Application of Acoustic Tomographic Data in Short-Term Forecasting of Streamflow Using Combinatorial GMDH Algorithm (CGA)

Yousef Olfatmiri[1], Ebrahim Jabbari[1], Masoud Bahreinimotlagh[2*], Hossein Alizadeh[1], Amirhosein Hasanabadi[1]

[1] Department of Civil Engineering, Iran University of Science and Technology, Tehran, Iran
[2] Department of Water Resources Studies and Research, Water Research Institute, Tehran, Iran

ABSTRACT

Short-term forecasting of streamflow is one of the most important goals in water resources management and flood control. However, one of the problems that researchers always face in this type of prediction is the Lack of an accurate and high-resolution database. The Fluvial Acoustic Tomography (FAT) is an innovative technology that acquires streamflow data. Therefore, by using the data collected from this technology with a suitable forecast model, accurate short-term streamflow forecasting can be achieved. In this research, the effect of FAT data on short-term streamflow forecasting by Combinatorial GMDH Algorithm (CGA) has been investigated and compared with one obtained from the Rating Curve method. The k-fold cross-validation criterion has been used to prevent over-fitting. The results showed that the FAT data increases the accuracy of short-term forecasting. As an example, the Nash-Sutcliffe coefficient (E_{NS}) for the 1, 6, 12, 24, 48, and 72 hours forecast horizons were 0.98, 0.96, 0.94, 0.88, 0.73, and 0.54, respectively. While these values for the Rating Curve ones were 0.97, 0.84, 0.61, 0.27, 0.12, and 0.11, respectively.

KEYWORDS

GMDH Algorithm, Short-Term Forecasting, Streamflow Forecasting, FAT.

* Corresponding Author: Email: m.bahreini@wri.ac.ir

1. Introduction

Predicting the parameters of the hydrological cycle is one of the most important factors for the successful management of water resources. One of these important parameters is streamflow. Streamflow Forecasting can help engineers in water supply, flood warning systems, and reservoir management [1]. Researchers have always sought to forecast streamflow. For this purpose, they have tried to estimate the streamflow using different hydrological models [2].

To forecast the streamflow of rivers; Hydrological models are generally divided into two categories: physical models and data-driven models. The complexity of the relationship between environmental variables and flow rate limits the use of physical models [3]. Also, as an advantage of data-driven models, these models require fewer input variables and can be developed using existing data records [4].

The main purpose of this study is to forecast the short-term streamflow of the Gono River in Hiroshima province, Japan, using data collected by the Fluvial Acoustic Tomography (FAT) method and compared with the Rating-Curves method. In this regard, an accurate database with a high time resolution (30 seconds) has been prepared for six months; Then, by one of the classical models of group method of data handling (GMDH), under the title of Combinatorial GMDH algorithm (CGA), short-term streamflow is forecasted by FAT method and compared with Rating-Curves method (1-hour time resolution). Finally, the effect of the application of FAT technology on increasing the accuracy of short-term streamflow forecasting has been investigated. The innovation of this research is the study of the effect of FAT data on the accuracy of short-term forecasts as 1, 6, 12, 24, 48, and 72 hours-ahead.

2. Methodology

The group method of data handling (GMDH) is a self-organizing learning method presented by Ivakhnenko (1968) to overcome the dead-end problem of equation complexity and linear dependency present in standard regression equations. One of the classic methods of data group classification is the Combinatorial GMDH algorithm (CGA).

The combinatorial GMDH algorithm (CGA) is the main single-layer algorithm for GMDH, to prevent over-fitting. First Shuffle the input Data then from this shuffled Data, 70% of the data will be selected as training data and the remaining 30% will be selected as testing data. According to the k-fold cross-validation criteria, the test data set is randomly divided into (K) sub-samples (In this study K≤5) of the same size. In each sub-sample, using the validation process, the number (K-1) of the sub-sample is considered as training data and one as the validation data set.

The input data matrix contains (N) observational data related to the (M) parameter. Training data are used to generate the optimal model. The optimal model is constructed by minimizing the following fitness function:

$$AR_{(B)} = \frac{1}{N_B}\sum_{i=l}^{N_B}(y_i - y_{i(B)})^2 \rightarrow \min \qquad (1)$$

(AR) fitness function, (N_B) Training data, (y_i) output in each layer, and ($y_{i(B)}$) is Testing data.

In each layer, a new input variable is applied and controlled using the fitness function the process of increasing accuracy is controlled. If no improvement in the performance of the algorithm is observed in the model with the addition of new variables, the model is stopped. The first layer is defined as the following equation.

$$y = a_0 + a_1 x_i \qquad i = 1, 2, ...M \qquad (2)$$

Non-linear individuals can also be added to polynomials in subsequent layers. The coefficients in these equations for each layer are obtained by the least-squares method. The general state of the equations in layer (m) is defined as the following form:

$$y = a_0 + a_1 x_i + a_2 x_j + ... + a_m x_l$$
$$i = j = l = 1, 2, ..., M \qquad (3)$$

Using the above algorithm for two different databases prepared by Rating-Curves and FAT, short-term forecasts are made for time intervals of 1, 6, 12, 24, 48, and 72 hours-ahead. To perform the above calculations, Wolfram Mathematica 12.1 software has been used.

3. Results and Discussion

In this study, the type of input data is the discharge collected in the river; therefore, the input data matrix is created by time delays on the discharge time series. The target data matrix also includes the discharge collected in the river. Then, with the combinatorial GMDH algorithm (CGA), short-term forecasts are made for the lead time of 1, 6, 12, 24, 48and 72 hours-ahead. Finally, to investigate the effect of data collection on prediction accuracy, all processes related to the forecasting algorithm; in addition to the data extracted using FAT technology, were repeated on the data extracted by Rating-Curves. To evaluate the obtained information, different statistical

criteria are used to evaluate the performance of each relationship; because statistical criteria alone cannot be a good criterion of the accuracy of different methods. In this research, R, RMSE, MAE, MARE, and Nash-Sutcliffe (ENS) criteria have been used.

Table 1. Results of short-term forecasts performed on FAT and Rating-Curve data.

E_{NS}	RMSE (M^3/S)	Lead Time (Hour)	Input Data Type
0.98	**2.105**	**1**	**FAT**
0.096	5.952	6	
0.94	9.207	12	
0.88	14.637	24	
0.73	22.375	48	
0.54	29.361	72	
0.97	**4.12**	**1**	**Rating-Curve**
0.84	19.677	6	
0.61	31.179	12	
0.27	41.661	24	
0.12	47.60	48	
0.11	45.96	72	

According to Table 1, the modeling performed by the CGA method showed that the forecast accuracy for FAT data is higher than the data prepared by the Rating-Curves data. For example, the RMSE for a 1h-ahead forecast of FAT data is 2.105 m^3/s; while this criterion for 1-hour forecast on Rating-Curves data events has almost doubled to 4.12m^3/s. As the lead time increases, although the forecast error increases for both databases; But the prediction error on FAT data is still lower. For a 24-hour forecast, the RMSE difference between the two databases almost triples from 14.637 m^3/s to 41.661m^3/s. This difference can be seen by other evaluation criteria. According to the Nash-Sutcliffe criterion, with increasing the lead time up to 12 h-ahead for FAT data, the forecast accuracy decreases slightly. For 24 h-ahead, the accuracy of the forecast decreases significantly. For Rating-Curve data with an increased lead time, the intensity of the accuracy decrease increases. So that the Nash-Sutcliffe coefficient for 12 and 24 hours-ahead, decreases from 0.94 and 0.88 for FAT data to 0.61 and 0.27 for Rating-Curve data, respectively.

According to the Nash-Sutcliffe criterion, with increasing the lead time up to 12 h-ahead for FAT data, the forecast accuracy decreases slightly. For 24 h-ahead, the accuracy of the forecast decreases significantly. For Rating-Curve data with an increased lead time, the intensity of the accuracy decrease increases. So that the Nash-Sutcliffe coefficient for 12 and 24 hours-ahead, decreases from 0.94 and 0.88 for FAT data to 0.61 and 0.27 for Rating-Curve data, respectively. The Nash-Sutcliffe (E_{NS}) figures for the 48 and 72 hours-ahead forecasts indicate that the forecasts for Rating-Curve data during this lead time are not accurate. As the Nash-Sutcliffe (E_{NS}) is (0.12 and 0.11), respectively, while this figures for the forecasts on the FAT are (0.73 and 0.54) and have relatively good accuracy.

4. Conclusion

The results of short-term forecasting performed on two different databases of FAT, and Rating-Curves showed that in general, the combinatorial GMDH algorithm (CGA) algorithm. Has good accuracy in short-term forecasting. Forecast accuracy regardless of the database used to predict reduced by increasing the lead time. The highest accuracy is related to 1-hour ahead and the lowest accuracy is related to 72-hour ahead forecasts. The accuracy of short-term forecasting performed on FAT data is higher than the Rating-Curves data. The reason for this high accuracy of prediction is the high time resolution and high accuracy of data collected. Therefore, it can be concluded that the application of FAT makes the algorithm used to predict better and more accurate results. The use of FAT can also help the forecasting model to increase the prediction accuracy for flood peaks.

5. References

[1] M. Abbasi, A. Farokhnia, M. Bahreinimotlagh, R. Roozbahani, A hybrid of Random Forest and Deep Auto-Encoder with support vector regression methods for accuracy improvement and uncertainty reduction of long-term streamflow prediction, Journal of Hydrology, 597 (2021) 125717.

[2] M. Ehteram, F. Binti Othman, Z. Mundher Yaseen, H. Abdulmohsin Afan, M. Falah Allawi, A. Najah Ahmed, S. Shahid, V. P Singh, A. El-Shafie, Improving the Muskingum flood routing method using a hybrid of particle swarm optimization and bat algorithm, Water, 10(6) (2018) 807.

[3] M.S. Khan, P. Coulibaly, Bayesian neural network for rainfall-runoff modeling, Water Resources Research, 42(7) (2006).

[4] A. Mosavi, P. Ozturk, K.-w. Chau, Flood prediction using machine learning models: Literature review, Water, 10(11) (2018) 1536.