are mainly in the form of percentages with 95% confidence intervals. The lower and upper limits of groundwater levels with 95% confidence interval are shown in the Fig.8. This Fig. 8 shows there is a decline and rise of groundwater table elevations for the entire period and the model compares well with observations with the highest groundwater level in the early February 2015 and 2016.

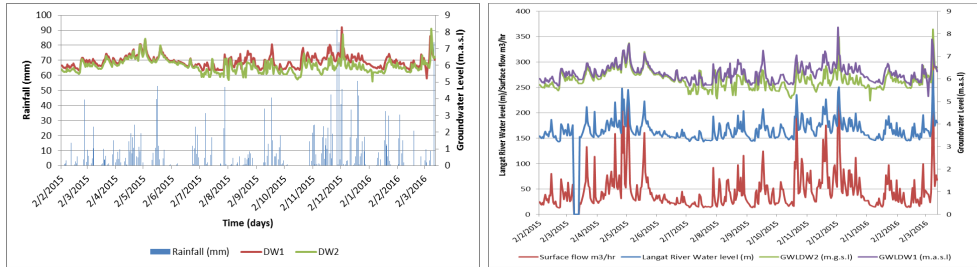


Fig. 3. (a) Daily groundwater level fluctuations and (b) Well hydrographs at sites DW1 and DW2 with river stage and surface water flow hydrograph at Jenderam Hilir

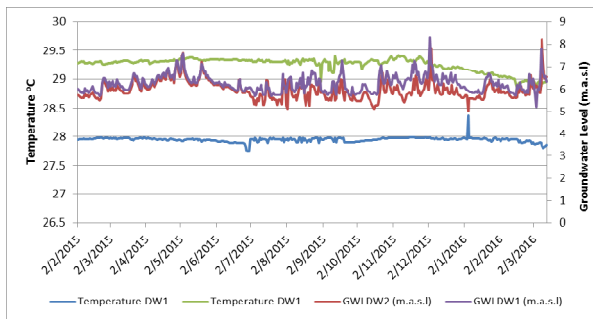


Fig. 4. Daily groundwater level fluctuations with groundwater temperature

Table 1. Performance statistic for model 1 and model 2.

	Model 1		Model 2	
	DW1	DW2	DW1	DW2
R	0.8649	0.7392	0.8006	0.7765
MSE	0.0356	0.0781	0.0513	0.0682
RMSE	0.01	0.0139	0.0113	0.013
R ²	0.748	0.546	0.64	0.602

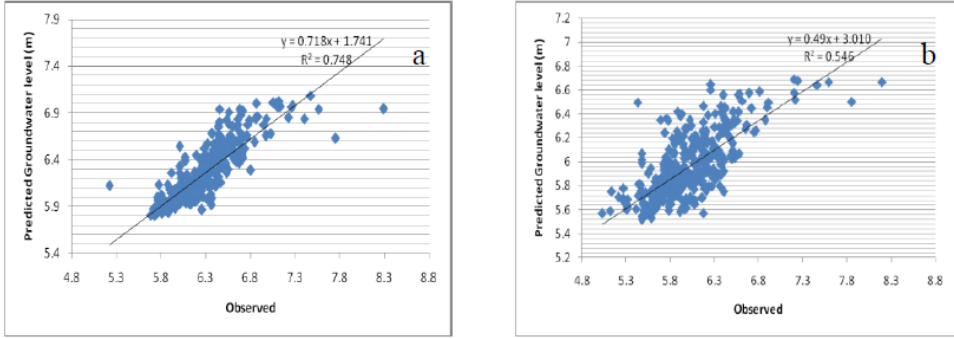


Fig. 5. Scatter plot for observation and predicted value for DW1 (a) and DW2 (b) in Model 1

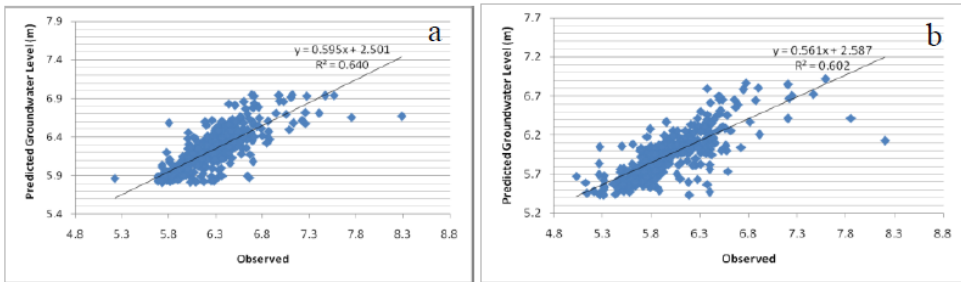


Fig. 6. Scatter plot for observation and predicted value for DW1 (a) and DW2 (b) in Model 2

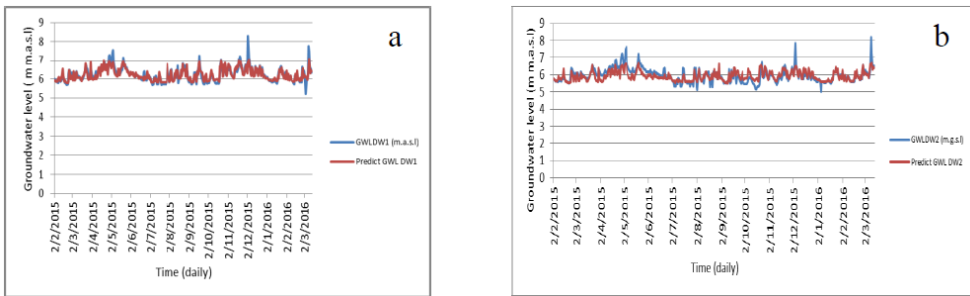


Fig. 7. Groundwater level prediction for DW1 (a) and DW2 (b) in a dynamics model

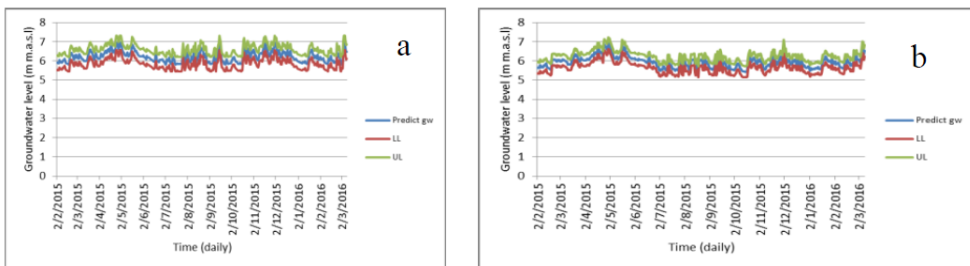


Fig. 8. The lower and upper limits are also shown in Model (DW1 (a) and DW2 (b))

4 Conclusion

In this study, the potential of neural network computing technique for forecasting groundwater level was investigated by developing ANN models for a shallow aquifer in Jenderam Hilir, Selangor in Malaysia. The result indicates that the capability of neural networks models in modelling of daily groundwater level using rainfall, temperature, stream flow and river water level data as inputs as well as the past value of the groundwater level. The two modelling techniques applied show that both models can perform well. This can be explained by referring to the correlation coefficient analysis, coefficient determination and RMSE. It was noted that the contribution of past values of groundwater level was very important in neural networks modelling as well as the value of the rainfall, temperature, stream flow and river water level. This was not observed in the time series modelling techniques where it only use past values of the groundwater level. This was the main reason, both the dynamic and time series modelling techniques was performed in this study. Therefore the appropriate techniques to model the groundwater level in study is by using the values of rainfall, temperature, stream flow and river water level as well as the values of groundwater level. A significant advantage of this model is that is can be provide good predictions with limitations of groundwater table record.

References

- [1] J. Schubert, Hydraulic aspects of riverbank filtration-field studies. *Journal of Hydrology*, **266**, 145–161, (2000)
- [2] R.A. Sheets, R.A. Darner and B.L. Whitteberry, Lag times of bank filtration at a well field, Cincinnati, OH, USA. *J. of Hydrology*, **266**, 162–174, (2002)
- [3] P.J. Dillon, M. Miller, H. Fallowfield and J. Hutson, The potential of riverbank filtration for drinking water supplies in relation to microcystin removal in brackish aquifers. *J. Hydrology*, **266**, 209–221, (2002)
- [4] C. Ray, G. Melin and R.B. Linsky, *Riverbank Filtration Improving Source-Water Quality*. Kluwer Academic Publishers, Dordrecht. (2002)
- [5] J.Z. Wang, S.A. Hubbs and R. Song, *Evaluation of Riverbank Filtration as a Drinking Water Treatment Process*. AWWA Research Foundation and American Water Works Association, USA, (2002)
- [6] B. Wett, H. Jarosch and K. Ingerle, Flood induced infiltration affecting a bank filtrate well at the River Enns, Austria. *J. of Hydrology*, **266**, 222–234, (2002)
- [7] M.K.N. Shamsuddin, W.N.A. Sulaiman and S. Suratman, *Hydrogeol J.*, **22**, 543. (2014)
- [8] Nourani V and Mano A.. Semi-distributed flood runoff model at the sub continental scale for southwestern Iran. *Hydrological Processes*, **21**, 3173–3180, (2007)
- [9] E.A. Coppola, C.F. McLane, M.M. Poulton, Szidarovszky F. and Magelky R.D.. Predicting conductance due to upconing using neural networks, *Ground Water* **43**: 827–836, (2005a)
- [10] Lallahem S., Mania J., Hani A. and Najjar Y. On the use of neural networks to evaluate groundwater levels in fractured media, *J. of Hydrology*, **307**, 92–111, (2005)
- [11] Coppola E.A., Rana A.J., Poulton M. M., Szidarovszky F. and Uhl V.W.. A neural network model for predicting aquifer water level elevations. *Ground Water*, **43**, 231–241, (2005b)

- [12]Tayfur G., Swiatek D., Wita A. and Singh V.P.. Case study: finite element method and artificial neural network models for flow through Jeziorsko earth dam in Poland. *J. of Hydraulic Engineering*, **131**, 431–440, (2005)
- [13]Gobindraju R.S. and Ramachandra Rao. A., *Artificial neural network in hydrology*. Kluwer, Dordrecht (2000)
- [14]Coulibaly P, Anctil F., Aravena R., Bobee B., Artificial neural network modeling of water table depth fluctuations. *Water Resources Research*, **37**(4), 885–896, (2001)
- [15]Coulibaly P., Anctil F. and Bobee B. Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. *J. of Hydrology*, **230**, 244–257, (2000)
- [16]Fausett L, *Fundamentals of neural networks*. Prentice Hall, Englewood Cliffs. . (1994)
- [17]Haykin S. *Neural networks, a comprehensive foundation*, 2nd edn. Prentice Hall, Englewood Cliffs. (1999).