Nonparametric methods for modeling GCM and scenario uncertainty in drought assessment

Subimal Ghosh[1] and P. P. Mujumdar[1]

Received 17 July 2006; revised 26 January 2007; accepted 22 February 2007; published 6 July 2007.

[1] Hydrologic implications of global climate change are usually assessed by downscaling appropriate predictors simulated by general circulation models (GCMs). Results from GCM simulations are subjected to a number of uncertainties due to incomplete knowledge about the underlying geophysical processes of global change (GCM uncertainties) and due to uncertain future scenarios (scenario uncertainties). With a relatively small number of GCMs available and a finite number of scenarios simulated by them, uncertainties in the hydrologic impacts at a smaller spatial scale become particularly pronounced. In this paper, a methodology is developed to address such uncertainties for a specific problem of drought impact assessment with results from GCM simulations. Samples of a drought indicator are generated with downscaled precipitation from available GCMs and scenarios. Since it is very unlikely that such small samples resulting from GCM scenarios fit a known parametric distribution, nonparametric methods such as kernel density estimation and orthonormal series methods are used to determine the probability distribution function (PDF) of the drought indicator. Principal component analysis, fuzzy clustering, and statistical regression are used for downscaling the mean sea level pressure (MSLP) output from the GCMs to precipitation at a smaller spatial scale. Reanalysis data from the National Center for Environmental Prediction (NCEP) are used in relating precipitation with MSLP. The information generated through the PDF of the drought indicator in a future year may be used in long-term planning decisions. The methodology is demonstrated with a case study of the drought-prone Orissa meteorological subdivision in India.

Citation: Ghosh, S., and P. P. Mujumdar (2007), Nonparametric methods for modeling GCM and scenario uncertainty in drought assessment, *Water Resour. Res.*, *43*, W07405, doi:10.1029/2006WR005351.

1. Introduction

[2] General circulation models (GCMs) are tools designed to simulate time series of climate variables for the world, accounting for the effects of the concentration of greenhouse gases in the atmosphere [*Prudhomme et al.*, 2003]. Coupled with projections of CO_2 emission rates, they produce climate scenarios that can be described as ''pertinent, plausible representations of the future emissions of greenhouse gases and with the understanding of the effect of increased atmospheric concentration of the gases on global climate'' [*IPCC-TGCIA*, 1999]. They are currently the most credible tools available for simulating the response of the global climate system to increasing greenhouse gas concentrations, and they provide estimates of climate variables (for example, air temperature, precipitation, wind speed, pressure, etc.) on a global scale. 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 [*Hughes and Guttorp*, 1994]. Additionally, the spatial scale on which a GCM can operate [for example, 3.75° longitude × 3.75° latitude for coupled global climate model (CGCM2)] is very coarse for hydrologic applications [*Prudhomme et al.*, 2003]. Downscaling is therefore necessary to model the hydrologic variables (for example, precipitation) at a smaller scale based on larger-scale GCM outputs. Dynamic downscaling uses complex algorithms at a fine-grid scale (typically of the order of 50 × 50 km) describing atmospheric processes nested within the GCM outputs [*Jones et al.*, 1995] to result typically in limited-area models or regional climate models (RCM), whereas statistical downscaling produces future scenarios based on statistical relationships between large-scale climate features (for example, circulation pattern) and hydrologic variables [*Wilby et al.*, 1998]. A major assumption in the 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 climate [*Easterling*, 1999]. Compared with dynamic downscaling, statistical downscaling has the advantage of being computationally simple and easily adjustable to new areas. The method generally requires very few parameters, and this makes it attractive for many hydrologic applications [*Wilby et al.*, 2000].

[1]Department of Civil Engineering, Indian Institute of Science, Bangalore, India.

0043-1397/07/2006WR005351$09.00

A comparison of statistical and dynamic downscaling in climate change impact assessment on precipitation may be found in the work of *Haylock et al.* [2006].

[3] Statistical downscaling methodologies can be broadly classified into three categories: weather generators, weather typing, and transfer function. Weather generators are statistical models of observed sequences of weather variables. They can also be regarded as complex random number generators, the output of which resembles daily weather data at a particular location. There are two fundamental types of daily weather generators based on the approach to model daily precipitation occurrence: the Markov chain approach [*Hughes et al.*, 1993; *Hughes and Guttorp*, 1994] and the spell-length approach [*Wilks*, 1999]. Weather-typing approaches involve grouping of local, meteorological variables in relation to different classes of atmospheric circulation. Future regional climate scenarios are constructed either by resampling from the observed variable distribution (conditional on the circulation pattern produced by a GCM) or by first generating synthetic sequences of weather pattern using Monte Carlo techniques and resampling from the generated data. The mean, or frequency distribution of the local climate, is then derived by weighting the local climate states with the relative frequencies of the weather classes. *Bardossy et al.* [1995] used a fuzzy rule-based technique for the classification of circulation patterns into different states. Stochastic models such as Markov chains may be used to predict precipitation from different states of classified circulation patterns [*Bardossy and Plate*, 1991; *Bardossy and Plate*, 1992; *Stehlik and Bardossy*, 2002]. The most popular approach of downscaling is the use of transfer function which is a regression-based downscaling method that relies on direct quantitative relationship between the local-scale climate variable (predictand) and the variables containing the large-scale climate information (predictors) through some form of regression. Individual downscaling schemes differ according to the choice of mathematical transfer function, predictor variables, or statistical fitting procedure. To date, linear and nonlinear regressions [*Wilby et al.*, 1998], artificial neural network [*Wilby et al.*, 1998; *Tripathi and Srinivas*, 2005], fuzzy rule-based system [*Bardossy et al.*, 2005], support vector machine [*Tripathi et al.*, 2006], analogue method [*Wetterhall et al.*, 2005, *Gutierrez et al.*, 2004], etc. have been used to derive predictor-preditand relationship. A combination of classification-based weather typing and transfer function method for downscaling may be found in the work of *Ghosh and Mujumdar* [2006], where principal component analysis (PCA), fuzzy clustering, and linear regression with seasonality term have been used for downscaling mean sea level pressure (MSLP) to precipitation. A completely different and unique approach of inverse modeling may be found in the papers of *Cunderlik and Simonovic* [2004] and *Prodanovic et al.* [2005], where critical meteorological situations are found from critical hydrologic events. In the final stage, the frequency of critical weather situations is investigated under future climatic conditions obtained from GCM. Since the analysis of GCM outputs is one of the last steps in this methodology, the approach allows easy updating when new and improved GCM outputs become available. Detailed discussions on different models used for downscaling may be found in the studies of *Leavesley* [1994] and *Prudhomme et al.* [2002].

[4] Climate change impact assessment models developed based on GCM output are subjected to a range of uncertainties due to both ''incomplete knowledge'' and ''unknowable future scenario'' [*Hulme and Carter*, 1999; *New and Hulme*, 2000]. ''Incomplete knowledge'' mainly arises from inadequate information and understanding about the underlying geophysical process of global change, leading to limitations in the accuracy of GCMs. This can also be termed as GCM uncertainty. Uncertainty due to ''unknowable future scenario'' is associated with the unpredictability in the forecast of future socioeconomic and human behavior resulting in future greenhouse gas (GHG) emission scenarios and can also be termed as scenario uncertainty. Scenarios are alternative images of how the future might unfold and are an appropriate tool with which to analyze how driving forces may influence future emission outcomes and to assess the associated uncertainties. A basic assumption in the development of a scenario is that all such scenarios are equally possible in the future. The choice of impact model (structure and parameterization) is also an another important source of uncertainty that is increasingly recognized. Downscaled outputs of a single GCM with a single climate change scenario represents a single trajectory among a number of realizations derived using various scenarios with GCMs. Such a single trajectory alone, therefore, cannot represent a future hydrologic scenario and will not be useful in assessing hydrologic impacts due to climate change. No quantified probability is attached to the simulated outcome of a single GCM for a single scenario, and thus the approach of downscaling a single GCM output is not particularly useful for risk adaptation studies [*New and Hulme*, 2000]. In the study of *Benestad* [2004], regional temperature scenarios were presented for northern Europe in the form of probability distributions, based on spatially interpolated empirically downscaled trends, derived using a multimodel ensemble as well as various downscaling options, and it was found that spatial warming rate patterns, derived from the individual models, exhibit large differences. *Simonovic and Li* [2003, 2004] have shown the uncertainty lying in climate change impact studies on flood protection resulting from selection of GCMs and scenarios. Available GCM outputs have been used by them for assessing effectiveness of flood protection system, and it has been concluded that different GCMs provide different estimates of the hydrologic parameters. Using several climate change scenarios with several GCMs provides the user of the impact study with a range of possible outcomes but, again, with no attached probabilities [*New and Hulme*, 2000].

[5] *New and Hulme* [2000] developed a model for scenario uncertainty using Bayesian Monte Carlo approach assuming a prior distribution of the uncertain parameters of the climate models. GCM and scenario uncertainty is presented in terms of sensitivity of climate change model outputs to streamflow. Similar methodology for sensitivity analysis and risk assessment of irrigation demand may be found in the work of *Jones* [2000]. A simple probabilistic energy balance model, which samples the uncertainty in greenhouse gas emissions, the climate sensitivity, the

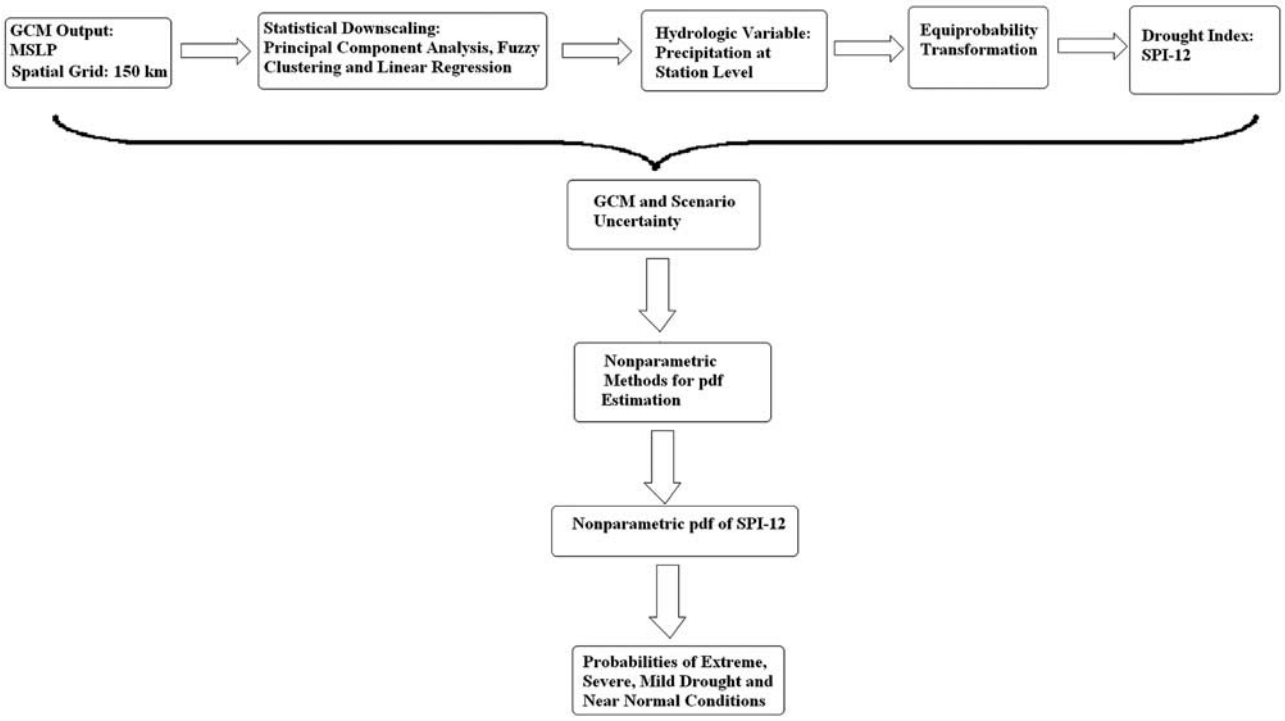

Figure 1. Overview of the method.

carbon cycle, the ocean mixing, and the aerosol forcing, has been used by *Dessai et al.* [2005] to quantify uncertainty in regional climate change projections. Assignment of global mean temperature probabilities in GCMs through pattern-scaling technique has been suggested in that study. In order to combine resulting probabilities, the regional skill scores for each GCM, season, and climate variable (surface temperature and precipitation) are devised in 23 world regions based on model performance and model convergence. A range of sensitivity experiments are carried out with different skill score schemes, climate sensitivities, and emission scenarios for performing sensitivity analysis of regional climate change probabilities.

[6] The above mentioned literature on modeling GCM and scenario uncertainty limit themselves in representing uncertainty by performing sensitivity analysis of hydrologic events to climatic parameters. However, implications of such uncertainty in estimating the severity of future extreme events, such as floods and droughts, with a probabilistic approach have not been addressed there. Research into probabilistic forecasts of climate change has been advancing rapidly on several fronts. For example, there have been systematic evaluations of uncertainties due to climate model projections using multimodel ensembles [*Raisanen and Palmer*, 2001; *Giorgi and Mearns*, 2003]; multiensemble experiments with one GCM [*Murphy et al.*, 2004]. Bayesian methods have been applied to multimodel ensembles to characterize uncertainty and probability distribution functions (PDFs) for future climate changes at regional scales [*Tebaldi et al.*, 2004, 2005]. In a more recent study, *Wilby and Harris* [2006] developed a probabilistic framework for modeling GCM and scenario uncertainty, where GCMs were weighted according to an index of reliability for downscaled effective rainfall. A Monte Carlo approach was then used to explore components of uncertainty affecting projections for the River Thames by the 2080s. It was found that the resulting cumulative distribution functions (CDFs) of low flows were most sensitive to uncertainty in the climate change scenarios and downscaling of different GCMs.

[7] The present study attempts to answer the specific question of interpreting the available outputs from GCMs with different scenarios in assessing the severity of future drought, addressing both GCM and scenario uncertainty. Uncertainty due to structure and parameterization is not considered in this work to keep the focus of the work on modeling GCM and scenario uncertainty. An overview of the methodology proposed in this work is presented in Figure 1. Fuzzy clustering-based downscaling [*Ghosh and Mujumdar*, 2006] is used for modeling future precipitation using circulation pattern, projected with the available GCM outputs. Standardized precipitation index (SPI) developed by *McKee et al.* [1993] is used as a drought index which requires precipitation as an input variable. Assuming future SPI to be a random variable at every time step, methodologies based on kernel density and orthonormal systems are used to determine the nonparametric PDF of SPI, as it is very unlikely that the small sample of available GCM outputs will follow a particular parametric distribution. Probabilities for different categories of future drought are computed from the estimated PDF. The methodology is applied to the case study of Orissa meteorological subdivision in India to analyze the severity of different degrees of drought in the future.

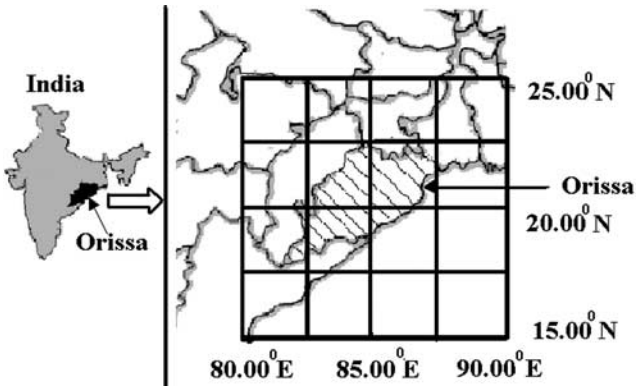

Figure 2. NCEP grids superposed on map of Orissa.

[8] The following section presents details of the case study area, data extraction, and downscaling technique used for the analysis.

2. Data Extraction and Statistical Downscaling

[9] The Orissa meteorological subdivision located on the eastern coast of India, extends from 17° to 22°N in latitude and from 82° to 87°E in longitude. The monthly area-weighted precipitation data of Orissa meteorological subdivision in India, from January 1950 to December 2002, is obtained from Indian Institute of Tropical Meteorology, Pune (http://www.tropmet.res.in). This data set is used in the downscaling as predictand. The primary source of this data is the India Meteorological Department. Selection of predictor is an important step in statistical downscaling. The predictors used for downscaling [*Wilby et al.*, 1999; *Wetterhall et al.*, 2005; *Tripathi et al.*, 2006] should be: (1) reliably simulated by GCMs, (2) readily available from archives of GCM outputs, and (3) strongly correlated with the surface variables of interest. Precipitation can be related to air mass transport and thus can be related to atmospheric circulation, which is a consequence of pressure differences and anomalies [*Bardossy*, 1997], and thus circulation pattern is used as the predictor for downscaling in most of the earlier models [e.g., *Bardossy and Plate*, 1991; *Hughes and Guttorp*, 1994; *Bardossy et al.*, 1995; *Wetterhall et al.*, 2005]. On the basis of these studies, the present methodology uses MSLP as predictor for downscaling. Gridded MSLP data used in the downscaling are obtained from the National Center for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis project [*Kalnay et al.*, 1996; http://www.cdc.noaa.gov/cdc/reanalysis/reanalysis.shtml]. Reanalysis data are outputs from a high-resolution atmospheric model that has been run using data assimilated from surface observation stations, upper-air stations, and satellite-observing platforms. Results obtained using these fields therefore represent those that could be expected from an ideal GCM [*Cannon and Whitfield*, 2002]. Monthly average MSLP outputs from 1948 to 2002 were obtained for a region spanning 15°–25°N in latitude and 80°–90°E in longitude that encapsulates the study region. Figure 2 shows the NCEP grid points superposed on the map of Orissa meteorological subdivision. A statistical relationship based on fuzzy clustering and linear regression is developed between MSLP and precipitation, with reanalysis data of MSLP as predictor and observed precipitation as predictand. This relationship is used to model the future precipitation using available GCM projections of MSLP. Table 1 gives a list of GCMs with available scenarios. The outputs of MSLP of GCMs with scenarios, as given in Table 1, are extracted from the IPCC data distribution center (http://www.mad.zmaw.de/IPCC_DDC/html/ddc_gcmdata.html) for the region covering all the NCEP grid points.

[10] An overview of the statistical downscaling technique used here to model future precipitation from GCM-projected circulation pattern is presented in Figure 3. The method involves training NCEP data of circulation pattern with observed precipitation and the use of the resulting regression relationship in modeling future precipitation from GCM projections. The training involves three steps [*Ghosh and Mujumdar*, 2006]: PCA, fuzzy clustering, and linear regression with seasonality terms. Standardization [*Wilby et al.*, 2004] is used prior to statistical downscaling to reduce systematic biases in the mean and variances of GCM predictors relative to the observations or NCEP/NCAR data. The procedure typically involves subtraction of mean and division by standard deviation of the predictor variable for a predefined baseline period for both NCEP/NCAR and GCM outputs. The period 1961–1990 is used as a baseline because it is of sufficient duration to establish a reliable climatology and is yet not too long nor too contemporary to include a strong global change signal [*Wilby et al.*, 2004]. For the Orissa meteorological subdivision, MSLP values at 25 NCEP grid points are used as predictor, which are highly correlated with each other. PCA is used to convert them into a set of uncorrelated variables. It was found that 99.7% of the variability of original data set is explained by the first three

Table 1. GCMs Used and Available Scenarios

GCM	Organization	Scenarios Available
CCSR/NIES coupled GCM	Center for Climate Research Studies (CCSR) and National Institute for Environmental Studies (NIES), Japan	A1, A2, B1, B2
Second-generation coupled global climate model (CGCM2)	Canadian Center for Climate Modelling and Analysis, Canada	IS92a, A2, B2
HadCM3	Hadley Centre for Climate Prediction and Research (HCCPR), UK	IS95a, (GHG + ozone + sulphate), A2
ECHAM4/OPYC3	Max Planck Institute für Meteorologie, Germany	IS92a, A2, B2
CSIRO-MK2	Australia's Commonwealth Scientific and Industrial Research Organisation (CSIRO)	(IS92a + sulphate), IS92a, A1, A2, B1, B2

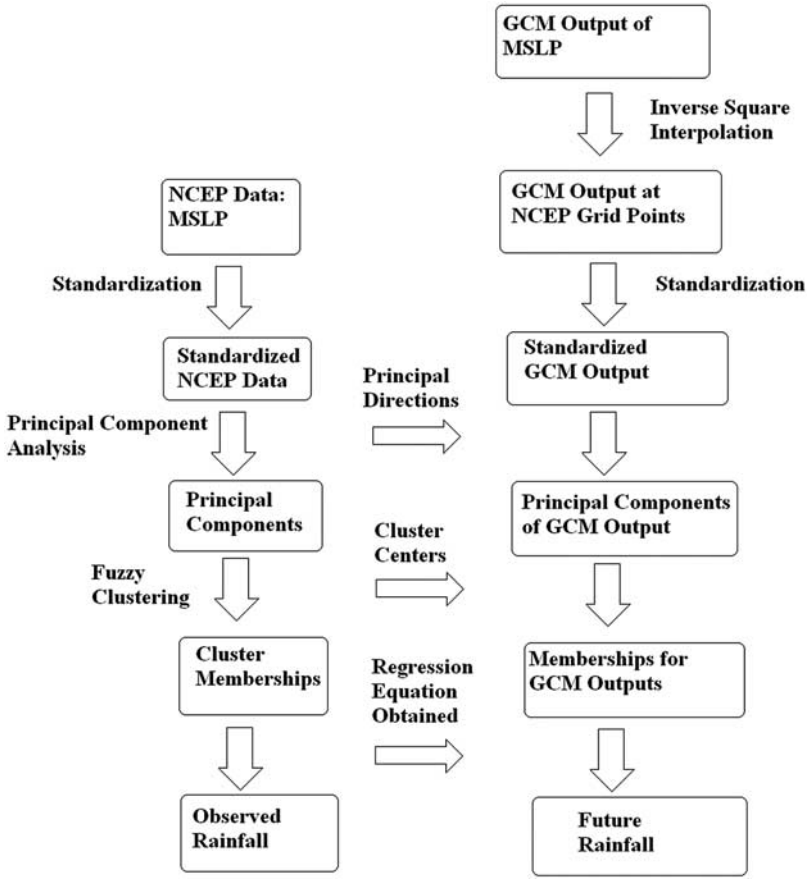

Figure 3. Statistical downscaling with fuzzy clustering.

principal components, and therefore only the first three principal components are used for modeling precipitation. Fuzzy clustering is used to classify the principal components into classes or clusters. Fuzzy clustering assigns membership values of the classes to various data points, and it is more generalized and useful to describe a point by its membership values, not by a crisp cluster, in all the clusters.

[11] The important parameters required for the fuzzy clustering algorithm are the number clusters (c) and the fuzzification parameter (m). The fuzzification parameter controls the degree of the fuzziness of the resulting classification, which is the degree of overlap between clusters. The minimum value of m is 1 which implies hard clustering. The number of clusters and the fuzzification parameter are determined from cluster validity indices like fuzziness performance index (FPI) and normalized classification entropy (NCE) [*Roubens*, 1982]. FPI estimates the degree of fuzziness generated by a specified number of classes and is given by:

$$\mathrm{FPI} = 1 - \frac{cF - 1}{c - 1} \tag{1}$$

where

$$F = \frac{1}{T}\sum_{i=1}^{c}\sum_{t=1}^{T}(\mu_{it})^2 \tag{2}$$

μ_{it} is the membership in cluster i of the principal components in time t. NCE estimates the degree of disorganization created by a specified number of classes and is given as:

$$\mathrm{NCE} = \frac{H}{\log c} \tag{3}$$

where

$$H = \frac{1}{T}\sum_{i=1}^{c}\sum_{t=1}^{T} -\mu_{it} \times \log(\mu_{it}) \tag{4}$$

[12] The optimum number of classes/clusters is established on the basis of minimizing these two measures. The clustering becomes nonfuzzy when FPI = 0 and turns into fully fuzzy when FPI = 1 [*Güler and Thyne*, 2004]. The value of FPI should be chosen in such a way that the resulting clustering is neither too fuzzy nor too hard. *Güler and Thyne* [2004] have recommended an FPI value of 0.25 for the purpose of selection of number of clusters and fuzzification parameter in fuzzy clustering. In this work, FPI and NCE are plotted with the number of clusters c for the different values of fuzzification parameter m (Figure 4). It is found that FPI value of 0.25 is achieved for $m = 2.0$ and $c = 2$. *Ross* [1997] also recommends a default value of $m = 2.0$. Therefore the number of clusters is selected as 2, and in clustering algorithm, the value of m is considered as 2.0.

[13] Linear regression is used to model the monthly precipitation with principal components, membership values

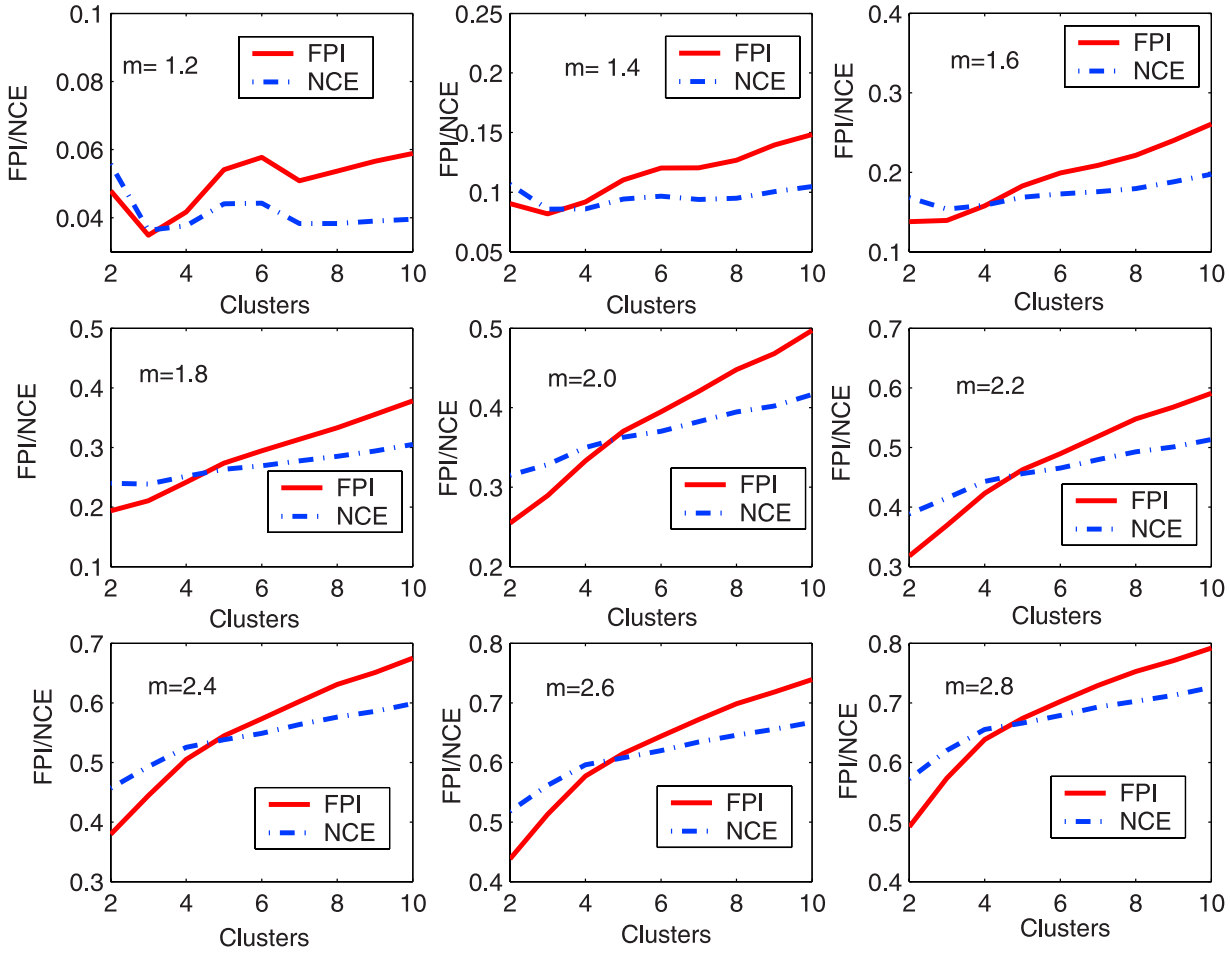

Figure 4. Cluster validity test (color).

of the principal components in each of the clusters, and the cross product of membership values and principal components as regressors. An appropriate seasonality term is used to capture the seasonality. The linear regression equation is given by:

$$P_t = C + \sum_{i=1}^{I-1} \beta_i \times \mu_{it} + \sum_{k=1}^{K} \gamma_k \times pc_{kt} + \sum_{i=1}^{I-1} \sum_{k=1}^{K} \rho_{ik} \times \mu_{it} \times pc_{kt} \quad (5)$$

with

$$C = C^0 + C^1 \times \sin(2\pi p/12) + C^2 \times \cos(2\pi p/12) \quad (6)$$

$$\beta_i = \beta_i^0 + \beta_i^1 \times \sin(2\pi p/12) + \beta_i^2 \times \cos(2\pi p/12) \quad (7)$$

$$\gamma_k = \gamma_k^0 + \gamma_k^1 \times \sin(2\pi p/12) + \gamma_k^2 \times \cos(2\pi p/12) \quad (8)$$

$$\rho_{ik} = \rho_{ik}^0 + \rho_k^1 \times \sin(2\pi p/12) + \rho_{ik}^2 \times \cos(2\pi p/12) \quad (9)$$

where, P_t is the precipitation in time t, pc_{kt} is the kth principal component of circulation pattern in time t, and μ_{it} is the membership in cluster i of the principal components in time t. K and I are the number of principal components used and the number of clusters, respectively. β_i, γ_k, and ρ_{ik} are the coefficients of μ_{it}, pc_{kt}, and their product terms, respectively. C is the constant term used in the equation. The membership values μ_{it} in each cluster are assigned to the different points based on fuzzy c-means algorithm. These membership values lie between 0 and 1. The $(I - 1)$ number of clusters is adequate to model the regression equation as the sum of the membership values in all the clusters at time t is 1, and thus $(I - 1)$ memberships will automatically fix the value of Ith membership and therefore the Ith membership will be a redundant input variable to the regression model. Seasonality is incorporated by equations (6)–(9), where p is the serial number of the month within a year (p = 1, 2, 3, …, 12). Correlation coefficient (r) between the observed and predicted precipitation is considered as the goodness of fit of the regression model. Here the r value is obtained as 0.924. Linear regression without fuzzy clustering, i.e., only with the principal components obtained from NCEP reanalysis data of MSLP, results in a lower r value of 0.803, which shows the importance of fuzzy clustering in the improvement of downscaling model fit.

[14] A drawback of the model is that a large number of regressors are used in the regression, which may lead to high multicollinearity and insignificant t statistic of the

Table 2. Bias of Downscaled Observed Precipitation Relative to Observed Data

Mean of Observed Annual Precipitation (1961–1990), mm	Mean of NCEP Downscaled Annual Precipitation (1961–1990), mm	GCM	Mean of GCM Downscaled Annual Precipitation (1961–1990), mm	Bias = Observed Mean Annual Precipitation − Downscaled Mean Annual Precipitation, mm
1394.2	1441.6	CGCM2	1558.3	−164.1
		CCSR/NIES	1542.2	−148.0
		HadCM3	1121.4	272.8
		ECHAM4/OPYC3	1387.8	6.4
		CSIRO-MK2	1322.6	71.6

regressors, with a chance of overfitting. Some of the regressors are found to have insignificant t statistic. A thumb rule for any linear regression without multicollinearity is that the condition index, which can be defined as the ratio of maximum to minimum eigenvalues of the matrix formed by the explanatory variables, should be less than 30 [*Gujrati*, 2004]. The condition index for this regression is 75.793 which shows high multicollinearity. This limitation is overcome by removing regressors having insignificant t statistic one by one from the regression equation without a significant change in R^2 value, which can be tested by checking the F statistic. SPSS 9.05, a data modeling tool (www.spss.com), is used to perform such regression based on F statistic. The resulting model gives the value of r as 0.924. The condition index of this model is 37.394. Although there is still a little multicollinearity, it has been significantly reduced from the previous model considering all the regressors. Also, the t statistic values of the regressors are significant that reduces the possibility of overfitting. To verify the model further, that there is no overfitting, a k-fold cross validation (k = 10) is performed, where the r values for training and testing are obtained as 0.9241 and 0.9225, respectively. Low difference between the r values of training and testing proves that the model is not characterized by overfitting.

[15] The goodness of fit of the model is also tested with the *Nash and Sutcliffe* [1970] coefficient, which has been recommended by the *ASCE Task Committee on Definition of Criteria for Evaluation of Watershed Models of the Watershed Management Committee, Irrigation and Drainage Division* [1993]. The Nash-Sutcliffe coefficient (E) is given by:

$$E = 1 - \frac{\sum_t \left(P_{ot} - \bar{P}_{pt}\right)^2}{\sum_t \left(P_{ot} - \bar{P}_{o}\right)^2} \tag{10}$$

where P_{ot} and P_{pt} are the observed and predicted precipitation in time t, respectively, and $\overline{P_o}$ is the mean observed precipitation. Nash-Sutcliffe coefficient can vary from 0 to 1, with 0 indicating that the model predicts no better than the average of the observed data, and with 1 indicating a perfect fit. It is obtained as 0.83 for the present model which is satisfactory. *Wetterhall et al.* [2005] have tested the long-term seasonal mean for verification of a downscaling model. In the present analysis also, similar test has been performed. For the wet period (June, July, August, and September), the long-term mean and median of observed precipitation are 281.4 and 281.9 mm/month, respectively, and those of predicted precipitation are 281.5 and 283.3 mm/month, respectively, which show a good match. Similar results are also obtained for the dry period. For the dry period (months other than June, July, August, and September), the long-term mean and median of observed precipitation are 74.9 and 73.8 mm/month, respectively, and those of predicted precipitation are 74.3 and 73.6 mm/month, respectively. After this verification, the model [equations (5), (6), (7), (8), and (9)] is used for modeling of future precipitation time series for different GCMs with different scenarios.

[16] GCM grid points do not match with NCEP grid points, and thus interpolation is required to obtain the GCM output at NCEP grid points. Interpolation is performed with a linear inverse square procedure using spherical distances [*Willmott et al.*, 1985]. For example, for GCM developed by CCSR/NIES, Japan, the grid size is 5.5° latitude × 5.625° longitude. The MSLP output is extracted for Orissa meteorological subdivision at 16 grid points extending from 13.8445° to 30.4576°N in latitude and from 78.7500° to 95.6250°E in longitude. These MSLP values are then interpolated to the 25 NCEP grid points. Statistical relationship [equations (5), (6), (7), (8), and (9)] obtained between MSLP and precipitation is then applied to these interpolated NCEP gridded GCM outputs to model precipitation. Similar to CCSR/NIES, all other GCMs, as given in Table 1, are used to simulate the precipitation for historic period and to project future precipitation of Orissa meteorological subdivision. The eigenvectors or principal directions obtained from NCEP data are used as reference to convert the gridded standardized GCM output to the corresponding principal components. Therefore weights, or principal directions/eigenvectors, are not calculated separately with the output of each of the GCMs; rather, the same reference principal directions or eigenvectors as obtained from NCEP output are used for all of the GCMs. Another alternative approach may be to blend GCM outputs with the NCEP reanalysis data for obtaining a universal set of principal components. For validation, the bias of annual mean of precipitation as downscaled from different standardized GCM outputs relative to observed data for baseline period is presented in Table 2. It is seen that even after standardization, the bias is not significantly reduced because the methodology may reduce the bias in the mean and variance of the predictor variable, but it is much harder to accommodate the biases in large-scale patterns of atmospheric circulation in GCMs (for example, shifts in the dominant storm track relative to observed data) or unrealistic intervariable relationships [*Wilby and Dawson*, 2004]. Discussion on biases of different GCMs after downscaling may also be found in

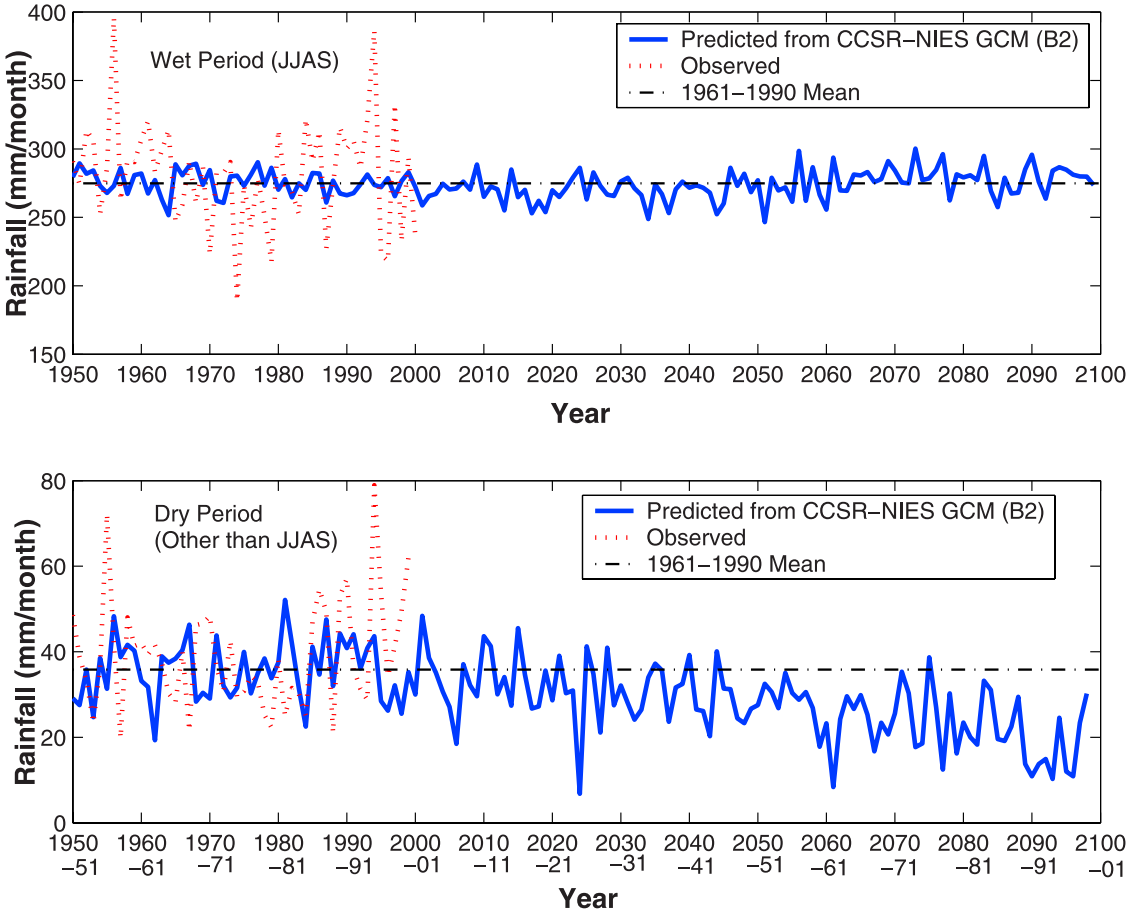

Figure 5. Rainfall for wet and dry periods with CCSR/NIES-B2 projection (color).

the work of *Wilby and Harris* [2006]. To remove the biases, 1961–1990 simulated mean is subtracted, and the observed baseline period mean is added, so that all the models have the same mean in the historic period, and thus the resulting uncertainty is solely due to GCM and scenario uncertainty and not due to biases present in the GCMs. Figure 5 shows the future projection of precipitation for wet (June, July, August, and September) and dry periods separately for CCSR/NIES GCM with B2 scenario. It is observed that the downscaling model significantly underestimates the interannual variability most notably in the wet season. A reason for this may be the insensitivity of MSLP in correctly modeling precipitation. MSLP can partially explain historic rainfall variation, but an improvement of the model is possible if moisture content or humidity is incorporated. In the present study the analysis is only limited with MSLP because for most of the GCMs listed in Table 1, the outputs of moisture content or humidity are not available. Figure 5 clearly indicates a slight increase in wet period precipitation and a severe decrease in dry-period precipitation for the particular scenario. The precipitation, thus computed for all the GCMs with scenarios, is converted into suitable drought indicator for examining future drought scenario.

3. Drought Indicators

[17] A drought indicator, briefly defined, is a variable to identify and assess drought conditions [*Steinemann*, 2003]. Common indicators are based on meteorologic and hydrologic variables such as precipitation, streamflow, soil moisture, reservoir storage, and groundwater levels. A drought trigger is a threshold value of the drought indicator that distinguishes a drought category and determines when drought response actions should begin or end. Drought categories typically represent levels of severity, such as mild, moderate, severe, or extreme drought. Commonly used drought indicators include standardized precipitation index (SPI), Palmer drought severity index (PDSI), crop moisture index (CMI), surface water supply index (SWSI), reclamation drought index (RDI), and deciles (http://www.drought.unl.edu/whatis/indices.htm).

[18] Most of the drought indicators stated above require multiple input data such as precipitation, available water content of soil, temperature, snowpack, reservoir storage, etc. The SPI is the simplest one which requires only precipitation as input and is generally computed for 3, 6, 12, and 48 months with notations of SPI-3, SPI-6, SPI-12, and SPI-48, respectively. Because of the computational simplicity and least input requirement, SPI is used for drought assessment in the present work. The analysis is performed for annual drought, and thus SPI-12 is used for examining the drought scenario.

[19] *McKee et al.* [1993] developed the SPI for the purpose of defining and monitoring drought. SPI is based on the probability distribution of precipitation and

Table 3. Drought Categories

Drought Category	SPI Values
Near normal	0 to −0.99
Mild-to-moderate drought	−1.00 to −1.49
Severe drought	−1.50 to −1.99
Extreme drought	−2.00 or less

requires only precipitation as the input data. SPI can be defined by the value of standard normal deviate corresponding to the cumulative distribution function (CDF) value of a precipitation event with a known probability distribution. A common procedure adopted for computing SPI is to fit a gamma distribution to the precipitation data, although the Pearson Type III has also been recommended, and then to transform the data to an equivalent SPI value based on the standard normal distribution [*Steinemann*, 2003]. Details of the methodology for calculation of SPI may be found in the Website of Colorado Climate Center, Colorado State University (http://ccc.atmos.colostate.edu/pub/spi.pdf). The standard procedure is as follows:

[20] 1. Fit a gamma distribution to the time series of nonzero precipitation for each timescale of interest (for example, 3, 12, 24, and 48 months, etc.) without overlapping of data segments. Compute the parameters of the gamma distribution.

[21] 2. Compute the value of CDF ($G(x)$) corresponding to each value of nonzero precipitation (x).

[22] 3. Compute the zero precipitation probability (q) from the historical time series. The value of CDF ($H(x)$) for a specific precipitation (x) will be:

$$H(x) = q + (1 - q) \times G(x) \tag{11}$$

[23] 4. Compute the value of standard normal deviate corresponding to the value of CDF ($H(x)$). This is the SPI value for the precipitation (x).

[24] On the basis of the value, the severity of drought can be assessed and categorized into different classes. Table 3 presents the categories of drought corresponding to their SPI values [*McKee et al.*, 1993; *Steinemann*, 2003].

[25] The parameters required for estimation of SPI, viz., parameters of gamma distribution and nonzero precipitation probability, are estimated based on the observed annual precipitation by fitting it to gamma distribution. Using these parameters, the future annual precipitation (computed from monthly precipitation), downscaled from GCM output, is converted into SPI-12. The SPI-12 is calculated for all GCMs for available scenarios. The projected SPI-12 is thus computed in Figure 6, which shows that SPI-12 time series downscaled from one GCM is entirely different from that of another and also a considerable dissimilarity exists among two scenarios of any particular GCM. The box plot presented in Figure 6 presents the sparseness of the SPI-12 values computed from different GCMs with scenarios for the years 2020, 2040, 2060, and 2080. A single time series of SPI-12 generated from a GCM for a particular scenario represents a single trajectory among a number of realizations derived using various scenarios with GCMs and cannot by itself represent the future drought condition. Such

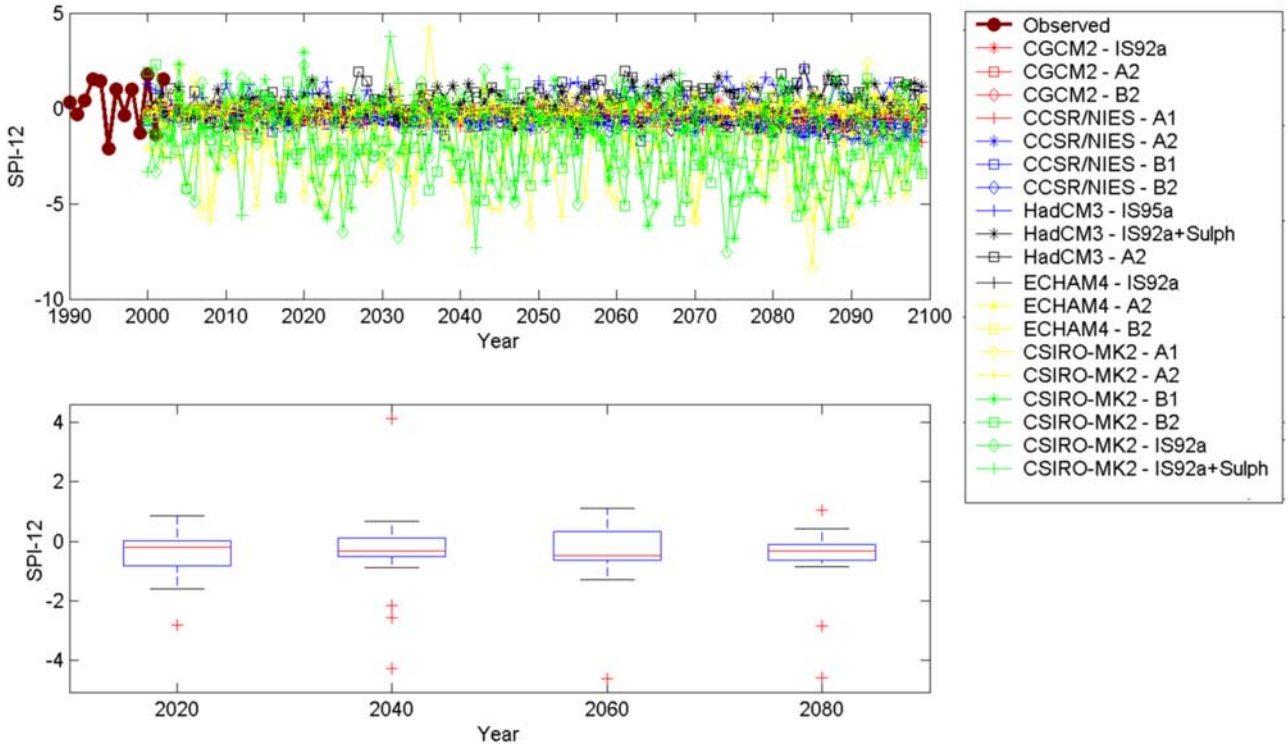

Figure 6. Predicted SPI-12 from GCM projections with different scenarios. The lower figure gives the box plots for 4 years (color).

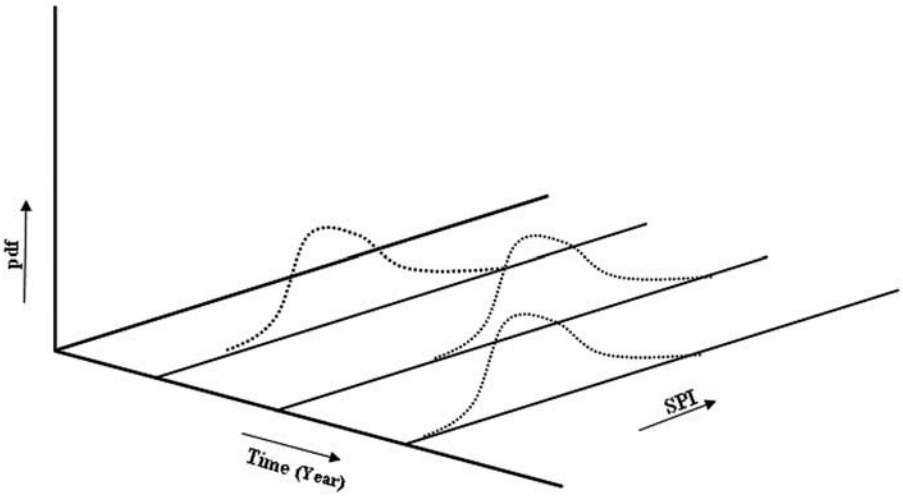

Figure 7. PDF of SPI-12 at each time step.

uncertainty due to GCMs and scenarios are modeled in a probabilistic framework to assess the severity of possible droughts in the future.

4. Modeling GCM and Scenario Uncertainty

[26] Climate change impact studies on hydrology, based on GCMs, are characterized by GCM and scenario uncertainty. The source of GCM uncertainty lies in inadequate information and understanding about the underlying geophysical process of global change leading to varied assumptions and limitations in GCM outputs. The unpredictability in the forecast of future socioeconomic and human behavior resulting in different greenhouse gas (GHG) emission scenarios leads to scenario uncertainty. Modeling of GCM and scenario uncertainty necessitates the use of a number of GCM outputs of different scenarios for risk-based studies of future hydrologic extremes.

[27] In the present work the SPI-12 values computed with downscaled outputs from GCMs are considered as the realizations of the random variable SPI-12 in each year where there exists a PDF of SPI-12 in each year (Figure 7). The severity of future drought may be studied by estimating the evolution of the PDF of a drought indicator. The simplest methodology of such analysis is based on the assumption of normal distribution for future SPI-12 in each year. However, it is very unlikely that SPI-12 will follow a normal distribution, and thus such analysis may lead to erroneous conclusion. In such cases, nonparametric PDFs estimated by a kernel density function with a suitable smoothing parameter are useful, as prior assumption of the data to follow a particular distribution can be avoided [*Lall et al.*, 1993]. Applications of kernel density estimation for determination of PDF for hydrologic variables may be found in the studies of *Lall* [1995], *Lall et al.* [1996], *Sharma et al.* [1997], and *Tarboton et al.* [1998]. Small sample size, however, may not result in accurate estimation of nonparametric PDF using kernel function. The methodology based on orthonormal series [*Efromovich*, 1999] for determination of nonparametric PDF from a small sample may be used to overcome this drawback. Here we discuss the use of all the three methods (viz., use of a normal distribution, kernel density estimation, and orthonormal series) for examining implications on future drought scenarios.

4.1. Assumption of Normal Distribution

[28] The simplest method of modeling a sample of data without prior knowledge of distribution is with an assumption of normal distribution. In the present case, we assume no prior information regarding the future distribution of SPI-12 and, for simplicity, assume a normal distribution. The results for each GCM and emission scenario is taken as the set of independent realizations of SPI-12 and that this set is used at each time step to establish the probability distribution. The values of the parameters of the normal distribution, i.e., mean and variance, are considered as the sample estimates and are obtained from the of SPI-12 projected from different GCMs with scenarios at a particular year. As SPI-12, with less than -2 value indicating extreme drought, the CDF value of SPI-12 at -2 will give the probability of extreme drought.

$$P(\text{extreme drought}) = F_{\text{SPI}}(-2) \tag{12}$$

[29] Similarly, the probability of other categories of drought at a particular year can be estimated from the CDF of the SPI-12 at that time. The probabilities of severe drought, mild-to-moderate drought, and near-normal condition are given by:

$$P(\text{severe drought}) = F_{\text{SPI}}(-1.5) - F_{\text{SPI}}(-2) \tag{13}$$

$$P(\text{mild drought}) = F_{\text{SPI}}(-1.0) - F_{\text{SPI}}(-1.5) \tag{14}$$

$$P(\text{near normal}) = F_{\text{SPI}}(0) - F_{\text{SPI}}(-1.0) \tag{15}$$

where $P(E)$ denotes the probability of an event E, and $F_{\text{SPI}}(x)$ denotes the value of CDF of SPI at x. A major limitation of this method is that there is no guarantee that SPI will follow a normal distribution. This may lead to erroneous results, but an idea about the trend of severity, i.e., whether the probability of extreme events increases or decreases, may be gathered from this analysis.

[30] Figure 8 shows the average probabilities of drought events for three time slices, years 2000–2010, 2040–2050,

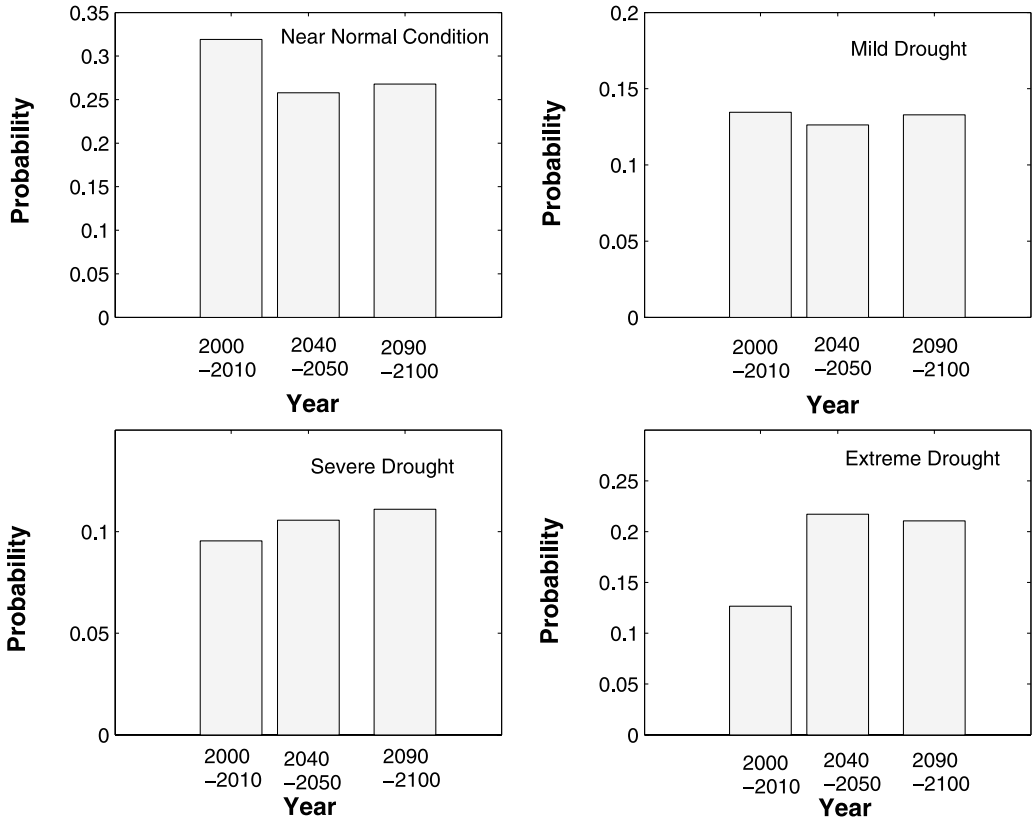

Figure 8. Probability of droughts with normal distribution for SPI-12.

and 2090–2100. Considerable variations in the probabilities of near-normal condition and extreme drought are seen from years 2000–2010 to 2040–2050. The probability of near-normal condition is reduced, and that of extreme drought is increased significantly in the years 2040–2050. Probabilities for mild and severe droughts remain almost same. Variations in the probabilities of different droughts are not significant in the later years, 2040–2050 to 2090–2100. This may mean that the assumption of normal distribution does not result in a correct assessment of drought impacts of climate change years farther in the future. Figure 9 presents the normal probability plot of SPI-12 for three arbitrarily chosen years 2007, 2041, and 2093 from the three time slices. For all the cases, the SPI-12 values deviate significantly from the normal distribution. A similar observation may be expected for other years also, and thus the probability represented in Figure 8 is not accurate. To determine the PDF of SPI-12 in a year more accurately, kernel density estimation method is used to obtain the nonparametric PDF of SPI-12 for each year in the future.

4.2. Kernel Density Estimation

[31] Kernel density estimation entails a weighted moving average of the empirical frequency distribution of the data. Most nonparametric density estimators can be expressed as kernel density estimators [*Scott*, 1992; *Tarboton et al.*, 1998]. It involves the use of kernel function ($K(x)$), defined by a function having the following property:

$$\int_{-\infty}^{\infty} K(x)\mathrm{d}x = 1 \tag{16}$$

[32] A PDF can therefore be used as a kernel function. A normal kernel (i.e., a Gaussian function with mean 0 and variance 1) is used here. A kernel density estimator ($\hat{f}(x)$) of a PDF at x is defined by:

$$\hat{f}(x) = (nh)^{-1} \sum_{l=1}^{n} K((x - x_l)/h) \tag{17}$$

where n is the number of observations (here number of available GCM outputs), x_l is the lth observation (here SPI-12), and h is the smoothing parameter known as bandwidth, which is used for smoothening the shape of the estimated PDF. The selection of bandwidth is an important step in kernel estimation method. A change in bandwidth may dramatically change the shape of the kernel estimate [*Efromovich*, 1999].

[33] Bandwidth for kernel estimation may be evaluated by minimizing the deviation of the estimated PDF from the actual one. When the actual PDF is unknown, the conventional method is to assume a normal distribution. Although there are other methods like plug-in estimates [*Polansky and Baker*, 2000] and least squares cross validation [*Scott*, 1992; *Tarboton et al.*, 1998], in the present study the bandwidth is estimated based on normal distribution for computational simplicity at each of the time steps. Thus the optimal bandwidth (h_0) is given by:

$$h_0 = (1.587)\sigma n^{-\frac{1}{3}} \tag{18}$$

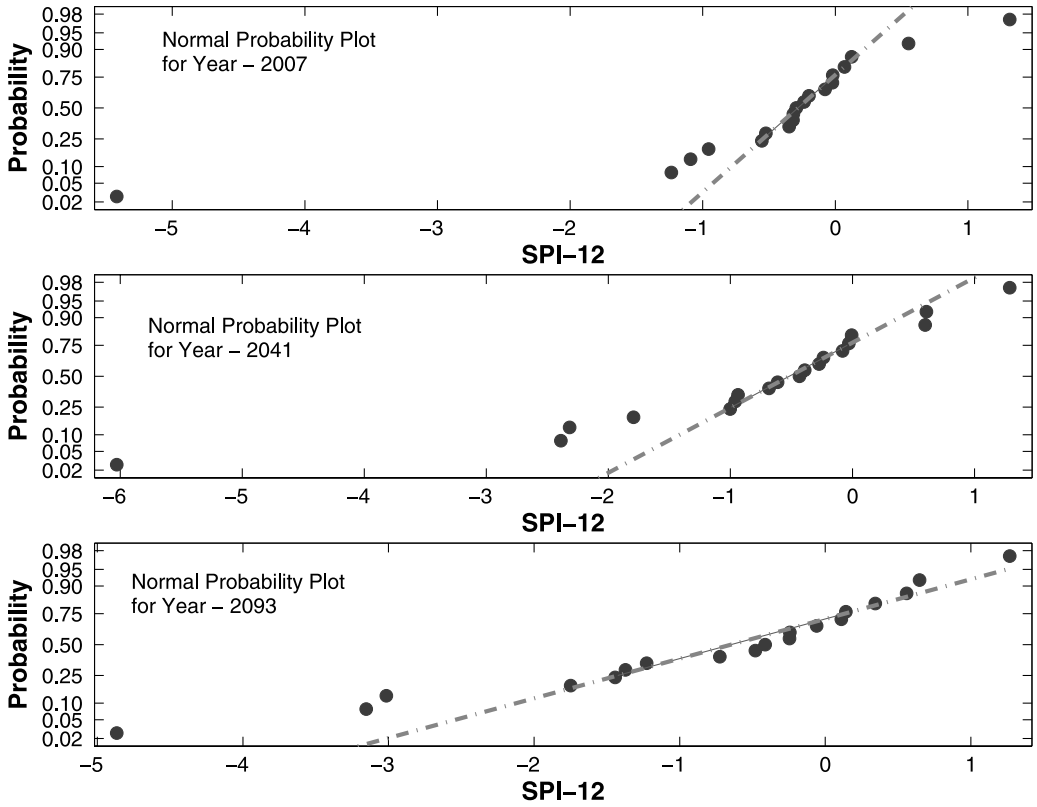

Figure 9. Normal probability plot of SPI-12 for years 2007, 2041, and 2093.

[34] For non-normal densities, σ is given by [*Silverman*, 1986]:

$$\sigma = \min\{S, IQR/1.349\} \tag{19}$$

where S is the sample standard deviation, and IQR is the interquartile range. The value of bandwidth thus evaluated is used to estimate the PDF of the data series using equation (17). This methodology is used to derive the nonparametric PDF for SPI at different time steps. By numerical integration, the CDF values at SPIs of -2, -1.5, -1.0, and 0 are estimated. These are used for finding out the probability of different classes of drought in the future.

[35] Figure 10 presents the probabilities of drought conditions in the years 2000–2010, 2040–2050, and 2090-2100, as obtained using the kernel density estimation. Although an apparent increase in the probability of extreme drought is observed from both methodologies (viz., methods based on the assumption of normal distribution and the kernel density estimation), the resulting probabilities are quite different. The difference between the probabilities of near-normal condition for the years 2000–2010 and 2040–2050 using the method based on kernel density estimation is not significant, whereas a larger change has been found for the model based on the assumption of normal distribution. A significant change is found for the years 2090–2100 from the years 2040–2050 in the probabilities of near-normal condition from kernel density estimation method, which is absent in the plots obtained from the model based on the assumption of normal distribution. The probability of extreme drought has a continuous increasing trend in Figure 10, which was absent in Figure 8. Significant changes are also observed in the probabilities of mild and severe droughts. Although the methodology of kernel estimation used in the present work is computationally simple, there are some drawbacks that include the following:

[36] 1. A large sample can give a better estimate of kernel density estimator. In the present analysis, the sample size (19) is small, consisting only of the downscaled SPI of the available GCM output.

[37] 2. The bandwidth is estimated by assuming that the actual density is normal, which may not be valid.

[38] To overcome these drawbacks, a methodology based on orthonormal series is used, which is an ideal method for estimation of nonparametric PDF from a small sample [*Efromovich*, 1999]. The next subsections present the details of the methodology for estimation of PDF using orthonormal series.

4.3. Method of Orthonormal Series

[39] A PDF from a small sample can be estimated using orthonormal series method, which is essentially a series of orthonormal functions obtained from the sample. The summation of the series with coefficients results in the desired PDF. The following subsection presents the details of the concept and methodology of the method.

4.3.1. Concept of Orthonormal Series and Density Estimation

[40] The methodology based on orthonormal series is used to estimate univariate density of data set with small sample size. Mathematical development of the methodology presented in this subsection is taken from the work of *Efromovich* [1999]. An orthonormal series is a series of

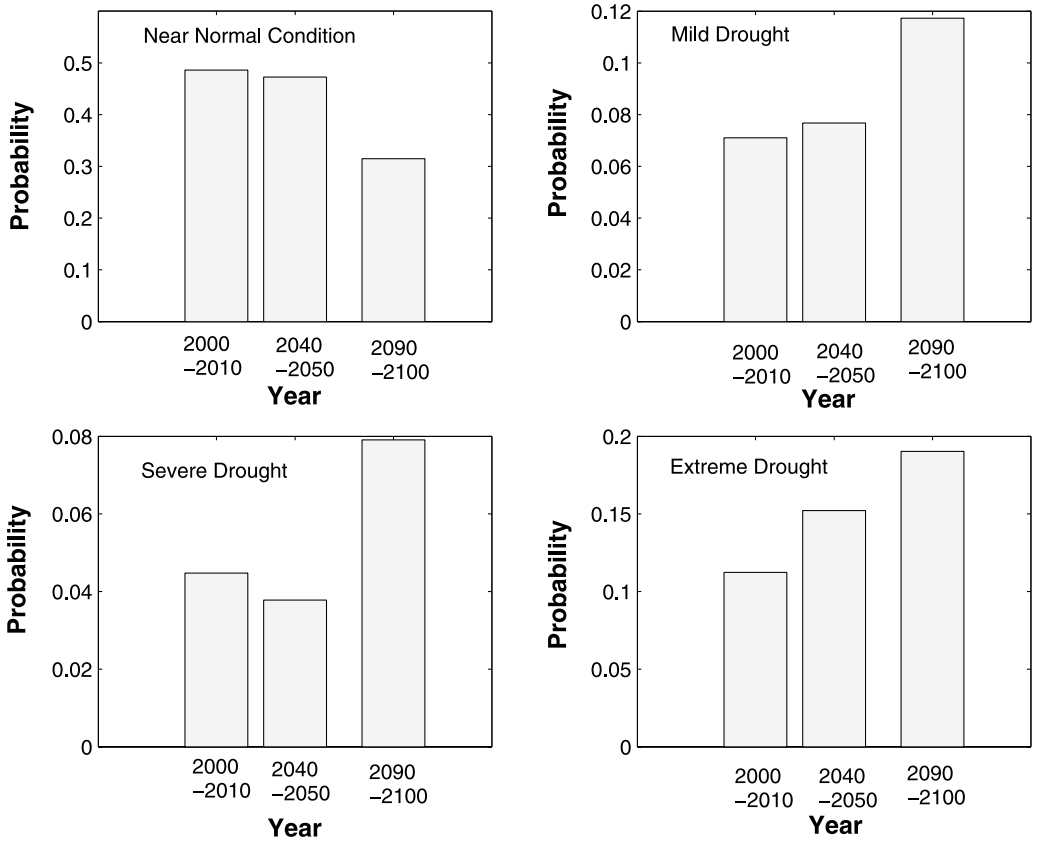

Figure 10. Probability of droughts by kernel density estimation.

orthonormal functions, $\Phi_s(x)$ and $\Phi_j(x)$, satisfying equations (20) and (21):

$$\int \Phi_s(x)\Phi_j(x)\mathrm{d}x = 0 \quad \forall s \neq j \tag{20}$$

$$\int \left(\Phi_j(x)\right)^2 \mathrm{d}x = 1 \quad \forall j \tag{21}$$

[41] Typically, a univariate density function of a random variable X may be well approximated by an orthonormal series $\tilde{f}_J(x)$:

$$\tilde{f}_J(x) = \sum_{j=0}^{J} \theta_j \Phi_j(x) \tag{22}$$

where J is called the cutoff, $\Phi_j(x)$, $j = 0, 1, \ldots$ are the functions of orthonormal system, and θ_j, $j = 0, 1, \ldots$ are the coefficients corresponding to each function. In our case, X is an SPI-12 simulated value from climate models. For this work, we select the orthonormal series as the subset of the Fourier series consisting of cosine functions:

$$\Phi_0(x) = 1 \tag{23}$$

$$\Phi_j(x) = \sqrt{2}\cos(\pi j x), \quad j = 1, 2, 3, \ldots \tag{24}$$

[42] The algorithm involved in estimating probability density function based on orthonormal series is presented in Figure 11. The detailed methodology is discussed in Appendix 1.

[43] After estimating the PDF, numerical integration is performed for evaluating the CDF values at critical points of SPI-12 equal to −2, −1.5, −1.0, and 0.0. These are used for examining the severity of future drought.

4.3.2. Application and Results

[44] The PDF of SPI-12 computed using the orthonormal series method is presented in Figure 12 along with frequency distribution of the sample and the PDFs resulting from the other two methods for three arbitrarily chosen years 2007, 2041, and 2093 selected from the three time slices of years 2000–2010, 2040–2050, and 2090–2100. For all the cases, it is clear from the figure that a normal PDF fails to model the samples of SPI-12, especially the feature of multimodality, in all the three cases. The PDF obtained using orthogonal series closely resembles the shape generated by the frequency distribution. As an example, for the year 2007, around zero value of SPI-12 the kernel density estimator overestimates the PDF, whereas the PDF generated by orthonormal series estimates it reasonably accurately. A similar result is also obtained for the year 2041. For the year 2093, the PDFs obtained from both the nonparametric methods are nearly the same. One possible reason for the difference between the PDFs obtained from kernel density estimation and orthonormal series is the improper selection of bandwidth in kernel density estimator, which may oversmooth or undersmooth the generated PDF.

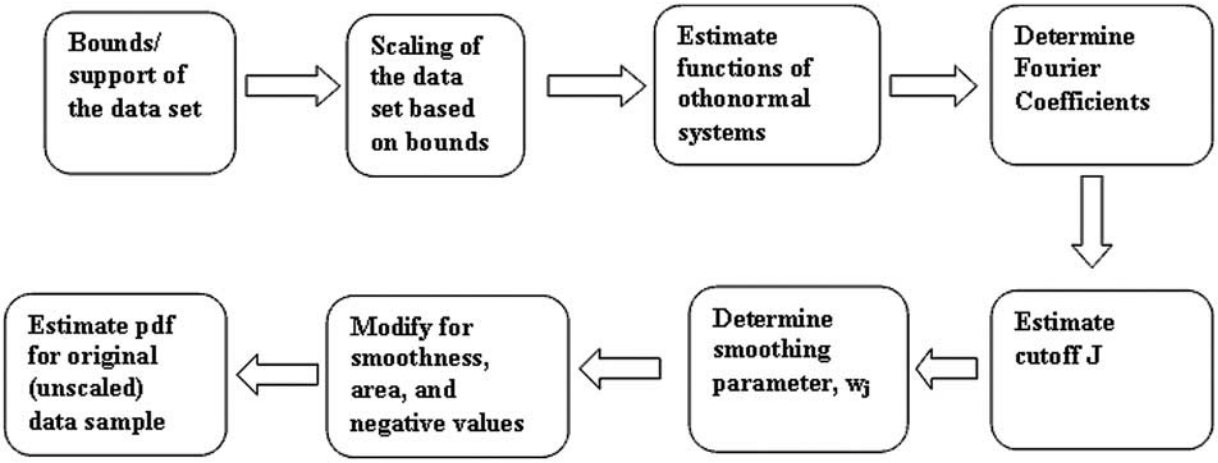

Figure 11. Algorithm for PDF estimation using orthonormal series method.

[45] Figure 13 presents the probabilities of drought conditions in the years 2000–2010, 2040–2050, and 2090–2100 as obtained using orthonormal series-based density estimation. The results are by and large similar to those of kernel density estimation except for the probabilities of mild drought. Kernel density estimation procedure projects a sudden increase in the probability of mild drought for the years 2090–2100, whereas such significant change is not observed in the results obtained from orthonormal series method. From the overall trend in probabilities of all categories of drought, it may be concluded that the probability of near-normal condition will decrease, and the probabilities of mild, severe, and extreme droughts will increase over time, which projects the Orissa meteorological subdivision to be more drought-prone in the future. From Figure 13, it is observed that significant increases in probabilities are indicated in cases of severe and extreme droughts only, which implies that climate change impact is more prominent on the extreme hydrologic events. As indicated by these results, the impact of climate change

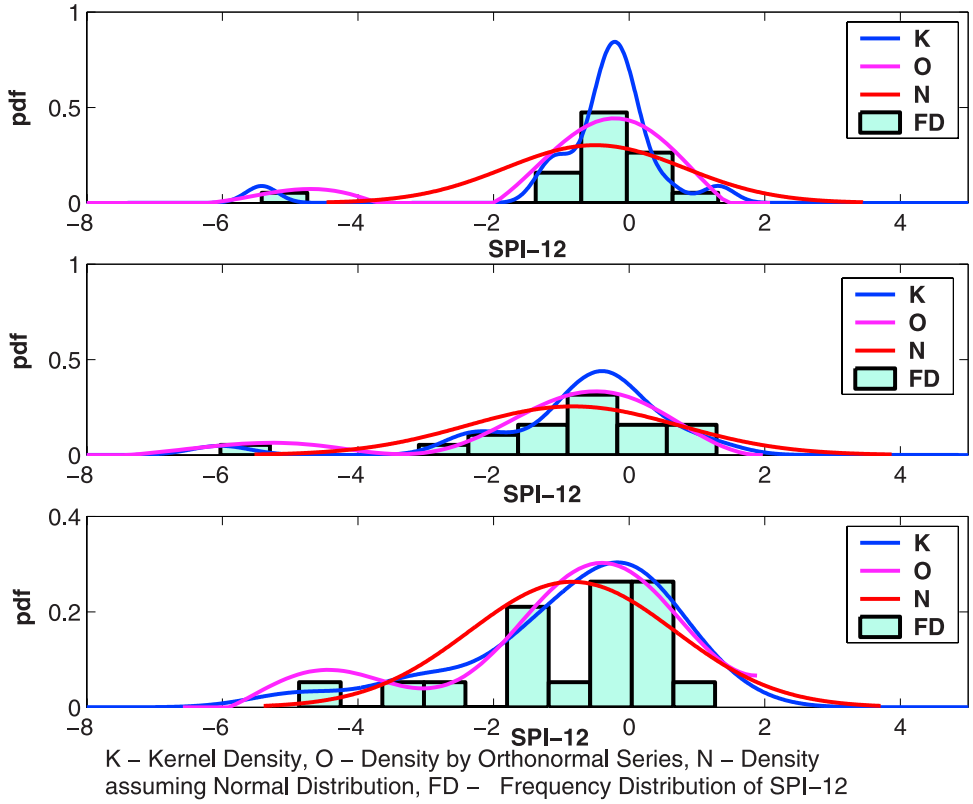

Figure 12. PDF of SPI-12 for years 2007, 2041, and 2093 (color).

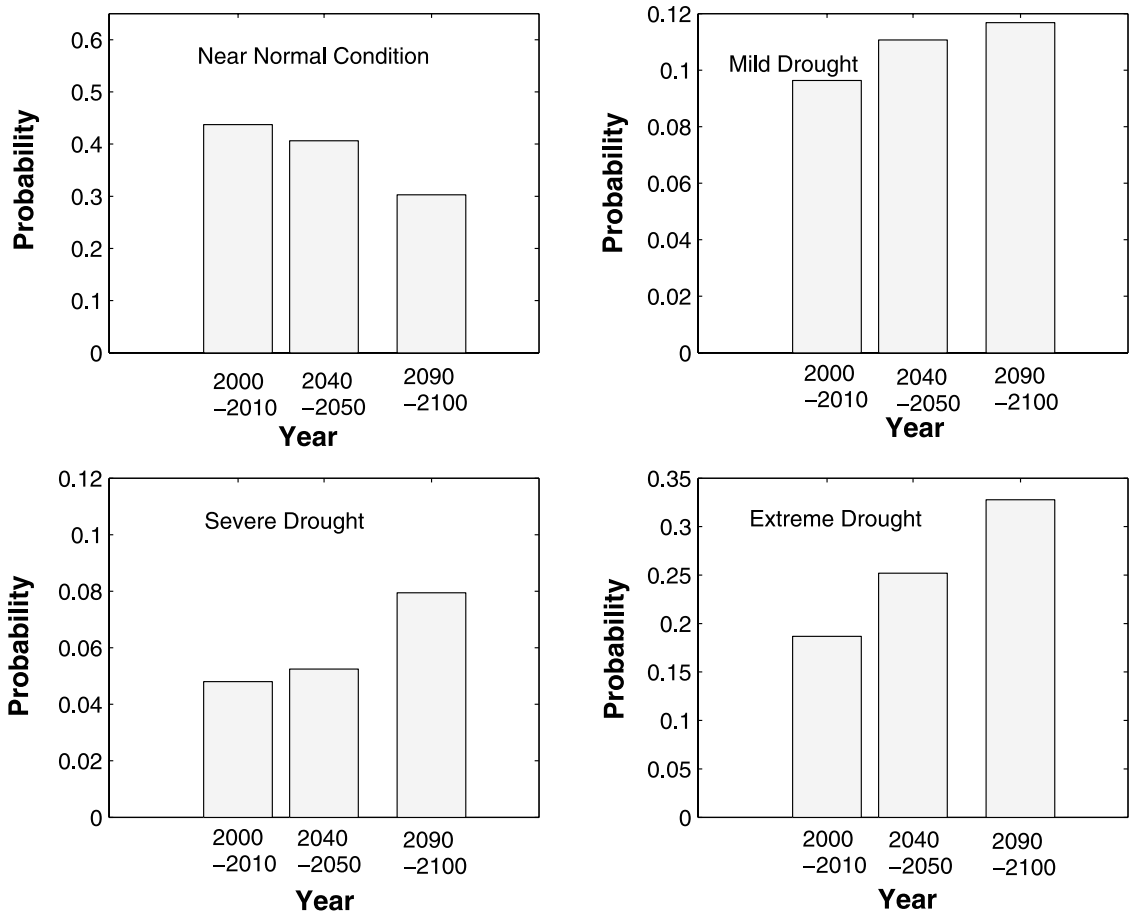

Figure 13. Probability of droughts using orthonormal series.

may also be more severe for the Orissa meteorological subdivision because of its position at the coast of the Bay of Bengal. A slight change in the pressure anomaly of the sea can have a severe impact on the precipitation of Orissa, which results in an increase of hydrologic extremes in that region. Recent past records of Orissa with a fluctuating weather condition and the high occurrence of hydrologic extremes show that this is the most affected region of India due to climate change (www.cseindia.org/programme/geg/pdf/orissa.pdf).

[46] Some recent studies on the trend of severity of drought for the different regions around the world suggest that drought-prone areas are increasing worldwide. *Andreadis and Lettenmaier* [2006] have examined drought characteristics over the conterminous United States and found an increasing trend of drought in the southwest and parts of the interior of the West US. From the monthly data set of global Palmer drought severity index (PDSI), *Dai et al.* [2004] have observed that most parts of Eurasia, Africa, Canada, Alaska, and eastern Australia became drier from 1950 to 2002 as large surface warming has occurred since 1950 over these regions. Dry area has more than doubled (from 12 to 30%) since the 1970s, with a large jump in the early 1980s due to precipitation decreases and subsequent expansion primarily due to surface warming. Without the warming, the PDSI decreases would have been much smaller and less pervasive. Surface warming due to climate change and greenhouse gas effect may lead to a severe drought condition in the present case study area Orissa. *Dai et al.* [1997] and *Dai and Wigley* [2000] have found that the decrease in precipitation is occurring mainly over El Niño Southern Oscillation (ENSO)-sensitive regions. There is an established evidence of climatic teleconnection between ENSO and Indian rainfall [*Maity and Nagesh Kumar*, 2006], and thus the impact is more severe for India, especially for the Orissa region, because of its coastal position. Therefore global warming with high surface warming in Orissa, sensitivity of precipitation to ENSO, and coastal position are the possible reasons for the trend in the probabilities obtained in this study for the different categories of drought. The probabilities obtained from the analysis will be useful in computing the expected future damage due to drought and in preparing the policy makers in generating appropriate responses.

5. General Remarks

[47] The methodology presented in this paper deals with the problem of uncertainties due to GCMs and scenarios in a climate change impact assessment study. For examining the future drought scenario of Orissa meteorological subdivision in India, time series of SPI-12 is obtained from the projections of available outputs of several GCMs with several available scenarios. To model the uncertainty in a

probabilistic framework, it is assumed that there exists a PDF of the SPI-12 in each year of simulation. The PDF is estimated using nonparametric methods to obtain the probability of different categories of drought in the future. The results show an increasing trend in the probabilities of extreme and severe droughts in the region in the future.

[48] A limitation of the methodology presented here is that it does not consider the uncertainty due to parameterization and the structure of the impact model (GCM) itself, which is increasingly recognized in recent years. The other two sources of uncertainty not considered here are those due to starting conditions used in GCM simulations and the downscaling techniques. Given the hydrologic variable of interest projected from climate model runs with different parameter values of impact models or by using different downscaling techniques, the basic concepts of the proposed methodology may be used for uncertainty modeling. It may be noted that all scenarios are not available under all GCMs or, in other words, in the IPCC data distribution center does not provide outputs for all the scenarios for all GCMs. This leads to some implicit weighting of the GCMs.

[49] Nonparametric PDF is used in the present study to estimate the probability of occurrence of different categories of droughts under future climate change scenarios. The use of parametric PDF, such as the normal distribution, does not lead to precise or accurate estimate of such probabilities. Modeling SPI assuming normal probability distribution should be considered only when the resulting imprecision is modeled. Uses of nonparametric PDF by kernel function, orthonormal system have also imprecision associated with the smoothing of kernel estimate and the determination of support of orthonormal system. Theoretically, SPI-12 can vary from $-\infty$ to $+\infty$, but in the present analysis, by using a heuristic method, its support is fixed and used for PDF estimation. This may lead to imprecision in its estimate which is not modeled in the present analysis. In such cases, simultaneous accounting of randomness and imprecision or fuzziness, in a single integrated model by using the concept of imprecise probability [*Zadeh*, 2002], is useful, where the parameters of the probability distribution are considered as interval numbers or fuzzy numbers. Modeling GCM and scenario uncertainty with imprecise probability may give more generalized estimates of the probabilities of drought conditions, which are useful in examining future conditions.

[50] Future scope of research of the present study includes the use of probabilities estimated for different categories of drought in water resource systems planning and operation. *Steinemann* [2003] has pointed out that drought indicators and triggers often lack statistical integrity, consistency among drought categories, and correspondence with desired management goals, and thus evaluation of indicators for compatibility, consistency, and applicability is a must before using them in water resource systems models. For example, for direct minimization of expected damage through reservoir operation, surface water supply index (SWSI) is more useful than SPI-12 as SWSI includes reservoir storage as input variable, but such analysis may involve a downscaling model to predict the reservoir inflow also. The use of imprecise probability in modeling GCM and scenario uncertainty along with the uncertainties due to parameterization, structure, and initial value, and the use of suitable drought indicator in decision making will be a useful direction of research on climate change impact assessment.

6. Conclusions

[51] A methodology of modeling GCM and scenario uncertainty for examining the severity of future drought is presented in this paper. The drought indicator, SPI-12, projected from GCMs is considered to have a PDF in each year in the future. The PDFs are estimated with nonparametric statistical techniques of kernel estimates and orthonormal system. Results are presented in terms of probabilities of different categories of drought in the future. The methodology does not only represent uncertainty due to different GCM projections but also incorporates it in examining the future drought scenario. Models based on software such as MAGICC (www.cgd.ucar.edu/cas/ACACIA/projects/magicc.html) require an assumption of prior probability distribution of GHG emission and concentration, which may also lead to high imprecision. The methodology presented here does not assume such prior probability. The methodology results in an increasing trend in the probability of severe and extreme droughts for Orissa meteorological subdivision with a decrease in the probability of near-normal condition. It may be concluded from the results that the region will be more drought-prone due to the effect of climate change. Sources of uncertainty other than GCMs and scenarios (for example, uncertainty due to parameterization, initial value, structure, and downscaling method) are ignored in the present study. The methodology presented here does not limit its usefulness only for drought prediction. Given a suitable index for a hydrologic event, the methodology may be used to examine the hydrologic implications of the GCM simulations for the particular event in the future.

Appendix A: Algorithm for Density Estimation Using Orthonormal Series

[73] The algorithm for density estimation using orthonormal series involves the following steps:

A1. Determination of Support and Scaling of Data Set

[74] The methodology presented for estimation of PDF using orthonormal systems is valid when the bound of random variable is [0, 1]. The random variable of interest may have different bounds (say, [a, b]) and thus may need to be converted to a variable y having an interval of [0, 1] by scaling the data. The scaled variable y may be given by:

$$y = (x - a)/(b - a) \tag{25}$$

[75] Considering the minimum and maximum values from the data set as the two bounds ([a, b]) are not a realistic method, there is no guarantee that unobserved values will not cross these bounds. Methodology for determination of support from a data set may be found in the paper of *Efromovich* [1999]. According to that methodology, if $x_1, x_2, \ldots, x_n$ are ordered observations ($x_1 \le x_2 \le \ldots \le x_n$), then

$$a = x_1 - d_1 \tag{26}$$

$$b = x_n + d_2 \tag{27}$$

where,

$$d_1 = (x_{1+s} - x_1)/s; \quad \text{and} \quad d_2 = (x_n - x_{n-s})/s \tag{28}$$

s is a small positive integer, assuming that the density is flat near the boundaries of its support. The default value of s is 1, which is considered in the present work.

A2. Estimation of Orthonormal Series With Coefficients

[76] The functions involved in the orthonormal series can be obtained using equations (23) and (24). The coefficients θ_j, $j = 0, 1, \ldots$ presented in equation (22) can be given by equation (29), where $f(y)$ is the PDF of the scaled random variable Y.

$$\theta_j \int_{-\infty}^{\infty} f(y)\Phi_j(y)\mathrm{d}y \tag{29}$$

[77] It follows that θ_j is the expected value of $\Phi_j(y)$, equation (30), which in turn may be approximated from a finite sample of n observations (y_j, $j = 1, n$) as equation (31).

$$\theta_j = E\left[\Phi_j(y)\right] \tag{30}$$

$$\theta_j = \frac{1}{n}\sum_{l=1}^{n} \Phi_j(y_l) \tag{31}$$

[78] In our case, where Y is the scaled SPI-12 with a bound [0, 1], the y_j are the scaled values of SPI-12 simulated from each climate model.

A3. Estimation of Cutoff J

[79] Determination of an appropriate cutoff J [equation (22)] is important in the method based on orthonormal series. The choice of J depends on the goodness of fit, which can be determined by mean integrated square error using Parseval's inequality. Following *Efromovich* [1999] J can be computed as:

$$J = \underset{0 \le \tilde{J} \le J_n}{\operatorname{argmin}} \sum_{j=0}^{\tilde{J}} \left(2d_j n^{-1} - \theta_j^2\right) \tag{32}$$

where $J_n = [C_{J0} + C_{J1}\ln(n)]$. The default values of C_{J0} and C_{J1} are 4 and 0.5, respectively [*Efromovich*, 1999].

A4. Smoothing of Estimated PDF

[80] In many cases, it is worthwhile to smooth the Fourier coefficients by multiplying them with some constants that take values between 0 and 1. After smoothing, a modified PDF is given by:

$$\tilde{f}_J(y) = \sum_{j=0}^{J} w_j\theta_j\Phi_j(y),\ 0 \le y \le 1,\ 0 \le w_j \le 1 \tag{33}$$

[81] The weights (w_j) used for smoothing Fourier coefficients may be given by:

$$w_0 = 1; \tag{34}$$

$$w_j = \left(1 - /\left(n\theta_j^2\right)\right)_+ \quad \forall_j > 0 \tag{35}$$

[82] Here $(1 - d/(n\theta_j^2))_+ = \max(0,(1 - d/(n\theta_j^2)))$, i.e., the positive part of $(1 - d/(n\theta_j^2))$. Other than the first J number of Fourier coefficients, a density function also requires a relatively large number of coefficients for a fair visualization. Thus high-frequency terms are added, which are shrunk by a hard threshold procedure. After adding these extra terms, equation (33) is modified to:

$$\tilde{f}_J(y) = \sum_{j=0}^{J} w_j\theta_j\Phi_j(y) + \sum_{j=J+1}^{C_{JM}\times J_n} I_{\left\{\theta_j^2 > C_T d\ \ln(n)/n\right\}}\theta_j\Phi_j(y) \tag{36}$$

where, C_{JM} and C_T are parameters for hard threshold procedure that define the maximal number of elements included in the estimate of PDF. The default values are 6 and 4, respectively [*Efromovich*, 1999]. I is an indicator variable such that $I_{\{A\}}$ has the value of 1, if A is true, but 0 otherwise. A high-frequency term is included only if the corresponding Fourier coefficient is extremely large, and thus the procedure does not reduce the smoothness of the estimate.

A5. Modification for Area Under PDF and Negative Values

[83] An improvement in the PDF thus estimated is necessary when it takes negative values at some of the points/regions. For such cases, the following algorithm is developed in the present study, which ensures that the properties of PDF are satisfied by the estimated PDF.

[84] 1. If there exist negative values at some points, find the maximum negative value.

[85] 2. Add this to $\tilde{f}_J(y)$ to make the value of the function positive everywhere.

[86] 3. Check the area under the curve. If it is less than 1, find out c by numerical methods such that

$$\int_{-\infty}^{\infty} \left(\tilde{f}_J(y) + c\right)\mathrm{d}y = 1 \tag{37}$$

$\hat{f}_J(y)$ is now modified by adding c, as obtained from equation (39), to it.

$$\tilde{f}_J(y) = \tilde{f}_J(y) + c \tag{38}$$

[87] 4. If the area under the curve is greater than 1, find out $c1$ in a similar procedure:

$$\int_{-\infty}^{\infty} \left(\tilde{f}_J(y) - c1\right)\mathrm{d}y = 1 \tag{39}$$

[88] Subtracting $c1$ from $\tilde{f}_J(y)$ may lead to a negative value of PDF at some points, which is not desirable. In such cases after, the subtraction takes only the positive part of $\tilde{f}_J(y)$.

$$\tilde{f}_J(y) = \left(\tilde{f}_J(y) - c1\right)_+ \tag{40}$$

where $(\tilde{f}_J(y) - c1)_+ = \max(0, (\tilde{f}_J(y) - c1))$. Check the area again, and if it is not nearly equal to 1, go to step 3, else, stop.

[89] The other way of making adjustment in the estimated PDF to make the area under the curve equal to 1 is the use

of multiplicative factor. In either case, the result will come almost similar, as this adjustment methodology is not much sensitive to the final PDF.

A6. Estimation of PDF for Unscaled Data Set

[90] The scaled data/observations are distributed according to a density $f_Y(y)$, where $y \in [0, 1]$.

[91] The PDF thus obtained corresponds to the scaled data set y over the interval of [0, 1]. The estimate of $f_X(x)$ of original data set x over interval $[a, b]$ may be given by:

$$f_X(x) = (b-a)^{-1} f_Y(y); \quad x \in [a, b]. \tag{41}$$

Notation

- $\Phi(x)$ function in orthonormal series
- β_i coefficient of membership value in cluster i in the regression equation
- γ_k coefficient of kth principal component in regression equation
- ρ_{ik} coefficient of the product of membership value in cluster i and kth principal component in regression equation
- μ_{it} membership in cluster i of the principal components in month t
- $\theta_j(x)$ Fourier coefficient of the functions in orthonormal series
- C constant term in regression equation
- F_{SPI} CDF of SPI-12
- $G(x)$ CDF of the precipitation without including zero values
- $H(x)$ CDF of precipitation including zero values
- J cutoff in orthonormal series
- $K(x)$ kernel density function
- P_t precipitation in month t
- $[a, b]$ support of random variable
- c number of clusters
- $f(x)$ probability density function
- h_0 optimal bandwidth for kernel density estimation
- pc_{kt} kth principal reference component in month t
- q zero precipitation probability
- w weights for smoothing in orthonormal series estimation

[92] **Acknowledgments.** The authors sincerely thank the three anonymous reviewers and associate editor Dennis P. Lettenmaier, University of Washington, USA for reviewing the manuscript and providing critical comments to improve the paper. The work presented in this paper was financially supported by the Department of Science and Technology (Earth Systems and Science Division) and the Ministry of Water Resources (Indian National Committee on Hydrology), Government of India.

References

Andreadis, K. M., and D. P. Lettenmaier (2006), Trends in 20th century drought over the continental United States, *Geophys. Res. Lett.*, *33*(10), L10403, doi:10.1029/2006GL025711.

ASCE Task Committee on Definition of Criteria for Evaluation of Watershed Models of the Watershed Management Committee, Irrigation and Drainage Division (1993), Criteria for evaluation of watershed models, *J. Irrig. Drain. Eng.*, *119*(3), 429–442.

Bardossy, A. (1997), Downscaling from GCM to local climate through stochastic linkages, *J. Environ. Manag.*, *49*, 7–17.

Bardossy, A., and E. J. Plate (1991), Modeling daily rainfall using a semi-Markov representation of circulation pattern occurrence, *J. Hydrol.*, *122*, 33–47.

Bardossy, A., and E. J. Plate (1992), Space-time model for daily rainfall using atmospheric circulation patterns, *Water Resour. Res.*, *28*(5), 1247–1259.

Bardossy, A., L. Duckstein, and I. Bogardi (1995), Fuzzy rule-based classification of atmospheric circulation patterns, *Int. J. Climatol.*, *15*, 1087–1097.

Bardossy, A., I. Bogardi, and I. Matyasovszky (2005), Fuzzy rule-based downscaling of precipitation, *Theor. Appl. Climatol.*, *82*, 119–129.

Benestad, R. E. (2004), Tentative probabilistic temperature scenarios for northern Europe, *Tellus Ser. A Dyn. Meteorol. Oceanogr.*, *56*, 89–101.

Bezdek, J. (1981), *Pattern Recognition with Fuzzy Objective Function Algorithm*, Springer, New York.

Cannon, A. J., and P. H. Whitfield (2002), Downscaling recent stream flow conditions in British Columbia, Canada using ensemble neural network models, *J. Hydrol.*, *259*, 136–151.

Cencov, N. N. (1962), Evaluation of an unknown distribution density from observations, *Sov. Med. Dokl.*, *3*, 1559–1562.

Cunderlik, J. M., and S. P. Simonovic (2004), Inverse modeling of water resources risk and vulnerability to changing climatic conditions, in *Proceedings of the 57th Canadian Water Resources Association Annual Congress. Water and Climate Change: Knowledge for Better Adaptation*, CWRA, Montreal.

Dai, A. G., and T. M. L. Wigley (2000), Global patterns of ENSO induced precipitation, *Geophys. Res. Lett.*, *27*, 1283–1286.

Dai, A. G., I. Y. Fung, and A. D. Del Genio (1997), Surface observed global land precipitation variations during 1900–88, *J. Clim.*, *10*, 2943–2962.

Dai, A., K. E. Trenberth, and T. Qian (2004), A global data set of Palmer Drought Severity Index for 1870–2002: Relationship with soil moisture and effects of surface warming, *J. Hydrometeorol.*, *5*(6), 1117–1130.

Dessai, S., X. Lu, and M. Hulme (2005), Limited sensitivity analysis of regional climate change probabilities for the 21st century, *J. Geophys. Res.*, *110*, D19108, doi:10.1029/2005JD005919.

Easterling, D. R. (1999), Development of regional climate scenarios using a downscaling approach, *Clim. Change*, *41*, 615–634.

Efromovich, S. (1999), *Nonparametric Curve Estimation: Methods, Theory and Applications*, Springer, New York.

Ghosh, S., and P. P. Mujumdar (2006), Future rainfall scenario over Orissa with GCM projections by statistical downscaling, *Curr. Sci.*, *90*(3), 396–404.

Giorgi, F., and L. O. Mearns (2003), Probability of regional climate change calculated using the reliability ensemble averaging (REA) method, *Geophys. Res. Lett.*, *30*(12), 1629, doi:10.1029/2003GL017130.

Gujrati, D. N. (2004), *Basic Econometrics*, McGraw-Hill, New York.

Güler, C., and G. D. Thyne (2004), Delineation of hydrochemical facies distribution in a regional groundwater system by means of fuzzy c-means clustering, *Water Resour. Res.*, *40*, W12503, doi:10.1029/2004WR003299.

Gutierrez, J. M., A. S. Cofino, R. Cano, and M. A. Rodriguez (2004), Clustering methods for statistical downscaling in short-range weather forecasts, *Mon. Weather Rev.*, *132*, 2169–2183.

Haylock, M. R., G. C. Cawley, C. Harpham, R. L. Wilby, and C. M. Goodess (2006), Downscaling heavy precipitation over the UK: A comparison of dynamical and statistical methods and their future scenarios, *Int. J. Climatol.*, *26*, 1397–1415.

Hughes, J. P., and P. Guttorp (1994), A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena, *Water Resour. Res.*, *30*(5), 1535–1546.

Hughes, J. P., D. P. Lettenmaier, and P. Guttorp (1993), A stochastic approach for assessing the effect of changes in synoptic circulation patterns on gauge precipitation, *Water Resour. Res.*, *29*(10), 3303–3315.

Hulme, M., and T. C. Carter (1999), Representing uncertainty in climate change scenarios and impact studies, *Proceedings of the ECLAT-2 Helsinki Workshop*, edited by T. Carter, M. Hulme, and D. Viner, Climatic Research Unit, Norwich.

IPCC (2000), *Emission Scenarios: A Special Report of IPCC Working Group III*, Cambridge Univ. Press, New York.

IPCC-TGCIA (1999), Guidelines on the use of scenario data for climate impact and adaptation assessment, Version 1, 69. (Available at http://ipcc-ddc.cru.uea.ac.uk/)

Jones, R. N. (2000), Analyzing the risk of climate change using an irrigation demand model, *Clim. Res.*, *14*, 89–100.

Jones, P. D., J. M. Murphy, and M. Noguer (1995), Simulation of climate change over Europe using a nested regional-climate model. 1: Assessment of control climate, including sensitivity to location of lateral boundaries, *Q. J. R. Meteorol. Soc.*, *121*, 1413–1449.

Kalnay, E., et al. (1996), The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*(3), 437–471.

Lall, U. (1995), Recent advances in nonparametric function estimation, *U.S. Natl. Rep. Int. Union Geod. Geophys. 1991–1994, Rev. Geophys., 33*, supplement, 1093–1102.

Lall, U., Y.-I. Moon, and K. Bosworth (1993), Kernel flood frequency estimators: Bandwidth selection and kernel choice, *Water Resour. Res.*, *29*(4), 1003–1016.

Lall, U., B. Rajagopalan, and D. G. Tarboton (1996), A nonparametric wet/dry spell model for resampling daily precipitation, *Water Resour. Res.*, *32*(9), 2803–2823.

Leavesley, G. H. (1994), Modeling the effects of climate change on water resources—A review, *Clim. Change*, *28*, 159–177.

Maity, R., and D. Nagesh Kumar (2006), Bayesian dynamic modeling for monthly Indian summer monsoon rainfall using El Niño-Southern Oscillation (ENSO) and Equatorial Indian Ocean Oscillation (EQUINOO), *J. Geophys. Res.*, *111*, D07104, doi:10.1029/2005JD006539.

McKee, T. B., N. J. Doesken, and J. Kleist (1993), The relationship of drought frequency and duration to time scale, in *Proceedings of the Eighth Conference on Applied Climatology, American Meteorological Society*, 179–184.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth (2004), Quantification of modelling uncertainties in a large ensemble of climate change simulations, *Nature*, *430*, 768–772.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models. Part 1: A discussion of principles, *J. Hydrol.*, *10*, 282–290.

New, M., and M. Hulme (2000), Representing uncertainty in climate change scenarios: A Monte Carlo approach, *Integrated Assessment*, *1*, 203–213.

Polansky, A. M., and E. R. Baker (2000), Multistage plug-in bandwidth selection for kernel distribution function estimates, *J. Stat. Comput. Simul.*, *65*, 63–80.

Prodanovic, P., J. Cunderlik, and S. P. Simonovic (2005), Synthetic storm model for climate change impact modelling, in *Proceedings of the 17th Canadian Hydrotechnical Conference, CSCE, Edmonton*, 8.

Prudhomme, C., N. Reynard, and S. Crooks (2002), Downscaling of global climate models for flood frequency analysis: Where are we now?, *Hydrol. Process.*, *16*, 1137–1150.

Prudhomme, C., D. Jakob, and C. Svensson (2003), Uncertainty and climate change impact on the flood regime of small UK catchments, *J. Hydrol.*, *277*, 1–23.

Raisanen, J., and T. N. Palmer (2001), A probability and decision-model analysis of a multimodel ensemble of climate change simulations, *J. Clim.*, *14*, 3212–3226.

Ross, T. J. (1997), *Fuzzy Logic With Engineering Applications*, 379–396, McGraw-Hill, New York.

Roubens, M. (1982), Fuzzy clustering algorithms and their cluster validity, *Eur. J. Oper. Res.*, *10*, 294–301.

Scott, D. W. (1992), *Multivariate Density Estimation, Theory, Practice, And Visualization*, John Wiley, Hoboken, N. J.

Sharma, A., D. G. Tarboton, and U. Lall (1997), Streamflow simulation: A nonparametric approach, *Water Resour. Res.*, *33*(2), 291–308.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, 1st ed., CRC Press, Boca Raton, Fla.

Simonovic, S. P., and L. Li (2003), Methodology for assessment of climate change impacts on large-scale flood protection system, *J. Water Resour. Plan. Manage.*, *129*(5), 361–371.

Simonovic, S. P., and L. Li (2004), Sensitivity of the red river basin flood protection system to climate variability and change, *Water Resour. Manag.*, *18*, 89–110.

Stehlik, J., and A. Bardossy (2002), Multivariate stochastic downscaling model for generating daily precipitation series based on atmospheric circulation, *J. Hydrol.*, *28*(5), 1247–1259.

Steinemann, A. (2003), Drought indicators and triggers: A stochastic approach to evaluation, *J. Am. Water Resour. Assoc.*, *39*(5), 1217–1233.

Tarboton, D. G., A. Sharma, and U. Lall (1998), Disaggregation procedures for stochastic hydrology based on nonparametric density estimation, *Water Resour. Res.*, *34*(1), 107–119.

Tebaldi, C., L. O. Mearns, D. Nychka, and R. L. Smith (2004), Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations, *Geophys. Res. Lett.*, *31*, L24213, doi:10.1029/2004GL021276.

Tebaldi, C., R. Smith, D. Nychka, and L. O. Mearns (2005), Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multi-model ensembles, *J. Clim.*, *18*, 1524–1540.

Tripathi, S., and V. V. Srinivas (2005), Downscaling of General Circulation Models to assess the impact of climate change on rainfall of India, in *Proceedings of International Conference on Hydrological Perspectives for Sustainable Development (HYPESD -2005)*, 23–25 February, IIT Roorkee, India, 509–517.

Tripathi, S., V. V. Srinivas, and R. S. Nanjundiah (2006), Downscaling of precipitation for climate change scenarios: A support vector machine approach, *J. Hydrol.*, *330*, 621–640.

Wetterhall, F., S. Halldin, and C. Xu (2005), Statistical precipitation downscaling in central Sweden with the analogue method, *J. Hydrol.*, *306*, 174–190.

Wilby, R. L., and C. W. Dawson (2004), *Using SDSM Version 3.1—A decision support tool for the assessment of regional climate change impacts. User manual*, 67 pp.

Wilby, R. L., and I. Harris (2006), A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the River Thames, UK, *Water Resour. Res.*, *42*, W02419, doi:10.1029/2005WR004065.

Wilby, R. L., H. Hassan, and K. Hanaki (1998), Statistical downscaling of hydrometeorological variables using general circulation model output, *J. Hydrol.*, *205*, 1–19.

Wilby, R. L., L. E. Hay, and G. H. Leavesly (1999), A comparison of downscaled and raw GCM output: Implications for climate change scenarios in the San Juan River Basin, Colorado, *J. Hydrol.*, *225*, 67–91.

Wilby, R. L., L. E. Hay, W. J. Gutowski, R. W. Arritt, E. S. Takle, Z. T. Pan, G. H. Leavesley, and M. P. Clark (2000), Hydrological responses to dynamically and statistically downscaled climate model output, *Geophys. Res. Lett.*, *27*(8), 1199–1202.

Wilby, R. L., S. P. Charles, E. Zorita, B. Timbal, P. Whetton, and L. O. Mearns (2004), The guidelines for use of climate scenarios developed from statistical downscaling methods, *Supporting material of the Intergovernmental Panel on Climate Change (IPCC), prepared on behalf of Task Group on Data and Scenario Support for Impacts and Climate Analysis (TGICA)* (Available at http://ipccddc.cru.uea.ac.uk/guidelines/StatDownGuide.pdf).

Wilks, D. S. (1999), Multisite downscaling of daily precipitation with a stochastic weather generator, *Clim. Res.*, *11*, 125–136.

Willmott, C. J., C. M. Rowe, and W. D. Philpot (1985), Small-scale climate map: A sensitivity analysis of some common assumptions associated with the grid point interpolation and contouring, *Am. Cartogr.*, *12*, 5–16.

S. Ghosh and P. P. Mujumdar, Department of Civil Engineering, Indian Institute of Science, Bangalore - 560012, India. (subimal@civil.iisc.ernet.in; pradeep@civil.iisc.ernet.in)