

Statistical Feature Selection for Hydrologic Prediction Models

Shivam Tripathi and Rao S. Govindaraju

School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA,
PH (765) 496 3402; FAX (765) 496 1988, govind@ecn.purdue.edu

Abstract

Hydrological processes at regional scales are known to be linked to global climate patterns. State-of-the-art statistical models utilize indicators of atmospheric and oceanic circulation for hydrologic prediction. Sea surface temperature (SST) that controls the exchange of heat and momentum between the ocean and the atmosphere is often used as a key input variable in these models. The accuracy of these models depends largely on the selection of location and timing of SST observations from several thousand possible combinations. Most of the studies to date have used linear correlation between global SST observations and hydrological variables as a measure for feature selection or have relied on principal component analysis for feature extraction. The availability of new and reliable SST data from remote sensing and other sources offers an opportunity to investigate the possibility of finding new features in global SST data that can be used in hydrologic prediction models. To this end, a two step approach to feature selection is proposed in this work. In the first step, probabilistic principal component analysis (PPCA) is used to extract features, and in the second step support vector machine-based recursive feature elimination algorithm (SVM-RFE) is used to select best features form the extracted principal components (PCs). The effectiveness of the proposed approach is illustrated through its application to Indian summer monsoon rainfall (ISMR) data. The results indicate that the proposed method can successfully obtain a small subset of SST patterns that can correctly classify extreme states in ISMR. Some of the identified patterns have resemblance with patterns that are well established in literature. Investigation on the relationship between selected SST patterns and ISMR indicate that they have a dynamical relationship. A static forecasting model, therefore, cannot achieve high forecasting skills. Instead a dynamic forecasting model is needed to fully exploit the potential of selected features in forecasting ISMR.

Keywords: Sea surface temperature, probabilistic principal components, Bayesian model selection, feature selection, support vector machine

Introduction

Long range seasonal forecast of Indian summer monsoon rainfall (ISMR) has been a topic of scientific studies for more than a century. Research in this field is driven by both scientific curiosity and economic necessity. In the last two decades, the most significant contribution towards successful forecasting of ISMR has been the identification of its link with El Niño-Southern Oscillation (ENSO) phenomenon. Another development has been the creation of large archives of ocean, atmosphere and land surface data that have greatly assisted in calibrating and validating models. In spite of these achievements, the ability to forecast ISMR has remained more or less unchanged (Gadgil et al., 2005). Empirical models that have been the mainstay for long range forecasting of ISMR have failed to predict two major droughts of this decade. Numerical models that have shown considerable prediction skill over different regions of the world could not achieve the same level of performance for Indian monsoon (Gadgil and Sajani, 1998). Thus, understanding and accurate forecasting of Indian monsoon and understanding its fundamental processes remain an open scientific problem.

Recent advances in remote sensing and satellite technology have laid put enormous and sometimes bewildering, amounts of hydrologic data (more in space then in time) at our disposal. Some of these hydrologic variables are suspected of being precursors of ISMR. However the important predictors have not all been identified, nor has their relationship to ISMR have been established. A key issue now is to search for patterns and discover regularities in the data to further corroborate existing relationships, and more importantly, to reveal hitherto unknown relationships between the predictors and ISMR. Among the various datasets that are available, sea surface temperature (SST) data are most commonly used in the literature. It is a primary variable that governs the slow varying surface boundary conditions from which seasonal climate estimation has a good chance of being accomplished. Further, it is the only variable for which consistent records are available for relatively long periods of time.

The identification of SST patterns that can explain the variability in ISMR is of fundamental and practical interest. The literature already abounds with such studies (Pai and Rajeevan, 2006; Sahai at al., 2003). However, most of these studies have used either linear correlation analysis or principal component analysis (PCA) as tools for finding relevant SST patterns. These tools, although powerful in their own right, have serious limitations when applied to this problem. PCA is perhaps the most commonly used technique for feature extraction and dimensionality reduction in hydro-meteorology. However, being unsupervised in nature, it is overly concerned with faithful representation of data. The greatest emphasis is on preserving variability rather than identifying patterns that can discriminate the signals in ISMR. Another disadvantage is the subjectivity involved in deciding number of principal components. Similarly, correlation analysis can identify individual patterns that can explain variability in ISMR. However, it cannot yield a compact set of best patterns because the patterns identified individually are often highly correlated amongst

themselves. Moreover, a pattern deemed useless when judged alone can be very useful in combination with other patterns.

In order to address these limitations, this study explores the potential of alternative tools for identifying useful SST patterns. In particular, we propose a two step procedure for identifying SST patterns that can discriminate extreme states (floods and droughts) in ISMR data. In the first step, probabilistic principal component analysis (PPCA) is used to extract features from SST data. The number of principal components is obtained objectively using Bayesian model selection. In the second step, support vector machine-based recursive feature elimination (SVM-RFE) algorithm is implemented to select best features form among the principal components (PCs) extracted in the first step. Further, to investigate the relationship between ISMR and identified SST patterns, a forecasting model was developed and tested.

The remainder of this paper is structured as follows: First, the mathematical formulations of PPCA, Bayesian model selection, and SVM-RFE are presented. Following this, data used in the study are described and the details of the methodology proposed for selection of features is presented. Finally, a set of conclusions is drawn following discussion of results obtained from the proposed methodology.

Feature extraction by probabilistic principal component analysis (PPCA)

Principal component analysis is defined as an orthogonal projection of the data on to a lower dimensional linear space such that variance of projected data is maximized. Equivalently, it is also defined as a linear projection which has minimum squared error in reconstructing original data. It is the second definition which is used to develop the concept of PPCA. The algorithm of PPCA was independently proposed by Roweis (1998) and Tipping and Bishop (1999).

Given p dimensional observed data vector $\boldsymbol{x}$ the goal of PPCA is to find a q dimensional principal vector $\boldsymbol{z}$ such that the number of principal components q is less than p. Assuming q is known, the reconstruction of data vector from principal vector is given by

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

where $\boldsymbol{\varepsilon}$ is a p dimensional Gaussian noise with zero mean and covariance $\sigma^2\boldsymbol{I}$, $\boldsymbol{\mu}$ is a p dimensional parameter vector that permits model to have non-zero mean and $\boldsymbol{W}$ is a $p \times q$ transformation matrix. Due to the assumption of Gaussian noise the distribution of observed variable $\boldsymbol{x}$ conditioned on $\boldsymbol{z}$ is given by

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

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

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

then the marginal distribution of the observed variable $p(\boldsymbol{x})$ also becomes Gaussian. The model parameters $\boldsymbol{W}$, $\boldsymbol{\mu}$ and σ can be estimated by maximizing the likelihood corresponding to this marginal distribution.

The maximum likelihood estimate of $\boldsymbol{\mu}$ is given by the mean of the data. The parameters $\boldsymbol{W}$ and σ can be estimated explicitly by using analytical expressions given in Tipping and Bishop (1999) or by using expectation maximization (EM) algorithm. For very high dimensional datasets like SST, EM algorithm has significant computational advantages and is therefore used in this study. Finally, the posterior distribution of the principal vector $\boldsymbol{z}$ given observed data $\boldsymbol{x}$ can be calculated by Bayes' rule and is given by

$$p(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{M}^{-1}\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{\mu}), \sigma^{-2}\boldsymbol{M}\right) \tag{4}$$

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

Choosing number of principal components by 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 it is obtained by either observing eigen value spectrum of the observed data or by using subjective criteria like retaining 95% variance in the original data. 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 $p(\boldsymbol{x} \mid q)$ 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 (2000) proposed non informative prior distribution for $\boldsymbol{\mu}$ and conjugate priors for $\boldsymbol{W}$ and σ. Using these priors he derived an analytical expression for $p(\boldsymbol{x} \mid q)$. The optimal number of principal component $\hat{q}$ can then be obtained by using Bayesian model selection rule as

$$\hat{q} = \text{argmax}_q \left[p\left(\boldsymbol{x} \mid q \right) \right], \;\; 1 < q < p \tag{5}$$

Feature selection by support vector machine

SVM based feature selection was used in this work to select the best features from PCs extracted in the foregoing analysis. The SVM is a constructive learning procedure based on statistical learning theory. Unlike other learning procedures that emphasize only on minimizing training error, SVM implements structural risk minimization principle that attempts to minimize an upper bound on the training error.

Consider finite training samples of N patterns $\left\{ \left(\boldsymbol{z}_i, y_i \right), i = 1,2,\mathbf{K}, N \right\}$, where $\boldsymbol{z}_i$ denotes the i^{th} pattern in q dimensional space (i.e. $\boldsymbol{z}_i \in \Re^q$) and y_i $\left(y_i \in \{-1,+1\} \right)$ represents the class label of that pattern. SVM algorithm learns a linear hyper-plane

$$f\left(\boldsymbol{z} \right) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{z} + b \tag{6}$$

such that

$$y_i \left[\boldsymbol{w}^{\mathrm{T}} \boldsymbol{z}_i + b \right] \geq 1 - \xi_i \quad i = 1,2,\mathbf{K}, N \tag{7}$$

$$\xi_i \geq 0 \tag{8}$$

where ξ_i is positive slack variable, $\boldsymbol{w} \in \Re^q$ and $b \in \Re$ are adjustable model parameters. Among all possible linear hyper-planes that satisfy Eqs. (6), (7) and (8), SVM finds the one that maximizes the margin, i.e. the minimum distance between the hyper-plane and the closest data points. Results from statistical learning theory states that the hyper-plane that maximizes the margin also minimizes the generalization error (Scholkopf and Smola, 2002). In SVM construct, this is done by minimizing the following cost function

$$\Psi\left(\boldsymbol{w}, \xi \right) = \frac{1}{2} \boldsymbol{w}^{\mathrm{T}} \boldsymbol{w} + \sum_{i=1}^{N} C_i \xi_i \tag{9}$$

subject to constraints given by Eqs. (7) and (8). The first term in Eq. (9) is the reciprocal of the margin and it controls the model complexity, while the second term is the penalty term that penalizes misclassification in the training data. The parameter C_i is a positive real constant that controls the trade-off between these two terms of the cost function: model complexity (variance) and training error (bias). It is also the only free parameter to be chosen while developing an SVM. The procedure used for selecting C_i in this study is described later in the paper.

The constrained quadratic optimization problem given by Eqs. (7), (8) and (9) can be solved by the method of Lagrange multipliers, from which model parameters $\boldsymbol{w}$ and b are obtained as

$$\boldsymbol{w} = \sum_{i=1}^{N} \left(\alpha_i y_i \boldsymbol{x}_i \right) \quad \text{and} \quad b = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i \right) \tag{10}$$

where, α_i is Lagrange multiplier.

SVM is used in the work as a tool to select best features from the features extracted by PPCA. This is done by using the SVM recursive feature elimination (SVM-RFE) algorithm proposed by Guyon et al. (2002). The algorithm is based on the argument that the magnitude of the weights multiplying input features of a trained SVM can be used to rank those features. The feature that is weighted by the largest value has the maximum influence on the classification decision by SVM. Therefore, if the SVM performs well, the features corresponding to largest weights are also the most informative features.

SVM-RFE is an iterative algorithm. In each iteration, an SVM is trained and a feature corresponding to the smallest magnitude of weight is eliminated from the training set. During this process individual features get ranked. The features that are eliminated in the start get low ranks while features that are removed in the last get top ranks. However, it should be emphasized that the top rank features (eliminated in the last) are not necessarily the best ones individually. Only when taken together as a subset are they optimal. In this study, the size of subset was determined using leave-one-out cross-validation. Readers are refereed to Guyon et al. (2002) for implementation details of SVM-RFE algorithm.

It is to be mentioned that SVM can handle non-linear hyper-planes using kernel functions. However, this study was restricted to linear SVMs because of the nature of data sets under investigation.

Data used in this study

Monthly 5° resolution global SST data from 1850 onwards was obtained from the Second Hadley Centre Sea Surface Temperature dataset (HadSST2) (Rayner et al., 2006). This data set is updated on the 10th day of every month and can be obtained from Hadley Centre's website http://hadobs.metoffice.com/hadsst2/. The data is based on recently created International Comprehensive Ocean Atmosphere Data Set (ICOADS) database. Prior to 1950, ICOADS data is mostly based on observations from ships at sea. Due to lack of standard measurement techniques at that time, the SST data for that period is less coherent. However from 1950s onward, ICOADS contains data from the World Ocean database which was collected from ocean profilers and ocean stations using standard methods. Therefore SST data after 1950s is more reliable.

The area weighted rainfall data for homogenous regions in India as well as for all India (Parthasarathy et al., 1994), was extracted from Indian Institute of Tropical Meteorology, Pune, web site http://www.tropmet.res.in. The data extends from January 1948 to December 2004. Primary source of the data is India Meteorological Department. The time series of summer monsoon rainfall for each region was derived by adding monthly rainfall values for monsoon months (JJAS). The geographic location of homogeneous regions in India is shown in Figure 1.

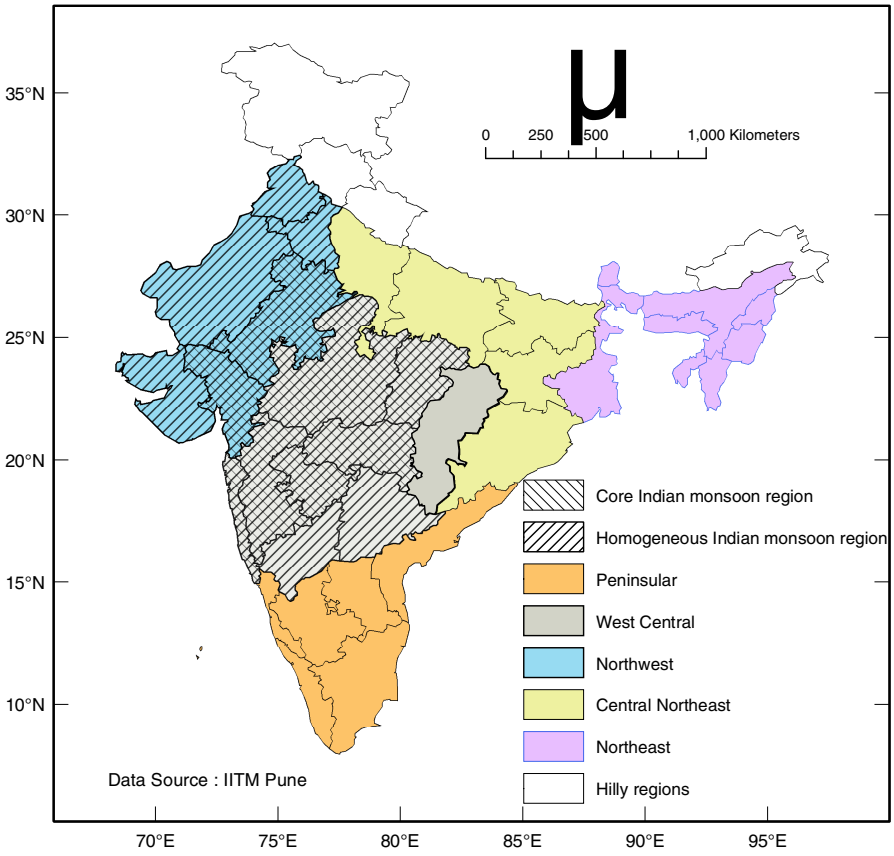

Figure 1. Homogeneous regions in India

Following India Meteorological Department definition, a year in each region was classified as a drought year if the rainfall in that year is less than 90% of the long term mean and as a flood (excess) year if it exceeds 110% of the long term mean.

Methodology

This section outlines the procedure involved in processing SST data to identify SST patterns that explain the variability associated with ISMR data.

As a first step, SST data were screened, to remove those grid points that have less than 50 years of record. The screened data were then centered and standardized. After that they were segregated into a training set (1971-1980) and a test set (1980-2004).

Further, following Lau et al. (2002), the screened data were partitioned into five non-overlapping sectors - Tropical Pacific sector (30°S - 30°N), North Pacific sector (north of 30°N), Tropical Atlantic sector (30°S - 30°N), North Atlantic sector (north of 30°N), and the Indian Ocean sector (north of 30°S). 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. SST data in the training set of each sector were then processed using PPCA to extract PCs. The number of PCs was determined using Bayesian model selection.

SVM-RFE algorithm was then applied on the extracted PCs to find most relevant features. The application of the algorithm involves selection of model parameter C. Since the SST data prior to 1950 is less reliable a small value (C=40) was assigned to those years while a larger value (C=80) was used for the years after 1950.

Finally, to study the nature of relationship between the features selected by SVM-RFE algorithm and ISMR, an SVM based forecasting model was developed. The selected features form the input to the forecasting model, while the rainfall series classified as drought/non-drought or flood/non-flood year forms the output. The generalization performance of the developed model was measured on the independent test set.

Results and Discussion

The number of grid points over different sectors that remained after screening analysis and were used further in the study are listed in Table 1. Typical results of PPCA are provided in Figure 2. It is evident form the figure that Bayesian model selection can reasonably determine the number of principal components. The second row in Table 1 gives the number of principal components extracted for different sectors for January month. It is evident from the table that the dimensionality of principal vector space is substantially smaller than the original dimensionality of the SST data. This reduction in dimensionality helps in making feature selection algorithm (SVM-RFE) more stable while reducing the computational burden at the same time.

Table 1. First row gives the spatial distribution of SST data over different sectors. Second row presents the number of PCs identified by the Bayesian model selection for January month.

	South Pacific	North Pacific	South Atlantic	North Atlantic	Indian Ocean
Number of grid points	320	112	161	143	165
Number of PCs	21	28	19	23	16

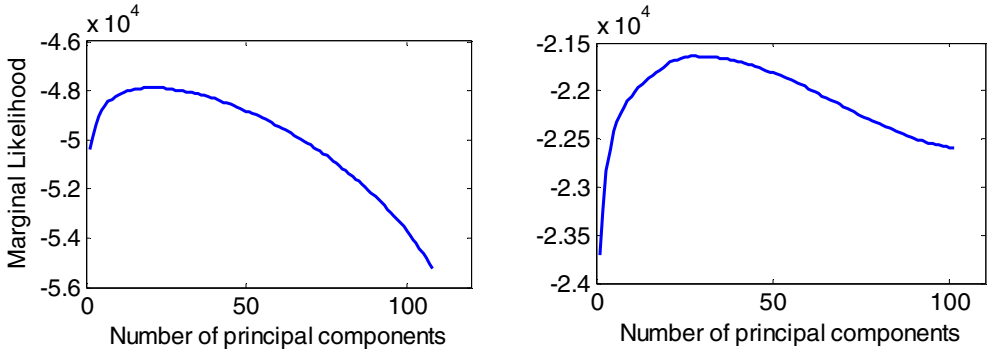

Figure 2. Identification of number of principal components using Bayesian model selection for (a) South Pacific sector (b) North Pacific sector for January month.

Typical results from the SVM-RFE based feature selection algorithm are presented in Figures 3 and 4. These figures show the spatial pattern of selected features using shaded contour plots. Spatial pattern of a selected feature used for discriminating drought year from a non-drought year in all India summer monsoon rainfall series is illustrated in Figure 3. It corresponds to SST values over Indian Ocean region in the month of March i.e. two months prior to the start of Monsoon season. The pattern indicates that the SST anomalies over Arabian Sea have a significant role in deciding whether the forthcoming all India summer monsoon will be a drought or a non-drought year. It is interesting to note that similar patterns were also deemed relevant for other homogenous regions of India. Apart from Indian Ocean sector, the features from North Atlantic sector in winter months were also frequently selected by the algorithm for identifying drought years.

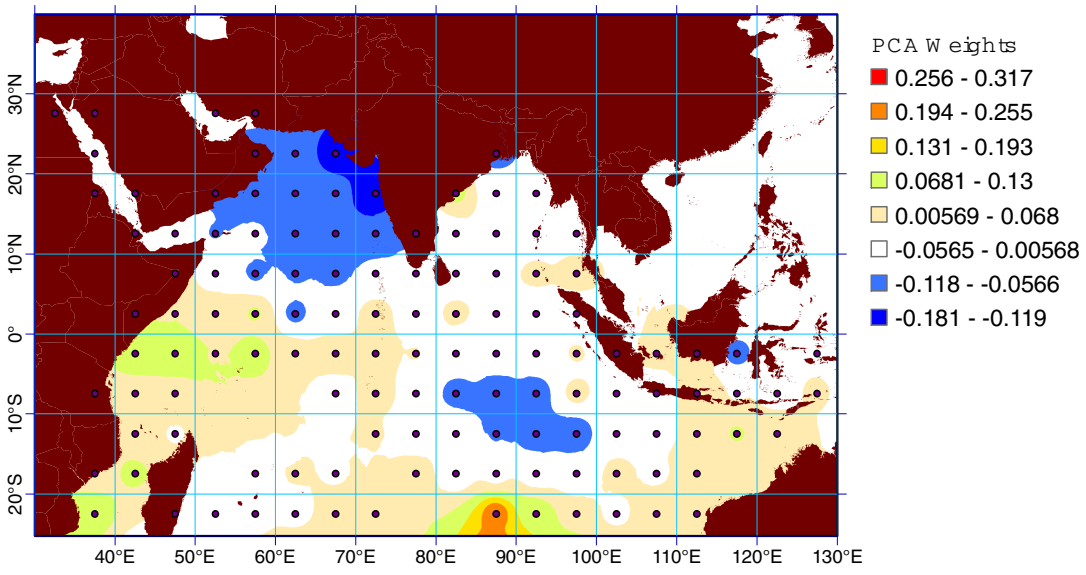

Figure 3. Spatial pattern of a selected feature used for discriminating drought year from a non-drought year, in all India summer monsoon rainfall series (AISMR). The pattern corresponds to the month of March.

For discriminating the flood years from the non-flood years, the proposed method invariably gave top ranks to the features from Pacific Ocean. This was true for all homogenous regions in India. A typical spatial pattern for the selected features is shown in Figure 4. The pattern in the figure indicates that a see-saw type behavior occurring in South Pacific Ocean during April can be useful to separate flood years from a non-flood years. Another commonly selected feature for this case was from North Atlantic sector.

To investigate the potential of selected features in forecasting ISMR, an SVM based classification model was developed for each homogeneous region in India. The model was trained using 109 years (1872-1980) of historical record and was tested on 24 years (1981-2004). The results from the analysis are given in Table 2. The results show that although the model performs very well in training phase, it performs poorly during testing phase. There can be two possible reasons for it. Either, the model is over-trained or the relationship between the identified features and ISMR is dynamical. The number of support vectors is a good indicator of the generalization performance of an SVM model (Scholkopf and Smola 2002). Smaller the number of support vectors, better the generalization performance. For the SVMs trained in this study, the number support vectors were around 10% of the total training vectors. This indicates that the SVMs are not over-trained, but it would appear that the relationship between ISMR and SST patterns is changing with time.

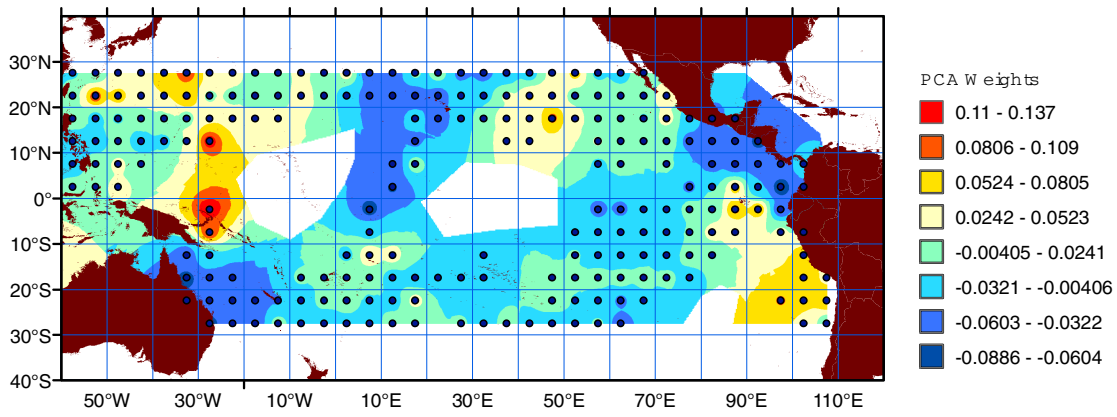

Figure 4. Spatial pattern of a selected feature used for discriminating flood year from a non-flood year, in all India summer monsoon rainfall series. The pattern corresponds to the month of April.

Table 2. SVM forecasting results for all India summer monsoon rainfall (AISMR) along with seven homogeneous regions of India. *nPC* denotes the number of principal components selected by feature selection algorithm. In Drought cell, column *pos* refers to drought years while *neg* represents non-drought years. Similarly in Flood cell, columns *pos* and *neg* means flood years and non-flood years, respectively. Quantity *a*/*b* indicate that out of *b* events SVM can classify *a* events correctly.

Homogeneous Regions	Drought					Flood				
		Training		Testing			Training		Testing	
	nPC	*pos*	*neg*	*pos*	*neg*	*nPC*	*pos*	*neg*	*pos*	*neg*
AISMR	7	16/17	86/92	3/6	16/18	4	16/16	80/93	1/3	18/21
Homogeneous Monsoon	8	20/20	86/89	4/10	11/14	7	28/35	64/74	3/4	15/20
Core Monsoon	7	23/24	74/85	2/10	13/14	6	27/34	61/75	3/4	11/20
North West	5	29/37	56/72	5/9	9/15	12	42/42	67/67	5/7	12/17
West Central	6	19/21	76/88	2/10	8/14	7	29/32	65/77	3/4	11/20
Central Northeast	6	18/19	70/90	3/4	13/20	9	15/15	94/94	2/4	16/20
Northeast	3	13/14	72/95	3/7	14/17	4	11/11	88/98	2/4	14/20
Peninsular	13	26/26	83/83	4/9	11/15	6	22/25	63/84	4/7	11/17

Conclusions

In this paper a two step approach was adopted to identify SST patterns that can discriminate extreme states in hydrologic time series. The effectiveness of the approach is illustrated through its application to ISMR data. The proposed method can select a small subset of SST patterns that can distinguish drought from a non-drought and flood from a non-flood year. Some of the features selected by the method closely resemble patterns well established in literature while others do not. Future studies will be directed towards exploring physical interpretation for the selected patterns.

A preliminary investigation into the relationship between selected features and ISMR indicate that the relationship between them is dynamical in nature. Therefore dynamic forecasting model is needed to harness the potential of selected features in forecasting ISMR. Extended research work in this direction is underway

References

Gadgil, S., and Sajani, S. (1998). "Monsoon precipitation in the AMIP runs." *Clim. Dynam.*, 14(9), 659-689.

Gadgil, S., Rajeevan. M., and Nanjundiah. R. (2005). "Monsoon prediction - Why yet another failure?" *Curr. Sci. India.*, 88(9), 1389-1400.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). "Gene selection for cancer classification using support vector machines." *Mach. Learn.*, 46, 389-422.

Lau, K.M., Kim, K.M. and Shen, S.S.P. (2002). "Potential predictability of seasonal precipitation over the United States from canonical ensemble correlation predictions." *Geophys. Res. Lett.*, 29(7), 1097, doi:10.1029/2001GL014263

Minka, T.P. (2000). "Automatic choice of dimensionality for PCA," *Advances in Neural. Information Proces. Systems*, In T.K. Leen, T.G. Dietterich, and V. Tresp, eds., vol. 13, 598-604.

Parthasarathy, B., Munot, A. A., and Kothawale, D.R. (1994). "All India monthly and seasonal rainfall series: 1871-1993." *Theor. Appl. Climatol.*, 49, 217-224.

Pai, D.S., and Rajeevan, M. (2006). "Empirical prediction of Indian summer monsoon rainfall with different lead periods based on global SST anomalies." *Meteorol. Atmos. Phys.*, 92(1-2), 33-43.

Rayner, N.A., Brohan, P., Parker, D.E., Folland, C.K., Kennedy, J.J., Vanicek, M., Ansell, T., and Tett, S.F.B. (2006). "Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: The HadSST2 dataset." *J. Climate*, 19(3): 446-469.

Roweis, S. (1998) "EM algorithms for PCA and SPCA." In M.I. Jordan, M.J. Kearns, and S.A., Solla, eds, *NIPS*, vol 10, 626-632, The MIT Press.

Sahai, A.K., Grimm, A.V., Satyan, V., and Pant, G.B. (2003). "Long-lead prediction of Indian summer monsoon rainfall from global SST evolution." *Clim. Dynam.*, 20 (7-8), 855-863.

Scholkopf, B., and Smola, A.J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Mass.

Tipping, M.E., and Bishop, C.M. (1999). "Probabilistic principal component analysis." *J. Roy. Stat. Soc. B*, 61(3), 611-622.