Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/issn/15375110

Research Paper: SW—Soil and Water

Application and analysis of support vector machine based simulation for runoff and sediment yield

Debasmita Misra[a,*], *Thomas Oommen*[b,1], *Avinash Agarwal*[c,2], *Surendra K. Mishra*[d,3], *Anita M. Thompson*[e,4]

[a]*Geological Engineering, College of Engineering and Mines, University of Alaska Fairbanks, P.O. Box 755800, Fairbanks, AK 99775, USA*
[b]*Tufts University, Department of Civil & Environmental Engineering, 200 College Avenue, Medford, MA 02155, USA*
[c]*National Institute of Hydrology, Roorkee-247667, Uttaranchal, India*
[d]*Indian Institute of Technology, Water Resources Development Training Centre, Roorkee-247667, Uttaranchal, India*
[e]*Department of Biological Systems Engineering, University of Wisconsin – Madison, 230 Ag. Eng. Building, 460 Henry Mall, Madison, WI 53706, USA*

ARTICLE INFO

Article history:
Received 26 August 2008
Received in revised form
14 April 2009
Accepted 27 April 2009
Published online 17 June 2009

The objective of the study was to use Support Vector Machines (SVM) to simulate runoff and sediment yield from watersheds. Recently, pattern-recognition algorithms such as artificial neural networks (ANN) have gained popularity in simulating rainfall-runoff-sediment yield processes producing comparable accuracy to physics-based models. We have simulated daily, weekly, and monthly runoff and sediment yield from an Indian watershed, with monsoon period data, using SVM, a relatively new pattern-recognition algorithm. Model performance was evaluated using correlation coefficient for evaluating variability, coefficient of efficiency for evaluating efficiency, and the difference of slope of a best-fit line from observed-estimated scatter plots to 1:1 line for evaluating predictability. Time-series data were split into training, calibration and validation sets. The results of SVM were compared to those of ANN. An alternate method, the Multiple Regressive Pattern Recognition Technique (MRPRT), was used for runoff estimation only. The MRPRT did not improve the results significantly compared to SVM, hence, it was not used to simulate sediment yield. We concluded that SVM provided significant improvement in training, calibration and validation as compared to ANN. SVM could be an efficient alternative to ANN, a computationally intensive method, for runoff and sediment yield predictions providing at least comparable accuracy.

* *Corresponding author.*
E-mail addresses: debu.misra@uaf.edu (D. Misra), thomas.oommen@tufts.edu (T. Oommen), avinash@nih.ernet.in (A. Agarwal), skm61fwt@iitr.ernet.in (S.K. Mishra), amthompson2@wisc.edu (A.M. Thompson).
[1] Tel.: +1 215 435 0867; Fax: +1 617 627 3994.
[2] Tel.: +91 1332 272906; Fax: +91 1332 272123.
[3] Tel.: +91 1332 285457; Fax: +91 1332 271073.
[4] Tel.: +1 608 262 0604; Fax: +1 608 262 1228.

1537-5110/$ – see front matter © 2009 IAgrE. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.biosystemseng.2009.04.017

1. Introduction

The need for short-term and long-term simulation of runoff and sediment yield is important for watershed management that includes increasing infiltration into soil, controlling excess runoff, managing and utilizing runoff for specific purposes, and reducing soil erosion. The complex nature of the processes such as runoff and sediment yield, their variability depending on catchment characteristics and precipitation patterns, and their dependence on various other factors make it difficult to predict and estimate them with desirable accuracy. However, over the years, hydrologists have developed several models ranging from empirical to physically based relationships. The physically based models have proved to be better for the simulation of runoff and sediment yield, but their data requirements are very high and often intensively monitored watersheds lack sufficient input data for these models. Therefore, the need to develop alternative models to simulate runoff and sediment yield using available data has taken priority. Recently, pattern-recognition algorithms such as artificial neural networks (ANN) have shown promise in simulating the rainfall-runoff-sediment yield processes producing equivalent accuracy to those of the physically based models (Rajurkar *et al.*, 2004; Agarwal *et al.*, 2006; Raghuwanshi *et al.*, 2006; Ardiclioglu *et al.*, 2007; Cimen, 2008).

Support Vector Machines (SVM) (Vapnik, 1998; Kecman, 2000) provide a contemporary pattern-recognition technique (based on statistical learning theory) that has provided highly accurate estimates compared to ANN for spatial data analyses (e.g., Twarakavi *et al.*, 2006). SVM utilizes the structural risk minimization principle, which has been shown to be superior to the empirical risk minimization principle employed by ANN. To the best of our knowledge, SVM has not been used in time-series data analysis in hydrology. The *objective* of this paper is to analyze the applicability of SVM to simulate daily, weekly and monthly runoff and sediment yield from an Indian watershed (area = 7820 km^2), with data from the monsoon period. The simulation results obtained using SVM have been compared to those of ANN [developed and estimated by Agarwal (2002) and reported also in Agarwal *et al.* (2006)] to assess the relative outcome from both the models. We also analyzed the application of a new method, Multiple Regressive Pattern Recognition Technique (MRPRT), for estimation of runoff, since both SVM and ANN provided counterintuitive results, as discussed later.

2. Regression using support vector machines

We provide a brief overview of the theoretical concepts of Support Vector Regression (SVR), one of the techniques of SVM used in our analysis. However, a detailed depiction of SVR is beyond the scope of this paper and can be obtained from Vapnik (1995), Kecman (2000), and Hastie *et al.* (2003).

The data used to develop the regression model in machine learning, is called the training data. Suppose we have training data $\{(x_{11}, x_{12}, \ldots x_{1n}, y_1), \ldots, (x_{l1}, x_{l2}, \ldots x_{ln}, y_l)\} \subset X \times R$, where ($x_{l1} \ldots x_{1n}$) represent the predictor variables and y_l represents observed runoff/sediment yield at that location. The goal in SVR is to find a function $f(x)$ that has the most ϵ deviation from the observed lateral displacements y_l for all the training data, and at the same time, is as flat as possible. In other words, what Vapnik (1995) introduced through the ϵ-insensitive loss function is that errors less than ϵ are acceptable, but those deviations larger than ϵ are unacceptable. Mathematically,

$$f(x) = \langle w, x \rangle + b \quad \text{with } w \in X, \;\; b \in R \tag{1}$$

where $\langle w, x \rangle$ denotes the dot product in X. Flatness in Eq. 1 means a small value of w, and it can be obtained by minimizing the Euclidean norm, i.e., $\|w\|^2$. Thus the SVR problem can be formulated as shown in Eq. 2.

$$\text{minimise}, \; \frac{1}{2}\|w\|^2 \tag{2}$$

$$\text{subject to} \begin{cases} (\langle w, x_i \rangle + b) - y_i \leq \varepsilon \\ y_i - (\langle w, x_i \rangle + b) \leq \varepsilon \end{cases}$$

However, in some cases having a function f that is flat with errors less than ϵ is not feasible. To deal with these infeasible situations a constant C and slack variables ξ_i^-, ξ_i^+ are introduced which leads to the formulation (Eq. 3) as stated in Vapnik (1995).

$$\text{minimize} \; \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i^- + \xi_i^+) \tag{3}$$

$$\text{subject to} \begin{cases} (\langle w, x_i \rangle + b) - y_i \leq \varepsilon + \xi_i^- \\ y_i - (\langle w, x_i \rangle + b) \leq \varepsilon + \xi_i^+ \\ \xi_i^-, \xi_i^+ \geq 0 \end{cases}$$

where C is the pre-specified term that controls the magnitude of penalty associated with errors outside the error margin, and ξ_i^-, ξ_i^+ are slack variables representing upper and lower constraints on the output system. The constant $C > 0$ determines the trade-off between the flatness of the function and the amount to which deviations larger than ϵ are tolerated (Smola and Schölkopf, 2004).

Lagrange multipliers that employ the Kharush–Kuhn–Tucker (KKT) method are then used to solve the optimization problem in Eq. 3. The KKT method converts the inequality constraint into an equation of the form $h(x) = 0$ by adding or subtracting slack variables and then solving the corresponding equality-constrained quadratic optimization problem. Solution of the optimization problem results in a dual pair variable Lagrangian L_d (α_i, α_i^*), one for each of the training patterns. The pairs that result in non-zero α_i or α_i^* are termed the support vectors. When the SVR model is developed (training) it is the support vectors that define the hyper-plane (regression line) and fall on the optimal margin. Any data point in the training set that falls outside the optimal margin and within the error margin, does not contribute to the definition of the regression line.

Often in complex nonlinear problems the original input space (predictor variable) is non-linearly related to the predicted variable (lateral spread displacement). This limits a linear formulation of the problem as shown in Eq. 3. In SVR, this limitation is overcome by mapping the input space onto some higher dimensional space (feature space) using

a nonlinear mapping function (kernel function). The advantage of the kernel function is that it enables us to implicitly work in a higher dimensional feature space and overcome the issues of dimensionality. Commonly used kernel functions include the linear, polynomial, Gaussian radial basis, and sigmoid kernel functions. Keerthi and Lin (2003) demonstrated that a linear kernel is a special case of the Gaussian radial basis kernel and that the sigmoid kernel behaves like a Gaussian radial basis kernel for certain parameters. It can be concluded that the Gaussian radial basis kernel is a more generalized kernel function. Therefore, in this study we use a Gaussian radial basis kernel function (Eq. 4).

$$K(x,x') = \exp\left(-\frac{\|x-x'\|}{2\gamma^2}\right) \quad (4)$$

In any SVR problem, when we use the Gaussian radial basis function as the kernel function, we have three parameters to optimize during training: they are the Gaussian radial basis function parameter γ, magnitude of penalty term C, and the width/deviation of the error margin ϵ. We have used the described method in estimation of the daily, weekly, and monthly runoff and sediment yield in the study area described in the following section.

3. Study area

The study area chosen for this study is the Vamsadhara river basin, situated in between the Mahanadi and Godavari river basins of south India (Fig. 1; Agarwal, 2002; Agarwal *et al.*, 2006). The area is located between 18°15′ to 19°55′ north latitudes and 83°20′ to 84°20′ east longitudes. The precipitation in the basin is influenced by the occasional cyclones formed due to the depression in the Bay of Bengal and the south-west monsoon from June to October. The basin has six rain gauge stations and the weighted rainfall for the study area was estimated using the Thiessen polygons as shown in Fig. 1.

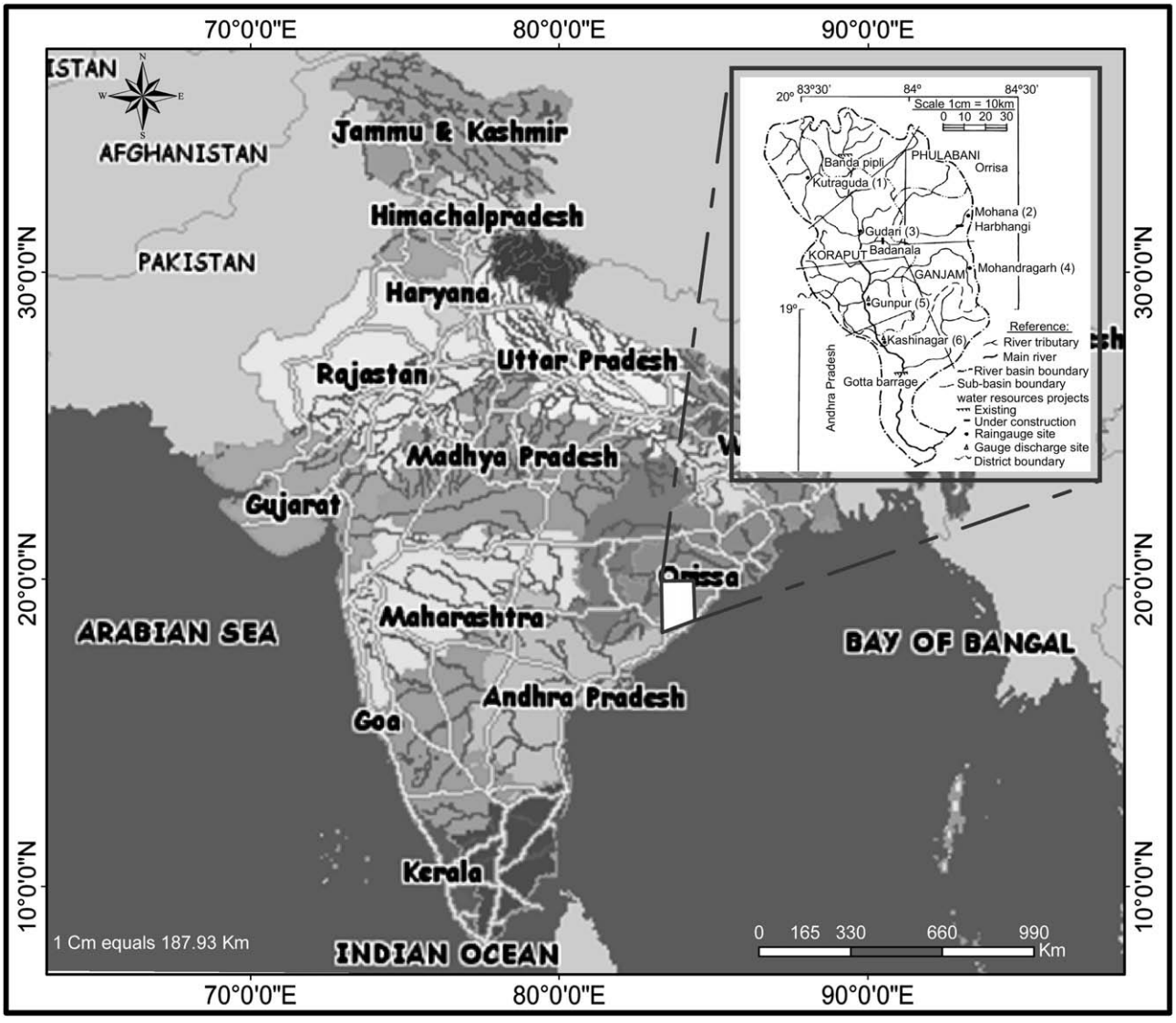

Fig. 1 – The location of the study area in India. Inset shows the position of the hydro-meteorological observation stations, district boundary, thiessen polygon and other detail information within the study area.

Table 1 – Input parameters used for the SVM and ANN models

Model	Variable	Model parameter associated with input time memory					
		Runoff				Sediment yield	
		t	t − 1	t − 2	t − 3	t	t − 1
Daily	Ri	I/P	I/P	–	–	I/P	I/P
	Qu	O/P	I/P	I/P	I/P	I/P	–
	Sy	–	–	–	–	O/P	I/P
Weekly	Ri	I/P	I/P	–	–	I/P	–
	Qu	O/P	I/P	–	–	I/P	–
	Sy	–	–	–	–	O/P	–
Monthly	Ri	I/P	I/P	I/P	–	I/P	–
	Qu	O/P	–	–	–	I/P	–
	Sy	–	–	–	–	O/P	–

Ri = total rainfall in mm, Qu = runoff in $m^3 s^{-1}$, Sy = sediment yield in $kg s^{-1}$, t = current day, I/P = input and O/P = output.

Similar to ANN models, the rainfall, runoff and sediment yield data of monsoon period (June 1–October 31) for 1984–87 was used for training the SVM model, and the data of 1988–89 and 1992–95 for calibration and validation. The input parameters were used for both the ANN and SVM models for runoff and sediment yield (daily, weekly and monthly) are shown in Table 1.

4. Method of model outcome assessment

The comparative evaluation of the outcome of both the models was done using correlation coefficient (r), coefficient of efficiency (E), and the difference of slope (SDiff). Out of all the various performance measures, in the past the most widely used evaluation for the validation of models is the correlation-based measures i.e., the r and R^2. However, they suffer from several limitations such as insensitivity towards additive and proportional difference occurring between the observed and the predicted data, and the over-sensitivity to outliers leading to a bias towards extreme events (Legates and McCabe, 1999). These limitations of the correlation-based measures are well documented (Willmott, 1981; Willmott *et al.*, 1985; Kessler and Neas, 1994; Legates and Davis, 1997; Legates and McCabe, 1999).

Coefficient of efficiency (E) is a non-dimensional criterion proposed by Nash and Sutcliffe (1970) and widely used to evaluate the performance of hydrologic models. $E = 100$ indicates a perfect agreement between the observed and the estimated values. $E = 0$ indicates that all the estimated values are equal to the mean of the observed values. A negative E indicates that the mean of the observed data is a better predictor than the estimated values. The coefficient of efficiency was an improvement over the correlation-based measures because it is sensitive to the observed and predicted means and variances but is also limited in the case of over-sensitivity to outliers (Nash and Sutcliffe 1970; Legates and McCabe, 1999).

Additionally, we have also used a new performance evaluation measure called Slope Difference (SDiff). The idea behind using SDiff is that while r is used to indicate the variational accountability of a model and E the efficiency, there is no comparative measure for the degree of predictability of a best-fit model to the 1:1 line when observed *vs.* predicted values are compared to each other. Hence, we have used SDiff as a measure of how different the slope of a best-fit line of the scatter plot of the predicted *vs.* observed data for a particular model is from the 1:1 line. SDiff of 0% means the best-fit line of a scatter plot is parallel to the 1:1 line thus ensuring perfect predictability of the best-fit linear model. SDiff of 100% means the best-fit line is the average line with a zero slope. SDiff between 0% and 100% would suggest that the best-fit linear model of the scatter plot would overestimate the low observed values and underestimate the high ones. A negative SDiff measure would suggest that the best-fit linear model of the scatter plot would underestimate the low observed values and overestimate the high ones.

5. Results and discussion

We obtained daily, weekly and monthly estimates of *runoff* and *sediment yield* using SVM (using SVR). In each case, 4 years of data (1984–87) were used to train the SVR model, 2 years of data (1988–89) were used for calibration to ensure similar model performance as ANN. Then the SVM model was applied to estimate 4 years worth of future data (1992–95). The observed and estimated data of 1992–95, as obtained from SVM were compared using r, E, and SDiff performance evaluation measures. Finally, outcome of both SVM and ANN were

Table 2 – Comparison of the performance of ANN and SVM models for daily, weekly and monthly runoff for calibration (1988–89) and validation (1992–95) periods. The optimal model parameters for SVM and MRPRT obtained from the training and calibration is presented below the corresponding model

Runoff models		Performance %, calibration (1988–89)		Performance %, validation (1992–95)*		
		r	E	r	E	Sdiff (%)
Daily	ANN	86.3	73.3	90.1	72.7	40.15
	SVM (c = 45, ε = .0004, γ = .75)	86.1	72.6	92.10	**80.02**	**33.79**
	MRPRT (c = 45, ε = .0004, γ = .75)	87.87	75.43	**92.72**	76.40	40.15
Weekly	ANN	79.6	60.3	87.4	54.6	56.95
	SVM (C = 176, ε = .0032, γ = .042)	80.67	62.03	91.16	67.55	47.0
	MRPRT (C = 50, ε = .02, γ = .255)	89.26	75.35	**91.98**	**80.48**	**26.17**
Monthly	ANN	77.4	26.0	79.3	−4.2	74.98
	SVM (C = 15, ε = .0091, γ = .05)	81.47	24.31	**86.33**	14.13	**60.09**
	MRPRT (C = 15, ε = .0124, γ = .625)	77.38	44.32	78.27	**22.61**	72.78

* The best performance index for a particular model for the validation data has been emboldened.

also compared. In case of *runoff* estimation, we obtained model performance that was counterintuitive. Hence, we also applied a new method called MRPRT (Oommen *et al.*, 2006), which is discussed later in this paper, to compare with the outcomes of SVM and ANN. The results are presented separately for *runoff* and *sediment yield* estimations.

5.1. *Runoff estimation using SVM*

SVM was used to obtain the output of daily, weekly and monthly runoff estimates using the input parameters as shown in Table 1. In our following discussion, we will emphasize how SVM has been proven to be a robust method despite poor estimates being obtained for the weekly and monthly runoffs using both ANN and SVM.

As stated before, time-series data of 1984–87 were used to train the SVM model. In order to obtain a comparable validation or estimate, the model was calibrated using the time-series data of 1988–89 (Table 2). We have developed an SVM model that is comparable to the ANN models developed by Agarwal (2002) and Agarwal *et al.* (2006) as is evident from the close (or similar) calibration performance measures (r and E) in Table 2.

In Table 2, we provide the validation or estimation performance measures for all the three SVM models as compared to the ANN models. Simply reviewing the coefficient of correlation (r) measure, it can be concluded that SVM provides improved estimates over ANN. It may be noted that the difference in the performance measures from the daily to the monthly estimates using SVM was reduced by 5.77%, while that of ANN was reduced by 10.8%. Hence, SVM provided a relatively robust estimation if r was only considered as a measure of performance.

Reviewing the E measures in Table 2, it may be concluded that the reliability of SVM estimates reduce from daily through monthly estimations. However, the same trend is also observed for the ANN models. The major difference between SVM and ANN estimates as obtained from E is that all the values of E in SVM are positive and hence provide a better estimate than the mean of the observed values. In case of ANN estimates, the daily and weekly E values are positive but the monthly value of E is negative thus depicting a poor estimation of monthly runoff as compared to the observed values.

It is intriguing to realize such reduction in model estimation performance with time-series data that have reduced non-linearity (such as monthly runoff data) as opposed to the daily runoff data. Despite the fact that all three models were trained with the same time-series data and were calibrated to close performance measures, such discrepancy was surprising. A careful review of the input parameters used for the estimation of the runoff (as provided in Table 1) revealed that the rainfall and runoff of the previous week were used as input to estimate the runoff of the current week. Similarly, the rainfall of the prior two months was used as input parameters for estimation of the runoff of current month. These input parameters did not make hydrological sense to us although such parameters were obtained from the best-fit model of a Linear Transfer Function (LTF) estimator (see Agarwal, 2002; Agarwal *et al.*, 2006). Hence, we conclude that the lack of improved estimates from the weekly and monthly models with both SVM and ANN were caused by improper input parameters used in their estimation.

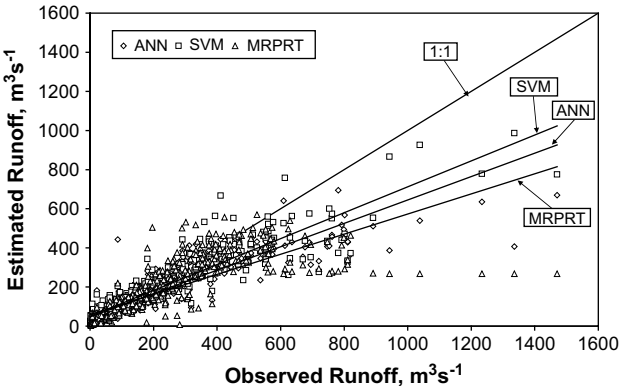

Fig. 2 – Scatter plot of the observed *vs.* estimated daily runoff (1992–95) using MRPRT, SVM and ANN methods.

Figs. 2–4 illustrate that the validation (1992–95) of the daily, weekly and monthly runoff models provided an improved estimate using the SVM over the ANN. While both SVM and ANN accounted for the variability of the data to a good extent (as observed from the r & E performance measures), outcomes of both the methods under-predicted the high values as seen from their deviation from the 1:1 line in daily, weekly and monthly estimates. Reviewing the SDiff measures in Table 2, it can be observed that SVM outperformed ANN in all three cases of runoff estimation.

Two aspects of our estimates of runoff using SVM forced us to seek alternate methods that could improve the daily, weekly and monthly runoff estimates. The first aspect was that the performance measures r and E showed a decline with SVM outcomes, albeit not as much as with ANN, for the validation data set when we went from daily to monthly estimations. While this may be a possibility in a statistical sense because the number of data used for the model development decreased from daily to monthly, what we considered to be counterintuitive was the fact that as we averaged the data from daily through weekly to monthly, we were also decreasing the degree of non-linearity in the data through this process. Hence, our expectation was for the SVM model to improve in performance as the data was averaged. The second aspect was the deviation of the best-fit line

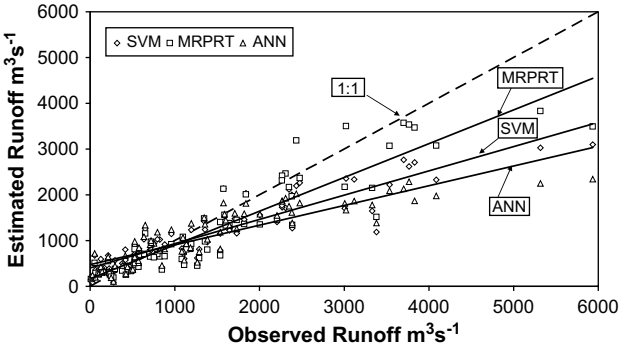

Fig. 3 – Scatter plot of the observed *vs.* estimated weekly runoff (1992–95) using MRPRT, SVM and ANN methods.

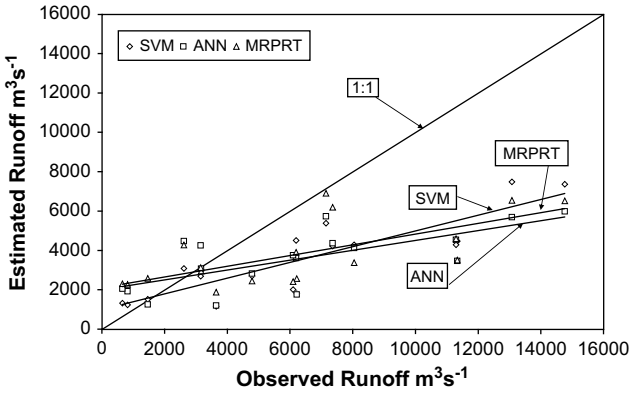

Fig. 4 – Scatter plot of the observed *vs.* estimated monthly runoff (1992–95) using MRPRT, SVM and ANN methods.

of each model outcome from the 1:1 predictor as shown in Figs. 2–4.

5.2. *Runoff estimation using MRPRT*

In order to improve our runoff estimates we considered using an alternate and newly developed method called the MRPRT (Oommen *et al.*, 2008a; Oommen *et al.*, 2006). The basic principle of the MRPRT is explained using the scatter plot shown in Fig. 5. The *x* and *y* axis of the scatter plot represent the actual and predicted values of the three models (individual regression techniques). The coefficients of correlation (Fig. 5) for the models 1, 2, 3 are 0.28, 0.09 and 0.10, respectively. It is evident from Fig. 5 that each of the three models individually have

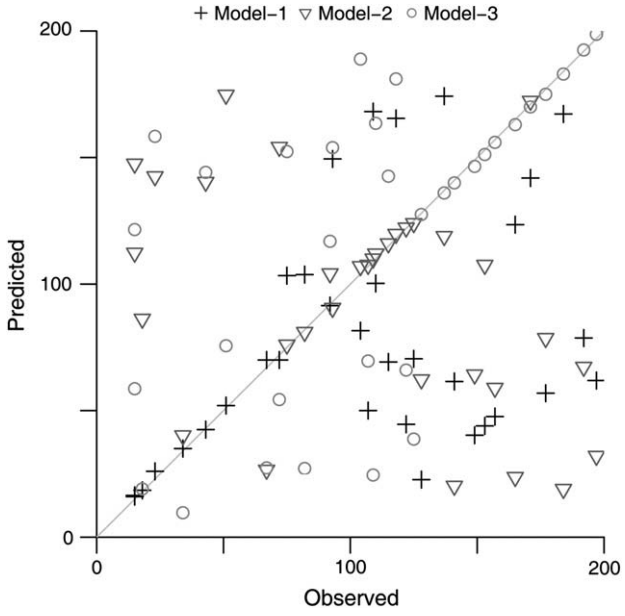

Fig. 5 – Scatter plot showing the observed *vs.* predicted values for three hypothetical datasets represented as models 1, 2 and 3. Model-1 predicts the low values accurately, model-2 the median and model-3 the high values (adopted from Oommen *et al.*, 2006).

overall poor estimative capability. However, it is learnt from Fig. 5 that model-1 is able to predict the low values accurately, model-2 the median values and model-3 the high values. Now the question is how could the accurate prediction of these different models be captured into a single model? In order to achieve this, a pattern-recognition technique, such as the SVM was used, where the technique would learn from the output of the three models and capture the accurate prediction of each individual model for an improved overall prediction of the combined data (Oommen *et al.*, 2006; Oommen *et al.*, 2008a).

When several techniques are used to learn/model a problem, the usual approach is to choose the one that performs the best on an independent validation set. However, learning/modelling is an ill-posed problem with finite data, each algorithm converges to a different solution and fails under different circumstances. Therefore, if we can obtain learning algorithms that produce different solutions that complement each other, then they can be combined to develop a single model that would perform better than an individual learning algorithm. This is the underlying principle of MRPRT.

The MRPRT is based on the technique of stacked generalization proposed by Wolpert (1992) in which the outputs of the base learners is combined and learned through another combiner system (another learning algorithm). The combiner system learns what the correct output is when the base learners produce a certain output combination. In MRPRT we combine both time-series statistics and machine learning techniques and further extend stacked generalization for regression problems.

Table 2 provides the model output for daily, weekly and monthly runoff estimates using MRPRT and as compared to the ANN and SVM model outputs. In Table 2, it may be observed that the calibrated MRPRT model is comparable to those of the SVM and ANN models for the daily, weekly and monthly runoff simulations. Comparing the *r* and *E* performance measures of the validation outputs from MRPRT with SVM (Table 2), it may be concluded that MRPRT did not provide substantial improvement in prediction of runoff over SVM in all three cases (daily, weekly and monthly estimates). However, it may be observed from the *E* performance measures alone that the MRPRT method provided significant improved estimates over SVM for weekly and monthly outputs.

MRPRT did not provide improved estimate in comparison to SVM as evident from the scatter plot illustrated in Fig. 2 for daily runoff prediction, even though the predicted runoff in both SVM and MRPRT were under-predicted for high values. MRPRT definitely provided significant improved estimates over SVM and ANN for weekly runoff as observed in Fig. 3, especially for higher values of runoff. In the case of monthly runoff estimates, none of the three models had any significant predictability (Fig. 4). All the three models grossly underestimated the observed data as is also evident from Table 2. A review of the SDiff measure for MRPRT (Table 2) reveals that the best-fit line to the model outcomes was closer to the 1:1 line only in the case of weekly runoff estimates. In case of daily runoff, SVM had a smaller SDiff than MRPRT and in case of monthly runoff all three models (ANN, SVM, and MRPRT) had a large SDiff value.

Table 3 – Comparison of the performance of ANN and SVM models for daily, weekly and monthly sediment yield for calibration (1988–89) and validation (1992–95) periods. The optimal model parameters for SVM and MRPRT obtained from the training and calibration is presented below the corresponding model

Sediment yield models		Performance %, calibration (1988–89)		Performance %, validation (1992–95)*		
		r	*E*	*r*	*E*	Sdiff (%)
Daily	ANN	79.30	62.80	83.20	68.00	**21.18**
	SVM (C = 15, ε = .001, γ = .05)	80.02	62.94	**87.87**	**75.68**	30.46
Weekly	ANN	80.20	64.10	75.10	51.80	45.90
	SVM (C = 30, ε = .00068, γ = .04)	78.43	60.26	**88.08**	**74.62**	**34.84**
Monthly	ANN	89.40	79.10	74.10	53.70	52.06
	SVM (C = 80, ε = .0015, γ = .03)	81.32	62.44	**87.66**	**74.49**	**24.76**

* The best performance index for a particular model for the validation data has been emboldened.

5.3. *Sediment yield estimation using SVM*

Models similar to those for runoff were developed using SVM to estimate daily, weekly, and monthly sediment yields from the basin. The input parameters used to develop these models (see Table 1) as obtained from the best-fit LTF model (see Agarwal, 2002; Agarwal *et al.*, 2006) were found to be reasonable and made hydrological sense. All the three SVM models were calibrated using the time-series data of 1988–89 to obtain comparable performance measures to those of the ANN models (Table 3). The performance measures obtained from the validation data time-series of 1992–95 for SVM and ANN models are compared to the observed sediment yield values in Table 3.

Comparing the *r* measures of the validation or estimation performance, it can be seen that the ANN models provided reduced performance of 9.1% between daily and weekly estimations. Although such reduction in performance is counterintuitive with the proper input parameters used to obtain the estimates, choice of an appropriate transfer function in ANN might affect estimation of variables. A similar trend is observed with the *E* performance measure with ANN. On the contrary, SVM proved to be a robust estimator of sediment yield with almost consistent performance in both *r* and *E* measures as shown in Table 3.

Figs. 6–8 illustrate that the validation output (1992–95) of the daily, weekly and monthly sediment yield models provided either comparable (daily) or considerably improved (weekly and monthly) estimates using the SVM over the ANN.

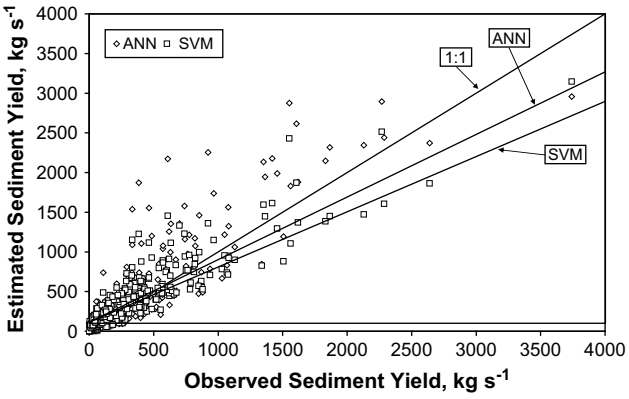

Fig. 6 – Scatter plot of the observed *vs.* estimated daily sediment yield (1992–95) using SVM and ANN methods.

In fact, estimates of monthly sediment yield using SVM were significantly closer to the observed values. As the non-linearity in the time units were reduced (from daily through monthly), the performance accuracy of the ANN model was considerably reduced, however, that of SVM demonstrated consistent performance accuracy. From Table 3, it may be observed that while *r* and *E* measures showed consistently that SVM provided better model outcome than ANN, the SDiff measure demonstrated that ANN provided closer predictability for daily sediment yield values while outcome of SVM was closer to the 1:1 line for weekly and monthly estimates. We did not use MRPRT to simulate sediment yield based on our experience in simulation of runoff.

6. Conclusions

SVM was used as a pattern-recognition (artificial intelligence) predictor to simulate daily, weekly and monthly runoff and sediment yield from an Indian watershed. The simulated results using SVM were compared to those obtained by Agarwal *et al.* (2006) using ANN models. SVM is a relatively new pattern-recognition algorithm and has been hardly applied to simulate time-series data, while it has been successfully used in predicting spatial distribution for resource and hydrologic estimations (Oommen *et al.*, 2008b).

From our results we specifically, arrive at the following conclusions:

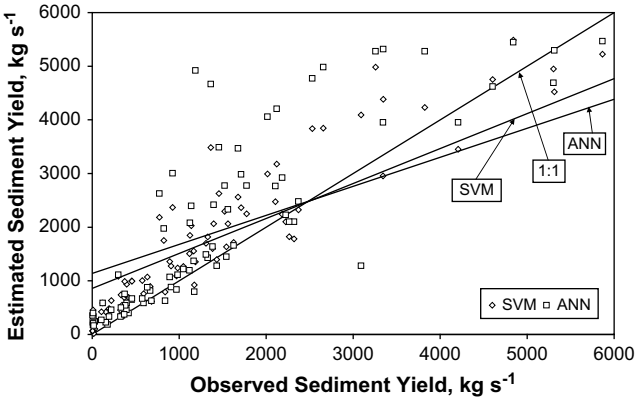

Fig. 7 – Scatter plot of the observed *vs.* estimated weekly sediment yield (1992–95) using SVM and ANN methods.

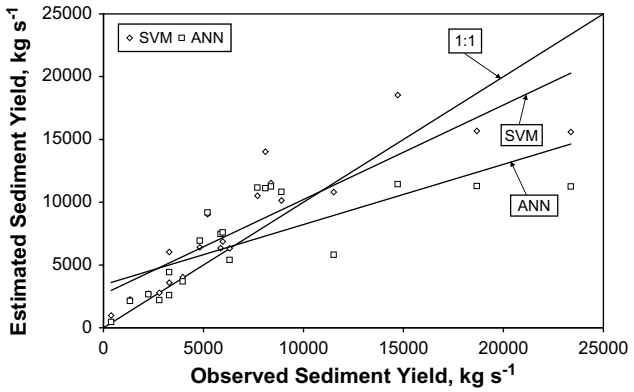

Fig. 8 – Scatter plot of the observed *vs.* estimated monthly sediment yield (1992–95) using SVM and ANN methods.

1. SVM provided a robust and better estimation in comparison to ANN for both runoff and sediment yield in the Indian watershed.
2. As the non-linearity in the time-series data were reduced (from daily through monthly), the performance accuracy of ANN was significantly reduced in the case of both runoff and sediment yield estimation. Whereas, SVM provided consistent performance accuracy for sediment yield estimation and reduced reduction in the performance of runoff estimation compared to ANN. It indicates that SVM is better for modelling sparse data conditions compared to ANN.
3. We believe that the reduction in the estimates of runoff performance accuracy using SVM from daily through monthly was due to the improper use of input parameters for weekly and monthly predictions. However, using different sets of input parameters that make hydrological sense for runoff estimation (especially for weekly & monthly estimates) might improve performance accuracy using SVM.
4. MRPRT did not provide significant improvement over SVM for runoff estimation probably because the inputs (ANN and SVM outputs) did not provide complementary information.
5. Using *r* and *E* performance measures as model evaluation tools for runoff estimation, it may be concluded that SVM or MRPRT could be used successfully in time-series simulation of runoff. However, using the SDiff performance measure, one may realize that neither ANN nor SVM nor MRPRT are suitable for monthly runoff simulation.
6. SVM turned out to be a clear choice in sediment yield estimation using *r* and *E* as performance measures. However, using SDiff, one may conclude that SVM was unsuitable for daily prediction of sediment yield as compared to ANN, although the SDiff values between the two models for daily estimation were close. In case of weekly and monthly sediment yield estimation, SDiff suggested SVM as the clear choice.

Overall we conclude that with ANN being computationally intensive, SVM is an efficient alternative and robust approach to model complex hydrological time-series data. From our results we may conclude additionally that daily and weekly runoff predictions using SVM or MRPRT may be successfully used along with observed precipitation data to predict future sediment yield in the watershed. While using SVM for hydrologic time-series data estimation, one needs to use input parameters that make hydrological sense, must optimize the model parameters with care, and use the three performance measures, *r*, *E*, and SDiff to evaluate the model outcome so that the performance of the model is evaluated appropriately for accountability of variability, efficiency and predictability.

REFERENCES

Ardiclioglu M; Kisi O; Haktanir T (2007). Suspended sediment prediction by using two different feed-forward back–propagation algorithms. Canadian Journal of Civil Engineering, **34**(1), 120–125.

Agarwal A (2002). Artificial Neural Networks and their application for simulation and prediction of runoff and sediment yield. Ph. D. thesis, Department of soil and Water Conservation Engineering. G.B. Pant University of Agriculture and Technology, Pantnagar, India.

Agarwal A; Mishra S K; Ram S; Singh J K (2006). Simulation of runoff and sediment yield using artificial neural networks. Biosystems Engineering, **94**(4), 597–613.

Cimen M (2008). Estimation of daily suspended sediments using support vector machines. Hydrological Science Journal, **53**(3), 656–666.

Hastie T; Tibshirani R; Friedman J (2003). The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer-Verlag, New York, ISBN 0-387-95284-5. 533p.

Kecman V (2000). Learning and Soft Computing, Support Vector Machines, Neural Network and Fuzzy Logic Models. MIT Press, ISBN 0-262-11255-8. 608p.

Kessler E; Neas B (1994). On correlation, with applications to the radar and raingage measurement of rainfall. Atmospheric Research, **34**, 217–229.

Keerthi S S; Lin C J (2003). Asymptotic behaviors of support vector machines with Gaussian Kernel. Neural Computation, **15**, 1667–1689.

Legates D R; Davis R E (1997). The continuing search for an anthropogenic climate change signal – limitations of correlation-based approaches. Geophysical Research Letters, **24**(18), 2319–2322.

Legates D R; McCabe G J (1999). Evaluating the use of ''Goodness of Fit'' measures in hydrologic and hydroclimatic model validation. Water Resources Research, **35**(1), 233–241.

Nash J E; Sutcliffe J V (1970). River flow forecasting through conceptual models. Journal of Hydrology, **10**, 282–290.

Oommen T; Prakash A; Misra D; Kelley J J; Naidu S; Bandopadhyay S (2008a). GIS based marine platinum exploration, goodnews bay, southwest Alaska. Marine Georesources & Geotechnology, **26**(1), 1–18.

Oommen T; Misra D; Twarakavi N K C; Prakash A; Sahoo B; Bandopadhyay S (2008b). An objective analysis of support vector machine based classification for remote sensing. Mathematical Geosciences, **40**(4), 409–424.

Oommen T; Misra D; Prakash A; Bandopadhyay S; Naidu S; Kelley J J (2006) Marine Geodatabase and Multiple Regressive Pattern Recognition Technique: A New Approach To Marine Placer Resource Assessment. American Geophysical Union Fall Annual Meeting, San Francisco, CA, Poster (IN14A).

Rajurkar M P; Kothyari U C; Chaube U C (2004). Modeling of the daily rainfall-runoff relationship with artificial neural network. Journal of Hydrology, **285**, 96–113.

Raghuwanshi N S; Singh R; Reddy L S (2006). Runoff and sediment yield modeling using artificial neural networks: upper siwane river, India. Journal of Hydrologic Engineering, **1**(71). doi:10.1061/(ASCE)1084-0699(2006)11.

Smola A J; Schölkopf B (2004). A tutorial on support vector regression. Statistics and Computing, **14**, 199–222.

Twarakavi N K; Misra D; Bandopadhyay S (2006). Prediction of arsenic in bedrock derived stream sediments at a gold mine site under conditions of sparse data. Natural Resources Research. doi:10.1007/s11053-006-9013-6.

Vapnik V N (1995). The Nature of Statistical Learning Theory. Springer, New York.

Vapnik V N (1998). Statistical Learning Theory. John Wiley and Sons, New York.

Willmott C J (1981). On the validation of models. Physical Geography, **2**, 184–194, p. 179.

Willmott C J; Ackleson S G; Davis R E; Feddema J J; Klink K M; Legates D R; O'Donnell J; Rowe C M (1985). Statistics for the evaluation and comparison of models. Journal of Geophysical Research, **90**, 8995–9005.

Wolpert D H (1992). Stacked generalization. Neural Networks, **5**, 241–259.