World Academy of Science, Engineering and Technology
International Journal of Environmental and Ecological Engineering
Vol:2, No:3, 2008

A Preliminary Study on the Suitability of Data Driven Approach for Continuous Water Level Modeling

Muhammad Aqil, Ichiro Kita, and Moses Macalinao

Abstract**—**Reliable water level forecasts are particularly important for warning against dangerous flood and inundation. The current study aims at investigating the suitability of the adaptive network based fuzzy inference system for continuous water level modeling. A hybrid learning algorithm, which combines the least square method and the back propagation algorithm, is used to identify the parameters of the network. For this study, water levels data are available for a hydrological year of 2002 with a sampling interval of 1-hour. The number of antecedent water level that should be included in the input variables is determined by two statistical methods, i.e. autocorrelation function and partial autocorrelation function between the variables. Forecasting was done for 1-hour until 12-hour ahead in order to compare the models generalization at higher horizons. The results demonstrate that the adaptive network-based fuzzy inference system model can be applied successfully and provide high accuracy and reliability for river water level estimation. In general, the adaptive network-based fuzzy inference system provides accurate and reliable water level prediction for 1-hour ahead where the MAPE=1.15% and correlation=0.98 was achieved. Up to 12-hour ahead prediction, the model still shows relatively good performance where the error of prediction resulted was less than 9.65%. The information gathered from the preliminary results provide a useful guidance or reference for flood early warning system design in which the magnitude and the timing of a potential extreme flood are indicated.

Keywords**—** Neural Network, Fuzzy, River, Forecasting

I. Introduction

THE design, planning and operation of river systems depend largely on relevant information derived from extreme events forecasting and estimation. Reliable flood forecasts are particularly important for warning against dangerous flood and inundation as well as in the case of multi-purpose reservoirs. Hydrological data estimation is also significant since the time series data often exhibits some form of deficiency due to the presence of gaps, discontinuities and inadequate length. There are many forecasting techniques have been developed to simulate the hydrological time series such as empirical black box, conceptual, and physically based distributed models. Conceptual and physically based distributed models are designed to simulate the physical mechanisms that determine the hydrological circle, and use to involve water transference physical laws, and parameters associated with the characteristics of the catchment area [1]. Such models may require sophisticated mathematical tools, a significant amount of calibration data, and some degree of expertise and experience with the model [2]. Therefore in practical situations, the uses of a simple model such as linear system models or black box models more commonly used. However, these simpler models normally fail to represent the non-linear dynamics, which are inherent in the process of rainfall-discharge transformation. By considering the complexity of phenomena involved there is a strong need to explore alternative solutions through modeling direct relationship between the input and output data without having the complete physical understanding of the system. While data-driven models do not provide any physics of the hydrologic processes, they are in particular, very useful for modeling hydrological time series where the main concern is to predict accurate flows at specific watershed locations [3].

Recently, intelligent computation methods have been adopted in water resources forecasting studies as a powerful alternative modeling tools. These methods offer advantages over conventional modeling, including the ability to handle large amounts of noisy data from dynamic and nonlinear systems, especially where the underlying physical relationships are not fully understood. Other associated benefits include improvement of model performance, faster model development and calculation times, and improved opportunities to provide estimates of prediction confidence through comprehensive bootstrapping operations [4]. Successful applications of adaptive network-based fuzzy inference system based modeling in water resources forecasting have been widely reported such as used neuro-fuzzy and neural networks model for short-term water level prediction [5], a fuzzy neural network model for inflow forecast into electric power plant [6]. Also, the performance of the adaptive network-based fuzzy inference system is significantly improved if the input data are transformed into the normal domain prior to model building [7]. Demonstration

Manuscript received February 16, 2006. This work was supported in part by the Ministry of Agriculture, Indonesia.

Muhammad Aqil is with the Indonesian Agency for Agricultural Research and Development, Indonesia (e-mail: akilshimane@gmail.com).
Ichiro Kita is with the Faculty of Life and Environmental Science Shimane University, Japan.
Moses Macalinao is with the Irrigation Engineering Department, Luzon State University, Philippines.

on the use of Takagi-Sugeno models for predicting discharge from rainfall time series by comparing grid partitioning, subtractive clustering and Gustafson-Kessel clustering identification method for constructing the models [8].

The current paper reports an outcome of a study aims at investigating the suitability of the adaptive network based fuzzy inference system for continuous water level modeling. The information gathered from the preliminary results will be used to design the warning system in which the magnitude and the timing of a potential extreme flood are indicated.

II. ADAPTIVE NETWORK INFERENCE SYSTEM

Adaptive network-based fuzzy inference system (ANFIS) used a feed forward network to search for fuzzy decision rules that perform well on a given task. Using a given input-output data set, ANFIS creates a FIS whose membership function parameters are adjusted using a backpropagation algorithm alone or combination between a backpropagation algorithm with a least squares method. This allows the fuzzy systems to learn from the data being modeled. ANFIS provide a method for the fuzzy modeling procedure to learn information from the data set, followed by creating the membership function parameters that best allow the associated FIS to well perform the given task [9]. Consider a first order Takagi-Sugeno fuzzy model with a two input, one output system having two membership functions for each input as shown in Fig. 1(a). Then the equivalent ANFIS architecture of the first order Takagi-Sugeno inference system is shown in Fig. 1(b). The functioning of ANFIS is a five layered feed forward neural structure and the functionality of the nodes in these layers can be summarized as follows:

Layer 1: Every node *i* in this layer is an adaptive node with a node output defined by:

$$O_{1,i} = \mu_{A_i}(x), \tag{1}$$

$$O_{1,i} = \mu_{B_{i-2}}(y), \tag{2}$$

where *x*(or *y*) is the input to the node; A_i (or B_{i-2}) is a fuzzy set associated with this node, characterized by the shape of the membership function in this node and can be any appropriate functions that are continuous and piecewise differentiable such as Gaussian, generalized bell shaped, trapezoidal shaped and triangular shaped functions. Assuming a Gaussian function as the membership function, A_i can be computed as,

$$\mu_{A_i}(x) = \exp\left[-0.5\left((x-c_i)/\sigma_i\right)^2\right] \tag{3}$$

where $\{\sigma_i, c_i\}$ are the parameter set. Parameters in this layer are referred to as premise (antecedent) parameters.

Layer 2: Every node in this layer is a fixed node labeled Π, which multiplies the incoming signals and outputs the product. For instance,

$$O_{2,i} = w_i = \mu_{A_i}(x) \times \mu_{B_i}(y) \tag{4}$$

Each node output represents the firing strength of a rule.

Layer 3: Every node in this layer is a circle node labeled *N*. The *i*th node calculates the ratio of the *i*th rule's firing strength to the sum of all rule's firing strengths. Output of this layer will be called normalized firing strengths.

$$O_{3,i} = \overline{w} = \frac{w_i}{w_1 + w_2}, \quad i = 1, 2 \tag{5}$$

Layer 4: Node *i* in this layer compute the contribution of the *i*th rule towards the model output, with the following node functions:

$$O_{4,i} = \overline{w}_i f_i = \overline{w}_i (p_i x + q_i y + r_i) \tag{6}$$

Where $\overline{w_i}$ is the output of layer 3 and $\{p_i, q_i, r_i\}$ is the parameter set. Parameters in this layer will be referred to as consequent parameters.

Layer 5: The single node in this layer is a fixed node labeled $\sum$ that computes the overall output as the summation of all incoming signals.

$$\text{Overall output} = O_{5,i} = \sum_i \overline{w}_i f_i \tag{7}$$

The learning algorithm for ANFIS is a hybrid algorithm, which is a combination between gradient descent and least-squares method [9]. For simplicity, the adaptive network has

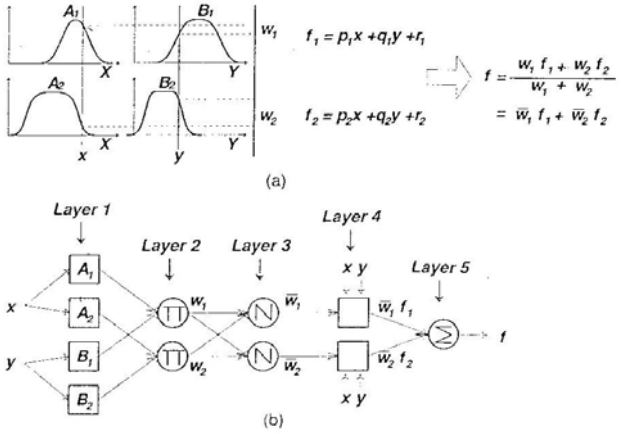

Fig. 1 (a) Fuzzy inference system. (b) Equivalent ANFIS architecture

only one output and is assumed to be

$$\text{Output} = F(\vec{I}, S) \tag{8}$$

where $\vec{I}$ is the set of input variables and *S* is the set of parameters. If there exists a function *H* such that the composite function $H \circ F$ is linear in some of the elements of *S*, then these elements can be identified by the least squares method. More formally, if the parameter set *S* can be decomposed into two sets

$$S = S_1 \oplus S_2 \tag{9}$$

(where $\oplus$ represents direct sum) such that $H \circ F$ is linear in the element S_2, then upon applying *H* to Eq. (8), we have

$$H(\text{output}) = H \circ F(\vec{I}, S) \tag{10}$$

which is linear in the elements of S_2. Now given values of elements of S_1, the *P* training data can be plugged into Eq. (9) and obtain a matrix equation

$$AX = B \tag{11}$$

where X is an unknown vector whose elements are parameters in S_2. Let $|S_2|=M$, then the dimensions of A, X and B are $P\times M$, $M\times 1$ and $P\times 1$, respectively. Since the number of training data pairs (P) is usually greater than the number of linear parameters (M), this is an over-determined problem and generally there is no exact solution to Eq. (11). Instead, a least squares estimate of X can be sought that minimizes the squared error $\|AX-B\|^2$.

Based on the ANFIS architecture shown in the Fig. 1, we observe that the values of the premise parameters are fixed, and the overall output can be expressed as a linear combination of the consequent parameters. In symbols, the output f in the Fig. 1 can be rewritten as

$$\begin{aligned} f &= \frac{w_1}{w_1+w_2} f_1 + \frac{w_2}{w_1+w_2} f_2 \\ &= \overline{w_1} f_1 + \overline{w_2} f_2 \\ &= (\overline{w_1} x) p_1 + (\overline{w_1} y) q_1 + (\overline{w_1}) r_1 + (\overline{w_2} x) p_2 + (\overline{w_2} y) q_2 + (\overline{w_2}) r_2 \end{aligned} \quad (12)$$

which is linear in the consequent parameters (p_1, q_1, r_1 ; p_2, q_2, and r_2). As a result, the set of total parameters (S) can be separated into two such that S_1 = set of premise parameters and S_2 = set of consequent parameters. Consequently the hybrid-learning algorithm can be used for an effective search of the optimal parameters of the ANFIS. More specifically, in the forward pass of the hybrid learning algorithm, node outputs go forward until layer 4 and the consequent parameters are identified by the least-squares method. In the backward pass, the error signals propagate backward and the premise parameters are updated by gradient descent.

III. Model Development

A. Study Area and Data Set

The study area is located in the Purwakarta Regency, West Java Indonesia. The total drainage area of the river basin is approximately 60.17 km^2. The climate of the catchment is generally dry, except during the monsoon months from December to April. Mean annual precipitation is 3000 mm but varies considerably from one year to another. Average air temperature ranges from 26.0 - 28.0°C during the rainy season of the northwest monsoon and 27.0 - 30.0°C during the dry season of the southeast monsoon. Satellite image of the river is shown in Fig. 2. For this study, data were available for a hydrological year of 2002 with a sampling interval of 6 minute. In this study, the performance of the Takagi-Sugeno fuzzy model was examined on hourly intervals. To achieve this, the 6-minute data series was converted first into hourly data before proceeding into the network. The data were divided into three independent subsets: a training subset includes 4000 data sets; the verification subset has 2500 data sets; and the testing subset has the remaining 2000 data sets.

B. Model Development

The current study employed two statistical methods, i.e. autocorrelation (ACF) and partial autocorrelation (PACF) to identify the appropriate input parameters. The ACF and PACF are generally used to gather information about the autoregressive process of the data series [10]. The number of antecedent water level to be included that should be included in the input variables are usually determined by placing a 95% confidence interval on the autocorrelation and partial autocorrelation plots.

Fig. 2 Satellite image of the river (acquired on Sep 4 2005)

The ACF and the corresponding 95% confidence intervals of the river water level series for lag 0 to lag 20 are presented in Fig. 3. Similarly, the PACF and the corresponding 95% confidence intervals of the river water level series are presented in Fig. 4. The ACF (Fig. 3) showed a significant correlation at 95% confidence level interval up to 14 hours of water level lag. In addition, the PACF showed significant correlation up to lag of 3 (3 hours). Result of correlogram plots of the data series shown in Figs. 3 and 4 imply that incorporating the water level values up to lag 3 hours can best represent the process in the catchment area under examination. Therefore, in this study, three antecedent values of water level are selected as input variables.

Using a given input/output data set, we construct a fuzzy inference in which partition the input space to reflect the premise part of the fuzzy inference system. As there are no preferable membership functions, we created an initial set of membership functions using grid partition method. In the common grid partitioning method, at the beginning of training, a uniformly partitioned grid is taken as the initial state. In this study, grid partition method was used to create the initial membership function matrix using the global bell functions for each of the input variables. We selected three membership functions for water level at *t-2*, *t-1*, and *t* respectively. As the parameters in the premise membership functions are adjusted, the grid evolves. After computing the gradient vector of the parameters of the membership functions, ANFIS employed an optimization technique to adjust the parameters to reduce some error measure (usually defined by sum of the squared difference between actual and

desired outputs).

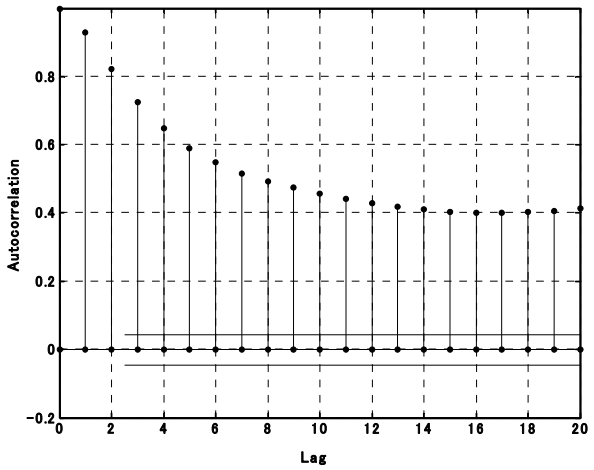

Fig. 3 Autocorrelation plot of the river water level

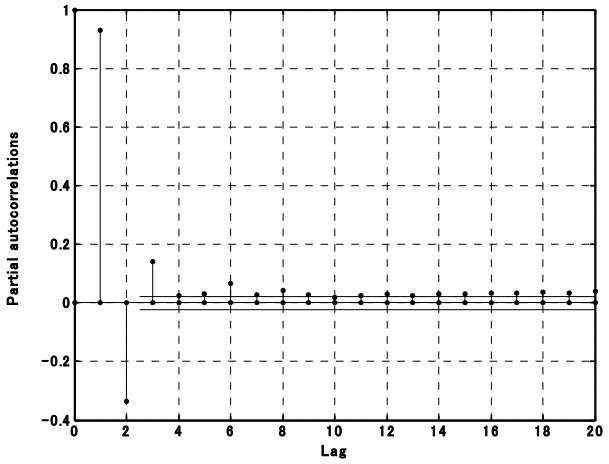

Fig. 4 Partial autocorrelation plot of the river water level

The ANFIS model applied in this study uses the hybrid learning algorithm, a combination of least square estimation and backpropagation (the gradient descent model), for membership function parameter estimation. In the forward pass of the hybrid learning algorithm, node outputs go forward until layer 4 and the consequent parameters are identified by the least-squares method. In the backward pass, the error signals propagate backward and the premise parameters are updated by gradient descent. The final fuzzy inference system model would ordinary be the one associated with minimum training error.

The performances of the models developed in this study were assessed using various standard statistical performance evaluation criteria. The statistical measures considered were correlation coefficient (r) and mean absolute percentage error (MAPE).

$$r = \left[\frac{\sum_{i=1}^{n} \left(WL_i^o - \overline{WL^o} \right) \left(W_i^p - \overline{WL^p} \right)}{\sqrt{\sum_{i=1}^{n} \left(WL_i^o - \overline{WL^o} \right)^2} \sqrt{\sum_{i=1}^{n} \left(WL_i^p - \overline{WL^p} \right)^2}} \right] \quad (13)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{WL_i^p - WL_i^o}{WL_i^o} \right| x100 \quad (14)$$

Where WL_i^o and WL_i^p are the observed and predicted water level at time t respectively; $\overline{WL^o}$ and $\overline{WL^p}$ are the mean of the observed and predicted water level; and n is the number of data.

IV. Result and Discussions

The fuzzy inference system has a total of 8 rules, and all of the rules include all the three input variables. In order to find the optimum membership parameters for the input models, the Sugeno-style ANFIS is employed. Therefore, the model contains 8 (2x2x2) rules.

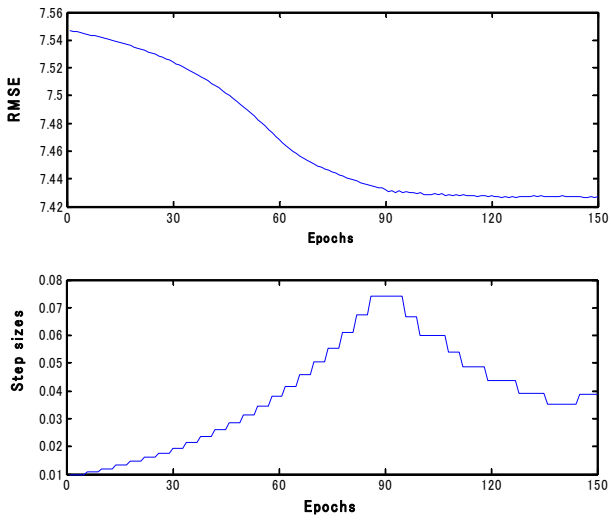

Fig. 5 Error curves during the learning process

The training RMSE as well as the variation of step-size for the input model at each-iteration is shown in Fig. 5. Usually the step size profile should be a curve, which goes uphill initially and reaches some maximum, then goes downhill till the end of training. This ideal step size is achieved by adjusting the initial step size and the increase and decrease rates. The membership functions of water level after training are shown in Fig. 6. From Fig. 6 we see how the final membership functions are trying to catch the local features of the training set.

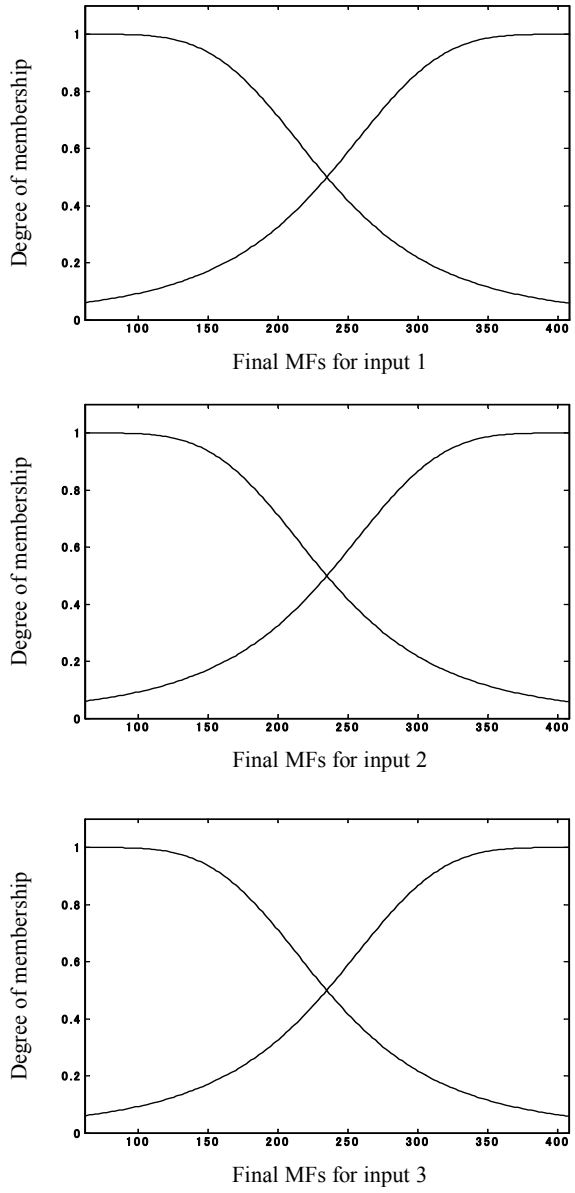

Fig. 6 The final membership functions for input parameters

Performance of the ANFIS model is compared in three data sets: (1) training sets, (2) verification sets, and (3) testing sets. Fig. 7 shows the observed water level on *x*-axis against the forecasted value on *y*-axis during the training, verification and testing respectively. In each of the scatter diagrams, the more perfectly the model was tested, the closer the points fall on the straight line. As could be concluded from those figures, the ANFIS was successful in learning the relationship between the input and output data. As shown in Fig. 7, the result of forecasting using the ANFIS model falls relatively close to the 45^{o} line. The results indicate that the generalization properties of the ANFIS model during the training, verification, and testing are comparable.

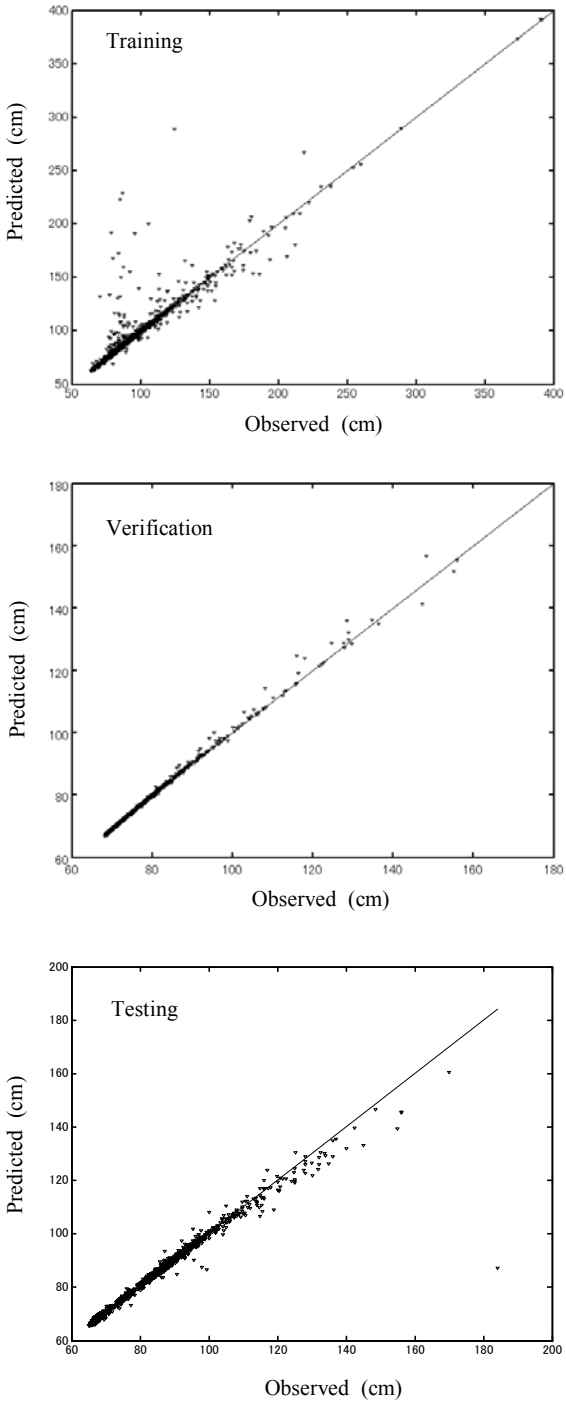

Fig. 7 Scatter plot of observed and forecasted water level during the training, verification and testing

A comparative prediction accuracy of the ANFIS model using two statistical indices (MAPE and correlation) at lead time 1-hour indicates that the ANFIS model is accurate and consistent in different subsets, where the value of MAPE are smaller (training = 2.15%, verification = 1.34% and testing = 1.15%) whereas all correlation values are also close to unity (training = 0.95, verification = 0.97 and testing = 0.98).

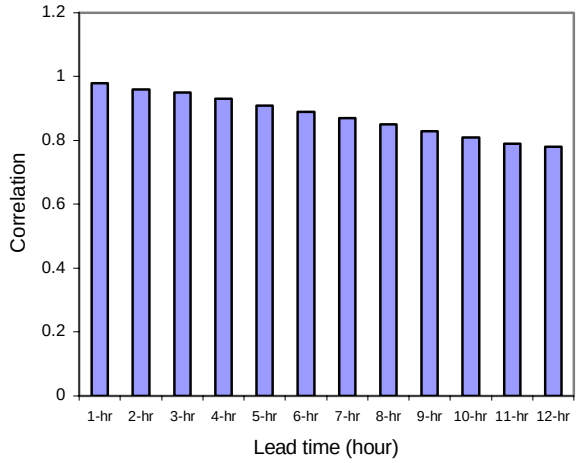

Fig. 8 The correlation values at different lead-time

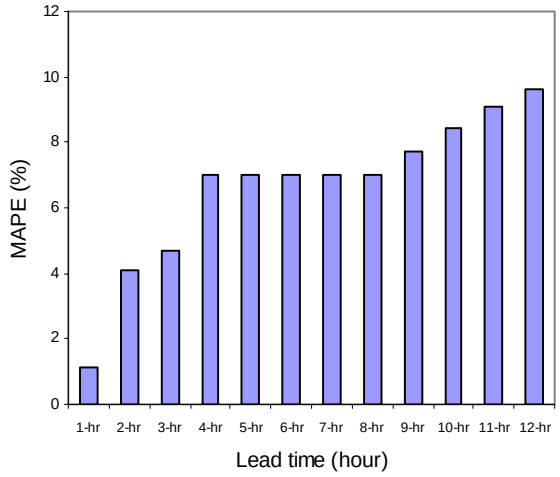

Fig. 9 The MAPE values at different lead-time

As a next task, a multi-step ahead prediction (12-hours ahead) is performed. Although the model has been trained in a one-step ahead forecast, it is desirable to investigate their performance in a multi-step ahead prediction. In this particular case, the predicted outputs were feed back into the networks to predict more values. It is to be noted that as the number of steps ahead increases, it is expected that the prediction error variance should also increase. For a prediction of shorter lead time, the performance indices of each model were comparable to each other and resulted in a high prediction accuracy as shown in Fig. 7. However, as the lead times increases, the performance of all models diminishes rapidly particularly at lead times >8 hour (Figs. 8 and 9). For 12-hour ahead forecast, the MAPE < 9.65 and correlation > 0.78 were obtained by ANFIS model. This result might suggest that the ANFIS has a great ability to learn from input-output patterns, which only represent three antecedent value of water level to produce a good generalization.

V. Conclusion

The current paper reports an outcome of a study aims at investigating the suitability of the adaptive network based fuzzy inference system for continuous water level modeling. A hybrid learning algorithm, which combines the least square method and the back propagation algorithm, is used to identify the parameters of the ANFIS. To illustrate the practical application of the adaptive network-based fuzzy inference system, the Cilalawi River was used as a case study. The Cilalawi River is located in the West Java Province, Indonesia. For this study, water levels data were available for a hydrological year of 2002 with a sampling interval of 1 hour. The number of antecedent water level that should be included in the input variables is determined by two statistical methods, i.e. autocorrelation function and partial autocorrelation function between the variables. Forecasting was done for 1-h until 12-h ahead in order to compare the models generalization at higher horizons.

The results demonstrate that the adaptive network-based fuzzy inference system model can be applied successfully and provide high accuracy and reliability for river water level estimation. In general, the adaptive network-based fuzzy inference system provide accurate and reliable water level prediction for 1-hour ahead where the MAPE=1.15% and correlation=0.98 was achieved. However the model accuracy deteriorates as the lead-time increases. Up to 12-hour ahead prediction, the adaptive network-based fuzzy inference system still shows relatively good performance where the error of prediction resulted was less than 9.65%. Therefore we conclude that the information gathered from the preliminary results provide a useful guidance or reference for flood early warning system design in which the magnitude and the timing of a potential extreme flood in the study area are indicated.

Acknowledgment

The authors would like to extend grateful thanks to Perusahaan Umum Jasa Tirta II (PJT II), West Java Indonesia for the cooperation and assistance during the data collection exercise. In a special way the authors wish to acknowledge Mr. Andri Sewoko assistance in providing the data on which this research was based.

References

[1] S. Sorooshian, and V.K. Gupta, "Model calibration. In: Singh, V.P. (Eds.), *Computer Models of Watershed Hydrology*," Water Resour. Publications, Colorado, 1995.

[2] Q. Duan, S. Sorooshian, and V.K. Gupta, "Effective and efficient global optimization for conceptual rainfall runoff models," *Water Resour. Res*., vol. 28, pp. 1015–1031, 1992.

[3] P.C. Nayak, K.P. Sudheer, D.M. Rangan, and K.S. Ramasastri, "Short-term flood forecasting with a neurofuzzy model," Water Resour. Res. vol. 41, pp. 2517–2530, 2005.

[4] S. Openshaw, and C. Openshaw, "*Artificial Intelligence in Geography,"* Chichester : John Wiley & Sons Ltd, 1997.

[5] B. Bazartseren, G. Hildebrandt, and K.P. Holz, "Short-term water level prediction using neural networks and neuro-fuzzy approach," *Neurocomputting*., vol. 55, pp. 439-450, 2003.

[6] M. Valenca, and T. Ludermir, "Monthly streamflow forecasting using an neural fuzzy network model," *Proceedings of the Sixth Brazilian Symposium on Neural Networks*., vol. 6, pp. 117-119, 2000.

[7] P.C. Nayak, K.P. Sudheer, and K.S. Ramasastri, "A neuro-fuzzy computing technique for modeling hydrological time series," *J. Hydrol*., vol. 291, pp. 52-66, 2004.
[8] H. Vernieuwe, O. Georgieva, B.D. Baets, V.R.N. Pauwels, N.E.C. Verhoest, and P.D. Troch, "Comparison of data-driven Takagi-Sugeno models of rainfall-discharge dynamics," *J. Hydrol*., vol. 291, pp. 173-186, 2005.
[9] J.S.R. Jang, C.T. Sun, and E. Mizutani, "*Neuro-fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*," Prentice Hall, New Jersey, 1997.
[10] K. P. Sudheer, A. K. Gosain, and K. S. Ramasastri, "A data-driven algorithm for constructing artificial neural network rainfall–runoff models," Hydrol. Process. vol. 16, 1325–1330, 2002.