Article type : Technical Paper

Streamflow Forecasting Using Singular Value Decomposition and Support Vector Machine for the Upper Rio Grande River Basin

Swastik Bhandari, Balbhadra Thakur, Ajay Kalra, William P. Miller, Venkat Lakshmi, and Pratik Pathak

Department of Civil and Environmental Engineering (**Bhandari, Thakur, Kalra**), Southern Illinois University Carbondale, Carbondale, Illinois, USA; Weather Forecast (**Miller**), NOAA Colorado Basin River Forecast Center, Salt Lake City, Utah, USA; Department of Engineering Systems and Environment (**Lakshmi**), University of Virginia, Charlottesville, Virginia, USA; and Water Resources (**Pathak**), FTN Associates, Ltd., Little Rock, Arkansas, USA (Correspondence to Kalra: kalraa@siu.edu).

Research Impact Statement: Long-term streamflow forecasting utilizing climate information is useful for resource planning and management in water stressed regions.

ABSTRACT: The current study improves streamflow forecast lead-time by coupling climate information in a data driven modeling framework. The spatial-temporal correlation between streamflow and oceanic-atmospheric variability represented by sea surface temperature (SST), 500-mbar geopotential height (Z_{500}), 500-mbar specific humidity (SH_{500}), and 500-mbar east-west wind (U_{500}) of the Pacific and the Atlantic Ocean is obtained through singular value decomposition (SVD). SVD significant regions are weighted using a non-parametric method and

utilized as input in a support vector machine (SVM) framework. The Upper Rio Grande River Basin (URGRB) is selected to test the applicability of the proposed model for the period of 1965-2014. The April-August streamflow volume is forecasted using previous year climate variability, creating a lagged relationship of 1-13 months. SVD results showed the streamflow variability was better explained by SST and U_{500} as compared to Z_{500} and SH_{500}. The SVM model showed satisfactory forecasting ability with best results achieved using a 1-month lead to forecast the following 4-month period. Overall, the SVM results showed excellent predictive ability with average correlation coefficient of 0.89 and Nash-Sutcliffe efficiency of 0.79. This study contributes towards identifying new SVD significant regions and improving streamflow forecast lead-time of the URGRB.

(KEYWORDS: Oceanic-atmospheric variability; streamflow; forecast; singular value decomposition; support vector machine.)

INTRODUCTION

Water has become a major natural commodity in the Western United States, where limited water availability has been exacerbated by past frequent droughts (Willey and Graff, 1984; Rice et al., 2009). Extreme hydrologic events such as floods and droughts are associated with hydro-climatic variability; improved knowledge of that variability in response to climatic fluctuations is crucial to mitigating social and economic impacts (Redmond and Koch, 1991). Several studies (e.g., Christensen et al., 2004; Stewart et al., 2004; Nijssen et al., 2001) have shown that climate change can result in increased uncertainty of water availability ranging from the watershed to global scale. In 2016, the United States Army Corps of Engineers issued Engineering and Construction Bulletin No. 2016-25 (ECB 2016-25) incorporated that climate change should be considered for all federally funded projects in planning stages. ECB 2016-25 provisioned qualitative analysis of historical climate trends, as well as assessment of future projections. As the impacts of climate change to the hydrologic characteristics of a basin are realized, streamflow forecasting can become difficult for hydrologists and climatologists as past hydrologic conditions are no longer representative of future conditions (Thakali et al., 2016; Pathak et al., 2016; Tamaddun et al., 2017). It is important to understand the relationship between climate variability and the hydrologic response of a basin such that sustainable and

efficient management of water related systems can be implemented (Middelkoop et al., 2001; Pahl-Wostl, 2007; Kundzewicz et al., 2009).

The dominant drivers of climatic variability affecting the hydrologic cycle all over the world and primarily in the U.S. include the El Niño Southern Oscillation (ENSO), Pacific Decadal Oscillation (PDO), Atlantic Multi-decadal Oscillation (AMO), North Atlantic Oscillation, Artic Oscillation, and Pacific-North America Pattern. Throughout the U. S., these teleconnection patterns are significant predictors of hydrologic response (Dettinger et al., 1998; McCabe et al., 2004). Sea surface temperature (SST), atmospheric pressure, humidity, and wind are the major ocean-atmospheric variables that have wide influence in explaining the hydrologic variability of a region (Woodruff et al., 1987). SST variability has been utilized to find teleconnections between streamflow, precipitation, and snowpack. Traditional predefined indices have shown consistent results in specific areas such as El Niño phase influences on the southwest, southeast, and northwest U.S. regions (Kahya and Dracup, 1993). Although the identification of predefined SST regions in the Pacific and Atlantic aid in forecasting streamflow in a certain basin, it may not influence hydrology over all basins (Tootle and Piechota, 2006). Consideration of the entire Pacific and Atlantic Ocean SST avoids regional biases and may lead to improved streamflow estimates (Tootle and Piechota 2006). Studies have associated 500-mbar geopotential height (Z_{500}) anomalies with climate change (Wallace and Gutzler, 1981). Z_{500} is the elevation above mean sea level at which atmospheric pressure is 500-mbar. Z_{500} has been used as a significant predictor in climate forecasting models and has performed well (Grantz et al., 2005; Soukup et al., 2009; Sagarika et al., 2015). Precipitation is related to ocean evaporation and the movement of clouds; these components of the hydrological cycle are primarily impacted by humidity, wind speed, and air temperature. In order to fully address these components, two additional climate data included in this analysis are: zonal wind stress (U_{500}) (i.e., east-west wind force per unit area parallel to the surface of water bodies corresponding to 500-mbar atmospheric pressure) and specific humidity (SH_{500}), corresponding to 500-mbar pressure of both the Pacific and Atlantic Ocean. Munot and Kumar (2007) have utilized the zonal wind at different pressure level including 500 mbar pressure level to predict long range Indian summer monsoon rainfall and found the zonal wind was as important predictor as the temperature in forecasting the rainfall. Pathak et al. (2018) have used the oceanic east-west zonal wind at 500 mbar pressure to find the association between western U.S. snowpack and zonal wind and the study showed

significant relationship between wind speed and snow water equivalent of the considered region. Similarly, Bhandari et al. (2018) have used both zonal wind and specific humidity to evaluate the correlation between these ocean-atmospheric variables with the regional streamflow of the continental United States and found that both wind speed and specific humidity are strongly correlated with the streamflow variability of the United States.

Principal component analysis, singular value decomposition (SVD), canonical correlation analysis, and combined principal correlation analysis are some of the techniques commonly used to find interrelationship between two spatial and temporal fields (Wallace et al., 1992). Bretherton et al. (1992) applied afore-mentioned statistical methods to find the coupled relationship between two spatial-temporal variables and opted for SVD for its simplicity and robustness. Wallace et al. (1992) also concluded that SVD extracts the most significant modes of variability in comparison to other tools. Several studies (Wallace et al., 1992; Tootle and Piechota, 2006; Soukop et al., 2009) have been conducted to find the linkage and forecasting ability between large scale climate data and streamflow, snowpack or precipitation using SVD technique. Popular predefined indices such as ENSO, PDO, and AMO are conventionally used as predictors of streamflow while these predefined indices are the source of spatial biases. Utilization of SVD subsides the use of these predefined indices by obtaining unique spatial-temporal correlation pertinent to the considered study area. In order to improve the forecasting ability of a model, several data preprocessing techniques are available. In conjunction with data-driven modeling, singular spectrum analysis (SSA) and discrete wavelet transform (DWT) are most common preprocessing tools and these are efficient in eliminating discontinuity of data and reducing forecasting errors (Marques et al., 2006; Nourani et al., 2009). However, recent research by Du et al. (2017) presented the incorrect usage of SSA and DWT in developing hybrid models and showed that those models may cause significant forecasting errors.

Various conventional forecasting models such as conceptual and time series models have been employed for streamflow prediction. Multiple Linear Regression, Auto Regressive Integrated Moving Average are some of the conventional model extensively used for prediction of hydrological time series. However, these models do not represent the non-linear processes involved in precipitation-streamflow transformation (Zealand et al., 1999). These time series models utilize the concept of data stationarity and hence provide little applicability when dealing with non-stationary data. Artificial Neural Network (ANN) has emerged as a dynamic, self-

learning model capable of utilizing noisy, non-linear data in predicting hydrological time series without knowing the physical relationship between input and output data (Nourani et al., 2009). ANNs have been applied and performed well in non-linear processes involved in multivariable conditions. Recently, support vector machines (SVM) have received growing attention as a novel regression technique (Mukharjee et al., 1997; Pai and Lin, 2005). SVM uses a statistical machine learning approach in which available data are trained to predict series of data (Liong and Sivapragasam, 2002). It can minimize prediction error and reduce model complexity (Vapnik, 1995, 1998). SVMs evolve incorporating the noise and non-linearity in the training data without assuming the stationarity proving it ideal while analyzing hydrologic parameters affected by climate change. SVM uses the principal of structural risk minimization unlike the empirical risk minimization principle used by ANNs. SVMs have been extensively applied in various hydrological forecasting problems and have outperformed ANNs approach (Dibike, 2000; Babovic et al., 2000; Cimen and Kisi, 2009). SVM has shown superior generalization ability and it is successful in reducing the overfitting problem compared to ANN (Cimen and Kisi, 2009). Astuti et al. (2014) used SVD for preprocessing and feature extraction and the extracted data were used to forecast location, time, and magnitude of earthquakes using SVM approach and concluded that the proposed methods were relatively better than the other hybrid forecasting models.

Several of the previous data driven modeling studies using climate information to improve streamflow forecasts have focused on pre-defined oceanic indices rather than entire SST regions that do not introduce spatial bias. To overcome this limitation, this research proposes a novel-modeling framework that would couple a large-scale climate variability into a data driven model and that would eliminate the spatial bias at a regional scale. First, SVD is used to determine a lagged spatial-temporal correlation between April-August streamflow and oceanic-atmospheric variabilities represented by SST, Z_{500}, SH_{500}, and U_{500} of the Pacific and the Atlantic Oceans. SVD significant regions are weighted using non-parametric approach formulated by Piechota et al. (2001) and utilized as input in SVM framework. The study is conducted in the Upper Rio Grande River Basin (URGRB) for the period of 1965-2014 and the lagged relationship is computed for 1-13 months.

This study is expected to investigate the time-lagged relationship of the URGRB streamflow variability with the ocean-atmospheric variability of the Pacific and the Atlantic

Ocean. This research further aims to address the following research questions: (1) How is streamflow within the URGRB associated with ocean-atmospheric variables? (2) What are the dominant predictors among oceanic-atmospheric variables that best describe the streamflow variability of the basin? and (3) How does the proposed modeling framework improve the lead-time of the streamflow forecast? Previous studies on streamflow forecasting in the URGRB have primarily focused on SST influence while the current research includes Z_{500}, SH_{500}, and U_{500} data for the analysis. Including these additional variables broadens the scope of the forecasting ability presented here and identifies significant SH_{500} and U_{500} regions in Pacific and Atlantic Ocean.

STUDY AREA AND DATA

Study area

The Rio Grande River is one of the major rivers in the United States, which originates in southwestern Colorado, flows through New Mexico and Texas in a southeasterly direction, and discharges into the Gulf of Mexico. The Rio Grande River, which is approximately 3,051 kilometers in length with a catchment area of 472,000 square kilometers, is a major source of water in southern states. More than three million people, agriculture, industries, and wildlife in Colorado, New Mexico, and Texas have been supported by the Rio Grande water supply (Michelsen and Wood, 2003; Booker et al., 2005). During drought conditions, the water allocation conflict among the users is considered among the most intense in the United States (US Department of Interior, 2003). Increased demand, over-allocation of water, and vulnerability to drought and climate change have created and added complexity in active water regulation and allocation in the URGRB region (Booker et al., 2005). The socio-economic importance of the river motivates the need for improved streamflow prediction several months in advance.

Data

The primary datasets used in analysis are streamflow data for six unimpaired gages in the URGRB and oceanic-atmospheric climate data represented by SST, Z_{500}, U_{500}, and SH_{500}. United States Geological Survey (USGS) Hydro-Climatic Data Network 2009 (HCDN-2009) provides the list of streamflow stations which have minimal impact from human activities such as construction of diversion, artificial dams or any activities which can affect the natural flow of

streams. The streamflow data from these stations are suitable for the analysis of hydrologic variations and trends for the present climatic context (Lins, 2012). Slack and Landwehr (1992) identified 1659 unimpaired streamflow stations in the United States (Lins, 2012). However, for the RGRB, it has been found that only six streamflow stations have minimal impact from human activities which are located in the upper region of the Rio Grande River Basin. These six stations from the upper region of the basin are the reason for selection of the Upper Rio Grande River Basin. The mean monthly streamflow values from those streamflow stations are extracted from USGS website (http://www.usgs.gov/) for 1965 to 2014. Monthly streamflow volumes from April through August are summed to develop seasonal streamflow volumes for the analysis. Figure 1 illustrates the location of six unimpaired streamflow stations. It is commonly observed that the daily streamflow has high uncertainty and it is difficult to find a time lagged relationship between oceanic-atmospheric data and daily streamflow data. To have higher accuracy in the prediction and to have a lump sum idea about the seasonal streamflow volume, April-August streamflow volume is used since seasonal variation of streamflow is typical in snow-fed rivers of the United States. Further, spring-summer streamflow accounts for the major flow volume of the year and can help water managers to create balance between annual future water demand and annual water availability. Additionally, seasonal analysis of streamflow with climate variability is preferred to water-year analysis because the water-year analysis does not effectively capture the seasonal interaction of streamflow and climatic variables (Sagarika et al., 2015). The analysis, therefore, aims to capture the seasonal relationship of streamflow and climate variability adequately.

National Oceanic and Atmospheric Administration (NOAA) Physical Sciences Division (http://www.esrl.noaa.gov/psd/data/gridded/) is the source of SST data for both the Pacific and Atlantic Oceans. The mean monthly SST data is extracted from 2^o by 2^o grid cells and the spatial extent of SST data in the Pacific Ocean is 100^oE to 80^oW longitude and 30^oS to 70^oN latitude. The extent for the Atlantic Ocean is 80^oW to 20^oW longitude and 30^oS to 70^oN latitude. The mean monthly SST data was divided into three periods: December to February of the previous year, September to November of the previous year, and December to February of the current year covering a period of 50 years (1964-2013). For example, if streamflow is predicted for April-August of 2010, monthly average SST data for December 2008 to February 2009, September to November of 2009, and December 2009 to February 2010 are considered in the

analysis for the three periods. The lead-time in the analysis is defined as the time lag from the last month of SST period to the first month of streamflow period. 1-month lead-time i.e., February to April, 4-month lead-time i.e., November to April, and 13-month lead-time i.e., previous year's February to current year's April are considered as the three forecast lead-times in this study.

In addition to SST, other data representing the ocean-atmospheric variability are Z_{500}, U_{500}, and SH_{500} and these data the product of National Centers for Environmental Prediction /National Center for Atmospheric Research Reanalysis Project (Kalnay et al., 1996). NOAA Physical Science Center (http://www.esrl.noaa.gov/psd/data/gridded) provided the mean monthly Z_{500}, U_{500}, and SH_{500} data from 1964 to 2013. These data are obtained from 2.5^{o} by 2.5^{o} grid cell for both oceans and the spatial extent and division of data is kept the same as that of SST data.

METHODOLOGY

The methods used here are divided into four steps:

1. Establishing correlation between two variables using SVD
2. Screening of predictors
3. Predicting streamflow using SVM
4. Model evaluation

The flowchart in Figure 2 summarizes the model algorithm to forecast the streamflow from the ocean-atmospheric variables with different lead times. In first step, SVD is applied to find the spatial-temporal correlations between the streamflow data and the climate variables that results in the temporal expansion series (TES) of significant modes explained later. These TES are screened in the second step. The screened predictors are used as the input for the SVM model of each streamflow station independently. Next, the forecasted streamflow is evaluated by comparing the forecasted and observed streamflow using statistical and graphical aspects. A brief description of the methods abstracted from several sources is provided in the ensuing sections. Interested readers are referred to original references for detailed descriptions (Bretherton et al., 1992; Piechota et al., 2001; Vapnik, 1995).

Establishing correlation between two variables using SVD

SVD is a simple and robust statistical technique primarily useful for differentiating major modes of variability out of extensive series of data. SVD evaluates a cross-covariance matrix between two fields and identifies the correlation between these fields (Bretherton et al., 1992). Each matrix has spatial component represented by $SST/Z_{500}/SH_{500}/U_{500}$ cells or streamflow stations while temporal component is represented by total number of years of data in which temporal dimension of each matrix must be equal. As the SVD approach evaluates the association of streamflow data and climate data in both space and time, the obtained correlation is generally referred as spatial-temporal correlation in the study. First of all, standardized $SST/Z_{500}/SH_{500}/U_{500}$ anomalies matrix and standardized streamflow matrix are developed and a cross-covariance matrix (A) is obtained by multiplying $SST/Z_{500}/SH_{500}/U_{500}$ matrix with the transpose of streamflow matrix (Q^T) and divided by total number of years of data period (N).

$$A = \frac{SST \times Q^T}{N} \tag{1}$$

The cross-covariance matrix is then decomposed into three matrices by SVD as :

$$SVD\ of\ A = USV^T \tag{2}$$

where, $U^T U = I$ and $V^T V = I$ meaning U and V are orthogonal and normalized matrices whereas S is a diagonal matrix with non-negative values. A left singular vector and right singular vector are derived from the columns of those orthogonal and normalized matrices. First columns and rows of these orthogonal matrices explain more of the correlation between variables compared to subsequent rows/columns. The diagonal matrix provides the singular value of the parent matrix in non-increasing order and these values provide information about the properties of a matrix. SVD approach to data unfolding. https://arxiv.org/pdf/hep-ph/9509307.pdf. Accessed 25 September 1995). Isolation of the most important modes of data is calculated based on squared covariance fraction (SCF). SCF value shows the degree of variability explained by SVD analysis, which is defined as:

$$SCF_i = \frac{C_i^2}{\sum C^2} \tag{3}$$

where, C is the singular value for i-th mode. The SCF values more than 10% only are considered for the analysis. Similarly, normalized squared covariance (NSC) indicates the correlation between two fields averaged over all the grid points (Wallace et al. 1992). *NSC* is defined as:

$$NSC = \frac{C^2}{N_S \times N_Z} \quad (4)$$

where, where C^2 is the sum of singular values and N_S is the number of grid points while N_Z is the number of streamflow stations. The *NSC* value ranges from 0 to 1 with maximum value for perfect correlation between two variables. Next, temporal expansion series of left field (*LTES*) is obtained by multiplied by left singular vector (*L*) with $SST/Z_{500}/SH_{500}/U_{500}$ matrix, and similar procedure is followed for temporal expansion series of right field (*RTES*).

$$LTES = L \times SST \quad (5)$$

Finally, heterogeneous correlation map of left (right) field is developed by correlating $SST/Z_{500}/SH_{500}/U_{500}$ (streamflow) matrix with *RTES* (*LTES*) at 90% significance level using Pearson-r correlation coefficient. The heterogeneous correlation map shows the influential regions of the ocean-atmospheric variables with streamflow for different lead-time cases. Each streamflow station can have either positive or negative correlation with climate variables which is known as station significance. Station significance are obtained from the SVD analysis but neither their signs (station significance) nor streamflow stations are shown in the heterogeneous correlation map. Only the signs of climate regions are shown in the map. Based on the station significance, positive or negative correlation of streamflow with the climate variables can be known from the heterogeneous correlation map. If the station significance and a particular region of climate variables are showing the same sign in the map, then the streamflow and the climate variables of the region are positively correlated and if the station significance and the region have different sign then there exists negative correlation between the variables.

Screening of predictors

The temporal expansion series of four different ocean-atmospheric variables obtained from SVD analysis are the possible predictors of streamflow in the Rio Grande. For each climate variable/predictor, a continuous exceedance probability is developed using the procedures from Piechota et al. (2001). First, temporal expansion series of each variable for each year is arranged with corresponding streamflow value. For an observed streamflow value Q_i, a greater than and less than streamflow category are created and corresponding to those categories, predictors are separated into different subsets. Bayes probability theorem is then applied to find the forecast probability of each category from the predictor values.

$$Prob(Q_i \,/\, X) = \frac{p_i f_i(x)}{\sum_{i=1}^{k} p_i f_i(x)} \tag{6}$$

where, X = predictor value; Q_i = streamflow value of category i; p_i = prior probability of streamflow of category i; $f_i(x)$ = probability density function (PDF) of prior X value of category i.

For each subset of predictors, a probability distribution is fitted to calculate the PDF $f_i(x)$. A nonparametric approach is employed by using Kernel density estimator to calculate the PDF where Kernel density estimation is associated with a histogram (Silverman, 1998; Piechota et al., 1998). Kernel density estimation is defined as in equation (7).

$$f(x) = \frac{1}{hn} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{7}$$

where, x_1 to x_i = set of n observations; $K(\,)$ = kernel function; h = bandwidth, which is calculated as:

$$h_i = 0.9 A_i n_i^{-1/5} \tag{8}$$

$$A_i = min\left(\sigma_i, \frac{interquartile\ range}{1.34}\right) \tag{9}$$

where, σ_i = standard deviation of predictors in subset i; n_i = number of observations in each subset. Next, for each predictor value a unique probability is estimated and a forecast curve is developed by plotting probability value for all predictor values against corresponding streamflow values. A final exceedance probability forecast is obtained by combining the exceedance forecast of all variables. The skill of probability forecast is measured by Linear Error in Probability Space (LEPS) score approach introduced by Ward and Folland (1991). The LEPS score determines the distance between the forecasted and observed value over the cumulative probability distribution, which is defined as:

$$S = 3\left(1 - \left|P_f - P_v\right| + P_f^2 - P_f + P_v^2 - P_v\right) - 1 \tag{10}$$

where, P_f and P_v are the cumulative probability of forecasted and observed value respectively. A climatology or no-skill forecast is also developed through exceedance probability curve of observed streamflow values. The value of P_f is obtained from exceedance probability curve mentioned earlier while P_v is obtained from climatology exceedance curve. For a given predictor

value and corresponding streamflow value, P_f and P_v can be obtained and LEPS score is calculated as in equation (10).The LEPS score for each year is then calculated for all predictor values and the average skill (SK) for all years is calculated as:

$$LEPS\ SK = \frac{\sum 100S}{\sum S_m} \tag{11}$$

where, S_m is the sum of best or worst possible forecast depending whether S is positive or negative respectively. Best possible forecast occurs when $P_f = P_v$ while worst possible forecast occurs when $P_{f=}$ 1 or 0. Similar process is applied for different predictors and for each streamflow station, skillful variables/predictors which give the highest LEPS SK score are then finally selected. LEPS score gives more weightage to those forecasts which predict high or low streamflow or extreme value in general while less weight is given to those forecasts which predict average streamflow value. A skillful forecast has a 10% or higher LEPS SK score (Potts et al., 1996). A flowchart for the predictor screening process is shown in Figure 3.

Predicting streamflow using SVM

The best combinations of predictors selected are then taken as input for SVM modeling. Unlike traditional learning methods that use an empirical risk minimization principle, SVM uses a machine-learning approach, and this formulation involves a structural risk minimization principle. The application of support vector regression (SVR) is briefly described here. The descriptions and equations are abstracted from Ahmad et al. (2010).

Suppose a training data set with input and output variable represented as, $\{x_i, y_i\}^N$ where $x_i \epsilon R^p$ represents independent input variable, and $y_i \epsilon R$ represents dependent output variable. We need to find a function $y = f(x)$ that provides the dependency relationship of these two variables. The function can be written as in equation (12):

$$y = f(x) = \langle w, x \rangle + b \tag{12}$$

where, $\langle w, x \rangle$ is the dot product of weighting vector w and input vector x; b is a bias. In addition, the optimization problem and equality constraints are formulated and shown below in equation (13).

$$\text{Minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

Subject to:

$$\begin{cases} y_i - \sum_{j=1}^{K}\sum_{i=1}^{N} w_j x_{ji} - b \le \varepsilon + \xi_i \\ \sum_{j=1}^{K}\sum_{i=1}^{N} w_j x_{ji} + b - \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \quad (13)$$

where, ε represents Vapnik's insensitive loss function. *C* is the capacity parameter cost, ξ_i and ξ_i^* are slack variables, and *K* is the number of support vectors are represented in the formulation. The goal is to determine optimal parameters, which minimizes the forecasting error for the SVR model. The optimization of SVM is based on the selection of a kernel function that utilizes non-linear mapping in the feature space (Dibike et al., 2001). Radial basis function kernel is used in the current framework which shows superior efficiency by minimizing test error (Scholkopf et al., 1997). For the detail description of support vector machine, interested readers are advised to go through Vapnik (1995, 1998).

The performance of the model is tested by training the data and validating to the remaining data sets. The training phase intends to find the optimal values of the parameters and attain the best possible generalization conditions. This research utilizes the leave-one-out cross validation approach commonly known as special case of *k*-fold cross validation that overcomes the data splitting problem when limited data sets are available for training and testing (Kalra et al., 2012). In this approach one data point is selected to test the model while the remaining data points are used for training phase. This process is then applied to next data point and repeated for all data sets accordingly.

Model Evaluation

SVM performance is evaluated based on various statistical and graphical measures. Time-series plots are used to depict the trend of observed and forecasted streamflow over the years while scatter plots demonstrate the correlation between observed and predicted streamflow values. Similarly, box plots show the statistical variation of streamflow values and non-exceedance probability plots are used to visualize the estimation error at different probability scenarios. The statistical measures utilize correlation coefficient (r), Nash Sutcliffe model efficiency (NSE), percent bias (PBIAS), and LEPS SK values to evaluate model efficiency. Collinearity between observed and predicted streamflow values are accessed through correlation coefficient. Higher correlation depicts less variance in the data. NSE determines the forecasting

ability of models by comparing the measured data variance with residual variance of observed data and determines the accuracy of a model (Nash and Sutcliffe, 1970). The skill of fitting predicted and measured data to a 1:1 line is explained by NSE. It is calculated as shown in equation 8:

$$NSE = 1 - \left[\frac{\sum_{i=1}^{n}\left(X_i^{meas} - X_i^{predi}\right)^2}{\sum_{i=1}^{n}(X_i^{meas} - X^{mean})^2}\right] \quad (14)$$

where, X_i^{meas} is the measured quantity for *i*-th item and X_i^{predi} is the corresponding predicted amount by the model and X^{mean} is the mean of measured quantity for *n* number of observations. The range of NSE is -∞ to 1with 1 as the ideal value. The recommended range of NSE values indicating a satisfactory model is NSE >0.5 (Moriasi et al. 2007). The percent bias is a measure of the average tendency of forecasted value higher or lower than observed value (Gupta et al. 1999). PBIAS value is calculated as shown in equation (9):

$$PBIAS = \left[\frac{\sum_{i=1}^{n}\left(X_i^{meas} - X_i^{predi}\right) \times 100}{\sum_{i=1}^{n}(X_i^{meas})}\right] \quad (15)$$

where, X_i^{meas} is the measured quantity for *i*-th item and X_i^{predi} is the corresponding predicted amount by the model and X^{mean} is the mean of measured quantity for *n* number of observations. The ideal value for PBAIS is 0 and smaller values show good model simulation while negative values show overestimation and positive values show underestimation (Gupta et. al. 1999). The LEPS SK score is also utilized to evaluate the model performance by determining the distance between measured and predicted streamflow values in the cumulative probability distribution, which has been already described in the methodology section.

RESULTS AND DISCUSSION

The results and discussion are described in three different sections. The SVD and SVM analysis are presented in three different sub-sections for each lead-time scenario while predictor screening analysis is discussed in a single section. The first section discusses the SVD spatial-temporal correlation of streamflow with oceanic-atmospheric variables. Next, results of predictor screening are presented followed by the SVM analysis.

SVD analysis

1-month lead-time. The SVD analysis of the Rio Grande streamflow with Pacific/Atlantic SST, Z_{500}, SHU_{500} and U_{500} resulted in the identification of significantly correlated regions. Most of the variability of the streamflow in the URGRB was explained by the first mode of SVD and therefore, only the first mode SVD results are reported throughout the section. Table 1 presents the SCF and NSC values obtained for different lead-time scenarios.

Figure 4a represents the relationship between Pacific Ocean ocean-atmospheric variability with URGRB streamflow through heterogeneous correlation map at the 90% significance level for the 1-month lead-time period. Resulting significant SST regions are shown in red and blue color. The red and blue color in the map indicates positive or negative correlation of streamflow with climatic variables in the significant regions. One of the two key significant regions identified for that period are the regions off the coast of Japan, Indonesia and Australia, which has a horseshoe shape, is negatively correlated with the April-August streamflow in URGRB. This identified region is similar to ENSO including the popular Niño 3.4 region previously identified by Trenberth (1997). The identified region also include northwestern US coastal regions representing PDO. Previously, Khedun et al. (2012) has also identified similar results – indicating ENSO and PDO being positively correlated with winter and spring precipitation which is the source of spring summer streamflow. The obtained SST regions-mostly ENSO are affirmed by previous literatures signifying the URGRB being wetter and colder during El Niño years because of the modifications in the mid latitude jet streams. The reason of ENSO being positively correlated with the streamflow can be attributed to the feeding of moisture to the Jet streams moving towards east from the Pacific as a result of above normal SST in ENSO regions during El Niño years. Another dominant region of SST that has a strong positive correlation is the region extending from West to Central Pacific Ocean bounded in between 90^0 W to 180^0 W latitude, and this region shows conformity with Niño Index as demonstrated by other researchers (Rajagopalan et al., 2000). Figure 5a shows a heterogeneous correlation map of Atlantic SST significant regions for the 1-month lead-time period. The identified significant area is separated into two zones, one is near the east coast of Canada and US resembling the AMO region and the other is near the north shore of Brazil. These regions have a negative correlation with streamflow variability. This is also verified by previous literature that cold north Atlantic SST in winter and spring favors the spring summer streamflow. (Trenberth et al., 1998; Pascolini-Campbell et al., 2017).

The second column of Figure 4a and Figure 5a show the heterogeneous correlation map of significant Z_{500} regions for both oceans. The significant regions for Pacific are more concentrated in equatorial regions of the Pacific Ocean and southeastern Asia. The altitude of Z_{500} is approximately 18000 feet above the sea level, and this has been associated with diverse weather phenomena (Soukup et al., 2009). The jet stream formation is related to locations where the Z_{500} contour lines are concentrated. Z_{500} is found to be more influential during wintertime. The shortwave train as shown in the second column of Figure 4a with red arrow head as a result of warming of SST in ENSO region signifies the fueling of Jetstream with moisture also responsible for the precipitation and streamflow of URGRB region. This physically explains the reliability of obtained teleconnection between Z_{500} regions and the streamflow of URGRB. For the Atlantic Ocean, the significant regions identified are clustered at Eastern Canada and north to mid-South America. All those identified regions show the positive correlation with the streamflow of the URGRB.

Five significant SH_{500} regions were prominent in the Pacific Ocean. Three regions with positive correlation are identified at the equatorial region and above the mid-United States while two negatively correlated specific humidity regions were identified at the eastern side of China, Indonesia, and Japan. As mentioned earlier, the warmer than average ENSO region fuels the jet stream with moisture over Pacific moving in east directions responsible for URGRB streamflow. With the extra fuel and pressure the Jetstream shift eastward with higher than normal precipitation in URGRB region. This can be verified in Figure 4a third column, the specific humidity in ENSO region being positively correlated with streamflow of URGRB region. Positive correlated SH_{500} regions are found near the basin area that may be linked to the direct relationship of distance with humidity influence making further regions less influential in streamflow variation. For the Atlantic Ocean, the significant regions were identified at eastern Canada and northern South America. These regions have shown the positive correlation with the streamflow of the basin.

Several significant U_{500} regions are established in the Pacific Ocean as probable predictors of streamflow. The U_{500} regions signifies below/above normal wind in east west directions which are the key cause of circulation of moisture from the ocean to any watersheds. Here, these complex U_{500} regions are the result of complex interactions of pressure, geographic features, temperature gradients and other climate variables. One significant region in the Atlantic

is found near the northeast coast of Canada, which has shown the negative correlation. The small spatial extent of the region indicates a little influence of Atlantic U_{500} on the streamflow variability.

4-month lead-time. The SCF and NSC values are both comparable for 1- and 4-month lead-time scenarios. The first plot of Figure 4b and Figure 5b show the heterogeneous correlation map for September-November SST of Pacific and the Atlantic Ocean respectively. The significant Pacific SST regions are almost the same for 4-month lead-time when compared to 1-month lead-time. However, for the Atlantic, SST regions are continuously extended in a greater area in 4-month lead-time, unlike 1-month lead-time where the significant regions are widely separated in two zones. The SCF and NSC values also dropped slightly in comparison to 1-month lead-time. Significant regions are located in the west to the east region of the northern Pacific Ocean in smaller groups. Furthermore, the spatial extent of significant Z_{500} regions also decreases in this period compared with a previous period in the Pacific Ocean while the spatial extent slightly increases for the Atlantic Ocean. It can be noted that SCF/NSC values for Pacific are smaller than that of Atlantic in this period. The identified indices for 4-month lead-time were similar to the identified indices for 1-month lead-time but the signals were weaker with the expense of lead times. As observed in Figure 5b, the indices like ENSO and the climatic phenomenon like short wave terrain and impacts of mid latitude jet stream are still evident like 1-month lead-time.

Compared to 1-month lead-time, the SCF and NSC for SH_{500} decrease for the Pacific while increases for the Atlantic Ocean in 4-month lead-time. The third plot of Figure 4b and Figure 5b show heterogeneous correlation map for Pacific and the Atlantic Ocean respectively. Similar to SST and Z_{500}, significant SH_{500} regions become separated from each other and smaller in areal extent in this period as compared to 1-month lead-time period. Nearer SH_{500} regions to the basin show positive correlation as in the previous 1-month led time case. The SCF and NSC values were higher in 4-month lead-time in comparison to 1-month lead-time. Furthermore, the majority of positively correlated associated regions are prevalent both in Pacific and in Atlantic regions that are clear from Figure 4b and Figure 5b. All other predictors, with the exception of U_{500}, have shown comparatively better results for 1-month lead-time when compared with 4-month lead-time.

13-month lead-time. The first plot of Figure 4c and Figure 5c show the heterogeneous correlation map of SST for Pacific and the Atlantic Ocean respectively. From these figures, it is clear that the significant SST areas become smaller and sparser when compared to smaller lead-time cases. It may be due to the influence of longer lead-time SST period is less effective than that of a shorter lead-time period. A similar drop of SCF and NSC values were obtained for SVD analysis of Z_{500}. The number of significant regions decreases and these regions move farther away from the ocean as compared to 4-month lead-time as clearly seen in Figure 4c and Figure 5c. For the Atlantic, only one significant region near the northeast coast of Canada is identified, with negative correlation with streamflow. As seen in Figure 4c and Figure 5c, the spatial extent of significantly teleconnected regions for 13-month lead-time was smaller than other lead times however, these 13-month long lead spatial-temporal associations can also help water managers by providing longer time window for planning and mitigation measures.

The heterogeneous correlation map for Pacific and Atlantic Ocean SH_{500} are shown in third plot of Figure 4c and Figure 5c. The spatial extent of the significant regions decrease considerably for 13-month lead-time case in comparison to 1- and 4-month lead-time cases; furthermore, the drop in SCF and NSC values also suggest that the 1- and 4-month lead-time period can have better forecasting abilities as compared to 13-month lead-time period. The forth plot of Figure 4c and Figure 5c show the heterogeneous correlation map for U_{500} for the Pacific and the Atlantic Ocean, respectively. In this period, number and spatial extent of significant regions decrease considerably similar to SST, Z_{500}, and SH_{500}. All the predictor variables have shown better results for smaller lead-time cases than longer lead-time cases.

The SVD results depict the identification of significant regions of SST, Z_{500}, SH_{500}, and U_{500} in the Pacific and the Atlantic Ocean, which are teleconnected with the streamflow stations in the Rio Grande River. The identification of various significant regions of climatic parameters in this study indicate a dominant influence of these regions on the streamflow variability and can provide better predictive capabilities than other regions. Moreover, these identified regions were found to be similar with results conducted by previous researchers. The lagged SVD analysis clearly showed that smaller lead-time analysis has better forecasting ability as compared to longer lead-time analysis. The inclusion of entire Pacific and Atlantic Ocean for SST, Z_{500}, U_{500} and SH_{500} data has eliminated the regional biases and the dependence on the existing indices to explain the hydrologic variability has reduced.

Predictor screening analysis

The abovementioned SVD analysis presented all the possible correlations of streamflow with the oceanic-atmospheric climate variables whereas this predictor screening analysis focuses on best correlation for each of the stations leading to screen the best predictor variables out of the eight climate variables. Table 2 shows two best streamflow predictors for each streamflow station for each scenario. The result shows that each variable is one of the best predictors at least once. It is clear from the table that different predictors are dominant at different lead-times. For 1-month lead-time, Atlantic SH_{500} was found to be the best overall predictor while Pacific U_{500} was found to be the best overall predictor for both 4- and 13-month lead-time scenario. Similarly, Atlantic Z_{500} was the least dominant variable followed by Pacific Z_{500}. Previous researchers have primarily focused on SST and Z_{500} as the important variables in explaining various hydrologic processes. However, this research has included two more climatic variables i.e. U_{500} and SH_{500} apart from SST and Z_{500} for broader scope. These included variables have shown valuable predictive information. For example, for stations 4 and 6, SH_{500} of Pacific Ocean could play a key role in forecasting streamflow at those locations, while Atlantic Ocean SH_{500} could be an important predictor of streamflow for station 1. The inclusion of these variables can be justified and supported as both U_{500} and SH_{500} explain the majority of streamflow variability in the URGRB. Results also suggest that the U wind over the Pacific Ocean is one of the major climatic factors that drive the variability of streamflow in the URGRB. Satisfactory performance of U_{500} and SH_{500} indicate that these climate variables have greater potential in providing finer results if they are extensively studied and understood. Thus, these variables have possibility of drawing research attention from climatologists and hydrologists in coming days.

The positive skill shown by continuous exceedance probability forecast depicts the improvement of prediction skill compared against a climatological forecast where the temporal aspect of historical data is taken into account. The continuous exceedance probability forecast labeled as good, fair, and poor compares climatology forecast, modeled forecast, and observed streamflow value and gives an idea about the availability of streamflow at different risk levels in a simple and efficient way. The continuous exceedance probability forecasts labeled as good, fair, and poor forecast are shown in Figure 6. For good forecast, the difference between observed and predicted streamflow value at certain probability is minimum while for poor forecast the difference becomes larger. From Figure 6c, it is clear that when streamflow was predicted for

2014 at 50% exceedance probability, the model predicts 12 million cubic meters (MCM) while the climatology predicts 18 MCM and the observed value is 11 MCM. However, for poor forecasts certain amount of risk is also present as the forecasted value deviates from the observed value. For majority of years higher number of good forecasts were observed compared to the poor forecasts. In addition, most of the stations showed higher LEPS SK for 1-month lead-time implying greater confidence in forecasting with smaller lead-time.

SVM Analysis

SVM analysis was used to predict streamflow volumes using input variables from three lead-time cases for all the six stations for a 50-year period. The predicted streamflow is then compared with observed streamflow, and the performance of the SVM model is described in coming sections.

1-month lead-time. Figure 7a shows time-series graphs showing the volume of simulated and observed streamflow for the 1-month lead-time scenario. Table 3 presents the values of various model performance parameters obtained for different lead-time cases. The dotted line represents observed streamflow values, and the solid line represents predicted streamflow values. It can be seen that the observed and predicted streamflow volume are fitted well, but some small discrepancies are also present. In addition, the predicted and measured streamflow have similar volumes. Simulated streamflow is found to be in almost perfect correlation with the measured streamflow for the year 2000 for all the stations. Similarly, Figure 7b shows the scatter plots for the 1-month lead-time scenario. The points lying above the bisector line indicate the prediction is overestimated, while those lying below the line indicate the predictions are underestimated, and points along the line represent perfect predictions. It can be seen that most points lie along the 45^{o} diagonal showing perfect correlation. This indicates that the forecasted and observed streamflow are in good correlation with each other. Better performance at low flows compared to high flows is apparent on the plots. It can be inferred that model applicability is best achieved during low flow events, which may indicate that the model is well suited for drought conditions.

The scatter plot also illustrates the PBIAS value and correlation coefficient. It is clear from the Figure 7b that each of the stations has a PBIAS value less than 10%. For the 1-month lead-time period, PBIAS value has an average value of 2.11 for all streamflow stations. The

average correlation coefficient for observed and simulated streamflow for the stations was 0.89. The higher correlation coefficient further strengthens the forecasting capability of the SVM model. It implies that the model performs well with less error variance. The effectiveness of SVM model is also evaluated through interpretation of the NSE value. Figure 7b shows the NSE values for all the six stations for the 1-month lead-time case. The average NSE value for all the stations was 0.79. These higher values of NSE statistics indicate that the SVM predicted streamflow was satisfactory.

The box plot of observed and simulated streamflow volume of the model for the December-February period is illustrated in Figure 7c. The horizontal line is the median streamflow while the interquartile range in between 25th percentile and 75th percentile is indicated by the box height. The whiskers in the plot represent extreme 5th and 95th percentiles. The median value for both observed and simulated streamflow values are similar at all sites. Although the interquartile range of estimated streamflow is slightly smaller than observed value, the fifth percentile of both datasets has a closer match. It is clear from the figures that the interquartile range of measured streamflow is wider as compared to that of predicted streamflow. This illustrates the uncertainty in forecasting ability for high flow range. Furthermore, the model underestimated the high flow, as most of the predicted high flows are smaller than the observed high flow.

Figure 7d shows the non-exceedance probability plot for the 1-month lead-time scenario. The y-axis represents the percentage of cumulative estimation error and the x-axis is the percentage of predicted data sample which is less than or equal to the value on the x-axis. The dotted line in the plot represents the cumulative modeling error value of 10%. Based on the plot it is clear that at 60% estimate of streamflow, the probabilistic absolute error is around 2% for almost all sites. As per the non-exceedance probability plot, site 3 gives best result, as the absolute error is just 10% at 80% estimate while other sites have more than 10% error for 80% sample estimates. Based on the plot, it is clear that smaller prediction error is achieved at higher estimation percentage, which in turns implies the greater confidence in prediction of streamflow for the water managers. The plot also tells the average skill score to evaluate the performance of the model forecast using LEPS approach. All the stations have LEPS SK more than 60%. The average LEPS SK score value was 72.2% for the streamflow stations. These higher value of LEPS SK score further support good forecasting capability of the model.

4-month lead-time. Figure 8a and Figure 8b show the time-series plot and scatter plot of simulated and observed streamflow values for September to November period, respectively. Similar to 1-month lead-time the time series plot for this period show similar trends. Scatter plots for this period are also similar to those shown at the 1-month lead-time period. Most of the points lie over the bisector line while high flow points are below the line implying high flow values are underestimated by the model. In the September-November case, average PBIAS value was 2.29 at all stations. Compared to the 1-month lead-0time, the PBIAS value has slightly increased at the 4-month lead-time. The correlation coefficient for the stations has an average value of 0.89. The average value of correlation coefficient for 1-month lead-time is equal to that of 4-month lead-time. . Four stations have NSE values higher than 0.8. The average NSE value was 0.78, which is slightly lower than the 4-month lead-time NSE average. In this period, high flows were also underestimated by the model as seen from the box plot of Figure 8c. The interquartile range of forecasted streamflow for station 6 was the smallest among the stations while the ranges were comparable to one another for rest of the stations. The non-exceedance probability plot for this period is shown in Figure 8d. The average LEPS SK score was 74% that is slightly higher than the 1-month lead-time case.

13-month lead-time. Figure 9a shows the time-series plot of simulated and observed streamflow values for the 13-month lead-time case. This graph also tells the higher prediction skill of the model. Despite the longest lead-time scenario, the 13-month lead-time scenario shows satisfactory forecasting results. Figure 9b is the scatter plot for this period. The average PBAIS value for this period was 2.87. When PBIAS values are compared, the 1-month lead yields the best results followed by 4- and 13-month lead-times respectively. Even though the smallest lead-time forecast demonstrates the best forecasting ability, the 13-month lead-time forecasting ability of the model is still comparatively satisfactory. The correlation coefficient also follows the same trend. The average R-value in this period was 0.87, slightly smaller than 1- and 4-month lead-time scenarios. When the model capability was measured based on NSE value, the 13-month lead-time case gave the smallest average value at 0.74, a NSE value that is considered satisfactory. It was observed that the NSE value improves as the lead-time period decreases, further supporting anticipated better forecasting ability for the lowest lead-time scenario.

Figure 9c shows the box plot at the 13-month lead-time depicting the comparison of modeled and observed streamflow for six streamflow stations. The boxplot for this period showed similar results from previous lead-time periods because, for all three scenarios, the interquartile range of measured streamflow is bigger as compared to that of predicted streamflow. This signifies that all the three scenarios captured both high and low flows, but underestimated high flow, as most of the predicted high flows are smaller than the observed. It can be inferred from the box plot that the model is efficient at predicting the low flow of URGRB as compared to high flow and may be effective for water management during drought seasons. Figure 9d shows the non-exceedance probability plot for the 13-month lead-time case. The average LEPS SK score value was 69.8%. The average values for this lead-time scenario were the lowest among the three lead-time scenarios. The anticipated higher LEPS SK score for smallest lead-time is not seen here because the 4-month lead-time case has highest average value followed by 1-month lead-time case. The LEPS SK score at all the three lead times indicates the satisfactory forecasting capability of the model.

The SVM model incorporates important oceanic-atmospheric climate variables for improving prediction of URGRB streamflow. The SVM model evaluation by different graphical and statistical analysis resulted in satisfactory results. Scatter plots clearly indicate improved forecast ability. The higher R and NSE values derived between measured and predicted streamflow with consistently smaller PBIAS values further support model forecasting capability. NSE, PBIAS, and R values all indicate that the best forecasting could be achieved for the smallest lead-time while LEPS SK value showed better forecasting ability at the 4-month lead-time followed by 1-month lead-time. Box plots, scatter plot, and time-series plots also suggest the adequate predictability of the model. These plots indicate SVM analysis was able to perform well in capturing low flow and intermediate flow as compared to high flow. Overall, the SVM model performance for high flows is not much impressive as compared to the low flows. However, low flow is more critical in comparison to high flow considering the water management prospect because water resources scarcity is not a serious problem during high flow period (Sharma et al., 2015). One of the reasons for underperformance of SVM model in predicting some streamflow value, extreme value in general, is may be due to the presence of outliers and erroneous data in the training phase. It is found that the association of training data and output data is instrumental in the model performance and longer period of data may show

higher generalization ability (Ahmad et al., 2010). Another reason of underperformance of SVM model could be due to the input variables used in the model that might not have adequately represented the physical system governing the generation of streamflow. Possible future research may further investigate in improving forecasting of high flow by exploring underlying hydro-climatic processes.

CONCLUSION

The primary goal of the study was to develop a modeling framework for improving streamflow lead-time in the URGRB using large-scale climate information of the Pacific and Atlantic Oceans. Spatial-temporal relationship of streamflow with each climate variable represented by SST, Z_{500}, U_{500}, and SH_{500} was analyzed by SVD approach for three lead-time cases. SVD temporal expansion series for each variable was weighted and screened by a non-parametric approach. These screened variables were used as input in SVM model to predict streamflow at six unimpaired streamflow stations within the URGRB. Overall the proposed research framework of combining several statistical approaches coupled with climate information to improve streamflow forecast lead time provides useful insights in regional hydrology.

The first research question was answered by the SVD analysis as it showed the association of URGRB streamflow with ocean-atmospheric variables of the Pacific and the Atlantic Ocean SVD analysis resulted in new significant SH_{500} and U_{500} regions in the Pacific and Atlantic Oceans in addition to SST and Z_{500} regions. The SVD analysis presented an extensive idea about all the possible associations between streamflow of the basin and ocean-atmospheric variability. Predictor screening analysis showed that Pacific SST and Pacific U_{500} are the two most dominant predictors for streamflow forecasting in the URGRB, which were the answer for the second research question. The teleconnection of streamflow with climate variables was sufficiently captured by the SVD study. The inclusion of SH_{500} and U_{500} climate variables led to identify associated significant regions for URGRB and showed equally competent potential for explaining streamflow variability of the basin. These variables have received little attention in previous research efforts. Moreover, the higher correlation of streamflow with U_{500} and SH_{500} shows that several other climate variables can be considered

together and studied extensively to fully understand the streamflow variability in a basin leading to better water resource management.

The study has shown that SVM model can be a useful method in streamflow forecasting by coupling an extensive range of climate variability with different lead-times. SVM model showed satisfactory forecast results for all the three lead-time cases. The best streamflow forecasting was achieved at the 1-month lead-time followed by 4-month lead-time scenario. The capability to improve long lead-time prediction can be helpful in efficient decision making process and various water management issues when the context of climate change are considered in the basin where snowmelt is the primary source of water. The model showed forecasting ability over the entire flow range, whereas forecasts at the low flow range were excellent. The third research question regarding the performance of proposed modeling framework was also answered as the SVM model satisfactorily predicted streamflow as supported by various performance parameters. The Rio Grande River heavily supports domestic use, agriculture, and industry. This river is highly utilized for water supply, and downstream supplies are significantly decreasing over the years. The basin has experienced low flows since 2000, and frequent droughts have been reported over the years. The rainfall pattern and water demand also differ considerably as more precipitation is observed in summer while the peak demand occurs in spring. The ability to capture low flows efficiently aids in water management during drought seasons and below average periods. Better forecasting of low flow events several months ahead may aid in the better allocation of water to competing users during dry periods.

The study doesn't assume stationary climate system and makes the assumption that stationarity is not valid. Relying on the appraisal of past climate by inferring from extreme events or changes in mean is not beneficial as the climate is constantly changing. At this moment, based on the results we obtained, conclusions can be drawn about the magnitude of change in streamflow in the future. While viewing these results, it must be looked at with a range of uncertainties as it is a statistical analysis. But, these uncertainties need to be looked at while making infrastructural investments as these decisions are irreversible. The streamflow predictions that the present study has made in terms of volume should be utilized by the water managers by making decisions so that these infrastructures can effectively respond to conditions that are changing and completely unknown. Detecting changes in past and future is not sufficient to make policy decisions and is the subject that needs more research on. Future work, may

explore extended lead-time scenarios. Additionally, the application of paleo data may provide promising results as data-driven models show higher efficiency for wide range of input data.

ACKNOWLEDGMENTS

The authors would like to thank three anonymous reviewers for providing valuable comments that helped in improving the overall quality of the manuscript. The authors are grateful to the Office of the Vice Chancellor for Research at Southern Illinois University Carbondale for providing support for the current research.

LITERATURE CITED

Ahmad, S., A. Kalra, and H. Stephen, 2010. Estimating soil moisture using remote sensing data: A machine learning approach. *Advances in Water Resources* 33(1): 69-80.

Astuti, W., R. Akmeliawati, W. Sediono, and M.J.E.Salami, 2014. Hybrid technique using singular value decomposition (SVD) and support vector machine (SVM) approach for earthquake prediction. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(5), 1719-1728.

Babovic, V., M. Keijzer, and M. Bundzel, 2000. From global to local modelling: a case study in error correction of deterministic models. In Proceedings of Hydroinformatics 2000, Vol. 4, No. 5.

Bhandari, S., A. Kalra, K. Tamaddun, and S. Ahmad, 2018. Relationship between Ocean-Atmospheric Climate Variables and Regional Streamflow of the Conterminous United States. *Hydrology*, 5(2), 1-24.

Booker, J.F., A.M. Michelsen, and F.A. Ward, 2005. Economic impact of alternative policy responses to prolonged and severe drought in the Rio Grande Basin. *Water Resources Research* 41(2).

Bretherton, C.S., C. Smith, and J.M. Wallace, 1992. An intercomparison of methods for finding coupled patterns in climate data. *Journal of Climate* 5(6): 541-560.

Çimen, M., and O. Kisi, 2009. Comparison of two different data-driven techniques in modeling lake level fluctuations in Turkey. *Journal of hydrology*, 378(3), 253-262.

Christensen, N.S., A.W. Wood, N. Voisin, D.P. Lettenmaier, and R.N. Palmer, 2004. The effects of climate change on the hydrology and water resources of the Colorado River basin. *Climatic change* 62(1-3): 337-363.

Dettinger, M.D., D.R. Cayan, H.F. Diaz, and D.M. Meko, 1998. North-south precipitation patterns in western North America on interannual-to-decadal timescales. *Journal of Climate* 11(12): 3095-3111.

Dibike, Y. B., 2000. Machine learning paradigms for rainfall-runoff modelling. *Hydroinformatics 2000*.

Dibike, Y. B., S. Velickov, D. Solomatine, and M. B. Abbott, 2001. Model induction with support vector machines: introduction and applications. *Journal of Computing in Civil Engineering*, 15(3), 208-216.

Du, K., Y. Zhao, and J. Lei, 2017. The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series. *Journal of Hydrology*, 552, 44-51.

Grantz, K., B. Rajagopalan, M. Clark, and E. Zagona, 2005. A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts. *Water Resources Research* 41(10).

Gupta, H.V., S. Sorooshian, and P.O. Yapo, 1999. Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. *Journal of Hydrologic Engineering* 4(2): 135-143.

Kahya, E., and J. A. Dracup, 1993. US streamflow patterns in relation to the El Niño/Southern Oscillation. *Water Resources Research*, 29(8), 2491-2503.

Kalra, A., L. Li, X. Li, X., and S. Ahmad, 2012. Improving streamflow forecast lead time using oceanic-atmospheric oscillations for Kaidu River Basin, Xinjiang, China. *Journal of Hydrologic Engineering*, *18*(8), 1031-1040.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, L., ... and Y. Zhu, 1996. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American meteorological Society*, 77(3), 437-472.

Khedun, C. P., A.K. Mishra, J.D. Bolten, H.K. Beaudoing, R.A. Kaiser, J.R. Giardino, and V.P. Singh, 2012. Understanding changes in water availability in the Rio Grande/Río Bravo del Norte basin under the influence of large-scale circulation indices using the Noah land surface model. *Journal of Geophysical Research: Atmospheres*, 117(D5).

Kundzewicz, Z.W., L.J. Mata, N.W. Arnell, P. Döll, B. Jimenez, K. Miller, and I. Shiklomanov, 2008. The implications of projected climate change for freshwater resources and their management. *Hydrological Sciences Journal* 53:1, 2-10

Lins, H.F., 2012. USGS hydro-climatic data network 2009 (HCDN-2009) (No. 2012-3047). US Geological Survey.

Liong, S.Y., and C. Sivapragasam, 2002. Flood stage forecasting with support vector machines. *Journal of the American Water Resources Association* (JAWRA) 38(1): 173-186.

Marques, C. A. F., J. A. Ferreira, A. Rocha, J. M. Castanheira, P. Melo-Goncalves, N. Vaz, and J. M. Dias, 2006. Singular spectrum analysis and forecasting of hydrological time series. *Physics and Chemistry of the Earth*, Parts A/B/C, 31(18), 1172-1179.

McCabe, G.J., M.A. Palecki, and J.L. Betancourt, 2004. Pacific and Atlantic Ocean influences on multidecadal drought frequency in the United States. Proceedings of the National Academy of Sciences 101(12): 4136-4141.

Michelsen, A.M., and K. Wood, 2003. Water demand in the Paso del Notre region. Paper presented at Weather and Water on the Border: A Forum on Drought, Paso del Notre Water Task Force. El Paso, Tex.

Middelkoop, H., K. Daamen, D. Gellens, W. Grabs, J.C. Kwadijk, H. Lang, and K. Wilke, 2001. Impact of climate change on hydrological regimes and water resources management in the Rhine basin. *Climatic change* 49(1-2): 105-128.

Moriasi, D.N., J.G. Arnold, M.W. Van Liew, R.L. Bingner, R.D. Harmel, and T.L. Veith, 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE* 50(3): 885-900.

Mukherjee, S., E. Osuna, and F. Girosi, 1997. Nonlinear prediction of chaotic time series using support vector machines. In *Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop* (pp. 511-520).

Munot, A.A., and K.K. Kumar, 2007. Long range prediction of Indian summer monsoon rainfall. *Journal of earth system science*, 116(1), 73-79.

Nash, J.E., and J.V. Sutcliffe, 1970. River flow forecasting through conceptual models part I—A discussion of principles. *Journal of hydrology* 10(3): 282-290.

Nijssen, B., G.M. O'Donnell, A.F. Hamlet, and D.P. Lettenmaier, 2001. Hydrologic sensitivity of global rivers to climate change. *Climatic change* 50(1-2): 143-175.

Nourani, V., M. Komasi, and A. Mano, 2009. A multivariate ANN-wavelet approach for rainfall–runoff modeling. *Water resources management*, 23(14), 2877-2894.

Pahl-Wostl, C., 2007. Transitions towards adaptive management of water facing climate and global change. *Water resources management* 21(1): 49-62.

Pascolini-Campbell, M., R. Seager, A. Pinson, and B.I. Cook, 2017. Covariability of climate and streamflow in the Upper Rio Grande from interannual to interdecadal timescales. *Journal of Hydrology: Regional Studies*, *13*, 58-71.

Pathak, P., A. Kalra, S. Ahmad, and M. Bernardez, 2016. Wavelet-aided analysis to estimate seasonal variability and dominant periodicities in temperature, precipitation, and streamflow in the Midwestern United States. *Water resources management*, 30(13), 4649-4665.

Pathak, P., A. Kalra, K.W. Lamb, W.P. Miller, S. Ahmad, R. Amerineni, and D.P. Ponugoti, 2018. Climatic variability of the Pacific and Atlantic Oceans and western US snowpack. *International Journal of Climatology*, 38(3), 1257-1269.

Pai, P. F. and C.S. Lin, 2005. A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega*, 33(6), 497-505.

Piechota, T. C., F.H. Chiew, J.A. Dracup, and T.A. McMahon, 1998. Seasonal streamflow forecasting in eastern Australia and the El Niño–Southern Oscillation. *Water Resources Research*, 34(11), 3035-3044.

Piechota, T.C., F.H. Chiew, J.A. Dracup, and T.A. McMahon, 2001. Development of exceedance probability streamflow forecast. *Journal of Hydrologic Engineering* 6(1): 20-28.

Potts, J.M., C.K. Folland, I.T. Jolliffe, and D. Sexton, 1996. Revised "LEPS" scores for assessing climate model simulations and long-range forecasts. *Journal of Climate* 9(1): 34-53.

Rajagopalan, B., E. Cook, U. Lall, and B.K. Ray, 2000. Spatiotemporal variability of ENSO and SST teleconnections to summer drought over the United States during the twentieth century. *Journal of Climate* 13(24): 4244-4255.

Redmond, K.T., and R.W. Koch, 1991. Surface climate and streamflow variability in the western United States and their relationship to large-scale circulation indices. *Water Resources Research* 27(9): 2381-2399.

Rice, J.L., C.A. Woodhouse, and J.J. Lukas, 2009. Science and Decision Making: Water Management and Tree-Ring Data in the Western United States 1. *JAWRA Journal of the American Water Resources Association*, 45(5), 1248-1259.

Sagarika, S., A. Kalra, and S. Ahmad, 2015. Interconnections between oceanic–atmospheric indices and variability in the US streamflow. *Journal of Hydrology* 525: 724-736.

Scholkopf, B., K. K. Sung, C. J. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, 1997. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE transactions on Signal Processing*, 45(11), 2758-2765.

Sharma, S., P. Srivastava, X. Fang, X., and L. Kalin, 2015. Long-range hydrologic forecasting in El Niño Southern Oscillation-affected coastal watersheds: Comparison of climate model and weather generator approach. *Journal of Hydrologic Engineering*, 20(12), 06015006.

Silverman, B.W., 1998. Density Estimation for Statistics and Data Analysis. *Chapman and Hall*, New York: Routledge.

Slack, J.R., and J.M. Landwehr, 1992. Hydro-climatic data network (HCDN); a U.S. Geological Survey streamflow data set for the United States for the study of climate variations, 1874-1988. U.S. Geological Survey, Open-File Report 92-129.

Soukup, T.L., O.A. Aziz, G.A. Tootle, T.C. Piechota, and S.S. Wulff, 2009. Long lead-time streamflow forecasting of the North Platte River incorporating oceanic–atmospheric climate variability. *Journal of Hydrology* 368(1): 131-142.

Stewart, I.T., D.R. Cayan, and M.D. Dettinger, 2004. Changes in snowmelt runoff timing in western North America under abusiness as usual'climate change scenario. *Climatic Change* 62(1-3): 217-232.

Tamaddun, K.A., A. Kalra, M. Bernardez, and S. Ahmad, 2017. Multi-Scale Correlation between the Western US Snow Water Equivalent and ENSO/PDO Using Wavelet Analyses. *Water Resources Management* 31(9): 2745-2759.

Thakali, R., A. Kalra, and S. Ahmad, 2016. Understanding the Effects of Climate Change on Urban Stormwater Infrastructures in the Las Vegas Valley. *Hydrology*, 3(4), 34.

Tootle, G.A., and T.C. Piechota, 2006. Relationships between Pacific and Atlantic Ocean sea surface temperatures and US streamflow variability. *Water Resources Research* 42(7).

U.S. Department of Interior, 2003. Water 2025: Preventing crises and conflict. Bur. of Reclam., Washington, D. C.

Vapnik, V., 1995. The nature of statistical learning theory. *Springer*, New York.

Vapnik, V., 1998. Statistical learning theory. *John Wiley*, New York.

Wallace, J.M., and D.S. Gutzler, 1981. Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Monthly Weather Review* 109(4): 784-812.

Wallace, J.M., C. Smith, and C.S. Bretherton, 1992. Singular value decomposition of wintertime sea surface temperature and 500-mb height anomalies. *Journal of climate* 5(6): 561-576.

Ward, M.N., and C.K. Folland, 1991. Prediction of seasonal rainfall in the north nordeste of Brazil using eigenvectors of sea-surface temperature. *International Journal of Climatology* 11(7): 711-743.

Willey, Z and T. Graff, 1984. Water is a commodity, so let's treat it a one. *Los Angeles Times* (February 5): Part IV, 5

Woodruff, S. D., R.J. Slutz, R.L. Jenne, and P.M. Steurer, 1987. A comprehensive ocean-atmosphere data set. *Bulletin of the American meteorological society*, 68(10), 1239-1250.

Zealand, C.M., D.H. Burn, and S.P. Simonovic, 1999. Short-term streamflow forecasting using artificial neural networks. *Journal of hydrology* 214(1): 32-48.

TABLES

Table 1. SVD results for different lead-time cases.

Climate variability	Lead-time	SST		Z_{500}		SH_{500}		U_{500}	
	Months	SCF (%)	NSC (%)	SCF (%)	NSC (%)	SCF (%)	NSC (%)	SCF (%)	NSC (%)
Pacific Ocean	1	97.3	6.9	96.4	4.0	96.3	4.6	95.3	4.6
	4	97.4	5.1	91.4	2.5	95.6	4.3	95.4	3.9
	13	92.4	2.2	89.1	1.6	88.6	2.1	90.5	2.4
Atlantic Ocean	1	96.9	4.5	96.0	2.5	92.5	3.7	90.5	3.0
	4	96.2	4.4	95.4	3.1	95.1	3.8	94.5	3.4
	13	94.6	1.9	92.0	1.5	87.1	2.2	89.5	1.7

Table 2. Best streamflow predictor variables for different lead-time scenarios .

	Best streamflow predictors		
Station	1-month lead-time	4-month lead-time	13-month lead-time
1	Atlantic SST Atlantic SH_{500}	Pacific U_{500} Atlantic U_{500}	Pacific SST Atlantic U_{500}
2	Pacific SST Atlantic Z_{500}	Pacific SST Pacific U_{500}	Pacific SST Pacific U_{500}
3	Atlantic SST Atlantic U_{500}	Pacific SST Pacific U_{500}	Atlantic SST Pacific U_{500}
4	Atlantic SH_{500} Atlantic U_{500}	Pacific Z_{500} Pacific U_{500}	Pacific SH_{500} Pacific U_{500}
5	Atlantic SH_{500} Atlantic U_{500}	Pacific Z_{500} Pacific U_{500}	Pacific SH_{500} Pacific U_{500}
6	Pacific SH_{500} Atlantic SH_{500}	Pacific SST Pacific U_{500}	Pacific SST Pacific SH_{500}

Table 3. SVM model performance for different stations for different lead-times.

Streamflow Station	Lead-time Months	Model performance parameter r	PBIAS (%)	NSE	LEPS SK (%)
1	1	0.87	0.79	0.72	64.1
	4	0.95	-0.53	0.87	78.3
	13	0.85	3.84	0.71	69.8
2	1	0.83	2.02	0.67	63.1
	4	0.93	2.32	0.83	75.0
	13	0.85	4.72	0.69	68.4
3	1	0.91	1.8	0.81	78.9
	4	0.93	3.07	0.86	79.6
	13	0.81	1.26	0.63	61.8
4	1	0.94	-1.48	0.87	75.9
	4	0.87	2.05	0.73	68.7
	13	0.86	2.98	0.72	67.9
5	1	0.92	-2.72	0.84	75.8
	4	0.91	0.84	0.80	76.0
	13	0.94	-0.36	0.86	73.6
6	1	0.90	3.87	0.79	75.6
	4	0.80	5.54	0.60	67.3
	13	0.92	4.11	0.84	77.4

LIST OF FIGURES

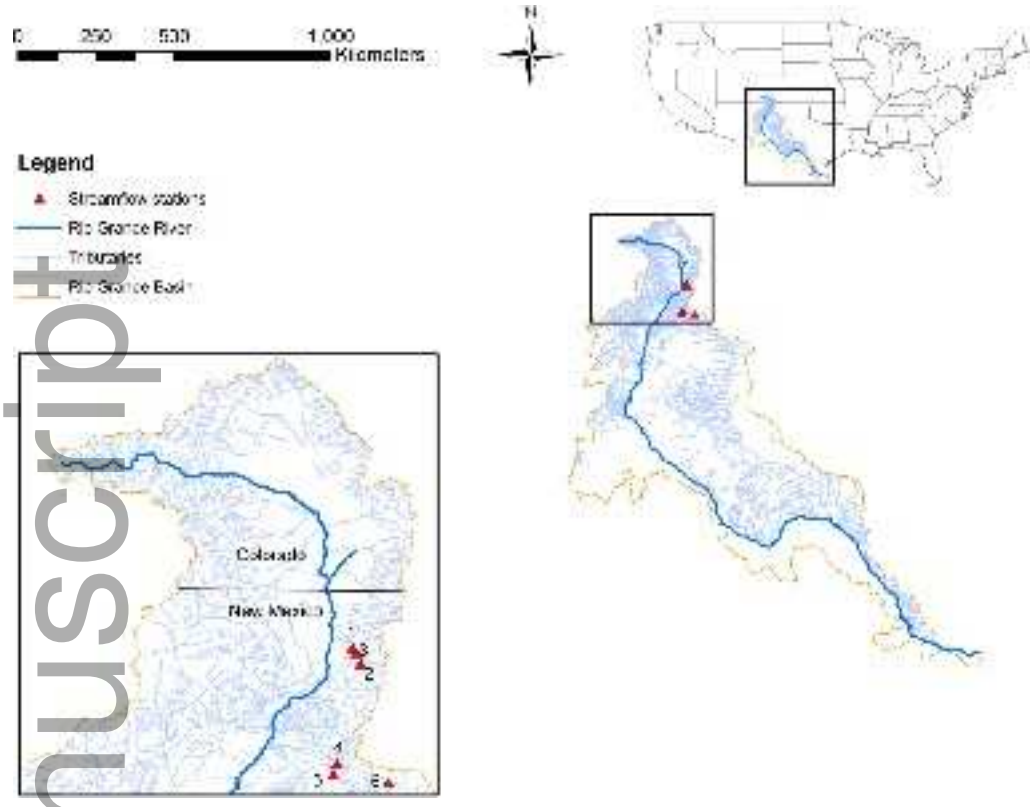
Figure 5. Heterogeneous correlation map for Atlantic Ocean (a) 1-month lead-time (b) 4-month lead-time (c) 13-month lead-time SST, Z_{500}, SH_{500}, and U_{500} with April-August streamflow.

Figure 6. Map showing examples of the poor, fair, and good continuous exceedance probability forecast for (a) 1-month lead-time (b) 4-month lead-time (c) 13-month lead-time

Figure 7. (a) Time series plot where dotted line represents measured streamflow and solid line represents the predicted streamflow (b) Scatter plot (c) Box plot (d) Non-exceedance probability plot for 1-month lead-time depicting the comparison between measured and forecasted streamflow for six streamflow stations.

Figure 8. (a) Time series plot where dotted line represents measured streamflow and solid line represents the predicted streamflow (b) Scatter plot (c) Box plot (d) Non-exceedance probability plot for 4-month lead-time depicting the comparison between measured and forecasted streamflow for six streamflow stations.

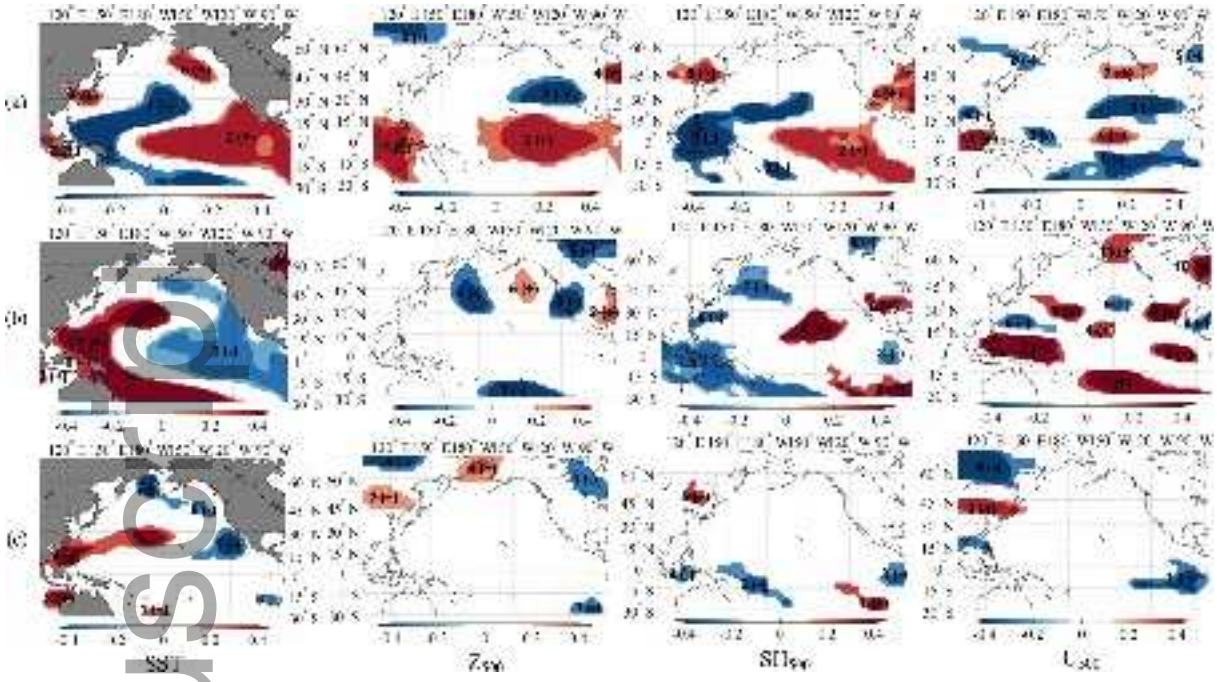
Figure 9. (a) Time series plot where dotted line represents measured streamflow and solid line represents the predicted streamflow (b) Scatter plot (c) Box plot (d) Non-exceedance probability plot for 13-month lead-time depicting the comparison between measured and forecasted streamflow for six streamflow stations



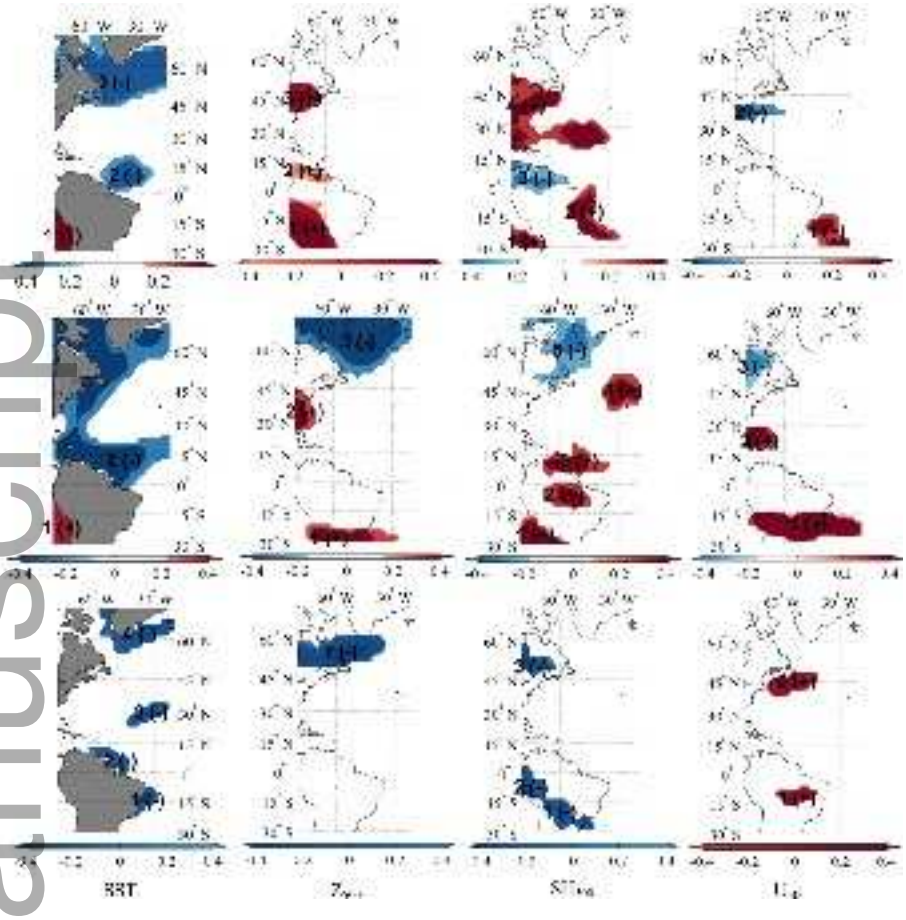
jawra_12733-18-0057_f1.tif

jawra_12733-18-0057_f2.tif

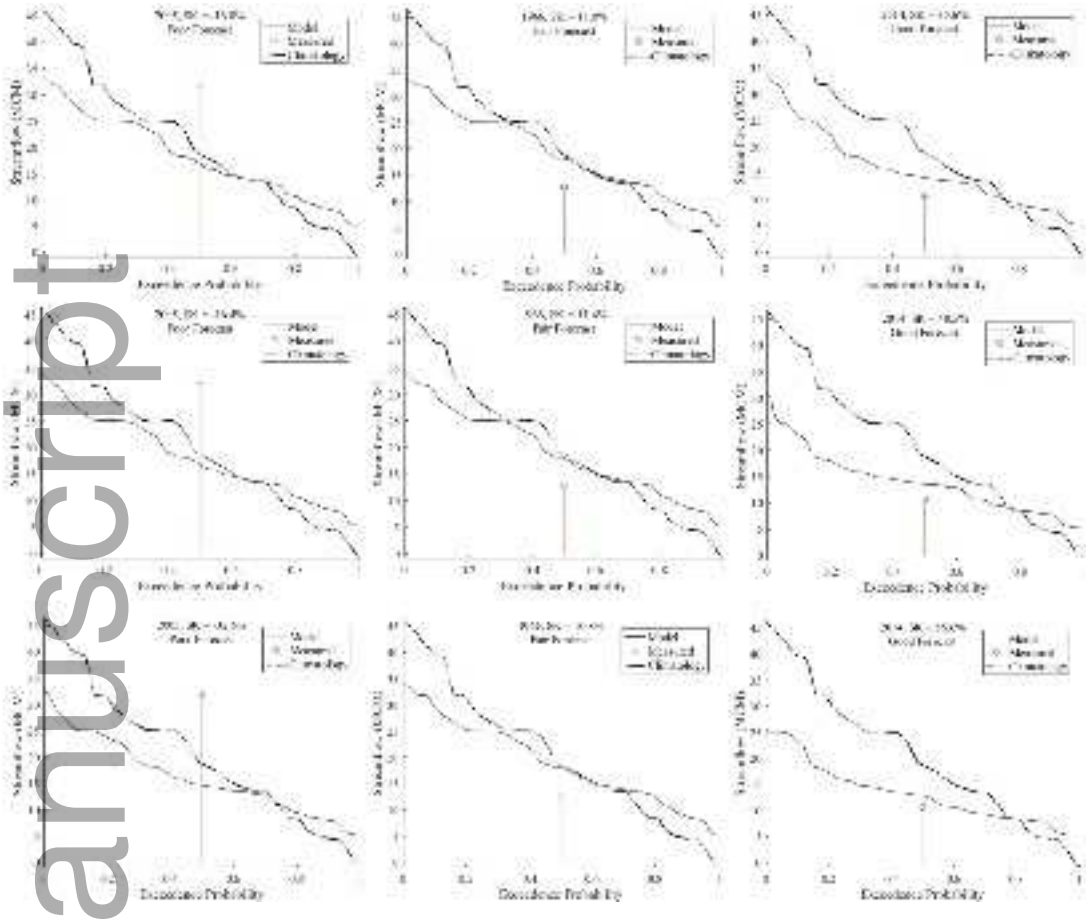
jawra_12733-18-0057_f3.tif



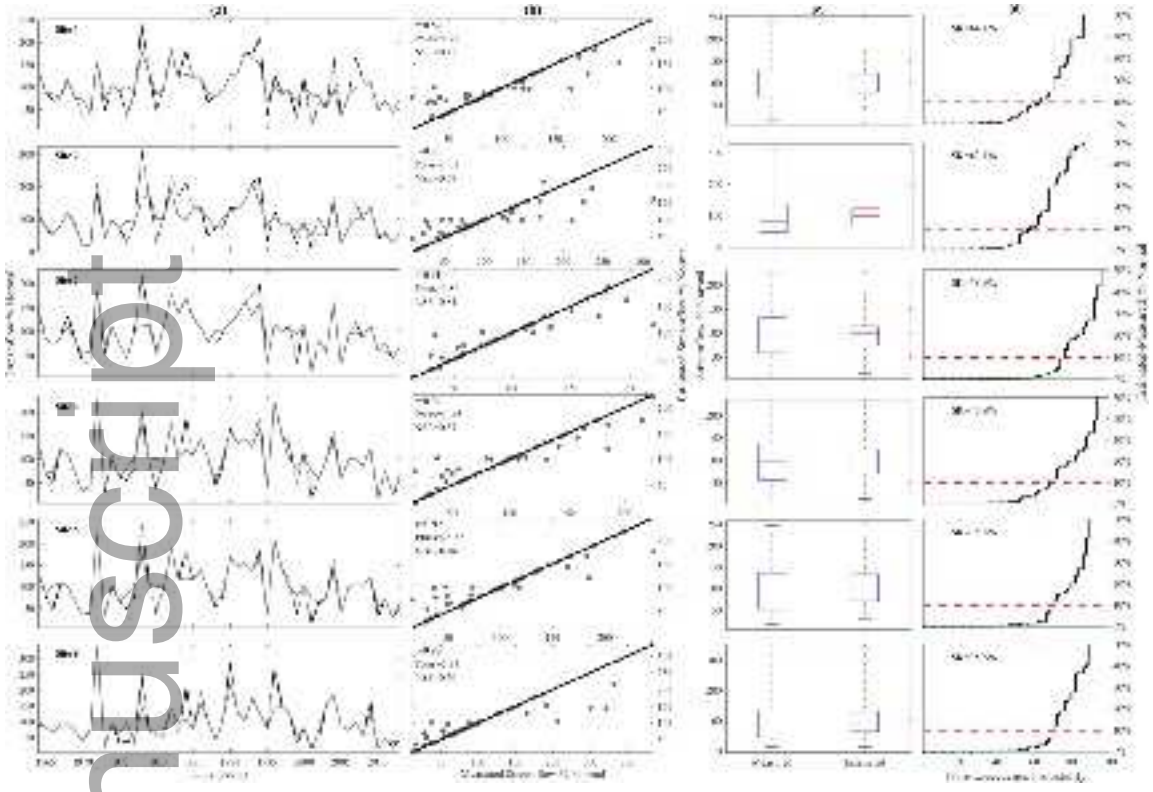
jawra_12733-18-0057_f4.tif



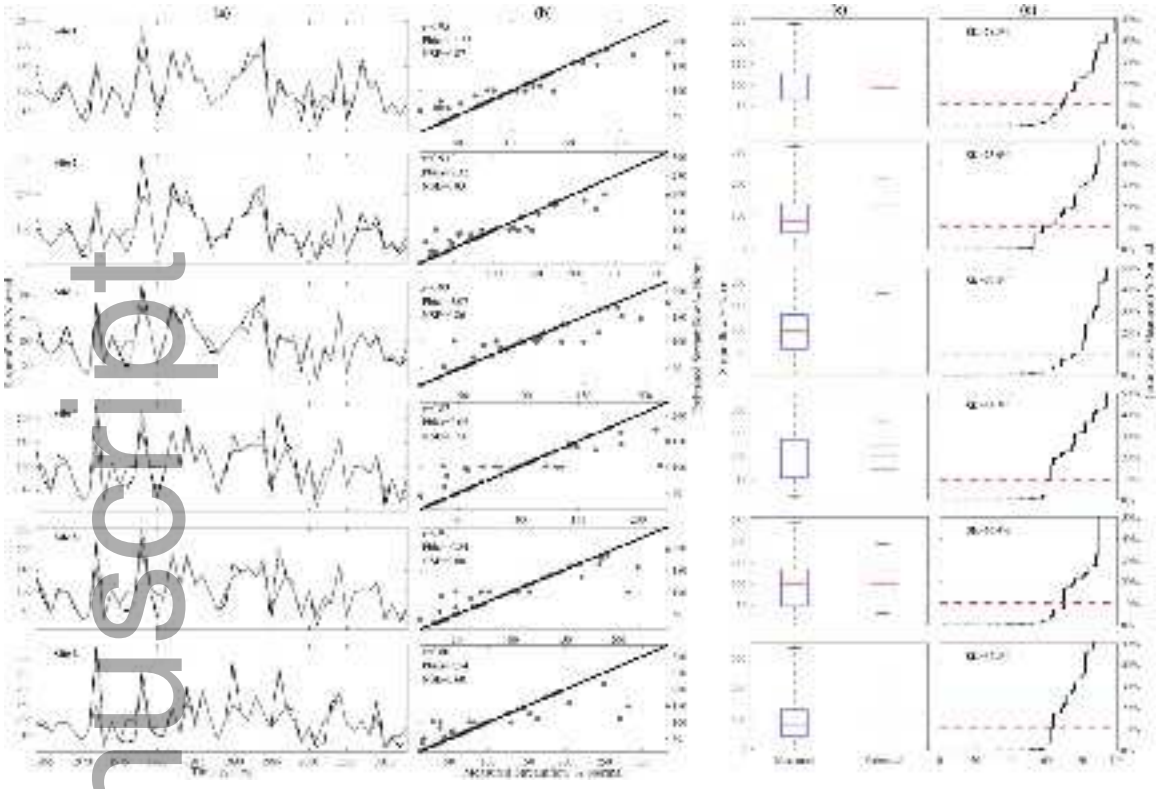
jawra_12733-18-0057_f5.tif



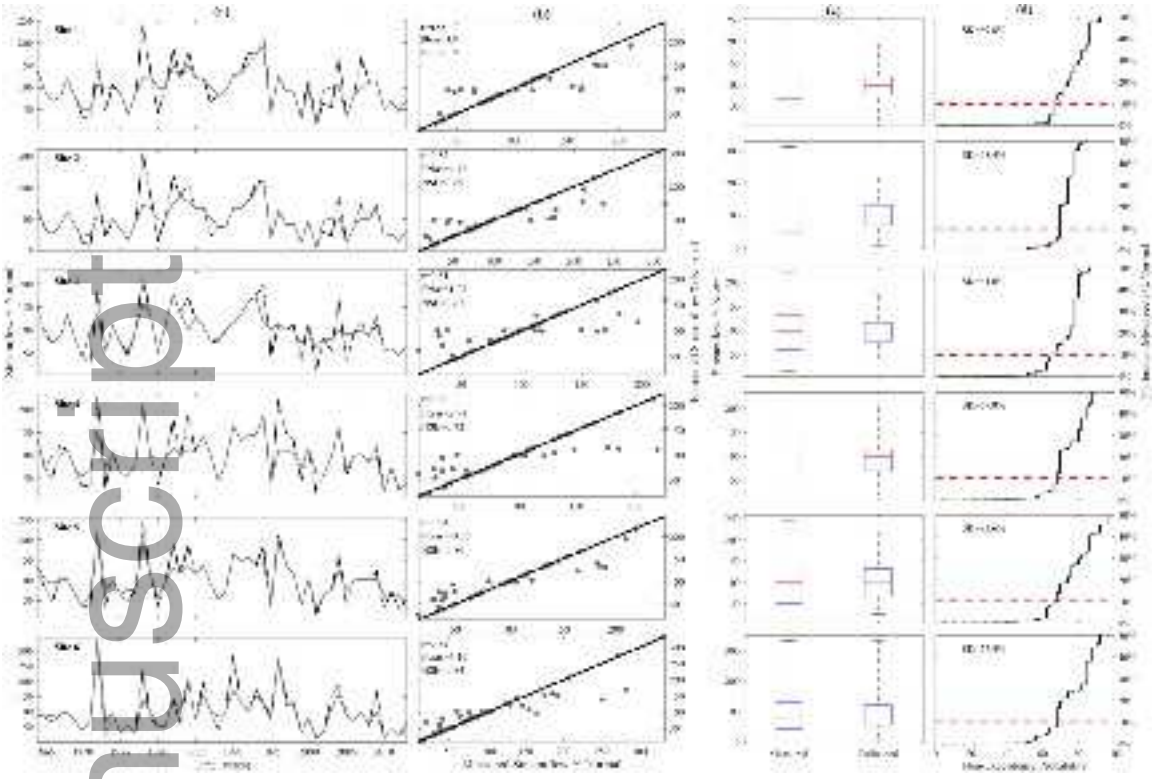
jawra_12733-18-0057_f6.tif



jawra_12733-18-0057_f7.tif



jawra_12733-18-0057_f8.tif



jawra_12733-18-0057_f9.tif