Proceedings WARMICE-2012
Sudhir Kumar, et al. (eds)
I.A.H., 2012

IV-7

Downscaling of Climate Variables using Support Vector Machine

Kuji Murtiningrum, SK Jain and ML Kansal
Department of WRD & M, IIT Roorkee, Uttarakhand, India

Abstract : World's climate is showing the changes in a number of components of the hydrological cycle and hydrological systems. Thus it is very important that scientist try to predict the future climate so that we can prepare strategies as part of mitigation and adaptation. Global Climate Models (GCMs) are the best tool to predict future climate but have resolution of hundreds of kilometer. However, many impact applications require the local scale climate variations. Statistical downscaling is one method to feed the large-scale output of GCM simulation into a statistical model to estimate the corresponding local and regional climate characteristics. In this paper, Multi Linear Regression (MLR) and Support Vector Machine (SVM) approaches were applied for statistical downscaling for precipitation and temperature variables in Roorkee area. The results are encouraging.

Keywords: Climate; Downscaling; GCM; MLR; SVM.

INTRODUCTION

Warming of the climate system is unequivocal and is already impacting a range of human and natural systems. Scientists have observed changes in the timing of seasons; the range of plant and animal species; regional pattern of precipitation, flooding, and drought. Sea levels are rising and glaciers and Arctic sea ice are forging a steady retreat, as is now evident from observations of increase in global average air and ocean temperatures, widespread melting of snow and ice, rising sea levels. Since 1980s, eleven of the last twelve years (1995-2006) rank among the twelve warmest years in the instrumental record of global surface temperature [1].

The Intergovernmental Panel on Climate Change (IPCC) has concluded that this warming is pri-marily the result of human activities [2]. Since the time of the Industrial Revolution (1850s), activities including deforestation and the burning of fossil fuels have released increasing quantities of greenhouse gases (GHGs) into our atmosphere. These gases, which in-clude carbon dioxide and methane, among others, trap heat that would otherwise escape into space. As such, the gases which have accumulated in the Earth's atmosphere have intensified the natural effect and now are causing climate change.

IPCC Special Report on Emissions Scenarios (SRES) predicts that temperatures will rise by 1.1 to 6.4° C by the end of 21 century, with range largely dependent on future GHG emissions. The type and severity of impacts that is associated with such temperature increases will vary by region, but on the whole they are expected to be negative and in some cases disastrous. Furthermore, the greater the temperature increase, the greater the impacts we can expect. Fragile ecosystems, coastal areas and low–lying islands will be destroyed. Species unable to adapt to changing conditions will go extinct. Agricultural pests and vector-borne diseases will spread, and people will suffer as droughts, floods, and storms may become more frequent and more intense. The world's poor will be hit first, and hardest, as changing climatic conditions exacerbate problems of food security, water scarcity, and sanitation [2].

It is almost certain that the world is experiencing climate change and hence additional risks will arise in the future. Thus it is very important that

scientists predict future climate. This is necessary so that we can prepare ourselves to face the future climate and make strategies as part of mitigation planning and adaptation.

General Circulation Model or Global Climate Models (GCM) have been developed to simulate the present climate and predict future climatic change. These are designed to simulate time series of climate variables globally, accounting for the effects of GHGs in the atmosphere. GCMs perform reasonably well in simulating climatic variables at larger spatial scale, but poorly at the smaller space and time scales relevant to regional impact analyses, especially in the important area of hydrology. Therefore the output from a GCM has to be downscaled to obtain the information relevant to hydrologic studies [3]. Downscaling climate data is a strategy for generating locally relevant data from GCM. The overarching strategy is to connect global scale predictions and regional dynamics to generate regionally specific forecasts. Basically, downscaling technique is a movement from large scale to small scale. One way to connect the GCM large scale with a smaller scale (study area) is to use Statistical Downscaling (SD) technique. SD is a process of downscaling where data on large-scale grids in the period and particular time is used as the basis for determining the data on the smaller grid scale.

SUPPORT VECTOR MACHINE

Support vector machine (SVM) is a relatively new technique to make prediction, both in the case of classification and regression, which is becoming very popular lately. The foundation of SVM has been developed by Vapnik [4] and is gaining popularity due to many attractive features, and promising empirical performance. The formulation embodies the Structural Risk Minimization principle, which has been shown to be superior [5] to traditional Empirical Risk Minimization principle, employed by conventional neural networks. SRM minimizes an upper bound on the expected risk, as opposed to ERM that minimizes the error on the training data. It is this difference which equips SVM with a greater ability to generalize, which is the goal in statistical learning. SVM was developed to solve the classification problem, but recently they have been extended to the domain of regression problems.

The term SVM is typically used to describe classification with support vector methods and support vector regression is used to describe regression with support vector methods. In this paper, the term SVM will refer to both classification and regression methods, and the terms Support Vector Classification (SVC) and Support Vector Regression (SVR) will be used for specification.

Statistical Learning Theory

This section is a very brief review some of Vapnik's statistical learning theory which based on learning examples. As is the case, learning is a stochastic process, with the training data being drawn from two sets of variables: an Input vector $x_i \in X \subseteq \Re^n$ and the response or Output y_i Y.

The relationship between X and Y is probabilistic: an element X does not map uniquely to an element of Y; rather it defines a probability distribution on X. Alternatively for x_i drawn from every X with probability $P(x_i)$ (called the marginal probability), the output y_i is observed with probability $P(y_i|x_i)$ (called the conditional probability of y_i given x_i). In other word, an unknown probability distribution $p(x,y)$ defined on $X \times Y$ determines the probability of observing a training data point (x_i,y_i). Therefore the training data set T = which we have been using time and again, is actually generated by sampling the cross space $X \times Y$, Q times in accordance with the distribution $p(x,y)$. This learning problem is searching for appropriate estimator function f: X Y which can then be used in predictive mode to generate a value y in output to an unseen input x.

To solve the regression or classification task, a learning machine learns an approximating function *f(x,á)* (also referred to as a hypothesis) which is a function of both inputs ***x*** and the parameter or weights ***á*** as the notation emphasizes.

Empirical Risk Minimization and Structural Risk Minimization.

The risk functional is the expected value of the loss due to the classification or estimation error. It employs a loss function *L* to measure the average error, and then searches out the estimator from the space hypotheses, that minimizes this risk. If the desired value is *y* and the predicted value is *f(x, á)*, then the expected risk is defined as:

R(á) = equation here

$$\int L\left(y, f(x,\alpha)\right) dP(x,y) \qquad (1)$$

As the probability *P*(x,y) is unknown, the risk *R(á)* cannot directly be minimized therefore an induction principle for risk minimization is required. This inductive principle is called *Empirical Risk Minimization (ERM)* which computes the empirical risk function as:

$$R_{emp}\ (\alpha) = \frac{1}{N} \sum_{i=1}^{N} L(y_i - f(x_i, \alpha)) \qquad (2)$$

However this $R_{emp}(á)$ will not able to guarantee a small actual risk if the number of training examples (N) is limited. In other words, a smaller error on the training set does not necessary implies higher generalization ability (i.e., smaller error on an independent test data). To make the most out of limited data, a statistical technique called *Structural Risk Minimization (SRM)* has been developed by Vapnik [4].

The theory of uniform convergence in probability developed in 1974 by Vapnik and Chervonenkis (VC) provides bounds on the deviation of the empirical risk from the expected risk. This theory shows that it is crucial to restrict the class of function that the learning machine can implement to one with a capacity that suitable for the amount of available training data. For á $\in$ Λ and N > h, a typical uniform VC bound which holds with probability $1 - \eta$, has the following inductive principle *SRM* form:

$$R(á)d"R_{emp}(á)+ \sqrt{\frac{h\left(\log\frac{2N}{h}+1\right)-\log\left(\frac{\eta}{4}\right)}{N}} \qquad (3)$$

Here the second term on the right is called *VC Confidence*. The parameter *h* is called the *VC dimension* of a set of function and it describes the capacity of a set function to represent the data set. When N/h is small, a small empirical risk does not guarantee a small value of the actual risk. In this case, to minimize the actual risk *R(á)*, the inequality on right hand in (3) should be minimized simultaneously over both terms; the empirical risk and the VC confidence interval.

The VC confidence term in (3) depends on the chosen class of the function, whereas the empirical risk depends on one particular function chosen by the training procedure. The objective here is to find that subset of the chosen set of the function, such that the risk bound for that subset is minimized. This is done by simply training a series of machines, one of each subset; where for a given subset the goal of training is simply to minimize the empirical risk. One then takes that trained machines in the series whose sum of empirical risk and VC confidence is minimal.

Feature Space

In case of non-linear separable data (which become the base of SVR), SVM formula should modified by construct a mapping into a high dimensional feature space. The input *x* is first mapped onto a *p*-dimensional feature space using some fixed (nonlinear) mapping, and then a linear model is constructed in this feature space (see figure 1).

Usually feature space have higher dimension from the input vector (input space), and this makes the computation in feature space become bigger,

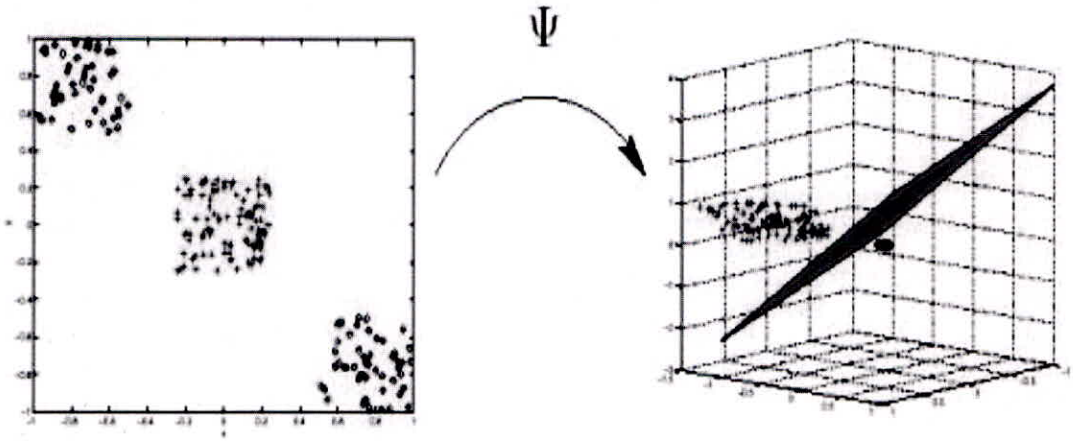

Fig. 1. Non-linear mapping of input examples into high dimensional feature space. (Classification case, however the same stands for regression as well)

because there is possibility that feature space has infinite number of feature. Besides that it is difficult to know the appropriate transformation function. To solve this problem, SVM uses *"kernel trick"*. Kernel Functions that are commonly used are as follows:

Linear Kernel: $K(x_i,x) = x^t,x$ (4)

Polynomial Kernel $K(x_i,x) = (\gamma' x_i^T x + r)^p, >0$ (5)

Radial Basis Function (RBF) $K(x_i,x) = \exp(-|x_i - x|^2) > 0$ (6)

Sigmoid Kernel $K(x_i,x) = \tanh(x + r)$ (7)

According to Tripathi [6], the RBF is computationally simpler than polynomial kernel, which has more parameters. Moreover, the advantage with RBF kernel is that it nonlinearly maps the training data into a possibly infinite dimensional space, thus it can effectively handle the situations when the relationship between predictors and predictand is non-linear.

Support Vector Regression (SVR)

SVM can be applied to regression problems by the introduction of an alternative loss function [7]. The loss function must be modified to include a distance measure. Figure 2 illustrates four possible loss functions. Figure 2(a) is the ϵ insensitive loss function that ensures existence of global minimum and at the same time optimization of reliable generalization bound. In figure 2(b) is the quadratic loss function which corresponds to the conventional least squares error criterion. Huber proposed the loss function shown in figure 2(c) that has optimal properties when the underlying distribution of the data is unknown. Figure 2(d) is a Laplacian loss function that is less sensitive to outliers than the quadratic loss function (Figure 2b).

SVR is based on the non-linear SVM that implicitly apply kernel functions which map the data to a higher dimensional feature space. A linear solution in the higher dimensional feature space corresponds to a non-linear solution in the original, lower dimensional input space. One method using the RBF and is called Least Square Support Vector Machine (LS-SVM) [7]. LS-SVM is computationally more efficient than the standard SVM method, since the training of LS-SVM requires only the solution of a set of linear equations instead of the long and computationally demanding quadratic programming problem involved in the standard SVM [8].

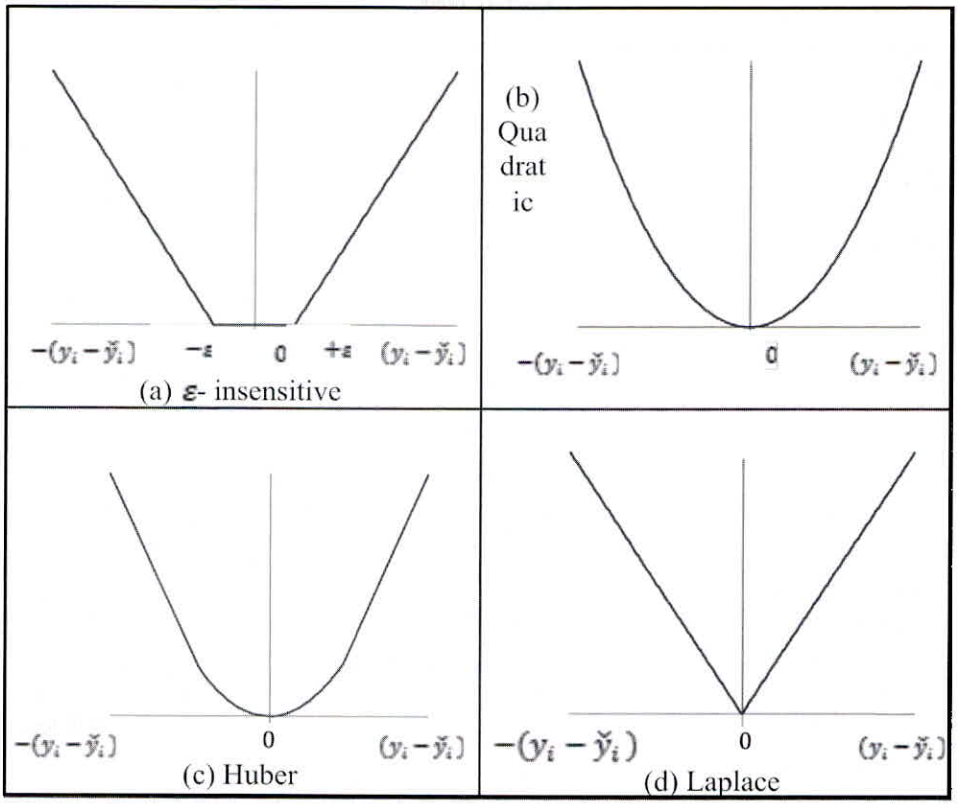

Fig. 2. Some Loss Function used in SVR

In SVR, $\{x_i, y_i\}_{i=1}^{N}$ is considered as a training set, in which $x_i \in \Re^p$ represents a p-dimensional input vector and y_i is a scalar measured output, which represents the system output. The goal is to construct a function $y = f(x)$ which represents the dependence of output y_i on input x_i. The form of this function is:

$$y = w^T \phi(x) + b \quad (8)$$

where w is known as the weight vector and b the bias. This regression model can be constructed using a nonlinear mapping function . By mapping the original input data into a high – dimensional space, the non-linear separable problem becomes linearly separable in space.

The function : is a mostly non-linear function which maps the data into a higher, possibly infinite, dimensional feature space. The LS-SVM involves equality constraints, and works with a least squares cost function. The optimization problem and the equality constraints are defined by the following equations:

$$\min \psi L(w,e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} e_i^2 \quad (9)$$

Subjected to equality constraint:

$$y_i - y = e_i, \; i = 1, \ldots, N \quad (10)$$

Or by substitution in equation (8):

$$y_i = w^T \phi(x_i) + b + e_i, \; i = 1, \ldots, N \quad (11)$$

Where is the quadratic loss term and is a regularization parameter in optimizing the trade-off between minimizing the training errors and minimizing the model's complexity. The objective is now to find the optimal parameters that minimize the prediction error of the regression model. The optimal model will be chosen by minimizing the cost function where the errors are minimized. This formulation corresponds to the regression in the feature space and since the dimension of the feature space is high, possibly infinite, this problem is difficult to solve. Therefore, to solve this optimization problem, the following Lagrange function is given:

$$\min_{w,b} L_p(w,b,e;\alpha) = \psi L(w,e) - \sum_{i=1}^{N} \alpha_i \{w^T \phi(x_i) + b + e_i - y_i\} \quad (12)$$

The solution of equation (12) can be obtained by partially differentiating with respect to w, b, e_i and $á_i$, i.e.

$$\frac{\partial L}{\partial w} = 0 \longrightarrow w = \sum_{i=1}^{N} \alpha_i \phi(x_i) \quad (13)$$

$$\frac{\partial L}{\partial b} = 0 \longrightarrow b = \sum_{i=1}^{N} \alpha_i = 0 \quad (14)$$

$$\frac{\partial L}{\partial e_i} = 0 \longrightarrow \alpha_i = \gamma . e_i,\ i = 1, \ldots . N \quad (15)$$

$$\frac{\partial L}{\partial x_i} = 0 \longrightarrow w^T \phi(x_i) + b + e_i - y_i = 0;\ i = 1, \ldots, N \quad (16)$$

Finally, the estimated values of b and á$_i$ i.e., And , can be obtained by solving the linear system and the resulting LS-SVM model can be expressed as:

$$y = f(x) = \sum_{i=1}^{N} \breve{\alpha}_i K(x, x_i) + \breve{b} \quad (17)$$

Where K(x,x$_i$) is a kernel function in the non-linear RBF (equation 6).

The regularization parameter is also necessary in LS-SVM model and determines the trade-off between the fitting error minimization and smoothness of the estimated function. It is not known beforehand which and are the best for a particular application problem to achieve the maximum performance with LS-SVM models. Thus, the regularization parameter and the value of from the kernel function have to be tuned during model calibration.

STUDY AREA

Roorkee is a city in the state of Uttarakhand. It is a part of the district of Haridwar and is located between the rivers Ganga and Yamuna; the Upper Ganga Canal flows through the city. Roorkee lies at 29 52 'N Latitude and 77 53 'E Longitude (see figure 3). It has an elevation of 274 meters above mean sea level. The climate of Roorkee is typical of Northwestern India. All three predominant seasons - summer, winter, and monsoon - are witnessed in Roorkee, with very hot summers and very cold winters. Being a submontanic district, with higher latitude than any other portion of the plains, it has longer spells of cold weather. Though the heat in May and June is considerable, relief is occasionally offered by the cooling effect of moderate Himalayan storms.

In terms of average annual precipitation (103.2 cm), Roorkee is semi-arid. The South-West monsoon generally breaks in mid-June and the North-East during November-December. Winters begin from October and continue through February. The coldest months are generally December and January, when the minimum temperature approaches zero. A rise in temperature is experienced from the beginning of March, which heralds the onset of summer. Temperature ranges from 0° C to 20° C in Winter (December to March), 25° C to 40° C in Summer (April to June) when warm winds blow frequently and 20° C to 40° C in Rainy season (June to September). !!

DATA REQUIRED

The data required is divided into two types, the predictand (Observed data) and the predictors (GCM data).

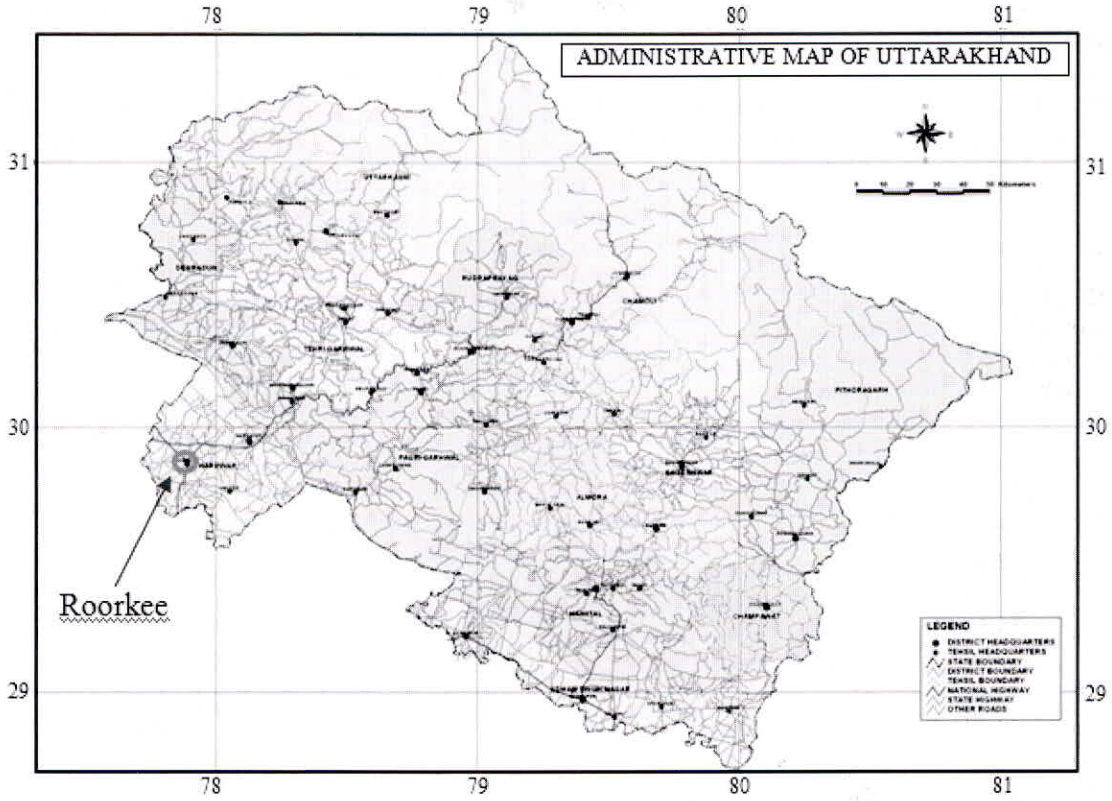

Fig. 3. Administrative Map of Uttarakhand

Predictand (Observed Data)

Observed data from January 1981- December 2010 is taken from Hydrology Department Station of IIT Roorkee. The observed data that will be used for downscaling calculation are:

- Mean monthly precipitation.
- Mean monthly minimum temperature.
- Mean monthly maximum temperature.

Predictors (GCM Data)

In this paper the mean monthly data which extends from January 1981 to December 2040 is extracted from Canadian Center for Climate Modeling and Analysis (CCCma) web site http:// www.cccma.bc.ec.gc.ca/. The simulated monthly climate data is taken from scenario IS92a of the second generation Coupled General Circulation Model (CGCM2). The extracted data pertains to 4 grid points whose latitude ranges from 27.83° N to 31.54° N and longitude ranges from 75° E to 78.75° E covering entire Roorkee Area. The CGCM2 grid is uniform along the longitude with grid box size of 3.75° and nearly uniform along the latitude (approximately 3.75°).

DEVELOPMENT OF DOWNSCALING MODEL

To develop the downscaling model, the available data set is partitioned into a training set and a test set. 50% of the available data from 1981-1995 (15 years) is selected for training (calibration) while

the remaining 50% from 1996 – 2010 (15 years) is used for testing (validation). The first step in developing downscaling model is selection of the predictors; the choice of predictor variables for downscaling is based on statistical analysis. The highest r (including negative value) for each type of predictors is chosen as the candidate predictors.

Selection of Predictors

The predictors were selected by computing the correlation between the observed and GCM data. Pearson product-moment correlation coefficient (PMCC, denoted by r) method was chosen. The r value can range between +1 and "1. The interpretation of r value according to Wang [9] is provided in this table 1. According to Tripathi [6] the choice of predictors could vary from one region to another. Since there are no general guidelines for selection of predictors, a comprehensive search is necessary. In general, the values of the climate variables at earth's surface (which corresponds to approximately 1000 mb), 850 mb, 500 mb and 200 mb pressure levels are found to be representative of circulation pattern in the study region [10]. Table 2 shows the correlation in various candidate predictors with observed data.

Table 1. The Interpretation of *r* value

Correlation Coefficient Value	Interpretation
0	No Relationship
Larger than 0 but smaller than 0.500	Weak Positive Relationship
From 0.500 to 0.699	Moderate Positive Relationship
From 0.700 to 0.999	Strong Positive Relationship
1	Perfect Positive Relationship
-1	Perfect Negative Relationship
From - 0.700 to -0.999	Strong Negative Relationship
From -0.500 to -0.699	Moderate Negative Relationship
Smaller than 0 but larger than -0.5000	Weak Negative Relationship

Table 2. The Correlation Coefficient (*r*) value between Observed and GCM data

GCM DATA	Correlation With		
	Observed Mean Monthly Precipitation	Observed Mean monthly MIN Temp	Observed Mean Monthly MAX Temp
200 Mb Temp	0.728	0.451	0.39
500 Mb Temp	0.58	0.853	0.686
850 Mb Temp	-0.125	-0.481	-0.517
200 Mb GPH	0.643	0.834	0.61
500 Mb GPH	0.425	0.807	0.714
850 Mb GPH	-0.714	-0.84	-0.608
200 Mb SpecHum	0.706	0.651	0.38
500 Mb SpecHum	0.651	0.766	0.564
850 Mb SpecHum	0.616	0.822	0.645
200 Mb U wind	-0.554	-0.569	-0.406
500 Mb U wind	-0.333	-0.437	-0.311
850 Mb U wind	0.376	0.273	0.131
200 Mb V wind	0.169	0.058	-0.075
500 Mb V wind	-0.333	-0.437	-0.311
850 Mb V wind	0.235	0.058	0.026
Rainfall	-0.116	-	-
Evaporation	-0.092	-	-

MLR Model

Multiple linear regression is a form of regression analysis in which the regression function establishes the relationship between one dependent variable y and more than one independent variables (x_1, x_2... x_n). A linear regression equation is in the following form:

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n \quad (18)$$

Parameters a (intercept) and b_1, b_2, ... b_n (coefficient of x_n) are estimated using the least squares method. In MLR method, the choice of predictor variables for downscaling is based on the statistical analysis. The highest r (including negative value) for each type of predictors is chosen as the candidate of predictors. The best statistic result from the combination predictors in calibration part is taken as the predictors in validation model.

SVR Model

Downscaling by SVR was carried out by using MATLAB LS-SVMLab toolbox Version 1.7. The data was divided into two periods and the steps of using the model in this study are given below:
Calibration period

1. Upload the data
2. Downscale calibration data
3. Determine the trial value of gam (gamma) and sig2 (sigma square).
4. Train the model to get the value of alpha and b (bias).
5. Using the values of alpha and b, get the simulated precipitation.
6. Check the errors and correlation.

Repeat Steps 3 to 6 until the value of gam and sig2 give the best model (smallest error and biggest correlation). Now, using the computed values of gam and sig2, downscale validation period data and check the errors and correlation.

The length of data series used for the calibration and validation is the same, 15 years. For the calibration, the data is from 1981 – 1995 and for the validation it is from 1996 -2010. SVR model from LS-SVMLab toolbox has two parameters γ(gam) and σ^2(sig2) to be determined. In this study, their near optimal values were obtained by a trial-and-error method. For both calibration and validation periods, the computed variable was compared to the observed variable and the error parameter is measured by using correlation coefficient (r), RMSE and NSE.

Performance Indices

Besides r value, others indices that were used to measure model performance are Nash-Sutcliffe Efficiency (NSE) and Root Mean Square Error (RMSE). NSE is defined as:

$$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - O_{Avg})^2} \quad (19)$$

RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \quad (20)$$

Where O_i is the observed value, P_i is predicted output, O_{avg} is the average of measured value and i equals the number of values and n is the number of data. For NSE, the closer the value is to 1, the more accurate the model is.

RESULTS AND ANALYSIS

Precipitation

From the result of various combination of predictors, 200Mb GCM Temp, 200Mb GPH and 200Mb SpecHum are chosen as the predictors for downscaling precipitation variable. The correlation coefficient between the observed and computed precipitation when MLR model was used was 0.714. NSE and RMSE were also measured in validation part and the values are 0.492 and 3.002 respectively. From the graph in figure 4, it is clearly seen that MLR cannot mimic the lower part of observed precipitation and also for the extreme precipitation.

For downscaling of precipitation by SVR model, the best combination for the prediction of precipitation is γ = 0.33 and σ^2 = 1.8 with the value of r = 0.747, RMSE = 2.838 and NSE = 0.546. Even though SVR can make 4.68 % improvement (see table 3) in correlation (r) and is able to reach lower part in some point, but the upper part or high values still cannot be well replicated by the model particularly the extreme precipitation. Figure 5 shows the graph between of observed precipitation and computed precipitation by SVR (SVR PPTn).

Minimum Temperature

From the results of various combination of predictors, the chosen predictors to downscale minimum temperatures are 500Mb Temp, 850Mb GPH, 850Mb SpecHum and 500Mb V wind. Downscaling of minimum temperature by MLR model results in r = 0.882, NSE = 0.766 and RMSE = 3.450. From the graph in figure 6 we can see that MLR overestimated almost all the upper part of observed temperature and underestimated lower part or small values.

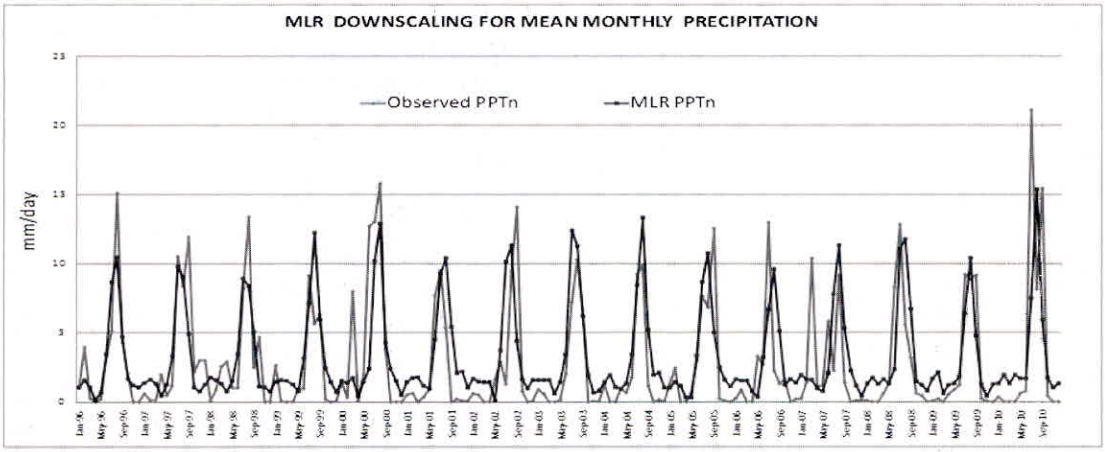

Fig. 4. MLR Graph for Precipitation

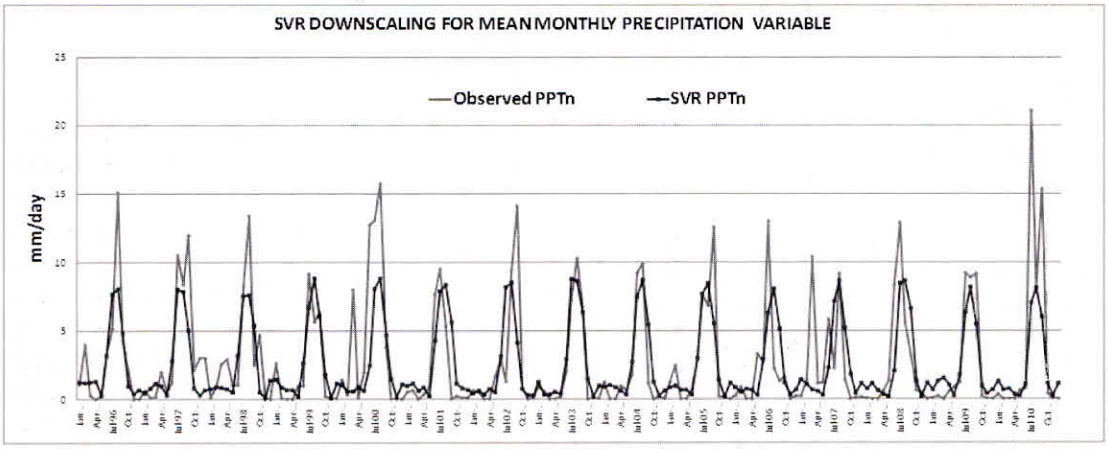

Fig. 5. SVR Graph for Precipitation

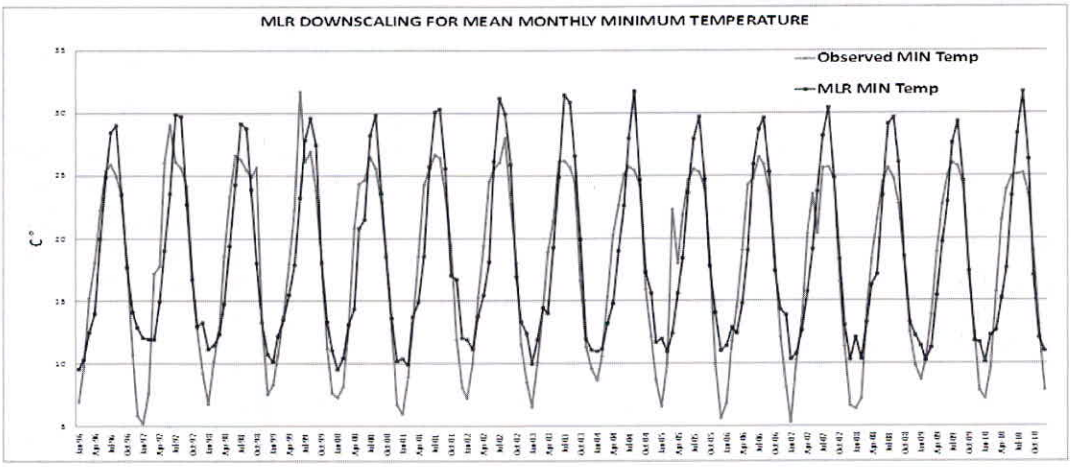

Fig. 6. MLR Graph for Minimum Temperature

The best combination for the predicted minimum temperature by SVR model is $\gamma = 0.45$ and $\sigma^2 = 2$ with the value of $r = 0.920$, RMSE = 2.880 and NSE = 0.837. Again if we compare the r value in table 4, then SVR makes slight improvement (4.331 %) over MLR. SVR model can reach the upper part of observed value however in some point but it was still difficult to predict the higher values. Figure 7 shows the graph between observed minimum temperature and computed minimum temperature by SVR (SVR MIN Temp). The results showing performance indices for downscaling minimum temperature are provide in table 4.

Table 4. The Performance Measures for Downscaling Minimum Temperature

Variables	r value			NSE			RMSE		
	MLR	SVR	% of improvement	MLR	SVR	% of improvement	MLR	SVR	% of improvement
Minimum Temperature	0.882	0.920	4.331	0.766	0.837	9.243	3.450	2.881	16.504

Maximum Temperature

From the result of various combination predictors, the chosen predictors for maximum temperatures downscaling are 500Mb Temp, 500Mb GPH, 850Mb SpecHum and 200Mb U wind.

The MLR for maximum temperature validation results $r = 0.732$, NSE = 0.529 and RMSE = 3.941. From figure 8, we can see that MLR cannot mimic the low values of observed temperature but better replicates the upper parts of maximum temperature. The best combination for the computed maximum temperature by SVR method is $\gamma = 0.9$ and $\sigma^2 = 0.2$ with the value is $r = 0.838$, RMSE = 3.171 and NSE = 0.695. From the comparison of r value in table 5, the SVR makes better prediction than MLR and the improvement is 14.44 %. Figure 9 shows the graph between of observed maximum temperature and computed maximum temperature by SVR (SVR MAX Temp). The model well replicated the upper part but still underestimated almost all lower part.

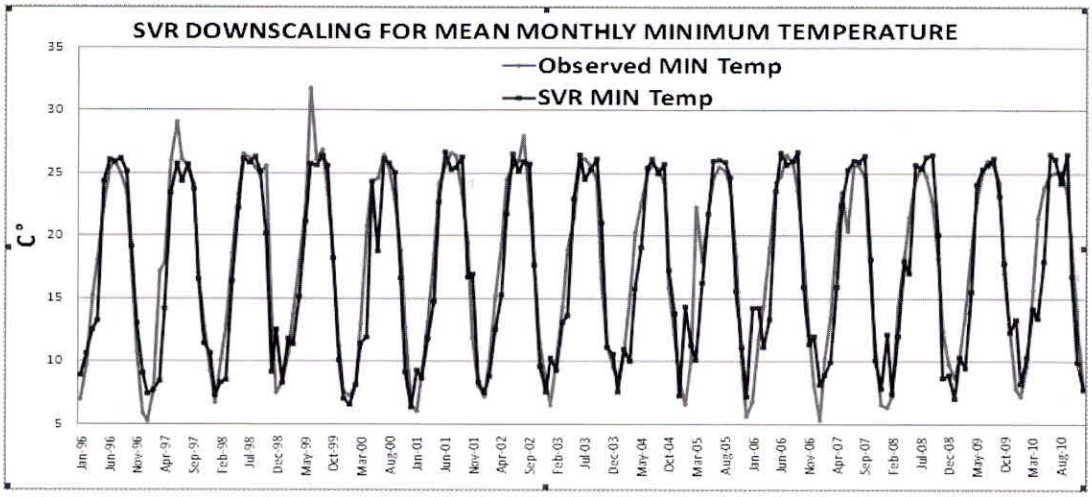

Fig. 7. SVR Graph for Minimum Temperature

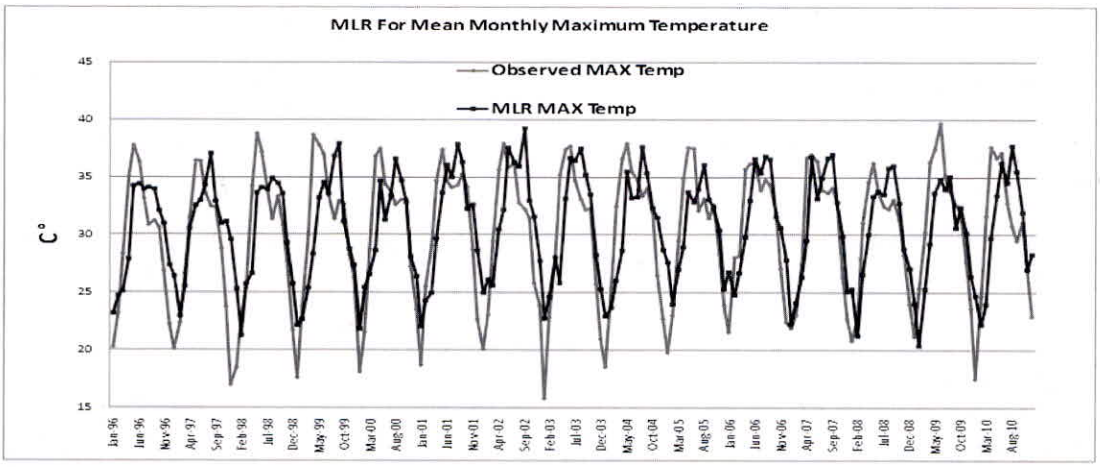

Fig. 8. MLR Graph for Maximum Temperature

Table 5. The Comparison Result of error Measurement for Maximum Temperature Downscaling

Variables	r value			NSE			RMSE		
	MLR	SVR	% of improvement	MLR	SVR	% of improvement	MLR	SVR	% of improvement
Maximum Temperature	0.732	0.838	14.440	0.529	0.695	31.437	3.941	3.171	19.541

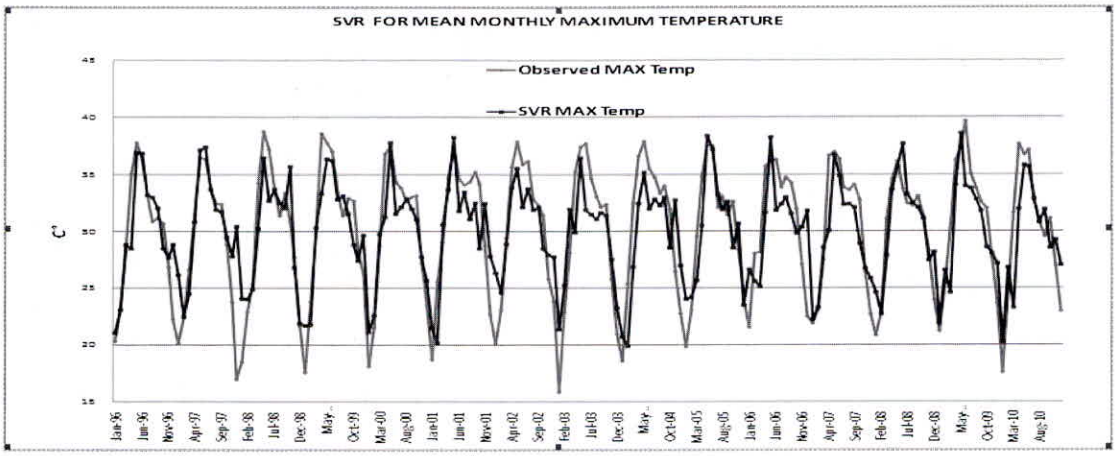

Fig. 9. SVR Graph for Maximum Temperature

FUTURE PROJECTIONS

To develop future projection, GCM data are divided into two groups with 15 years span for each group. The first group is from 2011-2025, second group from 2026-2040. For SVR model the data are computed by SVR validated model with the same value of gamma and sigma for each variable. For MLR model the data are computed by MLR validated model with the same formula for each variable.

Precipitation Projection

Descriptive statistic in table 6 shows that mean value of SVR projection for precipitation is between 2.5 - 2.9 mm/day and 2.4 -3.2 mm/day for MLR model. From total amount of precipitation per year, one can see that for SVR model there will be increase of precipitation approximately by 0.5 – 2%. This is in accordance with Tritpathi [6] which predicted that precipitation will increase in North India (Punjab, Haryana and Uttar Pradesh), while MLR predicts mixed trend in precipitation.

Table 6. Descriptive Statistic of Precipitation Projection

Descriptive statistic of precipitation (mm/day)	Observed 1996-2010	1996-2010		2011-2025		2026-2040	
		MLR	SVR	MLR	SVR	MLR	SVR
Mean	2.78	3.25	2.52	2.41	2.88	2.96	2.92
Standard Error	0.21	0.25	0.21	0.26	0.19	0.26	0.22
Median	1.01	1.61	1.09	0.82	1.39	1.30	1.35
Standard Deviation	4.05	3.39	2.83	3.54	2.60	3.54	2.94
Range	21.14	15.26	8.71	15.57	7.87	14.81	8.84
Minimum	0.00	0.11	0.12	-0.94	0.82	-0.32	0.60
Maximum	21.14	15.36	8.83	14.63	8.69	14.49	9.44
Sum (mm/year)	1048.32	1191.85	922.63	888.18	1054.52	1084.98	1070.174

Figure 9(a) shows that MLR projections contain some negative values and the projection has more variation in the values. The maximum value can reach 14 mm/day and the trend line is flat. The MLR seems to underestimate the lower part. Figure 9(b) shows less variation in SVR results. The maximum value is always below 10 mm/day. Even though there are no negative values in projection but still SVR cannot well compute the lowest amount of precipitation (zero value). The trend line make small positive slope. The SVR overestimate the lower part.

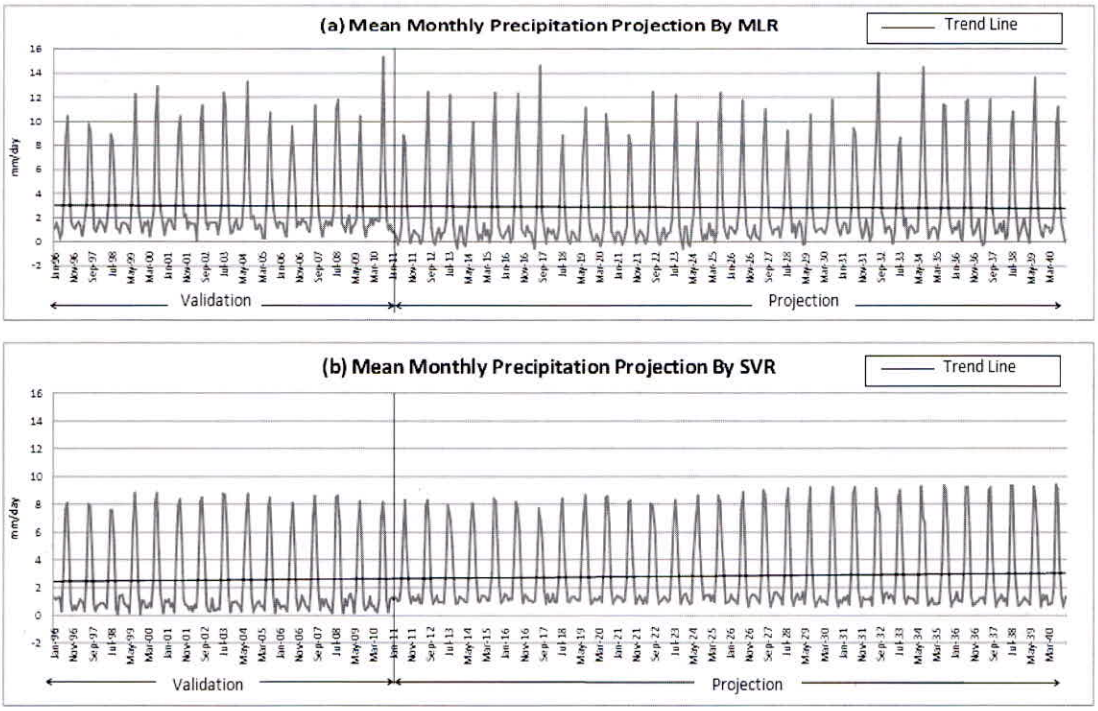

Fig. 9. Graphical depiction of projected Precipitation. a) Using MLR, b) Using SVR

Minimum Temperature Projection

From the projection of mean monthly minimum temperature by the SVR model, one can see that there will not be much change in temperature. The mean value of SVR in table 7 is between 16.9 - 17.3°C and for MLR, the mean value increases from 18.3°C to 18.8°C. The maximum value of MLR is approximately 31°C (nearby observed value). The minimum value is 9°C, 4°C higher than observed value. SVR has the minimum value 5-6°C (nearby the observed value). For the maximum value, the highest prediction from SVR is only 27.68°C, 4°C lower than observed.

Maximum Temperature Projection

From the projection of mean monthly maximum temperature by SVR model, the mean value of the maximum temperature is between 29.8 – 29.9 °C (see table 8) while for the MLR model, the mean value is between 28 – 31°C. The SVR predicts that the highest value of mean monthly maximum temperature until 2040 will be 39.25°C, while MLR predicts that it will reach 42.44°C.

Table 7. Descriptive Statistic of Minimum Temperature Projection

Descriptive Statistic of minimum temperature C°	Observed 1996-2010	1996-2010		2011-2025		2026-2040	
		MLR	SVR	MLR	SVR	MLR	SVR
Mean	17.75	18.31	17.29	18.50	16.92	18.89	17.32
Standard Error	0.53	0.51	0.53	0.51	0.57	0.53	0.56
Median	18.49	16.70	16.48	16.52	15.81	16.78	16.45
Standard Deviation	7.15	6.81	7.12	6.82	7.63	7.06	7.49
Range	26.55	22.24	20.46	22.31	21.56	23.01	21.24
Minimum	5.18	9.51	6.29	9.41	5.87	9.47	6.44
Maximum	31.72	31.75	26.74	31.71	27.43	32.47	27.68

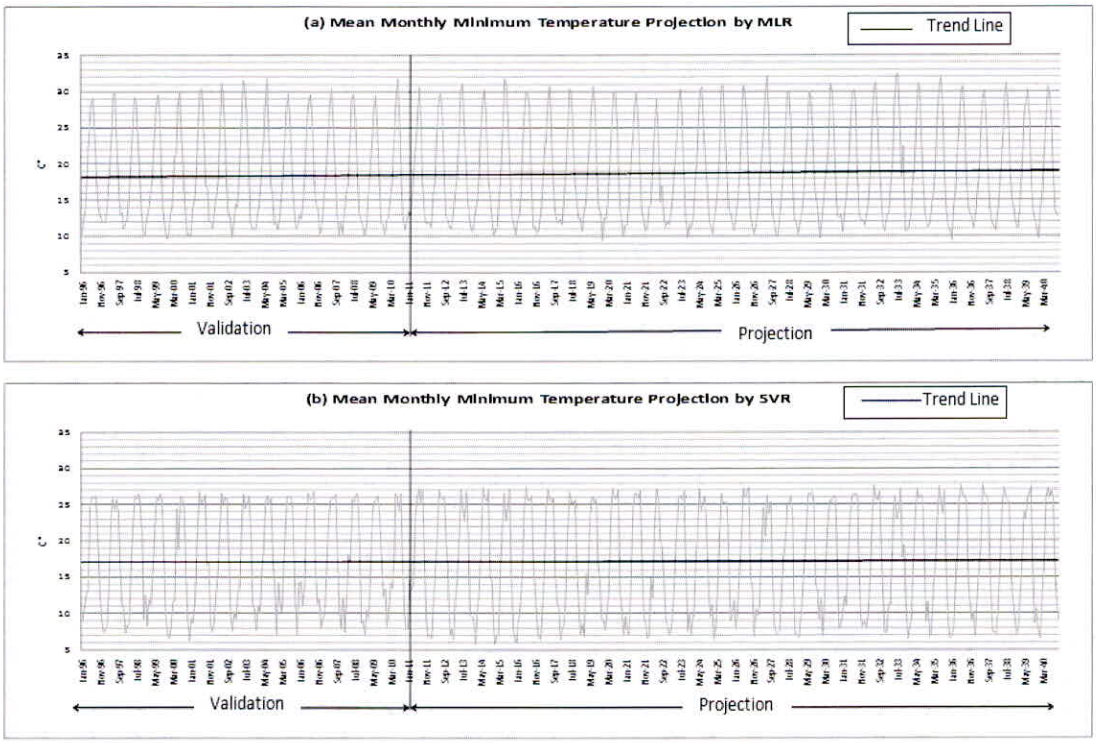

Fig. 10. Graphical depiction of projected minimum temperature. a) Using MLR, b) Using SVR

Fig. 10(a) shows the projections for minimum temperature by MLR. The highest value in projection part reaches 32° and the lowest is around 7°C. The trend line increase 1°C until 2040. Figure 10(b) shows that SVR have more variation in the upper part. SVR projection shows not much change and the trend line is flat.

Table 8. Descriptive Statistic of Maximum Temperature Projection

Descriptive statistic of mean monthly maximum temperature °C	observed 1996-2010	1996-2010		2011-2025		2026-2040	
		MLR	SVR	MLR	SVR	MLR	SVR
Mean	30.20	28.24	29.95	31.22	29.81	32.01	29.96
Standard Error	0.43	0.50	0.33	0.34	0.40	0.36	0.42
Median	32.00	29.00	31.05	31.58	31.60	32.42	31.35
Standard Deviation	5.76	6.75	4.44	4.59	5.30	4.84	5.61
Range	23.86	23.67	18.74	20.64	19.59	20.10	20.14
Minimum	15.82	15.58	19.84	21.09	19.43	22.35	19.12
Maximum	39.68	39.25	38.58	41.72	39.02	42.44	39.25

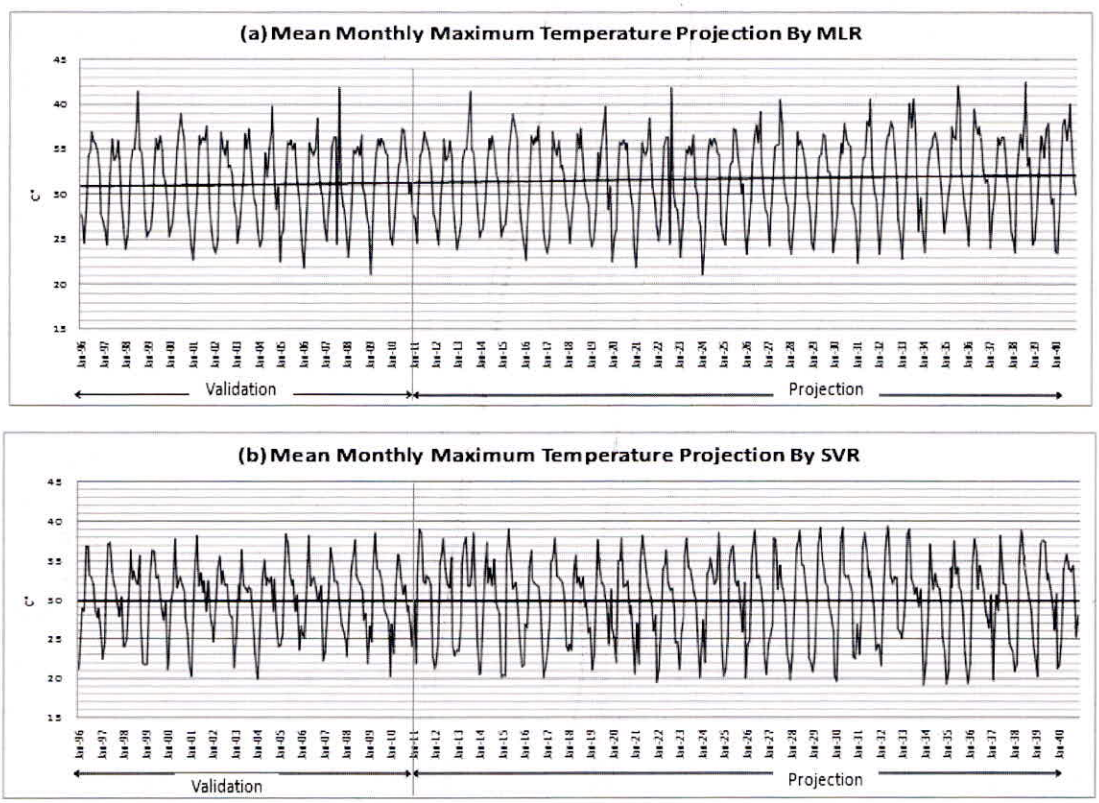

Fig. 11. Graphical depiction of projected maximum temperature. a) Using MLR, b) Using SVR

CONCLUSIONS

This work has downscaled climate variables in Roorkee area: mean monthly precipitation, minimum and maximum temperature by using SVR and MLR methods. Based on the results, the following conclusions can be drawn:

- The best combination of predictors for downscaling precipitation for Roorkee area among the available variables are temperature, geopotential height, and specific humidity at 200 mb (this means that the values refer to approximately 12,000 m height where the cumulonimbus clouds are formed). This finding is in accordance with Gadgil [11] that most of the rain over the Indian region comes from Cumulus and Cumulonimbus clouds.

- V (Vertical) or Meridional wind influences the computations when downscaling the minimum temperatures. This probably happens because in the summer, south-west monsoon brings heavy rain between July and September in Roorkee area. In the winter, north-east monsoon sweeps down from the plateaus of Asia and the Himalayas and brings rain and cooler weather between October and December.

- U (horizontal) or Zonal wind influences the computation of the maximum temperature. This probably happens because zonal wind flows in west–east direction and brings strong, hot "loo" and dry summer wind from the large desert regions of the northwestern Indian subcontinent [12].

- The result of downscaling for precipitation shows that SVR performs better than the MLR as seen by the improvement of error measurements which are 4.678 % for r, 10.931 % for NSE and 5.447 % for RMSE. However, it can be seen that both MLR and SVR could not well downscale precipitation in Roorkee Area. These results have been obtained possibly because regression based statistical downscaling model often cannot explain entire variance of the downscaled variable [3]. The other reason could be that, by nature, precipitation is much more erratic and dependant on very local factors [13]. Also, for precipitation the spatial variation is very large and it has very poor temporal correlation. Downscaling of precipitation is a challenge and more studies are needed.

- The result of SVR downscaling for minimum temperature shows a 4.33 % improvement in r, 9.24 % in NSE and 16.50 % in RMSE as compared to MLR.

- The result of SVR downscaling for maximum temperature shows a 14.44 % improvement in r, 31.44 % in NSE and 19.54 % in RMSE as compared to MLR.

- The results of downscaling show better improvement in the maximum temperature when the SVR model is used rather than the minimum temperature.

- It can be concluded that use of SVR can improve correlation in the range 4 – 14 % than MLR, for NSE by 9 - 31 % and for RMSE by 5 – 19 %.

- Future projection until 2040 for precipitation by SVR model shows that there will be little increase of precipitation and the future projection for temperature shows that there will not be much change in temperature in Roorkee area.

- More research is needed to confirm these conclusions.

REFERENCES

IPCC Technical Paper VI. Climate Change and Water, 2008.

Downie David.L, Brash Kate,Vaughan Kate. Climate Change: a reference handbook, 2009.

R.L Wilby, SP Charles, E Zorita, B Timbal, P Whetton, LO Mearns. Guidelines for Use of Climate Scenarios Developed From Statistical Methods, 2004.

Vapnik, V. The Nature of Statistical Learning Theory.Springer-Verlag. New York, 1995.

Gunn Steve.R. Support Vector Machines for Classification and Regression, 1998.

Shivam Tripathi, V.V Srinivas, Ravi S.Nanjundiah. Downscaling of Precipitation for climate change scenario: A support Vector Machine Approach, 2006.

Smola A. J. Regression estimation with support vector learning machines. Master's thesis, Technische Universit¨at M¨unchen, 1996.

Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Processing Letters, 1999.

Wang Xiao Hu. Financial Management in the Public Sector, ME Sharpe New York, USA, 2006.

Maity.P, Bhagwat et al. Potential of support vector regression for prediction of monthly streamflow using endogenous property, 2010.

Gadgil Sulochana. The Indian Monsoon, Part 1. Variations in Space and Time, Resonance, Vol.11, No.8, 2006.

Rana S.V.S. Essentials of Ecology and Environmental Science, Prentice Hall of India, ISBN 8120333004, 2007.

Berastegi G. Ibarra, Saenz.J, Ezcurra.A, Elias.A, Argandona.J, Errasati. Downscaling of surface moisture flux and precipitation in the Ebro Valley (spain) using analogues followed by random forests and multiple linear regression, 2011.