Statistical Forecasting of Indian Summer Monsoon Rainfall: An Enduring Challenge

Shivam Tripathi and Rao S. Govindaraju

School of Civil Engineering, Purdue University, West Lafayette, IN 47906, USA

1 Introduction

Forecasting All India Summer Monsoon Rainfall (AISMR), one or more seasons in advance, has been an elusive goal for hydrologists, meteorologists, and astrologers alike. In spite of advances in data collection facilities, improvements in computational capabilities, and progress in our understanding of the physics of the monsoon system, our ability to forecast AISMR has remained more or less unchanged in past several decades. On one hand, physically based *numerical prediction models* that are considered a panacea for daily weather forecasting have not evolved to a stage where they can realistically predict or even simulate annual variations in Indian monsoon. On the other hand, *statistical models* that have traditionally been used for making operational forecasts have failed in forecasting extreme monsoon rainfall years. It has been suggested that, in future, physically based models may improve to an extent where they can produce useful forecasts. However, until then, it would be prudent to develop statistical forecast models using state-of-the-art soft-computing techniques.

Statistical forecasting of AISMR has a long, venerable, and vulnerable history. The first ever scientific forecast of AISMR was made by H. F. Blanford for year 1878, after the great famine of 1877 that took a heavy toll on human lives. During initial years, forecasts issued were subjective and met limited success. Later, in early twentieth century the forecast skill of AISMR improved significantly mainly due to initiatives taken by Sir Gilbert Thomas Walker. Sir Walker, who was then the Director General of India Meteorology Department (IMD), collected and analyzed vast amounts of weather data from India and abroad. He discovered Southern Oscillation (SO), a major atmosphere phenomenon over tropical Pacific Ocean that was later linked to El Niño (Bjerknes 1969). The discovery of link between AISMR and El Niño / SO (ENSO) led to rapid development in statistical forecasting models. However, initial encouraging performance did not last long because the link between AISMR and ENSO started weakening in the 1980s (Kumar et al. 1999).

The work of Sir Walker encouraged researchers to find other atmospheric and oceanic variables over different parts of the world that can be used as potential predictors for AISMR. Some of the important predictors that came out these endeavors are : (i) global sea surface temperature (Sahai et al. 2003; Pai and Rajeevan 2006), (ii) Himalayan and Eurasian snow cover (Fasullo 2004; Kripalani et al. 2003), (iii) atmospheric circulation patterns like position of 500 hPa ridge over India (Prasad and Singh 1992), and wind anomalies (Bhalme et al. 1987; Gadgil et al. 2004), (iv) land surface conditions over Northern Hemisphere (Rajeevan 2002; Robock et al. 2003), and (v) the previous values of AISMR series (Kishtawal et al. 2003; Iyengar and Raghu Kanth

B. Prasad (Ed.): Soft Computing Applications in Industry, STUDFUZZ 226, pp. 207–224, 2008.
springerlink.com

2005). However, recent evidences suggest that the relationship between most of these predictors and AISMR is not stationary but varies on decadal to interdecadal time scales (Rajeevan et al. 2007; Gadgil et al. 2002; Kumar et al. 1999; Clark et al. 2000). Further, studies indicate that some of the predictors have even lost their importance over the course of time.

The facts mentioned in the previous paragraphs indicate that the successful forecasting of AISMR needs a statistical model that not only updates the relationship between the predictors and the predictand in light of new data but also dynamically selects the appropriate set of predictors for making forecasts. However, to date, little has been done to address this problem. The current strategy is to constantly monitor the performance of the model, and subjectively change the structure of the model, its inputs, and even training period in case of model failure (Rajeevan et al. 2007). Obviously, updating the model in this way leaves no possibility for model validation, and consequently no confidence can be assigned to the model predictions.

This study seeks to address the above mentioned issue by developing a dynamic forecasting model under Bayesian framework. The model not only updates the relationship between the predictors and the predictand as and when new data become available, but also dynamically selects the appropriate set of predictors. Further, we also suggest a way for assessing forecasting skill of the model. The model has only one adjustable parameter that helps in reducing the level of subjectivity in making forecasts. The efficacy of the model is assessed by evaluating its performance in predicting AISMR from 1951 to 2005, using global sea surface temperature (SST) dataset as the only predictor.

The proposed model, in its first step, uses *probabilistic principal component analysis* (PPCA) (Roweis 1998; Tipping and Bishop 1999) in combination with *Bayesian model selection* (Minka 2001) to reduce the dimensionality of the SST data. In the second step, the model uses a sparse Bayesian learning algorithm (Tipping 2001) to select the appropriate set of predictors, and to learn the relationship between selected predictors and AISMR, the predictand. The sparse Bayesian model is known as *relevance vector machine* (RVM) owing to its capability to identify most relevant patterns or predictors for making forecast. The parameters of the RVM model are estimated using a sequential learning algorithm. It is to be emphasized that the RVM can be used to learn non-linear relationships between predictors and predictand, however for the sake of simplicity and interpretability of results, only linear relationships are investigated in this work.

The remainder of this paper is structured as follows. The mathematical formulations of PPCA, Bayesian model selection, and RVM are presented in Sect. 2. Following this, data used in the study are described in Sect. 3. Details of the proposed methodology are given in Sec. 4, and the results obtained are presented and discussed in Sect. 5. Finally, a set of concluding remarks and our recommendations are provided in Sect. 6.

2 Mathematical Formulation

This section presents the mathematical formulation for probabilistic principal component analysis (PPCA), Bayesian model selection, and relevance vector machine (RVM) in the context of forecasting AISMR.

2.1 Probabilistic Principal Component Analysis (PPCA)

Principal component analysis (PCA) and its variants like empirical orthogonal teleconnections (van den Dool et al. 2000), nonlinear principal component analysis (Monahan 2001), rotational techniques (Horel 1981), space-time principal component (Vautard et al. 1996) and closely related methods such as canonical component analysis (Shabbar and Barnston 1996), are arguably the most commonly used methods for data compression, data reconstruction, and developing prediction models in hydrologic and meteorologic literature. PCA has also been successfully used as a feature extraction method in developing forecast models for AISMR (Cannon and McKendry 1999; Rajeevan et al. 2000; Pai and Rajeevan 2006).

However, a serious limitation of conventional PCA, when applied to inherently noisy hydro-meteorologic data, is the absence of an associated probabilistic model for the observed data. Tipping and Bishop (1999) proposed a probabilistic approach to PCA that can overcome this limitation. Besides this, probabilistic principal component analysis (PPCA) offers a number of other advantages, including a principled way for handling missing values in the data, and an objective way to decide the optimum number of principal components.

Given a p dimensional observed data variable $\boldsymbol{x}$ $\left(\boldsymbol{x} \in \Re^p\right)$, the goal of PPCA is to find a q dimensional principal variable $\boldsymbol{z}$ $\left(\boldsymbol{z} \in \Re^q\right)$, such that the number of principal components q is less than p. *Assuming q is known*, the reconstruction of data variable from principal variable is given by[1]

$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{1}$$

where, $\boldsymbol{\varepsilon}$ is a p dimensional Gaussian noise, with zero mean and covariance $\sigma^2 \boldsymbol{I}_p$, and $\boldsymbol{\mu}$ is a p dimensional vector. $\boldsymbol{W}$ is a $p \times q$ transformation matrix whose columns span a linear subspace within the p dimensional observed data variable space.

Due to the assumption of Gaussian noise, the distribution of observed data variable $\boldsymbol{x}$ conditioned on $\boldsymbol{z}$ is given by

$$p\left(\boldsymbol{x} \mid \boldsymbol{z}\right) = \mathcal{N}\left(\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_p\right) \tag{2}$$

If we assign zero mean, unit covariance Gaussian prior distribution to the principal vector $\boldsymbol{z}$, i.e.

$$p\left(\boldsymbol{z}\right) = \mathcal{N}\left(0, \boldsymbol{I}_q\right) \tag{3}$$

then the marginal distribution of the observed variable $p(\boldsymbol{x})$ also becomes Gaussian and is given by

$$p\left(\boldsymbol{x}\right) = \int p\left(\boldsymbol{x} \mid \boldsymbol{z}\right) p\left(\boldsymbol{z}\right) d\boldsymbol{z} = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{C}\right) \tag{4}$$

where, the covariance matrix $\boldsymbol{C} = \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} + \sigma^2 \boldsymbol{I}_p$.

[1] All vectors are column vectors. A d dimensional identity matrix is represented by $\boldsymbol{I}_d$.

Now consider a dataset with N observed data points i.e. $\boldsymbol{X} = \left[\boldsymbol{x}_n\right], n = 1,\ldots,N$. The log likelihood of the observed dataset, given the model (Eq. 1) is

$$\mathcal{L}\left(\boldsymbol{\mu},\boldsymbol{W},\sigma^2\right) = \sum_{n=1}^{N} \ln p\left(\boldsymbol{x}_n\right) = -\frac{N}{2}\left\{p\ln(2\pi) + \ln\left(\left|\boldsymbol{C}\right|\right) + \mathrm{tr}\left(\boldsymbol{C}^{-1}\boldsymbol{S}\right)\right\} \tag{5}$$

where $\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^{\mathrm{T}}$ is the data covariance matrix.

The model parameters $\boldsymbol{W}$, $\boldsymbol{\mu}$, and σ can be estimated by maximizing the likelihood function (Eq. 5) corresponding to these parameters. Tipping and Bishop (1999) showed that the maximum likelihood solution corresponds to principal component analysis of the dataset $\boldsymbol{X}$. The principal directions in $\boldsymbol{X}$ are contained in the columns of $\boldsymbol{W}$, while the principal components $\boldsymbol{z}$ can be calculated by Bayes' rule as

$$p\left(\boldsymbol{z} \mid \boldsymbol{x}\right) = \mathcal{N}\left(\boldsymbol{M}^{-1}\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{\mu}), \sigma^{-2}\boldsymbol{M}\right) \tag{6}$$

where $\boldsymbol{M} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} + \sigma^2\boldsymbol{I}_q$.

The maximum likelihood estimate of the parameters can be either obtained explicitly by using analytical expressions or by using expectation maximization (EM) algorithm (Roweis 1998; Tipping and Bishop 1999). For very high dimensional datasets like sea surface temperature (SST), EM algorithm has significant computational advantages and is therefore used in this study.

2.2 Bayesian Model Selection

In the above discussion of PPCA we have assumed that the dimensionality q of the principal vector $\boldsymbol{z}$ is known. In practice, subjective criteria like retaining certain percentage of the variance in the data or the point where the eigen value spectrum takes an elbow turn, are often used. However, these arbitrary thresholds cannot determine the true dimensionality of the principal vector and may result in models that are significantly different among different users.

An important advantage offered by the probabilistic interpretation of PCA is that an objective approach of Bayesian model selection can be used to determine the number of PCs. In Bayesian approach to model selection, the best model is the one which has maximum marginal likelihood over all possible values of model parameters.

In PPCA, for a given value of q, the marginal likelihood $p\left(\boldsymbol{x} \mid q\right)$ can be calculated by integrating out the model parameters $\boldsymbol{W}$, $\boldsymbol{\mu}$, and σ. To do this, suitable prior probabilities are assigned to model parameters. Minka (2001) proposed non informative prior distribution for $\boldsymbol{\mu}$ and conjugate priors for $\boldsymbol{W}$ and σ. Using these priors, an analytical expression for $p\left(\boldsymbol{x} \mid q\right)$ was derived. However, the estimate of marginal likelihood involves an integral over the Stiefel manifold that is difficult to compute exactly. To provide a practical solution, Minka (2001) applied Laplace's method that aims to find a Gaussian approximation of the marginal likelihood.

The optimal number of principal component $\hat{q}$ can then be obtained by using Bayesian model selection rule as

$$\hat{q} = \text{argmax}_q \left[p(\boldsymbol{x} \mid q) \right], \quad 1 \le q \le p \tag{7}$$

2.3 Relevance Vector Machine (RVM)

The principal components obtained from PPCA (Sect. 2.1) provide a compact representation of the raw data. However, out of the many PCs extracted from the data, only few are expected to be useful for a forecasting model. The next crucial step is therefore to select the relevant PCs by using a feature selection algorithm. These selected PCs form the predictor set for the model.

There are many feature selection algorithms available in the literature and most of them are restricted to static datasets. However, Indian summer monsoon is a time evolving system, where the predictors as well as their relationship with AISMR constantly changes with time. Therefore a *dynamic feature selection* algorithm is required. To this end, RVM model with sequential learning algorithm is adopted in this study. RVM algorithm was proposed by Tipping (2001). It has excellent generalization properties and has been successfully used in many real world applications including hydrology (Khalil et al. 2005, 2006). However, in this study RVM was selected because of the following two properties: (i) automatic relevance determination that selects the most relevant predictors for making forecasts, and (ii) sequential learning that allows model to progressively update itself as more and more data becomes available. In passing, it is worth mentioning that, to our knowledge, RVM has not been used in the context of dynamic feature selection.

Detailed mathematical formulation of RVM is available in Scholkopf and Smola (2001), Tipping (2001), and Bishop (2006). Here we provide a brief overview of RVM algorithm.

The PCs $\boldsymbol{Z} = \left[\boldsymbol{z}_n \right], n = 1,\ldots,N$ extracted from PPCA form the input to the RVM. The n^{th} member of the input set $\boldsymbol{z}_n \{ \boldsymbol{z}_n \in \Re^q, \boldsymbol{z}_n = [z_{n1}, z_{n2},\ldots, z_{nq}]^{\text{T}} \}$ constitutes the potential pool of predictors for making forecast at step n, that corresponds to time t. Further, the target value at step n is given by $y_n \{ y_n \in \Re; y = [y_1, y_2,\ldots, y_N]^T \}$, and it corresponds to standardized values of AISMR at time $t + \Delta t$, where Δt is the lead time of the forecast.

In linear RVM, the target value y_n is approximated as

$$y_n = \sum_{i=1}^{q} w_i z_{ni} + w_0 + \varepsilon_n \tag{8}$$

where, $\boldsymbol{w} = \left[w_0, w_1,\ldots, w_q \right]^{\text{T}}$ is a weight vector, and $\varepsilon = \left[\varepsilon_1,\ldots, \varepsilon_N \right]^{\text{T}}$ is independent zero mean Gaussian noise with variance σ^2. In this setting, likelihood of the observed data can be written as

$$p\left(\boldsymbol{y} \mid \boldsymbol{w}, \sigma^2\right)=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N exp\left\{-\frac{1}{2\sigma^2}\left\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{w}\right\|^2\right\} \tag{9}$$

where $\boldsymbol{\Phi}$ is defined as

$$\boldsymbol{\Phi}=\begin{bmatrix} 1 & z_{11} & \cdots & z_{1q} \\ 1 & z_{21} & \cdots & z_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{N1} & \cdots & z_{Nq} \end{bmatrix} \tag{10}$$

Under a Bayesian perspective, weights $\boldsymbol{w}$ can be estimated by first assigning prior distribution to it and then estimating its posterior distribution using likelihood of the observed data. In this study, following Tipping (2001) a zero mean Gaussian prior of the form given by Eq. 11 is used.

$$p\left(\boldsymbol{w} \mid \boldsymbol{\alpha}\right)=\prod_{i=0}^{q}\mathcal{N}\left(0, \alpha_i^{-1}\right) \tag{11}$$

In Eq. 11, $\boldsymbol{\alpha}=\left[\alpha_0,\ldots,\alpha_q\right]^T$ is a hyperparameter vector. The inverse of hyperparameter $\left(\alpha_i^{-1}\right)$ represents the importance of feature i in the model. It turns out that during the process of maximizing the likelihood of the observed data with respect to the hyperparameters, many of the hyperparameters go to infinity, and the corresponding features are removed from the model achieving sparsity in the model.

Having defined the likelihood (Eq. 9) and the prior (Eq. 11), the next step is to find the posterior distribution of the parameters. The posterior distribution of the weight vector $\boldsymbol{w}$ can be obtained analytically given hyperparameter values. However, there is no closed form equation for the posterior distribution of the hyperparameters. Nevertheless, it can be reasonably approximated by maximizing the log likelihood of the observed data with respect to hyperparameters (Tipping 2001) as given by Eq. 12

$$\mathcal{L}(\boldsymbol{\alpha})=\ln\left(p\left(\boldsymbol{y} \mid \boldsymbol{\alpha}, \sigma^2\right)\right)=-\frac{1}{2}\left(N\ln 2\pi+\ln\left|\boldsymbol{R}\right|+\boldsymbol{y}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{y}\right) \tag{12}$$

where, $\boldsymbol{R}=\sigma^2\boldsymbol{I}_N+\boldsymbol{\Phi}\boldsymbol{A}^{-1}\boldsymbol{\Phi}^T$, and $\boldsymbol{A}=diag(\alpha_0,\alpha_1,\ldots,\alpha_q)$.

The log likelihood function (Eq. 12) can be maximized by either using *type 2 maximum likelihood* (Berger 1985) or by using *sequential learning* (Li et al. 2002; Tipping and Faul 2003). Here the latter approach is used, wherein the log likelihood of observed data is maximized with respect to hyperparameter of each feature separately. The gradient of likelihood is given by

$$\frac{\partial\mathcal{L}(\boldsymbol{\alpha})}{\partial\alpha_i}=\frac{\alpha_i^{-1}S_i^{\,2}-(Q_i^{\,2}-S_i)}{2(\alpha_i+S_i)^2} \tag{13}$$

where Q_i^2 is known as the *quality* , and S_i is known as the *sparsity* of the i^{th} feature. They are calculated using Eq. 14 and Eq. 15, respectively.

$$Q_i = \boldsymbol{\varphi}_i{}^T \boldsymbol{R}_{-i}^{-1} \boldsymbol{y} \tag{14}$$

$$S_i = \boldsymbol{\varphi}_i{}^T \boldsymbol{R}_{-i}^{-1} \boldsymbol{\varphi}_i \tag{15}$$

In Eqs. 14 and 15, $\boldsymbol{\varphi}_i$ is the i^{th} column of $\boldsymbol{\Phi}$ (Eq. 10) and $\boldsymbol{R}_{-i}^{-1} = \boldsymbol{R}_i - \alpha_i^{-1} \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i{}^T$. The quality term Q_i^2 measures the increment in log likelihood of the observed data $\boldsymbol{y}$ (Eq. 12) due to inclusion of feature *i* and it thus indicates the benefit of including that feature in the model. The sparsity term S_i, on the other hand, measures the decrease in log likelihood value due increase in the covariance term $\boldsymbol{R}$. It thus indicates the cost of including an irrelevant feature in the model.

In sequential learning algorithm, the model is initialized with the bias term. In each successive iteration, the 'quality' and the 'sparsity' of all the features are calculated. The feature which is not in the model and for which the value of Q_i^2 relative to the value of S_i is greatest, is included in the model. Similarly, the feature/s that are in the model but for which the value S_i is greater the value of Q_i^2 are removed from the model. The iteration terminates when no feature can be included or excluded from the model, or the improvement in log likelihood function (Eq. 12) is below a threshold value (~ machine precision).

At convergence, the algorithm yields a set of features (F) that are deemed most relevant for making predictions, along with the posterior distribution of the associated weight vectors w_i, $\forall i \in \mathsf{F}$. The algorithm also provides an estimate of hyperparameter $\alpha_i \left(\forall i \in \mathsf{F}\right)$, and the noise variance σ^2. The weight w_i and hyperparameter α_i, $\forall i \notin \mathsf{F}$, are notionally set to zero and infinity, respectively.

The distribution of predictand y^* for new set of predictors $\boldsymbol{z}^*$ is obtained as

$$p\left(y^* \mid \boldsymbol{z}^*\right) = \mathcal{N}\left(\mu_{y^*}, \sigma_{y^*}^2\right) \tag{16}$$

where the mean and variance of the predicted value are, respectively,

$$\mu_{y^*} = \boldsymbol{\mu}_w^T \boldsymbol{\varphi}\left(\boldsymbol{z}^*\right) \tag{17}$$

$$\sigma_{y^*}^2 = \hat{\sigma}^2 \boldsymbol{\varphi}\left(\boldsymbol{z}^*\right)^{\mathrm{T}} \boldsymbol{\Sigma}_w \boldsymbol{\varphi}\left(\boldsymbol{z}^*\right) \tag{18}$$

Here, vector $\boldsymbol{\varphi}\left(\boldsymbol{z}^*\right) = \left[1, \boldsymbol{z}^*\right]^{\mathrm{T}}$, $\boldsymbol{\mu}_w$ and $\boldsymbol{\Sigma}_w$ are the mean and covariance of the posterior weight distribution and $\hat{\sigma}^2$ is the estimated error variance.

3 Data Used in This Study

3.1 Predictors

It has been suggested in the literature that the seasonal climate predictability is mainly derived from slow varying *surface boundary conditions*, like SST, snowcover, vegetation, and soil moisture that influence global atmospheric circulation and thus global surface climate (Charney and Shukla 1981). Among surface boundary conditions, SST is the most popular predictor for seasonal forecast. It is the principal surface boundary condition that influences the atmospheric seasonal variability (Barnston et al. 2005), and it is the only variable for which long term consistent records are available.

SST has long been used as a predictor for AISMR. Sahai et al. (2003) and recently Pai and Rajeevan (2006) used only global SST data to develop long range forecasting model for AISMR. They reported high correlation ($\approx$ 0.8 to 0.9) between observed and predicted values during validation period. Further, there is a plethora of studies that have investigated the link between AISMR and SST in different regions of the world including (i) Pacific Ocean (Mooley and Munot 1997; Kumar et al. 2006; Krishnan and Sugi 2003), (ii) Atlantic Ocean (Srivastava et al. 2002; Goswami et al. 2006), (iii) Indian Ocean (Kucharski et al. 2006; Li et al. 2001; Clark et al. 2000), (iv) Arabian Sea (Rao and Goswami 1988; Kothawale et al. 2007), and (v) regions surrounding Australia and Indonesia (Nicholls 1983, 1995). These studies indicate that a substantial portion of the interannual variability in AISMR can be explained by SST alone and hence the potential predictors were derived only form global SST data in this study.

Monthly 1^0 resolution global SST data from 1870 onwards is available from the Hadley Centre Sea Ice and Sea Surface Temperature dataset (HadISST1) (Rayner et al. 2003). The dataset is based on interpolation of measured SST values compiled in International Comprehensive Ocean Atmosphere Data Set (ICOADS) database, and Met Office Marine Data Bank (MDB). The dataset is constructed using a reduced space optimal interpolation procedure. This dataset is updated every month and can be obtained from Hadley Centre's website *http://hadobs.metoffice.com/hadisst/*.

3.2 Predictand

The predictand used in the study is All India Summer Monsoon Rainfall (AISMR). The monthly area weighted summer monsoon rainfall data over India (Parthasarathy et al. 1994), which extends from 1871 to 2004, is extracted from Indian Institute of Tropical Meteorology, Pune, web site *http://www.tropmet.res.in*. Primary source of the data is India Meteorological Department.

4 Methodology

This section outlines the procedure involved in processing SST and rainfall data, identifying SST patterns that are good predictors for AISMR, and developing RVM model to forecast AISMR.

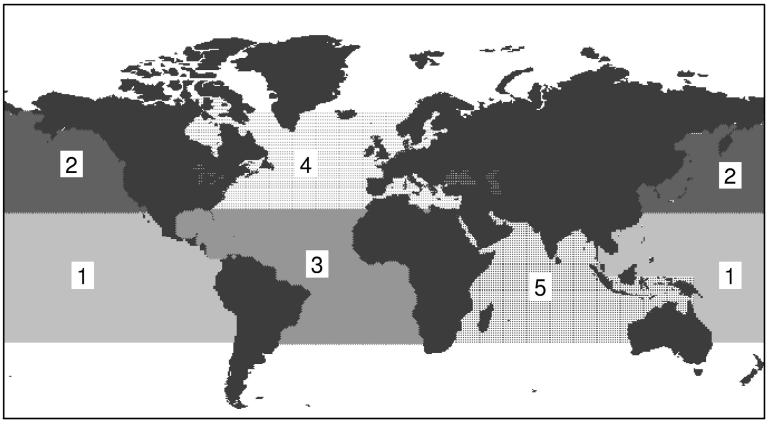

Fig. 1. Partitioning of HadISST1 data into five oceanic sectors: 1-Tropical Pacific sector (30^0S – 30^0N), 2-North Pacific sector (north of 30^0N), 3-Tropical Atlantic sector (30^0S – 30^0N), 4-North Atlantic sector (north of 30^0N), and 5-The Indian Ocean sector (north of 30^0S).

As the first step, the rainfall data were standardized (subtracted by the long term mean and divided by the standard deviation). Following Lau et al. (2002), the SST data were partitioned into five non-overlapping sectors- Tropical Pacific sector (30^0S–30^0N), North Pacific sector (north of 30^0N), Tropical Atlantic sector (30^0S–30^0N), North Atlantic sector (north of 30^0N), and Indian Ocean sector (north of 30^0S) as shown in Fig. 1. This was done because intrinsic ocean variability outside of the tropical Pacific Ocean is known to be frequently obscured by strong ENSO signal. Partitioning of data allows studying SST variability in all sectors separately.

The partitioned SST data was then used to calculate the SST *anomalies* (SSTa) and *tendencies of SST anomalies* (SSTt). SSTa at a grid point for a given season is defined as the deviation of SST value from its long term average. SSTt for a given season is defined as the change in SSTa from the previous season. SSTt values represent the evolution of SST data over time and are reported to have better predictive information than SSTa (Sahai et al. 2003). In this study, the time lag of SSTa and SSTt were varied from 1 to 12 months behind the start of monsoon month (June).

After preliminary exploratory data analysis (Tukey 1977), the rainfall, and the SST data were divided into a training set (1900-1950) and a test set (1951-2005). SST data prior to 1900 is less reliable (Smith and Reynolds 2003) and therefore not considered in the analysis. Following this, Pearson product-moment correlation coefficients between SST data (SSTa and SSTt) and rainfall data (AISMR) in training set were calculated. Potential predictors among SSTa and SSTt from different oceanic sectors were screened by imposing various thresholds (0.15 to 0.45) on the absolute value of correlation coefficients. The screening step is based on the assumption that the relevant features will exhibit some correlations on their own so that they can be segregated from the irrelevant features. Screening thus reduces the noise in the raw data, and is vital for the success of the following steps. The screened variables were then processed through probabilistic principal component analysis (PPCA) to extract principal components that preserve maximum variance in the screened data. The number

of principal components to be retained for subsequent analysis was decided by Bayesian model selection method described in Sect. 2.2.

The principal components (PCs) obtained from the PPCA serve as inputs to the relevance vector machine (RVM). RVM uses automatic relevance determination to identify the most relevant features (PCs) for forecasting AISMR. It also builds a linear relationship between identified predictors and AISMR of the form given by Eq. 8. RVM, by virtue of its Bayesian formulation, progressively updates the predictor set, and its relationship with predictand. In the sequential learning algorithm adopted in this work (Sect. 2.3), the posterior distribution of model parameters at a given step become prior distribution of the parameters for the next step. Bayes' rule is then used to update the posterior distribution in light of new data.

5 Results and Discussion

Exploratory data analysis was done with a view to understand the relationship between SST data [SST anomalies (SSTa) and tendency of SST anomalies (SSTt)] and All India Summer Monsoon Rainfall (AISMR). To this end, a 30 year moving window was used to calculate the correlation between AISMR and principal components of SSTa and SSTt over different oceanic sectors for different time lags. Pearson product-moment correlation and Kendall's tau rank correlation coefficient were calculated. Typical results of the analysis are shown in Fig. 2. It is evident from the figure that the relationship between AISMR and SST is not static but changes with time. These results are in agreement with the previous studies discussed in Sect. 1. Furthermore, the results highlight the need for developing a dynamic feature selection and learning model for forecasting AISMR.

The SSTa and SSTt values in the *training data* were used to screen the set of potential predictors following the procedure outlined in Sect. 4. The screened variables were processed through probabilistic principal component analysis (PPCA). The number of principal components that were to be retained for the subsequent analysis was estimated by using Bayesian model selection method (Sect. 2.2). A typical result of this selection method is given in Fig. 3. The figure shows the marginal likelihood computed by varying the number of principal components. The optimum number corresponds to maximum value of the marginal likelihood.

The PCs obtained in the foregoing step serve as inputs to the RVM model, while standardized values of AISMR formed its output. The model was initially trained for the period 1901 to 1950. The trained model was then used to forecast the value of AISMR for 1951. After the forecast was made, the observed value of AISMR for year 1951 was used to update the RVM parameters. During an update operation, the predictors that have lost their relevance are removed form the model, new relevant predictors are added, and the relationship between existing predictors and AISMR is revised. The steps are repeated to sequentially generate forecast for the test period (1951 to 2005). The correlation between the forecasted and the observed AISMR during the test period was computed. The analysis was repeated number of times by varying the threshold for screening predictors in the range 0.15 to 0.4. The results obtained are presented as a solid line in the group of curves labeled 'C' in Fig. 4. The dashed line in the 'C-group' are the results obtained from the same analysis, but this time

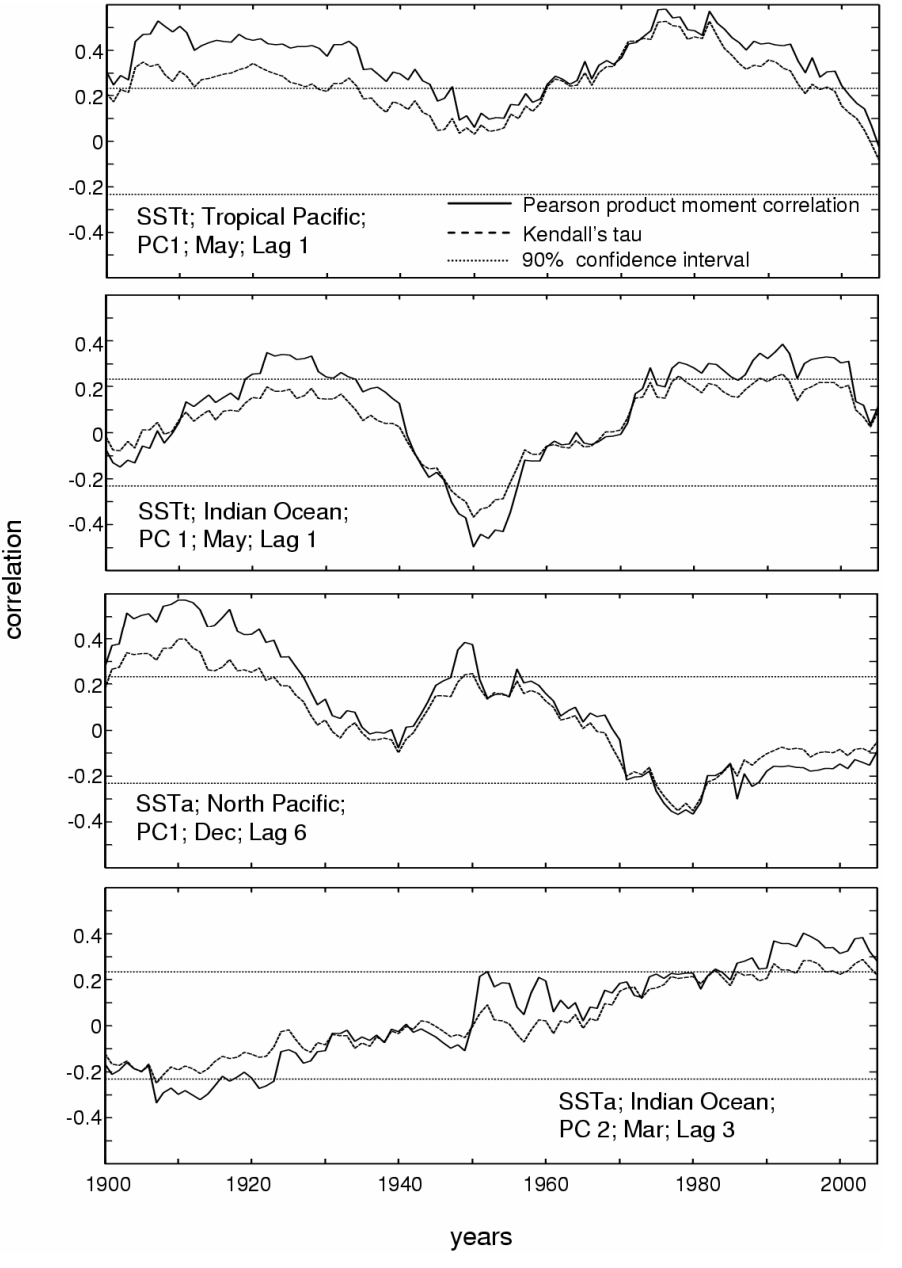

Fig. 2. Thirty year moving window correlation between AISMR and SST data. Text on each panel denotes the type of variable [SST anomaly (SSTa) / tendency of SST anomaly (SSTt)], oceanic sector, order of principal component, month of observation, and time lag (in months) from the start of monsoon season (June).

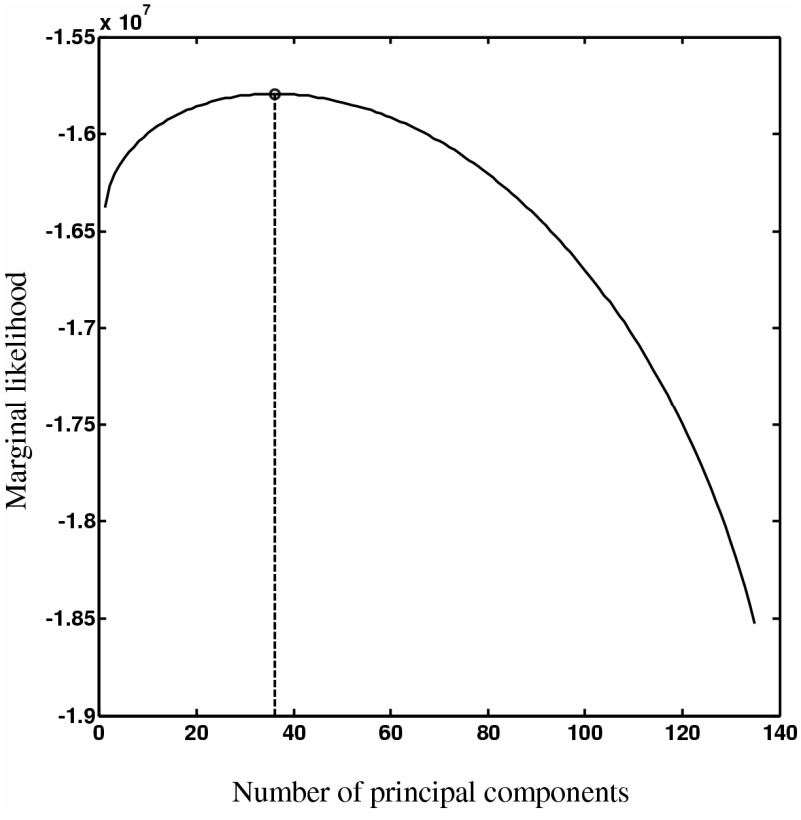

Fig. 3. Identification of number of principal components (PCs) using Bayesian model selection. The figure corresponds to the case for which the threshold on correlation for screening predictors was set to 0.15. The model selects 36 as the optimum number of PCs.

calculating PCs using standard principal component analysis (SPCA) instead of PPCA. Number of PCs was chosen to be same as the optimum number of PCs yielded by Bayesian model selection method for PPCA. It is evident for the figure that the model performed poorly in forecasting AISMR values. The results obtained are in contradiction with earlier findings that about 80% of inter annual variability in AISMR can be explained by global sea surface temperatures prior to monsoon months. It should be pointed out that qualitatively similar results were obtained for different choices of training and testing periods.

How can this miss-match be explained? Here is a possible explanation: In the literature, high correlation between forecasted and observed rainfall during independent test period has been reported without allowance for the *selection bias*. The selection bias is induced because the test data are used at the first instance to select the predictors, which are then used to develop the forecast model. But what can be the *magnitude of selection bias*? To answer this question, the following analysis was done. The *entire data* from 1901 to 2005 (i.e. both training data and test data) were used to screen the set of potential predictors. The screened variables were then processed

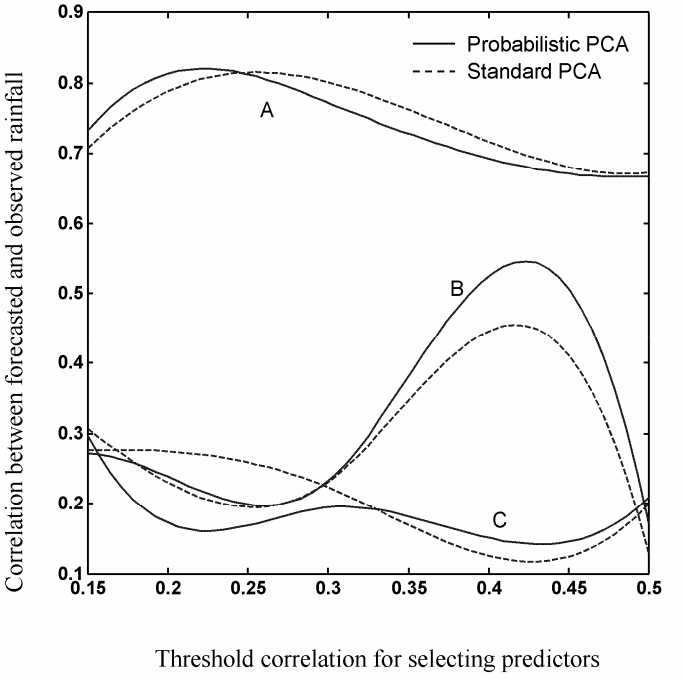

Fig. 4. Comparison of observed and forecasted rainfall values for testing period (1951 to 2005). The correlations were calculated for three strategies of screening potential predictors: (A) using entire data, (B) using training and alternate test data, and (C) using only test data.

through principal component analysis, the output of which forms the input to RVM model. The RVM model was trained in the same way as done before. The results obtained are shown as lines in the group of curves labeled 'A' in Fig. 4. The solid line corresponds to analysis using PPCA, whereas dashed line corresponds to SPCA. Clearly, a very high correlation can be obtained in this way. The results indicate that the selection bias can be very high and the results obtained without making allowance for it can be misleading. The selection bias, in general increases as the number of potential predictors increases. We further note that the concerns of getting deceptive results due to selection bias have been reported in other contexts. For example, Ambroise and McLachlan (2002) brought out the effects of selection bias in the context of cancer diagnosis and treatment.

Cross-validation and bootstrap estimates are the commonly used methods for assessing the performance of a model, when the number of data points is relatively small and the number of possible predictors large. However, these methods are not directly applicable to the dataset where the relationship between predictors and predictands is evolving over time. To address this problem a simple strategy is devised. Alternate years from the test data set are used along with the training dataset to screen the possible set of predictors. The performance of model is then assessed on forecasting the values of left over years in the test data set. The performance measured using this strategy is more realistic for an operational forecast model. The results obtained using this method are presented as group of curves labeled 'B' in Fig. 4. As expected

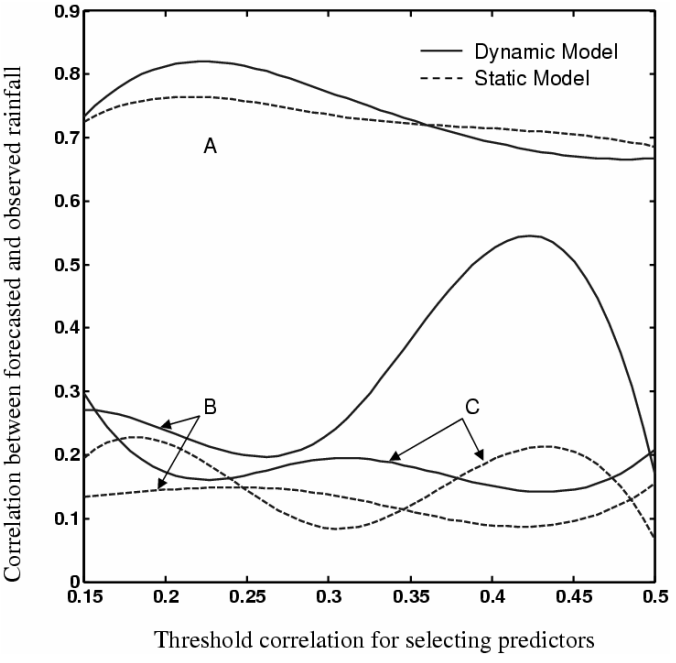

Fig. 5. Comparison of static and dynamic forecast models in forecasting rainfall values for testing period (1951 to 2005). The correlations were calculated for three strategies of screening potential predictors: (A) using entire data, (B) using training and alternate test data, and (C) using only training data.

the correlation between forecasted and observed rainfall values is in between two previous cases. Further, form Fig. 4 it can be inferred that the performance of forecasting model using PPCA or SPCA are similar. Nonetheless, PPCA is preferable because it provides an objective way of deciding the number of PCs.

The performance of a dynamic and a static model are also compared and the results are presented in Fig. 5. Both the static and the dynamic models were trained using data from 1901 to 1950. The dynamic model updates itself at each step as new data becomes available, while the static model remains same. The correlation between the forecasted and the observed values for the two models, for all three ways of screening potential predictors i.e. using: (A) entire data, (B) training and alternate test data, and (C) only training data are shown in Fig. 5. As expected the performance of dynamic model is better than static model. The advantages of dynamic model over static model are more pronounced for case B. The results corroborate the earlier findings that the relationship between SST and AISMR are continuously changing over time and that the dynamic forecast model is more suitable for forecasting AISMR.

6 Concluding Remarks

In this study, an attempt has been made to explore the links between All India Summer Monsoon Rainfall (AISMR) and global sea surface temperature (SST). The

exploratory data analysis indicated that the relationship between AISMR and SST over different oceanic sectors is not static, but continuously changes with time. Not only that, the analysis revealed that the set of predictors for AISMR also changes with time. These findings indicate that a reliable statistical forecast of AISMR can be obtained only by a model that progressively updates itself by accounting for these changing relationships. Further, it was pointed out that any sort of ad-hoc or handcrafted method of updating the model is unlikely to be optimal.

To address the above mentioned problem in a principled way, a Bayesian framework was adopted in this study. A methodology involving probabilistic principal component analysis, Bayesian model selection, and a sparse Bayesian learning algorithm (relevance vector machine) was introduced. The methodology automatically selects the best predictors in the global SST data for a forecasting model using principle of automatic relevance determination. Further, it progressively updates the predictor set and the relationship between predictors and predictand (AISMR) as new data becomes available.

The application of the proposed methodology to the forecasting of AISMR indicated that the strategy of constantly updating the model consistently provided better results than its static counterpart. However, in contrast to the results reported in the literature, the model developed in this work could only partially explain the interannual variability in the AISMR series. This discrepancy in the results indicated towards the problem of selection bias. It was found that, in the literature, high correlation between forecasted and observed rainfall during independent test period has been reported without allowance for the selection bias (i.e. test data are used to select the predictors for the model). To further understand the implications of selection bias in forecasting AISMR, its magnitude was estimated. The magnitude of selection bias was found to be strikingly high. The correlation between observed and forecasted value of AISMR jumped from ~ 0.25 to ~ 0.8 for unbiased to biased forecasts. These findings provide an inkling as to why there are frequent failures in forecasting AISMR even when model prediction error is small. It further points out the need to develop a strategy for estimating model prediction error for the systems that evolve over time. Towards this end, a simple strategy is recommended that considers alternate years in the test data for estimating model prediction error.

The methodology developed in this study is limited in several ways. Firstly, the methodology does not account for the measurement errors in SST data which is ubiquitous and significantly vary in space and time. Secondly, the sequential learning algorithm of relevance vector machine (RVM) is sensitive to the noise in the data. The problem is particularly pronounced for estimating noise term σ^2. Thirdly, the sequential learning algorithm for RVM tries to maximize the marginal likelihood of the data; however it does not guarantee a global optimum solution. The last two problems can be partly addressed by performing Monte Carlo simulations. Nevertheless, in spite of many limitations, the preliminary results obtained from the proposed methodology are promising. However, several avenues should be explored to further refine this attempt.

The true unbiased skill in forecasting AISMR using global SST data is very low. Even the use of sophisticated dynamic forecasting model can only marginally improve the forecasting performance. This brings to the fore the fact that the vagaries of Indian monsoon are difficulty to predict. The skillful forecasting of AISMR still

remains a challenge. Perhaps more advanced soft-computing techniques in association with improved physical understanding of the monsoon system will improve the quality of forecasts.

References

Ambroise, C., McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of National Academy of Science USA 99(10), 6562–6566 (2002), doi:10.1073/pnas.102102699

Barnston, A.G., et al.: Improving seasonal prediction practices through attribution of climate variability. Bulletin of the American Meteorological Society 86(1), 59–72 (2005)

Berger, J.O.: Statistical decision theory and Bayesian analysis, 2nd edn. Springer, New York (1985)

Bhalme, H.N., Rahalkar, S.S., Sikder, A.B.: Tropical Quasi-Biennial Oscillation of the 10-mb wind and Indian monsoon rainfall-implications for forecasting. International Journal of Climatology 7(4), 345–353 (1987), doi:10.1002/joc.3370070403

Bishop, C.M.: Pattern Recognition and Machine Learning. In: Information Science and Statistics, 1st edn., Springer, New York, USA (2006)

Bjerknes, J.: Atmospheric teleconnections from the equitorial Pacific. Monthly Weather Review 97(3), 163–172 (1969)

Cannon, A.J., McKendry, I.G.: Forecasting all-India summer monsoon rainfall using regional circulation principal components: a comparison between neural network and multiple regression models. International Journal of Climatology 19(14), 1561–1578 (1999)

Charney, J.G., Shukla, J.: Predictability of monsoon. In: Lighthill, J., Pearce, R.P. (eds.) Monsoon Dynamics, pp. 99–108. Cambridge University Press, Cambridge (1981)

Clark, C.O., Cole, J.E., Webster, P.J.: Indian ocean SST and Indian summer rainfall: Predictive relationships and their decadal variability. Journal of Climate 13(14), 2503–2519 (2000)

Fasullo, J.: A stratified diagnosis of the Indian monsoon – Eurasian snow cover relationship. Journal of Climate 17(5), 1110–1122 (2004)

Gadgil, S., et al.: On forecasting the Indian summer monsoon: the intriguing season of 2002. Current Science 83(4), 394–403 (2002)

Gadgil, S. et al.: Extremes of the Indian summer monsoon rainfall, ENSO and equatorial Indian Ocean oscillation. Geophysical Research Letters 31, L2213 (2004), doi: 10.1029/2004GL019733

Goswami, B.N., et al.: A physical mechanism for North Atlantic SST influence on the Indian summer monsoon. Geophysical Research Letters 33, L02706 (2006), doi: 10.1029/2005GL024803

Horel, J.D.: A rotated principal component analysis of the interannual variability of the Northern Hemisphere 500 mb height field. Monthly Weather Review 109(10), 2080–2092 (1981)

Iyengar, R.N., Raghu, K.S.: Intrinsic mode functions and a strategy for forecasting Indian monsoon rainfall. Meteorology and Atmospheric Physics 90(1–2), 17–36 (2005)

Khalil, A., et al.: Sparse Bayesian learning machine for real-time management of reservoir releases. Water Resources Research 41, W11401 (2005), doi: 10.1029/2004WR003891

Khalil, A.F., et al.: Multiobjective analysis of chaotic dynamic systems with sparse learning machines. Advances in Water Resources 29(1), 72–88 (2006)

Kishtawal, C.M., et al.: Forecasting summer rainfall over India using genetic algorithm. Geophysical Research Letters 30(23), 2203 (2003), doi: 10.1029/2003GL018504

Kothawale, D.R., Munot, A.A., Borgaonkar, H.P.: Temperature variability over the Indian Ocean and its relationship with Indian summer monsoon rainfall. Theoretical and Applied Climatolog (in press, 2007), doi: 10.1007/s00704-006-0291-z

Kripalani, R.H., Kulkarni, A., Sabade, S.S.: Western Himalayan snow cover and Indian monsoon rainfall: A re-examination with INSAT and NCEP/NCAR data. Theoretical and Applied Climatology 74(1), 1–18 (2003)

Krishnan, R., Sugi, M.: Pacific decadal oscillation and variability of the Indian summer monsoon rainfall. Climate Dynamics 21(3), 233–242 (2003)

Kucharski, F., Molteni, F., Yoo, J.H.: SST forcing of decadal Indian monsoon rainfall variability. Geophysical Research Letters 33, L03709 (2006), doi: 10.1029/2005GL025371

Kumar, K.K., Rajagopalan, B., Cane, M.A.: On the weakening relationship between the Indian Monsoon and ENSO. Science 284(5423), 2156–2159 (1999)

Kumar, K.K., et al.: Unraveling the mystery of Indian monsoon failure during El-Niño. Science 314(5796), 115–119 (2006)

Lau, K.M., Kim, K.M., Shen, S.S.P.: Potential predictability of seasonal precipitation over the United States from canonical ensemble correlation predictions. Geophysical Research Letters 29(7), 1097 (2002), doi: 10.1029/2001GL014263

Li, T., et al.: On the relationship between Indian Ocean sea surface temperature and Asian summer monsoon. Geophysical Research Letters 28(14), 2843–2846 (2001)

Li, Y., Campbell, C., Tipping, M.E.: Bayesian automatic relevance determination algorithms for classifying gene expression data. Bioinformatics 18(10), 1332–1339 (2002)

Minka, T.P.: Automatic choice of dimensionality for PCA. In: Leen, T.K., Dietterich, T.G., Tresp, V. (eds.) Advances in neural information processing systems, vol. 13, pp. 598–604. MIT Press, Cambridge (2001)

Monahan, A.H.: Nonlinear principal component analysis: Tropical Indo-Pacific sea surface temperature and sea level pressure. Journal of Climate 14(2), 219–233 (2001)

Mooley, D.A., Munot, A.A.: Relationships between Indian summer monsoon and pacific SST/SOI tendency from winter to spring and their stability. Theoretical and Applied Climatology 56(3), 187–197 (1997)

Nicholls, N.: Predicting Indian monsoon rainfall from sea-surface temperature in the Indonesia-north Australia area. Nature 306(5943), 576–577 (1983)

Nicholls, N.: All-India summer monsoon rainfall and sea surface temperatures around Northern Australia and Indonesia. Journal of Climate 8(5), 1463–1467 (1995)

Pai, D.S., Rajeevan, M.: Empirical prediction of Indian summer monsoon rainfall with different lead periods based on global SST anomalies. Meteorology and Atmospheric Physics 92(1), 33–43 (2006)

Parthasarathy, B., Munot, A.A., Kothawale, D.R.: All-India monthly and seasonal rainfall series: 1871–1993. Theoretical and Applied Climatology 49(4), 217–224 (1994)

Prasad, K.D., Singh, S.V.: Possibility of predicting Indian monsoon rainfall on reduced spatial and temporal scales. Journal of Climate 5(11), 1357–1361 (1992)

Rajeevan, M.: Winter surface pressure anomalies over Eurasia and Indian summer monsoon. Geophysical Research Letters 29(10), 1454 (2002), doi: 10.1029/2001GL014363

Rajeevan, M., Guhathakurta, P., Thapliyal, V.: New models for long range forecasts of summer monsoon rainfall over North West and Peninsular India. Meteorology and Atmospheric Physics 73(3), 211–225 (2000)

Rajeevan, M., et al.: New statistical models for long-range forecasting of southwest monsoon rainfall over India. Climate Dynamics 28(7–8), 813–828 (2007)

Rao, K.G., Goswami, B.N.: Interannual variations of sea surface temperature over the Arabian Sea and the Indian monsoon: A new perspective. Monthly Weather Review 116(3), 558–568 (1988)

Rayner, N.A., et al.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. Journal of Geophysical Research 108(D14), 4407,1–37 (2003)

Robock, A., et al.: Land surface conditions over Eurasia and Indian summer monsoon rainfall. Journal of Geophysical Research 108(D4), 4131,1–17 (2003), doi: 10.1029/2002JD002286

Roweis, S.T.: EM algorithms for PCA and SPCA. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (eds.) Advances in neural information processing systems, vol. 10, pp. 626–632. MIT Press, Cambridge (1998)

Sahai, A.K., et al.: Long-lead prediction of Indian summer monsoon rainfall from global SST evolution. Climate Dynamics 20(7), 855–863 (2003)

Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. In: Adaptive Computation and Machine Learning, 1st edn., MIT Press, Cambridge, MA, USA (2001)

Shabbar, A., Barnston, A.G.: Skill of seasonal climate forecasts in Canada using canonical correlation analysis. Monthly Weather Review 124(10), 2370–2385 (1996)

Smith, T.M., Reynolds, R.W.: Extended reconstruction of global sea surface temperatures based on COADS data. Journal of Climate 16(10), 1495–1510 (2003)

Srivastava, A.K., Rajeevan, M., Kulkarni, R.: Teleconnection of OLR and SST anomalies over Atlantic Ocean with Indian summer monsoon. Geophysical Research Letters 29(8), 1284 (2002), doi: 10.1029/2001GL013837

Tipping, M.E.: Sparse Bayesian learning and the Relevance vector machine. Journal of Machine Learning Research 1(3), 211–244 (2001)

Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society, Series B 61(3), 611–622 (1999)

Tipping, M.E., Faul, A.: Fast marginal likelihood maximization for sparse Bayesian models. In: Bishop, C.M., Frey, B.J. (eds.) Proceedings of the ninth international workshop on artificial intelligence and statistics, January 3–6, Key West, Florida, 8 pages (2003)

Tukey, J.W.: Exploratory data analysis. Addison-Wesley, Reading, MA, USA (1977)

van den Dool, H.M., Saha, S., Johansson, A.: Empirical orthogonal teleconnections. Journal of Climate 13, 1421–1435 (2000)

Vautard, R., Pires, C., Plaut, G.: Long-range atmospheric predictability using Space-Time principal components. Monthly Weather Review 124(2), 288–307 (1996)