Title

Application of Data Mining Tools for Long-Term Quantitative and Qualitative Prediction of Streamflow

Permalink

https://escholarship.org/uc/item/2sj8k4mq

Journal

Journal of Irrigation and Drainage Engineering, 142(12)

ISSN

0733-9437

Authors

Mirzaei-Nodoushan, Fahimeh
Bozorg-Haddad, Omid
Fallah-Mehdipour, Elahe
et al.

Publication Date

2016-12-01

DOI

10.1061/(asce)ir.1943-4774.0001096

Peer reviewed

Application of Data Mining Tools for Long-Term Quantitative and Qualitative Prediction of Streamflow

Fahimeh Mirzaei-Nodoushan[1]; Omid Bozorg-Haddad[2];
Elahe Fallah-Mehdipour, Ph.D.[3]; and Hugo A. Loáiciga, F.ASCE[4]

Abstract: This paper evaluates the performances of two long-term prediction approaches for streamflow and riverine total dissolve solids (TDS) and compares their results with observed data and with short-term predicted values. The future values predicted by the first, long-term, prediction approach (Approach 1) depend on data corresponding to time steps prior to the prediction time step. The future values predicted by the second, long-term, prediction approach (Approach 2) depend on data comprised within the observational period. Each long-term prediction approach calculates streamflow and TDS over a 12-month period ranging from April through March (Scheme 1) and by agricultural water year (December through November, Scheme 2). Genetic programming (GP) is implemented for long-term prediction. Prediction is applied to the streamflow and TDS of the Karoon River in southwestern Iran. The long-term Approach 1 was found to be more accurate than the long-term Approach 2 judged by the values of several diagnostic statistics. The root mean square error (*RMSE*), correlation coefficient (R^2), and Nash-Sutcliffe efficiency (E) statistics of long-term predictions of streamflow and TDS with Approach 1 are lower than those obtained with the long-term prediction Approach 2 for April–March and for the agricultural water-year predictions. It is concluded that prediction of the Karoon River's streamflow and TDS is best accomplished using GP in combination with the long-term prediction Approach 1. **DOI: 10.1061/(ASCE)IR.1943-4774.0001096.** *© 2016 American Society of Civil Engineers.*

Author keywords: Short-term and long-term prediction; Streamflow; Total dissolved solids (TDS); Genetic programming.

Introduction

Water is a vital factor inasmuch as any variation in its quantity and quality affects other resources such as food, energy, wildlife, and forests. Prediction of hydrologic processes such as precipitation, evaporation, and runoff plays an important role in many activities associated with the planning and operation of water resource systems. Several time series approaches have been employed for the prediction of hydrologic processes. Time series prediction models include: (1) statistical-based models using statistical concepts, such as autoregressive (AR), autoregressive moving average (ARMA), and autoregressive integrated moving average (ARIMA) (e.g., Yu and Tseng 1996; Kothyari and Singh 1999); and (2) models based on artificial intelligence (AI) that use simulation and learning time series patterns, such as artificial neural network (ANN), support vectors machine (SVM), adaptive neural-based fuzzy inference system (ANFIS), and genetic programming (GP) (e.g., Elshorbagy et al. 2002; Nagesh Kumar et al. 2004; Yoon et al. 2011).

Data mining, which is a method of data processing and a branch of AI, explores patterns and relations among data by using computerized recognition and analysis algorithms. Fu (2011) reported a general perspective on the development of time series and data mining. In addition, Liao et al. (2012) reviewed data-mining techniques, indicating that the development of those techniques is mainly expertise-oriented while their applications are primarily problem-centered.

Another data-mining technique used in prediction of time series is GP, which is an evolutionary computational method based on random search and a subset of genetic algorithm (GA). Savic et al. (1999) introduced GP to rainfall-runoff modeling and compared results with those of two optimally calibrated conceptual models and ANN. Results showed the superiority of GP with respect to ANN. Additional studies employing GP addressed real-time runoff forecasting (Khu et al. 2001) and determination of a basin's unit hydrograph (Rabuñal et al. 2007). Makkeasorn et al. (2008) compared ANN and GP models predicting short-term streamflow considering climate change. Charhate et al. (2009) introduced five separate GP models to predict streamflow and highlighted the high efficiency of the GP models in predicting streamflow peaks and GP's extrapolation capability, which is lacking in ANN models. Wang et al. (2009) applied ARMA, ANNs, ANFIS, GP, and SVM methods comparatively for monthly prediction of river flow discharges and confirmed the superior performance of ANFIS, GP, and SVM. Ni et al. (2010) modeled the relation between streamflow and the impact of climate change in China by employing the GP technique. They compared its results with three statistical methods and justified the better capability of GP compared with those of other methods. The results also indicated the adequacy of GP for estimating the effects of climate change on streamflow when large data sets are not available. Izadifar and Elshorbagy (2010)

[1] M.Sc. Graduate, Dept. of Irrigation and Reclamation Engineering, Faculty of Agricultural Engineering and Technology, College of Agriculture and Natural Resources, Univ. of Tehran, Karaj, 3158777871 Tehran, Iran. E-mail: Fhmnodoushan@ut.ac.ir

[2] Associate Professor, Dept. of Irrigation and Reclamation Engineering, Faculty of Agricultural Engineering and Technology, College of Agriculture and Natural Resources, Univ. of Tehran, Karaj, 3158777871 Tehran, Iran (corresponding author). E-mail: OBHaddad@ut.ac.ir

[3] Postdoctoral Researcher, Dept. of Irrigation and Reclamation Engineering, Faculty of Agricultural Engineering and Technology, College of Agriculture and Natural Resources, Univ. of Tehran, Karaj, 3158777871 Tehran, Iran. E-mail: Falah@ut.ac.ir

[4] Professor, Dept. of Geography, Univ. of California, Santa Barbara, CA 93106. E-mail: Hugo.Loaiciga@ucsb.edu

Note. This manuscript was submitted on January 28, 2016; approved on May 10, 2016; published online on July 18, 2016. Discussion period open until December 18, 2016; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Irrigation and Drainage Engineering***, © ASCE, ISSN 0733-9437.

implemented techniques including ANN, GP, and multiple regression to model hourly actual evapotranspiration (AET) with meteorological variables. They also employed the *HYDRUS-1D* model for AET estimation, and the results were compared with those of the latter three methods. The best solution was generated by GP, and the authors indicated that data-driven models in such predictions might be superior to a physically based model like *HYDRUS-1D*.

Nasseri et al. (2011) implemented a hybrid model combining an extended Kalman filter (EKF) and GP to forecast monthly water demand. Results attributed the notable impact of observation accuracy on water demand prediction, which could help to reduce the risks of online water demand forecasting and optimal operation of urban water systems. Sreekanth and Datta (2011) conducted a comparison between ANN and GP to calculate optimal groundwater extraction rates. Their results showed that the GP is more advantageous than the ANN due to several reasons such as simplicity and having fewer parameters and higher efficiency in achieving optimized structure. Fallah-Mehdipour et al. (2013d) applied GP as a hydrologic method instead of using hydraulic flow routing methods, which require a large number of data to calculate routed stage hydrograph in simple and compound channels. They found GP a capable method with fewer data and lower cost than other models and demonstrated the accuracy of GP. Orouji et al. (2014) compared two hydrologic methods based on an extended version of the Muskingum method and GP for attaining a routed flood hydrograph in a branched river. They concluded that GP, in addition to being more effective with excellent performance in hydrograph routing in branched rivers, is easier to use and needs fewer input data. Havlíček et al. (2013) attempted to improve rainfall-runoff forecasts by a prediction method combining GP and basic hydrological modeling concepts. They compared results with the ANN model, and the GP model proved its excellent performance. Fallah-Mehdipour et al. (2013a) extracted optimal operational decision rules employing GP and compared results with those of common linear and nonlinear decision rules. Their results demonstrated that the objective function value improved considerably for both the training and testing data when using GP. Those authors also reported the effectiveness of the proposed rule based on GP for optimizing the rule curves of reservoirs. Fallah-Mehdipour et al. (2013b) developed a fixed-length gene GP (FLGGP) rule, which computed a more effective operation rule to calculate a better objective function value than that obtained with the GA in an aquifer-dam system. Fallah-Mehdipour et al. (2013c) probed the sufficiency of ANFIS and GP for predicting and simulating groundwater levels. Their results found GP a more effective tool than ANFIS to determine groundwater levels. Several pieces of research dealing with prediction applications have demonstrated that GP has better capabilities than many statistical and AI methods in predicting hydrologic processes, in particular ANN, which is commonly used for that purpose.

Many studies have been reported concerning water quality prediction using different tools. Ahmad et al. (2001) compared three stochastic modeling approaches accounting for the effect of seasonality with the multiplicative ARIMA model, a deseasonalized model and the Thomas–Fiering model to forecast riverine water quality. The deseasonalized model was recommended to forecast riverine water quality parameters. Chau (2006) reviewed the use of artificial intelligence in water quality modeling. Palani et al. (2008) applied ANN to predict water quality variables at various locations. Results showed high accuracy of ANN simulation in modeling water quality variables where the available data set is limited. Liu et al. (2011) presented real-value GA support vector regression (RGA-SVR) as a hybrid approach to solve aqua cultural water quality prediction. They found excellent performance of their method compared to the traditional SVR and back-propagation (BP) neural network models. Tan et al. (2012) investigated the prediction of water-quality variables with a least squares SVM (LS-SVM) model and compared its results with BP and radial basis function (RBF) neural network prediction. They found better performance of the LS-SVM than those of the two other methods. Xu and Liu (2013) introduced the wavelet neural network model to predict water quality and compared it with the BP neural network and the Elman neural network. The wavelet neural network model had faster learning, better prediction accuracy than those of the other models, and high robustness. Orouji et al. (2013) compared the ANFIS and GP data-mining methods to predict water quality. Their results demonstrated that prediction of water quality with GP to be more efficient than that of the ANFIS model. Evidently, water-quality prediction has received much attention by the hydrologic community.

Several quantitative and qualitative water resources pieces of research have been conducted to manage future water supply (Ashofteh et al. 2013, 2015b, a, c; Beygi et al. 2014; Bozorg-Haddad et al. 2013, 2014, 2015b, a; Bolouri-Yazdeli et al. 2014; Orouji et al. 2014; Shokri et al. 2013, 2014; Soltanjalili et al. 2013). However, such studies usually rely on historical data, and long-term inflow prediction is neglected. Generally, predictions of hydrologic time series are performed with short-term and long-term approaches. The short-term approach predicts event values over a short horizon. The long-term prediction approach, on the other hand, predicts hydrologic variables over extended periods that are pertinent to water resources applications, such as the operation of water supply utilities, optimal reservoir operation, environment protection, and drought management. The main objectives of this study are assessing two long-term prediction approaches for streamflow and riverine water quality and modeling and comparing results with those obtained with short-term prediction. This work implements long-term prediction Approaches 1 and 2. Long-term Approach 1 uses as predictor variables of future values one or several values observed or predicted in previous time steps. Long-term Approach 2's predictions, on the other hand, depend on one or several values of the presently available (or current) time series. The root mean square error (*RMSE*), the correlation coefficient (R^2), and Nash-Sutcliffe efficiency (E) statistical performance indices are herein employed to determine the performance of the implemented prediction approaches. Streamflow was chosen for hydrologic quantitative prediction. TDS was selected as the predicted water-quality variable. TDS is widely used to characterize the quality of water for municipal, industrial, and agricultural uses.

Tools and Approaches

Many models, including statistical and AI methods, have been proposed for hydrologic time series prediction. AI tools, being capable of analyzing long-series and large-scale data, have been frequently used in water resources studies. A variant of AI, namely GP, is employed in this work for the prediction of streamflow and TDS in the Karoon River of Iran.

Genetic Programming

GP is an evolutionary computation technique based on random search, and a variant of the GA, which finds solutions to optimization problems by generating logical and mathematical expressions. GP searches a problem solution space to find a solution that best fits observation data. GP has a tree structure with nodes and branches, where each node implies an operator, variable, or number, and each branch describes the links between nodes. GP

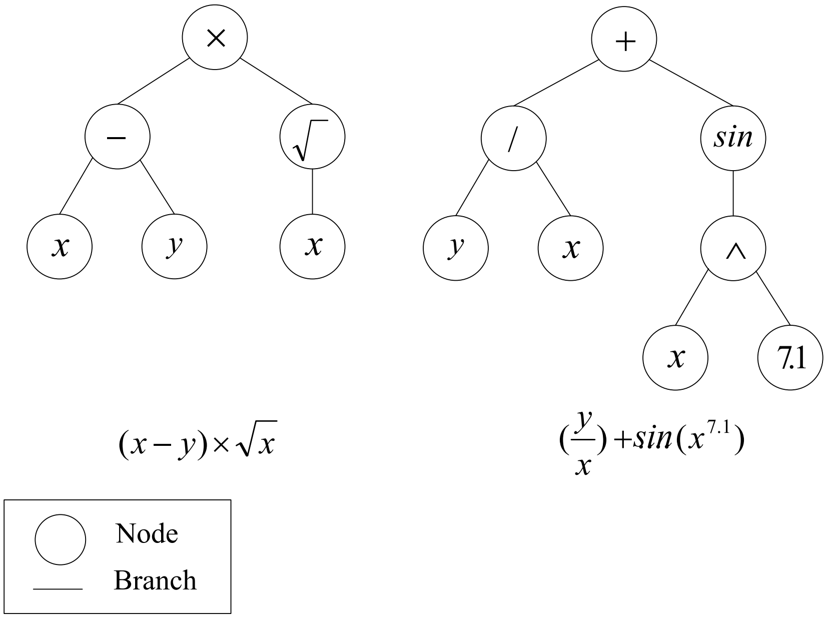

Fig. 1. GP structure and relation between variables

can construct functions between inputs and output, a capability that sets it apart advantageously with respect to AI methods, such as ANN and SVM. A function set involves operators such as arithmetic operators (+, −, ×, /), mathematical functions (e.g., *sin*, *cos*), logical expressions (e.g., if-then-else), Boolean operators (e.g., and, or) and also random numbers to construct an optimal function between inputs and output (Fig. 1). GP is provided with two sets of data: (1) input (independent) data, and (2) output (dependent) data. GP divides data in two parts: training data to find patterns in the observation data and testing data to examine patterns from extracted training data. The basic procedure of GP is summarized in the following steps:

1. Generating a set of random initial individuals (trees);
2. Determining objective functions of individuals (typically, the error between the estimated and observed data);
3. Assessing stopping criteria (number of iterations, runtime, error value or number of evaluations of objective function);
4. Selecting superior trees using techniques such as the roulette wheel, tournament, or ranking method;
5. Applying genetic operators (crossover and mutation) and generating new trees of new generation; and
6. Returning to Step 3 to proceed with the iterative search for the solution of optimization problems.

This process is repeated until a stopping criterion is fulfilled.

In crossover, some subtrees of two selected trees are randomly chosen, and two new trees are created by replacing subtrees from parents (Fig. 2). In mutation, one or more random nodes according to mutation probability are selected and changed with another random operator, variable, or number, and new trees are produced (Fig. 3).

GP relies on random optimization. Therefore, it converges to a different solution each time it is run starting with a different initial random population of solutions. For this reason, several runs must be made, sufficient in number to determine minimum, maximum, average, standard deviation, and coefficient of variation of the objective function. A low coefficient of variation of the GP solution is a sign of convergence to a global optimum.

Prediction Approaches

The prediction of hydrologic variables is most useful in water resources planning and management. Short-term predictions have

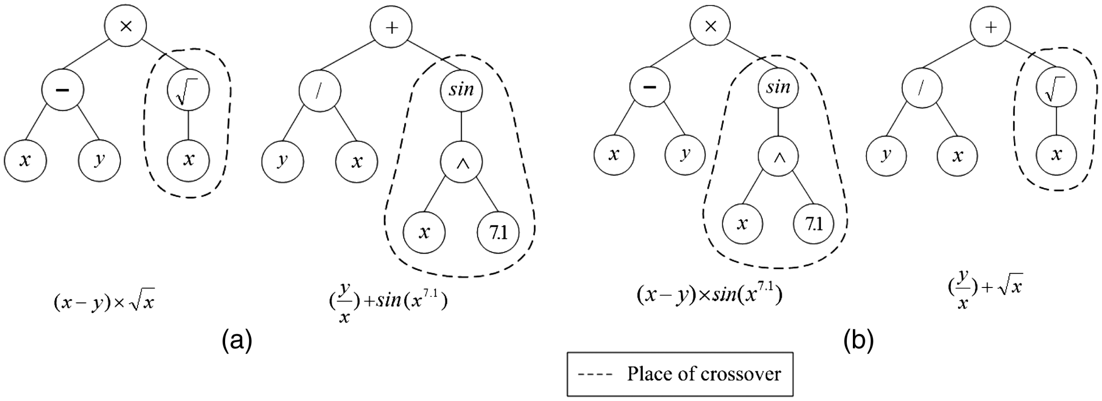

Fig. 2. Tree structure of GP: (a) before; (b) after crossover

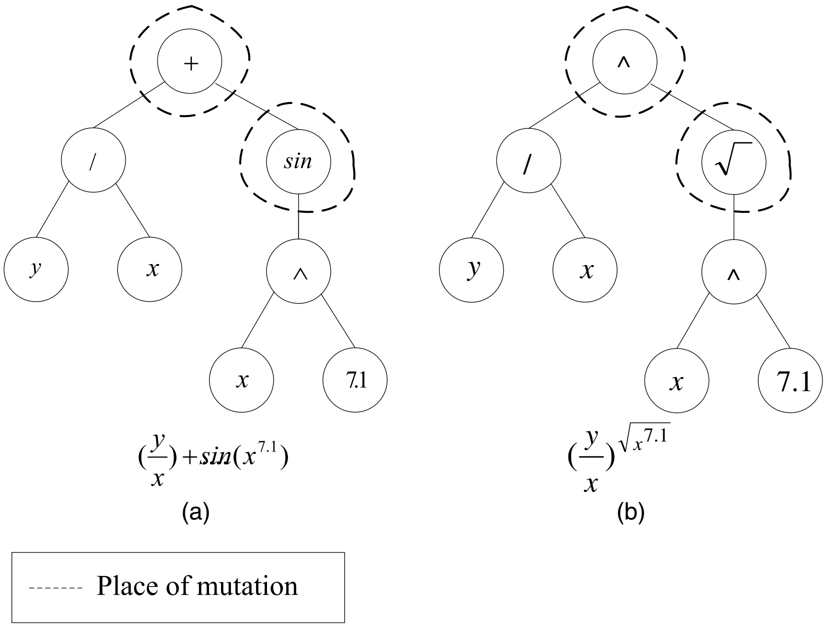
+
/
sin
y
x
^
x
7.1
$(\frac{y}{x})+sin(x^{7.1})$
(a)
^
/
√
y
x
^
x
7.1
$(\frac{y}{x})^{\sqrt{x^{7.1}}}$
(b)
Place of mutation

Fig. 3. Tree structure of GP: (a) before; (b) after mutation

been common (e.g., Smith et al. 1985; Billings and Agthe 1998; Liu et al. 2002; Bazartseren et al. 2003; Nayak et al. 2005; Gato et al. 2007; Makkeasorn et al. 2008; Xu and Liu 2013) but are somewhat limited in scope for many planning purposes that require relatively long-term forecasts. Long-term approach, on the other hand, predicts event values over many periods into the future. Two long-term Approaches 1 and 2 applied in this work are defined as follows:

1. Long-term Approach 1: In this approach the value of a predicted variable in a future time step is calculated using the values of the variable in one or several previous time steps (month or season). According to Fig. 4(a) the current time series includes time steps for which there are available data, and the future time series includes steps in which data are predicted. To illustrate, let t be the present time step in the current time series, and the number of input data be equal to $n+1$ time steps. The prediction of the value in time step $t+1$ in the future time series relies on the input data (current time series) of event values in time step $t-n$ to t, and the output (predicted) data is the event value in time step $t+1$. Furthermore, the prediction of the event value in time step $t+2$ relies on the input data (current time series) of event values in time steps $t-n+1$ to $t+1$, and the event value of time step $t+2$ is the output data (future time series), and so on and so forth. Therefore $n+1$ previous time steps are considered as input data in making future predictions according to this scheme. Each time a prediction is made for the next future step, that prediction becomes a predictor variable for the next future prediction.
2. Long-term Approach 2: In this approach, event values for all time steps in the future time series are predicted using data from the last step or several previous time steps in the present time series. It is seen in Fig. 4(b) that the prediction of values in future time steps $t+1$ through $t+n'$ relies on the event values for time step $t-n$ to t.

The results obtained from the two long-term prediction approaches were compared with observation data and with the results of the short-term prediction approach, described next.

3. Short-term approach: The procedure for this approach is similar to that of the long-term prediction Approach 1 except that short-term prediction is based entirely on observed input data instead of predicted data. In other words, the prediction of the future value in time step $t+1$ is based on observed values up to time step t; subsequently, the prediction of the future value in time step $t+2$ is based on observed values up to time step $t+1$, and so on.

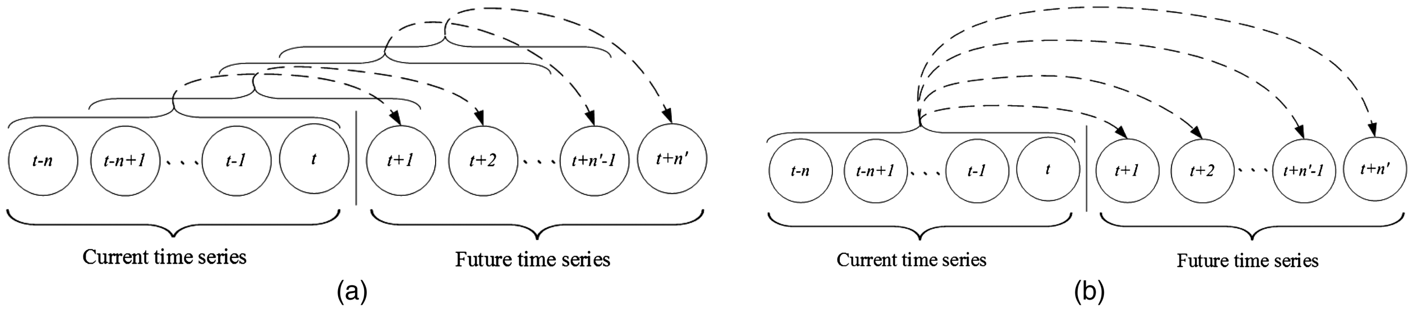

Fig. 4. Prediction approaches: (a) short-term; (b) long-term

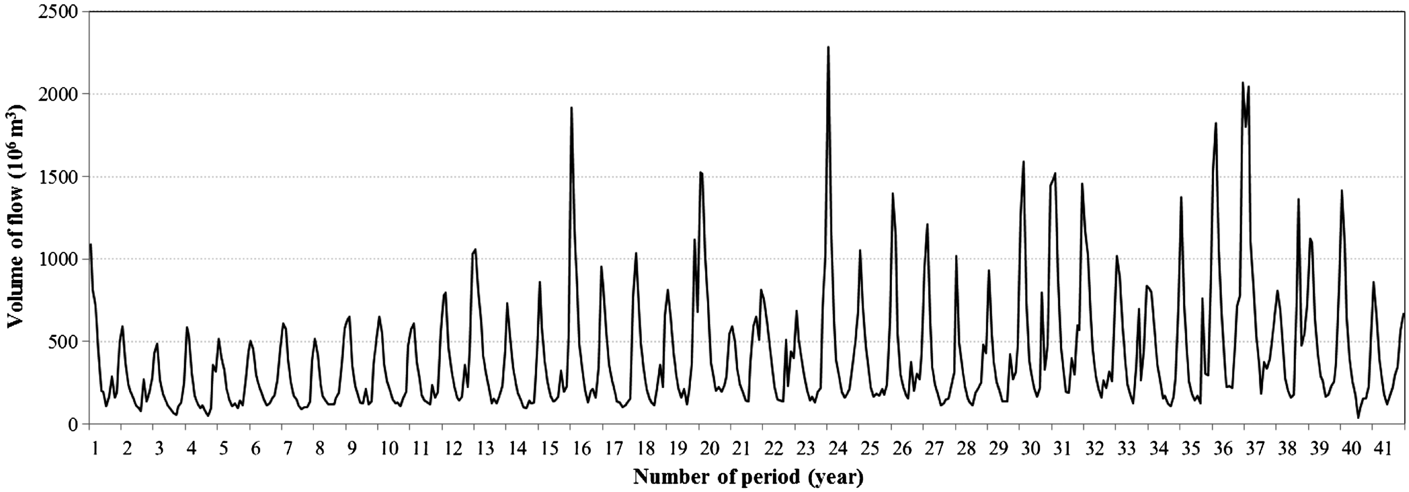

Fig. 5. Time series of monthly inflow for the Karoon 4 reservoir

Performance Measures

Several techniques are herein recommended to assess the performance of the prediction approaches by comparing observation data with estimated data. Three performance measures used in this study are computed as follows:

1. Root Mean Square Error

   The *RMSE* is defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - x_i)^2}{N}} \tag{1}$$

   where i = index for data values in time steps $i = 1, 2, \ldots, N$, x_i = ith observed data; y_i = ith estimated data; and N = total number of data values. The larger the *RMSE*, the poorer the predictive skill of a prediction approach.

2. Correlation coefficient

   R^2 describes the degree of statistical association between two variables. It is defined as

$$R^2 = \left[\frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}\right]^2 \tag{2}$$

   where $\bar{x}$ = average of observed data; and $\bar{y}$ = average of estimated data. This coefficient ranges from 0 to 1. There is a perfect positive or negative statistical association between the observed and estimated data when $R^2 = 1$. There is no statistical association between the variables x and y when R^2 equals 0.

3. Nash-Sutcliffe Efficiency

   The coefficient is a statistic that determines the relative magnitude of the residual variance compared to the estimated data. Its formula is

$$E = 1 - \frac{\sum_{i=1}^{N}(y_i - x_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{3}$$

   E ranges between 0 and 1. A value of 1 corresponds to a perfect fit between estimated and observed data. A value of 0 indicates that predictions equal the mean value of the observed data.

Case Study

Data from two stations on the Karoon River were used to model river streamflow and TDS. The Karoon River is Iran's longest river at 950 km long and the largest in the same country with an average discharge equal to 575 m^3/s. Reservoirs on the Karoon River serve flood control and power generation functions. A 41-year-long time series (1957–1997) of monthly inflow were used to predict streamflow (Fig. 5). A 34-year-long time series (1969–2002) of Karoon River's seasonal values of TDS at the Godarlandar station were used for water-quality modeling (Fig. 6).

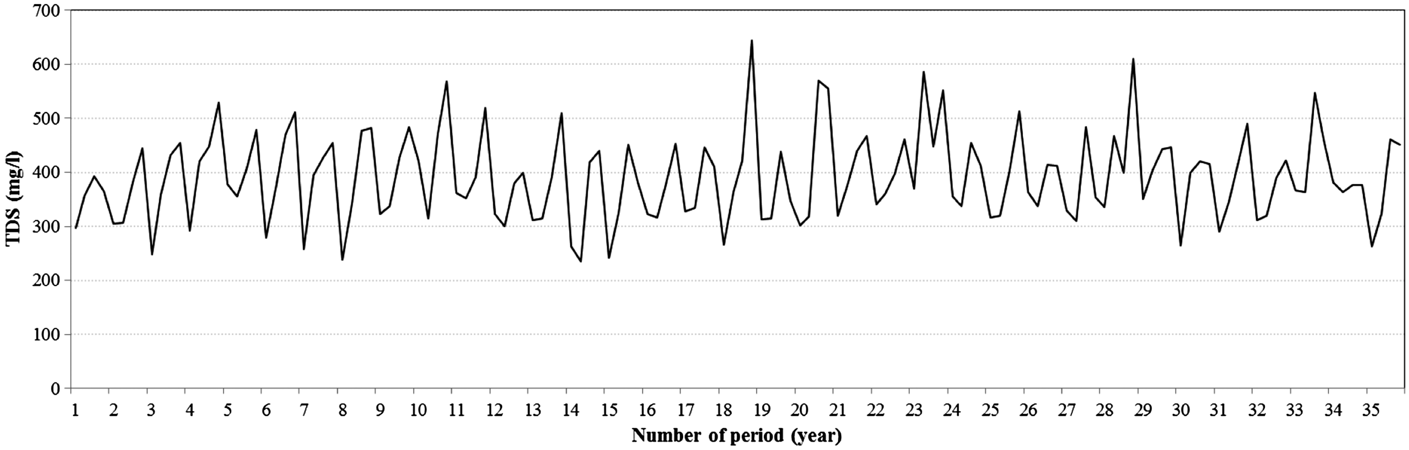

Fig. 6. Time series of seasonal TDS values at the Godarlandar station

Table 1. Statistical Values of the Objective Function after the Last Iteration from 10 Runs to Predict December Streamflow and Summer TDS with Scheme 1 of Approach 1

	Number of run														
Prediction	1	2	3	4	5	6	7	8	9	10	Minimum (mg/L)	Average (mg/L)	Maximum (mg/L)	Standard deviation (mg/L)	Coefficient of variation
December streamflow (RMSE)	28.32	28.83	27.68	25.08	34.87	29.82	48.12	31.22	35.58	26.05	25.08	31.56	48.12	6.75	0.21
Summer TDS (RMSE)	52.64	46.93	44.95	55.44	62.50	56.94	53.57	44.64	56.52	51.30	44.64	52.54	62.50	5.74	0.11

Table 2. Streamflow Prediction Performance Statistics

		Performance statistics					
		Training			Testing		
Scheme	Approach	*RMSE* (10^6m^3)	R^2	*E*	*RMSE* (10^6m^3)	R^2	*E*
1	Short-term	**774.09**	**0.771**	**0.767**	**862.25**	**0.696**	**0.672**
	Long-term 1	**774.09**	**0.771**	**0.767**	1,044.18	0.586	0.520
	Long-term 2	847.57	0.727	0.720	1,103.78	0.556	0.463
2	Short-term	**774.09**	**0.771**	**0.767**	**862.25**	**0.696**	**0.672**
	Long-term 1	**774.09**	**0.771**	**0.767**	1,003.41	0.695	0.556
	Long-term 2	948.83	0.650	0.649	1,156.27	0.527	0.411

Note: Bold values indicate the best values in each column.

The total data set was divided into two parts by considering 70% and 30% for training (calibration) and testing purposes, respectively. The training set includes a 29-year-long streamflow time series (1957–1985) and a 24-year TDS time series (1969–1992). It is essential in long-term prediction to choose the predictor variables. This study considers two 12-month periods for prediction purposes, namely April–March (Scheme 1) and an agricultural water year comprising months December–November (Scheme 2). The aim of considering these schemes is to investigate the impact of wet (Scheme 1) and dry (Scheme 2) seasons data on prediction. The input data (independent or predictor variables) used for long-term approach are the months of January, February, and March for Scheme 1 and July, August, and September for Scheme 2.

The GP runs involved 800 iterations and 10 trees. Applied operators in GP include +, −, /, ^, $\sqrt{}$, *sin*, and *cos*. The optimizing functions minimize the *RMSE*. The R^2 and *E* were calculated, also, to compare the performance of the implemented prediction approaches.

Results and Discussion

This study assessed the capability of two long-term approaches in predicting streamflow and TDS with the GP method. Ten runs of GP were performed to acquire multiple predictions of streamflow and TDS. Minimum, maximum, average, standard deviation, and coefficient of variation of the objective function for Scheme 1 in Approach 1 were calculated from the runs for streamflow and TDS and are listed in Table 1. The coefficients of variation (CVs) of the objective functions predicting streamflow and TDS were 0.21 and 0.11, respectively. These CVs are deemed acceptable, and their low values demonstrate a high reliability of GP to achieve accurate predictions.

Results of Streamflow Prediction

The *RMSE*, R^2, and *E* were calculated for the training phase and testing phase of the two long-term approaches and the short-term approach (Table 2). Figs. 7 and 8 present the results of streamflow training for two 12-month schemes. Results for the testing phase are portrayed in Figs. 9 and 10.

The calculated streamflow results shown in Table 2 indicate that the short-term approach obtained more accurate results than the two long-term approaches based on its *RMSE*, R^2, and *E* equal to 774.09, 0.771, and 0.767 in the training phase and to 862.25,

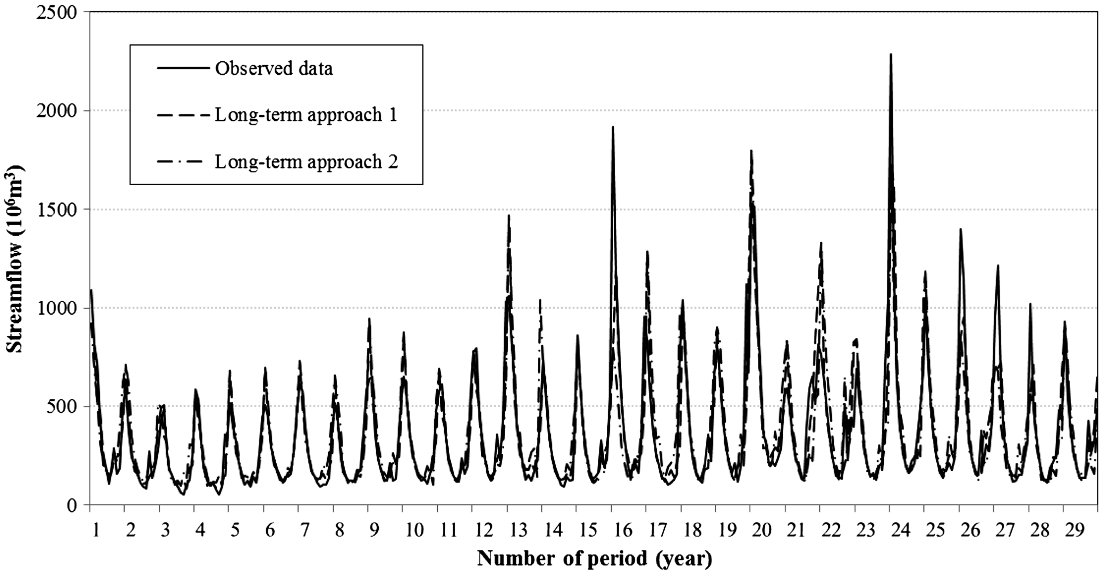

Fig. 7. Streamflow data used for training in Scheme 1

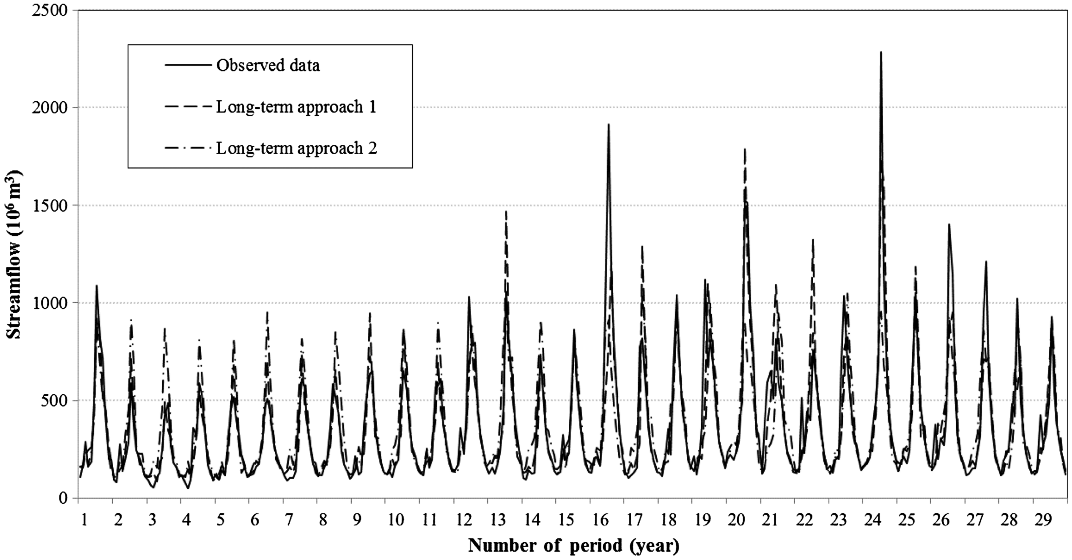

Fig. 8. Streamflow data used for training in Scheme 2

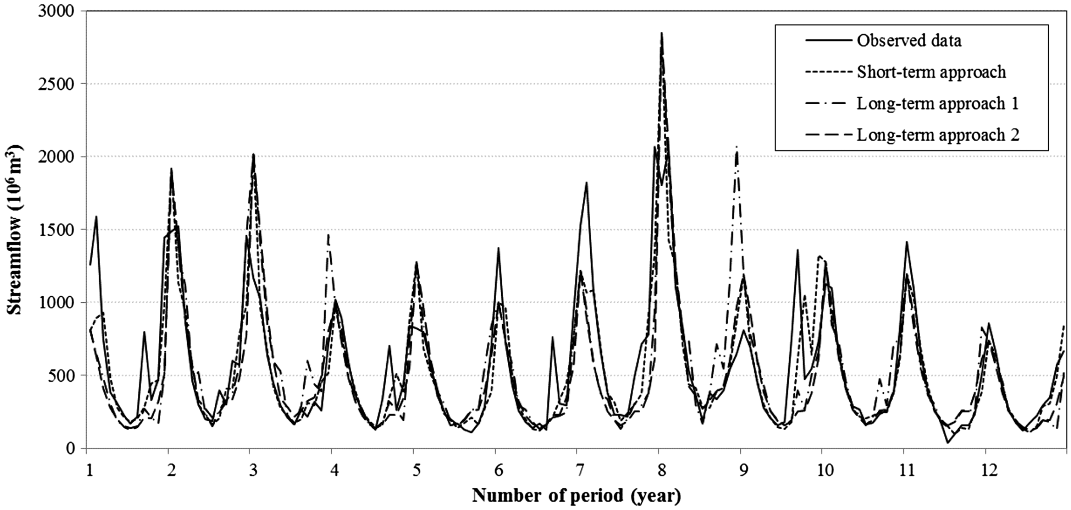

Fig. 9. Streamflow used for testing in Scheme 1

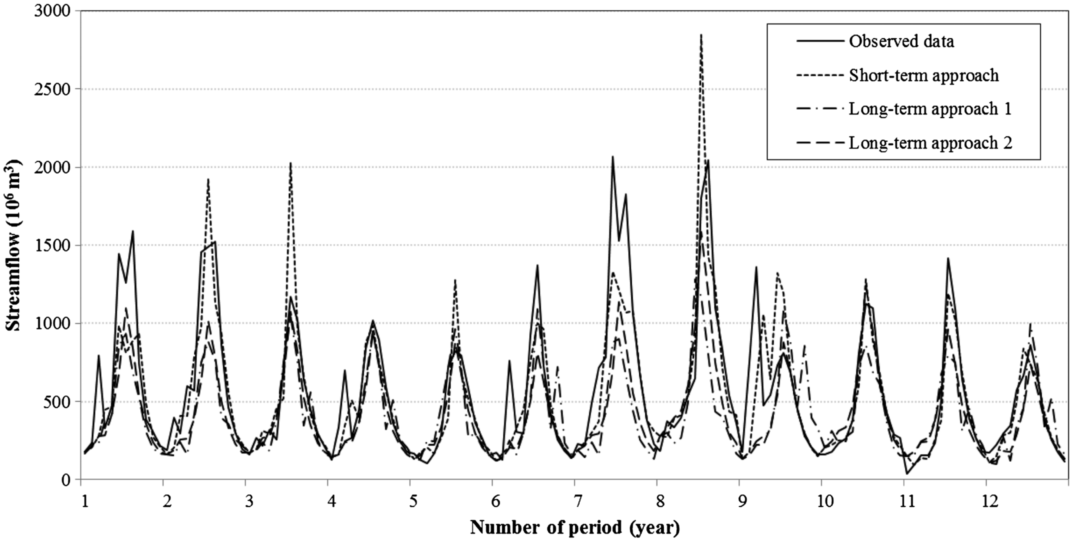

Fig. 10. Streamflow used for testing in Scheme 2

Table 3. TDS Prediction Performance Statistics

		Performance statistics					
		Training			Testing		
Scheme	Approach	*RMSE* (mg/L)	R^2	*E*	*RMSE* (mg/L)	R^2	*E*
1	Short-term	**157.01**	**0.586**	**0.577**	**91.20**	**0.523**	**0.510**
	Long-term 1	**157.01**	**0.586**	**0.577**	98.02	0.466	0.434
	Long-term 2	160.86	0.560	0.556	104.22	0.449	0.360
2	Short-term	**157.01**	**0.586**	**0.577**	**91.20**	**0.523**	**0.510**
	Long-term 1	**157.01**	**0.586**	**0.577**	100.10	0.450	0.410
	Long-term 2	171.27	0.530	0.496	113.66	0.424	0.239

Note: Bold values indicate the best values in each column.

0.696, and 0.672 in the testing phase for both duration schemes. This result was predictable because of the near-term prediction nature of the short-term approach. Yet, the main purpose of this study is to compare the two long-term prediction approaches because of their relevance to water-resources management. The results in Table 2 indicate that the long-term Approach 1 is more accurate than the long-term Approach 2 according to the *RMSE*, R^2, and *E* statistics obtained with Schemes 1 and 2, which means it is independent to type of input data (dry or wet season).

On the other hand, the long-term Approach 1 using dry season input data (Scheme 2) has better results than wet season input data (Scheme 1). Also, better prediction in the long-term Approach 2 was by wet season input data (Scheme 1).

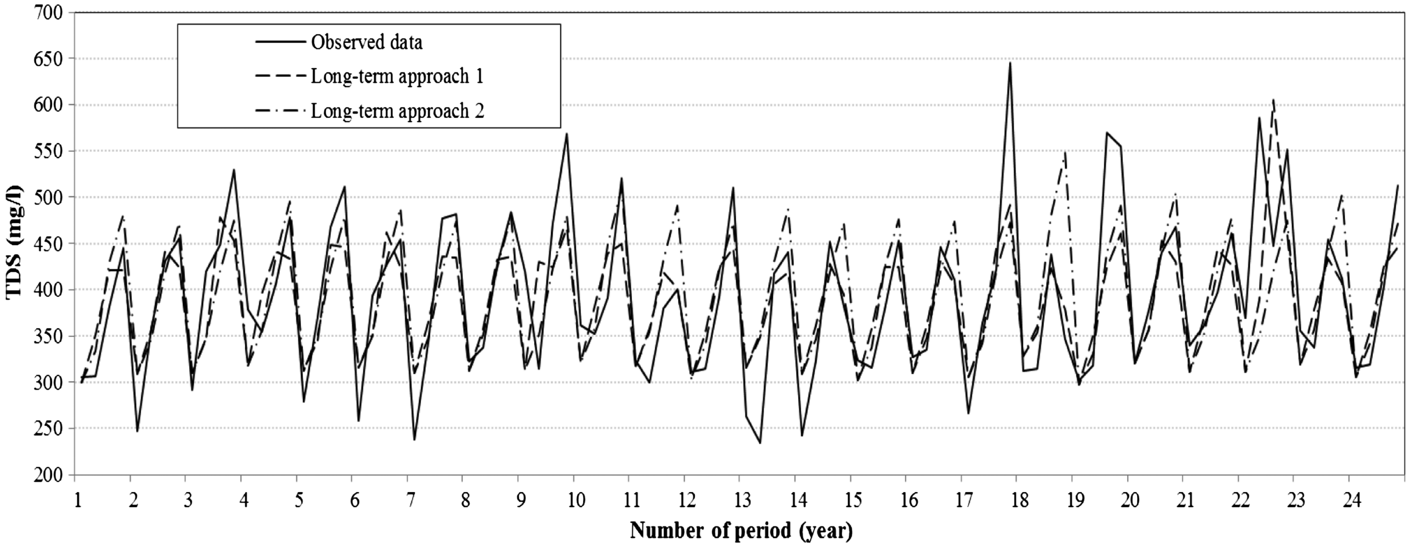

Fig. 11. TDS data used for training in Scheme 1

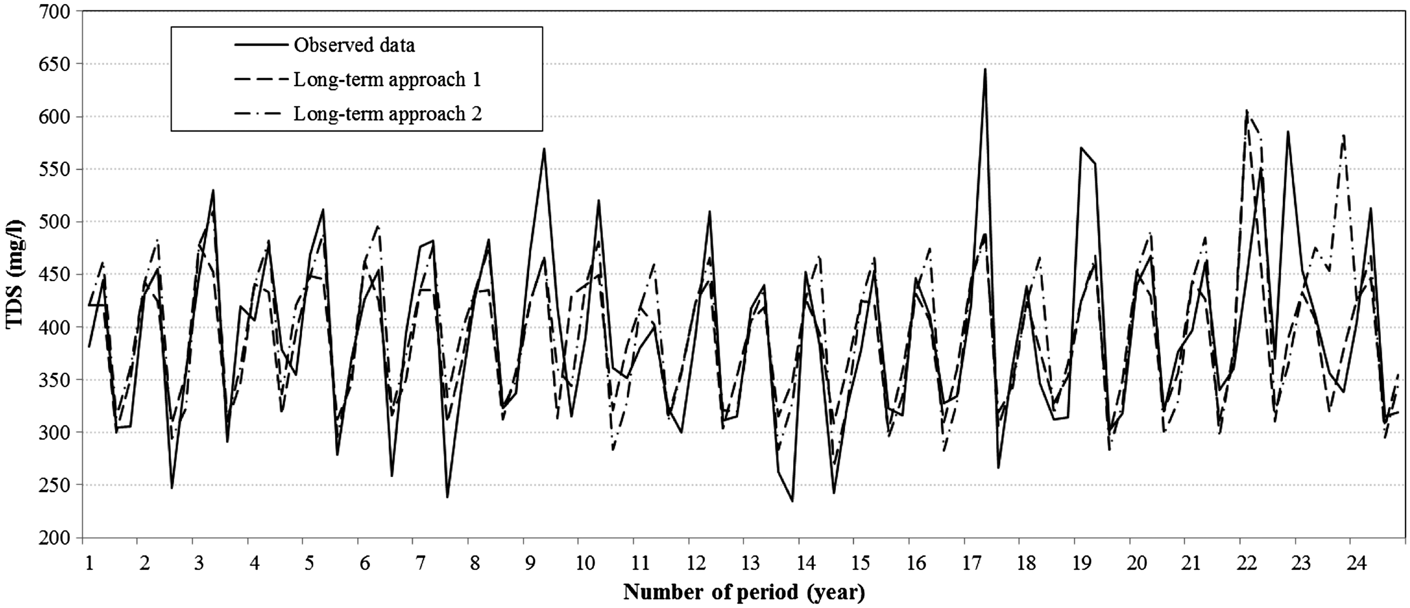

Fig. 12. TDS used for training in Scheme 2

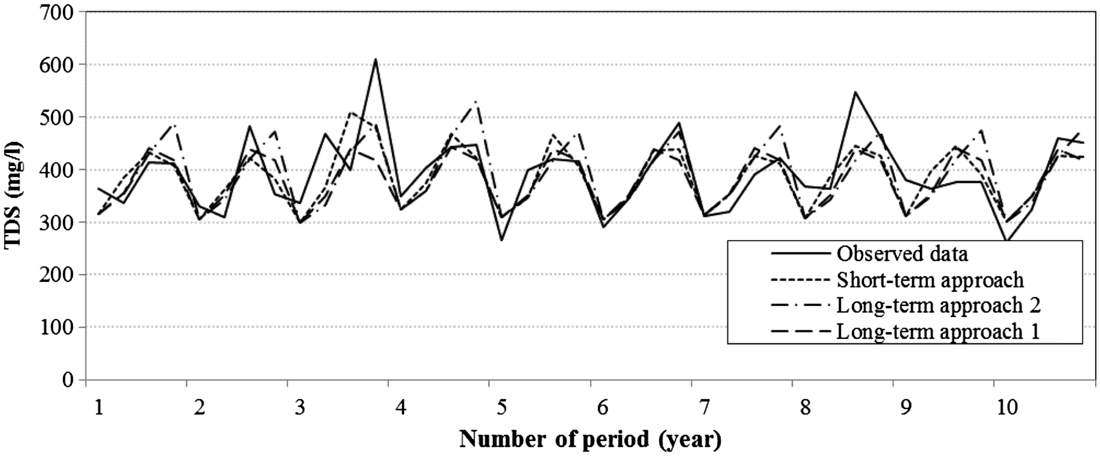

Fig. 13. TDS data used for testing in Scheme 1

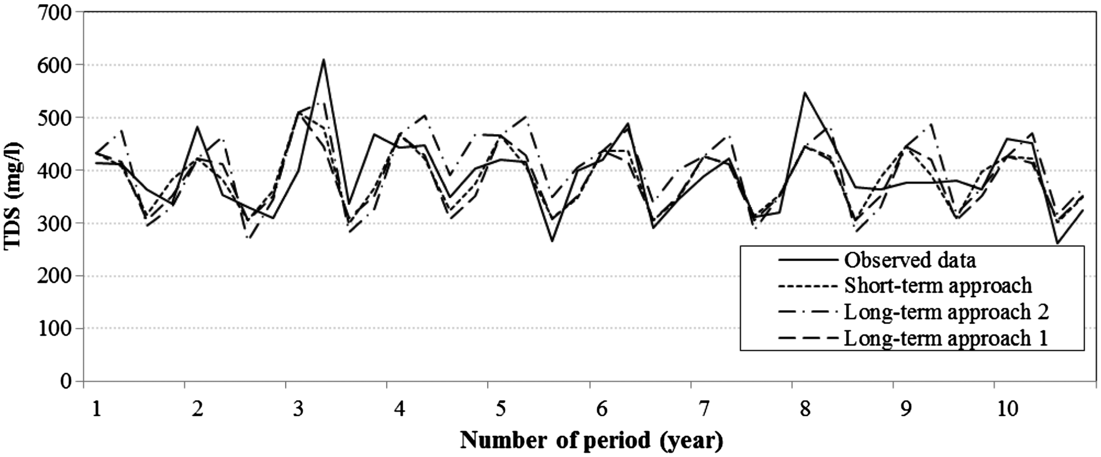

Fig. 14. TDS data used for testing in Scheme 2

In general, both long-term approaches predicted streamflow with acceptable accuracy and produced results close to those of the short-term approach. Therefore, the two long-term prediction approaches with GP predict future streamflow accurately, although the long-term prediction Approach 1 proved to be the most accurate one of the two.

Results of TDS Prediction

The diagnostic statistics *RMSE*, R^2, and *E* were calculated to assess the training and testing performances of the prediction approaches and are shown in Table 3. The training results for TDS for the two duration schemes are depicted in Figs. 11 and 12, and the testing results are presented in Figs. 13 and 14.

The results listed in Table 3 established that the short-term predictions of TDS yielded a *RMSE*, R^2, and *E* equaled 157.01, 0.586, and 0.577 in the training phase and 91.20, 0.523 and 0.510 in the testing phase. Overall, the short-term prediction approach for TDS performed better than the two long-term prediction approaches. Comparing the two long-term approaches, regardless of type of input data (dry season or wet season), Approach 1 outperformed Approach 2 based on the diagnostic statistics *RMSE*, R^2, and *E* for these two approaches with Schemes 1 and 2.

In addition, both long-term Approach 1 and 2 prediction applying wet season input data (Scheme 1) are more accurate than using dry season input data (Scheme 2).

Our results indicated that the long-term prediction approaches predicted riverine TDS accurately, yet, the long-term prediction Approach 1 proved to be more accurate than Approach 2.

Concluding Remarks

This study assessed two long-term prediction approaches and one short-term prediction approach for streamflow and TDS. The long-term prediction Approach 1 calculates future values with data corresponding to previous time steps, either from the observational time series or from the predicted time series. The long-term prediction Approach 2 calculates future event values with data from the observational period. Monthly observed data and TDS seasonal data were used in conjunction with GP to calculate the streamflow and TDS predictions. Performance measures *RMSE*, R^2, and *E* were computed to compare observed and predicted data.

TDS was selected as a parameter with independent variations to streamflow variations to assess two long-term approaches more accurately using two independent parameters of river. The short-term approach obtained more accurate results than the long-term approaches in predicting streamflow and TDS. In streamflow prediction, the long-term Approach 1 yielded more accurate results than long-term Approach 2, the former approach's *RMSE*, R^2, and *E* being 6%, 5%, and 11% better than those of the latter approach with the first scheme, and 15%, 24%, and 26% superior to the second scheme. The long-term prediction Approach 1 proved more accurate than the long-term prediction Approach 2 in predicting TDS, also, with the former approach's *RMSE*, R^2, and *E* being 6%, 3%, and 17% superior to those of the latter approach with the first scheme, and 14%, 6%, and 42% better with Scheme 2.

Concerning streamflow prediction, the long-term Approach 1 using dry season input data (Scheme 2) is more accurate than with wet season input data (Scheme 1), the former scheme's *RMSE*, R^2, and *E* being 4%, 16%, and 7% better than those of the latter scheme. In the long-term Approach 2 yielded more accurate results by wet season input data (Scheme 1), with the former scheme's *RMSE*, R^2, and *E* being 5%, 5%, and 11% better than those of the latter scheme. As for TDS prediction, both long-term Approaches 1 and 2 predictions applying wet season input data (Scheme 1) proved to be more accurate than using dry season input data (Scheme 2), with the former scheme's *RMSE*, R^2, and *E* being 2%, 3%, and 6% better than those of the latter scheme in the first approach, and 9%, 6%, and 34% superior to Approach 2. As a result, determining which scheme is more accurate is not possible, and it differs in different variables and approaches.

GP was found to perform very well as an optimizer of streamflow and TDS predictions with short-term and long-term prediction approaches. Among these approaches, the long-term prediction Approach 1 seems particularly well suited for predicting whether quantitative or qualitative hydrologic variables of importance in water resources management.

References

Ahmad, S., Khan, I., and Parida, B. P. (2001). "Performance of stochastic approaches for forecasting river water quality." *Water Res.*, 35(18), 4261–4266.

Ashofteh, P. S., Bozorg-Haddad, O., and Loáiciga, H. A. (2015a). "Evaluation of climatic-change impacts on multi-objective reservoir operation with multiobjective genetic programming." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)WR.1943-5452.0000540, 04015030.

Ashofteh, P.-S., Bozorg-Haddad, O., Akbari-Alashti, H., and Mariño, M. A. (2015b). "Determination of irrigation allocation policy under climate change by genetic programming." *J. Irrig. Drain. Eng.*, 10.1061/(ASCE)IR.1943-4774.0000807, 04014059.

Ashofteh, P.-S., Bozorg-Haddad, O., Mariño, M. A. (2013). "Scenario assessment of streamflow simulation and its transition probability in future periods under climate change." *Water Resour. Manage.*, 27(1), 255–274.

Ashofteh, P.-S., Bozorg-Haddad, O., and Mariño, M. A. (2015c). "Risk analysis of water demand for agricultural crops under climate change." *J. Hydrol. Eng.*, 10.1061/(ASCE)HE.1943-5584.0001053, 04014060.

Bazartseren, B., Hildebrandt, G., Holz, K. P. (2003). "Short-term water level prediction using neural networks and neuro-fuzzy approach." *Neurocomputing*, 55(3–4), 439–450.

Beygi, S., Bozorg-Haddad, O., Fallah-Mehdipour, E., and Mariño, M. A. (2014). "Bargaining models for optimal design of water distribution networks." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)WR.1943-5452.0000324, 92–99.

Billings, R. B. and Agthe, D. E. (1998). "State-space versus multiple regression forforecasting urban water demand." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)0733-9496(1998)124:2(113), 113–117.

Bolouri-Yazdeli, Y., Bozorg-Haddad, O., Fallah-Mehdipour, E., and Mariño, M. A. (2014). "Evaluation of real-time operation rules in reservoir systems operation." *Water Resour. Manage.*, 28(3), 715–729.

Bozorg-Haddad, O., Ashofteh, P.-S., Ali-Hamzeh, M., and Mariño, M. A. (2015a). "Investigation of reservoir qualitative behavior resulting from biological pollutant sudden entry." *J. Irrig. Drain. Eng.*, 10.1061/(ASCE)IR.1943-4774.0000865, 04015003.

Bozorg-Haddad, O., Ashofteh, P.-S., and Mariño, M. A. (2015b). "Levee's layout and design optimization in protection of flood areas." *J. Irrig. Drain. Eng.*, 10.1061/(ASCE)IR.1943-4774.0000864, 04015004.

Bozorg-Haddad, O., Ashofteh, P.-S., Rasoulzadeh-Gharibdousti, S., and Mariño, M. A. (2014). "Optimization model for design-operation of pumped-storage and hydropower systems." *J. Energy Eng.*, 10.1061/(ASCE)EY.1943-7897.0000169, 04013016.

Bozorg-Haddad, O., Rezapour Tabari, M. M., Fallah-Mehdipour, E., and Mariño, M. A. (2013). "Groundwater model calibration by metaheuristic algorithms." *Water Resour. Manage.*, 27(7), 2515–2529.

Charhate, S. B., Dandawate, Y. H., and Londhe, S. N. (2009). "Genetic programming to forecast stream flow." *Advances in water resources and hydraulic engineering*, Springer, Berlin, 29–34.

Chau, K. (2006). "A review on integration of artificial intelligence into water quality modelling." *Marine Pollut. Bull.*, 52(7), 726–733.

Elshorbagy, A., Simonovic, S. P., and Panu, U. S. (2002). "Estimation of missing stream flow data using principles of chaos theory." *J. Hydrol.*, 255(1–4), 123–133.

Fallah-Mehdipour, E., Bozorg-Haddad, O., and Mariño, M. A. (2013a). "Developing reservoir operational decision rule by genetic programming." *J. Hydroinform.*, 15(1), 103–119.

Fallah-Mehdipour, E., Bozorg-Haddad, O., and Mariño, M. A. (2013b). "Extraction of optimal operation rules in aquifer-dam system: A genetic programming approach." *J. Irrig. Drain. Eng.*, 10.1061/(ASCE)IR.1943-4774.0000628, 872–879.

Fallah-Mehdipour, E., Bozorg-Haddad, O., and Mariño, M. A. (2013c). "Prediction and simulation of monthly groundwater levels by genetic programming." *J. Hydro-Environ. Res.*, 7(4), 253–260.

Fallah-Mehdipour, E., Bozorg-Haddad, O., Orouji, H. and Mariño, M. A. (2013d). "Application of genetic programming in stage hydrograph routing of open channels." *Water Resour. Manage.*, 27(9), 3261–3272.

Fu, T. (2011). "A review on time series data mining." *Eng. Appl. Artif. Intell.*, 24(1), 164–181.

Gato, S., Jayasuriya, N., and Roberts, P. (2007). "Temperature and rainfall thresholds for base use urban water demand modeling." *J. Hydrol.*, 337(3-4), 364–376.

Havlíček, V., Hanel, M., Máca, P., Kuráž, M., and Pech, P. (2013). "Incorporating basic hydrological concepts into genetic programming for rainfall-runoff forecasting." *Computing*, 95(1), 363–380.

Izadifar, Z. and Elshorbagy, A. (2010). "Prediction of hourly actual evapotranspiration using neural network, genetic programming, and statistical models." *Hydrol. Process.*, 24(23), 3413–3425.

Khu, S. T., Liong, S., Babovic, V., Madsen, H., and Muttil, N. (2001). "Genetic programming and its application in real-time runoff forecasting." *J. Am. Water Resour. Assoc.*, 37(2), 439–451.

Kothyari, U. C. and Singh, V. P. (1999). "A multiple-input single-output model for flow forecasting." *J. Hydrol.*, 220(1), 12–26.

Liao, S., Chu, P., and Hsiao, P. (2012). "Data mining techniques and applications—A decade review from 2000 to 2011." *Exp. Sys. Appl.*, 39(12), 11303–11311.

Liu, J., Savenije, H. G., and Xu, J. (2002). "Forecast of water demand in Weinan city in China using WDF-ANN model." *Phys. Chem. Earth*, 28(4–5), 219–224.

Liu, S., Tai, H., Ding, Q., Li, D., Xu, L., and Wei, Y. (2011). "A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction." *Math. Comput. Modell.*, 53(3–4), 458–470.

Makkeasorn, A., Chang, N. B., and Zhou, X. (2008). "Short-term streamflow forecasting with global climate change implications—A comparative study between genetic programming and neural network models." *J. Hydrol.*, 352(3–4), 336–354.

Nagesh Kumar, D., Srinivasa Raju, K., and Sathish, T. (2004). "River flow forecasting using recurrent neural networks." *Water Resour. Manage.*, 18(2), 143–161.

Nasseri, M., Moeini, A., and Tabesh, M. (2011). "Forecasting monthly urban water demand using extended kalman filter and genetic programming." *Exp. Syst. Appl.*, 38(6), 7387–7395.

Nayak, P. C., Sudheer, K. P., Rangan, D. M., Ramasastri, K. S. (2005). "Short-term flood forecasting with a neuro-fuzzy model." *Water Resour. Res.*, 41(4).

Ni, Q., Wang, L., Ye, R., Yang, F., and Sivakumar, M. (2010). "Evolutionary modeling for streamflow forecasting with minimal datasets: A case study in the West Malian river, China." *Environ. Eng. Sci.*, 27(5), 377–385.

Orouji, H., Bozorg-Haddad, O., Fallah-Mehdipour, E., and Mariño, M. A. (2013). "Modeling of water quality parameters using data-driven models." *J. Environ. Eng.*, 10.1061/(ASCE)EE.1943-7870.0000706, 947–957.

Orouji, H., Bozorg-Haddad, O., Fallah-Mehdipour, E., and Mariño, M. A. (2014). "Extraction of decision alternatives in project management: Application of hybrid PSO-SFLA." *J. Manage. Eng.*, 10.1061/(ASCE)ME.1943-5479.0000186, 50–59.

Orouji, H., Bozorg-Haddad, O., Fallah-Mehdipour, E., and Mariño, M. A. (2014). "Flood routing in branched river by genetic programming." *Proc. Inst. Civ. Eng.: Water Manage.*, 167(2), 115–123.

Palani, S., Liong, S., and Tkalich, P. (2008). "An ANN application for water quality forecasting." *Mar. Pollut. Bull.*, 56(9), 1586–1597.

Rabuñal, J. R., Puertas, J., Suárez, J., and Rivero, D. (2007). "Determination of the unit hydrograph of a typical urban basin genetic programming and artificial neural networks." *Hydrol. Process.*, 21(4), 476–485.

Savic, D. A., Walters, G. A., and Davidson, J. W. (1999). "A genetic programming approach to rainfall-runoff modelling." *Water Resour. Manage.*, 13(3), 219–231.

Shokri, A., Bozorg-Haddad, O., and Mariño, M. A. (2013). "Reservoir operation for simultaneously meeting water demand and sediment flushing: A stochastic dynamic programming approach with two uncertainties." *J. Water Resour. Plann. Manage.*, 139(3), 277–289.

Shokri, A., Bozorg-Haddad, O., and Mariño, M. A. (2014). "Multi-objective quantity-quality reservoir operation in sudden pollution." *Water Resour. Manage.*, 28(2), 567–586.

Smith, R. C. G., Steiner, J. L., Meyer, W. S., and Erskine, D. (1985). "Influence of season to season variability in weather on irrigation scheduling of wheat: A simulation study." *Irrig. Sci.*, 6(4), 241–251.

Soltanjalili, M., Bozorg-Haddad, O., and Mariño, M. A. (2013). "Operating water distribution networks during water shortage conditions using hedging and intermittent water supply concepts." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)WR.1943-5452.0000315, 644–659.

Sreekanth, J. and Datta, B. (2011). "Comparative evaluation of genetic programming and neural network as potential surrogate models for coastal aquifer management." *Water Resour. Manage.*, 25(13), 3201–3218.

Tan, G., Yan, J., Gao, C., and Yang, S. (2012). "Prediction of water quality time series data based on least squares support vector machine." *Proc. Eng.*, 31, 1194–1199.

Wang, W., Chau, K., Cheng, C., and Qiu, L. (2009). "A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series." *J. Hydrol.*, 374(3–4), 294–306.

Xu, L., and Liu, S. (2013). "Study of short-term water quality prediction model based on wavelet neural network." *Math. Comput. Modell.*, 58(3–4), 807–813.

Yoon, H., Jun, S., Hyun, Y., Bae, G., and Lee, K. (2011). "A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer." *J. Hydrol.*, 396(1–2), 128–138.

Yu, P. S. and Tseng, T. Y. (1996). "A model to forecast flow with uncertainty analysis." *Hydrol. Sci. J.*, 41(3), 327–344.