International Journal of Structural and Civil Engineering
ISSN : 2277-7032 Volume 1 Issue 2 (February 2012)
http://www.ijsce.com/ https://sites.google.com/site/ijscejournal

Application of Chaos Theory and Genetic Programming in Runoff Time Series

Mohammad Ali Ghorbani[1], Hossein Jabbari Khamnei[2], Hakimeh Asadi[3*], Peyman Yousefi[4]

[1]Associate Professor, Department of Water Engineering, Tabriz University, Tabriz, Iran, E-mail: cusp2004@yahoo.com
[2]Assistant Professor, Department of Statistics, Tabriz University, Tabriz, Iran, E-mail: h_jabbari@tabrizu.ac.ir
[3*]Master of Science, Department of Water Engineering, Tabriz University, Tabriz, Iran, E-mail: tu.i2006@yahoo.com
[4]Master of Science, Department of Civil Engineering, Tabriz University, Tabriz, Iran, E-mail: pe.yousefi@yahoo.com

Abstract. Nowadays, prediction of runoff is very important in water resources management and their permanent development. Along with scientific advances in recent years, various intelligent methods and regression and mathematical methods have been presented to estimate the runoff. In this paper, Two different methods are used, Chaos analysis and genetic programming. The performances of models are analyzed and result showed that runoff have had chaotic behavior. Application of genetic programming models in estimating the runoff is also studied in this paper. The data that has been used has chaotic behavior and a mathematical model of genetic programming with rainfall and runoff as model inputs was result.

Keywords: Chaos, Genetic Programming, Runoff, Lighvan Basin

1 Introduction

Chaotic behaviors refer to the time history of a single variable of a deterministic dynamical system undergoing a loss of temporal correlation in response to small perturbations in initial conditions. Hydrologic and other water-related time series have been modeled by chaos theory over the past two decades and specific applications outlined as follows. (a) The presence of low-dimensional deterministic behaviors in the river flow processes were investigated by Jayawardena and Lai [1], Porporato and Ridolfi [2], Krasovskaia *et a*l. [3], Stehlik[4], Sivakumar *et al.* [5]. (b) Nonlinear deterministic approaches were used to detect the presence of chaos and achieve more accurate river flow predictions by Islam and Sivakumar [6], Lisi and Villi [7], Liu *et al.* [8]. (c) Alternative mathematical formulations have been developed to investigate

water-related problem, e.g. Qingfang and Yuhua who developed a new local linear prediction model for chaotic river flow series [9]. Genetic Programming (GP) is among heuristic algorithms all of which are based on Darwin's evolution theory. The mentioned algorithms define a target function in the shape of qualitative standards and then make use of the mentioned function to measure and compare different solution methods in a step by step process of data source correction and finally present the appropriate solution method. Kalra and Deo applied the GP for the completion of missing data in wave records along the west coast of India [10]. Ustoorikar and Deo used the GP for filling up gaps in datasets of wave heights [11]. Aytek and Kishi used GP approach to suspended sediment modeling for two stations on the Tongue River in Montana, USA, and indicate that the GP formulation performs quite well compared to sediment rating curves and multi linear regression models [12]. Gaur and Deo applied the GP for real-time wave forecasting [13].

2 Materials and Methods

2.1 Chaos

Chaos theory is a method of nonlinear time series analysis and involves a host of methods, essentially based on the phase-space reconstruction of the process, from scalar or multivariate measurements of physical observables.

Phase Space Reconstruction. One way of characterizing dynamical systems is by the concept of phase-space, according to which given a set of physical variables and an analytical model describing their interactions where each of its points corresponds to a state of the system. The delay embedding method reconstructs phase-space from a univariate or multivariate time series, which is assumed to be generated by a deterministic dynamical system [14]. The Takens theorem states that the underlying dynamics can be fully recovered by building a *m*-dimensional space wherein the components of each state vector $\vec{Y}_t$ are defined through the delay coordinates:

$$\vec{Y}_t = \left(X_t, X_{t-\tau}, X_{t-2\tau}, \ldots, X_{t-(m-1)\tau}\right) \quad \textbf{(1)}$$

where m is known as embedding dimension, τ as delay time and $X_t=\{x_1, x_2, \ldots x_N\}$ with N-observed values. This delay-embedding method is sensitive to both embedding parameters of τ and m, which are unknown a-priori. As suggested by Cellucci et al

[15], Average Mutual Information (AMI) is used to estimate τ . AMI defines how the measurements $X(t)$ at time t are related, from an information theoretic point of view, to measurements $X(t+\tau)$ at time $t+\tau$. The average mutual information is defined as [16]:

$$I(\tau) = \sum_{X(i),X(i+\tau)} P(X(i), X(i+\tau)) \log \left[\frac{P(X(i), X(i+\tau))}{P(X(i))\, P(X(i+\tau))} \right] \quad \textbf{(2)}$$

where the sum is extended to the total number of samples in the times series. $P(X(i))$ and $P(X(i+\tau))$ are the marginal probabilities for measurements $X(i)$ and $X(i+\tau)$, respectively, whereas $P(X(i), X(i+\tau))$ is their joint probability. The optimal delay time τ minimizes the function $I(\tau)$*: for* $t = \tau$, $X(i+\tau)$ adds the maximum information on $X(i)$.

Correlation Dimension. Correlation dimension is a nonlinear measure of the correlation between pairs lying on the attractor. For time series whose underlying dynamics is chaotic whereas for stochastic systems it is infinite. For an m -dimensional phase-space, the correlation function $C_m(r)$ is defined as the fraction of states closer than r [17, 18]:

$$C_m(r) = \lim_{N\to\infty} \frac{2}{(N-w)(N-w-1)} \sum_{i=1}^{N} \sum_{j=i+1+w}^{N} H(r - |\vec{Y}_i - \vec{Y}_j|) \quad \textbf{(3)}$$

where H is the Heaviside step function, $\vec{Y}_i$ is the i^{th} state vector, and N is the number of points on the reconstructed attractor. For stochastic time series $C_m(r) \propto r^m$ holds, whereas for chaotic time series the correlation function scales with r as:

$$C_m(r) \propto r^{D_2} \quad \textbf{(4)}$$

where D_2, correlation exponent, quantifies the degrees of freedom of the process, and defined by:

$$D_2 = \lim_{r \to 0} \frac{\ln C_m(r)}{\ln r} \quad \textbf{(5)}$$

and can be reliably estimated as the slope in the $\ln C_m(r)$ vs. $\ln(r)$ plot.

2.2 Genetic Programming

The GP is similar to Genetic Algorithm (GA) but employs a "parse tree" structure for the search of its solutions, whereas the GA employs bite strips. The technique is truly a "bottom up" process, as there is no assumption made on the structure of the relationship between the independent and dependent variables but an appropriate relationship is identified for any given time series. The relationship can be logical statements or it is normally a mathematical expression, which may be in some familiar mathematical format or it may assemble a mathematical function in a completely unfamiliar format. The construction of the relationship is made possible by two components: (i) a parse tree, which is a functional set of basic operators and those selected in this study are:

$$\{+, -, \times\} \quad \textbf{(6-a)}$$

$$\{+, -, \times, x, x^2\} \quad \textbf{(7-b)}$$

which emulates the role of RNA; and (ii) the actual components of the functions and their parameters (referred to as the terminal set), which emulates which emulates the role of RNA.

2.3 Study Area and Data

The runoff time series of Lighvan basin, Iran ($46°-20'-30''$ to $46°-27'-30''$ east latitude and $37°-42'-55''$ to $37°-49'-30''$ north longitude) was used in the study. This watershed with a drainage area 76.19 Km^2 is important part of the catchment of Talkheh River watershed. The maximum and minimum elevations of the area are around 3500 and 2000 m, respectively. The length of longest stream is 17 km. The

average stream slope is 11%. The Lighvan River drains into Talkheh River and Urmia Lake, respectively.

For the present investigation, rainfall-runoff data observed for composed storm. Figure 1 shows the variations of rainfall-runoff data. The entire dataset was divided into two parts. The first 80% of data was used in training for the phase space reconstruction, but the subsequent 20% of data was used as observed data in the prediction phase.

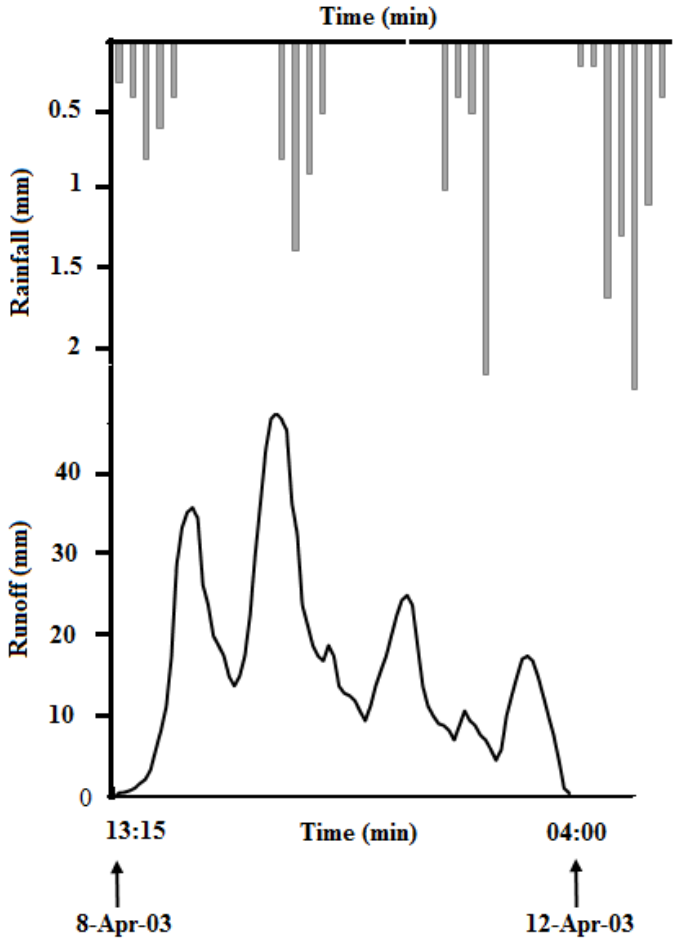

Fig. 1. Time series plot of runoff data in the Lighvan basin

3 Results

Two methods are used to identify a possible existence of chaos in the runoff time series in the Lighvan basin. Using the AMI method, the delay time, τ is estimated for the time series in the Lighvan basin as the intercept with the x-axis of the curves by plotting the values of the AMI evaluated by the TISEAN package against delay times progressively increased from 1 to 100 [19]. As shown in Figure 2 this method shows well-defined first minima at delay time of 10.

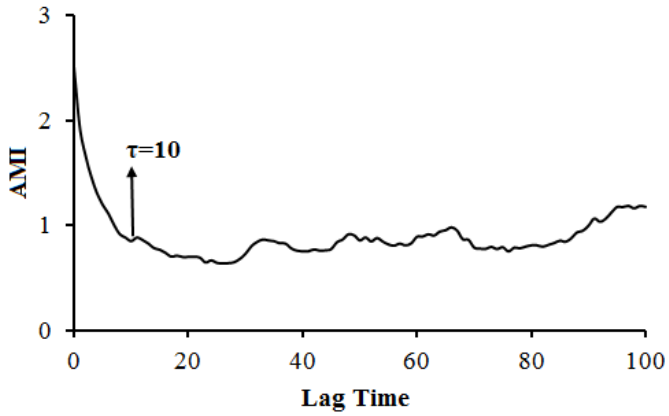

Fig. 2. Average mutual information function of runoff data from the Lighvan basin

The correlation function is calculated for the dataset using the delay times (τ =10), determined by the AMI method in the previous section, and embedding dimensions, m, by allowing it to vary from 1 to 20. Figure 3-a shows the relationship between correlation function *C(r)* and radius r (i.e. ln*C(r)* versus ln*(r)*) for increasing m, whereas Figure 3-b shows the relationship between the correlation dimension values $D_2(m)$ and the embedding dimension values m. It can be seen from Figure 3-b. that the value of correlation exponent increases with the embedding dimension up to certain value and then saturates beyond it. The saturation of the correlation exponent is an indication of the existence of deterministic dynamics. The saturated correlation dimension is 1.5, (D_2=1.5). The value of correlation dimension also suggests the possible presence of chaotic behavior in the dataset.

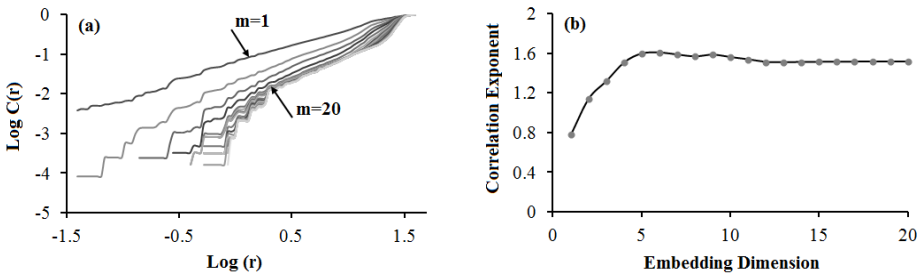

Fig. 3. a) Convergence of log$C(r)$ versus log(r) b) Saturation of correlation dimension $D_2(m)$ with embedding dimension m—saturation signifies chaotic signals in the Lighvan basin

Two different combinations of arithmetic function set were used to this problem: a) The first set $\{+,-,\times\}$; b) The second set $\{+,-,\times,x,x^2\}$. Genetic programming with two combination of arithmetic function set and three different combination of runoff time series;1: $\{P_t, P_{t-1}, Q_{t-1}, Q_t\}$; 2: $\{P_t, P_{t-1}, P_{t-2}, Q_{t-1}, Q_t\}$ and 3: $\{P_t, P_{t-1}, P_{t-2}, Q_{t-1}, Q_{t-2}, Q_t\}$ applied for training and testing data. Comparison of the statistical parameters (RMSE=0.1446, R^2=0.9989) of GP resulted from training and testing step shows that combination no.3 of runoff time series with first (6-a) arithmetic function set is the best combination of input data and arithmetic function where shown in Table 1. The comparison between observed and computed runoff from GP model is shown in Figure 4.

Table 1. The results of GP model for the training and testing steps

Model	Operators		R^2	RMSE
1	$\{+,-,\times\}$	Train	0.9948	0.7158
		Test	0.9891	0.4584
2		Train	0.9947	0.7205
		Test	0.9891	0.4584
3		Train	**0.9992**	**0.2720**
		Test	**0.9989**	**0.1446**
1	$\{+,-,\times,x,x^2\}$	Train	0.9948	0.7119
		Test	0.9891	0.4584
2		Train	0.9947	0.7182
		Test	0.9891	0.4568
3		Train	0.9992	0.2733
		Test	0.9989	0.1446

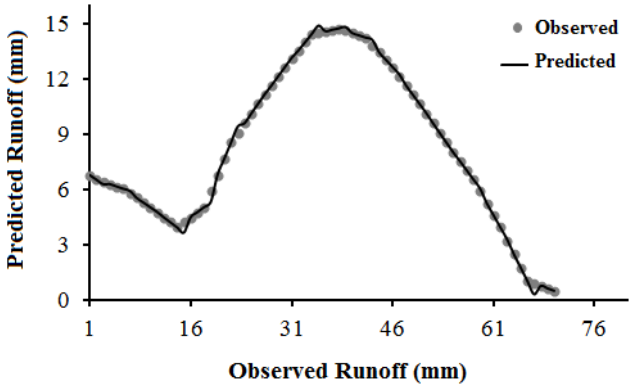

Fig. 4. Comparison between observed and computed runoff time series by GP model

The mathematical model that obtained by GP, According to Equation 7 is combination of the best input and output parameters with P_t, P_{t-1}, P_{t-2}, Q_{t-1}, Q_{t-2} as input and Q_t as output parameters.

$$Q_t = 2Q_{t-1} - 0.007843(-0.007843 - P_{t-1})P_{t-2}{}^2 Q_{t-1} - Q_{t-2} \quad \textbf{(7)}$$

4 Conclusion

This paper investigated possible chaotic behaviours in the runoff time series of the Lighvan basin. The analysis was performed on runoff time series records on an event. The analysis was based on using phase space reconstruction, correlation dimension method. The correlation dimension value is 1.5. Predictions of these time series using a GP are found to be acceptable. In addition, results obtained from GP were compared with observed data. It was documented that prediction with GP is good.

References

1. Jayawardena, W., Lai, F.: Analysis and prediction of chaos in rainfall and streamflow time series, J. of Hydrology. 153, 23--52 (1994).
2. Porporato, Ridolfi, L.: Nonlinear analysis of river flow time sequences, Water Resources Research. 33(6), 1353--1367 (1997).
3. Krasovskaia, L., Gottsehalk, Z., Kundzewicz, W.: Dimensionality of Scandinavian river flow regimes, Hydrol. Sci. J. 44(5),705--723 (1999).

4. Stehlík, J.: Deterministic chaos in runoff series, J. of Hydrology and Hydrodynamics. 47(4), 271--287 (1999).
5. Sivakumar, B., Berndtsson, R., Persson, M.: Monthly runoff prediction using phase-space reconstruction, Hydrological Sciences Journal. 46(3), 377--388 (2001b).
6. Sivakumar, B.: A phase-space reconstruction approach to prediction of suspended sediment concentration in rivers, J. of Hydrology. 258(1-4), 149--162 (2002).
7. Lisi, V., Villi, V.: Chaotic forecasting of discharge time series: A case study, J. of the American Water Resources Association. 37(2), 271--279 (2001).
8. Liu, Q., Islam, S., Rodriguez-Iturbe, V., Lee, Y.: Phase-space analysis of daily streamflow: characterization and prediction, Adv. Wat. Resour. 21, 463--475 (1998).
9. M. Qingfang and P. Yuhua P.: A new local linear prediction model for chaotic time series, Physics Letters A. 370: 465-470 (2007).
10. Kalra, R., Deo, M. C.: Genetic programming for retrieving missing information in wave records along the west coast of India, App. Ocean Res. 29(3):99--111(2007).
11. Ustoorikar, V., Deo, M. C.: Filling up gaps in wave data with genetic programming, Marine Structures. 21:177--195 (2008).
12. Aytek, A., Kisi, O.: A genetic programming approach to suspended sediment modelling, J. of Hydrology. 351: 288--298 (2008).
13. Gaur, V., Deo, M. C.: Real-time wave forecasting using genetic programming, Ocean Engineering. 35 (11–12):1166--1172 (2008).
14. Takens, F.: Detecting strange attractors in turbulence, in Lectures Notes in Mathematics, edited by D.A. Rand and L.S. Young, 898, 366--381, Springer-Verlag, New York (1981).
15. Cellucci, V., Albano, A. M., Rapp, P. E.: Comparative study of embedding methods, Physical Review E. 67, No. 6, 66210 (2003).
16. Fraser, M., Swinney, H. L.: Independent coordinates for strange attractors from mutual information, Physical Review A, 33(2), 1134--1140 (1986).
17. Grassberger, P., Procaccia, I.: Characterization of strange attractors, Physical review letters. Vol. 50, No. 5, 346--349 (1983).
18. Theiler, J.: Spurious dimension from correlation algorithms applied to limited timeseries data, Phys. Rev. A 34, 2427--2432 (1986).
19. Hegger, R., Kantz, V., Schreiber, T.: Practical implementation of nonlinear time series methods: The TISEAN package. Chaos. 9, 413--435(1999).