Journal of Hydrology 563 (2018) 669–678

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

An improved gene expression programming model for streamflow forecasting in intermittent streams

Ali Danandeh Mehr

Department of Civil Engineering, Antalya Bilim University, Antalya, Turkey

ARTICLE INFO

This manuscript was handled by Corrado Corradini, Editor-in-Chief, with the assistance of Gokmen Tayfur, Associate Editor

Keywords:
Streamflow forecasting
Gene expression programming
Genetic algorithm
Evolutionary optimization
Intermittent streams

ABSTRACT

Skilful forecasting of monthly streamflow in intermittent rivers is a challenging task in stochastic hydrology. In this study, genetic algorithm (GA) was combined with gene expression programming (GEP) as a new hybrid model for month ahead streamflow forecasting in an intermittent stream. The hybrid model was named GEP-GA in which sub-expression trees of the best evolved GEP model were rescaled by appropriate weighting coefficients through the use of GA optimizer. Auto-correlation and partial auto-correlation functions of the streamflow records as well as evolutionary search of GEP were used to identify the optimum predictors (i.e., number of lags) for the model. The proposed methodology was demonstrated using monthly streamflow data from the Shavir Creek in Iran. Performance of the GEP-GA was compared to that of classic genetic programming (GP), GEP, multiple linear regression and GEP-linear regression models developed in the present study as the benchmarks. The results showed that the GEP-GA outperforms all the benchmarks and motivated to be used in practice.

1. Introduction

Accurate streamflow forecasting is an important task for variety of issues in basin hydrology including (but not limited to) reservoir operation, irrigation planning, food production, flood damage mitigation and environmental protection. A number of models have been suggested to simulate this complex process either conceptually or through data-driven methods (Aksoy and Bayazit, 2000; Wang et al., 2009, Yaseen et al., 2017). Intermittent streams are those that may experience dry spells occasionally. This is often the case in arid and semi-arid regions (Salas, 1993), particularly in the tributaries of mountainous rivers or snow-fed streams. Because of the paucity of gauging stations in mountainous regions, the commonly used rainfall-runoff approaches may not be applicable to forecast streamflow in intermittent streams. In such situations, data-driven techniques could be implemented to model streamflow time series if a continuous set of streamflow measurements is available. Then, the evolved model could be applied for neighbouring tributaries using regionalization techniques. In recent literature, due to the advances in data-driven techniques, a number of cross-station, single-station, and successive-station monthly streamflow forecasting models have been developed and their successful results have been reported Danandeh Mehr et al. (2013).

Gene expression programming (GEP) is relatively a new data-driven method that uses population of individuals (programs), improves according to fitness, and obtains the best solution using one or more genetic operators (Ferreira 2001). However, there is foremost differences between genetic programming (GP) and GEP algorithms mainly reside in the nature of their programs. In both, programs are nonlinear entities with different size and shape. While programs are encoded as parse tree in GP, they are encoded as linear strings of fixed length in GEP which are afterwards expressed as the chromosomes. Details about GP and GEP are provided in Section 2.

In recent years, different variants of GP such as GEP, multigene GP (MGGP), and linear GP (LGP) have been used for streamflow prediction (Babovic and Keijzer, 2002; Meshgi et al., 2015; Ravansalar et al., 2017). For example, Guven (2009) compared LGP with two versions of artificial neural networks (ANNs) to predict daily streamflow of Schuylkill River in the USA. The author demonstrated that the performance of LGP is higher than ANNs. Danandeh Mehr et al. (2013) used LGP for monthly streamflow prediction between successive-stations at Çoruh River, a perennial river in Turkey and showed that LGP is superior to neuro-wavelet model. Shoaib et al. (2015) integrated GEP model with discrete wavelet transform pre-processing approach to predict streamflow using rainfall data. The main contribution of the study was the introducing a novel wavelet-GEP model applicable over four watersheds. Worth to mention, the aim of applying wavelet transform on the streamflow time series was to extract their temporal and spectral information. The authors used the sequential time series approach to determine the input vector matrix that built the predictive model. The proposed wavelet-GEP model outperformed the individual GEP model in all case study catchments during both training and testing phases. Using rainfall, potential evapotranspiration and streamflow

E-mail address: ali.danandeh@antalya.edu.tr.

https://doi.org/10.1016/j.jhydrol.2018.06.049
Received 12 June 2018; Accepted 19 June 2018
Available online 20 June 2018
0022-1694/

from Moselle River basin in France, Danandeh Mehr and Demirel (2016) showed that MGGP can be satisfactorily used for one-day ahead low flow prediction. More recently, Danandeh Mehr and Kahya (2017) developed a Pareto-optimal moving MGGP model for daily streamflow prediction and demonstrate that their hybrid model can overcome the timing error in time series analysing of daily streamflow models.

Focusing on the implementation of GP/GEP in wider range of hydrological studies, the author's review showed that they have been frequently used to distil knowledge from natural or experimental observations (e.g., Khu et al., 2001; Kisi et al., 2012b; Meshgi et al., 2014; Johari and Nejad, 2015; Danandeh Mehr, 2018). These are techniques which generate symbolic expressions that can be interpreted and combined with domain knowledge (Babovic, 2005 and 2009). Thus, motivating to be used in practice. Until recently, only a few studies focused on the application of GEP for monthly streamflow forecasting. For example, Karimi et al. (2016) forecasted river flow for both daily and monthly time scales using GEP model integrated with wavelet data pre-processing approach at Filyos River, which is a perennial river in Mediterranean region of Turkey. For comparison purpose, traditional auto regressive moving average model together with two other soft computing methods, ANNs and adaptive neuro-fuzzy inference system, were used in the study. The authors showed that wavelet-GEP was superior to its counterparts. Al-Juboori and Guven (2016) developed a GEP-based stepwise monthly streamflow prediction model and demonstrated that their model precisely forecasts monthly flows at the perennial Hurman River in Turkey as well as Diyalah and Lesser Zab Rivers in Iraq.

Table 1 has listed some of the studies that implemented at least one GP variant for time series modelling of streamflow data. As shown in the table, Karimi et al. (2016) as well as Al-Juboori and Guven's (2016) papers are dealing with generating GEP-based monthly streamflow forecasting model for perennial rivers, whereas the present study focuses on the calibrating GEP for intermittent rivers. The main difference between the methodology of this study and those of Karimi et al. (2016) and Al-Juboori and Guven (2016) is the inclusion of seasonality effect in the selection of potential predictors which is the major pattern in the intermittent streamflow series. Moreover, the present study puts forward a new strategy to enhance the accuracy of GEP forecasts.

On the other hand, the documented studies related to the streamflow forecasting in intermittent rivers are quite limited owing to the complexity of time series modelling of intermittent flows (Kisi et al., 2012b). Although one might find a few studies that suggest the implementation of soft computing methods for intermittent streamflow forecasting (e.g., Cigizoglu, 2005; Kişi, 2009; Kisi et al., 2012b), to the best of the author's knowledge, the present study is the first study in the literature that applies GEP for monthly streamflow forecasting in an intermittent stream. Under the lights of the abovementioned literature, a new hybridization procedure is suggested in order to augment GEP prediction accuracy. This is a new procedure by which the coefficients of the best GEP induced expression are optimized through genetic algorithm (GA). The proposed hybrid GEP-GA methodology is applied for single-station monthly streamflow forecasting at Shavir Creek, an intermittent stream located at North West of Iran. The efficiency results of the new model are compared with those of classic GP, standalone GEP as well as multi-linear regression (MLR) and hybrid GEP-linear regression (GEP-LR) models developed in the present study as the benchmarks.

Table 1
List of papers that implemented GP for trim series modelling of streamflow time series.

Authors	GP variant	Time scale
Guven (2009)	LGP	Daily
Wang et al. (2009)	GP	Monthly
Danandeh Mehr et al. (2013)	LGP	Monthly
Karimi et al. (2016)	GEP	Daily, Monthly
Al-Juboori and Guven (2016)	GEP	Monthly
Danandeh Mehr and Kahya (2017)	MGGP	Daily
Ravansalar et al. (2017)	LGP	Monthly

2. Materials and methods

2.1. Study area and data

The task of intermittent streamflow forecasting in arid and semi-arid regions is more complicated than in moist tropical and subtropical climates. A first order tributary of Shavir stream, an intermittent stream in Sefidrood River Basin, located in a semi-arid region in North West of Iran, was selected as the case study in the present study (Fig. 1). The stream catchment covers an area of approximately 55.5 km^2, which is about 0.03% territory of Ardabil Province, Iran. The stream springs from Shavirdagh Mountains in Ardabil and reaches to Caspian Sea in Kiashahr City of Gilan Province, after a course of 300 km. Location of the stream gauge station, Givi hydrometric station, used in the present study was also shown in Fig. 1, and the historical monthly streamflow measurements at the station were depicted in Fig. 2. Fig. 2a shows 30-year mean monthly streamflow time series during the 1978–2008 period (local water year). The first 20 years (Fig. 2b) and the remaining 10 years (Fig. 2c) of the observations were respectively used to train and validate the standalone models (i.e., GP, GEP and MLR). The statistical characteristics of the entire data as well as the training and validation sub-series were presented in Table 2.

2.2. Performance evaluation

There are several ways to assess the performance of a model, some of which are (i) line plot to visually inspect the trend between measured data and model output, (ii) scatter plot of measured data versus model output, and (iii) error measures such as mean absolute error, root mean square error (RMSE), mean absolute relative error, Nash-Sutcliffe coefficient of efficiency (NSE), discrepancy ratio and others (Tayfur, 2012). In this study, performance of the proposed model and benchmarks are evaluated on the basis of line plots of trends between measured data and the models' output together with two error measures including NSE and RMSE. The former (Equation (1) is a normalized statistic that shows how well the scatter plot of observed and modeled data lie around the 1:1 perfect model straight line. The higher NSE (one for the perfect model), the better the model. The latter (Eq. (2)) is the conventional quantitative goodness-of-fit indicator that measures the average magnitude of the error with the same dimension of the predictand variable. The lower RMSE (zero for the perfect model), the better the model.

$$NSE = 1 - \frac{\sum_{i=1}^{n} (X_i^{obs} - X_i^{pre})^2}{\sum_{i=1}^{n} (X_i^{obs} - X_{mean}^{obs})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_i^{obs} - X_i^{pre})^2}{n}} \quad (2)$$

where *and* and *arethepredictedobservedvaluesofX* (*he remonthlystreamflow*), *respectively*. are the predicted observed values of X (here monthly streamflow), respectively. *ismeanamountofobserveddata*, *andndenotesthenumberofobservations*. is mean amount of observed data, and n denotes the number of observations.

2.3. Overview of GA, GP, and GEP

GA (Holland, 1975; Goldberg, 1989) is an evolutionally optimization technique that frequently used in hydrology (e.g., McKinney and