Journal of Hydrology: Regional Studies 27 (2020) 100651

Contents lists available at ScienceDirect

Journal of Hydrology: Regional Studies

journal homepage: www.elsevier.com/locate/ejrh

Bringing statistical learning machines together for hydro-climatological predictions - Case study for Sacramento San joaquin River Basin, California

Balbhadra Thakur[a], Ajay Kalra[a,*], Sajjad Ahmad[b], Kenneth W. Lamb[c], Venkat Lakshmi[d]

[a] Department of Civil and Environmental Engineering, Southern Illinois University, 1230 Lincoln Drive, Carbondale, IL, USA
[b] Department of Civil and Environmental Engineering and Construction, University of Nevada, Las Vegas, NV, USA
[c] Department of Civil Engineering, California State Polytechnic University Pomona, CA, USA
[d] Department of Engineering Systems and Environment, University of Virginia Charlottesville, VA, USA

ARTICLE INFO

Keywords:
Streamflow
Forecast
Climate variability
SVD
SVM
KNN
Teleconnections

ABSTRACT

Study region: Sacramento San Joaquin River Basin, California
Study focus: The study forecasts the streamflow at a regional scale within SSJ river basin with largescale climate variables. The proposed approach eliminates the bias resulting from pre-defined indices at regional scale. The study was performed for eight unimpaired streamflow stations from 1962–2016. First, the Singular Valued Decomposition (SVD) teleconnections of the streamflow corresponding to 500 mbar geopotential height, sea surface temperature, 500 mbar specific humidity ($SHUM_{500}$), and 500 mbar U-wind (U_{500}) were obtained. Second, the skillful SVD teleconnections were screened non-parametrically. Finally, the screened teleconnections were used as the streamflow predictors in the non-linear regression models (K-nearest neighbor regression and data-driven support vector machine).
New hydrological insights: The SVD results identified new spatial regions that have not been included in existing predefined indices. The nonparametric model indicated the teleconnections of $SHUM_{500}$ and U_{500} being better streamflow predictors compared to other climate variables. The regression models were capable to apprehend most of the sustained low flows, proving the model to be effective for drought-affected regions. It was also observed that the proposed approach showed better forecasting skills with preprocessed large scale climate variables rather than using the predefined indices. The proposed study is simple, yet robust in providing qualitative streamflow forecasts that may assist water managers in making policy-related decisions when planning and managing watersheds.

1. Introduction

Streamflow forecast has been given notable focus by the hydrologists and water managers in the past decades (Chang and Chen, 2003; Besaw et al., 2010; Yaseen et al., 2016). The uncertainty in the streamflow resulting from the hydro-climatic variability demands skillful forecasting of streamflow. Further, water stress resulting from increasing population and climate change makes

* Corresponding author at: Department of Civil and Environmental Engineering, Southern Illinois University, 1230 Lincoln Drive, Carbondale, IL 62901-6603, USA.
E-mail address: kalraa@siu.edu (A. Kalra).

https://doi.org/10.1016/j.ejrh.2019.100651
Received 7 November 2018; Received in revised form 30 November 2019; Accepted 1 December 2019
Available online 10 December 2019

streamflow forecasting more important to enhance future water management decisions. Several water scarcity problems recently have been encountered due to climate variability and change, reduction in available water, and increased water demand (Trenberth, 2001; Schewe et al., 2014; Mehran et al., 2015). Additionally, watersheds already facing water scarcity are more vulnerable to further climatic variations (Mehran et al., 2015). Increase in the temperature and change in precipitation patterns have imposed severe droughts in the western US (Crockett and Westerling, 2018) such as the California drought in 2012. Improvements on seasonal streamflow forecasting at regional scales in hydrologically sensitive regions of California can be beneficial to the community and literature.

Comprehending future streamflow behavior and improving long lead forecasts have been a major goal for hydrologists and researchers for solving problems related to rising water demand and reducing water availability. However, streamflow prediction is a challenging task because streamflow has a complicated, nonlinear, and nonstationary nature, resulting from uncertainties in the underlying hydrologic mechanisms (Zealand et al., 1999; Wang et al., 2006). The fluctuations in streamflow greatly depend on climate variability and change, and climatic fluctuations and its hydrologic effects pose increased uncertainties associated with water supply and its management (Frederick and Major, 1997; Middelkoop et al., 2001; Maurer, 2007). Since a major proportion of domestic water depends on the surface water in streams and rivers, increase in population adds to the demand for freshwater and generates management issues of freshwater streams. Thus, determining the interconnection amidst the streamflow and the climatic indices and using them for flow forecasts can help water managers in decision-making and can minimize the effects of extreme hydrologic events in the future.

While forecasting any hydrologic variables the first step is to evaluate the historic correlations between the variables and the climate predictors. Studies have shown the teleconnection between continental U.S. hydrology and large-scale climatic indices (Enfield et al., 2001; Wei and Watkins, 2011; Tamaddun et al., 2017). The existence of these teleconnection patterns has appealed to scientists in order to improve long-lead forecasts. Some popular spatial teleconnections among water resource personnel, also known as predefined indices, are El-Niño southern oscillations (ENSO), Pacific decadal oscillations (PDO), and Atlantic multidecadal oscillations (AMO). Various studies such as Enfield et al. (2001); Hidalgo-Muñoz et al. (2015); and Ehteram et al. (2019) have identified the teleconnections of these predefined indices with the hydrology of various regions, and often are utilized in hydrologic predictions at various scales (Hamlet and Lettenmaier, 1999, 2007; Kalra and Ahmad, 2012; Ehteram et al., 2019). Although these indices have proved to be effective in forecasting, they have associated regional biases. Instead of using predefined indices, the use of large scale climate variables can reduce the spatial biases related to the regional effects of these predefined indices (Tootle and Piechota, 2006). Thus, identifying and utilizing new teleconnected regions that affect regional hydrology is more beneficial.

Sea surface temperature (SST) influences evaporation in oceans and pressure variations above the sea's surface, which leads to atmospheric circulation (Omondi et al., 2013). 500-mbar geopotential height (Z_{500}), an atmospheric variable used for referencing pressure regimes (Wallace and Gutzler, 1981), provides an understanding about the pressure gradient acting as the driving force for the moisture circulation. The winter snowpack and resulting spring streamflow in the western US highly depend upon the circulation of moisture from the Pacific thus, Z_{500} can be streamflow forecaster within western United States. Wallace and Gutzler (1981) used geopotential regimes for various mbar pressure regimes to teleconnect the winter of the northern hemisphere. Studies also examined the effects of SST and Z_{500} on hydrological parameters (Sagarika et al., 2016; Pathak et al., 2018). In addition to SST and Z_{500}, other climate variables, such as specific humidity corresponding to Z_{500} height ($SHUM_{500}$), and east-west wind (i.e., U-wind) corresponding to Z_{500} height (U_{500}), are directly related the atmospheric moisture circulation and may improve the streamflow forecasts. U_{500} is the U-wind force per unit area present in the atmosphere, while U_{500} sheds light on the circulation of atmospheric moisture. Previously, Pathak et al. (2018) have Additionally, $SHUM_{500}$ which entails the information about atmospheric humidity is also associated with the winter snow precipitation. As the western US streamflow is primarily driven by winter snowpack testing the prospect of regional streamflow predictions utilizing newly evaluated SST and Z_{500} teleconnections in coordination with that of U_{500} and $SHUM_{500}$ can add to the previous literatures.

Evaluating the teleconnection amongst aforementioned large scale climate variables (SST, Z_{500}, $SHUM_{500}$ and U_{500}) and streamflow can remove the regional bias in the streamflow prediction at regional scale. Singular Valued Decomposition (SVD) is widely used as a reliable statistical tool to find the teleconnections of two different spatiotemporal climate variables, and can be used to relate oceanic-atmospheric variabilities with hydrologic variables. It is one of the preferred methods of principal component analysis (Bretherton et al., 1992) to evaluate the teleconnections in hydro-climatological parameter. Unlike other principal component analysis (PCA), SVD is beneficial as it can be applied to a rectangular cross covariance matrix making it able to evaluate the primary modes of correlation between two variables of different spatiotemporal dimensions. SVD mostly is used to interrelate hydrologic parameter like streamflow and snowpack to various oceanic-atmospheric parameters (Sagarika et al., 2016; Pathak et al., 2018). SVD is a powerful multivariate tool that can simultaneously recognize the oscillations in spatiotemporal data that can be used to find frequencies in order to determine the occurrence of narrowband variance (Rojsiraphisal et al., 2009). Utilization of SVD can effectively substitute the use of predefined indices.

Statistical prediction of hydrologic time series are mostly associated with the historic observations. Multiple statistical regression techniques such as multiple linear regression, autoregressive moving average models are not effective in considering non-linear nature of hydro-climatologic phenomenon associated to streamflow. Artificial intelligence such as Artificial Neural Network (ANN) and Support Vector Machine (SVM) performs better while performing nonlinear time series analysis (Zhang et al., 2018). Furthermore, studies have shown that SVM are more skillful than ANN as it less prone to the overfitting (Kalra et al., 2013). SVMs are data-driven machine-learning models that use structural risk minimization to reduce anticipated error during learning and minimizes the problem of overfitting (Li et al., 2009). SVMs perform nonlinear mapping of the input data to a higher dimension feature space where the regression is performed. SVM initially was formulated for classification and pattern recognition problems in statistical datasets

(Vapnik, 1995, 1998, 2013). Later, this method evolved to include nonlinear regression, and has been applied to hydrologic forecasting by many researchers (Sivapragasam et al., 2001; Asefa et al., 2006; Hong, 2008; Simões et al., 2011; Kalra and Ahmad, 2012; Nikam and Gupta, 2013). Like SVMs another widely used regression technique is K-nearest neighbor (KNN) which fits the data locally, making it able to capture both linear and nonlinear variability. Unlike most of the linear regression models which assumes the Gaussian distribution of errors, KNN regression is a non-parametric technique which estimates the predictand locally making it able to capture both linear and nonlinear relations between predictors and predictand (Grantz et al., 2005).

Previous studies have extensively utilized predefined indices to forecast the streamflow which can result in the spatial bias while forecasting the streamflow at regional scale. The current research focuses on forecasting the regional streamflow with largescale climate variables. First, SVD was utilized to identify spatial teleconnections between the spring streamflow and largescale climate variables. Next, the modes, as indicated by SVD, were weighted and screened using a kernel density estimator model to generate continuous exceedance probability curves. Finally, weighted predictors were used as input in the regression models- KNN and data-driven SVM to provide regional scale streamflow forecasts. The major research questions addressed by the current study are: (1) What are the spatiotemporal teleconnections between the largescale climate variables and the streamflow at regional scale? (2) Which climate variables are more skillful for streamflow forecasting? (3) Is the proposed forecasting approach skillful as compared to the streamflow forecast with predefined indices? SST and Z500 are more popular for streamflow forecast while the current study also tests the applicability of $SHUM_{500}$ and U_{500} for predicting the streamflow.

2. Materials and methods

2.1. Study area and data

The hydrology of the western U.S. is expected to be affected mostly due to changing climate as the region's runoff cycle is influenced mostly by snowmelt; specifically, it is anticipated that winter snow accumulation will vary due to the change in climate and melt early in the spring. A projected temperature rise of 2.1 °C by 2090 is expected to reduce April snowpack, thereby reducing spring runoff by 5.6 km^3 in Sacramento and San Joaquin regions of California (Knowles and Cayan, 2002). The streamflow forecast in a region like California, which has a greater chance of prolonged droughts, facilitates planning the management of reservoirs, agriculture, flood, and drought (Moradkhani and Meier, 2010). The current study was performed in two U.S. Geological Survey

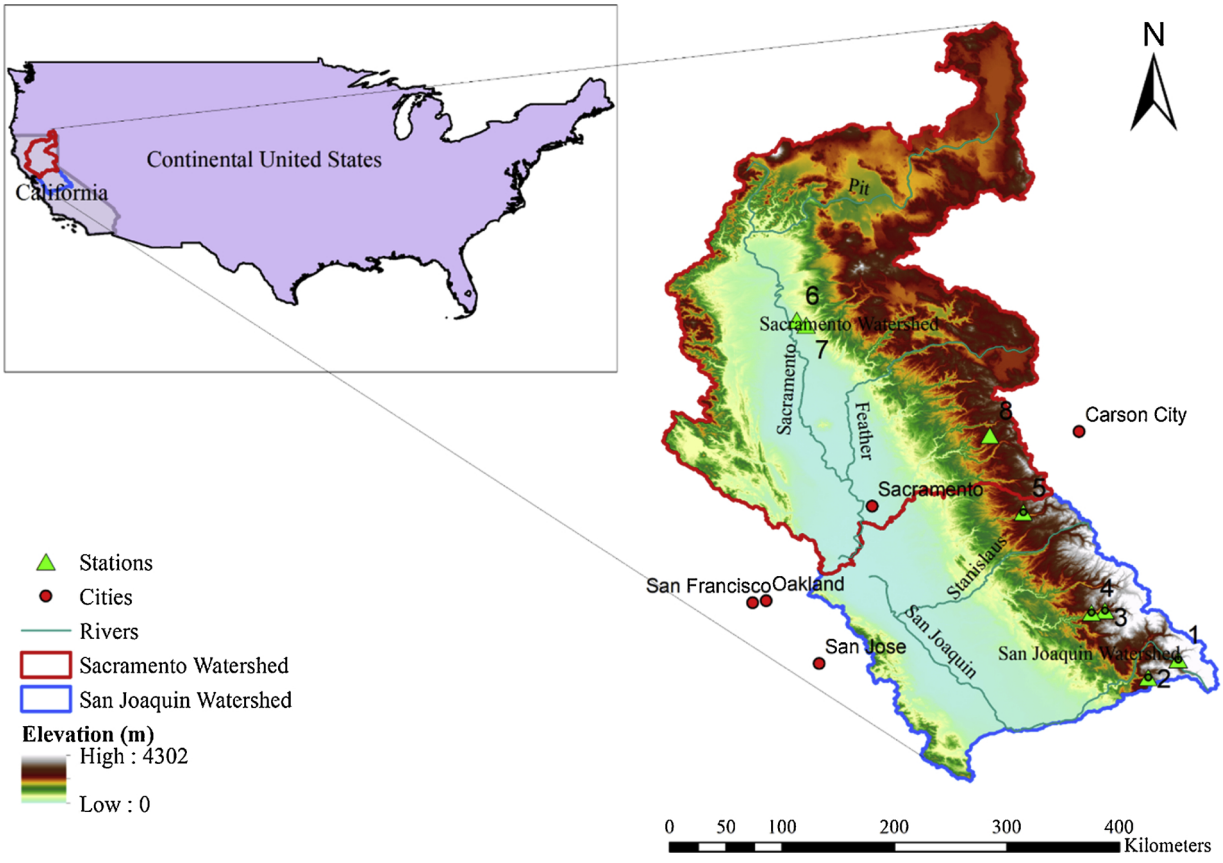

Fig. 1. Study area, encompassing the Sacramento-San Joaquin River Basin, streamflow stations, and major cities within the watershed.

(USGS) hydrologic regions, Sacramento and San Joaquin (SSJ) watershed in California, as shown in the Fig. 1. The Sacramento watershed is the upper region in the figure, and the San Joaquin watershed is the lower region. SSJ watershed is major water resource to the San Francisco estuarine system. SSJ watershed, encompassing a major region of cropland in California, is driven mostly by winter precipitation resulting in spring-summer streamflow. Being a snowpack driven watershed SSJ watershed may be affected significantly by climate-related flow variations (VanRheenen et al., 2004). Forecasting the season-ahead streamflow could help in developing strategies to mitigate the effects of droughts in this region.

Data for the hydrologic parameter (i.e., streamflow) of the study area was procured from the Hydro-Climatic Data Network (HCDN-2009) (Lins, 2012). The gaging stations included new stations, which are part of the Gage II dataset (Falcone, 2011). The stations were unimpaired; in other words, the watersheds above the gages did not have any anthropogenic impacts. Data for monthly average streamflow were retrieved from the online database of USGS (http://www.usgs.gov/). The obtained data was monthly streamflow data (i.e., April through June) from 1962–2016. These were used to determine the seasonal spring streamflow volume which is used in the current study. The data for monthly averaged winter SST from December to February of 1962–2016 (i.e., for 55 years) was obtained from the Physical Science Division of the National Oceanic and Atmospheric (NOAA) Administration's Earth System Research Laboratory (ESRL) (http://www.esrl.noaa.gov/psd/data/gridded). The included SST is the most recent dataset of its type available as “NOAA extended and reconstructed V5″ in the aforementioned website.

Previous researchers, such as Hidalgo (2004) and Ellis et al. (2010) have correlated variabilities of the Atlantic Ocean to the hydrology of the western U.S. This current study also incorporated the Atlantic oceanic-atmospheric parameters in addition to Pacific variabilities. The SST data were obtained for each 2 × 2-degree grid cell of the Pacific Ocean (30S to 70 N and 100E to 80 W) and the Atlantic Ocean (30S to 70 N and 80 W to 20 W). Altogether, there were 4581 such grids whose SST data were extracted.

The most recent average monthly Z_{500}, $SHUM_{500}$, and U_{500} (December - February) from 1962–2016 (i.e., for 55 years) were obtained from (https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.derived.pressure.html). The Z_{500}, $SHUM_{500}$, and U_{500} data were obtained for each 2.5 × 2.5-degree grid cells of the Pacific Ocean and the Atlantic Ocean. Altogether, there were 4018 such grids in the Pacific and Atlantic Oceans whose Z_{500}/$SHUM_{500}$/U_{500} data were extracted.

First, the spatiotemporal teleconnections were made of various climate predictors (SST, Z_{500}, $SHUM_{500}$, and U_{500}) with the hydrologic predictands (i.e., streamflow), using SVD. SVD analysis yielded the predictors of streamflow as a temporal expansion series (TES). Next, the TES predictors were screened based on nonparametric screening. Next, the probabilistic forecasts were made with the screened TES. Finally, forecasting of the streamflow was conducted using SVM/KNN with the screened predictors. All the aforementioned methods are discussed in the following sections.

2.2. SVD analysis

Bretherton et al. (1992) compared various approaches utilized in relating multiple fields of climate data, and found that singular valued decomposition (SVD) is easy to implement and performs well for large time series. SVD was used as a robust statistical tool to detect the spatiotemporal interrelationship between a predictor (i.e., climatic variability- SST, Z_{500}, $SHUM_{500}$, and U_{500}) and a predictand (i.e., the hydrologic parameters of streamflow and precipitation) from their cross-covariance matrix (Sagarika et al., 2015). Details regarding SVD can be obtained from Bretherton et al. (1992). The standardized SST, Z_{500}, $SHUM_{500}$, and U_{500} anomalies matrices along with standardized streamflow matrices were developed first before evaluating cross-covariance matrices over which the SVD analysis was performed. The standardized climate variable matrix were of the order M × T where T –the number of columns was equal to the temporal dimension of the data here, 55 years. Similarly, M is equivalent to the number of grid cells. The standardized streamflow matrix was of the order N × T where T is the temporal dimension and the total number of streamflow stations is expressed by N. The temporal dimensions of standardized SST, Z_{500}, $SHUM_{500}$, and U_{500} anomalies matrices along with standardized streamflow matrices were same even though they varied in size spatially.

SVD decomposes the covariance matrix M as USV^T, where U and V often are the representation for the left matrix and the right matrix. S is a central diagonal matrix that comprising non-zero singular values. Eight different left matrices and eight different right matrices were obtained as a result of an independent SVD analysis of each variable for both the Pacific and Atlantic Ocean with the streamflow. The singular values were organized in decreasing order; the first value was termed as the first mode, and so on (Bretherton et.al, 1992; Sagarika et.al, 2015). The importance of different SVD modes can be explained by the Squared Covariance Factor (SCF). SCF is the ratio of a corresponding singular value squared to the squared sum of singular values of all the modes (Bretherton et al., 1992) and is summarized in Eq. (1).

$$SCF_i = \frac{S_i^2}{\sum_{j=1}^{r} S_i^2} \tag{1}$$

Where, the i^{th} mode SCF is indicated by SCF_i and S_i is the i^{th} mode singular value. Similar to SCF, Normalized Square Covariance (NSC) gives the idea about the correlation among the significantly teleconnected predictors and predictand for each mode. It gives the idea about how well each correlated grid cells of climatic variable is linearly associated to significantly teleconnected streamflow stations. Mathematically, NSC is the ratio between singular value squared sums and product of significantly teleconnected number of streamflow stations and grid cells of corresponding climate variables as expressed in Eq. (2).

$$NSC = \frac{C^2}{N_s N_P} \tag{2}$$

where, C signifies the squared singular values sum, N_s is the number of streamflow sites and N_P is the total number of gridded data in any predictor sets.

After the SVD results were obtained, left TES of first mode were obtained projecting U matrix on standardized climate variable matrix corresponding to first mode. Similarly, right TES of the first mode were obtained from V matrix and standardized streamflow matrix. Left TES and Right TES together holds the consolidated information about primary modes of correlations between streamflow and climate data. Further, Left TES retains the primary modes SST, Z500 $SHUM_{500}$, and U_{500} variability which are associated to the streamflow and which may be unique as compared to the predefined indices. Thus, utilizing the Left TES may improve the forecast at the regional scale and is also utilized by previous studies (Soukup et al., 2009). In this study, the first mode Left TES for most of the predictor had an SCF of more than 90% thus, explaining more than 90% of the variability. Thus, Left TES for the first mode – having the information about correlated indices of SST, Z_{500}, $SHUM_{500}$, and U_{500} – was utilized for forecasting. Altogether, SVD resulted in eight Left TES for all the climate variables considered.

2.3. Nonparametric approach for predictor screening

The nonparametric algorithm adopted in the current study is based on the method to forecast the streamflow proposed by Piechota et al. (1998). For more details readers are about the method referred to Piechota et al. (1998) and Soukup et al. (2009). This approach yields the continuous relation among the predictand and the predictors and the approach also doesn't presume any model structure. While the forecast is probabilistic and reports the forecasts at different risk level; the shortcoming of this approach is that it assumes the available data to represent the entire population. Thus, the predictor sets showing good skills with this model were selected for final predictions with SVM yielding quantitative yearly forecasts. As this non-parametric model is computationally less complex, it can handle multiple inputs in a single instant in shorter time duration as compared to SVM proving it to be more convenient screening algorithm. The steps involving nonparametric screening is discussed below.

Fig. 2 shows how SVD and SVM are coupled with nonparametric screening algorithm. The nonparametric screening of the predictors is summarized in the box with dotted boundaries of Fig. 2. The inputs to the nonparametric model is the TES resulting from decomposition of cross-covariance matrix with SVD and output of the nonparametric screening is the input for SVM. First, the probability distribution function (PDF) was generated for each independent predictor Left TES from SVD of each predictors Z_{500}, SST, $SHUM_{500}$, and U_{500} of both Atlantic and Pacific and Ocean utilizing the kernel density estimator (KDE) adopted by Piechota et al. (1998). To generate the pdf, the streamflow and their respective predictors Pi are ranked (highest to lowest) based on the streamflow (Qi) values. The first data point is the 6^{th} ranked as minimum of five data points are required to obtain the PDF based on the KDE. KDE is expressed as the following Eq. 3.

$$f(x) = \frac{1}{hn}\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right) \tag{3}$$

$where for gaussian kernel density estimator,\ h_i = 0.9A_i h_i^{-1/5}$

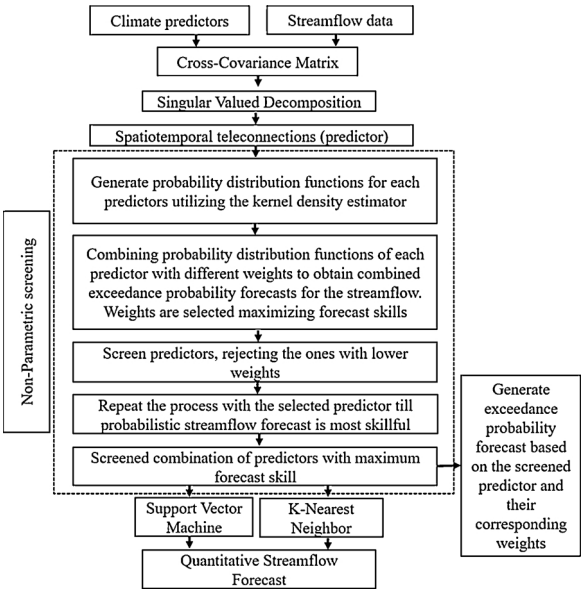

Fig. 2. A flowchart representation of adopted hybrid statistical modelling framework obtained by coupling singular valued decomposition and support vector machine/k-nearest neighbor with nonparametric screening approach.

$$A_i = min\left(\sigma_i, \frac{interquartilerange}{1.34}\right)$$

Here, bandwidth of the data is h; the number of data is expressed as n and K () is the kernel function. σ_i is the standard deviation of predictor data in i^{th} subset. Eight predictor's PDFs for each year's streamflow were obtained with KDE.

Once, the predictor PDFs were obtained the exceedance probability curve of the streamflow were obtained from the Bayes theorem for conditional probabilities as the following Eq. 4.

$$P\left(\frac{Q_i}{x}\right) = \frac{P_i f_i(x)}{\sum_{i=1}^{n} P_i f_i(x)} \tag{4}$$

P_i being the probability of occurrence of the streamflow Q_i based on the observed streamflow dataset. $f_i(x)$ is the probability of the predictor x based on the PDFs. $P\left(\frac{Q_i}{x}\right)$ is the probability of the streamflow Q_i for the predictor x. Later, the exceedance probability of the forecast is generated for any given year. The skill of the probabilistic exceedance forecast is evaluated with the linear error in probability space (LEPS) score. The skill for the entire forecast for all years is computed with the average of LEPS scores (SK) values. The mathematical representation of both LEPS and LEPS SK is presented in Section 3.4.

For screening the predictors, the continuous exceedance probability curve of each predictor was combined by assigning random weights to each predictor, and the sum of their weights were '1'. The forecasts were made with a combined continuous exceedance probability curve in order to optimize the weights assigned to each predictor. Optimization was done by using the LEPS SK scores of the forecast, which were higher for the better nonparametric forecasts (Ward and Folland, 1991; Potts et al., 1996). LEPS SK is further described in Section 2.6.

The predictors with lower weights (i.e., less than 0.1) were rejected. The combined continuous exceedance probabilistic forecasts were made again incorporating only the selected predictors, and the weights were again optimized for the higher LEPS SK scores.

The process was continued until the combination of predictors for each streamflow station was obtained with the maximum LEPS SK score. The final combination of predictors for each station was used in the SVM/KNN.

2.4. SVM modelling

The generalized expression for SVM regression technique was abstracted from Vapnik (1998) is expressed as:

$$f = w^T \Phi(x_i) + b, \tag{5}$$

where $\Phi(x_i)$ is the input and the coefficients w^T (weight factor) and b (bias term) are obtained by optimizing the risk function $R(f)$, which is expressed as Eq. (6).

$$R(f) = \frac{C}{N}\sum_{i=1}^{N}(\xi + \xi^*) + \frac{1}{2}\|w\|^2$$

$$Subjectto: y_i - \sum_{j=1}^{K}\sum_{i=1}^{L} w_j x_{ji} - b \le \varepsilon + \xi_i \tag{6}$$

$$\sum_{j=1}^{K}\sum_{i=1}^{L} w_j x_{ji} + b - y_i \le \varepsilon + \xi_i$$

$$\xi_i, \xi_i^* \ge 0$$

where, the cost (C) that is used for generating agreement between the tolerance of the prediction error and the complexity of the function. The number of support vectors are denoted by K, ξ_i, and ξ_i^*, which are the slack variables that are the deciding parameters that indicate the level to which the samples should be penalized for the error larger than Vapanik's insensitive loss function, ε.

The above discussed standard SVM models have associated problem of model complexity. To overcome this complexity issue least-square SVM (LSSVM) was incorporated in the current study. LSSVM improves upon the standard SVM by including equality constraints instead of the inequality one (Suykens and Vandewalle, 1999; Suykens et al., 2002; Kuh, 2004). The Eq. (6) of standard SVM gets reduced in LSSVM to the following Eq. (7).

$$\min J(w, e) = \frac{C}{2}\sum_{i=1}^{N} e_i^2 + \frac{1}{2}W^T W$$

$$Subjectto: y_i - (w^T \Phi(x_i) + b) = e_i; whereivariesfrom1toN \tag{7}$$

Eq. (7) shows the optimization expression in which $\frac{1}{2}W^T W$ helps to adjust the weight, and $\frac{C}{2}\sum_{i=1}^{N} e_i^2$ is the penalty function. The error in the i^{th} prediction of y_i is e_i. The optimization of Eq. (7) is done utilizing the Langrangian multipliers as shown in Eq. (8) below

$$f(x) = \sum_{i=1}^{N} a_i k(x, x_i) + b \tag{8}$$

Where, a_i is the Langrangian multiplier, and $k(x, x_i)$ is the kernel function, responsible for operating the data in higher dimension feature space without finding the data coordinates in the higher dimension. Some of the examples of the kernel functions are linear kernel, Gaussian kernel, radial basis kernel function (RBF). RBF kernel was utilized in the current study which have better skills over other kernel functions (Scholkopf et al., 1997; Kalra et al., 2013). The RBF kernel on samples x and x_i is expressed as the following Eq. (9).

$$k(x, x_i) = exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \tag{9}$$

The hyper-parameter σ is obtained by optimizing the error and improving the forecast in the training and testing period. The hyper-parameters are evaluated with grid-based search (Twarakavi et al., 2009) this removes uncertainty, which estimates best combination with minimum error as in Eq. (3). All sets of possible hyper-parameters is considered to calculate the error. The reasonable feature space is generated for each sets of hyper-parameter based on the possible lowest and highest values (cost function ranging from 0.001–1000 and the slack variable ranging from 0.001–100). For optimizing the grid based hyper-parametrization and for efficient computation the increment of 0.01 was selected and the optimization was done for the minimal mean square error. This hyper-parametrization results in dense clusters and also reduces the prediction uncertainty (Asefa et al., 2006; Kalra and Ahmad, 2012). Optimizing the hyper-parameter have large number of support vectors but a smooth model which does not mean overfitting.

Validating a data driven model is crucial in determining the forecast reliability. Traditional data driven models are divided into training and testing period which has its own disadvantage while dealing with periodic time series. As oceanic-atmospheric variables manifest different types of periodicity, breaking the data into training and testing periods may not reproduce comparable results for various training and testing data sets. Thus, to control the error caused by breaking the samples into training and testing periods for support vector regression, moving-period SVM was implemented thus, reducing the model sensitivity and prediction uncertainty (Kalra and Ahmad, 2012). For more details about moving period cross validation approach, readers are referred to Stone, 1974 and Geisser (1975). Instead of dividing the entire data into specific training and testing period k-fold moving period algorithm is utilized with $k = 1$ also known as leave-one-out technique. For the time series of N years, the data for one year was selected for testing while other mutually exclusive (N-1) year's data were used for training purpose. The process was repeated for the entire time series i.e. N times with mutually exclusive training period during each iteration. Final predictions were obtained by pooling all the predictions made by SVM with all training and testing periods.

In order to evaluate the strength of the SVM approach, and to access whether multiple training and testing samples do not reproduce varying outputs, the bootstrap technique was incorporated in this current study. The bootstrap approach, which was first developed by Efron (1979) and described further by Efron and Tibshirani (1993). Addition of bootstrap technique adds extra layer of validation to the SVM model.

It was implemented in this study to evaluate how well the SVM model results generalized while using an independent dataset. SVM was trained with a bootstrapped data sample, and then tested on the entire measured dataset. For a given dataset, if the size of the dataset is *n*, the bootstrap sample was generated by sampling *n* instances from the data with replacement. The bootstrap resampling process was iterated to obtain 100 combinations, corresponding to the measured value of each year for each station. Finally, the average of 100 combination results was calculated as the final estimation. For a reliable forecast, the SVM results should be consistent with and without implementation of bootstrap resampling approach.

2.5. KNN regression

Like SVMs another widely used regression technique is K-nearest neighbor. Unlike conventionally used linear regression models which assumes the Gaussian distribution of errors, KNN regression is a non-parametric technique which estimates the predictand locally making it able to capture both linear and nonlinear relations between predictors and predictand (Grantz et al., 2005).The linear regression can be expressed as the following Eq. (10).

$$Y = f(x_1, x_2, x_3, \ldots x_1) + e \tag{10}$$

where, f is the function fitting the predictor indices - $x_1, x_2, x_3, \ldots x_p$ to the predictand like streamflow (Y) with the error e assumed to be following Gaussian distribution. The non-parametric regression like KNN (Lall and Sharma, 1996; Rajagopalan and Lall, 1999); kernel based regression (Piechota et al., 1998); and Locally weighted polynomials (LWP) (Loader, 1999) on the other hand estimate f locally making them capable to model both local and global variations in the datasets.

The KNN regression in the current research utilized the properties of both KNN and LWP. First, the order of the polynomial and the neighborhood size are obtained based on generalized cross-validation function (GCV). The GCV and its application is expressed in following Eq. (11).

$$GCV(k, p) = \frac{\sum_{i=1}^{n} \frac{e_i^2}{N}}{(1 - \frac{m}{N})^2} \tag{11}$$

where, regression error in i^{th} data is represented as e_i, total data points number is represented by N, and number of parameters is represented by m. The nearest neighbors for each of the data point were obtained and the polynomials were fitted locally. The streamflow Y_{new} was obtained from the local fit for each i^{th} input sets. The error between i^{th} Y_{new} and i^{th} Y data is e_i. Higher weight is then assigned to the nearest neighbor as compared to less near neighbor. The weight for k nearest neighbor is distributed as shown in

following Eq. (12).

$$W(j) = \frac{1}{j\sum_{i=1}^{k}(1/i)} \tag{12}$$

For any new value of predictor say, X_{new}; Y_{new} is predicted as mentioned above after finding the nearest neighbor and so on. The error for the nearest neighbor X_n is then calculated as e_n. The error en is then summed to the forecast mean ($Y_{new} + e_n$) to evaluate one member of the ensemble forecast. The number of neighbors adopted to generate the ensemble should may vary as compared to the number of neighbors adopted while evaluating local polynomials. Selection procedures for the number of nearest neighbor can be referred to Rajagopalan and Lall (1999).

Apart from other statistics, the efficacy of the ensemble forecasts is evaluated by incorporating ranked probability skill score (RPSS). RPSS signifies whether the model's performance to apprehend categorical probability is better than the categorical probability of the historic data (climatology). For the current study the ensemble forecast and the climatological data are categorized as terciles (0-33rd percentile; 34-66th percentile; 67-100th percentile). Here, the categorical probability is the proportion of the data out of the total falling in each category (tercile). The ranked probability score (RPS) for the forecast and the climatology is evaluated first before evaluating the RPSS. Eq. (13) given below was utilized for calculation of RPS of each year.

$$RPS = \sum_{m=1}^{k}\left[\left(\sum_{i=1}^{m} P_i - \sum_{i=1}^{m} d_i\right)^2\right] \tag{13}$$

For the current study, k was three as the ensemble was divided into three categories. The categorical probability of the ensemble forecast for ith category (P_i) is ratio of the number of members of ensemble forecast falling in the ith category to the total number of ensemble members. The parameter d_i is one if the observed streamflow falls in ith category else it is zero. After evaluating RPS for forecast and the climatology RPSS is evaluated as following Eq. (14).

$$RPSS = 1 - \frac{RPS(forecast)}{RPS(Climatology} \tag{14}$$

Positive RPSS suggests that the forecast is skillful as compared to the climatology while the RPSS of 1 shows the perfect forecast. The negative RPSS signifies that the model performs worse than the climatology. The RPS and RPSS were calculated for each year independently.

2.6. Assessment of forecasted streamflow reliability using statistical parameters

The agreement between the observed and estimated streamflow in this study was evaluated with the correlation coefficient (*r*), percentage bias (P_{bias}), and Nash-Sutcliffe coefficient of efficiency (NSE). All above mentioned statistical performance measures are described in Moriasi et al. (2007). Further, the cumulative non-exceedance probability was estimated using LEPS Skill (SK), which was used to access the forecast skill as compared to 'climatological' mean.

The expression of r, P_{bias}, and NSE is expressed in following Eqs. (15),(16), and (17) respectively.

$$r = \frac{\sum_{i=1}^{n}(X_i^{mes} - X_{mean}^{mes})(X_i^{est} - X_{mean}^{est})}{\sqrt{\sum_{i=1}^{n}(X_i^{mes} - X_{mean}^{mes})^2}\sqrt{\sum_{i=1}^{n}(X_i^{est} - X_{mean}^{est})^2}} \tag{15}$$

$$P_{bias} = \frac{\sum_{i=1}^{n}(X_i^{mes} - X_i^{est})}{\sum_{i=1}^{n} X_i^{mes}} \tag{16}$$

$$NSE = \frac{\sqrt{\sum_{i=1}^{n}(X_i^{mes} - X_i^{est})^2}}{\sqrt{\sum_{i=1}^{n}(X_i^{mes} - X_{mean}^{mes})^2}} \tag{17}$$

Where, X_i^{mes} is the observed/measured streamflow and X_i^{est} is the forecasted/estimated streamflow. X_{mean}^{mes} is the average observed/measured streamflow and X_{mean}^{est} is the average forecasted/estimated streamflow.

The correlation coefficient, *r*, signifies the linear relationship among the measured and estimated streamflow. It ranges between negative and positive '1'; a higher magnitude of *r* shows the good relationship between the two variables. The streamflow forecast is said to be reliable if *r* is close to positive '1', signifying a low amount of errors in the forecast. The percentage bias is the measure of average tendency of biasness of the estimated forecasts as compared to the actual one. The optimal P_{bias} is '0', signifying the minimum biasness of the forecast. The Nash-Sutcliffe coefficient of efficiency (NSE) is popular when assessing the predictive capabilities of the hydrologic model (Nash and Sutcliffe, 1970). NSE ranges from -∞ to 1, where '1' is the best NSE score, signifying a perfect agreement among the estimated and observed streamflow.

The LEPS skill parameterizes the difference between the observed and estimated streamflow in terms of the cumulative probability distribution (Potts et al., 1996). The expression for the LEPS score derived from Potts et al. (1996) is summarized in Eq. (18).

$$S = 3 \times (1 - |p_f - p_0| + p_f^2 - p_f + p_0^2) - 1 \tag{18}$$

where p_f is the cumulative probability for the forecasted streamflow and p_0 is the cumulative probability for the observed streamflow. An accurate forecast far from the climatological mean will have a higher LEPS score, and are considered to be good forecasts. The average LEPS SK is computed with Eq. (19).

$$SK = \frac{\sum S}{\sum S_m} \times 100 \tag{19}$$

where S is the total LEPS score for all years. The LEPS score is the sum for the best forecast, represented by S_m. When S is positive, the cumulative probability distribution of observed and predicted streamflow are same. For negative S, S_m is the sum of LEPS score for the poor forecasts. The model is considered to have good performance if LEPS SK is greater than 10 and its value ranges from -100 to +100 (Potts et al., 1996).

3. Results and discussion

First, the results obtained by tele-connecting the predictors and predictands using SVD is presented in this section, and their physical significance is discussed. Next, results of screening of the predictors based on LEPS score and exceedance probability curve is explained. Finally, the SVM forecast results and cross-validated results are presented and discussed.

3.1. SVD

After individual SVD analysis for both river basins – the Sacramento and the San Joaquin – the variability of the streamflow gages were explained by the first modes of all the predictors with, in each case, SCFs greater than 90%. After applying the SVD separately on both watersheds, the correlations generated as Left TES were used for further analysis. The SVD results, which are explained in this section, are shown as significant regions. In Figs. 3 and 4 , these are regions with '+' signs (regions with '-' signs), which are the regions of increasing (decreasing) oceanic-atmospheric variabilities.

For the first mode of SVD for the Sacramento River Basin, Fig. 3a shows the 90% significance correlation map representing Pacific climate variabilities and Fig. 4a shows the 90% significant region for the Atlantic Ocean's climate variabilities. 90% significant regions signify that there is a 10% chance of the results to be false positive, so the results have 90% confidence. Similarly, results for the San Joaquin watershed is shown in Fig. 3b for the Pacific and Fig. 4b for the Atlantic. SST for the first mode explains most of the variability in each watershed of the study area as compared to other modes of SST. The SST teleconnected regions for Sacramento river basins are positively correlated which means the increase in SST in the teleconnected regions with (+) sign results in the increase in seasonal flow and vice versa. The increase in Z_{500}, $SHUM_{500}$, and U_{500} in the teleconnected regions in Figs. 3a and 4 a with (+) sign results in the decrease in streamflow in Sacramento River basin and vice versa as these regions are negatively correlated with the streamflow. Similarly, SST, Z_{500} and U_{500} teleconnected regions of Pacific Ocean shown in Fig. 3b are positively correlated with the streamflow of San Joaquin River basin. The increase in these regions +/- result in the increase/decrease in streamflow of San Joaquin River basin respectively. On the other hand, $SHUM_{500}$ teleconnected regions shown in Fig. 3b are negatively correlated with the streamflow.

The Z_{500} region below Alaska was out of phase with the streamflow stations for both the Sacramento and San Joaquin watersheds; in other words, these regions had negative correlations to the streamflow. This shows a connection to variability in the Aleutian Low (AL). Lower than average pressures in the AL region signify an increase in the frequency or intensity of storm tracks moving across the Pacific Ocean toward North America. This, in turn, leads to increased precipitation along the Pacific Northwest of the U.S., down to

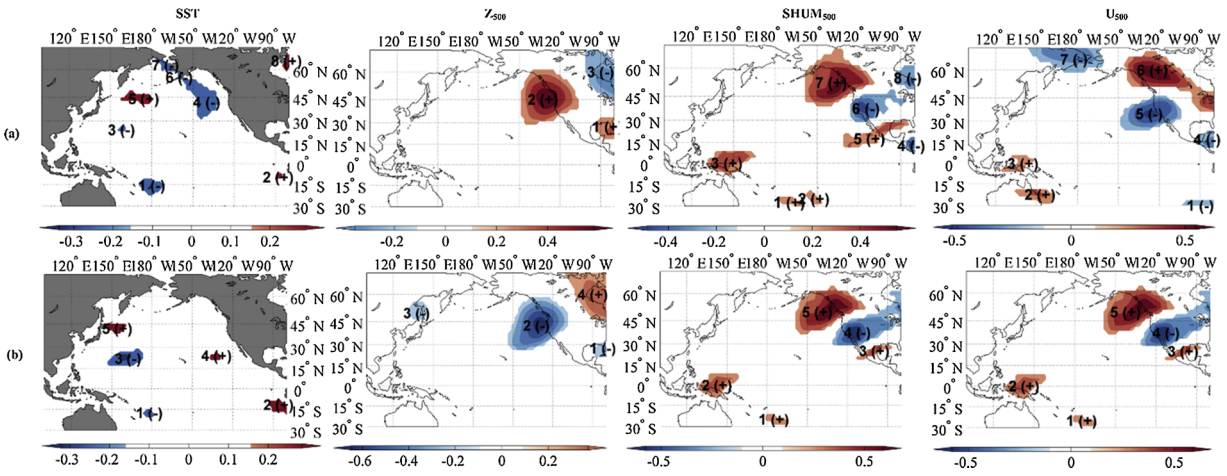

Fig. 3. SVD plots showing teleconnected regions of the Pacific Ocean SST, Z_{500}/ $SHUM_{500}$, and U_{500} with streamflow of the (a) Sacramento River Basin and (b) the San Joaquin River Basin.

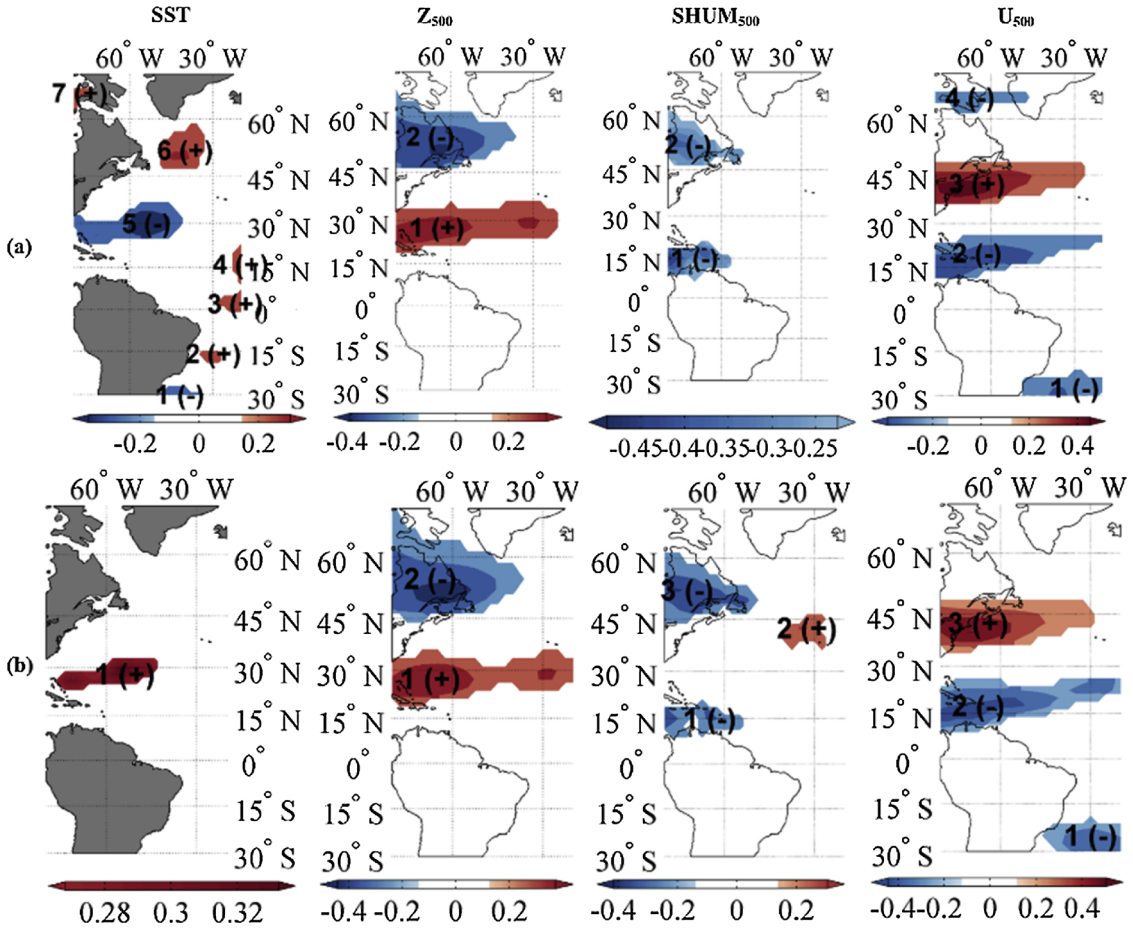

Fig. 4. SVD plots showing teleconnected regions of the Atlantic Ocean SST, Z_{500}, $SHUM_{500}$, and U_{500} with streamflow for (a) the Sacramento River Basin and (b) the San Joaquin River Basin.

the Sacramento and San Joaquin watersheds. Furthermore, the close contours of Z500 signify a rapid change of geopotential index levels, which can be taken as the origin of the polar jet streams. The polar jet streams play a key role in the advection of moisture to the west coast of the U.S. from the Pacific Ocean. Physically, a higher $SHUM_{500}$ in the Alaskan Gulf signifies a greater amount of Dec-Feb precipitation in the U.S. Pacific Northwest, suggesting the effect of jet-stream that reduces the precipitation and runoff in the SSJ River Basin. The lower $SHUM_{500}$ in the California region, resulting in lower streamflow, signifies less Dec-Feb winter precipitation and low Apr-Jun runoff.

U_{500} teleconnected regions shown in Figs. 3a and b are similar to the pattern of the jet stream moving to the western coast of the U.S. through the Pacific. This signifies that the moisture that gets transformed to streamflow in the region is driven by the jet stream in addition to the westerlies. The trough of the jet stream shifts eastwards and westwards in the Pacific during various phases of ENSO; thus, the jet stream sometimes has a higher probability of passing through the SSJ watershed, resulting in higher than normal precipitation and runoff. In other instances, the jet stream moves storm tracks along the Alaskan and Canadian coasts, and passes through the northwestern U.S. above California; this increases the precipitation and streamflow in the northwestern U.S., but lowers the amount of runoff and precipitation in California.

The standardized Left TES for the first mode climate variables (of SST and Z_{500} along with $SHUM_{500}$ and U_{500}) of both Pacific Ocean and Atlantic Ocean obtained from the SVD analysis are summarized in Fig. 5. Column (a) of Fig. 3 represents the Left TES for Sacramento watershed while the LTES results for San Joaquin river basin is shown in column (b). The solid line in Fig. 5 represents the Left TES corresponding to the Pacific Ocean and the one corresponding to the Atlantic Ocean is represented by dotted lines. Table 1 summarizes the SCF and NSC values expressed in percentage corresponding to all predictors of SSJ watershed. As seen in the figure SCFs are mostly greater than 90% showing that the first mode of teleconnections captured most of the teleconnections between the streamflow and the climate variables. Similarly, NSC is 100% for the perfect correlation which signifies 100% correlation among all streamflow values and the predictor values for each grid. 100% NSC is impractical to anticipate while the NSC values tabulated in the Table 1 is comparable to the previous studies. For the current study the NSCs ranged from1.5 to 3.5 % while Pathak et al. (2018) evaluated the NSCs ranging from 1.78 to 3.87 for western US snowpack. Similarly, Sagarika et al. (2015) evaluated the correlation coefficient ranging from 0.009 to 0.64 for western US Apr-Jun streamflow.

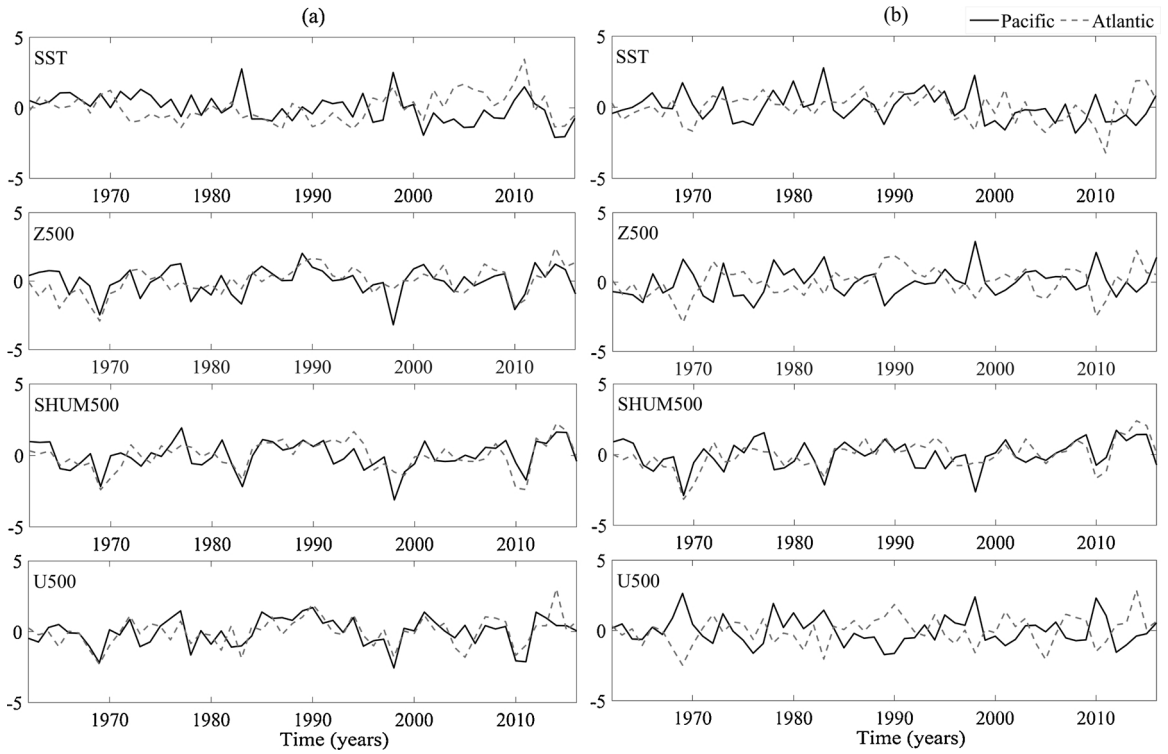

Fig. 5. Time series plot of standardized Left TES plot for the first mode of SST, Z_{500}, $SHUM_{500}$, and U_{500} of both Pacific Ocean and Atlantic Ocean corresponding to (a) Sacramento and (b) San Joaquin river basin. The solid lines represent Pacific Ocean's Left TES and the dotted lines show Atlantic Ocean's Left TES.

Table 1
Squared Covariance Factor and Normalized Squared Covariance values in percentage corresponding to of SST, Z_{500}, $SHUM_{500}$, and U_{500} of both Pacific Ocean and Atlantic Ocean for SSJ watershed.

	Sacramento Watershed		San Joaquin Watershed	
Climate Variables	Squared Covariance Factor	Normalized Squared Covariance	Squared Covariance Factor	Normalized Squared Covariance
Pacific SST	88%	1.3%	90%	1.5%
Atlantic SST	96%	2.7%	96%	2.1%
Pacific Z_{500}	90%	1.8%	96%	2.3%
Atlantic Z_{500}	95%	2.6%	97%	2.9%
Pacific $SHUM_{500}$	95%	2.4%	96%	2.6%
Atlantic $SHUM_{500}$	94%	1.6%	95%	1.9%
Pacific U_{500}	95%	2.6%	97%	2.9%
Atlantic U_{500}	97%	3.1%	98%	3.5%

3.2. Nonparametric screening

SVD results show that the first mode SCFs are highly significant, more than any other modes, in terms of explaining the variability of the streamflow. Thus, taking into account the first mode Left TES's for all oceanic-atmospheric variables could prove to be a skillful predictor for the streamflow stations under consideration. Further, based on the exceedance results (i.e., the LEPS skill for the models), the best combination of predictors for each station were evaluated.

In addition, they can be used to make probabilistic forecasts by using a continuous exceedance probability curve, as shown in Fig. 6. Additionally, a good model should be able to forecast the flow more accurately than can the climatology (1981–2010). To evaluate the model skill relative to the climatology (1981–2010), LEPS Skill was evaluated. The LEPS SK above 20% is categorized as good forecast as shown in Fig. 6. Similarly, as shown in the Fig. 6 the LEPS SK between 10% and 20% is categorized as fair forecast and the LEPS SK less than 10% is categorized as poor forecast.

Fig. 6 shows the exceedance probability forecasts for Station 1 for 2013, 2016, and 2015 which were the good, fair and poor forecast years respectively. The forecast year 2013 for station 1 was a good year, because at a 50% exceedance level (i.e., 50% risk), the forecast shown as solid line in Fig. 6 lies close to the observed streamflow volume shown as a circle in Fig. 6 as compared to the climatology shown as the dotted line, with LEPS Skill scores 24.56%. On the other hand, there are risks in predicting the flow in a

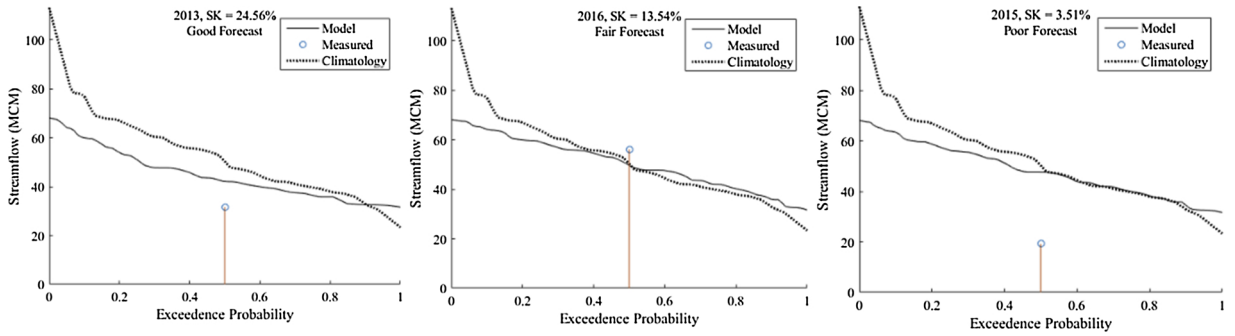

Fig. 6. Sample exceedance plots for the seasonal streamflow volume of station 1 in million cubic meters (MCM). Figure shows poor, fair and good forecast years based on the LEPS SK.

poor forecast year, such as 2015 for station 1 as shown in Fig. 6. However, the model behaves well and is comparable to the climatology at 50% exceedance level for even these years. Although the model had associated errors, by averaging the entire period of 54 years, the model was skillful, with the LEPS score approaching +10 %. Thus, the models with highest LEPS scores were selected, and their predictors were used for further SVM modeling.

The best combination of the predictors were the first mode Left TES of $SHUM_{500}$ Pacific and U_{500} Pacific for streamflow Stations 2, 3, 4, 5, 6, 7, and 8 yielding the maximum LEPS score. Similarly, $SHUM_{500}$ Pacific, $SHUM_{500}$ Atlantic, and U_{500} Pacific were found to be the best predictors for Site 1, which had an optimum LEPS score.

Based on the % weight for different predictors of the best model for each station, $SHUM_{500}$, and U_{500} of Pacific were more skillful for the streamflow forecast of the SSJ watershed. Previously, SST was widely used by various researchers to predict the precipitation and streamflow (Uvo et al., 1998; Westra and Sharma, 2010); in addition, Z_{500} was widely used for the same purpose (Serreze et al., 1998). The current study identified that one month lagged winter variability of Pacific $SHUM_{500}$ and Pacific U_{500} proved to be more effective in predicting the April to June streamflow of SSJ watershed compared to SST, and Z_{500}.

Although individual SVD analysis with each variables resulted in the teleconnected regions but $SHUM_{500}$ and U_{500} were skillful predictor for the streamflow based on the non-parametric screening which can be explained physically. The Apr-Jun streamflow in SSJ watershed is driven primarily by the winter (Dec-Feb) snow precipitation within the region. Furthermore, winter snow precipitation highly depends upon the Pacific $SHUM_{500}$ and U_{500}. Both $SHUM_{500}$ and U_{500} variabilities corresponds to the moisture circulation from the Pacific toward western United States during the winter months. $SHUM_{500}$ captures the amount of moisture circulated from the Pacific. Similarly, U_{500} provides the information about jet streams which drives the moist air rising from the Pacific Ocean towards the Western United States.

3.3. SVM

The results of SVM in Fig. 7a shows the time series plots for individual streamflow stations of the SSJ River basin. The statistical parameters (i.e., r, Pbias, and NSE) for the forecasted streamflow with SVM are presented in Fig. 7b. Fig. 7c shows the distribution of measured and estimated streamflow and Fig. 7d represents the LEPS SK of the forecast. Bootstrap cross-validation result of the SVM model is presented in Fig. 8. The unit of streamflow represented as % Normal is the percentage of the average measured (observed) streamflow.

The time series plots in Fig. 7a incorporate the measured (observed) streamflow and SVM estimates (estimated streamflow) as the %normal of the mean observed flow of the streamflow at each gage stations. In Fig. 7a, dashed lines show measured and the solid lines show the SVM estimated streamflow..

The time series plots indicate that the current approach was able to predict most of the high flows and low flows. Although some events were missed, the magnitudes of error were not very high. The model was good at predicting most of the low flows as well as sustained low flows. The capability of capturing sustained low flows makes the model effective for water managers for sustainable flow management in drought-affected regions.

A robust harmony between the observed and the predicted values was observed from the time series plots. According to Moriasi et al. (2007), the simulated model was considered to have very good performance in terms of Pbias, which was less than 10 for each streamflow stations of the SSJ watershed. The current model had a minimum r of 0.87, ranging to 0.96; the model was assumed to have a very good performance rating as r was greater than 0.85. This indicates that the correlation coefficient ranks the model as a very good one. For SSJ watershed NSE ranged between 0.73 and 0.82. Further, Moriasi et al. (2007) ranked the model as a very good one if NSE ranged in between 0.75 and 1, which was mostly true for this hybrid model. Thus, based on NSE, r, and Pbias, the hybrid model was reliable, with very good performance. These statistical parameters are well described in Moriasi et al. (2007). Further, Moriasi et al. (2007) established model evaluation guidelines based on the range of these parameters. These parameters showed that the model forecasted results in good correlation with the observed values.

The scatterplots in Fig. 7b represent good agreement among the measured and estimated streamflow volumes for the SSJ River Basins, respectively. As in all scatterplots, the major number points were aligned and were close to the 45° bisector line. The flow

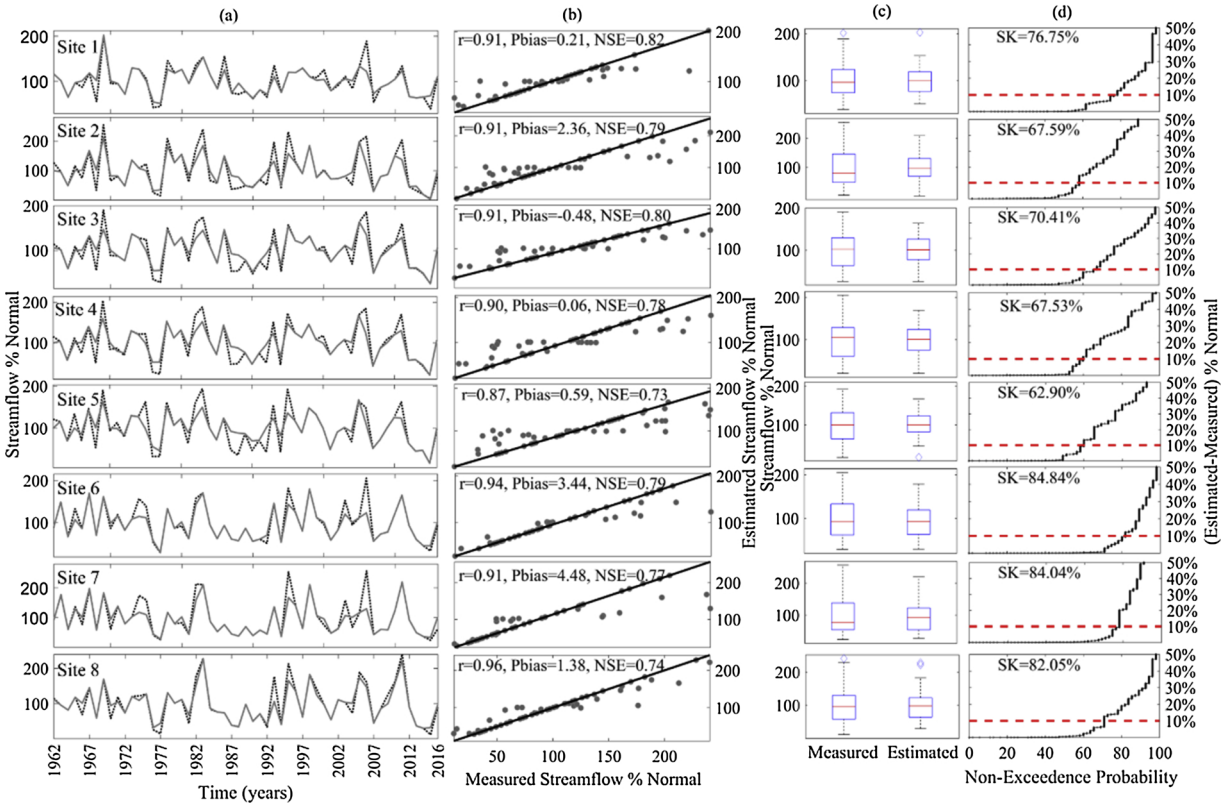

Fig. 7. (a) Time series (dashed line shows measured and the solid one shows SVM estimated streamflow), (b) scatterplots, (c) box plots, and (d) non-exceedance probability plots of the Sacramento and San Joaquin River Basin comparing the measured (observed) and estimated (forecasted) streamflow at each stations.

volume for spring (April to June) is presented as the % normal of the mean observed streamflow. From the scatterplot, this can be visualized that the model forecast is in agreement to the observed streamflow. Using boxplots, Fig. 7c compare the measured and estimated spring streamflow for the SSJ watershed. For most of the stations, the model was able to apprehend low flows as compared to the high flows; as can be seen from the box plots, the 5th percentile of the estimated flow resembled that of the measured flow. The model ability to capture the low flow makes it useful for the water managers working on drought affective.

The accuracy of the SVM forecast relative to the climatological mean was evaluated as the cumulative non-exceedance probability error, which was calculated and evaluated as the LEPS SK score. The probabilistic cumulative error among measured and estimated streamflow is shown in Fig. 7d for the SSJ River Basin. According to Potts et al. (1996), the outcome is said to be random if SK = 0, and the model is said to perform well if SK > 10%. In this study, the SK > 60% for all stations, with a minimum of 62.9% and a maximum of 80.4%. From Fig. 7d, nearly 80% of the predictions for Site 6 had 10% error; thus, the performance of the model was better for Station 6 in the Sacramento watershed as compared to other stations of SSJ watershed. These LEPS SK scores ascertained the robustness of the model as a forecasting tool for water managers, as the model skill was better than when using climatology techniques.

Fig. 8 shows the time series plots for measured, estimated, and cross-validated streamflow values for the study period for each station of the SSJ watershed. Again, good agreement among the measured, predicted, and bootstrapped time series was observed. The bootstrapped cross-validation outcome reaffirms the robustness of the SVM model. The results of the model were stable, and did not vary with changes in the training and testing datasets. The bootstrapped results showed agreement with the estimated time series, following a similar trend as that of the estimated time-series plots.

After screening the predictors and using them to forecast using SVM, it was observed that the model performance was acceptable, based on the accessed statistical skills. As evident from the time-series plots, the cross-validated and estimated flows were similar to the observed streamflow. The model performed well, as the estimated flow followed the 45̊ bisector line, indicating the good fit results. The overall model performance was acceptable, as the predictions were associated with low errors; there was good agreement to the measured flow, although sometimes that was difficult to capture during such events as the high flows of Sacramento in 2005 and 2006. Based on the evaluations of model performance against climatology (1981–2010), using LEPS Skill scores, the model predictor seemed to serve well in forecasting the streamflow as compared to climatology. This indicates that the model performance worked adequately well for SSJ watershed, especially when capturing the low flows, as evident from the time-series plots. The model's ability to capture sustained low flows may prove to be a promising resource for water managers, especially in regions facing

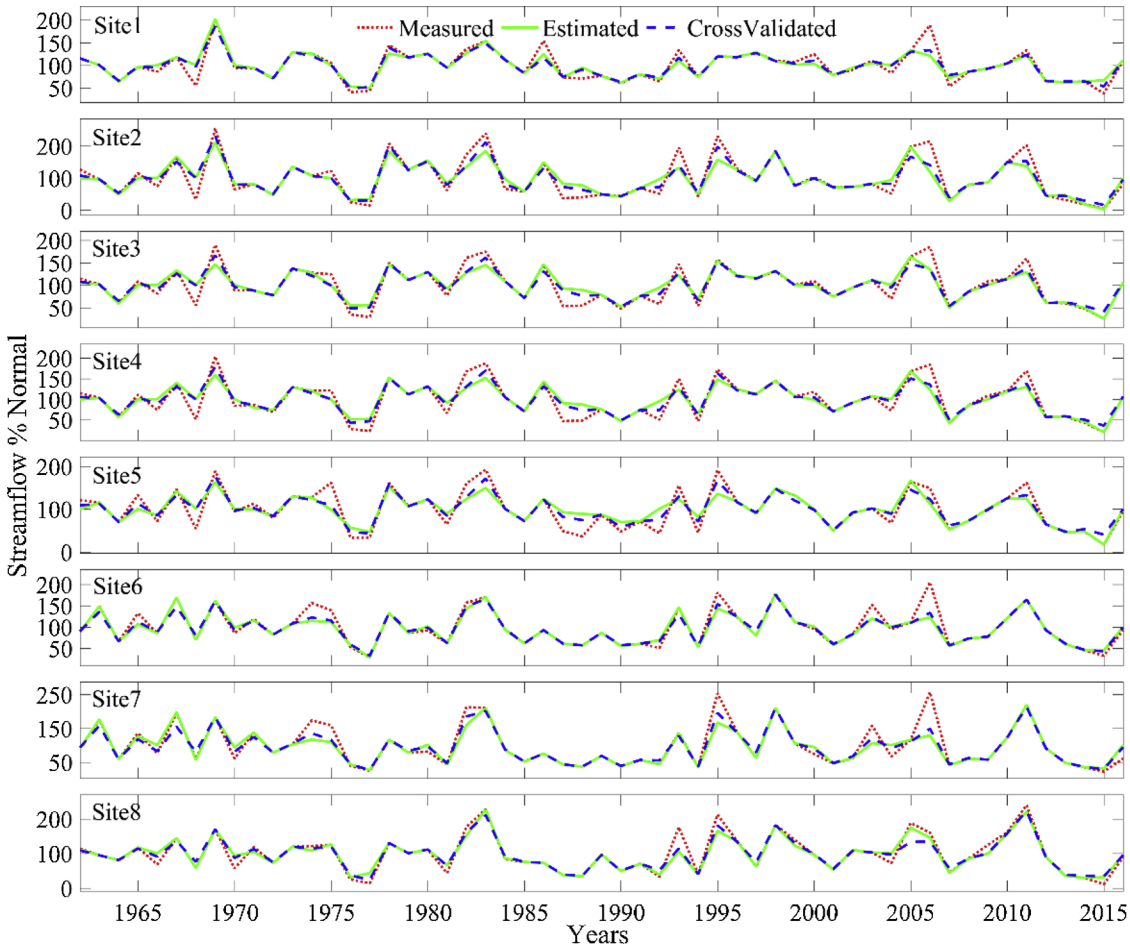

Fig. 8. Time-series plots for measured streamflow, estimated streamflow, and bootstrapped cross-validated streamflow for each stations of the SSJ watershed.

prolonged droughts, such as the western U.S.

The performed SVM regression incorporated primary modes of four different types of largescale climate variables to improve the streamflow at the regional scale in the SSJ river basin of the drought affect northern California. The skill of the forecast based on all the presented statistical evaluations were within the acceptable limits. Larger correlation coefficients and larger NSE accompanied with smaller PBIAS confirms the model accuracy. Higher skills of SVM can be attributed to both preprocessing the large scale climate variables with SVD and the non-parametric screening of the SVD teleconnections. To establish the hypothesis that the use of large scale climate variable have improved the forecast skill, forecast were also made with predefined indices for the evaluation of the proposed approach which is presented below.

The robustness of the proposed approach of utilizing the primary modes of large scale teleconnections for streamflow forecasting is further reaffirmed with the streamflow forecast results based on the predefined indices as shown in Fig. 9. Fig. 9 shows streamflow forecasts with three combinations Pacific Ocean's predefined indices – Niño 3.4, PDO, and Niño 3.4 + PDO corresponding to site 5. Where, Niño 3.4 an ENSO index ranging from 120th to 170th west longitudes on both sides of equator by 5 degrees. The first column of Fig. 9 shows the time series of the forecast utilizing Niño 3.4 predefined index along with the scatter plots, box plots and non-exceedance probability curve. Similarly, second and third column of Fig. 9 shows the results of streamflow forecast utilizing PDO index and both PDO and Niño 3.4 respectively. The forecast with Niño 3.4 and PDO indices but the proposed approach of utilizing the primary modes of large scale teleconnections results had better statistical skills. The maximum NSE as shown in Fig. 9b for site 5 was 0.66 while the NSE utilizing the large scale climate variable was 0.73 as shown in Fig. 7b. Similarly, the correlation coefficient between the measured and observed streamflow volume forecasted with large scale climate variable was 0.87 which was higher than the maximum correlation coefficient of 0.82 for the forecast obtained utilizing the predefined indices. Similarly the LEPS SK of the model utilizing the predefined indices for site 5 as shown in Fig. 7d is less as compared to the LEPS SK for site 5 shown in Fig. 7d corresponding to the forecast with large scale climate variable.

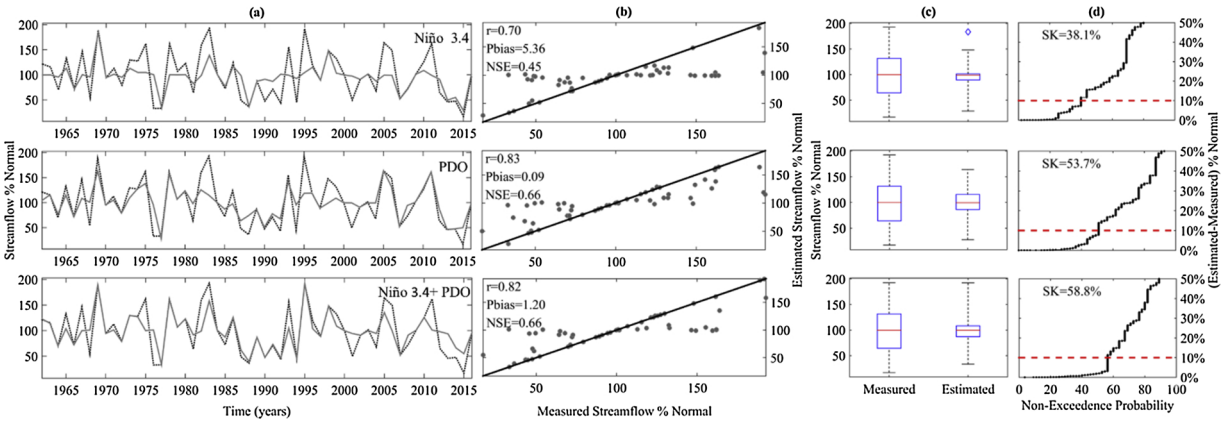

Fig. 9. SVM forecast with the predefined indices- Niño 3.4, PDO, and Niño 3.4 + PDO (a) Time series plots (dashed lines show measured and the solid lines show SVM estimated streamflow), (b) scatterplots, (c) box plots, and (d) non-exceedance probability plots for Site 5 comparing the measured streamflow and estimated streamflow utilizing predefined indices.

3.4. KNN

After SVM regression, the rationale to include KNN was to establish the confidence that the teleconnections of the regional streamflow values with the large scale climate variables are skillful in any regression other than SVM. The KNN forecast of the SSJ River Basin is presented in Fig. 10 as the boxplot representing the medians, quartiles and the outliers with 100 forecasts for each year at each station. The figures also include the time series plots of the observed streamflow to compare the ensemble forecast with the observed streamflow values. All streamflow in Fig. 10 are presented as the % normal of the mean observed streamflow at each gage stations. From Fig. 10 it is evident that observed/measured streamflow lies within the boxes of the ensemble forecasts suggesting good model performance capable of capturing most of the flows including high flows and low flows in the interquartile range. The ability of KNN model to capture the streamflow peaks showed the effectiveness of the coupled SVD and non-parametric screening algorithm. The predictors evaluated with coupled SVD and non-parametric screening were able to capture the streamflow variabilities. Asymmetry of the boxes around the median shows the skewness of the forecast captured due to resampling of the data.

The quantitative evaluation of the KNN is performed with RPSS, correlation coefficient (*r*), percentage bias (P_{bias}), and Nash-Sutcliffe coefficient of efficiency (NSE). The parameters r, P_{bias}, and NSE are calculated with the mean of the ensemble forecast evaluating the mean ensemble forecast. The aforementioned parameters are tabulated in the Table 2. The simulated KNN model was considered to have very good performance in terms of Pbias, which was less than 10 for each streamflow stations of the SSJ watershed. The current model had a minimum r of 0.83, ranging to 0.90; the model was assumed to have a very good performance rating as r was greater than 0.85 except for one station. This indicates that the correlation coefficient ranks the model as a very good one. For SSJ watershed NSE ranged between 0.67 and 0.79. Further, Moriasi et al. (2007) ranked the model as a good one if NSE ranged in between 0.65 and 0.75 and very good for the stations with NSE ranged in between 0.75 and 1. The RPSS presented in the table is the median value of the 55 years RPSS obtained for each streamflow stations. The positive RPSS indicates the ensemble forecast having better skill than climatology. Further, the RPSS closer to 1 signifies more skillful forecast. Similar to the SVM model KNN was also capable to capture the variations in the streamflow like high flows and the low flows as seen in the Fig. 10 where the peaks of the observed time series lies within the box plots of the ensemble forecasts. Further, it was also able to capture the sustained low flows making the model useful in drought prone regions.

Forecasting the streamflow with climate indices can result in errors in the forecast because of the inability of climate indices to capture every underlying physical process. To overcome this the use of largescale climate variable teleconnections with the regional streamflow can lead to better inclusion of regional climate systems. Further, to improve the model results extensive care was taken while choosing the largescale climate predictors based on the regional geography climate processes. The climate variables utilized were skillful in capturing local climate features affecting the streamflow. The approach is purely statistical and it can perform well at the regional scales but the underlying physical drivers of the streamflow cannot be fully understood. While the advantage of using the statistical model is that, it is easy to implement and these models are not prone to parameter uncertainty. Future studies can be done in other regions with the clear overview of different sets of climate variables effecting the streamflow within the region. In the current approach, both SVM and KNN were able to capture the extreme hydrological extremes with minor deviation between the forecast and the observed streamflow, overall the both model's forecast were skillful during the stud period and can lead to in making month ahead seasonal water management strategies.

4. Conclusions

In this study, a new approach to forecast seasonal streamflow volume –was developed as an improvement to those developed in previous researches. In order to improve streamflow forecast with predefined indices, this study utilized teleconnections obtained

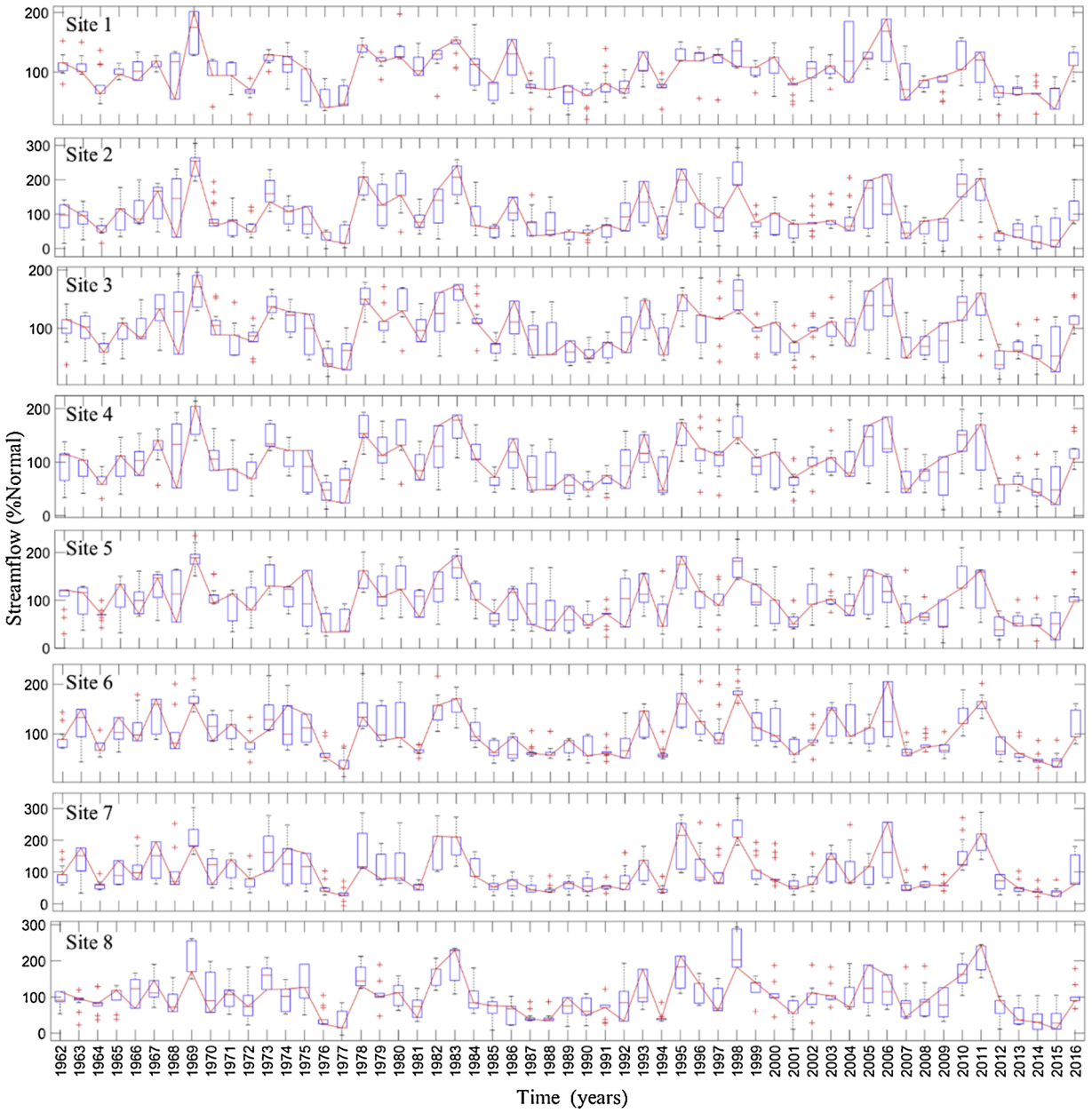

Fig. 10. Box plots of KNN ensemble forecast and the time series plots of the observed streamflow from 1962 to 2017 expressed as % normal of the mean observed streamflow of Sacramento-San Joaquin River Basin.

Table 2
Performance evaluation of KNN ensemble forecast based on Ranked probability skill score (RPSS), correlation coefficient (*r*), percentage bias (P_{bias}), and Nash-Sutcliffe coefficient of efficiency (NSE).

Watershed	Station	RPSS	r	P_{bias}	NSE
San Joaquin Watershed	1	0.9	0.88	−0.65	0.77
	2	0.7	0.87	1.05	0.75
	3	0.8	0.86	0.37	0.72
	4	0.7	0.86	0.35	0.72
	5	0.6	0.83	0.60	0.67
Sacramento Watershed	6	0.7	0.90	−0.60	0.79
	7	0.8	0.88	−0.58	0.77
	8	0.7	0.89	−0.79	0.78

from SVD analysis to predict streamflow volume in conjunction with nonparametric screening and the nonlinear regression model like SVM and KNN. The first advantage of the adopted framework was that it subsides the use of predefined indices with the aid of SVD. The second advantage was utilizing the non-parametric approach to screen the SVD relations. This lead to segregate the skillful SVD relations that would enhance the input of the regression model. Finally, using the SVM and KNN as a regression tool established the robustness of the forecast. Four different oceanic-atmospheric variables– SST and Z_{500} along with $SHUM_{500}$ and U_{500} over the Pacific Ocean and Atlantic Ocean were utilized to improve the streamflow forecast at the regional scale.

The first research question was addressed with the help of SVD analysis by finding the new teleconnected regions corresponding to$SHUM_{500}$, and U_{500} in addition to SST and Z_{500}over Pacific and Atlantic Oceans for the SSJ watershed streamflow. The second research question was addressed with the non-parametric screening showing that the $SHUM_{500}$ and U_{500} of Pacific were the best predictors for the streamflow of SSJ watershed. The proposed forecasting approach utilizing the large scale climate variables were skillful based on the statistical evaluations. The proposed forecast approach were more reliable than streamflow forecasts performed using the predefined indices in the selected study area. This was the answer for the final research question.

As the forecasts is truly based on the statistical principles the physical processes involved in the variations of streamflow cannot be understood. The forecasts made by data driven models improve with increase in the training period. The current study incorporated 55 years of available streamflow data, the training period of the current model could be improved in future by including reconstruction data in addition to the instrumental records. Forecasts for different lead times also could be evaluated in the future to determine the lead time at which the model performs the best in a selected study area. Finally, using this proposed method in different regions with different types of datasets may lead to a better understanding of the implications of the model.

The current research is important for season-ahead prediction of streamflow values in the SSJ watersheds, a major cropland of California, with the teleconnected parameters of the Pacific and Atlantic Oceans. The study used standard global datasets, along with the standard statistical tools for this research. The seasonal forecast volume obtained from the proposed forecast approach was reliable, as the robustness of the forecast was established with the bootstrap cross-validation technique. The results were verified using different measures for statistical performance. This study might be beneficial to the water-resource scholars and water managers in understanding and managing hydrology at a watershed scale.

Acknowledgement

The authors acknowledge the valuable comments provided by reviewers that helped in improving the overall quality of the manuscript. The authors would like to acknowledge the Office of the Vice Chancellor for Research at Southern Illinois University Carbondale for providing the research support. The information relating to dataset used in the analysis is provided in the manuscript.

References

Asefa, T., Kemblowski, M., McKee, M., Khalil, A., 2006. Multi-time scale stream flow predictions: the support vector machines approach. J. Hydrol. 318 (1), 7–16.
Besaw, L.E., Rizzo, D.M., Bierman, P.R., Hackett, W.R., 2010. Advances in ungauged streamflow prediction using artificial neural networks. J. Hydrol. 386 (1–4), 27–37.
Bretherton, C.S., Smith, C., Wallace, J.M., 1992. An intercomparison of methods for finding coupled patterns in climate data. J. Clim. 5 (6), 541–560.
Crockett, J.L., Westerling, A.L., 2018. Greater temperature and precipitation extremes intensify Western US droughts, wildfire severity, and Sierra Nevada tree mortality. J. Clim. 31 (1), 341–354.
Chang, F.J., Chen, Y.C., 2003. Estuary water-stage forecasting by using radial basis function neural network. J. Hydrol. 270 (1–2), 158–166.
Efron, B., 1979. Bootstrap methods: another look at the jacknife. Ann. Stat. 7, 1–26.
Efron, B., Tibshirani, R.J., 1993. An introduction to the bootstrap. CRC press.
Ehteram, M., Afan, H.A., Dianatikhah, M., Ahmed, A.N., Fai, C.M., Hossain, M.S., Allawi, M.F., Elshafie, A., 2019. Assessing the predictability of an improved ANFIS model for monthly streamflow using lagged climate indices as predictors. Water 11 (6), 1130.
Ellis, A.W., et al., 2010. A hydroclimatic index for examining patterns of drought in the Colorado River Basin. Int. J. Climatol. 30 (2), 236–255.
Enfield, D.B., Mestas-Nuñez, A.M., Trimble, P.J., 2001. The Atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental US. Geophys. Res. Lett. 28 (10), 2077–2080.
Falcone, J.A., 2011. GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow. Retrieved from < https://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml > . .
Frederick, K.D., Major, D.C., 1997. Climate change and water resources. Clim. Change 37 (1), 7–23.
Geisser, S., 1975. The predictive sample reuse method with applications. J. Am. Stat. Assoc. 70, 320–328.
Grantz, K., Rajagopalan, B., Clark, M., Zagona, E., 2005. A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts. Water Resour. Res. 41 (10).
Hamlet, A.F., Lettenmaier, D.P., 1999. Columbia River streamflow forecasting based on ENSO and PDO climate signals. J. Water Resour. Plan. Manag. 125 (6), 333–341.
Hamlet, A.F., Lettenmaier, D.P., 2007. Effects of 20th century warming and climate variability on flood risk in the western US. Water Resour. Res. 43 (6).
Hidalgo, H.G., 2004. Climate precursors of multidecadal drought variability in the western United States. Water Resour. Res. 40 (12).
Hidalgo-Muñoz, J.M., Gámiz-Fortis, S.R., Castro-Díez, Y., Argüeso, D., Esteban-Parra, M.J., 2015. Long-range seasonal streamflow forecasting over the Iberian Peninsula using large-scale atmospheric and oceanic information. Water Resour. Res. 51 (5), 3543–3567.
Hong, W.C., 2008. Rainfall forecasting by technological machine learning models. Appl. Math. Comput. 200 (1), 41–57.
Kalra, A., Ahmad, S., 2012. Estimating annual precipitation for the Colorado River Basin using oceanic-atmospheric oscillations. Water Resour. Res. 48 (6).
Kalra, A., Miller, W.P., Lamb, K.W., Ahmad, S., Piechota, T., 2013. Using large-scale climatic patterns for improving long lead time streamflow forecasts for Gunnison and San Juan River Basins. Hydrol. Process. 27 (11), 1543–1559.
Knowles, N., Cayan, D.R., 2002. Potential effects of global warming on the Sacramento/San Joaquin watershed and the San Francisco estuary. Geophys. Res. Lett. 29 (18).
Kuh, A., 2004. “Least Squares Kernel Methods and Applications.” Soft computing in Communications. Springer, Berlin, Heidelberg, pp. 365–387.
Lall, U., Sharma, A., 1996. A nearest neighbor bootstrap for resampling hydrologic time series. Water Resour. Res. 32, 679–693.
Li, P.H., Kwon, H.H., Sun, L., Lall, U., Kao, J.J., 2009. A modified support vector machine based prediction model on streamflow at the Shihmen Reservoir. Taiwan. International Journal of Climatology 30, 1256–1268.

Lins, H.F., 2012. USGS Hydro-Climatic Data Network 2009 (HCDN-2009): U.S. Geological Survey Fact Sheet 2012–3047. U.S. Geological Survey. p. 4, https://pubs.usgs.gov/fs/2012/3047/.
Loader, C., 1999. Statistics and Computing: Local Regression and Likelihood. Springer, New York.
Maurer, E.P., 2007. Uncertainty in hydrologic impacts of climate change in the Sierra Nevada, California, under two emissions scenarios. Clim. Change 82 (3–4), 309–325.
Mehran, A., Mazdiyasni, O., AghaKouchak, A., 2015. A hybrid framework for assessing socioeconomic drought: linking climate variability, local resilience, and demand. J. Geophys. Res. Atmos. 120 (15), 7520–7533.
Middelkoop, H., Daamen, K., Gellens, D., Grabs, W., Kwadijk, J.C., Lang, H., Wilke, K., 2001. Impact of climate change on hydrological regimes and water resources management in the Rhine basin. Clim. Change 49 (1–2), 105–128.
Moradkhani, H., Meier, M., 2010. Long-lead water supply forecast using large-scale climate predictors and independent component analysis. J. Hydrol. Eng. 15 (10), 744–762.
Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. Asabe 50 (3), 885–900.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. J. Hodrol. 10 (3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.
Nikam, V., Gupta, K., 2013. SVM-based model for short-term rainfall forecasts at a local scale in the Mumbai urban area, India. J. Hydrol. Eng. 19 (5), 1048–1052.
Omondi, P., Awange, J.L., Ogallo, L.A.,Ininda, J., Forootan, E., 2013. The influence of low frequency sea surface temperature modes on delineated decadal rainfall zones in Eastern Africa region. Adv. Water Resour. 54, 161–180.
Pathak, P., Kalra, A., Lamb, K.W., Miller, W.P., Ahmad, S., Amerineni, R., Ponugoti, D.P., 2018. Climatic variability of the Pacific and Atlantic Oceans and western US snowpack. Int. J. Climatol. 38 (3), 1257–1269.
Piechota, T.C., Chiew, F.H., Dracup, J.A., McMahon, T.A., 1998. Seasonal streamflow forecasting in eastern Australia and the El Niño–Southern Oscillation. Water Resour. Res. 34 (11), 3035–3044.
Potts, J., Folland, C., Jolliffe, I., Sexton, D., 1996. Revised “LEPS” scores for assessing climate model simulations and long-range forecasts. J. Clim. 9 (1), 34–53.
Rajagopalan, B., Lall, U., 1999. A k-nearest-neighbor simulator for daily precipitation and other weather variables. Water Resour. Res. 35 (10), 3089–3101.
Rojsiraphisal, T., Rajagopalan, B., Kantha, L., 2009. The Use of MTM-SVD Technique to Explore the Joint Spatiotemporal Modes of Wind and Sea Surface Variability in the North Indian Ocean during 1993–2005. Int. J. Oceanogr. 2009.
Sagarika, S., Kalra, A., Ahmad, S., 2015. Interconnections between oceanic–atmospheric indices and variability in the US streamflow. J. Hydrol. 525, 724–736.
Sagarika, S., Kalra, A., Ahmad, S., 2016. Pacific Ocean SST and Z500 climate variability and western US seasonal streamflow. Int. J. Climatol. 36 (3), 1515–1533.
Schewe, J., Heinke, J., Gerten, D., Haddeland, I., et al., 2014. Multimodel assessment of water scarcity under climate change. Proceedings of the National Academy of Sciences 111 (9), 3245–3250.
Scholkopf, B., Sung, K.K., Burges, C.J., Girosi, F., Niyogi, P., Poggio, T., Vapnik, V., 1997. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. Ieee Trans. Signal Process. 45 (11), 2758–2765.
Serreze, M.C., Clark, M.P., McGinnis, D.L., Robinson, D.A., 1998. Characteristics of snowfall over the eastern half of the United States and relationships with principal modes of low-frequency atmospheric variability. J. Clim. 11 (2), 234–250.
Simões, N., Wang, L., Ochoa-Rodriguez, S., Leitão, J., Pina, R., Onof, C., Maksimović, Č., 2011. A Coupled SSA-SVM Technique for Stochastic Short-term Rainfall Forecasting.
Sivapragasam, C., Liong, S.-Y., Pasha, M., 2001. Rainfall and runoff forecasting with SSA–SVM approach. J. Hydroinformatics 3 (3), 141–152.
Soukup, T.L., Aziz, O.A., Tootle, G.A., Piechota, T.C., Wulff, S.S., 2009. Long lead-time streamflow forecasting of the North Platte River incorporating oceanic–atmospheric climate variability. J. Hydrol. 368 (1–4), 131–142.
Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. 36, 111–147.
Suykens, J.A., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural Process. Lett. 9 (3), 293–300.
Suykens, J.A., Van Gestel, T., De Brabanter, J., 2002. Least Squares Support Vector Machines. World Scientific.
Tamaddun, K.A., Kalra, A., Ahmad, S., 2017. Wavelet analyses of western US streamflow with ENSO and PDO. J. Water Clim. Chang. 8 (1), 26–39.
Tootle, G.A., Piechota, T.C., 2006. Relationships between Pacific and Atlantic ocean sea surface temperatures and US streamflow variability. Water Resour. Res. 42 (7).
Trenberth, K.E., 2001. Climate variability and global warming. Science 293 (5527), 48–49.
Twarakavi, N.K., Šimůnek, J., Schaap, M.G., 2009. Development of pedotransfer functions for estimation of soil hydraulic parameters using support vector machines. Soil Sci. Soc. Am. J. 73 (5), 1443–1452.
VanRheenen, N.T., Wood, A.W., Palmer, R.N., Lettenmaier, D.P., 2004. Potential implications of PCM climate change scenarios for Sacramento–San Joaquin River Basin hydrology and water resources. Clim. Change 62 (1), 257–281.
Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York.
Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York.
Vapnik, V., 2013. The Nature of Statistical Learning Theory: Springer Science & Business Media.
Wallace, J.M., Gutzler, D.S., 1981. Teleconnections in the geopotential height field during the Northern Hemisphere winter. Mon. Weather. Rev. 109 (4), 784–812.
Wang, W., Vrijling, J.K., Van Gelder, P.H., Ma, J., 2006. Testing for nonlinearity of streamflow processes at different timescales. J. Hydrol. 322 (1), 247–268.
Ward, M.N., Folland, C.K., 1991. Prediction of seasonal rainfall in the north nordeste of Brazil using eigenvectors of sea-surface temperature. Int. J. Climatol. 11 (7), 711–743.
Wei, W., Watkins, D.W., 2011. Data mining methods for hydroclimatic forecasting. Adv. Water Resour. 34 (11), 1390–1400.
Yaseen, Z.M., Jaafar, O., Deo, R.C., Kisi, O., Adamowski, J., Quilty, J., El-Shafie, A., 2016. Stream-flow forecasting using extreme learning machines: a case study in a semi-arid region in Iraq. J. Hydrol. 542, 603–614.
Zealand, C.M., Burn, D.H., Simonovic, S.P., 1999. Short term streamflow forecasting using artificial neural networks. J. Hydrol. 214 (1–4), 32–48.
Zhang, Z., Zhang, Q., Singh, V.P., Shi, P., 2018. River flow modelling: comparison of performance and evaluation of uncertainty using data-driven models and conceptual hydrological model. Stoch. Environ. Res. Risk Assess. 1–16.