Journal of Civil Engineering and Environmental Technology
Print ISSN: 2349-8404; Online ISSN: 2349-879X; Volume 1, Number 5; August, 2014 pp. 42 – 50

http://www.krishisanskriti.org/jceet.html

Downscaling Monthly Rainfall Using Support Vector Regression and M5P Model Tree

Karan Singh[1], **Parveen Sihag**[2], **Ritesh Kumar**[3]

[1]*M. Tech Scholar*
[2]*Ph.D Scholar*
[1, 2]*Department of Civil Engineering*
N.I.T. Kurukshetra, Haryana
[3]*Department of Civil Engineering*
THDCIHET, TehriGharwal

Abstract **: General Circulation Models are Hybrid Mathematical Model used for Climate Change Impact Studies. It works on a coarser scale and downscaling is necessary for climate change impact at regional scale. Monthly Rainfall in Haryana is downscaled in this study using Machine Learning Techniques: Support Vector Regression and Data Driven Method: M5P Model Tree. In this study the future monthly rainfall in Haryana is downscaled using outputs of CGCM3 for A2 emission scenario. The data for the base period of 30 years (1971-2000) and for the future period of 100 years (2001-2100) has been employed in downscaling. Gridded data set of rainfall from National Climate Centre, IMD Pune has been taken as observed and a relationship is generated between predictand and predictors of NCEP/NCAR Reanalysis 1and the generated relationship are then used for downscaling rainfall for future period. Models are generated using different parameters and best model of SVR and M5P are compared. M5P Model has performed better than SVR in terms of statistical performance parameters such as correlation coefficient and root mean square error.**

Keywords: GCM, Climate Change, SVR, M5P Model Tree, CGCM3, NCEP/NCAR, Haryana, IMD.

INTRODUCTION

The change in the global climate has been observed recent years due to increasing green house gases (GHGs) in the atmosphere. According to WMO, rainfall is an essential atmospheric climate variable. It is necessary to study the pattern of rainfall and its variability for assessing the impact of climate change on various surface processes i.e. hydrology, agriculture, forestry, water resources management (Anandhi et al, 2008, 2009).

The GCM models are very advance mathematical model which are widely used for Climate change Impact studies. A general circulation model (also known as a global climate model, both labels are abbreviated as GCM) uses the same equations of motion as a numerical weather prediction (NWP) model, but the purpose is to numerically simulate changes in climate as a result of slow changes in some boundary conditions (such as the solar constant) or physical parameters (such as the greenhouse gas concentration). (*B. Geerts and E. Linacre)* (*http://www.das.uwyo.edu/~geerts/cwx/notes/chap12/nwp_gcm.html*)

There are numbers of GCMs are available for different emission scenario. In this study Canadian coupled global climate model (CGCM3) has been used. It works on a coarser scale with a 3.75 degree grid cell size for atmospheric horizontal resolution. Hence downscaling is necessary to obtain local-scale surface variable from global-scale atmospheric variables that are provided by GCMs

DOWNSCALING

There are two types of downscaling techniques i.e. dynamic downscaling and statistical downscaling. Dynamical downscaling involves the nesting of a higher resolution Regional Climate Model (RCM) within a coarser resolution GCM. The RCM uses the GCM to define time–varying atmospheric boundary conditions around a finite domain, within which the physical dynamics of the atmosphere are modelled using horizontal grid spacing of 20–50 km. The main limitation of RCMs is that they are as computationally demanding as GCMs (placing constraints on the feasible domain size, number of experiments and duration of simulations) (Wilby and Wigley, 1997). In statistical downscaling technique a relationship is generated between

local scale surface variable (predictand) and large scale atmospheric variables (predictors). There are three types of statistical downscaling namely regression methods, weather pattern-based approaches, stochastic weather generators (Wilby and Wigley, 1997). Among these approaches regression methods are preferred because of its ease of implementation and low computation requirements.

STUDY AREA

The state of Haryana is situated in the northern part of India and It is surrounded by Uttar Pradesh (UP) on the east, Punjab on the west, Uttaranchal, Himachal Pradesh &Shivalik Hills on the north and Delhi, Rajasthan and Aravalli Hills on the south. The river yamuna defines its eastern border with Uttarakhand and Uttar Pradesh. Haryana is a landlocked state in northern India. It is located between 27°39' to 30°35' N latitude and between 74°28' and 77°36' E longitude. The altitude of Haryana varies between 700 to 3600 ft (200 metres to 1200 metres) above sea level.

The climate of Haryana is similar to other states of India lying in the northern plains. It is very hot in the summer and markedly cold in winter; maximum temperatures in May and June may exceed 110 °F (43 °C), and in January, the coldest month, low temperatures may drop below the freezing point. Rainfall is varied, with the Shivalik Hills region being the wettest and the Aravali Hills region being the driest. About 80% of the rainfall occurs in the monsoon season (July–September) and sometimes causes local flooding.

METHODOLOGY

The study carried out for 9 grid points for rainfall analysis. The rainfall is downscaled using support vector regression and M5P Model Tree. This section outlines the procedures.

DATA EXTRACTION

NCEP/NCAR Re-analysis 1 data are extracted from http://www.esrl.noaa.govfor the grid points whose latitude ranges from 27.5 to 32.5 N and longitude ranges from 72.5 to 77.5 E. The climate variables extracted are mean sea level pressure, specific humidity at different pressure level, air temperature, zonal wind velocity and meridional velocity at different pressure level, geo-potential height for the period of 1971-2000 (30 years).

For observed data, IMD gridded Rainfall at (0.5° x 0.5°) spatial resolution from 1969-2005 (37 years) are collected from National Climate Centre IMD, Pune. 30 years data has been extracted from 1971 to 2000 and daily data are converted into monthly data.

GCM outputs of Canadian centre for environmental prediction model CGCM3.1 are obtained from website http://www.cccma.ec.gc.ca for the base period of 1971-2000(30 years) and for the future period of 2001-2100 (100 years) under A2 scenario.

The data extracted from GCM are re-gridded to NCEP/NCAR grid. Re-gridding is often needed because the grid spacing or co-ordinate system is GCM do not correspond to the grid-spacing and co-ordinate system of the re-analysis data set (NCEP/NCAR).For example the NCEP/NCAR re-analysis 1 has a grid spacing of 2.5 latitude by 2.5 longitude whereas the CGCM3.1 model has a coarser resolutions of 3.75 latitude by 3.75 longitude. The re-gridding is done using linear/bilinear interpolation.

Support Vector Regression

Support vector regression aims to find a function $f(\bar{x}) = \bar{w}.\bar{x}+b$, that approximates target values $(y_1, y_2, \ldots\ldots\ldots y_N)$ given input data $x_1, x_2, \ldots\ldots x \in R^n$

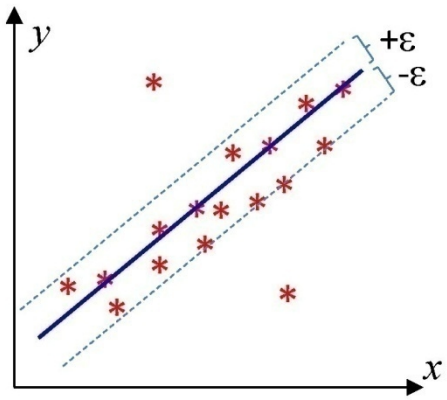


Fig. 2.ε-Support Vector Regression

The main idea of SVR is to find a function that has at most ε deviation from the true value $y_i$and it is as flat as possible (to avoid over fitting). The different between y_i and fitted function should be smaller than $+\varepsilon$ and larger than $-\varepsilon$

The flatness of the function can be ensured by smaller value of w which can be achieved by minimizing the $\frac{1}{2}\|w\|^2$

Subjected to $\begin{cases} y_i - (w_i.x_i) - b \leq \varepsilon \\ (w_i.x_i) + b - y_i \leq \varepsilon \end{cases}$

It is a convex quadratic programming (QP) optimization problem.

If the data has some outliners or noise then slack variable $\xi\xi^*$ are assigned. Now to find a function the optimization problem can be written as

$$minimize \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{k}(\xi + \xi^*)$$

Subjected to

$$\begin{cases} y_i - (w_i . x_i) - b \le \varepsilon + \xi \\ (w_i . x_i) + b - y_i \le \varepsilon + \xi^* \end{cases}$$

$$\xi\xi^* \ge 0$$
$$for i = 1, \ldots . k$$

Constant C, which influences a trade-off between an approximation error and the weight vector norm$\|w\|$, is a design parameter chosen by the user.
The performance of the SVR model depends on the parameter C, types of kernel, kernel parameters. Kernel tricks are used for transforming non linear data in original dimension input space into linear separable data in higher dimensional feature space. Three types of kernel used
(i) Linear $k(xy) = (x'^*y)$
(ii) Polynomial $k(xy) = (gamma^*x'^*y + coef0)\text{^}degree$
(iii) Gaussian RBF $k(xy) = \exp(-gamma^*|x - y|\text{^}2$

The above problem can be solved by Lagrange multiplier in a dual space

$$f(x) = \sum_{i=1}^{k}(\alpha_i - \alpha_i^*)k(x_i x) + b$$

Where $\alpha_i \& \alpha_i^*$ are the Lagrangian multiplier.

The optimal desired weight vector of the regression hyperplane can be found as

$$w = \sum_{i=1}^{k}(\alpha_i - \alpha_i^*)\emptyset(x_i)$$

In this study, the support vector regression is carried out using a Data Mining Software WEKA v3.7.9

M5P MODEL TREE

It is important for any model construction, that the predicted value should be as close as possible to the actual output. The errors should be minimized. The reliability of the model depends on how it predicts the target value of unseen data with accuracy.

Model Tree is a data driven method, in which a complex problem can be solved by dividing it into a number of simple problems and combining the solutions of these problems. (R. Arunkumar et al, 2013)

The splitting in the M5 model tree approach follows the idea of a decision tree, but instead of the class labels, it has linear regression functions at the leaves, which can predict continuous numerical attributes. (K ksingh et al, 2009)

M5P is a reconstruction of Quinlan's M5 algorithm for inducing trees of regression models. M5P combines a conventional decision tree with the possibility of linear regression functions at the nodes. A decision-tree induction algorithm is used to build a tree, but instead of maximizing the information gain at each inner node, a splitting criterion is used that minimizes the intra-subset variation in the class values down each branch. The splitting procedure in M5P stops if the class values of all instances that reach a node vary very slightly, or only a few instances remain.The tree is pruned back from each leaf when pruning an inner node is turned into a leaf with a regression plane. (www.opentox.org, steven Kramer)

The first step in building a model tree is to compute the standard deviation of the target value of cases in T. Unless T containes very few cases or their values vary only slightly, T is split on the outcomes of a test. Every potential test is evaluated by determining the subset of cases associated with each outcome; let T_i denote the subset of cases that have the i^{th} outcome of the potential test. If we treat the standard deviation $sd(T_i)$ of the target values of cases in T_i as a measure of error, the expected reduction in error as a result of this test can be written as

$$\Delta error = sd(T) - \sum \frac{|T_i|}{|T|} \times sd(T_i)$$

After examining all possible tests, M5 choose one that maximises this expected error reduction. (Quinlan, 2006)
The study by (jyotiprakash&kote, 2011a) shows robustness of unpruned and unsmoothed model tree in hydrological studies because pruned and smoothing cuts the peak & trough processes.

PERFORMANCE EVALUATION

The performance of the best model of SVR and M5P model tree was evaluated using statistical performance evaluation measures such as correlation coefficient (R-value) and root mean square error (RMSE).

The correlation coefficient (R-Value) is a measure of the linear regression between the predicted and the target of models

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The root mean square error (RMSE) is a measure of the difference between values predicted by model and the actual observed values.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}}$$

The model is better when RMSE is smaller and R value is high.

SELECTION OF PREDICTORS

The most important steps in a downscaling is the selection of appropriate predictors. The variable should be selected such that it should exhibit a relationship with the predictand. In climate impact studies such predictors should be chosen that are (1) reliably simulated by GCM outputs and reanalysis data, (2) strongly correlated with the predictand (3), based on previous studies. (A. Anandhi et al, 2008)

For the rainfall analysis, the potential predictors selected are shum, shum850, shum500, Ua500, Va_ns and Va925 having correlation coefficient above 0.6 with the pridictand. Total data for 30 years of each grid are divided into two parts, 70% of the data was used for training the model and 30% of the remaining data was used for the testing of trained model in the WEKA software.

RESULTS

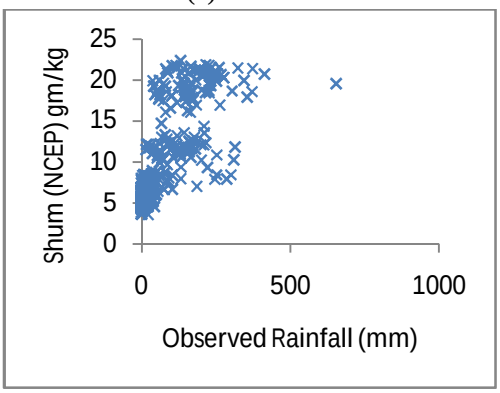
After the selection of probable predictors the scatter plots between the observed rainfall and NCEP climate variables are prepared to see the relationship between predictand and predictors.Total 9 grids grid points are selected for analysis as given in **Table 1.** Due to the large number of plots and graphs, the results are shown in this paper only for a particular grid.

Table 1. Grid Location of IMD Data

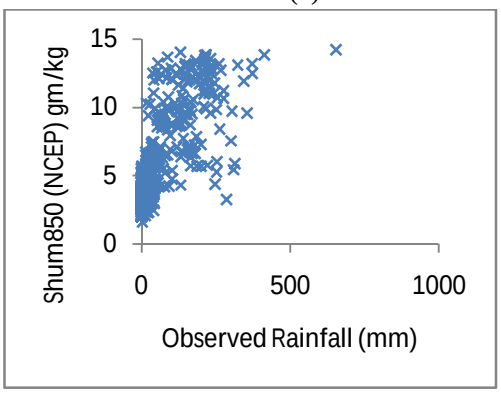
Grid No	Latitude	Longitude
1	27°30'N	77°00'E
2	28°00'N	77°00'E
3	28°30'N	77°00'E
4	29°00'N	77°00'E
5	29°30'N	77°00'E
6	27°30'N	77°30'E
7	28°00'N	77°30'E
8	28°30'N	77°30'E
9	30°00'N	77°30'E

The scatter plots from **Fig. 3** (a) to (d) shows that the rainfall varies nonlinearly with the predictors of NCEP i.e. specific humidty, zonal velocity &meridional velocity. The scatter plots also prepared among NCEP and GCM variable to verify whether predictors are realistically simulated by the GCM as shown in **Fig. 4** (a) to (d). The performance of the model were analysed by statistical performance parameters i.e. root mean square error (RMSE) and correlation coefficient (R) are given in **Table 2**.

(a)



(b)



(c) (d)

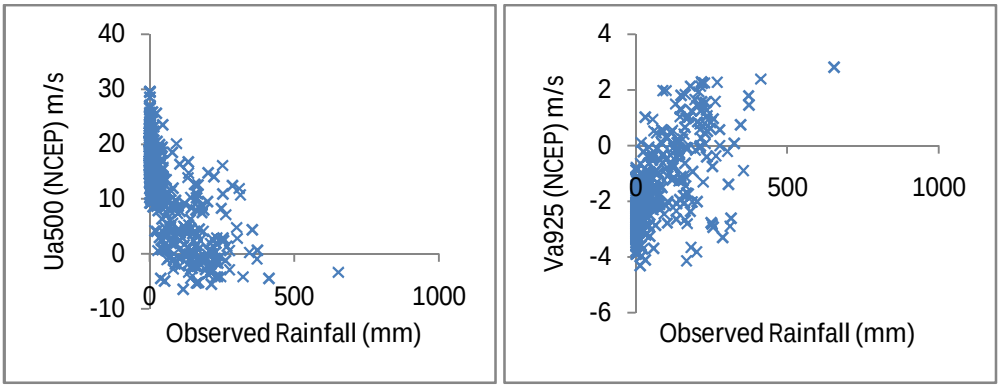


Fig. 3. Scatter Plot between NCEP predictors and Predictand

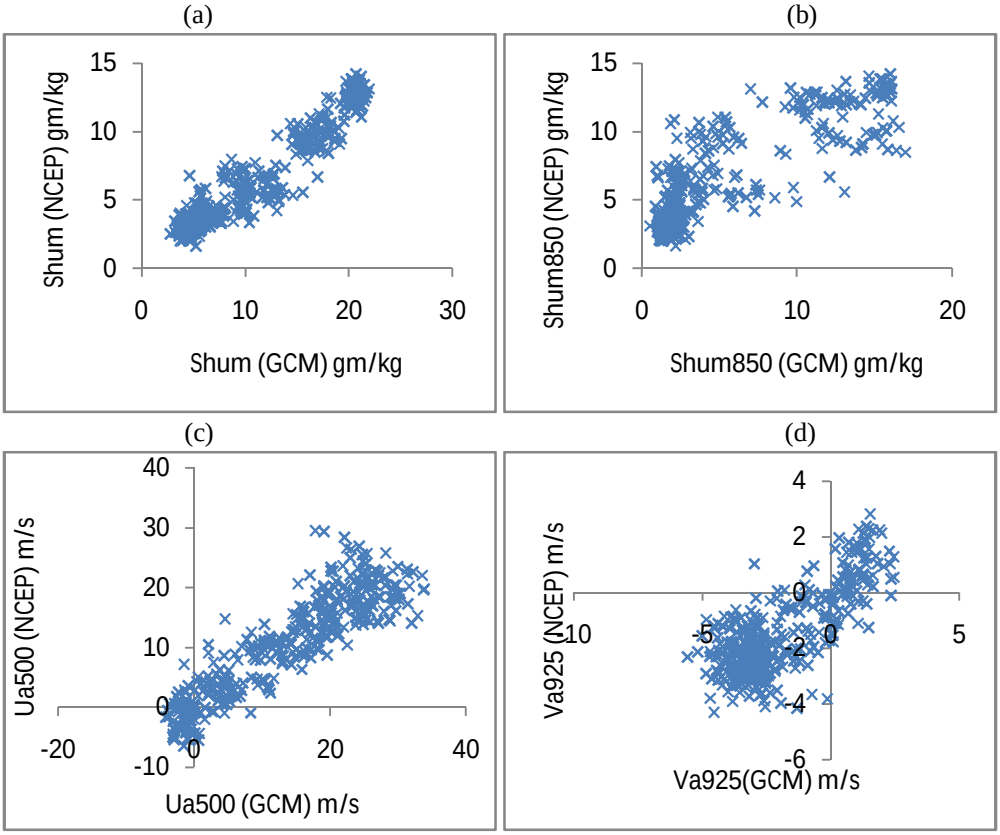


Fig. 4. Scatter Plot between GCM and NCEP predictors

Table 2. Performance of SVR and M5P Models

Grid	SVR				M5P Model Tree			
	Training		Testing		Training		Testing	
	RMSE	R	RMSE	R	RMSE	R	RMSE	R
1	56.01	0.77	67.71	0.67	49.97	0.83	66.94	0.67
2	62.69	0.78	59.29	0.76	52.71	0.85	60.00	0.74
3	61.30	0.79	55.21	0.79	61.16	0.79	54.27	0.80
4	80.25	0.71	67.63	0.79	76.27	0.75	66.72	0.80
5	89.35	0.70	99.92	0.69	79.64	0.78	104.63	0.66
6	67.90	0.79	45.91	0.81	63.14	0.80	47.32	0.81
7	44.63	0.83	46.72	0.79	44.67	0.82	48.92	0.76
8	58.15	0.74	57.98	0.73	56.44	0.75	57.69	0.72
9	57.63	0.85	56.55	0.82	57.10	0.85	53.42	0.84

The M5P model was tried for pruned and unpruned tree. The unpruned model tree showed better results, hence the best model is selected ieunpruned models tree for the comparison with the best SVR model. The final model parameters of SVR and M5P model in WEKA are given in **Table 3**.

Table 3. Model Parameters of SVR and M5P

S.No.	SVR Parameters	M5P Parameters
1	RBF Regressor, numfunction = 3	Unpruned
2	e-SVR, cost = 120, RBF Kernel, gamma = 0.007	Unpruned
3	RBF Regressor, numfunction = 4	Unpruned
4	RBF Regressor, numfunction = 4	Unpruned
5	RBF Regressor, numfunction = 2	Unpruned
6	e-SVR, cost = 55, RBF Kernel, gamma = 0.01	Unpruned
7	RBF Regressor, numfunction = 3	Unpruned
8	nu-SVR, cost = 60, RBF Kernel, gamma = 0.01	Unpruned
9	RBF Regressor, numfunction = 5	Unpruned

The model which gives higher correlation coefficient and least root mean square error is considered. The generalised model is also considered rather than over-fitting or under-fitting model which performs equally for the training as well as test data.

Observed and simulated rainfall for the base period (1971-2000) has shown in **Fig. 5.**and**Fig. 6.**for a particular grid. It has been observed from the figure that model fails to capture extreme events.

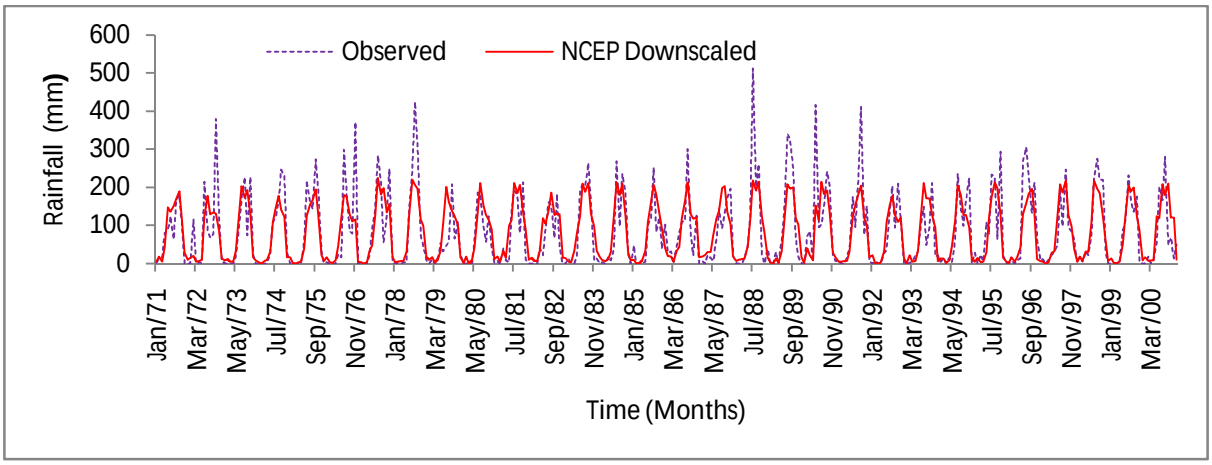


Fig. 5. Observed and NCEP Downscaled monthly rainfall for base period (1971-2001)

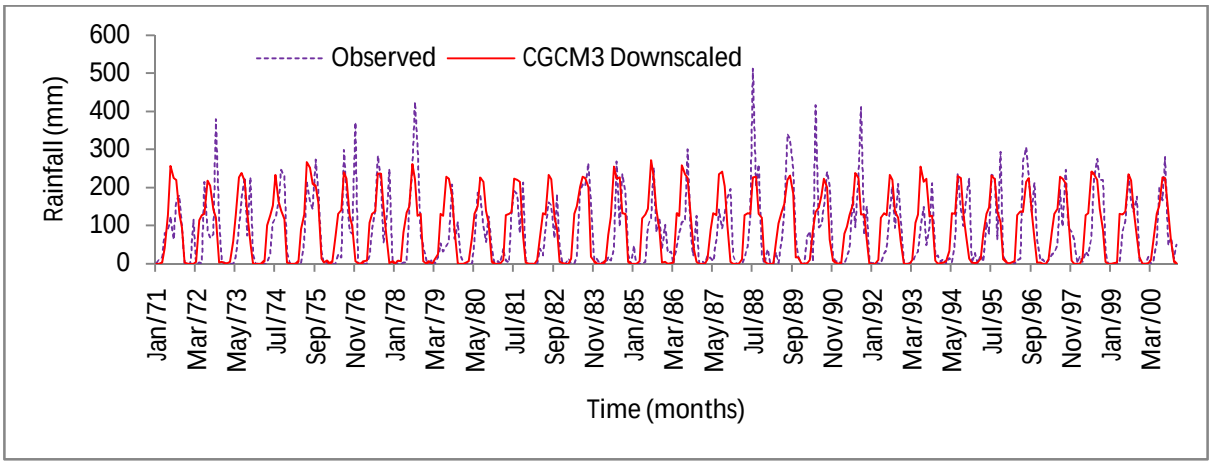


Fig. 6. Observed and CGCM3 Downscaled monthly rainfall for base period (1971-2001)

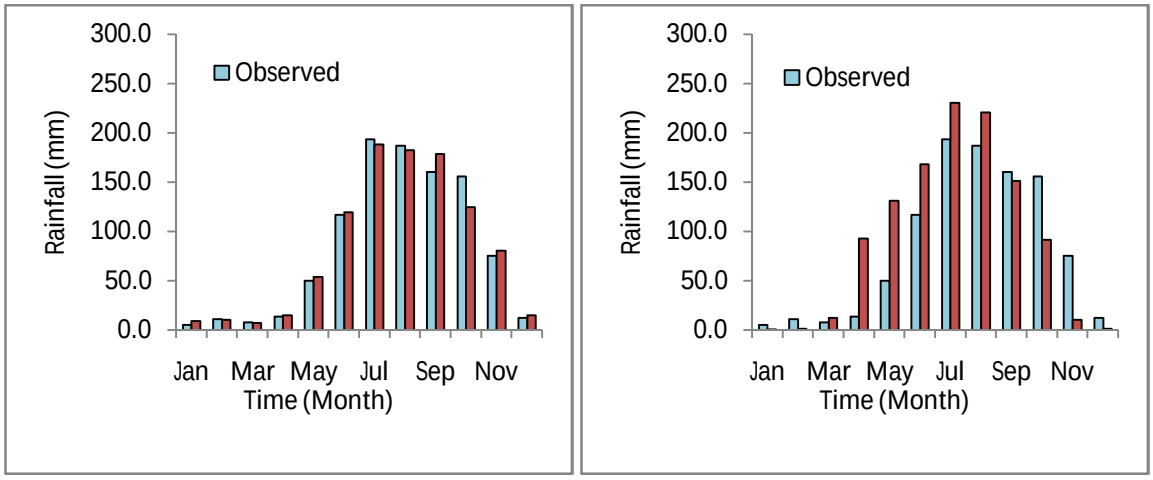


Fig. 7. Observed and NCEP Downscaled mean monthly rainfall for base period (1971-2000)

Fig. 8. Observed and CGCM3 Downscaled mean monthly rainfall for base period (1971-200

The simulated and observed monthly mean rainfall for the base period (1971-2000) is shown in **Fig. 7**. And **Fig. 8**.The annual trend of observed rainfall and simulated rainfall for base as well as future period for a particular grid are shown in **Fig. 9**. The overall result shows that the predicted annual rainfall is higher than the observed and the trend of the annual observed rainfall and predicted rainfall is overall increasing. **Fig. 10.** Shows the average rainfall is increasing.

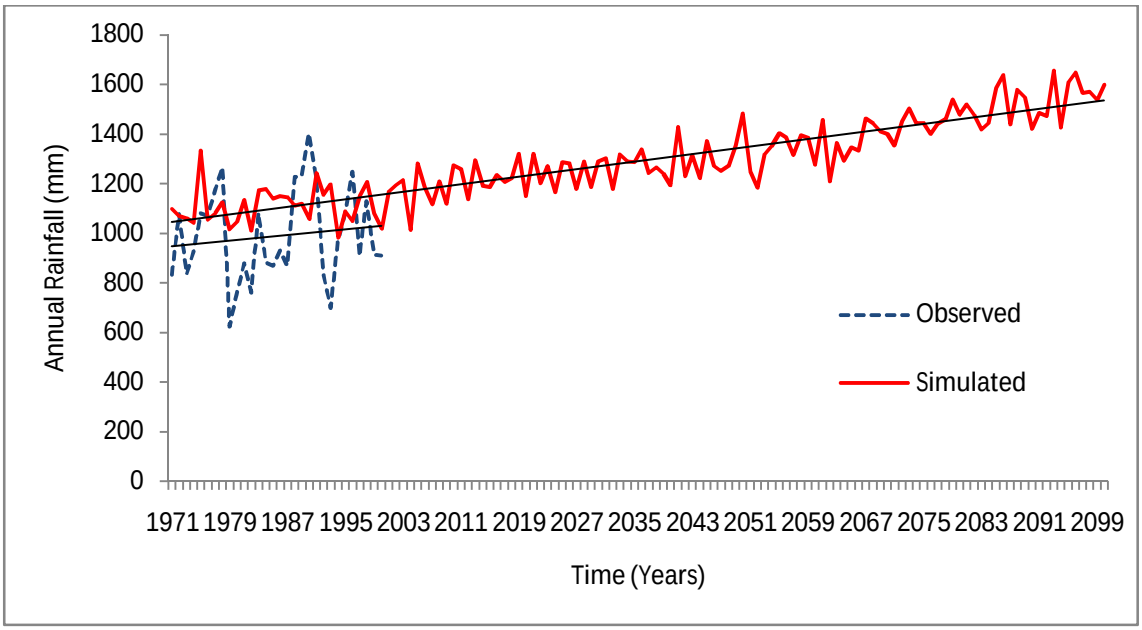


Fig. 9. Observed and GCM downscaled annual rainfall trend

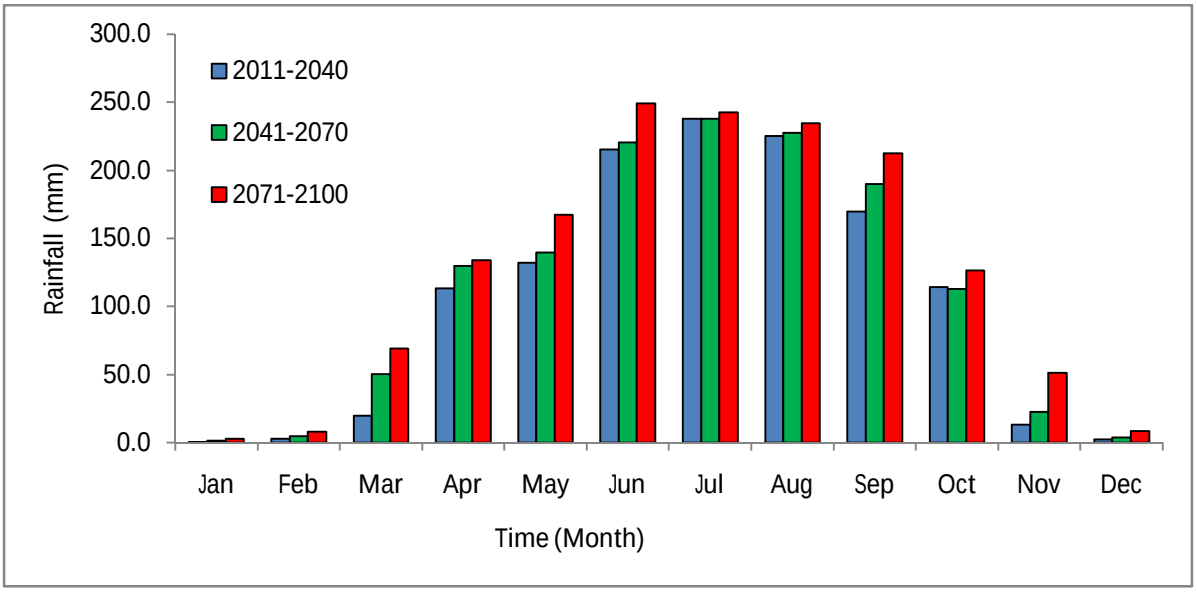


Fig. 10. Mean Monthly rainfall downscaled from CGCM3 A2

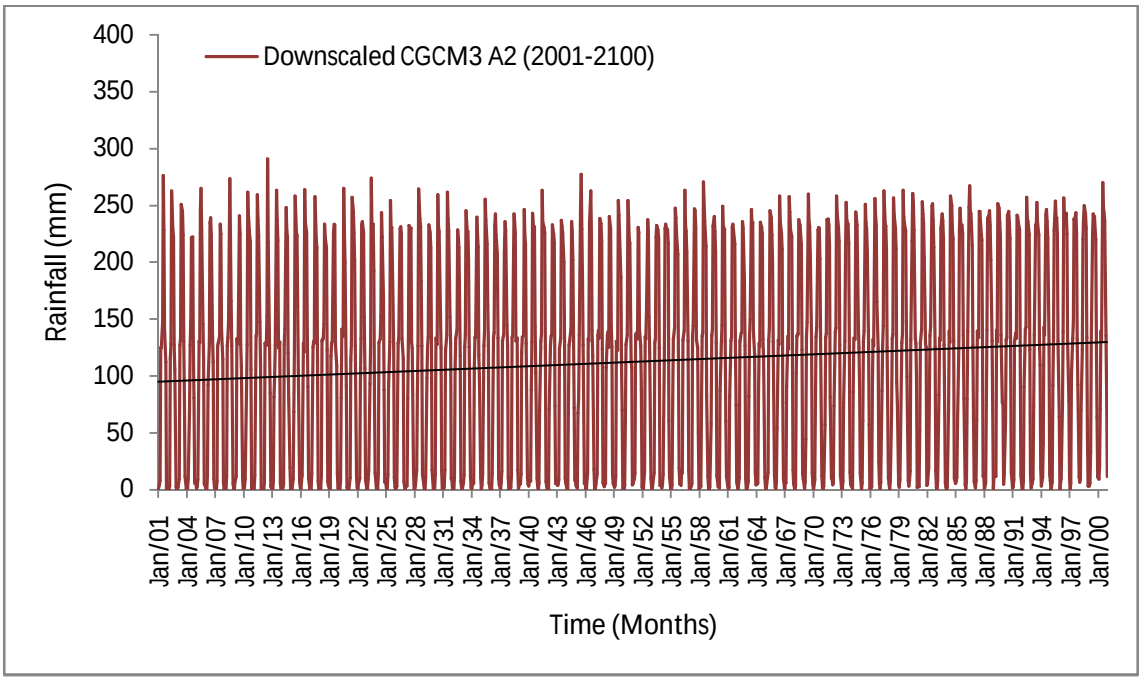


Fig. 11. Downscaled monthly rainfall for period (2001-2100) for CGCM3 A2

Typical results of predicted rainfall for the future period (2001-2100) for a particular grid are shown in **Fig. 11**. It has been observed from the results, that the overall trend of monthly rainfall predicted for future period (2001-2100) is increasing.

SUMMARY AND DISCUSSION

This study shows the ability of GCM simulation output for climate change impact studies. This studies shows that the M5P model tress which has less parameters than SVR performed better. It is also observed that Unpruned M5P model is better for hydrological process i.e. rainfall which has peak and trough. But models fails to capture extreme events this may be due to the presence of inter and intra variability in monthly rainfall. Result shows that the model predicts more rainy days and less extreme events. Most of the predicted rainfall data set for future period shows an increasing trend.

REFERENCE

Kripalani, R.H., Oh, J.H., Kulkarni, A., Sabade, S.S.andChaudhari, H.S., (2007). "South Asian summer monsoon precipitation variability: Coupled climate model simulations and projections under IPCC AR4" Theor. Appl. Climatol. 90, 133–159.

Lal, Murari et al (2001). "Future climate change: Implications for Indian summer monsoon and its variability".Current Science, vol. 81, no. 9.

Goyal et al. (2011)."Nonparametric Statistical Downscaling of Temperature, Precipitation, and Evaporation in a Semiarid Region in India. Journal of Hydrologic Engineering, Vol. 17, No. 5

Rupa Kumar et al. (2006) "High-resolution climate change scenarios for India for the 21st century". CURRENT SCIENCE, VOL. 90, NO. 3

Anandhi et al. (2008) "Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine". *Int. J. Climatol.* **28**: 401–420

IPCC 2007. Climate Change 2007: The Physical Science Basis *Working Group I Constribution to the Fourth Assessment Report of the International Panel on Climate Change.* Cambridge University Press: Cambridge.

IPCC 2000. Emissions Scenarios : Intergovernmental Panel on Climate Change. Cambridge University Press: Cambridge.

Ghosh, Subimal and Mujumdar, P.P., (2006) "Future rainfall scenario over Orissa with GCM projection by statistical downscaling". CURRENT SCIENCE, VOL. 90, NO. 3