Major Comments:

Comment #1.

The description of the study area should be more precise. Is it right, that the upstream gauge (input data) is RantauPanjang (fig. 3) and the target time series are from Kota Tinggi? The catchment size of 2636 km2 refers to any of these stations? For the evaluation of the model performance it would be appropriate to indicate both areas. The water levels at the target gauge result from both the discharge at the upstream gauge and the precipitation at / runoff from the area between the two gauges. The runoff from this area is certainly not independent from the runoff in the upper catchment. But however - the size of this 'unobserved' catchment (in the sense of not been represented by any input data) is quite interesting for a valuation of the model.

Reply to comment #1:

The target water level forecasting in this study is Rantau Panjang station using historical water level data at the station itself.

Initially, Kota Tinggi is the target forecasting area (high population town) but due to non-available of data at Kota Tinggi, the study focus to Rantau Panjang which is on the upper stream of Kota Tinggi. The reason Kota Tinggi is described in the study to show the relevant of Rantau Panjang station in the study.

The catchment size of 2636 km2 refers to Johor River Basin. Since the study area is Rantau Panjang station, other data is not the concerned by the study since ANN independent of physical condition of the river basin.

The case study area might have misled to the misunderstanding of this study thus a revised case study area have been made in the revised manuscript.

Revised case study area;

The initial concerned study area in the development of the forecasting model is Kota Tinggi, Johor, Malaysia. However, due to the lack of historical water level data at Kota Tinggi station, Rantau Panjang station, which is on the upper stream of Kota Tinggi has been selected as the study area. Both stations are in the Johor River basin which is shown in Fig. 1. The reason that Kota Tinggi is the concerned area is because it is an administrative town with a large population and it has been hit with 13 flood events since 1978, making the development of the forecasting model relevant. Rantau Panjang has been hit with 17 flood events above 8 m since 1963. <u>There is a correlation between the flooding events that occurred at the two stations that will not be described in this paper, but successful forecasting at Rantau Panjang could later assist with other studies of flood events at Kota Tinggi.</u> All of the flooding events at Rantau Panjang occurred during the Northeast monsoon which happens between November and March, shown in Fig. 3. The distance between the two stations is about 40 km and the lag time for river flow from Rantau Panjang to Kota Tinggi is about 24 hours. The normal water levels at Rantau Panjang and Kota Tinggi are 4 m and 2 m respectively, while the flood level at the two stations are 9 m and 2.1 m. Historically, there are cases of flood events above 8 m in Jan-1965 and Jan-1967 before some measure of flood mitigation are carried out in the area.Figure 2 shows the normal, alert and danger water levels at Rantau Panjang station. The Johor River is about 122.7 km in length and drains an area of 2,636 km2. Its main tributaries are Sayong River and Linggiu River. The river originates from

Mount Gemuruh (109 m) and discharges into the Straits of Johor (0 m). The average annual precipitation for the Johor River catchment is 2.47 m.

- The authors hope that the explanation adequately answers comment #1. If further question, please make comment. Thank you.

Comment #2.

Regarding the modelled system, it seems more adequate to deal with discharges. The riverbed geometry might change (within some 50 years!). And so might change the water level resulting from a certain discharge. What is the reason for modelling the water levels rather than the discharges?

Reply to comment #1:

In this study, the data inputs and the target forecasting are based on water level. One of the reasons that modeling is based on the water levels is that the monitoring of flood level by water authorities is based on water level. Thus, direct comparison can be made between the results of the forecasting model and the flood events.

To overcome of the perception that there are changes in river bed geometry in the future where the forecasting of the water level might not able to produce a reliable result, the validation datasetselected in this study has water level data that have changes in the geometry of the river bed.

There are excavation of the river depth and clearing of the sedimentation of the river before the year 2000 which is the starting date of the validation dataset.A strong forecasting resultbased on modeling of water level can overcome the wrong perception that the water level is not reliable in future events when there are changes in the river bed. This is another reason why the modeling of the water level in this study.

These are the reasons why the use of water level instead of discharge in this paper. The authors believe of the special capability of the ANN in capturing nature events.

- Modificationshave been made to the revised manuscript in the Dataset section. The section is now changes to Data and Dataset which include the above description after the section title. Thanks for the comment.

Comment #3.

Is it necessary to use that many performance measures? The authors don't use all of them for interpreting the results and the model performance.

Reply to comment #1:

Many performance measures are used to strengthen the high accuracy results of this study. This can be helpful to provide in the confidence of developed model. This is described in introduction section - “The reliability of forecasting results is also a major focus in this study, since this will build confidence in the forecasting model that is developed. Thus, several performance measures such as Nash-Sutcliffe efficiency, the correlation coefficient, root mean square error, the scatter index and three offset errors are used to evaluate forecasting results."

The authors believe all performance measures NSC, R2, RMSE, SI, and Offset errors are used in the paper in the results and discussion section. Maybe some parts are not described in detail but this is to avoid too much description on particular results.

- The authors hope if the reviewer can specify the performance measure described.

Comment #4.

To illustrate the issue and to get the modelling problem more imaginable, the referee would appreciate a plot of time series of a single event (like fig. 10) showing both input and target data. As ANNs are purely empirical models they are only valid within the range of the training data. And, unfortunately, extreme flood events are rare. To both guarantee a sufficient database and restrict the training/validation data to only flood events, the choice of the lower limit of the ZMA of 8000 mm seems very reasonable (regarding fig. 5). But, to my opinion, for an application of the model the distinction has to be made for the input data. In a word: The authors should refer more to the model input.

Reply to comment #1:

The authors agreed with the reviewer. One more figure and explanation regarding illustration of input-target data have been reported in the revised version.

Comment #5.

It would be desirable to make clearer the choice of the input variables (time lag to the forecast value), as this is somehow the heart of the approach. Please, specify

Reply to comment #1:

The authors quite agree that the time lag also play a greater role in this ZMA study which the authors unaware that could have leads to high forecasting results. Below are the reasons that the author use the time lag based on fig. 6.

The reason that the interval hours is the same as the lead-time is so that the rate of change of interval hours between time steps is measured on a consistent scale. This is consistent with the ANN capability which is to find hidden pattern. The pattern is not consistent if the time steps for data output (forecasted) is different from the data inputs. This is the author's first impression. Thus, the study make data pattern based on the same interval for data inputs and forecasted time. However, other studies might have successfully applied different time step than the forecasted time. Thus, the study also make pre-testing using 1 hours interval for forecasted time of 3 hours. But the results of the same interval hours of the data inputs and forecasted produced better results.

Modifications are made to the manuscript to clarify this issue.

- Thanks for the comment which definitely help to improve the paper.

Comment #6.

For an application by the authorities, as suggested in the conclusions section, the authors should suggest a range of validity of the model.

Reply to comment #1:

This has been answered in reply to reviewer #1 Major comment #5. The range of validity of the model or restriction of the 3 hr model already described after the first revised manuscript. However, the authors have made some more modification based on this comments. The range of validity is forecasting is for water level above 7000 mm and the detail reliability of the model should be referred to Table 4. Other models, 1 hr and 2 hrs lead-time can also be used for high accuracy forecasting results.

- Thanks for the suggestion.

Technical comments:

Comment #1.
Which software was used for the setup of the ANN?

Reply to comment #1:

The ANN program is developed using visual basic. That's the reason why we know that the forecasted water level is function of water level inputs and also weights in the neural network. In fact, the general equation that is described in Sulaiman et al. (2011) is based on the understanding of how the ANN model worked. The equation was presented in Mar 2010 in Young Water Professional international conference. Since, we have the capability to make computer program, we also have the flexibility of testing many scenarios of data inputs such as 1 and 3 hours interval data inputs for 3 hours ahead forecasting.

Comment #2.
fig. 3: a bar plot would be more appropriate

Reply to comment #1:

The authors agreed with the reviewer that the bar plot is more appropriate. Changes has been made in the revised manuscript.

- Thanks for the suggestion for improvement of the paper.

Comment #3.
3.6 and fig. 6: Is the time step always 3h? Has the forecast time to be equal to the time step? If it is so, then why?

Reply to comment #1:

Based on figure 6, yes the time step is 1 hr for 1 hr forecasted time, 2 hr for 2 hr forecasted time, and 3 hr for 3 hr forecasted time.

- Please refer to Major comments - comment #5 for the description.

Comment #4.
3.1 line 13-14: The neurons do not receive the weights from the adjacent layer! line 23: The bias does not stabilize the output between 0 and 1. The sigmoid transfer function does! next page, eq. 3: The water levels are not a function of the weights; the weights are parameters of this function. Further, it would be preferable to distinguish somehow between WL as a target value and WL from the upstream gauge. line 20: The weights aren't initialized randomly to speed the training process up!

Reply to comment #4:

3.1 Line 13-14 – The sentence has been changed in the revised version based on the comment by reviewer #2. Part of the sentence is "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)."

Line 23 –It is meant to stabilize data input to sigmoid function so that the output of sigmoid function between 0 and 1 are shifted to the threshold value. The sentence has been rewritten in more general form in revised manuscript.

Eq. 3 – The equation is correct. The general ANN modeling equation is also described in Maier et al. (2010) equation 1: Y = f(X.W) + e where X is the inputs and W is the weights and Y is the model output. Equation 3 is derived based on the understanding of development of ANN program. The authors are surprised that the general equation has not been described before 2010. Please correct the authors on this if the equation has been described in any paper. Similar to equation 3was first presented in Mar-2010 at young water professional conference. The equation described by Sulaiman et al. (2010) is without calibration process and the Maier et al. (2010) does include the calibration process. However, the equation provided by Sulaiman et al. (2010) includes ANN architecture to support the equation.

WL as a target and WL from the upstream gauge –
Line 20 –All the water levels refer to water level at the upstream which is RantauPanjang station.

- The authors appreciate all the comments.

Comment #5.
In the conclusions the authors refer to DNN. It doesn't get clear why this should be advantageous. But maybe it is not necessary to introduce this type of ANN in this chapter, if was not applied in the paper

Reply to comment #1:
The authors agree that the DNN better to be removed since it might require extensive introduction which is not the focus of this study. Two sentences that describe the DNN are removed in the revised manuscript.

- Thanks for the suggestion for improvement of the paper.

Additional notes:
Correction in Table 4. The first column title should be lead-time instead of network model.