International Journal of Civil Engineering Research.
ISSN 2278-3652 Volume 5, Number 2 (2014), pp. 111-120

http://www.ripublication.com/ijcer.htm

Downscaling of Precipitation in Mahanadi Basin, India

Manjula Devak[1] and C.T. Dhanya[2]

[1]Department of Civil Engineering, Indian Institute of Technology, Hauz Khas, New Delhi–110016, INDIA.
[2]Department of Civil Engineering, Indian Institute of Technology, Hauz Khas, New Delhi–110016, INDIA.

Abstract

Downscaling is a technique for obtaining local - scale hydrological variables from coarser - scale atmospheric variables that are generated by General Circulation Models. Mainly there are two downscaling methods i.e. Dynamic Downscaling & Statistical Downscaling. In this study Statistical Downscaling is used as it offers less computational work as compared to Dynamic Downscaling & it also provide us a platform to use ensemble GCM outputs. In this paper, the result generated from two methods are compared i.e. by Support Vector Machine (SVM) and K- Nearest Neighbour (KNN), which covers some parts of Chhattisgarh, Orissa , Bihar and Maharashtra state. The above two models are applied at five different locations in Mahanadi Basin. Bias correction by Equidistant CDF matching method is also applied to the future projection. Calibration & validation of the model incorporates the result from Canadian global climate model (CanCM4) for historical scenario and future projections are done using predictors from RCP 4.5 scenario. Various performance measures like, Normalized Mean Square Error (NMSE) & Correlation Coefficient is also taken into account. Kolmogorov Smirnov test is also performed for the two models.

Keywords: Downscaling, General Circulation Models, Support Vector Machine, K- Nearest Neighbour, Equidistant CDF matching method.

1. Introduction

From the decades to millions of years the most symbolic, continual and indelible change in the statistical dispensation of climate stencil is climate change. The changes

may be in the average climatic situations or its diffusion around average conditions. Variability of variables such as temperature and precipitation over a period defines climate and changes associates with it. General Circulation Models (GCMs) are one of the best means or tools which are used to study the impact of climate change. The purpose of GCM is to predict changes in the weather occurring due to the changes in boundary conditions or physical parameters. These are three dimensional models. It provides a spatial coverage at global scale. Generally resolution of GCMs ranges from 250 to 600 kms. As this model gives the simulations at a very coarser scale which is of the order of 2.8° by 2.8° which corresponds to many thousands of kilometres and many times there is a need to know the variables at finer scales that is at few hundreds of kilometres, keeping this prospect in mind many downscaling methods are developed.

Downscaling is an art of getting finer grid data from coarser grid data. Basically downscaling is a method for obtaining high-resolution climate or climate change information from relatively coarse-resolution Global Climate Models (GCMs). The gap between the simulations obtained from global climate models and the information which is needed at local scales that is what is needed by decision makers and impact assessors is minimised by downscaling techniques. Basically downscaling are of two types, viz, dynamic and statistical downscaling. In dynamic downscaling Regional Climate Model (RCM) is embedded into GCM. Rather than using equations to bring global-scale projections down to a regional level, dynamic downscaling involves using numerical meteorological modelling to reflect how global patterns affect local weather conditions. Despite the number of advantages dynamic downscaling is computationally demanding; may require considerable effort to adapt a regional model for a source of lateral boundary data; parameterizations may not be adequate for ranges beyond which they were developed. Statistical downscaling set ups statistical relationships between predictors, predictands and observed data by formulating equations which converts global scale variables to local scales. These relationships are ultimately used to forecast climate information for future period. This approach takes the input from GCMs for particular region to relate global climate aberrations to regional climate aberrations. Wilby and Wigley (1997); von storch *et al* (2000); gives some implicit assumptions in statistical downscaling. They are 1. The predictors are variables of relevance and are realistically modelled by the GCM. 2. The predictors employed fully represent the climate change signal and 3. The relationship is valid under altered climate conditions. In this paper two statistical downscaling methods are used, viz, Support Vector Machine (SVM) and K-Nearest Neighbour (KNN).Support vector machine is based on the statistical learning and structural risk minimization principle. The K-nearest neighbour is simplest among all machine learning algorithms. Equidistant CDF matching method is employed for the removal of biases. Kolmogorov smirnov test is applied to the two models to compare distributions of the modelled values to the distribution of the observed data set. All the techniques described above is applied to Mahanadi basin in India, at five different locations and results are compared for both the models.

2. Study Area and Data Discription

The study region is the catchment area of Mahanadi River, which is located between 19˚N and 23.7˚N latitude and 80.4˚E and 86.9˚E longitude. Mahanadi Basin extends over an area of 141589 Sq km. The basin lies in the states of Chhattisgarh (75,858.45 sq. km), Orissa (65,580 sq. km), Bihar (635 sq. km) and Maharashtra (238 sq. km). Mahanadi River rises from Raipur district of Chhattisgarh and flows for about 851 km before its outfall into the Bay of Bengal. The normal time of onset of monsoon over the basin is the first week of June. The bulk of the precipitation (800mm to over 1600mm) over the basin falls in the period from June to September while precipitation received in January to February is less than 50 mm. The south west monsoon (June-October) accounts for nearly 91% of the annual rainfall. December is the driest month contributing less than 10% of the annual rainfall. Eleven predictor variables, viz, Air temperature (at 200, 500, 700, 925 mb), Geo-potential height (at 200, 500, 925 mb), Eastward wind (at 200, 925 mb) and Northward wind (at 200, 925 mb), are selected for downscaling precipitation in Mahanadi basin. These predictors were selected based on the study conducted by Anandhi *et al*., (2008). Predictor data set for the period from January 1961 to December 2004 for nine grid points is extracted from the NCEP. National Centre for Environmental Protection (NCEP) prepared a gridded data at monthly time scale, which is extracted from http://www.cdc.noaa.gov/.

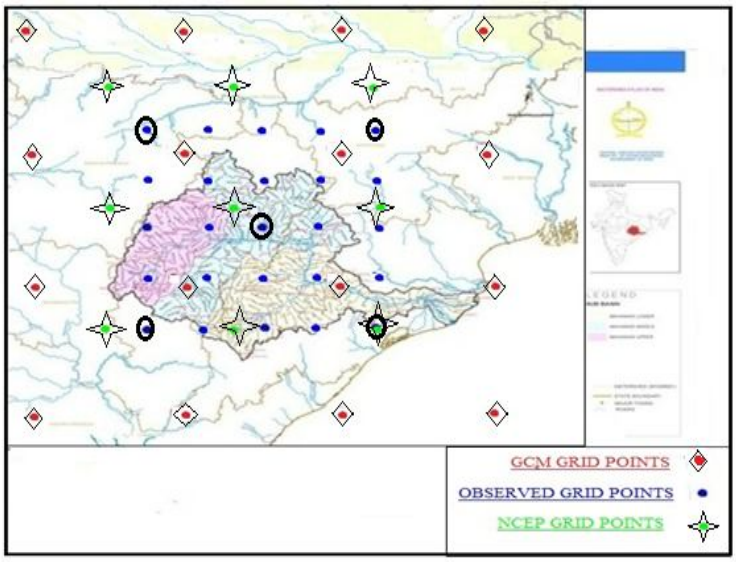

Figure 1: Mahanadi Basin.

Latitude of NCEP data ranges from 20°N to 25°N and longitudes ranges from 80°E to 85°E with 2.5° by 2.5° spatial resolution. Gridded precipitation data of resolution 1° by 1° is taken from Indian Meteorological Department (IMD). For the study region GCM data is taken from the website www.cccma.bc.ec.gc.ca/. These are the simulations from CanCM4 model of the Canadian centre for climate modelling and

analysis of Coupled modelled Intercomparison project-5 (CMIP-5). Historical and RCP 4.5 scenarios are used in the study. The predictor set for historical scenario is extracted from January 1961 to December 2004 and for RCP 4.5 scenario predictor set is extracted from January 2016 to December 2035 for sixteen grid points. A latitude of GCM data ranges from 18.149°N to 26.5108°N and longitudes ranges from 78.75°E to 87.1875°E which is 2.8125° by 2.8125° spatial resolution. The GCM data points are interpolated to NCEP data points for study purpose. Figure 1 shows the five encircled locations where downscaling is done.

3. Methodology

Statistical downscaling is performed in this study by the use of Support Vector Machine and K-Nearest Neighbour approach.

3.1 Support Vector Machine

The Least-Square Support Vector Machine (LS-SVM) has been used in this study for downscaling (Anandhi *et al.*, 2008). The whole procedure of downscaling precipitation using SVM is explained in Figure 2.

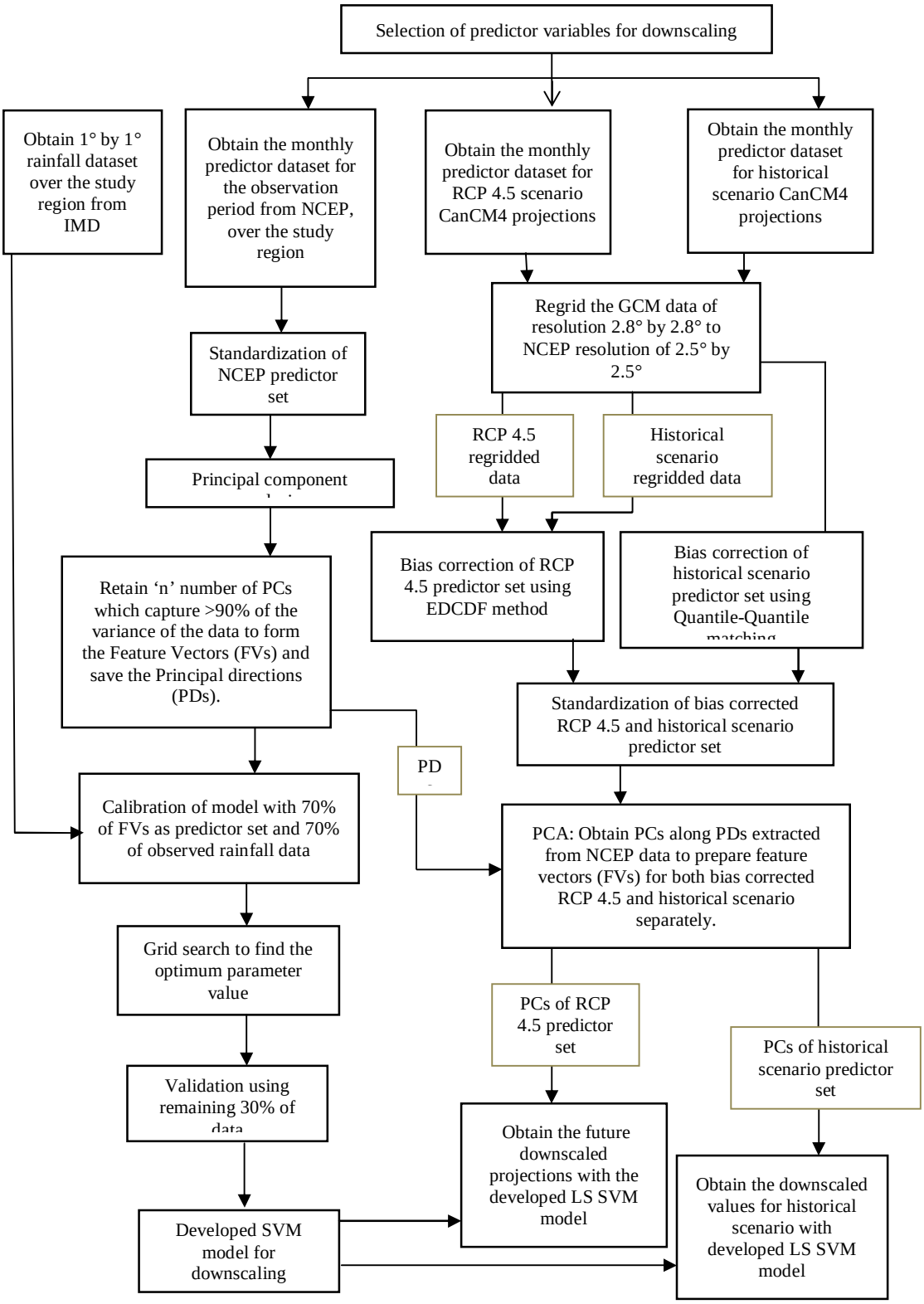

Figure 2: Methodology for SVM downscaling.

3.2. K- nearest neighbour approach

The K-nearest neighbour is simplest among all machine learning algorithms. It finds the K- neighbour nearest to the new sample data from the training set based on distance metric and respective similarity, when new sample data is introduced.

The methodology adopted for downscaling the precipitation using KNN is as follows:

1. Compile all predictor variables for nine grid points.
2. Perform principal component analysis and preserve the principal directions (PDs) for future reference.
3. Standardise the testing set by subtracting the mean and dividing by standard deviation as done for SVM model before.
4. Multiply the standardised set with PDs preserved earlier to form feature vectors.
5. For each time element *i*, calculate the Euclidian distance between the feature vectors and the PCs. The equation of Euclidian distance (d_i) is given below.

$$d_i=[\sum_{j=1}^{n} \frac{\lambda j}{tr[W]}(x'_j-p_{ij})^2]^{0.5} \quad (1)$$

Where, λ_j= Eigen values

tr[W]= trace of the matrix W which is diagonal matrix of Eigen values

n= number of principle components retained.

6. Retain only first K distances after sorting the distance d_i in ascending order. The choice of K can be made by equation K= $(t)^{0.5}$ where t is the total time element in calibrating set.
7. From the observed data of precipitation select the K (= 20) neighbours corresponding to the particular Euclidian distance for each day.
8. Bisquare weighting function is used to assign the weights to each of the 20 neighbours. The relationship is given below:

$$w_i= [\{1-(d_i/d_k)^2\}^2] / \sum_{i=1}^{K} [\{1-(d_i/d_k)^2\}^2] \quad (2)$$

9. Where, d_k is the distance of Kth neighbour.
10. Then apply weighted average technique to find the precipitation for particular month.
11. Repeat steps 5 to 9 for each time element.
12. Repeat steps 3 to 10 for each node point.

4. Results and Discussion

4.1 Calibration and Validation of SVM model

For obtaining the optimal range of each of the SVM parameters the grid search procedure is used. A range of kernel width (σ) and penalty term (C) is selected and optimal range of the parameters is obtained by the domain search. The value of σ and C having the least NMSE (Normalized Mean Square Error) and highest value of correlation coefficient is selected as the optimum parameter. The optimal values of SVM parameters C and σ for all the five grid points thus obtained are 50 and 50000 respectively. Once the calibration of model is done, LS SVM model is validated with

the remaining 30% data. NMSE value and R^2 value gives better result for complete testing set than for monsoon months. Complete testing set shows less NMSE values and higher R^2 values than monsoon months, this can be the explanation that model is unable to capture the extreme values of precipitation but follows good behaviour for non- monsoon months.

4.2 Calibration and Validation of KNN model

The optimum number of nearest neighbour, K is fixed through an empirical formula, K =(length of dataset)^0.5= (420)^0.5≈ 20. Once the model is calibrated (by preserving PCs and PDs for future references), KNN model is validated with the remaining data. Although KNN model is able to capture the extreme values at most of the locations, but there exists a time lag between observed values and KNN simulated values of precipitation. The results for Normalized Mean Square error and Correlation coefficient (C=50 and σ= 50000) between observed precipitation testing data and simulations from KNN model is better for complete testing set than for monsoon months (June, July, August and September).

4.3 Comparison of results

Simulations from Support vector machine (SVM) and K-nearest neighbour method is compared with respect to observed data by the help of box plots and cumulative distribution functions for both historical period and future projections at all five locations. Box plots for historical period reveals that K-nn method gives results close to the observed one while SVM sometimes underestimates and overestimates the results and also unable to predict peak values. Figure 3 shows Box-plot and Figure 4 shows CDF at one of the five locations. **KOLMOGOROV- SMIRNOV TEST**, a non parametric test, which is used to compare two sample data sets (weather it belongs to the same distribution or not) is also conducted. The K-S test is conducted first on observed data with simulation from SVM model and second on observed data with simulations from KNN model. Both reject the null hypothesis which signifies that it belongs to the same family of distribution.

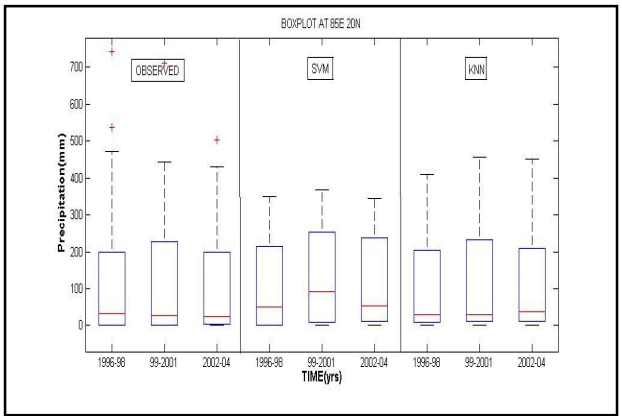

Figure 3: Box plot for testing period (at 85E 20N, Location 4)

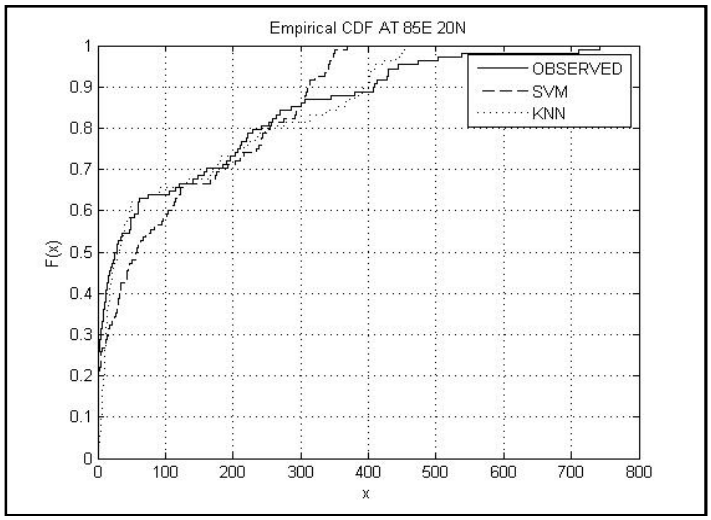

Figure 4: CDF for testing period (at 85E 20N).

4.4 Future Projections

Figure 5 shows box-plot for future projections at one of the five locations. In projecting the future value, KNN model generates the peak values higher than the values generated by SVM model.

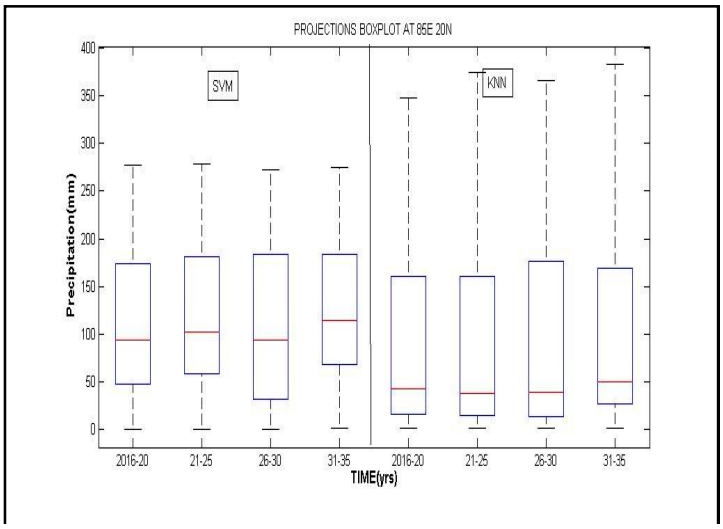

Figure 5: Future projections (85E 20N).

6.5 Simulations of precipitation values for predictor values taken from General Circulation Model (GCMs)

Figure shows box-plot for historical scenario at one of the five locations.

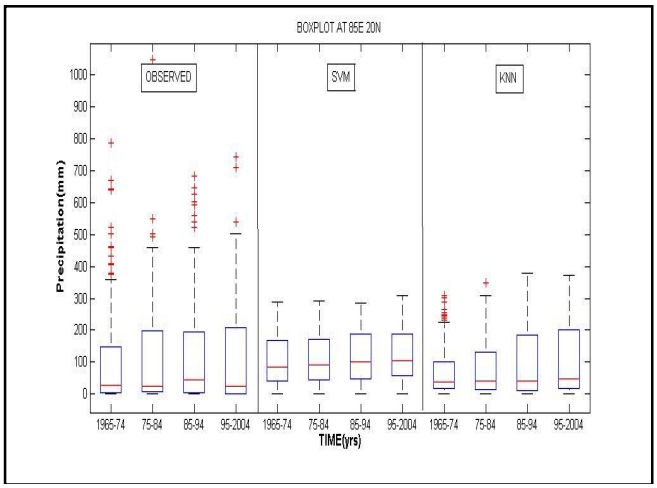

Figure 6: Box plot for simulations (precipitation) by using GCM predictors for historical scenario (85E 20N).

5. Conclusion

Outputs from model were examined at five locations. After calibrating both the model it was found that SVM model is unable to simulate peak values than KNN model. As SVM is regression based model, unable to capture the whole variance of the data set, is one of the reason for the above stated argument. SVM model is computationally more efficient than KNN model. It is seen that SVM model overestimates the mean at some locations as compared to KNN model and underestimates the standard deviation at some locations. Also there is the presence of regional variation in the results. Comparison between the results from both the models is made by means of box-plots, CDFs and quantile- quantile plots. Kolmogorov- smirnov test is also conducted for observed data set with both SVM and KNN results separately which rejects the null hypothesis and signifies that all the results belong to the same distribution. SVM model becomes difficult to formulate when it comes the case of large dataset. It is seen from the above results in some parameters KNN proved to be a good model over SVM while in some parameters SVM shows its good productivity. When we consider both the model location wise, it is observed that at some location SVM simulates results close to observed data while at some locations KNN predict good results. Therefore sometimes the applicability of the model also depends on the location of site. These models can also be used to downscale other parameters like maximum temperature, specific humidity, minimum temperature etc which further helps in the valuation of climate changes with time and location both. Also these models can be used for downscaling at daily basis all we have to take care with the SVM model is the size of the data set.

References

[1] Anandhi, A., V.V. Srinivas, R.S.Nanjundiah,(2008), Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine, International journal of climatology 28: 401-420.

[2] Buishand, T.A.., and T. Brandsma (2001), Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest- neighbour resampling, Water Resources Research, 37(11), 2761-2776.

[3] Coles, S., (2001), An Introduction to Statistical Modeling of Extreme Values, Springer, London.

[4] Cortes, C., V. Vapnik, (1995), Support vector networks, Machine Learning 20,273-297.

[5] Dibike, Y.B., P. **Coulibaly,** (2006), Temporal neural networks for downscaling climate variability and extremes, *Special Issue of Neural Networks*, 19(2), 135-144.

[6] Gangopadhyay, S., M. Clark, (2005), Statistical downscaling using K-nearest neighbours, Water Resources Research, 41, W02024.

[7] Ghosh, S., P.P. Mujumdar, (2008), Statistical downscaling of GCM simulations to streamflow using relevance vector machine, Advances in water resources, 31:132-146.

[8] Ghosh, S., P.P. Mujumdar, (2008), Statistical downscaling of GCM simulations to streamflow using relevance vector machine, Advances in water resources, 31; 132-146.

[9] Khan M.S., P. Coulibaly, Y. Dibike, (2006), Uncertainty analysis of statistical downscaling methods, Journal of Hydrology 319 357 -382.

[10] Lall, U., A. Sharma, (1996), A nearest neighbour bootstrap for time series resampling, Water Resources Research, 32(3), 679-693.

[11] Lek, S., J.F. Guegan, (2000), Artificial Neuronal Networks: Application to Ecology and Evolution, Springer, Berlin.

[12] Li H., J. Sheffield, E.F. Wood,(2010), Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching, Journal of geophysical research, vol. 115.

[13] Rajagopalan, B., U. Lall, (1999), A K- nearest neighbour simulator for daily precipitation and other variables, Water Resources Research, 35(10), 3089-101.

[14] Tripathi, S., V.V.Srinivas, R.S.Nanjundiah, (2006),Downscaling precipitation for climate change scenarios: A support vector machine, Journal of hydrology 330:621-640.

[15] Vapnik V.N., (1998), Statistical Learning Theory, Wiley, New York.

[16] Wetterhall F, A. Bardossy, D. Chen, S. Halldin, C. Xu, (2007), Daily precipitation-downscaling techniques in three Chinese regions, Water Resources Research 42.

[17] Wilby RL, T.M.L. Wigley, (1997), Downscaling general circulation model: a review of methods and limitations. Progress in Physical Geography 21:530-548.

[18] Yates,D., S. Gangopadhyay, B. Rajagopalan, K. Strzepek, (2003), A technique for generating regional climate scenarios using a nearest neighbour algorithm, Water Resources Research, 39(7), 1199.