Fine Resolution Indian Summer Monsoon Rainfall Projection with Statistical Downscaling

K.Shashikanth, and Subimal Ghosh

***Abstract*—** Kernel Regression based statistical downscaling is performed on recently developed and much improved CMIP5 model data product and derive Indian summer rainfall (ISMR) on to a very fine scale (25km) on daily time step. Statistical models are based on statistical relationship between large scale climate features as simulated from General Circulation Models (GCMs) and hydrological variable of interest namely Rainfall. Statistical downscaling which is not only computationally easy but also adds enormous value to the projections derived from global climate models. Statistical downscaling consists of two steps viz. derivation of rainfall state using Classification And Regression Tree (CART) and rainfall amounts are estimated conditioned on rainfall state using non parametric Kernel Regression estimator. Concept of common rainfall state for the study area has improved cross correlation of different rainfalls at various locations. NCC (Norway Climate Center) GCM model has projected heterogeneous changes in future rainfall for the period 2081-2100 under RCP 8.5 scenario. The model shows an increase in rainfall along west coast, North East and western Regions of India. Further decrease in rainfall in the rest of country. This study highlights usefulness of proper water resources planning and management in the future.

***Keywords*—** Climate Change, Indian Monsoon, GCM, Statistical Downscaling, Kernel Regression

I. Introduction

INDIAN summer Monsoon Rainfall (ISMR) is an important component of Asian Monsoon system which contributes 70% of total rainfall; studying the impact of climate change on ISMR is a challenge facing the climate science community currently. Climate change study basically involves the downscaling of coarse resolution variables data which are well simulated by GCMs by using two well known techniques viz. Dynamic downscaling and Statistical Downscaling [2] Kastubh et al 2013). Without downscaling the GCM data is of no value as it not only contains bias but also it works on high resolution of the order of more than 1.8 °x 1.8° making it very difficult to directly use for impact assessment studies at that resolution. Therefore downscaling adds considerable value to the from the GCM projection [1]. In the present study, we have used recently developed and much improved GCM Norway Climate center (NCC) from CMIP5 data archive from which ISMR at fine resolution (25km) on daily time are developed. For future projections we have used A2 scenario for the period 2081-2100 time slice. These projections are important so that appropriate decisions on adaption and mitigation strategies can be worked under climate change.

Shashikanth. Kulkarni Department of Civil Engineering, Indian Institute of Technology Bombay, Mumbai, India- 400076

Subiama.Ghosh Department of Civil Engineering, IIT Bombay, Mumbai, India (corresponding author's phone: +91 9930663969

(E-mail: subimal@civil.iitb.ac.in).

II. Observed Rainfall Data- Aphrodite

Asian Precipitation Highly Resolved Observational Data Integration towards Evaluation of Water Resources (APHRODITE)), Japan is a daily gridded precipitation data set created from 1961-2004. The rainfall product is based on data collected from dense network of daily rain gauge data from all across Asia including the data from sparse areas like Himalayas and Mountainous areas of Middle East. The state of art of daily precipitation data is available at 0.5 ^{o} x 0.5 ^{o} and 0.25 ^{o} and 0.25 ^{o} resolution. The Aphrodite body has used an improved interpolation scheme which gives proper weightage to local topographical features to improve the orographic precipitation [3]. The website address http://chikyu.ac.jp/precip

III. III Selection of Predictors

The selection of predictors is an important step in the statistical modeling. The final results heavily depends on the variables. The "ideal" downscaling variable is (Wilby et al 2004) the one which is

- Strongly correlated with local variable(s) of interest.
- Physically and/or conceptually sensible.
- Able to preserve covariance between local variables.
- Accurately described by the GCM.
- Archived at the same temporal resolution as the local variable(s).

In the present study the following variables are chosen; air temperature, wind velocities both U and V wind at surface and 500hpa, Mean sea level Pressure (MSLP) and humidity at 500hpa. For selection of predictors see the literature [4,6].

IV. NCEP/ NCAR Reanalysis Data

Reanalysis data is surrogate for observed data for any predictor variable. As the numerically solved fundamental equations in case of GCMs result in systematic errors known as bias, it is necessary to correct them. This correction is done based on the observed data.

The NCEP-NCAR Reanalysis data set is a continually updating gridded data set representing the state of the Earth's atmosphere, incorporating observations and numerical weather

prediction (NWP) model output dating back to 1948. It is a joint product from the National Centers for Environmental Prediction (NCEP) and the National Center for Atmospheric Research (NCAR), NOAA. For the current projections, the reanalysis data was downloaded. The resolution is 2.5º lat × 2.5º long. The base line period considered for the present study is from 1961-2000 because it is of sufficient duration to establish a reliable climatology. The NCEP/NCAR reanalysis-I data [8] provide global atmospheric data which is a mixture of physical observations and model forecasts. *Kalnay et al.*, [8] have used different data assimilated systems such as global rawinsonde data, aircraft data, satellite data, and surface land synoptic data, advanced microwave surface wind speed data etc. to with a T62 resolution and 28 vertical sigma levels to calculate the reanalysis data products for various climate variables. Climatic variables are categorized in to three types [8]. Category A variables are those that are strongly influenced by observations eg. Zonal and meridional wind. Category B variables are influenced by the model and observations, but are not as accurate as category A level variable. Eg. Specific humidity. Category C variables are directly determined by model e.g. Precipitation flux. For more information visit the site:
http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html.

V. Statistical Downscaling Methodology

Statistical downscaling involves development of statistical relationship between large scale climate variables, (which are reliably simulated by GCMs and known as predictors), and fine resolution rainfall (predictands) and then applying the relationship to the GCM simulations.

Here we use the statistical downscaling method developed by *Kannan and Ghosh* [6] and *Salvi et al. [4]* to project Indian Summer Monsoon Rainfall from NCC GCM simulation. The generation of multisite rainfall consists of two parts viz. the first part includes Classification and Regression Tree (CART) based rainfall state occurrence model conditioned on climate predictors and second part includes the generation of multisite rainfall amounts conditioned on the rainfall state using Kernel regression Model. The methodology is presented in supplementary Figure 1.The methodology is applied separately to seven meteorologically homogeneous zones [13] in India.

The GCM simulated variables have systematic deviations (bias) with respect to observed NCEP/NCAR reanalysis data. **For the current study,** quantile based remapping method (*Li et al.*, 2010) is used for bias correction. The Bias corrected data needs further mathematical treatment because of multidimensionality and multicollinearity [4].Principal Component Analysis *(PCA)* is used to reduce the dimensionality and multicollinearity.

K means clustering is first used with the observed rainfall data to classify the spatial pattern of rainfall in a zone to different classes. The classes are then simulated with Classification and Regression Tree (CART) [6] from large scale climate variables. The training of CART is performed for 1961-1980. Conditional on the derived classes, the non-parametric kernel regression is performed between large scale predictors and fine resolution rainfall.

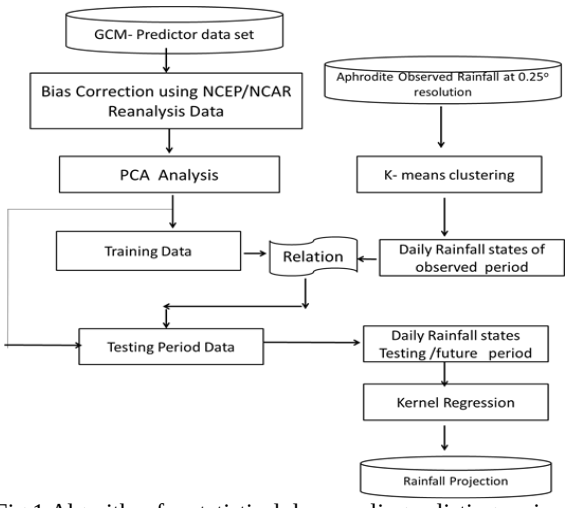

Fig.1 Algorithm for statistical downscaling enlisting various mathematical operations to be performed on GCM predictors and predictand (rainfall) are shown. Kernel regression methodology developed [6, 7, 4] is used for present work.

Nonparametric downscaling techniques are smoothing filters and are widely used for multisite downscaling. Generally, the relationships between predictors and predictand are highly nonlinear and to capture the relationship, nonparametric techniques such as k-nearest neighborhood [9,10] and kernel regression ([6], [4]) are used. For the present study, kernel regression technique is used for obtaining rainfall projections conditioned on the future states. The kernel regression is general form of conditional expectation. Exponential kernel function is used to assign the probabilistic weights based on the difference between predictor vector for testing period and training period for the day chosen. If the difference is small very high weight is assigned and vice versa. Details of the expression for kernel regression, bandwidth calculation may be found in ***Kannan and Ghosh*[6].**

A. Bias Correction

For the Bias removal, a method devised by Li et al. [5] is used which is a Quantile based mapping method. In this method fit same distributions to NCEP/NCAR data, GCM train and test data and find its corresponding parameters respectively. Compute the CDFs to all three data sets using their respective parameters. Generate new data set using the CDF gcm test and taking the parameters of GCM train. Subtract this series from the GCM simulated GCM test series. This is treated as Climate change signal. This difference is added to the new series to be generated for CDF GCM test by using the observed parameters. This is the Bias corrected series for test period or future.

The GCM data is interpolated to NCEP/NCAR level by using bilinear interpolation technique. For the predictors namely temperature, specific humidity and MSL gamma distribution and for u wind and v wind normal distribution is used before applying bias correction principle. Once the bias is removed

from the GCM predicted data, PCA is used to get rid of correlation among the predictors.

B. Principal Component Analysis

Principal Component Analysis (PCA) is a way of transforming a set of correlated N-dimensional vectors of observations into another set of N-dimensional uncorrelated vectors, such that most of the information content of the original observations is stored in the first few dimensions of the new set. Therefore, the objective of PCA is to identify the patterns of multidimensional variables and to transfer the set of correlated variables into a set of uncorrelated variables, called principal components, with reduction in dimensionality. The advantage of PCA is that by using a small number of principal components it is possible to represent the variability of the original multivariate data set. At the same time, the principal components are uncorrelated and therefore there is no redundant information. For further details on the PCA technique the works of Salvi et al [4].

C. K-means Clustering Technique

India extends from 5 ° to 40 ° N in latitude and from 60 ° to 105 ° E. A very high resolution (0.22 ° x 0.22° lat/long) gridded daily rainfall data containing daily rainfall time series for 4971 grid points covering entire India from 1961 -2000 developed by APHRODITE is used [3] to identify rainfall states using K – means clustering technique. In this method, in order to find the future rainfall states, the K- means algorithm is run to cluster the gridded rainfall. As this is an unsupervised clustering technique without any target output, validation/ testing is not required/ possible for K-means algorithm[6].

D. Classification and Regression Tree Analysis

Supervised classification is a machine learning technique for learning a function from training data. The training data consist of pairs of input objects (typically vectors) and desired outputs. The output of the function can be a continuous value or can predict a class label of the input object. The task of the supervised learner is to predict the value of the function for any valid input object after having seen a number of training examples (i.e. pairs of input and target output). The CART model builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). In most general terms, the purpose of the analysis via tree-building algorithms is to determine a set of if-then logical (split) conditions that permit accurate prediction or classification of cases. The tree is built through a process known as binary recursive partitioning algorithm. This is an iterative process of splitting the data into partitions, and then splitting it up further on each of the branches. In the current study, the CART Analysis is used to (1) Obtain the relationship between the PCA processed predictors and rainfall states for training period (2) Generate future rainfall states using the developed relationship and PCA processed predictors for testing period (simulated by GCM) with as assumption that the relationship developed for training period holds good for future as well. Once the future states are simulated, using Kernel regression the future rainfall is projected.

E. Kernel Regression

In a linear regression problem the mean of a dependent variable Y is related to a set of independent variables X1,X2,……Xd as

$$E(Y\bullet X)= X1^{2}\,1+X2^{2}\,2+\&\ \&\ \&\ \&\ ..Xd\ (^{2}\,d= X^{T_2}) \qquad (1)$$

Multivariate kernel regression [11,12] belongs to non-parametric smoothing technique which uses weighted sum of the observed responses with the use of kernel density function weights.

The general form:

$$E\,(Y|X) = m(x) = \int y\ f(y|x)\,dy = \frac{\int y\ f(y,x)\,dy}{f_X(x)} \qquad (2)$$

Where m(x) is the conditional expectation function X=[X1, X2, X3…….Xd] ^{T}

Y = Predictand, f (Y|X) is the conditional probability density function of Y given X=x.

$f_X(x)$ is the marginal Pdf of X. Nadaraya (1964) provided an estimator by replacing the multivariate pdf with kernel density estimates. Any kernel function is essentially a probability distribution function; an exponential Kernel function used in the study. Bandwidth is found using AMISE technique.

VI. Results And Analysis

The Fig.2 gives us the comparison between observed rainfall and projected rainfalls from NCC GCM model for testing period 1981-2000 are presented. The mean rainfall of JJAS months have been modeled from the above methodology has been able to model the characteristics of ISMR well. The method has able to capture the orography in the west and northeast part of India also properly.

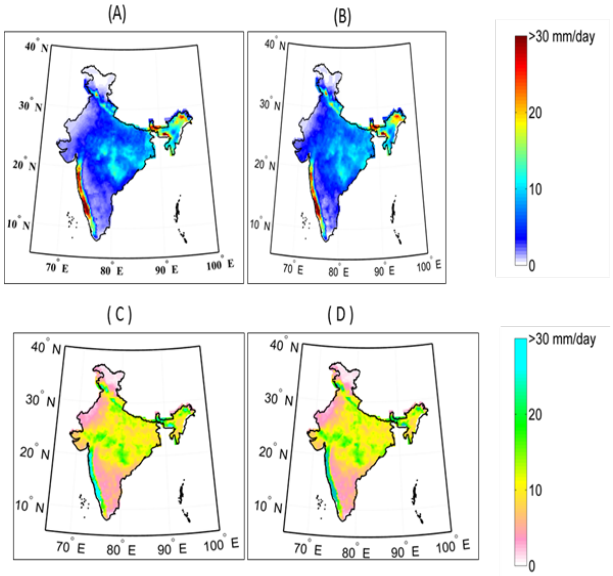

Fig. 2: The Plot shows the Observed mean rainfall (A) and the mean rainfall projected from NCC GCM (B) is almost matching the observed rainfall. The standard deviation of Observed rainfall (C) and NCC GCM (D) are shown.

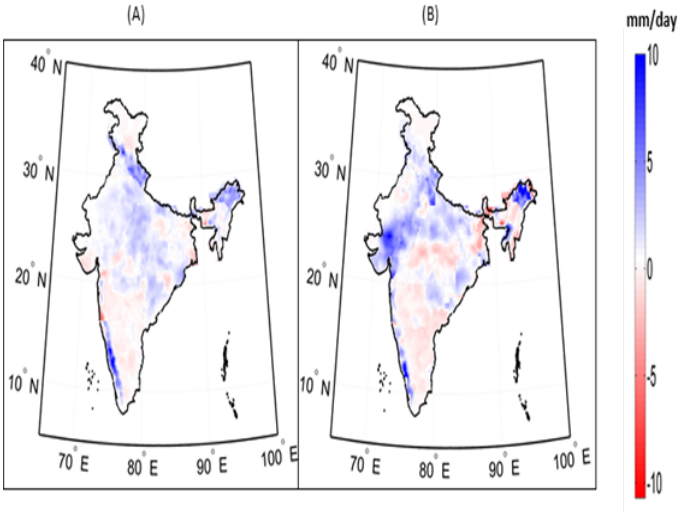

Fig 3: Plot shows the difference in mean plot for testing period Fig.(A) (1981-2000) and changes in mean for future projections for the period 2081-2100 (Fig (B).

The Fig 3 is an important one it shows capability of the model in simulating the ISMR. The Fig(A) shows that the difference in mean plot for the testing period is lying in between ±10mm. In the most of nodes it is in fact lying in between -2mm to +2mm. only which indicates that it is replicating very well the ISMR.

The Fig 3 (B) indicates that the changes in mean future rainfall projections where in there is increase in rainfall along west coast of India and in the Northeast part of India. There is decrease in rainfall in the south India. There is increase in rainfall in the borders of western India and in the rest of the country there is decrease in rainfall.

VII. Conclusion

For the present study the Norway Climate center (NCC) GCM is used for projections of ISMR, However in order to reduce the intra model uncertainty different GCM models can be selected so that better results can be obtained. With a single GCM the results can be inconclusive for proper planning and adaptation purposes. Statistical downscaling requires huge sums of data during projections and The other biggest limitation of this method includes assumption of stationarity, where in the relation between predictors and predictand holds good for the future . The quality and quantity of observed data set may influence the rainfall projections to a great extent.

Acknowledgment

We thank the World Climate Research Programme's working Group on coupled Modeling, which is responsible for CMIP, and the climate modeling for making available their model outputs. We would like to thank also APHRODITE, Japan for making available 25 km resolution observed data.

References

[1] Maraun, D., et al., Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user, *Rev. Geophys.,* 2010 48 RG3003, doi:10.1029/2009RG000314.

[2] Wilby R L , L.E. Hayc , G.H. Leavesleyc, A comparison of downscaled and raw GCM output: implications for climate change scenarios in the San Juan River basin, Colorado *J. Hydrol.,* 1999 225 ,67–91

[3] Yatagai A., Osamu Arakawa, Kenji Kamiguchi, Haraku Kawamoto, Masato I Nodzu and Atsushi Hamada , A 44 year daily gridded precipitation data set for Asia based on a dense network of rain gauges *SOLA* Vol., 5, 137-140, doi; 10.2151/sola.2009-035.

[4] Salvi, K., S. Kannan, and S. Ghosh, High-resolution multisite daily rainfall projections in India with statistical downscaling for climate change impacts assessment, *J. Geophys. Res.Atmos.*,2013 118, doi:10.1002/jgrd.50280

[5] Li, Haibin., Justin Sheffield, and Eric, F. Wood., Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching, *J. Geophys.Res.,* 2010,doi:10.1029/2009JD012882.

[6] Kannan, S., and S. Ghosh , A nonparametric kernel regression model for downscaling multisite daily precipitation in the Mahanadi basin, *Water Resour. Res.,* 49, 2013, doi:10.1002/wrcr.20118.

[7] Kannan, S., and S. Ghosh, Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output, *Stochastic Environ.Res. Risk Assess.*, 25, 457–474.2011, doi:10.1007/s00477-010-0415-y.

[8] Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deavan, L. Gandin, M. Iredell, S. Saha, G. White, J. Wollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K.C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, D. Joseph The NCEP/NCAR 40-years reanalysis project, *Bull. Amer. Meteor. Soc.,* 77(3), 437471,1996.

[9] Mehrotra, R., and A. Sharma , A nonparametric stochastic downscaling framework for daily rainfall at multiple locations, *J. Geophys. Res.*, 111, D15101, 2006,doi:10.1029/2005JD006637

[10] Lall, U. and A.Sharma , A nearest neighbor bootstrap for time series resampling, *Water Resour. Res.*, 32(3): 679-693, 1996.

[11] Wand, M.P., and M.C.Jones , *Kernel Smoothing* Chapman and Hall London 1995.

[12] Nadayara E.A. on estimating regression *Theor.Probob. Appl*.10, 186-190, 1964

[13] Parthasarathy, B., K. Rupa Kumar, and A. A. Munot, Homogeneous regional summer monsoon rainfall over India : Interannaul variability and teleconnections Research Report no. RR-070, 1996

K.Shashikanth, presently working as Asst.Professor in the Department of Civil Engineering at Osmania University and pursuing his doctoral program in Indian Institute of Technology Bombay. Subimal Ghosh presently working as Asst. Professor in the Department of Civil Engineering Indian Institute of Technology Bombay, Mumbai, India. Main areas of research includes statistical downscaling, uncertainty modeling, impacts of urbanization and Urban Heat island on rainfall, Data driven methods for analysis of hydro-climatic data.