METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **19**: 179–188 (2012)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/met.1333

Prediction of Indian summer monsoon rainfall (ISMR) using canonical correlation analysis of global circulation model products

Ankita Singh,[a]* Makarand A. Kulkarni,[b] U. C. Mohanty,[a] S. C. Kar,[c] Andrew W. Robertson[d] and G. Mishra[e]

[a] *Centre for Atmospheric Sciences, Indian Institute of Technology, New Delhi, India*
[b] *Department of Atmospheric and Space Sciences, University of Pune, India*
[c] *National Center for Medium Range Weather Forecasting, Noida, India*
[d] *International Research Institute for Climate and Society, New York, NY, USA*
[e] *Department of Statistics, Utkal University, Bhubaneshwar, India*

ABSTRACT: The Canonical Correlation Analysis (CCA) method has been used in this study for improving General Circulation Model (GCM) predicted rainfall over India during the southwest monsoon season. Hindcast runs for 27 years (1982–2008) from six GCM outputs are used. This statistical technique relates the pattern of multivariate predictor field (model rainfall) to the pattern of predictand fields (observed rainfall). It is found that the CCA method improves the skill of three of the GCMs at the all-India level. A noticeable improvement is also observed in the composite prediction with CCA as compared to the simple mean of raw GCM products. The skill of the composite prediction after applying CCA is higher compared to the simple mean of raw model products in several homogeneous zones such as the hilly areas, west central area and over some parts of northwest India. The possible reason for the improvement in the skill of some of the GCMs may be the similarity between the loading patterns of model predictions and the observed rainfall. Copyright © 2012 Royal Meteorological Society

KEY WORDS Canonical correlation analysis; prediction; General Circulation Model; season; rainfall

Received 23 September 2011; Revised 2 March 2012; Accepted 23 April 2012

1. Introduction

The southwest monsoon is the principal rainy season in India which contributes to about 80% of the annual rainfall. Some statistical characteristics of the southwest monsoon rainfall have been evaluated in previous studies, e.g. Parthasarathy *et al.* (1995) and Pai and Rajeevan (2007). The inter-annual standard deviation of the all Indian monsoon rainfall is about 84.7 mm with a mean of 852.4 mm: thus, a small variation affects the Indian agriculture to a large extent and impacts on the social and economic condition of the country. Therefore, it is important to develop a better forecast methodology which will provide an estimation of the monsoon rainfall in advance (monthly to seasonal scale) for planning purposes.

Several statistical models have been developed in the past for the long range prediction of the Indian monsoon rainfall ever since Walker (1924) and Gowariker *et al.* (1989). These models have been developed taking into account teleconnections of several atmospheric variables with the Indian summer monsoon rainfall. The India Meteorological Department (IMD) provides the forecast for seasonal rainfall in advance in two phases. The first phase forecast is issued in the middle of April using an eight parameter regression model. It is updated around the end of June using a 10 parameter regression model (Rajeevan and McPhaden, 2004). Presently, the ensemble multiple linear regression (EMR) and projection pursuit regression (PPR) are used for the long range prediction of summer monsoon rainfall (Rajeevan *et al.*, 2007). All these statistical models have certain limitations due to their dependence on the interrelationship of variables which should remain the same for future (Rajeevan, 2001). However, some studies, such as Mooley and Munot (1993) and Krishna Kumar *et al.* (1999), have examined the variation of relationships between some global variables with the Indian monsoon rainfall and showed that the relationship changes with time. In spite of all these efforts the seasonal prediction of monsoon rainfall has remained a challenging task for forecasters.

After the availability of state-of-the-art General Circulation Models (GCMs), various studies have been conducted for forecasting the Indian summer monsoon rainfall using GCMs (e.g. Kang *et al.*, 2004; Krishna Kumar *et al.*, 2005; Sahai *et al.*, 2008; Pattanaik and Kumar, 2010; Acharya *et al.*, 2011; Janakiraman *et al.*, 2011; Kar *et al.*, 2011). Preethi *et al.* (2010) have analysed the skill of some of the coupled models for the hind-cast run from 1959 to 1979 and found the skill of the models to be positive. It is generally noted that the GCMs have large variations in simulating the observed climatology and the inter-annual variability (evaluated in terms of standard deviation). Recently, the performances of some of the coupled and atmospheric models have also been evaluated by Acharya *et al.* (2011). It was shown that although the GCM-simulated values of both of the monsoon rainfall climatology and inter annual variability (IAV) at the all India level are in good agreement with the observed values, the models underestimate the IAV at the homogeneous zones as compared to the observed IAV. Uncertainty to the initial conditions, the non linearity in atmospheric dynamics and model errors contribute to the inaccuracy of the GCM forecast products. Therefore,

*Correspondence to: A. Singh, Centre for Atmospheric Sciences, Indian Institute of Technology, Delhi, New Delhi-110016.
E-mail: ankita.stats@gmail.com

the use of direct GCM model output may not be appropriate (Sahai and Chattopadhyay, 2006). Therefore, various methods of post processing should be used with these GCM outputs before making a seasonal prediction. Some of the statistical post-processing methods are multi-model ensemble (MME), principal component regression (PCR) and canonical correlation analysis (CCA). Krishnamurti *et al.* (2000, 2006), Sahai *et al.* (2008), and Acharya *et al.* (2011) have used the MME for the seasonal prediction.

Canonical correlation analysis (Hotelling, 1936) is defined as a multivariate statistical technique which relates the pattern of the multivariate predictor field to the pattern of predictand field. In other words, it finds a set of linear combinations of data sets which are highly correlated. This property of CCA can be used to correlate the pattern of GCM rainfall product with the pattern of observed rainfall, which can be used for prediction purposes. Earlier, CCA has been used for monthly as well as seasonal prediction by Barnett and Preisendorfer (1987) and Barnston (1994). Further, the technique was used by many forecasters for the long range prediction of sea surface temperature and precipitation all over the globe (e.g. Barnston and Smith, 1996). Yu *et al.* (1997) have used both CCA and PCR for the prediction of rainfall fluctuations and have used Pacific SST as a predictor because it provides a good estimate of seasonal climate variations. Further, they have found the usefulness of both the models by analyzing their high skill for the winter months. This analysis was also used for the operational long lead forecast of South African rainfall (Landman and Mason, 1999). CCA can also be used to study predictability of rainfall extremes (Landman *et al.*, 2005). This analysis has been applied to the GCM forecasts for SST in order to correct the model biases (Tippett *et al.*, 2005). Further, Lim *et al.* (2011) introduced regularized CCA and implemented it for the predictions of precipitation over East Asia. They found that the forecast obtained from this method was more skilful compared to the results from the GCMs.

The potential of the CCA method to correct the model forecasts over the Indian monsoon region has not been fully exploited. Prasad and Singh (1996) have used this analysis to estimate the monsoon rainfall over 29 meteorological subdivisions using global variables such as the 500 hPa ridge axis position in April and the Darwin surface pressure tendency. Recently, Sinha *et al.* (pers. comm., 2011) have used the CCA technique to improve seasonal forecasts by developing model output statistics (MOS) which consider several meteorological variables from a global model as predictors.

Hagedorn *et al.* (2005) have described the rationale behind the success of multi-model ensembles. The skill of the simple arithmetical average of all the model products tends to yield higher skill than the skill of individual participating models. Kang *et al.* (2004) have examined the systematic errors of a dynamical seasonal prediction system and estimated potential predictability of summer mean precipitation with correction of such systematic errors. Kar *et al.* (2006) have used several multi-model approaches to estimate the economic values of the forecasts and have found that the multi-model ensemble schemes improve the value of the forecasts over the single model. Kug *et al.* (2008) have described the skill of several MME methods for seasonal prediction and proposed a step-wise pattern projection scheme for MME. The performance of multi-model techniques for precipitation forecasting over India have been examined in some recent studies (e.g. Chakraborty and Krishnamurti, 2009). However, the success of such methods in monsoon rainfall prediction is very limited (Kar *et al.*, 2011). If the skill of individual models is improved through some statistical post processing method then it is expected that skill of forecasts of the composite made out of such improved forecasts (MME of improved forecasts) shall be higher than individual improved forecasts or the MME of raw model forecasts. There has been no study of the application of the CCA technique to improve the Indian monsoon rainfall forecasts from GCMs and no attempt has been made to document the skill of such improved forecasts after applying MME.

Therefore, the main objective of the present study is to use the CCA technique to improve the forecast skill of a set of GCM products individually and then to estimate the skill of composite forecasts prepared using such CCA improved forecasts. Comparative skill assessment has also been made between such composite forecasts and MME forecasts made using raw model products. The region-wise impact of the seasonal forecast is also important, along with the high resolution gridded forecast. Therefore, in the present study an attempt has also been made to evaluate the skill of the CCA improved forecast at homogeneous rainfall regions (Parthasarathy *et al.*, 1995) along with the 1°× 1° grid boxes. Section 2 of this article consists of a brief description on the data and methodology. The results and related discussion are elaborated in Section 3. Finally, the conclusion of the study is presented in Section 4.

2. Data and methodology

2.1. Model and observed data

The lead-1 prediction of precipitation for monsoon season (June to September, JJAS) with May start (model runs use observations up to May 1) of six GCMs have been used in this study. The prediction products used are from 1982 to 2008 and all the GCM hindcasts are collected from the data library of the International Research Institute for Climate and Society (IRI), Columbia University, USA. The GCMs used in this study are now briefly introduced. The two fully coupled versions of IRI models, referred to as MOM3AC1 and MOM3DC2, have the European Centre-Hamburg Model (ECHAM version 4.5) as the atmospheric component coupled with the Modular Ocean Model (version 3). MOM3AC1 is anomaly-coupled, while MOM3DC2 is direct-coupled. IRI's mixed layer coupled model, referred to as ECHGML (Roeckner *et al.*, 1996; Pacanowski and Griffes, 1998), has also been used. The fourth model, referred to as CFS, is the National Center for Environmental Prediction (NCEP)'s climate forecast system version-1 model (Saha *et al.*, 2006). The last two models, referred to as ECHcasst and ECHcfssst, are two two-tier versions of ECHAM4.5 forced with constructed analog SST and CFS forecast SST (Li and Goddard, 2005), respectively. All these models are also discussed in detail in Kar *et al.* (2011). The number of ensemble members and spatial resolution of these models are given in Table 1.

The predictand field is the observed rainfall values for the summer monsoon period. These observed values are extracted over the extended Indian domain from 10 °S to 50 °N and 50 °E to 120 °E. Observed rainfall data over India are taken from the IMD at 1° × 1° resolution (Rajeevan *et al.*, 2006). It may be noted here that there are several gridded rainfall datasets available at 0.5°, such as the University of Delaware rainfall data (Matsuura and Willmott, 2009) and the data sets from the Climatic Research Unit (CRU) archive (Mitchell and Jones, 2005). Many of the same observational (rain-gauge) records

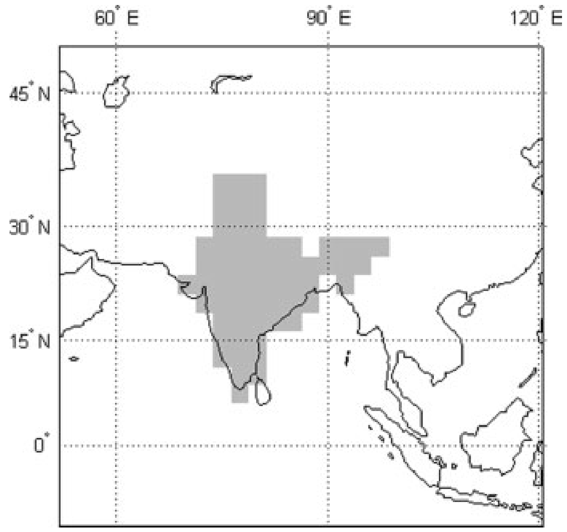

Figure 1. Selected domain for the analysis (grey shaded are the Indian land points for which data are taken from the IMD and merged with CMAP data for the white shaded area).

Table 1. The Global Circulation Models used in the study.

Serial no.	Model	Ensemble members	Resolution	Type
1	CFS	15	T62	Fully coupled
2	GML	12	T42	Semi-coupled
3	MOM3AC1	24	T42	Anomaly coupled
4	MOM3DC2	12	T42	Direct coupled
5	ECHcasst	24	T42	Two-tier
6	ECHcfssst	24	T42	Two-tier

are used in these data sets. However, each data set is not based on exactly the same set of rain-gauge records. In these datasets, the number of rain gauge station data used from India is very limited. In contrast, the 1° × 1° Indian rainfall data (Rajeevan *et al.*, 2006) used in the present study considers 1803 stations which had a minimum of 90% of data availability during the analysis period 1951–2003. This rainfall data set was compared with other similar global gridded rainfall data sets and this dataset better represents rainfall over the Indian region. The correlation co-efficients between this rainfall time series and other global data sets are more than 0.80 (Rajeevan *et al.*, 2006). The observed data outside of India are taken from CPC (Climate Prediction Centre) Merged Analysis of Precipitation (CMAP) estimated precipitation (Xie and Arkin, 1995) which is available at 2.5° × 2.5°. Further, these values are merged to obtain the rainfall observation over the entire domain shown in Figure 1. In the figure, the IMD observed values are obtained for the grey shaded area and the CMAP data is merged for the remaining part of the domain. The GCM precipitation outputs are also extracted for the above selected domain.

2.2. Canonical correlation analysis (CCA)

In the CCA technique, the multivariate predictors (patterns) are linearly related to the multivariate predictands (patterns), i.e. a set of weights for predictors are linearly related to the set of weights for predictand. These weighting sets are called the loading pattern for predictors and predicands respectively. These patterns may represent the physical processes. CCA is also known as a specialized version of an empirical orthogonal function analysis (EOF) in which the correlation matrix of predictor and predictand is analysed (Barnston and Smith, 1996). Each of the successive CCA modes defines more completely the correlation structure between predictor and predictand. It contains the patterns of canonical variables (predictor and predictand) having maximum correlation. It may be noted here that in the traditional manner the predictors and predictands are augmented to obtain the CCA loadings. However, some earlier studies such as Barnston and Smith (1996) and Yu *et al.* (1997) have shown that it is possible to estimate the CCA loadings using principal components of predictors and predictands instead of using the full dataset. This makes finding the CCA rather easy and also gives a way of choosing the data based on the explained variance. This procedure and not the extended empirical orthogonal function (EOF) analysis method used in Singh and Kripalani (1986) has been used in the present study. Details about the CCA are available in Graham *et al.* (1987) and Wilks (1995). In the present study, rainfall from the six GCMs is the predictor in the CCA and each model enters in the analysis separately.

2.2.1. *Pre-orthogonalization and standardization of data*

Before performing the CCA, the data sets (predictor and predictand) have been properly standardized and orthogonalized separately (Barnston and Smith, 1996; Yu *et al.*, 1997). This step is essential in cases (such as here) where the length of the historical record is smaller than the dimensions of the predictor and predictand fields, and some regularization method is necessary to invert the singular predictor covariance matrix (Tippett *et al.*, 2003). The orthogonalization compresses the datasets using the concept of standard EOF analysis as it reduces the large number of spatial dimension into a smaller number which explains maximum variability within that variable. Secondly, the EOF analyses also filter out the incoherent variability (noise) as only a few EOF modes are retained. The EOF analysis is thus performed on each of the predictors (GCM outputs) and predictand. All preprocessing is done for all the model data separately (similarly for the predictand). As there is no universally agreed upon procedure for determining how many EOF modes should be retained (Yu *et al.*, 1997), in this study the mode truncation has been selected such that maximum variance explained by the chosen number of modes is 85% for the predictor and 70% for predictand. Henceforth, the leading 11 or 12 modes for the predictor and 12 modes for predictand are used for further analysis. These selected EOFs (temporal) are then cross-correlated. Further, the cross-covariance matrix for predictor predictand is the input for the CCA.

Assuming that $X_{m,t}$, $Y_{o,t}$ are the matrices for predictor and predictand where m, o are the number of grid boxes for the predictor and predictand, respectively, and t is the number of years, separate EOFs $E_{m,m}$ and $E_{o,o}$ are evaluated for X and Y as:

$$X_{m,t} = E_{m,m}T_{m,t} \tag{1}$$

$$Y_{o,t} = E_{o,o}T_{o,t} \tag{2}$$

In equations (1) and (2), $E_{m,m}$ and $E_{o,o}$ are the spatial modes, whereas $T_{m,t}$ and $T_{o,t}$ are the corresponding time co-efficients (Yu *et al.*, 1997). Assume that i and j are the retained EOF modes for the predictor and predictand variable. The canonical variables Z and W are defined as the linear combination of canonical vectors U and V:

$$Z = u'T_{i,t}, \text{ and } W = v'T_{j,t} \tag{3}$$

The predicted value of the canonical predictand following Wilks (1995) is defined as:

$$\hat{W} = A_{q,q}Z \tag{4}$$

Here, $A_{q,q}$ is the diagonal matrix of correlation between the canonical variables known as the matrix of canonical correlation of order $q \times q$, where $q = minimum\ (i, j)$. The predicted value of the original predictand, that is the rainfall observed at o spatial points, is obtained by using the property of EOFs and the inverse transformations predicted rainfall value for time $t + l$ is given below where l is the lead time:

$$\hat{Y}_{o,t+l} = E_{o,j}(V')^{-1}A_{q,q}U'(E_{m,i})'X_{m,t+l} \tag{5}$$

A brief derivation of the prediction equation used here is available in Yu *et al.* (1997). The predicted values of rainfall over the interested domain are obtained using each GCM output separately as explained above. After carrying out the statistical post processing of each GCM product using the CCA technique composite of all corrected predictions is obtained.

3. Results and discussion

The detailed discussion of the results regarding the developed CCA model is divided in the forthcoming sub-sections. Firstly, the canonical patterns are discussed in detail with the canonical component time series. As discussed earlier, the models have large variability in the simulation of ISMR. Therefore, in the present study the post-processing of the GCMs is done using the CCA technique. The post-processed GCMs are further combined to obtain a single predicted series for further analysis and the evaluation. The skill assessment of the developed CCA model is done at each grid point as well as at the regional level in leave-one-out cross validation mode.

3.1. CCA patterns

It is well known that the predictand loadings will change with the choice of predictor. The canonical loading patterns for mode-1 of different predictors (models) are shown in Figure 2. Plots 2(a) to 2(d) show the loading patterns for the coupled models used in this study: plots 2(e) and 2(f)

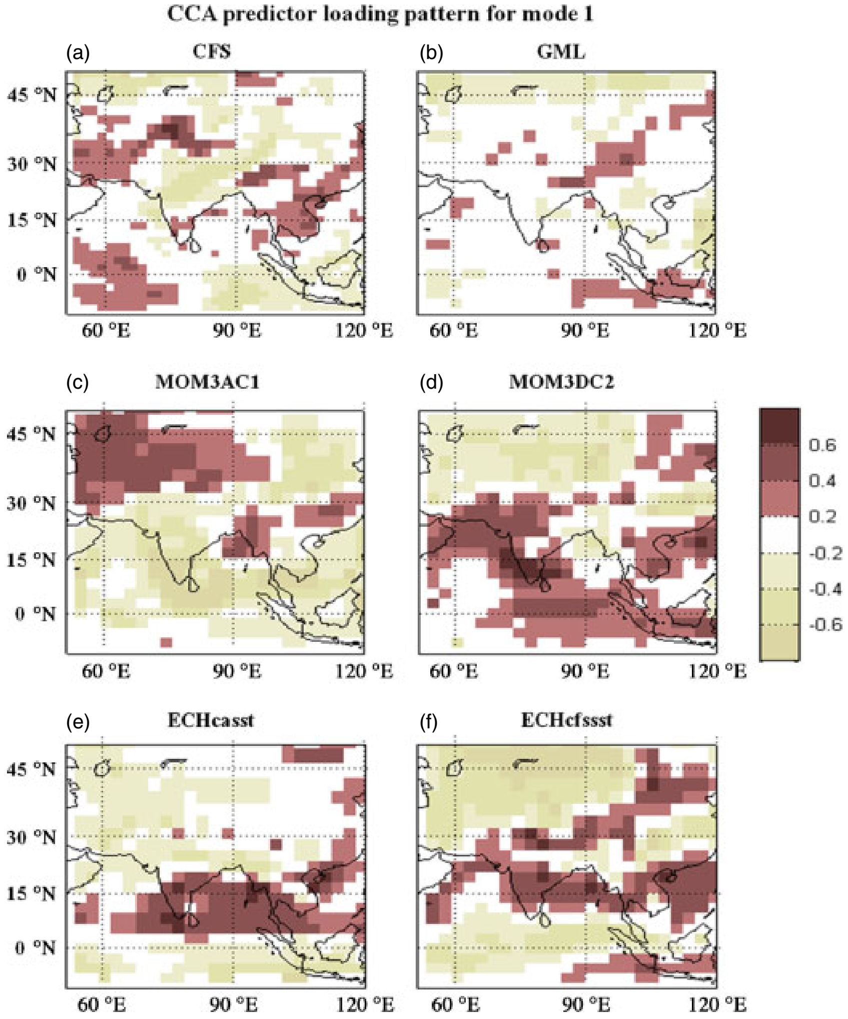

Figure 2. CCA predictor loading pattern for mode 1. Plots (a) to (f) are the pattern for the different GCM outputs. This figure is available in colour online at wileyonlinelibrary.com/journal/met

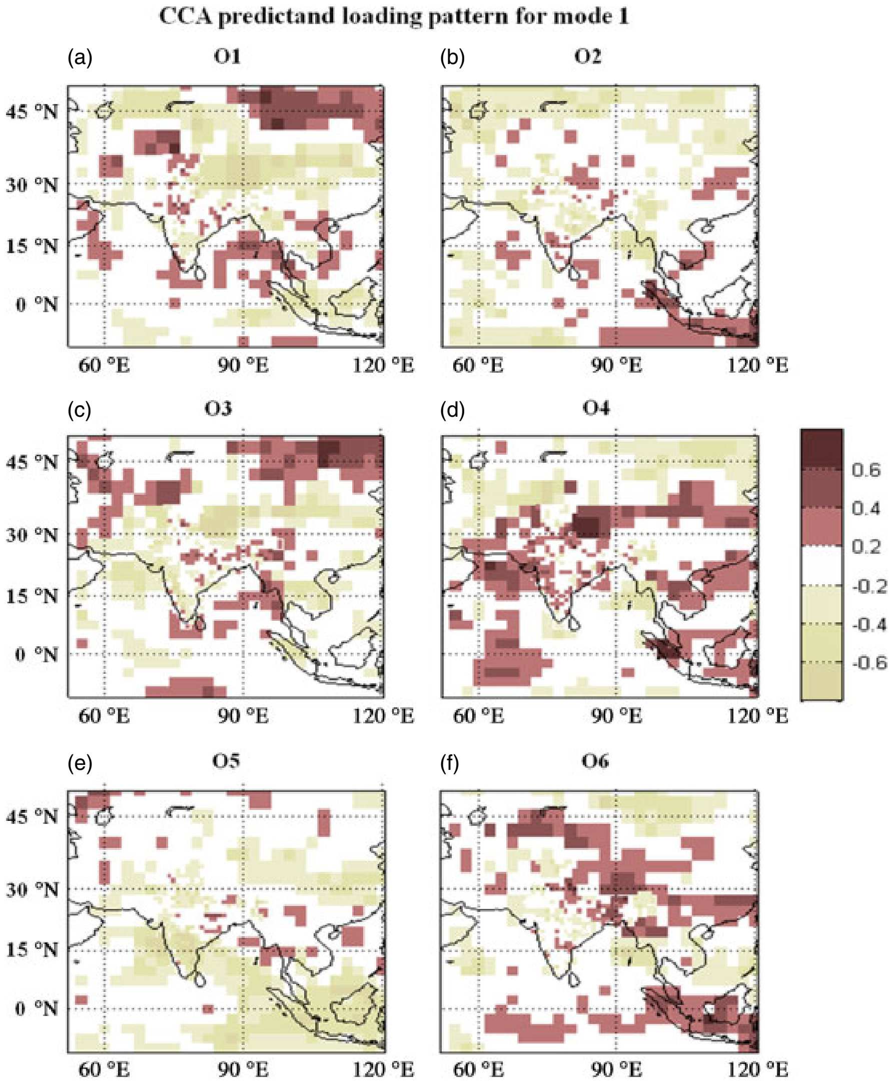

Figure 3. CCA predictand loading pattern for mode 1. Plots (a) to (f) are the pattern of predictand loading corresponding the different GCM outputs. This figure is available in colour online at wileyonlinelibrary.com/journal/met

display the two-tier models. The first canonical loading pattern for predictand (observed rainfall) corresponding to different predictors (models) is shown in Figure 3(a)–(f) in the same order as that of Figure 2. It is seen that the loading patterns for predictor and predictand are different for different GCMs. The mode 1 for the coupled models explains about 20–24% of the total variability whereas the leading mode of predictand (Figure 3) explains about 11–12% of the total variability. The details about the percentage variance explained and the retained modes are given in Table 2. In Figure 2(a), the canonical loadings for the coupled model CFS show high negative values over the Indian land points mainly in the central part of India. However, there are some patches of positive values over the hilly areas and some parts of northeast and southern India which show similar kind of variation over the entire country while the northeastern part behaves in a different phase (Kulkarni *et al.*, 1992). Out of phase loadings over the western and eastern equatorial Indian Ocean resemble the rainfall pattern due to the Indian Ocean dipole mode. These predictor loadings are compared with the predictand loadings for the same model. It can be clearly noticed in Figure 3(a) that over some parts of the hilly regions and the west central region the predictor loadings are comparable (i.e. they show similar kinds of rainfall variations over the country) with the predictand loadings, but over some areas such as the central part of India the patterns are opposite. Moreover, over the Indian Ocean (mainly over the western Indian Ocean and the north Bay of Bengal) the loadings for predictor and predictand show patterns of opposite sign. The other coupled GCM, the GML (shown in Figure 2(b)), shows similar loading patterns but their magnitude is lower than that of the CFS. These loadings are also similar to the pattern of the predictand loadings shown in Figure 3(b). The loadings for MOM3AC1 and MOM3DC2 (Figure 3 parts (c) and (d)) show large positive loadings over the Indian Ocean and over the hilly areas, while the predictand loadings show negative values over these areas. Moreover, predictand loadings for this mode are positive over some parts of the central northeast but no such signal is evident for the predictor loadings. MOM3DC2 is able to capture the pattern of the observed loading in some parts of the domain but overestimate their magnitude. The leading modes of the two-tier models explain about 30–34% of the total variability. Both of the atmospheric models show similar kind of loadings over the southern part of India as well as over the Indian Ocean.

Figure 4 shows the time series of the canonical component for all the predictors with the predictand. These time series

Table 2. The variance explained by EOFs and the retained canonical modes for the analysis.

Serial no.	Model	Explained variance in % (mode 1 for predcitor)	Explained variance in % (mode 1 for predcitand)	Canonical correlation and retained CCA modes
1	CFS	20	12	0.97, 12
2	GML	26	12	0.98, 12
3	MOM3AC1	31	11	0.97, 12
4	MOM3DC2	23	12	0.98, 12
5	ECHcasst	33	10	0.97, 12
6	ECHcfssst	32	10	0.97, 12

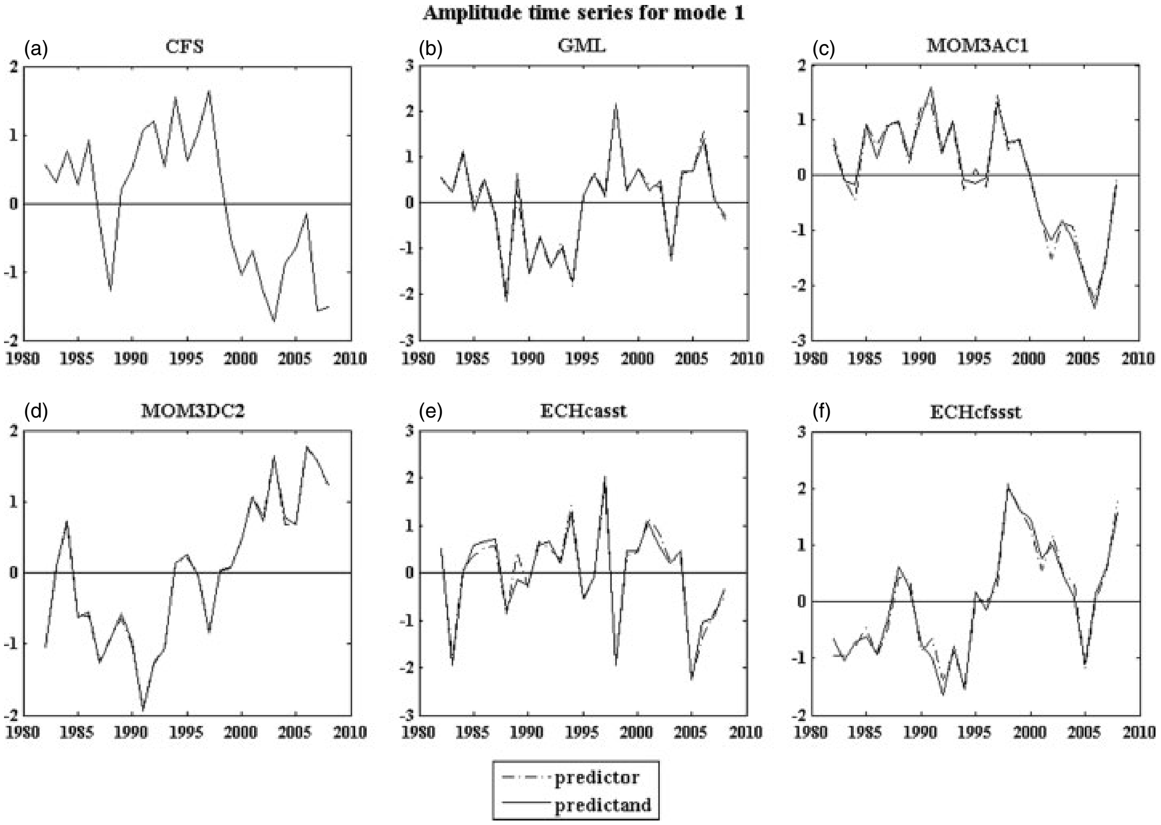

Figure 4. Amplitude time series for the canonical components for mode 1. Subplots (a) to (d) are the canonical pairs for the coupled models, (e) and (f) are for the atmospheric models.

indicate the year-to-year variation in the amplitude of the predictor loading patterns, which in turn are related to the predictand loading pattern for the respective mode (Hwang *et al.*, 2001). The canonical correlation for the corresponding mode 1 is very high (of the order of 0.97) for almost all of the models. From the figure it is clear that the temporal patterns of these two series are almost similar for the entire domain. Similarly for the CFS model, the two series are exactly similar for the domain.

The patterns in the observation and predictor are in opposite phase for almost all the models for mode 2 (figure not shown). Moreover, the amplitude time series for predictors is also of opposite sign compared to the observed time series for the majority of the years (figure not shown). As the actual contribution of a particular mode is obtained from multiplication of the pattern loadings and the time series, it can be said that, for most of the models, canonical loading pattern for predictor and predictand for the second mode are very similar.

In short, it can be said that some of the models are able to show a similar kind of loadings for both the predictor and predictand. As an example, the GML model (Figure 4(b)) best captures the predictand loadings over the entire domain whereas the other coupled models have missed some parts of the domain. Among the atmospheric models, ECHcasst (Figure 4(c)) has missed some parts for mode-1 as well as for mode 2.

3.2. CCA cross-validated skill

In the present study, as the number of hindcast years is 27, the dataset cannot be separated into two independent periods in order to develop a model from one part (training data) and then to verify it on the independent data set (verification data). Therefore, a leave-one-out cross-validation scheme has been used for the verification of the developed CCA model. For this, each year has been successively withheld from the training dataset and the remaining 26 years have been used for development of the CCA model. This model is then used for

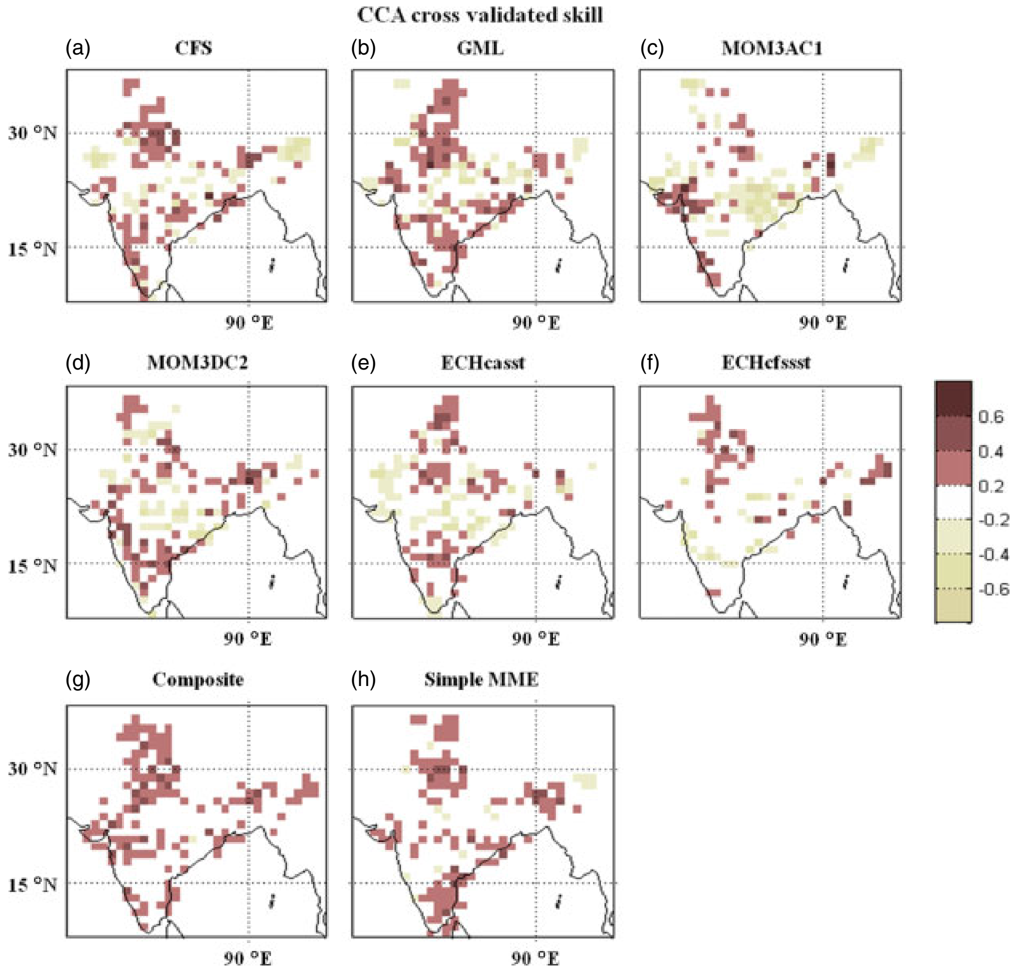

Figure 5. CCA cross validated skill for all the models ((a) to (f)) at each grid point. Plots (g) and (h) are the skill map for the mean of the models after CCA (Composite) and the raw mean (Simple MME). This figure is available in colour online at wileyonlinelibrary.com/journal/met

calculating the forecast for the verification year (the year that was withheld). This procedure is repeated to generate a cross-validated sample at all grid points. These predicted values are then correlated with the observed value to obtain the skill of the CCA model.

The leave-one-out correlation skill at each grid point is shown in Figure 5, where panels (a) to (f) display the CCA post-processed predictions from each participating model. The skill in the prediction of area averaged JJAS rainfall is evaluated for the country as a whole for each individual model. In Figure 6(a) the skill for both raw and CCA post-processed model are shown. CCA exhibits a noticeable improvement in the skill of the prediction at all-India level. The post-processed output of the ECHcasst exhibits a positive correlation which was negative in the raw model. Also, out of the four coupled GCMs used in the study, the CCA technique has improved skill

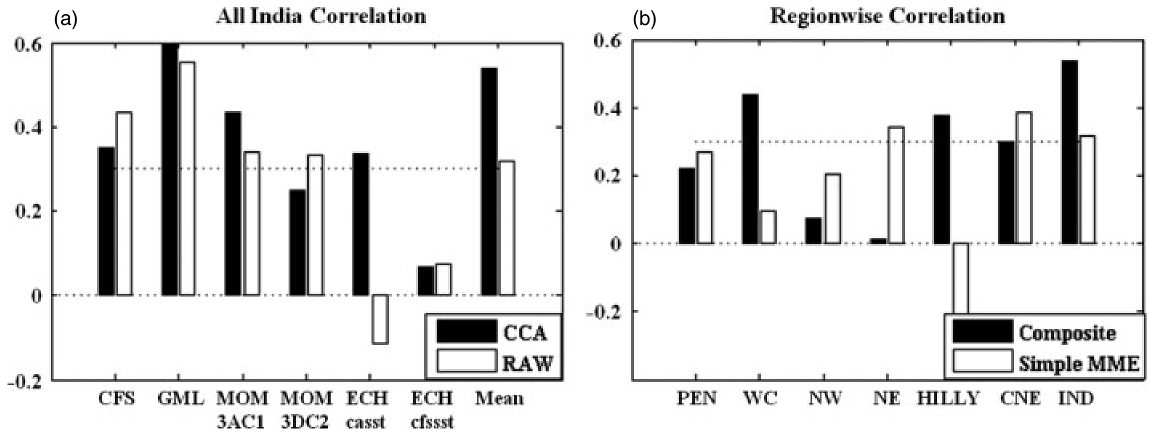

Figure 6. Correlation skill of CCA (a) correlation over the homogeneous regions for CCA (black), for raw mean of models (white), (b) skill for individual GCM averaged over the Indian land points (white), GCM output skill after the postprocessing (CCA in black bars).

of two of them at all-India level. The possible reason for the improvement in the skill of some of the GCMs may be the similarity between the loading patterns of model predictions and the observed rainfall.

In order to examine whether the multi-model ensemble of these improved products has higher skill than the simple MME of raw models, skill assessment has been made based on the leave-one-out cross validation scheme. First, multi-model ensemble of output from all the models (without applying any correction) has been carried out using a simple arithmetic average. In the following text, this is referred to as MME (referred to as Simple MME in figures). A simple arithmetic average of all the improved products (CCA model output) has then been carried out and this is referred to as Composite in the following text. Figure 5(g) shows the forecast skill of the Composite of all the CCA post-processed models and Figure 5(h) represents the skill of the MME. A comparison shows that there are several parts of India where the Composite exhibits better skill compared to the MME. The CCA composite predictions possess some high positive correlation values over the hilly areas and west central part of the country. It can also be noticed that the composite prediction has better skill in some parts of northwest India, west Uttar Pradesh, Haryana, Chandigarh and Delhi than the MME does. From Figure 6(a) it is seen that the noticeable skill improvement of individual models through CCA is further improved in Composite forecasts at the all-India level. Similarly, the skill of the Composite is compared with the MME at a regional level and is shown in Figure 6(b). A noticeable improvement can be noticed in composite CCA over the MME in the regions such as the west central and the hilly areas. This indicates that after the post processing the model performance has improved at the all-India level, and in addition there are some regions where the Composite shows better skill compared to the simple MME.

The observed rainfall anomaly time series of the CCA corrected MME rainfall at the all-India level are shown in Figure 7. These are cross-validated anomaly time series for the country as a whole for the entire period from 1982 to 2008, for the predicted rainfall (white bars) and the observed rainfall (black bars). From the figure it can be clearly observed that the CCA model is able to capture the observed features in almost 70% of the total cases (in terms of positive/negative departures). It is also seen that the CCA model is able to provide indications of the drought and flood years (e.g. rainfall with negative anomalies in 1987, 2002 and 2004) and with positive anomalies (e.g. 1988, 1994, 1998 and 2005). The model is able to predict the negative/positive departures as observed although the magnitude of the predicted departures is very low.

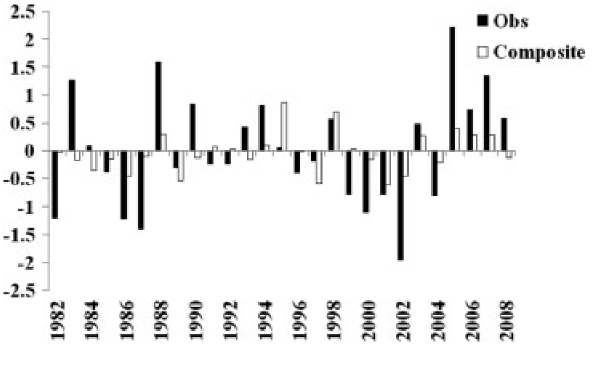

Figure 7. Anomaly time series for the observed and the Composite at the all India level. The black bars are for the observed anomaly for the year, while the white bars corresponds to the Composite.

The above findings show that the CCA has some potential to improve the forecast skill over some of the homogeneous regions as well as noticeable improvement is observed at the all India level as far as the Indian monsoon rainfall prediction is concerned. Further, the role of the CCA method in preparation of rainfall anomaly magnitudes for individual years has been examined. Forecasts for various regions of India, as well as for the country as a whole, are prepared independently for 3 years (2006, 2007 and 2008). The leave-one-out and independent prediction are slightly different procedures. In the leave-one-out procedure, the forecast for a year (say 2006) is made using the data from all the other years (1982–2005 and 2007–2008) leaving out the selected forecast period. On the contrary, the independent prediction procedure uses only the data available up to the selected forecast period, e.g. if the year selected for the forecast is 2006 then only data up to and excluding 2006 will be used. The forecast values from the composite are evaluated in terms of percentage departure and compared with the observed departure (Figure 8). In this figure, the black and white bars represent the departures for observed and forecast, respectively. For 2006 the sign of the predicted and observed percentage departure of rainfall for all the regions are in the same direction. Similar results are also noticed for the 2007 prediction for all the regions except the northeast region where the forecast skill is also not satisfactory. Over this region for 2007 the observed rainfall departure is 20% while the predicted rainfall departure is −2%. Similarly, in 2008, the forecast departures are of the same sign except for the central northeast part, in which correct sign of anomaly is obtained but the magnitude is smaller. In conclusion, the CCA corrected predictions are able to capture most of the observed features in the years 2006, 2007 and 2008.

4. Summary and conclusion

The main objective of the present study is to develop a forecast model for the improvement of the GCM's forecast of rainfall during the summer monsoon season. The lead-1 predictions for monsoon seasons (June to September) with May start (model runs use observations up to May 1) have been used in this study. It is found that these model outputs do not have significant skill over all the homogeneous regions as well as for the country as whole. To improve the forecast skill, a postprocessing technique, the Canonical Correlation Analysis (CCA), has been used. The CCA has been applied on each of the GCM outputs for rainfall in order to project the observed rainfall pattern onto the GCM predicted rainfall over the selected domain. Finally, a composite of all the post processed GCM has been made. The skill of such postprocessed products has been estimated in a leave-one-out cross validation mode and it was found that the atmospheric models having poor skill have shown noticeable improvement after the CCA corrections. It is also seen that the forecast skill of two coupled models has also improved after the post processing. The possible reason for the improvement in the skill of some of the GCMs may be the similarity between the loading patterns of model predictions and the observed rainfall. The skill of composite forecasts made using the improved products is higher than the MME over some parts of northwest India such as the plains of west Uttar Pradesh, Haryana, Chandigarh and Delhi. Prediction skill over the homogeneous regions such as the west central, hilly region as well as the country as a whole improved noticeably

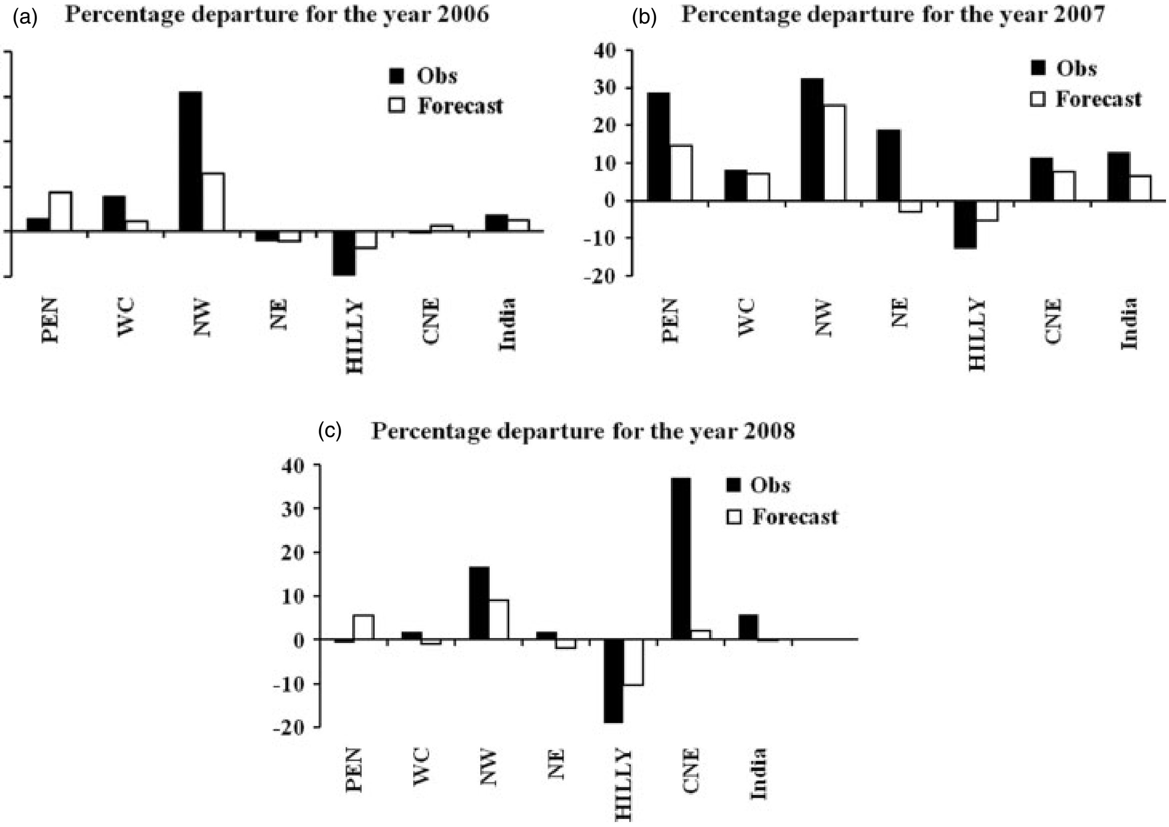

Figure 8. Forecast rainfall percentage departure (%) using CCA (in white bars) and the observed departure (in black bars) (a) for 2006, (b) for 2007, (c) for 2008.

with the Composite forecasts as compared to the MME. Prasad and Singh (1996) have used the CCA method to estimate the monsoon rainfall using some global observed variables, such as the 500 hPa ridge axis position in April and the Darwin surface pressure tendency, and had obtained significant positive skill for the large contiguous meteorological subdivisions of India with high skill score ($\geq$0.3), particularly for the meteorological subdivisions lying in west-central India. The present study, using GCM products, has been able to achieve better skill over most of the homogeneous regions considered. Therefore, it can be concluded that canonical correlation analysis has some potential to improve the forecast skill over most parts of the country, whereas the simple mean of raw model products does not exhibit satisfactory skill.

Acknowledgements

The authors are grateful to the anonymous reviewers for the valuable comments and suggestion those made substantial improvement in the present research work. This research has been conducted as part of the project titled 'Development and Application of Extended Range Weather Forecasting System for Climate Risk Management in Agriculture' at IIT, Delhi, sponsored by the Department of Agriculture and Cooperation, Government of India. We gratefully acknowledge the International Research Institute for Climate and Society (IRI), modelling and prediction group led by D. Dewitt for making five of their GCM-based seasonal forecasting systems available to this study, as well as the IRI Data Library group led by B. Blumenthal. The computing for the GCM simulations made by IRI was partially provided by a grant from the NCAR Climate System Laboratory (CSL) program to the IRI. The IRI represents a cooperative agreement between the U.S. National Oceanic and Atmospheric Administration (NOAA) Office of Global Programs and Columbia University. Gridded rain data have been obtained from India Meteorological Department (IMD). Also, the authors are thankful to National Oceanic and Atmospheric Administration National Centers for Environmental Prediction Climate Prediction Center Merged Analysis of Precipitation (CMAP) for providing the rainfall data.

References

Acharya N, Kar SC, Mohanty UC, Kulkarni MA, Dash SK. 2011. Performance of GCMs for seasonal prediction over India – a case study for 2009 monsoon. *Theor. Appl. Climatol.* **105**: 505–520.

Barnett TP, Preisendorfer R. 1987. Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by Canonical correlation analysis. *Mon. Weather Rev.* **115**: 1825–1850.

Barnston AG. 1994. Linear statistical short-term climate predictive skill in the northern hemisphere. *J. Clim.* **7**: 1513–1564.

Barnston AG, Smith TM. 1996. Specification and prediction of global surface temperature and precipitation from global SST using CCA. *J. Clim.* **9**: 2660–2697.

Chakraborty A, Krishnamurti TN. 2009. Improving global model precipitation forecasts over india using downscaling and the FSU superensemble. Part II: seasonal climate. *Mon. Weather Rev.* **137**: 2736–2757.

Gowariker V, Thapliyal V, Sarker RP, Mandal GS, Sikka DR. 1989. Parametric and power regression models: new approach to long range forecasting of monsoon rainfall in India. *Mausam* **40**: 115–122.

Graham NE, Michaelsen J, Barnett TP. 1987. An investigation of the El Niño–Southern Oscillation cycle with statistical models. Part 1. Predictor field characteristics. *J. Geophys. Res.* **92**: 14251–14270.

Hagedorn R, Doblas-Reyes FJ, Palmer TN. 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting-I. Basic concept. *Tellus* **57A**: 219–213.

Hotelling H. 1936. Relations between two sets of variates. *Biometrika* **28**(3/4): 321–377.

Hwang S-O, Schemm J-KE, Barnston AG, Kwon W-T. 2001. Long-lead seasonal forecast skill in far Eastern Asia using canonical correlation analysis. *J. Clim.* **14**: 3005–3016.

Janakiraman S, Ved M, Laveti RN, Yadav P, Gadgil S. 2011. Prediction of the Indian summer monsoon rainfall using a state-of-the-art coupled ocean-atmosphere model. *Curr. Sci.* **100**(3): 354–362.

Kang I-S, Lee J, Park CK. 2004. Potential predictability of summer mean precipitation in a dynamical seasonal prediction system with systematic error correction. *J. Clim.* **17**: 834–844.

Kar SC, Acharya N, Mohanty UC, Kulkarni MA. 2011. Skill of mean of distribution of monthly rainfall over India during July using multi-model ensemble schemes. *Int. J. Clim.* DOI: 10.1002/joc.2334.

Kar SC, Hovsepyan A, Park CK. 2006. Economic values of the APCN multi-model ensemble categorical seasonal predictions. *Meteor. Appl.* **13**(3): 267–277.

Krishna Kumar K, Hoerling M, Rajagopalan B. 2005. Advancing dynamical prediction of Indian monsoon rainfall. *Geophys. Res. Lett.* **32**: L08704, DOI: 10.1029/2004GL021979.

Krishna Kumar K, Rajagopalan B, Cane MA. 1999. On the weakening relationship between the Indian Monsoon and ENSO. *Science* **284**: DOI: 10.1126/science.284.5423.2156.

Krishnamurti TN, Kishtawal CM, Shin DW, Williford CE. 2000. Improving tropical precipitation forecasts from a multi-analysis superensemble. *J. Clim.* **13**: 4217–4227.

Krishnamurti TN, Mitra AK, Yun W-T, Kumar TSVV. 2006. Seasonal climate forecasts of the Asian monsoon using multiple coupled models. *Tellus* **58A**: 487–507.

Kug JS, Lee J, Kang I-S, Wang B, Park CK. 2008. Optimal multi-model ensemble method in seasonal prediction. *Asia Pac. J. Atmos. Sci.* **44**(3): 259–267.

Kulkarni A, Kripalani RH, Singh SV. 1992. Classification of summer monsoon rainfall patterns over India. *Int. J. Climatol.* **12**: 269–280.

Landman WA, Botes S, Goddard L, Shongwe M. 2005. Assessing the predictability of extreme rainfall seasons over southern Africa. *Geophys. Res. Lett.* **32**: L23818, DOI: 10.1029/2005GL023965.

Landman WA, Mason SJ. 1999. Operational long-lead prediction of south african rainfall using canonical correlation analysis. *Int. J. Climatol.* **19**: 1073–1090.

Li S, Goddard L. 2005. Retrospective forecasts with the ECHAM4.5 AGCM IRI. Technical Report 05-02, The International Research Institute for Climate and Society, New York, NY. December 2005. http://iri.columbia.edu/outreach/publication/report/05-02/report05-02.pdf

Lim Y, Jo S, Lee J, Oh HS, Kang HS. 2011. An improvement of seasonal climate prediction by regularized canonical correlation analysis. *Int. J. Climatol.* DOI: 10.1002/joc.2368.

Matsuura K, Willmott CJ. 2009. Terrestrial precipitation: 1900–2008 gridded monthly time series. http://climate.geog.udel.edu/∼climate/ (accessed 5 September 2009).

Mitchell TD, Jones PD. 2005. An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int. J. Climatol.* **25**: 693–712.

Mooley DA, Munot AA. 1993. Variation in the relationship of the Indian summer with global factors. *Earth Planet. Sci.* **102**(1): 89–104.

Pacanowski RC, Griffes SM. 1998. *MOM 3.0 Manual*. NOAA/ Geophysical Fluid Dynamics Laboratory: Princeton, NJ; 608 pp.

Pai DS, Rajeevan M. 2007. Indian summer monsoon onset: variability and prediction. National Climate Centre Research Report 6, India Meteorological Dept: Pune, India.

Parthasarathy B, Munot AA, Kothawale DR. 1995. Monthly and seasonal rainfall series for all India homogeneous regions and meteorological sub-divisions: 1871–1994. Research Report RR-065, Indian Institution of Tropical Meteorology: Pune, 113 pp.

Pattanaik DR, Kumar A. 2010. Prediction of summer monsoon rainfall over India using the NCEP climate forecast system. *Clim. Dyn.* **34**: 557–572.

Prasad KD, Singh SV. 1996. Forecasting the spatial variability of the Indian monsoon rainfall using canonical correlation. *Int. J. Climatol.* **16**: 1379–1390.

Preethi B, Kripalani RH, Kumar KK. 2010. Indian summer monsoon rainfall variability in global coupled ocean-atmosphere models. *Clim. Dyn.* **35**: 1521–1539.

Rajeevan M. 2001. Prediction of Indian summer monsoon: status, problems and prospects. *Curr. Sci.* **81**: 1451–1457.

Rajeevan M, Bhate J, Kale J, Lal B. 2006. High resolution daily gridded rainfall data for the Indian region: analysis of break and active monsoon spells. *Curr. Sci.* **91**: 296–306.

Rajeevan M, McPhaden MJ. 2004. Tropical Pacific upper ocean heat content variations and Indian summer monsoon rainfall. *Geophys. Res. Lett.* **31**: L18203, DOI: 10.1029/2004GL020631.

Rajeevan M, Pai DS, Anil Kumar R, Lal B. 2007. New statistical models for long-range forecasting of southwest monsoon rainfall over India. *Clim. Dyn.* **28**: 813–828, DOI: 10.1007/s00382-006-019706.

Roeckner E, Arpe K, Bengtsson L, Christoph M, Claussen M, Dumenil L, Esch M, Giorgetta M, Schlese U, Schulzweida U. 1996. The atmospheric general circulation model ECHAM4: model description and simulation of present-day climate. Report 218, Max-Planck-Institut fur Meteorologie: Hamburg; 90.

Saha S, Nadiga S, Thiaw C, Wang J, Wang W, Zhang Q, Van Den Dool HM, Pan H-L, Moorthi S, Behringer D, Stokes D, Pena M, Lord S, White G, Ebisuzaki W, Peng P, Xie P. 2006. The NCEP climate forecast system. *J. Clim.* **19**(15): 3483–3517.

Sahai AK, Chattopadhyay R. 2006. An objective study of Indian summer monsoon variability using the self organizing map algorithms. Research Report No. RR-113, ISSN 0252–1075. Indian Institute of Tropical Meteorology, Pune, India.

Sahai AK, Chattopadhyay R, Goswami BN. 2008. SST based large multi-model ensemble forecasting system for Indian summer monsoon rainfall. *Geophys. Res. Lett.* **35**: L19705, 1–9, DOI: 10.1029/2008GL035461.

Singh SV, Kripalani RH. 1986. Application of Extended Empirical Orthogonal Function Analysis to inter-relationships and sequential evolution of monsoon fields. *Mon. Wea. Rev.* **114**: 1603–1610.

Tippett M, Anderson J, Bishop C, Hamill T, Whitaker J. 2003. Ensemble square 726 root filters. *Mon. Wea. Rev.* **131**: 1485–1490.

Tippett MK, Barnston AG, Dewitt DG. 2005. Statistical correction of tropical pacific sea surface temperature forecast. *J. Clim.* **18**: 5141–5162.

Walker GT. 1924. Correlation in seasonal variations of weather. IX, A further study of world weather. *IMD Mem.* **XXIV**: (Part IX): 75–131.

Wilks DS. 1995. *Statistical Methods in Atmospheric Sciences*. 2nd edn, Vol. 59. Academic Press: San Diego, CA; 467 p.

Xie P, Arkin PA. 1995. Analysis of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions. *J. Clim.* **9**: 840–858.

Yu Z, Chu PS, Schroeder T. 1997. Predictive skills of seasonal to annual rainfall variations in the U.S. Affiliated Pacific Islands: canonical correlation analysis and multivariate principal component regression approaches. *J. Clim.* **10**: 2586–2599.