

Effect of noise in nonlinear hydrological time series analysis and prediction

A. W. JAYAWARDENA & A. B. GURUNG

Department of Civil Engineering, The University of Hong Kong, Hong Kong, China
e-mail: hrecjaw@hkucc.hku.hk

Abstract Hydrological time series are considered as the outcome of deterministic systems, which may become “chaotic” at times. Identification of such systems requires the data to be noise free. In this study three nonlinear noise reduction techniques have been applied to two sets of hydrological data: daily river flow and sea surface temperature anomaly index (*S*-index). The correlation dimensions have been computed and predictions made before and after noise reduction. The results show that convergence in the correlation dimension as well as increase in the prediction accuracy can be achieved in the noise-reduced data.

INTRODUCTION

Linear approaches using autoregressive moving average (ARMA) type of models have traditionally been used for the analysis and prediction of hydrological time series. Recently however, it has been realised that certain types of time series, which appear to be evolving from stochastic processes, can in fact be the outcome of fully deterministic processes. By treating the system that generates the time series as a deterministic one, it is possible to gain an understanding of the associated complicated dynamics, and to make more realistic short-term predictions. Such systems can exhibit stable properties, which are predictable with certainty at times but may become “chaotic” under certain initial conditions. The study of “chaotic” systems has drawn the attention of many researchers in many disciplines in the recent past.

Many of the developments related to the analysis and prediction of chaotic systems have come from disciplines like mathematics and physics. The application to hydrology is still in its infancy. In this study, an attempt is made to highlight the relevance, application and associated problems from the point of view of hydrological prediction.

The first step in treating a time series as chaotic is to diagnose the system; i.e. to determine whether the time series is driven by a low dimensional deterministic system. It can be done by computing several invariant measures, such as the correlation dimension, the Lyapunov exponent, the Kolmogorov entropy among others. Once a particular time series has been identified as chaotic, short-term predictions can be obtained by a number of modelling techniques.

There are, however, problems associated with the computation of invariant measures. Many of the practical time series are often corrupted by “noise”, which must first be removed before any diagnostic tests can be done. The focus of attention in this study is on how to deal with noise present in a time series and to examine noise-reduction measures. The efficiency of several nonlinear noise-reduction techniques is

first analysed using artificially corrupted data series by computing the noise level and the signal-to-noise ratio. Series which are known to be chaotic have been used as examples. The procedures are then applied to a river flow data series and a sea surface temperature anomaly index. Predictions made after noise reduction when compared with raw/noisy data sets show significant improvement.

PHASE-SPACE AND CORRELATION DIMENSION

A dynamical system can be described by a phase-space diagram whose trajectories describe its evolution from some initial state, which is assumed to be known. If the trajectories converge to a single sub-space regardless of the initial conditions, then it is called an *attractor*, which can be multi-dimensional, and lies in an m -dimensional phase-space but has dimension less than m .

The dynamics of a time series x_1, x_2, \dots, x_n are fully captured or embedded in the m -dimensional phase-space ($m \geq d$ where d is the dimension of the *attractor*) defined by the vector:

$$\mathbf{Y}_t = \{x_t, x_{t-\tau}, x_{t-2\tau}, \dots, x_{t-(m-1)\tau}\} \quad (1)$$

where τ is delay time. According to the embedding theorem (Takens, 1981), a d -dimensional attractor can be embedded into a $(2d+1)$ -dimensional phase-space to evaluate the characteristics of the dynamical system.

The delay time τ may be chosen as the lag time at which the autocorrelation falls below a threshold value, which is commonly defined as $1/e$, specially if the autocorrelation function is approximately exponential (Tsonis & Elsner, 1988). Another method is to use the lag time at which the autocorrelation first becomes zero if it crosses the zero line (Mpitsos *et al.*, 1987).

Grassberger & Procaccia (1983a,b) defined the correlation integral $C(r)$ as:

$$C(r) = \frac{1}{N_{\text{ref}}} \sum_j \frac{1}{N} \sum_i H(r - \|\mathbf{Y}_i - \mathbf{Y}_j\|) \quad i \neq j \quad (2)$$

where H is the heaviside step function with $H(u) = 1$ for $u > 0$, and $H(u) = 0$ for $u \leq 0$; N is the number of points in the vector time series $\{\mathbf{Y}(t)\}$; $N_{\text{ref}} (\leq N)$ is the number of reference points taken from the vector time series $\mathbf{Y}(t)$; r is the radius of sphere centred on either of the points $\{\mathbf{Y}_i\}$ or $\{\mathbf{Y}_j\}$. The norm $\|\mathbf{Y}_i - \mathbf{Y}_j\|$ may be any of the three usual norms, the maximum norm, the diamond norm, or the Euclidean norm. The maximum norm is used in this study for the computation of the correlation dimension. Correlation integrals are calculated for a series of embedding dimensions.

If an attractor for the system exists, then, for small r , it can be shown that

$$C(r) \cong r^d \quad (3)$$

where d is the correlation exponent. It may be estimated by the slope of a straight line in the plot of $\log(C(r))$ vs $\log(r)$ for each value of m . For random processes, d varies linearly with increasing m without reaching a saturation value, whereas for deterministic processes, the value of d levels off after a certain m . The saturation value of d is defined as the correlation dimension of the attractor or a time series. The nearest

integer above the saturation value d provides the minimum number of embedding dimensions of the phase-space necessary to model the dynamics of the attractor.

Although it is quite easy to implement the algorithm of Grassberger & Procaccia (1983a,b) for noise-free data, it is not the case for noisy data sets. Noise blurs the lower region of the length and causes deviation of the trajectory in all lengths increasing the inter-point distance in the plot of correlation integral vs length. The usual procedure for reconstructing the phase space also would not work in the presence of noise.

Other methods of determining the correlation dimension include that proposed by Theiler (1987) in which the whole attractor's extent is divided into several boxes (grids) and all points are assigned to the boxes, and the method proposed by Schouten *et al.* (1994) built into the software RRCHAOS in which the increase in inter-point distance is assumed to be due to noise.

NOISE REDUCTION

In practice, all experimental or field observed time series are contaminated by noise. Most techniques for computing invariant measures such as the correlation dimension fail if the data contains as little as 2% noise (Schreiber, 1993a). The noise present in time series is measured by the absolute noise level $\langle v^2 \rangle$, where $\langle v^2 \rangle$ is defined as:

$$\langle v^2 \rangle = \frac{1}{N} \sum_{n=1}^N v_n^2 \quad (4)$$

or by the signal-to-noise ratio (SNR) in dB units, defined as:

$$SNR = 20 \log \left(\frac{\langle y^2 \rangle}{\langle v^2 \rangle} \right)^{1/2} \quad (5)$$

where v_n is the noise component (additive) of a noisy data series x_n , of which the clean or deterministic component is y_n .

Linear approaches of noise reduction such as Fourier transformation, differencing etc. are not useful for nonlinear time series. In this study, several existing techniques of nonlinear noise reduction were used. They include the "0th order" method of noise reduction (Schreiber, 1993b) in which the time series is embedded with the embedding vector to include equal number of past and future values and a present value, and for which a set of nearest neighbours are found inside a specific radius, the linear approximation method of Schreiber & Grassberger (1991), and the "local projection" method of Grassberger *et al.* (1993) which has been encoded as a Fortran program by Kantz & Schreiber (1997).

PREDICTION

In a deterministic system, predictions can generally be made using an evolutionary equation of the form:

$$y_{n+1} = f_T(y_n) \quad (6)$$

where y_{n+1} is the predicted value which is dependent upon present and past values y_n . The prediction process therefore involves an accurate estimation of f_T . In the case of chaotic systems, the predictive power is lost very quickly because of sensitivity to the initial conditions.

The function f_T can be estimated using local models in which the function approximation at each time step is done from data sets of the local neighbourhood only in a piecewise manner, or global models in which the function approximation is done for the whole domain. Local models include linear or polynomial function approximations in the local neighbourhoods whereas global models are generally of the polynomial type although radial basis functions also have been used. In this study, the function approximation is done using the "0th order" predictor (Farmer & Sidorowich, 1987) in which the prediction is done on the basis of the behaviour of the series in the closest neighbourhood of the vector time series \mathbf{x}_t which contains the current value x_t .

APPLICATION AND RESULTS

Some preliminary studies on the noise-reduction techniques referred to earlier were carried out on artificially corrupted series which are otherwise fully deterministic. They include the Sine map, which is a regular periodic series, Henon map and Lorenz system which are known to be chaotic and have well defined phase space trajectories and known correlation dimensions. All the methods referred to above were able to reduce the noise efficiently as evidenced by their phase space trajectories and the correlation dimensions. Substantial increases in the *SNR* were also observed for these series after noise removal.

Similar attempts were made for two real data series as well. The first series consists of average daily stream flow data at Nakhon Sawan gauging station (15.67°N, 100.12°E) of the Chao Phraya River in Thailand (catchment area at the gauging station: 110 569 km²) for the period April 1978–March 1994 ($N = 5844$). The second data series is the monthly mean sea surface temperature (*SST*) anomaly from 1872–1986 ($N = 1380$), averaged over the region bounded approximately between 6°N–6°S and 180°W–90°W. This has been defined as the *S* index (Wright, 1989) to identify climatic anomalies attributed to the El Niño Southern Oscillation. The former data series was taken from the Global Runoff Data Centre (GRDC) in Germany while the latter was taken from a table compiled by Wright (1989).

The software RRCHAOS (Shouten *et al.*, 1994), which does not yield the clean series but computes the correlation dimension from a noisy data set, was used with the raw data. These would be used as reference values for comparison purposes. Noise reduction was carried out by three methods (referred to as Methods I, II, and III in Table 1, which are respectively based on the approaches proposed by Schreiber (1993b), Schreiber & Grassberger (1991) and Grassberger *et al.* (1993)). The correlation dimensions of the "clean" series were again computed by RRCHAOS and are shown in Table 1. An attempt to compute the correlation dimensions by the method proposed by Theiler (1987) did not lead to saturation values for the raw data but gave slightly lower values for the "clean" data (Table 1).

Table 1 Correlation dimension of data series.

Data series	Correlation dimension						
	Raw data using (RRCHAOS)	Noise reduced data using: Method I (RRCHAOS) Theiler		Method II (RRCHAOS) Theiler		Method III (RRCHAOS) Theiler	
Chao Phraya	2.90	4.633	2.93	3.976	2.33	4.956	1.87
S-index	3.85	5.336	2.67	2.428	1.90	4.314	1.96

Table 2 Noise level and signal to noise ratio, *SNR*.

Data series	Noise level of raw data	Noise reduced by: Method I		Method II		Method III	
		Noise level	<i>SNR</i>	Noise level	<i>SNR</i>	Noise level	<i>SNR</i>
Chao Phraya	22.0	15.0	30.38	8.0	36.897	0.65	17.077
S-index	28.6	4.90	2.976	1.25	3.996	0.37	0.0361

Table 2 shows noise levels of the raw and cleaned series by the three methods of noise reduction. The noise levels were determined by the method of Schreiber (1993a). One of the problems in dealing with real data is that the clean data are never known. Some noise reduction techniques may not remove the noise completely, whereas others may overdo and remove parts of the clean series thereby distorting the deterministic features. The latter effect is more serious than not removing the noise at all. To the best of the authors' knowledge, there is no solution to this problem at the present time. However, a measure of the *SNR* is estimated in this study on the assumption that the estimated "clean" series is the real clean series. The values calculated using equation (5) are given in Table 2. A real test of the reliability of a noise-reduction method would need to be based on the prediction accuracy. Methods II and III appear to be performing equally well in this respect.

The "0th order" predictor was used to make predictions of the two time series. The original method proposed by Farmer & Sidorowich (1987) used the nearest neighbour as the predictor. Prediction based on averaging more than one nearest neighbour was adopted in this study as it could be expected to give better results. This approach requires the determination of the number of neighbours, the embedding dimension and the neighbourhood size to be fixed in advance. The embedding dimension was adopted according to Taken's (1981) theorem. The other two parameters were adjusted by trial and error to give the maximum correlation between the predicted and the observed. Predictions were made for raw data as well as noise reduced data. Some typical comparisons are shown in Figs 1–4. An assessment of the accuracy of prediction was made by computing the correlation coefficients between the predicted and the observed values, and the normalized mean square error (*NMSE*) given by:

$$NMSE = \frac{1}{P\sigma^2} \sum_{n=i}^P \left\{ (x_n)_{\text{obs}} - (x_n)_{\text{prd}} \right\}^2 \quad (7)$$

in which P is the range of prediction, σ is the standard deviation, $(x_n)_{\text{obs}}$ and $(x_n)_{\text{prd}}$ are the observed and predicted values. Table 3 gives a summary of these statistics.

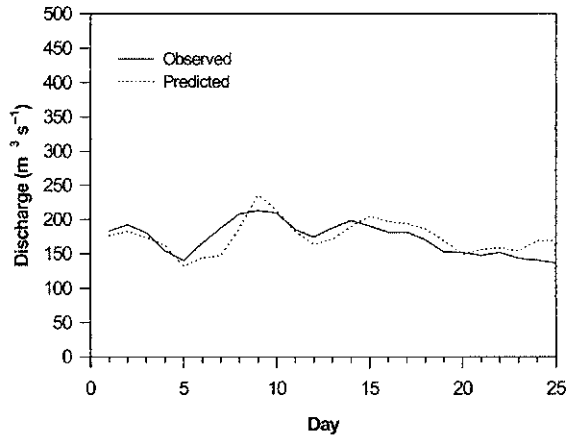


Fig. 1 Prediction of daily discharge in Chao Phraya basin using raw data set ($N = 5844$) (day 1 corresponds to 7 March 1994).

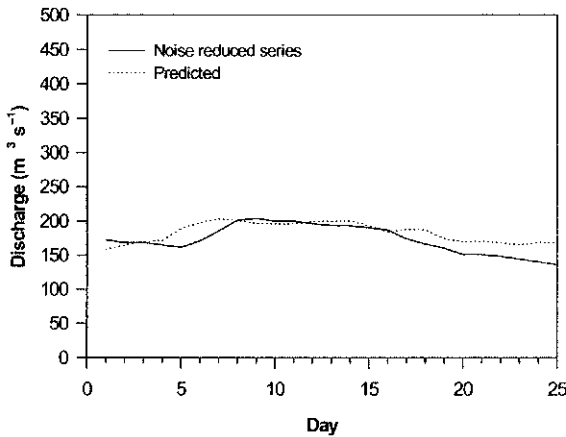


Fig. 2 Prediction of daily discharge in Chao Phraya basin using noise-reduced data set ($N = 5844$) (day 1 corresponds to 7 March 1994; noise reduced by the method of Schreiber & Grassberger, 1991).

Table 3 Statistics of prediction accuracy.

Data series	Raw data		Noise reduced by:					
	Correlation coefficient	NMSE (10^{-3})	Method I		Method II		Method III	
			Correlation coefficient	NMSE (10^{-3})	Correlation coefficient	NMSE (10^{-3})	Correlation coefficient	NMSE (10^{-3})
Chao Phraya	0.735	1.27	0.78	4.63	0.793	1.18	0.715	2.05
S-index	0.881	263.8	0.972	73.0	0.966	250.5	0.988	213.0

CONCLUSION

The dynamical systems approach of modelling and prediction of hydrological time series has been introduced and the importance of dealing with inherent noise associated with all experimental and field observed time series is highlighted. Noise

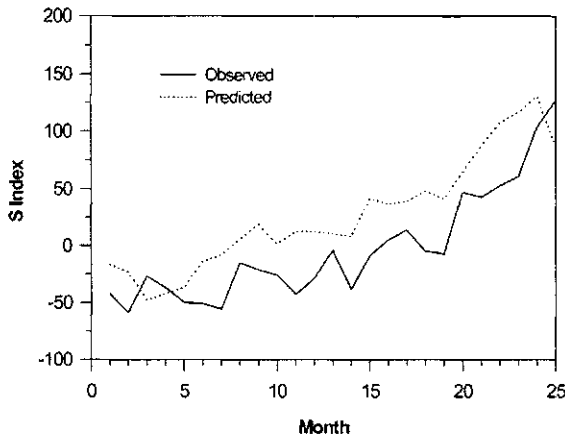


Fig. 3 Prediction of monthly sea surface temperature anomaly index using raw data set ($N = 1380$) (month 1 corresponds to December 1984).

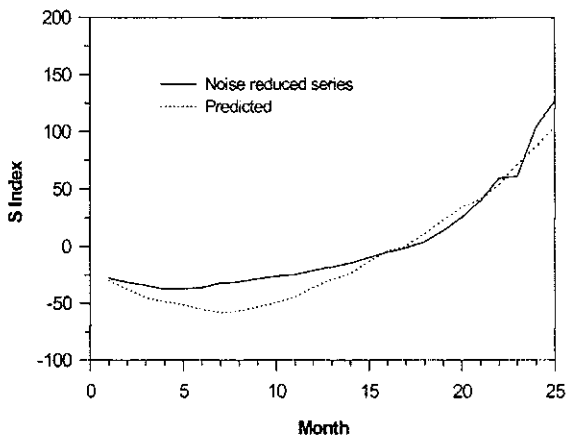


Fig. 4 Prediction of monthly sea surface temperature anomaly index using noise reduced data set ($N = 1380$) (month 1 corresponds to December 1984; noise reduced by the method of Schreiber, 1993b).

removal by different methods and their impact on the determination of invariants such as the correlation dimension are illustrated with reference to two real hydrological data series. Comparisons of predictions made by the “0th order” method before and after noise removal show significant improvement as measured by the correlation coefficient and the normalized mean square error.

Acknowledgements The Fortran program for noise reduction by the “0th order” method was provided by Dr T. Schreiber, University of Wuppertal, Germany.

REFERENCES

- Farmer, J. D. & Sidorowich, J. J. (1987) Predicting chaotic time series. *Phys. Rev. Lett.* **59**, 845-848.
 Grassberger, P. & Procaccia, I. (1983a) Measuring the strangeness of strange attractors. *Physica* **D9**, 189-208.

- Grassberger, P. & Procaccia, I. (1983b) Generalized dimensions of strange attractors. *Phys. Rev. Lett.* **50**, 346–349.
- Grassberger, P., Hegger, R., Kantz, H., Schaffrath, C. & Schreiber, T. (1993) On noise reduction methods for chaotic data. *Chaos* **3**, 127–141.
- Kantz, H. & Schreiber, T. (1997) *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge, UK.
- Mpitsos, G. J., Creech, H. C., Cohan, C. S. & Mendelson, M. (1987) Variability and chaos: neurointegrative principles in self-organization of motor patterns. In: *Directions in Chaos* (ed. by B. Hao), 163–190.
- Schouten, J. C., Takens, F. & van den Bleek, C. M. (1994) Estimation of the dimension of a noisy attractor. *Phys. Rev. E* **50**, 1851–1861.
- Schreiber, T. (1993a) Determination of the noise level of chaotic time series. *Phys. Rev. E* **48**, R13–R16.
- Schreiber, T. (1993b) Extremely simple noise reduction method. *Phys. Rev. E* **47**, 2401–2404.
- Schreiber, T. & Grassberger, P. (1991) A simple noise-reduction method for real data. *Phys. Lett. A* **160**, 411–418.
- Takens, F. (1981) Detecting strange attractors in turbulence. In: *Lecture Notes in Mathematics: Dynamical systems and Turbulence* (ed. by D. A. Rand & L. S. Young) (Proc. Symp. Univ. of Warwick 1979/80) **898**, 366–391.
- Theiler, J. (1987) Efficient algorithm for estimating the correlation dimension from a set of discrete points. *Phys. Rev. A* **36**, 4456–4462.
- Tsonis, A. A. & Elsner, J. B. (1988) The weather attractor over very short time scales. *Nature* **333**, 545–547.
- Wright, P. B. (1989) Homogenized long-period southern oscillation indices. *Int. J. Clim.* **9**, 33–54.