Derivation of effective and efficient data set with subtractive clustering method and genetic algorithm

C. D. Doan, S. Y. Liong and Dulakshi S. K. Karunasinghe

ABSTRACT

Success of any forecasting model depends heavily on reliable historical data, among others. Data are needed to calibrate, fine tune and verify any simulation model. However, data are very often contaminated with noise of different levels originating from different sources. This study proposes a scheme that extracts the most representative data from a raw data set. Subtractive Clustering Method (SCM) and Micro Genetic Algorithm (mGA) were used for this purpose. SCM does (a) remove outliers and (b) discard unnecessary or superfluous points while mGA, a search engine, determines the optimal values of the SCM's parameter set. The scheme was demonstrated in: (1) Bangladesh water level forecasting with Neural Network and Fuzzy Logic and (2) forecasting of two chaotic river flow series (Wabash River at Mt. Carmel and Mississippi River at Vicksburg) with the phase space prediction method. The scheme was able to significantly reduce the data set with which the forecasting models yield either equally high or higher prediction accuracy than models trained with the whole original data set. The resulting fuzzy logic model, for example, yields a smaller number of rules which are easier for human interpretation. In phase space prediction of chaotic time series, which is known to require a long data record, a data reduction of up to 40% almost does not affect the prediction accuracy.

Key words | Chaos, fuzzy inference system, genetic algorithm, neural networks, subtractive clustering method

C. D. Doan
S. Y. Liong (corresponding author)
Dulakshi S. K. Karunasinghe
Civil Engineering Department,
National University of Singapore,
Singapore 117576,
Tel: +65 6874 3081
Fax: 65 6779 1635
E-mail: *cvelsy@nus.edu.sg*

NOTATION

d	dimension of the state space
k	number of nearest neighbors
k'	modified number of nearest neighbors
T	lead time
x_t	observed value of the time series at time t
$\bar{x}$	mean value of the time series
$\hat{x}_i$	predicted value of x_i
$d_{\min}$	shortest distances between the point under consideration and all previously found cluster centers.
F	objective function
f_T	mapping function
F_T	predictor which approximate f_t
m	embedding dimension
n	total number of data
P	potential value, penalty term
P_1^*	potential value of the first cluster center
P_k^*	potential value of the k^{th} cluster center;
R^2	Nash index, coefficient of efficiency
R_a	influence range
R_b	neighborhood range
X_t	current state/phase space vector
X_{t+T}	future state
τ	time delay

ABBREVIATIONS

ANN	Artificial Neural Networks
AR	Accept Ratio
CAL	Calibration Set

DoS	Degree of Support
EA	Evolutionary Algorithm
FIS	Fuzzy Inference System
FL	Fuzzy Logic
GA	Genetic Algorithm
GEATbx	Genetic And Evolutionary Algorithm Toolbox
mGA	Micro-Genetic Algorithm
MLP	Multi-Layer Perceptrons
MSE	Mean Squared Error
NLP	Nonlinear Prediction
NRMSE	Normalized Root Mean Square Error
PROD	Production Set
RR	Reject Ratio
SCM	Subtractive Clustering Method
SF	Squash Factor
SSR	State Space Reconstruction
ST	Gauging Station

INTRODUCTION

Although the performance capability of the model depends very much on the model itself, it is a known fact that data play a very crucial role in building quality models. This is particularly more so for a data-driven model since this type of model, as the name suggests, relies heavily on data. There are, however, many problems related to data as data often suffer from noise, inconsistency, redundancy, etc. Noisy data will result in an inappropriately constructed model. Thus, the resulting model should not be expected to yield good and reliable predictions. Apart from noise, a data set could also contain conflicting data – the same input with differing output (Figure 1); these may confuse models, in particular data-driven models, during the training period.

A data clustering technique, the Subtractive Clustering Method or SCM (Chiu, 1994), is considered in this study to select only the most representative patterns, from the given data record, to be used for forecasting models. SCM could (a) remove outliers, (b) resolve conflicting data and (c) discard unnecessary or superfluous patterns. Since SCM contains several parameters whose values need to be calibrated to yield an acceptable model prediction error, a Micro Genetic Algorithm (mGA) is used to determine its optimal parameter set. This study shows the derivation of the compact data and its applications in neural network, fuzzy logic and chaos theory.

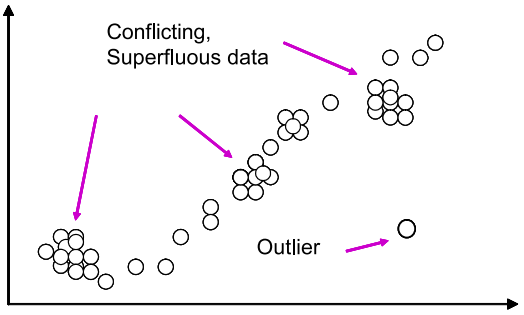

Figure 1 | Conflicting, superfluous data and outlier.

In this study, three data sets are used: (1) water level at Dhaka, Bangladesh; (2) discharge in the Wabash River at Mt. Carmel, USA and (3) discharge in the Mississippi River at Vicksburg, USA. The first data set is applied to Neural Network and Fuzzy Logic forecasting models. The second and third data sets are used with a forecasting model based on chaos theory. The prediction accuracy obtained with the assistance of the compact data set is then compared with that obtained from a model trained (or calibrated) with the whole original data set.

METHODOLOGIES

Subtractive clustering method

The Subtractive Clustering Method was developed by Chiu (1994). In Chiu's work, SCM was used as the basis of a fast and robust algorithm for identifying fuzzy models. The fuzzy logic model was then benchmarked with other fuzzy models; the comparison was done using chaotic time series.

Chiu's method of clustering is a modification of mountain clustering (Yager & Filev, 1994) which was also used to identify the number of fuzzy rules. The differences between the two clustering methods are mainly in the way of estimating potential values and the influence of a neighboring data point. In addition, the process of acquiring a new cluster center and revising potentials was amended to ease the difficulty in establishing a very sensitive parameter.

The subtractive clustering method assumes that each data point is a potential cluster center. A data point with more neighboring data will have a higher opportunity to

become a cluster center than points with fewer neighboring data. Based on the density of surrounding data points, the potential value for each data point is calculated as follows:

$$P_i = \sum_{j=1}^{n} e^{-4\|x_i - x_j\|^2/R_a^2} \quad (1)$$

where x_i, x_j are data points and R_a is a positive constant defining a neighborhood. Data outside this range have little influence on the potential.

After the potential of every data point has been computed, the data point with the highest potential is chosen as the first cluster center. Let x_1^* be the location of the first cluster center and P_1^* be its potential value. The potential of the remaining data points x_i is then revised by

$$P_i \Rightarrow P_i - P_1^* e^{-4\|x_i - x_1^*\|^2/R_b^2} \quad (2)$$

where R_b is a positive constant ($R_b > R_a$). Generally, after the kth cluster center has been obtained, the potential of each data point is revised by the formula

$$P_i \Rightarrow P_i - P_k^* e^{-4\|x_i - x_k^*\|^2/R_b^2} \quad (3)$$

Thus, the data points near the first cluster center will have greatly reduced potential, and therefore are unlikely to be selected as the next cluster center. The constant R_b is the radius defining the neighborhood that will have measurable reductions in potential. To avoid obtaining closely spaced cluster centers, R_b is set to be greater than R_a. Since the parameters R_a and R_b are closely related to each other and R_b is always greater than R_a, the parameter R_b can be replaced by another parameter called the Squash Factor (SF) which is the ratio between R_a and R_b:

$$SF = \frac{R_b}{R_a} \quad (4)$$

The process described in Equation (3) continues until no further cluster center is found. As for whether a data point is chosen as a cluster center or not, there are two parameters involved, the Accept Ratio (AR) and the Reject Ratio (RR). These two parameters, together with the influence range and squash factor, set the four criteria for the selection of cluster centers.

The first criteria states that, if the potential value ratio of the current data point to the original first cluster center is larger than the Accept Ratio, then the current data point is chosen as a cluster center. Therefore, the larger the value of the Accept Ratio, the fewer the number of chosen cluster centers.

If the potential value falls in between that of the Accept and Reject Ratios, then the compensation between the magnitude of that potential value and the distance from this point to all the previous chosen cluster centers is taken into consideration. The data point is only accepted as a cluster center if the sum of the potential value and the ratio of the shortest distance between the current data point and all other previously found cluster centers to the influence range is more than or equal to 1. Otherwise, the potential value of that data point is revised to zero and the data point with the next highest potential is tested.

It should be noted that, in general, the value of RR should be less than AR. But if the RR value is greater than or equal to the AR value, all the cluster centers chosen will be from the first criteria and none will be in the region defined by the trade-off between the magnitude and distance criteria.

The criteria used in SCM for accepting or rejecting data points as cluster centers and the corresponding parameters are summarized in Table 1.

Although there are four parameters in the SCM method, Chiu (1994) in his work used a trial-and-error approach to vary only the cluster radius R_a in determining the number of Takagi–Sugeno type rules (Sugeno and Kang 1988). He then optimized the consequences of these rules by applying the least-square estimation on the coefficients of the first-order fuzzy model. The computational time was found to be fast and the accuracy level was also relatively high in comparison to several other fuzzy models.

In this paper, the merits of the SCM method, such as robustness to noisy data, cluster centers being a subset of the actual data, etc., are employed not only to define the number of fuzzy rules but, more generally, to extract the efficient and effective data set readily for use in any application, for instance, Neural Network, Fuzzy Logic or Chaos Technique. The ability of the SCM method is exploited to the fullest by integrating it with an optimization scheme, Genetic Algorithm. All four SCM parameters are optimized simultaneously in extracting the representative data. The performance of the extracted data set in any application is assessed on both the training and verification sets.

Table 1 | Summary of criteria for accepting or rejecting points as cluster centers

Criterion	Detail	Note
First	Potential value $> AR \times P_1^*$	
Second	$RR \times P_1^* <$ potential value $< AR \times P_1^*$ and $\frac{d_{\min}}{R_a} + \frac{P_k^*}{P_1^*} \geq 1$	Criteria for accepting points as cluster centers
Third	$RR \times P_1^* <$ potential value $< AR \times P_1^*$ and $\frac{d_{\min}}{R_a} + \frac{P_k^*}{P_1^*} < 1$	Criteria for rejecting points as cluster centers
Fourth	Potential value $< RR \times P_1^*$	

Note: P_1^*: the potential value of the first cluster center; P_k* : potential value of the kth cluster center; $d_{\min}$: the shortest distances between the point under consideration and all previously found cluster centers.

Genetic algorithm

To choose the optimal set of SCM parameters, SCM is coupled with micro-GA or mGA (Krishnakumar 1989), a variant of GA (Holland 1975, Goldberg 1989). Unlike traditional optimization methods, GAs operates with more than one parameter set. With a large number of parameter sets, it gives GAs the advantage of avoiding the possibility of getting trapped in the local optima. Another advantage of GAs over the traditional optimization method is that GAs do not require the objective function to be explicit or differentiable. It can also be used when the objective function is discontinuous.

Each parameter set, known as a chromosome, comprises a series of parameters known as genes. There is a fitness associated with each chromosome depending on the chromosome's performance. Each parameter value and its fitness are coded, generally in binary form. The total number of chromosomes in each GA iteration is known as the population size. The chromosomes in each iteration are collectively known as a generation. In GAs, the size of a population is usually the same from one generation to the next. The chromosomes of the very first generation are usually generated through a random process. However, the chromosomes of the subsequent generations are generated through four basic mechanisms, namely:

(1) fitness based selection of parent chromosomes for mating,
(2) recombination/crossover of parents to produce offspring,
(3) mutation of offspring,
(4) re-insertion of offspring into population.

As the population evolves, new chromosomes replace the older ones and are supposed to perform better. The chromosomes in a population associated with the very fit individuals will, on average, be produced more often than the less-fit chromosomes. This is known as the principle of the "survival of the fittest". This process of generating new individuals is repeated until a predetermined stopping criterion is met.

mGA, in principle, uses the same algorithm as GA without implementing the conventional mutation and has a much smaller population size than that of the conventional GA. Whenever convergence in mGA occurs, a totally new population set, with the exception of the best chromosome (elitism), is randomly generated. The flowchart describing the operational procedure of the coupled SCM and mGA is drawn in Figure 2.

APPLICATIONS ON NEURAL NETWORK AND FUZZY LOGIC

The data used in this section are the daily water levels in Bangladesh. Bangladesh, a land area of 145,000 km^2, is located on the world's largest delta comprising three of the world's most unstable rivers, the Ganges, the Brahmaputra and the Meghna. The rivers flowing into Bangladesh drain some of the wettest catchment areas on Earth with average annual rainfalls as high as 1100 cm. The major rivers and their tributaries have their origins outside Bangladesh and only about 7.5% of their total catchment area of about 1500,000 km^2 lies within Bangladesh (Liong *et al.*, 2000).

In this study data from 5 stations during the monsoon season from 1991–1996 are considered. The monsoon

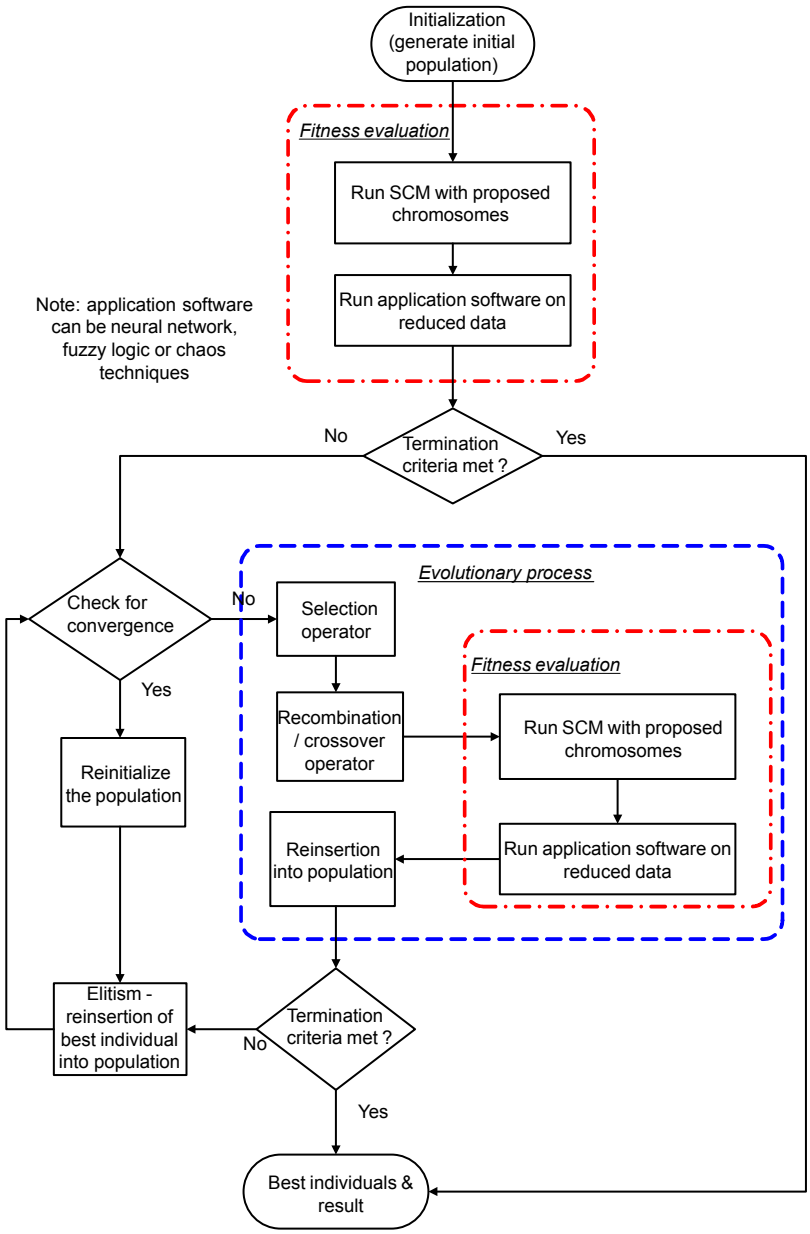

Figure 2 | Flowcharts of coupled SCM and GA and their application.

period is between May and September. The total number of data patterns is 841. The lowest water level in the period was recorded as 1.28 m at ST12 (Dhaka) while the highest water level was recorded as 24.71 m at ST11. The average water levels for ST33, ST11, ST14, ST17 and ST12 were 13.7, 22.1, 11.5, 7.4 and 3.9 m, respectively. The data set is divided into two parts. The first part containing 467 patterns (originated from 1992, 1993 and 1995) is used for training while the second part containing 374 patterns (originated from 1991, 1994 and 1996) is used for verification. The schematic diagram of the river network and the positions of the 5 water level stations are shown in Figure 3. It should be noted that all 5 stations, with the exception of the Dhaka station (Station 12), are located close to the boundary between Bangladesh and India.

To extract the most representative data set, SCM, mGA and Multilayer Perceptron Neural Network (MLP) are coupled. The SCM and MLP used are the toolboxes in

Matlab while the main principle of mGA (Krishnakumar, 1989) is incorporated in the Genetic and Evolutionary Algorithm Toolbox (GEATbx) of Pohlheim (2000) in Matlab. In this study, 200 generations and a population size of 10 per generation is employed. The parameter set of SCM (R_a, SF, AR and RR) are genes of the chromosomes of mGA. The range of each of these real parameters is as follows:

(1) the influence range, R_a: [0.001,1];
(2) the squash factor, SF: [1,2];
(3) the accept ratio, AR: [0,1];
(4) the reject ratio, RR: [0,1].

The objective function (F) in mGA is to minimize the sum of a mean-squared-error (MSE) term and a penalty term (P). The penalty term is introduced to yield a reduction in the number of training patterns:

$$\text{Minimize } F = MSE + P \tag{5}$$

The procedure of estimating the objective function is as follows:

(1) Extract the training set with SCM based on the chromosome (parameter set) suggested by mGA.
(2) Train the artificial neural network (ANN) with the reduced training data set.
(3) Utilize the original entire training data set (before reduction) as the test set. The trained ANN is tested and its simulated output is compared with the measured data. The mean square error (MSE) is then computed as

$$MSE = \frac{\sum_{i=1}^{n}(\hat{x}_i - x_i)}{n} \tag{6}$$

It should be noted that $\hat{x}_i$ is the ith output simulated by a model trained with the extracted data set while x_i is the ith measured data and n is the number of data considered.

(4) Compute the objective function F. This value is used by the evolutionary algorithm to guide the search for more fit chromosomes:

$$F = MSE + P \tag{7}$$

where P is the penalty term computed as follows:

$$P = max\_penalty \times \frac{\text{number of patterns extracted by SCM}}{467} \tag{8a}$$

$$max\_penalty = 0.2 \tag{8b}$$

The purpose of the penalty term (P) is to penalize the fitness/objective value of the parameter set if the resulting data set size extracted is large. This term increases proportionately with the size of the data set and has the largest penalty value (0.2 in this study) when the entire data set is used for training. The linear penalty term is graphically represented in Figure 4.

It should be noted that if *max_ penalty* is set too high, this will result in a P term dominating the objective function, Equation (7). Setting it too low, however, will result in the domination of the MSE in the objective function. Since the MSE ranges between 0.005 and 0.5 in this study, limited values of *max_ penalty* (0.1, 0.2, 0.5) are tried. The best result is obtained when the *max_ penalty* is set as 0.2. Hence, the value of 0.2 is recommended for this study. An overview flow of the processes within the objective function is given in Figure 5.

A "goodness-of-fit" of the Nash–Sutcliffe statistical measure (R^2), recommended by ASCE (1993) for hydrological studies, is used:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{9}$$

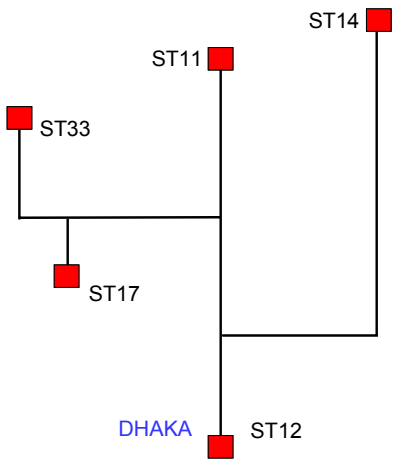

Figure 3 | Schematic diagram of river system and water level stations.

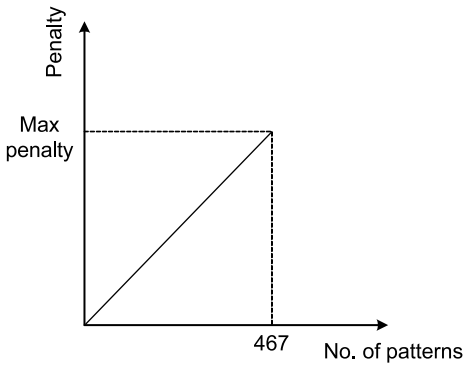

Figure 4 | Graphical representation of the linear form of the penalty term.

where x_i and $\hat{x}_i$ are the ith measured and simulated data, $\bar{x}$ is the mean value of the measured data and n is the number of data considered. When $R^2 = 1$, it implies a perfect fit between the predicted and the observed data; when $R^2 = 0$ it implies that the model is predicting no better than the average of the observed data.

Applying a derived effective and efficient data set to a neural network

Results from the mGA suggest one optimal set of parameters satisfying two criteria: (1) a high R^2 value and (2) a significantly smaller number of patterns extracted. The optimal set ($R_a = 0.232\ 34$, $SF = 1.690\ 17$, $AR = 0.171\ 88$, $RR = 0.316\ 34$) yields only 13 (out of 467) most representative patterns.

Table 2 lists the prediction accuracies of two models. Model 1 is trained with 467 patterns while Model 2 is trained with only 13 patterns. The performances of the two models are applied to both the training and the verification data sets. Results show that the neural network model trained with the smaller, yet most representative, number of patterns (13) yields an equally high prediction accuracy (with the difference only in the third and four digits of the Nash index) as that trained with the entire number of patterns (467). It is also noted that the "trivial" or "naive" forecasting gives a R^2 value of 0.9584 for the verification set (374 patterns). The comparison between the observed and simulated hydrographs and the scatter plot of Model 1 is given in Figure 6. Figure 7 shows similar comparisons resulting from Model 2.

Applying a derived effective and efficient data set to a fuzzy inference system

The result of the extracted data set in the previous section could be utilized in establishing a fuzzy model. However, as for Mamdani fuzzy rule types (Mandani and Assilian 1975, 1999), each data pattern is required to build one rule in the

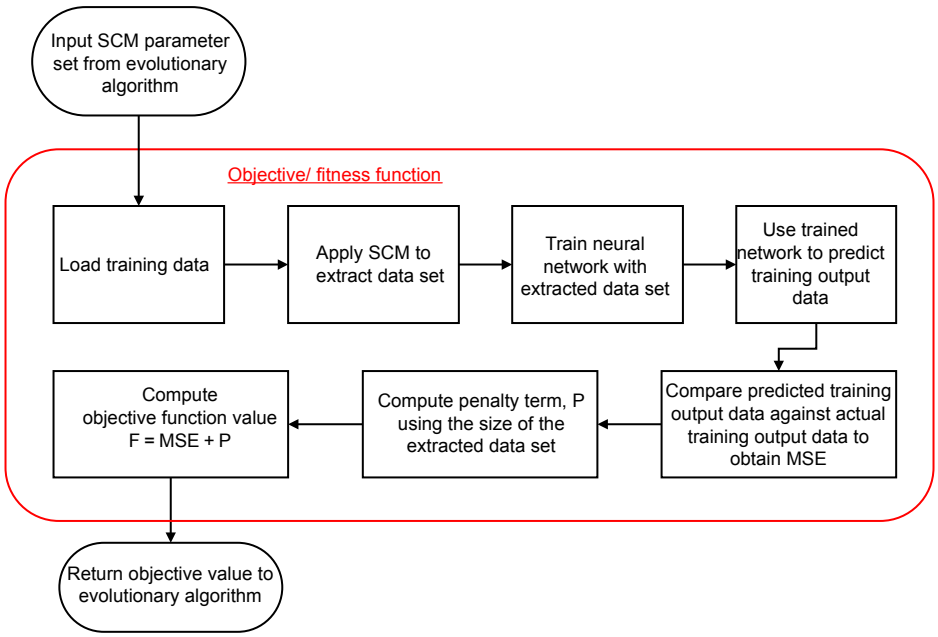

Figure 5 | Structure of the objective function.

Table 2 | Prediction accuracy of two Multi-Layer Perceptron Neural Networks

	Goodness-of-fit: R^2 value	
Data set used for testing	**Model 1(trained with 467 patterns)**	**Model 2(trained with 13 patterns)**
Applied on the whole original data set (467 patterns)	0.9949	0.9943
Applied on the verification data set (374 patterns)	0.9911	0.9937

fuzzy model. Therefore, if too little data are utilized it will result in an insufficient fuzzy rule base. Consequently, the system described by this rule base is underdetermined.

With the need for having a sufficient number of training data patterns in certain situations, fuzzy logic in this case, the penalty term (P) has to be altered so that the proposed scheme will extract the user-specified number of patterns. The following piecewise form of penalty term (Figure 8) is proposed:

$$P = \begin{cases} max\_penalty \times \frac{Patterns_{min} - Patterns}{Patterns_{min}} & Patterns \le Patterns_{min} \\ 0 & Patterns_{min} \le Patterns \le Patterns_{max} \\ max\_penalty \times \frac{Patterns - Patterns_{max}}{467 - Patterns_{max}} & Patterns \ge Patterns_{max} \end{cases} \quad (10)$$

where

$$Patterns_{min} = 0.95 \times \text{targetted number of patterns}$$
$$Patterns_{max} = 1.05 \times \text{targetted number of patterns.}$$

The penalty function, Equation (10), penalizes chromosomes with a number of patterns deviating from ±5% of the desired number of patterns. The more the deviation is the heavier is the penalty.

The experiment conducted tried several numbers of patterns (50, 100, 150 and 200). The study shows the best result is when 150 patterns are chosen. The following paragraph describes the experiment with 150 patterns.

With 150 patterns, chromosomes with the number of patterns ranging from 142 to 157 will not be penalized. The optimal parameter set obtained from SCM and mGA with this piecewise penalty term ($R_a = 0.0838$, $SF = 1.6884$, $AR = 0.4736$ and $RR = 0.1312$) ultimately yields 145 patterns. Two models are then built. One model, Model 1, uses the whole original training data set (467 patterns) to construct the rule base while the other model, Model 2, is constructed only with the extracted data set (145 patterns).

Membership functions with standard triangular shapes and high shoulders are used for each variable in both models. The approach of Wang & Mendel (1992) is used to construct the rule-base. In this approach, each data pattern is represented by one rule. The value of Degree of Support (DoS) for each rule is then calculated. The DoS value originates from the product of all maximum membership degree values of every linguistic variable in both the antecedent and the consequent parts of each rule. There are, however, many rules that are identical in both the

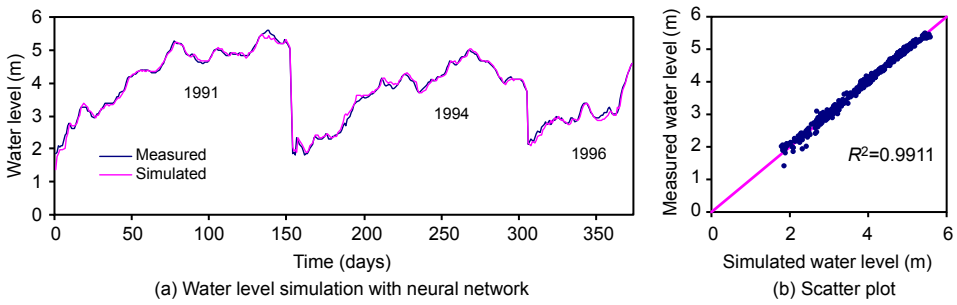

Figure 6 | Simulated and measured daily water levels of Model 1 (467 data): verification.

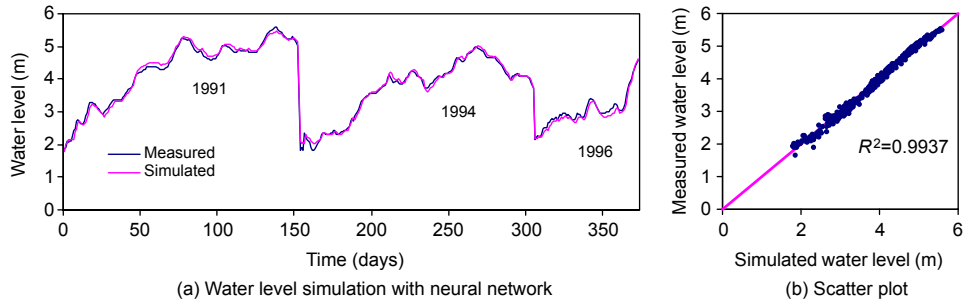

Figure 7 | Simulated and measured daily water levels of Model 2 (13 data): verification.

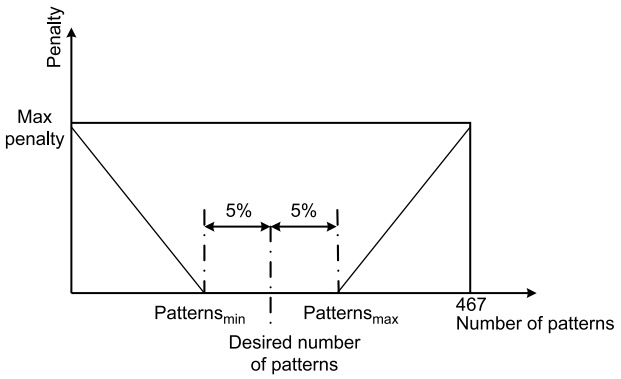

Figure 8 | Graphical representation of the piecewise form of the penalty term.

antecedent and consequent parts and yet have different DoS values. To overcome these conflicting rules, only the rules with maximum DoS values are kept in the rule-base. The constructed models are then applied on the verification data set to test both their prediction accuracy and their complexity (number of rules).

Table 3 summarizes the performance and the complexity of each Fuzzy Inference System, FIS. It shows that Model 2, the model trained with the extracted data set, yields equally high performance accuracy ($R^2 = 0.9596$) as that of Model 1 ($R^2 = 0.9577$). In addition, the compactness of the extracted data set, and hence the number of rules, makes the constructed FIS more manageable. The number of rules in Model 2 is 85 which is lower than that of Model 1 (110 rules).

The fuzzy inference system can be further enhanced by adjusting the number of membership functions of each variable according to the important degree of each variable. Liong & Doan (2002) conducted a study investigating the important degrees of the stations in the Bangladesh river system using a General Regression Neural Network. Based on that study, the number of membership functions for each station is adjusted and the result is shown in Table 4.

Table 3 | Performance of two Fuzzy Logic models and their number of rules

FIS model	Configuration	R^2	Number of rules
Model 1	467 patterns, 5 labels for each variable	0.9577	110
Model 2	145 patterns, 5 labels for each variable	0.9596	85

Table 4 shows that in Model 4, after extracting the number of data and properly assigning the number of membership functions, the number of rules has been greatly reduced to 35 rules (from 110 rules in Model 1). At the same time, an equally good prediction accuracy is obtained from Model 4 (0.9639) as from Model 1 (0.9577).

It is noted that the performance of the Fuzzy Logic is lower than that of the Neural Network. However, this is expected since Fuzzy Logic "fuzzifies" crisp values into linguistic terms through a membership function. This feature does not give an advantage to Fuzzy Logic if the crisp values are used. When crisp values are not available (due to equipment failure, for example) Fuzzy Logic will have advantages over other methods.

APPLICATION ON PHASE SPACE PREDICTION OF CHAOTIC TIME SERIES

Introduction

Prediction of time series using the theory of dynamical systems has gained much interest lately. However, one of

Table 4 | Performance of adjusted two Fuzzy Logic Models and their number of rules

FIS Model	Configuration	R^2	Number of rules
Model 3	145 patterns,	0.9590	58
	5 labels for ST12, ST11		
	3 labels for ST14, ST17, ST33		
Model 4	145 patterns, 5 labels for each variable	0.9639	35
	5 labels for ST12, ST11		
	2 labels for ST14, ST17, ST33		

the difficulties with this approach is the large amount of past records to be handled in computations. It is generally assumed that the larger the number of past records the better the predictions could be made. However, no investigation has been undertaken to see whether all data contribute valuable information for phase space prediction. This section investigates the possibility of using the above data reduction procedure to extract an informative, dynamics preserved small data set from a large data set for phase space prediction.

River flow prediction: phase space prediction approach

The embedding theorem of Takens (1981) is the stepping-stone for the dynamical systems approach for analysis of chaotic time series. The theorem establishes that, using only a single variable of a dynamical system, it is possible to reconstruct a state space that is equivalent to the original (but unknown) state space of all the dynamical variables. The state space is defined as the multidimensional space whose axes consist of variables of a dynamical system. When the state space is reconstructed from a time series, it is called a phase space. Thus, if a time series is identified as chaotic, a phase space can be reconstructed and the phase space based prediction methods can be employed to forecast the time series. In phase space prediction of chaotic river flow time series, no explicit governing equations, explaining the dynamics of river flow, exist. Instead, the predictive model is directly constructed from river flow time series. There are many methods to reconstruct the phase space. The time delay coordinate method (e.g. Packard *et al.*, 1980; Takens, 1981) is currently the most popular choice and is used in this study.

Phase space reconstruction

Let $x_1, x_2, \ldots x_n$ be the scalar time series of a variable. The dynamics of the time series can be embedded in an m-dimensional phase space ($m > d$, where d is the dimension of the state space) given by

$$X_t = \{x_t, x_{t-\tau}, \ldots, x_{(t-(m-1)\tau)}\} \tag{11}$$

where τ is the time delay. In chaos literature, there is more than one method to locate suitable values for m and τ. But they do not necessarily provide reconstructions which lead to the best forecasting accuracies. Therefore, various researchers have suggested the use of inverse approaches to find optimal phase space parameters (e.g. Babovic *et al.*, 2000; Phoon *et al.*, 2002). In this study, the phase space parameters (m, τ) and the number of nearest neighbours, k (a prediction parameter) are determined by an inverse approach so that the prediction accuracy is optimal.

Phase space prediction method

In phase space prediction, the basic idea is to set a functional relationship between the current state X_t and a future state X_{t+T} in the form

$$X_{t+T} = f_T(X_t) \tag{12}$$

where T is referred to as the lead time. At time t, for an observation value x_t, the current state of the system is X_t, where $X_t = (x_t, x_{t-\tau}, \ldots, x_{t-(m-1)\tau})$, and the future state at time $t+T$ is X_{t+T}, where $X_{t+T} = (x_{t+T}, x_{t+T-\tau}, \ldots, x_{t+T-(m-1)\tau})$. For a chaotic system, the predictor F_T that approximates f_T is necessarily nonlinear. There are two strategies to obtain F_T; (1) global approximation (e.g. neural network, polynomial and rational function, etc.) and (2) local approximation (e.g. Farmer & Sidorowich, 1987; Casdagli, 1989). In global approximation a function F_T, which is valid over the entire state space, is approximated.

Nonlinear prediction method (NLP) with local approximation

In local approximation, only the states near the current state are used to make predictions. To predict a future state

X_{t+T}, an Euclidean metric is imposed on the phase space to find the k nearest neighbours of the current state X_t. Once the nearest neighbours, X_i, are found, one can project each of these states X_i to their respective future states X_{i+T} and construct a local predictor using this group of future states. A local predictor can be constructed in several ways. Among them, the averaging technique is the most popular way (e.g. Liu *et al.*, 1998; Sivakumar *et al.*, 1999; Jayawardena & Gurung, 2000). Here, the estimate of a future state $\hat{X}_{t+T}$ is calculated as

$$\hat{X}_{t+T} = \left(\sum X_{i+T}\right)/k \tag{13}$$

Data

Two real daily river flow time series are used in this study: (1) the Mississippi river at Vicksberg (1975–1993) and (2) the Wabash River at Mt. Carmel (1960–1978). A fairly long set of data, approximately 6900 data points from each time series, is used in the analysis. The data were downloaded from the US Geological Survey website. The two rivers were selected as they represent a large range of flow rates. The flow rates of the Mississippi river is quite large (mean at around 18,500 m^3/s) while the Wabash river is more moderate (mean at around 750 m^3/s). The Mississippi and Wabash river flow data used in the study are shown in Figure 9. Each data set is divided into three parts: a state space reconstruction set (SSR), a calibration set (CAL) and a production set (PROD). The SSR set and the CAL set are incorporated in calibrating (1) the optimal phase space parameters with an inverse approach and (2) the optimal SCM parameters. The PROD set is used as the verification set. Out of 19 years of total data in each series, the first 15 years (approximately 5480 data points) are used for the state space reconstruction set (SSR), the next two years (approximately 730 data points) as the calibration set (CAL) and the last two years (approximately 730 data points) the production set (PROD).

Methodology

First the phase space parameters are determined using an inverse approach so that the prediction error is lowest. The phase space reconstructed using these parameters is then used for data reduction. The data reduction procedure for the chaotic river flow time series is the same as presented in the above section (for the Bangladesh data) with the exception of the prediction tool used in the calibration of optimum SCM parameters. A nonlinear prediction method with local approximation is used as the prediction tool in chaotic river flow time series analysis. In this paper, it will be referred to, from here on, as the nonlinear prediction method (NLP).

Calibration

First, the phase space is reconstructed with the optimal phase space parameters obtained using the inverse approach. Then SCM parameters suggested by mGA are used with SCM to extract a reduced set of phase space vectors from the phase space reconstructed. The reduced phase space vectors are then used for prediction of the calibration set (CAL) using the nonlinear prediction method (NLP). The Normalized Root Mean Square Error (NRMSE) is used as the error indicator. NRMSE is expressed as

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}} \tag{14}$$

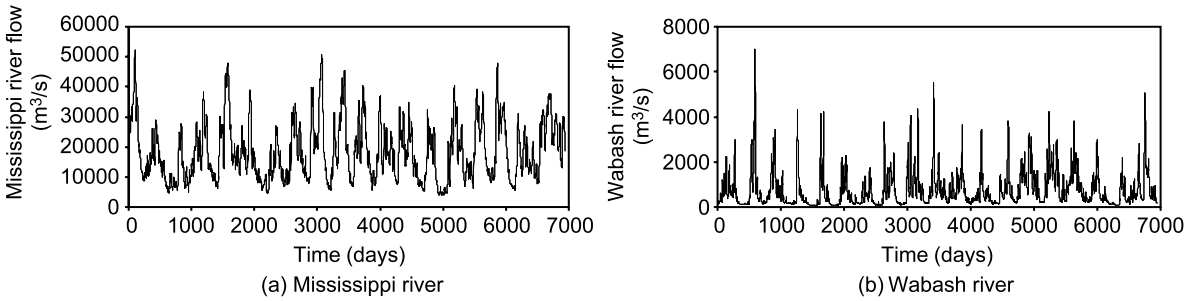

Figure 9 | Daily river flow time series.

A zero value of NRMSE indicates a perfect prediction, while a value close to 1 indicates that the predictions are no better than using the mean value of the time series ($\bar{x}$). The generation of SCM parameters and evaluation of the prediction error on the CAL set is repeated until a predetermined stopping criterion is reached. The optimal SCM parameters are selected based on the least prediction error resulting from the calibration set. The calibration procedure is illustrated in the schematic diagram given in Figure 10(a).

Validation

After obtaining the optimal SCM parameters, they are applied to the data set comprising the SSR and CAL sets. The reduced number of phase space vectors obtained from the SSR and CAL data sets are then applied to predict the PROD set. The prediction error *NRMSE* on the production set is then calculated. The validation procedure is shown in Figure 10(b).

The parameters of mGA used in this study are: a population size of 10, 100 generations, uniform crossover and binary tournament selection. The number of bits representing each variable varies according to their sensitivity to data reduction. Thus, it is decided to use 10 bits each to represent the highly sensitive *RR* and R_a, and 5 bits each for the less sensitive *AR* and *SF*. Although the present problem has two objectives: (1) a small number of patterns and (2) high prediction accuracy, only the prediction error, measured using *NRMSE*, is used as the fitness indicator in mGA. The ranges of SCM parameters used are the same as before: $0.001 \le R_a \le 1.0$; $1.0 \le SF \le 2.0$; $0 \le AR \le 1.0$; $0 \le RR \le 1.0$.

When the nonlinear prediction method is used for prediction using the reduced phase space vectors, it is necessary that a modified value of nearest neighbors, k', is chosen. k' is defined as

$$k' = \frac{k}{\text{total number of patterns used to determine } k} \times \text{reduced number of patterns} \quad (15)$$

where k is the optimal number of nearest neighbours found using the inverse approach. Thus the number of patterns used to determine k is the total number of patterns in the SSR set. The analysis on chaotic river flow time series is performed for three different prediction horizons (lead-times): 1, 3 and 5 days. The optimal phase space parameters and the prediction errors on CAL and PROD sets using the total number of patterns are shown in Table 5. The naïve prediction performance is also shown.

Results

The mGA solutions, from the calibration set, with prediction errors less than 120% of that resulting from the total data sets, have been selected as optimal solutions. These optimal solutions (optimal SCM parameters) are then

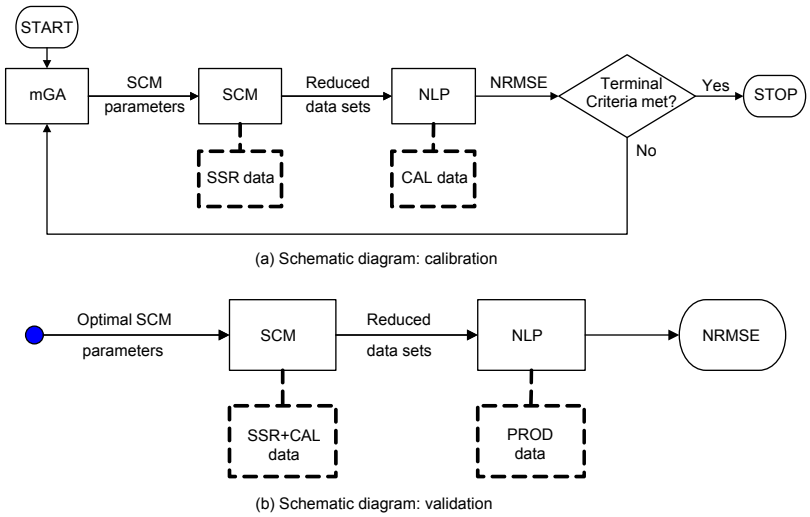

Figure 10 | Schematic diagrams for calibration and validation.

Table 5 | Optimal phase space parameter sets and prediction errors at various lead times with NLP and naïve forecasting techniques

		Mississippi river		Wabash river			
		Lead 1	Lead 3	Lead 5	Lead 1	Lead 3	Lead 5
Parameter set (m, τ, k)		(2, 1, 5)	(2, 1, 9)	(2, 1, 8)	(2, 6, 16)	(3, 1, 21)	(2, 1, 25)
NRMSE on	NLP	0.0354	0.1102	0.1947	0.1252	0.2935	0.4505
Calibration set	Naive	0.0599	0.1682	0.2656	0.1384	0.3755	0.5552
NRMSE on	NLP	0.0454	0.1437	0.2644	0.1163	0.2883	0.4390
Production set	Naive	0.0771	0.2162	0.3392	0.1218	0.3316	0.5038

incorporated to predict the production sets using the nonlinear prediction method. The prediction error of the reduced data sets is expressed as a percentage of the prediction error resulting from the total number of patterns (Equation (16)). Similarly, the reduced number of patterns is expressed as a percentage of the total number of patterns used (Equation (17)):

$$NRMSE(\%) = \frac{NMRSE_{\text{reduced number of patterns}}}{NMRSE_{\text{total number of patterns}}} \times 100 \qquad (16)$$

$$\text{Number of patterns } (\%) = \frac{\text{reduced number of patterns}}{\text{total number of patterns}} \times 100. \qquad (17)$$

For the Mississippi and Wabash river flow series, the performance of reduced data sets at lead-times 1, 3 and 5 days prediction are shown in Figures 11 and 12, respectively. Prediction accuracies of the reduced data sets

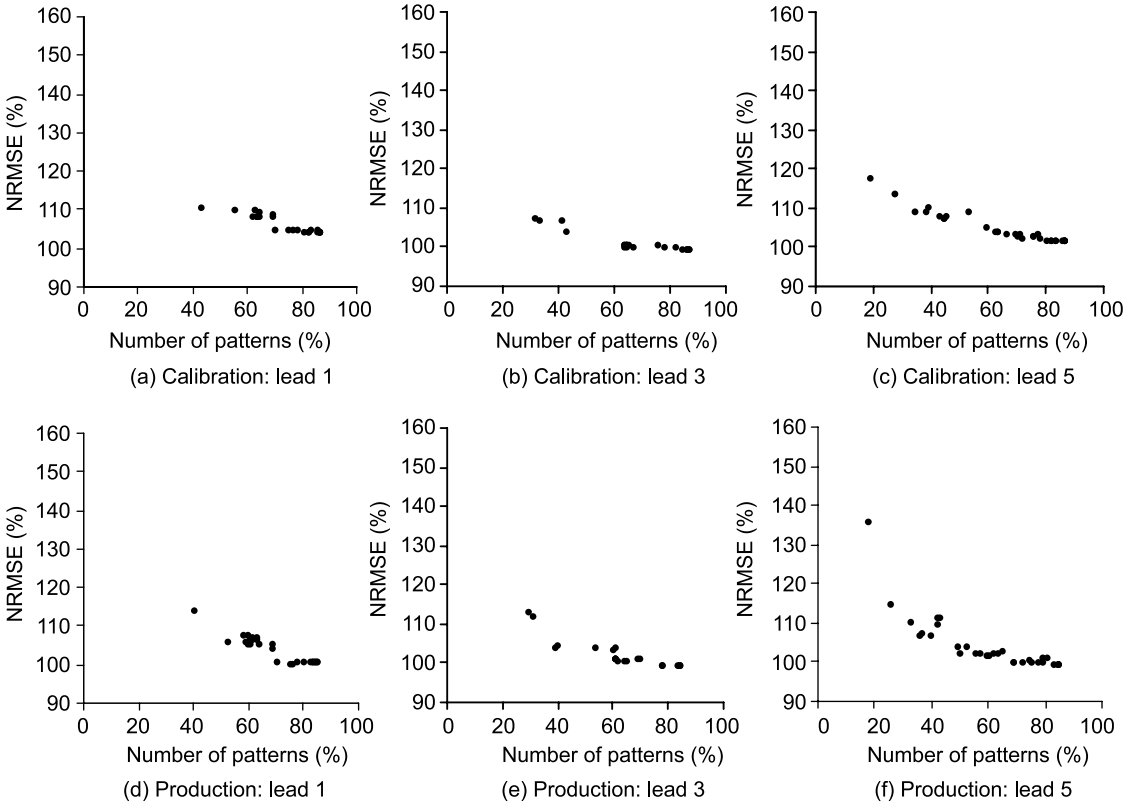

Figure 11 | Prediction errors at different lead times with various reduced numbers of patterns: Mississippi river.

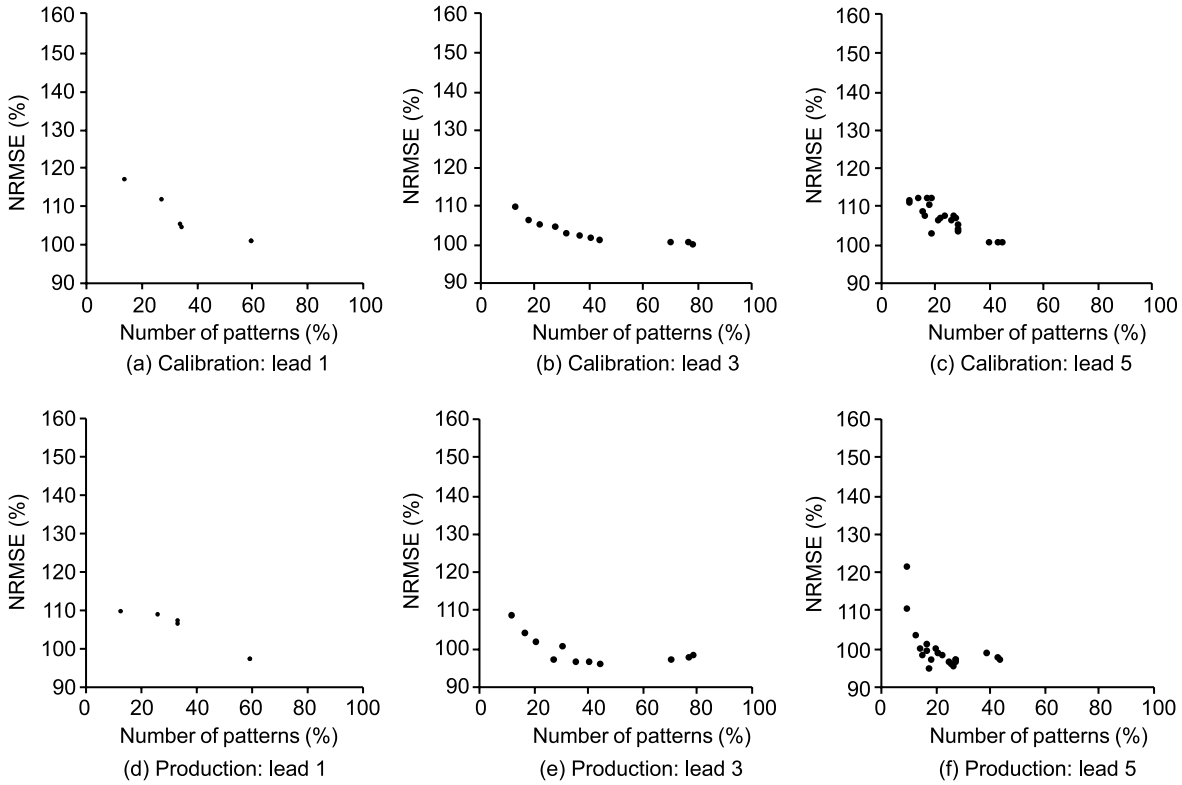

Figure 12 | Prediction errors at different lead times with various reduced numbers of patterns: Wabash river.

decrease with the decreasing number of selected patterns. For both river flow time series, the total data sets can be reduced up to about 40% almost without any effect on the prediction accuracy. A further 20% reduction in the data set will affect only about 10% loss in prediction accuracy. The SCM parameters, which have been selected considering the prediction error on calibration sets, have shown equally low prediction errors on production sets. Furthermore the percentage reduction in patterns gained on SSR and CAL data sets in the validation stage is nearly equal to the reduction in patterns in the calibration stage. These show that the proposed data extraction procedure produces robust SCM parameters.

CONCLUSIONS

A new scheme to derive an effective and efficient data set, using SCM and mGA, has been proposed. A data set containing only the "most representative" patterns of the original data set is extracted. Several advantages are identified. The storage capacity required for the training data set and the time required to train forecasting models are greatly reduced. Furthermore, models trained with the compact data set have been shown to yield equally high prediction accuracy as, if not higher than, models trained with the whole original data set.

This study linked the compact data set with several soft-computing techniques: Neural Network, Fuzzy Logic and phase space prediction of chaotic time series. With Neural Network, the model trained with a very small data set (only 3%) produces an even higher (although only slightly higher) prediction accuracy than that trained with the whole data set. A Fuzzy Inference System trained with the compact data set yields equally high prediction accuracy but a smaller number of rules than its counterpart trained with the entire original data set. In phase space prediction, which is known to require long data sets, the river flow time series reduction of up to 40% produces a prediction accuracy as good as that obtained from the whole original data set.

ACKNOWLEDGEMENTS

CDD and DSKK wish to thank the National University of Singapore for granting the research scholarship. Appreciation is also extended to DHI – Water and Environment, Denmark, for providing the water level (Bangladesh) data. Mississippi and Wabash river data were downloaded from http://water.usgs.gov/pubs/wri/wri934076/.

REFERENCES

ASCE Task Committee on Definition of Criteria for Evaluation of Watershed Models of the Watershed Management Committee, Irrigation and Drainage Division, 1993 Criteria for evaluation of watershed models. *J. Irrig. Drainage Engng.* **199** (3).

Babovic, V., Keijzer, M. & Stefansson, M. 2000 Optimal embedding using evolutionary algorithms. In *Proc. 4th International Conference on Hydroinformatics, Iowa City, IA, July*. Iowa University Press. Iowa, CD-ROM Proceedings.

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

Chiu, S. L. 1994 Fuzzy model identification based on cluster estimation. *J. Intell. Fuzzy Syst.* **2**, 267–278.

Farmer, J. D. & Sidorowich, J. J. 1987 Predicting chaotic time series. *Phys. Rev. Lett.* **59** (8), 845–848.

Goldberg, D. E. 1989 *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley. Reading, MA.

Holland, J. H. 1975 *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. University of Michigan Press. Ann Arbor.

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**, 242–264.

Krishnakumar, K. 1989 Micro-genetic algorithms for stationary and non-stationary function optimization. In *SPIE Proc. on Intelligent Control and Adaptive Systems*, SPIE, Philadelphia, PA **1196**, pp. 289–296.

Liong, S. Y. & Doan, C. D. 2002 Derivation of effective and efficient data set for training forecasting model. *Proc. 13th Congress of the Asia and Pacific Division (APD) of the International Association for Hydraulic Engineering and Research (IAHR), 6–8 August, Singapore*. World Scientific, **2**, 681–686

Liong, S. Y., Lim, W. H. & Paudyal, G. N. 2000 River stage forecasting in Bangladesh: neural network approach. *J. Comput. Civil Engng* **14** (1), 1–7.

Liu, Q., Islam, S., Rodriguez-Iturbe, I. & Le, Y. 1998 Phase-space analysis of daily streamflow: characterization and prediction. *Adv. Wat. Resources* **21**, 463–475.

Mamdani, E. H. & Assilian, S. 1975 An experiment in linguistic synthesis with a fuzzy logic controller. *Int. J. Man-Machine Stud.* **7**, 1–13.

Mamdani, E. H. & Assilian, S. 1999 An experiment in linguistic synthesis with a fuzzy logic controller. *Int. J. Human-Comput. Stud.* **51**, 135–147.

Packard, N. H., Crutchfield, J. P., Farmer, J. D. & Shaw, R. S. 1980 Geometry from a time series. *Phys. Rev. Lett.* **45** (9), 712–716.

Phoon, K. K., Islam, M. N., Liaw, C. Y. & Liong, S. Y. 2002 Practical inverse approach for forecasting nonlinear hydrological time series. *J. Hydrol. Engng* **7** (2), 116–128.

Pohlheim, H. 2000 *The Genetic and Evolutionary Algorithm Toolbox* (GEATbx) for use with Matlab: http://www.systemtechnik.tu-ilmenau.de/ ~ pohlheim/GA_Toolbox

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

Sugeno, M. & Kang, G. T. 1988 Structure identification of fuzzy model. *Fuzzy Sets Syst.* **28**, 15–33.

Takens, F. 1981 Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence. Lecture Notes in Mathematics*, vol 898 (D. A. Rand & L. S. Young), (eds.), Springer-Verlag Berlin, pp. 366–381.

Wang, L. & Mendel, J. M. 1992 Generating fuzzy rules by learning from examples. *IEEE Trans. Systems, Man, Cybern.* **30** (6), 1414–1423.

Yager, R. R. & Filev, D. P. 1994 Generation of fuzzy rules by mountain clustering. *J. Intell. Fuzzy Syst.* **2**, 209–219.