Statistical downscaling of multisite daily precipitation for Tapi basin using kernel regression model

Sadhana Singh[1], S. Kannan[2] and P. V. Timbadiya[1,*]

[1]Department of Civil Engineering, S.V. National Institute of Technology-Surat, Surat 395 007, India
[2]Central Water and Power Research Station, Pune 411 024, India

The study presents fine resolution multisite daily precipitation projection for the Tapi basin using the kernel-regression (KR) based statistical downscaling methodology developed by Kannan and Ghosh with and without conditioned on the estimated rainfall state. The models are applied in downscaling of daily monsoon precipitation at a fine resolution of 0.25° comprising 351 grid points in and around the basin. The air temperature, specific humidity, zonal and meridional wind (at surface, 250, 500 and 850 hPa); mean sea level pressure and geopotential height at surface are utilized as predictors from five GCMs under CMIP-5 for two future scenarios, viz. RCP4.5 and RCP8.5. The performance of the downscaling model examined with respect to reproduction of various statistics for training period and indicated the better performance of KR model conditioned on the rainfall state than KR model without conditioned on the rainfall state of the basin. The KR model conditioned on the rainfall state is employed for future projections from GCMs outputs. The statistically downscaled daily precipitation from GCM (MPI-M) and CORDEX (COSMO-CLM) data is compared to quantify uncertainty. The statistically downscaled daily precipitation performs better than corresponding CORDEX data for the present study area. The study also revealed a possibility of decrease in the occurrences of extreme events with an increase in the medium rainfall events in the basin for future.

Keywords: Climate change, daily precipitation, general circulation models, statistical downscaling.

THE use of general circulation model (GCM) outputs (air temperature, pressure, specific humidity, wind speed) to quantify consequences of climate change on water resources of river basins has grown in the past decade. Recent studies[1–12] reported the projection of rainfall at basin/local/regional scale using statistical downscaling from GCM outputs. The output obtained using GCMs is accurate in simulating the large-scale atmospheric variables, but fails to model local-scale hydrologic variables such as precipitation[2,13], which necessitates the use of downscaling techniques. The downscaling methods are broadly classified into two types – statistical and dynamical. Details of these downscaling techniques and their comparisons have been widely reported[14–16]. As statistical downscaling is simple to compute and the area of domain can be easily transferred from one place to another is preferred over the dynamical downscaling. The success of any statistical downscaling method depends on how powerful the predictor is and the predictand relationship. However, the time invariant relationship between the predictor and the predictand is a major limitation of the statistical downscaling methods.

Further, the credibility of statistical downscaling under nonstationary climate is discussed by Salvi *et al.*[17]. Precipitation is a highly uncertain and heterogeneous spatial phenomenon, occurring as a result of compound interaction between different climate variables. Therefore, spatial modelling of precipitation is a challenging task, when modelling the multisite rainfall which suffers from the problem of inter-site cross-correlations. Kannan and Ghosh[18] successfully tackled the problem of multisite precipitation downscaling by means of the common rainfall state (pattern of the multisite rainfall) of a river basin. They have developed a model based on *k*-means clustering combined with supervized data classification technique, i.e. classification and regression tree (CART), for the rainfall state of a river basin using the large-scale atmospheric variables. Further, the non-parametric kernel regression (KR) model has been developed for the rainfall projections using the derived common rainfall state within a river basin to overcome the problem of inter-site cross-correlations. The aforesaid study concluded that KR-based downscaling model overcomes the limitations of other downscaling models developed in the past, including the problem of negative rainfall amount[19]. The aforementioned methodology was applied by Salvi *et al.*[9] to project Indian monsoon rainfall at 0.5° resolution using the CMIP (Coupled Model Inter-comparison Project)-3 data for the GCM developed by Canadian Centre for Climate Modelling and Analysis (CCCma). Most of the studies carried out for statistical downscaling use the CMIP-3 model data[1,9,17,19,20]. Recently, Knutti and

*For correspondence. (e-mail: pvtimbadiya@ced.svnit.ac.in)

Sedláček[21] reported that the mean temperature from CMIP-5 results is less uncertain than CMIP-3, while by Shashikanth *et al.*[20] concluded that the Indian summer monsoon rainfall results from CMIP-5 are less uncertain than the CMIP-3 results.

The present study is an attempt to downscale the multisite daily precipitation over the Tapi Basin at a fine resolution of 0.25° from CMIP-5 GCM data using the KR-based statistical downscaling methodology developed by Kannan and Ghosh[1]. The four scenarios of CMIP-5 experiments are based on representative concentration pathways (RCPs), i.e. RCP2.6, RCP4.5, RCP6 and RCP8.5 (ref. 22) and represent radiative forcing levels of 2.6, 4.5, 6 and 8.5 Wm^{-2} respectively, by the end of the century[23]. The study is an attempt to downscale precipitation at a very fine resolution of 0.25° using CMIP-5 GCM data of scenario RCP4.5 (considering this as the most probable scenario)[22,23] and RCP8.5 (considering this as the worst case scenario)[23,24] to quantify impact of climate change on water resources of the Tapi Basin up to the end of the 21st century. Furthermore, the downscaled daily precipitation from CMIP-5 GCM (MPI-M) and daily precipitation from Co-ordinated Regional Climate Downscaling Experiment (CORDEX) South Asia Regional Climate Model (RCM) were compared to evaluate the uncertainty in the KR-based statistical downscaling model for the study area under consideration. The study supports strength of the KR-model in capturing the highly heterogeneous rainfall as well as the inter-station cross-correlations in the basin.

Study area and data sources

The Tapi Basin lies in the northern part of the Deccan Plateau extending over an area of 65,145 sq. km, which is nearly about 2% of the total geographical area of India[25]. The Tapi River, the second largest west-flowing river of the Indian Peninsula, having catchment area up to Ukai dam is 62,225 sq. km. Figure 1 shows the index map of the Tapi basin, including the India Meteorological Department (IMD)-operated rain-gauge stations. The basin lies between 72°38′–78°17′E long. and 20°5′–22°3′N lat. (ref. 25). The Tapi River has 14 major tributaries of length more than 50 km; the Purna and Girna River basins together account for nearly 45% of the Tapi Basin[25]. The basin has an elongated shape with a maximum length of 587 km from east to west, and the maximum width of 210 km from north to south with perimeter of about 1840 km (ref. 24). Over 90% of the total rainfall occurs during summer monsoon (June–September) in the basin and falls within the zone of severe rainstorm lasting 1–2 days with maximum 24 h heavy rainfall ranging from 86 to 459 mm (ref. 26).

The GCMs are credible tools for projecting the future time series of the large-scale atmospheric variables. They are three-dimensional mathematical models based on the physical principles of fluid dynamics, thermodynamics and radiative heat transfer[16]. They represent the climate system in a simplified form using combinations of models for different components of the climate system[27]. There is no single rule available for the selection of GCMs for downscaling of precipitation. They are generally selected on the basis of the availability of large-scale predictor variables and the time period. The GCMs given in the Table 1 have been selected from CMIP-5 experiments (scenario RCP4.5 and RCP8.5) for the present study due to their reliability to simulate precipitation extremes for India[28], and availability of relevant data required for multisite precipitation for the area under consideration.

The predictors are the climate variables that are well simulated by the GCMs and are available from their archives. There is no single approach or method available for selection of predictors for precipitation downscaling. One should use the predictors which are strongly correlated with the surface variable of interest (daily precipitation in the present study). The air temperature, mean sea-level pressure, specific humidity, zonal and meridional winds and geopotential height (taken by Kannan and Ghosh)[18] at the surface, 250, 500 and 850 hPa, along with MSLP and geopotential height at the surface are considered as predictor variables in the present study.

The all-India gridded daily precipitation data at a resolution of 0.25° from APHRODITE (Asian Precipitation Highly-Resolved Observational Data Integration Towards Evaluation of the Water Resources) has been downloaded from http://www.chikyu.ac.jp and used as predictand for the present study. The APHRO_MA_V1003R1 dataset is also used, which is a newly available gauge-based high-quality dataset for Asia[29]. Due to better grid resolution, APHRODITE data are presently being used in many climate studies. Rajeevan and Bhate[30] recently compared APHRODITE v0804 (based on about 2000 station observations over the Indian subcontinent) with the rain gauges from IMD (comprising more than 6000 stations). The study revealed that APHRODITE data have high correlation with the IMD rainfall data over India. The use of APHRODITE data is also recommended by Krishnamurti *et al.*[31] as an enhancement to the Tropical Rainfall Measuring Mission (TRMM) over India.

Due to the unavailability of adequate observed climatological data at fine resolution, the reanalysis data are utilized in climate change studies. The climatological variables of NCEP/NCAR (National Centre for Atmospheric Research) reanalysis-I data[32] were downloaded from http://www.esrl.noaa.gov/psd/data/reanalysis/reanalysis.shtml and used. These data are outputs from a high-resolution climate model, run using assimilated data from sources such as surface observation stations and satellites, and hence can be considered as outputs from an ideal GCM[1]. The NCEP/NCAR reanalysis-I data provide global atmospheric information from 1948 to the present,

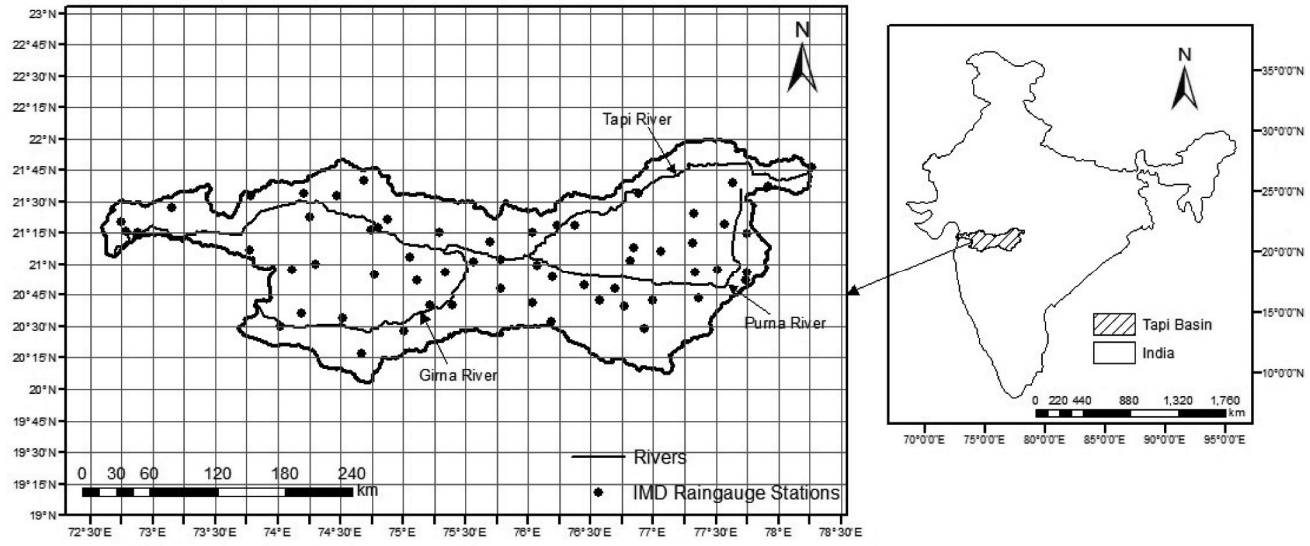

Figure 1. Index map of Tapi Basin.

Table 1. General circulation models (GCMs) used in the present study

Modelling Centre	Model	Institution	Resolution (lat. × long.), level
CCCma	CanESM2	Canadian Centre for Climate Modelling and Analysis	2.8° × 2.8°, L35
CMCC	CMCC-CMS	Centro Euro-Mediterraneo per I Cambiamenti Climatici	1.87° × 1.87°, L95
BCC	bcc-csm 1-1	Beijing Climate Center, China Meteorological Administration	2.8° × 2.8°, L26
CNRM-CERFACS	CNRM-CM5	Centre National de Recherches Meteorologiques/Centre Europeen de Recherche et Formation Avancees en Calcul Scientifique	1.4° × 1.4°, L31
MPI-M	MPI-ESM-LR	Max Planck Institute for Meteorology	1.9° × 1.9°, L47

which is a mixture of physical observations and model simulations. The data assimilation system used in the NCEP/NCAR reanalysis includes the global model and a three-dimensional analysis scheme that incorporates land, surface, ship, rawinsonde, satellite and other data with a T62 horizontal resolution and 28 vertical sigma levels[1]. Kalnay *et al.*[32] classified the global output variables into three types depending upon their dependencies on the observations and model. Type 'A' variables, for example, zonal and meridional winds are strongly influenced by the observations. Type 'B' variables, including specific humidity and temperature are influenced by both the observations and the model, while type 'C' variables are influenced by the model and no observation directly affects the variable, which includes precipitation. For the present study, the NCEP/NCAR reanalysis-I data provide the variables, i.e. air temperature, specific humidity, MSLP, zonal and meridional winds and geopotential height at the surface, 250, 500 and 850 hPa, for an area extending between 5°–40°N lat. and 60°–120°E long., covering the complete study area for a period of 25 years (1981–2005). Out of total 25 years' data, the first 15 years' data from 1981 to 1995 are used for training and the remaining 10 years' data from 1996 to 2005 are utilized for validation of the proposed downscaling model. If the grid spacing of GCMs and NCEP/NCAR does not match, the data of different GCMs are interpolated to NCEP/NCAR grid points.

To evaluate the uncertainty of the KR-based statistical downscaling model for the study area, CORDEX RCM-based COSMO-CLM daily precipitation data have been downloaded from ftp://cccr.tropmet.res.in. These data have been chosen, as their host-GCM is CMIP-5 MPI-M (MPI-ESM-LR). The CORDEX data are available at a resolution of 0.5°; they are interpolated to 0.25° resolution using bi-linear interpolation.

Development of model and results

For the present statistical downscaling of daily precipitation over the Tapi Basin, the methodology developed by Kannan and Ghosh[1] has been adopted. The predictors and predictand data need to be processed before being utilized for downscaling. The first task to be performed on the predictors is the bias correction, which involves the

removal of the systematic biases from their mean and standard deviation. The standardization method of bias correction is used for a predefined baseline period. Principal component analysis (PCA) is carried out on the bias-corrected predictor variables to diminish dimensionality and multicollinearity in a highly uncorrelated dataset. It is the application of orthogonal transformation on a set of correlated predictor variables, producing principal components (PCs) which retain almost the same variability as that of original dataset[1]. For modelling the rainfall occurrence and for deriving the daily rainfall state from the predictand dataset, an unsupervised clustering technique, viz. *k*-means clustering is employed. A supervised data classification algorithm, CART (Classification And Regression Tree) is applied for establishment of statistical relationship between PCs (arrived from bias-corrected predictors) and rainfall states (arrived using the *k*-means clustering). The daily rainfall states and PCs of the predictor variables are used as input in the KR model for the generation of daily rainfall amount. The rainfall projections are obtained for the following three rationales:

1. For validation period, the rainfall projections are obtained using the NCEP/NCAR reanalysis-I data. The data of time slice from 1981 to 1995 (baseline period) are used for training period for which the statistical relationship is established. The time slice from 1996 to 2005 is used for testing/validation period for which the projections are obtained and are being validated with that of the predictand.
2. The historical projections validation period, i.e. for the time period 1996–2005 are obtained from five GCMs, viz. CCCma (CanESM2), CMCC (CMCC-CMS), BCC (bcc-csm 1-1), CNRM-CERFACS (CNRM-CM5) and MPI-M (MPI-ESM-LR).
3. The future projections are also obtained from the same five GCMs for the time period 2006–2100 for the two scenarios, RCP4.5 and RCP8.5. The CORDEX (COSMO-CLM) rainfall projections are obtained for the validation period (1996–2005) and future period (2006–2099). For the validation period, the daily precipitation from CORDEX data and statistical downscaled data is compared with the APHRODITE rainfall data in order to check uncertainty in the downscaling model. Also, the statistically downscaled data and CORDEX data are compared for future time period.

Selection of spatial extent

The choice of spatial extent for predictors required for multisite statistical downscaling is based on the linear association between the predictor and predictand. The Pearson correlation coefficient (PCC) is determined and plotted between the key predictor variables and the average rainfall data of the Tapi Basin between 5°–40°N lat. and 60°–120°E long. Figures 2 and 3 show the contour plots of PCC of predictor variables at the surface and 500 hPa respectively. The spatial domain is selected in such a way that the Indian summer monsoon activity is captured well in the downscaling framework. After analysing the contour plots of all the predictor variables with the observed mean rainfall for the basin, the spatial extent of 10°–35°N lat. and 65°–95°E long. has been selected for the present study.

Bias correction

The GCMs involve parameterization of many processes due to incomplete knowledge of the underlying physics[16]. The parameterization leads to the systematic difference between the observed and GCM-simulated variables, which is defined as bias. It is required to be removed for correct future projections from the GCM-simulated variables. The standardization method of bias correction is used in this study, before statistical downscaling, taking the NCEP/NCAR reanalysis-I data as the reference data. The standardization of NCEP/NCAR reanalysis-I data is performed by subtraction of the mean and division by the standard deviation of the predictor variable of NCEP/NCAR for a predefined baseline period. Similarly, the standardization of GCM data is carried out using historical experimental data for the same baseline period. Thus, for a single day, the total dimension of the standardized data available for modelling is 715 (five variables at 143 grid points).

Principal component analysis

The data of predictor variables obtained after bias correction suffer from the problem of multidimensionality and multicollinearity. The use of high-dimensional correlated predictor data in the present downscaling framework is costly with regard to computation. Hence, there is a need to convert the highly correlated multidimensional predictor into a set of uncorrelated variables with reduced dimensionality, and PCA is used for this purpose. In the present study, the standardized predictor containing 715 dimensions (five variables at 143 grid points) is reduced to a predictor set containing 100 dimensions, without ignoring the significant information contained in the original data.

Identification of daily rainfall states

Capturing the cross-correlation between the rain gauge stations in close proximity is one of the problems in the multisite statistical downscaling. This can be solved using common rainfall states in the river basin[9]. Kannan and Ghosh[18] employed *k*-means clustering for the identification of daily rainfall state, which is an unsupervized

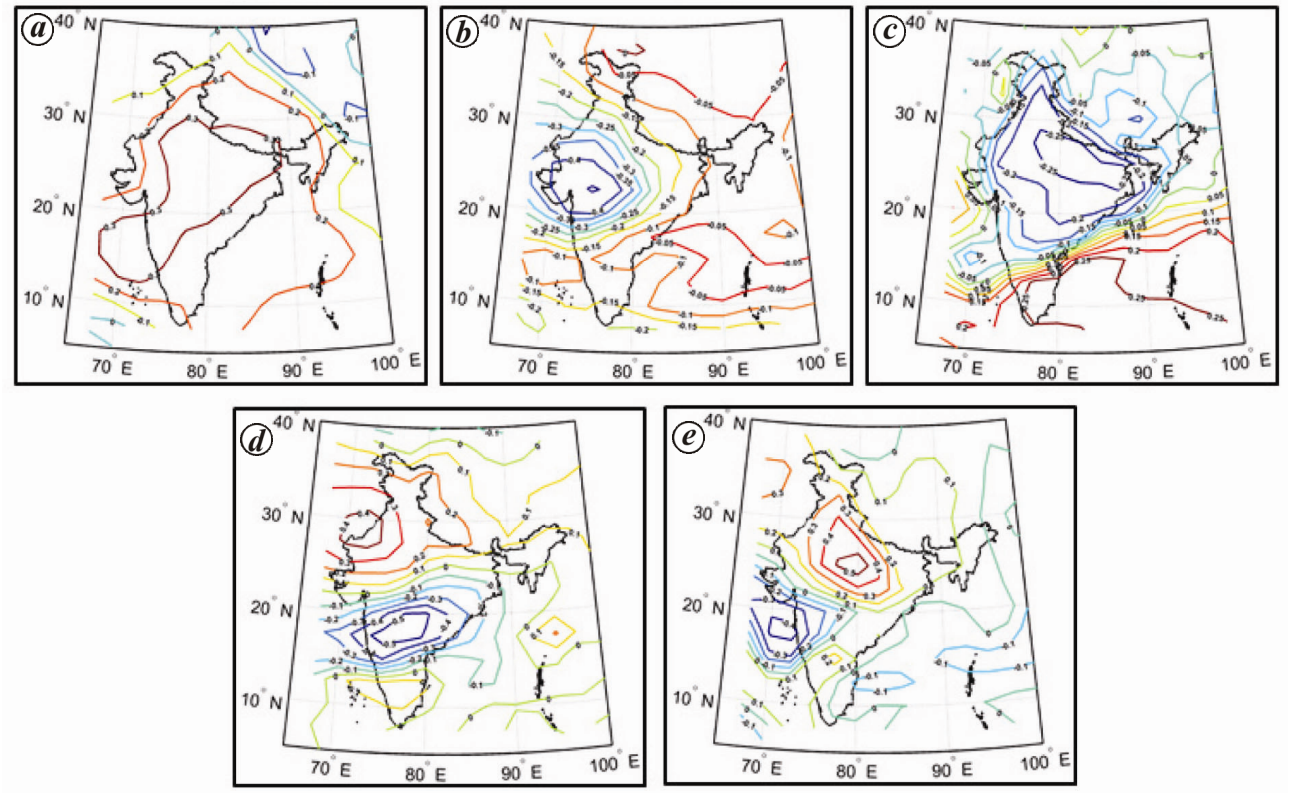

Figure 2. Contour plot of Pearson correlation coefficient (PCC) between surface predictor variables and basin average rainfall. ***a***, Specific humidity at the surface (kg/kg); ***b***, Mean sea-level pressure (Pa); ***c***, Air temperature at the surface (K); ***d***, *U* wind at the surface (m/s); ***e***, V wind at the surface (m/s).

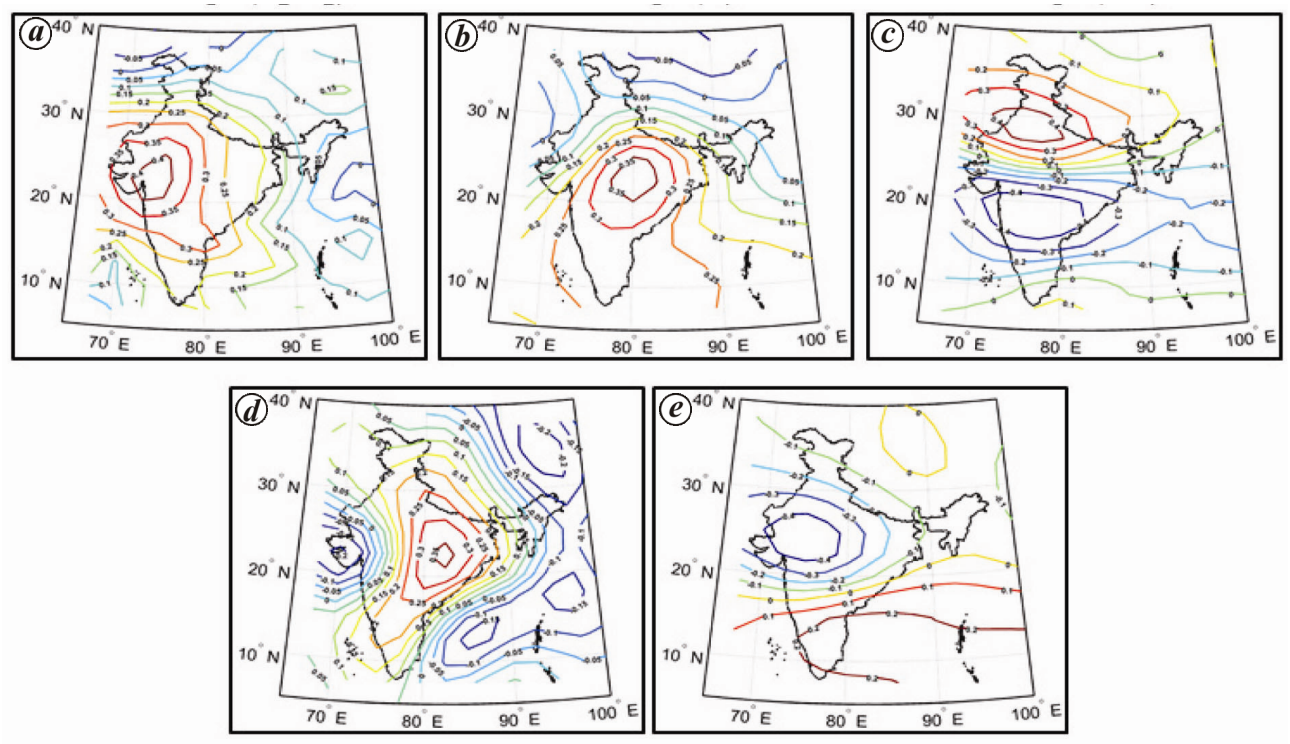

Figure 3. Contour plot of PCC between 500 hPa predictor variables and basin average rainfall. ***a***, Specific humidity at 500 hPa (kg/kg); ***b***, Air temperature at 500 hPa (K); ***c***, U wind at 500 hPa (m/s); ***d***, V wind at 500 hPa (m/s); ***e***, Geopotential height at 500 hPa (m).

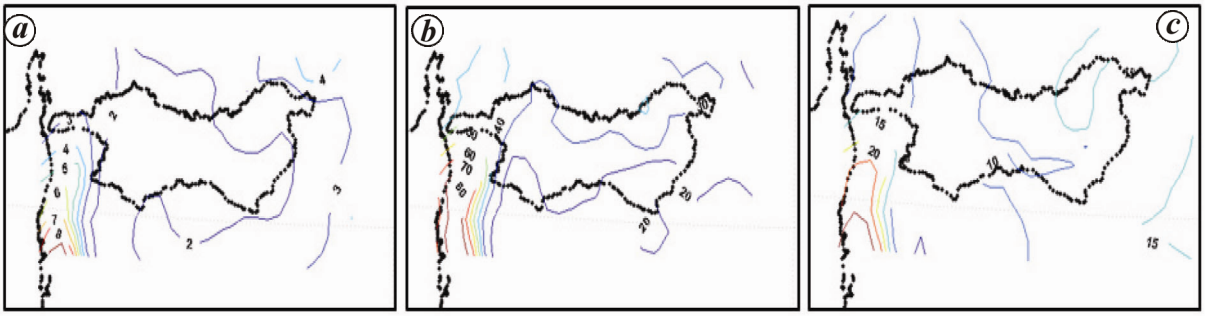

Figure 4. Cluster centroids derived from *k*-means clustering technique for three clusters. ***a***, Rainfall state 1: almost dry; ***b***, Rainfall state 2: high; ***c***, Rainfall state 3: medium.

data classification technique classifying n observations into k-clusters. In this study, the gridded daily rainfall data from APHRODITE extracted at 351 grid points, representing the Tapi basin, are clustered to obtain daily rainfall state of the river basin. The clustering is used since it requires no assumptions to be made on the number of groups (clusters) or the group structure. The groupings are carried out on the basis of similarities or distances (dissimilarities). The total number of clusters (in which data are grouped) is an input parameter for the clustering algorithm, but it is not known. Hence careful evaluation of the results of the clustering algorithm is necessary. The acceptability of clustering results is also not defined; thus several cluster validity techniques are used to identify the optimum number of clusters. Three cluster validation measures (i.e. Dunn's index[33], Davies–Bouldin index[34] and Silhouette index[35]) are computed for each cluster obtained under k-means clustering technique. The results of the validity measures computed for different sets of clusters are presented in the Supplementary Material (see online). The highest value for Dunn's index and Silhouette index and the minimum value for Davies–Bouldin index present the optimum number of clusters. For obtaining the optimum number of clusters, the total number of clusters is varied from 2 to 10. It can be concluded that the optimum number of clusters for the present study is three. Figure 4 shows cluster centroids arrived using k-means clustering technique for the three clusters. The three rainfall states derived are 'almost dry', 'medium' and 'high', on the basis of rainfall amounts represented by the cluster centroids[18]. Amongst the three cluster centroids, the one for the 'almost dry' condition is found to be well separated. Therefore, the results obtained by k-means clustering algorithm for $k = 3$ are adjudged the best partition.

Modelling rainfall occurrence using CART

Modelling of the nonlinear relationship between large-scale atmospheric predictors and the river basin-wise rainfall state is carried out using CART. The inherent ability of CART to select variables automatically and to select the appropriate tree structure in the absence of prior information regarding variable selection enable us to come up with different models[1]. Therefore, in the present study three models are proposed for training and validation purpose. These are:

$$R(t) = \{m(t), m(t-1), R(t-1)\}, \quad (1)$$

$$R(t) = \{m(t), m(t-1), R(t-1), R(t-2)\}, \quad (2)$$

$$R(t) = \{m(t), m(t-1), R(t-1), R(t-2), R(t-3)\}, \quad (3)$$

where $R(t)$, $R(t-1)$, $R(t-2)$ and $R(t-3)$ are the rainfall occurrence on the tth, $(t-1)$th, $(t-2)$th and $(t-3)$th day respectively. $m(t)$ and $m(t-1)$ are the set of large-scale atmospheric variables on the tth and $(t-1)$th day respectively.

In the present study, CART model is developed using standardized and dimensionally reduced NCEP/NCAR predictor data and concurrent rainfall states. The data from 1981 to 1995 (a period of 15 years) are used as the training set for construction of the classification tree, while the remaining data for a period of 10 years (from 1995 to 2005) are utilized for validation of the CART model. A total of three skill score measures, i.e. SRMP (success rate of model prediction), HSS (Heidke skill score) and χ^2 goodness-of-fit statistic are computed from the validation data to assess reliability of the model for future uses[18]. Details of the calculation of SRMP, HSS and χ^2 statistic can be found in Kannan and Ghosh[18]. The value of SRMP ranges from 0 to 100, with 0 for poor and 100 for perfect forecast. The perfect forecasts have HSS equal to one, forecast equivalent to reference forecasts receive zero scores, and forecasts worse than reference forecasts are have negative scores. According to Maity and Nagesh Kumar[36], a reasonably good forecast has HSS value greater than 0.15. The χ^2 distribution is used to decide whether there is any association between the observed and forecasted rainfall occurrences. The null hypothesis H_0 for the present case can be defined as

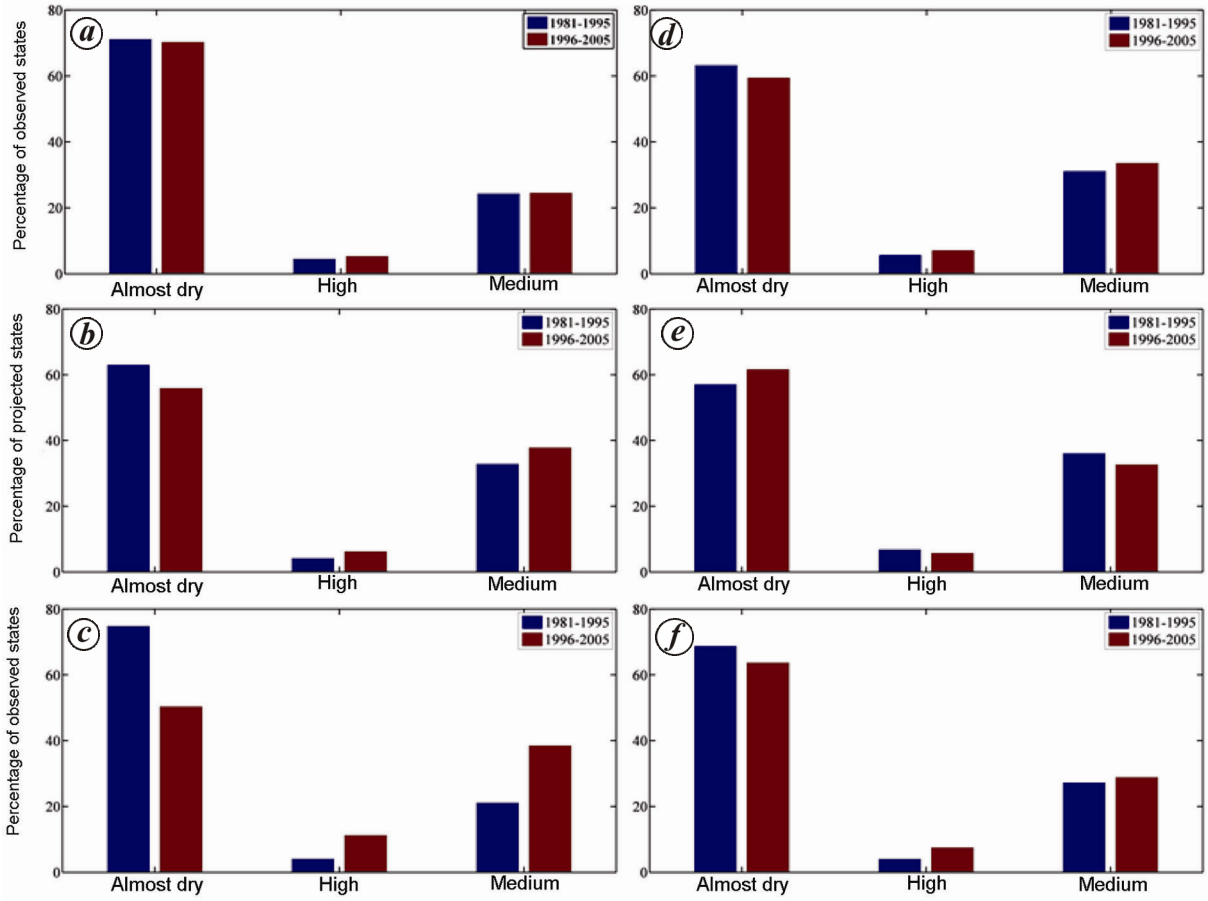

Figure 5. Percentage-wise break-up of occurrence of rainfall states for the training and validation periods. ***a***, *k*-Means cluster; ***b***, CCCma projections; ***c***, CMCC projections; ***d***, BCC projections; ***e***, CNRM-CERFACS projections; ***f***, MPI-M projections.

follows: 'There is no association between the observed and forecasted rainfall occurrences'[18].

The results of computed SRMP (percentage), HSS and χ^2 goodness-of-fit statistic for the three CART models under consideration are tabulated in the Supplementary Material (see online). It is observed that the SRMP, HSS and χ^2 goodness-of-fit computed for the results of model-III (see eq. (3)) validation runs are fairly good compared to model-I (see eq. (1)) and model-II (see eq. (2)). Thus, model-III gives a reasonably good simulation for present case and it is selected for projecting daily rainfall states using the GCM outputs. After the simulation of future states, KR is used conditionally on the derived states in order to project future rainfall amount.

Projection of future daily rainfall state

The PCs of all the five GCMs under consideration (CCCma, BCC, CNRM-CERFACS, CMCC and MPI-M outputs for experiments such as historical, RCP4.5 and RCP8.5) are taken as input for projecting rainfall states of future climate change scenarios using trained CART model. Figure 5 shows the occurrences of rainfall states (percentage-wise) for all five GCMs for training and validation, while Figure 6 presents the same for future projection of RCP4.5. From Figure 5, it is observed that the rainfall states projected for four GCMs models (viz. BCC, CCCma, CNRM-CERFACS and MPI-M) almost match with the observed rainfall states obtained from *k*-means clustering with no considerable increase or decrease in 'almost dry', 'high' and 'medium' rainfall states. While for CMCC, a considerable decrease is observed from training period (1981–1995) to validation period (1996–2005) for 'almost dry' rainfall state and an increasing trend is observed for 'high' and 'medium' rainfall states from training period to validation period.

The model results obtained for the RCP4.5 experiment show a moderate decrease in the 'almost dry' rainfall state for the period 2041–2070 and 2071–2100 and a moderate increase in the 'high' rainfall state for the same period, with no trend for 'medium' rainfall state for CCCma (Figure 6 *a*). The rainfall state for CMCC (Figure 6 *b*) shows a decreasing and increasing trend for 'almost dry' and 'high' rainfall states respectively, with a slight increase and decrease in the 'medium' rainfall state for the period 2041–2070 and 2071–2100 respectively. There is a slight decrease in the 'almost dry' rainfall state for the period 2041–2070 with a small increase in 'almost

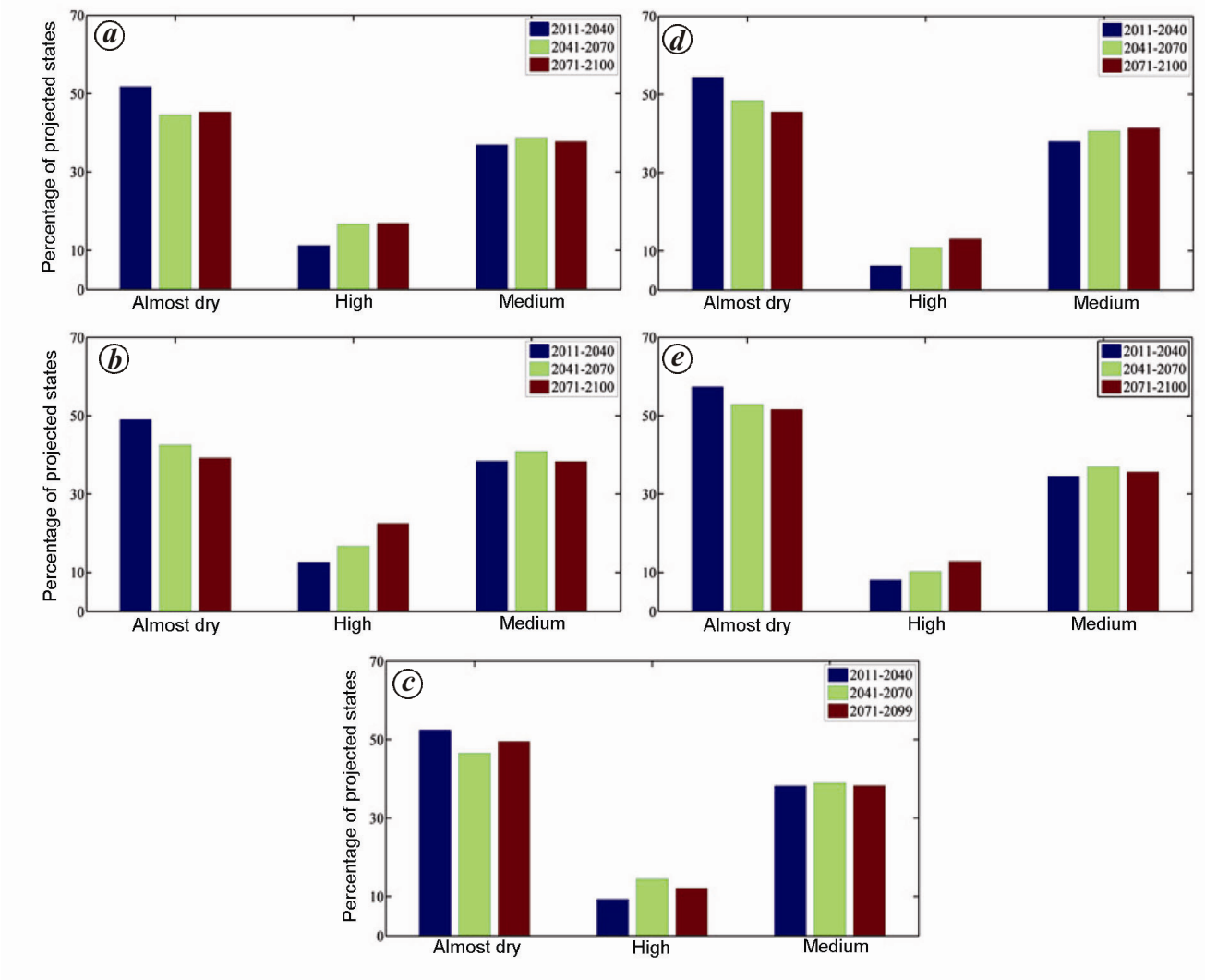

Figure 6. Percentage-wise break-up of occurrence of rainfall states for future scenario RCP4.5 of GCMs: ***a***, CCCma; ***b***, CMCC; ***c***, BCC; ***d***, CNRM–CERFACS; ***e***, MPI-M for time different slices.

Dry' rainfall states for the period 2071–2100 with no significant trend for 'medium' rainfall states for BCC (Figure 6 *c*). There is also a slight increase and decrease in the 'high' rainfall state for the period 2041–2070 and 2071–2099 respectively. From Figure 6 *d* and *e*, it can be inferred that the rainfall state projections for CNRM-CERFACS and MPI-M are almost similar with decreasing, increasing and no significant trend for 'almost dry', 'high' and 'medium' rainfall states respectively.

The occurrence of rainfall states constitutes a dynamic system that changes randomly between one state and another. These changes are called transitions, and the probabilities associated with the changing state are called transition probabilities. The plots of changes in 'almost dry–almost dry', 'high–high' and 'medium–medium' rainfall state transition probabilities obtained for all GCM models for scenarios RCP4.5 and RCP8.5 are given in the Supplementary Material (see online).

Modelling the multisite rainfall amount and its spatial dependence

A non-parametric KR-based downscaling model (Figure 7) developed by Kannan and Ghosh[1], is used in the present study for modelling the multisite rainfall amount, which considers the rainfall state of the region as one of the predictors. Moreover, the methodology is successful in overcoming the different models developed so far for downscaling the multisite daily precipitation.

The generation of multisite rainfall is achieved by modelling the relationship between the large-scale atmospheric circulation climate variables (predictors) and the local-scale precipitation. The functional form of the relationship is given as

$$R_t = f_R(X_t/S_t), \tag{4}$$

where R_t is the rainfall at an individual grid point at time t, X_t the climate predictor at time t, and S_t is the rainfall

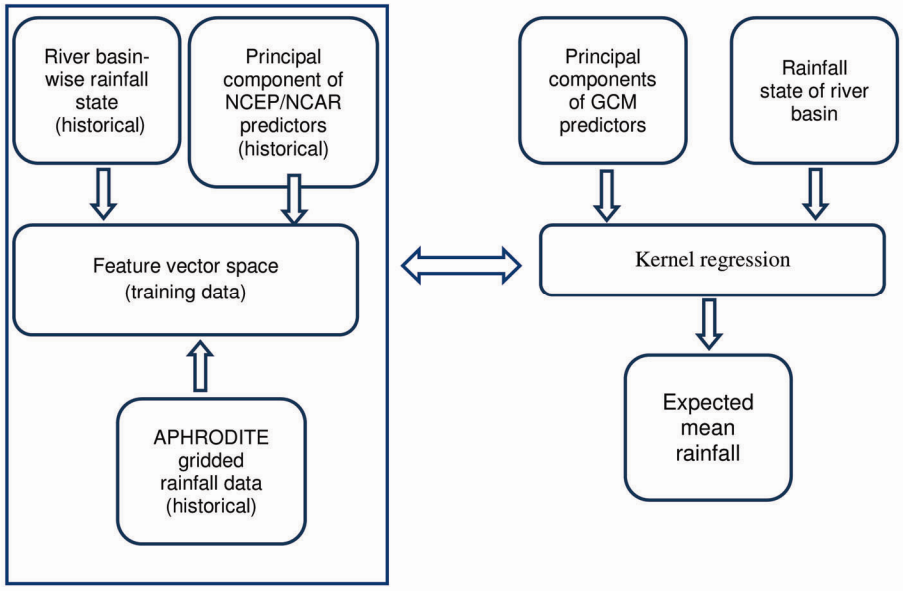

Figure 7. Methodology for downscaling multisite rainfall amount.

State of the river basin at time t. The conditional probability of a random variable can be estimated using KR. The multivariate KR uses a weighted sum of the observed responses with kernel density functions for weights[1]. The weight of each training input is computed using a kernel function which typically decays rapidly with the distance between itself and the test input. The estimates are obtained such that the estimated test point value has the strongest dependence on nearby training points. A common approach in KR combines a Euclidean distance metric with Gaussian kernel, which decays exponentially with the squared distance rescaled by a kernel bandwidth. The general form of conditional expectation function $E(Y/X)$ for a multivariate distribution of the variables Y and X is given below

$$E\left(\frac{Y}{X}\right) = m(X) = \frac{\int yf(y,x)\mathrm{d}y}{f_X(x)}. \quad (5)$$

Nadaraya[37] has replaced the multivariate pdf (probability density function) by the kernel density estimate[1]

$$\hat{m}_H(x) = \frac{\sum_{i=1}^{n} \kappa_H(X_i - x)Y_i}{\sum_{i=1}^{n} \kappa_H(X_i - x)}, \quad (6)$$

where $\hat{m}_H(x)$ is the expected value of Y for a condition of $X_i = x$ and κ_H is the kernel with bandwidth H.

For the present study, selection of kernel bandwidth and stochastic simulation of rainfall amounts has been carried out according to the procedure described by Kannan and Ghosh[1]. The rainfall state is given by the rainfall occurrence model for the basin under consideration. It considers the previous day rainfall state as one of the predictors which explain long-term persistence pass by daily short-term persistence. The spatial dependence of the rainfall is maintained by the combined use of common rainfall state of the river basin and a data-driven KR approach as defined by Kannan and Ghosh[1] for the downscaling technique. The approach as defined by them explicitly considers the effect of large-scale circulation conditioned on the rainfall state, and therefore provides refined and logical approach of incorporating spatial and temporal dependency.

Application of the model

The KR-based model uses the feature vector space as training sample for generation of daily rainfall amount. The standardized and dimensionally reduced NCEP/NCAR reanalysis-I daily rainfall data along with concurrent daily precipitation are used for the construction of feature vector space for a period of 15 years from 1981 to 1995. Validation of the KR model is carried out with the remaining 10 years from 1996 to 2005 using standardized and dimensionally reduced NCEP/NCAR climate variables. The performance of the KR model (with conditioning on weather states) is tested by comparing the simulation of KRWS model (kernel regression model without conditioning on weather states). The KR-based downscaling approach is used to simulate rainfall for near future (2041–2070) and far future (2071–2100) using the standardized and dimensionally reduced predictor variables for RCP4.5 and RCP8.5 scenarios of all the five GCMs (i.e. CCCma, CMCC, CNRM-CERFACS, BCC and MPI-M).

Table 2. Observed and downscaled (5th, median (50th), and 95th percentile estimates) monthly wet days and rainfall for the testing period: 1996–2005

	Wet days					Rainfall (mm)			
Percentile	5	50	95	Mean	Percentile	5	50	95	Mean
Observed					Observed				
June	3	7	15	8	June	75.1	161.0	257.9	159.9
July	1	14	18	12	July	60.2	247.9	333.6	243.5
August	8	13	15	12	August	156.9	227.8	295.7	227.2
September	0	6	16	7	September	31.3	139.7	312.5	153.5
Modelled with rainfall state					Modelled with rainfall state				
June	4	8	13	8	June	108.4	193.9	236.6	181.8
July	1	13	17	12	July	86.2	239.5	299.4	228.3
August	8	13	17	12	August	144.9	227.0	364.7	235.1
September	0	6	14	6	September	57.7	148.2	311.4	161.2
Modelled without rainfall state					Modelled without rainfall state				
June	2	9	15	9	June	115.4	173.2	229.3	180.0
July	5	11	16	11	July	139.3	189.1	293.5	197.8
August	4	9	17	9	August	108.7	169.1	255.9	181.3
September	1	6	12	5	September	66.0	130.4	233.5	136.1

Model validation: Validation of the KR model over the baseline period is performed through the statistical comparison of calculation of basin-averaged wet days and rainfall amount, basin-averaged wet/dry spell length and their conditional probabilities, as well as KR model's ability to capture spatial and temporal dependence.

Statistical comparison: The performance of KR-based model mentioned in Table 2, in regenerating the different statistical characteristics with reference to the observed data is analysed in this section. The contour plots of the mean and standard deviation for the observed rainfall, rainfall projected by NCEP/NCAR and rainfall projections from the five GCMs using the KR method conditioned on weather states used in the present study are given in the Supplementary Material (see online). It is observed that mean and standard deviation of generated rainfall series are captured well by the KR-based approach. The correlation plots for R^2 values for the KR model with and without conditioning on weather states are also presented in the Supplementary Material (see online), which show that the R^2 value for the KR model is greater than the KRWS model, thus indicating that the rainfall values are well nalyzi by the former model.

Wet days averaged over the basin and rainfall amount: The wet spells (WS) and dry spells (DS) play an important role in the planning and management of water resources of the basin. In spite of many definitions for the identification of WS (DS), the one by Singh and Ranade[38] has been adopted for the present study. According to them, WS (DS) is identified as a 'continuous period with daily rainfall equal to or greater than (less than) daily mean rainfall (DMR) of the climatological monsoon period over the area of interest'. The present study area is located in the western part of India, for which Singh and Ranade[38] have worked out a DMR of 12.1 mm/day. Table 2 gives the monthly observed and downscaled monsoon wet days and rainfall amounts over the Tapi Basin. From the table, it can be observed that the KRWS model overestimates the number of wet days and the rainfall amount, while the KR model performs better than the KRWS model both in capturing the wet days and reproducing the total rainfall amount.

Wet/dry spell length averaged over the basin and conditional probabilities: The KR and KRWS models were also evaluated for wet/dry spell length probabilities and conditional probabilities of rainfall, in addition to the total number of wet days. In the following discussion, a wet spell is defined as a continuous sequence of days when daily rainfall is greater than or equal to DMR. Similarly, a dry spell represents a sequence of days with daily rainfall being less than DMR[5]. It is intended to verify the distribution and shape of wet/dry spell lengths and the conditional probabilities of the predictand captured by the downscaling model for the validation period (1996–2005). The plots of basin-averaged wet and dry spell length probabilities for the rainfall observed and predicted by the model for the validation period are presented in the Supplementary Material (see online). It is observed that the KR model exhibits good performance in comparison to the KRWS model, as the spell length probability curves for both wet and dry spells closely matching with the observed case, whereas a noticeable deviation can be seen in case of the KRWS model. The wet and dry spell length probabilities for all the five GCMs for the validation period are also presented in the Supplementary Material (see online). It is observed that all the five GCMs underestimate the wet spell length probabilities and overestimate the dry spell length probabilities, except for BCC which is slightly underestimating the dry spell length when compared with the observed

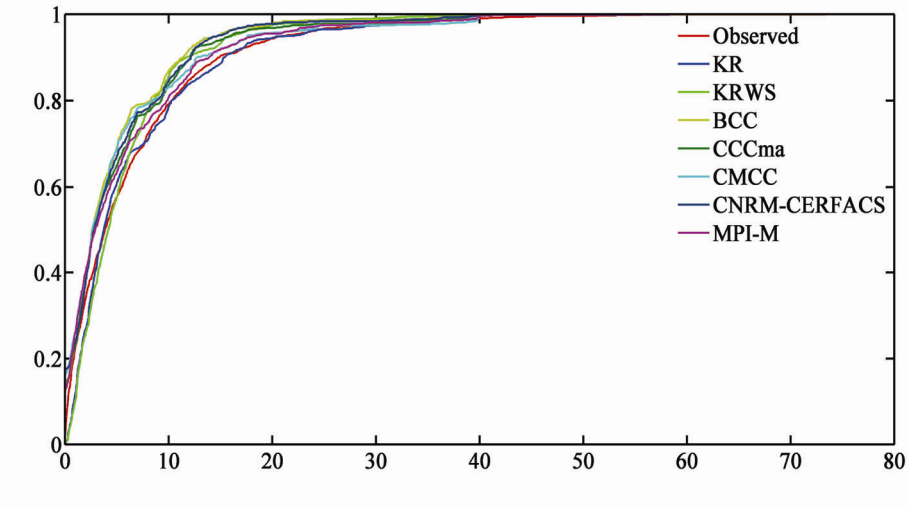

Figure 8. Cumulative distribution function of basin-averaged rainfall for the validation period.

rainfall. Figure 8 compares the cumulative distribution obtained from basin-averaged observed rainfall series with those obtained with five models used in the present study for the validation period. It is observed that the KR model shows minimum deviation from observed rainfall and therefore the KR method is selected for rainfall projections.

Rainfall variability (temporal) and spatial dependence: A distributed hydrologic model requires rainfall variability both in terms of space and time. Therefore, the projected rainfall from both the KR and KRWS models is evaluated in terms of their ability to capture both temporal and spatial dependence. The inter-station correlation coefficients are computed for the observed and the predicted rainfall series obtained from KR and KRWS models for the validation period. The scatter plots of cross-correlation coefficients obtained from the observed and model-simulated rainfall series for both the KR and KRWS models is given in the Supplementary Material (see online). It can be inferred that the spatial structure of rainfall field is captured well by the KR model, while the KRWS model overestimates the same.

Model selection: For the present study, after analysing the KR and KRWS models in terms of their ability to capture basin-averaged wet days and rainfall amount, basin-averaged wet/dry spell length probabilities, and spatial and temporal variability, the former model is found better than the latter model and is therefore selected for future rainfall projections. However, Kannan and Ghosh[1] concluded that the performance of both models is good when compared with other downscaling models.

Future projections with GCM simulations

The KR model along with standardized and dimensionally reduced data corresponding to RCP4.5 and RCP8.5 scenarios from CMIP-5 experiments of the five GCMs, including CCCma, CMCC, BCC, CNRM-CERFACS and MPI-M outputs conditioned on the rainfall states of their respective experiments, is used for the projection of the future JJAS (June, July, August and September month) daily rainfall. For investigating the changes of the global warming on the precipitation characteristics, three time slices (2011–2040, 2041–2070 and 2071–2100) have been selected in the future. The changes in the precipitation characteristics are evaluated in terms of changes in shape of wet and dry spell lengths, and their conditional probabilities for the aforementioned three time slices. Further, the 50-year return period extreme daily rainfall for future time slices are presented for Tapi Basin for both the future scenarios (RCP4.5 and RCP8.5) of the selected GCMs.

Wet/dry spell length probabilities and conditional probability under the changed climate: For efficient reservoir operation and flood management applications, it is important that the model accurately reproduces the number of wet days and rainfall amounts in the downscaled simulations[4]. According to Kannan and Ghosh[1], 'change in the number of wet days in any downscaling location results in the changes in WS length and its corresponding occurrence probabilities'. Also, sustained increase/decrease in the amount of predicted rainfall changes the shape of cumulative distribution function (CDF). Therefore, the downscaled results are evaluated for changes in wet/dry spell length probabilities and shape of CDF during the change in climate for three time slices (2011–2040, 2041–2070 and 2071–2100). The plots of wet and dry spell length probabilities obtained from the model results of all the five GCMs (CCCma, CMCC, BCC, CNRM-CERFACS and MPI-M) for both the scenarios (RCP 4.5 and RCP 8.5) are included in the Supplementary Material (see online). For CCCma, a plot for wet and dry spell length probabilities shows no significant change in shape during two time periods

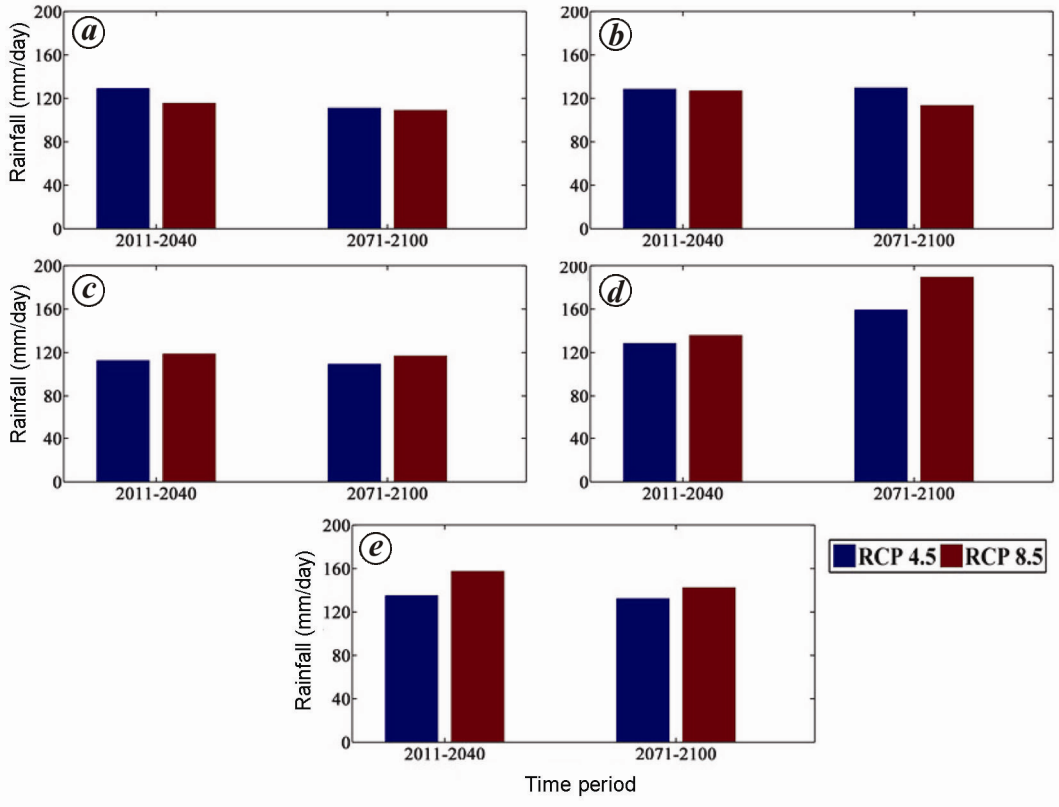

Figure 9. Fifty-year return period of extreme daily rainfall for near future (2011–2040) and far future (2071–2100) for: ***a***, CCCma; ***b***, CMCC; ***c***, BCC; ***d***, CNRM-CERFACS; ***e***, MPI-M.

(2011–2040 and 2071–2100). However, a slight increase in the shape of wet and dry spell length probabilities is observed for four or less days, and seven or less days respectively. For CMCC plots of wet and dry spell length probabilities for all the three time periods (2011–2040, 2041–2070 and 2071–2100), there is a moderate increase in the wet and dry spell length probabilities for five or less days, and six or less days respectively, for RCP 4.5 projections. For BCC, plots of wet spell length probabilities for the two time-periods, i.e. 2041–2070 and 2071–2100 and dry spell length during 2041–2070 show no change in shape of CDF for both the scenarios. However, slight increase in wet spell length probability for RCP4.5 is observed for four or less days during 2011–2040 and a slight increase in dry spell length probability plot for RCP4.5 for seven or less days. The plots for wet and dry spell length probabilities show no change in shape for CNRM-CERFACS during 2011–2040. The plots for wet and dry spell length probabilities for RCP8.5 show a slight increase for five or less days, and seven or less days respectively, during 2041–2070 and 2071–2100. However, during 2041–2070 and 2071–2100, the dry spell length probabilities show a marginal increase for RCP8.5 for four or less days, and three or less days respectively. It can be inferred that the plots for wet spell length for the two time-periods 2011–2040 and 2041–2070, and dry spell length during 2011–2040 show no change in shape for both the scenarios, RCP4.5 and RCP8.5. However, dry spell length probabilities for RCP8.5 show a slight increase for five or less days during 2041–2070 and 2071–2100, along with a slight increase in the wet spell length probability for five or less days for RCP 8.5 during 2071–2100.

The CDFs are derived from model-generated rainfall series for both scenarios, i.e. RCP4.5 and RCP8.5, in order to detect changes in the frequency of occurrence of high/low rainfall over the period of time under consideration. The CDF plots obtained for RCP4.5 and RCP8.5 for CCCma, CMCC, BCC, CNRM-CERFACS and MPI-M respectively, are given in Supplementary Material (see online). It can be inferred that for CCCma, CDFs obtained for RCP4.5 and RCP8.5 almost match with each other. For CMCC, the CDFs obtained for RCP4.5 and RCP8.5 almost match with each other during 2011–2040. However, an upward shift is observed for RCP 8.5 during 2041–2070 and 2071–2100, indicating the occurrence of high rainfall events during these time slices. For BCC, it can be inferred that the CDFs obtained for both RCP4.5 and RCP8.5 almost match with each other for two time slices (2011–2040 and 2071–2100). However, an upward shift is observed for RCP4.5 during the 2041–2070. For CNRM-CERFACS and MPI-M, the CDF plots for all the

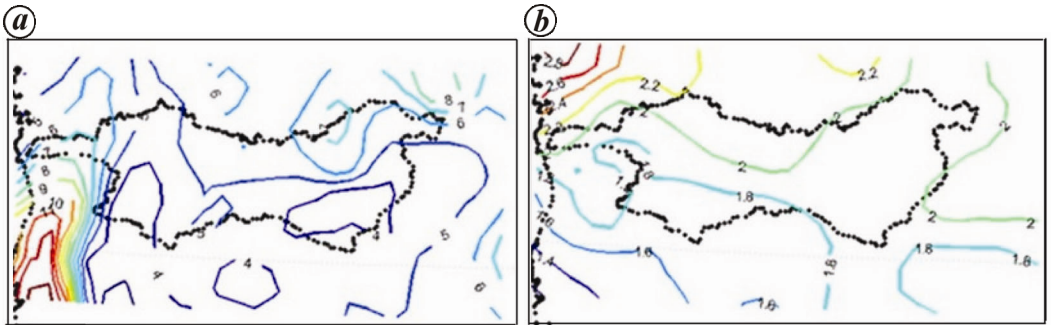

Figure 10. Contour plots of mean for validation period: ***a***, statistically downscaled data; ***b***, CORDEX data.

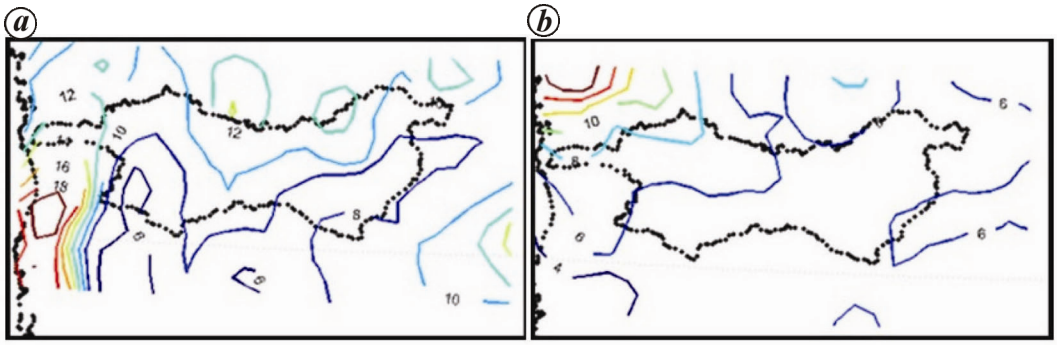

Figure 11. Contour plots of standard deviation for validation: ***a***, statistically downscaled; ***b***, CORDEX data.

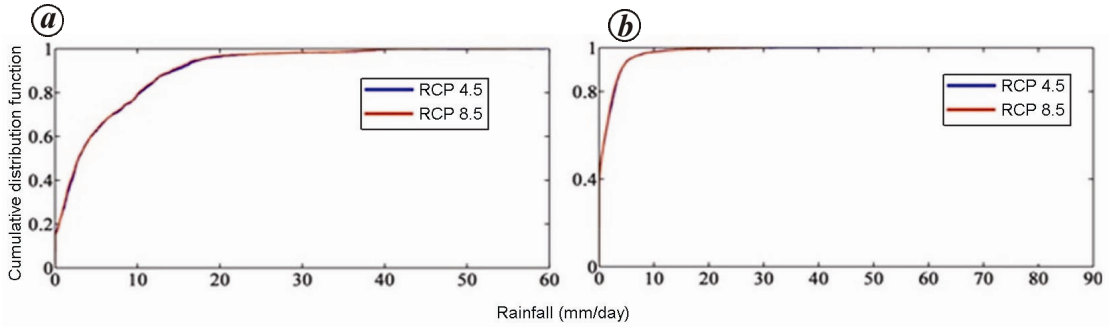

Figure 12. CDF of daily rainfall for MPI-M projections during 2011–2040; ***a***, statistically downscaled data; ***b***, CORDEX data.

Three time slices are almost the same. The CDF plots for RCP4.5 and RCP8.5 almost match with each other during 2011–2040. However, a slight upward shift is observed for RCP4.5 during 2041–2070 and 2071–2100, indicating an increased frequency of high rainfall events during these time slices.

The extreme events may increase as a result of the warming environment[39]. For analyzing the extreme rainfall for the 50-year return period for the future, Gumbel's extreme value distribution[40] has been used on extracted annual daily rainfall maxima for two time-periods during 2011–2040 and 2071–2100 for both future scenarios. Figure 9 compares results for trends in extreme daily rainfall for 50-year return period for both the future scenarios for two time-periods during 2011–2040 and 2071–2100 for all the five GCMs used in the present study. From the figure, it can be observed that for CCCma, RCP4.5 exhibits a decreasing trend. Moreover, it can be seen that for far future, RCP4.5 and RCP8.5 have a slight difference in their projections. For CMCC, there is no significant trend in RCP4.5, while decreasing trend is observed for RCP8.5 projections. For BCC, the RCP4.5 as well as RCP8.5 projections are the same for near and far future. CNRM-CERFACS shows an increasing trend for both the scenarios (RCP4.5 and RCP8.5) and for both the time slices (2011–2040 and 2071–2100). For MPI-M, RCP8.5 exhibits a decreasing trend in the amount of 50-year return period extreme daily rainfall. It is also observed that RCP4.5 projections are the same for both time periods.

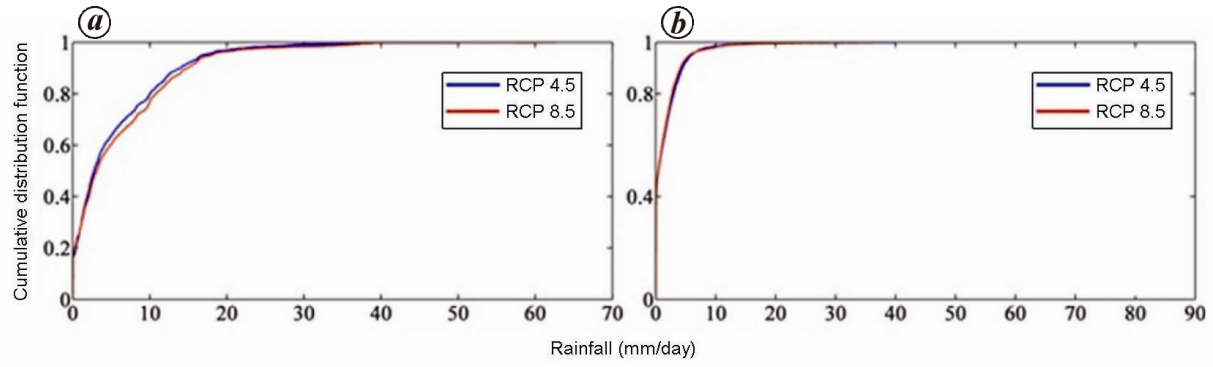

Figure 13. CDF of daily rainfall for MPI-M projections during 2041–2070. ***a***, Statistically downscaled data; ***b***, CORDEX data.

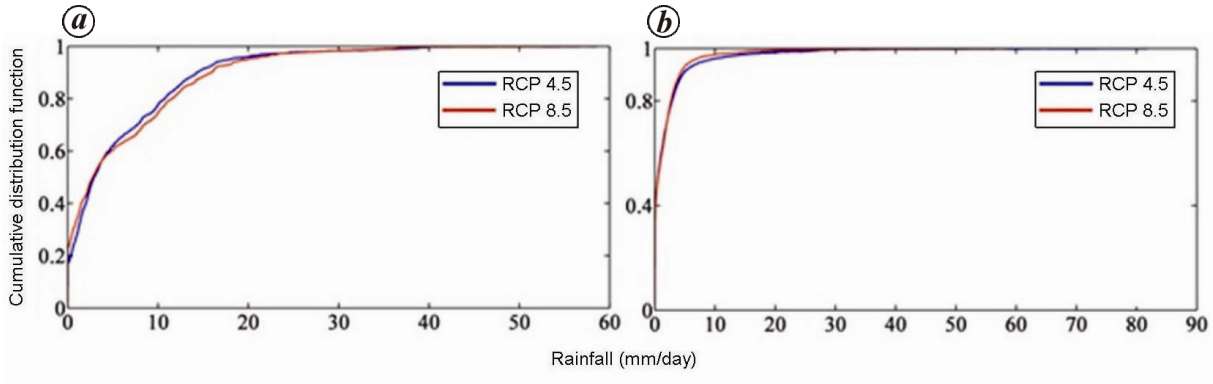

Figure 14. CDF of daily rainfall for MPI-M projections during 2071–2100: ***a***, Statistically downscaled data; ***b***, CORDEX data.

Comparison between statistically downscaled and CORDEX rainfall

For the present study, the daily rainfall is projected for the future from all the five GCM models at 0.25° resolution. To evaluate the uncertainty in statistical downscaled daily precipitation obtained from CMIP-5 MPI-M GCM through the KR conditioned on weather states-based statistical downscaling model, it is compared with the daily precipitation from CORDEX COSMO-CLM (host-GCM is CMIP-5 MPI-ESM-LR) taking the predictand (APHRODITE rainfall data-observed rainfall) as reference data for the validation period. Table 3 presents the results for R^2 using statistically downscaled data and CORDEX dataset for the validation period (1996–2005). It can be inferred from the table that the statistically downscaled projections are able to explain 89.12% of variation from the observed data, whereas projections using the CORDEX data explain only 55.5% of the variation. However, it may be noted that the results obtained from CORDEX are not directly comparable with observations, because they (CORDEX) describe a mean value over a volume rather than a point measurement, while statistically downscaled results may be directly comparable with the observed values used to calibrate the statistical models in statistical downscaling[41]. Furthermore, for future period, results from GCMs and CORDEX are compared using CDF plots for the three time slices, i.e. 2011–2040, 2041–2070 and 2071–2100, and the 50-year return period extreme daily rainfall for both the future scenarios (RCP4.5 and RCP8.5) for near and far future (viz. 2011–2040 and 2071–2100) for the present study area. Figure 10 presents the contour plots of daily mean rainfall for statistically downscaled data and CORDEX data. It can be observed from the figure that the observed mean rainfall is captured well using the statistically downscaled data. The contour plots of standard deviation for daily mean rainfall are presented in Figure 11, which also reveals that the standard deviation of observed mean rainfall is captured well using the statistically downscaled data. The CORDEX data do not model accurately the mean and standard deviation for the observed rainfall for the study area under consideration. Figures 12–14 show the CDF plots for three time slices, i.e. 2011–2040, 2041–2070, 2070–2100 respectively, using the statistically downscaled data and CORDEX data. It can be observed from the figures that the CDF plots for RCP4.5 and RCP8.5 match each other for all the three time slices for projections obtained using CORDEX data, revealing the same frequency of rainfall events from both the scenarios.

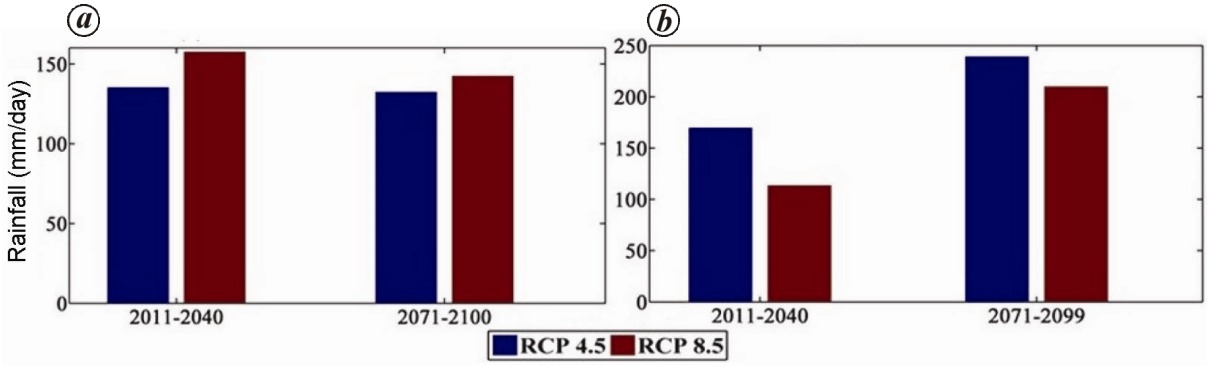

Figure 15. Fifty-year return period of extreme daily rainfall for near future (2011–2040) and far future (2071–2100): ***a***, Statistically downscaled data; ***b***, CORDEX data.

Table 3. Comparison of coefficient of determination for the precipitation for the validation period

Data	Coefficient of determination (R^2)
Statistically downscaled projections	0.8912
CORDEX projections	0.5554

However, projections using statistically downscaled data show no change in CDF plots during 2011–2040, and an upward shift for RCP4.5 projections during 2041–2070 and 2071–2100. The extreme rainfall for the 50-year return period of statistically downscaled data and CORDEX data for the future is calculated using Gumbel's extreme value distribution to further check the uncertainty. Figure 15 compares results for trends of extreme daily rainfall for 50-year return period for both the future scenarios for two time periods during 2011–2040 and 2071–2100 for statistically downscaled data and CORDEX data. The projections using statistically downscaled data show a decreasing trend for RCP8.5, with no trend for RCP4.5. While the projections using CORDEX data show an increasing trend for both the scenarios in the 50-year return period of extreme rainfall amount.

Conclusion

The statistical downscaling of multisite daily precipitation for Tapi basin, has been carried out using KR and KRWS models. The KR model performs better than the KRWS model for the validation period due to the availability of more data in terms of rainfall state. The KR model has been further used for projecting the rainfall for both the climate scenarios, RCP4.5 and RCP8.5, for all the five GCM models, i.e. CCCma, CMCC, BCC, CNRM-CERFACS and MPI-M. The projected rainfall series are used to identify the impacts of climate change in the Tapi Basin in terms of change in the shape of CDF and the occurrence of extreme rainfall event. For future projections, there are differences between the projected results for different scenarios and GCM models. It is broadly observed that for CCCma and BCC, almost dry, medium and high rainfall states have no significant trend. For CMCC, CNRM-CERFACS and MPI-M, medium rainfall state shows no significant trend, while almost dry and high rainfall states depict decreasing and increasing trend respectively. In general, it can be concluded that for future projections there is a possibility of decrease in the occurrences of extreme events with an increase in the medium rainfall events in Tapi Basin. For the study area under consideration, the statistically downscaled rainfall is better than the corresponding CORDEX rainfall for the validation period. It can also be concluded that the projected multisite daily precipitation based on KR conditioned on weather states is less uncertain than the CORDEX daily precipitation.

Limitations of the present study

In the present study, two RCPs (i.e. 4.5 and 8.5) which represent the most probable and worst-case scenario respectively, are considered for statistical downscaling of rainfall of Tapi basin. The downscaled rainfall for the remaining RCPs (i.e. 2.6 and 6.5) may be different from the results obtained in this study, and may be the future topic of study. A total of five GCMs (i.e. CCCma, CMCC, BCC, CNRM-CERFACS and MPI-M) were considered for the present study and the results show the variation in rainfall for future time periods. The remaining GCM models may be used to downscale the future rainfall and quantify the uncertainty in the projections in future. There are four reanalysis data generally found in the literature[13] and are freely downloadable. In the present study, the NCEP/NCAR reanalysis-I data are used. However, while carrying out such a study, the reader should be aware about the resolution and reliability of the data. The downscaling result may vary due to use of other reanalysis data.

The methodology used for downscaling daily rainfall data is based on the *k*-means clustering, which is an

unsupervized clustering technique; unsupervized clustering may sometimes misclassify the rainfall state, owing to the rainfall field falling on the fringes of two clusters. A better weather-typing scheme may be used in future to identify the exact weather state of the region. The KR-based model used in the present study considers the present-day rainfall state for prediction of multisite rainfall in the Tapi Basin. However, there is scope to include the previous-day rainfall state in the present downscaling scheme for better prediction. The bandwidth estimation in KR is important for realistic simulation of rainfall amount. The present KR-based downscaling technique uses the simple bandwidth formula. There is scope for improving the kernel bandwidth formula in this study. The KR-based downscaling technique assumes stationary relationship between the reanalysis predictor and predictand (rainfall) for projection of future rainfall. There is scope to address the changing relationship between large-scale predictors of GCM and original precipitation simulated by the predictand. Also, there is uncertainty resulting from the use of different models (sources) which needs further studies.

1. Kannan, S. and Ghosh, S., A nonparametric kernel regression model for downscaling multisite daily precipitation in the Mahanadi basin. *Water Resour. Res*., 2013, **49**(3), 1360–1385.
2. Ghosh, S. and Mujumdar, P. P., Future rainfall scenario over Orissa with GCM projections by statistical downscaling. *Curr. Sci.*, 2006, **90**(3), 396–404.
3. Anandhi, A., Srinivas, V. V., Nanjundiah, R. S. and Kumar, D. N., Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine. *Int. J. Climatol.*, 2008, **28**, 401–420.
4. Ghosh, S. and Mujumdar, P. P., Statistical downscaling of GCM simulations to streamflow using relevance vector machine. *Adv. Water Resour.*, 2008, **31**(1), 132–146.
5. Mehrotra, R. and Sharma, A., Development and application of a multisite rainfall stochastic downscaling framework for climate change impact assessment. *Water Resour. Res*., 2010, **46**(7); doi: 10.1029/2009WR008423.
6. Ghosh, S., SVM-PGSL coupled approach for statistical downscaling to predict rainfall from GCM output. *J. Geophys. Res.: Atmos.*, 2010, **115**(D22), doi: 10.1029/2009JD013548.
7. Groppelli, B., Bocchiola, D. and Rosso, R., Spatial downscaling of precipitation from GCMs for climate change projections using random cascades: a case study in Italy. *Water Resour. Res.*, 2011, **47**(3); doi: 10.1029/2010WR009437.
8. King, L. M., Irwin, S., Sarwar, R., McLeod, A. I. A. and Simonovic, S. P., The effects of climate change on extreme precipitation events in the Upper Thames River Basin: a comparison of downscaling approaches. *Can. Water Resour. J.*, 2012, **37**(3), 253–274.
9. Salvi, K., Kannan, S. and Ghosh, S., High-resolution multisite daily rainfall projections in India with statistical downscaling for climate change impacts assessment. *J. Geophys. Res.: Atmos.*, 2013, **118**(9), 3557–3578.
10. Okkan, U. and Inan, G., Bayesian learning and relevance vector machines approach for downscaling of monthly precipitation. *J. Hydrol. Eng.*, 2014; 10.1061/(ASCE)HE.1943-5584.0001024.
11. Okkan, U. and Fistikoglu, O., Evaluating climate change effects on runoff by statistical downscaling and hydrological model GR2M. *Theor. Appl. Climatol.*, 2014, **117**(1–2), 343–361.
12. Gutmann, E., Pruitt, T., Clark, M. P., Brekke, L., Arnold, J. R., Raff, D. A. and Rasmussen, R. M., An intercomparison of statistical downscaling methods used for water resource assessments in the United States. *Water Resour. Res.*, 2014, **50**(9), 7167–7186.
13. Bárdossy, A., Downscaling from GCMs to local climate through stochastic linkages. *J. Environ. Manage.*, 1997, **49**(1), 7–17.
14. Hewitson, B. C. and Crane, R. G., Climate downscaling: techniques and application. *Climate Res.*, 1996, **7**, 85–95.
15. Murphy, J., An evaluation of statistical and dynamical techniques for downscaling local climate. *J. Climate*, 1999, **12**(8), 2256–2284.
16. Mujumdar, P. P. and Nagesh Kumar, D., *Floods in Changing Climate: Hydrologic Modelling*, International Hydrology Series, Cambridge, 2013, pp. 43–87.
17. Salvi, K., Ghosh., S. and Ganguly, A. R., Credibility of statistical downscaling under nonstationary climate. *Climate Dyn.*, 2015; doi: 10.1007/s00382-015-2688-9.
18. Kannan, S. and Ghosh, S., Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output. *Stoch. Environ. Res. Risk Assess.*, 2011, **25**(4), 457–474.
19. Raje, D. and Mujumdar, P. P., A conditional random field-based downscaling method for assessment of climate change impact on multisite daily precipitation in the Mahanadi basin. *Water Resour. Res.*, 2009, **45**(10); doi: 10.1029/2008WR007487.
20. Shashikanth, K., Salvi, K., Ghosh, S. and Rajendran, K., Do CMIP5 simulations of Indian summer monsoon rainfall differ from those of CMIP3? *Atmos. Res. Lett.*, 2013; doi: 10.1002/as12.466.
21. Knutti, R. and Sedláček, J., Robustness and uncertainities in the new CMIP5 climate model projections. *Nat. Climate Change*, 2012, **3**(4), 369–373.
22. Zhou, B., Wen, Q. H., Xu, Y., Song, L. and Zhang, X., Projected changes in temperature and precipitation extremes in China by the CMIP5 multimodel ensembles. *J. Climate*, 2014, **27**(17), 6591–6611.
23. Vuuren, D. P. *et al.*, The representative concentration pathways: an overview. *Climatic Change*, 2011, **109**, 5–31.
24. Murari, K. K., Ghosh, S., Patwardhan, A., Daly, E. and Salvi, K., Intensification future severe heat waves in India and their effect on heat stress and mortality. *Reg. Environ. Change*, 2014; doi: 10.1007/s10113-014-0660-6.
25. Jain, S. K., Agarwal, P. K. and Singh, V. P., *Hydrol. Water Resour. India*, Springer, The Netherlands, 2007, pp. 561–595.
26. Timbadiya, P. V., Patel, P. L. and Porey, P. D., One-dimensional hydrodynamic modelling of flooding and stage hydrographs in the lower Tapi River in India. *Curr. Sci.*, 2014, **106**(5), 708–716.
27. Raje, D., Ghosh, S. and Mujumdar, P. P., Hydrologic impacts of climate change: quantification of uncertainties. In *Climate Change Modeling, Mitigation and Adaptation* (eds Surampalli, R. Y. *et al.*), American Society of Civil Engineers, Reston, Virginia, USA, 2013, pp. 178–218.
28. Mishra, V., Kumar, D., Ganguly, A. R., Sanjay, J., Mujumdar, M., Krishnan, R. and Shah, R. D., Reliability of regional and global climate models to simulate precipitation extremes over India. *J. Geophys. Res.: Atmos.*, 2014; doi: 10.1002/2014JD021636.
29. Yatagai, A., Arakawa, O., Kamiguchi, K., Kawamoto, H., Nodzu, M. I. and Hamada, A., A 44-year daily gridded precipitation dataset for Asia based on a dense network of rain gauges. *Sola*, 2009, **5**, 137–140.
30. Rajeevan, M. and Bhate, J., A high resolution daily gridded rainfall dataset (1971–2005) for mesoscale meteorological studies. *Curr. Sci.*, 2009, **96**(4), 558–562.
31. Krishnamurti, T. N., Mishra, A. K., Simon, A. and Yatagai, A., Use of a dense rain-gauge network over India for improving blended TRMM products and downscaled weather models. *J. Meteorol. Soc. Japan*, 2009, **87**, 393–412.

32. Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L. and Joseph, D., The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.*, 1996, **77**(3), 437–471.
33. Dunn, J. C., A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybernetics*, 1973, **3**(3), 32–57.
34. Davies, D. L. and Bouldin, D. W., A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1979, **2**, 224–227.
35. Rousseeuw, P. J., Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 1987, **20**, 53–65.
36. Maity, R. and Nagesh Kumar, D., Basin-scale stream-flow forecasting using the information of large-scale atmospheric circulation phenomena. *Hydrol. Processes*, 2008, **22**(5), 643–650.
37. Nadaraya, E. A., On estimating regression. *Theory Prob. Appl.*, 1964, **9**(1), 141–142.
38. Singh, N. and Ranade, A., The wet and dry spells across India during 1951–2007. *J. Hydrometeorol.*, 2010, **11**(1), 26–45.
39. Goswami, B. N., Venugopal, V., Sengupta, D., Madhusoodanan, M. S. and Xavier, P. K., Increasing trend of extreme rain events over India in a warming environment. *Science*, 2006, **314**(5804), 1442–1445.
40. Chow, V. T., Maidment, D. R. and Mays, L. W., *Applied Hydrology*, McGraw-Hill Company, New York, 1988, pp. 380–415.
41. Takayabu, I., Kanamaru, H., Dairaku, K., Benestad, R., von Storch, H. and Christensen, J. H., Reconsidering the quality and utility of downscaling. *J. Meteorol. Soc. Jpn. Ser. II*, 2016, **94A**, 31–45; doi: 10.2151/jmsj.2015-042.

ACKNOWLEDGEMENTS. This work is partly funded by S. V. National Institute of Technology (SVNIT), Surat and the Centre of Excellence on Water Resources and Flood Management at SVNIT, Surat under the Technical Education Quality Improvement Programme (TEQIP-II). S.K. and P.V.T. thank the Director, SVNIT, Surat and Director, Central Water and Power Research Station, Pune for permitting for joint guidance to S.S. P.V.T. thanks Dr Subimal Ghosh, Department of Civil Engineering, IIT-Bombay for fruitful interactions on methodology of the work. The data for this work were obtained from http://www.chikyu.ac.jp/precip, http://www.esrl.noaa.gov/psd/data/reanalysis/reanalysis.shtml and CMIP-5 experiment site. We also thank Dr Milind Mujumdar, Indian Institute of Tropical Meteorology, Pune for providing access to CORDEX data, and the two anonymous reviewers for critical comments that helped improve the manuscript.

Received 14 February 2015; revised accepted 24 October 2015

doi: 10.18520/cs/v110/i8/1468-1484