Comparison of a data-driven model and a physical model for flood forecasting

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

Abstract

Flooding from torrential rain occurs in a short amount of time, while drought lasts for a longer period; the former may inflict huge losses in terms of both life and property. For these reasons, considerable research has been performed in the field of flood control system development. A physical model is mainly used for flood forecasting and warning. However, physical rainfall-runoff models for the conventional flood forecasting process require extensive information and data, and include uncertainties that can accumulate errors during the modeling process. On the other hand, ANFIS, which is a data-driven model combining the neural network and fuzzy techniques, can decrease the amount of physical data required for the construction of a conventional model and easily construct and evaluate a flood forecasting model using only rainfall and water level data. However, data-driven models have the disadvantage that they do not provide mathematical and physical logic, so that there are no logical correlations between the input and output data of the model. This study analyzes the characteristics of a data-driven model, ANFIS, according to its functional options and input data, such as changes in the clustering radius and the training data length. In addition, the suitability of ANFIS is evaluated through comparison with the results of HEC-HMS, which is widely used for rainfall-runoff models. In this study, the neuro-fuzzy technique is applied to the Cheongmicheon Basin using the observed precipitation and stream level data from 2008 to 2011.
Keywords: ANFIS, HEC-HMS, neuro-fuzzy technique, flood forecasting.

1 Introduction

Flooding occurs when a river overflows, usually due to torrential rain that dramatically increases runoff in a short time. Unlike drought, flooding occurs

doi:10.2495/FRIAR120111

quickly, and may cause severe damage to life and property. In Korea, precipitation is most concentrated in the summer season, and flooding frequently occurs between June and September. Steady efforts have been made to alleviate this problem, and accurate flood forecasting can be useful in this application. Some of the widely used physical models for flood forecasting include a storage function, HEC-HMS, and a unit hydrograph; however, these models have the disadvantages that the parameters require complex calculation, and many errors accumulate in the modeling process. In comparison, a data-driven model only requires rainfall and water level data from a basin for flood forecasting. Furthermore, once the model is established, it can promptly yield reliable outcomes with data input.

Advanced studies in this field are underway, including comparative analyses of flood forecasting outcomes using various methods, as well as analysis of the accuracy of the ANFIS model in relation to changing membership functions. Vernieuwe *et al.* [1] recently conducted a study to examine outcomes of a Takagi-Sugeno-type model using a changing clustering method; such an approached is a widely used rainfall-runoff model based on the data-driven technique. Chen *et al.* [2] applied the ANFIS model to forecast flooding of the Choshui River in Taiwan. They used rainfall and water level data from the precipitation observatory, and the result showed that the lasting effect and the upstream water level were essential elements in flood forecasting. Shin *et al.* [3] developed a forecasting simulator for water demand based on mathematical and neural network methods as linear and non-linear models to implement optimal water demand forecasting. It was shown that multilayer perceptron (MLP) and ANFIS, respectively, can be applied to obtain better forecasting results in multi-regional water supply systems with a large scale and local water supply systems with a small or medium scale than conventional methods.

In this study, the ANFIS model was established for flood forecasting at Wonbugyo point, which is located downstream of the Cheongmicheon Basin, a branch of the Namhan River. The precipitation and water level data between 2007 and 2011 were used as input for five rainfall events. Simulation was conducted to forecast the water level for three hours, at 10 minute intervals, starting from time T (from T+1 to T+18), and the results were compared with those of HEC-HMS.

2 Fuzzy set theory

The term *fuzzy* indicates an uncertain condition where it cannot be clearly decided whether a member of a set belongs to the set or not [4]. Fuzzy inference induces a statement from a number of fuzzy statements, and works similarly to the general inference method. For example, there are many situations that require clarification, such as “the water level of the reservoir seems *high*” and “the runoff is *substantial*.” These are hard to handle with the binary codes of computer languages. Zadeh [4] developed fuzzy set theory to logically approach these ambiguous circumstances through quantitative mathematics and computer languages.

In a fuzzy set, the membership of the set has continuity. This membership can be described in three ways:

i) x belongs to fuzzy set A ($\mu_{\tilde{A}}(x) = 1$);

ii) x does not belong to fuzzy set A ($\mu_{\tilde{A}}(x) = 0$);

iii) x belongs to fuzzy set A to some degree ($\mu_{\tilde{A}}(x) = 0{\sim}1$).

ANFIS allows a user to construct a membership function with input and output data when the data are insufficient to determine the format of the membership function. Jang [5] developed ANFIS by combining the benefits of the Sugeno-type fuzzy inference system (FIS) and neural networks. ANFIS is particularly suitable for the analysis of advanced, nonlinear systems such as the rainfall-runoff model.

3 Test basin application

Rainfall and water level data from the Cheongmicheon Basin were used to compare the suitability of a data-driven model, ANFIS, and a broadly used physical model, HEC-HMS. The Cheongmicheon Basin covers the central area of the Han River system, ranging in longitude from 127°20′ to 127°44′E and in latitude from 36°56′ to 37°13′N. The basin has a triangular shape with a mild slope. The total basin area is 595.70 km^2, and the entire length is 60.8 km. The maximum east-west span is 39 km, and maximum north-south span is 31 km (Figure 1). Annual precipitation amounts to 1,061.9 mm on average, mostly concentrated between June and September. To calculate average rainfall, six

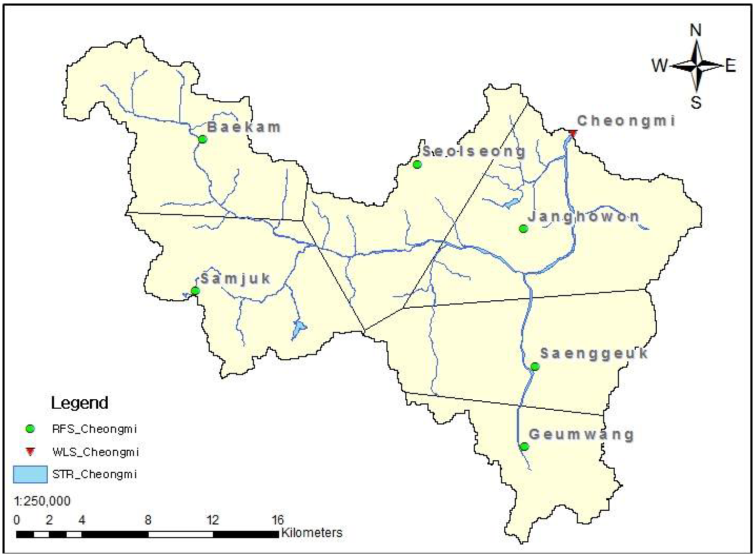

Figure 1: Map of Tancheon basin.

observatories – Saenggeuk, Seolseong, Samjuk, Janghowon, Geumwang, and Baekam – were selected, and their impact was calculated based on the Thiessen weighting method (Figure 1, Table 1). Water level data were drawn from the Cheongmi observatory located in the Wonbugyo area.

Table 1: Rainfall station of the Cheongmi basin.

Observatory	Coordinates		Thiessen Coefficient
	Longitude	Latitude	
Baekam	127° 22′ 48″	37° 9′ 36″	0.174
Samjuk	127° 22′ 28″	37° 4′ 36″	0.164
Seolseong	127° 31′ 39″	37° 8′ 45″	0.140
Janghowon	127° 36′	37° 6′ 36″	0.236
Saenggeuk	127° 36′ 25″	37° 2′ 4″	0.186
Geumwang	127° 36′	36° 59′ 24″	0.100

3.1 Model development

To construct a flood forecasting model, five datasets were selected for rainfall events that yielded substantial runoff at the Cheongmicheon Basin between 2007 and 2011. In order to examine changes due to training data length, Data 1 was further divided into Data 1_1, Data 1_2, and Data 1_3 for different lengths of input data (Table 2).

Table 2: Characteristics of selected data.

	Data Length (day)	Max Water Level (m)	Max Discharge (㎥/sec)
Data 1_1	15	4.50	588.98
Data 1_2	11	4.50	588.98
Data 1_3	5	4.50	588.98
Data 2	5	4.76	656.93
Data 3	4	3.36	326.97
Data 4	1	4.55	601.82
Data 5	2	4.74	651.60

Ten minute interval data were used for both rainfall and water level. To examine the influence of data construction, five model compositions were established for precipitation data at lead time between t and t-3, and water level data at lead times between t and t-2 (Table 3). Discharge error and RMSE were examined to compare the results (Equations (1) and (2)).

Table 3: Model composition.

	Observed Precipitation	Observed Discharge	Estimated Discharge
Model A	P_t, P_{t-1}	D_t	$D_{t+1} \sim D_{t+18}$
Model B	P_t, P_{t-1}	D_t, D_{t-1}	$D_{t+1} \sim D_{t+18}$
Model C	P_t, P_{t-1}, P_{t-2}	D_t, D_{t-1}	$D_{t+1} \sim D_{t+18}$
Model D	P_t, P_{t-1}, P_{t-2}	D_t, D_{t-1}, D_{t-2}	$D_{t+1} \sim D_{t+18}$
Model E	$P_t, P_{t-1}, P_{t-2}, P_{t-3}$	$D_t, D_{t-1,} D_{t-2}$	$D_{t+1} \sim D_{t+18}$

$$DE = \frac{|DP_{obs.} - DP_{est.}|}{DP_{obs.}} \times 100(\%) \tag{1}$$

DE indicates discharge error ratio (%); $DP_{obs.}$, observed peak flow; $DP_{est.}$, estimated peak flow.

$$RMSE = \sqrt{\frac{\sum(D_{est.} - D_{obs.})^2}{n-1}} \tag{2}$$

$D_{est.}$ represents estimated water level data; $D_{obs.}$, observed water level data; n, data length.

3.2 Changes due to different clustering radii

Clustering is a basic element in many system modeling algorithms and classifications. Its purpose is to produce compact and accurate results by grouping an extensive amount of input data by certain criteria. In the ANFIS model, a user can determine the clustering radius; the smaller the radius, the more finely the input data are grouped, yielding a greater number of rules. In this study, the clustering radius varied between 0.2 and 0.6 at intervals of 0.1; the aim of selecting these radii was to examine changes in discharge error and RMSE.

The results show that there is no significant difference in RMSE; as for discharge error, its average value was lowest (4.23%) at a clustering radius of 0.5, and the deviation from the median was also least using this parameter. The deviation was largest at a clustering radius of 0.2; as the radius increased, the deviation also gradually decreased. When the clustering radius exceeded 0.6, an error occurred, so that the system failed to produce the rules that are required to construct the ANFIS model.

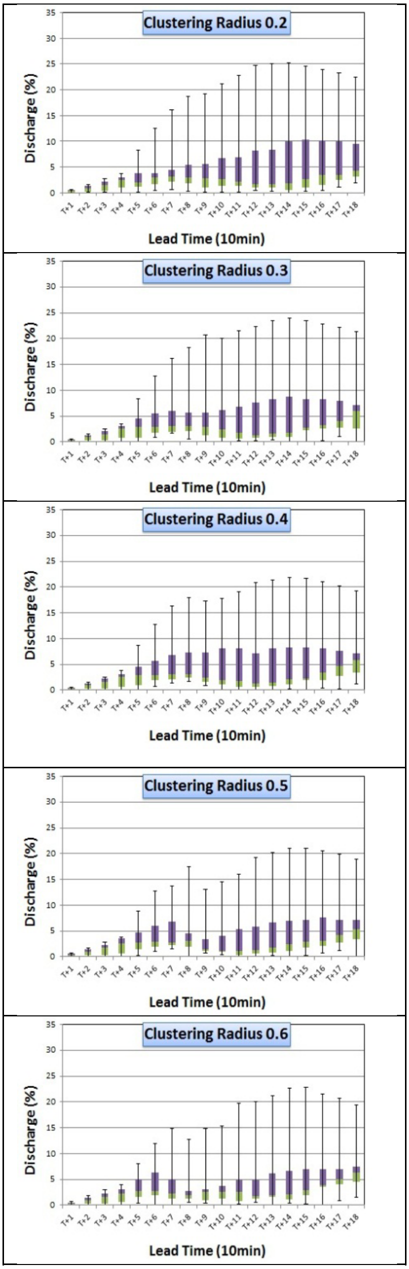

Figure 2: Discharge error comparison.

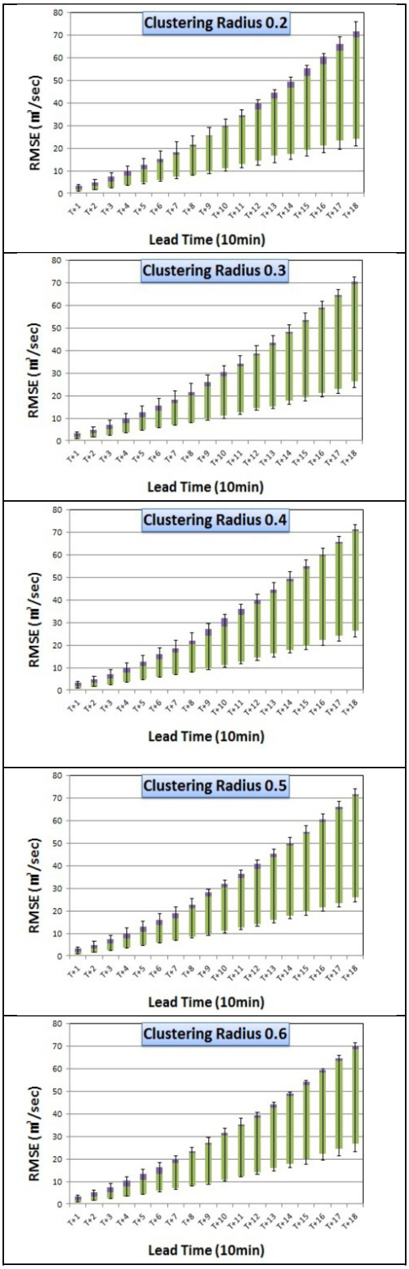

Figure 3: RMSE comparison.

Table 4: Clustering radius comparison.

	Discharge Error Average (%)	Discharge Error Median (%)
Clustering Radius 0.2	4.820	2.232
Clustering Radius 0.3	4.687	2.199
Clustering Radius 0.4	4.614	2.173
Clustering Radius 0.5	4.229	2.190
Clustering Radius 0.6	4.431	2.202

The experimental results showed that in the ANFIS model, reducing the clustering radius of the training data and producing many rules did not necessarily lead to better results in flood forecasting. It will be desirable to calculate the optimal radius for the model based on the particular characteristics of precipitation and the basin area.

3.3 Changes due to different training data lengths

As a data-driven model, ANFIS depends on training data for accuracy. It is generally known that longer training data builds a superior model. However, the observation data show that the water level may rise even when there is no precipitation due to the travel time of rainfall. To examine whether such an event adversely affects the ANFIS model, the accuracy of the model was evaluated for different lengths of training data.

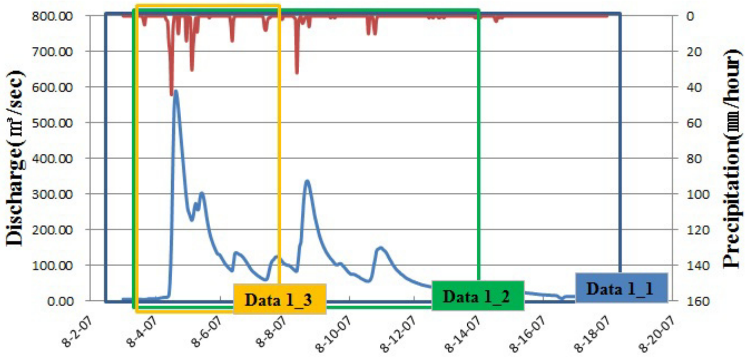

Figure 4: Training data.

Data 1_1 reflected all data for 15 days, while Data 1_2 and Data 1_3 reflected smaller amounts of data, including rainfall that showed peak flow.

The results did not show significant variance in RMSE. As for the discharge error, Data 1_3 showed a more accurate median value and a smaller range of deviation. Presumably, this is because if the water level changes when there is no precipitation, it causes noise that interrupt the rule making of the ANFIS model.

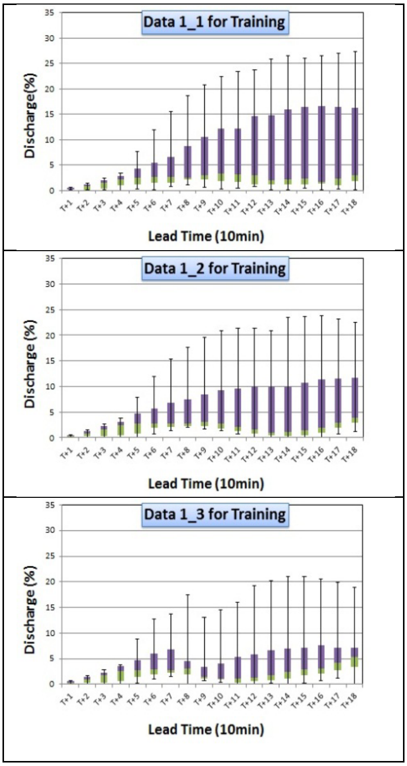

Figure 5: Discharge error comparison.

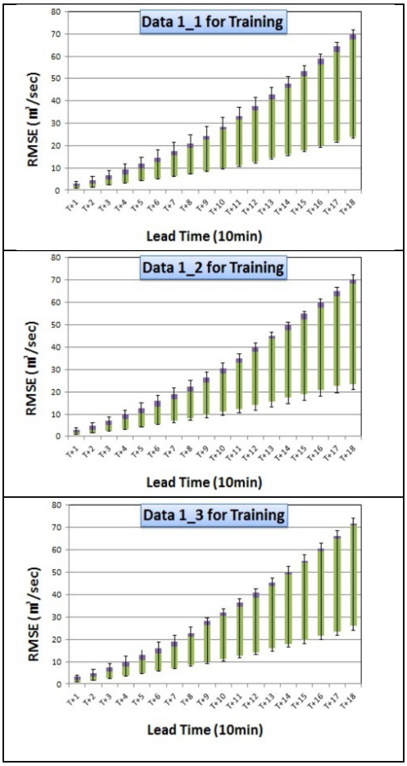

Figure 6: RMSE comparison.

3.4 Comparison of the data-driven model and physical model

Compared to the physical model, the data-driven model is relatively easy to construct and produces excellent estimates; on the other hand, it has the disadvantage of lacking mathematical and physical logic. To evaluate the suitability of the ANFIS model, its results were compared with those of HEC-HMS, a widely used physical model.

Unlike HEC-HMS, the ANFIS model makes rules with training data, constructs a model, and provides forecasts based on the observed rainfall and water level data. In this study, three rainfall events during 2011 were selected, and their observation data were compared with the forecasted outcomes of the ANFIS model and HEC-HMS. The first event took place between June 22 and 25, 2011, and was a lengthy downpour. The second event occurred on July 3, 2011 – its data were only partially taken. The last event was also concentrated in time, taking place between August 16 and 17, 2011. Figure 7 illustrates the observed discharge data and the HEC-HMS and ANFIS models' forecasts for the last event.

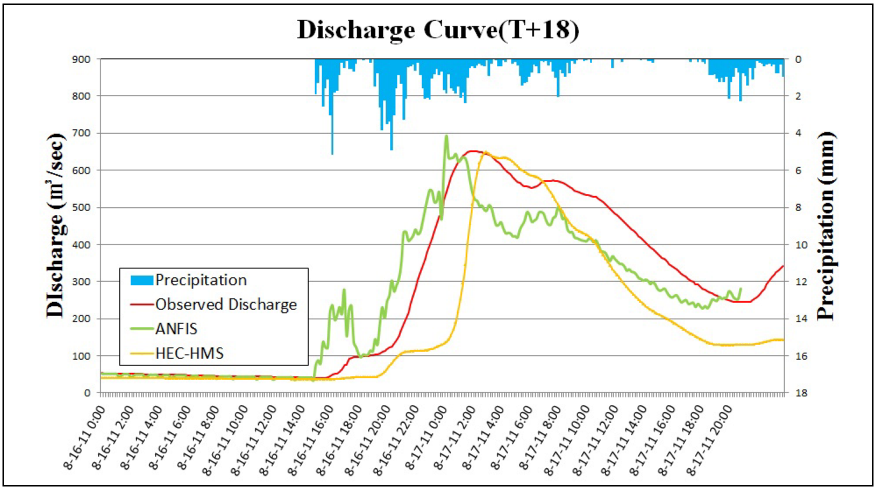

Figure 7: Discharge curve (T+18).

As the figure shows, the peak flow estimate of ANFIS at T+18 deviated further from the observation data than the discharge curve of HEC-HMS. However, overall, the ANFIS model provided more accurate estimates, as these were closer to the observed data. Table 5 shows the RMSE estimates and discharge error of the HEC-HMS and ANFIS models at T+9 and T+18.

Table 5: Comparison of ANFIS and HEC-HMS.

	June 22-25		July 3		Aug. 16–17	
	Discharge Error (%)	RMSE (m^3/sec)	Discharge Error (%)	RMSE (m^3/sec)	Discharge Error (%)	RMSE (m^3/sec)
HEC-HMS	5.25	28.25	7.25	37.39	0.40	90.15
ANFIS T+9	0.52	9.29	1.20	26.83	2.87	26.36
ANFIS T+18	4.38	23.28	6.25	70.82	6.35	71.30

As the table shows, the discharge estimate of the ANFIS model at T+9 showed excellent outcomes in most indicators, and at T+18, the estimate was similar to that of HEC-HMS.

4 Conclusion

In this study, the neuro-fuzzy technique was applied to forecast the water level at Wonbugyo point, downstream of the Cheongmicheon Basin, based on observed rainfall and water level data. Five rainfall events between 2007 and 2011 were selected to derive input data, and a model simulation was conducted for three hours (from T+1 to T+18) at 10 minute intervals. The changes in forecasting due to different clustering radii and training data lengths were examined and analyzed. The best model was compared with a widely used physical model, HEC-HMS, to identify the potentials and limitations of the data-driven model.

Changes in clustering radius did not lead to a significant difference; the result was best at a clustering radius of 0.5 in terms of discharge error. This is contrary to a general belief that a smaller clustering radius is more advantageous, as it produces more rules and thus provides a more precise forecast. The experiment suggests that the optimal clustering radius might depend on particular circumstances.

In terms of training data length, the outcome was best for the dataset that included the shortest period of non-precipitation. Generally, it is known that training data from a longer period help to build a superior forecast model. However, the experimental results suggest that water level may either rise or fall even when there is no precipitation, and such data might adversely affect the accuracy of the forecast.

In comparison with HEC-HMS, the ANFIS model produced equally excellent outcomes in terms of discharge error and RMSE; at T+9, the estimates were generally better than those of HEC-HMS. Given these results, it can be considered that the ANFIS model—as a supplementary tool to the physical model—could help to reduce uncertainties and address problems of the current flood forecast and warning system, as well as improving the accuracy of discharge forecasts.

Acknowledgements

This work was supported by a 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. and 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. and 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] Shin, G.W., Kim, J.H., Yang, J.L. and Hong, S.T., Development of water demand forecasting simulator and performance evaluation. *Journal of the Korean Society of Water and Wastewater*, 25(4), pp.581-589, 2011.

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

[5] Jang, J.S., ANFIS: adaptive-networkbased fuzzy inference systems, *IEEE Transactions on Systems, Man, and Cybernetics*, 23(3), pp.665-685.