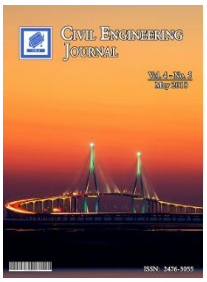
Available online at www.CivileJournal.org

Civil Engineering Journal

Vol. 4, No. 5, May, 2018



Developing an ANN Based Streamflow Forecast Model Utilizing Data-Mining Techniques to Improve Reservoir Streamflow Prediction Accuracy: A Case Study

Hamed Zamani Sabzi [a*], James Phillip King [b], P.E., M. ASCE, Naci Dilekli [c], Bahareh Shoghli [d], Shalamu Abudu [e]

[a] *Postdoctoral research associate, Dept. of Geography and Environmental Sustainability, University of Oklahoma, 100 East Boyd St, SEC Suite 662, Norman, OK 73019.*

[b] *Professor, Dept. of Civil Engineering, New Mexico State University, MSC 3CE, PO Box 30001, Las Cruces, NM, USA 88003, and member of the Engineering Research Center for Re-inventing Urban Water Infrastructure, Stanford University.*

[c] *Research Scientist, Dept. of Geography and Environmental Sustainability, University of Oklahoma, 100 East Boyd St., EC Suite 562, Norman, OK 73019, USA, 88003.*

[d] *Ph.D., Dept. of Civil Engineering, University of North Dakota.*

[e] *Postdoctoral Research Associate, Texas AgriLife Research & Extension Center at El Paso, Texas A&M University System, 1380 A&M Circle, El Paso, TX 79927.*

Received 31 March 2018; Accepted 27 May 2018

Abstract

This study illustrates the benefits of data pre-processing through supervised data-mining techniques and utilizing those processed data in an artificial neural networks (ANNs) for streamflow prediction. Two major categories of physical parameters such as snowpack data and time-dependent trend indices were utilized as predictors of streamflow values. Correlation analysis of different models indicate that, for the period of January to June, using fewer predictors led to simpler modeling with equivalent accuracy on daily prediction models. This did not hold in all periods. For monthly prediction models, accuracy was improved compared to earlier works done to predict monthly streamflow for the same case of Elephant Butte Reservoir (EB), NM. Overall, superior prediction performance was achieved by utilizing data-mining techniques for pre-processing historical data, extracting the most effective predictors, correlation analysis, extracting and utilizing combined climate variability indices, physical indices, and employing several developed ANNs for different prediction periods of the year.

Keywords: Artificial Neural Networks; Data Mining; Streamflow Prediction; Reservoir Management.

1. Introduction

Elephant Butte Reservoir, a multi-objective reservoir, provides electrical power and water for south-central New Mexico and West Texas, including irrigation water for 68,708.25 ha (169, 650 acres) of farmland. Caballo Reservoir, located 40.25 km (25 miles) downstream of Elephant Butte Reservoir, is fed primarily by water released from Elephant Butte Reservoir and provides direct release of irrigation water downstream into the Rio Grande Project during the primary cropping season. Consequently, the two reservoirs can be represented as an integrated control volume, whose streamflows, outflows, and storage volume can be simulated simultaneously. A reliable release plan from Caballo Reservoir to meet existing water demands downstream in the Rio Grande Project directly depends on the Caballo

* Corresponding author: hamedzs@ou.edu

http://dx.doi.org/10.28991/cej-0309163

Reservoir's storage level, whose optimal operation plan must be developed under uncertain conditions. In addition, providing the optimal storage levels in both reservoirs can minimize total evaporation loss from both reservoirs. Such optimization requires accurately estimating streamflow volume to the reservoir and considering the control volume elements in both reservoirs. Based on the geometric characteristics of both Elephant Butte and Caballo Reservoirs, increasing the water storage volume has different effects on the surface areas of each reservoir. Therefore, since water surface area directly correlates to evaporation loss, selecting the appropriate reservoir to store specific water volumes in operational periods is a critical decision. In addition, the streamflows to the reservoirs should be predicted to anticipate the storage level variation in the reservoirs. The initial operation plans based on optimization are defined for seven days ahead, 15 days ahead, and one month ahead, based on the predicted values for streamflow to the Elephant Butte Reservoir and related parameters of the control volume. These parameters include evaporation volumes from both Elephant Butte and Caballo Reservoir; tributary streamflows to the reservoirs; and seepage volume from both reservoirs. The developed prediction models are incorporated in a decision support system that uses the predicted values for 7 days ahead, 15 days ahead, and one month ahead to provide the optimal release plans from Elephant Butte Reservoir into the Caballo Reservoir. The study area has been shown in Figure 1.

Neural networks are widely used to predict dependent variables and they are suitable for addressing objectives identified in this study. Liao et al. (2012) investigated applied data mining techniques through research from 2000 to 2011, and their study shows that artificial intelligence has been utilized significantly and successfully for different types of prediction models. Generally, neural networks utilize several effective criteria (observations) for a dependent target parameter to conclude its value [1]. The application of neural networks has been widely studied by several researchers who investigate how different neural network characteristics would change the processing cost and computational time [1, 2, and 3].

Neural networks are mainly used for linear and nonlinear regressions and as classification methods where prediction is done based on the linear and nonlinear combinations of the input variables. After the training process which optimizes the ANNs, the output values are estimated as a non-linear function of the weighted sum of the input values. Labadie's review study (2004) shows that ANNs have been widely used as an alternative to multiple regression models, and ANNs are an appropriate method for finding patterns in data and classifying nonlinear systems [4]. Although some classification techniques such as Decision Trees (DTs) produce decision rules, which are more transparent for categorizing and prediction purposes, ANNs are more powerful in mapping nonlinear relationships between the effective parameters and the dependent target values [4, 5].

ANN models have long been used by several researchers to predict daily streamflow, monthly streamflow, water quality, water level on the river, rainfall-runoff relationships, calculating solitary wave run-up, and several specific hydraulic characteristics [4, 6-14]. Researching the same study area of this study, Abudu et al. (2010) utilized ANN to forecast monthly streamflow for Rio Grande basin through spring-summer runoff season. Stedinger et al. (1984) found that utilizing a long period of historical data results in more accurate streamflow prediction models. In addition, they showed that adding the snowpack data significantly improves the prediction accuracy [15]. Along with the ANN, hybrid models have been widely used to predict the streamflow and future strategic planning of different resources [16-20]. Humphery et al. (2016) utilized a hybrid approach of Bayesian networks and neural networks for streamflow prediction [12]. Faruk (2010) used a hybrid ARIMA and neural networks to predict the water quality time series data [21].

In most of the previous studies, where an ANN were applied, the effective parameters on the target values were recognized and utilized as predictors. Since there are hundreds of available potential input variables for stream streamflow forecasting, more robust and complex statistical analysis are required in selecting the effective parameters as predictors and designing the structure of the neural networks to improve prediction accuracy. In this study, we utilized physically-based conceptual relationships between the effective parameters on the target values and extracted the existing time-dependent's trend parameters through the historical stream streamflow data. Applying these two concepts improved the prediction accuracy. Different configurations of predictors were utilized to build dozens of Artificial Neural Networks models, and the predictors included daily observed stream streamflow values; temperature; month number; season number; snowpack indices, including snow telemetry precipitation data and snow water equivalent (measured in 12 snow telemetry stations); previous year's streamflow (previous year for the predicted year); average streamflow of previous 2 years (previous 2 years for the predicted year; the average streamflow of previous 5 years (previous 5 years for the predicted year); and number of days in the predicted month. In the materials and methods section, the procedure of extracting the trend parameters is described. Finally, the effective parameters including physical parameters, time parameters, and trend parameters were utilized to develop the streamflow prediction models. Extracting and utilizing the climate variability indices along with neural networks led us to a superior prediction performance for stream flow values through different times of the year.

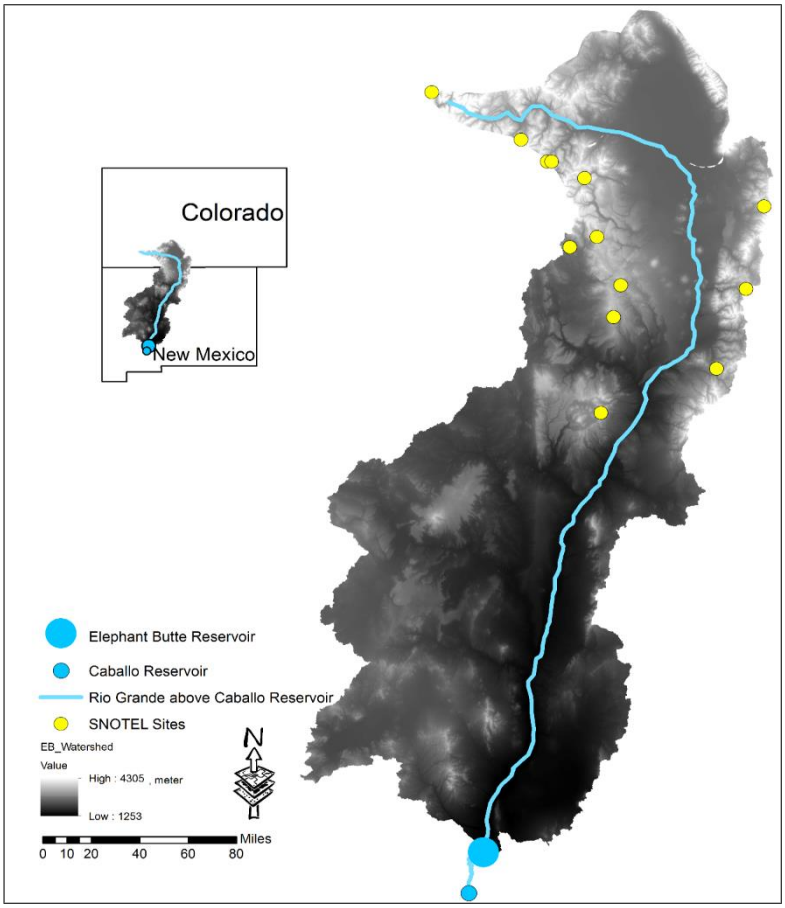


Figure. 1. Study watershed area, SNOTEL sites, and Rio Grande upstream of Elephant Butte Reservoir

In addition, prediction analysis for a specific period of the year is crucially important: having an annually accurate prediction model does not mean it is an accurate model for each specific period of the year (each individual month). Most of the time, because of agricultural, environmental and industrial purposes, a separate model should be developed for each individual prediction period. Therefore, this study develops several prediction functional models for several considered prediction periods. We utilized Artificial Neural Networks (ANN) as an appropriate forecast model for this case of study since reviewed ANN studies that produced superior modeling results. We employ ANNs to develop prediction models that forecast the daily (7 days ahead and 15 days ahead) and monthly stream flows to Elephant Butte Reservoir. The developed prediction model builds on a large body of work in data mining techniques, which are utilized to classify, predict or extract existing knowledge from historical data. The remainder of the paper is organized as follows. Section 2 describes the material and methods. Then we report the numerical results in Section 3. In the last two sections we present our discussions and conclusions.

2. Material and Methods

After considering the conceptual relationships between the physical parameters and the streamflow magnitude, we applied several data mining techniques to select the most appropriate predictors and prediction models. Neurosolution 6.0 was utilized to develop the ANN as predictive models. To develop the ANN model, the effective variables were selected through both correlation analysis and a network modeling process. Historical data from 1961 to 2015 were processed and utilized to develop both monthly and daily streamflow prediction models for the Elephant Butte Reservoir. Two major physical parameters and time-dependent trend indices were utilized as predictors of streamflow values. Time-dependent trend indices were extracted through data processing and data mining techniques. The selected physical variables were snow water equivalent (SWE), temperature, precipitation indices, and daily and monthly streamflow values; the temporal variables were the month and the day of the month. Additionally, to model the existing trend in the historical data, three additional indices were defined, based on the magnitudes of the yearly streamflow of the past year, the average streamflow of the past 2 years, and the average streamflows of past 5 years. The magnitudes for these indices were classified through 10 different classes. The values of 10 and 1 represent the highest streamflow magnitude and the lowest streamflow magnitude, respectively. Based on the feedforward NN, the input variables were fed to the nodes (neurons) on the input layer. Considering the accuracy analysis, the optimal network was designed for different prediction models, for which a more detailed procedure is described in sections 2.1, 2.2, and 3.

2.1. Artificial Neural Networks

Neural networks are computational prediction models, which were developed based on inspiration from functionality of the human brain, which can identify the complex relationships between input and output values and recognize patterns among the input and output values [22- 26]. In prediction models, neural networks are generally presented through interconnected neurons, which utilize portion of some data as training data to compute future values. In this context, historical data may be used to train the model to predict future outcomes. To develop an accurate neural network, the most meaningful information should be utilized. Depending on the simplicity or complexity of the relationships, simple and complex models can be examined to find the best model representing the relationships between the predictors and predicted values [22, 23]. Basically, the best model of the network depends on not just the quality of the utilized data, but on informed trial and error process to obtain the best model that reveals the hidden relationships in the data. Considering these requirements, we utilized a feed-forward network of ANN, which is the most commonly used ANN model. Figure 2 illustrates a typical structure of an ANN model with one input layer, one hidden layer, and one output layer.

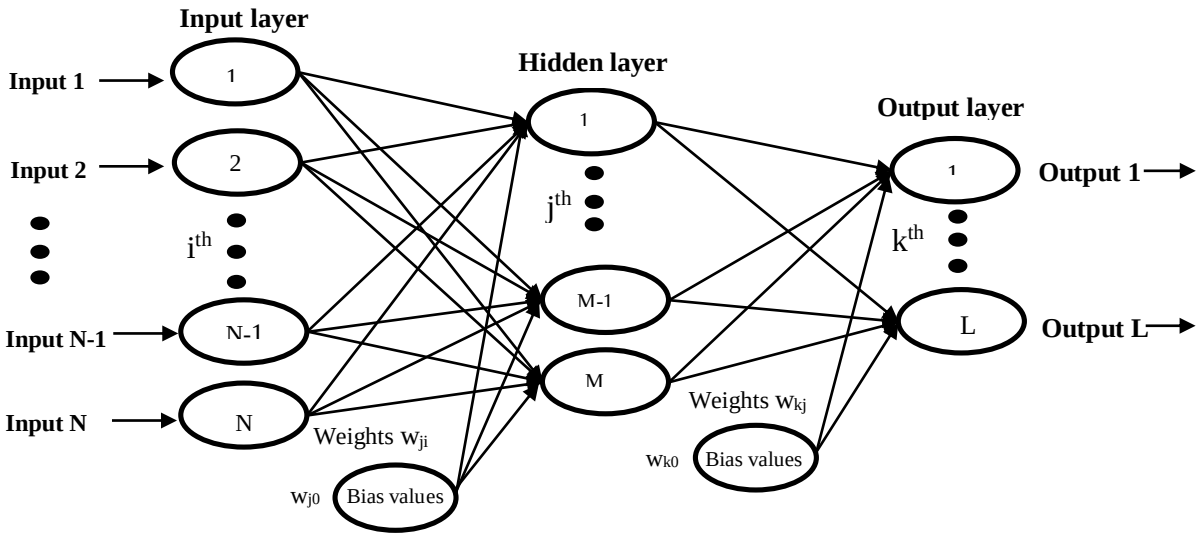


Figure 2. Typical structure of a feedforward neural networks (FFNN) model including one input layer, one hidden layer, and with one output layer [22, 23]

The output values in three-layered feedforward neural networks (FFNNs) are generally driven based on the nonlinear transformation of linear combinations of input parameters as described by Kim and Valdes (2003), in which the output values explicitly expressed as follows [23, 24]:

$$\hat{y}_k = f_0[\sum_{j=1}^{M} w_{kj}.f_h\left(\sum_{i=1}^{N} w_{ji}x_i + w_{j0}\right) + w_{k0}] \quad (1)$$

Where w_{ji} is the assigned weight that connects the ith neuron from input layer to the jth neuron of the hidden layer; w_{j0} is the bias value assigned to the jth neuron in the hidden layer; f_h is the assigned activation function applied on the hidden neurons; w_{kj} is the assigned weight that connects the jth neuron of the hidden layer to the kth neuron of the output layer; w_{k0} is the bias value assigned to the kth neuron in the output layer; and f_0 is the assigned activation function applied on the output neurons:

$$E(n) = \frac{1}{2}\sum_{p=1}^{N}\sum_{k=1}^{L}[y_{pk}(n) - \hat{y}_{pk}(n)]^2 \quad (2)$$

Where N is the number of inputs (observations); L is the number of predicted outputs; $y_{pk}(n)$ is the actual or desired target values; and $\hat{y}_{pk}(n)$ is the value predicted by the network for the kth neuron through nth iteration [22, 23].

2.2. The Characteristics of Developed ANNs

In this study, we utilized available daily and monthly data from 1961 to 2015. In most of the analysis, 60%, 15%, and 25% of data were used as training, cross validation (CV), and testing datasets, respectively. The hyperbolic tangent function was utilized as activation function through the hidden layer, and a linear function was selected as the activation function in the output layer. Considering the convergence speed, stability in the learning process, and learning error values, the Momentum was selected as a learning rule for updating the weights through all the designed networks. For the training process, step size and momentum were taken as 1 and 0.7, respectively.

In most of the designed ANNs, defining one hidden layer derived higher accuracy. Based on the training process monitoring, the maximum number of epochs were selected as 1000 to 5000 epochs in separate models. The appropriate number of nodes (neurons) on the hidden layer was obtained through a trial and error process. Then, based on the accuracy performance (error magnitude), in most of the designed ANNs, 12 to 15 nodes were assigned to the hidden layer. Based on the accuracy performance of the developed models, training, cross validation, and testing procedures, the optimal ANN prediction models were selected. Finally, three categories of daily prediction models for the annual period, the January to June season, and the March to April season were introduced.

3. Numerical Results

3.1. Correlation analysis

Regression analysis were performed between each individual predictor and predicted future daily streamflow values to investigate the existing significant correlation between the predictors and the future predicted values [22, 23]. To investigate the effectiveness of different predictors on predicted streamflow values, we analyzed several linear regression models between each individual predictor and the predicted streamflow values as suggested in the literature [23]. Table 1 illustrates the statistical test results of the developed linear regression models. The results indicate whether each individual predictor has a statistical significant effect on the predicted value or not. In addition, the effects of the SWE for the months of January to June on the monthly streamflow values were investigated.

Table 1. Examined T-test analysis for regression analysis between each individual considered predictor and predicted daily streamflow values

P-value	I_{t1}	I_{t2}	I_{t3}	I_{t4}	I_{t5}	I_{t6}	I_{t7}	I_{t8}	I_{t9}	I_{t10}	I_{t11}	I_{t12}	I_{t13}	I_{t14}	I_{t15}
MI	3.53E-05	1.14E-06	2.03E-08	1.23E-10	1.88E-13	7.08E-17	2.36E-20	1.37E-24	5.2E-29	7.73E-34	3.57E-38	4.6E-42	5.66E-46	3.78E-50	4.25E-54
DI	0.27680	0.21027	0.08612	0.01013	0.00141	0.00013	7.26E-05	0.00082	0.00682	0.04620	0.16539	0.45628	0.98537	0.45779	0.20748
PY	0.003	7.46E-06	9.03E-09	1.35E-11	1.48E-14	1.24E-17	1.01E-20	4.58E-23	1.61E-25	4.34E-28	9.74E-31	4.31E-33	2.83E-35	1.89E-37	1.43E-39
P2Y	0.10976	0.01473	0.00220	0.00029	4.99E-05	7.99E-06	1.53E-06	4.03E-07	1.21E-07	3.76E-08	1.3E-08	4.92E-09	1.94E-09	8.62E-10	3.97E-10
P5Y	0.03902	0.00471	0.00057	7.29E-05	9.16E-06	1.14E-06	1.64E-07	2.58E-08	4.18E-09	6.53E-10	1.04E-10	1.98E-11	3.6E-12	7.35E-13	1.88E-13
SI	2.31E-05	1.46E-07	1.23E-09	6.06E-12	9.03E-15	5.26E-18	2.56E-21	2.72E-25	1.67E-29	3.98E-34	2.15E-38	5.72E-42	2.08E-45	4.07E-49	1.66E-52
P_{M-1}	0.80880	0.90004	0.91807	0.96412	0.99423	0.97066	0.90145	0.74373	0.65389	0.65105	0.702657	0.816241	0.967637	0.994966	0.9839
P_{M-2}	0.222195	0.132627	0.056644	0.017911	0.002793	0.000425	6.48E-05	1.18E-05	2.9E-06	1.13E-06	5.93E-07	6.04E-07	5.46E-07	3.48E-07	2.47E-07
SWE	5.75E-20	7.74E-37	3.6E-55	2.38E-75	6.49E-96	8.9E-118	8.8E-140	2.7E-155	1.9E-171	7.4E-188	2.3E-203	4.6E-217	3.6E-230	2.6E-242	2.5E-253
$SWEE_{m^*}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I_t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Note: the bold values indicate the non-significant correlation between the predictors and the daily streamflow values through different future time steps ahead. The P-values are compared with the Alpha-level of 0.05. P-values less than Alpha-level = 0.05 indicate the significant existing correlation between the predictors and the predicted streamflow values. The P-values greater than the Alpha-level = 0.05 indicate that the evaluated parameter statistically does not have significant correlation with the related predicted streamflow value.

Table 2 shows the correlation coefficients between SWE indices through the months of January to June of the basin an monthly streamflow.

Table 2. Correlation coefficients between the SWE indices through months of January to June and monthly streamflow (used historical data: 1961–2015)

Month	Jan. 1	Feb. 1	Mar. 1	Apr. 1	1-May	Jun. 1
Jan.	0.138					
Feb.	0.154	0.169				
Mar.	0.341	0.407	0.468			
Apr.	0.403	0.575	0.658	0.662		
May	0.429	0.666	0.715	0.864	0.782	
Jun.	0.416	0.664	0.695	0.810	0.762	0.446
Jul.	0.353	0.457	0.460	0.587	0.602	0.506
Aug.	0.154	0.072	*	0.012	0.114	0.094
Sep.	0.108	0.153	0.150	0.144	0.152	0.062
Oct.	0.185	0.232	0.188	0.157	0.161	*
Nov.	0.345	0.311	0.247	0.358	0.476	0.371
Dec.	0.415	0.479	0.463	0.583	0.664	0.548

Note: In Table 2 and 3, the star sign (*) indicates that there is not significant correlation between the SWE values and related daily values in different months.

Some models can be accurate in the training process with a lower amount of error and a higher regression coefficient amount but exhibit lower accuracy in testing the test dataset. Therefore, the error and regression coefficient from the testing data set should be evaluated for each variable for its inclusion in the final prediction model. Table 3 shows the numerical results (obtained coefficients of determination) of some of developed networks as prediction models with acceptable performances in term of accuracy.

As shown in Figure 1, the SNOTEL (snow precipitation indices and snow water equivalent) data were obtained and processed from 12 stations on the headwater of the watershed. In the developed hybrid model, the most significant autoregressive lags along with the meaningful predictors of developed ANN models were utilized [22, 23].

Considering the results from Tables 1, 2, and 3, the best daily streamflow prediction models for specific prediction periods were obtained as follows (Sabzi et al., 2017):

$$I_{t+1,t+2,t+3,t+4,t+5,t+6,t+7} = f(I_t, SWE, P_{M-2}, S_i, M_i, P5Y_i, P2Y_i, PY_i) \quad (1)$$

$$I_{t+1,t+2,t+3,t+4,t+5,t+6,t+7} = f(I_t, SWE_{m^*}, P_{M-1}, P_{M-2}, S_i, M_i, D_i, P5Y_i, P2Y_i, PY_i) \quad (2)$$

$$I_{t+1,t+2,t+3,t+4,t+5,t+6,t+7} = f(I_t, SWE_{m^*}, SWEE_{m^*}, P_{M-1}, P_{M-2}, S_i, M_i, D_i, P5Y_i, P2Y_i, PY_i) \quad (3)$$

$$I_{t+1,t+2,t+3,t+4,t+5,t+6,t+7} = f(I_t, SWE_{t^*}, P_{M-2}, M_i) \quad (4)$$

where I_t indicates the observed daily value at a specific day, SWE represents the snow water equivalent at the first day of the prediction month, $SWEE_{m^*}$ represents the snow water equivalent at the first day of the month with higher correlation of the SWE and predicted streamflow values in the prediction month, P_{M-2} represents the precipitation index at two previous month of the predicted month, S_i represents season index, M_i stands for the number of the predicted month, $P5Y_i$ represents the average of annual streamflow value in the past 5 years of the predicted year, $P2Y_i$ represents the average of annual streamflow value in the past 2 years of the predicted year, and PY_i represents the average of annual streamflow value in the past year of the predicted year [22, 23].

The hybrid model is defined through Eq. (7) as follows:

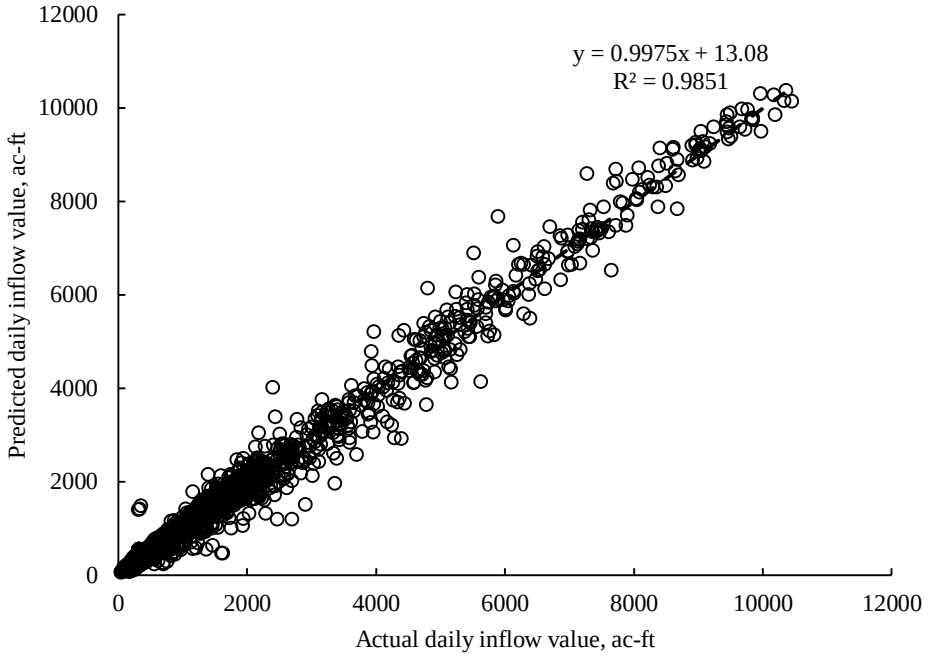
$$I_{t+1,t+2,t+3,t+4,t+5,t+6,t+7} = f(I_t, I_{t-1}, I_{t-2}, I_{t-7}, I_{t-8}, I_{t-24}, I_{t-30}, SWE_{t^*}, P_{M-2}, M_i) \quad (7)$$

Where $I_{t-1}, I_{t-2}, I_{t-7}, I_{t-8}, I_{t-24}, I_{t-30}$ are the lagged daily streamflow values at 1, 2, 7, 8, 24, and 30 days in the past [23].

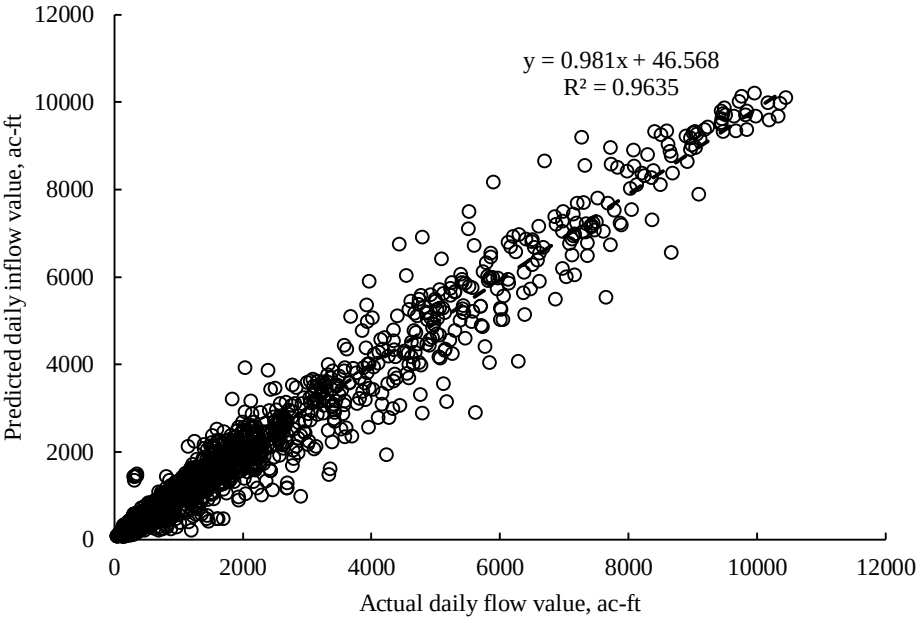
Table 3. Numerical results of the developed networks for prediction models through different periods of the year

Input variables (Predictors)															Correlation coefficient (r) between predicted and observed daily streamflow values							
MI	DI	PYI	P2YI	P5YI	SI	P_M	P_{M-1}	P_{M-2}	SWE_{EM}	SWE_t	I_t	T	No. of Month	Prediction period	I_{t1}	I_{t2}	I_{t3}	I_{t4}	I_{t5}	I_{t6}	I_{t7}	
*	*	*	*	*	*	_	*	*	_	*	*	-	January to June	7 days	0.989	0.981	0.969	0.957	0.946	0.938	0.928	
*	*	*	*	*	*	_	*	*	*	*	*	-	January to June	7 days	0.991	0.980	0.968	0.958	0.945	0.934	0.925	
*	_	*	*	*	*	_	_	*	_	*	*	-	January to June	7 days	0.993	0.982	0.971	0.959	0.948	0.936	0.927	
*	-	-	-	-	-	-	-	*	-	*	*	-	January to June	7 days	0.992	0.982	0.971	0.960	0.948	0.937	0.926	
*	-	-	-	-	-	-	*	*	-	*	*	-	January to June	7 days	0.993	0.982	0.971	0.958	0.947	0.937	0.926	
*	-	-	-	-	-	-	*	*	-	*	*	-	January to June	15 days	I_{t1} 0.990	I_{t2} 0.975	I_{t3} 0.967	I_{t4} 0.953	I_{t5} 0.940	I_{t6} 0.927	I_{t7} 0.917	
															I_{t8} 0.908	I_{t9} 0.903	I_{t10} 0.893	I_{t11} 0.888	I_{t12} 0.874	I_{t13} 0.870	I_{t14} 0.871	I_{t15} 0.856
*	*	*	*	*	*	_	*	*	*	*	*	-	12 Month	7 days	0.982	0.965	0.948	0.929	0.911	0.897	0.884	
*	-	-	-	-	-	-	*	*	-	*	*	-	12 Month	15 days	I_{t1} 0.982	I_{t2} 0.962	I_{t3} 0.937	I_{t4} 0.927	I_{t5} 0.910	I_{t6} 0.895	I_{t7} 0.879	
															I_{t8} 0.869	I_{t9} 0.857	I_{t10} 0.847	I_{t11} 0.838	I_{t12} 0.828	I_{t13} 0.822	I_{t14} 0.818	I_{t15} 0.812
*	-	-	-	-	-	-	*	*	*	*	*	-	March to July	7 days	0.988	0.973	0.956	0.941	0.924	0.908	0.893	
*	-	-	-	-	-	-	*	*	*	*	*	-	March to July	15 days	I_{t1} 0.985	I_{t2} 0.968	I_{t3} 0.946	I_{t4} 0.940	I_{t5} 0.918	I_{t6} 0.907	I_{t7} 0.896	
															I_{t8} 0.884	I_{t9} 0.871	I_{t10} 0.861	I_{t11} 0.840	I_{t12} 0.832	I_{t13} 0.822	I_{t14} 0.816	I_{t15} 0.810
*	-	-	-	-	-	-	-	*	-	$I_{t, t-1, t-2, t-7, t-8, t-24, t-30}$		_	January to June (Hybrid Model)	7 days	0.983	0.973	0.962	0.948	0.934	0.921	0.910	

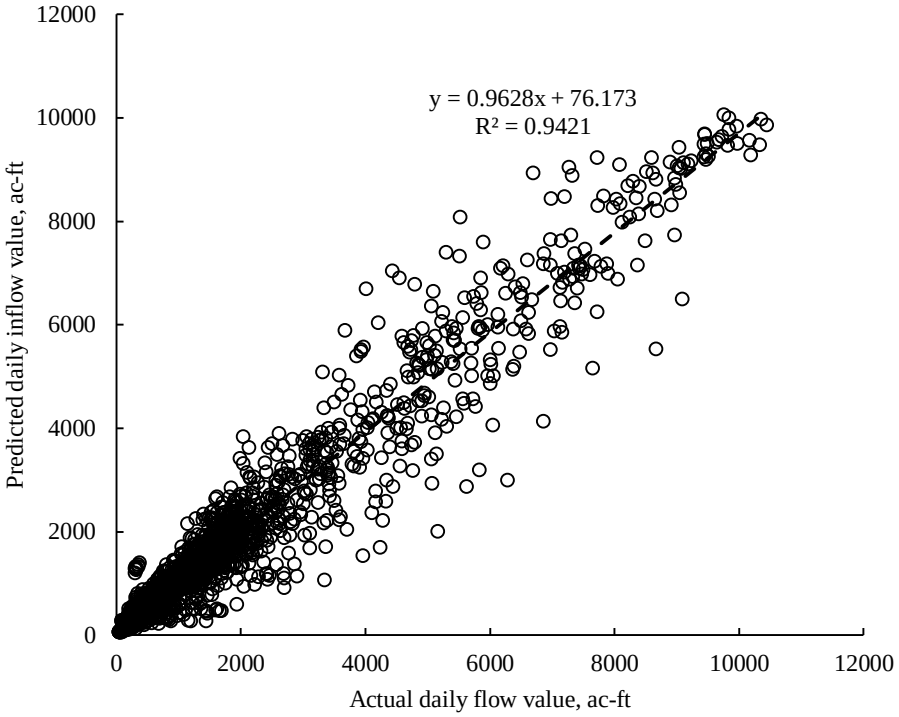
Figure 3. shows the graphical representation of the predicted values versus actual values for 7-day-ahead streamflow predictions to the Elephant Butte reservoir developed based on Eq. (4). Through this paper, we use US system of units, based on which 1 acre-foot (ac-ft) = 1,233.482 cubic meters. Table 4 shows the accuracy performance of the developed model based on the Eq. (6) versus different predicted streamflow values through different time steps ahead.



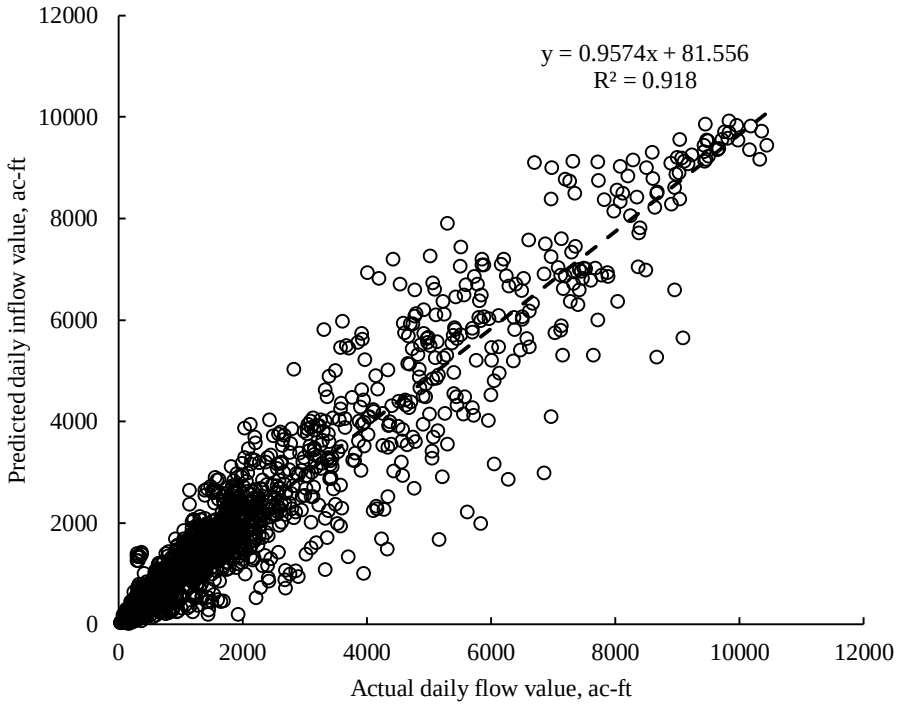
3. a. One day ahead daily streamflow prediction



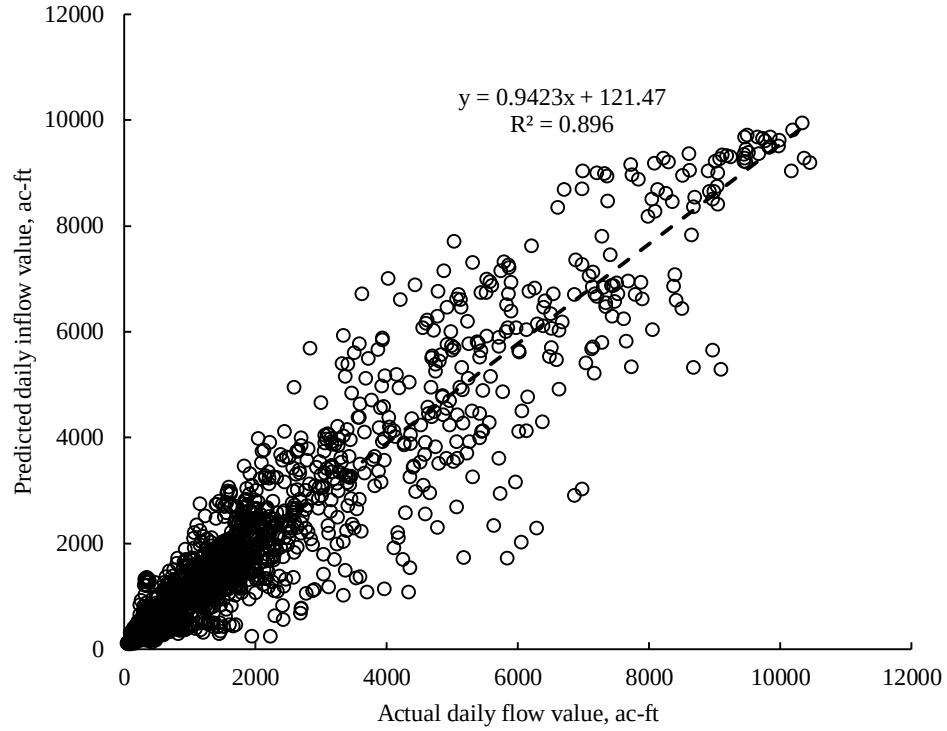
3. b. Two days ahead daily streamflow prediction



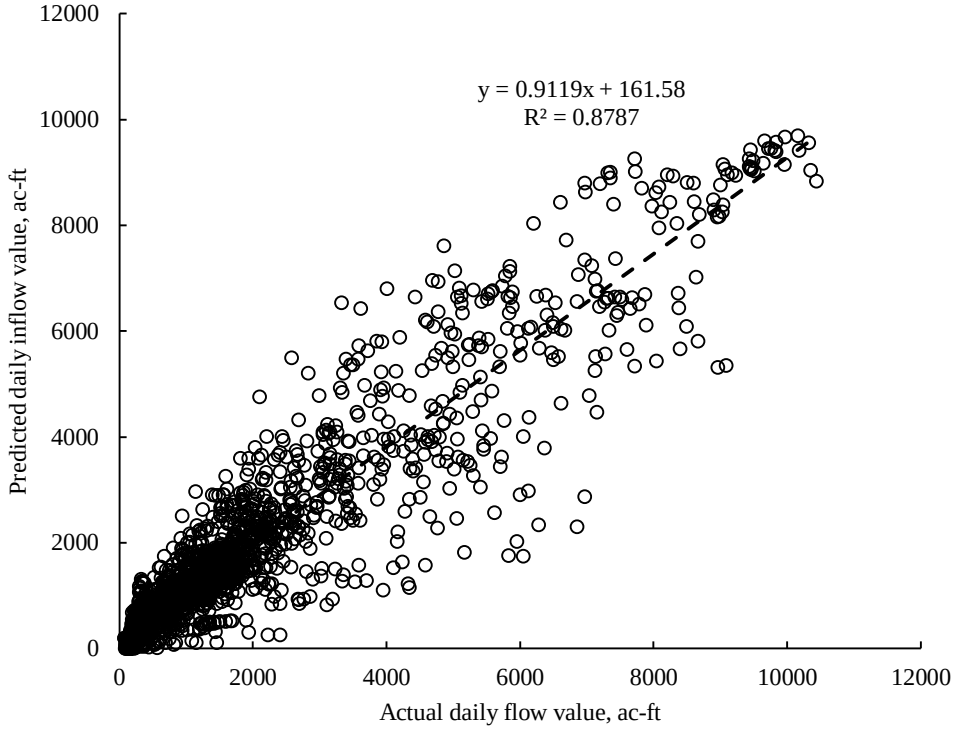
3. c. Three days ahead daily streamflow prediction



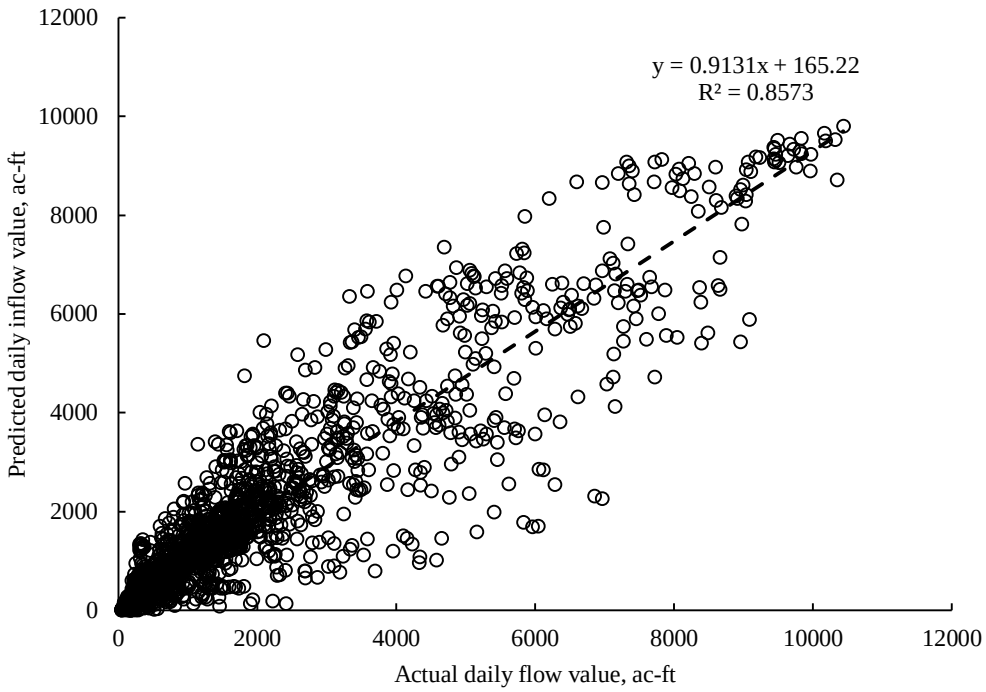
3. d. Four days ahead daily streamflow prediction



3. e. Five days ahead daily streamflow prediction



3. f. Six days ahead daily streamflow prediction



3. g. Seven days ahead daily streamflow prediction

Figure 3. One day to seven days ahead daily streamflow prediction based on the 6 months data for January-June, 1961-2015, CV: 15%, Test: 25%, and Train: 60%

Table 4. Accuracy performance of the model versus different predicted streamflow values through different time steps ahead

Performance	I_{t1}	I_{t2}	I_{t3}	I_{t4}	I_{t5}	I_{t6}	I_{t7}
MSE	51,836	126,980	201,925	290,175	370,151	426,975	511,016
NMSE	0.015	0.037	0.058	0.083	0.107	0.122	0.146
MAE	133.909	205.941	253.910	301.517	343.963	388.089	408.393
Min Abs Error	0.083	0.020	0.108	0.009	0.152	0.112	0.005
Max Abs Error	1,793.183	2,708.357	3,285.706	3,882.992	4,115.903	4,559.752	4,709.103
r	0.993	0.982	0.971	0.958	0.947	0.937	0.926

According to the Table 4, the parameters of MSE, MAE, and NMSE represent the Mean Squared Error, Mean of Absolute Error, and Normalized Mean Squared Error respectively.

3.2. Importance degrees of predictors

The predictors' importance degrees represent the reduction of the target variance for each predictor. Considering the variance of targets, sensitivity measure of the predictors is utilized to rank their importance degrees. Considering a prediction model including *k* predictors, $E\left(V(Y|X_i)\right)$ is the variance over X_{-i} (a (*k-1*) dimensional vector representing all involved predictors except X_i where E is over X_i. As a result, $E\left(V(Y|X_i)\right)$ is an appropriate measure to indicate that how influential is X_i. As a result, smaller amounts of $E\left(V(Y|X_i)\right)$ stand for more influential predictors. Unconditional total variance, V_Y is computed as two complement variances through Eq. (8) as follows:

$$V_Y = E\left(V(Y|X_i)\right) + V(E(Y|X_i)) \quad (8)$$

Where $V(E(Y|X_i))$ stands for main effect of X_i on Y, and $E\big(V(Y|X_i)\big)$ considers the residuals. The sensitivity measure (importance degree) was introduced by Saltelli et al (2004) and for each predictor, its measure of sensitivity is defined as a ratio of $V(E(Y|X_i))$ to the total variance V_Y. The sensitivity measure is calculated through Eq. 9) as follows [24]:

$$S_i = \frac{V_i}{V_Y} = \frac{V(E(Y|X_i))}{V_Y} \tag{9}$$

Where Y represents the target value, in which is defined as a function of predictors ($Y = f(X_1, X_2, \dots, X_k)$), X_i is the predictor, and k is the number of predictors. Then, normalized sensitivities represent the predictors' importance degrees, which are calculated through Eq. (10) as follows:

$$VI_i = \frac{S_i}{\sum_{i=1}^{k} S_i} \tag{10}$$

Where S_i was calculated in previous step through Eq. (9). Tables 5 and 6 represent the importance degrees of predictors in forecasting streamflow values for different time steps ahead based on the models as represented through Eqs. (4) and (6) respectively.

Table 5. Importance degrees for different predictors of streamflow for the developed model based on the Eq. (5).

Nodes	**Importance degrees of predictor on predicted streamflow values**						
Predictors	I_{t1}	I_{t2}	I_{t3}	I_{t4}	I_{t5}	I_{t6}	I_{t7}
DI	0.009	0.007	0.009	0.010	0.017	0.021	0.031
PYI	0.014	0.017	0.024	0.034	0.044	0.064	0.111
P2YI	0.020	0.012	0.017	0.016	0.018	0.018	0.024
P5YI	0.025	0.012	0.018	0.025	0.022	0.025	0.020
PRCP (m-1)	0.026	0.014	0.017	0.020	0.016	0.017	0.042
SI	0.029	0.023	0.028	0.033	0.038	0.052	0.036
MI	0.030	0.027	0.029	0.036	0.041	0.057	0.044
SWE	0.035	0.028	0.049	0.062	0.059	0.049	0.060
PRCP (m-2)	0.038	0.037	0.037	0.046	0.058	0.058	0.066
I_t	0.774	0.824	0.772	0.719	0.687	0.639	0.566

Table 6. Importance degrees for different predictors of streamflow for the developed model based on the Eq. (4).

Nodes	**Importance degrees of predictor on predicted streamflow values**						
Predictors	I_{t1}	I_{t2}	I_{t3}	I_{t4}	I_{t5}	I_{t6}	I_{t7}
PRCP (m-2)	0.009	0.015	0.018	0.017	0.038	0.039	0.051
MI	0.017	0.017	0.020	0.035	0.033	0.037	0.018
SWE	0.018	0.055	0.054	0.097	0.088	0.118	0.160
PRCP (m-1)	0.019	0.012	0.029	0.042	0.058	0.038	0.073
I_t	0.938	0.902	0.878	0.808	0.783	0.768	0.698

Figure 4 shows the desired output and actual output of the network for test dataset based on 6 months' data (January to June). In all developed models, test datasets were selected based on the recent dataset. Testing based on the latest observed data assures the applicability of the developed networks based on recent predictors' values for predicting future datasets.

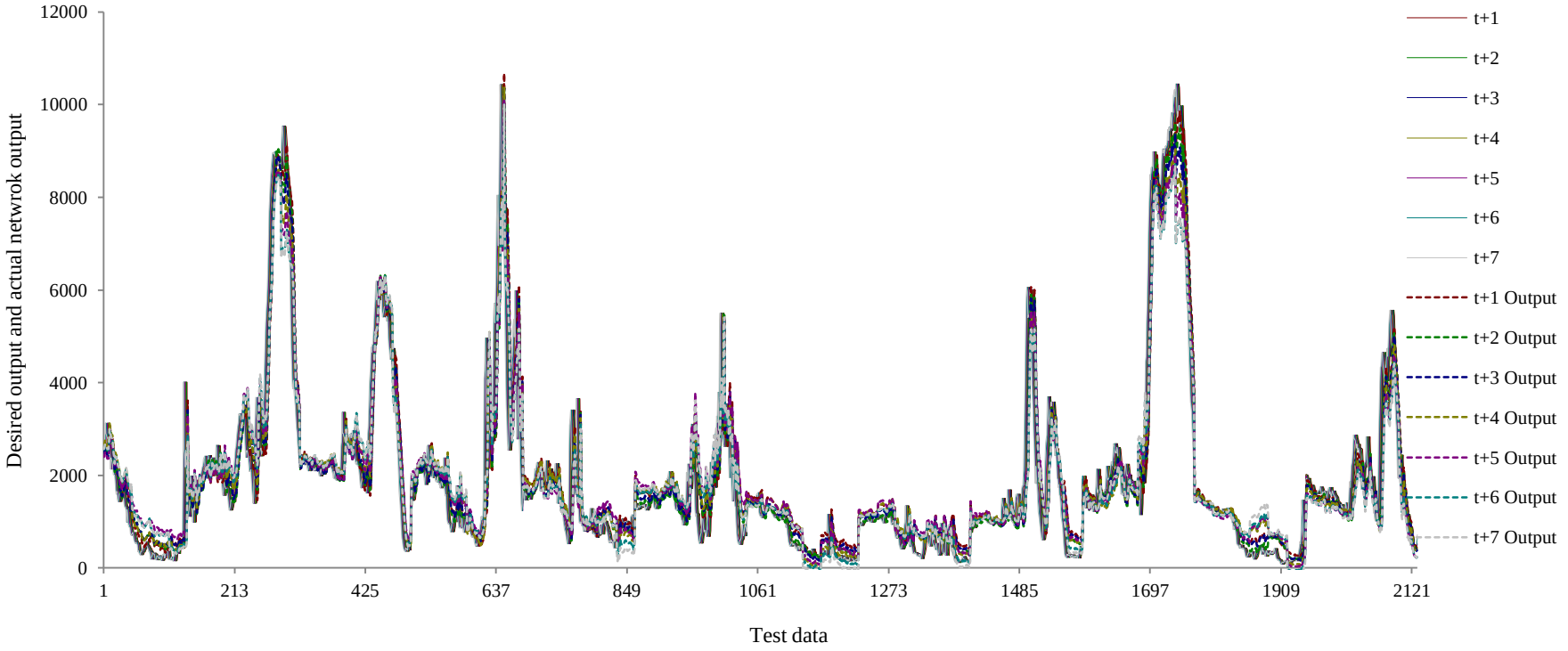
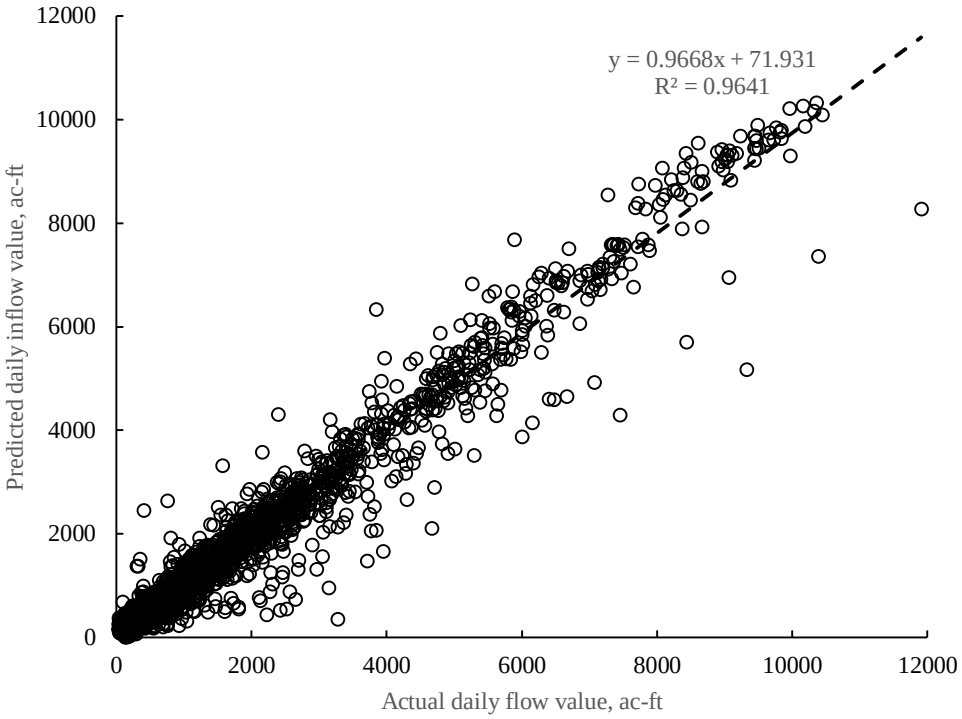
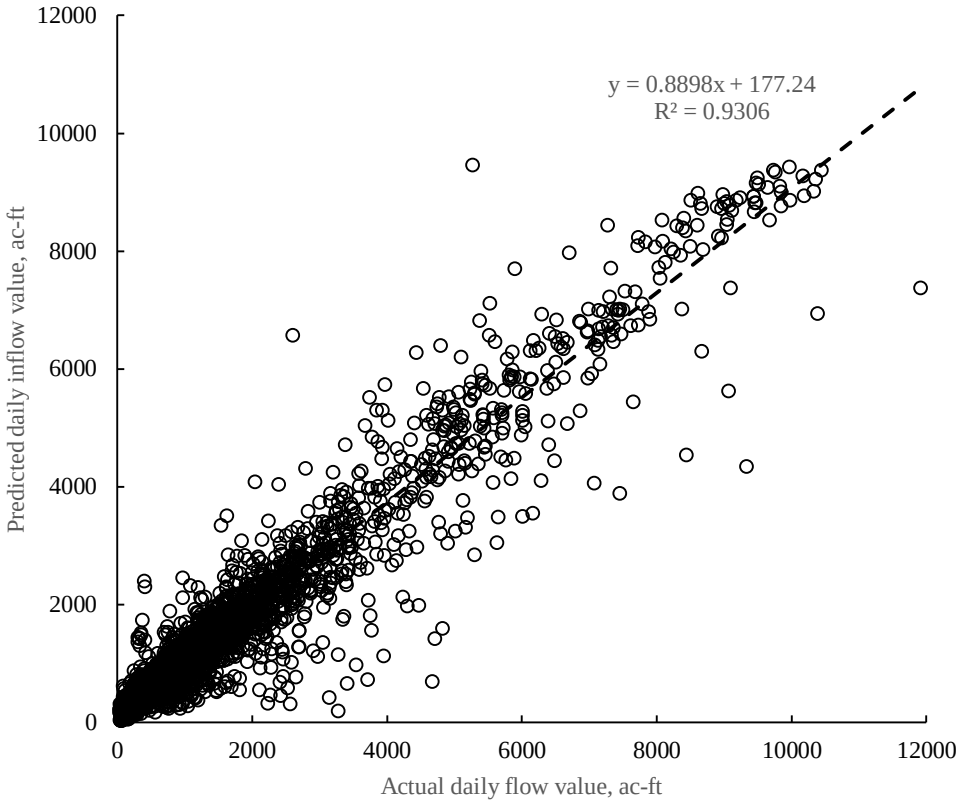


Figure 4. Seven days ahead daily streamflow prediction based on the 6 months' data (January to June), used data period: 1961-2015, Cross Validation (CV): %15, Testing: %25, and Training: %60

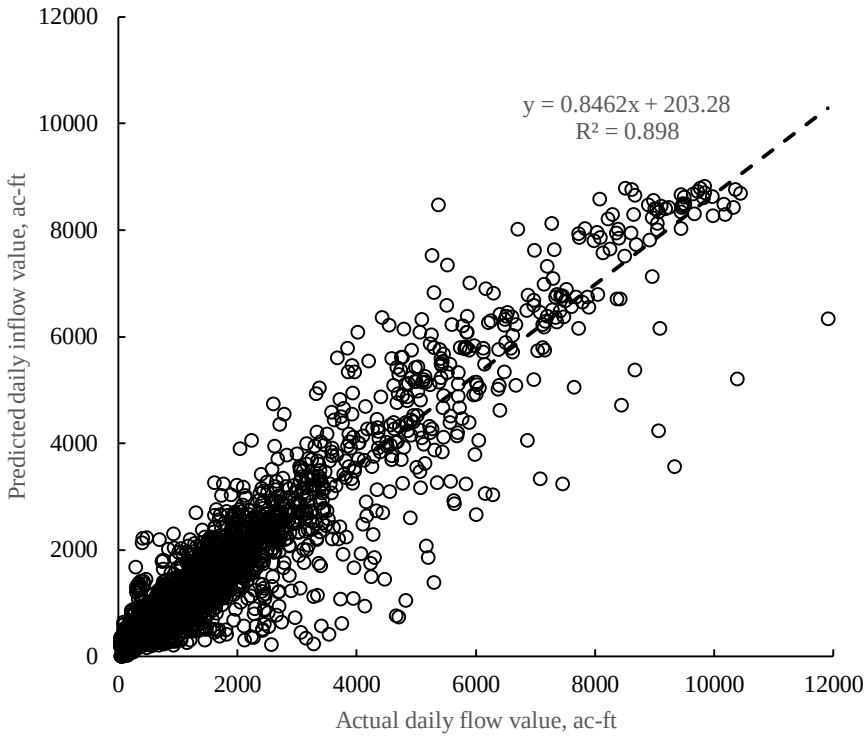
Figure 5 illustrates the graphical representation of the predicted values versus actual values for 7-day-ahead streamflow predictions to the Elephant Butte reservoir developed based on Eq. (4) and applied on annual data (6 months).



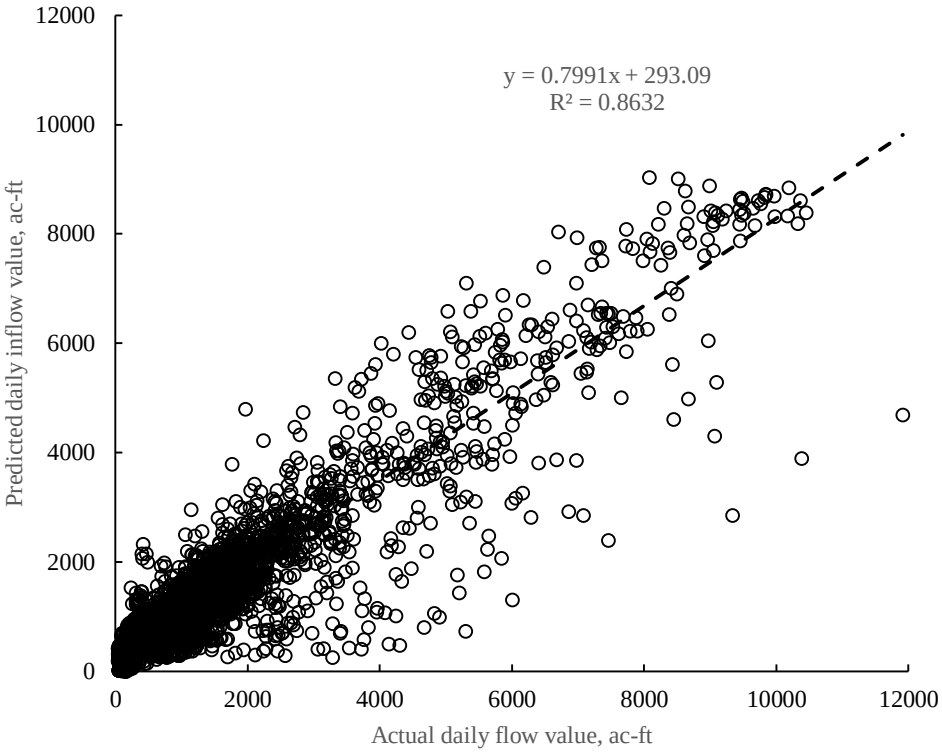
3. a. One day ahead daily streamflow prediction, for 12 months period



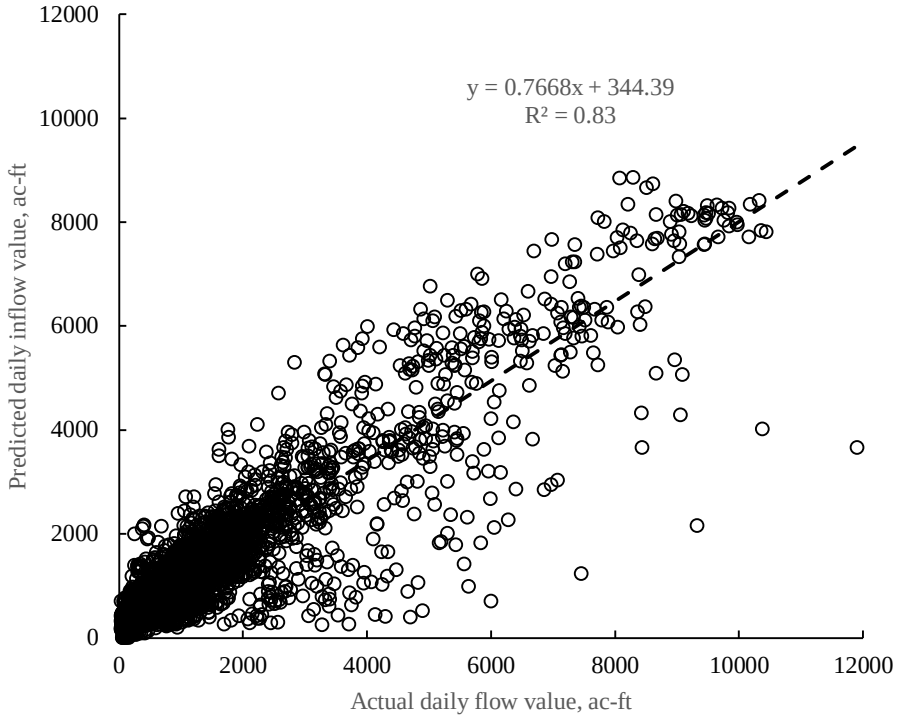
3. b. Two days ahead daily streamflow prediction, for 12 months period



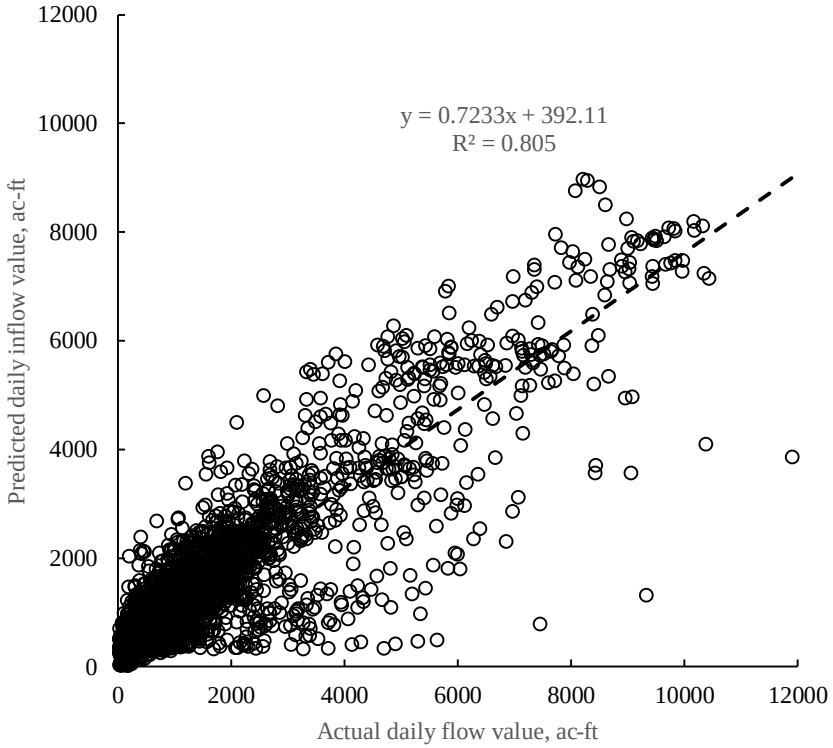
3. c. Three days ahead daily streamflow prediction, for 12 months period



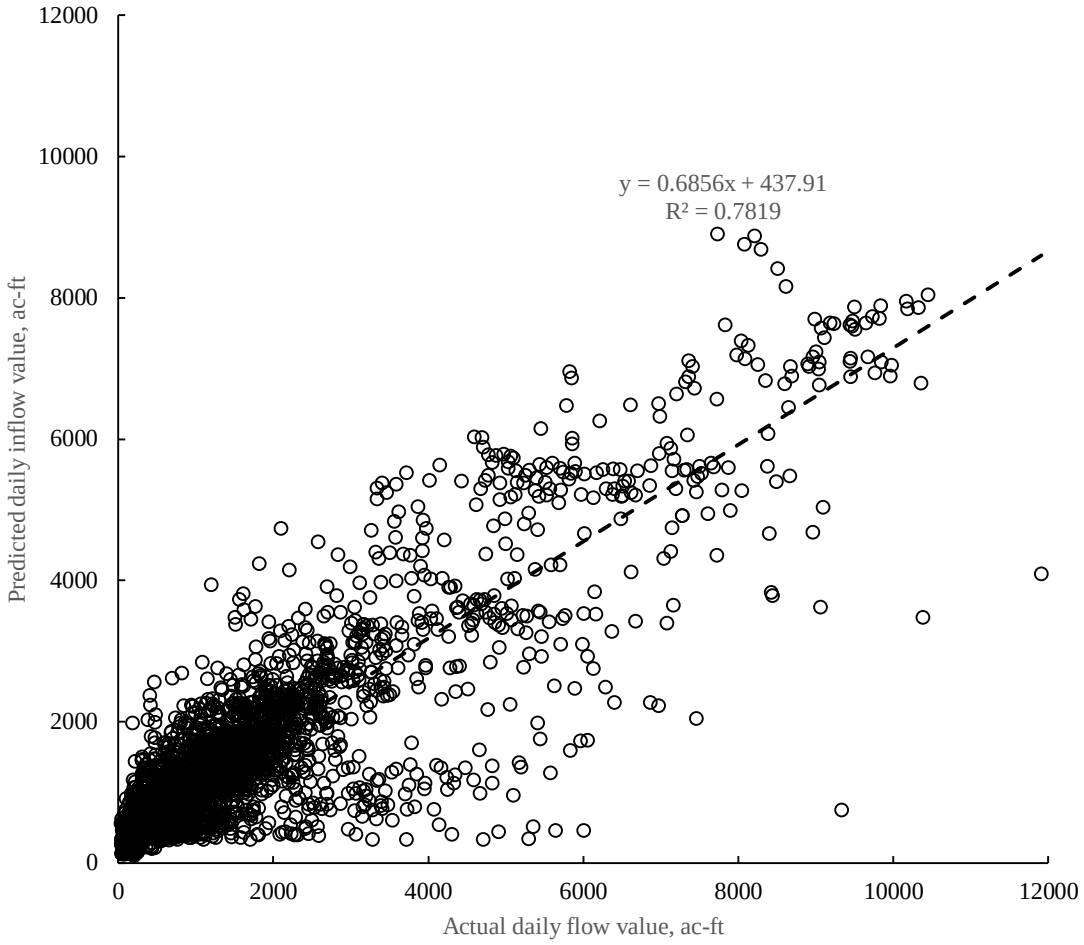
3. d. Four days ahead daily streamflow prediction, for 12 months period



3. e. Five days ahead daily streamflow prediction, for 12 months period



3. f. Six days ahead daily streamflow prediction, for 12 months period



3. g. Seven days ahead daily streamflow prediction, for 12 months period

Figure 5. Seven days ahead daily streamflow prediction based on the 6-month data, 1961-2015, Cross Validation (CV): %15, Testing: %25, and Training: %60

Figure 6 shows the desired output (observed values in test dataset) and actual output of the network for test dataset based on annual daily data.

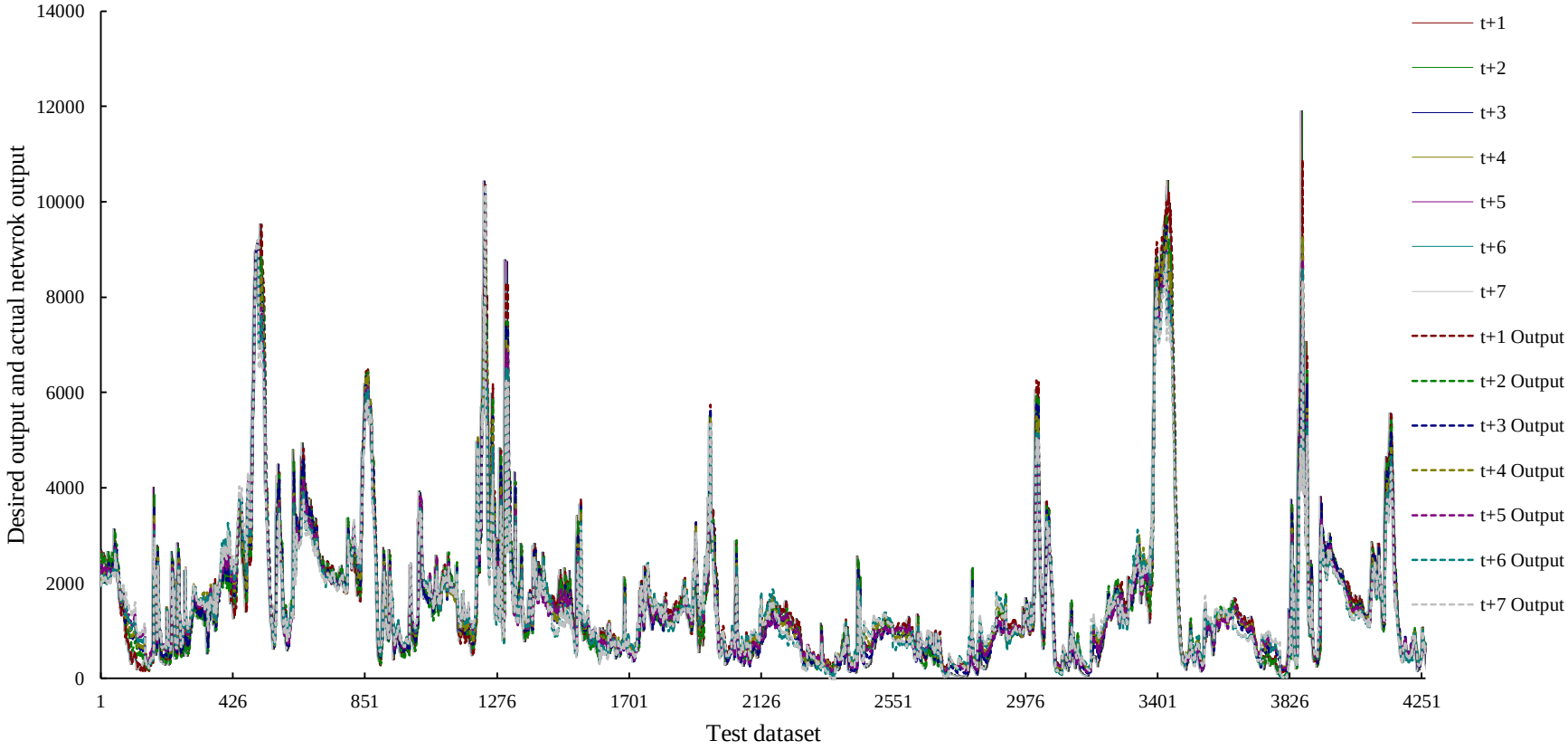


Figure 6. Desired output and actual network output for one to seven days ahead daily streamflow prediction based on the 12 months data, 1961-2007, Cross Validation (CV): %15, Testing: %25, and Training: %60

Table 7. Accuracy performance of the model versus different predicted streamflow values using different time steps ahead based on the 12-month data and developed model using Eq. 6.

Performance	I_{t1}	I_{t2}	I_{t3}	I_{t4}	I_{t5}	I_{t6}	I_{t7}
MSE	84638.477	166999.527	245012.308	331103.609	409315.532	475256.840	537477.879
NMSE	0.036	0.071	0.105	0.142	0.175	0.203	0.230
MAE	158.302	230.514	269.439	329.618	372.736	398.215	425.277
Min Abs Error	0.011	0.007	0.058	0.048	0.164	0.184	0.053
Max Abs Error	4167.692	4987.699	5770.438	7214.738	8242.337	8049.779	8584.266
r	**0.982**	**0.964**	**0.948**	**0.929**	**0.911**	**0.897**	**0.884**

4. Discussion

The predictors' importance degrees for daily and monthly streamflow values are analyzed by both Neurosolution and SPSS software packages. Tables 5 and 6 show that the daily and monthly streamflow values are the most dominant prediction element in predicting the streamflow values seven days one month in advance. This finding confirms Sabzi *et al.*'s (2017) results {22, 23]. The analysis of importance degrees along with the accuracy performance led us to select the optimal prediction models for each daily prediction period applicable for January to June, March to July, and an annual period (12 months). For example, the correlation analysis showed that precipitation indices of previous two months are more important than the precipitation index of the previous month. Therefore, it is reasonable to use the precipitation indices of the previous two months as one of the key predictors. In contradiction to standard engineering judgment, the incorporation of temperature did not show a significant improvement in model accuracy. This observed independency of streamflow values and average temperature should be explored further.

In this study, along with the developed ANN models, several hybrid models were developed, in which in the hybrid daily prediction models, the significant effects of different preceding time periods (past values of streamflow considering different previous timeframes) were recognized through ARIMA, which is a statistical univariate time series prediction model. As illustrated through Eq. (7), the effective preceding time values (effective lagged streamflow values) were incorporated in the prediction models. The accuracy performance analysis showed that a hybrid model did not improve the performance of the developed prediction models, however. Table 7 shows the accuracy performance of the developed model based on the Eq. (6) 6 versus different predicted streamflow values through different time steps ahead based on the 12 months data. Eq. (11) indicates the developed monthly streamflow prediction based on combination of physical, temporal, and time dependent streamflow trend indices. The time dependent stream flow trend indices were obtained through supervised data mining techniques as suggested in the literature [22, 23].

$$I_m = f(I_{m-1}, SWE_{Em}, SWE_m, P_{m-1}, P_{m-2}, S_i, M_i, P5Y_i, P2Y_i, PY_i) \quad (11)$$

where monthly streamflow I_m in mth depends on the streamflow I_{m-1} that is from one month before, SWE_{Em} at first day of the effective month (the month that its SWE index has significant effect on the streamflow at mth month), SWE_m at the first day of the predicted month, P_{m-1} precipitation index of the month before the predicted month, P_{m-2} precipitation index of the 2 months before the predicted month, S_i season number, M_i month number, $P5Y_i$ average of streamflow of the past 5 years, $P2Y_i$ average streamflow of the past 2 years, and PY_i average streamflow of the previous year [22].

Table 8 indicates the sensitivity analysis of the mean of the monthly streamflow to the variation of the utilized predictors. Sensitivity analysis of the mean of streamflow to its predictors indicates the importance degree of predictors. In this process, after developing the pre-trained network, through batch testing, each individual predictor is changed by ± one standard deviation while rest of the variables means are fixed. Then, the predicted value is computed based on the limited steps of ± standard deviation (usually 50 steps) below and above the predictor's mean. The same process is done for rest of the predictors. Then, the variability of the predicted value (I_m) affected by variability of the predictors is obtained. Finally, to rank the effectiveness of the input variables on the outputs (predicted values), relative importance indices were developed by normalizing the sensitivity of the variables [8].

Table 8. Sensitivity analysis of effective parameters on monthly streamflow prediction (April to September)

Predictor	M_i	PY_i	$P2Y_i$	$P5Y_i$	S_i	P_{m-1}	P_{m-2}	SWE_{Em}	SWE_m	I_{m-1}
Sensitivity on I_m	7379	4285	2308	12300	932	3658	11647	18187	23540	57297

Incorporating the time and time dependent trend indices of $S_i, M_i, P5Y_i, P2Y_i, and\ PY_i$ improved the prediction accuracy with higher coefficient of determination ($R^2 = 0.912$) compared to the models with the same data excluding those time dependent trends ($R^2 = 0.9$).

Figure 7 illustrates the estimated and observed monthly streamflow values for the test dataset. The model monthly streamflow predictive model was developed based on the Eq. (11).

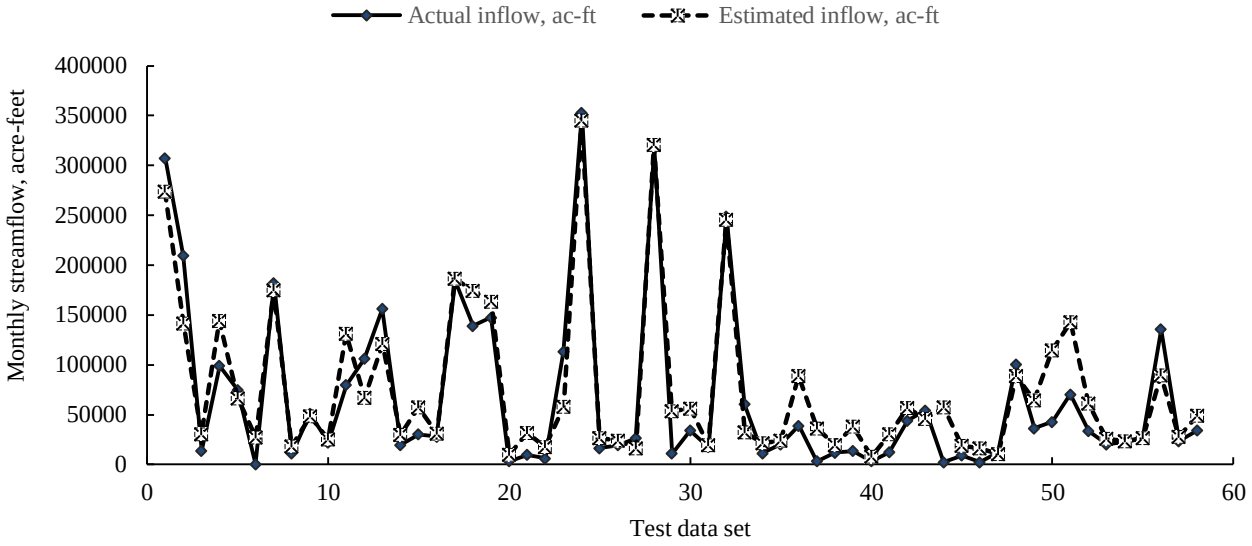


Figure 7. Monthly streamflow prediction model (April to September) with 10 input variables and one hidden layer

5. Conclusions

Considering the statistical analysis and accuracy performance, the key predictors in most of the developed streamflow prediction models were found to be daily observed streamflow value, SNOTEL precipitation indices, and SWE amounts. As shown in tables 1, 2, and 3, for the developed comprehensive daily streamflow prediction model applicable throughout the year (12 months), the effective SWE indices were utilized as a predictor along with other effective predictors. For example, for the months of July, November and December, the SWE indices are zero, but according to the correlation analysis results in table 2, the SWE amount in May has a significant correlation with the streamflow values in July, November, and December. Therefore, SWE of May would be an effective predictor of the streamflow amount in July, November and December. As a result, incorporating the results of correlation analysis as shown in tables 1 and 2 enabled us to utilize effective SWE (SWE_E) and SWE along with other effective predictors in the prediction model. Accuracy performances along with the higher coefficient of determination were two key elements in selecting the optimal prediction models for each specific prediction period. The correlation analysis, along with the accuracy performance analysis, led us to select the relatively parsimonious model that used fewer predictors. Comparison of the hybrid ANN with regular ANN showed that hybrid models of ANN with the statistical univariate ARIMA model did not improve the prediction performances. This can be because of the parsimonious concept of the model, where using fewer predictors can lead to a simpler model with equivalent prediction accuracy. Therefore, in this case, ANN would be a better prediction model than hybrid ANN. Cumulatively, the pre-processing of data with data mining techniques improved the prediction accuracy of the developed models. Although the prediction accuracy improvement was not significant for daily prediction models, the prediction accuracy improvement was significant for monthly streamflow predictions. Table 8 provides crucially beneficial importance degrees of the predictors which developed time dependent streamflow trend indices of $P5Y_i, P2Y_i, and\ PY_i$ improved prediction accuracy compared to the previous studies performed on the same case study. This suggests that prediction accuracy in ANNs depends on the optimal structure of ANNs, the intelligent selection of predictors, and the pre-processing of those predictors using supervised data mining techniques.

Developing the importance degrees of predictors provides an intelligent basis for optimal selection of predictors considering the availability of data on predictors. Finally, providing detailed correlation analysis and accuracy

performances of the developed daily streamflow models in Tables 3 along with developed importance degrees, as shown in Tables 5, 6, and 8, provides the valuable basis for developing diverse models for different periods of the year; the same utilized approach in this study is applicable and extendible for similar hydrological case studies.

6. Acknowledgment

The authors would like to thank the Stanford Engineering Research Center for Reinventing the Nation's Urban Water Infrastructure for partially supporting this research project.

7. References

[1] Abrahart, Robert J., and Linda See. "Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments."Hydrological processes 14, no. 11-12 (2000): 2157-2172. DOI: 10.1002/1099-1085(20000815/30)14:11/123.0.CO;2-S.

[2] Abudu, Shalamu, C. L. Cui, James Phillip King, and Kaiser Abudukadeer. "Comparison of performance of statistical models in forecasting monthly streamflow of Kizil River, China."Water Science and Engineering 3, no. 3 (2010): 269-281. DOI: 10.3882/j.issn.1674-2370.2010.03.003.

[3] Abudu, Shalamu, J. Phillip King, and A. Salim Bawazir. "Forecasting monthly streamflow of spring-summer runoff season in Rio Grande headwaters basin using stochastic hybrid modeling approach." Journal of Hydrologic Engineering 16, no. 4 (2010): 384-390. DOI: 10.1061/(ASCE)HE.1943-5584.0000322.

[4] Alizadeh, R., M. Majidpour, R. Maknoon, and J. Salimi. "Iranian energy and climate policies adaptation to the Kyoto protocol." International Journal of Environmental Research 9, no. 3 (2015): 853-864. DOI: 10.22059/ijer.2015.972.

[5] Alizadeh, Reza, Mehdi Majidpour, Reza Maknoon, and Saeed Shafiei Kaleibari. "Clean development mechanism in Iran: does it need a revival?." International journal of global warming 10, no. 1-3 (2016): 196-215. DOI: 10.1504/IJGW.2016.077913.

[6] Al-Jarrah, Omar Y., Paul D. Yoo, Sami Muhaidat, George K. Karagiannidis, and Kamal Taha. "Efficient machine learning for big data: A review." Big Data Research 2, no. 3 (2015): 87-93. DOI: 10.1016/j.bdr.2015.04.001.

[7] Chen, X. Y., K. W. Chau, and A. O. Busari. "A comparative study of population-based optimization algorithms for downstream river flow forecasting by a hybrid neural network model." Engineering Applications of Artificial Intelligence 46 (2015): 258-268. DOI: 10.1016/j.engappai.2015.09.010.

[8] Cheung, Sai On, C. M. Tam, and F. C. Harris. "Arbitration as an alternative dispute resolution method." *World Construction Conference, Global Challenges in Construction Industry*, (2012): 23–31.

[9] Damle, Chaitanya, and Ali Yalcin. "Flood prediction using time series data mining." Journal of Hydrology 333, no. 2-4 (2007): 305-316. DOI: 10.1016/j.jhydrol.2006.09.001.

[10] Dastourani, M. T., A. Habibipoor, M. R. Ekhtesasi, A. Talebi, and J. Mahjoobi. "Evaluation of the design tree model in precipitation prediction (case study: Yazd synoptic station)" (2013): 14-27.

[11] Faruk, Durdu Ömer. "A hybrid neural network and ARIMA model for water quality time series prediction." Engineering Applications of Artificial Intelligence 23, no. 4 (2010): 586-594. DOI: 10.1016/j.engappai.2009.09.015.

[12] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine 17, no. 3 (1996): 37. DOI: 10.1609/aimag.v17i3.1230.

[13] Jiang, Zhu, Hui-yan Wang, and Wen-wu Song. "Discharge estimation based on machine learning." Water Science and Engineering 6, no. 2 (2013): 145-152. DOI: 10.3882/j.issn.1674-2370.2013.02.003.

[14] He, Zhibin, Xiaohu Wen, Hu Liu, and Jun Du. "A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region." Journal of Hydrology 509 (2014): 379-386. DOI: 10.1016/j.jhydrol.2013.11.054.

[15] Humphrey, Greer B., Matthew S. Gibbs, Graeme C. Dandy, and Holger R. Maier. "A hybrid approach to monthly streamflow forecasting: integrating hydrological model outputs into a Bayesian artificial neural network." Journal of Hydrology 540 (2016): 623-640. DOI: 10.1016/j.jhydrol.2016.06.026.

[16] Kim, Tae-Woong, and Juan B. Valdés. "Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks." Journal of Hydrologic Engineering 8, no. 6 (2003): 319-328. DOI: 10.1061/(ASCE)1084-0699(2003)8:6(319).

[17] Kasiviswanathan, K. S., Jianxun He, K. P. Sudheer, and Joo-Hwa Tay. "Potential application of wavelet neural network ensemble to forecast streamflow for flood management." Journal of Hydrology 536 (2016): 161-173. DOI: 10.1016/j.jhydrol.2016.02.044.

[18] Labadie, John W. "Optimal operation of multireservoir systems: state-of-the-art review." Journal of water resources planning

and management 130, no. 2 (2004): 93-111. DOI: 10.1061/(ASCE)0733-9496(2004)130:2(93).

[19] Liao, Shu-Hsien, Pei-Hui Chu, and Pei-Yuan Hsiao. "Data mining techniques and applications–A decade review from 2000 to 2011." Expert systems with applications 39, no. 12 (2012): 11303-11311. DOI: 10.1016/j.eswa.2012.02.063.

[20] Moradkhani, Hamid, Kuo-lin Hsu, Hoshin V. Gupta, and Soroosh Sorooshian. "Improved streamflow forecasting using self-organizing radial basis function artificial neural networks." Journal of Hydrology 295, no. 1-4 (2004): 246-262. DOI: 10.1016/j.jhydrol.2004.03.027.

[21] Moreno, Jimmy, Shalamu Abudu, A. Salim Bawazir, and J. Phillip King. "Comment on 'Kişi Ö. 2009. Daily pan evaporation modelling using multi-layer perceptrons and radial basis neural networks". Hydrological Processes 24, no. 21 (2010): 3115-3118. DOI: 10.1002/hyp.7713.

[22] Sabzi, Hamed Zamani, James Phillip King, and Shalamu Abudu. "Developing an intelligent expert system for streamflow prediction, integrated in a dynamic decision support system for managing multiple reservoirs: A case study." Expert Systems with Applications 83 (2017): 145-163. DOI: 10.1016/j.eswa.2017.04.039.

[23] Sabzi, Hamed Zamani. Artificial intelligence and time series based forecasting in water resources, decision modeling, and optimal selection using ranking techniques. New Mexico State University, 2016.

[24] Saltelli, Andrea, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. Sensitivity analysis in practice: a guide to assessing scientific models. John Wiley & Sons, 2004. DOI:10.1002/0470870958.

[25] Stedinger, Jery R., Bola F. Sule, and Daniel P. Loucks. "Stochastic dynamic programming models for reservoir operation optimization." Water resources research 20, no. 11 (1984): 1499-1505. DOI: 10.1029/WR020i011p01499.

[26] Wei, You-xing, Deng-ting Wang, and Qing-jun Liu. "Application of artificial neural network to calculation of solitary wave run-up." Water Science and Engineering 3, no. 3 (2010): 304-312. DOI: :10.3882/j.issn.1674-2370.2010.03.006.

[27] Zamani Sabzi, Hamed, Shalamu Abudu, Reza Alizadeh, Leili Soltanisehat, Naci Dilekli, and James Phillip King. "Integration of Time Series Forecasting in a Dynamic Decision Support System for Multiple Reservoir Management to Conserve Water Sources."Energy Sources, Part A: Recovery, Utilization, and Environmental Effects Doi:10.1080/15567036.2018.1476934.