

Research Article

A New Adaptive Local Linear Prediction Method and Its Application in Hydrological Time Series

Dunxian She and Xiaohua Yang

State Key Laboratory of Water Environment Simulation, School of Environment, Beijing Normal University, Beijing 100875, China

Correspondence should be addressed to Xiaohua Yang, xiaohuayang@bnu.edu.cn

Received 31 January 2010; Revised 2 April 2010; Accepted 7 May 2010

Academic Editor: Ming Li

Copyright © 2010 D. She and X. Yang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The embedding dimension and the number of nearest neighbors are very important parameters in the prediction of a chaotic time series. In order to reduce the uncertainties in the determination of the forgoing two parameters, a new adaptive local linear prediction method is proposed in this study. In the new method, the embedding dimension and the number of nearest neighbors are combined as a parameter set and change adaptively in the process of prediction. The generalized degree of freedom is used to help select the optimal parameters. Real hydrological time series are taken to examine the performance of the new method. The prediction results indicate that the new method can choose the optimal parameters of embedding dimension and the nearest neighbor number adaptively in the prediction process. And the nonlinear hydrological time series perhaps could be modeled better by the new method.

1. Introduction

The global and regional climates have already begun changing [1], and meteorological-driven processes have been studied by some researchers just as the study in the signal analysis and other fields [2–8]. Many hydrological processes, such as runoff, are usually nonlinear, complex, dynamic processes because of the involved physical process and the considerable spatial and temporal variability [9]. The simulation of the nonlinear hydrological time series was turned out to be very difficult with the traditional deterministic mathematic models. However, the emergence of chaos theory provides a new way to study this kind of highly complex system and makes it possible to extract deterministic regulation from the seemingly disordered hydrological phenomenon. Chaos theory is an important part of nonlinear science and some scientists have done research on its theories and applications [10, 11]. From the last decades, a series of theories and methods identifying chaotic essences

in dynamic systems have gradually been established [12–17]. The first application of chaos theory in hydrological processes could be traced back to the analysis of a series of 1008 monthly rainfall values recorded in Nauru Island by Hense in 1987 [18], and then the chaos theory became a more and more reliable tool for the study of hydrological processes.

With the development of chaos theory and research on its application technique, many methods have been proposed to predict chaotic time series, which can broadly be divided into two main categories: global method [19] and local method [20, 21]. The global method is made an attempt to approximate the whole time series on all attractors and seek a function valid at every point. It is obvious that this kind of method has the disadvantage that if new information is added into the model, the parameters may be changed and a lot of time will be wasted in the parameter estimation. However, the local method overcomes this drawback by building model only on the local attractor and utilizing only part of the past information. Farmer and Sidorowich [20] have already proved that the local prediction method is better than the global prediction method. On the other hand, by using some combining techniques, the forecast accuracy can be improved both in the global method and in the local method [21].

In the process of the local prediction, the phase space reconstruction is the first step and three parameters (the embedding dimension m , the time delay τ , and the number of the nearest neighbors q) should be determined. The studies of the nonlinear predictions indicate that the selection of delay time does not affect the reconstructed attractor reflecting the dynamics of the system unambiguously; so the key problem is the determination of the optimal m and q for prediction. (Usually, the number of the nearest neighbors q is taken greater than m in general condition.)

Although there are many discussions on how to determine the optimal embedding dimension, three basic methods presented below are most usually used. The first kind of methods is based on calculating some geometrical invariants of the attractor, such as Grassberger-Procaccia method (G-P method) [16]. By increasing the embedding dimension, the minimum embedding dimension can be selected when the value of the invariant becomes saturated. The typical problems of this method are time-consuming for computation, unsuited for a short time series, and certainly subjective and sensitive to noise. The second kind of methods, such as the false nearest neighbors method (FNN) [22, 23] and Aleksić method [24], is based on the theory of false neighbors that the far apart points in the original phase space will moving closer together in the reconstruction phase space when a too low embedding dimension is selected. But the criterion is subjective in the judgment of the false neighbors, and Cao has improved the method [25]. The third kind of methods, such as principal component analysis (PCA) [26], singular-spectrum analysis (SSA) [27] and singular-value decomposition (SVD) [28], is based on a singular value decomposition proposed by Broomhead and King [26]. But the singular value decomposition is essentially a linear method based on the covariance matrix which reflects the linear dependence [29]. This approach is also subjective to some extent that the number of large singular values may depend on the details of the embedding process and the accuracy of data. Numerical experience has led several researchers [28, 30] to suspect the authenticity of this method in the analysis of nonlinear time series [31]. All of the methods mentioned above have some shortcomings like inapplicability to a short time series, more or less subjective, sensitivity to noise, and so forth [31].

The uncertainties in the determination of the embedding dimension and the number of the nearest neighbors can affect the forecast precision in the process of the traditional local prediction. So as to improve the prediction accuracy, the problem that arises then is how to

reduce the parameter uncertainties in the prediction method. As we know, the time series data are continually updated as the prediction process proceeded; the parameter estimated by the phase space reconstruction using the original time sequence may not be able to reconstruct the chaotic attractor of the new sequence. In order to reduce the uncertainties caused by the time sequence data updating, the embedding dimension and the number of nearest neighbors should be changed adaptively in the forecast procedure. For the purpose of getting a better prediction results, Jayawardena et al. [32] proposed a method using the generalized degrees of freedom (GDF) to determine the optimal number of the neighbors for the prediction. But in his method, the embedding dimensions take the same values in the whole forecast process like the traditional local prediction; so this method still needs to be improved.

In this study, a parameter set (m, q) is established on the basis of the two uncertain parameters, that is, embedding dimension m and the number of nearest neighbors q . Then a new adaptive local prediction method is proposed, in which m and q are not fixed as certain values but are changed as the prediction steps developed. In the selection of the optimal parameter set (m, q) for each prediction step, the generalized degrees of freedom (GDF) are used and different error variances are calculated under different combinations of (m, q) . The optimal parameter set (m, q) is chosen when the variance obtains the smallest value. In order to examine the validity of the new method, some real hydrological time series are used.

2. The New Local Linear Prediction Method

2.1. The Phase Space Reconstruction

For a scalar time series x_1, x_2, \dots, x_n , the multidimensional phase-space Y_i can be reconstructed by the Takens embedding theory [15], according to

$$Y_i = (x_i, x_{i+\tau}, \dots, x_{i+(m-1)\tau}), \quad (2.1)$$

where $i = 1, 2, \dots, M$, M is the total points number of the phase-space, and $M = n - (m - 1)\tau$, τ is the delay time, m is the dimension of the vector Y_i , called as embedding dimension, and n is the length of the time series.

2.1.1. The Determination of Delay Time

Many methods have been developed in estimating the delay time. And the autocorrelation function [15, 16, 33, 34], the most widely used tool in determining the delay time, is employed in this study. The autocorrelation function can be described as

$$C(t) = \frac{1}{n-t} \frac{\sum_{i=1}^{n-t} (x_i - \mu)(x_{i+t} - \mu)}{\sigma^2(x)}, \quad (2.2)$$

where $C(t)$ is the autocorrelation coefficient, t is the lag time, and μ and σ are the mean and standard variation of the time series, respectively. The delay time τ is always selected when the autocorrelation coefficient has dropped to $1 - 1/e$ of its initial value (e is the base of natural logarithm).

2.1.2. The Determination of Embedding Dimension

Among a number of embedding dimension calculation methods, the false nearest neighbors (FNN) method [22] is most widely used. Cao has proposed a new method on the basis of FNN method. The Cao method [25] can be described as follows. Assume that

$$a(i, m) = \frac{\|Y_i^{(m+1)} - Y_j^{(m+1)}\|}{\|Y_i^{(m)} - Y_j^{(m)}\|} \quad (i = 1, 2, \dots, n - m\tau), \quad (2.3)$$

where $\|\cdot\|$ is some measurements of Euclidian distance, $Y_i^{(m+1)}$ is the i th phase points in the $(m+1)$ -dimension reconstructed phase space, and $Y_j^{(m)}$ ($j = 1, 2, \dots, q$) is the nearest neighbor of $Y_i^{(m)}$. The mean value of all $a(i, m)$ is

$$E(m) = \frac{1}{n - m\tau} \sum_{i=1}^{n-m\tau} a(i, m). \quad (2.4)$$

When the embedding dimension changes from m to $m + 1$, $E_1(m)$ can be defined as

$$E_1(m) = \frac{E(m+1)}{E(m)}. \quad (2.5)$$

If $E_1(m)$ stops changing when m is greater than some value m_0 , then $m_0 + 1$ can be taken as the minimum embedding dimension to reconstruct the phase space.

2.2. The Traditional Local Linear Prediction Method

Many local prediction methods have been developed during the last decades, and the local linear prediction method is considered in this study. The first step of the local linear prediction is to find the nearest neighbor points of the current phase point Y_M in the reconstructed phase space. The Euclidean distances $d(i) = \|Y_i - Y_M\|$ ($i = 1, 2, \dots, M - 1$) between the current vector Y_M and its $M - 1$ preceding delay vectors Y_i ($i = 1, 2, \dots, M - 1$) will be calculated and then q nearest neighbor points Y_{M_k} ($k = 1, 2, \dots, q$, q is taken value great than m in general condition) will be selected.

The local linear prediction model in an m -dimensional reconstructed phase space is an autoregressive model and the prediction value, which is a linear superposition of the m elements in the delay vector Y_M , can be given as follows:

$$\hat{x}(n+1) = AX_1 = a_0 + \sum_{j=1}^m a_j x_{M+(j-1)\tau}, \quad (2.6)$$

where $A = [a_0, a_1, a_2, \dots, a_m]$ is a coefficient vector that needs to be determined, and

$$X_1 = [1, Y_M]^T = [1, x_M, x_{M+\tau}, \dots, x_{M+(m-1)\tau}]^T. \quad (2.7)$$

Deterministic predictions assume that if the phase space point Y_t was similar to the current point Y_M , then the future point Y_{t+1} will also be close to the future point Y_{M+1} . The coefficient vector A can be estimated by the current phase vector Y_M and its q nearest neighbor points Y_{M_k} ($k = 1, 2, \dots, q$) through the following equation:

$$AX = Y, \quad (2.8)$$

where $Y = [x_{M_1+(m-1)\tau+1}, x_{M_2+(m-1)\tau+1}, \dots, x_{M_q+(m-1)\tau+1}]^T$ is the next series value of the points Y_{M_k} ($k = 1, 2, \dots, q$), and

$$X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{M_1} & x_{M_2} & \cdots & x_{M_q} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M_1+(m-1)\tau} & x_{M_2+(m-1)\tau} & \cdots & x_{M_q+(m-1)\tau} \end{bmatrix}. \quad (2.9)$$

Because the value of X and Y is known, so the estimation of coefficient vector A can be obtained by the least squares method as $A = YX^T(XX^T)^{-1}$ from (2.8), and then by using (2.6), the prediction value $\hat{x}(n+1)$ can be calculated. The new prediction value $\hat{x}(n+1)$ is added to the original time series as the prediction steps developed, and the last phase point is Y_{M+1} by the phase space reconstruction now. Following the same scheme in the forecast of $\hat{x}(n+1)$ and reestimating the coefficient vector A , $\hat{x}(n+2)$ can be computed.

2.3. The Determination of the Optimal Parameters (m, q)

In the regression analysis, the degrees of freedom are often used as a model complexity measure in various model selection criteria, such as Mallows C_p , Akaike information criteria (AIC), and Bayesian information criterion (BIC). Yet these model selection criterion are asymptotic in nature and do not take into account the modeling procedure which can often be very complex [32]. So Ye [35] developed a concept of generalized degrees of freedom (GDF) that is applicable for evaluation of the model selection.

The GDF is defined as the sum of the sensitivities of each fitted value of the model to perturbations in the corresponding observed value [32]. It is nonasymptotic in nature and thus is free of the sample-size constraint [35]. In the process of chaotic local linear prediction, the GDF can be viewed as the cost of the modeling process. Considering the uncertainty in the determination of the optimal embedding dimension and the number of nearest neighbors for the prediction, different unbiased estimates of the error variance can be obtained under different parameter sets (m, q). A better (m, q) for the model prediction is the one that has a smaller error variance. So the optimal parameters (m, q) can be selected by comparing the different error variances that are calculated. In every step of the prediction, the GDF can be calculated and the optimal parameters (m, q) can be selected; so the GDF can provide a novel way to guide the (m, q) changing adaptively.

Using the local linear prediction method introduced earlier, the future values of the data series can be obtained. Here we will show how to choose the optimal (m, q) for the prediction by using the GDF method.

From the foregoing description, the estimate of the coefficient vector A in (2.8) can also be given as follows:

$$Y = AX + \varepsilon, \quad (2.10)$$

where Y, A, X are the same in (2.8) and ε is a row vector of error values.

Using the least squares method, A and $\mu_Y = AX$ (μ_Y is the mean vector of Y) can be estimated as

$$A = YX^T (XX^T)^{-1}, \quad (2.11)$$

$$\hat{\mu}_Y = AX = YX^T (XX^T)^{-1} X. \quad (2.12)$$

The sum of squared residuals for a fixed embedding dimension RSS is then expressed as

$$\text{RSS} = (Y - \hat{\mu}_Y)(Y - \hat{\mu}_Y)^T. \quad (2.13)$$

Then an unbiased estimation of the error variance σ_{GDF}^2 is given as

$$\sigma_{\text{GDF}}^2 = \frac{\text{RSS}}{q - D} = \frac{(Y - \hat{\mu}_Y)(Y - \hat{\mu}_Y)^T}{q - D}, \quad (2.14)$$

where q is the number of nearest neighbors. D is the GDF and can be estimated by $D = \text{tr}(H) = \sum_i h_{ii}$, $H = (h_{ii})_{(m+1) \times (m+1)} = X(X^T X)^{-1} X^T$.

This provides a tool to evaluate the goodness of the model with the chosen (m, q) . With different values of the parameter set (m, q) , the matrix X and fitted vector functions $\hat{\mu}_Y$ will be different, then different error variances can be obtained. An optimal (m, q) for the prediction is the one that has the smallest variance.

2.4. The New Local Linear Prediction Method

By comparing the estimations of error variances σ_{GDF}^2 for different (m, q) , the optimal parameter set can be selected. It is then used for prediction. The new local linear method proposed in this study is described as follows.

Step 1. Phase space reconstruction. Determine the embedding dimension m_c and the time delay τ for the original time series.

Step 2. Calculate the error variances. Let the embedding dimension m change from m_{\min} to m_{\max} , and the number of the nearest neighbor change from q_{\min} to q_{\max} (In this study, m_{\min} and m_{\max} are selected as 2 and m_c , resp. This is not the only choice. q_{\min} and q_{\max} are taken as $2m+1$ and $2m+10$, the same as Jayawardena's method [32].) $(m_{\max} - m_{\min} + 1) \times (q_{\max} - q_{\min} + 1)$

Table 1: Statistics indices of daily discharge time series at Peschanaya and Tim.

Statistical indices	Peschanaya	Tim
Total number of data	16437	15340
Mean value (m ³ /s)	32.3357	192.5562
Standard of deviation (m ³ /s)	44.6106	215.4076
Maximum value (m ³ /s)	661.0000	1360.0000
Minimum value (m ³ /s)	1.7400	28.4000
Coefficient of variation	1.3796	1.1187

local linear models can be obtained. For each model, the error variance can be estimated using (2.13) and (2.14).

Step 3. Choose the optimal $(m_{\text{opt}}, q_{\text{opt}})$ which has the minimum variance.

Step 4. Use $(m_{\text{opt}}, q_{\text{opt}})$ to reconstruct the original time series and construct a local model to predict the next value $\hat{x}(n+1)$.

Step 5. Then $\hat{x}(n+2)$ can be computed by the same scheme following the last point; now it is Y_{M+1} .

Some real hydrological time series are chosen to examine the performance of the new local linear method (NLLP-2), and the new method is compared with the traditional local linear prediction method (TLLP) and Jayawardena's method (NLLP-1) [32].

3. Application in Hydrological Time Series

3.1. Study Area and Data Description

The real hydrological time series in this study are chosen as the daily discharges of Peschanaya at Tochil'noye (52.17°N and 85.17°E, basin area, 4720 km²) in Russian for the period January 1936–December 1980, and the daily discharges of Tim at Napas (59.85°N and 81.95°E, basin area, 24500 km²) in Russian for the period January 1953–December 1994. Some important statistical values are shown in Table 1.

For the daily discharge from Peschanaya, the time series is divided into two data sets, the first training data set is the data during the period January 1936–December 1979, and the second prediction data set is the data during the period January 1, 1980–February 9, 1980. In the same way, the first training data set for Tim is the data during the period January 1953–December 1993, the second prediction data set is the data during the period January 1, 1994–February 9, 1994. The prediction lead time for both of the time series is 40 days.

3.2. Phase Space Reconstruction

In this study, the delay time and the embedding dimension of the above two daily discharge time series are determined by the autocorrelation function method and the Cao method, respectively. The changes of the autocorrelation functions for both of the time series are presented in Figure 1 and the values of $E_1(m)$ are shown in Figure 2.

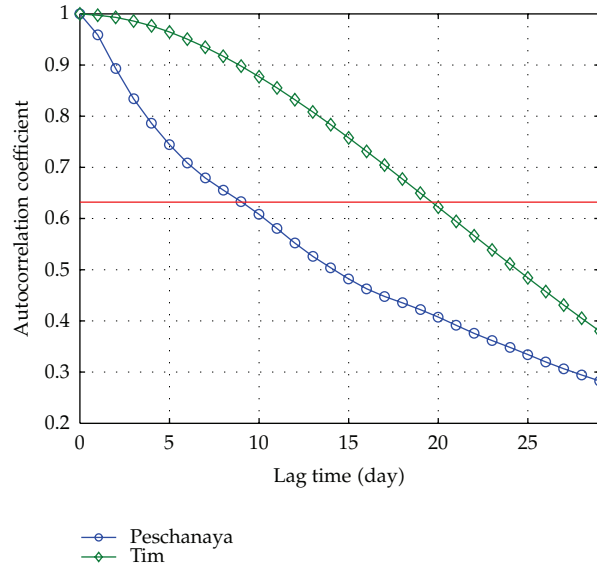


Figure 1: Autocorrelation functions for daily discharge time series observed at Peschanaya and Tim. (The red line presents the situation that the autocorrelation function value is $1 - 1/e$ of its initial value.)

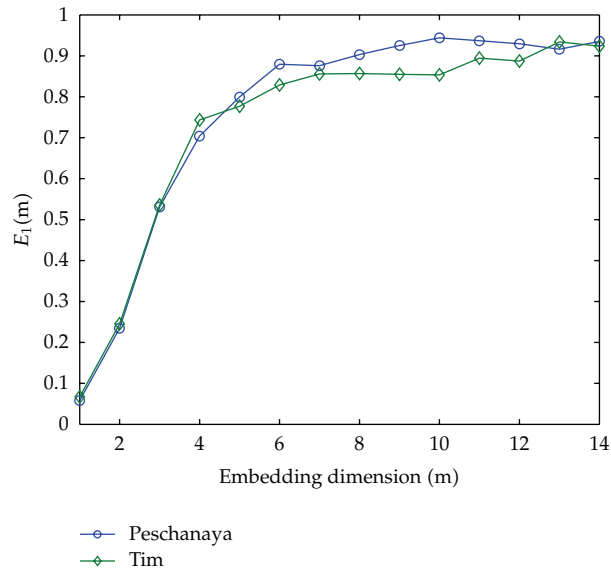


Figure 2: The values of $E_1(m)$ for daily discharge time series observed at Peschanaya and Tim.

The delay time is selected when the autocorrelation coefficient has dropped to $1 - 1/e$ of its initial value. From Figure 1, the delay time for the Peschanaya and Tim can be determined as 9 and 20, respectively. (The red line in Figure 1 presents the situation that the autocorrelation function value is $1 - 1/e$ of its initial value.)

From Figure 2, the value of $E_1(m)$ becomes saturated from $m_0 = 5$; so the embedding dimension can be determined as $m_0 + 1 = 6$ for Peschanaya. In the same way, the embedding dimension for Tim can be determined as 6 too.

The third parameter in the local linear prediction of the daily discharge time series is the number of the nearest neighbors. In this study, only the traditional prediction method (TLLP) needs to give this parameter as a prior; the rest two methods (NLLP-1 and NLLP-2) can obtain the number of nearest neighbors adaptively in the prediction procedure.

3.3. The Prediction Results and Analysis

Both of the time series were predicted by TLLP, NLLP-1, and NLLP-2, respectively. In order to obtain the best forecast results with the TLLP, the number of nearest neighbors q for TLLP is changed from $m + 1$ to $m + 19$. By comparing the prediction results under different q , the optimal forecast results by TLLP in different q can be selected. The prediction results by TLLP were shown in Table 2.

The accuracy of prediction is evaluated by three measurement indices in this study, which are the *Mean absolute error (MAE)*, *Root mean square error (RMSE)*, and *Correlation coefficient (CC)*. The definitions of the three indices can be given as follows:

$$\begin{aligned} \text{MAE} &= \frac{1}{n} \sum_{i=1}^n |y_i - x_i|, \\ \text{RMSE} &= \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2}, \\ \text{CC} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \end{aligned} \quad (3.1)$$

where n is the number of the time series under investigation. The y_i and x_i are the predicted values and observed values in the time series, respectively, and \bar{x} and \bar{y} are the mean of observed values and predicted values, respectively.

Generally, in the above measurement indices, lower values with the MAE and RMSE indicate better agreement, and higher positive values with the CC indicate better agreement between the observed values and predicted values.

From Table 2, it can be seen that the best forecast results for Peschanaya by TLLP method are obtained when the number of nearest neighbors is $m + 7$, that is, 13, (with MAE = 0.4781, RMSE = 0.6878, CC = 0.9092). And the best forecast results for Tim by TLLP method are obtained when the number of nearest neighbors is $m + 16$, that is, 22, (with MAE = 7.6517, RMSE = 9.4725, CC = 0.9089).

The results of NLLP-1, NLLP-2, and the optimal results of the TLLP are showed in Figure 3. The comparisons of the three methods between 20 lead time steps and 40 lead time steps are shown in Tables 3 and 4, respectively.

From Figure 3(a) and Table 3, it can be seen that NLLP-1 gets the best prediction results at the beginning of the prediction process for Peschanaya (with MAE = 0.2367, RMSE = 0.2767 and CC = 0.9636 for NLLP-1 when the lead time step is 20), and under this condition, both of the results by NLLP-1 and NLLP-2 are better than those by TLLP. But as the prediction step

Table 2: The prediction results for the daily discharge time series observed at Peschanaya and Tim by TLLP method.

Number of nearest neighbors	Peschanaya			Tim		
	MAE (m ³ /s)	RMSE (m ³ /s)	CC	MAE (m ³ /s)	RMSE (m ³ /s)	CC
$m + 1$	9.6882×10^2	2.1893×10^3	0.6966	1.6088×10^{10}	4.5795×10^{10}	0.2904
$m + 2$	4.2862×10^5	1.4036×10^6	0.5548	47.9224	78.8664	0.0600
$m + 3$	5.3151	6.2056	0.4293	8.1494×10^4	1.4928×10^5	0.5736
$m + 4$	1.4779	1.8310	0.9503	8.4746×10^{11}	3.0884×10^{12}	0.3439
$m + 5$	2.4415	4.2853	0.7696	2.0832×10^{17}	9.1309×10^{17}	0.2899
$m + 6$	0.8428	1.1907	0.8307	2.1107×10^{11}	7.5627×10^{11}	0.3491
$m + 7$	0.4781	0.6878	0.9092	22.4636	24.4395	-0.3345
$m + 8$	0.5362	0.6917	0.8891	19.9365	20.9531	-0.5286
$m + 9$	0.5599	0.7638	0.8127	19.1111	20.0202	-0.5056
$m + 10$	0.4978	0.6815	0.9236	11.1223	14.2123	0.7822
$m + 11$	0.7445	0.9671	0.8399	18.2638	24.1865	0.8162
$m + 12$	2.6259	5.6787	-0.6568	13.6905	17.0975	0.8609
$m + 13$	2.7742	6.0485	-0.6574	11.0150	13.6199	0.8799
$m + 14$	1.9536	3.7148	-0.6846	11.7535	14.7329	0.8702
$m + 15$	2.6877	5.8700	-0.6505	11.7260	14.4979	0.8819
$m + 16$	1.8358	3.4853	-0.6612	7.6517	9.4725	0.9089
$m + 17$	1.8763	3.3413	-0.7216	11.9486	14.7064	0.8939
$m + 18$	2.5108	5.5012	-0.6394	13.2781	16.1363	0.8916
$m + 19$	2.0266	4.0896	-0.6516	14.8208	17.9076	0.8850

Table 3: The comparison of the prediction results by TLLP ($m = 6, q = 13$), NLLP-1, and NLLP-2 for Peschanaya.

Method	20 steps			40 steps		
	MAE (m ³ /s)	RMSE (m ³ /s)	CC	MAE (m ³ /s)	RMSE (m ³ /s)	CC
TLLP	0.4723	0.6005	0.8747	0.4781	0.6878	0.9092
NLLP-1	0.2367	0.2767	0.9636	0.8613	1.1324	0.7158
NLLP-2	0.3511	0.4455	0.6942	0.9658	1.1894	0.9516

Table 4: The comparison of the prediction results by TLLP ($m = 6, q = 22$), NLLP-1, and NLLP-2 for Tim.

Method	20 steps			40 steps		
	MAE (m ³ /s)	RMSE (m ³ /s)	CC	MAE (m ³ /s)	RMSE (m ³ /s)	CC
TLLP	2.5799	3.5369	0.9539	7.6517	9.4725	0.9089
NLLP-1	7.4430	9.2437	-0.5338	8.8872	9.8291	-0.1647
NLLP-2	2.6050	2.9928	0.9113	1.7600	2.2413	0.9054

increased, the forecast precision of NLLP-1 and NLLP-2 becomes worse than that of TLLP. When the lead time step becomes 40, the TLLP (with MAE = 0.4781, RMSE = 0.6878 and CC = 0.9092) obtains the best results.

From Figure 3(b) and Table 4, it can be seen that, not only at the beginning of the prediction procedure but also along the whole procedure when the lead time step is 40, the NLLP-2 gets the best prediction results (with MAE = 2.6050, RMSE = 2.9928, and CC = 0.9113

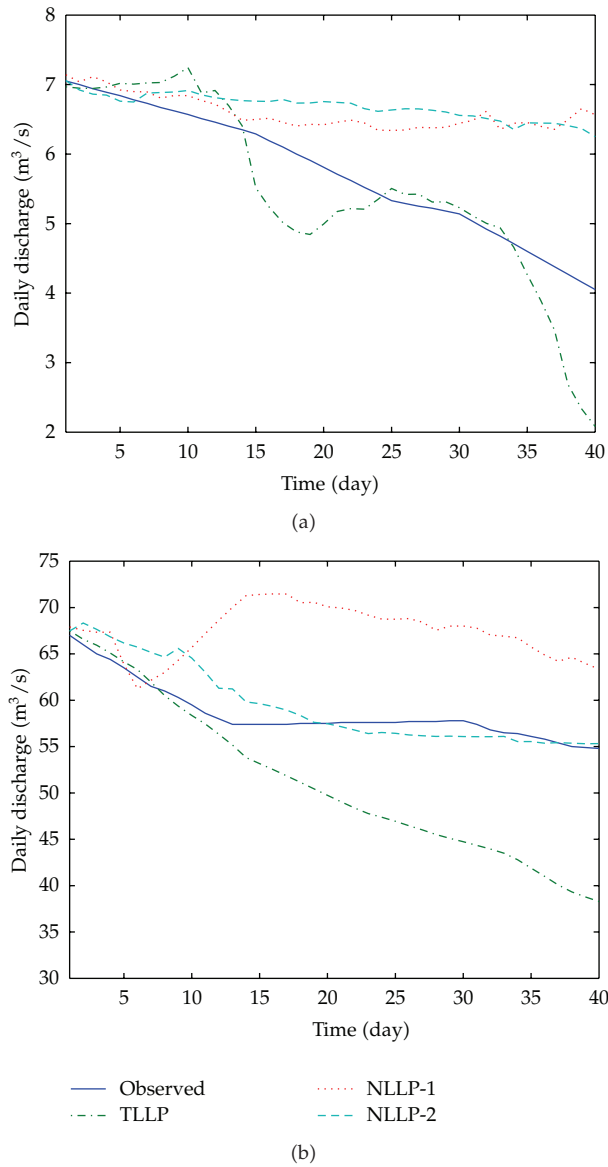
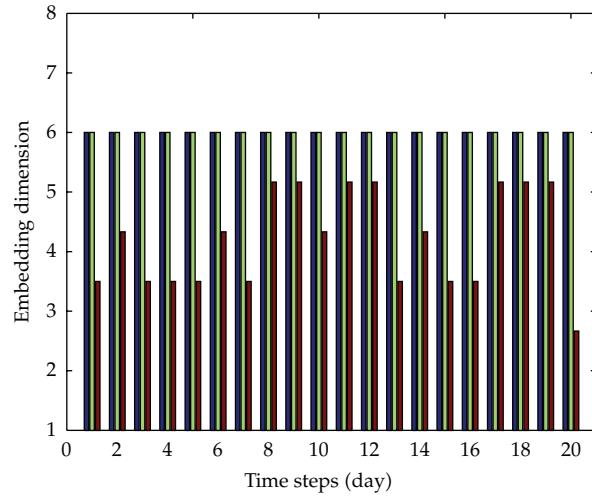


Figure 3: The prediction results of the hydrological time series. (a) The prediction results by TLLP ($m = 6$, $q = 13$), NLLP-1 and NLLP-2 for Peschanaya. (b) The prediction results by TLLP ($m = 6$, $q = 22$), NLLP-1 and NLLP-2 for Tim.

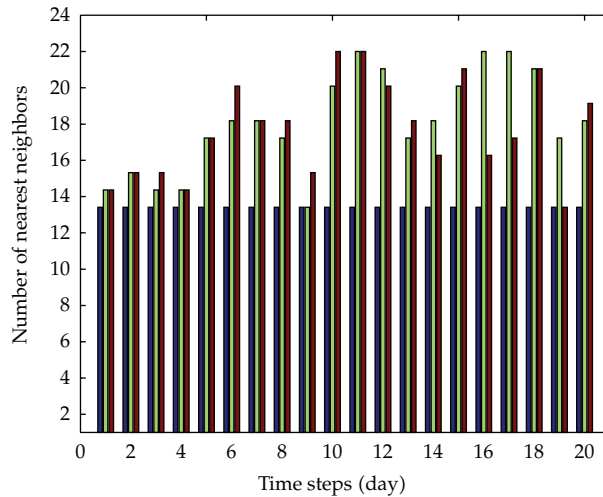
when the lead time step is 20 and MAE = 1.7600, RMSE = 2.2413, and CC = 0.9054 for TLLP when the lead time step is 40).

From the above analysis, the prediction performance of the new method proposed in this study (NLLP-2) is better than that of TLLP at the beginning of prediction process for Peschanaya. For the daily discharge from Tim, the NLLP-2 is working superior than both TLLP and NLLP-1.

The changes of the embedding dimension and the number of nearest neighbors in the foregoing three methods are shown in Figures 4 and 5 (the lead time step is 20 days). It can



(a)



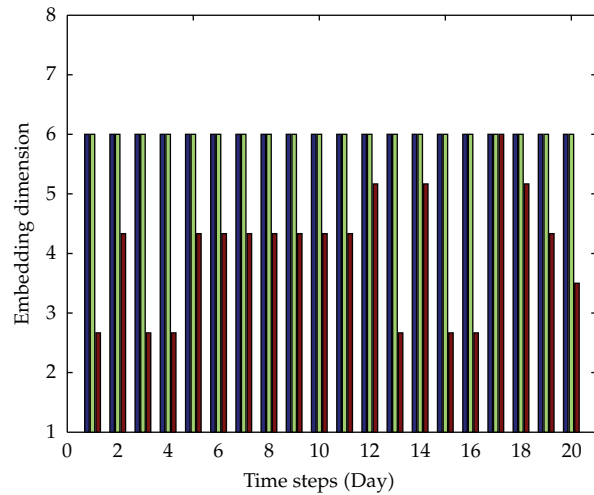
(b)

Figure 4: The changes of embedding dimension (a) and number of nearest neighbors (b) in the prediction process for Peschanaya (the lead time step is 20 days).

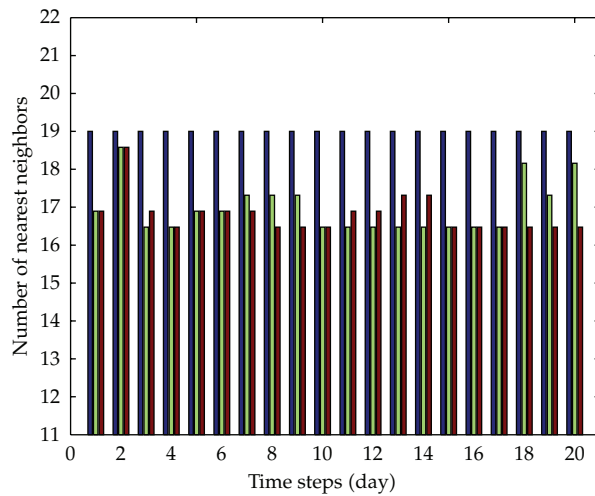
be found that the embedding dimension and the number of nearest neighbors do change adaptively in the prediction processes.

4. Conclusion and Prospect

To obtain the optimum prediction results, a new adaptive local linear prediction method is proposed in which the embedding dimension and the number of nearest neighbors are



(a)



(b)

Figure 5: The changes of embedding dimension (a) and number of nearest neighbors (b) in the prediction process for Tim (the lead time step is 20 days).

combined as a parameter set (m, q) and change adaptively as the forecast steps increased in this study. The main results are given as follows.

- (i) The optimal parameters (m, q) can be gotten by using the GDF method. In the process of the selection of the optimal parameters (m, q) for each prediction step, the generalized degree of freedom (GDF) is used and different error variances are calculated under different combinations of (m, q) . The optimal parameters (m, q)

are chosen when the variance is the smallest. Then the optimal parameters are used to reconstruct the phase space and predict the next value of the time series.

- (ii) For the sake of comparing the performance of the different prediction method, real nonlinear hydrological time series are chosen to be examined. The results from the used example indicate that the new adaptive local linear prediction method proposed in this study can choose the embedding dimension and the number of the nearest neighbors adaptively in the prediction process. Compared with the methods of TLLP and NLLP-1, the new method is better than TLLP at the beginning of the prediction steps for both of the time series, and NLLP-2 is better than TLLP and NLLP-1 for the time series from Tim during the whole forecast period.
- (iii) The new adaptive local linear prediction method can be used in predicting other nonlinear time series in the future and its theory will be further studied.

Acknowledgments

This work was supported by the Key Program of the National Science Foundation of China (Grant no. 50939001), the National Basic Research Program of China (Grant no. 2010CB429003), and the National Key Technology R&D Program of China (Grant no. 2006BAB04A09)

References

- [1] G. B. Sahoo and S. G. Schladow, "Impacts of climate change on lakes and reservoirs dynamics and restoration policies," *Sustainability Science*, vol. 3, no. 2, pp. 189–199, 2008.
- [2] C. A. Lin, L. Wen, G. Lu et al., "Real-time forecast of the 2005 and 2007 summer severe floods in the Huaihe River Basin of China," *Journal of Hydrology*, vol. 381, no. 1-2, pp. 33–41, 2010.
- [3] C. Cattani and A. Kudreyko, "On the discrete harmonic wavelet transform," *Mathematical Problems in Engineering*, vol. 2008, Article ID 687318, 7 pages, 2008.
- [4] C. Cattani, "Harmonic wavelet approximation of random, fractal and high frequency signals," *Telecommunication Systems*, vol. 43, no. 3-4, pp. 207–217, 2010.
- [5] M. Li and W. Zhao, "Representation of a stochastic traffic bound," *IEEE Transactions on Parallel and Distributed Systems*. In press.
- [6] M. Li, "Modeling autocorrelation functions of long-range dependent teletraffic series based on optimal approximation in Hilbert space—a further study," *Applied Mathematical Modelling*, vol. 31, no. 3, pp. 625–631, 2007.
- [7] M. Li and S. C. Lim, "Modeling network traffic using generalized Cauchy process," *Physica A*, vol. 387, no. 11, pp. 2584–2594, 2008.
- [8] M. Li and W. Zhao, "Variance bound of ACF estimation of one block of fGn with LRD," *Mathematical Problems in Engineering*, vol. 2010, Article ID 560429, 14 pages, 2010.
- [9] P. Shang, X. Na, and S. Kamae, "Chaotic analysis of time series in the sediment transport phenomenon," *Chaos, Solitons and Fractals*, vol. 41, no. 1, pp. 368–379, 2009.
- [10] E. N. Lorenz, "Deterministic nonperiodic flow," *Journal of the Atmospheric Sciences*, vol. 20, pp. 130–141, 1963.
- [11] R. M. May, "Simple mathematical models with very complicated dynamics," *Nature*, vol. 261, no. 5560, pp. 459–467, 1976.
- [12] M. Hénon, "A two-dimensional mapping with a strange attractor," *Communications in Mathematical Physics*, vol. 50, no. 1, pp. 69–77, 1976.
- [13] M. C. Mackey and L. Glass, "Oscillation and chaos in physiological control systems," *Science*, vol. 197, no. 4300, pp. 287–289, 1977.
- [14] E. N. Lorenz, "Atmospheric predictability as revealed by naturally occurring analogues," *Journal of the Atmospheric Sciences*, vol. 26, no. 4, pp. 636–646, 1969.

- [15] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence*, D. A. Rand and L. S. Young, Eds., vol. 898 of *Lecture Notes in Mathematics*, pp. 366–381, Springer, Berlin, Germany, 1981.
- [16] P. Grassberger and I. Procaccia, "Characterization of strange attractors," *Physical Review Letters*, vol. 50, no. 5, pp. 346–349, 1983.
- [17] G. B. Sahoo, S. G. Schladow, and J. E. Reuter, "Forecasting stream water temperature using regression analysis, artificial neural network, and chaotic non-linear dynamic models," *Journal of Hydrology*, vol. 378, no. 3-4, pp. 325–342, 2009.
- [18] A. Hense, "On the possible existence of a strange attractor for the southern oscillation," *Beitraege zur Physik der Atmosphaere*, vol. 60, no. 1, pp. 34–47, 1987.
- [19] L. Cao, Y. Hong, H. Fang, and G. He, "Predicting chaotic time series with wavelet networks," *Physica D*, vol. 85, no. 1-2, pp. 225–238, 1995.
- [20] J. D. Farmer and J. J. Sidorowich, "Predicting chaotic time series," *Physical Review Letters*, vol. 59, no. 8, pp. 845–848, 1987.
- [21] D. I. Jeong and Y.-O. Kim, "Combining single-value streamflow forecasts—a review and guidelines for selecting techniques," *Journal of Hydrology*, vol. 377, no. 3-4, pp. 284–299, 2009.
- [22] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data*, Institute for Nonlinear Science, Springer, New York, NY, USA, 2nd edition, 1996.
- [23] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Physical Review A*, vol. 45, no. 6, pp. 3403–3411, 1992.
- [24] Z. Aleksić, "Estimating the embedding dimension," *Physica D*, vol. 52, no. 2-3, pp. 362–368, 1991.
- [25] L. Cao, "Practical method for determining the minimum embedding dimension of a scalar time series," *Physica D*, vol. 110, no. 1-2, pp. 43–50, 1997.
- [26] D. S. Broomhead and G. P. King, "Extracting qualitative dynamics from experimental data," *Physica D*, vol. 20, no. 2-3, pp. 217–236, 1986.
- [27] R. Vautard, P. Yiou, and M. Ghil, "Singular-spectrum analysis: a toolkit for short, noisy chaotic signals," *Physica D*, vol. 58, no. 1–4, pp. 95–126, 1992.
- [28] A. I. Mees, P. E. Rapp, and L. S. Jennings, "Singular-value decomposition and embedding dimension," *Physical Review A*, vol. 36, no. 1, pp. 340–346, 1987.
- [29] M. Paluš and I. Dvořák, "Singular-value decomposition in attractor reconstruction: pitfalls and precautions," *Physica D*, vol. 55, no. 1-2, pp. 221–234, 1992.
- [30] A. M. Fraser, "Reconstructing attractors from scalar time series: a comparison of singular system and redundancy criteria," *Physica D*, vol. 34, no. 3, pp. 391–404, 1989.
- [31] Q.-F. Meng, Y.-H. Peng, and P.-J. Xue, "A new method of determining the optimal embedding dimension based on nonlinear prediction," *Chinese Physics*, vol. 16, no. 5, pp. 1252–1257, 2007.
- [32] A. W. Jayawardena, W. K. Li, and P. Xu, "Neighbourhood selection for local modelling and prediction of hydrological time series," *Journal of Hydrology*, vol. 258, no. 1–4, pp. 40–57, 2002.
- [33] X. H. Yang, D. X. She, Z. F. Yang, Q. H. Tang, and J. Q. Li, "Chaotic bayesian method based on multiple criteria decision making (MCDM) for forecasting nonlinear hydrological time series," *International Journal of Nonlinear Sciences and Numerical Simulation*, vol. 10, no. 11-12, pp. 1595–1610, 2009.
- [34] B. Sivakumar, A. W. Jayawardena, and T. M. Fernando, "River flow forecasting: use of phase-space reconstruction and artificial neural networks approaches," *Journal of Hydrology*, vol. 265, no. 1–4, pp. 225–245, 2002.
- [35] J. Ye, "On measuring and correcting the effects of data mining and model selection," *Journal of the American Statistical Association*, vol. 93, no. 441, pp. 120–131, 1998.