# Urban flood forecasting using a neuro-fuzzy technique

C. Choi, J. Ji, M. Yu, T. Lee, M. Kang & J. Yi
*Department of Civil & Transportation Engineering, Ajou University, South Korea*

## Abstract

In the conventional flood forecasting process, a rainfall–runoff model is used to predict runoff at a specific location. However, the process of determining the required parameters for the model is sometimes very complicated and requires extensive information and data. In addition, considerable amount of uncertainties may be included during the parameter estimation processes. Errors can occur during the pre-processing and main processing stages of the modeling, and errors from each step accumulate into the model result. In this study, a neuro-fuzzy technique is used to minimize the amount of uncertainties included in a conventional flood forecasting model for more accurate forecasting of floods. The adaptive neuro-fuzzy inference system (ANFIS), which is a data-driven model that combines a neural network and the fuzzy technique, can decrease the amount of physical data required for constructing a conventional model. By using only rainfall and water level data, ANFIS can easily construct and evaluate a flood forecasting model. Furthermore, the model construction process is relatively simple, and reliable results can be efficiently obtained in a reasonably short time once the model is developed. The developed model is applied to the Tancheon basin in Korea. The water level at the Daegok Bridge, which is located downstream of Tancheon, is forecast by the neuro-fuzzy method. The applicability and suitability of the model are studied by comparing the result with the observed stream level data from 2007 to 2011 in the Tancheon basin area. Tancheon is a tributary of the Han River and begins from the city of Yongin in Gyeonggi-do. It has a total length of 35.6 km and an area of 302 $km^2$. The water level data from t + 1 to t + 18 is estimated by ANFIS using 10-min interval data. The results showed that the average height error was 24.48% and the average RMSE was 0.367 m.
*Keywords: ANFIS, neuro-fuzzy technique, flood forecasting.*

WIT Transactions on The Built Environment, Vol 122, © 2012 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/UW120221

## 1 Introduction

Generally, physical models used in predicting floods, such as a rainfall–runoff model require an extremely large amount of data to determine the model parameters, which causes them to inherit a factor of uncertainty.

In this research, to minimize the problems and uncertainty of the flood forecast system, a neuro-fuzzy reasoning method was used to predict the river's water level for better flood estimation. The adaptive neuro-fuzzy inference system (ANFIS) can predict the water level and create a model using only the observed rainfall and water level data in a basin; it does not need the massive amount of physical data necessary for the construction step of the physical model. Moreover, once the model is created, simply inputting data can produce a reliable result in a very short period of time.

Recently, as model development has become easier and prediction accuracy has become better, neuro-fuzzy techniques such as ANFIS have become widely used in the water research. Studies on neuro-fuzzy reasoning techniques not only try to increase the accuracy of the flood volume prediction but also compare the prediction results of various techniques. Vernieuwe *et al.* [1] recently compared the clustering technique results for the Takagi–Sugeno type model, which is generally used in data-driven techniques that use rainfall–runoff analysis. Grid partitioning, subtractive clustering, and Gustafson-Kessel (GK) clustering have been used for the clustering technique, with GK clustering showing the best results. Chen *et al.* [2] used the ANFIS model and estimated the flood volume for the Choshui River in Taiwan. The rainfall and runoff data from the rainfall observation station in the basin were used. The results showed that the consistency effects and upstream runoff information impact the flood estimation model. Dastorani *et al.* [3] compared an artificial neural network (ANN) with ANFIS by estimating the rainfall amount using weather observation data from the Yazd Meteorological Station in central Iran as input data. Although the ANN and ANFIS model results appeared to be very different, both models showed extremely good rainfall estimation results. Valizadeh *et al.* [4] applied the ANFIS model to the Klang Gate Reservoir in Malaysia for reservoir inflow estimation. Their results showed that it was possible to precisely estimate the water level when a sufficient time interval is given between the input and output data.

An ANFIS model for water level estimation at the Daegok Bridge, located downstream of Tancheon, was developed in this study. For the model, the input data were rainfall and water level observation data from 2007 to 2011, and seven flood occurrences where the water level was 5 $m$ or more were used. Data collected at 10-min intervals were used; during the simulation, water levels from 10 min to 3 h (t + 1 to t + 18) after the rainfall event were simulated. The seven selected flood occurrences were arranged in various combinations to create the ANFIS training, checking, and testing data. The peak water level ratio and mean square error of 40 sets of models and 720 testing results were compared. In particular, the change in training data and testing data results were compared, and the main focus was placed on the effect of the training data.

## 2 Fuzzy set theory

Ambiguity and uncertainty that cannot be expressed numerically constantly surround human beings and the natural world. In particular, ambiguous statements in the water resource sector such as "the reservoir water level is high," "there was a large amount of rainfall in the upstream area," and "runoff was large" need more explicit definitions. Although these are common phrases, mathematical calculations and computer languages use binary systems, which make them difficult to describe in the systems. In water resource situations, problem solving in a binary system such as "yes" or "no" is extremely difficult. Zadeh [5] quantified these ambiguities and developed the fuzzy set theory for logical accessibility of mathematical calculations and computer language. Mamdani and Assilian [6] used the fuzzy algorithm for language modeling of a complicated system, after which fuzzy control became widely applied.

The basic concept of fuzzy set theory can be explained by comparing it to a conventional set theory. If the state or membership of element $x$ belonging to set $A$ is represented as $\mu_{\bar{A}}$, then it can be defined as either "$x$ is in set $(\mu_{\bar{A}}(x) = 1)$" or "$x$ is not in set $A$ $(\mu_{\bar{A}}(x) = 0)$" in a conventional set. This means that element $x$ can only belong or not belong to set $A$.

$$\mu_{\bar{A}}(x) = \begin{cases} 1, & if\ x \in A \\ 0, & if\ x \notin A \end{cases} \tag{1}$$

where $x$ belongs to $A$ and $A$ is a subset of $X$. In contrast, a fuzzy set has a membership continuity using $\mu_{\tilde{A}}$ that can be defined by three states: "$x$ is completely included in fuzzy set $(\mu_{\tilde{A}}(x) = 1)$," "$x$ is not included in fuzzy set $(\mu_{\tilde{A}}(x) = 0)$," or "$x$ is partially included in fuzzy set $A$ $(\mu_{\tilde{A}}(x) = 0{\sim}1)$."

$$\mu_{\tilde{A}}(x) = \begin{cases} 1, & if\ x \in A \\ 0{\sim}1, & if\ x \in A \\ 0, & if\ x \notin A \end{cases} \tag{2}$$

where $\tilde{A}$ is a fuzzy set defined within domain $X$. Reservoir state evaluation using the fuzzy set is similar to a human thought process, and it is useful for quantitative evaluation of a reservoir.

ANFIS is a technique that uses a neural network and fuzzy theory simultaneously and automatically controls the input and output information to fit the rules by using the neural network structure and learning ability. A backpropagation algorithm is generally used for learning; this algorithm has the structure of a multi-layered perceptron and operates with a learning stage and calculation step. In the learning stage, the input and goal conditions are mostly given as input–output pairs; output is calculated first for each input condition. Furthermore, to decrease the difference between the desired and actual outputs, the connection intensity is altered (reverse direction process). Once the alteration process ends, the same learning process is repeated until the optimal connection

intensity is reached. For the calculation step, once the input is given, the appropriate output is calculated depending on the connection intensity.

## 3 Test basin application

Data on the Tancheon basin rainfall and water level were used to evaluate the suitability of using ANFIS for water level estimation and to analyze the effects of using the data for model development. Tancheon begins at Gyeonggi-do and flows into the Han River by going through Seongnam-si and Seoul; more than half the length, approximately 25 $km$, flows through Seongnam-si, which makes it a typical urban river. The basin has a total area of 302 $km^2$ and a total length of 35.6 $km$, as shown in fig. 1. According to the last 11 years of rainfall data obtained from the Tancheon basin, the mean annual rainfall amount is approximately 1238.3 $mm$, and approximately 959 $mm$ of rainfall (77.4%) is concentrated between June and September. Since the city developed around the river, it is especially sensitive to flood damage.

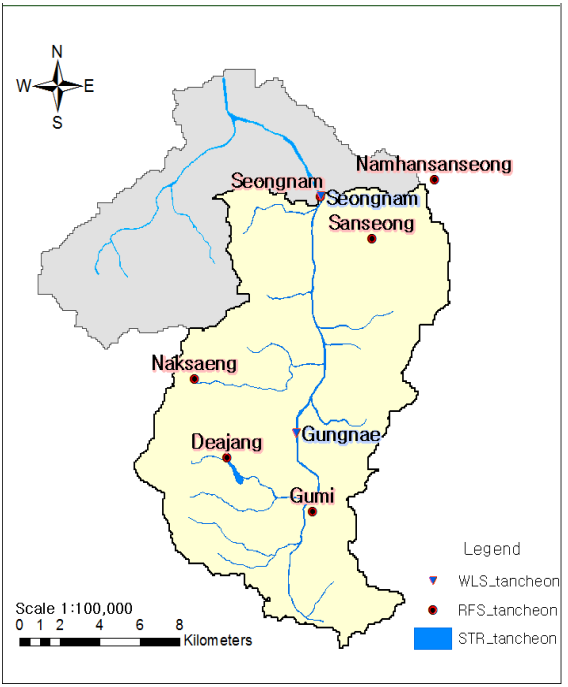


Figure 1: Map of Tancheon Basin.

Seven rainfall observation stations were selected to estimate the basin's average rainfall amount. The Seongnam water level observation station located near Daegok Bridge was used for water level estimation.

### 3.1 Model development

Relatively large runoff events for the Tancheon basin from 2007 to 2011 were selected for the ANFIS model development; they were denoted. To compare the model estimation results according to the training data essential for model learning, data with different numbers of peaks representing the data length, maximum water level, and flood trend were selected table 1.

Table 1: Characteristics of selected data.

| | Data length(day) | Max water level (m) | Number of peak |
|---|---|---|---|
| Data 1 | 10 | 5.38 | 6 |
| Data 2 | 4 | 6.74 | 4 |
| Data 3 | 2.5 | 5.12 | 1 |
| Data 4 | 4 | 4.92 | 3 |
| Data 5 | 1.5 | 5.73 | 1 |
| Data 6 | 2.5 | 5.11 | 5 |
| Data 7 | 5.5 | 6.13 | 6 |

To compare the results according to changes in the data for model development, the seven selected data values were divided into 40 sets. Sets 1–3 used data 1 as the training data. Sets 4–6, 7–9, 10–12, and 13 used data 2, 3, 6, and 7, respectively, as the training data. To compare the simulation results for each data set, data 4, 5, and 7 were used for testing.

Table 2: Model composition.

| | Observed Precipitation | Observed Water Level | Estimated Water Level |
|---|---|---|---|
| Model A | $P_t$ , $P_{t-1}$ | $H_t$ | $H_{t+1} \sim H_{t+18}$ |
| Model B | $P_t$ , $P_{t-1}$ | $H_t$ , $H_{t-1}$ | $H_{t+1} \sim H_{t+18}$ |
| Model C | $P_t$ , $P_{t-1}$ | $H_t$ , $H_{t-1}$ | $H_{t+1} \sim H_{t+18}$ |
| Model D | $P_t$ , $P_{t-1}$ | $H_t$ , $H_{t-1}$ | $H_{t+1} \sim H_{t+18}$ |
| Model E | $P_t$ , $P_{t-1}$ | $H_t$ , $H_{t-1}$ | $H_{t+1} \sim H_{t+18}$ |

The rainfall and water level data were collected at 10 min intervals for the model. To compare the results according to the input data combinations, t to t - 3 (30 min before the current time) rainfall data and t to t - 2 (20 min before the current time) water level data were used to develop five models, as shown in table 2. The height error and RMSE were used to compare the results.

$$HE = \frac{|HP_{obs.} - HP_{est.}|}{HP_{obs.}} \times 100(\%) \quad (3)$$

where $HE$ is the height error ratio calculated as a percentage, $HP_{obs.}$ is the actual peak water level, and $HP_{est.}$ is the estimated peak level.

$$RMSE = \sqrt{\frac{\sum(H_{est.} - H_{obs.})^2}{n-1}} \quad (4)$$

where $H_{est.}$ is the estimated water level, $H_{obs.}$ is the actual water level, and $n$ is the number of data.

### 3.2 Comparison of results according to the configuration of the model input data

The height error and RMSE are shown in figs. 2. and 3. The height error does not show any special tendency, but the RMSE shows that Model A appeared to produce excellent results. This is because the size of the basin was relatively small, and the effects of the observed rainfall and water level had the biggest impacts on the runoff for the first 10–20 min. The average height error and the average RMS were 24.48% and 0.367 m, respectively.

### 3.3 Comparison of results according to the training data

The effects of the changes in training data were compared. The time lengths of data 1, 2, and 7 were 10, 4, and 5.5 days, respectively, and the maximum water levels were 5.38, 6.74, and 6.13 $m$, respectively. Also, each data had six, four, and six peak water levels, respectively. When comparing the three training data values, although data 1 had the longest duration of 10 days and six peak numbers, the maximum water level was lower than those of data 2 and 7 by 1.36 m (25.3%) and 0.75 m (13.9%), respectively. Although data 2 had a time duration of 4 days, which was 6 and 1.5 days shorter than data 1 and 7, respectively, its maximum water level of 6.74 m was the highest. Data 7 had the second-longest duration of 5.5 days and six peaks, and its maximum water level was 6.13 $m$, table 3.

Table 3: Characteristics of data.

| | Data Length (day) | Max. Water Level (m) | Number of Peak |
|---|---|---|---|
| Data1 | 10 | 5.38 | 6 |
| Data2 | 4 | 6.74 | 4 |
| Data7 | 5.5 | 6.13 | 6 |

According to the training data, when the water level estimation results were compared by the height error, data 7 had the smallest average height error of 12.2% and smallest deviation range of 12.5% from the median, table 4, fig. 4. When comparing data 1 and 2, using the latter as the training data produced a

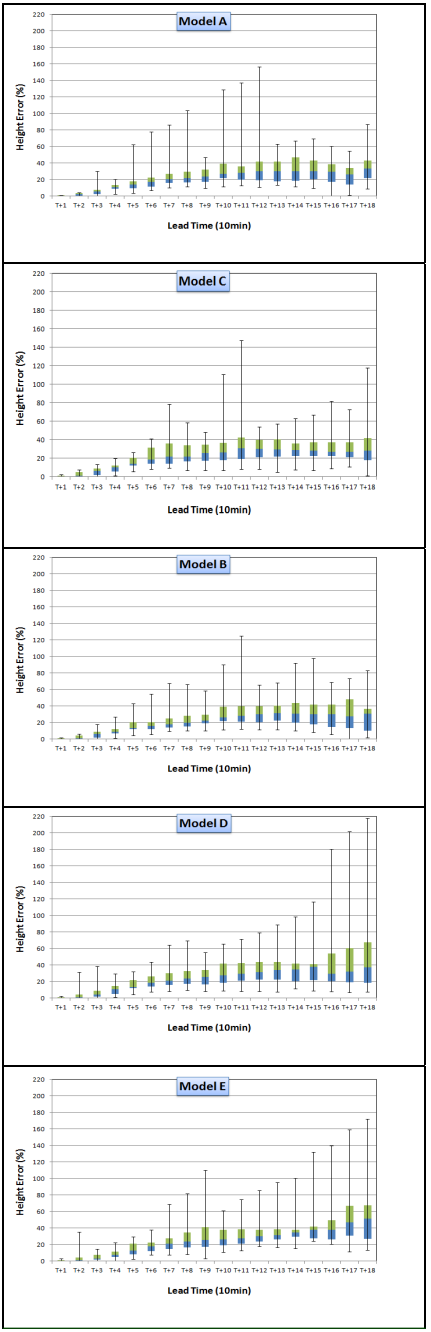


Figure 2: Height error comparison.

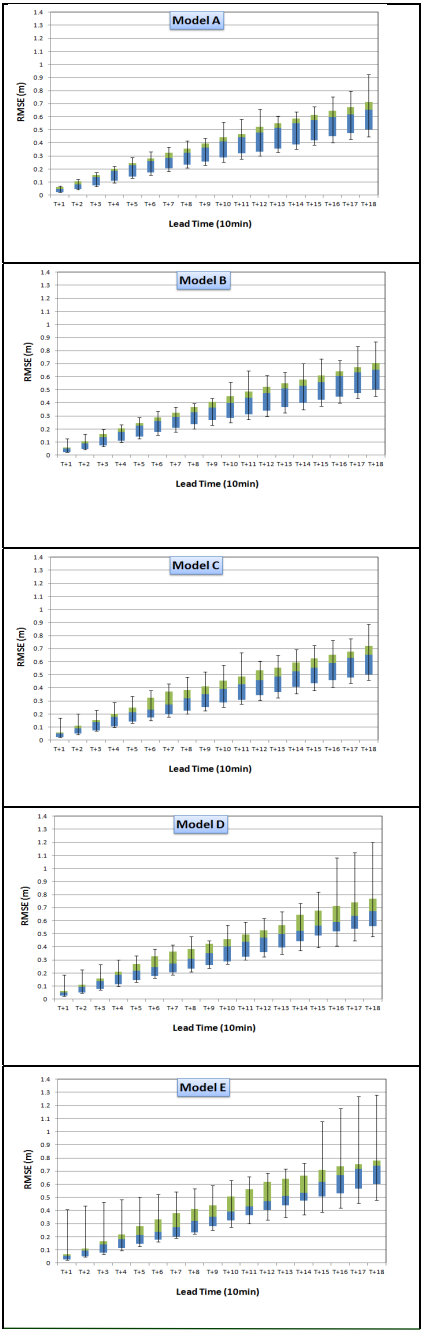


Figure 3: RMSE comparison.

Table 4: Comparison of training data.

| | Height Error (%) | | RMSE (m) | |
|---|---|---|---|---|
| | Average | Median | Average | Median |
| Data 1 for Training | 26.7 | 29.0 | 0.358 | 0.344 |
| Data 2 for Training | 26.9 | 23.9 | 0.376 | 0.365 |
| Data 7 for Training | 12.2 | 12.5 | 0.327 | 0.318 |

smaller height error up to a t + 14 lead time. On the other hand, when data 1 was used as the training data, the height error was smaller for lead times from t + 15 to t + 18. The deviation from the median for data 1 showed a narrower range, whereas the deviation from the median for data 2 showed a considerably wider range.

When compared according to the RMSE, the model using data 7 as the training data showed the best overall results in terms of the average and median lead times, table 4, fig. 5. Also, the deviation range from the median showed a

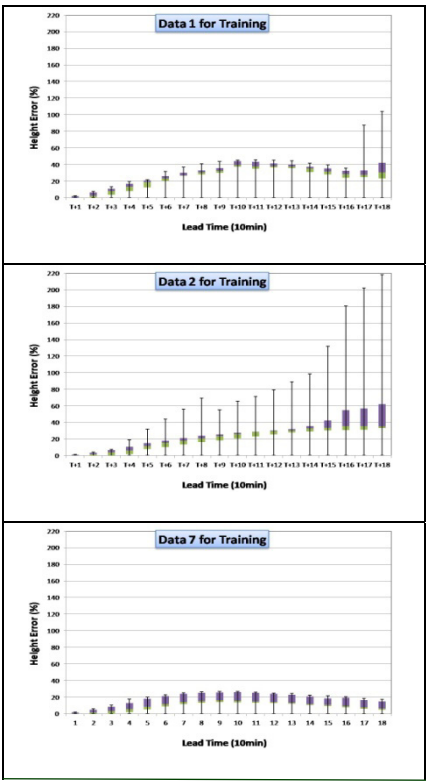


Figure 4: Height error comparison.

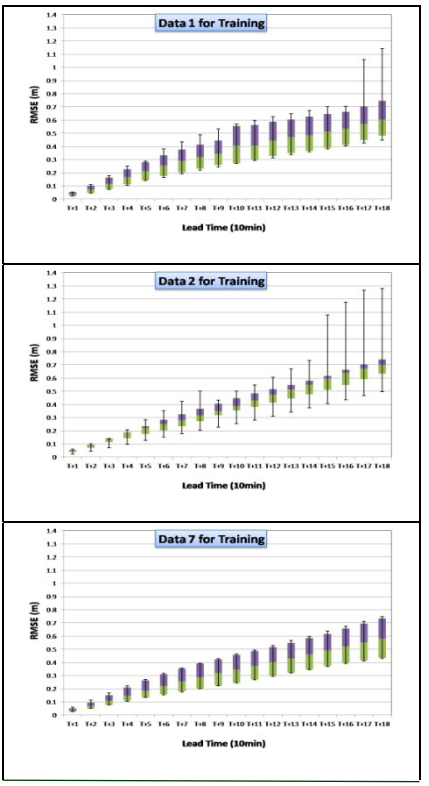


Figure 5: RMSE comparison.

narrower range for data 7 than for data 1 or 2. Data 2 showed a better median than data 1 for each lead time. Data 2 showed a narrower 25% deviation from the median but a larger overall deviation.

If the training data length was too short, the deviation of the estimated value became large, and the large maximum flood enhanced the estimation accuracy. If the peak value occurred frequently and increased and decreased the water level, it could also help develop a better model.

### 3.4 Comparison of results according to the testing data

The three testing data were compared, and the maximum water level increases were in the order of data 4, 5, and 6. Data 5 showed a relatively short data length and simple form. Data 4 had a slightly complex form with three peaks and a relatively low maximum water level. Data 7 had the longest duration, the most complex form, and the highest water level. In particular, the maximum water level of data 7 was higher than those of all the other data, including the training and checking data, table 5.

Table 5: Characteristics of selected data.

| | Data length (day) | Max. water level (m) | Number of peak |
|---|---|---|---|
| Data 4 | 4 | 4.92 | 3 |
| Data 5 | 1.5 | 5.73 | 1 |
| Data 7 | 5.5 | 6.13 | 6 |

When compared by the height error, the water level estimation according to the testing data showed the best results, table 6, fig. 6. A 25% deviation from the median was concentrated in a relatively narrow range, and the total range of deviation was also modest. Although data 5 showed the best median and good 25% deviation range for lead times of t + 14 to t + 18, it showed the largest deviation overall. Data 7 showed a significantly large median compared to data 4 and 5 and a relatively wide range of deviation, fig. 6.

Table 6: Comparison of testing data.

| | Height Error (%) | | RMSE (m) | |
|---|---|---|---|---|
| | Average | Median | Average | Median |
| Data 4 for Training | 18.50 | 17.59 | 0.28 | 0.26 |
| Data 5 for Training | 22.14 | 19.25 | 0.44 | 0.42 |
| Data 7 for Training | 32.81 | 30.91 | 0.40 | 0.38 |

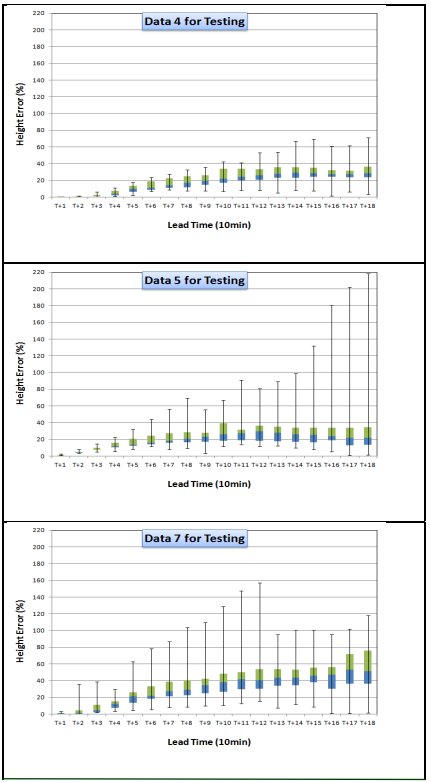


Figure 6: Height error comparison.

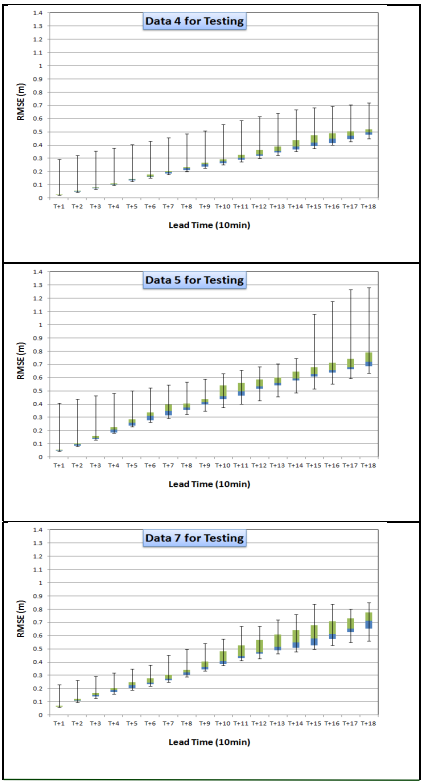


Figure 7: RMSE comparison.

When compared by the RMSE and height error, the model using data 4 as the testing data showed the best results, table 6, fig 7. The 25% deviation range was significantly concentrated and showed very high accuracy. Also, when comparing the medians of data 7 and 5, the former showed better estimation results than the latter, unlike the case for the height error. Data 7 also showed a more concentrated deviation range. However, data 5 showed a slightly more concentrated 25% deviation range, fig. 7.

In summary, if testing data larger than the training data and checking data are used to develop a model, the accuracy may decrease. Therefore, building a model with the largest observed rainfall and flood events can increase the accuracy of estimating future floods.

## 4 Conclusion

In this study, neuro-fuzzy techniques were applied to observe rainfall and water level data for predicting the water levels at Daegok Bridge in the Tancheon basin. Seven data values were extracted from flood events of 2007–2011 as

model input data. Data with 10-min intervals were used to simulate water levels from t + 1 to t + 18. The training, checking, and testing data were combined to obtain 40 sets and 720 testing results, and the results were compared according to the RMSE and peak levels ratio.

Although no trend was found when the rainfall and water level data were compared, model A showed the smallest RMSE. This is because the Tancheon basin has a relatively small area, and the rainfall and the water level data observed 10–20 min prior appeared to have the largest impact.

If the data used for training are over a time duration that is too short, the deviation of the estimation value may be large. It is better to use data with a large maximum flood volume as training data to increase the estimation accuracy. Also, data with frequent occurrences of the peak water level can enhance the development of a better model. Therefore, when developing an ANFIS model, using the largest and most complex observed data can help develop models with improved flood estimation accuracy.

## Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 2011-0029851).

## References

[1] Vernieuwe, H., Georgieva, O., De Baets, B., Pauwels, V.R., Verhoest, N.E. & De Troch, F.P.. Comparison of data-driven Takagi-Sugeno models of rainfall-discharge dynamics. *Journal of Hydrology*, **302(1-4)**, pp. 173-186, 2005.

[2] Chen, S.H., Lin, Y.H., Chang, L.C. & Chang, F.J., The strategy of building a flood forecast model by neuro-fuzzy network. *Hydrological Processes*, 2**0(7)**, pp. 1525-1540, 2006.

[3] Dastorani, M.T., Afkhami, H., Sharifidarani, H. & Dastorani, M., Application of ANN and ANFIS Models on Dryland Precipitation Prediction (Case Study: Yazd in Central Iran). *Journal of Applied Sciences*, pp. 2387-2394, 2010.

[4] Valizadeh, N., El-Shafie, A., Mukhlisin, M. & El-Shafie, A.H., Daily water level forecasting using adaptive neuro-fuzzy interface system with different scenarios: Klang Gate, Malaysia. *International Journal of Physical Sciences*, Vol. 6, Issue 32, pp. 7379-7389, 2011.

[5] Zadeh, L.A., Fuzzy Sets. Information and Control, pp. 338-353, 1965.

[6] Mamdani, E.H. & Assilian, S., An Experiment in Linguistic Synthesis with a Fuzzy Logic Controller. *International Journal of Man-Machine Studies*, **7(1)**, pp. 1-13, 1975.