SVM-PGSL coupled approach for statistical downscaling to predict rainfall from GCM output

Subimal Ghosh[1]

Received 13 November 2009; revised 10 July 2010; accepted 28 July 2010; published 16 November 2010.

[1] Hydrological impacts of climate change are assessed by downscaling the General Circulation Model (GCM) outputs of predictor variables to local or regional scale hydrologic variables (predictand). Support Vector Machine (SVM) is a machine learning technique which is capable of capturing highly nonlinear relationship between predictor and predictand and thus performs better than conventional linear regression in transfer function-based downscaling modeling. SVM has certain parameters the values of which need to be fixed appropriately for controlling undertraining and overtraining. In this study, an optimization model is proposed to estimate the values of these parameters. As the optimization model, for selection of parameters, contains SVM as one of its constraints, analytical solution techniques are difficult to use in solving it. Probabilistic Global Search Algorithm (PGSL), a probabilistic search technique, is used to compute the optimum parameters of SVM. With these optimum parameters, training of SVM is performed for statistical downscaling. The obtained relationship between large-scale atmospheric variables and local-scale hydrologic variables (e.g., rainfall) is used to compute the hydrologic scenarios for multiple GCMs. The uncertainty resulting from the use of multiple GCMs is further modeled with a modified reliability ensemble averaging method. The proposed methodology is demonstrated with the prediction of monsoon rainfall of Assam and Meghalaya meteorological subdivision of northeastern India. The results obtained from the proposed model are compared with earlier developed SVM-based downscaling models, and improved performance is observed.

Citation: Ghosh, S. (2010), SVM-PGSL coupled approach for statistical downscaling to predict rainfall from GCM output, *J. Geophys. Res.*, *115*, D22102, doi:10.1029/2009JD013548.

1. Introduction

[2] General Circulation Models (GCMs) are tools designed to simulate time series of climate variables globally, accounting for effects of greenhouse gases in the atmosphere. They attempt to represent the physical processes in the atmosphere, ocean, cryosphere, and land surface and provide estimates of climate variables (e.g., air temperature, precipitation, wind speed, pressure) on a global scale. GCMs demonstrate a significant skill at the continental and hemispheric spatial scales and incorporate a large proportion of the complexity of the global system; they are, however, inherently unable to represent local subgrid-scale features and dynamics, which is of interest to a hydrologist. Accuracy of GCMs, in general, decreases from climate-related variables such as wind, temperature, humidity, and air pressure to hydrologic variables such as precipitation, evapotranspiration, runoff, and soil moisture, which are also simulated by GCMs. These limitations of the GCMs restrict the direct use of their output in hydrology [*Hughes et al.*, 1993].

[3] Downscaling, in the context of hydrology, is a method to project the hydrologic variables (e.g., rainfall and streamflow) at a smaller scale based on large-scale climatological variables (e.g., mean sea level pressure) simulated by a GCM. Poor performances of GCMs at local and regional scales have led to the development of Limited Area Models (LAMs) in which a fine computational grid over a limited domain is nested within the coarse grid of a GCM [*Jones et al.*, 1995]. This procedure is also known as dynamic downscaling. Another approach to downscaling is statistical downscaling [*Wilby et al.*, 2004], in which regional or local information about a hydrologic variable is derived by first determining a statistical model which relates large-scale climate variables (or predictors) to regional- or local-scale hydrologic variables (or predictands). Then the large-scale output of a GCM simulation is fed into this statistical model to estimate the corresponding local or regional hydrologic characteristics [*Wilby et al.*, 2004]. Statistical downscaling methods are data- driven models and do not consider the physics between predictors and predictand. Statistical downscaling methods can be further classified into weather generators [*Hughes et al.*, 1993; *Wilks*, 1999], weather typing, and transfer functions based on the use of different statistical tools. Downscaling with the prediction of finer gridded weather/meteorologic variables from large-scale coarse-gridded climate variables is

[1] Assistant Professor, Department of Civil Engineering, Indian Institute of Technology Bombay, Mumbai, India.

0148-0227/10/2009JD013548

also used for weather or hydrometeorological applications. Details of such applications may be found in *Venugopal et al.* [1999], *Xue et al.* [2007], and *Tao and Barros* [2010].

[4] The transfer function-based downscaling method relies on a direct quantitative relationship between the local-scale climate variable (predictand) and the variables containing the large-scale climate information (predictors). To date, linear and nonlinear regression [*Wilby and Dawson*, 2004], canonical correlation [*Conway et al.*, 1996], and Artificial Neural Network (ANN) [*Hewitson and Crane*, 1992, 1996; *Crane and Hewitson*, 1998; *Trigo and Palutikof*, 1999; *Tripathi and Srinivas*, 2005] have been used to derive the predictor-preditand relationship.

[5] Despite a number of advantages, the neural network models have several drawbacks, including the possibility of getting trapped in local minima and subjectivity in the choice of model architecture [*Suykens*, 2001]. However, with technical advancements, the limitations may be overcome. *Vapnik* [1995, 1998] pioneered the development of a novel machine learning algorithm called Support Vector Machine (SVM), which provides an elegant solution to these problems. Although recurrent ANNs perform better than feed forward neural networks in many applications [e.g., *Nagesh Kumar et al.*, 2004], being a subset of neural networks, they involve numerical algorithms (back propagation or conjugate gradient) in training which sometimes do not result in global optimum values of the parameters. On the other hand, as SVM involves analytical methods such as quadratic programming, it always results in global optima [*Vapnik*, 1998]. ANN trains a model with the objective of empirical risk minimization which lacks in generalization of input-output relationship [*Gunn et al.*, 1997]. SVM, on the other hand, performs structural risk minimization which is more generalized and results in more credible solutions. The SVM has been used in a statistical downscaling model by *Tripathi et al.* [2006]. SVM has some drawbacks of rapid increase of basis functions with the size of training data set [*Govindaraju*, 2005], which may lead to overtraining (large difference between the system performance measure of training and testing data set). High overtraining suggests that a model is good for training data set but may not perform well with a new data set for present conditions (in a slightly different period) as well as for the future. This can be overcome by selecting the appropriate values of SVM parameters. In the present study, an optimization model is developed for selection of SVM parameter values where the objective is to maximize the correlation coefficient (r) (considered as a system performance measure for this study) between observed and predicted data for testing data set, which is independent of training data set. To control the difference between the r values of training and testing (high difference signifies overtraining) data set, a constraint is used in the optimization model for allowable difference. As the optimization model is nonlinear and it is difficult to represent the objective function and constraints in a functionable form of decision variables, a search algorithm is required to solve the optimization model. In the present study, Probabilistic Global Search Laussane (PGSL), a global search optimization model, is used for this purpose. Tests on benchmark problems having multiparameter nonlinear objective functions have revealed that PGSL performs better than Genetic Algorithm and advanced algorithms for Simulated Annealing [*Raphael and Smith*, 2003]. The algorithm is based on the assumption that better sets of points are more likely to be found in the neighborhood of good sets of points and therefore intensifying the search in the regions that contain good solutions. After the computation of SVM parameters and subsequent training, the model is applied to the bias corrected standardized GCM output for prediction of rainfall in the next century.

[6] To summarize the motivations and contributions of the present analysis with respect to the earlier developed models: the state of art methodology for downscaling with SVM includes a grid search method [*Tripathi et al.*, 2006] which selects the best SVM from a number of trained model on the basis of maximum performance for testing data set. However, it is observed in the present study, that such selection is also characterized by overtraining, as there is no control over the difference between the performances of SVM for training and testing. The present study develops an optimization model for selection of best SVM, which not only uses the criteria of highest performance in testing but also constrains the difference between training and testing performance to an allowable value. This results in minimum overtraining and is found to be significantly improved over the state of art modeling of grid search method. This is a generalized methodology which can be applied to any SVM model and, more specifically, is useful for downscaling as overtraining is an important factor in downscaling. The major assumption in downscaling is that the statistical relationship between the predictor and predictand will hold good in the future. High overtraining denotes that the relationship may fail in a changed condition for future, and for this specific reason the developed methodology will be very useful for statistical downscaling model applied to any case study.

[7] Modeling impacts of climate change is characterized by the uncertainty resulting from the use of multiple GCMs [*Wilby and Harris*, 2006; *Ghosh and Mujumdar*, 2007; *Mujumdar and Ghosh*, 2008]. During the past decade, research on modeling uncertainty in assessment of climate change impact has advanced on several fronts; some of them are *Raisanen and Palmer* [2001]; *Giorgi and Mearns* [2003]; *Tebaldi et al.* [2004, 2005]; *Wilby and Harris* [2006]; *Ghosh and Mujumdar* [2007]; and *Mujumdar and Ghosh* [2008]. The present study assigns weights to GCMs based on "model performance" and "model convergence" with a modified version of "Reliability Ensemble Averaging (REA)" [*Giorgi and Mearns*, 2003]. With the weights derived from modified REA, weighted mean CDFs are computed for Assam and Meghalaya meteorological subdivision, India, during three 30-year time slices: 2020s, 2050s, and 2080s. It should be noted that downscaling literally means conversion of coarse grid data to a finer grid data. However, in hydroclimatological context, downscaling is used normally for two reasons: (1) inability of GCMs in simulating rainfall accurately [*Hughes and Guttrop*, 1994] and (2) simulations of GCMs at coarse grid. Therefore, in the present context, downscaling refers to the prediction of local-scale hydrologic variable (rainfall) from larger-scale climatic pattern, which is simulated by GCM. It should be noted that the meteorological subdivisions in India have irregular shapes (as it is based on political boundaries) and rainfall amount (which is computed from rainfall at stations/points) in those subdivisions has huge implications on water sharing, planning, and management considering irregular borders. Regular grid points, on which

GCMs work, cannot be used as a representation of such subdivisions with irregular boundaries. Therefore, there is a need to predict rainfall in a subdivision from large-scale climatic pattern, which is performed in the present study. Applications of statistical downscaling for projection of rainfall in Indian meteorological subdivisions may also be found in *Tripathi et al.* [2006].

[8] It should be noted that until now, several statistical methods have been used for downscaling in projecting regional hydrologic variables; however, each method has its own limitation. It is now the high time to use the combination of models to overcome such limitations. For example, SVM is used with conventional grid search method, but does not overcome the problem of overtraining completely. The present study couples PGSL with SVM, with the development of a new optimization model for minimizing overtraining to the extent possible. Uncertainty resulting from the use of multiple GCMs is then combined for more reliable projections. None of the individual tools for downscaling are novel in terms of the methodology; the overall combination becomes new, at least in the context of the impacts and adaptation. The next section presents the details of the data and case study area.

2. Data and Case Study Area

[9] The Assam and Meghalaya meteorological subdivision, located in northeast India, extends from 90°E to 96°E and 24°N to 28°N. The monthly area weighted precipitation data of Assam and Meghalaya meteorological subdivision in India, for monsoon period (June, July, August, and September) from 1948 to 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 predictors used for downscaling [*Wilby et al.*, 1999;*Wetterhall et al.*, 2005] 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 (rainfall in the present case). Monsoon rainfall in northeast India is caused by high temperature in the land area and subsequent generation of low-pressure zone. This results in wind flow with moisture from the Bay of Bengal to the land area. This is considered in selection of predictors for the downscaling model. It has been reported in the literature [*Wilby et al.*, 1999; *Wetterhall et al.*, 2005; *Mujumdar and Ghosh*, 2008] that these variables can be simulated well at a larger scale by a GCM and may be used for downscaling. Considering this, the predictors preliminarily selected for the present study are Mean Sea Level Pressure (MSLP), surface specific humidity, near surface air temperature, zonal wind speed, and meridional wind speed. Overview of the statistical downscaling model is presented in Figure 1. Training (calibration) of the statistical downscaling model requires observed climate data. In the absence of adequate observed climatological data, the data from the National Center for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis project [*Kalnay et al.*, 1996] may be used. 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. NCEP/NCAR reanalysis data resemble the observed data, and therefore a usual practice in hydroclimatology is to use reanalysis data when observed data are not available. It should be noted that reanalysis data need further improvements as they do not match accurately the observed data for some cases reported in the literature [*Aihong et al.*, 2007; *Ma et al.*, 2009]. However, the present case study belongs to Himalayan region and northeast part of India, where the density of weather stations is much less. Owing to the shortage of observed data, reanalysis data are used as a proxy to the observed data. Use of reanalysis data for Indian subdivisions has also been reported in *Tripathi et al.* [2006]. In the present study, NCEP/NCAR reanalysis data are used for calibration of the downscaling model. Monthly (for monsoon period) average climate variables for 1951 to 2000 are obtained for a region spanning 5°N–40°N in latitude and 60°E–110°E in longitude. The third criterion is tested by plotting the contour plot of the correlation coefficient between the predictor variables at NCEP gridpoints and the predictand, monsoon rainfall in Assam and Meghalaya meteorological subdivision. Figure 2 shows the contour plots of correlation coefficient with monsoon rainfall for the predictor variables listed above. It shows that the monsoon rainfall in Assam and Meghalaya meteorological subdivision is correlated with the local predictor variables, selected preliminarily, except the surface temperature. The correlation between the rainfall in Assam and Meghalaya meteorological subdivision and local temperature is very low. However, correlation is high with the temperature at distant areas (northern and northwestern India), which is difficult to explain with geophysics. Such correlation may be spurious correlation and such unexplained relationship may not be valid for future under altered climatic condition. Therefore, the surface temperature is not used as a predictor, and for other predictor variables, the data are extracted for the region spanning 20°N–35°N in latitude and 85°E–105°E in longitude (constituting 63 grid points) that encapsulates the study region (Figure 3). The correlation between the climate variables and rainfall is observed to be around 0.2–0.4, which is relatively low. Therefore, a single climate variable alone cannot be considered a predictor for downscaling, and it is required to use a combination of multiple climate variables in downscaling, which results in simulating better the rainfall pattern. It should be noted that the correlation coefficient is based on a linear relationship, whereas the relationship between the predictors and predictand is nonlinear. Kendalls' Tau, which is capable of capturing a nonlinear relationship, is also used and similar results are obtained. The outputs (MSLP, surface specific humidity, zonal wind speed, and meridional wind speed) of GCMs are downloaded from IPCC data distribution center for AR4 [*IPCC*, 2007]. The GCMs considered, based on the availability of the output in IPCC data, are given in Table 1.

[10] Due to incomplete knowledge about the geophysical processes, assumptions are made in the development of a GCM in terms of parameterizations and empirical formulae. Because of these assumptions, a GCM may not simulate climate variables accurately, and there is a difference between the observed and simulated climate variable for almost all the GCMs. This difference is known as bias. It is important to remove the bias from the GCM output for projecting the future hydrologic and climatic scenario correctly. Standardization [*Wilby et al.*, 2004] is used prior to statistical down-

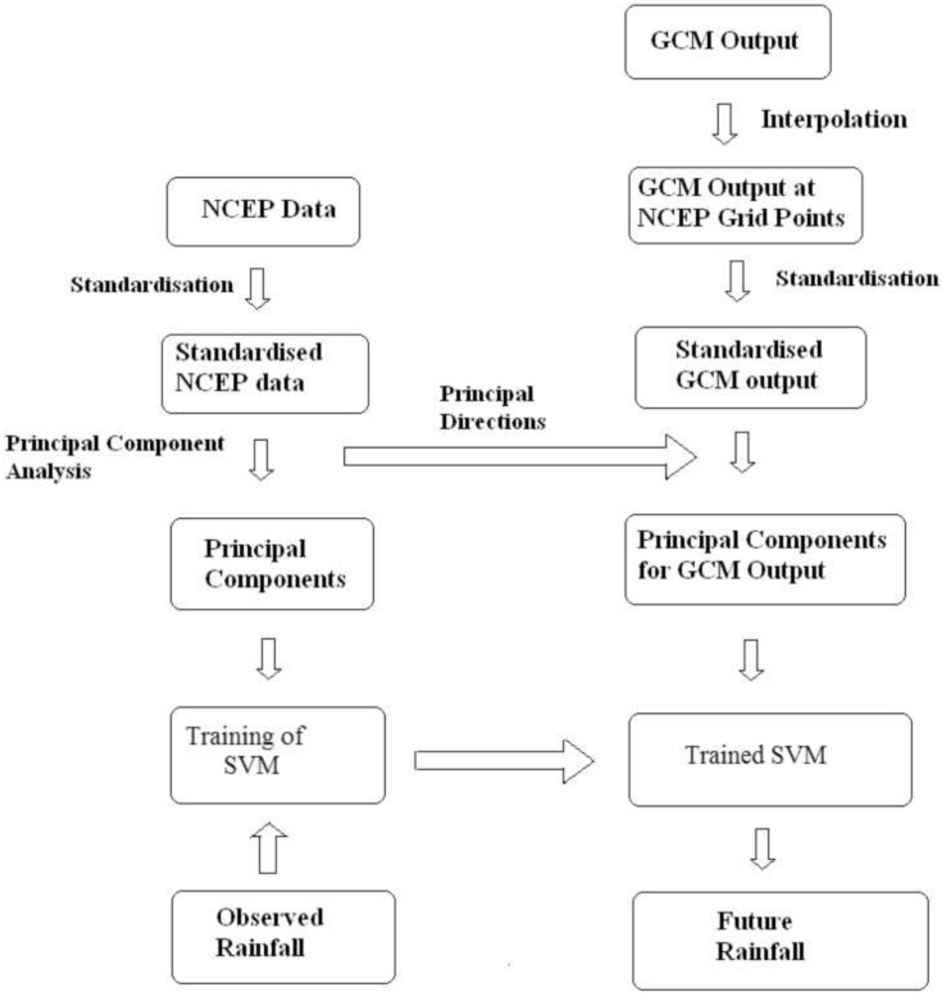

Figure 1. Flowchart of SVM-based statistical downscaling model.

scaling to reduce systematic bias in the mean and variance 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 output. The period 1961–1990 is used as a baseline period because it is of sufficient duration to establish a reliable climatology, yet not too long nor too contemporary to include a strong global change signal. Four climate variables at 63 grid points are used as predictors, and hence the dimension of the predictors is 252. Furthermore, predictors at a grid point are expected to be highly correlated with those of neighboring grid points. Therefore, direct use of the predictor variables, in statistical regression, may lead to multicollinearity and may be computationally expensive. Principal Component Analysis (PCA) is performed to reduce the dimensionality of the predictor variables. The principal components are selected on the basis of the percentage of variance of original data explained by them. Statistical downscaling model maps the variability of climate variables to the variability of rainfall using regression. Therefore, the principal components are selected on the basis of the variability of original climate variables explained by individual principal components. This criterion is also used by *Hughes and Guttrop* [1994] and *Tripathi et al.* [2006]. There are some other tests such as the weight method used by *Zorita et al.* [1995] and *Wetterhall et al.* [2005], which may also be used. It is observed that first 36 principal components represent 98% variability of the original data set and hence are used in the study.

[11] Bias-free principal components are used as regressors to predict the monthly monsoon rainfall of Assam and Meghalaya meteorological subdivision in the proposed SVM regression model. The first two thirds of the data set is used in training, and the rest of the data set is used in testing of the model. It should be noted that the downscaling model will be useful and valid if the statistical relationship between climate and hydrologic variable holds good in changed climatic condition. In the last one third data of 1950–1999, the signals of climate forcing are more visible compared to the first two thirds of data and therefore to test whether the model is valid for changed climatic condition, the last one third of the data is used for testing. The next section presents the mathematical background of support vector machine regression method.

3. Support Vector Machine Regression

[12] The Support Vector Machine (SVM) was developed by *Vapnik* [1995] and is gaining popularity due to many

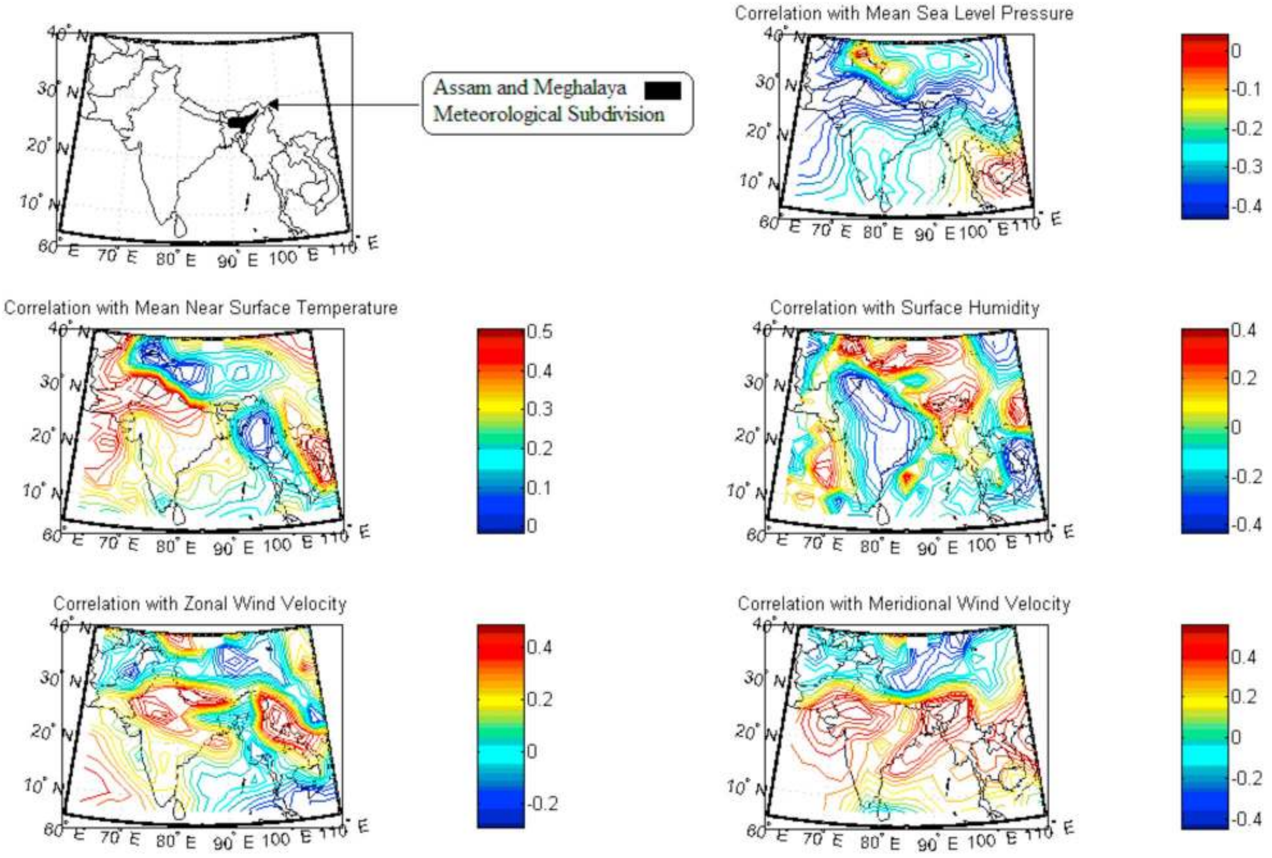
Assam and Meghalaya Meteorological Subdivision
Correlation with Mean Sea Level Pressure
Correlation with Mean Near Surface Temperature
Correlation with Surface Humidity
Correlation with Zonal Wind Velocity
Correlation with Meridional Wind Velocity
40° N
30° N
20° N
10° N
60° E 70° E 80° E 90° E 100° E 110° E

Figure 2. Contour plots of correlation between monthly climate variables and monthly monsoon (June, July, August, and September) rainfall of Assam and Meghalaya.

attractive features and its promising empirical performance. The formulation of SVM embodies Structural Risk Minimization (SRM) principle, which has been proved to be superior [*Gunn et al.*, 1997] to the traditional Empirical Risk Minimization (ERM) principle employed by conventional neural networks. SRM minimizes an upper bound on the expected risk, as opposed to ERM, which minimizes the error on the training data. This feature equips SVM with a greater ability to generalize, which is the goal of statistical learning. A brief introduction to statistical learning with the concept of SRM may be found in *Smola* [1996], *Vapnik* [1998], and *Dibike et al.* [2001].

[13] SVM-based regression method actually selects some points from the training vector and fixes the relationship between the predictands and predictor. When a new data point is fitted to the relationship, it is coupled with the selected point by kernel function and predicts the predictand.

[14] Given training data $\{(x_1, y_1), \ldots, (x_l, y_l), X \in \Re^n, Y \in \Re\}$, the Support Vector (SV) regression equation may be given by the following [*Smola*, 1996]:

$$f(x) = \sum_{i=1}^{l} w_i \times K(x_i, x) + b, \qquad (1)$$

where, $K(x_i, x)$ and w_i are the kernel functions and the corresponding weights used in the SV regression. b is a constant known as bias. The ith input x_i for training is called support vector if $w_i \neq 0$ for that particular i. x is the input variable of the SVM. The training process selects optimum number of points from the training data set which fix the relationship between predictors and predictand. These points are known

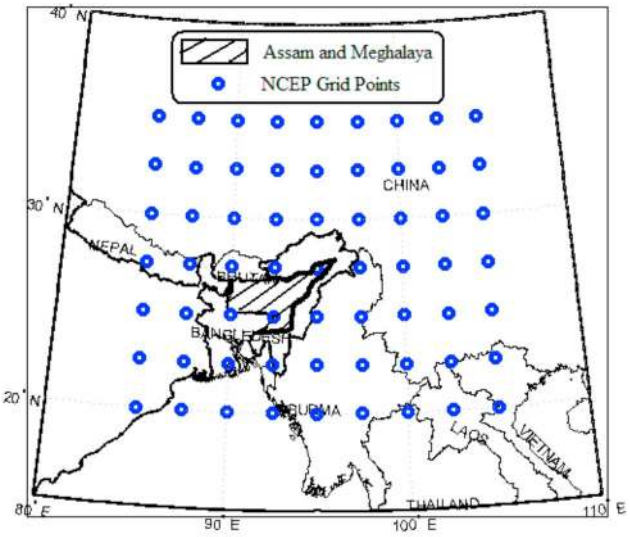

Figure 3. NCEP gridpoints superposed on the map of Assam and Meghalaya meteorological subdivision.

Table 1. GCMs Used in the Study

GCM	Institute	Spatial Resolution
BCCR	Bjerknes Centre for Climate Research, Norway	2.80° × 2.80°
CNRM	Centre National de Recherches Meteorologiques, France	2.80° × 2.80°
CM4	Institut Pierre Simon Laplace, France	2.50° × 3.75°
MIROC3.2 medres	National Institute for Environmental Studies, Japan	2.80° × 2.80°
CGCM2.3.2	Meteorological Research Institute, Japan	2.80° × 2.80°

as support vectors. The points in training data set, other than support vectors, are not required in the regression equation given in equation (1). Structural risk minimization in SVM computes the weights corresponding to the support vectors and bias, as given in equation (1). For the downscaling model developed in this chapter, x denotes a set of principal components, whereas $f(x)$ denotes the predicted values of monsoon rainfall. Details of SVM are presented in Appendix A.

[15] *Tripathi et al.* [2006] has pointed out that the performance of SVM depends on the selection of the values of C and σ (Appendix A). They have performed a grid search method to compute the best estimates of these two parameters. Performance of SVM also depends on the selection of the values of kernel parameter b (Appendix A) and loss function parameter ε (Appendix A), and therefore selection of the values of these parameters is equally important but has not gotten attention in the literature. Handling decision variables of dimension 4 is also difficult by a grid search method, and therefore a sophisticated search algorithm is required to use for evaluation of the values of these four parameters with proper development of optimization model. An effort is made in this regard in the present study, which is described in the next section.

4. SVM-PGSL Coupled Approach

[16] To compute the best estimates of the four parameters of SVM, viz., C, σ, b and ε, the following optimization model is developed:

$$Maximize \quad r(testing) \tag{2}$$

subject to

$$|r(training) - r(testing)| \leq d \tag{3}$$

$$r(training) = Correlation \ \ Coefficient\left(y_{train}, f(x)_{train}\right) \tag{4}$$

$$r(testing) = Correlation \ \ Coefficient\left(y_{test}, f(x)_{test}\right) \tag{5}$$

$$f(x) = g(C, \sigma, b, \varepsilon), \tag{6}$$

where, r(training) and r(testing) equations (4) and (5) are the correlation coefficients of observed (y) and predicted values (f(x)) for training (two thirds of the data set) and testing (one third of the data set) data set. The objective function equation (2) of the optimization model is to maximize r (testing), which is independent of training. Therefore, the objective function will select the model which fits best to a new data set independent of training, and thus it will perform well in altered climatic conditions. High difference between the r values for training and testing denotes overtraining, and therefore it is kept less than an allowable value (d) to control overtraining equation (3). The predicted values f(x) are computed from the SVM model and therefore can be considered to be the function of decision variables C, σ, b, and ε equation (6). As the objective function and the constraints used in the optimization model equations (2)–(6) are nonlinear and cannot be expressed in a functional form of decision variables, it is difficult to obtain an analytical solution. A search algorithm, Probabilistic Global Search Laussane (PGSL) is proposed to use for this purpose which will generate the values of decision variables from their domain and select the best solution leading to the best estimate of the objective function. As the problem is a constrained optimization problem, it is converted into an unconstrained problem by penalty function method. In the present analysis, bracket operator penalty term is used.

$$F = obj(x) + \zeta \sum_{j=1}^{k} \delta_j \, \nu_j^2 \tag{7}$$

where, F = modified objective function value,
obj(x) = objective function value, here test r value
k = total number of constraints,
$\zeta = -1$ (for maximization problem),
δ_j = penalty coefficient (a large value) for jth constraint,
v_j = amount of violation in jth constraint.
Whenever there is a constraint violation the penalty function value is added to the objective function value to make the solution inferior. The algorithm is presented in Figure 4. It should be noted that the optimization model equations (2)–(5) are not the structural risk minimization (similar to least square optimization of conventional regression) of SVM regression; rather, this is the optimization model which selects the best SVM with minimum overtraining, out of several trained SVM models. The next subsection presents a brief overview of PGSL.

4.1. Probabilistic Global Search Laussane

[17] Probabilistic Global Search Laussane (PGSL), a global search algorithm for the solution of nonlinear optimization, was developed starting from the observation that optimally directed solutions can be obtained efficiently through sampling the search space without using special operators. The principal assumption is that better points are likely to be found in the neighborhood of families of good points. PGSL has been developed at IMAC (informatique et de mecanique appliques la construction) [*Raphael and Smith*, 2000]. It has already been applied to several tasks in the field of structural engineering, such as optimization problem in timber structures [*Svanerudh et al.*, 2002]. Tests on benchmark problems having multiparameter nonlinear objective

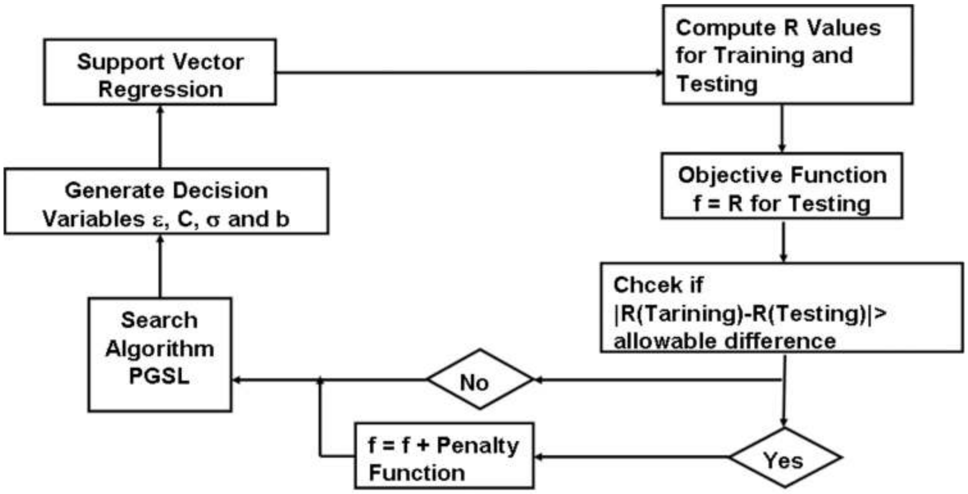
Support Vector Regression
Compute R Values for Training and Testing
Objective Function f = R for Testing
Chcek if |R(Tarining)-R(Testing)|> allowable difference
Yes
No
f = f + Penalty Function
Search Algorithm PGSL
Generate Decision Variables ε, C, σ and b

Figure 4. Search algorithms for selection of hyper-parameters used in SVM.

functions revealed that PGSL performs better than genetic algorithms and advanced algorithms for simulated annealing [*Raphael and Smith*, 2003].

[18] PGSL is basically a direct method which depends on the objective function only through ranking a countable set of function values. PGSL uses the assumption that better sets of points are more likely to be found in the neighborhood of good sets of points, therefore intensifying the search in regions that contain good solutions [*Domer et al.*, 2003]. PGSL algorithm consists of four nested cycles: sampling cycle, probability updating cycle, focusing cycle, and subdomain cycle [*Raphael and Smith*, 2003]. Initially uniform Probability Density Function (PDF) is assumed for all the decision variables of the optimization model to start the searching. In the sampling cycle number of points (say NS) is generated randomly by generating a value for each variable according to the PDF. Among them, the best sample is selected. In a probability updating cycle, the sampling cycle is invoked for a number of times (say NPUC). After each iteration, the PDF of each variable is modified. The interval containing the best solution is first selected, and then the probability of that interval is multiplied by a factor greater than 1. The PDF thus generated is then modified to make the area under the density function equal to unity. This ensures that the sampling frequencies in regions containing good points are increased. In a focusing cycle, probability updating cycle is repeated for NFC times. After each iteration, the search is increasingly focused on the interval containing the current best point. The interval containing the best point is divided into uniform subintervals. A 50% probability is assigned to this interval. The remaining probability is then distributed to the region outside this interval in such a way that the PDF decays exponentially from the best interval. In subdomain cycle, the focusing cycle is repeated NSDC times and at the end of each iteration, the current search space is modified. In the beginning, the entire space is searched, but in subsequent iterations a subdomain is selected for search. The size of the subdomain decreases gradually and the solution converges to a point. The flowchart for PGSL is presented in Figure 5. Details of the algorithms are available in *Raphael and Smith* [2003]. SVM coupled with global search algorithm PGSL is used to solve the statistical downscaling problem with the best possible values of SVM parameters. The next section presents details of the results obtained.

5. Results and Discussion

[19] The SVM-PGSL coupled approach is applied in the regression-based statistical downscaling model for forecasting of rainfall in Assam and Meghalaya meteorological subdivision. The principal components of standardized NCEP/NCAR reanalysis data are used as predictors, and the monsoon rainfall in Assam and Megahlaya meteorological subdivision is used as predictand (of duration 1950–1999). The first two thirds of the data set is used for training and the rest is used for testing of model. Conventional regression-based approaches consider three sets, training, and testing validation. However, for the present case, the same size is very small (200), and therefore it is split into only training and testing sets. The model is applied with the allowed difference of *r* values between training and testing as 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, and 0.40. The plots of *r* values (for training and testing) with the allowed difference are presented in Figure 6. It is observed that initially until 0.25, with the increase of allowed difference between the *r* values of training and testing, both the *r* values are increasing. This suggests that the improvement of model is possible in terms of testing *r* till the difference in *r* values is 0.25. When the allowed difference is made more than 0.25, no improvement is observed and therefore allowed difference between *r* values of training and testing more than 0.25 is not recommended. Linear regression model is also applied to the statistical downscaling model and the training and testing *r* values are obtained as 0.82 and 0.62. The results obtained from linear regression are observed to be inferior to SVM in terms of training and testing *r* when the allowable difference between them in SVM is greater than or equal to 0.2. Considering the allowable difference more than 0.2 leads to a model inferior to linear regression on the basis of the criterion "allowable difference." As an SVM model inferior to linear regression on any criterion is not desired, the allowable difference of *r* values from training and testing is recommended as 0.2. The SVM parameters, viz, C, σ, b, and ε are obtained as, 462.4774, 3.7145, 1.9995, and 0.1223, respectively, with around 500

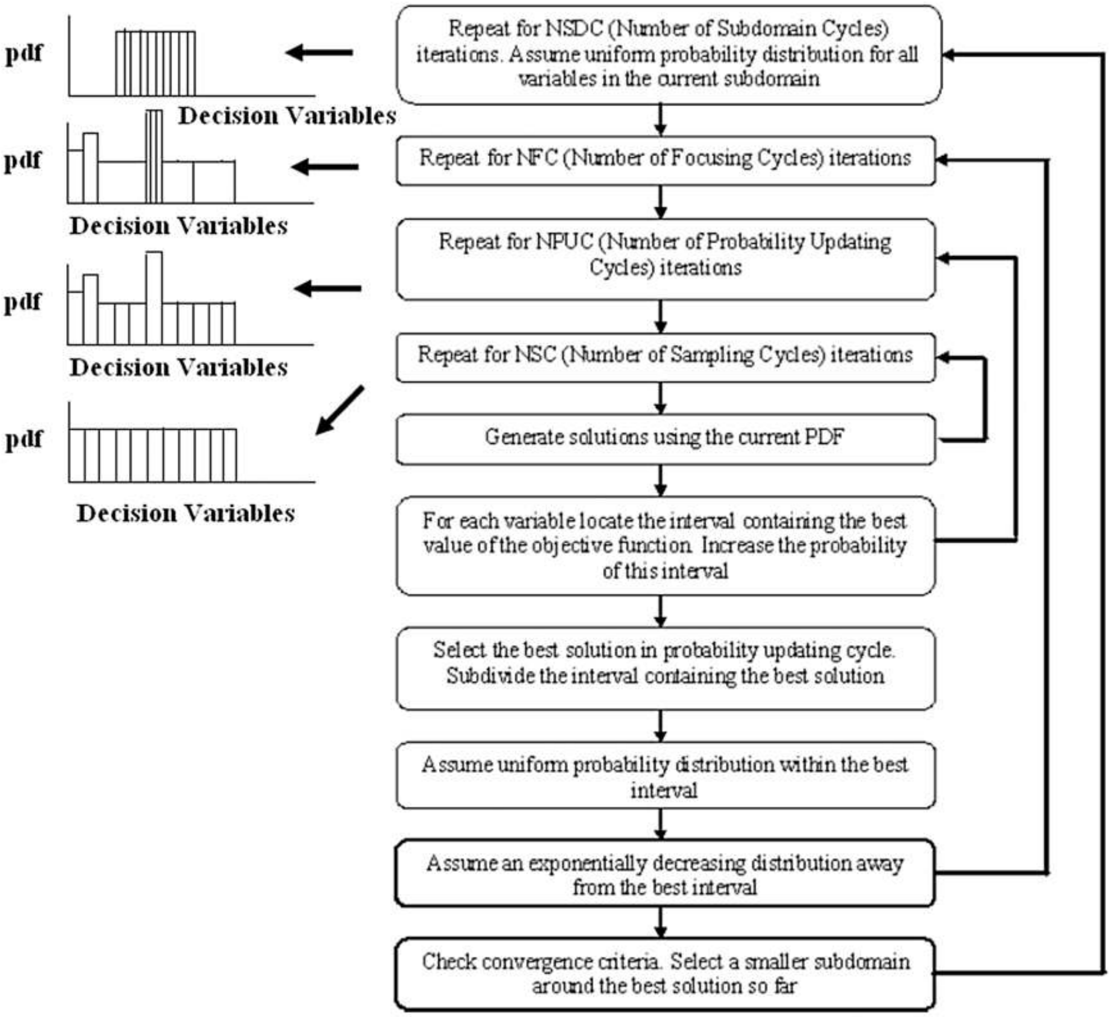
pdf
Decision Variables
pdf
Decision Variables
pdf
Decision Variables
pdf
Decision Variables
Repeat for NSDC (Number of Subdomain Cycles) iterations. Assume uniform probability distribution for all variables in the current subdomain
Repeat for NFC (Number of Focusing Cycles) iterations
Repeat for NPUC (Number of Probability Updating Cycles) iterations
Repeat for NSC (Number of Sampling Cycles) iterations
Generate solutions using the current PDF
For each variable locate the interval containing the best value of the objective function. Increase the probability of this interval
Select the best solution in probability updating cycle. Subdivide the interval containing the best solution
Assume uniform probability distribution within the best interval
Assume an exponentially decreasing distribution away from the best interval
Check convergence criteria. Select a smaller subdomain around the best solution so far

Figure 5. PGSL algorithm used for solving optimization model.

iterations to converge. All the parameters are unitless, as by standardization, the predictand has been made unitless. With these parameters, the training and testing *r* values are obtained as 0.87 and 0.67, which are slightly better than those of linear regression. Although the performance based on correlation coefficient for this case study does not show much improvement with respect to the linear regression, the Mean Square Error (MSE) of predicted value from observed values shows a significant improvement. The MSE for linear regression is obtained as 5.820×10^3 mm^2 and that for SVM is 3.450×10^3 mm^2. With respect to linear regression, the results of SVM show 11% improvement. For comparison purpose, Artificial Neural Network (ANN) is also used for downscaling. Several ANN structures (one and two hidden layers) with multiple transfer functions are tried, and for the best trial in terms of test errors, the training and testing *r* values are obtained as 0.86 and 0.64, which are slightly inferior to the results of SVM-PGSL coupled approach. Grid search method is also applied for comparison purpose, where the hyperparameters are selected based on the criterion minimum error for the test set. The correlation coefficients *r* for training and testing are obtained as 0.97 and 0.68. The performance measure for testing is slightly higher than that of SVM-PGSL coupled approach, but the difference is significantly larger for the methodology using test set. The huge difference between training and testing *r* signifies the possibility of overfitting. Such a method does not guarantee good performance in a changed climatic condition, and therefore the criterion of "allowable difference" plays an important role in the present study and shows the importance of the proposed approach. Furthermore, grid search algorithm for solving the present problem requires huge number of grids (five grids for each of the four variables, leading to 625 function evaluations) with a high computational effort in searching and therefore a logical search approach like PGSL is preferred in the present study. *Cherkassky and Ma* [2004] have provided an analytical approach based on empirical equation to solve C and ε with the fixed values of σ and b (which have significant impact on the performance of SVM). The present method does not consider any empirical equations and used logical search to solve for all the four parameters. For validation, σ and b are fixed to the values determined from SVM-PGSL coupled approach and the empirical equations from *Cherkassky and Ma* [2004] are applied. The training and testing *r* are obtained as 0.82 and 0.61. This inferiority of results is maybe because of the use of empirical equations which may not be valid for

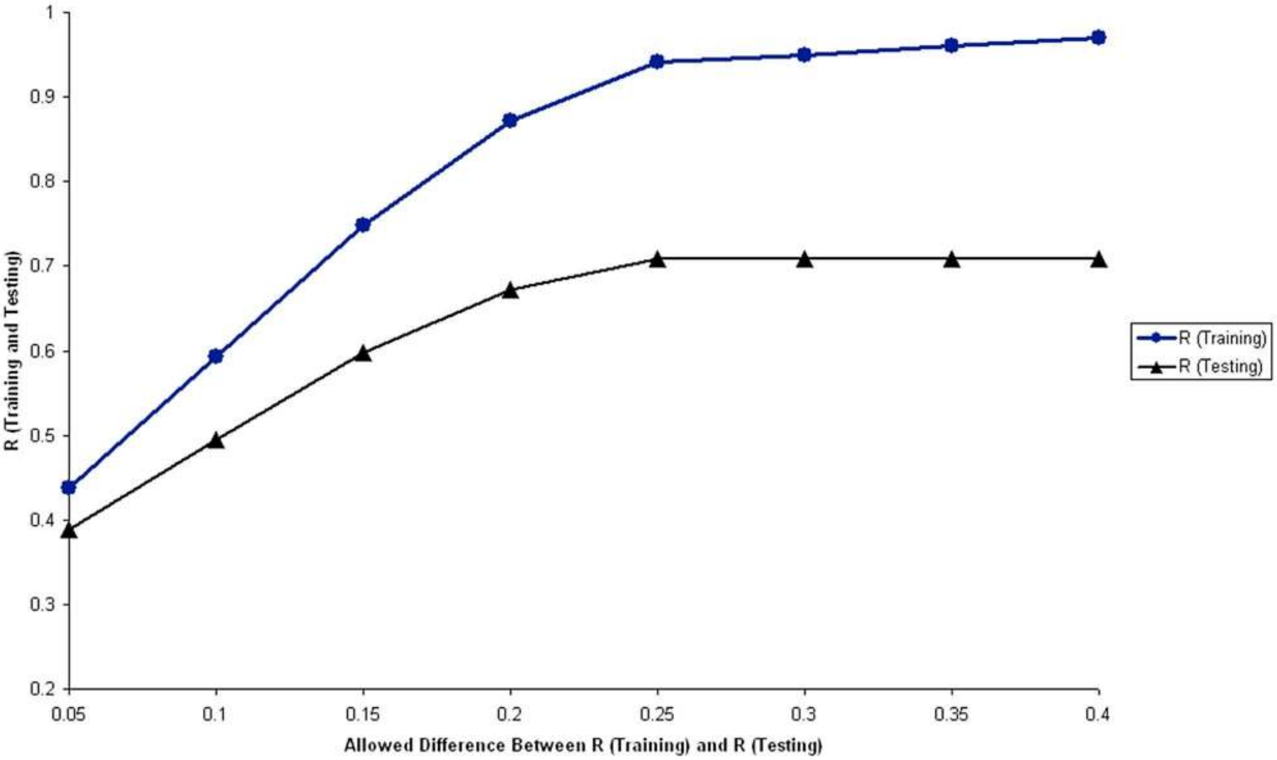

Figure 6. Variation of training and testing R with the allowable difference between them.

this case study. It should be noted that improvement in terms of *r* values by using SVM is not significant over those by ANN. The present study area is a typical case study where ANN, even linear regression, performs quite well and therefore significant improvement using SVM is not observed. However, for other Indian subdivisions, significant improvement using SVM (with grid search) over ANN is observed by *Tripathi et al.* [2006]. Therefore, it can be concluded that, although significant differences in results are not observed in the present case study, the usefulness of the present model (SVM coupled with PGSL) will be more visible in terms of system performance measure, *r* value, for other subdivisions.

[20] After selecting the SVM parameters, the monsoon data of Assam and Meghalaya meteorological subdivision for years 1950–1999 is used for calibration of the model and the results (predicted and observed) are presented in Figure 7. The goodness of fit of the model is also tested with Nash-Sutcliffe coefficient [*Nash and Sutcliffe*, 1970], which has been recommended by 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 the following:

$$E = 1 - \frac{\sum_t \left(P_{ot} - P_{pt}\right)^2}{\sum_t \left(P_{ot} - \overline{P_o}\right)^2} \tag{8}$$

where, P_{ot} and P_{pt} are the observed and predicted rainfall in time t, and $\overline{P_o}$ is the mean observed rainfall. Maximum value of Nash-Sutcliffe coefficient is 1. Value of E as 0 indicates that the model predicts no better than the average of the observed data, and 1 indicating a perfect fit. The value of *E* is obtained as 0.65 for the model, which is satisfactory [*Mujumdar and Ghosh*, 2008; *Dankers et al.*, 2007]. After validation, the SVM model is applied to the outputs of GCMs presented in Table 1. The outputs of the GCMs are first standardized with the individual means and standardization of 20C3M runs for 1950–1999. It should be noted that spatial resolutions of GCMs are different from NCEP grid settings, and hence linear inverse square interpolation [*Willmott et al.*, 1985] is used to obtain NCEP gridded GCM output. The principal components of GCMs are derived with the principal directions/eigen vectors obtained with reanalysis data. The trained SVM is applied to the principal components to obtain the future projections for A1B, A2, and B1 scenarios. For validation, the rainfall is also simulated for 20C3M runs (years 1950–1999).

[21] The results for statistical downscaling with the simulations of 20C3M for all the GCMs, mentioned in Table 1, are presented in Figure 8. The results are presented in terms of Cumulative Distribution Function (CDF). The CDF is derived with Weibull's plotting position formula. The CDFs derived with 20C3M projections for all the GCMs are not deviating significantly from that of observed. Thus the performances of all the GCMs for 20C3M are similar and quite well matching with that of observed data. However, downscaled GCM simulations show poor skill in capturing the extreme events during calibration as well as future periods. This is because standardization may reduce the bias in the mean and variance of the predictor variable, but it is much harder to accommodate the bias in large-scale patterns of atmospheric circulation in GCMs (e.g., shifts in the dominant

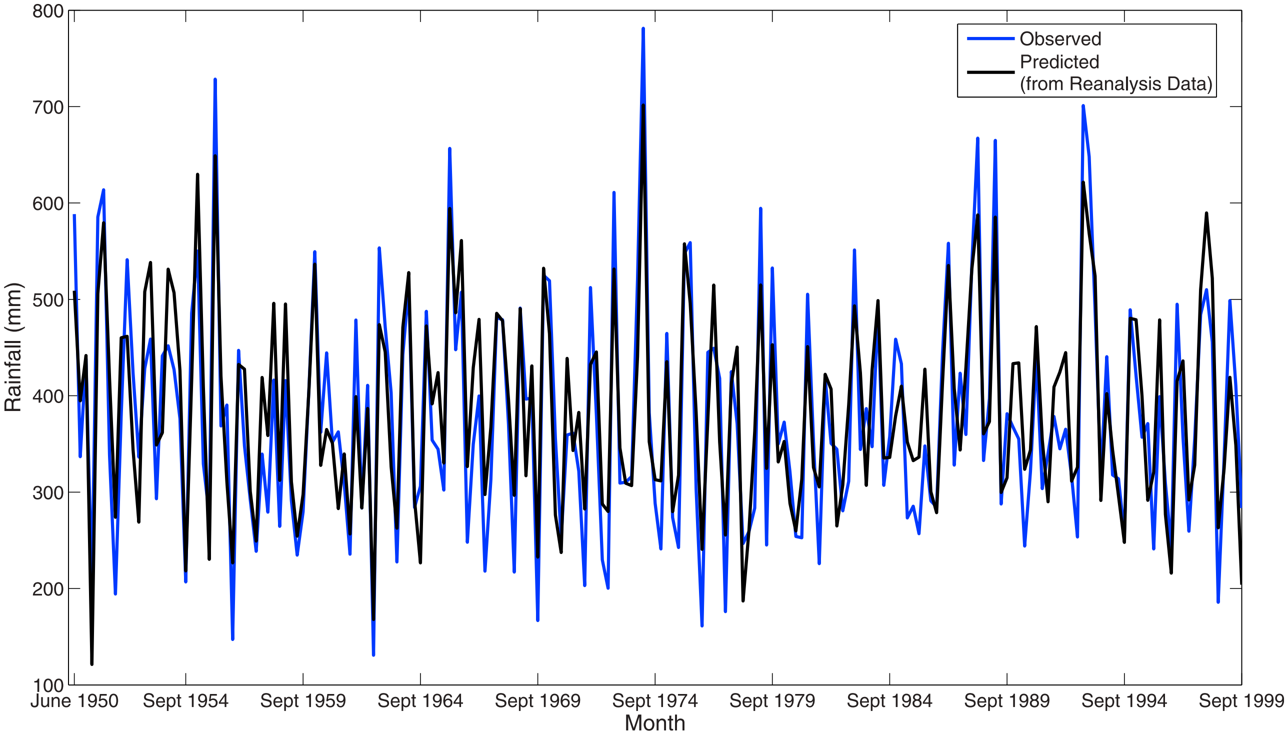

Figure 7. Time series plot of observed and predicted (from reanalysis data) monsoon (June, July, August and September) rainfall (monthly) in Assam and Meghalaya meteorological subdivision.

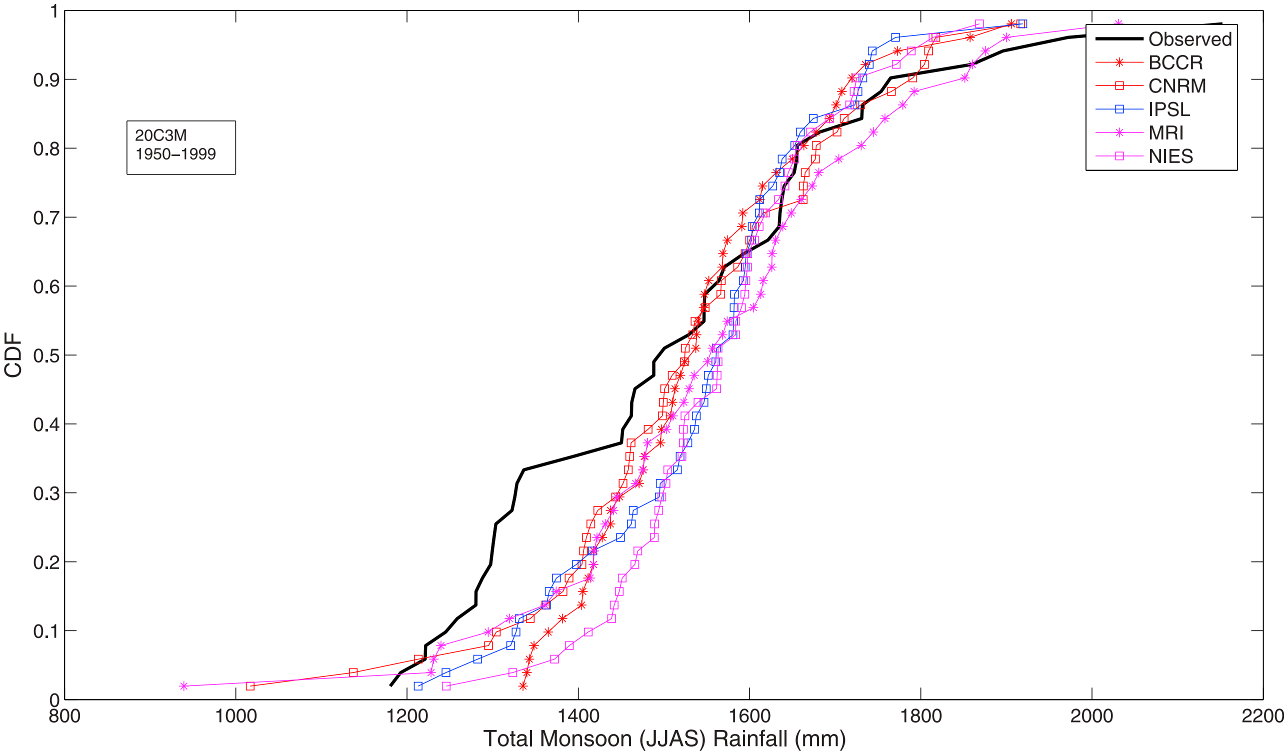

Figure 8. CDF of downscaled seasonal monsoon (June, July, August and September) rainfall from GCM output for 20C3M (duration 1950–1999).

Table 2. Validation With Downscaled GCM Output (1950–1999) [unit: mm/month]

GCM/Observed	Mean	Standard Deviation	Maximum	Minimum
Observed	377.12	119.08	781.2	130.8
BCCR	385.05	74.96	589.05	232.73
CNRM	382.31	104.04	638.19	101.89
CM4	385.96	82.10	573.10	154.67
CGCM2.3.2	388.88	89.89	606.64	163.27
MIROC3.2 medres	391.48	95.01	620.46	210.68

storm track relative to observed data) or unrealistic intervariable relationships between predictor variables [*Wilby and Dawson*, 2004]. Another reason may be that the GCMs are not able to reproduce the extreme value as many extreme events occur at a much smaller scale, which cannot be resolved by GCMs dynamically. Table 2 compares the mean, standard deviation, minimum, and maximum values of downscaled GCM output for 20C3M with observed data. It is observed that the mean rainfall for 1950–1999 is well simulated by the GCMs; however, the extreme low (minimum) and extreme high (maximum) events are not well simulated by all the GCMs. This does not result in very good match of standard deviation. This is because of GCM's inability to model local scale extreme events. It is true that after selection of the best model, verification is not performed with a data set independent of both training and testing. However, the model is applied with 20C3M scenario of GCM outputs (independent of training and testing data) and for 1950–1999, the results are found to resemble the observed data set. This verifies the applicability of the selected SVM for downscaling purposes. After verifications with 20C3M, the downscaling model is applied to A1B, A2, and B1 scenarios.

[22] The results for A1B, A2, and B1 scenarios are presented in Figures 9, 10, and 11, respectively. They are presented in terms of CDFs derived with Weibull's plotting position formula for three standard time slices, 2020s, 2050s, and 2080s. For all the scenarios, possible increases of monsoon rainfall are observed for almost all the GCMs. However, there are significant mismatches between the projections of GCMs for the future. This suggests that use of the output of a single GCM is not reliable. Downscaled outputs of a single GCM represents a single trajectory among a number of realizations derived using various GCMs. Such a single trajectory alone cannot represent the uncertainty related to future hydrologic condition 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, and thus downscaling a single GCM output is not particularly useful for risk adaptation studies [*New and Hulme*, 2000]. In the present study, uncertainty resulting from the use of multiple GCMs is modeled using a modified version of Reliability Ensemble Averaging (REA) proposed by *Giorgi and Mearns* [2003]. Details of uncertainty modeling are presented in the next subsection.

5.1. Uncertainty Modeling

[23] For modeling GCM uncertainty in climate change impact assessment, *Giorgi and Mearns* [2002, 2003] proposed the Reliability Ensemble Averaging (REA) method.

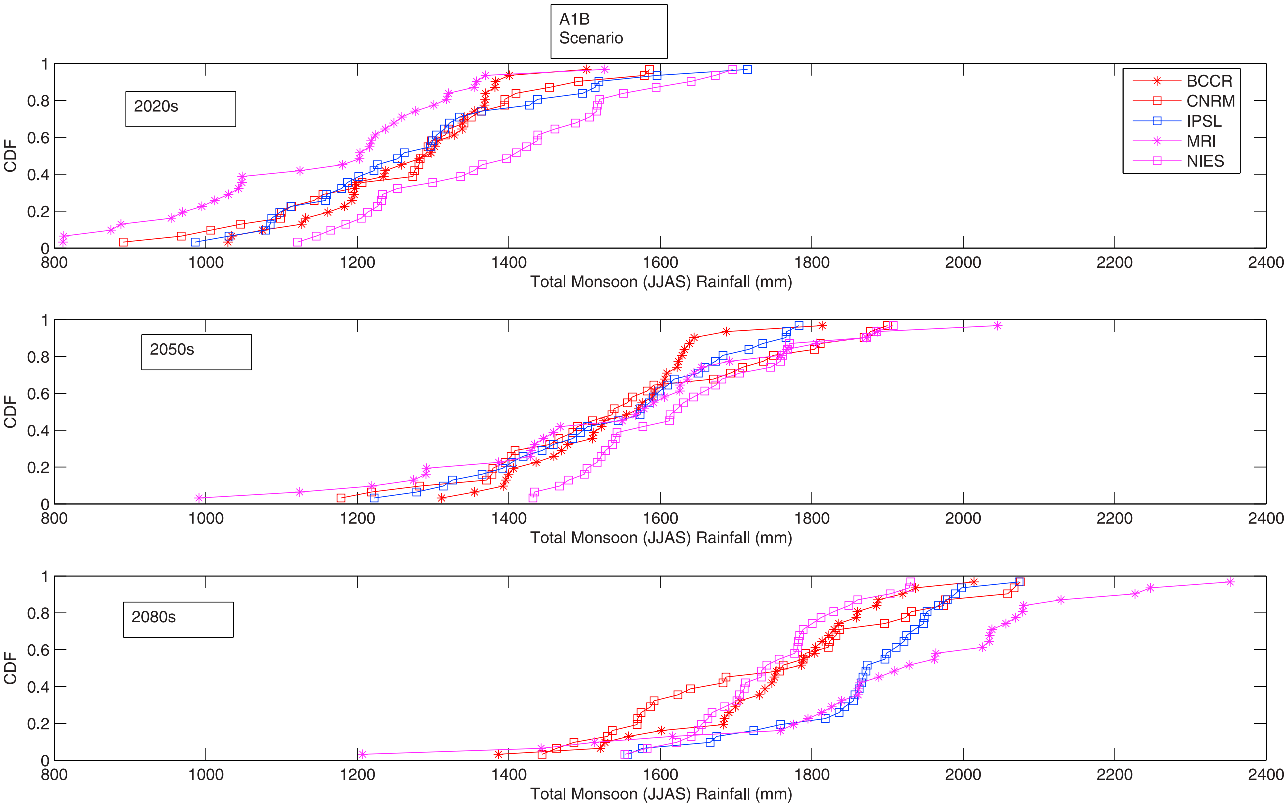

Figure 9. CDF of predicted seasonal monsoon (June, July, August and September) rainfall from multiple GCM output for A1B scenario.

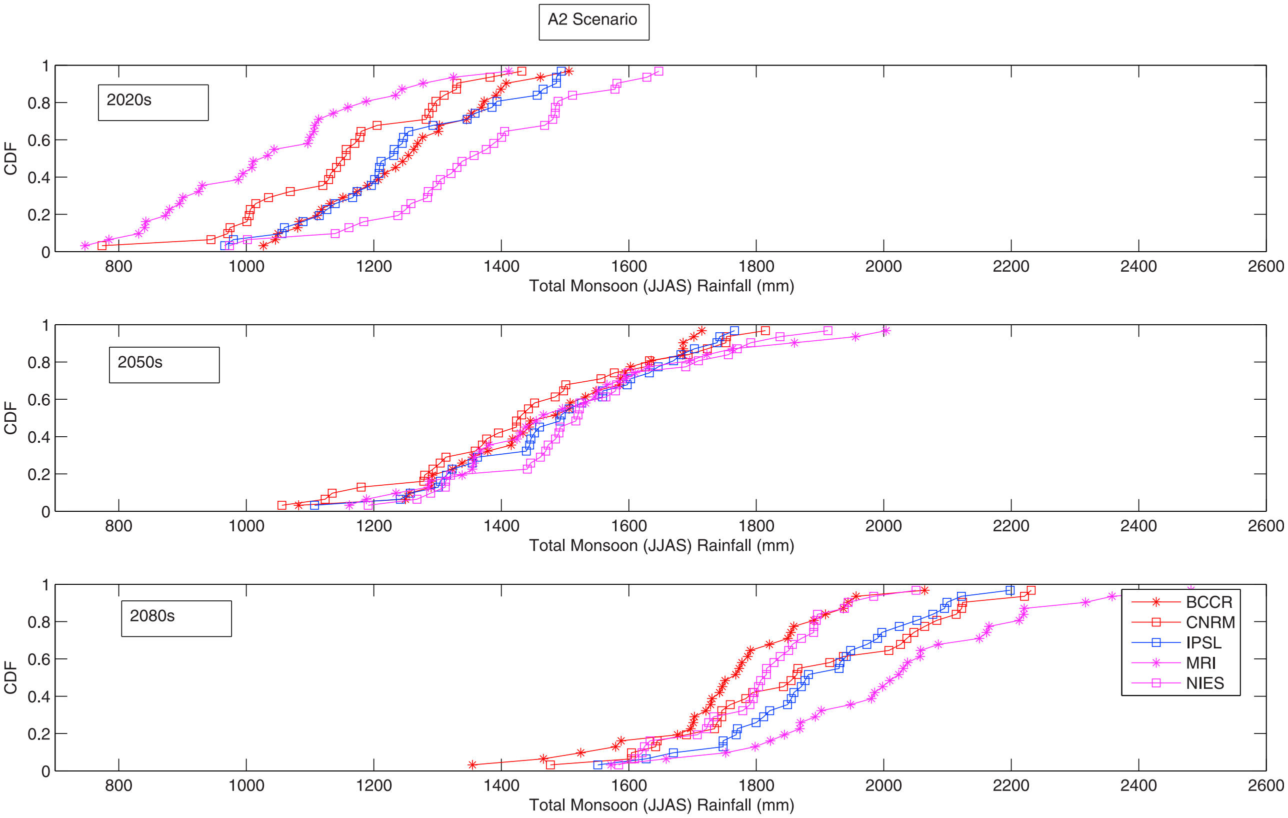

Figure 10. CDF of predicted seasonal monsoon (June, July, August and September) rainfall from multiple GCM output for A2 scenario.

The method takes into account two reliability criteria: the performance of the model in reproducing present-day climate ("model performance criterion") and the convergence of the simulated changes across the models ("model convergence criterion"). The first criterion is based on the ability of GCMs to reproduce present-day climate: the better the model performance, the higher the reliability of the GCM. The second criterion is based on the convergence of simulations by different models for a given forcing scenario for the future. As the observed climate time series is not available for the future, a factor is used as reliability indicator of a GCM, which measures the model reliability in terms of the deviation of simulations by that GCM from the REA average (weighted mean) simulations. High deviation denotes low model reliability. The philosophy underlying the REA approach is to minimize the contribution of simulations that either perform poorly in the representation of present-day climate over a region or provide outlier simulations for future with respect to the other models in the ensemble. In the present study, the deviation of the simulated variable (rainfall) with respect to the observed or REA average variable is computed with the deviations of mean and standard deviation.

[24] The REA method was applied by *Giorgi and Mearns* [2003] to mean seasonal temperature and precipitation changes over 22 land regions of the world at continental scales for A2 and B2 scenarios. In the present study, the objective is to model monsoon rainfall at subdivisional scale with an estimate of the temporal variation along 30-year time slices. Therefore, the earlier developed REA model is slightly modified here and performed with respect to mean and standard deviation of the monsoon rainfall and not only with respect to the mean condition. Model performance measure was evaluated by determining the total deviation of mean and standard deviation of GCM-simulated downscaled rainfall for 20C3M (duration, 1950–1999) with respect to those of observed rainfall. Model convergence measure is evaluated based on the deviation of mean and standard deviation of rainfall simulated with individual GCMs for future with respect to weighted means of mean and standard deviation (derived with the weighted projections of multiple GCMs). As weights are unknown and to be determined using REA, the algorithm used is an iterative method. The algorithm for the proposed approach is as follows.

[25] 1. Weights are assigned to GCMs based on the model performance. The deviation of the mean and standard deviation of GCM projected rainfall (downscaled), for 20C3M (duration, 1950–1999), from those of observed data for the same duration (years, 1950–1999) is computed. The inverse values of total deviations are proportionately used as weights so that the sum of weights across all the GCMs is equal to 1.

$$\mu_{dev_i} = \left|\mu_{GCM_i,20C3M} - \mu_{obs}\right| \tag{9}$$

$$\sigma_{dev_i} = \left|\sigma_{GCM_i,20C3M} - \sigma_{obs}\right| \tag{10}$$

$$dev_i = \mu_{dev_i} + \sigma_{dev_i} \tag{11}$$

$$w_i = \frac{1/dev_i}{\sum_{i=1}^{NG} 1/dev_i} \tag{12}$$

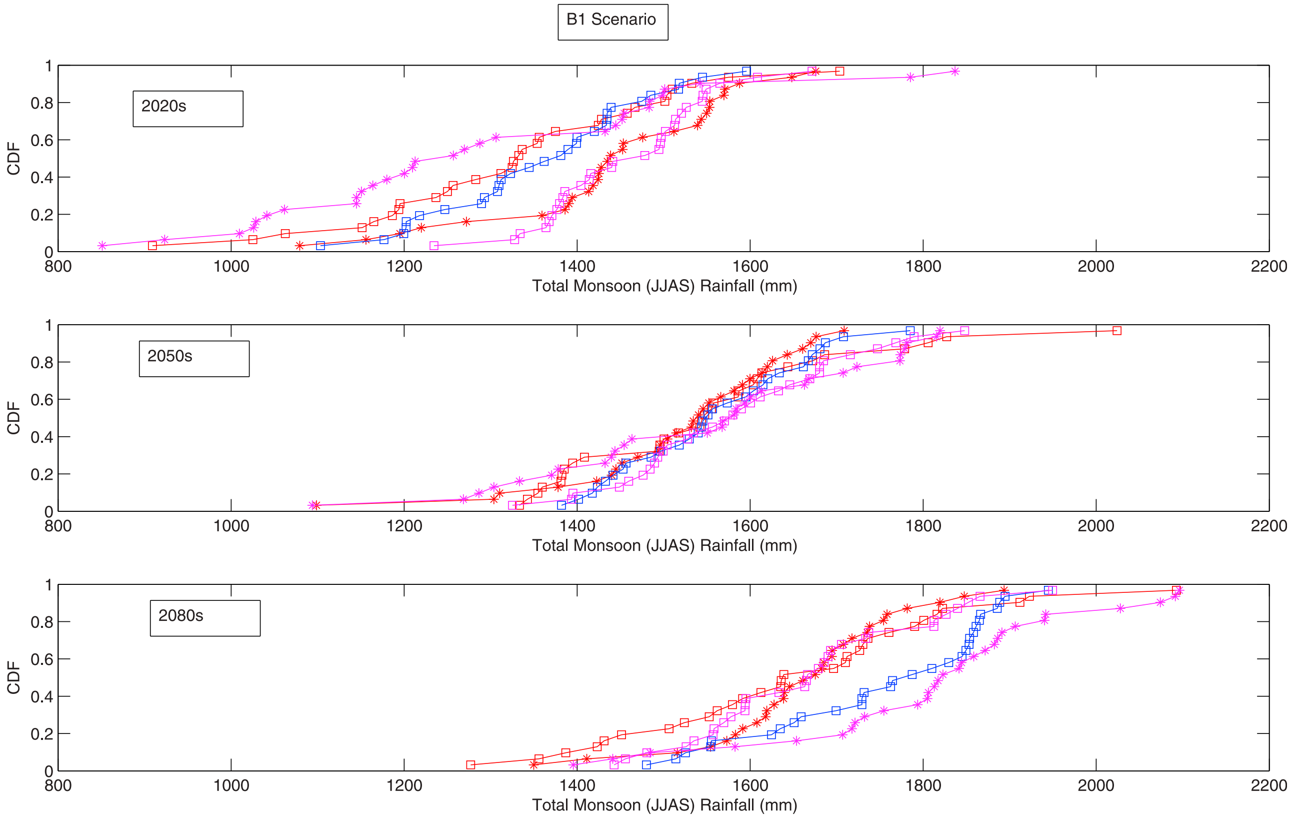

Figure 11. CDF of predicted seasonal monsoon (June, July, August and September) rainfall from multiple GCM output for B1 scenario.

where, μ_{devi}, $\mu_{GCM,\ 20C3M}$, μ_{obs} are mean deviation for ith GCM, mean of rainfall simulated for 20C3M by ith GCM and the observed mean, respectively. σ denotes standard deviation with same subscript notation. dev_i, w_i and NG denotes deviation for ith GCM, weight for ith GCM, and total number of GCMs respectively. Differences between the mean of observed and simulated data for 1950–1999 denotes the inverse of system performance of a particular GCM in reproducing mean condition. Differences between the standard deviation of observed and simulated data for 1950–1999 denotes the inverse of system performance of a particular GCM in reproducing temporal variability of hydrologic variable during the period. Both are important in hydrologic context and therefore equal weights are assigned to both the criteria by adding them. It should be noted that high standard deviation partially means that the hydrologic variable has high temporal variability, possibly with higher occurrences of extremes.

[26] 2. The weights thus computed are used as initial weights assigned to the GCMs.

[27] 3. With the weights and the mean and standard deviation of rainfall downscaled with GCM predictions, the weighted mean of mean and standard deviation of future monsoon rainfall is computed.

[28] 4. The deviation of the mean and standard deviation of future rainfall for all the GCMs are computed individually from the weighted means computed in step 3.

[29] 5. The average of the inverse of deviations (derived from steps 1 and 4) is computed and proportionately (maintaining the same ratio among the weights) used as new weights so that the sum of new weights across all the GCMs is equal to 1.

[30] 6. Steps 3 to 5 are repeated until convergence of the weights is achieved.

[31] The weights obtained using the above-mentioned algorithm for A1B, A2, and B1 scenarios are presented in Table 3. For all the scenarios, highest weight is assigned to the GCM, CNRM based on the results of modified REA. These weights are further used to compute the weighted mean CDF [*Mujumdar and Ghosh*, 2008; *Ghosh and Mujumdar*, 2009] for the three time slices in the future: 2020s, 2050s, and 2080s. The weighted mean CDFs for A1B, A2, and B1 are presented in Figures 12, 13, and 14 respectively. It is observed that for all the scenarios, there is a possibility of increase in summer monsoon rainfall of Assam and Meghalaya meteorological subdivision. A2 scenario shows the highest increase, whereas B1 scenario projects less severe changes. In India, during monsoon, Assam and Meghalaya meteorological subdivision receives highest rainfall and

Table 3. Weights Assigned to the GCMs

GCM	A1B	A2	B1
BCCR	0.1364	0.1726	0.1567
CNRM	0.4427	0.2882	0.3110
CM4	0.1959	0.2489	0.2009
MIROC3.2 medres	0.1403	0.1779	0.1892
CGCM2.3.2	0.0847	0.1124	0.1422

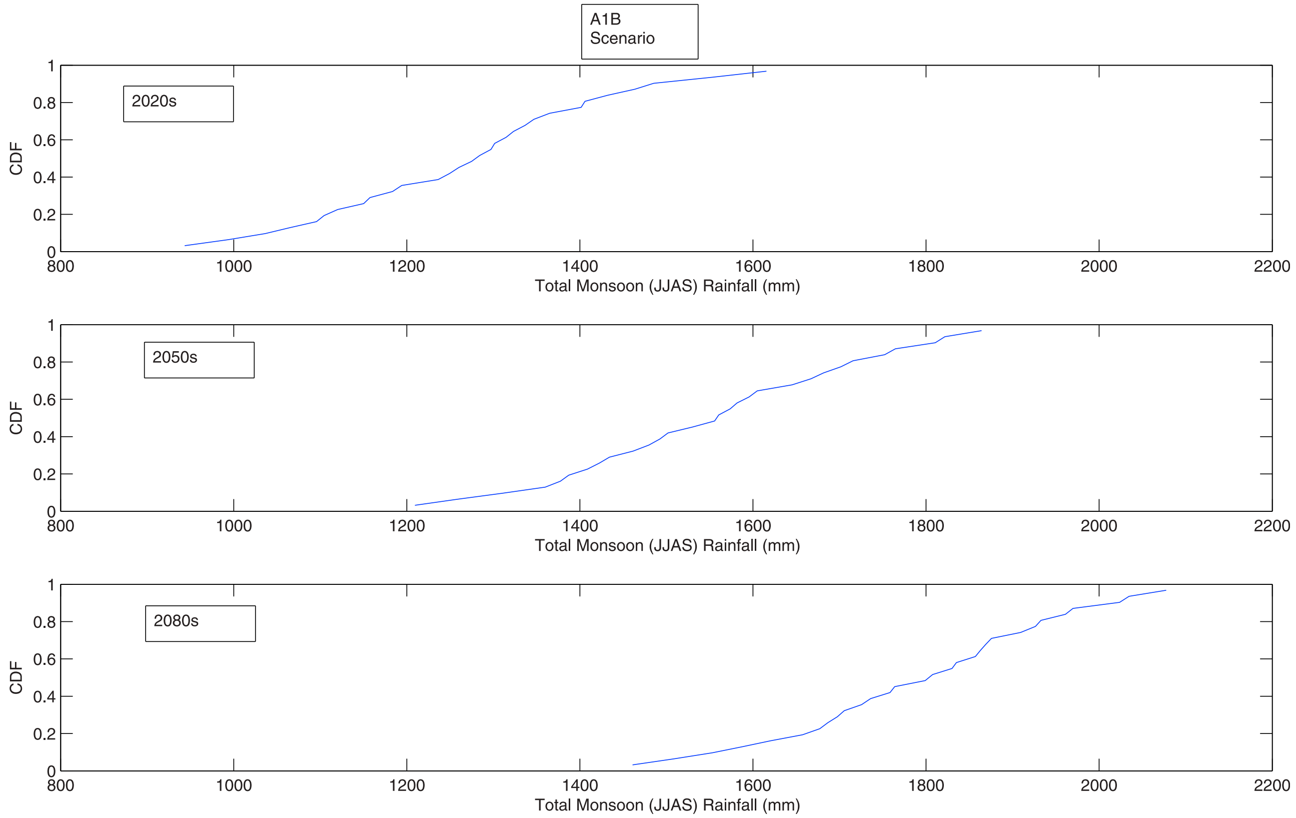

Figure 12. Weighted mean CDF of monsoon rainfall in Assam and Meghalaya meteorological subdivision for A1B scenario.

is flood prone because of high occurrences of flood in Brahmaputra River. With the possible increase in rainfall, the flood condition will be more severe in the future. Long-term planning of flood management using the weighted CDF is essential for the case study area.

6. Concluding Remarks

[32] Statistical downscaling model based on Support Vector Machine is developed to predict the monsoon rainfall of Assam and Meghalaya meteorological subdivision from GCM outputs. The parameters of support vector machine control overtraining of the model, which is a key factor when the trained model will be applied to a changed condition for future. With the objective of maximizing the system performance measure for testing data set as well as minimizing the overtraining, the parameters of SVM are selected using a global search algorithm, Probabilistic Global Search Laussane (PGSL). SVM coupled with PGSL is applied for calibration of the model, and the calibrated model is used for future prediction with GCM outputs, for A1B, A2, and B1 scenarios. The rainfall downscaled with different GCMs shows different projections. Therefore, relying on a single GCM may not be correct for adaptation purposes. Uncertainty associated with multiple GCMs is modeled with modified Reliability Ensemble Averaging (REA), which assigns weights to GCMs based on "model performance" and "model convergence." The weights are further used to derive the weighted mean CDF for future. The predictions show a possible increase of summer monsoon rainfall in Assam and Meghalaya meteorological subdivision for all the three scenarios. Increase of rainfall in flood-prone areas of northeast India requires proper flood management planning for future where the derived weighted mean CDF of future rainfall may be used. The limitations of the model are as follows.

[33] 1) The developed methodology is computationally intensive and use of this method, for daily scale downscaling, will be computationally more difficult. However, after obtaining the hyperparameters with the computationally intensive SVM-PGSL approach, the model does not take significant CPU time in deriving the relationship between climate and hydrologic variables and using it in climate change impact assessment.

[34] 2) The developed model uses correlation coefficient as the performance measure for SVM and uses it in the optimization model equations (2)–(6). However, it is reported in literature [*Krause et al.*, 2005; *Jain and Sudheer*, 2008], that any system performance index alone is not adequate in describing the performance of a model. Use of other performance measures such as the Nash Sutcliffe Coefficient along with correlation coefficient in the optimization model equations (2)–(6) may be considered as the future scope of the present work.

[35] 3) A major limitation of SVM regression is that the outputs are point estimates. It is not possible to derive the conditional distribution of predicted variable given input, and hence quantification of uncertainty in prediction is not possible [*Tipping*, 2001]. Recent developments of Relevance Vector Machine (RVM) [*Tipping*, 2001] based on Bayesian

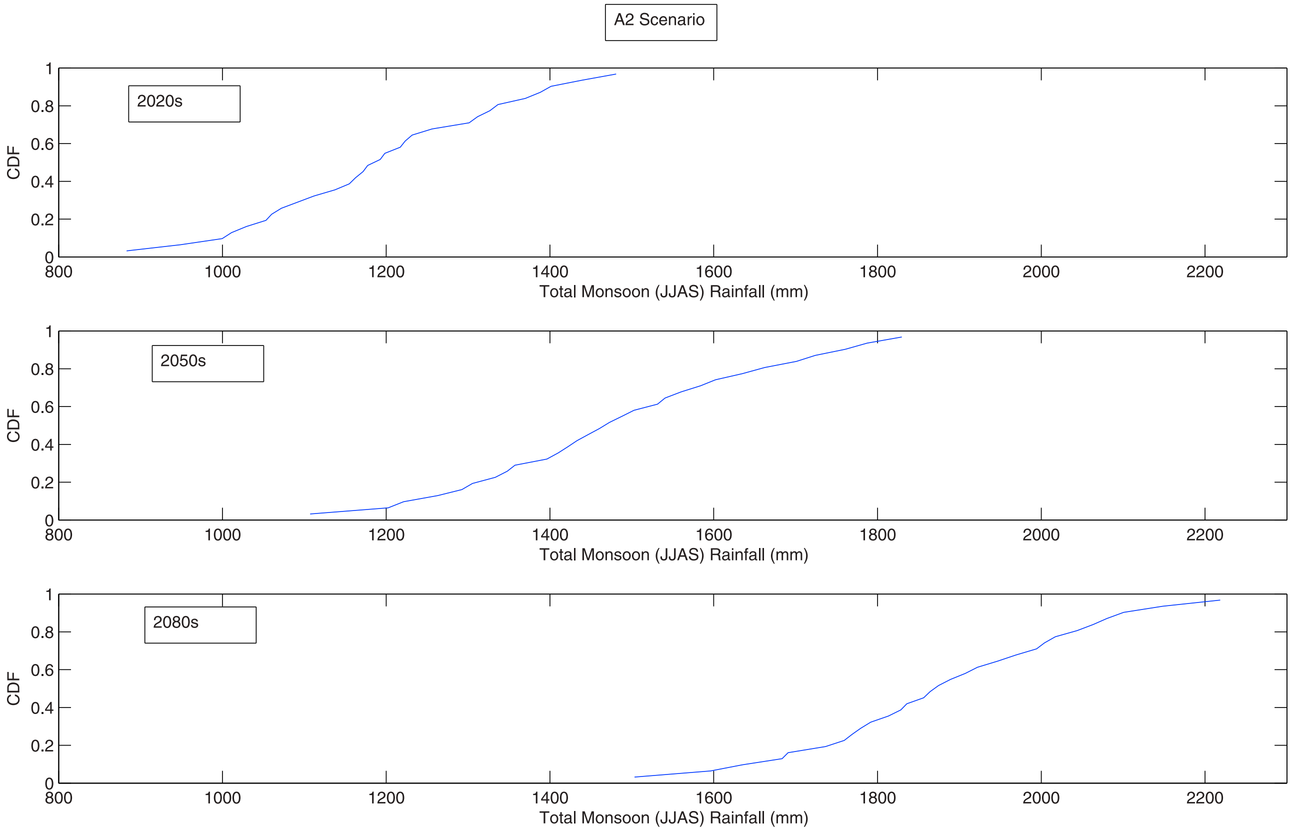

Figure 13. Weighted mean CDF of monsoon rainfall in Assam and Meghalaya meteorological subdivision for A2 scenario.

algorithm overcomes this limitation, and coupling of PGSL with RVM is the potential area of future research.

[36] Finally GCMs are still (AR4) not good for modeling climate and meteorological variables involved in tropical monsoon. Use of more reliable Assessment Report 5 (AR5), GCM simulated data, as predictors for downscaling, may lead to better projections of rainfall in Indian meteorological subdivisions.

Appendix A: Support Vector Regression

[37] Given training data $\{(x_1, y_1), \ldots, (x_l, y_l), X \in \Re^n, Y \in \Re\}$, the Support Vector (SV) regression equation may be given by equation (1).

[38] The basic concept of Support Vector (SV) regression is discussed in this section first with a linear model and then it is extended to a nonlinear model using kernels. Given a training data $\{(x_1, y_1), \ldots, (x_l, y_l), X \in \Re^n, Y \in \Re\}$, the linear model SV regression equation can be given by the following [*Smola*, 1996]:

$$f(x) = \langle w, x \rangle + b, \tag{A1}$$

where, $\langle .,. \rangle$ denoted the dot product in X. The loss function considered for SVM is an ε-insensitive loss function described as follows:

$$|\xi|_\varepsilon = |y - f(x)|_\varepsilon = \begin{cases} 0 & \text{if} \quad |y - f(x)| \le \varepsilon; \\ |y - f(x)| - \varepsilon & \text{otherwise} \end{cases} \tag{A2}$$

[39] The objective of SVM regression is to find the function $f(x)$ with minimum value of loss function and at the same time is as flat as possible [*Smola and Schoelkopf*, 1998]. Flatness mathematically denotes the smaller value of w, and one way to ensure this is to minimize the norm, i.e., $\|w\|^2 = \langle w, w \rangle$. Thus the model can be expressed as the following convex optimization problem:

$$\text{Minimize } \frac{1}{2} \| w \|^2 + C \left(\sum_i^l \xi_i{}^* + \sum_{i=1}^l \xi_i \right) \tag{A3}$$

subject to

$$y_i - \langle w, x \rangle - b \le \varepsilon + \xi_i \tag{A4}$$

$$\langle w, x \rangle + b - y_i \le \varepsilon + \xi_i{}^* \tag{A5}$$

$$\xi_i, \xi_i{}^* \ge 0, \tag{A6}$$

where C is a prespecified value which determines the trade-off between the flatness of $f(x)$ and the amount up to which deviations larger than ε are tolerated (ξ_i and $\xi_i{}^*$), which correspond to ε-insensitive loss function as presented in equation (A2). The optimization model presented in equations (A3)–(A6) can be solved using Lagrange multi-

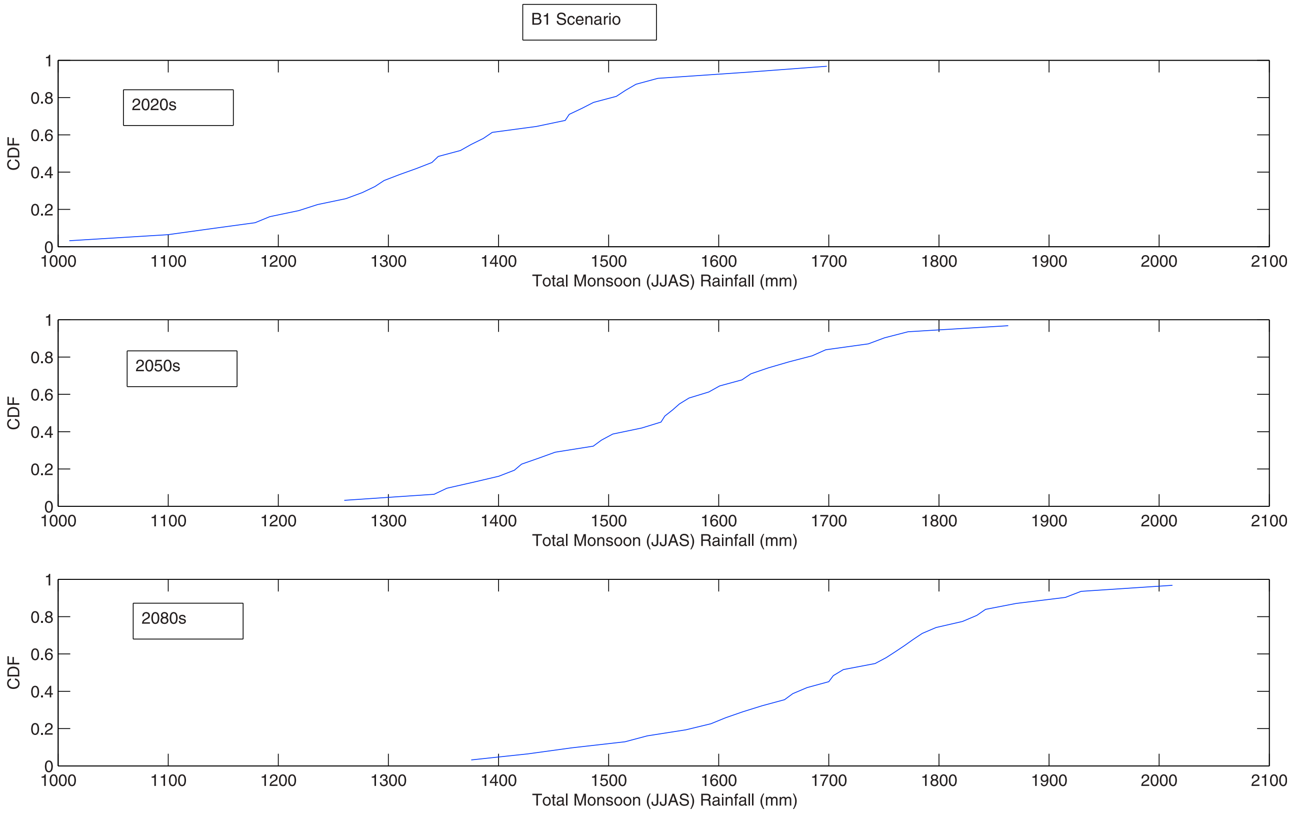

Figure 14. Weighted mean CDF of monsoon rainfall in Assam and Meghalaya meteorological subdivision for B1 scenario.

pliers. A dual set of variables are introduced to construct the Lagrange function, which is given below:

$$
\begin{aligned}
L = \frac{1}{2}\parallel w \parallel^2 + C\left(\sum\nolimits_{i=1}^{l} \xi_i^* + \sum\nolimits_{i=1}^{l} \xi_i\right) - \sum\nolimits_{i=1}^{l} (\eta_i \xi_i + \eta_i^* \xi_i^*) \\
- \sum\nolimits_{i=1}^{l} \alpha_i(\varepsilon + \xi_i - y_i + \langle w, x\rangle + b) \\
- \sum\nolimits_{i=1}^{l} \alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x\rangle - b)
\end{aligned}
\tag{A7}
$$

where L is the Lagrangian and η_i, η_i^*, α_i, α_i^* are Lagrangian multipliers satisfying the positivity constraints.

$$\eta_i, \eta_i^*, \alpha_i, \alpha_i^* \geq 0 \tag{A8}$$

[40] From the saddle point condition, the partial derivatives of L with respect to the primal variables (w, b, ξ_i, ξ_i^*) have to vanish for optimality.

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) = 0 \tag{A9}$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) x_i = 0 \tag{A10}$$

$$\frac{\partial L}{\partial \xi_i^{(*)}} = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0 \tag{A11}$$

where $\xi_i^{(*)}$, $\alpha_i^{(*)}$, $\eta_i^{(*)}$ refer to ξ_i and ξ_i^*; α_i and α_i^*; η_i and η_i^* respectively.

[41] Substituting equations (A9)–(A11) in equation (A7) the following dual optimization problem is formulated.

$$
\begin{aligned}
\mathit{Maximize} \; - \frac{1}{2}\sum\nolimits_{i,j=1}^{l} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle \\
- \; \varepsilon \sum\nolimits_{i=1}^{l} (\alpha_i + \alpha_i^*) + \sum\nolimits_{i=1}^{l} y_i(\alpha_i - \alpha_i^*)
\end{aligned}
\tag{A12}
$$

subject to

$$\sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0 \tag{A13}$$

$$\alpha_i, \alpha_i^* \in [0, C]. \tag{A14}$$

Equation (A10) can be rewritten as follows:

$$w = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) x_i \tag{A15}$$

and thus from equation (A1):

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b \tag{A16}$$

[42] This is called the Support Vector Expansion for linear model which is used in SV regression. b can be computed by using Karush Kuhn Tucker (KKT) condition [*Smola and Schoelkopf*, 1998].

[43] For downscaling model linear regression may not be appropriate, and therefore a nonlinear mapping using kernel K is performed to map the input data into a higher dimensional feature space. Using the kernel, the regression equation (equation (A16)) can be modified to equation (1) and the same algorithm can be used to compute the weights and bias. Details of kernel functions are presented in the next subsection.

A1. Kernel Functions

[44] Kernel functions are used in SVM for nonlinear mapping of the original data or input into a high-dimensional feature space. Kernel function used in a SVM should follow Mercer's theorem, according to which it can be written that:

$$\int_{X\times X} K(x,x')f(x)f(x')dxdx' \geq 0 \quad \forall f \in L_2(X) \tag{A17}$$

Some of the valid kernel functions satisfying the above mentioned condition are given below.

[45] 1) Linear kernel: The linear kernels are the simplest kernels used in SVM for linear regression. They can be given by:

Homogeneous Kernel:

$$K(x,x') = \langle x,x' \rangle \tag{A18}$$

Nonhomogeneous Kernel:

$$K(x,x') = (\langle x,x' \rangle + 1) \tag{A19}$$

[46] The performance of SVM with linear kernel function, being similar to that of linear regression, is not capable of modeling complicated and nonlinear relationship between climatological variables and monsoon rainfall, and therefore such kernels are not used in the present study.

[47] 2) Radial basis functions: Radial Basis Functions (RBFs) have received significant attention because of their excellent performance in capturing nonlinear relationship. A generalized RBF [*Chapelle et al.*, 1999] can be given by:

$$K(x,x') = \exp\left(-\frac{\parallel x^a - x'^a \parallel^b}{2\sigma^2}\right) \tag{A20}$$

where σ is the width of RBF kernel, giving an idea about the smoothness of the derived function. A large kernel width acts as a low-pass filter in frequency domain, attenuating higher-order frequencies and thus resulting in a smooth function. Alternatively, RBF kernel with small kernel width retains most of the higher-order frequencies, leading to an approximation of a complex function by learning machine [*Smola and Schoelkopf*, 1998]. An RBF will satisfy Mercer's condition if and only if $0 \leq b \leq 2$. The choice of a has no impact on Mercer's condition. Conventionally, a is selected as 1 as it does not have significant impacts on training performance [*Chapelle et al.*, 1999]. Different values of b, i.e., b = 0.5, 1 and 2 denote heavy-tailed RBF, Laplacian RBF and Gaussian RBF.

[48] **Acknowledgments.** The author sincerely thanks editor Dr. Steve Ghan and the anonymous reviewers for reviewing the manuscript and providing critical comments which have significantly improved the quality of contribution. The work reported in this paper was carried out as a part of the projects 07IR040 sponsored by IRCC, Indian Institute of Technology Bombay, and SR/FTP/ETA-16/08, sponsored by the Department of Science and Technology, India.

References

Aihong, X., R. Jiawen, Q. Xiang, and K. Shichang (2007), Reliability of NCEP/NCAR reanalysis data in the Himalayas/Tibetan Plateau, *J. Geographical Sci.*, *17*(4), 421–430.

Chapelle, O., P. Haffner, and V. N. Vapnik (1999), Support vector machines for histogram-based image classification, *IEEE Trans. Neural Networks*, *10*(5), 1055–1064.

Conway, D., R. L. Wilby, and P. D. Jones (1996), Precipitation and sir flow indices over British Isles, *Clim. Res.*, *7*, 169–183.

Crane, R. G., and B. C. Hewitson (1998), Doubled co_2 precipitation changes for the susquehanna basin: down-scaling from the genesis general circulation model, *Int. J. Climatol.*, *18*, 65–76.

Dankers, R., O. B. Christensen, L. Feyen, and M. Kalas (2007), Evaluation of very high-resolution climate model data for simulating flood hazards in the Upper Danube Basin, *J. Hydrol.*, *347*, 319–331.

Dibike, Y. B., S. Velickov, D. Solomatine, and M. B. Abbott (2001), Model induction with support vector machines: introduction and applications, *J. Computing Civil Eng.*, *15*(3), 208–216.

Domer, B., B. Raphael, K. Shea, and I. F. C. Smith (2003), A Study of Two Stochastic Search Methods for Structural Control, *J. Computing Civ. Eng.*, *17*(3), 132–141.

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.

Ghosh, S., and P. P. Mujumdar (2009), Climate change impact assessment: Uncertainty modeling with imprecise probability, *J. Geophys. Res.*, *114*, D18113, doi:10.1029/2008JD011648.

Giorgi, F., and L. O. Mearns (2002), Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "Reliability Ensemble Averaging" (REA) method, *J. Clim.*, *15*(10), 1141–1158.

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.

Govindaraju, R. S. (2005), Bayesian learning and relevance vector machines for hydrologic applications, In: *2nd Indian International Conference on Artificial Intelligence (IICAI-05)*, Pune, India.

Gunn, S. R., M. Brown, and K. M. Bossley (1997), Network performance assessment for neuro fuzzy data modelling, In: *Intelligent Data Analysis*, ed. By X. Liu, P. Cohen and M. Berthold, Lecture Notes in Computer Science, 1208, 313–323.

Hewitson, B. C., and R. G. Crane (1992), Large-scale atmospheric controls on local precipitation in tropical Mexico, *Geophys. Res. Lett.*, *19*(18), 1835–1838.

Hewitson, B. C., and R. G. Crane (1996), Climate downscaling: Techniques and application, *Clim. Res.*, *7*, 85–95.

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.

IPCC (2007), *Climate Change 2007 The physical science basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Ed. by S. Solomon, D. Qin, M. Manning, M. Marquis, K. Averyt, M. M. B. Tignor, H. L. R. Miller Jr. and Z. Chen, Cambridge University Press, Cambridge, UK.

Jain, S. K., and K. P. Sudheer (2008), Fitting of hydrologic models: A close look at the Nash-Sutcliffe Index, *J. Hydrol. Eng.*, *13*(10), 981–986.

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

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

Karamouz, M., B. Zahraie, and S. Araghinejad (2005), Decision support system for monthly operation of hydropower reservoirs: A case study, *J. Comp. Civ. Eng.*, *19*(2), 194–207.

Krause, P., D. P. Boyle, and F. Base (2005), Comparison of different efficiency criteria for hydrological model assessment, *Adv. Geosci.*, *5*, 89–97.

Ma, L., T. Zhang, O. W. Frauenfeld, B. Ye, D. Yang, and D. Qin (2009), Evaluation of precipitation from the ERA-40, NCEP-1, and NCEP-2 Reanalyses and CMAP-1, CMAP-2, and GPCP-2 with ground-based measurements in China, *J. Geophys. Res.*, *114*, D09105, doi:10.1029/2008JD011178.

Mujumdar, P. P., and S. Ghosh (2008), Modeling GCM and scenario uncertainty using a possibilistic approach: Application to the Mahanadi River, India, *Water Resour. Res.*, *44*, W06407, doi:10.1029/2007WR006137.

Nagesh Kumar, D., K. Srinivasa Raju, and T. SathishRiver (2004), Flow forecasting using recurrent neural networks, *Water Resour. Manage.*, *18*(2), 143–161.

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

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.

Raphael, B., and I. F. C. Smith (2000), A probabilistic search algorithm for finding optimally directed solutions, *Proceedings of Construction Information Technology 2000, Icelandic Building Research Institute, Reykjavik*, 708–721.

Raphael, B., and B. Smith (2003), A direct stochastic algorithm for global search, *Applied Mathematics and Computation*, *146*(3), 729–758.

Smola, A. J. (1996), *Regression Estimation with Support Vector Learning Machines*, Technische Universitat Munchen, Munich, Germany.

Smola, A. J., and B. Schoelkopf (1998), A tutorial on support vector regression, *NeuroCOLT2 Technical Report NC2-TR-1998-030*, Royal Holloway College, University of London, UK.

Suykens, J. A. K. (2001), Nonlinear modelling and support vector machines, In: *Proceedings of IEEE Instrumentation and Measurement Technology Conference*, Budapest, Hungary, 287–294.

Svanerudh, P., B. Raphael, and I. F. C. Smith (2002), Lowering costs of timber shear-wall design using global search, *Eng. Comput.*, *18*, 93–108.

Tao, K., and A. P. Barros (2010) Using fractal downscaling of satellite precipitation products for hydrometeorological applications. *J. Atmos. Oceanic Technol.*, *27*, 409–427 doi:10.1175/2009JTECHA1219.1.

Tatli, H., H. N. Dalfes, and S. Mentes (2004), A statistical downscaling method for monthly total precipitation over Turkey. *Int. J. Climatol.*, *24*(2), 161–180.

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.

Tipping, M. E. (2001), Sparse Bayesian learning and the relevance vector machine, *J. Machine Learning Res.*, *1*, 211–244.

Trigo, R. M., and J. P. Palutikof (1999), Simulation of daily temperatures for climate change scenarios over portugal: A neural network model approach, *Clim. Res.*, *13*, 45–59.

Tripathi, S., and V. V. Srinivas (2005), Downscaling of general circulation models to assess the impact of climate change on rainfall of India, *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.

Vapnik, V. N. (1995), *The Nature of Statistical Learning Theory*, Springer Verlag, New York.

Vapnik, V. N. (1998), *Statistical Learning Theory*, Wiley, New York.

Venugopal, V., E. Foufoula-Georgiou, and V. Sapozhnikov (1999), A Space time downscaling model for rainfall, *J. Geophys. Res.*, *104*(D16), 19,705–19,721.

Vrac, M., D. Paillard, and P. Naveau (2007), Non-linear statistical downscaling of present and LGM precipitation and temperatures over Europe, *Clim. Past Discuss.*, *3*, 899–933.

Wetterhall, F., S. Halldin, and C. Xu (2005), Statistical precipitation downscaling in central sweeden 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 Mannual*.

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., 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., 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)*.(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.

Wilks, D. S., and R. L. Wilby (1999), The weather generation game: a review of stochastic weather models, *Prog. Phys. Geog.*, *23*(3), 329–357.

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. Cartographer*, *12*, 5–16.

Xue, Y., R. Vasic, Z. Janjic, F. Masinger, and K. Mitchell (2007), Assessment of Dynamic Downscaling of the Continental U.S. Regional Climate Using the Eta/SSiB Regional Climate Model, *J. Clim.*, *20*, 4172–4193.

Zorita, E., J. P. Hughes, D. P. Lettenmaier, and H. von Storch (1995), Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation, *J. Clim.*, *13*, 223–234.

S. Ghosh, Assistant Professor, Department of Civil Engineering, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India. (subimal@civil.iitb.ac.in)