Journal of Hydrology (2007) 333, 305–316

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jhydrol

Flood prediction using Time Series Data Mining

Chaitanya Damle, Ali Yalcin *

Department of Industrial and Management Systems Engineering, University of South Florida, Tampa, FL 33620, USA

Received 20 June 2005; accepted 1 September 2006

KEYWORDS
River flood forecasting;
Time Series Data Mining;
Chaotic systems;
Event prediction

Summary This paper describes a novel approach to river flood prediction using Time Series Data Mining which combines chaos theory and data mining to characterize and predict events in complex, nonperiodic and chaotic time series. Geophysical phenomena, including earthquakes, floods and rainfall, represent a class of nonlinear systems termed chaotic, in which the relationships between variables in a system are dynamic and disproportionate, however completely deterministic. Chaos theory provides a structured explanation for irregular behavior and anomalies in systems that are not inherently stochastic. While nonlinear approaches such as Artificial Neural Networks, Hidden Markov Models and Nonlinear Prediction are useful in forecasting of daily discharge values in a river, the focus of these approaches is on forecasting magnitudes of future discharge values rather than the prediction of floods. The described Time Series Data Mining methodology focuses on the prediction of events where floods constitute the events in a river daily discharge time series. The methodology is demonstrated using data collected at the St. Louis gauging station located on the Mississippi River in the USA. Results associated with the impact of earliness of prediction and the acceptable risk-level vs. prediction accuracy are presented.

Introduction

Geophysical phenomena such as earthquakes, floods and rainfall represent complex systems that are generally governed by a large number of variables. These systems are nonlinear and are subject to high levels of uncertainty and unpredictability. The irregularity of these systems is not a transient phenomenon, but an intrinsic property (Galka, 2000). Nonlinear dynamics or chaos theory addresses the set of systems that may be apparently stochastic but also display correlations that are deterministic, and are known as deterministic chaotic systems. Deterministic chaos provides a structured explanation for irregular behavior and anomalies in systems which do not seem to be inherently stochastic. Thus, chaotic systems are treated as near accurately predicable in the short-term and can be studied under the framework of nonlinear system dynamics (Kantz and Schrieber, 1997).

In recent years, numerous studies from fields of hydrodynamics and civil engineering (Finnerty et al., 1997; Knebl et al., 2005), statistics (Xiong et al., 2001) and data mining

* Corresponding author. Tel.: +1 813 974 5590.
E-mail address: ayalcin@eng.usf.edu (A. Yalcin).

0022-1694/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2006.09.001

(Ayewah, 2003; Boogard et al., 1998; Coulibaly et al., 2000; Deo and Thirumalaiah, 2000; Laio et al., 2003; Sivakumar et al., 2002) have contributed to research in river flow prediction. Among these approaches, linear and nonlinear time series models are called black-box models in which prediction is based on single or multiple inputs such as flood discharge, weighted flood discharge, precipitation intensity, elevations, stream length, and main channel slope. The black-box models try to establish relationships between model inputs and outputs.

Linear Gaussian time series models are inadequate in the analysis and prediction of complex geophysical phenomena (Sivakumar, 2004). Linear methods, such as ARIMA approach, are limited by the requirements of time series stationarity and invertibility and independence of residuals (Povinelli, 1999). The time series observed in geophysical phenomena do not meet these requirements.

Nonlinear time series approaches such as Hidden Markov Models (HMM) (Ayewah, 2003), Artificial Neural Networks (ANN) (Boogard et al., 2000; Coulibaly et al., 2000; Deo and Thirumalaiah, 2000) and Nonlinear Prediction (NLP) (Islam and Sivakumar, 2002; Porporato and Ridolfi, 1997, 2001; Sivakumar et al., 2002) have been applied to discharge forecasting. Along with discharge forecasting, NLP has also been applied to the problem of flood forecasting in Laio et al. (2003). The HMM, ANN and NLP methods are useful in forecasting future values of discharge over the prediction horizon and their accuracy is measured over all the forecasted values and not on the accuracy of predicting floods. The prediction of floods requires a technique that can predict events (floods) in particular. This is where the event based data mining approach of Time Series Data Mining (TSDM) is useful because it focuses on the prediction of floods, rather than on forecasting future discharge values. The prediction accuracy can then be measured by how successful the approach is in predicting floods.

The rest of the paper is organized as follows: the following section describes the TSDM methodology and third section describes the application of the TSDM to the flood prediction problem. Fourth section presents the results and discussions regarding the prediction accuracy associated with the TSDM approach, and finally conclusions are presented in the last section.

Time Series Data Mining methodology

The TSDM methodology (Povinelli and Feng, 2003) follows the time delayed embedding procedure to predict future occurrences of events. TSDM combines the methods of phase space reconstruction and data mining to reveal hidden patterns predictive of future events in nonlinear and nonstationary time series. Fig. 1 outlines the steps involved in the TSDM methodology discussed in the following subsections.

Phase space reconstruction

Phase space reconstruction is a technique that provides a simplified, multi-dimensional representation of often a single-dimensional nonlinear time series. *Attractors* are the states towards which a system evolves when starting from certain initial conditions. Since the dynamic of the chaotic system is unknown, the original theoretical attractor that gives rise to the observed time series cannot be constructed. Instead, a phase space is created where the attractor is reconstructed from the scalar observed data that preserves the invariant characteristics of the original

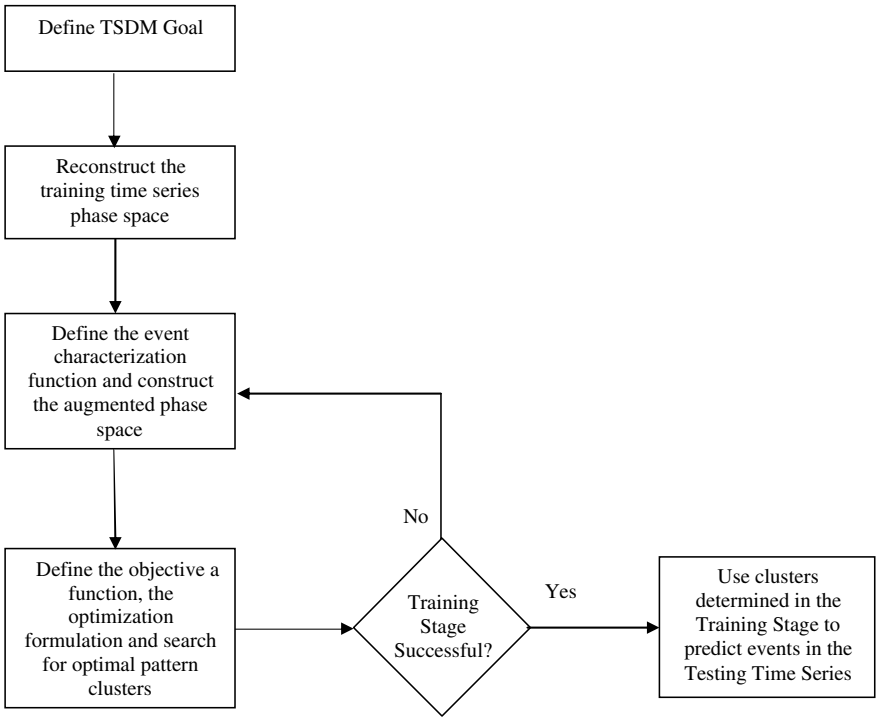

Figure 1 Flowchart of Time Series Data Mining methodology.

unknown attractor described by the time delay method to approximate the state space from a single time series.

The *reconstructed phase space* is a Q-dimensional metric space into which a time series is embedded. It is a vector space for the system such that specifying a point in this space specifies the state of the system at any given moment and vice versa. Time delayed embedding maps a set of Q observations taken from time series X onto $\boldsymbol{x}_t$, where $\boldsymbol{x}_t$ is a vector or point in the phase space. The time series is represented by $\{x_{t-(Q-1)\tau}, \ldots, x_{t-2\tau}, x_{t-\tau}, x_t\}$ where x_t represents the current observation, and $(x_{t-(Q-1)\tau}, \ldots, \boldsymbol{x}_{t-2\tau}, x_{t-\tau})$ are the past observations. If t is the current time index, then $t-\tau$ is a time index in the past, and $t+\tau$ is a time index in the future. The embedding delay (τ) is the time difference in number of time units between adjacent components of delay vectors and the embedding dimension (Q) is the number of dimensions of reconstructed phase space. Any further analysis of deterministic properties of a nonlinear time series depends on the precondition of a successful reconstruction of a state space of the underlying process (Galka, 2000).

Takens' Theorem (Takens, 1981) states that, when a system has a state space M that is Q dimensional, $\varphi : M \to M$ be a map that describes the dynamics of the system, and $y : M \to R$ its twice-differentiable function, which represents the observation of a single state variable; then the map $\Phi_{(\phi,y)} : M \to R^{2Q+1}$ defined by

$$\Phi_{(\varphi,y)}(\boldsymbol{x}) = (y(\boldsymbol{x}), y(\varphi(\boldsymbol{x})), \ldots, y(\varphi^{2Q}(\boldsymbol{x}))) \tag{1}$$

is an embedding that retains the structure of the original attractor from one topological space to another. Takens' Theorem guarantees that the reconstructed dynamics are topologically identical to the true dynamics of the system and provides the theoretical justification for the reconstruction of phase space. According to Takens' Embedding Theorem, the selection of any value for the delay (τ) will result in an embedding, given the fact that the data is infinitesimally accurate and does not contain any noise (Abarbanel, 1996; Galka, 2000). The data collected from naturally occurring dynamical systems hardly matches these specifications.

In regards to the embedding dimension (Q), an accurate value of Q is required when we want to exploit determinism with minimal computational effort. If a Q-dimensional embedding yields a faithful representation of the state space, every Q'-dimensional reconstruction with $Q' > Q$ does so as well. Selecting a large value of Q for chaotic data adds redundancy, degrades the performance of prediction algorithms and increases the computation time (Galka, 2000).

Many algorithms have been proposed that provide an estimation of optimal embedding dimension and time delay. Some of the methods for estimation of embedding dimension are the method of False Nearest Neighbors (Kantz and Schrieber, 1997), Fillfactor algorithm (Buzug et al., 1990; Buzug and Pfister, 1992), and Integral Local Deformation algorithm (Buzug et al., 1990; Buzug and Pfister, 1992). Some of the methods for estimation of time delay are Average Mutual Information (AMI) function (Fraser and Swinney, 1986), and Lyapunov exponents (Abarbanel, 1996). Selection of method for estimation of Q and τ is largely dependant on the application and the kind of analysis that is to be performed. The accurate calculation of time delay and number of embedding dimensions is not a requirement for TSDM. TSDM does not try to exploit the predictive structure of the reconstructed phase space. The purpose of a phase space reconstruction is to provide a simplified representation of the nonlinear time series. The identification of predictive patterns is accomplished by the data mining part of the TSDM methodology. However, selecting optimum time delay and embedding dimension would assist in reconstructing a more simplified phase space that closely reflects the true dynamics of the system. A simplified reconstruction should reduce the amount of computation to determine the optimal temporal pattern clusters as the number of dimensions to be searched is reduced.

Event characterization function

After an embedding is achieved, the next step is to define a function that helps identify the events of interest and projects them in the phase space to identify the temporal patterns and temporal pattern clusters. The event characterization function is an application dependent function where the value that represents the equation at time t correlates to the value of that event in the future. These types of event characterization functions are classified as *causal* and are useful in prediction type problems. For example, an event characterization function $g(t) = x_{t+1}$ is useful in predicting an event one-step before it actually occurs. Another example is

$$g(t) = \frac{x_{t+1} - x_t}{x_t} \tag{2}$$

where $g(t)$ represents the percentage change of values from time t to time $t+1$. The $g(t)$ value, referred to as *eventness*, is calculated for each phase space point. The choice of event characterization function has a significant effect on the prediction results.

Once the appropriate event characterization function is selected, an *augmented phase space* is generated. The augmented phase space is a $Q+1$ dimensional phase space that is created by adding one more dimension to the existing phase space. The additional dimension is represented by the event characterization function $g(t)$ and every phase space point is a vector defined by Eq. (3).

$$\langle \boldsymbol{x}_t, g(t) \rangle \in R^{Q+1} \tag{3}$$

Objective function

The objective function is used to determine which temporal pattern cluster is efficient in its ability to characterize events and is consistent with the TSDM goal. The objective function calculates a value for temporal pattern cluster P, which provides an ordering of the temporal pattern clusters according to their ability to characterize events. The number of temporal pattern clusters could be one or more. Before formulating the objective function, some basic definitions about *average eventness* and *variances* are necessary.

The index set Λ is a set of all time indices t of phase space points described by

$$\Lambda = \{t : t = (Q-1)\tau + 1, \ldots, N\} \quad (4)$$

where $(Q-1)\tau$ is the largest embedding time-delay, and N is the number of observations in the time series. The time index set M is the set of all time indices t when x_t is within the temporal pattern cluster (P), i.e. $M = \{t : \mathbf{x}_t \in P, t \in \Lambda\}$.

The average value of g, also called *average eventness*, of the phase space points within the temporal cluster P is given by

$$\mu_M = \frac{1}{c(M)} \sum_{t \in M} g(t) \quad (5)$$

where $c(M)$ is the *cardinality* of M. Cardinality of M represents the number of points located inside the temporal cluster. The average eventness of the phase space points not in P is

$$\mu_{\tilde{M}} = \frac{1}{c(\tilde{M})} \sum_{t \in \tilde{M}} g(t) \quad (6)$$

where $c(\tilde{M})$ is the number of points outside the temporal cluster. The average eventness of all the phase space points is given by

$$\mu_x = \frac{1}{c(\Lambda)} \sum_{t \in \Lambda} g(t) \quad (7)$$

The corresponding variances for respective average eventness functions are given in Eqs. (8)–(10).

$$\sigma_M^2 = \frac{1}{c(M)} \sum_{t \in M} (g(t) - \mu_M)^2 \quad (8)$$

$$\sigma_{\tilde{M}}^2 = \frac{1}{c(\tilde{M})} \sum_{t \in \tilde{M}} (g(t) - \mu_{\tilde{M}})^2 \quad (9)$$

$$\sigma_x^2 = \frac{1}{c(\Lambda)} \sum_{t \in \Lambda} (g(t) - \mu_x)^2 \quad (10)$$

From these definitions, many different objective functions can be created for the purpose of selecting one temporal pattern cluster over the other. One example is the t test for the difference between two means of points inside and outside the cluster as shown in the following equation:

$$f(P) = \frac{\mu_M - \mu_{\tilde{M}}}{\sqrt{\frac{\sigma_M^2}{c(M)} + \frac{\sigma_{\tilde{M}}^2}{c(\tilde{M})}}} \quad (11)$$

where P is a temporal pattern cluster. This function can be used for identifying a single statistically significant cluster that has high average eventness.

For identifying a single temporal pattern cluster where the purpose is to minimize the false predictions, an objective function of the type

$$f(P) = \frac{tp}{tp + fp} \quad (12)$$

is useful, where values of tp, tn, fp, and fn are described in Table 1. This type of objective function is only applicable to problems where the time series observations can be classified as tp, tn, fp, and fn.

When the accuracy of prediction is of primary importance, the efficacy of a collection of temporal pattern clusters in total prediction accuracy is given by

Table 1 Event categorization

	Actually an event	Actually a nonevent
Categorized as an event	True positive, *tp*	False positive, *fp*
Categorized as a nonevent	False negative, *fn*	True negative, *tn*

$$f(C) = \frac{tp + tn}{tp + tn + fp + fn} \quad (13)$$

where C is the collection of temporal pattern clusters.

An objective function of the type

$$f(P) = \begin{cases} \mu_M & \text{If } c(M)/c(\Lambda) \geqslant \beta \\ (\mu_M - g_0)\dfrac{c(M)}{\beta c(\Lambda)} + g_0 & \text{otherwise} \end{cases} \quad (14)$$

where β is the minimum proportion of points inside the cluster, and g_0 is the minimum eventness of the phase space which can be defined for the category of TSDM problems where the exact count of events and their magnitude are known,. This objective function orders temporal pattern clusters on the basis of time series observations with high eventness and characterizes at least a minimum number of events. The selection of objective function depends on the goal of TSDM.

Optimization formulation and search for optimal pattern clusters

Different temporal pattern clusters within the augmented phase space can contain the same set of phase space points. The optimization formulation is used to determine the optimal size of cluster by maximizing the objective function $f(P)$. Three types of biases, namely minimize, maximize or moderate, can be possibly placed on δ, the radius of the temporal pattern cluster *hypersphere*. Formulation by minimizing the δ subject to $f(P)$ remaining constant can be used to minimize the false positive prediction errors, i.e. the error of classifying a nonevent as an event. This ensures that the temporal pattern cluster has as small coverage as possible, keeping the value of objective function constant. The search for optimal temporal pattern cluster is performed using Genetic Algorithms.

Evaluation of training results and testing

If the optimal pattern clusters determined in the training stage are sufficiently accurate in predicting the events in the training time series, these clusters are used in the testing time series. Otherwise the training process is repeated for alternate event characterization functions, objective functions or optimization formulations.

In the testing stage, the testing time series is embedded in the reconstructed phase space using the same embedding parameters as the training time series. Whenever a point in the testing time series phase space falls inside the cluster identified in the training stage, an event is predicted. Testing results are evaluated by measuring the number of events correctly identified and predicted.

Application of TSDM to flood prediction

Study area and basic data set

The TSDM methodology is applied to flood forecasting at the St. Louis gauging station on the Mississippi River. The daily discharge time series is obtained from the United States Geological Survey website (http://www.usgs.gov), covering the period from April 1933 to September 2003, consisting of 25,750 data points. The daily discharge time series is shown in Fig. 2. Before the TSDM methodology is applied to flood forecasting, the presence of nonlinearity in the river discharge time series is confirmed using the Surrogate data method (Kantz and Schrieber, 1997). Surrogate data method is one among the large number of methods available for detecting nonlinearity in time series. It consists of computing a nonlinear statistic from the data being analyzed and from an ensemble of realizations of a linear stochastic process, which mimics linear properties of the data under study. If the computed statistic for the original data is significantly different from the values obtained for the surrogate data set, it can be inferred that the data was not generated by a linear process; otherwise the null hypothesis, that a linear model fully explains the data, cannot be rejected and the data can be analyzed, characterized and predicted using well-developed linear methods. Many issues need to be addressed before the surrogate data series is generated, making the method a very complex topic itself. More information regarding the formulation of hypotheses, preservation of the spectrum, autocorrelation in the generated data and the construction of Fourier Transform are found in Theiler et al. (1992) and Palus (1995). Using the surrogate data testing, the null hypothesis of a stationary, linear Gaussian random process is rejected at the 95% level of significance, since the prediction error of the river discharge data is found to be smaller than that of the surrogates.

The presence of nonlinearity at the St. Louis gauging station discharge time series has also been confirmed using the Correlation Dimension Method in Sivakumar and Jayawardena (2002). The Correlation Dimension Method is based on the fact that irregularity arising from deterministic dynamics will have limited degrees of freedom, which are equal to the least number of first order differential equations that capture most important features of the dynamics. Therefore, as phase spaces are reconstructed in higher embedding dimensions for an infinite data set, a point will be reached after which increasing the number of dimensions will not have any significant effect on the correlation dimension.

An important consideration in the accurate identification and prediction of flood events is the selection of thresholds (Sivakumar, 2005). Selection of a low threshold will result in a large number of false positives and conversely, a high threshold will cause events to be missed. Historical records indicate that floods have occurred during the years 1943, 1944, 1947, 1973, 1993 and 1995. The minimum value of discharge, during the reported flooding periods is 780,000 ft^3/s. This metric has been chosen as the threshold for identifying events, i.e. a value of discharge exceeding 780,000 ft^3/s is considered a flood and hence constitutes an event. The time series is split into two parts, the first 15,000 data points make up the training time series and, the rest is used in testing.

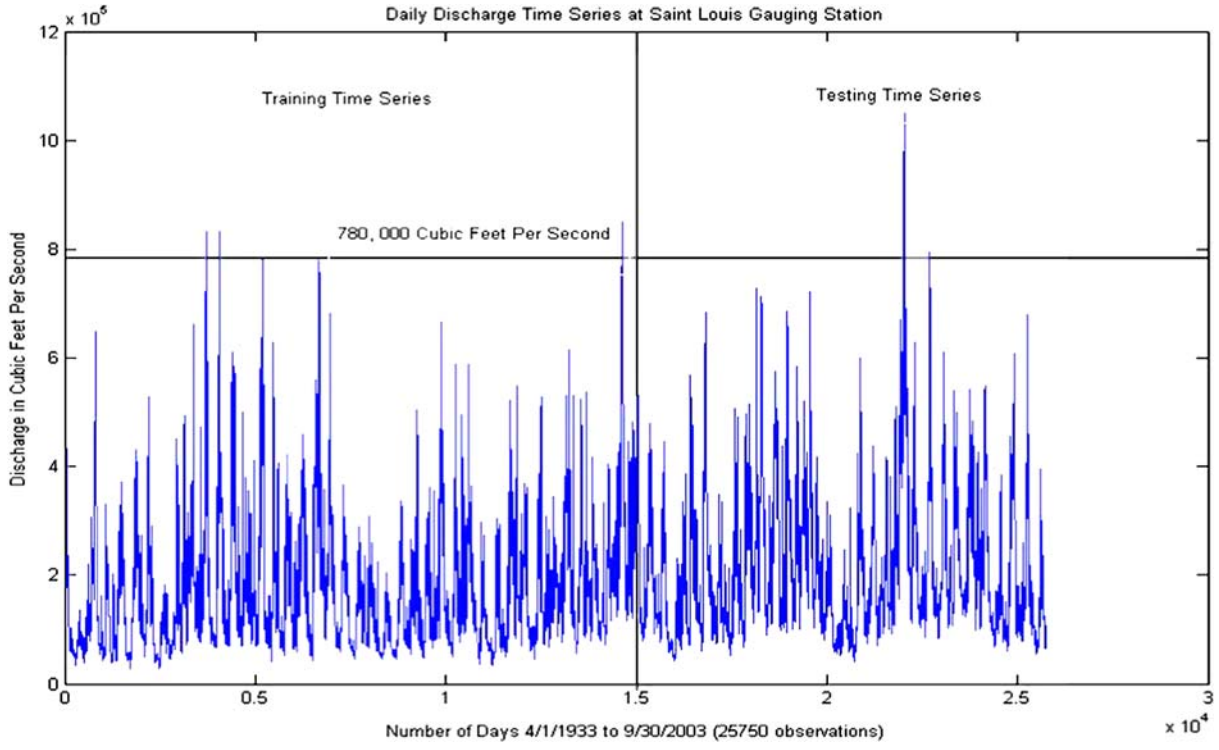

Figure 2 Daily discharge values at the St. Louis Gauging Station.

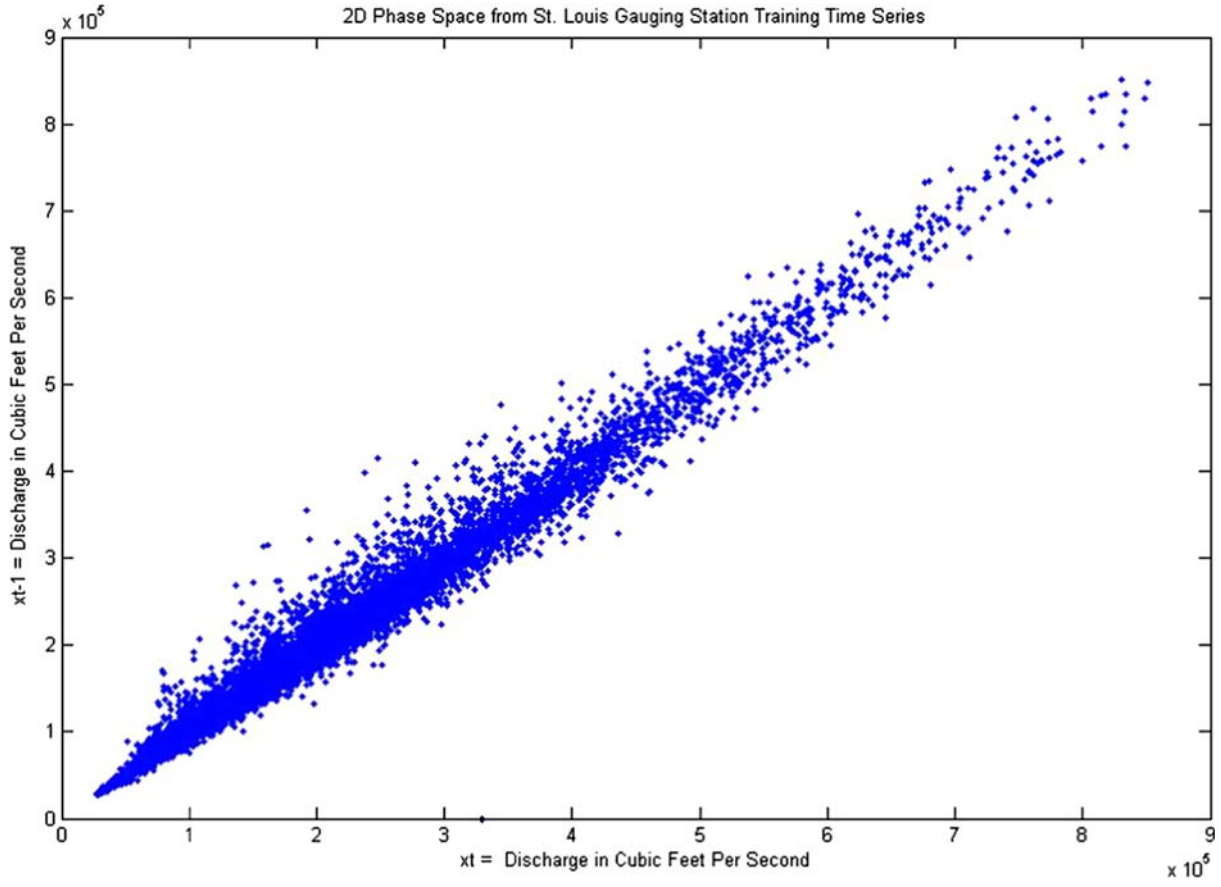

Figure 3 2D Embedding of training time series from St. Louis Gauging Station.

The time series is embedded in a two dimensional phase space with a delay of one. The reconstructed phase space for the training time series is shown in Fig. 3.

Event characterization function

The event characterization function $g(x_t) = x_{t+i}$, captures the goal of characterizing a flood i time steps in the future is selected. The reconstructed phase space is augmented by the $g(x_t)$ value for each corresponding phase space point as shown in Fig. 4. Since the event characterization function is a step-ahead function, the points that have the highest $g(x_t)$ value in the augmented phase space are the phase space points that have high discharge i steps ahead in the future depending on the step-ahead function used. The augmented phase space for $g(x_t) = x_{t+1}$ is shown in Fig. 4.

Objective function

The objective function is defined in a manner that incorporates two important aspects of the flood prediction problem as follows:

1. Based on the historical data, a minimum discharge value that results in a flood can be identified and each event can be classified as a *true positive (tp)* or a *false positive (fp)*. If an event identified by the cluster as a flood is actually a flood, then it is called a true positive. If an event identified by the cluster is not a flood, then it constitutes a false positive.
2. The demographic and economic impact of flood varies greatly based on its geographical location. If the risk associated with the demographic and economic impacts of the flood are incorporated into flood prediction models, the planner can utilize these models more effectively in their decision making process.

The objective function which incorporates these aspects is defined in Eq. (15) where g_i represents the eventness of the point i inside the cluster and β is a user defined parameter that corresponds to the proportion of false positives allowed in the cluster.

$$f(P) = \begin{cases} tp \times \sum_{i=1}^{M} g_i & \text{If } \dfrac{fp}{fp+tp} \leqslant \beta \quad 0 < \beta < 1 \\ -(\sum_{i=1}^{M} g_i) & \text{otherwise} \end{cases} \tag{15}$$

For every tp included in the cluster, the $f(P)$ function is rewarded by multiplying the summation of g_i values of points inside the cluster by the number of tp's inside the cluster increasing the value of the objective function. An optimization formulation that maximizes the objective value, would include all the points in the data set in the cluster, which in

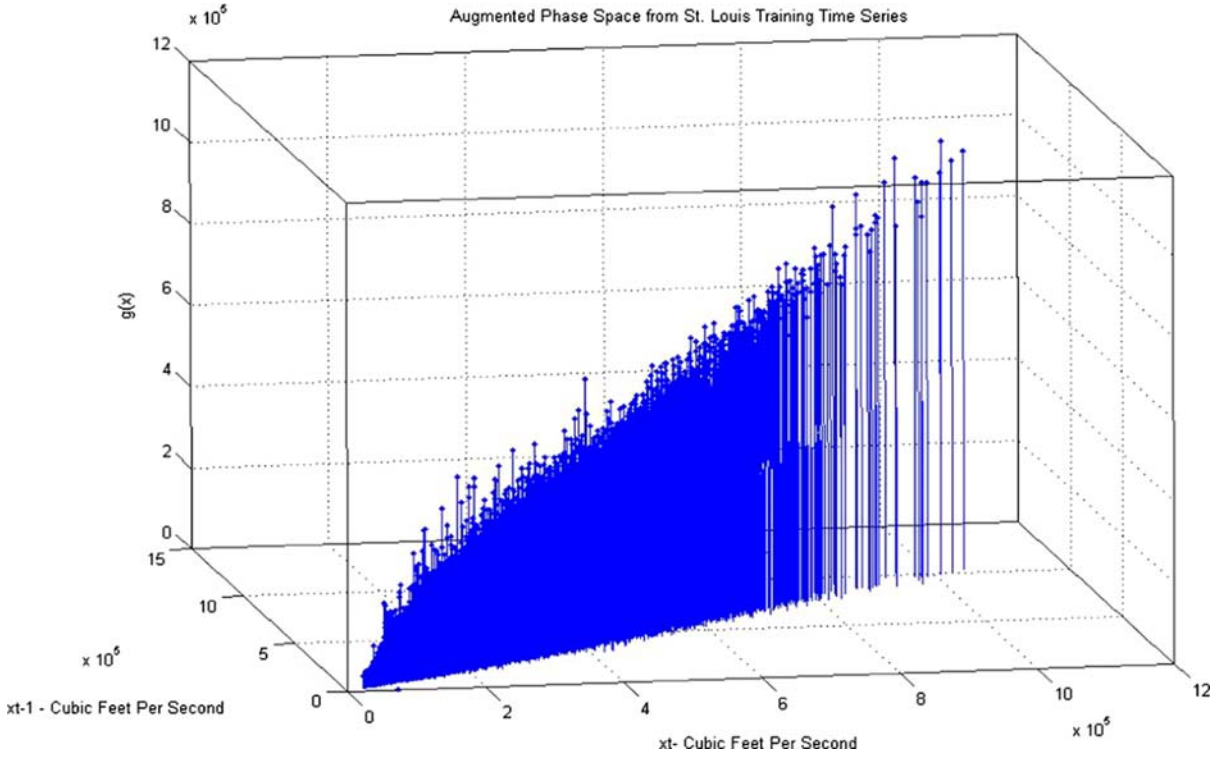

Figure 4 Augmented phase space for training series for St. Louis.

essence would predict every event as a flood. The user defined β value controls the number of points in the cluster based on the planners' choice of acceptable risk. In other words, a low value of β minimizes the number of points inside the cluster that are not actual floods, minimizing the number of false positives or false alarms, but at the same time increases the risk of missing an actual flood. On the other hand, a high value of β leads to a large number of false positives to be included in the cluster but decreases the risk of missing an actual flood, effectively incorporating the planners' desired level of risk. As a result, the maximization of the objective function results in a cluster that tries to include all true positives along with other phase space points with high g_i values, and the number of points in the cluster is restricted by β.

Optimization formulation

The optimization formulation is modeled as a multi-objective formulation with the two objectives as:

1. Maximize the value of objective function and
2. Minimize the radius of the cluster.

An unsupervised clustering technique, the Genetic Algorithm (Goldberg, 1989) is used in the search process for optimal temporal pattern cluster. The Genetic Algorithm (GA) searches for a global maximum, and identifies the optimal cluster. The priorities are assigned to the two objectives, with the maximization of objective function being the first priority and minimization of radius is the secondary priority. Thus, the GA searches for a temporal pattern cluster that maximizes the objective function value and if there are multiple clusters with equally high objective function value, it selects the cluster with minimum radius. The minimization objective is required to select the crispest cluster in order to minimize the number of false positives in the testing phase. The Genetic Algorithm Toolbox in Matlab, version 7.0.1 (Release 14) is used for the search and the output is the cluster center and its radius.

Evaluation of cluster prediction accuracy

To determine the prediction accuracy of the clusters and measure their ability to predict floods in the training and testing phases, the following set of performance parameters is measured:

1. *True positives (tp)*. If the event identified by the cluster as a flood is actually a flood it is called a true positive.
2. *False positive (fp)*. If an event identified by the cluster is not a flood it constitutes a false positive or a false alarm.
3. *Positive prediction accuracy (PPA)*. PPA is the percentage of true positives in the cluster. Since the events are classified either as true positives or false positives, the positive prediction accuracy of a cluster is calculated as $(tp/(tp + fp)) \times 100$. Note that PPA = $(1 - \beta_{\text{actual}}) \times 100$ where β_{actual} is the resultant β value based on the points in the cluster.
4. *Correct prediction percentage (CPP)*. CPP is the percentage of true positives predicted with respect to actual events and is calculated as (tp/# of actual events) $\times$ 100.
5. *Number of starts missed*. Since the goal is to predict occurrences of floods, it is important to predict the first instance when the discharge exceeds the flood threshold,

causing the river to overflow. Hence, along with CPP measured with respect to number of *tp*'s predicted, the number of starts of floods missed are also measured.

TSDM methodology results and discussion

This section presents the results of the application of the TSDM methodology to the river discharge data set from the St. Louis Gauging Station. The training stage results are summarized in Table 2.

The optimization process is able to identify clusters which include all 15 *tp*'s and have a 100% CPP. No cluster is identified for β value of 0.05. These results are indicative of appropriate event characterization and objective functions. The clusters for different values of β identified in the training stage are used to predict floods in the testing stage. These results are shown in Table 3.

Table 2 Training stage results at St. Louis Gauging Station

β	*tp*	*fp*	PPA	CPP
0.95	15	280	5.08	100
0.85	15	83	15.31	100
0.75	15	45	25.00	100
0.65	15	27	35.71	100
0.55	15	18	45.45	100
0.45	15	12	55.56	100
0.35	15	8	65.22	100
0.25	15	5	75.00	100
0.15	15	2	88.24	100
0.05	No cluster identified			

Number of actual events in the training time series = 15.

Table 3 Testing stage results for different β values at St. Louis Gauging Station

β	*tp*	*fp*	PPA	CPP	Number of starts missed
0.95	12	260	4.4	36.4	0
0.85	20	70	22.2	60.6	0
0.75	17	36	32.0	51.5	0
0.65	12	14	46.2	36.4	0
0.55	14	4	77.8	42.4	0
0.45	15	3	83.3	45.4	0
0.35	20	2	90.9	60.6	0
0.25	10	1	90.9	30.3	0
0.15	10	1	90.9	30.3	0

Number of floods in testing time series = 2.
Number of events in testing time series = 33.

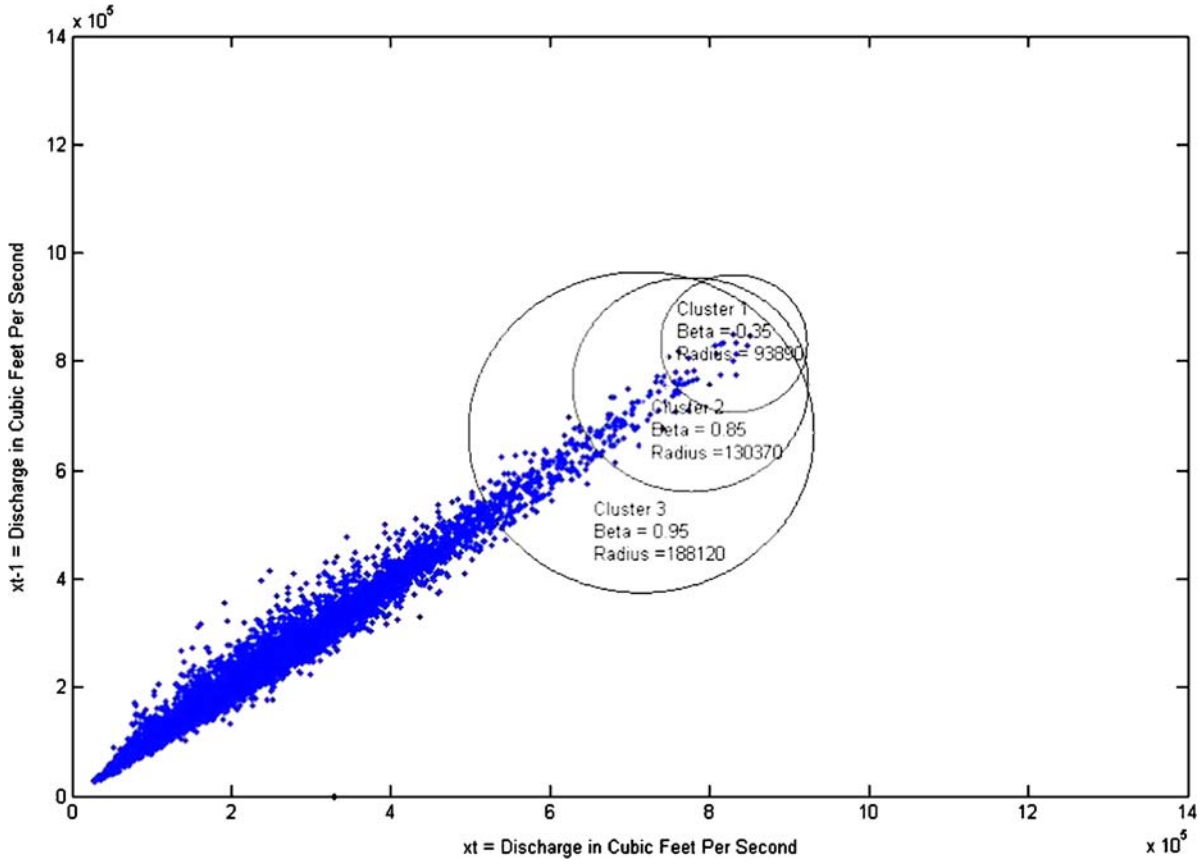

Figure 5 Clusters for different β values in training phase space.

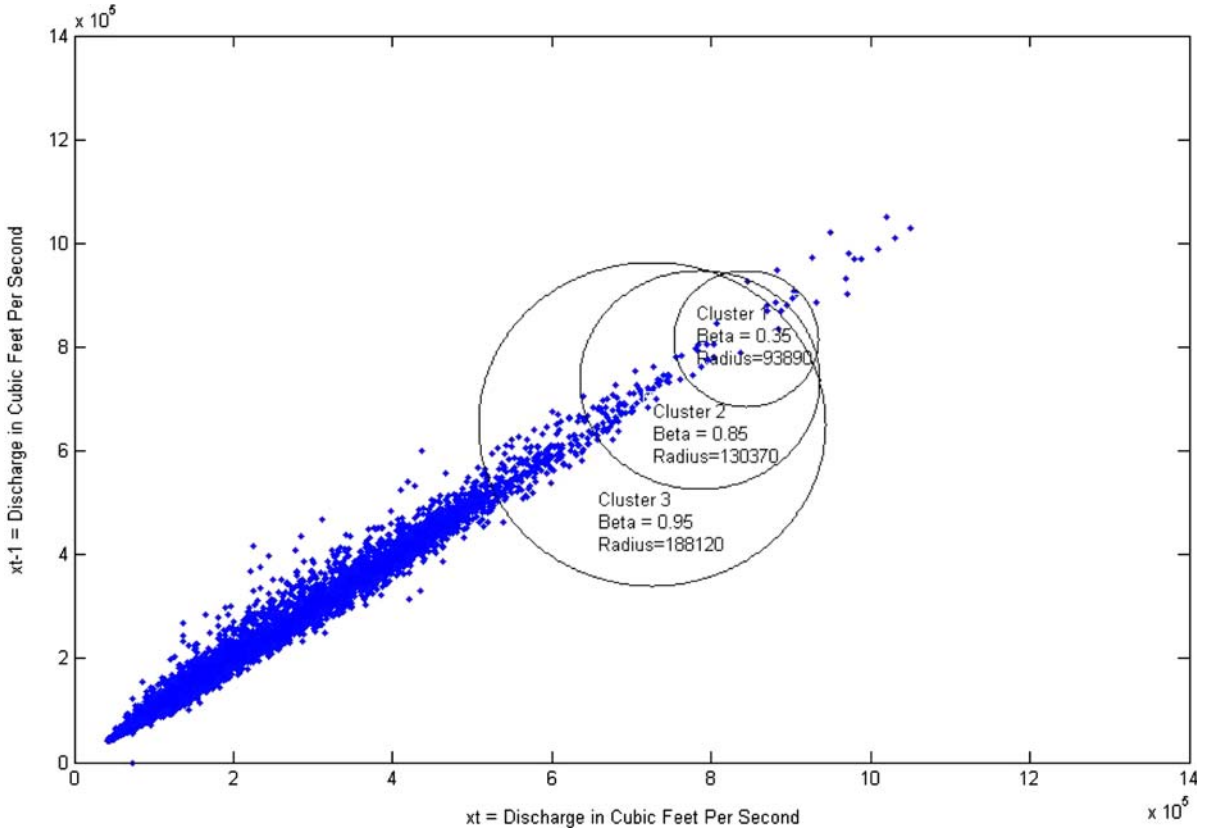

Figure 6 Clusters for different β values in testing phase space.

The testing stage results indicate variable success in terms of predicting all true positive events but capture all events corresponding to the start of the floods. The testing stage results are better interpreted by considering the affects of β on the optimal clusters characteristics in the training stage. Consider the three clusters determined in the training stage superimposed on the training data set phase space from the St. Louis Gauging Station as shown in Fig. 5. Specifying a small value for $\beta = 0.35$ identifies Cluster 1 as the optimal cluster. The location of the cluster is on the periphery of the point cloud where there are more true positives (actual floods) and few false positives (events with high discharge values but not floods) to accommodate the low β specified. On the other hand, specifying a large $\beta = 0.95$ leads to Cluster 3 located closer to the center of the point cloud since more false positives are allowed. Cluster 2 with $\beta = 0.85$ is located somewhere between the two other clusters. Note that all three clusters capture all true positives with CPP=100%. It is clear that β, by restricting the number and proportion of points in a cluster, affects the location of the cluster in the phase space. The clusters with high β values (lower risk of missing the start of floods) are located closer to the center of the point cloud. These clusters tend to miss the events corresponding to very high discharge values and contain a higher number of false positives. On the other hand, the clusters corresponding to low values of β are located in the periphery of the point cloud and contain those events with very high discharge values, but also have a higher probability of missing true flood events with low discharge values.

Fig. 6 depicts the same clusters superimposed on the testing stage data set. Note that while the point cloud

Table 4 Training stage earliness prediction accuracy for $\beta = 0.85$

Event characterization function	*tp*	*fp*	PPA	CPP
$\mathbf{x}_{t+1}$	15	83	15.3	100.0
$\mathbf{x}_{t+2}$	14	76	15.6	93.3
$\mathbf{x}_{t+3}$	14	67	17.3	93.3
$\mathbf{x}_{t+4}$	10	56	15.2	66.7
$\mathbf{x}_{t+5}$	7	36	16.3	46.7
$\mathbf{x}_{t+6}$	5	25	16.7	33.3
$\mathbf{x}_{t+7}$	3	10	23.1	20.0
$\mathbf{x}_{t+8}$	No cluster identified			

Number of actual events in the training time series = 15.

Table 5 Testing earliness prediction accuracy for $\beta = 0.85$ at St. Louis Gauging Station

Event characterization function	*tp*	*fp*	PPA	CPP	Number of starts missed
$\mathbf{x}_{t+1}$	20	70	22.2	60.6	0
$\mathbf{x}_{t+2}$	10	65	13.3	30.3	0
$\mathbf{x}_{t+3}$	10	63	13.7	30.3	0
$\mathbf{x}_{t+4}$	7	36	16.3	21.2	0
$\mathbf{x}_{t+5}$	4	32	11.1	12.1	0
$\mathbf{x}_{t+6}$	3	21	12.5	9.1	0
$\mathbf{x}_{t+7}$	1	15	6.3	3.0	1
$\mathbf{x}_{t+8}$	No cluster identified				

Number of actual events in the testing time series = 33.

exhibits a similar shape, the discharge values on the *y*-axis correspond to events with higher discharge values at the tip of the point cloud. These points correspond to the large flooding event in 1993 that corresponded to very high discharge values unlike the ones observed in the training series. In the testing stage the clusters with the highest CPP are Clusters 1 and 2 with $\beta = 0.35$ and $\beta = 0.85$ which capture 20 of the 33 true positives in the testing data set. Those points with the highest discharge values during the peak flooding period (therefore located in the periphery of the point cloud) are not captured by these clusters. Cluster 3 with $\beta = 0.95$ is located closer to the center of the point cloud and misses more of the true positives but still manages to capture events corresponding to the beginning of floods.

The lower CPP values associated with the testing stage draw attention to a key point in training data set selection. If the training set is not selected in a manner where the events of interest accurately represent the magnitude and the patterns leading to the events of interest to be pre-

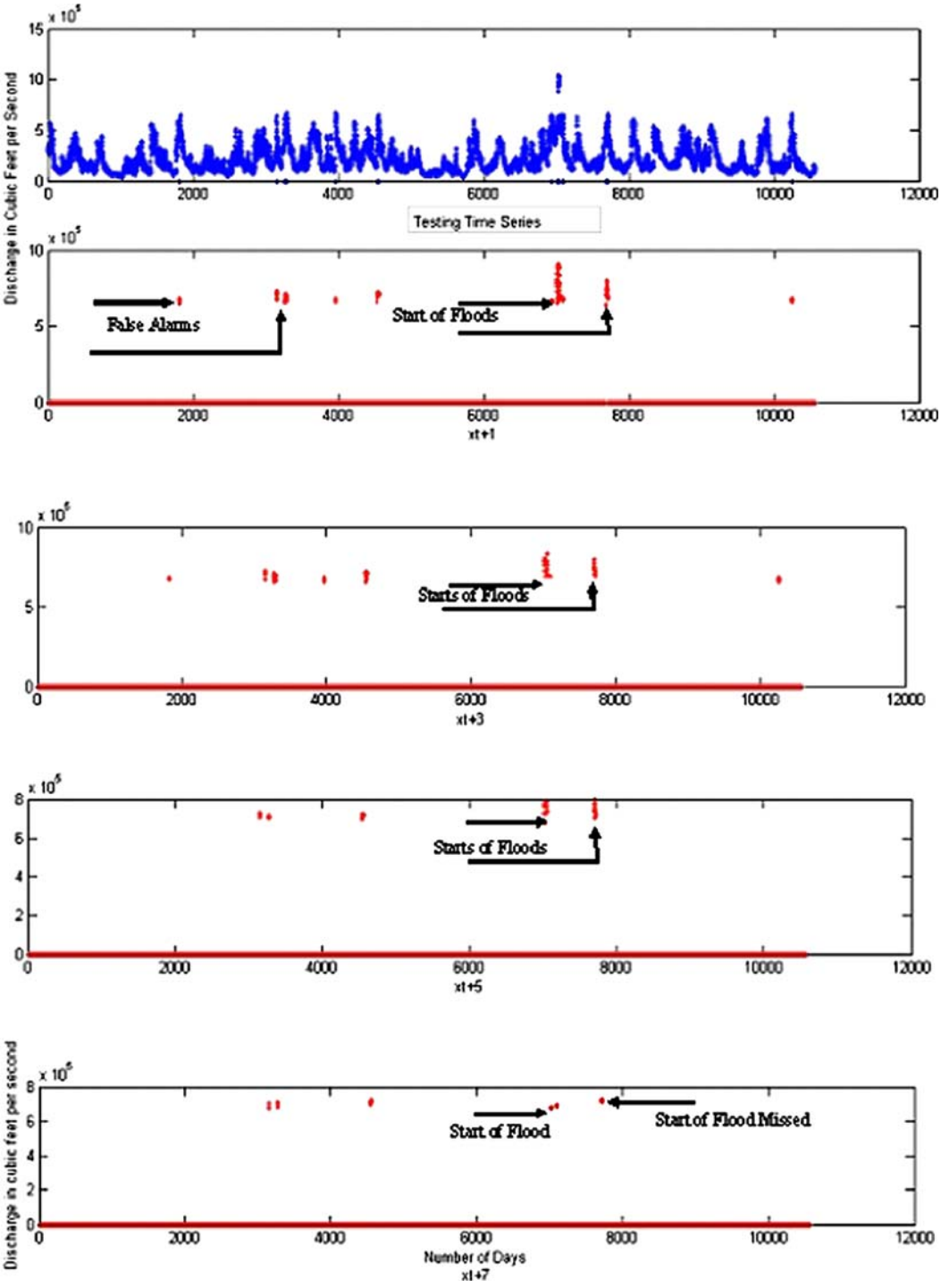

Figure 7 Earliness prediction results for $\beta = 0.85$ at St. Louis Gauging Station.

dicted, then the prediction accuracy of the proposed approach declines since the location and the size of the optimal clusters are determined by the data points corresponding to the events of interest in the training stage. On the other hand, even though the events with very high discharge values during a flood are outside of the optimal clusters, since this application considers a static threshold for the beginning of a flood, the optimal clusters are able to predict the beginning of floods.

Based on these observations, the testing stage results must be interpreted and used in a manner that exploits the advantages of several available clusters representing different risk levels. The testing time series of discharge values must be evaluated at several levels of β. Namely, if a time series results in points falling in a cluster corresponding to a high β value, the same time series must be tested at several lower β levels. If the same points are in the clusters determined for these lower β levels, then there is a higher chance that the time series is indicative of a flood.

Earliness prediction accuracy

Another performance measure of interest is the earliness prediction accuracy of TSDM using different step-ahead functions. The prediction accuracy is measured using different step-ahead event characterization functions for the same objective function with $\beta = 0.85$ and the same optimization formulation.

The training stage results for the St. Louis gauging station are presented in Table 4. The results indicate that the CPP values decrease as the prediction time horizon increases.

The clusters found in the training phase are then used to predict the floods in the testing phase. The results from the testing phase are displayed in Table 5. The same trend where the CPP decreases with the increase in prediction horizon is also observed in the testing stage results. For prediction horizon of seven days and more, the cluster misses the starts of floods.

Fig. 7 displays the events predicted in the testing time series using varying step-ahead functions. The first time series represents the testing time series and, the following time series display the predicted events for different step-ahead event characterization functions. It can be seen in Fig. 7 that, until seven-day-ahead prediction, beginning of the two floods in the testing data series is predicted. For a seven day prediction horizon, the cluster misses the start of one out of the two floods. These results are comparable to the results obtained in Sivakumar and Wallender (2005) where a nonlinear deterministic approach leads to good results up to five-day-ahead prediction for discharge values.

Conclusions

This paper presents the application of the TSDM methodology in the prediction of floods. The methodology is applied to the river discharge data at the St. Louis Gauging Station with the goal of predicting floods accurately and as early as possible.

The training stage results indicate that the selected event characterization and the objective functions are successful in predicting floods and the beginning of all floods are successfully predicted in the testing stage. An inverse relationship exists between the earliness of prediction and the Correct Prediction Percentage where, as the prediction horizon increases, the Correct Prediction Accuracy decreases. Ideally, one would like to predict a flood as early as possible, however associated with the earliness of prediction is the risk of missing the start of floods.

The described TSDM approach, through the user-defined β parameter in the objective function, incorporates the acceptable risk level of emergency operation plans dictated by the demographic and economic characteristics of the location under consideration. The affects of various levels of β on the location and the proportion of points in the clusters used for prediction are discussed and simultaneous use of multiple clusters is recommended to provide planners with the most useful information.

The successful application of the TSDM requires careful training time series selection to determine optimal clusters for event prediction. Associated with the system under consideration, the analyst must ascertain that the physical characteristics of the system at the time the training time series is observed must adequately resemble the conditions for prediction to ensure that the data indicative of the events of interest are comparable. An accurate and precise description of events of interest (selection of thresholds) is also necessary to ensure the highest possible accuracy of prediction by minimizing false positives and true negatives. Furthermore, the training time series must include sufficient variety in the data patterns that lead to events of interests. The variety in these data patterns determines the location and the size of the optimal clusters used in event prediction. The testing stage of the example considered in this paper is an example where the very high values of discharge associated with a large flood were outside of the optimal cluster and, while the beginnings of the floods were successfully predicted, the correct prediction percentages were lower than the ones in the training stage.

Future research directions based on the TSDM methodology introduced in this paper include selection of event characterization functions that incorporate multiple time series data for different factors (e.g. rainfall) that contribute to floods, investigating the use and accuracy of intersections of clusters for prediction accuracy, and development of software applications to integrate emergency planners' acceptable risk-levels, flood location demographics and the TSDM methodology.

References

Abarbanel, H.D.I., 1996. Analysis of Observed Chaotic Data. Institute for Nonlinear Science.

Ayewah, N., 2003. Prediction of spatial temporal events using a hidden markov model. Department of Computer Science and Engineering, Southern Methodist University, Dallas, TX.

van den Boogard, H.F.P., Gautam, D.K., Mynett, A.E., 1998. Autoregressive neural networks for the modeling of time series. In: Babovic, V., Larsen, C.L. (Eds.), Hydroinformatics98. Balkema, Rotterdam, pp. 41–748.

Buzug, T., Pfister, G., 1992. Comparison of algorithms calculating optimal parameters for delay time coordinates. Physica D 58, 127–137.

Buzug, T., Reamers, T., Pfister, G., 1990. Optimal reconstruction of strange attractors from purely geometrical arguments. Europhysics Letters 13, 605–610.

Coulibaly, P., Anctil, F., Bobée, B., 2000. Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. Journal of Hydrology 230, 244–257.

Deo, M.C., Thirumalaiah, K., 2000. Real time forecasting using neural networks. In: Govindaraju, R.S., Ramachandra Rao, A. (Eds.), Artificial Neural Networks in Hydrology. Kluwer Academic Publishers, Dordrecht, pp. 3–71.

Finnerty, B.D., Smith, M.B., Seo, D.J., Koren, V., Moglen, G.E., 1997. Space–time scale sensitivity of the Sacramento model to radar-gage precipitation inputs. Journal of Hydrology 203, 21–38.

Fraser, A.M., Swinney, H.L., 1986. Independent coordinates for strange attractors from mutual information. Physics Review A 33, 1134–1140.

Galka, A., 2000. Topics in Nonlinear Time Series Analysis – With Implications for EEG Analysis. World Scientific Publishing Company.

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

Internet Website URL: http://www.usgs.gov/index.html, United States Geological Survey, 2004.

Islam, M.N., Sivakumar, B., 2002. Characterization and prediction of runoff dynamics: a nonlinear dynamical view. Advances in Water Resources 25, 179–190.

Kantz, H., Schrieber, T., 1997. Nonlinear time series analysis. Nonlinear Science Series, No. 7, Cambridge.

Knebl, M.R., Yang, Z.L., Hutchison, K., Maidment, D.R., 2005. Regional scale flood modeling using NEXRAD rainfall, GIS, and HEC-HMS/RAS: a case study for the San Antonio River basin. Journal of Environmental Management 75, 325–336.

Laio, F., Porporato, A., Revelli, R., Ridolfi, L., 2003. A comparison of nonlinear forecasting methods. Water Resources Research 39 (5), 1129–1132.

Palus, M., 1995. Testing for nonlinearity using redundancies: quantitative and qualitative aspects. Physica D 80, 186–205.

Porporato, A., Ridolfi, L., 1997. Nonlinear analysis of river flow time sequences. Water Resources Research 33 (6), 1353–1367.

Porporato, A., Ridolfi, L., 2001. Multivariate nonlinear prediction of river flows. Journal of Hydrology 248, 109–122.

Povinelli, R.J., 1999. Time Series Data Mining: Identifying Temporal Patterns for Characterization and Prediction of Time Series Events. Ph.D. Dissertation, Marquette University, Milwaukee, WI.

Povinelli, R.J., Feng, X., 2003. A new temporal pattern identification method for characterization and prediction of complex time series events. IEEE Transactions on Knowledge and Data Engineering 15 (2), 339–352.

Sivakumar, B., 2004. Chaos theory in geophysics: past, present and future. Chaos, Solutions & Fractals 19, 441–462.

Sivakumar, B., 2005. Hydrologic modeling and forecasting: role of thresholds. Environmental Modeling and Software 20 (5), 515–519.

Sivakumar, B., Jayawardena, A.W., 2002. An investigation of the presence of low-dimensional chaotic behaviour in the sediment transport phenomenon. Hydrological Sciences Journal 47 (3), 405–416.

Sivakumar, B., Wallender, W.W., 2005. Predictability of river flow and suspended sediment transport in the Mississippi River basin: a non-linear deterministic approach. Earth Surface Processes and Landforms 30, 665–677.

Sivakumar, B., Jayawardena, A.W., Fernando, K.G., 2002. River flow forecasting: use of phase-space reconstruction and artificial neural networks approaches. Journal of Hydrology 265, 225–245.

Takens, F., 1981. Detecting strange attractors in fluid turbulence. In: Rand, D., Young, L.-S. (Eds.), Dynamical Systems and Turbulence. Springer, Berlin, pp. 366–381.

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Farmer, J.D., 1992. Testing for non linearity in time series: the method of surrogate data. Physica D 58, 77–94.

Xiong, L., Shamseldin, A.Y., O'Connor, K.M., 2001. A non-linear combination of the forecasts of rainfall-runoff models by the first-order Takagi-Sugeno fuzzy system. Journal of Hydrology 245, 196–217.