The authors gratefully thank to the Referee for the constructive comments and recommendations which definitely help to improve the readability and quality of the paper. All the comments are addressed accordingly and have been incorporated to the revised manuscript. Detailed responses to the comments and recommendations are as follows.
Please note that all the comments are bold-faced and authors reply follow immediately below the comments.

Major Comments:

Comment #1

- **In the Abstract Authors claim that the "optimal performance" of the ANN model is found in a few iterations, but they suggest that the results of forecasting are good for "normal water levels" only. For higher water levels, which are rarely measured, the results are of poor quality. The paper is focused on an improvement of prediction just for the higher water levels by means of ANN with Zone Matching Approach. However, neither objective function nor training methods are clearly presented, hence it is difficult to understand or discuss the reason of quick convergence of the algorithm and poor performance for higher water levels.**

Reply to comment #1:

There is no special training method or algorithm use in the study except is the use of MLP with BP and steeper steepness coefficient (Sulaiman et al., 2011). Any cases that use high volume of almost the same data pattern will result fast convergence of data training to optimal performance (within few cycle of data training). One training cycle means one epoch. The fast convergence which is based on the epoch does not mean faster convergence in term of time. The reason is quite basic if one understands of what an epoch means in data training. One epoch is a completion of computation of one cycle of computation of feed-forward and back propagation after processing the whole sets of data training records. This means that if there are a very high volume of the almost the same data pattern in one epoch, it as though training of smaller volume of almost the same data pattern records with many epochs. An example, this study use 287,100 (290,000 total records – 2,900 high water level data) of almost the same data pattern (that is mainly normal water level) in one epoch could be equivalent to 28.7 epochs if 10,000 almost the same data pattern records are used instead. Thus, 20 epochs that is achieved in this study could be represented by 20x28.7 that is 574 epochs for the 10,000 records. This is not a small number of epochs in data training. Many studies have limited historical data which is about 5 to 10 years. Here, there is 29 years of hourly water level data which is about 290,000 records for data training alone. That is the reason why the words 'short cycle of data training' and 'high volume normal water level data' is used in the abstract. The above is an explanation of the reason why the fast convergence of the ANN model when high volume of almost the same data pattern is feed to the ANN model.

Since there is confusion on the issue, we have included this discussion in revised manuscript at the result and discussion section. We appreciate the comment/suggestion on this issue which can help to clarify the issue more clearly.

Comment #2.

- **In general pure autoregressive models which take into account only historical flow/water level measurements at single cross-section cannot provide adequate forecasts. They simply lack any information about source of water in the catchment. Previously observed flow/water level values are useful for data-based models but only as additional input variables. To predict flows/water levels for a bit longer lead times exogenous variables are required. If only very short lead time is considered, the simplest linear models could be used..**

Reply to comment #2:

The authors agree with the reviewer in this comment, however, there are two major reasons that motivate the authors to develop the model in this architecture. The first, at the study area, Johor River, the data availability is ONLY the water level at certain cross-section, on the other hand. The second reason is that the proposed neural network model in our study is NOT mainly rely on the physical and/or hydrological behavior of the system in the study area, it is conceptually a time series forecasters with consideration of the water level pattern of consecutive certain time period to predict the time series of different systems' behavior that uses the previous and most recently behavior of a system to predict its future changes. The major advantage of this method is the ability to predict the behavior of systems without fully consideration or analytical prediction rules (hydrological/physical). In fact, all the other sources of water in the catchment that affect the water level at the cross-section under study could be considered that they already embedded in the historical data records of the water level. As a result, within this concept, the water level could be forecasted as long as the previous data records could help enhancing the forecasting skills and are available in the time series.

Comment #3.

- **The paper lacks focus. It is difficult to read and understand the presented argumentation. A number of ideas are disputable (for example discussion of drawbacks of rainfall-runoff models, see p. 9359, or motivations of application of ANN, p. 9358 and 9360) Some topics are described in a few different parts of the paper (for example the ANN training – see p. 9363 and 9365).**

Reply to comment #1:

There are two issues that the authors make argument on drawback rainfall-runoff (pg. 9359).

1- Synchronization of rainfall-runoff in real time to capture the rainfall and runoff data cannot be easily implemented. There must be application to capture from different sources of data from different location.In advanced countries the infrastructure might be available but not in many third countries. That could be the reason during the early computing, Dawson and Wilby (2001) described that the rainfall-runoff require infrastructure to manage different data inputs.

2- ANN is data sensitive means that any modification of the data inputs can affect the training results. Even, the normalization of observed data that are feed to the ANN with values range between 0 to 1, 0.9 to 0.1 or -0.1 to 0.1 will bring different training results. Thus, if two ANN models where one is trained by original observed runoff data and another trained by water level data that are converted through rating curve almost definitely produced different results. As described by the paper, to date there is no paper that has shown that the results between the two are comparable.

The authors wish that the reviewer have made a specific reply or discussion to the above arguments on the issue of drawback of rainfall-runoff models.

For ANN training, page 9363 is a discussion on ANN in general and for page 9365 describes the implementation of ANN – the network model used in the study. However, the authors agree that section 3.1 and 3.3 can be combined so that content is not in a few different parts of paper. Thus, section 3.1 and 3.2 are combined in the revised manuscript.

For the motivation of application of ANN in page 9358 and 9360, the authors believed that the two items below does provide strong motivational reasons for the application of ANN.

1- Extensive review in hydrological simulation and forecasting in ASCE (2000a, b)

2- ANN does not require data regarding physical characteristics (Dawson et al., 2001)

However, the authors agree that additional reasons could strengthen the motivational of the use of ANN.An additional motivational of the application of ANN is added in revised manuscript based on Maier et al. (2010) study based on 230 papers that have been published in the water forecasting studies. The reason to use ANN specifically MLP-BP is also described by Maier et al. (2010) is cited in the revised manuscript.

Reference:

Maier, H.R., Jain, A., Dandy, G.C., and Sudheer, K.P., (2010) Methods use d for the development of neural network s for the prediction ofwater resource variables in river systems: Current status and future directions, Environmental Modelling & Software 25, 891-909

- The authors appreciate this comments that can strengthen the arguments and make the paper more focus and organized.

Minor comment.

Comment #1

- **The description of neural network (section 3.1) is poorly presented. The equations 1 and 2 need improvement. For example, the notation of I (from i=0 to N) should be clarified. Note that xo is described but is not used in Eq. 1. Xij is used as input, but with two indices.**

Reply to comment #1:

The authors tried to simplify the description of neural networks based on previous comments by other reviewers on this topic. As described by the reviewers, this topic is readily available in many books and published papers. However, the authors agree that the content in this topic is not quite clearly presented. Thus, this section has been rewritten starting from line 5 to line 23 of page 9362. The two equations in previous manuscript have been reduced to one. The corrected sentences are below;

Corrected sentences:

An Artificial Neural Network (ANN) is a parallel-computing mathematical model for solving dynamic nonlinear time series problems. There are many types of ANN, the most common being the multilayer perceptron neural network (MLP-NN) (Zhang et al., 1998) that is used in this study. Figure 4 shows the basic architecture of the MLP-NN which is consists of three layersof neurons that are ordered in sequence. The first layer is an input layer, the last layer is an output layer and in between the layers is hidden layer. The architecture shown in this study has only one hidden layer but the MLP-NN can have more than one hidden layers. The function of the neurons in the input layer is to receive data input and pass this data as output to the neurons in the second layer which is the hidden layer. The function of neurons in the hidden layer is to receive the output and the weight of output from the neurons in the input layer and compute the received data using activation transfer function (ATF). The computed result is passed as output to each the neurons in the next layer which is the output layer. The process of computation in the output layer is the same as in in the hidden layer where output and weight of the output from each neuron from the previous layer are computed in the neuron in the output layer using activation transfer function.The computed output at the output neuron is the output of the network model. There are two additional neurons that are added in the input and hidden layers that are called bias which has a fixed value of 1. These neurons do not receive and compute data but passing the value of 1 and its weight to the neurons in the next layer.The function of the bias is to stabilize the computed output between 0 and 1.There are many types of ATF, and the most common use of ATF is sigmoid function (Zhang et al., 1998; Maier et al., 2000). The equation to compute the inputs to the computing neuron is;

$$Output = \frac{1}{1+e^{-k.\sum_{i=0}^{n} w_{ij}x_{ij}}} \qquad (1)$$

where x and w is the output and weight of output from each of the neurons from previous layer, i refer to the neuron in previouslayer to the computing neuron, j refer to the computingneuron in the current layer and k is the steepness coefficient of the sigmoid function.

Comment #2.

- **It is not clear why the forecasting by means of ANN "requires the input data with hourly time step" (section 3.2). In general neural networks do not need the same intervals between inputs.**

Reply to comment #2:

The authors agree that the keyword 'requires' in line 19 page 9364 seem to restrict the input data to the hourly time step. The sentence has been rewritten as below;

Corrected sentence:

To forecast the water level M hours ahead with N data inputs, the network model use data inputs of hourly water level data at times t, t−1 to t−N, where the interval between each time step is M hours.

Comment #3.

- **Discussion about overfitting is dubious (section 3.3). The more data we have, thelower overfitting should be, but this depend on the particular problem and frequently overfitting cannot be easily eliminated even if the number of data is large (see Geman et al., Neural networks and Bias/Variance dilemma, Neural Computation 4, 1-58, 1992). Authors suggest using the low number of hidden nodes but this is not always sufficient(see Giustolisi and Laucelli 2005, Hydrological Sciences Journal 50(3)). Moreover keeping number of hidden nodes equal to the number of inputs is not well justified requirement. Contrary to the Authors suggestion, this may lead to too large number of parameters.**

Reply to comment #3:

The first impression that might influence in the forecasting of water level with a single source of data (water level) is the number of data inputs. Thus, the study starts with variation of data inputs range from 2 to 8 with the number of hidden neurons also have the same number as the data inputs. The testing show that data input with less than 3 or greater than 7 does not provide strong forecasting performance as compared to 3 to 7. The study also tested several large numbers of hidden neurons to the 3 to 7 data inputs but the results show that there is not much difference in performance using the higher number of neurons to the same number of neurons in the input and hidden layers. The effect of using a larger number of hidden neurons is the ANNtraining becomes very slow. Thus, the study focuses on examining the 3 to 7 data inputs with the same number of hidden neurons. The use of the same number of data inputs and hidden neurons has been tested by Sulaiman et al. (2011) in daily water level forecasting and Jain et al. (2006) in the flow hydrograph modeling.After satisfied with the results, the study then focus on improving the forecasting of high water level events using the ZMA.

The authors agree that keeping the number of hidden neurons with the same number of data inputs can cause large parameters if the number of inputs is large. The authors in this paper do not suggest the use of the same number of input and hidden neurons. This study use the architecture and steepness coefficient suggested by Sulaiman et al. (2011). This part of comments should be replied or comments in the Sulaiman et al. (2011) paper. The authors agree that the Sulaiman et al. (2011) could be improved to limit the use of the same number of neurons input and hidden layer only in the case of the study and further study, should be examined for the case when the number of input are large.

The authors also agree that using small number is not always sufficient but it does not mean that it cannot be done. As again, the end results show that the low the number of neurons used in this study does succeed to produce high forecasting results.

Reference:

Jane, A. and Srinivasulu, S. (2006) Integrated approach to model decomposed flow hydrograph usingartificial neural network and conceptual techniques, Journal of hydrology, 317, 291–306

Comment #4.

- **Please, verify the scale on Fig. 3 and 7.**

Reply to comment #4:

The authors realized the mistake in Fig. 3 where the values should be between 0 to 14000 and not 1400 mm. It was spotted when doing correction after responding to the first reviewer. Thecorrection is in the revised manuscript.

- Thanks you for the comment.

Comment #5.

- **The Fig. 4 is not clear. Why the output values from the previous layers become indices of a parameter or variable in the subsequent layer?**

Reply to comment #5:

The i0 = 1 has been changed to i0 only and also the j0=1 has been changed to j0 only in Fig. 4. The i is to represents the neurons in input layer and j is to represents neurons in hidden layer. The reason the error occurred because the authors tried to show that the first neuron in the input and hidden layers are bias neuron that has a value of 1. The word bias is shown next to the neuron in the corrected figure.

- The authors acknowledge the mistake and thanks for highlighting the problem.

Comment #6.

- **In Table 4, network model 1, the Nash-Sutcliffe coefficient is above 1. How suchresult was obtained?**

Reply for comment #6:

The errors have been spotted by the first reviewer (Technical comment- #12). It is not that the Nash-Sutcliffe coefficient is above 1 but there are two data in that column. One is for the NSC and another is the epoch. The errors are cause by the shift-enter in that column. The column for the data in the original manuscript is small that require shift-enter to place the two data in two lines but when converted into HESSD format.It has been corrected by placing the epoch data in bracket.

Comment #7.

- **Note that sometimes not adequate literature is cited. For example the idea that ANN "mimics information processing in human brain" was not proposed by El-Shafie (2008) or El-Shafie and Noureldin (2011).**

Reply to comment #7:

The authors agree that to cite the authors, the words should be proposed by the authors. The sentence has been removed and replace with a more general statement.

Corrected sentence:

ANN is a parallel-computing model based on how human brain process information. The use of error propagation and computer to simulate the neural processes (Rummelhart and McClelland, 1986) has popularized the use of ANN in many data forecasting and data classification studies in recent years.

Additional notes.

The authors realized that equation 3 has wrong arguments. It is corrected in the revised manuscript.

- $WL_{t+1} = f(WL_t, WL_{t-1},.., WL_{t-N}, w_1,w_2,..w_k)$ (3)