Engineering Applications of Artificial Intelligence 23 (2010) 1350–1367

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Data-driven models for monthly streamflow time series prediction

C.L. Wu, K.W. Chau *

Department of Civil and Structural Engineering, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, People's Republic of China

ARTICLE INFO

Article history:
Received 23 August 2009
Received in revised form
26 February 2010
Accepted 12 April 2010
Available online 6 May 2010

Keywords:
Hydrologic time series
Auto-regressive moving average
K-nearest-neighbors
Artificial neural networks
Phase space reconstruction
False nearest neighbors
Dynamics of chaos

ABSTRACT

Data-driven techniques such as Auto-Regressive Moving Average (ARMA), K-Nearest-Neighbors (KNN), and Artificial Neural Networks (ANN), are widely applied to hydrologic time series prediction. This paper investigates different data-driven models to determine the optimal approach of predicting monthly streamflow time series. Four sets of data from different locations of People's Republic of China (Xiangjiaba, Cuntan, Manwan, and Danjiangkou) are applied for the investigation process. Correlation integral and False Nearest Neighbors (FNN) are first employed for Phase Space Reconstruction (PSR). Four models, ARMA, ANN, KNN, and Phase Space Reconstruction-based Artificial Neural Networks (ANN-PSR) are then compared by one-month-ahead forecast using Cuntan and Danjiangkou data. The KNN model performs the best among the four models, but only exhibits weak superiority to ARMA. Further analysis demonstrates that a low correlation between model inputs and outputs could be the main reason to restrict the power of ANN. A Moving Average Artificial Neural Networks (MA-ANN), using the moving average of streamflow series as inputs, is also proposed in this study. The results show that the MA-ANN has a significant improvement on the forecast accuracy compared with the original four models. This is mainly due to the improvement of correlation between inputs and outputs depending on the moving average operation. The optimal memory lengths of the moving average were three and six for Cuntan and Danjiangkou, respectively, when the optimal model inputs are recognized as the previous twelve months.

1. Introduction

Many data-driven models, including linear, nonparametric or nonlinear approaches, are developed for hydrologic discharge time series prediction in the past decades (Marques et al., 2006). Generally, there are two basic assumptions underlay different model techniques. The first assumption suggests that a time series is originated from a stochastic process with an infinite number of degrees of freedom. Under this assumption, linear models such as AutoRegressive (AR), AutoRegressive Moving Average (ARMA), AutoRegressive Integrated Moving Average (ARIMA), and Seasonal ARIMA (SARIMA) had made a great success in river flow prediction (Carlson et al., 1970; Salas et al., 1985; Haltiner and Salas, 1988; Yu and Tseng, 1996; Kothyari and Singh, 1999; Huang et al., 2004; María et al., 2004).

The second assumption is that a random-looking hydrologic time series is derived from a deterministic dynamic system such as chaos. In the past two decades, chaos-based streamflow prediction techniques have been increasingly obtaining interests of the hydrology community (Jayawardena and Lai, 1994; Jayawardena and Gurung, 2000; Elshorbagy et al., 2002; Wang et al., 2006b) although some doubts have been raised in terms of the existence of chaos in hydrologic data (Ghilardi and Rosso, 1990; Koutsoyiannis and Pachakis, 1996; Pasternack, 1999; Schertzer et al., 2002; Wang et al., 2006a). Generally, the prediction techniques for a dynamic system can be roughly divided into two approaches: local and global. Local approach uses only nearby states to make predictions whereas global approach involves all the states. K-Nearest-Neighbors (KNN) algorithm, Artificial Neural Networks (ANN) and Support Vectors Machine (SVM) are some typical forecast methods for dynamic systems (Sivapragasam et al., 2001; Laio et al., 2003; Wang et al., 2006b). Phase-Space-Reconstruction (PSR) is a precondition before performing any predictions of the dynamic system. Typical methods involved in PSR are correlation integral, singular-value decomposition of the sample covariance matrix, False Nearest Neighbors (FNN), and true vector fields (Grassberger and Procaccia, 1983; Abarbanel et al., 1993).

Comparative studies on the above prediction techniques have been further carried out by some researchers. Sivakumar et al. (2002)

Abbreviations: ACF, Autocorrelation function; AIC, Akaike information criterion; AMI, Average mutual information; ANN, Artificial neural networks; AR, Auto-regressive; ARMA, Auto-regressive moving average; ARIMA, Auto-regressive integrated moving average; BFGS, Broyden–Fltecher–Goldfarb–Shanno; BP, Back-propagation; CCF, Cross-correlation function; CE, Coefficient of Efficiency; CT, Cuntan; DJK, Danjiangkou; FNN, False nearest neighbors; FNNP, Percentage of FNN; GA, Genetic algorithm; KNN, K-nearest-neighbors; LM, Levenberg–Marquart; LM–GA, Levenberg–Marquart and genetic algorithm; MA-ANN, Moving average artificial neural networks; MW, Manwan; NNM, Nearest neighbor method; PI, Persistence index; RE, Relative error; PSO, Particle swarm optimization; PSR, Phase space reconstruction; ANN-PSR, Phase space reconstruction-based artificial neural networks; RMSE, Root mean square error; SARIMA, Seasonal ARIMA; SEC-UA, Shuffled complex evolution; SVM, Support vectors machine; XJB, Xiangjiaba

* Corresponding author.
E-mail address: cekwchau@polyu.edu.hk (K.W. Chau).

0952-1976/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2010.04.003

found that the performance of the KNN approach was consistently better than ANN in short-term river flow prediction. Laio et al. (2003) carried out a comparison of KNN and ANN for flood predictions and found that KNN performed slightly better at short forecast time while the situation was reversed for longer time. Similarly, Yu et al. (2004) proposed that KNN performed worse than ARIMA on the basis of daily streamflow prediction. The conclusions in literature are very inconsistent. It is difficult to justify which modeling technique is more suitable for a streamflow forecast.

The above two assumptions are in the extremes of a hydrologic streamflow series. Salas et al. (1985) suggested that a streamflow process should be treated as an integration of stochastic (or random) and deterministic components. Describing it as either a totally linear stochastic process or fully nonlinear deterministic chaos is not a practical approach (Elshorbagy et al., 2002). Therefore, the model based on either of two assumptions may not be the most suitable. An investigation on an optimal prediction model is worthy to further study with different real monthly streamflow data (Xiangjiaba, Cuntan, Manwan, and Danjiangkou).

The scope of this study is to compare four forecast models, ARMA, ANN, KNN, and ANN-PSR and develop an optimal model for monthly streamflow prediction. This paper is organized in the following manner. Section 2 presents the four sets of streamflow data used in this study. Section 3 first describes the principles of PSR and then identifies its parameters using the correlation integral approach and the FNN approach. The implementation of the forecast models, including data preparation and selection of parameters, is discussed in Section 4. Forecast results are described in Section 5 and conclusions of the study are presented in Section 6.

2. Streamflow data

Monthly streamflow series of three watersheds and one river, i.e. Xiangjiaba, Manwan, Danjiangkou, and Yangtze River, were analyzed in this study.

The largest watershed, Xiangjiaba, is at the upstream of Yangtze river with average yearly discharge of 4538 m^3/s. Monthly streamflow series were taken from the hydrological station near the Xiangjiaba Dam site located in Sichuan Province. The basin area contributed to the streamflow series is around 45.88×10^4 km^2. The period of the data was from January 1940 to December 1997.

The medium watershed, Manwan, is located in the Lancang River which originates from the Qinghai–Tibet Plateau. Monthly streamflow series were taken from the hydrological station near the Manwan Dam site located in Sichuan Province. The catchment area controlled by the station is 11.45×10^4 km^2, and the average yearly discharge is 1230 m^3/s based on a statistic of 30-year data (January 1974–December 2003).

The smallest watershed, Danjiangkou, lies at the upstream of Han river with average yearly discharge of 1203 m^3/s. Monthly streamflow data came from the hydrology station at the Danjiangkou Dam site which is located in Hubei Province. The catchment area at the dam site is around 9.5×10^4 km^2. The data range was from January 1930 to December 1981.

The last streamflow series is Yangtze River, the largest river in China. The selected monthly streamflow data were from the hydrology station of Cuntan located in the middle stream of the river. The stream flow series spanned from January 1893 to December 2007.

Four monthly streamflow series are shown in Fig. 1. Monthly streamflow data in Xiangjiaba, Manwan, and Cuntan are characterized by a smooth process whereas monthly streamflow data in Danjiangkou exhibits complex oscillations. The linear fits (dotted lines in Fig. 1) verify the consistency of the streamflow series. All series exhibit good consistency because the linear fits are close to horizontal. Since there were no large-scale hydraulic works such as dams built during the data collection period, the streamflow process is fairly pristine in each case.

3. Reconstruction of dynamics

3.1. Phase space reconstruction

To describe the temporal evolution of a dynamic system in a multi-dimensional phase space with a scale time series, it is essential to employ some techniques to unfold the

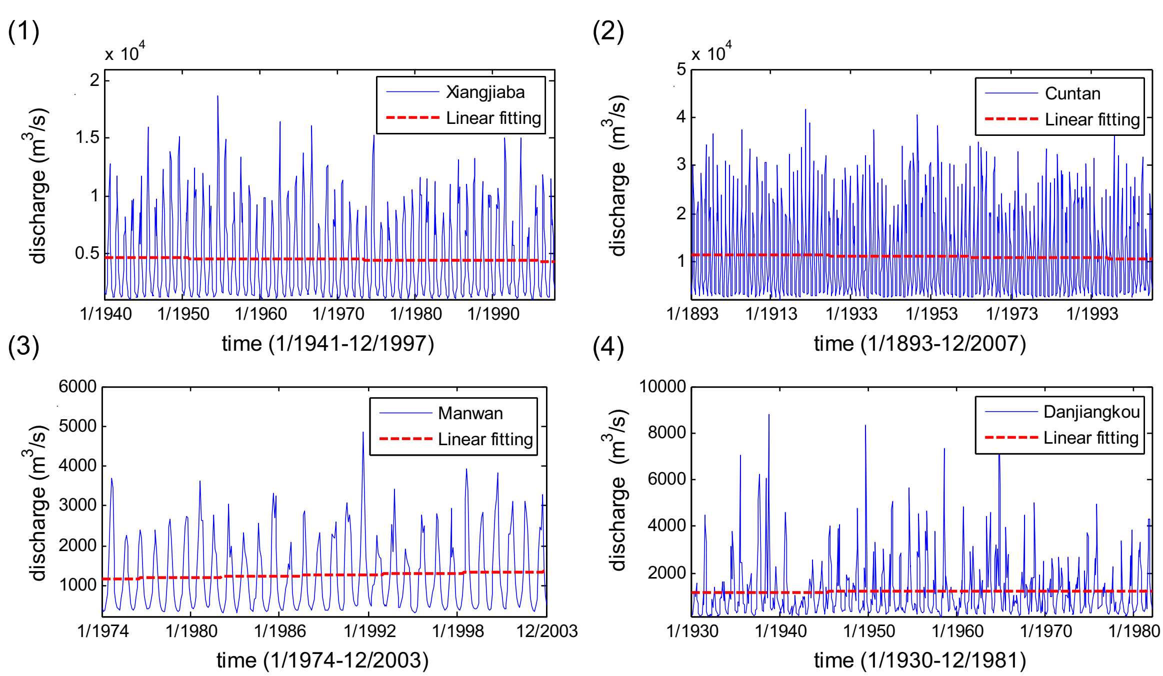

Fig. 1. Monthly discharge series of (1) Xiangjiaba, (2) Cuntan, (3) Manwan, and (4) Danjiangkou.