

Future rainfall scenario over Orissa with GCM projections by statistical downscaling

Subimal Ghosh and P. P. Mujumdar*

Department of Civil Engineering, Indian Institute of Science, Bangalore 560 012, India

The article presents a methodology for examining future rainfall scenario using fuzzy clustering technique from the General Circulation Model (GCM) projections. GCMs might capture large-scale circulation patterns and correctly model smoothly varying fields such as surface pressure, but it is extremely unlikely that these models properly reproduce nonsmooth fields such as precipitation. The model developed in the present study is a linear regression model for estimation of rainfall, using GCM outputs of mean sea-level pressure and geopotential height as explanatory variables/regressors. To reduce the dimensionality of the dataset, the Principal Component Analysis (PCA) is used. Fuzzy clustering technique is applied to classify the principal components identified by the PCA and the fuzzy membership values are used in the regression model, with an assumption that the effects of circulation patterns on precipitation in different clusters are different. The regression model is then modified with an appropriate seasonality term. A major advantage of the proposed methodology is that while being computationally simple, it can model rainfall with a high goodness-of-fit (R^2) value. The methodology is applied to forecast monthly rainfall over Orissa.

Keywords: Fuzzy clustering, General Circulation Model, Orissa, Principal Component Analysis, rainfall.

GENERAL Circulation Models (GCMs) are tools designed to simulate time series of climate variables globally, accounting for effects of greenhouse gases (GHGs) in the atmosphere¹. They attempt to represent the physical processes in the atmosphere, ocean, cryosphere and land surface. GCMs are currently the most credible tools available for simulating the response of the global climate system to increasing greenhouse gas concentrations, and to provide estimates of climate variables (e.g. air temperature, precipitation, wind speed, pressure, etc.) on a global scale. They are good for the prediction of large-scale circulation patterns, but unfortunately precipitation, which is the main input in hydrologic models, cannot be well modelled by GCMs². Another drawback of GCMs is that the spatial scale on which a GCM can operate (e.g. 3.75° long. \times 3.75° lat. for Coupled Global Climate Model (CGCM2)) is coarse compared to that of the hydrological process (e.g. precipitation in a region, streamflow

in a river, etc.) to be modelled in the climate change impact assessment studies¹. Methodologies to model the hydrologic variables (e.g. precipitation) at a smaller scale based on large-scale GCM outputs are known as downscaling. They include dynamic downscaling, which uses complex algorithms at a fine grid-scale (typically of order of $50 \text{ km} \times 50 \text{ km}$) describing atmospheric process nested within the GCM outputs³ (commonly known as Limited Area Models or Regional Climate Models (RCM)) and statistical downscaling, that produces future scenarios based on statistical relationship between large-scale climate features and hydrologic variables like precipitation^{4,5}. The assumption of statistical downscaling is that there are certain physical relationships underlying the statistical relationships developed, and these physical relationships hold, regardless of whether the model simulation is a control (stationary) experiment or an experiment incorporating changed climate⁶. Compared to dynamic downscaling, statistical downscaling has the advantage of being computationally simple and easily adjusted to new areas. It requires few parameters and this makes it attractive for many hydrological applications⁷. A comparative study of statistical and dynamic downscaling may be found in Murphy⁸. Detailed discussions on different models used for downscaling GCM outputs may be found in Prudhomme *et al.*⁹. A brief overview of statistical downscaling models developed earlier and used to study climate change impact on hydrology is now presented.

Models are available on downscaling based on classification of circulation patterns (CP) and using this classified CP in estimation of precipitation. Bardossy *et al.*¹⁰ used a fuzzy rule-based technique for classification of CP into different states. Stochastic models such as Markov chains may be used to model rainfall from different states of classified circulation patterns^{11,12}. Semi-Markov chain was used by Bardossy and Plate¹³ to model daily rainfall incorporating duration of the state of CP from daily GCM outputs. Hughes *et al.*¹⁴ used Classification and Regression Tree (CART) to classify the principal components obtained from CP into different weather states. Hughes and Guttorp¹⁵ have used Nonhomogeneous Hidden Markov Model (NHMM) to downscale CP, as obtained from GCM. The underlying feature of NHMM is the hypothesis of an unobserved weather state which transcends the differences in scales between the two processes, circulation patterns and precipitation. The weather state is not explicitly defined a priori; rather, the model attempts to find distinctive patterns in the at-

*For correspondence. (e-mail: pradeep@civil.iisc.ernet.in)

mospheric data that are predictive of particular patterns in the hydrologic process¹⁵. Wetterhall *et al.*¹⁶ used analogue method for downscaling GCM output of circulation patterns to rainfall. A hard clustering-based analogue method for short-term weather forecasting may be found in Gutierrez *et al.*¹⁷. A review of different models used to simulate the effects of climate change on water resources is presented by Leavesley¹⁸.

In this study, a linear regression model is used to downscale the GCM outputs for estimation of monthly rainfall over Orissa. The methodology is based on fuzzy clustering technique. Appropriate seasonal component is added to the regression model for improving the goodness of fit. Development of the methodology is presented in the following sections.

Data

The GCM model used for the analysis is Coupled Global Climate Model (CGCM2), developed in Canadian Center for Climate Modelling and Analysis (CCCma). IPCC-IS92a scenario is selected for the estimation of monthly rainfall in Orissa. The CCCma-coupled global climate model, CGCM2, represents the net radiative effect of all greenhouse gases (GHGs) by means of an equivalent CO₂ concentration. The equivalent CO₂ concentration is necessarily higher than the observed CO₂ concentration, since it represents the climate forcing due to CO₂ and also the forcing associated with all other GHGs. In transient climate change simulations, the change in GHG forcing is represented in the model as a perturbation relative to the 330 ppmv equivalent CO₂ concentration, used in the control simulation, i.e. 330 ppmv is taken as a reference value and climate change simulations involve changes relative to this value (http://www.cccma.bc.ec.gc.ca/data/cgcm/cgcm_forcing.shtml). The first report by Intergovernmental Panel on Climate Change (IPCC) was published in the year 1992, which describes six alternative scenarios (IS92a to f). These scenarios embodied a wide array of assumptions affecting how future GHG emission might evolve. Out of these scenarios, IS92a was widely adopted by the scientific community during the last decade. According to the scenario, population rises to 11.3 billion by 2100 and economic growth averages 2.3% per annum between 1990 and 2100, with a mix of conventional and renewable energy sources being used. CGCM2 output used in the present analysis considers IPCC-IS92a forcing scenario in which the change in GHG forcing corresponds to that observed from 1900 to 1990 and increases at a rate 1% per year thereafter, until 2100. The direct effect of sulphate aerosols is also included.

Although GCM runs are available at timescales as small as 15 minutes, there is little confidence in GCM outputs for timescales shorter than 1 month⁹. The selection of appropriate predictor or characteristics from the large-

scale atmospheric circulation, is one of the most important steps in downscaling using statistical methods. The predictors used for downscaling should be¹⁹ (1) reliably simulated by GCMs, (2) readily available from archives of GCM outputs, and (3) strongly correlated with the surface variables of interest. GCMs are more accurate for free atmospheric variables such as air pressure. Precipitation can be related to air mass transport and thus related to atmospheric circulation, which is a consequence of pressure differences and anomalies². Based on the literature^{2,10-16}, monthly data of mean sea-level pressure and geopotential height at 500 mb are used as predictors.

Data from January 1950 to December 2099 are extracted from the official website of CCCma (<http://www.cccma.bc.ec.gc.ca/>). Twelve grid points are selected for the study ranging from 16.70 to 24.12°N lat. and 78.75 to 90.00°E in long. Figure 1 shows the grid points superposed on the map of Orissa. The dimension of the GCM output dataset extracted is $12 \times 2 = 24$ (mean sea-level pressure and geopotential height (500 mb) at each of the twelve grid points). Multi-dimensionality of the predictors may lead to a computationally complicated and large-sized model with high multicollinearity (high correlation between the explanatory variables/regressors). To reduce the dimensionality of the explanatory dataset, Principal Component Analysis (PCA) is performed.

PCA is a statistical procedure to identify the patterns of multidimensional variables and to transfer correlated variables into a set of uncorrelated variables. Starting with the set of twenty-four variables (GCM outputs), the method generates a new set of variables called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information.

For performing principal component analysis²⁰, first the covariance matrix of the normalized variables is computed.

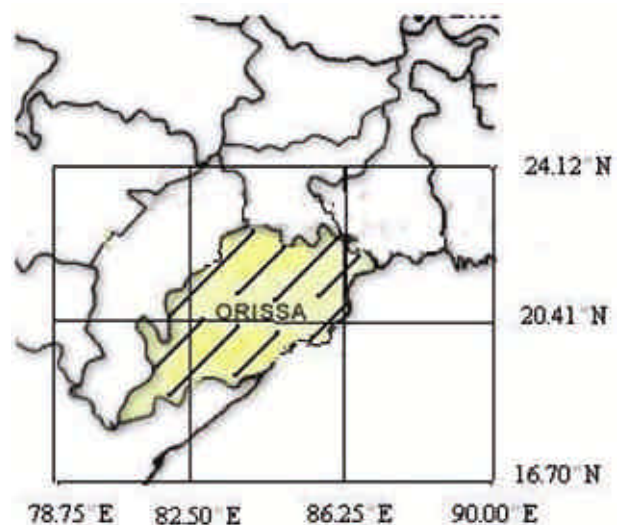


Figure 1. GCM grids superposed on map of Orissa.

Each variable is normalized by subtracting the mean from it and then dividing the result by the standard deviation of the original variable. Eigen vectors resulting from the covariance matrix are used for PCA. The eigen vectors are orthonormal and the indices are arranged so that the first eigen vector corresponds to the largest eigen value and in general the k th eigen vector to the k th largest eigen value I_k . The k th principal component at time t (pc_{kt}) is computed as:

$$pc_{kt} = \sum_q e_{kq} [(p_t(q) - \bar{p}(q))/S(q)], \tag{1}$$

where $p_t(q)$ is the value of q th variable (mean sea-level pressure/geopotential height at any node) at time t . $\bar{p}(q)$ and $S(q)$ are the mean and standard deviation of the variable $p(q)$. e_{kq} is the q th element of the eigen vector corresponding to k th eigen value. The percentage of total variance w_k explained by the k th principal component is given by:

$$w_k = \frac{I_k}{\sum_{m=1}^M I_m} \times 100, \tag{2}$$

where M is the dimensionality of the original dataset (twenty-four in the present study).

The advantage of PCA is that using a small number of principal components, it is possible to represent the variability of the original multivariate dataset. In the present study the number of principal components, which can together preserve more than 97% of the total variance of original dataset is used.

The monthly area-weighted rainfall data of Orissa, which extends from January 1950 to December 2003, is extracted from the website of the Indian Institute of Tropical Meteorology, Pune (<http://www.tropmet.res.in>). This dataset is used for regression analysis. Primary source of data is India Meteorological Department (IMD). Details of development of methodology for modelling rainfall using these datasets are presented in the following section.

Model development

Linear Regression Model is used to predict the monthly rainfall in Orissa using the principal components as explanatory variables. For IS92a scenario, the first three principal components together preserve more than 97% of the variability of the original 24-dimensional dataset (Table 1). Only the first three principal components are used as explanatory variables with the constant term in the linear regression model, which results in the following equation with R^2 value of 0.676.

$$\begin{aligned} \text{RAIN}_t = & 688.354 + 312.132 \times pc_{1t} + 114.136 \\ & \times pc_{2t} - 347.795 \times pc_{3t}, \end{aligned} \tag{3}$$

Table 1. Percentage of variance explained by principal components

Eigenvalue	Percentage variance explained
12.7076	52.9483
10.5460	43.9415
0.4861	2.0254
0.0999	0.4161
0.0657	0.2740
0.0291	0.1214
0.0209	0.0872
0.0158	0.0659
0.0076	0.0317
0.0073	0.0305
0.0047	0.0198
0.0026	0.0108
0.0019	0.0079
0.0015	0.0062
0.0011	0.0044
0.0007	0.0028
0.0005	0.0022
0.0003	0.0013
0.0003	0.0011
0.0001	0.0006
0.0001	0.0004
0.0001	0.0004
≈ 0.0000	≈ 0.0000
≈ 0.0000	≈ 0.0000

where RAIN_t is the monthly rainfall of Orissa in 10^{-1} mm/month. The same regression model is applied without a constant term using SPSS-9.05, a data-modelling tool. It gives R^2 value as 0.797, which is a measure of the proportion of the variability of the dependant variable about the origin explained by the regression model.

$$\begin{aligned} \text{RAIN}_t = & 324.753 \times pc_{1t} + 289.706 \times pc_{2t} \\ & - 707.706 \times pc_{3t}. \end{aligned} \tag{4}$$

The R^2 value thus evaluated in this regression model is unsatisfactory. The reason behind this may be the existence of different classes of atmospheric circulation patterns, and also the relationships of rainfall with the principal components different for different classes or seasons. Studies are available^{2,9-15} in support of the existence of different classes in the circulation patterns. Methods based on Classification and Regression Trees (CART), fuzzy rule-based systems have been applied to classify the atmospheric circulation in these studies. In the present analysis, the atmospheric circulation pattern is classified based on clustering technique.

Overview of clustering method

Clustering refers to partitioning of a dataset into a number of classes. The objective of clustering technique is to minimize the Euclidean distance between each datapoint in a cluster and its cluster centre, and to maximize the Euclid-

ean distance between cluster centres²¹. There are two broad methods of clustering: hard clustering; and fuzzy clustering. Hard clustering is used to classify data in a crisp sense. By this method each datapoint will be assigned to one and only one data cluster. If c partitions/classes/clusters of a sample set X , having n data samples, can be defined as a family of sets $\{A_i, i = 1, 2, \dots, c\}$, then the following set theoretic properties for hard clustering can be observed²¹:

$$\bigcup_{i=1}^c A_i = X, \tag{5}$$

$$A_i \cap A_j = \mathbf{f} \quad \forall i \neq j, \tag{6}$$

$$\mathbf{f} \subset A_i \subset X \quad \forall i, \tag{7}$$

where $2 \leq c < n$. $c = 1$ places all data samples into the same class and $c = n$ places each datum into its own class; neither case requires any effort in classification and is also of any use. Equation (5) expresses the fact that the set of all classes exhausts the universe of data samples. Equation (6) indicates that none of the classes overlaps, i.e a data sample cannot belong to more than one class. Equation (7) indicates that a class cannot be empty and it cannot contain all the data samples. Application of hard clustering in climatic sciences may be found in Gadgil and Iyengar²⁰.

In the fuzzy clustering technique, the crisp classification is extended to fuzzy classification using the concept of membership values. A family of fuzzy sets/clusters/classes may be defined as $\{\tilde{A}_i, i = 1, 2, \dots, c\}$, which differs from the crisp sets generated by hard clustering by the following properties²¹:

$$\tilde{A}_i \cap \tilde{A}_j \neq \mathbf{f}. \tag{8}$$

Membership values are assigned to the various datapoints for each fuzzy set/cluster/class. Hence a single point can have partial membership in more than one class. The membership value, \mathbf{m}_i , of the t th datapoint x_t in the i th class has the following restrictions²¹:

$$\mathbf{m}_i = \mathbf{m}_{\tilde{A}_i}(x_t) \in [0, 1], \tag{9}$$

$$\sum_{i=1}^c \mathbf{m}_i = 1. \tag{10}$$

Equation (10) ensures that the sum of membership values for a single datapoint in all classes is unity. The goal of the fuzzy clustering is to determine the clusters with their centres and to compute the membership value of all datapoints in each class. For example, at any time period t , the datapoint made up of the first three principal components (i.e. pc_{1t} , pc_{2t} and pc_{3t}) will have c membership values, one in each of the c clusters, with values ranging between

0 and 1 (eq. (9)) and with a sum of the c membership values equal to 1 (eq. (10)). Details of algorithms of hard clustering and fuzzy clustering may be found in Ross²¹.

Hard clustering technique is not used in the present analysis because a slight change in the mean sea-level pressure or geopotential height lying in a particular class, using this method may lead to a different class with a different regression equation, which may not be realistic. Furthermore, the circulation pattern generated by GCM for the future may constitute a new class, having few members of the past and present circulation patterns. This may lead to an erroneous regression model. On the other hand, fuzzy clustering assigns membership values of the classes to various datapoints, and it is more generalized and useful to describe a point not by a crisp cluster, but by its membership values in all the clusters. A brief overview of fuzzy c -means clustering, an algorithm widely used for fuzzy clustering, is given in Appendix 1. For the present study, three clusters have been used and membership of all the datapoints in each of the three clusters has been calculated. These values are used in the regression equation for modelling rainfall.

Regression with cluster membership

Regression analysis is performed using membership values obtained from the fuzzy clustering method. First, it is assumed that for different clusters only the intercept (constant terms) of the regression equation is different. The following equation is used for regression analysis:

$$\text{RAIN}_i = \sum_{i=1}^2 \mathbf{b}_i \times \mathbf{m}_i + \sum_{k=1}^3 \mathbf{g}_k \times pc_{kt}. \tag{11}$$

\mathbf{b}_i and \mathbf{g}_k are the coefficients of μ_{it} (i.e. membership in cluster i of t th data) and pc_{kt} . The membership values μ_{it} in each cluster are assigned to different points based on fuzzy c -means algorithm (Appendix 1, eqs (27)–(31)). These membership values lie between 0 and 1. The centre of a cluster will have a membership value of 1 in that cluster, and 0 in the other clusters. To demonstrate the effect of inclusion of membership values in the regression model, equations for the centres of different clusters are given here. The regression equation for the centre of cluster 1 having complete membership in cluster 1 ($\mathbf{m}_1 = 1$; $\mathbf{m}_2 = 0$; $\mathbf{m}_3 = 0$) can be given by:

$$\text{RAIN}_i = \mathbf{b}_1 + \sum_{k=1}^3 \mathbf{g}_k \times pc_{kt}. \tag{12}$$

The constant term/intercept in this regression equation is \mathbf{b}_1 . Similarly, for the centre of cluster 2 ($\mathbf{m}_1 = 0$; $\mathbf{m}_2 = 1$; $\mathbf{m}_3 = 0$), the intercept term can be given by \mathbf{b}_2 (eq. (13)).

$$\text{RAIN}_t = \mathbf{b}_2 + \sum_{k=1}^3 \mathbf{g}_k \times pc_{kt}. \tag{13}$$

The equation for the centre of cluster 3 ($\mathbf{m}_1 = 0$; $\mathbf{m}_2 = 0$; $\mathbf{m}_3 = 1$) can be given by:

$$\text{RAIN}_t = \sum_{k=1}^3 \mathbf{g}_k \times pc_{kt}. \tag{14}$$

For all of the cases shown in eqs (12)–(14), the slopes will be the same (i.e. \mathbf{g}_1 , \mathbf{g}_2 and \mathbf{g}_3) for pc_{1t} , pc_{2t} and pc_{3t} respectively). Using this equation R^2 value is obtained as 0.811, which is better than the regression equation with only the principal components as the explanatory variables ($R^2 = 0.797$). The values of \mathbf{b}_1 , \mathbf{b}_2 , \mathbf{g}_1 , \mathbf{g}_2 and \mathbf{g}_3 are obtained as 1051.839, 136.782, 411.611, 143.749 and -491.605 respectively. The assumption of difference only in the intercept term is now relaxed and it is considered that both the slope and intercept terms are different for different clusters. The following regression equation is used to predict the rainfall:

$$\text{RAIN}_t = \sum_{i=1}^2 \mathbf{b}_i \times \mathbf{m}_t + \sum_{k=1}^3 \mathbf{g}_k \times pc_{kt} + \sum_{i=1}^2 \sum_{k=1}^3 \mathbf{r}_{ik} \times \mathbf{m}_t \times pc_{kt}. \tag{15}$$

A new set of coefficients \mathbf{r}_{ik} is introduced in the equation, considering different slopes for different clusters, in the linear regression model. For the datapoint having full membership in cluster 1 ($\mathbf{m}_1 = 1$; $\mathbf{m}_2 = 0$; $\mathbf{m}_3 = 0$), the regression equation will be:

$$\text{RAIN}_t = \mathbf{b}_1 + \sum_{k=1}^3 (\mathbf{g}_k + \mathbf{r}_{1k}) \times pc_{kt}. \tag{16}$$

Similarly, the regression equations for datapoint having full membership in cluster 2 ($\mathbf{m}_1 = 0$; $\mathbf{m}_2 = 1$; $\mathbf{m}_3 = 0$) and cluster 3 ($\mathbf{m}_1 = 0$; $\mathbf{m}_2 = 0$; $\mathbf{m}_3 = 1$) respectively, are as follows:

$$\text{RAIN}_t = \mathbf{b}_2 + \sum_{k=1}^3 (\mathbf{g}_k + \mathbf{r}_{2k}) \times pc_{kt} \quad \text{for } \mathbf{m}_2 = 1 \tag{17}$$

$$\text{RAIN}_t = \sum_{k=1}^3 \mathbf{g}_k \times pc_{kt} \quad \text{for } \mathbf{m}_3 = 1. \tag{18}$$

The results of the regression analysis is given in Table 2. It gives an R^2 value as 0.841, a better one than that obtained from regression eq. (11); i.e. 0.811. To improve the model fitness or goodness-of-fit, the seasonal component is now introduced in the model. Details are presented as follows.

Table 2. Results for regression analysis considering only cluster membership ($R^2 = 0.841$)

Coefficients	Value	<i>t</i> -statistics	Significance
\mathbf{b}_1	299.989	1.274	0.203
\mathbf{b}_2	404.085	0.624	0.533
\mathbf{g}_1	527.232	12.323	0.000
\mathbf{g}_2	184.369	2.365	0.018
\mathbf{g}_3	-249.320	-2.060	0.040
\mathbf{r}_{11}	-491.716	-6.766	0.000
\mathbf{r}_{12}	-195.553	-1.962	0.050
\mathbf{r}_{13}	320.759	1.652	0.099
\mathbf{r}_{21}	-668.167	-3.789	0.000
\mathbf{r}_{22}	-3.925	-0.011	0.991
\mathbf{r}_{23}	-952.224	-1.652	0.099

Modification for seasonal component

The coefficient values in the regression equation should have some seasonal/periodic component. This should be considered and properly incorporated for correct estimation of rainfall. In the present analysis all the coefficients used in the regression equation (eq. (15)) are rewritten in terms of seasonal component. The seasonal component is assumed to be different for different months with a periodicity of 12.

$$\mathbf{b}_i = \mathbf{b}_i^0 + \mathbf{b}_i^1 \times (2\mathbf{p} t/12) + \mathbf{b}_i^2 \times \cos(2\mathbf{p} t/12), \tag{19}$$

where t is the serial number of the datapoint. The equation will take care of the seasonal/periodical term. Similarly, the modifications for other coefficients will be as follows:

$$\mathbf{g}_k = \mathbf{g}_k^0 + \mathbf{g}_k^1 \times \sin(2\mathbf{p} t/12) + \mathbf{g}_k^2 \times \cos(2\mathbf{p} t/12), \tag{20}$$

$$\mathbf{r}_{ik} = \mathbf{r}_{ik}^0 + \mathbf{r}_{ik}^1 \times \sin(2\mathbf{p} t/12) + \mathbf{r}_{ik}^2 \times \cos(2\mathbf{p} t/12). \tag{21}$$

The modified regression equation is fitted to model the monthly rainfall. Details of the model results are given in Table 3. The model is improved by incorporating the seasonal component with a good R^2 value of 0.900. But the major drawback of this regression equation is that the t -statistics is insignificant for the coefficients of most of the explanatory variables/regressors, and with 95% confidence, we cannot reject the null hypothesis that these coefficients are significantly different from zero. It may also lead to overfitting of the model. Furthermore, there may exist a high correlation between the regressors used in the regression equation leading to multicollinearity. It can be tested by the condition index, defined as the ratio of maximum to minimum eigen values of the matrix formed by the explanatory variables. A thumb rule²² for any linear regression model without multicollinearity, is that the condition index should be less than 30. A high condition index of 253.033 in the present case indicates that there exists multicollinearity, which may lead to high standard error in the estimation of coefficients. The following section

presents a methodology based on F -statistics, for screening of regressors used in the regression equation, without a significant loss in R^2 value.

Selection of explanatory variables using F-statistics

To get over this problem the backward method of SPSS-9.05 (a data modelling tool) based on F -statistics is used, which removes regressors having insignificant t -statistics one by one from the regression equation, without a significant change in R^2 value. The significance of the change in R^2 value is tested with a statistical measure (F), which follows F -distribution.

$$F = \frac{(R_1^2 - R_2^2) / n_1}{(1 - R_1^2) / df}, \tag{22}$$

Table 3. Results for regression analysis considering seasonality ($R^2 = 0.900$)

Coefficients	Value	t -statistics	Significance
b_1^0	-852.451	-1.057	0.291
b_1^1	260.944	0.254	0.800
b_1^2	615.669	0.680	0.497
b_2^0	5109.470	3.646	0.000
b_2^1	-230.709	-0.104	0.917
b_2^2	-2816.291	-1.290	0.198
g_1^0	-243.256	-1.237	0.217
g_1^1	-884.011	-5.651	0.000
g_1^2	-695.669	-3.550	0.000
g_2^0	262.348	0.952	0.341
g_2^1	284.945	1.072	0.284
g_2^2	284.812	0.954	0.341
g_3^0	888.144	1.638	0.102
g_3^1	1278.630	2.399	0.017
g_3^2	973.748	1.899	0.058
r_{11}^0	45.085	0.159	0.874
r_{11}^2	716.607	2.777	0.006
r_{11}^0	831.383	2.416	0.016
r_{12}^0	-50.236	-0.139	0.889
r_{12}^1	-531.498	-1.580	0.115
r_{12}^2	-499.944	-1.569	0.117
r_{13}^0	-1446.770	-1.963	0.050
r_{13}^1	-670.092	-0.979	0.328
r_{13}^2	-466.058	-0.635	0.526
r_{21}^0	346.953	0.515	0.607
r_{21}^1	1951.155	2.876	0.004
r_{21}^2	1784.396	3.506	0.000
r_{22}^0	-383.565	-0.419	0.676
r_{22}^1	-1647.925	-1.441	0.150
r_{22}^2	-309.752	-0.316	0.752
r_{23}^0	-2312.361	-1.222	0.222
r_{23}^1	-3171.496	-1.724	0.085
r_{23}^2	-537.540	-0.266	0.791

where R_1^2 is the R^2 value without any restriction (i.e. with all the regressors), R_2^2 is the R^2 value with restrictions (i.e.

Table 4. Results of the forecast model ($R^2 = 0.898$)

Coefficients	Value	t -statistics	Significance
b_2^0	2956.666	7.711	0.000
b_2^1	-2060.101	-5.462	0.000
g_1^1	-684.799	-28.392	0.000
g_1^2	-422.772	-18.805	0.000
g_3^1	359.749	4.385	0.000
g_3^2	296.648	2.782	0.006
r_{11}^1	577.439	13.134	0.000
r_{11}^2	427.741	14.055	0.000
r_{13}^0	-490.832	-3.856	0.000
r_{21}^1	1433.725	5.658	0.000
r_{21}^2	1021.475	-3.900	0.000
r_{22}^1	-610.797	-1.441	0.000

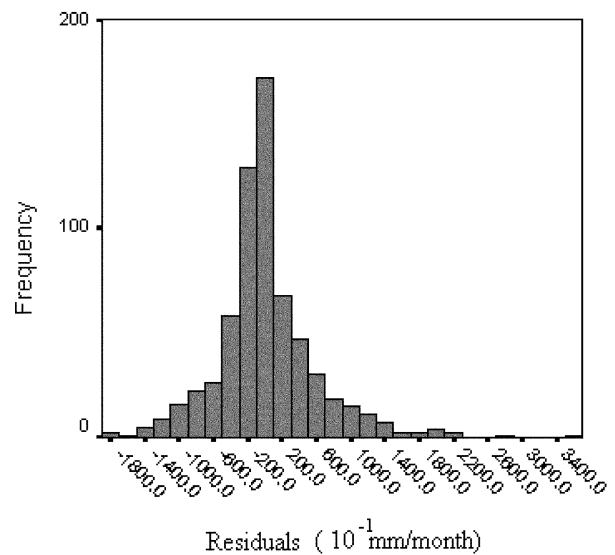


Figure 2. Frequency distribution of residuals.

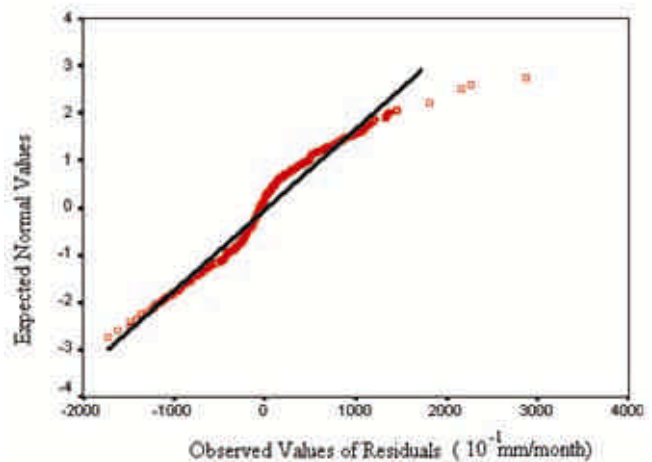


Figure 3. Normal Q-Q plot of residuals.

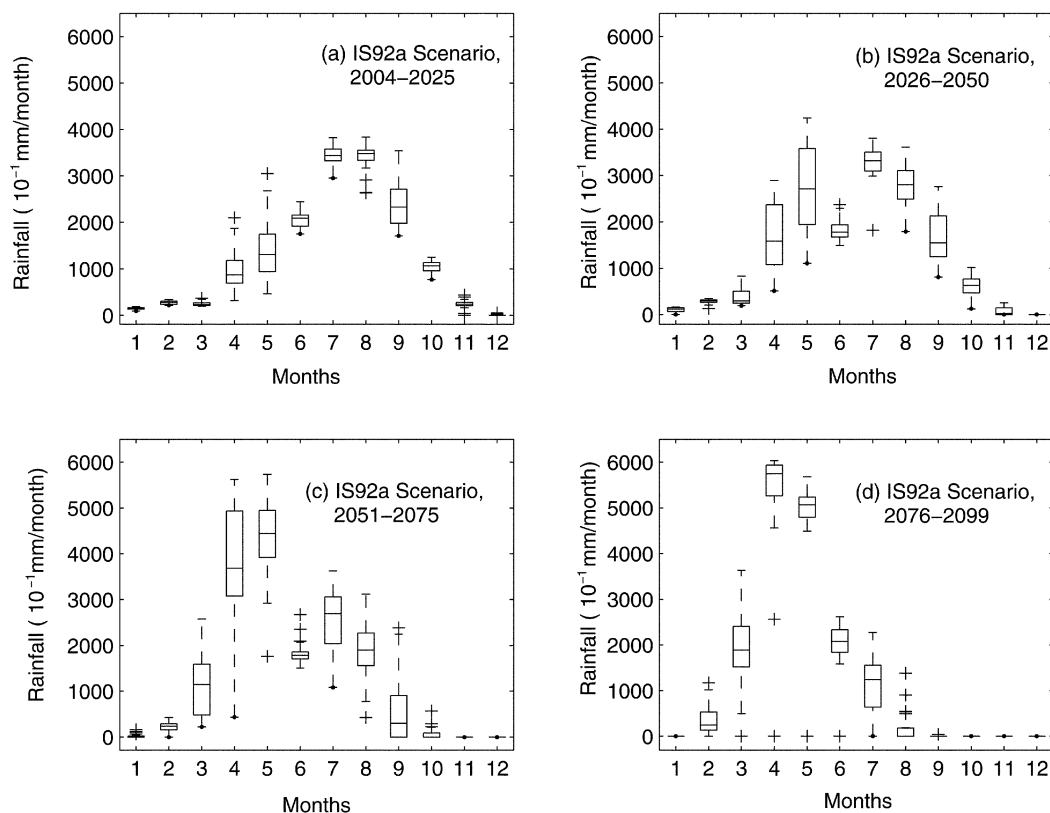


Figure 4. Box-plot for monthly predicted rainfall.

after removing some insignificant regressors), n_1 is the number of regressors removed, and df is the degree of freedom of the unrestricted model. The resulting F -statistics determines which of the regressors should be removed without significant loss in R^2 value.

Results for the coefficients of variables selected for the regression model with the backward method are given in Table 4. The final R^2 value is obtained as 0.898 which is a satisfactory one.

The condition index for the model is obtained as 7.524 which is less than 30, signifying that there is no multicollinearity. For testing autocorrelation in the residuals, Durbin–Watson test is performed. The Durbin–Watson statistics (d) of a regression model is given by:

$$d = \left(\sum_{t=2}^n (u_t - u_{t-1})^2 \right) / \sum_{t=2}^n u_t^2, \quad (23)$$

where u_t is the t th residual. For a model without a significant autocorrelation between the residuals, the d value should be close to 2 and it should lie between d_L and $4 - d_U$. The values of d_L and d_U depend on the number of observations and number of explanatory variables. In the present study, the d value is 1.875. The number of observations used here is 648. For any regression having observations more than 200 the standard d_L and d_U values are not

available, but here the d value is close to 2, which signifies that there is no significant autocorrelation in the residuals. Breusch–Godfrey test for autocorrelation also reveals that autocorrelation is absent among the residuals. Breusch–Pagan–Godfrey (BPG) test is performed for heteroscedasticity (variance of the residuals not being constant). As the pattern of residuals for different clusters is different, the BPG test reveals heteroscedasticity. But, heteroscedasticity has never been a reason to throw out an otherwise good model^{22,23}. The frequency distribution of the residuals is plotted in Figure 2, which shows that the distribution closely resembles a normal distribution. Figure 3 shows the normal $Q-Q$ plot of the residuals supporting the fit of normal distribution for the residuals. Based on these tests and the R^2 value of 0.898, the model given by eqs (15), (19)–(21) with values of coefficients given in Table 4, is selected for projection of monthly rainfall over Orissa in future.

Future rainfall scenario

The methodology based on regression and fuzzy clustering thus developed is used for modelling monthly rainfall in Orissa for 2004 to 2099 for IS92a scenario, with a basic assumption that this regression relationship will not change in the future. A negative value generated by the regression

is set to zero. The box-plot for different months has been plotted for the periods 2004–25, 2026–50, 2051–75, and 2076–99 (Figure 4). Figure 4 shows that there is a possibility of decrease in rainfall during the dry period (September–February). The summer and monsoon (March–August) rainfall has an increasing trend, with an increase in the maximum/peak rainfall. A possibility of increase in hydrologic extremes (droughts/floods) may be indicated by these results, based on the IS92a scenario of CCCma, CGCM-2. A recent study²⁴ on the impact of future climate change on Indian summer monsoon revealed that there is a possibility of decline in all-India rainfall in the winter season, which may lead to droughts. Also the possibility of increase in summer/monsoon rainfall is predicted in that analysis. The present analysis results in similar rainfall scenario for Orissa. Figure 5 *a* and *b* shows trends in the predicted dry season and wet season rainfall. Under this scenario, severe drought conditions are indicated during the period 2076–99.

Conclusion

The methodology described here for predicting future rainfall from GCM outputs, is based on linear regression and fuzzy clustering, and is computationally simple. Use of fuzzy clustering overcomes the limitation of rigidity of hard clusters. Membership values obtained from fuzzy clustering are used as dummy variables in regression analysis. Seasonal components have been taken care by appropriate sine and cosine components. SPSS-9.05 is used to perform the backward regression analysis based on *F*-statistics. The required tests like Durbin–Watson test, multicollinearity test, etc. have been performed to validate the assumptions of the linear regression model. The goodness-of-fit (*R*²) value of the regression model is comparable to similar

models used for downscaling of rainfall. The rainfall projection, based on IS92a scenario, shows that there is a possibility of increase in hydrologic extremes in Orissa in future. High value of goodness-of-fit and different tests on the assumptions on linear regression (multicollinearity, normality of residuals, etc.) suggest that this model can be used to realistically simulate precipitation at regional scale, and hence can be used for climate change impact studies.

Appendix 1. Fuzzy *c*-means clustering algorithm.

A brief overview of fuzzy *c*-means clustering algorithm²¹ is presented here. To determine the fuzzy partition matrix \tilde{U} for grouping a collection of *n* datasets with *m* dimensions/coordinates into *c* classes, an objective function *J_m*, which is to be minimized, is defined as:

$$J_m(\tilde{U}, \mathbf{v}) = \sum_{t=1}^n \sum_{i=1}^c (\mathbf{m}_{it})^{m'} (d_{it})^2, \tag{24}$$

$$d_{it} = \left[\sum_{k=1}^m (x_{kt} - v_{ik})^2 \right]^{1/2}. \tag{25}$$

Subscript *i* and *k* stand for the *i*th cluster and *k*th coordinate respectively. *x_{kt}* is the *k*th coordinate of the *t*th data-point. *m'* is a parameter called weighting factor and has a range [1, ∞). This parameter controls the amount of fuzziness in the classification process. In general, higher the *m'*, the fuzzier are the membership assignments of the clustering. Conversely, as *m' → 1*, the clustering values become hard, i.e. 0 or 1. There is no theoretical optimum choice of *m*, but it is suggested to use a value²¹ in between 1.25 and 2.00. **v** denotes the set of cluster centres. The *i*th cluster centre **v_i** can be described by *m* features (*m* coordinates) in a vector form **v_i** = {*v_{i1}*, *v_{i2}*, . . . , *v_{ik}*, . . . , *v_{im}*}. *v_{ik}* can be defined as:

$$v_{ik} = \frac{\sum_{t=1}^n \mathbf{m}_{it}^{m'} \cdot x_{kt}}{\sum_{t=1}^n \mathbf{m}_{it}^{m'}}. \tag{26}$$

An effective algorithm for fuzzy classification called iterative optimization, was proposed by Bezdek²⁵. The steps of the algorithm are as follows²¹:

- (i) Fix *c* (2 ≤ *c* ≤ *n*) and select a value of parameter *m'*. Initialize the partition matrix $e \tilde{U}^{(0)}$. Each iteration of the algorithm will be labelled *r*, where *r* = 0, 1, 2,
- (ii) Calculate the *c* centres {**v_i**^(*r*)} for each step.

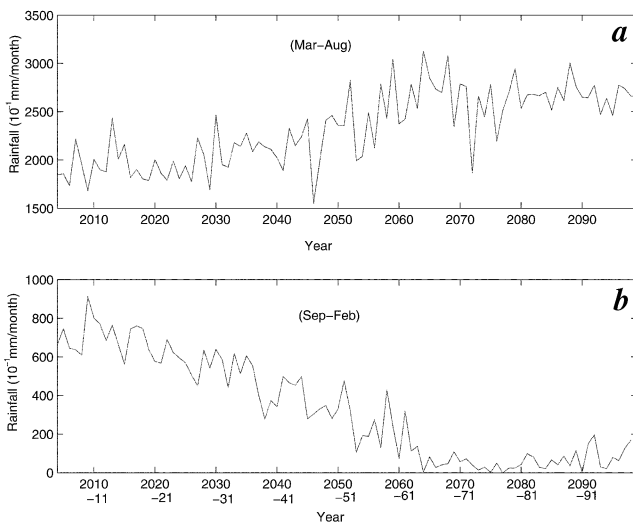


Figure 5. Predicted rainfall for wet (*a*) and dry (*b*) periods.

(iii) Update the partition matrix for the r th step, $\tilde{U}^{(r)}$ as follows:

$$\mathbf{m}_t^{(r+1)} = \left[\sum_{i=1}^c ((d_{it}^{(r)}) / (d_{it}^{(r)})^{2(m'-1)}) \right]^{-1} \text{ if } I_t = \mathbf{f}, \quad (27)$$

or

$$\mathbf{m}_t^{(r+1)} = 0 \quad \forall i \in I'_t, \quad (28)$$

where

$$I_t = \{i \mid 2 \leq c < n; d_{it}^{(r)} = 0\} \quad (29)$$

and

$$I'_t = \{1, 2, \dots, c\} - I_t \quad (30)$$

and

$$\sum_{j \in I_t} \mathbf{m}_t^{(r+1)} = \mathbf{1}. \quad (31)$$

(iv) If $\|U^{(r+1)} - U^{(r)}\| \leq \mathbf{e}_L$, stop; otherwise $r = r + 1$ and return to step (ii),

where \mathbf{e}_L is a small number used as a level of accuracy. In step (iii), eq. (27) will be invalid when the denominator of the fraction d_{it} will be zero. I_t and I'_t are used as a book-keeping parameter to handle this situation. Equations (29) and (30) are used to describe these parameters. In the present analysis, Fuzzy logic toolbox of MATLAB 6.5 is used for this algorithm.

1. Prudhomme, C., Jakob, D. and Svensson, C., Uncertainty and climate change impact on the flood regime of small UK catchments. *J. Hydrol.*, 2003, **277**, 1–23.
2. Bardossy, A., Downscaling from GCM to local climate through stochastic linkages. *J. Environ. Manage.*, 1997, **49**, 7–17.
3. Jones, P. D., Murphy, J. M. and Noguer, M., Simulation of climate change over Europe using a nested regional–climate model, I: Assessment of control climate, including sensitivity to location of lateral boundaries. *Q. J. R. Meteorol. Soc.*, 1995, **121**, 1413–1449.
4. Wilby, R. L., Hassan, H. and Hanaki, K., Statistical downscaling of hydrometeorological variables using general circulation model output. *J. Hydrol.*, 1998, **205**, 1–19.
5. Goldstein, J., Gachon, P., Milton, J. and Parishkura, D., Statistical downscaling models evaluation: A regional case study for Quebec regions, Canada. *Geophys. Res. Abstr.*, 2004, **6**, 04559.
6. Easterling, D. R., Development of regional climate scenarios using a downscaling approach. *Climate Change*, 1999, **41**, 615–634.

7. Wilby, R. L. *et al.*, Hydrological responses to dynamically and statistically downscaled climate model output. *Geophys. Res. Lett.*, 2000, **27**, 1199–1202.
8. Murphy, J., An evaluation of statistical and dynamical techniques for downscaling local climate. *J. Climate*, 1999, **12**, 2256–2284.
9. Prudhomme, C., Reynard, N. and Crooks, S., Downscaling of global climate models for flood frequency analysis: Where are we now? *Hydrol. Process.*, 2002, **16**, 1137–1150.
10. Bardossy, A., Duckstein, L. and Bogardi, I., Fuzzy rule-based classification of atmospheric circulation patterns. *Int. J. Climatol.*, 1995, **15**, 1087–1097.
11. Bardossy, A. and Plate, E. J., Space–time model for daily rainfall using atmospheric circulation patterns. *Water Resour. Res.*, 1992, **28**, 1247–1259.
12. Stehlik, J. and Bardossy, A., Multivariate stochastic downscaling model for generating daily precipitation series based on atmospheric circulation. *J. Hydrol.*, 2002, **28**, 1247–1259.
13. Bardossy, A. and Plate, E. J., Modeling daily rainfall using a semi-Markov representation of circulation pattern occurrence. *J. Hydrol.*, 1991, **122**, 33–47.
14. Hughes, J. P., Lettenmaier, D. P. and Guttorp, P., A stochastic approach for assessing the effect of changes in synoptic circulation patterns on gauge precipitation. *Water Resour. Res.*, 1993, **29**, 3303–3315.
15. Hughes, J. P. and Guttorp, P., A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resour. Res.*, 1993, **30**, 1535–1546.
16. Wetterhall, F., Halldin, S. and Xu, C., Statistical precipitation downscaling in Central Sweden with the analogue method. *J. Hydrol.*, 2005, **306**, 174–190.
17. Gutierrez, J. M., Cofino, A. S., Cano, R. and Rodriguez, M. A., Clustering methods for statistical downscaling in short-range weather forecasts. *Mon. Weather Rev.*, 2004, **132**, 2169–2183.
18. Leavesley, G. H., Modeling the effects of climate change on water resources – A review. *Climatic Change*, 1994, **28**, 159–177.
19. Wilby, R. L., Hay, L. E. and Leavesly, G. H., 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.
20. Gadgil, S. and Iyengar, R. I., Cluster analysis of rainfall stations of the Indian Peninsula. *Q. J. R. Meteorol. Soc.*, 1980, **106**, 873–886.
21. Ross, T. J., *Fuzzy Logic with Engineering Applications*, McGraw Hill International Edition, 1997, pp. 379–396.
22. Gujarati, D. N., *Basic Econometrics*, Tata McGraw Hill, 2004.
23. Mankiw, N. G., A quick refresher course in macroeconomics. *J. Econ. Lit.*, 1990, **XXVIII**, 1648.
24. Lal, M. *et al.*, Future climate change: Implications for Indian summer monsoon and its variability. *Curr. Sci.*, 2001, **81**, 1196–1207.
25. Bezdek, J., *Pattern Recognition with Fuzzy Objective Function Algorithm*, Plenum, New York, 1981.

ACKNOWLEDGEMENTS. The work reported here was carried out as a part of a project, sponsored by the Earth System Science Division of the Department of Science and Technology, New Delhi. The rainfall data used was obtained from the website of the Indian Institute of Tropical Meteorology, Pune (<http://www.tropmet.res.in>).

Received 27 June 2005; revised accepted 17 November 2005