Flow categorization model for improving forecasting

C. Sivapragasam[1] and Shie-Yui Liong[2]

[1] Department of Civil Engineering, Mepco Schlenk Engineering College, Virudhunagar (District), Tamilnadu, 626 005, India. E-mail: *sivapragasam@yahoo.com*
[2]National University of Singapore, Singapore.

Received 23 May 2003; accepted in revised form 16 April 2004

Abstract Prediction of high magnitude flows is of interest in many hydrological applications such as operation of flood control reservoirs, flood forecasting and gated spillways. Of the various types of existing streamflow prediction approaches, data driven models (such as ANN) are increasingly being preferred over the traditional conceptual models due to their simplicity, fast speed and ease of use. For models that consider only historical streamflow data, an attempt has been made to design a robust model over a wide range of streamflow magnitudes. The model inputs are the immediate past streamflow data which generally do not predict the typically high flows well, particularly for large lead times.

In this study, the flow range is divided into three regions (low, medium and high flow regions) and the attributes are decided based on the underlying hydrological process of the flow region. A flow forecasting model is applied for each flow region, using only the historical streamflow data as input. The proposed approach is implemented in Tryggevælde Catchment (Denmark) for 1- and 3-lead days, using the Support Vector Machine (SVM), which yields promising results, particularly for high flows in a 3-lead day model.

Keywords Artificial neural network; flow forecasting; support vector machine

Nordic Hydrology **Vol 36 No 1 pp 37–48** © IWA Publishing 2005

Introduction

The importance of having a highly accurate predictive streamflow model for water resources systems management which includes reservoir operation has been widely recognized (Mishalani and Palmer 1988; Georgakakos 1989). Reservoirs typically serve multiple purposes and, since flood control is generally one of the primary objectives, advance prediction of the high inflows into reservoirs becomes particularly significant. The reservoir system operators/managers need to decide to what extent reservoir releases will increase so as to safely accommodate the inflow and also optimally fulfil other water uses. Typically, 3 or 5 days' advance forecast can significantly aid the operators/managers in arriving at optimal performance of real-time operation. However, in practice, perfect forecasts are difficult to obtain and the error associated with forecasting always increases with the advance (lead) time.

In the past three decades or so, many approaches have been proposed for flow forecasting, which can be classified under two major categories: conceptual models and data driven models. Conceptual models are not ideal for real-time forecasting owing to many practical difficulties such as the availability of data (Yapo *et al.*1996), the time and effort involved in model development and implementation and the need for expertise and experience with the model (Hsu *et al.* 1995). On the other hand, the system analysis or data driven (black box) approach is based on identifying a relationship between input and output from the historical observed data without attempting to describe any of the internal mechanisms. This approach relies heavily on techniques of system theory. The system analysis models can be categorized by system theory into two types: statistical-based methods and artificial intelligence. Traditionally, statistical method, e.g. regression and Box–Jenkins techniques,

have been the most widely used for modelling water resources time series as recognized by Maier and Dandy (1996). Recently, however, with the advent of fast computing machines many researchers have shown the efficacy of Artificial Intelligence (AI) techniques like Artificial Neural Networks (ANN), Fuzzy Logic and Genetic Programming in forecasting and modelling problems as compared to the statistical methods. The popularity of AI techniques can be attributed to the fact that these models tackle the complex problems more easily than sophisticated statistical models constrained by many strict rules governing model development and thus making them difficult for real-world applications. Many applications of ANNs in streamflow forecasting have been carried out, e.g. Karunanithi *et al.* (1994); Thirumalaih and Deo (1998) and Liong *et al.* (1999). These demonstrate its high prediction accuracy and produce relationships between input and output variables.

When ANNs are used to model the rainfall–runoff process, inputs such as precipitation, infiltration, temperature, snowmelt and historical streamflows have often been included. Although incorporating inputs such as precipitation, temperature and other variables may improve the prediction accuracy, in practice such information is often either not available or difficult to obtain. Of those models which consider only historical streamflow data (without involving precipitation or other inputs), the modelling has used robust prediction techniques over a wide range of streamflow. Further, the most suitable input parameters for the model are identified by incorporating the immediate past data (usually by trial and error) to reflect the temporal effect (Karunanithi *et al.* 1994; Hsu *et al.* 1995; Thirumalaiah and Deo 1998). Such models often perform poorly in predicting typically high flow events (e.g. Karunanithi 1994; Thirumalaiah and Deo 1998). One possible reason for this is the fact that mechanisms governing different flow events (such as high flows or low flows) are characteristically different, as recognized by the ASCE Task Committee (2000).

This study attempts to improve the prediction of high flow events. The inputs to the model are derived entirely from the available streamflow records. The methodology, referred to as the Flow Region Specific Forecasting Model (FRSFM) approach, is applied in two stages, viz. categorizing the streamflow into high, medium and low flows (Table 1) and developing models each specifically for a flow region and for a specific lead time of forecast as well. This approach also has the advantage of a faster learning process with limited data in each flow type and is thus suitable under real-time operations. Further, the proposed approach is implemented with a newly emerging machine learning algorithm, the Support Vector Machine (SVM).

The SVM is an approximate implementation of the method of *structural risk minimization*. This induction principle minimizes an upper bound on the error rate of a learning machine on test data (i.e. generalization error) rather than minimizing the training

Table 1 Single SVM model: 1-lead day and 3-lead day verification

		Goodness-of-fit measures			
Lead day (1)	**Flow class (2)**	**COD (3)**	**RMSE (m) (4)**	**CC (5)**	**MAE (6)**
1	Single SVM model	0.87	0.64	0.92	0.24
	High	0.59	1.11	0.77	0.78
	Low	0.56	0.23	0.78	0.07
	Medium	0.30	0.99	0.63	0.49
3	Single SVM model	0.50	1.27	0.75	0.53
	High	−0.02	1.98	0.28	1.39
	Low	0.09	0.40	0.84	0.20
	Medium	−0.03	1.72	0.51	1.07

error itself (used in empirical risk minimization as in ANN). This helps them to generalize well on the unseen data. Liong and Sivapragasam (2002) give an elaborate introduction to SVM and compare its advantage over ANNs.

This paper is organized as follows. In the next section, a brief introduction to the support vector machine is given. Then, a single model (fitted to the entire flow range) based on SVM is implemented for the Tryggevælde Catchment runoff data (Denmark) for 1- and 3-lead day forecasting. This is followed by the proposed FRSFM approach. The results are compared to the SVM-based single model. Finally a brief discussion concludes the paper.

Support vector machine

According to the Structural Risk Minimization (SRM) principle, the generalization ability of learning machines depends more on capacity concepts than merely the dimensionality of the space or the number of free parameters of the loss function (as espoused by the classical paradigm of generalization). Thus, for a given set of observations $(x_1,y_1),\ldots,(x_n,y_n)$, the SRM principle chooses the function f_β^* in the subset $\{f_\beta: \beta \in \Lambda\}$, for which the guaranteed risk bound, as given by Eq. (1) below, is minimal. In other words, the actual risk is controlled by the two terms given in Eq. (1):

$$R(\beta) \leq R_{emp}(\beta) + \Omega\left(\frac{n}{h}\right) \tag{1}$$

where the first term is an estimate of the risk and the second term is the confidence interval for this estimate. The parameter h is called the VC dimension (named after Vapnik and Chervonenkis) of a set of functions. It can be seen as the measure of the capability of a set of functions implementable by the learning machine to best approximate the problem.

SVM is an approximate implementation of the SRM principle. The final approximating function used in SVM for regression is of the form

$$f(x) = \sum_{i=1}^{l} \left(\alpha_i - \alpha_i^*\right) K(x_i, x) + b \tag{2}$$

where $K(x_i, x) = \langle \phi(x).\phi(x_i) \rangle$ is called the kernel function, which performs the inner product in feature space, $\phi(x)$. α_i and α_i^* are Lagrange multipliers. To act as a kernel, a function needs to satisfy Mercer's condition. The kernel representation offers a powerful alternative for using linear machines in hypothesizing complex real world problems as opposed to Artificial Neural Network based learning paradigms, which use multiple layers of threshold linear functions (Cristianini and Shawe-Taylor 2000).

The approximating function is designed to have the smallest ε deviation (given as Vapnik's ε-insensitive loss function) from measured targets, d_i, for all training data. Slack variables, ξ and ξ^*, are introduced to account for outliers in the training data. The algorithm computes the value of Lagrange multipliers, α_i and α_i^*, by minimizing the following objective function:

$$Minimize \quad \frac{1}{2}\|a\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right) \tag{3}$$

$$Subject\ to \quad d_i - (a.x_i + b) \leq \varepsilon + \xi_i \tag{3a}$$

$$(a.x_i + b) - d_i \leq \varepsilon + \xi_i^* \tag{3b}$$

$$\xi_i \xi_i^* \geq 0$$

Eqs (3), (3a) and (3b), expressed in the dual form, are given as

maximize

$$-\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i-\alpha_i^*)\left(\alpha_j-\alpha_j^*\right)\langle\phi_i,\phi_j\rangle-\varepsilon\sum_{i=1}^{N}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{N}y_i(\alpha_i-\alpha_i^*) \tag{4}$$

$$\begin{aligned} &\textit{subject to} \quad \sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0 \\ &0\le\alpha_i\le C, \quad i=1,2,\ldots...,N \\ &0\le\alpha_i^*\le C, \quad i=1,2,\ldots...,N \end{aligned} \tag{4a}$$

where C is a user specified constant and it determines the trade-off between the flatness of $f(x)$ and the amount of deviation that can be tolerated. The value 'a' refers to the weight factor for obtaining the flattest decision function. It should be noted that the training patterns, appearing in both objective functions of Eq. (4) and in the approximating function of Eq. (2), are in the form of dot products.

The solution of the above problem yields α_i and α_i* for all i = 1 to N. It can be shown that all the training patterns within the ε-insensitive zone yield α_i and α_i* as zeros. The remaining non-zero coefficients essentially define the final decision function. The training examples corresponding to these non-vanishing coefficients are called Support Vectors.

Optimal values of ε, C and the kernel-specific parameters are to be used for the final regression estimation. Currently, identification of optimal values for these parameters is mainly conducted on a trial and error process. As well as the ε-insensitive loss function, a quadratic loss function (ε = 0) may also be used. In this study, the quadratic loss function is preferred over the ε-insensitive loss function as the former is less computer memory intensive. Details on SVM can be found in Vapnik (1995, 1999), Drucker *et al.* (1996), Smola and Scholkopf (1998), Haykin (1999) and Cristianini and Shawe-Taylor (2000).

Tryggevælde Catchment

In this study, the Tryggevælde Catchment in Denmark is used for runoff forecasting. The Tryggevælde catchment (with an area of 130.5 km^2) is situated in the eastern part of Sealand, north of the village Karise. The soils in the catchment are predominantly clay, implying a very flashy flow regime. Six years of average daily runoff data, 1 Jan 1986 to 31 Dec 1991, are used for training. One year of data, 1 Jan to 31 Dec 1993, is used for verification. The data used in this study are provided by the Danish Hydraulic Institute (DHI), Denmark.

The prediction performance is evaluated using four goodness-of-fit measures, the Root-Mean-Square-Error (RMSE), Coefficient-of-Determination (COD), the Correlation Coefficient (*CC*) and the Mean Absolute Error (*MAE*), as in Eqs (5)–(8):

$$RMSE=\sqrt{\frac{1}{n}\sum[(Q_m)_i-(Q_s)_i]^2} \tag{5}$$

$$COD=1-\frac{\sum_{i=1}^{n}[(Q_m)_i-(Q_s)_i]^2}{\sum_{i=1}^{n}[(Q_m)_i-(\bar{Q}_m)]^2} \tag{6}$$

$$CC=\frac{\sum_{i=1}^{n}[(Q_m)_i-(\bar{Q}_m)][(Q_s)_i-(\bar{Q}_s)]}{\sqrt{\sum_{i=1}^{n}[(Q_m)_i-(\bar{Q}_m)]^2}\quad\sqrt{\sum_{i=1}^{n}[(Q_s)_i-(\bar{Q}_s)]^2}} \tag{7}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|(Q_m)_i-(Q_s)_i| \tag{8}$$

where Q is the discharge, the subscripts m and s represent the measured and simulated values, the average value of the associated variable is represented with a 'bar' above it and n is the total number of patterns considered.

Runoff prediction with single SVM model

The Support Vector Machine based single model is used to train the runoff data for 1- and 3-lead day predictions. The input variables are the immediate past streamflow data. The total number of immediate past data is decided through a trial-and-error approach. The procedure is summarized below:

(1) A 1-lead day runoff model is proposed:

$$Q_{t+1} = f(Q_t, Q_{t-1}, \ldots, Q_{t-\beta}) \tag{9}$$

where β is an integer commencing from 1;

(2) Train the above model with $\beta = 1$ for various SVM architecture (the kernel parameter and C);

(3) Record the goodness-of-fit measures, as given in Eqs (5)–(8), resulting from the various combinations;

(4) Repeat Steps 2 and 3 for $\beta = 2$ and 3; and

(5) The model adopted is the one which yields the best goodness-of-fit measures.

The above procedure is repeated for a 3-lead day model, Q_{t+3}. The best prediction results are obtained with 3-lag inputs ($\beta = 2$) and 4-lag inputs ($\beta = 3$) for Q_{t+1} and Q_{t+3} models, respectively. The goodness-of-fit measures of the SVM (training and verifications) are shown in Table 1. As can be seen from Table 1, the prediction accuracy of the model reduces with the increase in prediction time horizon, which is to be expected.

For subsequent prediction analysis, a streamflow to be predicted (1- or 3-lead day) is assumed to be in the high, medium or low flow region based on the streamflow magnitude of the current time, t. The classification criteria of various flow regions are given in Table 2. For example, a streamflow at the current time 't' is 3.2 m^3/s (high flow region): this magnitude would assume that the flow to be predicted (either 1- or 3-lead day) would be in the same (high) flow region. The performance of a single SVM model for the various flow regions is summarized in Table 2 and their scatter plots are displayed in Figs 1 and 2. It is interesting to note the following:

(1) For 1-lead day prediction of high flows within each individual flow regions, the high flow magnitudes are predicted reasonably well, based on (high) flow observed at the current time t, by the single SVM model (Fig. 1(a)). However, the high flow magnitudes, corresponding to the low flows and medium flows observed at time t, are predicted poorly (Figs 1(b) and (c)).

(2) For 3-lead day prediction of high flows within each individual flow regions, the single SVM model predicts high flows, in all three flow regions, at 3-lead day very poorly (Figs 2(a)–(c)). The coefficient of determination (COD) values, for example, for the high flow and medium flow regions are -0.02 and -0.03, indicating an overall prediction worse than the mean value.

(3) The low value of COD in all three flow categories, for 1- or 3-lead day predictions, can be mainly attributed to the poor predictions of high flow magnitudes at each of the flow

Table 2 Criteria adopted for various flow regions for 1- and 3-lead day predictions

Measured flow at current time t	Flow regions	Notation	
		1-lead day ($t + 1$)	3-lead day ($t + 3$)
Flow $\geq$ 3.00 m^3/s	High flow	HF_{t+1}	HF_{t+3}
Flow between 1.00–3.0 m^3/s	Medium flow	MF_{t+1}	MF_{t+3}
Flow $\leq$ 1.00 m^3/s	Low flow	LF_{t+1}	LF_{t+3}

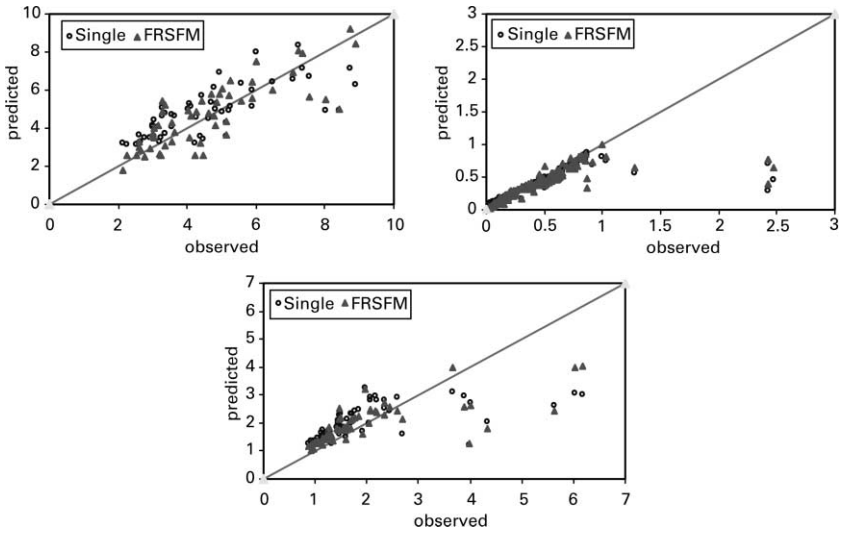

Figure 1 Single SVM and FRSFM model for 1-lead day prediction: verification

regions. Such occurrences of high flow events may require the use of different sets of input vectors for better prediction.

The goodness-of-fit measures, COD for example, when applied for the single model representing the whole flow range may be misleading. For example, in the case of the 1-lead day prediction, the COD value of the single model is as high as 0.87 (Table 2). However, in terms of the three categories of flows (high, medium and low), the corresponding COD is as low as 0.59, 0.30 and 0.56, respectively. The apparent high value of COD for the single model is due to the dominant presence of low flows which are also predicted quite well by the model. These reduce the mean of the observed flows appeared in the denominator of Eq. (6). The same observation is applied for the case of the 3-lead day prediction. The COD for the single model is as high as 0.50 whereas COD values for high, medium and low flows are as

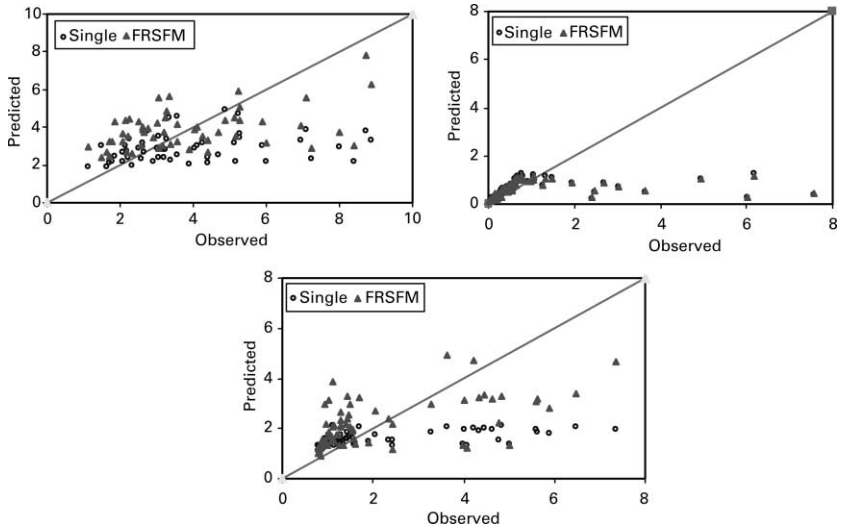

Figure 2 Single SVM and FRSFM model for 3-lead day prediction: verification

low as -0.02, -0.03 and 0.09, respectively. This difference is of significance and led to the development of separate flow models for flow regions based on magnitude.

The proposed method

The dynamical processes underlying the streamflow generation is very complex in nature and are influenced by various factors. It is generally believed that the dynamics of high flow and low flow events are different in nature. The high flows depend primarily upon the outburst of a heavy storm in the immediate past, catchment slope, vegetation cover and soil types. The low flows, on the other hand, are affected primarily by releases from ground water storage, which is a function of the distribution and infiltration characteristics of the soils, the hydraulic characteristics and the extent of the aquifers, the rate, frequency and amount of recharge.

In this approach, the flow time series is mapped to the pattern space (Fig. 3) consisting of three pattern classes, viz. high, medium and low flow regions. To each region, the streamflow data are assigned according to the criteria defined in Table 1. Representative attributes of each respective region must be identified. Within the same region (say, high flow), the attributes associated with the patterns may be different for different lead-day predictions. Thus, each model pertaining to a particular region and lead-day prediction will be unique in itself. In the real-time operation, depending upon the current day's average flow rate, a particular flow region model will be chosen.

The task of associating proper attributes (input variables) to the patterns is the most crucial. Attributes can be obtained either directly from the measured streamflow data (as is usually done by incorporating lag terms) or it can be derived (based on some understanding of the process) from the observed flow data or a combination of both. For example, a pattern vector, e.g. from the high flow region Q_{hf}, may be formed by associating the past 'n' days of streamflow data as

$$(Q_{HF})_{t+1} = [Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, \ldots, Q_{t-n}] \tag{10}$$

Once the appropriate attributes corresponding to each region and lead-day prediction are ascertained, the learning machine can be trained to yield the optimal prediction model.

The proposed approach is implemented with SVM for the Tryggevælde Catchment runoff. The nomenclature adopted for the models is: the first two letters in the model name refers to the flow region (HF for high flow, MF for medium flow and LF for low flow). The numeral represents the number of lead days under consideration. The results of the model are discussed to compare them with those obtained from Table 1 for the single SVM model for 1- and 3-lead days, respectively.

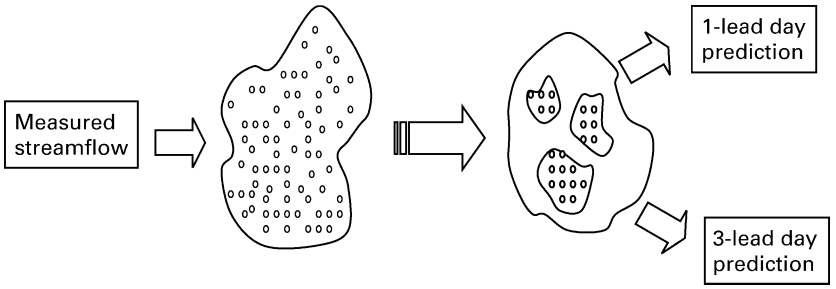

Figure 3 Object to pattern space mapping

Development of 1-lead day models

The trial-and-error procedure as described previously is adopted to ascertain the number of past streamflows controlling each of the high (HF_{t+1}), medium (MF_{t+1}) and low flow (LF_{t+1}) models. For high and medium flow models, 1-lead day prediction is mainly dependent on the immediate past streamflows, which represent surface runoff from storm events. However, this is not the same with low flows. The dynamics of low flow are very much dependent on zones in the vicinity of the river channel as opposed to the full range of hydrological processes occurring over the whole catchment during periods of relatively high flows. As such, it is important to note that the trial-and-error process adopted for the model development may not adequately capture the complexity of the subsurface flow movement in the aquifers. The final structure of the models adopted under each of these flow categories are summarized as below:

$$(Q_{HF})_{t+1} = [Q_t, Q_{t-l}, Q_{t-2}, Q_{t-3}] \tag{11}$$

$$(Q_{MF})_{t+1} = [Q_t, Q_{t-1}, Q_{t-2}] \tag{12}$$

$$(Q_{LF})_{t+1} = [Q_t, Q_{t-1}, Q_{t-2}, \ldots.., Q_{t-9}] \tag{13}$$

Discussion on model predictions

HF_{t+1} model. The model is implemented with FRSFM. When compared to the single SVM model, the HF_{t+1} model performs relatively better with a COD and RMSE of 0.63 and 1.05 (Table 3). As seen from the scatter plots, the typically high flows are predicted better, for the 1-lead day, than the single SVM model (Fig. 1(a)).

MF_{t+1} model. In comparison to the MF_{t+1} model (with a COD of 0.52), the single SVM model (with a COD of 0.52) performs poorly in predicting the typically high flow magnitudes (Table 3). Also, many low to medium magnitude flows are also predicted better with the MF_{t+1} model (Fig. 1(c)).

LF_{t+1} model. The best performance is obtained for a model with a total of 10 inputs. The model predicts poorly the high flow magnitudes with a COD and RMSE (on verification data) as 0.64 and 0.21. When compared to the single SVM model, the 10 input LF_{t+1} model performs marginally better in predicting the high flow. As observed from Table 3, the COD value improves from 0.56 to 0.64 for the single SVM model and the LF_{t+1} model.

From the scatter plot of SVM (Fig. 1(c)) it is observed that, although points A and B have nearly the same observed streamflow values, i.e. 3.978 m^3/s and 3.657 m^3/s, respectively, the predicted streamflows are very much different. An analysis is carried out to ascertain the reason and is tabulated in Table 4. It is noted that all the attributes of point A belong to medium flow ranges (i.e. between 1.00–3.00 m^3/s). The observed output for this point is

Table 3 Performance of flow region specific flow models: 1-lead day

	Goodness-of-fit measures			
Model (1)	**COD (2)**	**RMSE (m) (3)**	**CC (4)**	**MAE (5)**
Training				
HF_{t+1}	0.54	1.17	0.75	0.89
LF_{t+1}	0.67	0.21	0.82	0.08
MF_{t+1}	0.33	0.75	0.57	0.37
Verification				
HF_{t+1}	0.63	1.05	0.81	0.81
LF_{t+1}	0.64	0.21	0.80	0.06
MF_{t+1}	0.52	0.83	0.74	0.51

Table 4 MF_{t+1} model's prediction discrepancies

Points	Inputs (Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}) (m^3/s)	Attributes	Output Q_{t+1} (m^3/s)	Number of records with similar input to output patterns in the training set
A	(1.00, 1.04, 1.04, 1.00)	M,M,M,M	3.978	0
B	(0.36, 0.35, 0.50, 2.47)	L,L,L,M	3.657	12

3.978 m^3/s, which is a high flow. This type of input to output behaviour is not present in the training records, which makes the learning poor. In the case of point B there are as many as 12 training records with similar input–output relationships, resulting in good streamflow prediction for the 1-lead day model. In real-time applications, as more and more data come in every day, the prediction accuracy in the future is expected to improve.

Fig. 4 illustrates the hydrograph for the observed and predicted flow for 1-lead day prediction.

Development of 3-lead day models

Prediction accuracy will deteriorate, as expected, with the increase in the lead horizon. Results of the various flow regional models are discussed below.

HF_{t+3} model. Observation on the training data set reveals that a lot of combinations exist between the desired output streamflow Q_{t+3} and the immediate past four streamflow input data (Q_t, Q_{t-1}, Q_{t-2} and Q_{t-3}):

(a) "Weak" falling limb: the immediate past four streamflow data first show a rise and then a decline (Fig. 5(a)).
(b) "Weak" rising limb: this is characterized by a gradual rise (Fig. 5(b)) or sometimes a decrease first followed by a rise (Fig. 5(d)).
(c) "Strong" rising limb: this is characterized by "constant" high flows (Fig. 5(c)).

In all these cases, the observed flows at time $(t + 3)$ show wide variations from very high to very low flows for similar input patterns. This situation is best represented by two attributes, namely (1) the "minimum" of streamflow data at times $(t - 1)$, $(t - 2)$ and $(t - 3)$; and (2) the streamflow at time t. The final attributes adopted for this model are of the form

$$(Q_{HF})_{t+3} = [\min(Q_{t-1}, Q_{t-2}, Q_{t-3}), Q_t] \tag{14}$$

For example, a streamflow of approximately 5 units may have developed suddenly or gradually or may be coming from a "weak" falling limb hydrograph (see Table 5). When the

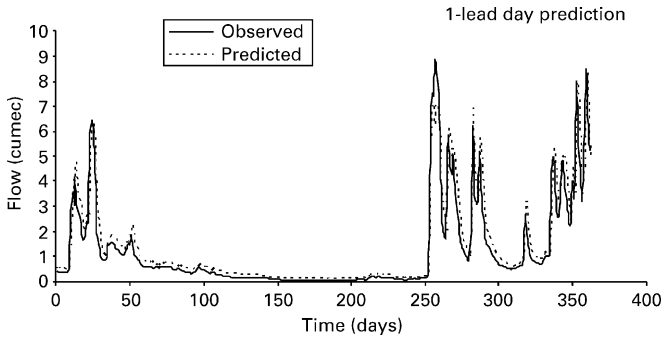

Figure 4 Observed and predicted hydrograph for 1-lead day prediction

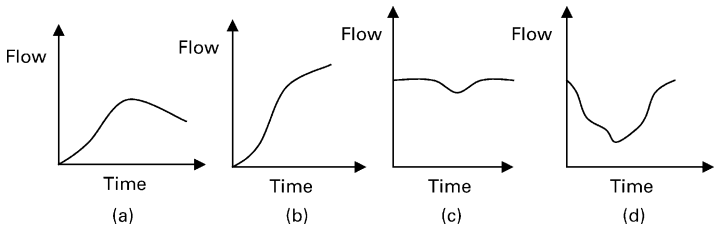

Figure 5 Various observed flow scenarios for predicting 3-lead day high flows

inputs are from the immediate past four streamflows, the number of similar input records is too few for good learning by the machine. However, when only two inputs [min (Q_{t-1},-Q_{t-2},Q_{t-3}), Q_t] are used to represent the situation, the description is more representative and hence more input records of the similar attributes may be found in the training set. Furthermore, these two attributes [min (Q_{t-1},Q_{t-2},Q_{t-3}), Q_t] assist in revealing the probable streamflow rate at time ($t + 3$).

The performance of SVM (as discussed above) is given in Table 6 together with a comparison of the results obtained with the trial-and-error procedure outlined in this paper. The proposed model improves the results significantly, with a COD and RMSE of 0.20 and 1.75, respectively. As shown in Tables 2 and 6, the single SVM model performs very poorly (with a COD of -0.02) in predicting the typically high flow events. Furthermore, the results obtained using the trial-and-error approach for FRSFM is only marginally better than that of the single model.

LF_{t+3} model. As observed earlier, the 3-lead day flow occurring in the low streamflows series is sometimes difficult to predict due to the difficulty in accounting for a possible rainfall event between the current day and the third day. This is particularly true for predicting high flow events observed on the ($i + 3$)th day. The procedure as outlined before is adopted. The final form of the model is

$$(Q_{LF})_{t+3} = [Q_t, Q_{t-1}, Q_{t-2}] \tag{15}$$

The prediction results in Table 6 and Fig. 2(b) show that the model predicts the high flow events at time ($t + 3$) very poorly. The low flow events at time ($t + 3$) are predicted reasonably well. The single model also fails to predict the high flow events at time ($t + 3$).

MF_{t+3} model. The effect of rainfall events of small storms that cause the medium flow may die out completely at the ($t + 3$) day. Therefore, the concept as applied for the model *HF*3 cannot be applied here. A sequential prediction approach is suggested as below:

(a) Determine the best ($t + 1$) day prediction (resulting from the *MF*1 model).
(b) Next, determine the ($t + 2$) day prediction with the ($t + 1$) day prediction as one of the inputs. The inputs to be included in the model are determined using the procedure described in a previous section. The model for 2-lead day prediction is:

$$Q_{MF_{2-t}} = [Q_{MF_{1-t}}, Q_t, Q_{t-1}] \tag{16}$$

where $x_{MF_{1-t}}$ is the 1-lead day prediction.

Table 5 Sample training data used for 3-lead day high flow prediction

Input to HF_{t+3} model (m^3/s)				Attributes of model HF_{t+3} (m^3/s)	
X_{t-3}	X_{t-2}	X_{t-1}	X_t	min(X_{t-1},X_{t-2},X_{t-3})	X_t
1.069	1.128	4.779	5.545	1.069	5.545
1.235	6.347	6.73	5.603	1.235	5.603
0.139	0.146	0.961	5.291	0.130	5.291

Table 6 Performance of flow region specific forecasting models: 3-lead day

Model (1)	COD (2)	Goodness-of-fit RMSE (m) (3)	CC (4)	MAE (5)
HF_{t+3}	0.20	1.75	0.45	1.38
LF_{t+3}	0.16	0.81	0.40	0.29
MF_{t+3}	0.32	1.38	0.57	1.05
HF_{t+3}*	0.05	1.92	0.26	1.45
MF_{t+3}*	0.19	1.51	0.57	1.06

* Refers to model arrived with trial-and-error procedure

(c) The 3-lead day flow is best found when the (t+1) and (t+2) predictions are included as the input variables. The proposed sequential prediction approach as outlined above (model *MF*3) is applied. The best prediction is obtained for a total of four input variables as shown below:

$$Q_{MF_{3-t}} = [Q_{MF_{2-t}}, Q_{MF_{1-t}}, Q_t, Q_{t-1}] \quad (17)$$

As can be seen from the scatter plot (Fig. 2(c)), the peak flow events are significantly improved by the MF_{t+3} model. A comparison is also done with the results obtained using the trial-and-error procedure (Table 6).

The single model predicts the high flow events observed on the 3-lead day very poorly. In general, the prediction is as good as the average of the flows as indicated by a very small COD. The proposed method with SVM predicts far better with a COD of 0.32 as opposed to -0.04 by the single model.

Fig. 6 illustrates the hydrograph for the observed and predicted flow for 3-lead day prediction.

Conclusions

A flow region specific flow forecasting approach is suggested in this study with a view to improving the prediction of typically high flows. The approach considers the streamflow time series in three separate regions, each uniquely constructed to give the best prediction for the different lead days considered. From this study, the following conclusions can be drawn:

(a) The FRSFM approach allows more flexibility in choosing the relevant input variables for different types of flows. For example, in the case of 1-lead day prediction, FRSFM indicates 4-, 3- and 10-lag inputs as the best for high, medium and low flows, respectively. In contrast, a single SVM model using 3-lag inputs is found to give the best prediction

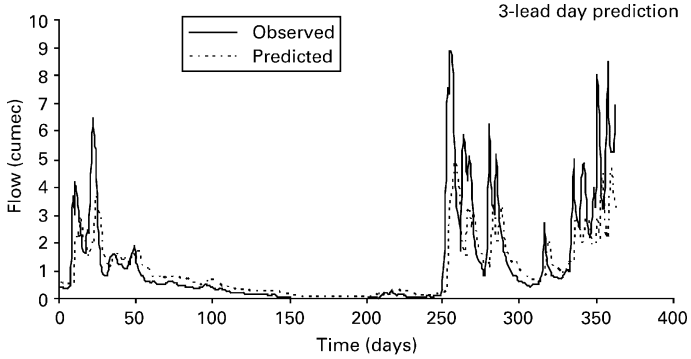

Figure 6 Observed and predicted hydrograph for 3-lead day prediction

over the entire range of flow. However, this model does not predict the typically high flow events well (observed at 1-lead day) for the low and medium flow categories.

(b) For higher lead-day predictions, a trial-and-error procedure for selecting the input variables results in a poor prediction both for the single model and the FRSFM based model. In such cases, some specialized models need to be explored. In this study, two such models are used for predicting medium and high flow regions.

(c) Since the FRSFM approach deals with small subsets of the total data (corresponding to high, medium and low flows), the training is very fast. This is particularly advantageous for real-time operation studies.

(d) Low flows for 1-lead and 3-lead days are poorly predicted because of the fact that the causative principle for low flow generation cannot be completely characterized by streamflow time series alone.

(e) Although the proposed FRSFM approach does improve the accuracy of the forecast, such an improvement was found to be not very significant. One possible reason could be that the use of long sampling times (daily, in this study) allow for situations in which flow in the low flow rate region, for example, becomes flow in the high flow rate region the following day. It is believed that should the sampling time interval be much smaller, say hourly, the drastic change in flow regime will not happen and the proposed method would demonstrate much more promising results.

(f) Self-organizing methods such as that based on Kohonen Neural Networks can be used for sorting the data before applying region specific models.

References

ASCE Task Committee (2000). Artificial neural networks in hydrology-2: hydrologic applications. *J. Hydrol. Engng.*, **5**(2), 124–137.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge.

Drucker, H., Burges, C., Kaufman, L., Smola, A. and Vapnik, V. (1996). Linear Support Vector Regression Machines. *NIPS* 96.

Georgakakos, A. (1989). The value of streamflow forecasting in reservoir operations. *Wat. Res. Bull.*, **25**(4), 789–800.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Englewood Cliffs, NJ.

Hsu, K.L., Gupta, H.V. and Sorooshian, S. (1995). Artificial neural network modeling of the rainfall-runoff process. *Wat. Res. Res.*, **31**(10), 2517–2530.

Karunanithi, N., Grenney, W.J., Whitley, D. and Bovee, K. (1994). Neural networks for river flow prediction. *J. Comput. Civil Engng.*, **8**(2), 201–220.

Liong, S.Y. and Sivapragasam, C. (2002). Flood stage forecasting with SVM. *J. Am. Wat. Res. Assoc.*, **38**(1), 173–186.

Liong, S.Y., Lim, W.H. and Paudyal, G. (1999). Real time river stage forecasting for flood stricken Bangladesh: neural network approach. *J. Comput. Civil Engng.*, **4**(1), 38–48.

Maier, H.R. and Dandy, G.C. (1996). The use of artificial neural networks for the prediction of water quality parameters. *Wat. Res. Res.*, **32**(4), 1013–1022.

Mishalani, N. and Palmer, N. (1988). Forecast uncertainty in water supply reservoir operation. *Wat. Res. Bull.*, **24**(6), 1237–1245.

Smola, A.J. and Scholkopf, B. (1998) A tutorial on support vector regression. *Technical Report* NeuroCOLT NC-TR-98-030. Royal Holloway College, University of London.

Thirumalaiah, K. and Deo, M.C. (1998). River stage forecasting using artificial neural networks. *J. Hydrol. Engng.*, **3**(1), 26–32.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin.

Vapnik, V. (1999). An overview of statistical learning theory. *IEEE Trans. Neural Networks*, **10**(5), 988–999.

Yapo, P., Gupta, V.K. and Sorooshian, S. (1996). Calibration of conceptual rainfall-runoff models: Sensitivity to calibration data. *J. Hydrol.*, **181**, 23–48.