ELSEVIER

Journal of Hydrology 258 (2002) 40–57

Journal of Hydrology

www.elsevier.com/locate/jhydrol

Neighbourhood selection for local modelling and prediction of hydrological time series

A.W. Jayawardena[a,*], W.K. Li[b], P. Xu[c]

[a]*Department of Civil Engineering, The University of Hong Kong, Polfulam Road, Hong Kong, People's Republic of China*
[b]*Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, People's Republic of China*
[c]*Institute of Applied Mathematics, Chinese Academy of Science, Beijing, People's Republic of China*

Received 25 April 2000; revised 2 October 2001; accepted 10 October 2001

Abstract

The prediction of a time series using the dynamical systems approach requires the knowledge of three parameters; the time delay, the embedding dimension and the number of nearest neighbours. In this paper, a new criterion, based on the generalized degrees of freedom, for the selection of the number of nearest neighbours needed for a better local model for time series prediction is presented. The validity of the proposed method is examined using time series, which are known to be chaotic under certain initial conditions (Lorenz map, Henon map and Logistic map), and real hydro meteorological time series (discharge data from Chao Phraya river in Thailand, Mekong river in Thailand and Laos, and sea surface temperature anomaly data). The predicted results are compared with observations, and with similar predictions obtained by using arbitrarily fixed numbers of neighbours. The results indicate superior predictive capability as measured by the mean square errors and coefficients of variation by the proposed approach when compared with the traditional approach of using a fixed number of neighbours.

Keywords: Local models; Chaos; Neighbourhood selection; Generalized degrees of freedom; Hydrological time series

1. Introduction

In the traditional approach of modelling hydrological time series, the series is considered to be originating from a stochastic process, which, at least in theory, has an infinite number of degrees of freedom. In such cases, linear models of the Box–Jenkins type autoregressive moving average (ARMA) have been used for the analysis and prediction by many researchers over the years (Lawrance and Kottegoda, 1977; Box et al., 1994; Jayawardena and Lai, 1989; amongst others). However, it has been realized recently 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, the available limited evidence suggests that it is possible 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 (Farmer and Sidorowich, 1987; Abarbanel et al., 1990; Sugihara and May, 1990; Smith, 1992; Jayawardena and Lai, 1994; Sivakumar et al., 1999; amongst others). The existence or otherwise of chaos

* Corresponding author. Fax: +852-25595337.
E-mail address: hrecjaw@hku.hk (A.W. Jayawardena).

0022-1694/02/$ - see front matter
PII: S0022-1694(01)00557-1

has also been a topic of debate (Ghilardi and Rosso, 1990; Koutsoyiannis and Pachakis, 1996), and subjected to different interpretations (Sivakumar, 2000).

In a deterministic system, predictions can generally be made using an evolutionary equation in which the future value is considered to be dependent upon present and past values. The prediction process therefore involves an accurate estimation of the mapping function, which transforms the present and past values to the future value. In a chaotic system, the predictive power is lost very quickly because of sensitivity to initial conditions.

The mapping function 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 have also been used. The scope in this study is restricted to local models only.

Earlier local models were based on the 'zeroth order' predictor (Farmer and Sidorowich, 1987; Sugihara and May, 1990) 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. It is believed that better predictions can be obtained if 'higher order' predictors are used instead (Xia and Li, 1999). An important question that arises then is how many neighbours would be needed for a better model? In this paper, a method to select the number of neighbours using the generalized degrees of freedom (GDF) (Ye, 1998) is presented and its validity is examined using time series, which are known to be chaotic as well as real hydrological time series. The results are encouraging to the extent that the model selection based on the GDF can be considered to be superior to the model based on an arbitrarily chosen number of neighbours.

2. Embedding dimension

Before a prediction for a given set of chaotic data can be made, its time delay and the embedding dimension should be known. There are several methods of estimating the embedding dimension (Grassberger and Procaccia, 1983; Abarbanel, 1996), but the false nearest neighbour (FNN) method (Abarbanel, 1996) is used in this study. The method works as follows.

For a point $\mathbf{X}(t)$ at time level t, defined as

$$\mathbf{X}(t) = (x(t), x(t-\tau), x(t-2\tau), \ldots, x(t-(d_e-1)\tau)), \tag{1}$$

in the reconstructed phase space of embedding dimension d_e, and time delay τ, there must exist another point $\mathbf{Y}(s)$, defined as

$$\mathbf{Y}(s) = (x(s), x(s-\tau), x(s-2\tau), \ldots, x(s-(d_e-1)\tau))$$

at time level $s \neq t$, such that for every point in the reconstructed phase space $\mathbf{Z}(u)$, defined as

$$\mathbf{Z}(u) = (x(u), x(u-\tau), x(u-2\tau), \ldots, x(u-(d_e-1)\tau))$$

at time level $u \neq t$

$$\|\mathbf{Y}(s) - \mathbf{X}(t)\| \leq \|\mathbf{Z}(u) - \mathbf{X}(t)\|.$$

The nearest neighbour $\mathbf{X}^{NN}(t)$ of $\mathbf{X}(t)$, is then $\mathbf{Y}(s)$, which can be written as

$$\mathbf{X}^{NN}(t) = (x^{NN}(t), x^{NN}(t-\tau), x^{NN}(t-2\tau), \ldots, x^{NN}(t-(d_e-1)\tau)). \tag{2}$$

The time level t of $\mathbf{X}^{NN}(t)$ has very little relation to the time level at which $\mathbf{X}(t)$ appears.

The point $\mathbf{X}^{NN}(t)$ is called an FNN of $\mathbf{X}(t)$ if it arrives in its neighbourhood by projection from a higher dimension. This means that the embedding dimension d_e cannot unfold the attractor. If most points in the phase space have false nearest neighbours, then, the number d_e would not be the embedding dimension for the chaotic data. By comparing the distance between the vectors $\mathbf{X}(t)$ and $\mathbf{X}^{NN}(t)$ in dimension d_e with that in dimension $d_e + 1$, it is possible to establish whether a nearest neighbour is true or false.

This is checked using the approximate condition that if $\mathbf{X}^{NN}(t)$ is the nearest neighbour of $\mathbf{X}(t)$, and if: (according to the definition of Abarbanel, 1996)

$$\frac{|x(t+d_e\tau) - x^{NN}(t+d_e\tau)|}{\|\mathbf{X}(t) - \mathbf{X}^{NN}(t)\|} > 15, \tag{3}$$

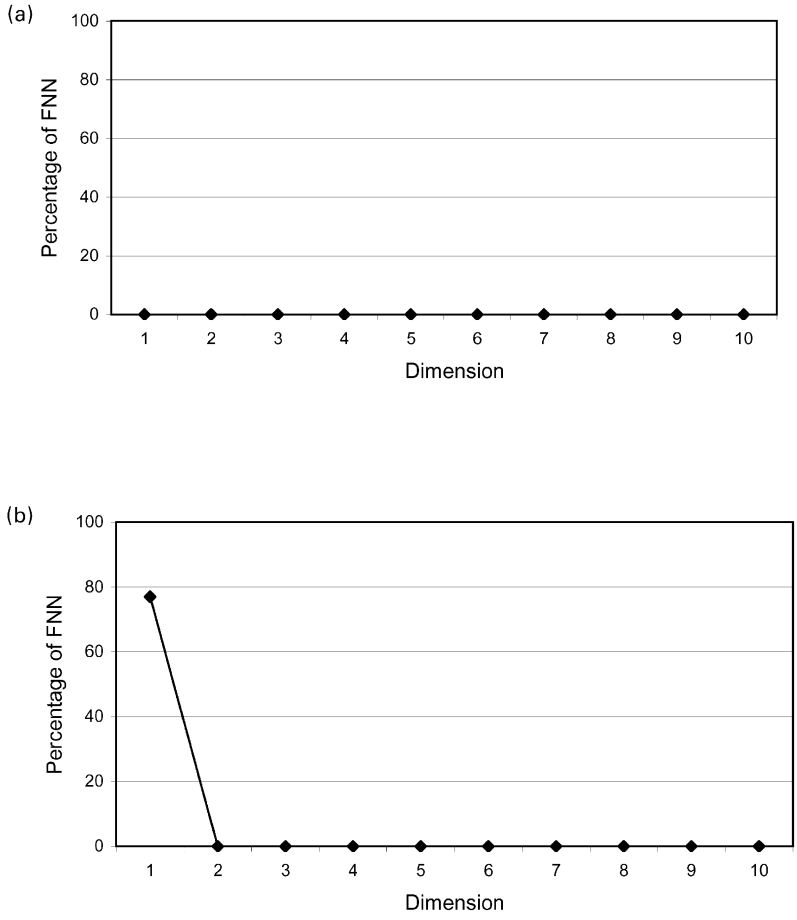


Fig. 1. (a) Percentage of global false nearest neighbours for Logistic data. (b) Percentage of global false nearest neighbours for Henon data. (c) Percentage of global false nearest neighbours for Lorenz data. (d) Percentage of global false nearest neighbours for Mekong data at Nong Khai. (e) Percentage of global false nearest neighbours for Mekong data at Pakse. (f) Percentage of global false nearest neighbours for Chao Phraya data at Nakhon Sawan. (g) Percentage of global false nearest neighbours for SST data.

then $\mathbf{X}^{NN}(t)$ is an FNN of $\mathbf{X}(t)$. By checking for every point in the phase space whether it has an FNN, the percentage of FNN points could be obtained. If for a certain d_e, the percentage of FNN points is less than 5%, it is accepted as the embedding dimension for the data set. For clean data from a chaotic system, it is expected that the percentage of false nearest neighbours drops from nearly 100 in dimension 1 to zero when d_e is reached. Illustrations of this are given in Fig. 1(a)–(c) for the Logistic, Henon and Lorenz data sets (Eqs. (14)–(16)) which are known to be chaotic under certain parameter conditions, and which have embedding dimensions of 1–3, respectively. Fig. 1(d)–(g) illustrates the corresponding behaviour for the real data used in this study.

3. Prediction by local method

To make a prediction in the neighbourhood of the observed point $\mathbf{X}(t)$, a sub-space of dimension d_l within the embedding space of dimension $d_e (d_l \leq d_e)$ is chosen (in this study, the equality condition is assumed). The next signal $x(t + \tau)$ is then obtained from the d_l dimension components of $\mathbf{X}(t)$ via the

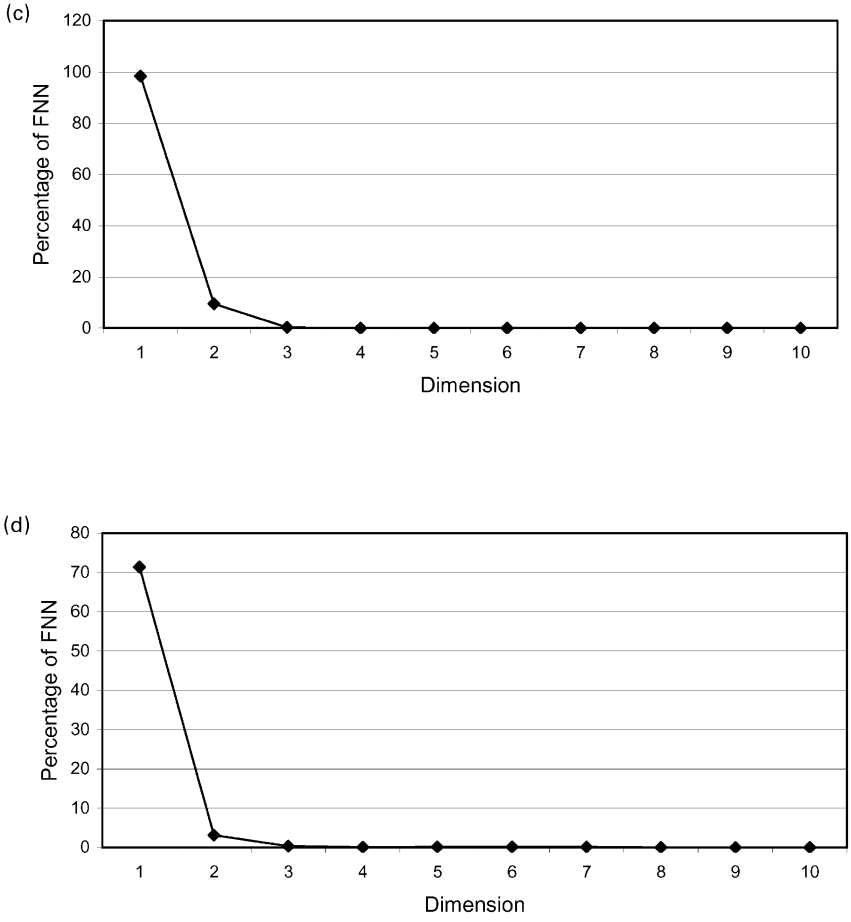


Fig. 1. (*continued*)

evolutionary relationship

$$x(t+\tau) = \mathbf{C}(t)\mathbf{\Phi}(\mathbf{X}(t)) = (\, c_1(t) \quad c_2(t) \quad \cdots \quad c_M(t)\,) \times \begin{pmatrix} \phi_1(\mathbf{X}) \\ \phi_2(\mathbf{X}) \\ \vdots \\ \phi_M(\mathbf{X}) \end{pmatrix}, \tag{4}$$

where $\mathbf{C}(t) = (c_1(t), c_2(t), \ldots, c_M(t))$ is a coefficient vector that needs to be determined.

Here $\mathbf{\Phi}(\mathbf{X}(t))$ is a vector of M local basis functions, which is assumed a priori. It could consist of polynomials, or, in the case of sparse data and high dimensions, radial basis functions. Linear basis functions are special cases of the polynomials, i.e.

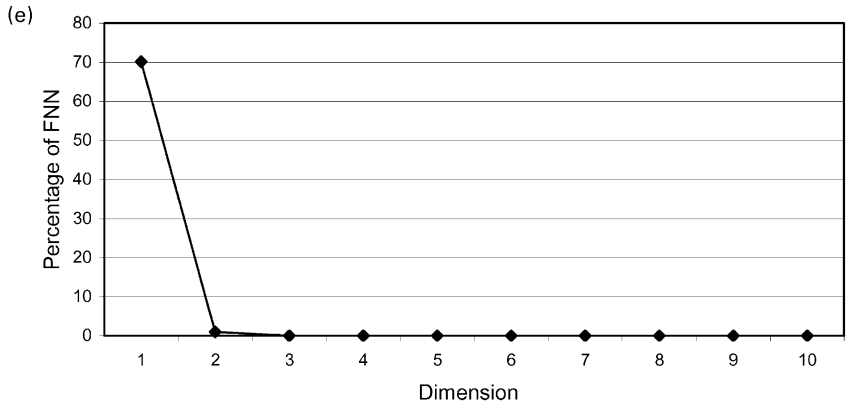
$$\mathbf{\Phi}(\mathbf{X}(t)) = (\phi_1(X), \phi_2(X), \phi_3(X), \ldots, \phi_M(X))^{\mathrm{T}} = (1, x(\tau), x(t-\tau), x(t-2\tau), \ldots, x(t-(d_1-1)\tau))^{\mathrm{T}}.$$

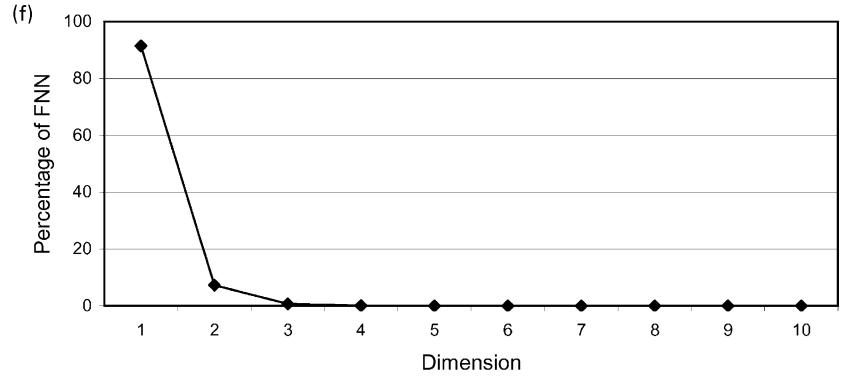
Here $M = d_1 + 1$.

To estimate the coefficient vector $\mathbf{C}(t)$, we employ a set of N nearest neighbours, $\mathbf{X}^r(t)$; $r = 1, 2, 3, \ldots, N$ of $\mathbf{X}(t)$

$$\mathbf{X}^r(t) = (x^r(t), x^r(t-\tau), \ldots, x^r(t-(d_1-1)\tau)). \tag{5}$$

These will, at time level $t+\tau$, evolve to $\mathbf{X}^r(t+\tau)$ which will be in the neighbourhood of $\mathbf{X}(t+\tau)$. The





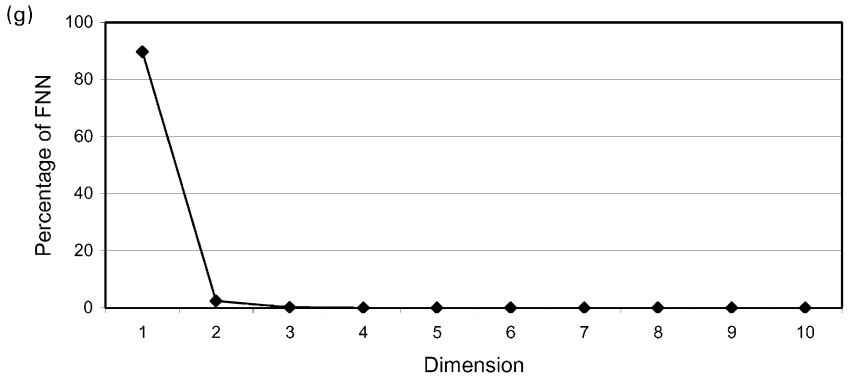


Fig. 1. (*continued*)

coefficient vector $\mathbf{C}(t)$ is then determined by minimizing

$$\sum_{r=1}^{N}\left|x^r(t+\tau)-\sum_{m=1}^{M}c_m(t)\phi_m(\mathbf{X}^r(t))\right|^2, \tag{6}$$

where $x^r(t+\tau)$ is the evolved point at time level $t+\tau$ of $x^r(t)$ at time level t. This is known.

Once the basis functions are known, the above minimization by least squares method is a linear problem. The problem then is to determine the number of neighbours N that will produce a superior prediction.

4. Selection of the nearest neighbours

In regression analysis, the degrees of freedom play a central role in selecting the appropriate statistical model. The degrees of freedom appear in many model selection criteria such as Mallows C_p (Mallows, 1973), Akaike information criterion (AIC) (Akaike, 1974; Wong and Li, 1998), and Bayesian information criterion (BIC) (Schwarz, 1978). Yet these model selection criteria are asymptotic in nature and do not take into account the modelling procedure which can often be very complex. Ye (1998) proposed the GDF, which can handle such complex modelling procedures of which the so-called 'data mining' approach is one example.

The GDF can be defined as the sum of the sensitivities of each fitted value of the model to perturbations in the corresponding observed value. Ye (1998) argued that the GDF is non-asymptotic in nature and is hence free of the sample size consideration. The GDF can be applied to many modelling situations such as artificial neural networks (ANN) and the prediction of chaotic data using the nearest neighbourhood approach. Given $\mathbf{X}(t)$, our objective in the latter application is to choose locally the best set of nearest neighbours in the prediction of $x(t+\tau)$ using the GDF. The assumed local relationship is linear. Contrary to common practice with a fixed number of nearest neighbours, we believe that for each $\mathbf{X}(t)$, a 'best' set of nearest neighbours should be able to provide superior predictions. The reasons are obvious since a fixed number of nearest neighbours either may not be enough or may be too many (in which case over-fitting the noisy data could occur) for a particular local environment.

Using the local prediction method introduced earlier, the future values of the data series can be obtained. Here we show how to choose the best number of neighbours by using the GDF method.

Suppose that we have the following relationships for N neighbours of $\mathbf{X}(t)$

$$\mathbf{Y}=\mathbf{C}\mathbf{X}+\mathbf{V}, \tag{7}$$

where

$$\mathbf{Y}=(x^1(t+\tau),x^2(t+\tau),\ldots,x^N(t+\tau))$$

is the response vector of unknown values of $\mathbf{X}(t+\tau)$ at time $t+\tau$, $\mathbf{V}$ is a row vector of error values

$$\mathbf{C}=(c_1(t),c_2(t),\ldots,c_M(t))$$

and

$$\mathbf{X}=\begin{bmatrix}\phi_1(\mathbf{X}^1) & \phi_1(\mathbf{X}^2) & \cdots & \phi_1(\mathbf{X}^N)\\ \phi_2(\mathbf{X}^1) & \phi_2(\mathbf{X}^2) & \cdots & \phi_2(\mathbf{X}^N)\\ \vdots & \vdots & \vdots & \vdots\\ \phi_M(\mathbf{X}^1) & \phi_M(\mathbf{X}^2) & \cdots & \phi_M(\mathbf{X}^N)\end{bmatrix}.$$

If the basis function is linear, then,

$$\mathbf{X}=\begin{bmatrix}1 & 1 & \cdots & 1\\ x^1(t) & x^2(t) & \cdots & x^N(t)\\ \vdots & \vdots & \vdots & \vdots\\ x^1(t-(d_1-1)\tau) & x^2(t-(d_1-1)\tau) & \cdots & x^N(t-(d_1-1)\tau)\end{bmatrix}.$$

If $\boldsymbol{\mu}$, the mean vector of $\mathbf{Y}$, is estimated by the fitted value $\hat{\boldsymbol{\mu}}$,

$$\hat{\boldsymbol{\mu}}=\mathbf{C}\mathbf{X}, \tag{8}$$

then the estimates of $\mathbf{C}$ and $\hat{\boldsymbol{\mu}}$ are obtained by the least squares method as

$$\mathbf{C}=\mathbf{Y}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1} \tag{9}$$

and

$$\hat{\boldsymbol{\mu}}=\mathbf{Y}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}. \tag{10}$$

For different values of the number of neighbours N, there will be different fitted vector functions $\hat{\boldsymbol{\mu}}$, which in general will be at variance with the observed values $\mathbf{Y}$. A better model is one, which has a smaller variance.

To estimate the value of σ^2 for a given number of neighbours, we need its GDF D. Ye (1998) defined it as

$$D = \sum_{i=1}^{N} h_i(\mu), \tag{11}$$

where

$$h_i(\mu) = \frac{\partial E_\mu[\hat{\mu}_i(Y)]}{\partial y_i}. \tag{12}$$

Here $\hat{\mu}_i(Y)$ and Y_i are the ith components of vectors $\hat{\boldsymbol{\mu}}(\mathbf{Y})$ and $\mathbf{Y}$, respectively. The function $E_\mu[\hat{\mu}_i(Y)]$ can be thought of as a smoothing of the fitted value $\hat{\mu}_i$. The GDF is the sum of the average sensitivities of the fitted value $\hat{\mu}_i(Y)$ to a small perturbation in y_i.

If the assumed local relationship is linear as in this study, then, the GDF simplify to the trace of the smoothing matrix as follows

$$D(\mu) = \mathrm{tr}(\mathbf{H}) = \sum_i h_{ii} = \sum_i \frac{\partial \hat{\mu}_i}{\partial y_i}, \tag{13}$$

where $\mathbf{H} = (h_{ii})_{n\times n} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, and y_i and $\hat{\mu}_i$ are the ith components of $\mathbf{Y}$ and $\hat{\boldsymbol{\mu}}$. Therefore, the GDF in this case are the sum of the sensitivities of the fitted values $\hat{\mu}_i$ with respect to the observed values y_i. Eq. (13) is easier to compute although Eqs. (11) and (12) have more generality. In this study, only Eq. (13) is used. A general algorithm to calculate D by Eqs. (11) and (12) has been given by Ye (1998).

The sum of squared residuals for a chosen number of neighbours, RSS, is then expressed as

$$\mathrm{RSS} = (\mathbf{Y} - \hat{\boldsymbol{\mu}})(\mathbf{Y} - \hat{\boldsymbol{\mu}})^{\mathrm{T}}. \tag{14}$$

An unbiased estimation of the variance σ^2 is given as

$$\sigma^2 = \frac{\mathrm{RSS}}{N - D} = \frac{(\mathbf{Y} - \hat{\boldsymbol{\mu}})(\mathbf{Y} - \hat{\boldsymbol{\mu}})^{\mathrm{T}}}{N - D}. \tag{15}$$

This provides a tool to evaluate the goodness of the model with the chosen number of neighbours.

By comparing the estimations of variances σ^2 for different number of neighbours, the best one can be selected. It is then used for prediction. The following procedure is used in this study:

- for $N = 2d_1 + 1, \ldots, 2d_1 + 10$, obtain 10 local models for every N;
- for each model, obtain D and σ^2, using Eqs. (13) and (15);
- choose N^* from among N which has minimum σ^2;
- use N^* to construct a local model and predict the next signal $x(t + \tau)$;
- obtain lead-time predictions recursively for the desired number of time steps.

The number of neighbours $N = 2d_1 + 1, \ldots, 2d_1 + 10$ used in our procedure is not the only choice. In general, N must be larger than d_1. However, if N is too large, there may be some neighbours that are far away from $\mathbf{X}(t)$ and would not lead to a better model. Our computing results show that the best range of N for a better model is between $2d_1 + 1$ and $2d_1 + 10$.

In this study, $\sigma^2(N)$ is used as the criterion to select the best model for prediction although there are other criteria to evaluate the goodness of a modelling procedure. The performance of this approach is illustrated by the following numerical examples.

5. Application

The proposed method is first used to predict some theoretical functions, which are known to become chaotic under certain parameter conditions. These include the Lorenz map, the Henon map and the Logistic map. They are, respectively, defined by the following equations:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \sigma(y - x), \qquad \frac{\mathrm{d}y}{\mathrm{d}t} = -xz + rx - y, \qquad \frac{\mathrm{d}z}{\mathrm{d}t} = xy - bz; \tag{16}$$

$$x_t = 1 - ax_{t-1}^2 + bx_{t-2}; \tag{17}$$

$$x_t = 3.85x_{t-1}(1 - x_{t-1}). \tag{18}$$

The Lorenz map (Eq. (16)) becomes chaotic for $\sigma = 16$, $r = 45.92$ and $b = 4$ and the Henon map (Eq. (17)) for $a = 1.4$ and $b = 0.3$. By solving the Lorenz equation for the x-component by the Runge–Kutta method with a time step of 0.01 for the assumed initial values of $x_0 = 12.5$, $y_0 = 2.5$ and $z_0 = 1.5$, a series of discrete data $[x(t),\ t = 1, 2, 3, \ldots 100,001]$ is generated. Then, by using a time delay $\tau = 10$, a new data series $[y(s) = x(1 + s\tau),\ s = 1, 2, \ldots, 10\,000]$ is obtained. This is the Lorenz data series used in this study. The remaining data series were generated using

Table 1
Summary of data and prediction error indicators ((Data)-1: predicted by fixed number of neighbours; (Data)-2: predicted by chosen number of neighbours; Mekong1: data at Nong Khai; Mekong2: data at Pakse)

Data set	Length of data, N	Time delay, τ	Embedding dimension	Origin for prediction	Mean square error for 20 steps	Mean square error for 50 steps	Coefficient of variation for 20 steps	Coefficient of variation for 50 steps
Henon-1	20 000	1	2	10 000	705×10^{-8}	584×10^{-3}	104×10^{-7}	961×10^{-3}
Henon-2	20 000	1	2	10 000	676×10^{-8}	533×10^{-3}	100×10^{-7}	872×10^{-3}
Logistic-1	20 000	1	1	10 000	262×10^{-3}	278×10^{-3}	562×10^{-3}	597×10^{-3}
Logistic-2	20 000	1	1	10 000	0	0	0	0
Lorenz-1	10 000	1	3	9000	130×10^{-3}	383	896×10^{-6}	273×10^{-2}
Lorenz-2	10 000	1	3	9000	416×10^{-3}	901×10^{-1}	287×10^{-5}	561×10^{-3}
Mekong1-1	4292	1	3	4000	496×10^{1}	496×10^{1}	454×10^{-5}	361×10^{-5}
Mekong1-2	4292	1	3	4000	420×10^{-1}	253×10^{2}	385×10^{-7}	184×10^{-4}
Mekong2-1	4292	1	2	4232	163×10^{4}	306×10^{6}	177×10^{-4}	586×10^{-2}
Mekong2-2	4292	1	2	4232	265×10^{3}	369×10^{3}	288×10^{-5}	707×10^{-5}
Chao-1	5844	1	3	5700	310×10^{1}	491×10^{1}	272×10^{-4}	547×10^{-4}
Chao-2	5844	1	3	5700	133×10^{1}	152×10^{2}	117×10^{-4}	169×10^{-3}
SST-1	1380	1	3	1100	216	969	629×10^{-3}	200×10^{-2}
SST-2	1380	1	3	1100	300×10^{-1}	301×10^{-1}	874×10^{-4}	621×10^{-4}

initial values $x_0 = 0.3$ and $x_1 = 1.2$ for the Henon map and $x_0 = 0.43$ for the Logistic map.

The hydro meteorological data sets used in this study include the daily discharges of Chao Phraya river at Nakhon Sawan (15.67°N and 100.2°E, basin area, 110 569 km^2, GRDC #2964100) in Thailand for the period April 1978–March 1994, the daily discharges of Mekong river at Nong Khai (17.87°N and 102.72°E, basin area, 302 000 km^2, GRDC #2969090) in Thailand, and of the same river at Pakse (15.12°N and 108.80°E, basin area, 545 000 km^2, GRDC #2469260) in Lao for the period April 1980–December 1991, and the monthly mean sea surface temperature (SST) anomaly over the region bounded approximately by 6°N–6°S and 180°–90°W for the period January 1872–December 1986, which has been defined as S-Index by Wright (1984) and used to identify climatic anomalies attributed to El-nino and southern oscillation. The first three data sets were obtained from the GRDC in Germany and the last one from a table compiled by Wright (1989). A few missing records of the data sets were replaced by the long-term averages. All the real data sets used were noise reduced by methods described in a separate study (Jayawardena and Gurung, 2000) before using them in the modelling and prediction described in this study.

6. Results and discussion

The prediction process of a time series by the dynamical systems approach requires knowledge of three parameters; the time delay τ, the embedding dimension, d_e, and the number of nearest neighbours, N_B. The time delay may be chosen as the lag time at which

Fig. 2. (a) Actual and predicted values of $\mathbf{X}(t)$ for Logistic map (prediction-1: using fixed number of neighbours; prediction-2: using chosen number of neighbours; origin for prediction; $t = 10\,000$). (b) Actual and predicted values of $x(t)$ in Henon map (prediction-1: using fixed number of neighbours; prediction-2: using chosen number of neighbours; origin for prediction: $t = 10\,000$). (c) Actual and predicted values of $x(t)$ in Lorenz equation (prediction-1: using fixed number of neighbours; prediction-2: using chosen number of neighbours; origin for prediction: $t = 9000$). (d) Observed and predicted discharges in Mekong river at Nong Khai (prediction-1: using fixed number of neighbours; prediction-2: using chosen number of neighbours; origin for prediction: $t = 4000$). (e) Observed and predicted discharges in Mekong river at Pakse (prediction-1: using fixed number of neighbours; prediction-2: using chosen number of neighbours: origin for prediction: $t = 4232$). (f) Observed and predicted discharges in Chao Phyraya river at Nakhon Sawan (prediction-1: using fixed number of neighbours; prediction-2: using chosen number of neighbours; origin for prediction: $t = 5700$). (g) Observed and predicted sea surface temperature anomaly (S-Index) (prediction-1: using fixed number of neighbours: prediction-2: using chosen number of neighbours; origin for prediction: $t = 1100$).

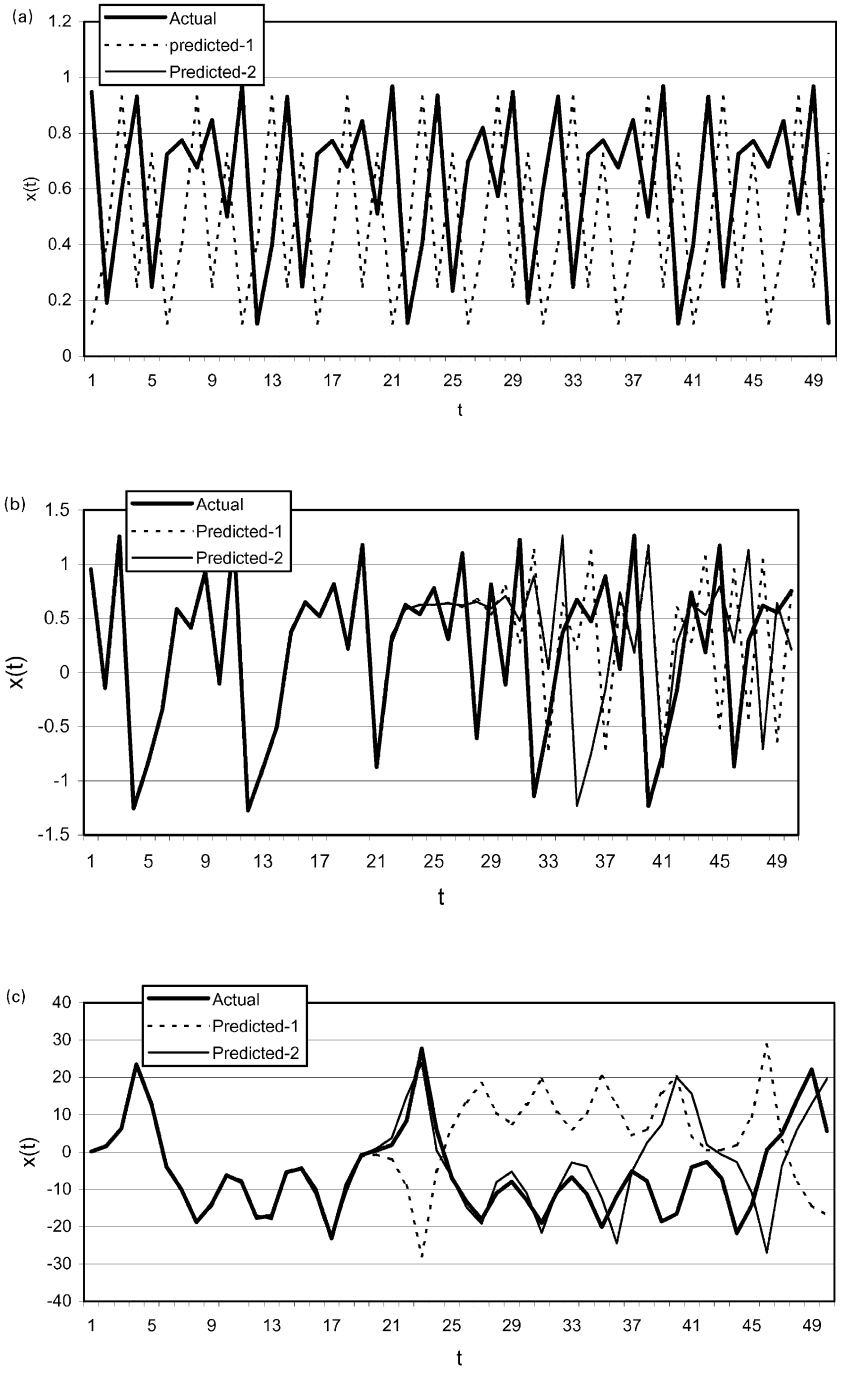

(a)
Actual
predicted-1
Predicted-2
x(t)
t
(b)
Actual
Predicted-1
Predicted-2
x(t)
t
(c)
Actual
Predicted-1
Predicted-2
x(t)
t

Fig. 2.

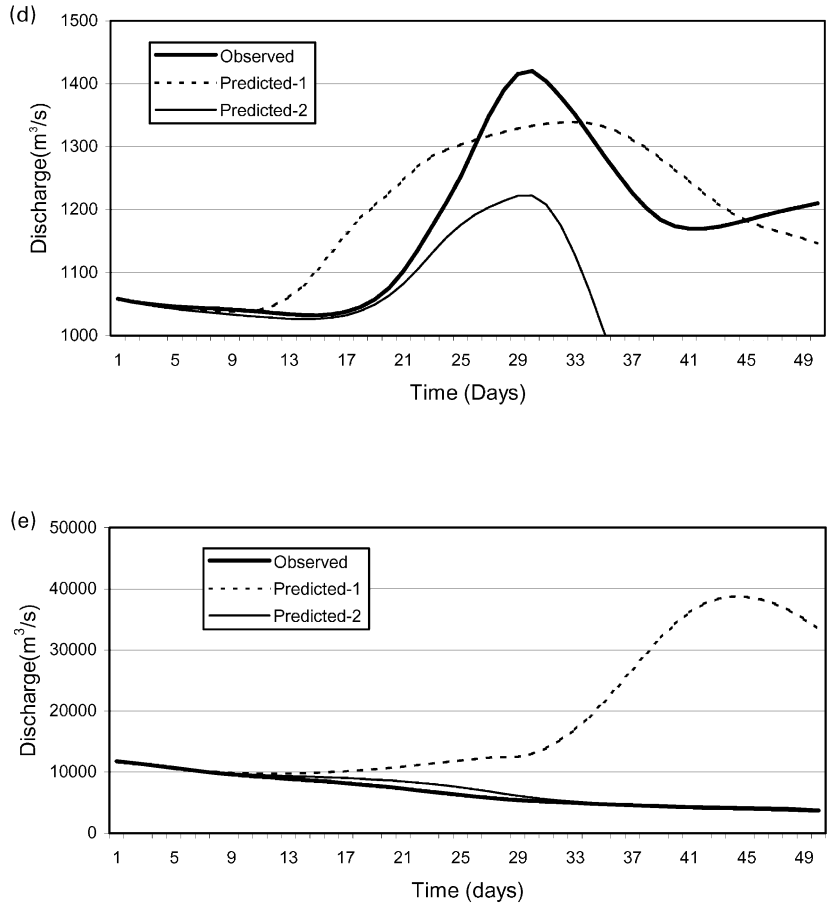


Fig. 2. (*continued*)

the auto-correlation falls below a threshold value, which is commonly defined as $1/e$, specially if the auto-correlation function is approximately exponential (Tsonis and Elsner, 1988). Another suggestion is to take it as the lag time at which the auto-correlation first becomes zero if it crosses the zero line (Mpitsos et al., 1987). From a mathematical point of view, τ, is arbitrary (Kantz and Schreiber, 1997, p. 130). Schouten et al. (1994) used a value of unity for τ, for convenience. A possible problem that may arise in the incorrect choice of the time delay is that if it is too small, the data would not be independent, and if it is too large, the series may become over-smooth thereby losing some information. Despite numerous suggestions, there is no rigorous method of determining an optimal value for τ. In this study, a value of 10 was assumed for the Lorenz data set ($N = 10\,000$), and a value of unity was assumed for all the other data sets.

A precise estimation of the embedding dimension is needed only when determinism has to be explored with minimum computational effort. For practical purposes, the product of the embedding dimension and the time delay is more important than their individual values (Kantz and Schreiber, 1997, p. 34). The embedding dimension gives only the lower limit of the number of dimensions needed for the faithful reconstruction of the phase space. Any dimension value greater than the embedding dimension would also be equally satisfactory. Choosing a large value however would be redundant. In this study, an estimate of the optimal value of the embedding dimension is obtained by the FNN method described earlier. Plots of the percentage of FNN points with the

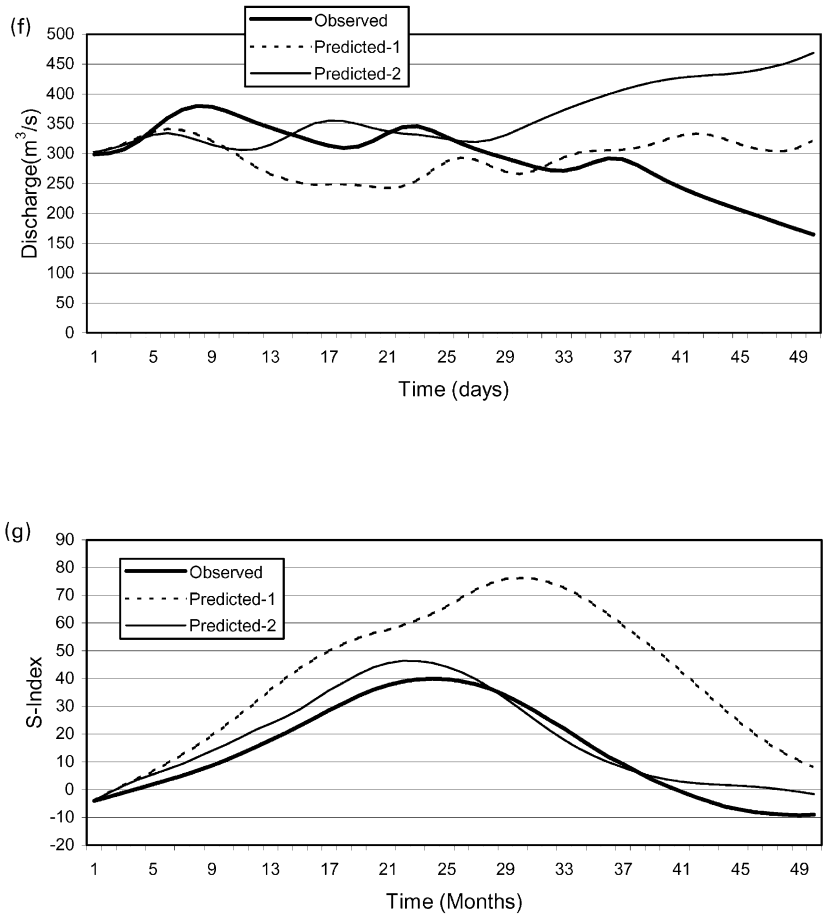


Fig. 2. (*continued*)

dimension for all the data sets used in this study are given in Fig. 1(a)–(g).

The third parameter, N_B, needed for the prediction process is also the focus of this study. It can vary from a single nearest neighbour which gives rise to the zeroth order model (Farmer and Sidorowich, 1987) to more than one nearest neighbour (Smith, 1992) in which case the number has to be determined by some means. In this study, arbitrarily fixed N_B (Fig. 3(a)–(g)) as well as N_B chosen on the basis of the GDF have been used. In the latter case, the number of neighbours for different prediction times varied over a wide range (Lorenz data, 7–14; Henon data, 6–13; Logistic data, 6–12; Chao Phraya data, 7–16; Mekong1 data, 7–16; Mekong2 data, 7–16 and SST data, 7–16); see also Fig. 3(a)–(g).

For all the data sets used, the method in which the number of neighbours is chosen on the basis of the GDF gives better predictions in the short term when compared with the method in which the number of neighbours is fixed. A summary of the results is given in Table 1, which shows the mean square errors and coefficients of variation for 20 steps as well as for 50 steps of lead-time prediction. For the 20-step lead-time, the values of the coefficient of variation using chosen N_B are generally one order less than those using fixed N_B. For the longer lead-time of 50 steps, the error indicators are larger in magnitude and there is no particular pattern of variation relative to the fixed neighbour method. This illustrates the inherent unpredictable characteristic of non-linear deterministic processes for long lead-times.

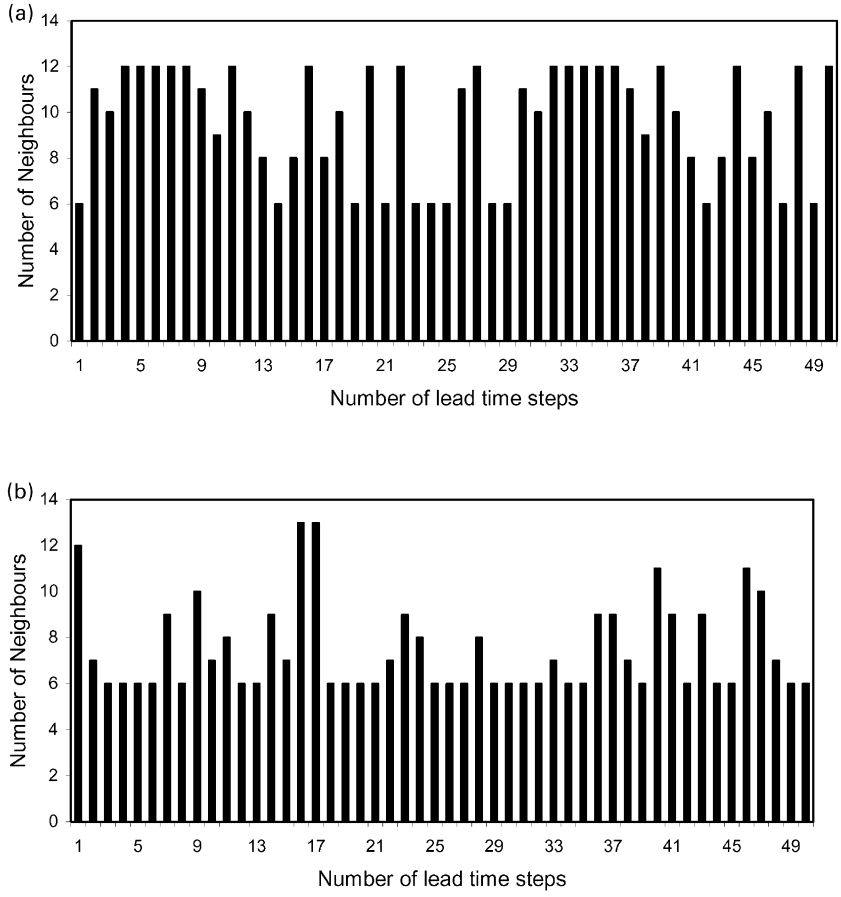


Fig. 3. (a) Number of neighbours used in prediction-2 for Logistic data (prediction-1: using fixed number of neighbours (six); prediction-2: using chosen number of neighbours; origin for prediction: $t = 10\,000$). (b) Number of neighbours used in prediction-2 for Henon data (prediction-1: using fixed number of neighbours (eight); prediction-2: using chosen number of neighbours; origin for prediction: $t = 10\,000$). (c) Number of neighbours used in prediction-2 for Lorenz data (prediction-1: using fixed number of neighbours (eight); prediction-2: using chosen number of neighbours; origin for prediction: $t = 9000$). (d) Number of neighbours used in prediction-2 for Mekong data at Nong Khai (prediction-1: using fixed number of neighbours (eight); prediction-2: using chosen number of neighbours; origin for prediction: $t = 4000$). (e) Number of neighbours used in prediction-2 for Mekong data at Pakse (prediction-1: using fixed number of neighbours (eight); prediction-2: using chosen number of neighbours; origin for prediction: $t = 4323$). (f) Number of neighbours used in prediction-2 for Chao Phraya data at Nakhon Sawan (prediction-1: using fixed number of neighbours (eight); prediction-2: using chosen number of neighbours; origin for prediction: $t = 5700$). (g) Number of neighbours used in prediction-2 for SST data (prediction-1: using fixed number of neighbours (eight); prediction-2: using chosen number of neighbours; origin for prediction: $t = 1100$).

The comparisons of predictions by the two methods are shown in Fig. 2(a)–(c) for the Henon, Logistic and Lorenz data, and Fig. 2(d)–(g) for the Chao Phraya data, Mekong data at Nong Khai, Mekong data at Pakse and SST data, respectively (In some cases, the actual and predicted-2 coincide in the scale of the graph, and therefore do not show two distinct lines. This can also be seen in Fig. 4.).

Fig. 4(a)–(g) shows the variation of the cumulative mean square error with the number of lead-time predictions for the different data sets. These clearly indicate that predictions are possible in the short term, but the predictive power is not long lasting. Another point of importance in this kind of predictions is the effect of the predicting origin. Several origins have been tried for the

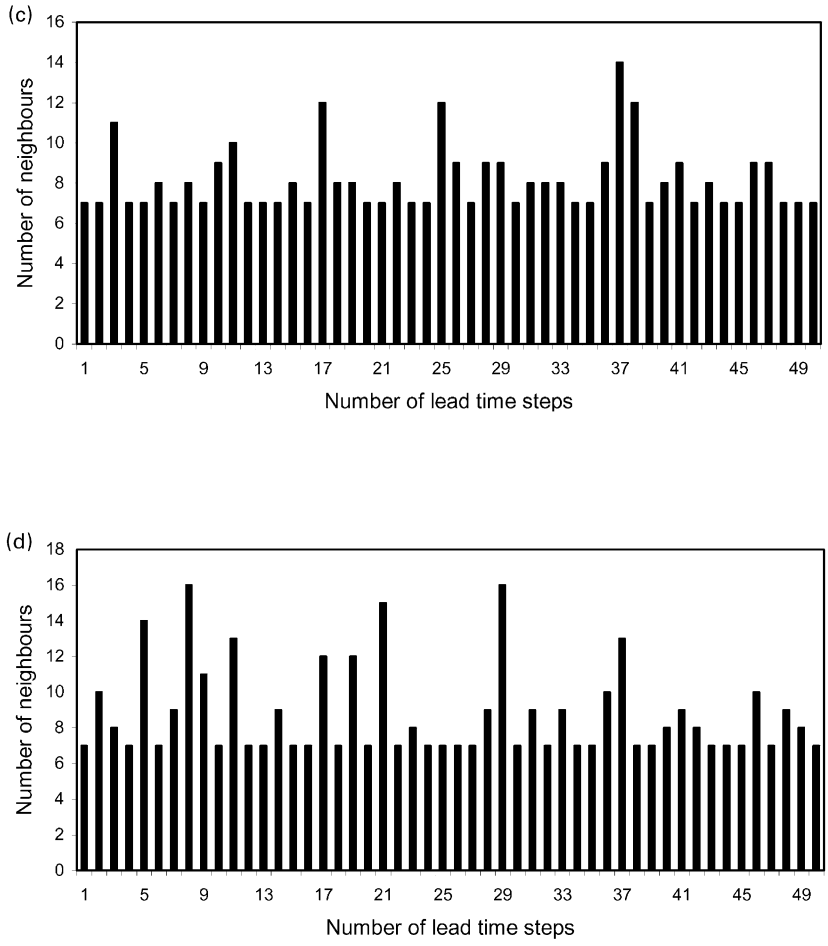


Fig. 3. (*continued*)

data sets, and for most cases, the results are still valid.

7. Conclusion

In this study, a new criterion, based on the GDF method, for the choice of the number of neighbours needed for a 'better' local model of a time series is introduced. With the examples used, it can be concluded that the GDF criterion certainly leads to a better model despite the fact that only linear basis functions have been used in this study.

The first three examples use time series, which are known to be chaotic and therefore deterministic. As shown by the error indicators, the superior predictive capability of the proposed approach is well demonstrated. The method is then applied to practical hydrological time series, which may have evolved from deterministic processes. The results again are quite convincing. The main finding in this study is that the number of neighbours needed for modelling in the phase space is best determined by the GDF method. It is also conjectured that hydrological time series could perhaps be modelled better by the dynamical systems approach. It can only be substantiated after comparison with others methods.

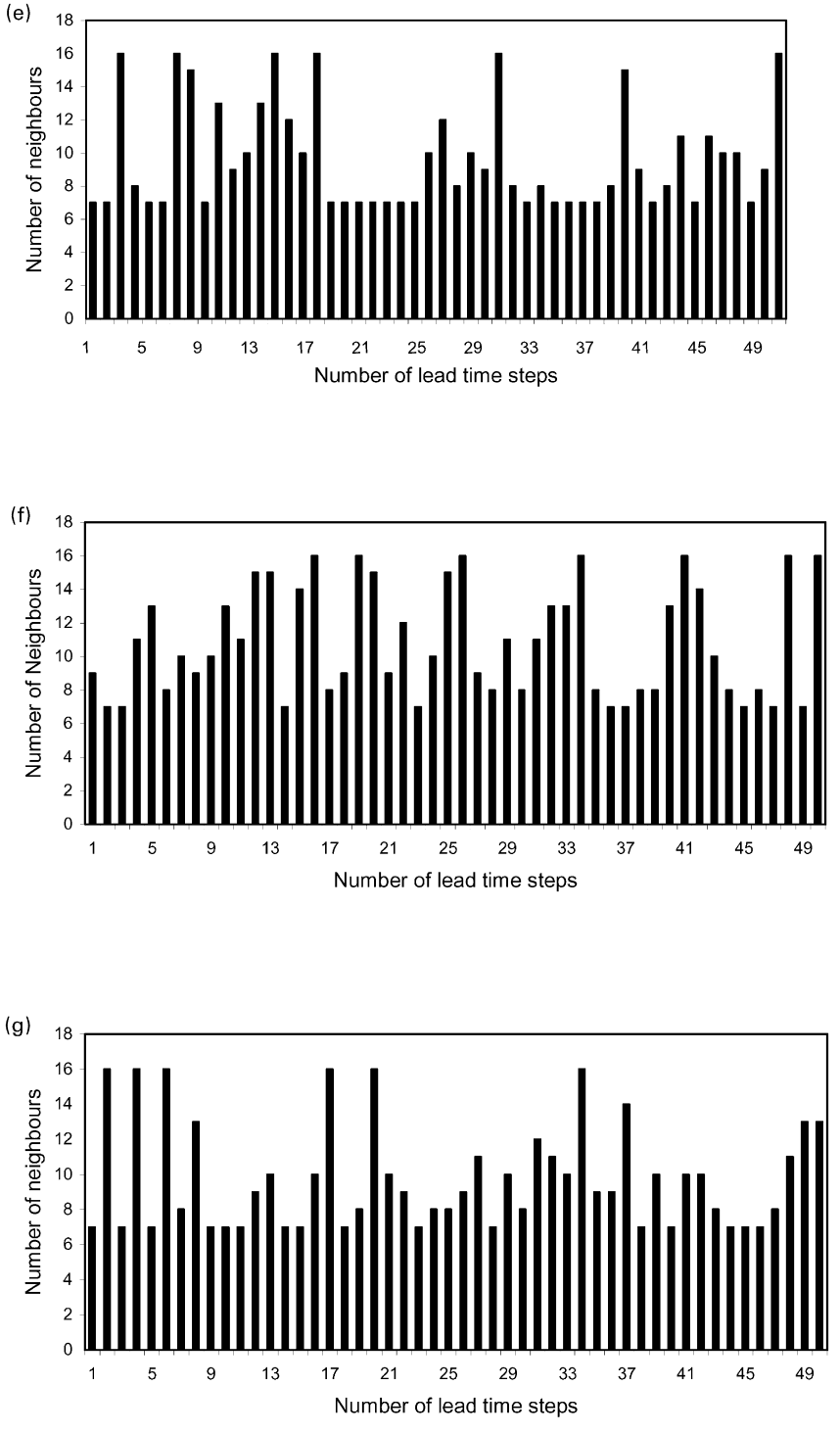
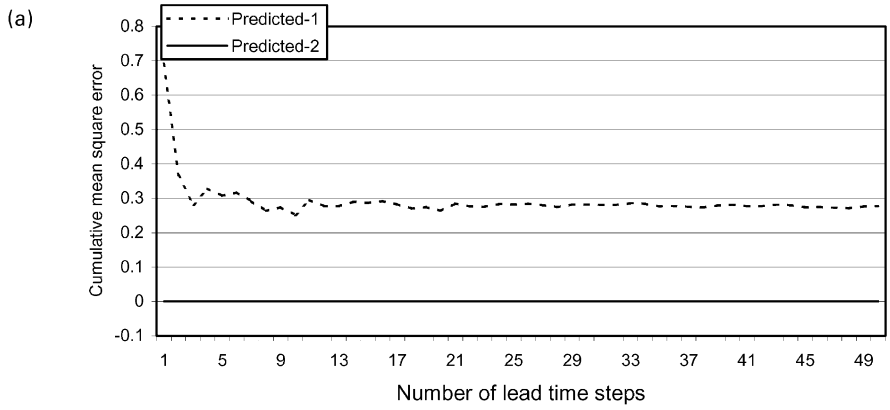
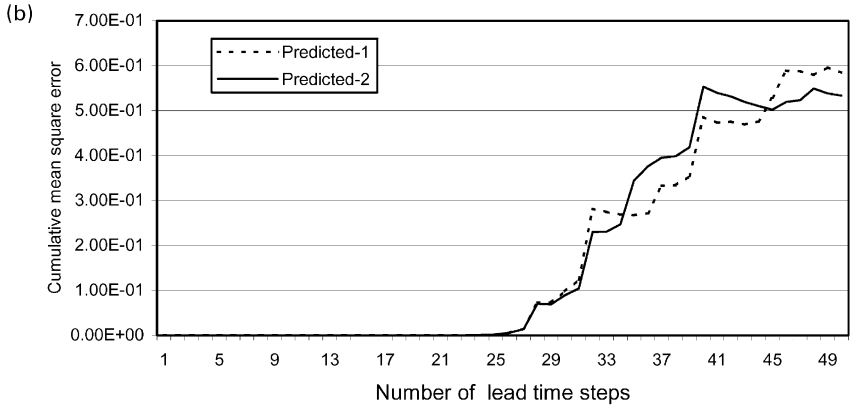


Fig. 3. (*continued*)


(a)
Predicted-1
Predicted-2
Cumulative mean square error
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
1 5 9 13 17 21 25 29 33 37 41 45 49
Number of lead time steps


(b)
Predicted-1
Predicted-2
Cumulative mean square error
7.00E-01
6.00E-01
5.00E-01
4.00E-01
3.00E-01
2.00E-01
1.00E-01
0.00E+00
1 5 9 13 17 21 25 29 33 37 41 45 49
Number of lead time steps

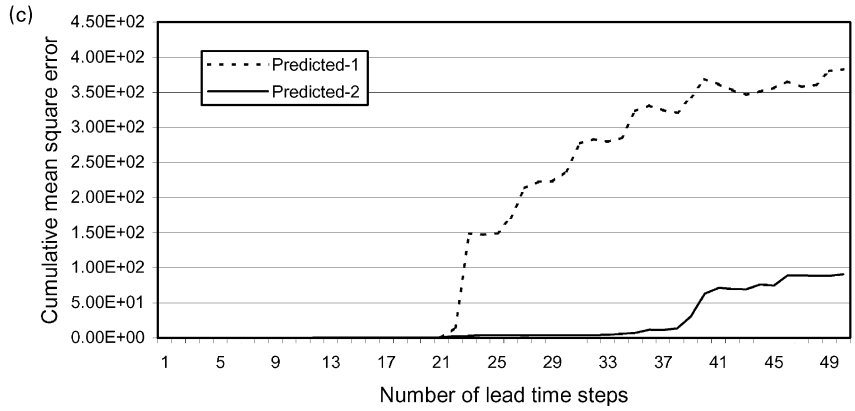
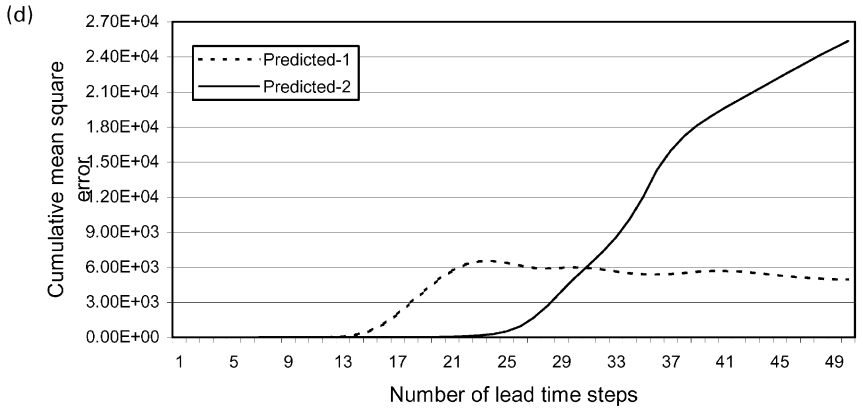

(c)
Predicted-1
Predicted-2
Cumulative mean square error
4.50E+02
4.00E+02
3.50E+02
3.00E+02
2.50E+02
2.00E+02
1.50E+02
1.00E+02
5.00E+01
0.00E+00
1 5 9 13 17 21 25 29 33 37 41 45 49
Number of lead time steps

Fig. 4.



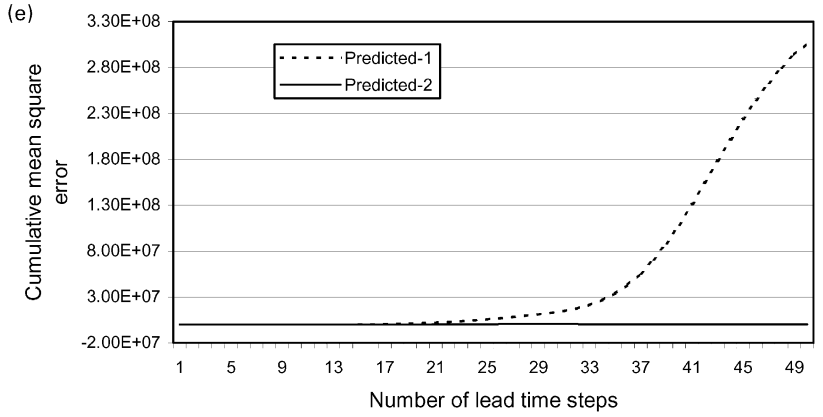


Fig. 4. (*continued*)

Fig. 4. (a) Cumulative mean square errors for Logistic data (prediction-1: using fixed number of neighbours; predicted-2: using chosen number of neighbours; origin for prediction: $t = 10\,000$). (b) Cumulative mean square errors for Henon data (predicted-1: using fixed number of neighbours; predicted-2: using chosen number of neighbours; origin for prediction: $t = 10\,000$). (c) Cumulative mean square error for Lorenz data (prediction-1: using fixed number of neighbours; prediction-2: using chosen number of neighbours; origin for prediction: $t = 9000$). (d) Cumulative mean square errors for Mekong data at Nong Khai (prediction-1: using fixed number of neighbours; prediction-2: using chosen number of neighbours; origin for prediction: $t = 4000$). (e) Cumulative mean square errors for Mekong data at Pakse (prediction-1: using fixed number of neighbours; prediction-2: using chosen number of neighbours; origin for prediction: $t = 4232$). (f) Cumulative mean square errors for Chao Phraya data at Nakhon Sawan (prediction-1: using fixed number of neighbours; prediction-2: using chosen number of neighbours; origin for prediction: $t = 5700$). (g) Cumulative mean square errors for SST data (prediction-1: using fixed number of neighbours; prediction-2: using chosen number of neighbours; origin for prediction: $t = 1100$).

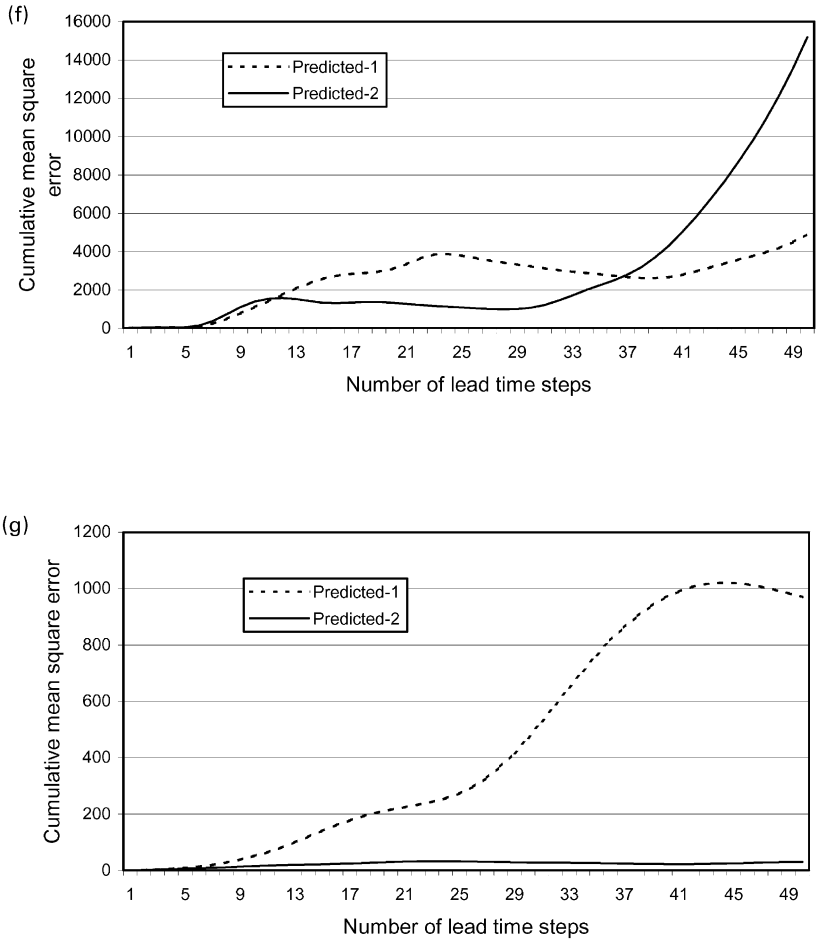


Fig. 4. (*continued*)

Acknowledgements

This work is partially supported by the Hong Kong Research Grants Council Grant No. HKU 7003/97E. Their financial support is gratefully appreciated.

References

Abarbanel, H.D.I., 1996. Analysis of Observed Chaotic Data. 2nd Edition Springer, New York 272 pp.

Abarbanel, H.D.I., Brown, R., Kadtke, J.B., 1990. Prediction in chaotic non-linear systems: methods for time series with broad band Fourier spectra. Phys. Rev. A 41 (4), 1782–1807.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control AC-19, 716–723.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 1994. Time Series Analysis: Forecasting and Control. 3rd Edition Prentice Hall, Englewood Cliffs.

Farmer, J.D., Sidorowich, J.J., 1987. Predicting chaotic time series. Phys. Rev. Lett. 59, 845–848.

Ghilardi, P., Rosso, R., 1990. Comments on chaos in rainfall by I. Rodriguez Iturbe et al. Water Resour. Res. 26 (8), 1837–1839.

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

Jayawardena, A.W., Gurung, A.B., 2000. Noise reduction and prediction of hydro meteorological time series: dynamical systems approach vs. stochastic approach. J. Hydrol. 228 (3–4), 242–264.

Jayawardena, A.W., Lai, F.Z., 1989. Time series analysis of water quality data in Pearl River, China. J. Environ. Engng, ASCE 115 (3), 590–607.

Jayawardena, A.W., Lai, F.Z., 1994. Analysis and prediction of

chaos in rainfall and stream flow time series. J. Hydrol. 153 (1–4), 23–52.

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

Koutsoyiannis, D., Pachakis, D., 1996. Deterministic chaos versus stochasticity in analysis and modelling of point rainfall series. J. Geophys. Res. Atmos. 101 (D21), 26 444–26 451.

Lawrance, A.J., Kottegoda, N.T., 1977. Stochastic modelling of river flow time series. J. R. Stat. Soc. 140, 1–47.

Mallows, C.L., 1973. Some comments on C_p. Technics 15, 661–675.

Mpitsos, G.J., Creech, H.C., Cohan, C.S., Mendelson, M., 1987. Variability and chaos: neurointegrative principles in self-organization of motor patterns. Hao, B. (Ed.). Directions in Chaos, 162–190.

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.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.

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

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

Smith, L.A., 1992. Identification and prediction of low dimensional dynamics. Physica D 58, 50–76.

Sugihara, J., May, R.M., 1990. Non-linear forecasting as a way of distinguishing chaos from measurement error in time series. Nature 344 (19), 734–741 reprinted in Ott et al., 1994.

Tsonis, A.A., Elsner, J.B., 1988. The weather attractor over very short timescales, Nature 333 (6173), 545–547.

Wong, C.S., Li, W.K., 1998. A note on the corrected Akaike information criterion for threshold autoregressive models. J. Time Ser. Anal. 19 (1), 113–124.

Wright, P.B., 1984. Relationship between the indices of the southern oscillation. Monthly Weather Rev. 112, 1913–1919.

Wright, P.B., 1989. Homogenized long-period southern oscillation indices. Int. J. Climatol. 9, 33–54.

Xia, Y., Li, W.K., 1999. On the estimation and testing of functional-coefficient linear models. Stat. Sinica 9 (3), 735–757.

Ye, Jianming, 1998. On measuring and correcting the effects of data mining and model selection. J. Am. Stat. Assoc. 93 (441), 120–131.