Stochastic Hydrol. Hydraul. 7 (1993) 205–212

Stochastic Hydrology
and Hydraulics

Forecasting river flow using nonlinear dynamics

G. Kember and A. C. Flower

Mathematical Institute, Oxford University, 24-29 St. Giles, Oxford OX1 3LB, England

J. Holubeshen

Beak Consultants Ltd., 14 Abacus Rd., Brampton, Ontario, Canada L6T 5B7

Abstract: A Nearest Neighbor Method (NNM) is used to forecast daily river flows that were measured at a single location over a time period spanning about seventy years. A parsimonious three parameter NNM is developed in the context of Nonlinear Dynamics and the dependence between forecast error and length of history used to construct forecasts is investigated. Comparison is made to Auto-Regressive Integrated Moving Average (ARIMA) models. The NNM is found to provide improved forecasts.

Key words:

1 Introduction

Rivers are used by man for the generation of hydroelectric power, the discharge of effluent from municipalities and industries and as a source of food and water. Management of this resource requires the simulation of the effects of continued use and/or proposed changes to that use. The forecasting of river flow is an important component in the simulation of the effects of our interaction with this resource.

Flow forecasting methods include watershed rainfall-runoff prediction models, parametric statistical models (Box and Jenkins, 1976; Olason and Watt, 1986) and nearest neighbor methods of non-parametric models. Mack and Rosenblatt (1979) use nearest neighbor methods to forecast daily flows generated by rainfall-runoff processes. Comparison to forecasts formed using Auto Regressive, Moving Average (ARMA) models (Yakowitz, 1987) and Transfer Function Noise (TFN) models (Galeati, 1990), have shown NNM to produce practically equal results. Non-parametric methods have also been combined with rainfall-runoff prediction models (Smith et al., 1992) to produce long-range streamflow forecasts.

A parsimonious three parameter nearest neighbor method has been developed and tested using a database containing the average daily flows that were measured at a location over a time period spanning seventy years and: (i) forecasts are found to be best calculated using the change in flow of the nearest neighbors as a correction to the present flow, contrary to the methods presented in Mack and Rosenblatt, 1979; Yakowitz, 1987; Galeati, 1990; and Smith et al., 1992 where a weighted average of the actual neighbor flows is used; (ii) using a large database the sensitivity of the forecasts to the length of history used and the embedding

dimension (see below for definition) is investigated; (iii) the NNM forecast error is significantly less than the ARIMA forecast error.

2 Method of analysis

Although the researchers in Olason and Watt, 1986; Mack and Rosenblatt, 1979; Yakowitz, 1987; Galeati, 1990 have developed the NNM as a *statistical nonparametric forecasting method*, it is useful to consider this method within the framework of *nonlinear dynamics* (see for example Sugihara and May, 1990 for a restatement of Mack and Rosenblatt, 1979; Yakowitz, 1987; Galeati, 1990; and Smith et al., 1992 in terms of nonlinear dynamics). A few definitions are provided to fix ideas.

The NNM is based upon the assumption that the time series of concern is the output of a fundamentally deterministic system with superimposed noise.

The basic concept is the *phase space*. Many deterministic systems can be thought of as being governed by a sequence of first order coupled differential equations. If there are n variables $x_1(t)$, $x_2(t)$, ... $x_n(t)$, then the solution of this system may be visualized as a trajectory in the *phase space* of dimension n with coordinates $(x_1(t), x_2(t) \ldots x_n(t))$. Studying a trajectory in phase space is a different way of understanding the system, and in fact is more revealing than a time series. As an instance, a periodic time series is a simple closed curve in phase space.

The method of reconstructing a phase space trajectory from a single time series is simply described. If we have a scalar time series $\{y_i\}_1^N$, $y_i = y(t_i)$, and where $t_{i+1} = t_i + \Delta$, then we can associate with y_n the d_E dimensional vector $x_n = (y_n, y_{n-1}, \ldots y_{n-d_E+1})$. The parameter Δ is the embedding delay time, and may be equal to, or a multiple of the sampling interval. Thus the time series enables a trajectory $x_{d_E}, x_{d_E+1}, \ldots x_N$ to be traced out in this *embedded phase space*. There are a variety of ways of embedding, of which this is just the simplest. If the embedding is 'properly' done, then the reconstruction described above does indeed give a true (topologically similar) portrait of the phase space trajectories. It has been shown (Takens, 1984) that for a d_E dimensional system a $2d_E + 1$ dimensional reconstruction will capture the dynamics of the actual system.

Even for systems which are governed by high (or infinite) dimensional equations, for example partial or delay differential equations, it is often the case that the solutions are attracted to a finite-dimensional set, on which the dynamics may be regular (periodic, quasi-periodic) or irregular (chaotic). Although daily river flows are unlikely to be chaotic, the embedding technique provides a superior basis on which to predict them.

3 The nearest neighbor method in terms of nonlinear dynamics

A trajectory in an embedded phase space is equivalent to the evolution of pattern in the time series. To forecast the evolution of a predictee point in phase space we consider the evolution of other points in phase space that are *closest* in the Euclidean sense. The resulting forecast of the time series is thus based upon an analysis of the evolution of past patterns that are *similar* to the current pattern. In Figures 1 and 2 the evolution of the Lorenz time series and its evolution in phase space are presented. A comparison of the two shows that the evolution of a pattern in Figure 1 is easily described in terms of the evolution of a point in phase space in Figure 2.

4 Development of a parsimonious nearest neighbor method (NNM)

Researchers in Mack and Rosenblatt, 1979; Yakowitz, 1987; Galeati, 1990; and Smith et al., 1992 have defined an L step ahead forecast in terms of the weighted average of the nearest neighbors in phase space as

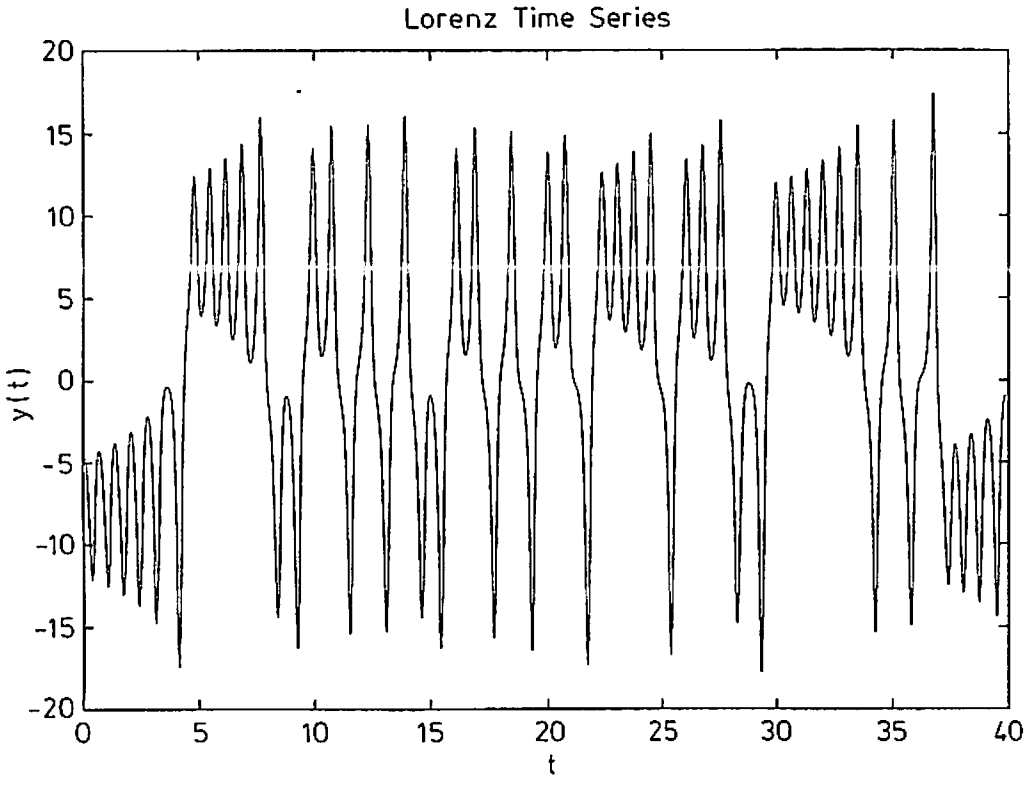

Figure 1. Time series of solution of the Lorenz equations: the Lorenz y variable versus t at r = 28, σ = 10, b = 8/3

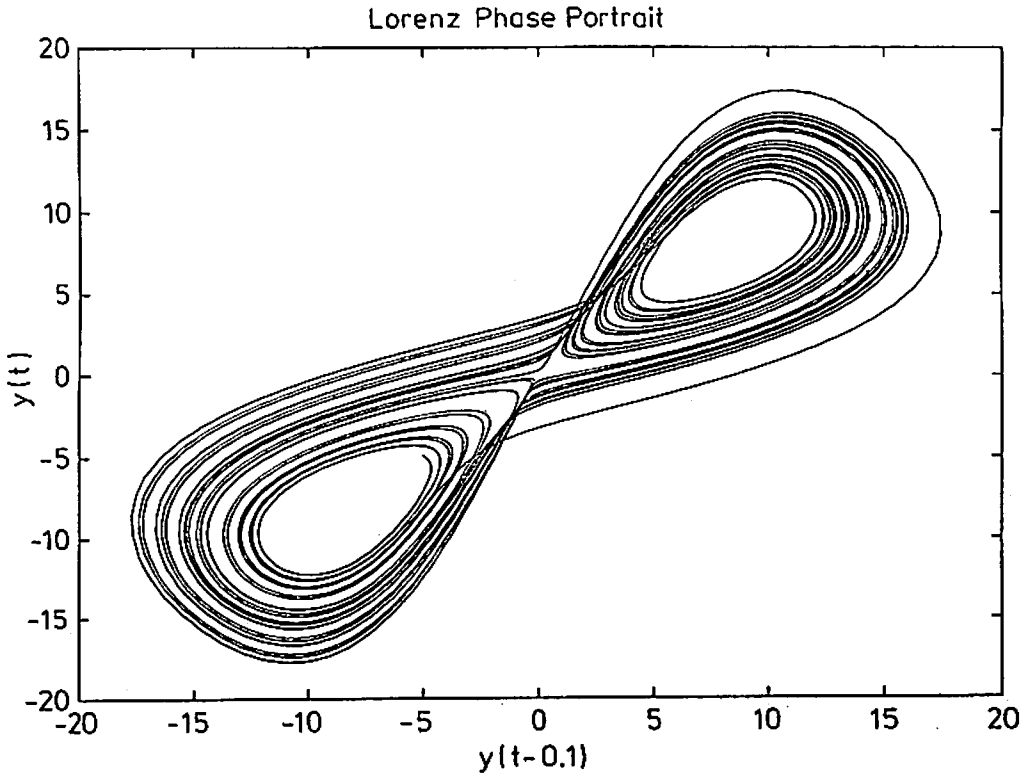

Figure 2. Two-dimensional projection of a seven-dimensional embedding of the time series shown in Figure 1, {y(t), y(t-δ), ... y(t-6δ)), where δ = 0.1. Shown is the section spanned by (y(t), y(t-δ))

$$\bar{x}_{n+L} = \frac{\sum_{i=1}^{k} w_j x_{j+L}^{(i)}}{k} \tag{1}$$

where the $x_j^{(i)}$, i = 1, 2, ... k are closest to x_n in the Euclidean sense, w_j are the weights, and we choose $k = d_E + 1$ (Sugihara and May, 1990). This forecasting definition does not recover x_n for a zero step ahead forecast and is thus biased toward the magnitude of the nearest neighbors values. If we use the *change* in the value of the nearest neighbors as a *correction* to the predictee then for a zero step ahead we recover the predictee and eliminate the above bias.

The accuracy of a forecast is also dependent in some sense on the proximity of the neighbors and therefore the neighbor corrections should be weighted in terms of the neighbor/predictee proximity. A simple weighting definition is to assign equal weights to the neighbor corrections. The group of weights is discounted uniformly, by multiplying the weights by an exponentially decaying function so that an L step ahead forecast is

$$\bar{x}_{n+L} = x_n + \frac{e^{-\lambda\rho}}{k} \sum_{i=1}^{k} (x_{j+L}^{(i)} - x_j^{(i)}), \tag{2}$$

where

$$\rho = \frac{1}{k} \sum_{i=1}^{k} \|x_j^{(i)} - x_n\|, \tag{3}$$

and $0 \leq \lambda \leq \infty$. The three parameters to be chosen are the embedding dimension d_E, the constant λ, and the embedding delay time Δ. It is useful to consider two limits in equation (2):

$$\begin{aligned} &\text{as } \lambda \to \infty\text{: } \bar{x}_{n+L} = x_n; \\ &\text{as } \lambda \to 0\text{: } \bar{x}_{n+L} = x_n + \frac{1}{k} \sum_{i=1}^{k} \left\{x_{j+L}^{(i)} - x_j^{(i)}\right\}. \end{aligned} \tag{4}$$

In the first limit, we obtain a simple forecast (i.e., present value is the forecast for the future value). The second limit $\lambda \to 0$ corresponds to the application of equal weights to the neighbor corrections for each of the predictees.

The effect of d_E is simply described. For a projection of the phase space where d_E is less than the true dimension, many trajectories will *cross* or occupy the same location. The crossing of trajectories introduces doubt as to the future evolution of a point in phase space. Increasing the dimension decreases the number of trajectory crossings. However a side effect of increasing the dimension d_E is that the trajectory becomes more spread out. It is to counter this effect that we introduced the parameter λ (see equation (2)). As the dimension d_E is increased we generally decrease the value of λ to allow for the increase in the average neighbor distance. The parameter Δ is the interval between the points taken from the time series to define the embedding (see section 2). As $\Delta \to 0$, the trajectory collapses to a line parallel to (1, 1, ... 1), and as Δ is increased from zero the amount of the embedding space explored by the trajectory generally increases. The resulting increase in resolution of the embedded series causes the prediction error to decrease initially as Δ increases. In practice

Δ is first set to the sampling width, and then increased until a minimum is encountered in the prediction error.

The calibration of the model involves the choice of the values of d_E, λ and Δ for which the forecast accuracy is optimized. For large values of N the time spent searching for neighbors in a training set increases rapidly. However Theiler, (1990) develops searching methods which reduce the number of computations involved in neighbor searches so that large values of N are not a constraint.

5 Prediction errors

The time series of daily river flows considered here is characterized by a large spring surge variation and smaller daily variation due to runoff from precipitation events (see Figure 3). The dimension d_E of the embedding phase space must be high enough to resolve spring surge and smaller daily variations in the flow.

To establish the robustness of the method it is important to investigate the dependence of forecast error on the length of history or *training set* used to construct the forecasts. We denote by N_s the number of embedded data points in the training set, taken backward from the current value, so that the training set is $\{x_{n-N_s-L+1}, ..., x_{n-L}\}$.

A portion of the time series of daily flows depicted in Figure 3 was forecasted three (L = 3) days ahead, after first taking the logarithm of the flows, to stabilize the variance. The entire historical database was searched for nearest neighbors to within L values prior to the present value, so that future information is not used. The NNM is calibrated by searching for the optimal values of d_E, λ and Δ. The prediction error proved always to be minimized for $\Delta = 1$ and $\lambda = 0$ so the calibration will be discussed in terms of the embedding dimension d_E only. The NNM forecast error ϵ is the average daily flow forecast error in $m^3 s^{-1}$. The relative improvement is depicted in Figure 4. This is defined in terms of the ratio of the error obtained using a simple forecast, ϵ_∞, to ϵ: specifically, it is $1 - (\epsilon/\epsilon_\infty)$. To simplify the presentation of the results, the error is depicted in Figure 4 as a function of the embedding dimension d_E for different training sets N_s. The forecast errors in Figure 4 are presented for the optimal value of λ and are calculated for actual flows greater than 20 $m^3 s^{-1}$. It is clear from Figure 4 that the forecast error is decreasing as the training set N_s increases. For each training set the forecast error has a global maximum which converges (for all histories) at a relatively low dimension of $d_E = 6$. Figure 5 shows the best prediction, while Figure 6 shows a typical one. It can be seen in both cases that the series is well predicted, and in particular the magnitude of the spring surge.

The statistical analysis involved the fitting of Box Jenkins, multiplicative seasonal ARIMA models. A range-mean analysis indicated a logarithmic transform to stabilize the variance and persistent large correlations in the autocorrelation function were removed by a first difference of the log-transformed data. The ARIMA model parameters were estimated with maximum likelihood for various histories ($N_s \geq 2000$) using the same data as the NNM and then forecasted for the same segment as the NNM within the calibration region. The ARIMA models were verified by testing model residuals using the Box-Ljung portmanteau statistic (Box and Ljung, 1978) and the residual autocorrelation of the residuals (McLeod, 1978) was considered. The relative gain of the ARIMA models was less than about 0.1 for all of the models tested. The ARIMA models were found to have forecast errors consistently greater than the NNM forecast errors within the range of parameters which we chose.

6 Conclusions

A parsimonious three parameter nearest neighbor method was developed and tested. The model robustness was demonstrated by investigating the length of history used to make forecasts. Comparison to an ARIMA model shows that the NNM gives consistently improved

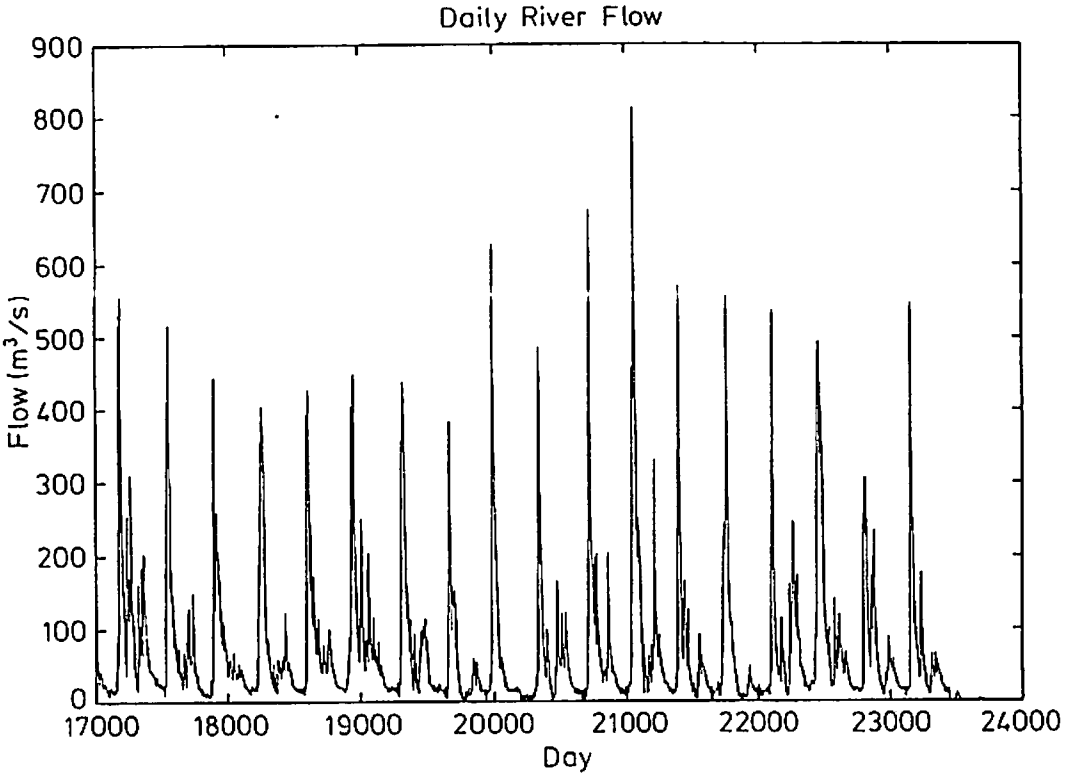

Figure 3. Time series of daily river flows at Spruce Falls, Northern Ontario, 1967-1986

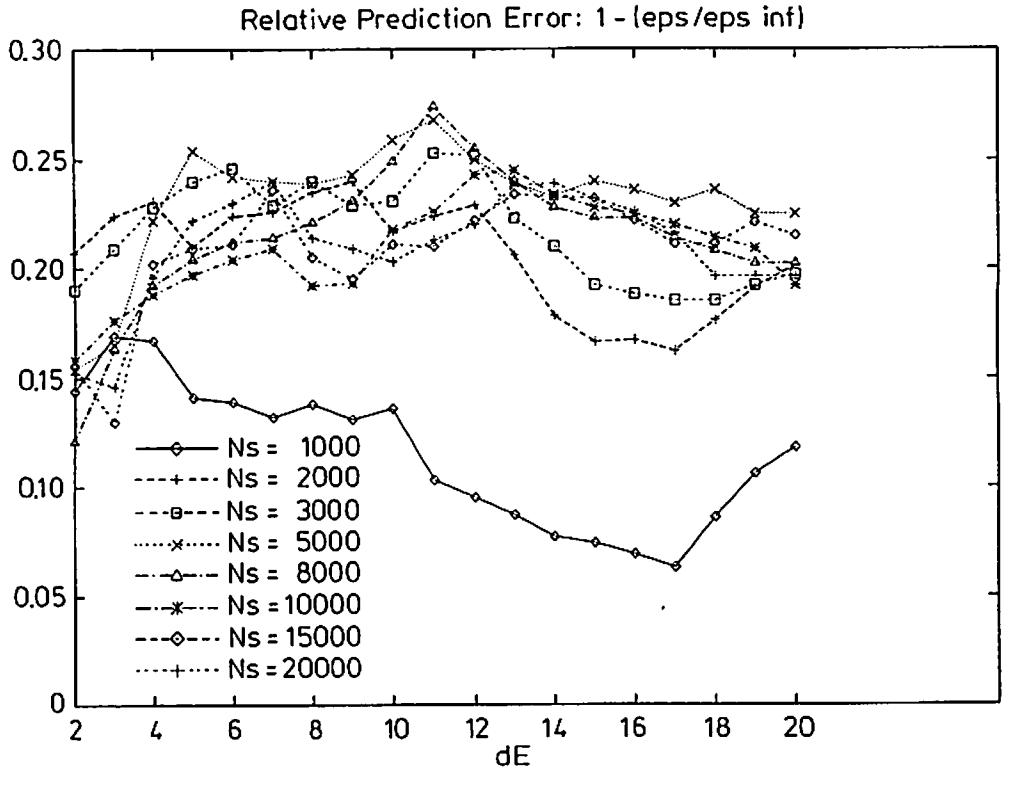

Figure 4. Relative improvement, defined as 1 - $(\epsilon/\epsilon_\infty)$, where ϵ is the forecast error, and ϵ_∞ is that obtained using a simple forecast, equation (4) ($\lambda \to \infty$). For the NNM ϵ is optimized with respect to λ, and the improvement is shown as a function of embedding dimension, d_E.

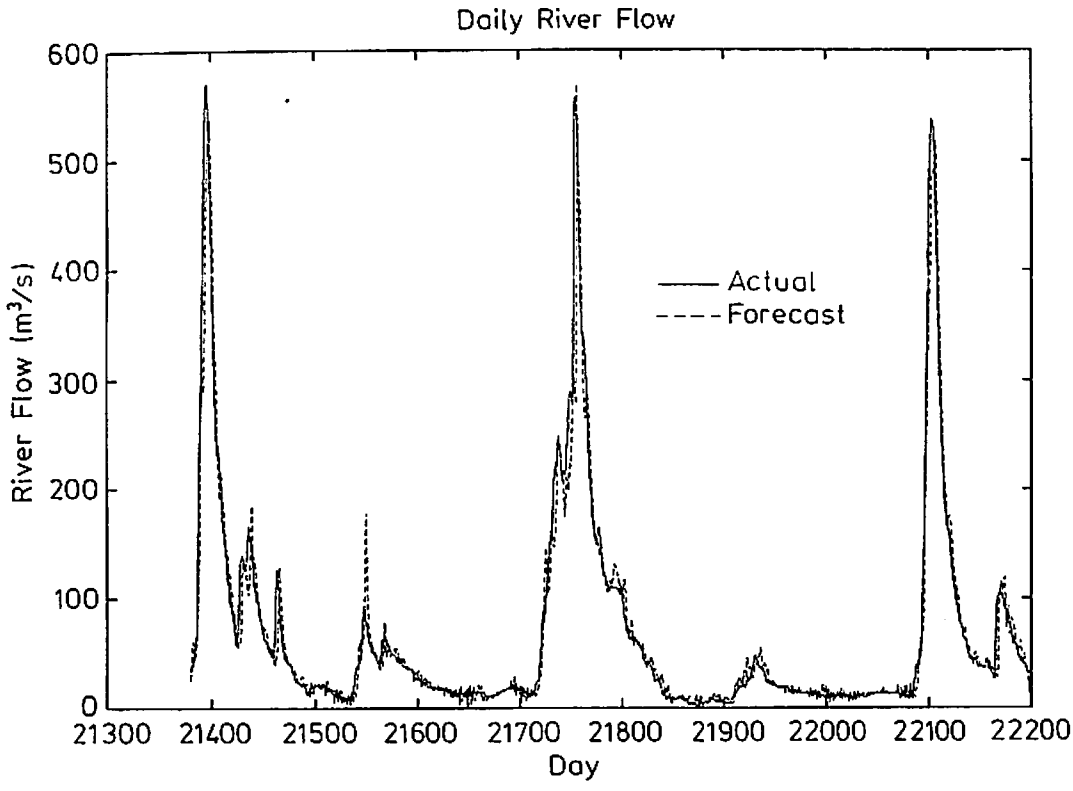

Figure 5. Actual data versus forecast for optimized NNM with $d_E = 11$, $\Delta = 1$, $\lambda = 0$, and $N_s = 8000$. The dashed curve represents the predictions

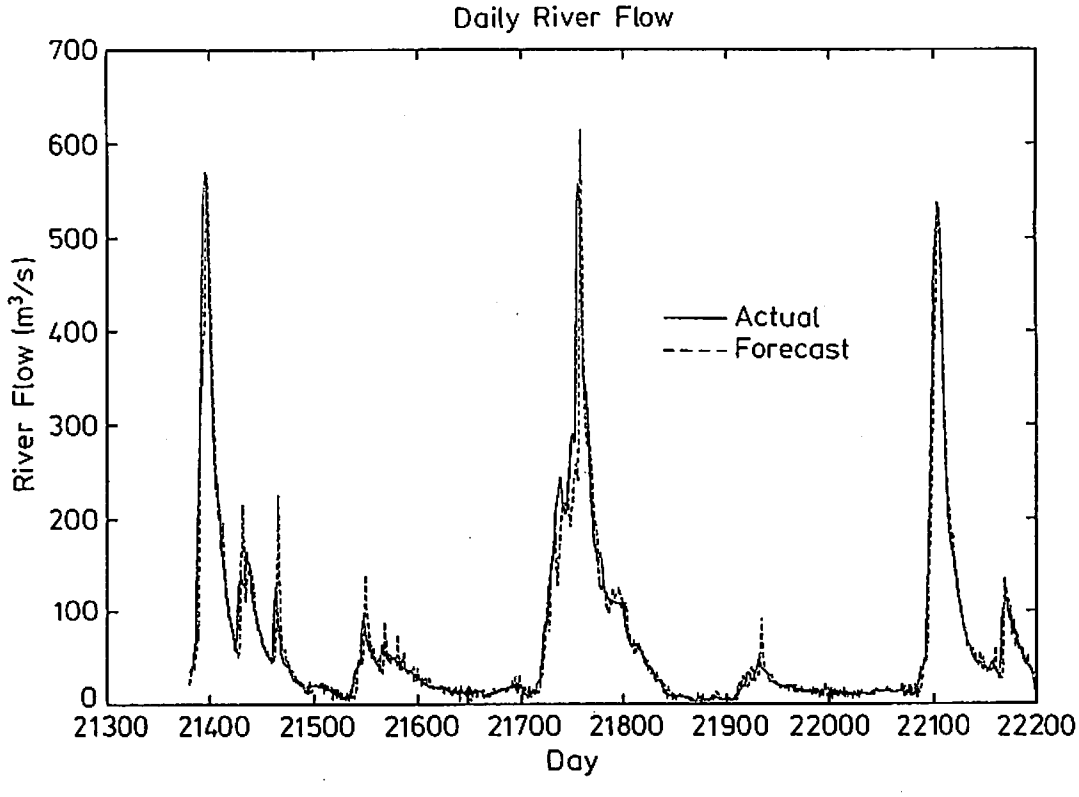

Figure 6. As for Figure 5, but with $d_E = 6$, $N_s = 2000$

forecasting ability (Galeati [3] and Yakowitz [2] show the method could provide practically the same results but do not investigate robustness). The intuitive clarity and the simplicity of the model implementation make it generally useful.

References

Box, G. E. P.; Jenkins, G. M. 1976: Time series analysis, forecasting and control. San Francisco: Holden-Day

Olason, T.; Watt, W. E. 1986: Multivariate transfer function-noise model of river flow for hydropower operation. Nordic Hydrology 17, 185-202

Mack, Y. P.; Rosenblatt, M. 1979: Multivariate k-nearest neighbor density estimates. Multivariate Anal, 9, 1-15

Yakowitz, S.; Karlsson, M. 1987: Nearest neighbor methods for time series with application to rainfall-runoff prediction. Stochastic Hydrology, 149-160

Galeati, G. 1990: A comparison of parametric and non-parametric methods for runoff forecasting. Hydrological Sciences Journal 35, 79-94

Smith, J. A.; Day, G. N.; Kane, M. D. 1992: Nonparametric framework for long-range streamflow forecasting. ASCE Journal of Water Resources Planning and Management 118, 82-92

Sugihara, G.; May, R. M. 1990: Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature, 734-741

Takens, F. 1981: Detecting strange attractors in fluid turbulence. In: D. Rand and L.-S. Young (eds.) Dynamical systems and turbulence. Springer-Verlag, pp. 366-381

Theiler, J. 1990: Efficient algorithm for estimating correlation dimension. Phys. Rev. A. 36, 4456-4462

Box, G. E. P.; Ljung, G. M. 1978: On a measure of lack of fit in time series models. Biometrika 65, 297-303

McLeod, A. I. 1978: On the distribution of the residual autocorrelations in Box-Jenkins models. J. Roy. Statist. Soc. Set. B 40, 296-302