Iranian Journal of Science & Technology, Transaction B, Engineering, Vol. 34, No. B1, pp 81-92
Printed in The Islamic Republic of Iran, 2010

MULTI MODEL DATA FUSION FOR HYDROLOGICAL FORECASTING USING *K*-NEAREST NEIGHBOUR METHOD*

M. AZMI, S. ARAGHINEJAD** AND M. KHOLGHI

Dept. of Irrigation and Reclamation Engineering, University of Tehran, Karaj, I. R. of Iran
Email: araghinejad@ut.ac.ir

Abstract– Hydrological forecasting is one the most important issues in water resources systems which helps in dealing with the real time operation, flood and drought warning, and irrigation scheduling. Recent studies have suggested that the use of data fusion approach instead of using a single forecast approach may improve the hydrological forecast skill. This paper presents a comparative assessment of five different methods of data fusions including simple and weighted averaging; relying on the user's experience; artificial neural networks; and error analysis, by applying them in two real case studies. Multiple linear regression, non-parametric *K*-nearest neighbour regression, conventional multilayer perceptron, and an artificial neural network improved for extreme value forecasting are used as individual forecasting methods at each case study. Conventional data fusing methods as well as a new proposed statistical method based on the non-parametric *K*- nearest neighbor model are used for hydrological forecasting. Results of data fusion approach in two contrasting case studies are thoroughly analyzed and discussed. The results demonstrate that the use of data fusion could significantly improve forecasts in comparison with the use of singe models. As a result of this study, it is concluded that time-varying combining methods which benefit from the use of real-time predictors in their fusion procedure could be more promising than others. Also, data fusion by *K*-NN method outperforms conventional methods by improving forecasts through decreasing the bandwidth of ensemble forecast and error of point forecast in both case studies.

Keywords– Data fusion, nearest neighbour method, streamflow forecasting, Zayande-rud River; Red River

1. INTRODUCTION

Data fusion is an emerging area of research that covers a broad spectrum of application areas ranging from ocean surveillance, strategic warning, and medical diagnosis [1]. The principal objective of data fusion, which is the process of combining or amalgamating information from multiple sensors and/or data sources, is to provide a solution that is either more accurate according to some measure of evaluation, or allows one to make additional inferences above and beyond those that could be achieved through the use of single source data alone [2]. Data fusion provides new modeling opportunities in the water resources and hydrology fields. Operational hydrological forecasting, in particular, might benefit from the ability to combine information derived from multiple sources, such as the individual outputs from different forecasting models. Data fusion researches are divided into two board groups. The first takes the view that data fusion is the amalgamation of raw information to produce an output, while the second advocate a more generalized view of data fusion in which both raw and processed information can be fused into useful outputs including higher level decision.

Recently, researches such as See and Abrahart [3], Abrahart and See [4], and Shu and Burn [5] have used model-fusion approaches in hydrological engineering. See and Abrahart [3] used data fusion approach for continuous river level forecasting where data fusion was the amalgamation of information

*Received by the editors December 31, 2005; Accepted September 15, 2006.
**Corresponding author

from multiple sensors and different data sources. Abrahart and See [4] evaluated six data fusion strategies and found that data fusion by an Artificial Neural Network (ANN) model provided the best solution. Shu and Burn [5] applied artificial neural network ensembles in pooled flood frequency analysis for estimating the index flood and the 10-year flood quintiles. The data fusion method was used to combine individual ANN models in order to enhance the final estimation.

This paper provides a comparative assessment of five general methods of data fusion in hydrological forecasting. Furthermore, a new statistical method based on the non-parametric *K*-nearest neighbour (K-NN) simulation is proposed for the purpose of data fusion. Application of these methods is tested in two contrasting case studies of long-term and short-term hydrological forecasting.

Data fusion is an emerging area of research that covers a broad spectrum of application areas ranging from ocean surveillance, strategic warning, and medical diagnosis [1]. The principal objective of data fusion, which is the process of combining or amalgamating information from multiple sensors and/or data sources, is to provide a solution that is either more accurate according to some measure of evaluation, or allows one to make additional inferences above and beyond those that could be achieved through the use of single source data alone [2]. Data fusion provides new modeling opportunities in the water resources and hydrology fields. Operational hydrological forecasting, in particular, might benefit from the ability to combine information derived from multiple sources, such as the individual outputs from different forecasting models. Data fusion researches are divided into two board groups. The first takes the view that data fusion is the amalgamation of raw information to produce an output, while the second advocate a more generalized view of data fusion in which both raw and processed information can be fused into useful outputs including higher level decision.

Recently, researches such as See and Abrahart [3], Abrahart and See [4], and Shu and Burn [5] have used model-fusion approaches in hydrological engineering. See and Abrahart [3] used data fusion approach for continuous river level forecasting where data fusion was the amalgamation of information from multiple sensors and different data sources. Abrahart and See [4] evaluated six data fusion strategies and found that data fusion by an Artificial Neural Network (ANN) model provided the best solution. Shu and Burn [5] applied artificial neural network ensembles in pooled flood frequency analysis for estimating the index flood and the 10-year flood quintiles. The data fusion method was used to combine individual ANN models in order to enhance the final estimation.

This paper provides a comparative assessment of five general methods of data fusion in hydrological forecasting. Furthermore, a new statistical method based on the non-parametric *K*-nearest neighbour (K-NN) simulation is proposed for the purpose of data fusion. Application of these methods is tested in two contrasting case studies of long-term and short-term hydrological forecasting.

2. DATA FUSION METHODS

The general equation of a hydrological event forecasting model is

$$\hat{y}_i = f(X_i) + \varepsilon_i \qquad i = 1, \cdots, n \tag{1}$$

where X = vector of predictors, $\hat{y}$ = forecast variable, ε = model error and n = number of observation data. In the case of using multiple models to forecast y , and considering similar predictors, Eq. (1) is changed to the following matrix form

$$\left[\hat{Y}_i\right] = \begin{bmatrix} \hat{y}_{i1} \\ \hat{y}_{i2} \\ \vdots \\ \hat{y}_{im} \end{bmatrix} = \begin{bmatrix} f_1(X_i) \\ f_2(X_i) \\ \vdots \\ f_m(X_i) \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{im} \end{bmatrix} \qquad i = 1, \cdots, n \tag{2}$$

where m = number of forecast models used to estimate y., $\left[\hat{Y}\right]$ = matrix of estimations of y provided by different individual models. Using the data fusion approach, $\left[\hat{Y}\right]$ is sum up to a unique estimation of $\hat{y}$. The following paragraphs describe some of the general methods of data fusion, presented and implemented by different researchers.

a) Simple and weighted averaging methods (MD1 and MD2)

Linear combination of the outputs of ensemble members is one of the most popular approaches for combining different outputs. A single output can be created from the combination of the outputs of a set of models via simple averaging, or a weighted average method that considers the relative performance of each model. Combining using simple average is defined as:

$$\hat{y}_i = 1/m(\sum_{j=1}^{m} \hat{y}_{ij}) \qquad i = 1,\cdots,n \tag{3}$$

Despite its simplicity, the simple averaging method suffers from the problem of considering equal weights for individual models. Obviously, the difference in the reliability of individual models is not considered in simple averaging as all of them are assigned by similar weights in this data fusion method. To overcome this shortcoming, the method of weighted averaging also known as *stacking* might be used. This method is presented by the following equation:

$$\hat{y}_i = \sum_{j=1}^{m} c_j \hat{y}_{ij} \qquad i = 1,\cdots,n \tag{4}$$

where, c = the weight of each individual model. Under "stacking" an additional attempt is used to learn how to combine the models by tuning the c weights over the calibration data. To derive c weights, Shu and Burn [5] suggested minimizing the following function:

$$w = \sum_{i=1}^{n} \left[\frac{y_i - \sum_{j=1}^{m} c_j \hat{y}_{ij}}{y_i} \right]^2 \qquad c_k > 0 \tag{5}$$

The stacking method uses constant weights over the time period of the calibration data set which reduces the flexibility of the method in facing different hydroclimatological situations that a system might experience during its operation.

b) Relying on the user's experience (MD3)

In this method, at each step of decision making, instead of combining outputs of different models the result of just one model is selected, relying on the experience of the last step. Obviously this method is limited to cases of time series forecasting where predictors are well-correlated. Using this method might not be suitable for event forecasting.

c) Using empirical models such as artificial neural networks (MD4)

Empirical models, particularly artificial neural networks (ANNs) are known as powerful tools for function mapping. See and Abrahart [3] have suggested the use of ANNs as a data fusion method. The general form of this method is

$$\hat{y}_i = g(\left[\hat{Y}_i\right]') \qquad \left[\hat{Y}\right] = \begin{bmatrix} \hat{y}_{i1} \\ \hat{y}_{i2} \\ \vdots \\ \hat{y}_{im} \end{bmatrix}; \; i = 1,\cdots,n \tag{6}$$

Where g is a non-linear function which maps outputs of different individual forecast models to a single output of $\hat{y}_i$ using an ANN model. Like most empirical methods, this method suffers from the lack of statistical sense in the mechanism of data processing.

d) Method of error analysis (MD5)

In this method, the historical forecasting error of individual models is analyzed to enhance the current forecast error. A common form of this method is presented as the following equation:

$$\hat{y}_i = y_{i-1} + g(E_{i-1}) \qquad [E] = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{im} \end{bmatrix};\ i = 1,\cdots,n \tag{7}$$

where, e_{im} =the error of mth model in ith forecast, E = set of errors of different models and g = a mapping function such as an ANN model. Performance of this method is limited to the ability of function mapping by g model.

3. DATA FUSION USING *K*-NN METHOD

The recognition of the nonlinearity of the underlying dynamics of hydrological processes has spurred the growth of nonparametric methods in testing streamflow characteristics [6] as well as hydrological forecasting. Nonparametric estimation of probability densities and regression functions are pursued through weighted local averages of the dependent variable. This is the foundation for nearest neighbor methods. *K*-Nearest Neighbour (*K*- NN) methods use the similarity (neighbourhood) between observations of predictors and similar sets of historical observations (successors) to obtain the best estimate for a dependent variable.

For each instant t, let $\lfloor X_j \rfloor$ (j=1,2,3,…, number of observed data) be a P-dimensional feature vector of past records. A feature vector is a vector that summarizes the past history in a smaller-dimension vector of observations that contains most of the information relevant to the forecast. To estimate the dependent variable in time t, y_t, the K-NN method imposes a metric on feature vectors to find the set of K past nearest neighbors of $\lfloor X_j \rfloor$.

The K vectors of past observations having the minimum norm among all candidates are obtained. The distance between the current and historical condition is calculated by the Euclidian [7] or Mahalanobis distance [8], between current and historical predictors. The most widely used metric to identify neighbors is the Euclidean norm, which, for a P-dimensional feature vector, is calculated as

$$Dis_{tj} = \sqrt{w_1(x_{1t} - x_{1j})^2 + w_2(x_{2t} - x_{2j})^2 \ldots + w_p(x_{pt} - x_{pj})^2} \tag{8}$$

where, w = the weight of predictors in calculating the distance between the current value of predictors and the neighbors. The forecast is then obtained as a weighted average of the nearest neighbors, such that greater weight is assigned to the nearer neighbors. A kernel function proposed by Lall &Sharma [9] defines the weights, leading to a K-NN regression estimate of:

$$y_t = \frac{\sum_{i=1}^{K} (1/i) y_i}{\sum_{i=1}^{K} (1/i)} \tag{9}$$

where i is the order of the nearest neighbors in which the nearest have the lowest order (i=1 to K) , and y_i is the magnitude of nearest neighbor i. The weights and the number of the neighbors that produce the lowest mean square error of the forecasting are found by computing the error for all combinations of weights and K values (from 1 to n-1) through Generalized Cross Validation (GCV) of the calibration data set, which is defined as [9]:

$$GCV = \frac{\sum_{i=1}^{n} e_i^2 / n}{\left(1 - 1/\sum_{j=1}^{K} 1/j\right)^2} \tag{10}$$

where e is the error between observed and predicted data, n is the number of data, and j is the order of the nearest neighbors based on their distance from the current condition in which the nearest have the lowest order. The GCV score function is used to choose both K and the weights of the predictors. The weights of each predictor control the distance between the successors and the current condition and subsequently the error of forecasting. Parameter K and the weights that provide the minimum GCV are selected.

Using the concept of *K-NN*, an algorithmic procedure is proposed for data fusion as follows (MD6):

1. Use m individual forecast models to produce $n \times m$ forecasts, where n is the number of observed data used for calibration.
2. Evaluate individual forecasting models in all n forecast experiences. Compute the matrix of $[A] = [a_{ij}]_{n \times m}$, where $a_{ij} = 1$ if mth model results in the best forecast during ith experience; otherwise $a_{ij} = 0$.
3. At the present time, t, compute m forecasts of y_t, using m individual forecast models and develop $[\hat{Y}] = \begin{bmatrix} \hat{y}_{i1} \\ \hat{y}_{i2} \\ \vdots \\ \hat{y}_{im} \end{bmatrix}$
4. Compute $[F] = [A] \times [\hat{Y}] = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}$
5. Use equation 8 to find the distance between present time predictors, $[X_t]$, and the historical predictors, $[X_j]$. Determine the nearest neighbors from n sets of observed data, such that the smaller distance is assigned to the nearest neighbour.
6. Forecast the dependent variable by the following equation

$$\hat{y}_r = \frac{1}{\sum_{i=1}^{K} 1/i} \sum_{i=1}^{K} (1/i) f_i \tag{11}$$

where i= the order of the nearest neighbors in which the nearest have the lowest order (i=1 to K), K= number of nearest neighbours obtained through the generalized cross validation (GCV), and f_i= the magnitude of nearest neighbor i.

4. DEVELOPMENT OF INDIVIDUAL FORECASTING MODELS

a) Linear multiple regression and K-NN regression models (LMR, and K-NN)

Linear Multiple Regression (MLR) is a multivariate method of forecasting which estimates unknown coefficients of predictors by performing a least squares fit. In this study, the linear multiple regression is applied as a forecasting method in the following form:

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_p x_p + c \tag{12}$$

where, $x_1 \, to \, x_p = p$ predictors of dependent variable y , $a_1 \, to \, a_p$ = coefficients of predictors, and c= constant of the model. Furthermore, a non-parametric regression model, *K-NN*, is used as described in the previous section by equations 8 to 10.

b) Conventional and enhanced multilayer perceptron model (MLP and MLP-E)

The application of the artificial neural networks in hydrology has grown rapidly in recent years. The ANN approach is an effective and efficient way to model forecasting problems. Three- layer feed-forward networks are known as the networks that are capable of approximating any continuous input-output mapping. The multi-layer feed-forward network uses supervised training procedure that consists of providing input-output examples to the network and minimizing the error function *E*, which is expressed as follows;

$$E = 1/2 \sum_{p=1}^{n} (y_p - \hat{y}_p)^2 \tag{13}$$

where *n* is the number of input/output data sets, and $\hat{y}_p$ and y_p are the observed and forecasted output of the *p*th set respectively. In the back propagation (BP) training method, *E* is minimized using the steepest descent method.

Coulibaly et al [10] applied a different form of performance function known as peak flow/low flow criterion (PLC) to improve the ability of the model in extreme value forecasting. The PLC for an input set k is specified as

$$PLC = P_k \times L_k \tag{14}$$

where P_k is the peak flow criterion given by

$$P_k = \frac{\left[\sum_{i=1}^{n_p} (Q_{pi} - \hat{Q}_{pi})^2 Q^2_{pi} \right]^{0.25}}{(\sum_{i=1}^{np} Q^2_{pi})^{0.5}} \tag{15}$$

where n_p is the number of peak flows greater than one-third of the mean peak flow observed, Q_p and $\hat{Q}_p$ are respectively the observed and the computed flows, and L_k is the low flow criterion, which is given by

$$L_k = \frac{\left[\sum_{i=1}^{n_l} (Q_{li} - \hat{Q}_{li})^2 Q_{li}^{\,2} \right]^{0.25}}{(\sum_{i=1}^{n_l} Q_{li}^{\,2})^{0.5}} \tag{16}$$

where n_i is the number of low flows lower than one-third of the mean low flow observed, and Q_l and $\hat{Q}_{li}$ are the observed and the computed flows, respectively. The p_k provides a more accurate measure of the model performance than the Eq. (14) for the peak flow periods, whereas the L_k is a better performance indictor for the low flow period. Three-layer feedforward networks trained by Eqs. (13) and (14) are used in this study.

The schematic of the procedure for data fusion, individual models used in the procedure, and the methods of data fusion are shown in Fig. 1.

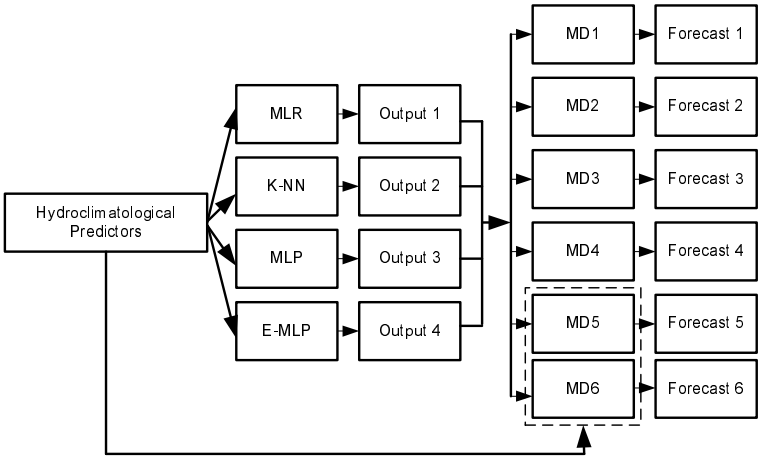


Fig. 1. The schematic of the procedure used in this study

5. CASE STUDY I : SEASONAL STREAMFLOW FORECASTING

a) Study area and data

Zayandeh-rud River is the main surface resource for irrigation demands in the central part of Iran, especially the Isfahan metropolitan area [11]. The Zayandeh-rud reservoir controls the streamflow upstream of Isfahan. It is the largest surface reservoir on the river with a volume of 1470 million cubic meters. The location of the Zayandeh-rud reservoir is shown in Fig. 2. Total annual average inflow to the Zayandeh-rud reservoir is about 1600 million cubic meters, of which, an average annual flow of 600 million cubic meters is transferred from the adjacent Karoon River basin. Seasonal inflow data to the Zayandeh-rud reservoir for a 32-year period from 1969 to 2001 are used in this study. Streamflow from April to September (spring and summer streamflow), which is result of the winter snow pack is used as a predicted variable in this study. Recently, the effect of large scale climate signals such as ENSO and NAO have been considered as the forcing factors on the climate of Iran [12]. The North Atlantic Oscillation climate signal, NAO, which seems to be effective for predicting climate variations of the central and southern parts of Iran, is considered as a predictor of Zayandeh-rud River streamflow [13]. The North Atlantic Oscillation (NAO) involves a negative correlation in winter months between sea-level pressures in the subtropical Atlantic high and the Icelandic low. It index is the difference between normalized sea level pressure over the Azores and Iceland. The usual index is given by the December to March average of this measure [14]. There is a significant relationship between averaged December to March NAO and spring (April to June) streamflow. In this study NAO as well as snow budget and winter streamflow, are used as predictors of spring and summer streamflow.

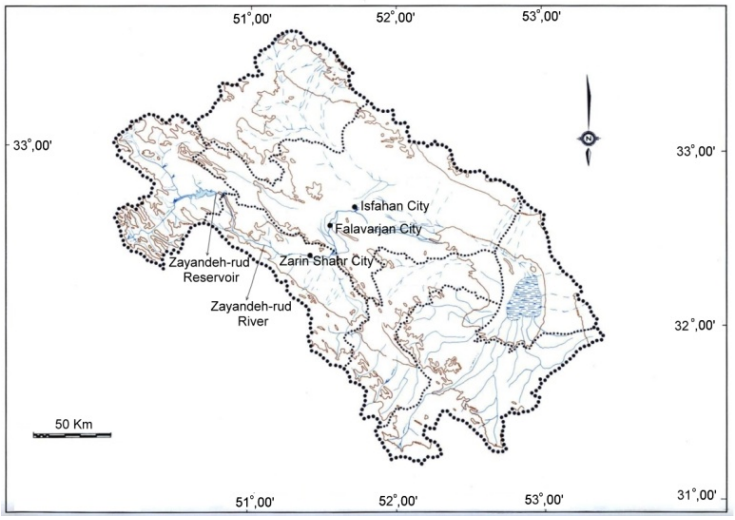


Fig. 2. The location map of Zayandeh-rud river and Zayandeh-rud reservoir

b) The results of forecasting in case study I

The results of applying individual models as well as the data fusion methods are shown in Table 1 and Table 2, respectively, in cross validation analysis for testing the models on data not used in the calibration. As shown in Table 1, MLP has resulted in the best point forecast as it has provided forecasts with minimum root mean square error (RMSE), percent of volume error (%VE), and maximum linear correlation between the observed and forecast data (CORR). This means that, although MLP-E is a better model in extreme value forecasting, MLP outperforms the other models according to the overall forecast of normal and extreme values. MLP has resulted in RMSE, %VE, and CORR, equal to 40, 15.7, and 0.8 respectively.

Data fusion methods have improved the results of individual models as demonstrated in Table 2. The best forecast is obtained by MD6 as it has resulted in RMSE, %VE, and CORR equal to 37, 14, and 0.95. It means that the data fusion approach has improved the accuracy of point forecast by 7.5, 11, and 19 percent, according to RMSE, %VE, and CORR statistics, respectively.

Maximum and minimum seasonal streamflow volume in the first study are 1576 *MCM* and 536 *MCM*, which have been observed in years 1993, and 2001, respectively. MLP has forecasted dependent variables in those years as 1853 and 537 *MCM*. The seasonal streamflow volumes at the same years are forecasted as 1827 *MCM* and 765*MCM* using data fusion approach. This shows that in the first study, individual models were more successful than data fusion methods in the case of forecasting extreme values.

Methods of MD6 and MD5 have resulted in the best forecasts among others. MD6 and MD5 benefit from using current-time predictors in addition to the output of single models as well as timely dynamic weighting of the single models in their fusion procedure. These benefits are most likely the reasons for the supremacy of these two methods.

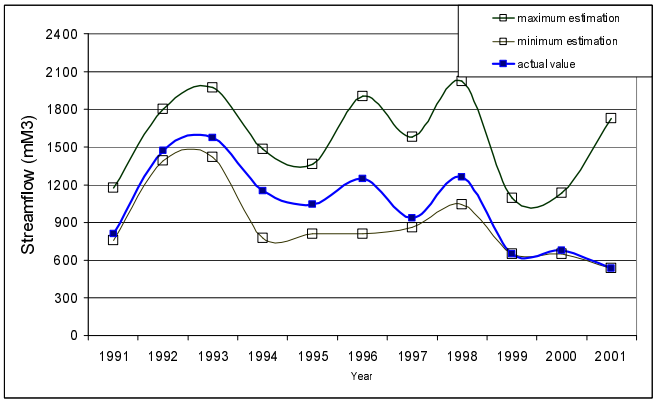
Table 1. The cross validation results of forecasting Zayandeh-rud streamflow using individual models

Models	Cross validation		
	RMSE	%VE	CORR
MLR	53	37.0	0.34
K-NN	43	15.0	0.74
MLP	40	15.7	0.80
MLP-E	49	18.4	0.70

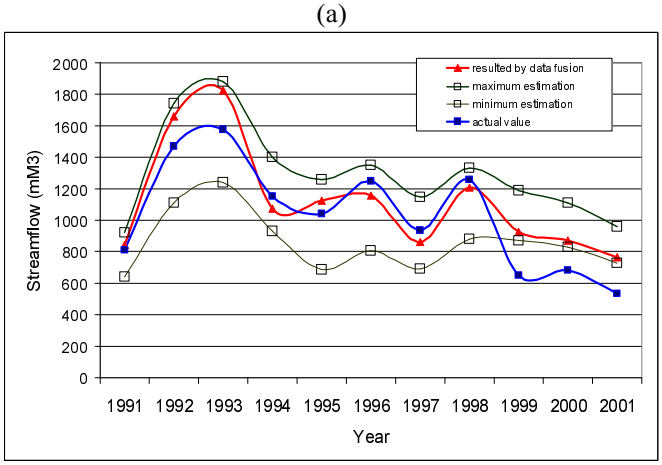
Table 2. The result of data fusion in forecasting Zayandeh-rud streamflow using different methods

Methods	Cross validation		
	RMSE	%VE	CORR
MD1	41	21	0.80
MD2	37	16	0.89
MD3	39	11	0.84
MD4	38	15	0.90
MD5	29	16	0.92
MD6	37	14	0.95

Furthermore, the maximum and the minimum estimation of the predicted variable developed by MD6 as well as what is developed by the other methods are shown in Fig. 3. The difference between the maximum and the minimum estimations, which is the bandwidth of the estimated variables, gives us scenes of reliability in operational forecasts. As shown in Fig. 3, the bandwidth of forecast developed by MD6 is narrower than what is provided by the other methods. The average statistic of $\left(\frac{\hat{y}_{\max} - \hat{y}_{\min}}{y}\right)$ (where $\hat{y}_{\max}$ and $\hat{y}_{\min}$ are the maximum and the minimum estimations of dependent variable, respectively; and y is the observed value of the dependent variable) is calculated as 39.5 percent using the MD6 method. Also, using the MD6 method, 3 out of 11 observed data are out of the range of the bandwidth provided in Fig. 3b. The same statistic for the MD1 to MD5 methods is 75.7 percent. Using MD1 and MD5 methods, 2 out of the 11 observed data are out of the range of forecasts as shown in Fig. 3a. The results indicate that MD6 is a better model in providing ensemble forecasts with narrower bands, resulting in more reliable estimation.



(a)



(b)

Fig. 3. The bandwidth of the forecasts developed by MD1 to MD5 (a), and MD6 (b) in case study I

Where methods of MD1 to MD5 could not make a preference between individual models in real time forecasting, MD6 is able to apply a selective choice among the individual models by assessing the experience of previous forecasts. It should be noticed that in contrast with the MD3 method, MD6 considers not only the forecast experienced in the previous time step, but also the forecasts experienced in the history of application of the individual models. This is the reason for the supremacy of MD6 in providing ensemble forecasts with a minimum bias and narrower-band.

6. CASE STUDY II: FLOOD PEAK DISCHARGE FORECASTING

Flooding of the Red River in Manitoba, Canada, typically results from snowmelt, often in combination with spring precipitation events. The nature of the red river, with its headwater in the USA, results in considerable warning of pending flood events within the Canadian portion of the watershed [15]. The low slope of the red river channel, and the consequent low water velocities, facilitates advanced warning of flood conditions. Major flood events in the Red River valley occurred in 1950, 1979 and 1996, in addition to the 1997 flood. In 1997, the red river experienced a major flood, which has increased the need for flood forecasting in the region.

Major casual parameters of the red river flood, based on previous flood studies, are [16]: 1) index of soil moisture at freeze-up the previous autumn, based on weighted basin precipitation from May to October; 2) Average degree-days per day at Grand Forks during the active melt period; 3) Total basin precipitation from 1 November of the previous year to the start of active melt during the flood year; 4) Total basin precipitation from the start of active spring melts to the date of the spring crest at the Emerson; and 5) the Index of the south-north time phasing of the runoff based on the percentage of tributary peaks experienced on the date of the mainstream peak at specific points from Halstad to the City of Winnipeg (percent of worst possible). The 40 years of annual peak discharge data from 1940 to 1979 are used for model calibration and 20 years of data from 1980 to 1999 are used for model validation.

a) The results of forecasting in case study II

Results of flood forecasting in the Red River using the individual models are shown in Table 3. The results are shown in both calibration and validation. As it is demonstrated in the table, conventional MLP and enhanced MLP (MLP-E) models have resulted in the best forecasts as they provided forecasts with minimum RMSE and %VE and maximum CORR. Conventional MLP is better than MLP-E in calibration, whereas MLP-E is a better model in validation. Minimum RMSE and % VE, and maximum CORR in the validation data set are obtained by MLP-E, which are 7, 18, and 0.9, respectively. This might be because of extreme values occurring in the validation period of the flood data.

Table 3. The results of forecasting the Red river floods using individual models

Models	Calibration			Validation		
	RMSE	%VE	CORR	RMSE	%VE	CORR
MLR	11	27.1	0.78	9	30.1	0.84
K-NN	26	43.1	0.48	23	28.5	0.68
MLP	6.2	14.0	0.85	9	43.0	0.86
MLP-E	6.4	17.5	0.84	7	18	0.90

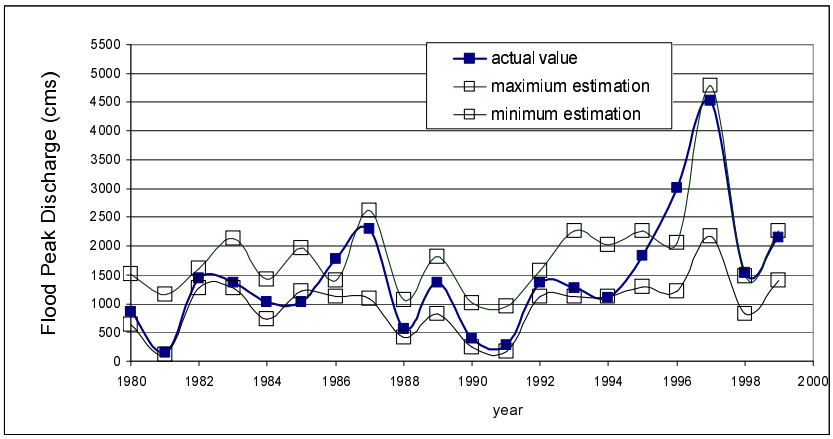
Results of data fusion using 6 different methods are shown in Table 4. As demonstrated in this table, the overall accuracy of the forecast has been significantly improved. The results of MD6 are better than others as it provides forecast with minimum RMSE and %VE, and Maximum CORR. MD6 resulted in RMSE, %VE, and CORR equal to 7, 16, and 0.92. This means that the data fusion approach has improved the accuracy of point forecast by 11 and 2 percent, according to %VE, and CORR statistics, respectively.

The bandwidth of forecasts using MD1 through MD5 as well as MD6 is shown in Fig. 4. The average statistic of $\left(\frac{\hat{y}_{\max} - \hat{y}_{\min}}{y}\right)$ (as discussed in a previous case study) is calculated as 72.8 percent using the

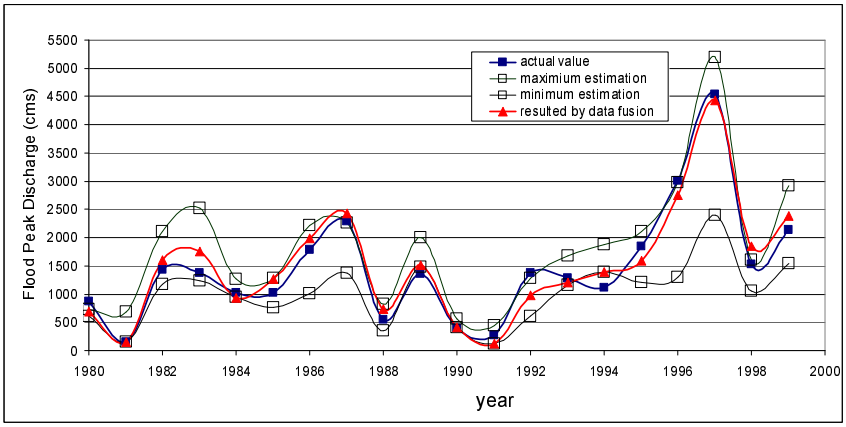
MD6 method. Using MD6 method, 1 out of 20 observed data are out of range of the bandwidth provided in Fig. 4b. The same statistic for the MD1 to MD5 methods is 108.5 percent. Using MD1 to MD5 methods, 4 out of 20 observed data are out of the range of forecasts as shown in Fig. 4b. The results indicate that MD6 is a better model in providing ensemble forecasts with narrower bands.

Table 4. The result of data fusion methods for flood forecasting in the Red river

Methods	Calibration			Validation		
	RMSE	%VE	CORR	RMSE	%VE	CORR
MD1	7.7	20.1	0.88	7.8	25.8	0.90
MD2	6.5	13.5	0.90	7	17	0.91
MD3	7.6	18.8	0.78	9	19.8	0.84
MD4	2.1	7.4	0.91	12	36.9	0.88
MD5	0.2	0.1	0.90	9	49.4	0.86
MD6	----	----	----	7	16	0.92



(a)



(b)

Fig. 4. The bandwidth of the forecasts developed by MD1 to MD5 (a), and MD6 (b) in case study II.

7. SUMMARY AND CONCLUSION

This paper investigated the application of multi-data fusion approach into the hydrological forecasting using six methods, including a new method based on the *K*-nearest neighbour algorithm. Two statistical and two empirical models were used to forecast seasonal inflow to the Zyandeh-rud reservoir in Iran and flood discharge in the Red River, Canada. Multiple linear regression, non-parametric K-nearest neighbour

regression, conventional multilayer perceptron, and an artificial neural network improved for extreme value forecasting was used for this purpose. The improved accuracy of forecasts resulted by data fusion methods confirmed that the combined forecasts outperform forecasts of individual models in both case studies. Another aim of the paper was to examine the ability of the proposed K-NN-based method in comparison with the other methods. The results demonstrated that the proposed method outperforms other methods in both cases. Also, the method based on the analysis of previous forecast errors was distinguished as another good method in data fusion. It is concluded that time-varying combining methods which benefit from the use of real-time predictors in their fusion procedure could be more promising than others. Furthermore, *K-NN* improved the reliability of real time forecasts by decreasing the bandwidth of the ensemble forecast in both cases. The proposed K-NN-based method has the potential to sum up individual models in a probabilistic manner, which could be considered in future studies.

REFERENCES

1. Hall, D. L. (1992). *Mathematical techniques in multisensor data fusion*. Artech House, Boston, MA.
2. Dasarathy, B. V. (1997). Sensor fusion potential exploitation-Innovative architechtures and illustrative applications. *Proceedings of the IEEE*, Vol. 85, pp. 24-38.
3. See, L. & Abrahart, R. J. (2001). Multi-model data fusion for hydrological forecasting, *Comput. Geosci*., 27, pp. 987– 994.
4. Abrahart R. & L. See. (2002). Multi-model data fusion for river flow forecasting: an valuation of six alternative methods based on two contrasting catchments, *Hydrology and Earth System Sciences*, Vol. 6, No. 4, pp. 655-670.
5. Shu, C. & Burn, D. H. (2004). Artificial neural network ensembles and their application in pooled flood frequency analysis. *Water Resour. Res.,* Vol. 40 W09301.
6. Aksoy, H. (2007). Hydrological variability of the European part of Turkey. *Iranian Journal of Science & Technology, Transaction B, Engineering*, Vol. 31, No. B2, pp. 225-236.
7. Karlsson, M. & Yakowitz, S. (1987). Nearest-neighbor methods for nonparametric rainfall-runoff forecasting, *Water Resour. Res.*, Vol. 23, No. 7, pp. 1300-1308.
8. Yates, D., Gangopadhyay, S., Rajagopalan, B. & Strzepek, K. (2003), A technique for generating regional climate scenarios using a nearest neighbor algorithm, *Water Resour. Res.,* Vol. 39, No. 7, 1199, doi:10.1029/2002WR001769.
9. Lall, U. & Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series*, Water Resour. Res.,* Vol. 32, No. 3, pp. 679-694.
10. Coulibaly, P., Bob´ee, B. & Antcil, F. (2001). Improving extreme hydrologic events forecasting using a new criterion for artificial neural network selection. *Hydrol. Process*. Vol. 15, pp. 1533–1536.
11. Modarres, R. & Eslamian, S. S. (2006). streamflow time series modeling of Zayandehrud river. *Iranian Journal of Science & Technology, Transaction B, Engineering*, Vol. 30, No. B4, pp. 567-570.
12. Nazemosadat, N. J., Samani, N., Barry D. A. & Molaii Niko, M. (2006). ENSO forcing on climate change in Iran: precipitation analysis. *Iranian Journal of Science & Technology, Transaction B, Engineering*, Vol. 30, No. B4, pp. 555-565.
13. Araghinejad. S., Burn, D. H. & Karamouz, M. (2006). Long-lead Streamflow Forecasting using Ocean-atmospheric and Hydrological Predictors. *Water Resour. Res.*, Vol. 42, doi:10.1029/2004WR003853.
14. Jones, P. D., Jonsson, T. & Wheeler, D. (1997). Extension to the North Atlantic Oscillation using early instrumental pressure observation from Gibraltar and south-west Iceland. *Int. J. Climatol*., Vol. 17, pp. 1433-1450.
15. Burn, D. H. (1999) Perceptions of flood risk: a case study of the Red River flood of 1997. *Water Resour. Res*., Vol. 35, No. 11, pp. 3451–3458.
16. Warkentin, A. A. (1999). *Red River at Winnipeg hydrometeorological parameter generated floods for design purposes. In: Red River Flooding Decreasing our Risks (Proc. Conf., Winnipeg, Manitoba, Canada)*, V-1–V-32, Canadian, Water Resources Association, Cambridge, Ontario, Canada.