RESERVOIR DAILY INFLOW SIMULATION USING DATA FUSION METHOD†

BEHNAM ABABAEI*, FARHAD MIRZAEI, TEYMOUR SOHRABI AND SHAHAB ARAGHINEJAD

Department of Irrigation and Reclamation Engineering, University of Tehran, Karaj, Iran

ABSTRACT

Information about the parameters defining water resource availability is a key factor in its management which improves the operation policies for water resource systems. One of the most important parameters in this area is river streamflow. In this research, two different strategies of data fusion were tested for daily inflow simulation of the Taleghan Reservoir. Four artificial neural network models as well as two Hammerstein–Wiener models were used as individual simulation models. The results showed that the data fusion method has the capacity to improve substantially the results of individual simulation models. The individual models were also tested in combination with a weather generator model which was used to generate 100 yr of daily temperature and precipitation data. The results demonstrated that although some models performed well in calibration and validation phases, in combination with a weather generator they could result in eccentric outcomes. This research also showed that the data fusion method can combine the results of single simulation models to improve the final estimate and decrease the bandwidth of errors. Copyright © 2013 John Wiley & Sons, Ltd.

KEY WORDS: data fusion; artificial neural networks; Hammerstein–Wiener models; Taleghan Reservoir; daily inflow

INTRODUCTION

Data fusion is an emerging research area that covers a broad range of applications ranging from surveillance, strategic warning to medical diagnosis (Abrahart and See, 2002). Data fusion (DF) is the process of combining information from multiple sensors or data sources to provide a solution that is either more accurate or allows one to make additional inferences beyond those that could be achieved through the use of a single source data alone (Azmi et al., 2010). Recently, researchers have used model fusion approaches in hydrological engineering. See and Abrahart (2001) used a data fusion approach for river level forecasting where data fusion was the combination of information from multiple...
sensors and different data sources. Azmi et al. (2010) compared five different methods of data fusion including simple and weighted averaging, relying on the user’s experience, artificial neural networks, and error analysis. These conventional methods were also compared with a new proposed statistical method based on the non-parametric K-nearest neighbour (KNN) model. Their results showed that use of the data fusion method could significantly improve hydrological forecasts in comparison with the use of a single model. Also, data fusion by the KNN method outperforms conventional methods by improving forecasting through decreasing the bandwidth of ensemble forecast error of point forecasts. Abrahart and See (2002) evaluated six data fusion strategies and found that data fusion by an artificial neural network (ANN) model provided the best solution. Shu and Burn (2004) applied ANN ensembles in pooled flood frequency analysis for estimating the index flood and 10-yr flood quintiles. The data fusion method was used to combine individual ANN models in order to enhance the final estimation. Tapiaidor et al. (2004) used a neural networks-based fusion technique to estimate half-hourly rainfall using satellite passive microwave and infrared data. The rainfall retrieval capabilities of this method were tested against Geostationary Operational Environmental Satellite (GOES) precipitation index (GPI), GOES precipitation index and a histogram-matching (HM), and gauge data which revealed improved temporal resolution of the passive microwave (PMW) estimates and provided reasonable discrimination capabilities. Ma et al. (2003) used a data fusion approach for soil erosion monitoring in the Upper Yangtze River Basin of China. They concluded that this method improved the accuracy for every parameter calculated using the Universal Soil Loss Equation (USLE) model.

This research provides assessment of two different methods of data fusion in hydrological simulation. These two strategies are based on the non-parametric KNN method. Application of these two methods is tested in daily inflow simulation of the Taleghan Reservoir.

METHODS AND MATERIALS

Data fusion methods

In the case of using multiple individual models to forecast or simulate hydrological processes, all having common predictors; the general equation of data fusion method is as below:

\[ [Y_i] = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{im} \end{bmatrix} = \begin{bmatrix} f_1(X_i) \\ f_2(X_i) \\ \vdots \\ f_m(X_i) \end{bmatrix} + \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{im} \end{bmatrix} \quad i = 1, \ldots, n \]

where \( m \) is the number of models, \([Y_i]\) the matrix of estimation of \( y \) by different individual models, \( \epsilon \) model error and \( n \) the number of observations. Using the DF method, \([Y_i]\) is eventually summed up to a unique estimation of \( y \). Some of the common approaches for DF are simple and weighting averaging (Shu and Burn, 2004), user’s judgement, using empirical models like ANNs (See and Abrahart, 2001) and the method of error analysis. Azmi et al. (2010) also proposed a new statistical method based on the non-parametric KNN simulation which is described here.

Data fusion using KNN method

Non-parametric estimation of probability densities and regression functions are carried out using weighted local averages of the dependent variable (Azmi et al., 2010). This is the foundation of nearest neighbour methods. KNN methods use the similarity between observations of predictors and similar sets of historical observations (successors) to obtain the best estimate for a dependent variable. The \( K \) vectors of historical observations with the minimum norm are selected. The most widely used formula to compute the distances and select nearest neighbours is the Euclidian norm (Karlsson and Yakowitz, 1987), which for a \( P \)-dimensional feature vector is calculated as

\[ \text{Dis}_{ij} = \sqrt{\sum_{i=1}^{P} w_i (x_{it} - x_{jt})^2} \]  

(2)

where \( w \) is the weight of each predictor in calculating the distance. The estimated or forecasted value of the dependent variable in time \( t \) is then estimated as a weighted average of the nearest neighbours in such a way that the greater weights are assigned to the nearest neighbours. A kernel function proposed by Lall and Sharma (1996) is used to estimate the weights and the value of dependent variable in time \( t \) is estimated as

\[ y_t = \frac{\sum_{i=1}^{K} (1/i) y_i}{\sum_{i=1}^{K} (1/i)} \]

(3)

where \( i \) is the order of the neighbours and the nearest has the lowest order \((i = 1 \text{ to } K)\) and \( y_i \) is the magnitude of the nearest neighbour \( i \). The weights and number of the nearest neighbours are selected such that they produce the lowest value of mean square error of the forecasting or simulation.

Azmi et al. (2010) proposed an algorithm using the concepts of the KNN method, to be used in the DF method. In the present study, two different algorithms are proposed based on the algorithm of Azmi et al. (2010). The first algorithm (AL1) is as follows:
1. The magnitude of the dependent variable is estimated using \( m \) individual simulation models for the entire historical period (\( N \): number of all observations).

2. Each model used in stage 1 is evaluated for the entire calibration period (\( n \): number of observations in calibration phase).

3. The distances between the predictor vector in current time, \( [X_t] \), and all the historical predictors \( [X_j] \), are calculated and K-nearest neighbours are selected among \( n \) calibration vectors such that the lowest distance is assigned to the nearest neighbour.

4. Matrix \( [Y] \) is formulated as below, where \( Y_{ti} \) is the magnitude of the dependent variable in time \( t \) from the model that is selected as the best model in the \( i \)th neighbour:

\[
[Y] = \begin{bmatrix}
Y_{t1} \\
Y_{t2} \\
\vdots \\
Y_{tK}
\end{bmatrix}
\] (4)

5. The dependent variable in time \( t \) is estimated as

\[
y_t = \frac{\sum_{i=1}^{K} (1/i) \times Y_i}{\sum_{i=1}^{K} (1/i)}
\] (5)

In this algorithm, all the independent models are evaluated in the calibration (historical) period and the best one is selected for each observation. The results from these best models are utilized in the next steps to estimate the magnitude of the dependent variable in current time.

A modified form of this algorithm is also assessed. In this algorithm (AL2), instead of selecting the best model for each observation in the historical period, all individual models are used. In each time step in the historical period, according to the error value of each of the \( m \) models, weights are assigned to the models and the magnitude of \( Y_{ti} \) is estimated as a weighting average between all \( m \) models. Hence, the only difference between these two algorithms is in step 4.

**Study area**

The Taleghan Reservoir is located in Qazvin province, Iran. This dam was constructed in 2007 and has been in operation to supply the Qazvin Irrigation and Drainage Network with 278 MCM (million cubic metres) each year. Also, 20 MCM is allocated for groundwater recharge, 150 MCM for the domestic demand of Tehran (Iran’s capital) and 12 MCM for ecosystem demand. The area of its hydrological basin is about 960 km\(^2\), average total yearly precipitation is about 615 mm and average daily river inflow is about 16.3 m\(^3\) s\(^{-1}\).

**DEVELOPMENT OF INDIVIDUAL SIMULATION MODELS**

**Artificial neural network models**

In this study, four different types of ANN models were used as individual simulation models. Table I summarizes the properties and performance indices of ANN models. The dependent variable was daily inflow of the Taleghan Reservoir. FeedF1 and FeedF2 (feed forward back propagation models with 1 and 2 hidden layers, respectively) models use the Levenberg-Marquardt back propagation (LM) learning function. GRNN2 is a generalized regression network and RB2 is an ANN model with radial basis function. Different combinations of inputs (precipitation and mean temperature lags) were analysed. In Table I, the signs ’P’ and ’T’ represent daily precipitation and mean temperature lags, respectively. They include lags 0 through lags 3. For example, PPPTT means the inputs include precipitation from two days ago until the current day and temperature of the last day and the current day.

\[
R^2, \text{ SRMSE and SMAE are calculated as follows:
}\]

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (O_i - P_i)^2}{\sum_{i=1}^{n} (O_i - \bar{O})^2} \times 100
\] (6)
Table I. Properties and performance indices of ANN models

<table>
<thead>
<tr>
<th>Model</th>
<th>Inputs</th>
<th>Learning function</th>
<th>Neurons or spread value</th>
<th>( R^2 (%) )</th>
<th>SRMSE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Train</td>
<td>Valid</td>
<td>Test</td>
</tr>
<tr>
<td>FeedF2</td>
<td>PPT</td>
<td>LM</td>
<td>20</td>
<td>82</td>
<td>86</td>
</tr>
<tr>
<td>FeedF1</td>
<td>PPT</td>
<td>LM</td>
<td>30</td>
<td>81</td>
<td>86</td>
</tr>
<tr>
<td>GRNN2</td>
<td>PPPTT</td>
<td>–</td>
<td>0.6</td>
<td>87</td>
<td>93</td>
</tr>
<tr>
<td>RB2</td>
<td>PPT</td>
<td>–</td>
<td>300</td>
<td>82</td>
<td>94</td>
</tr>
</tbody>
</table>

LM: Levenberg–Marquardt back propagation model; FeedF1 and FeedF2: feed forward back propagation models with 1 and 2 hidden layers; GRNN2: generalized regression network; RB2: ANN model with radial basis function; P: precipitation lags; T: temperature lags; \( R^2 \): linear correlation coefficient; SRMSE: standardized root mean squared error.

\[
\text{SRMSE} = \left[ \frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n} \right]^{0.5} \times \frac{100}{O} \quad (7)
\]

\[
\text{SMAE} = \left[ \frac{\sum_{i=1}^{n} |P_i - O_i|}{n} \right] \times \frac{100}{O} \quad (8)
\]

where \( R^2 \) is the linear correlation coefficient, SRMSE the standardized root mean squared error, SMAE the standardized mean absolute error, \( Q_i \) the observed value on day \( i \), \( P_i \) the simulated value on day \( i \), \( n \) the number of days and \( O \) the mean of the observed values.

**Hammerstein–Wiener models**

Hammerstein–Wiener (HW) systems often arise in the black box approach to identify nonlinear systems. Hammerstein–Wiener systems consist of linear dynamic blocks and static nonlinearities that are interconnected in series, parallel or in feedback (MATLAB, 2008). Figure 2 and Equations (9)–(11) show the general structure of HW models. This structure has been successfully used in order to simulate nonlinear systems in different fields of science (Eskinat et al., 1991; Pearson and Pottmann, 2000; Kalafatis et al., 1995, 1997; Celka et al., 2000), but not in hydrology.

\[
W(t) = G(U(t)) \quad (9)
\]

\[
X(t) = \frac{B_{ij}(q)}{F_{ij}(q)} W(t) \quad (10)
\]

\[
Y(t) = H(X(t)) \quad (11)
\]

where

- \( U(t) \) and \( Y(t) \) are the inputs and outputs for the system, respectively;
- \( G \) and \( H \) are nonlinear functions that correspond to the input and output nonlinearities, respectively. For multiple inputs and multiple outputs, \( G \) and \( H \) are defined independently for each input and output channel;
- \( W(t) \) and \( X(t) \) are internal variables that define the input and output of the linear block, respectively;
- \( B(q) \) and \( F(q) \) in the linear dynamic block are linear blocks similar to the polynomial in an output-error model (MATLAB, 2008).

In this research, two HW models are used as individual simulation models. HW1 is a Hammerstein–Wiener model with 45 Sigmoidnet terms for precipitation, 40 Sigmoidnet terms for temperature and 48 piecewise-linear nonlinear terms for the dependent variable (daily reservoir inflow). This model uses standardized inputs. HW2, which uses non-standardized inputs, has 25 Sigmoidnet terms for precipitation, 21 Sigmoidnet terms for temperature and 63 piecewise-linear nonlinear terms for the dependent variable. This number of terms were selected using a repetitive algorithm. Table II shows the results of these two HW models.
RESULTS AND DISCUSSION

Evaluation with historical data

The results of each individual model simulation of the validation period besides the results of the DF method are presented in Tables III and IV. From these tables, it can be concluded that only the HW1 model could outperform the DF method. Superiority of the DF method over the individual ANN models is reported by other researchers (e.g. Azmi et al., 2010; Shu and Burn, 2004). It can be seen that no significant difference can be detected between the two DF algorithms. However, the performance of algorithm AL2 is to some extent better than algorithm AL1. The lowest value of SRMSE for AL1 is related to a K value of 20 and lag time of 1 day. But this value is related to a K value of 10 for algorithm AL2. More increase in K increases the errors.

Evaluation in combination with a weather generator model

In this section, a weather generator (WG) model was used to generate 100 yr of precipitation and temperature daily time series. This model is capable of keeping the correlation between neighbouring weather stations and also is able to keep the daily and monthly autocorrelations between climatic variables. An extended Markov model is used in this model to simulate short-term autocorrelation in the precipitation occurrence time series of a weather station besides keeping the spatial correlation between neighbouring stations. Also, a non-parametric algorithm was developed and used in this model to generate climatic variables with the goal of keeping the correlations and autocorrelations between climatic variables in each station and also spatial correlation between neighbouring stations. The model is well tested, but the results are beyond the scope of this paper. Figure 3 shows the errors of different individual models in combination with the WG model in simulating the monthly mean of daily inflow values, and Figure 4 shows the errors of simulating daily peak inflows. Although models HW1 and HW2 show quite good performance in the calibration and validation periods, their errors increase when they are used in combination with the WG model. The greatest errors are related to the wet months and both models (especially HW1) overestimate absolute maximum inflows. The deviations of RB2 are also considerable (especially in August).

Table V summarizes the mean, standard deviation and maximum value of daily reservoir for each month. The best performance in simulating daily mean values is related to FeedF1 and the worst is related to HW1. Again, HW2 model outperforms the other models in simulating daily standard deviation values. However, the results of all models are unacceptable. Table VI presents the results of individual model assessments in combination with the WG model.

Evaluation of data fusion method

Tables VII and VIII show the results of DF methods in combination with the WG model. DF methods considerably improve the results of individual models. The best results are related to algorithm AL1 (K=5, Lag = 2). The SRMSE value for the simulation of daily mean and standard deviation values are about 6 and 21%, respectively.
The performance of algorithm AL2 is very similar to AL1 in simulating daily mean values. But its results are not very acceptable in simulating the daily standard deviation values. One of the reasons for this issue could be that in this algorithm, the weighting average of all models is used. This averaging process could result in lower variances. Another reason could be the short length of the study period. Occurrence of peak flows in this period increased the magnitude of daily variance.

Table IV. The results of DF algorithms for validation period ($n = 372$)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Lag</th>
<th>$K$</th>
<th>$R^2$ %</th>
<th>SRMSE %</th>
<th>SMAE %</th>
<th>Algorithm</th>
<th>Lag</th>
<th>$K$</th>
<th>$R^2$ %</th>
<th>SRMSE %</th>
<th>SMAE %</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL1</td>
<td>1</td>
<td>5</td>
<td>94</td>
<td>43</td>
<td>25</td>
<td>AL2</td>
<td>1</td>
<td>5</td>
<td>95</td>
<td>40</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>95</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td>20</td>
<td>95</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>30</td>
<td>95</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td>30</td>
<td>94</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5</td>
<td>94</td>
<td>43</td>
<td>25</td>
<td>2</td>
<td>5</td>
<td>95</td>
<td>40</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>94</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td>20</td>
<td>94</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>30</td>
<td>94</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td>30</td>
<td>94</td>
<td>44</td>
</tr>
</tbody>
</table>

$K$: number of nearest neighbours; $R^2$: linear correlation coefficient; SRMSE: standardized root mean squared error; SMAE: standardized mean absolute error.

Table V. Monthly mean, standard deviation and maximum values of individual models in combination with the WG model

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Model</th>
<th>Mean</th>
<th>Std</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td></td>
<td>13.83</td>
<td>17.04</td>
<td>109.54</td>
</tr>
<tr>
<td>PT</td>
<td>HW1</td>
<td>18.74</td>
<td>32.12</td>
<td>269.51</td>
</tr>
<tr>
<td>PT</td>
<td>HW2</td>
<td>17.70</td>
<td>21.18</td>
<td>107.81</td>
</tr>
<tr>
<td>PPT</td>
<td>FeedF1</td>
<td>14.98</td>
<td>15.55</td>
<td>101.51</td>
</tr>
<tr>
<td>PPT</td>
<td>FeedF2</td>
<td>15.53</td>
<td>58.99</td>
<td>8190</td>
</tr>
<tr>
<td>PPPTT</td>
<td>GRNN2</td>
<td>14.30</td>
<td>15.02</td>
<td>200.98</td>
</tr>
</tbody>
</table>

Table VI. Performance assessment of individual models in combination with the WG model

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>HW1</td>
<td>R2 %</td>
<td>99</td>
</tr>
<tr>
<td>HW2</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>FeedF1</td>
<td>99</td>
<td>6</td>
</tr>
<tr>
<td>FeedF2</td>
<td>99</td>
<td>16</td>
</tr>
<tr>
<td>RB2</td>
<td>99</td>
<td>22</td>
</tr>
<tr>
<td>GRNN2</td>
<td>99</td>
<td>8</td>
</tr>
</tbody>
</table>

The performance of algorithm AL2 is very similar to AL1 in simulating daily mean values. But its results are not very acceptable in simulating the daily standard deviation values. One of the reasons for this issue could be that in this algorithm, the weighting average of all models is used. This averaging process could result in lower variances. Another reason could be the short length of the study period. Occurrence of peak flows in this period increased the magnitude of daily variance.

Figure 5 and 6 depict the errors of DF methods in simulating the daily mean and standard deviation values of all 12 months with the AL1 algorithm. This method underestimates the...
observed values of monthly mean inflow values. A similar problem exists in simulating the standard deviation values, especially in April and May.

Figures 7 and 8 depict the errors of DF methods in simulating the daily mean and standard deviation values of all months with the AL2 algorithm. This method
overestimates the observed values of monthly mean inflow values and underestimates the standard deviation values, especially in April and May. The lowest error value of this algorithm is related to a $K$ value of 20 or 30 and the lag time of 1 day.

**SUMMARY AND CONCLUSION**

In this research, application of the data fusion method was assessed in simulating daily inflows into the Taleghan Reservoir in Iran. Two data fusion algorithms were proposed and analysed based on $K$-nearest neighbour. Four different ANN models and two different Hammerstein–Weiner models were used as individual simulation models. The results revealed that despite very good performance of some models in the calibration and validation phases, their combination with a weather generator model could lead to eccentric numeric results. This issue was very considerable in relation to HW models. Comparison of the results from individual models with the results of the DF method shows that the DF method can improve the results of individual models considerably. The performance of both DF algorithms.
was similar in simulating daily mean values. But algorithm AL1 resulted in greater values of daily standard deviation. These results suggest that using the DF method could increase the certainty of the results of individual models and can be used as a powerful tool in simulating and forecasting hydrological processes.

REFERENCES


