ELSEVIER

Journal of Hydrology 255 (2002) 123–133

Journal of Hydrology

www.elsevier.com/locate/jhydrol

Estimation of missing streamflow data using principles of chaos theory

Amin Elshorbagy[a,*], S.P. Simonovic[b], U.S. Panu[c]

[a]*Kentucky Water Research Institute, University of Kentucky, Lexington, KY 40506-0107, USA*
[b]*Department of Civil and Environmental Engineering, Institute for Catastrophic Loss Reduction, University of Western Ontario, London, Ont., Canada N6A 5B9*
[c]*Department of Civil Engineering, Lakehead University, Thunder Bay, Ont., Canada P7B 5E1*

Received 24 May 2000; revised 29 August 2001; accepted 31 August 2001

Abstract

In this paper, missing consecutive streamflows are estimated, using the principles of chaos theory, in two steps. First, the existence of chaotic behavior in the daily flows of the river is investigated. The time delay embedding method of reconstructing the phase space of a time series is utilized to identify the characteristics of the nonlinear deterministic dynamics. Second, the analysis of chaos is used to configure two models employed to estimate the missing data, artificial neural networks (ANNs) and *K*-nearest neighbor (*K*-nn). The results indicate the utility of using the analysis of chaos for configuring the models. ANN model is configured using the identified correlation dimension (measure of chaos), and (*K*-nn) technique is applied within a subspace of the reconstructed attractor. ANNs show some superiority over *K*-nn in estimating the missing data of the English River, which is used as a case study.

Keywords: Chaos theory; Missing data; Artificial neural networks; Nonlinear time series analysis

1. Introduction

During the past few decades, hydrologists have been conducting comprehensive research regarding the appropriate type of analysis for the hydrologic data. Statements such as linear versus nonlinear, deterministic versus stochastic, black box versus conceptual models, parametric versus nonparametric have become part of the common hydrologic vocabulary. Influenced by the fast-advancing research on chaotic behavior in the physics field, Rodriguez-Iturbe et al. (1989) have indirectly introduced to hydrologists, a new topic for comparative analysis. The last few years have witnessed the birth of the topic of stochastic versus chaotic time series analysis (e.g. Jayawardena and Gurung, 2000). The literature on chaos in water resources is limited and research is still in its infancy. However, various levels of research and applications are found among the few available publications. The only question of whether data are chaotic or not is addressed by Rodriguez-Iturbe et al. (1989), Sharifi et al. (1990), Islam et al. (1993), Angelbeck and Minkara (1994), Sangoyomi et al. (1996) and Sivakumar et al. (1998). Others (e.g. Jayawardena and Lai, 1994; Lall et al. 1996; Porporato and Ridolfi, 1997; Sivakumar et al. 1999a,b) have taken a step forward by trying to predict future values of the variable under consideration. Also, Puente and

* Corresponding author.

0022-1694/02/$ - see front matter
PII: S0022-1694(01)00513-3

Obregon (1996) have been able to fit a model to high-resolution rainfall time series via projections of fractal interpolating functions.

In many applications, nonlinear modeling tools have provided better results when used in hydrological time series analysis. Few examples, among others, are the superiority of nearest neighbor technique over ARMA models for predicting streamflows (Jayawardena and Lai, 1994), ANNs over ARMA (e.g. Hsu et al., 1995), ANNs over linear regression (Elshorbagy et al., 2000a; Panu et al., 2000). Further, the superiority of ANNs over nonlinear regression in predicting river flows has been attributed to the possible existence of nonlinear dynamics, which are not captured by the regression technique (Elshorbagy et al., 2000c). Daily rainfall and streamflows might show large dispersions from a mean motion similar to those exhibited by a stochastic process. This behavior might result either from a random probabilistic structure in the data or from a nonlinear deterministic system highly sensitive to the initial conditions (Rodriguez-Iturbe et al., 1989). In the latter case, the dynamics are deterministic although the appearance is similar to that of a stochastic process. Chaotic systems cannot be distinguished from stochastic processes using conventional statistical tools. Systems are said to be chaotic if they are nonperiodic, sensitive to initial conditions, and long predictability is lost (Grassberger and Procaccia, 1983; Procaccia, 1988).

It is worth to mention that some doubts have been raised about the existence of chaos in hydrologic data (Chilardi and Rosso, 1990). However, the importance of evaluating hydrologic data from the viewpoint of nonlinear deterministic dynamics is stressed by others (Sivakumar et al., 1999a). In reality, describing a time series as either a totally linear stochastic process or fully nonlinear deterministic chaos is not a practical approach. Any time series might have components of both systems. The analyst has to decide whether the process to be modeled is linear stochastic or nonlinear deterministic chaos (Kantz and Schreiber, 1997). On the basis of impossibility of long-term prediction in chaotic time series, the chaotic applications in water resources literature handled only short-term prediction. Cases where lengthy records or consecutive observations are missing, which are common in hydrologic data, have not been addressed.

Analysis of chaotic time series involves calculation of dynamic invariants or dimensions. These parameters are the best available measure to describe and quantify the underlying chaotic system (Angelbeck and Minkara, 1994). Hydrologists applying nonlinear dynamic analysis to their time series use these invariants as a test for chaoticity. Polynomials or nearest neighbors technique, used for prediction, do not benefit from the chaotic invariant measure (i.e. model parameters or configurations are not influenced by the chaotic measure). In stochastic modeling, one can say that type of ARMA models, for example, can be deciphered from autocorrelation and partial autocorrelation functions. In that sense, we can say that stochastic models are available but hydrologic chaotic models cannot be claimed existing.

In this paper, two issues are addressed, first, estimation of missing consecutive observations (missing segment of data) of a chaotic time series. This process has been known and addressed recently as ‘group approach’ (Panu et al., 2000; Elshorbagy et al., 2000a,b). Second is the use of chaotic invariants or dimensions for configuring the hydrologic model. In our application, the dimensions will be used to configure the artificial neural network (ANN) model (determining the number of input and output nodes). These two issues are the contribution of this paper. They have not been explicitly addressed in the available chaos-related water resources literature. In order to address the above-mentioned issues, two steps have to be followed. First, existence of chaos in the time series has to be investigated. Second, chaotic invariants (e.g. correlation dimension) have to be estimated. Third, information obtained from chaos analysis, such as the correlation dimension is used to configure the proposed models. Fourth, missing data are estimated using the configured models.

2. Characteristics of chaotic behavior

2.1. Definitions

A dynamic system can be described by a phase-space diagram that depicts the evolution of the system from some initial state. In fact, what really describe the evolution are the trajectories of the phase-space. The construction of such a phase-space for a time series will be explained in Section 2.2. If the

trajectories converge to a subspace regardless of the initial conditions, then it is called an 'attractor'. An attractor can lie in an m-dimensional phase-space and be a multi-dimensional but has a dimension less than m. Deterministic systems for which long term predictability is possible have attractors of integer dimension (Embrechts, 1994). When the dynamic system is sensitive to initial conditions, the attractors have non-integer or 'fractal' dimensions. Such attractors are called 'strange attractors' and systems containing them are called 'chaotic dynamic systems' (Jayawardena and Lai, 1994).

There are several algorithms available for the analysis of chaotic time series (Casdagli, 1989). The purpose of these algorithms is to calculate geometric and dynamic invariants of an underlying strange attractor, such as correlation dimension and Lyapunov exponents.

2.2. Reconstruction of the phase-space

The first step in the process of chaotic analysis is that of attempting to reconstruct the dynamics in phase-space. A method for reconstructing a phase-space from a single time series has been presented by Takens (1981). The dynamics of a scalar time series $\{x_1, x_2, \ldots x_n\}$ are embedded in the m-dimensional phase-space ($m > d$, where d is the dimension of the attractor). The phase-space is defined by:

$$Y_t = \{x_t, x_{t-\tau}, x_{t-2\tau}, \ldots x_{t-(m-1)\tau}\} \tag{1}$$

where τ is the time delay. Usually, the choice of τ is made with the help of the auto-correlation function or the mutual information content (Frazer and Swinney, 1986). It may be chosen as the lag time at which the autocorrelation becomes zero (Kantz and Schreiber, 1997) or at which the autocorrelation function falls below a threshold value commonly defined as $1/e$ (Tsonis and Elsner, 1988). However, considering various values of τ demonstrates that the results do not show a strong dependence on the actual value chosen (Porporato and Ridolfi, 1997). In this paper, as will be explained later, τ is used also as the maximum length of missing segment of observations that can be estimated in one step (i.e. estimation of a missing value does not depend on the previous missing one) by the proposed technique.

2.3. Correlation dimension

There are few distinct methods for computing fractal dimensions: rescaled range analysis, relative dispersion analysis, correlation analysis, Fourier analysis, maximum likelihood estimator analysis, and the 'Higuchi' method. To estimate the fractal dimension of a time series, the concept of correlation dimension is useful and often applied (Kantz and Schreiber, 1997).

The method of correlation dimension, as explained by Embrechts (1994), consists of centering a hyper sphere around a point in hyperspace or phase-space, letting the radius (r) of the hyper sphere grow until all points are enclosed, and keeping track of the number of data points that are enclosed by the hyper sphere. The slope of the line on a double logarithmic plot will be an estimate of the fractal dimension of the set of data points.

For an m-dimensional phase-space, the correlation integral $C(r)$ is given by Theiler (1986):

$$C(r) = \lim_{N\to\infty} \frac{2}{N(N-1)} = \sum_{\substack{i,j \\ 1\le i\le j\le N}} H(r - |Y_i - Y_j|) \tag{2}$$

where H is the Heaviside step function, with $H(u) = 1$ for $u > 0$, and $H(u) = 0$ for $u \le 0$; N is the number of points of the reconstructed attractor, r is the radius of the sphere centered on Y_i or Y_j. The most commonly used norms for $|Y_i - Y_j|$ are the maximum norm and the standard Euclidean norm. The maximum norm is the maximum absolute difference between the elements of Y_i and Y_j. The Euclidean norm, which is the distance between two points in the space, is adopted in this study, and it has also been used by others (e.g. Jayawardena and Lai, 1994).

If the phenomenon is chaotic, for a large number of points, beyond a certain m, the correlation integral follows the power law:

$$C(r) \underset{\substack{r\to 0 \\ N\to\infty}}{\sim} \alpha r^{\nu} \tag{3}$$

where α is constant; and ν is the correlation dimension, which represents generally good estimate of the fractal dimension d of the attractor. The slope of the $\log C(r)$ versus $\log r$ plot is given by:

$$\nu = \lim_{\substack{r\to 0 \\ N\to\infty}} \frac{\log C(r)}{\log r} \tag{4}$$

For random process, ν varies linearly with increasing m, without reaching a saturation value, whereas for deterministic process the value of ν saturates (levels off) after a certain m. The saturation value, d, is the fractal dimension of the attractor or the time series.

2.4. Lyapunov exponents

The Lyapunov exponents describe the rate at which close points in the phase-space diverge. There is one exponent for each dimension. If one or more Lyapunov exponents are positive, the system is chaotic (Frison, 1994). Therefore, one needs to compute only the maximal Lyapunov exponent. The Lyapunov exponents are invariants with respect to initial conditions. Therefore, they constitute another way of identifying a chaotic system. However, in water resources literature, Lyapunov exponents have been ignored by researchers as a necessary indication of chaotic behavior (Sivakumar et al., 1998, 1999a; Porporato and Ridolfi, 1997).

Consider the representation of the time series data as a trajectory in the embedding space. Assuming that, one observes a close return $s_{n'}$ to a previously visited point s_n, then one can consider the distance $\Delta_0 = s_n - s_{n'}$ as a small perturbation, which should grow exponentially in time. Its future can be read from the time series: $\Delta_l = s_{n+l} - s_{n'+l}$. If one finds that $|\Delta_l| \cong \Delta_0 \, e^{\lambda l}$, then λ is the maximum Lyapunov exponent. The routine given by Kantz and Schreiber (1997) is used to calculate the maximal Lyapunov exponent in this study.

2.5. Kolmogorov entropy

Entropy is a thermodynamic quantity describing the amount of disorder in the system. It can characterize the amount of information needed to predict the next measurement with a certain precision. The most popular one is the Kolmogorov entropy. The Kolmogorov entropy of a time series gives a lower bound to the sum of the positive Lyapunov exponents. An estimate of the Kolmogorov entropy (K) is K_2 (Jayawardena and Lai, 1994).

$$K_2(m) \cong \lim_{r \to 0} \left(\frac{1}{\Delta t} \{ \log [C_m(r)] - \log [C_{m+1}(r)] \} \right) \quad (5)$$

$$K_2 \cong \lim_{m \to \infty} [K_2(m)] \quad (6)$$

where Δt is the time interval between two successive observations, K_2 is expected to be zero for regular system (e.g. periodic), positive and finite for chaotic systems and infinite for stochastic process.

2.6. Method of surrogate data

Another way of supporting the argument that a specific data set is coming from a nonlinear deterministic system might be by rejecting the hypothesis that it is coming from a linear process. The method of surrogate data (Theiler et al., 1992) makes use of the substitute data generated in accordance to the probabilistic structure underlying the original data. The null hypothesis consists of a candidate linear process and the objective is to reject the hypothesis that the original data have come from a linear stochastic process. A null hypothesis is formulated, for example, that the data have been created by a stationary Gaussian linear process. Then, it is attempted to reject this hypothesis by comparing results for the data to appropriate realizations of the null hypothesis. Sivakumar et al. (1999a) follow the algorithm provided by Theiler et al. (1992) and the significance of a discriminating statistic obtained for surrogate data is judged. In Theiler et al. (1992), it is mentioned that a value of $\sim$2 of the statistic cannot be considered significant whereas a value of $\sim$10 is highly significant. It is not explained how the authors made an inference from these values and on what statistical basis they derived the conclusion.

Since the primary interest is to give a support to the identification of chaos, in time series, provided by the previous invariants, one of the discussed invariants would be used for visual inspection and comparison. In this study, the correlation dimension for different m of both original and surrogate data will be plotted and inspected. Since the null assumption leaves room for free parameters, the process of generating the surrogate data has to take these into account (Hegger et al., 1999). One approach is to construct constrained realizations of the null hypothesis. This approach of constrained realizations is adopted, for generating the surrogate data, using the algorithm provided by Kantz and Schreiber (1997). Constrained realizations are obtained by randomizing the data subject to the constraint that an appropriate set of parameters remains fixed (e.g. random data with a given periodogram can be made).

Table 1
Structure of the reconstructed time series data

	1	2	⋯	$m-1$	m
1	x_1	$x_{1+\tau}$		$x_{1+(m-2)\tau}$	$x_{1+(m-1)\tau}$
2	x_2	$x_{2+\tau}$		$x_{2+(m-2)\tau}$	$x_{2+(m-1)\tau}$
⋮					
k	x_k	$x_{k+\tau}$		$x_{k+(m-2)\tau}$	$x_{k+(m-1)\tau}$
⋮					⋮
$k+\tau-1$	$x_{k+\tau-1}$			$x_{k-1(m-1)\tau}$	$x_{k+\tau-1+(m-1)\tau}$
⋮					
$n-\tau(m-1)$	$x_{n-\tau(m-1)}$	$x_{n-\tau(m-2)}$		$x_{n-\tau}$	x_n

More details of this technique can be found in Kantz and Schreiber (1997) and Hegger et al. (1999).

3. Estimation of missing data

Researchers have been tackling the problem of missing data in different ways and from different perspectives as well. Even their definitions of 'missing data' and the expressions that they have used to describe the in-filling process are no less diversified than the different techniques that they have used. A group of researchers tackled the problem of intermediate missing data where data or observations before and after the missing observations are available (e.g. Bennis et al., 1997). Others consider the cases in which data are available only from one side of the gap or the gap is so lengthy that the data set is considered bounded from one side only. Generally, the in-filling process in this case is called 'extension' (e.g. Hughes and Smakhtin, 1996). This paper explores the first type, where data before and after the gap are available. Extensive literature review on the existing techniques for estimation of missing data can be found in Elshorbagy et al. (2000a) where the problem of estimation of missing groups (consecutive observations) in streamflow records has been reported and explained.

Following the traditional approach for estimating missing values will lead to the use of previously estimated observations to estimate the successive one. This approach, which is similar in concept to prediction of flows for few steps ahead, contradicts with the concept of chaotic behavior. When a time series is investigated and proved to be chaotic, such type of long-term prediction or estimation of missing segments should not be valid.

Using the technique of reconstruction of attractor, the time series is presented as shown in Table 1. A segment of missing consecutive data can be confined to the last column, m. The data set will be structured so that each observation in column m can be estimated based on the previous $m-1$ columns. The missing segment, indicated in Table 1 inside a rectangle, can be as lengthy as τ. In this way, the missing τ observations can be estimated using the values from x_k till $x_{k-1+(m-1)\tau}$, which do not include any of the observations from $x_{k+(m-1)\tau}$, till $x_{k+\tau-1+(m-1)\tau}$ (that are supposed to be missing).

In this study, ANNs, as global approximators, are used for estimating the missing data. The number of input nodes will be $m-1$, output nodes will be equal to one, and the hidden nodes will be obtained using trial and error. A conceptual advantage of this configuration is that the information obtained about 'm' from the correlation dimension computation is used for configuring the ANN model. Therefore, correlation dimension is not only used as an invariant indicator of chaos. The theoretical possibility of predicting the mth dimension, using the previous $m-1$ dimensions, is indicated by Kantz and Schreiber (1997).

Another technique for making nonlinear prediction or estimation of missing data is the K-nearest neighbor (K-nn) algorithm. It is a representative of the local approximation method, which uses only nearby states to make prediction. The local approximators are always believed to provide good results in chaotic time series. Therefore, it is widely used in chaos literature (e.g. in water resources, Porporato and Ridolfi, 1997; Sivakumar et al., 1999a). To estimate x_{i+t} based on Y_i (m-dimensional vector) and historical observations, K-nearest neighbors of Y_i are found

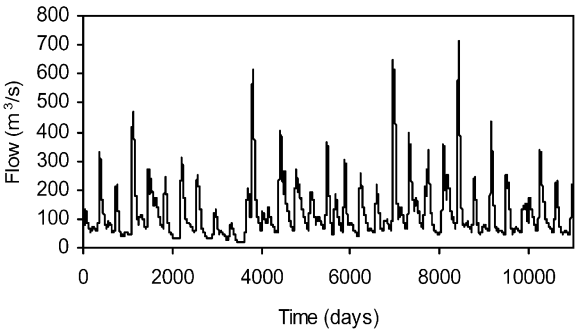

Fig. 1. Variation of daily flows of the English River.

based on the minimum distances $|Y_i - Y_j|$. For K number of neighbors, the estimation of x_{i+t} can be taken as the average of the K values of x_j.

Considering the data structure shown in Table 1, the subspace estimation method is employed here. The m-dimensional points are projected on the first, completely known, $m-1$ dimensions. The nearest neighbors to the vectors in the rows from (k) to $(k+\tau-1)$ can be calculated using Euclidean norm. Note that the mth dimension of the vectors from (k) to $(k+\tau-1)$ is the missing segment. A similar technique is used in pattern recognition problems and is known as subspace classification (Querios and Gelsema, 1988). In pattern recognition, the objective is only to classify objects, therefore the problem is projected on $m-1$ dimensions (features) and objects are classified based on the complete features regardless of the incomplete one. In our application, once the nearest neighbors are identified, their mth dimension (which is known) is used for estimation of the missing mth dimension of the vectors from (k) to $(k+\tau-1)$.

It is reported by Casdagli (1991) that when smaller number of neighbors give the most accurate estimation or short-term prediction compared to that of larger number of neighbors, then this may be considered as a strong evidence for low-dimensional chaos in the data.

4. Analysis and results

In this paper, the streamflow data from English River at Umferville, Ontario, Canada is used and analyzed to investigate the possible existence of chaotic behavior. The average daily flow is 124.3 m^3/s and the standard deviation is 92.3 m^3/s. Fig. 1 shows the variation of daily flows obtained from Umferville station. The issue related to length of the data record that is considered sufficient for chaos analysis has been addressed by many authors. While Frison (1994) suggests a minimum of ~10,000 data points for reliable results, others accepted number of points as low as 1200 observations (Jayawardena and Lai, 1994) and 1500 observations (Sivakumar et al., 1999a). In our application, 11,000 observations are used.

It should be noted that the complete data record of the English River is available. However, an intermediate data section of 1100 observations length (110 segments, each of 10 observations length) is arbitrary chosen to test the proposed models and the modeling

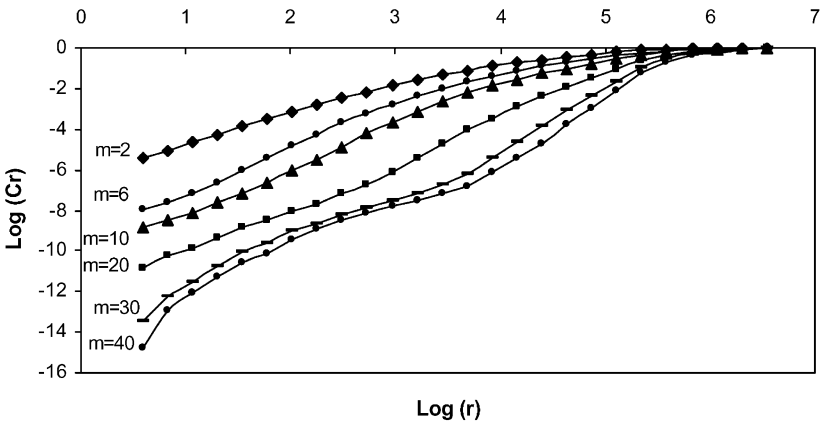

Fig. 2. Correlation integral of the English River.

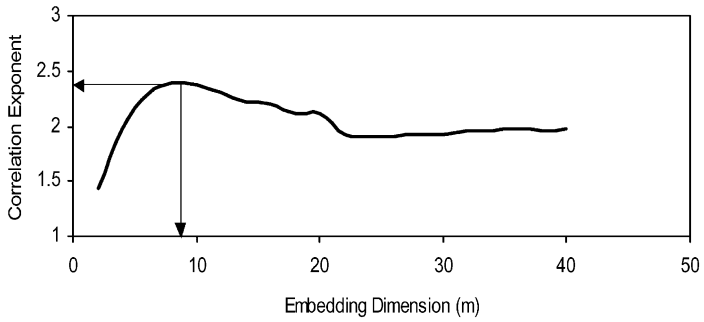

Fig. 3. Correlation exponent of the English River.

approach. The 110 segments are removed, each at a time, and assumed missing. After estimating the missing segments, each at a time, the accuracy of estimating the missing data is calculated by averaging the error over the 1100 estimated observations.

4.1. Correlation dimension

Daily flows comprising of 11,000 observations of the English River are considered for investigating the existence of chaotic behavior in the streamflow. Value of lag time 10 (i.e. 10 days) is considered in the analysis with the purpose of estimating (in-filling) 10 consecutive missing observations. The correlation integral (sum) $C(r)$ and the correlation exponent ν are computed, as explained earlier, from the data set. The relationship between the correlation integral $C(r)$ and the radius r for various values of embedding dimensions m is shown in Fig. 2. The correlation exponent increases with the increase in the embedding dimension up to a certain point ($m = 8$) and saturates beyond that point (see Fig. 3). The saturation value of the correlation exponent (dimension) is ~2.4. The nearest integer above the correlation dimension value ($d = 3$) is taken as the minimum dimension of the phase-space that can embed the attractor. The value of m at the saturation point ($m = 8$) is supposed to provide the sufficient number of variables to describe the dynamics of the attractor.

4.2. Lyapunov exponent

Using the algorithm of Kantz and Schreiber (1997), the largest Lyapunov exponent is found to be positive (9.1×10^{-3}). It should be noted that it can be positive also for some random and ARMA processes as observed by Jayawardena and Lai (1994) and Rodriguez-Iturbe et al. (1989), when calculated using the algorithm of Wolf et al. (1985). Therefore, results of Lyapunov exponent computation are not recommended to be taken as a sole indication of chaotic behavior.

4.3. Entropy

A positive finite K_2 value can be found for the data set under consideration. Fig. 4 shows the relationship

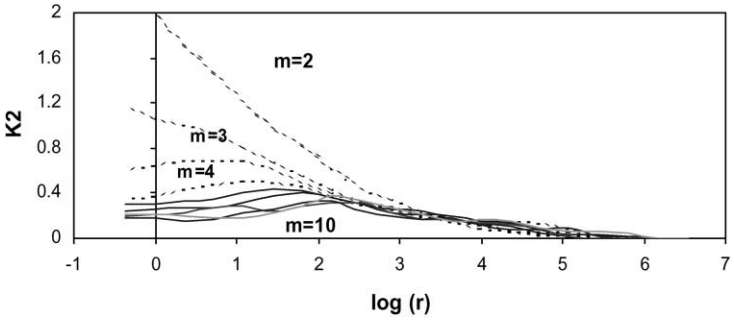

Fig. 4. Estimate of correlation entropy of the English River.

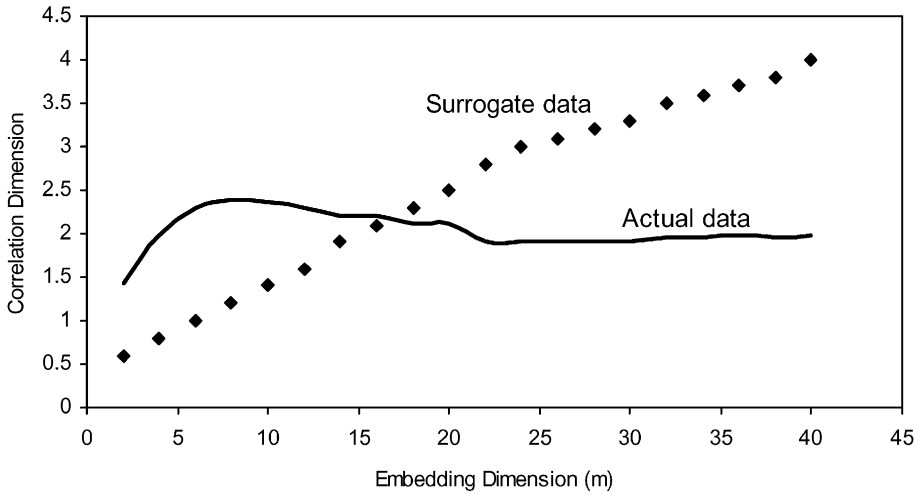

Fig. 5. Correlation dimensions of both English River data set and its surrogate.

between the correlation entropy and $\log(r)$ for various m. A plateau can be observed at K_2 value $\sim$0.2. This type of visual inspection of K_2 entropy is used and reported by Kantz and Schreiber (1997).

4.4. Method of surrogate data

Several realizations of surrogate data sets can be generated according to the probabilistic structure underlying the original data and also according to the null hypothesis as discussed earlier. The major aim of this step is to detect nonlinearity. Fig. 5 shows the relationship between correlation dimension and embedding dimension for the original data and one of the surrogate data sets. The purpose of this process is to demonstrate the difference in behavior between the two data sets. In the case of the original data of the English River, the correlation exponent curve levels off at a certain point (i.e. $m = 8$) whereas, the correlation exponents computed for the surrogate data continue increasing monotonically with the increase in embedding dimension. In this paper, the visual difference between the two data sets is considered sufficient to indicate that original data might not come from a linear stochastic process.

4.5. Estimation of missing segments

Two techniques are employed in this paper for estimating the missing data, ANNs as a global approximator and K-nearest neighbors (K-nn) as a local approximator.

4.5.1. Artificial neural networks

Feed forward neural networks that employ the back propagation technique, for training the network, are

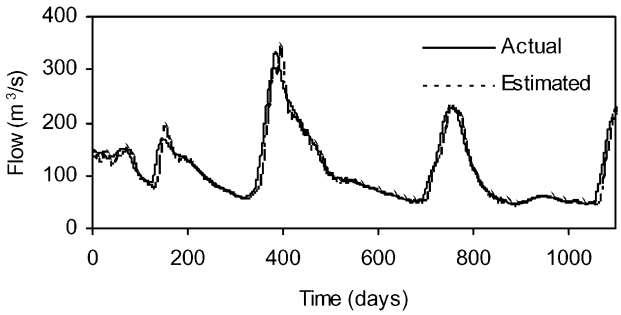

Fig. 6. Actual and estimated data using ANN technique.

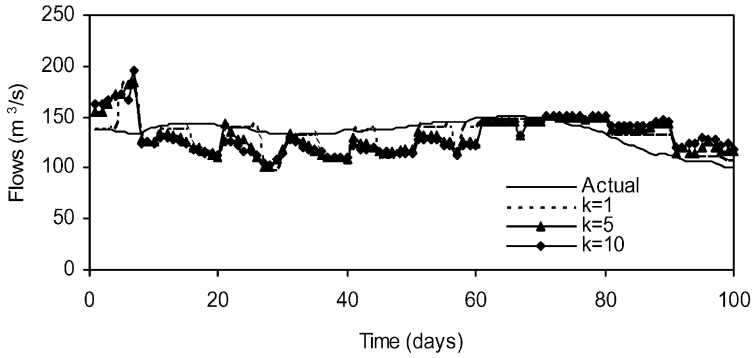

Fig. 7. Actual and estimated data using *K*-nn technique.

used. The structure of the input and output data of the ANNs is as shown in Table 1, seven ($m-1$) inputs and one (mth) output. Around 90% of the data are used for training and 10% for testing the results of the network. Many network configurations are tried and satisfactory results are achieved with the (7–3–1) configuration (i.e. seven input nodes, three hidden nodes, and one output node). Fig. 6 shows the actual and estimated flows using the ANN model. Note that the objective is to estimate 10 consecutive missing observations.

4.5.2. K-nearest neighbor (K-nn) technique

The *K*-nearest neighbors technique is used to estimate the missing data in the same way explained earlier. Many values of *K* can be tried, and a satisfying value of *K* is usually obtained by trial and error. In this paper, our purpose is to show the effect of increasing the value of *K* on the results rather than obtaining the optimum *K*. Therefore, three values of *K* are arbitrarily considered: 1, 5, and 10. Other values can be assumed but the selected three values are believed to be sufficient for representing the *K*-nn technique. Fig. 7 shows the actual and estimated flows using the *K*-nn technique. A segment of 100 observations are selected for the graphical illustration in Fig. 7. The estimated values using the *K*-nn technique may appear in the figure as broken line with consecutive points of constant magnitude that form a segment. The reason for such appearance is that *K*-nn estimates a value based on the closest point or points in space. Those points, which are the closest to consecutive observations, might be constant giving the appearance of that horizontal segments.

Table 2
Mean squared and mean relative error of estimated data

Technique	Mean squared error	Mean relative error
ANNs	80.2	0.06
K-nn ($K=1$)	255.5	0.08
K-nn ($K=5$)	365.9	0.12
K-nn ($K=10$)	441.5	0.14

4.5.3. Discussion

The results shown in Figs. 6 and 7 are summarized in Table 2. This table indicates that the lower the *K*-value, the better is the estimate of the missing observations (lower mean squared error and mean relative error). This result may give another support to the previously discussed invariants (e.g. correlation dimension) that indicates the existence of chaos. Fig. 8 shows the actual streamflows and those estimated by ANN and *K*-nn (using $K=1$) models. Both, Table 2 and Fig. 8 indicate that ANNs are superior to *K*-nn for estimating the missing observations. Such superiority of ANN may be problem-related and need extensive applications on various data sets to be generalized. However, one can say that the superiority of ANNs might be attributed to the ability of ANNs to capture the nonlinear dynamics of the data. Such a characteristic of ANNs is indicated by others (Panu et al., 2000; Elshorbagy et al., 2000a,b,c). Furthermore, the way ANNs are used in this paper makes use of the identified embedding dimension, which means that the attractor is modeled directly using its first seven dimensions. It seems that the ANNs are

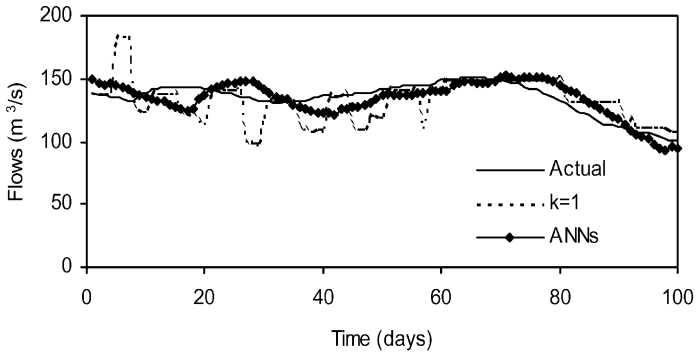

Fig. 8. Actual and estimated data using ANN and *K*-nn techniques.

able to generalize the structure of the attractor throughout the whole data set.

On the other hand, the subspace technique used with the *K*-nn could not benefit significantly from the information related to the embedding dimension ($m = 8$). This could be due to the aggregation that happens when the first seven dimensions of the attractor are represented by one number (Euclidean norm).

It should be noted that some recent studies have indicated that the noise that exists in hydrologic data may limit the performance of many modeling technique. Some methods have been proposed to reduce the level of noise in the data set (Sivakumar et al., 1999b; Sivakumar, 2000), which may lead to improvement in the accuracy of the estimation of missing data. The issue of noise in chaos analysis has been extensively discussed in Elshorbagy et al., (2001), where it is shown that the raw data should always be used for the hydrologic analysis.

5. Conclusion

In this paper, the existence of chaotic behavior (nonlinear deterministic dynamics) in the daily flows of the English River is investigated. The correlation dimension, the Lyapunov exponent, the Kolmogorov entropy, and the method of surrogate data are used in the analysis. There are sufficient indications to believe that the streamflows have some chaos and the data could be modeled by the time delay embedding method. On the basis of the attractor dimension, the minimum number of variables essential to model the dynamics of the daily flows of the English River is identified as three and the number of variables sufficient is eight. The sufficient number of variables is used to configure the ANN model. Seven input nodes (first seven dimensions) are used to estimate the output (eighth dimension). The data are structured in a way that facilitates the estimation of 10 consecutive missing observations in one step (previously estimated value is not used for estimating the following one). Also, the *K*-nearest neighbor technique is used, with $K = 1$, 5, and 10, to estimate the missing data. A subspace modeling approach is adopted by projecting the reconstructed attractor on a lower dimension scale (i.e. projecting the eight-dimension attractor on its first known seven-dimension space). The eighth dimension is estimated using the previous seven dimensions. The ANN model shows superiority in the accuracy of estimating the missing data, which is attributed to the capability of the ANNs to capture the nonlinear dynamics and generalize the structure of the attractor on the whole data set. Finally, this work is considered as an endeavor towards establishing hydrologic chaotic modeling by using the chaos indicators (correlation dimension) directly in the process of modeling or configuring the data model.

Acknowledgements

The authors wish to acknowledge the financial support given to this research by the Natural Sciences and Engineering Research Council (NSERC) of Canada through grant no. OGP-0004404. Also, the first author wishes to express his appreciation to the

University of Manitoba for the Graduate Fellowship awarded to him.

References

Angelbeck, D.I., Minkara, R.Y., 1994. Strange attractors and chaos in wastewater flow. J. Environ. Engng, ASCE 120 (1), 122–137.

Bennis, S., Berrada, F., Kang, N., 1997. Improving single-variable and multivariable techniques for estimating missing hydrological data. J. Hydrol. 191, 87–105.

Casdagli, M., 1989. Nonlinear prediction of chaotic time series. Physica D 35, 335–356.

Casdagli, M., 1991. Chaos and deterministic versus stochastic nonlinear modeling. J. Royal Stat. Soc. B 54 (2), 303–328.

Chilardi, P., Rosso, R., 1990. Comment on Chaos in rainfall by I. Rodriguez-Iturbe et al. Water Resour. Res. 26 (8), 1837–1839.

Elshorbagy, A., Panu, U.S., Simonovic, S.P., 2000a. Group-based estimation of missing hydrological data. I. Approach and general methodology. Hydrol. Sci. J. 45 (6), 849–866.

Elshorbagy, A., Panu, U.S., Simonovic, S.P., 2000b. Group-based estimation of missing hydrological data. II. Application to streamflows. Hydrol. Sci. J. 45 (6), 867–880.

Elshorbagy, A., Simonovic, S.P., Panu, U.S., 2000c. Performance evaluation of artificial neural networks for runoff prediction. J. Hydrol. Engng, ASCE 5 (4), 424–427.

Elshorbagy, A., Simonovic, S.P., Panu, U.S., 2001. Noise reduction in chaotic hydrologic time series: Facts and doubts. J. Hydrol., in press.

Embrechts, M., 1994. Basic concepts of nonlinear dynamics and chaos theory. In: Deboeck, G.J. (Ed.). Trading on the Edge: Neural, Genetic, and Fuzzy Systems for Chaotic Financial Markets. Wiley, New York, pp. 265–279.

Frazer, A.M., Swinney, H.L., 1986. Independent coordinates for strange attractors from mutual information. Phys. Rev. A 33 (2), 1134–1140.

Frison, T., 1994. Nonlinear data analysis techniques. In: Deboeck, G.J. (Ed.). Trading on the Edge: Neural, Genetic, and Fuzzy Systems for Chaotic Financial Markets. Wiley, New York, pp. 280–296.

Grassberger, P., Procaccia, I., 1983. Measuring the strangeness of strange attractors. Physica D 9, 189–208.

Hegger, R., Kantz, H., Schreiber, T., 1999. Practical implementation of nonlinear time series methods: the TISEAN package. Chaos 9, 413–440.

Hsu, K.L., Gupta, H.V., Sorooshian, S., 1995. Artificial neural network modeling of the rainfall–runoff process. Water Resour. Res. 31 (10), 2517–2530.

Hughes, D.A., Smakhtin, V., 1996. Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves. Hydrol. Sci. J. 41 (6), 851–871.

Islam, S., Bras, R.L., Rodriguez-Iturbe, I., 1993. A possible explanation for low correlation dimension estimates for the atmosphere. J. Appl. Meteor. 32, 203–208.

Jayawardena, A.W., Gurung, A.B., 2000. Noise reduction and prediction of hydrometeorological time series: dynamical systems approach vs. stochastic approach. J. Hydrol. 228, 242–264.

Jayawardena, A.W., Lai, F., 1994. Analysis and prediction of chaos in rainfall and streamflow time series. J. Hydrol. 153, 23–52.

Kantz, H., Schreiber, T., 1997. Nonlinear Time Series Analysis. Cambridge University Press, UK.

Lall, U., Sangoyomi, T., Abarbanel, H.D.I., 1996. Nonlinear dynamics of the great salt lake: nonparametric short-term forecasting. Water Resour. Res. 32 (4), 975–985.

Panu, U.S., Khalil, M., Elshorbagy, A., 2000. Streamflow data infilling techniques based on concepts of groups and neural networks. In: Govindaraju, R.S., Rao, R. (Eds.). Artificial Neural Networks in Hydrology. Kluwer, The Netherlands, pp. 235–258 chapter 12.

Porporato, A., Ridolfi, L., 1997. Nonlinear analysis of river flow time sequences. Water Resour. Res. 33 (6), 1353–1367.

Procaccia, I., 1988. Complex or just complicated?. Nature 333 (9), 498–499.

Puente, E., Obregon, N., 1996. A deterministic geometric representation of temporal rainfall: results for a storm in Boston. Water Resour. Res. 32 (9), 2825–2839.

Querios, C.E., Gelsema, E.S., 1988. Incomplete data sets. In: Gelsema, E.S., Kanal, L.N. (Eds.). Pattern Recognition and Artificial Intelligence. Elsevier, Amsterdam, pp. 237–255.

Rodriguez-Iturbe, I., De Power, B.F., Sharifi, M.B., Georgakakos, P.K., 1989. Chaos in rainfall. Water Resour. Res. 25 (7), 1667–1675.

Sangoyomi, T.B., Lall, U., Abarbanel, H.D.I., 1996. Nonlinear dynamics of the great salt lake: dimension estimation. Water Resour. Res. 32 (2), 149–159.

Sharifi, M.B., Georgakakos, K.P., Rodriguez-Iturbe, I., 1990. Evidence of deterministic chaos in the pulse of storm rainfall. J. Atmos. Sci. 47 (7), 888–893.

Sivakumar, B., 2000. Chaos theory in hydrology: important issues and interpretations. J. Hydrol. 227, 1–20.

Sivakumar, B., Liong, S., Liaw, C., 1998. Evidence of chaotic behaviour in Singapore Rainfall. J. Am. Water Resour. Assoc. 34 (2), 301–310.

Sivakumar, B., Liong, S., Liaw, C., Phoon, K., 1999a. Singapore rainfall behavior: chaotic?. J. Hydrol. Engng, ASCE 4 (1), 38–48.

Sivakumar, B., Phoon, K., Liong, S., Liaw, C., 1999b. A systematic approach to noise reduction in chaotic hydrological time series. J. Hydrol. 219, 103–135.

Takens, F., 1981. Detecting strange attractors in turbulence. Dynamical Systems and Turbulence, Rand, D.A., Young, L.S. (Eds.). Warwick, Lect. Notes Math., 366–381.

Theiler, J., 1986. Spurious dimension from correlation algorithms applied to limited time-series data. Phys. Rev. A 34, 2427–2432.

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Farmer, J.D., 1992. Testing for nonlinearity in time series: the method of surrogate data. Physica D 58, 77–94.

Tsonis, A.A., Elsner, J.B., 1988. The weather attractor over very short time scales. Nature 333, 545–547.

Wolf, A., Swift, J.B., Swinney, H.L., Vastano, J.A., 1985. Determining Lyapunov exponents from a time series. Physica D 16, 285–317.