Journal of
Water & Climate Change

IWA PUBLISHING

Journal of Water and Climate Change Vol 14 No 6, 1868 doi: 10.2166/wcc.2023.479

Seasonal precipitation forecasting for water management in the Kosi Basin, India using large-scale climate predictors

Manjeet Singh Dhillon [a,*], Mohammed Sharif[b], Henrik Madsen[c] and Flemming Jakobsen[d]

[a] Ganga Flood Control Commission, 3rd Floor, Sinchai Bhavan, Old Secretariat, Rajbansi Nagar, Patna, Bihar 800013, India
[b] Department of Civil Engineering, Faculty of Engineering and Technology, Jamia Millia Islamia Central University, NewDelhi, India
[c] Emerging Technologies, DHI, Agern Alle 5, Hørsholm DK-2970, Denmark
[d] COWI, Parallelvej 2, Lyngby 2800, Denmark
*Corresponding author. E-mail: msdcwc62@gmail.com

MSD, 0000-0003-4361-5796

ABSTRACT

A novel approach for qualitative seasonal forecast of precipitation at a basin scale is presented as significant enhancement in seasonal forecast at regional and country scales in India. The process utilizes empirical and typically lagged relationships between target variables of interest, namely precipitation at the basin level and various large-scale climate predictors (LSCPs). A total of 14 LSCPs have been considered for the seasonal forecast of precipitation with lead times of 1, 2, and 3 months in the Kosi Basin, India. Random split training and testing were conducted on seven machine-learning (ML) models using a potential predictor dataset for model selection. The Logistic Regression (LR) model was adopted since it had the highest mean accuracy score compared to the remaining six ML models. The LR model has been optimized by testing it on all possible combinations of potential predictors using Leave-One-Out Cross-Validation (CV) scheme. The resulting Seasonal Prediction Model (SPM) provides the probability of each tercile categorized as Above Normal (AN), Normal (N), and Below Normal (BN). The model has been evaluated using various metrics.

Key words: Kosi Basin, large-scale climate predictors, logistic regression, machine-learning, seasonal forecasting, seasonal forecast model

HIGHLIGHTS

- A basin-scale approach is presented instead of a larger country scale.
- Use of large number of large-scale climate predictors for the development of an ML-based categorical forecast model.
- The methodology is generic in nature and can be applied to any other basins.
- Directions for further research are suggested for the generation of weather ensembles, automatic climate predictors, and for model operationalization.

1. INTRODUCTION

The availability and demand of water vary spatially and temporally making water management a daunting task for the decision-makers. Reservoir operations, water allocation strategies, and drought and flood management require seasonal forecasting of precipitation at the basin scale. There is a wide range of methodologies that have been developed for seasonal forecasting that vary from pure dynamical modelling to pure statistical modelling or a combination of both. Statistical models exploit empirical relationships between a target variable of interest and one or more predictor variables. Such models are developed from historical data, and their performance depends upon the quality of the historical oceanic, atmospheric, and hydro-meteorological data (Anderson *et al.* 1999; Lavers 2011). Statistical models are less costly to develop and run compared with the dynamical models (Barnston *et al.* 1994; Bierkens & Van Beek 2009). In addition, statistical models are relatively flexible in model construction and have the potential to improve the forecast lead time and predictability depending on the accumulation of observations and related data, although their predictability is unstable compared to dynamic models.

In the present work, several Large-Scale Climate Predictors (LSCPs) have been investigated for seasonal forecasting of precipitation over the Kosi Basin, India. A number of studies in the past have emphasized the importance of understanding the influence of large-scale climatic patterns on precipitation for improving forecast accuracy (Feng *et al.* 2020), and therefore many studies have analyzed the relationship between precipitation and climatic patterns in India. This research has shown that the relevant patterns are the Indian Ocean Dipole (IOD) (Behera *et al.* 1999; Krishnan & Sugi 2003; Goswami *et al.* 2006), the El Niño–Southern Oscillation (ENSO) (Kumar *et al.* 2006; Mokhov *et al.* 2012; Feliks *et al.* 2013), the North Atlantic Oscillation (NAO) (Bharath & Srinivas 2015), the Pacific Decadal Oscillation (PDO) (Dong 2016), and the Atlantic Multidecadal Oscillation (AMO) (Krishnamurthy & Krishnamurthy 2016). These large-scale indices have been used to forecast seasonal precipitation and stream flows (Lavers 2011; Rasouli *et al.* 2012; Arnal *et al.* 2017; Apel *et al.* 2018).

Previous studies of Indian South West Monsoon Rainfall (ISMR) forecasting (Rajeevan *et al.* 2007; Kumar *et al.* 2012) have demonstrated the use of multiple LSCPs. Substantial variation across India and across timescales has been found (Kurths *et al.* 2019). In particular, ENSO and IOD influence precipitation in the southeast at inter annual and decadal scales, respectively. The NAO has a strong connection to precipitation, particularly in the northern regions. The effect of the PDO stretches across the whole country, whereas AMO influences precipitation, particularly in the central arid and semi-arid regions.

Over the years, linkages between climatic patterns and precipitation have been investigated by using a range of statistical methods, such as correlation (Abid *et al.* 2018), principal component analysis, and empirical orthogonal functions (Hannachi *et al.* 2007) among others. Linear regression, auto-regressive moving average (ARMA), auto-regressive integrated moving average (ARIMA), and multiple linear regression (MLR), for example, are the other most commonly implemented statistical techniques (Eldaw *et al.* 2003; Archer & Fowler 2008; Barlow & Tippett 2008; Gámiz-Fortis *et al.* 2010; Kirono Dewi *et al.* 2010; Purdie & Bardsley 2010). Kashid & Maity (2012) applied genetic programming (GP) to forecast ISMR based on LSCPs. The GP approach was found to adequately capture the complex relationship between the monthly ISMR and LSCPs.

During the last decade, machine-learning (ML) algorithms have received wide attention in both classification and regression tasks (Colomo *et al.* 2019). ML algorithms are capable of investigating hierarchical and non-linear relationships between the response variable and predictor variables, based on ensemble learning approaches (Shalev-Shwartz & Ben-David 2014). An artificial neural network (ANN)-based model was used to forecast summer rainfall in the Yangtze River basin, using LSCPs including SOI and the Scandinavia Pattern. Vathsala & Koolagudi (2017) presented an algorithm by integrating data mining and statistical techniques. The proposed technique predicted the rainfall in five different categories such as flood, excess, normal, deficit, and drought. Mishra *et al.* (2018) presented an ANN model for the forecast of rainfall with lead times of 1 and 2 months. The efficiency of the model was demonstrated through application to a dataset from multiple stations in north India. The performance of the ANN model was evaluated by using regression analysis, mean square error, and magnitude of relative error.

A study by Praveen *et al.* (2020) applied Artificial Neural Network-Multilayer Perceptron (ANN-MLP) to analyze and forecast the long-term spatio-temporal changes in rainfall using the data from 1901 to 2015 across India at the meteorological divisional level. The results of the analysis showed that the rainfall for the next 15 years exhibited a significant decline. Saha *et al.* (2021) employed a feature reduction approach based on non-linear deep learning to identify effective predictors for monsoon rainfall using climatic variables from different regions worldwide. The study found that certain predictors, such as sea surface temperature (SST) and zonal wind (ZW), were capable of forecasting the Indian summer monsoon 1 month in advance, while sea level pressure (SLP) could predict the season 10 months in advance. Additionally, the authors demonstrated that combining multiple climatic variables to derive predictors resulted in superior performance compared to using predictors derived from individual variables. A Gaussian Process Regression (GPR) approach, one of the ML methods, was employed by Subrahmanyam *et al.* (2021) on long time-series rainfall data for the determination of heavy and light rainfall days. Sharma & Goyal (2015) reported the application of the Bayesian network model for the forecast of rainfall at 21 stations in Assam, India. The efficiency of the forecast was found to be above 85% for most of the cases.

Although a few studies have been carried out at the national level, no significant research to demonstrate seasonal forecast skills of precipitation at the basin scale in India has been reported in the literature. Currently, the forecast of ISMR is made available before every monsoon. However, such forecasts have limited use due to significant variations in monsoonal rainfall over the country. The objective of the present research is to develop and evaluate a seasonal forecast model for forecasting ISMR at a basin scale using a ML model. The present research is novel in many ways: (1) a basin-scale approach is presented as the focus is on water management at the basin scale, (2) large numbers of LSCPs are used and analyzed for developing robust ML-based forecast models, and (3) the results are indicative of teleconnections between various LSCPs and seasonal

precipitation in the Kosi Basin. A distinct practical advantage of the methodology employed herein is that it is fairly generic in nature and can be applied to other basins with minimal changes to the model developed herein.

2. KOSI BASIN

The Kosi Basin as shown in Figure 1 is a representative river basin in the middle of the Himalayan range of mountains. The Kosi River is a trans-boundary river across China, Nepal, and India, and is also an important tributary of the River Ganges. Kosi Basin lies in 25° 18′–29° 09′ N, 85° 1′–85° 57′ E. The basin covers an area of approximately 88,000 km^2: 45% in Nepal, 32% in China, and 23% in India. The upstream basin is divided into three tributaries: the North branch Arun, which is the main tributary, the West branch Sun Kosi, and the East branch Tamor. The three come together to form the Kosi River, near Chatra Gorge, and finally flows into the Ganges River. Two more tributaries, Bagmati and Kamla Balan join Kosi in the Indian region. The part of the Kosi River valley in India is an alluvial plain, an important part of the Ganges Plain. Since 80% of rainfall in the river basin fall during the months of June, July, and August, the area is badly affected by frequent floods.

The part in Nepal has a great elevation drop from 8,848 m (Mt Everest) above mean sea level to 60 m above mean sea level. Precipitation is unevenly distributed throughout the basin. The annual precipitation is about 300–400 mm in the northern Himalayan region, 1,000–1,500 mm in the subtropical and tropical region, and 1,500–2,500 mm in the temperate region. Owing to such unique physiographical and topographical distribution, the basin features a variety of climates that range from the tropical savannah over the southern plains to the polar frost in the northern mountains.

The Kosi River Basin also has a severe drought problem characterized by vast affected areas. The drought mainly affects the middle Kosi River Basin, the Tibetan Himalayas, and the downstream plain. The Kosi Basin has several other water

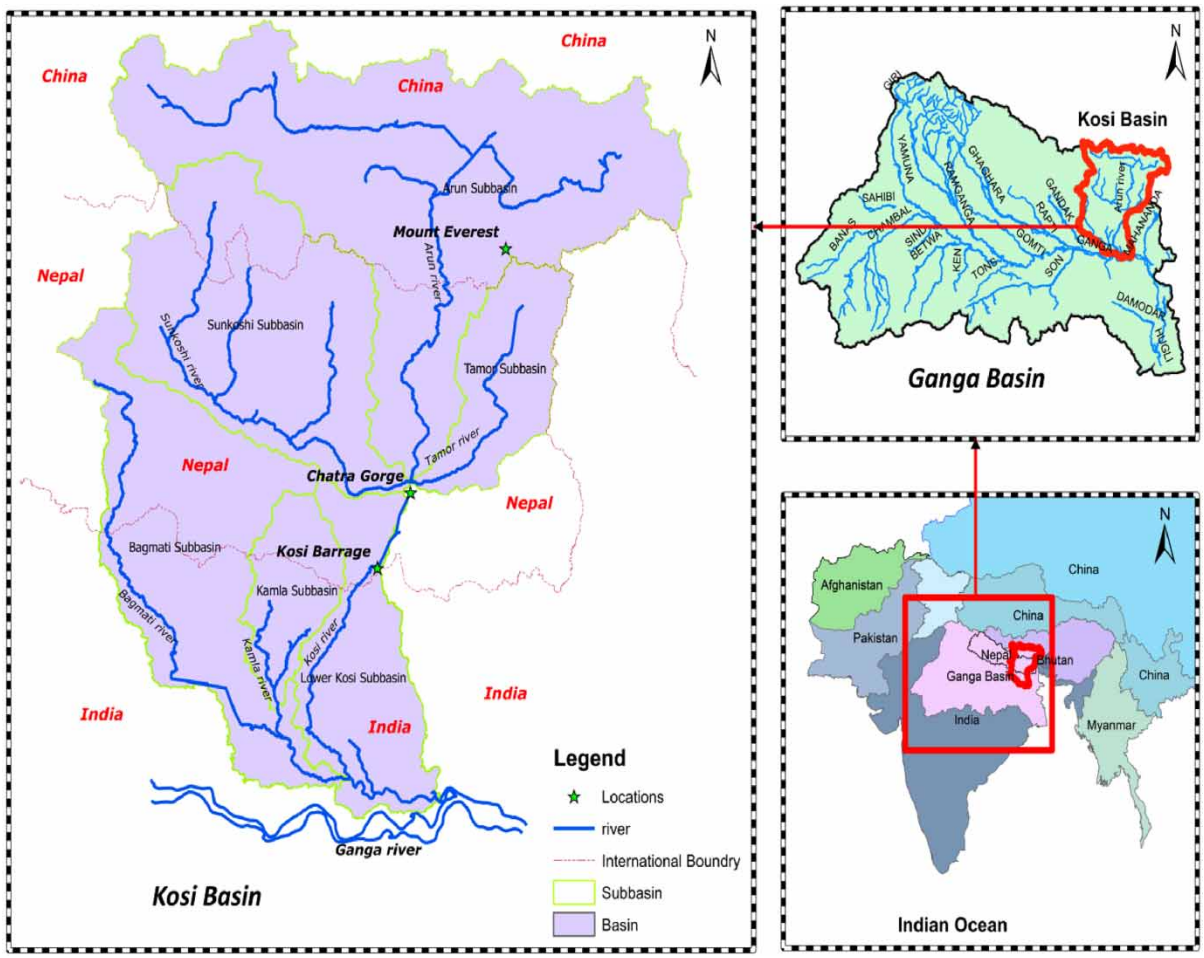

Figure 1 | Schematic diagram of the Kosi Basin.

management issues, which include flood control, irrigation, hydroelectric generation, embankment management, and rise in riverbed level due to heavy siltation and barrage operation. The water resources of the Kosi Basin are largely untapped. There are 11 proposed development projects for hydroelectric generation and water storage (Chinnasamy *et al.* 2015). Furthermore, Nepal receives 225 km^3 of surface water annually from the basin. However, less than 7% of water is utilized. Water is in surplus during monsoon months in Nepal's Tarai region and India's Bihar region, whereas there is shortage of water in pre- and post-monsoon seasons (Bharati *et al.* 2014). Seasonal precipitation forecasting for water management in the Kosi Basin will, therefore, facilitate better water management in both India and Nepal.

3. DATA

In total, 16 LSCPs relevant to ISMR have been adopted for the seasonal forecast of precipitation for the Kosi Basin. These predictors along with their spatial locations are shown in Table 1.

The LSCPs described in Table 1 have been adopted by the National Oceanic and Atmospheric Administration (NOAA) for the period from 1991 to 2020. In addition, accumulated non-monsoon precipitation (ANMP) from October to May at the basin level has also been included as a predictor to represent the local conditions before the onset of the monsoon. Due to the non-availability of actual precipitation observations in more than two-third of the basin, which lies in Nepal and China, satellite-based precipitation estimates from global precipitation measurement (GPM), available from 2000 onwards have been used. Two temporal products of GPM were used on a monthly and on a 30-min scale. Both products were obtained from the Google Earth Engine repository. The 30-min GPM data were resampled to daily and monthly scales. The monthly GPM product was compared month-wise with monthly resampled GPM data to determine the correction factors for every month. The factor for each month was then uniformly applied to all the daily resampled GPM values in that month.

4. METHODS

The analysis of precipitation data over the basin has been carried out to develop a statistical model for the seasonal forecast of precipitation. Monthly anomalies of LSCPs, which are perceived to impact the region of interest, have been derived. Subsequently, the linear relationship of monthly anomalies of predictors with the aggregated precipitation over the monsoon season for which the forecast is to be made has been ascertained using the Pearson correlation coefficient. The significant predictors have been adopted in lagging months in the non-monsoon season (October to May) for the forecast of monsoonal (June–September) rainfall with a lead time of 1 month (forecast by the end of May), 2 months (forecast by end April), and 3 months (forecast by end March).

The aggregated precipitation data over the monsoon season have been categorized into terciles, that is, 'Above Normal' (AN), 'Normal' (N), and 'Below Normal''(BN). Seven ML classification models have been fitted on the dataset. Based on the outcome of a random split train/test analysis, the most optimal ML model has been selected. Optimization of predictor combinations based on cross-validation (CV) and tuning of hyperparameters based on random search optimization technique has been carried out, so as to further improve the model performance. The performance of the model has been evaluated based on several metrics. The complete methodology is shown in Figure 2.

4.1. Precipitation data

The precipitation data over the basin were analyzed for the monsoon season (June–September) and the non-monsoon season (October–May). GPM rainfall data were processed for determining monsoon and pre-monsoon rainfall at the basin level and for the six sub-basins for the period 2000–2020. In general, the quantity of precipitation in the monsoon period is four times and the variability is three times that during the non-monsoon season. The length of the non-monsoon period is twice that of the monsoon period, but the precipitation in the non-monsoon season is around 22% of the annual precipitation. This indicates that seasonal forecast of monsoon precipitation is critical for long-term water management in the basin.

4.2 Monthly LSCP anomalies

The monthly anomalies of four of the LSCPs, that is, AMO, North Atlantic Oscillation (NAO), PDO, and South Indian Ocean SST Index (IOD) have been adopted directly as given by NOAA. For the remaining LSCPs representing SST, SLP, and ZW at 850 hpa monthly gridded data are provided by NOAA. Using the monthly gridded data provided by NOAA, time-series data at the defined spatial domain of each climate predictor were prepared. The anomalies of the predictors are assessed on a

Table 1 | Predictor and target variable along with spatial domain and its data source

S. No.	Predictor/target variable	Spatial domain	Data source
1	North Atlantic Sea Surface Temperature (NASST)	20N–30N, 100W–80W	NOAA Optimum Interpolation (OI) Sea Surface Temperature (SST)V2 https://psl.noaa.gov/data/gridded/data.noaa.oisst.v2.html
2	Equatorial Southeast Indian Ocean Sea Surface Temperature (ESEIOSST)	20S–10S, 100E–120E	
3	Arabian Sea Surface Temperature (ASST)	5° N–20° N, 50° E–80° E	
4	NINO3.4 Sea Surface Temperature (NINO3.4)	5S–5N, 170W–120W	https://www.cpc.ncep.noaa.gov/data/indices/sstoi.indices
5	NINO3 Sea Surface Temperature (NINO3)	5N–5S, 150W–90W	
6	NINO4 Sea Surface Temperature (NINO4)	5N–5S, 160E–150W	
7	NINO1 + 2 Sea Surface Temperature (NINO1 + 2)	0–10S, 90W–80W	
8	East Asia Sea Level Pressure (EASLP)	35N–45N, 120E–130E	*NCEP Reanalysis Derived data* (https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.derived.surface.html
9	Northwest Europe Sea Level Pressure (NWESLP)	65N–75N, 20E–40E	
10	North Atlantic Sea Level Pressure (NASLP)	35N–45N, 30W–10W	
11	North Central Pacific Zonal Wind (NCPZW)	5N–15N, 180E–150W	
12	Warm Water Volume (WWV)	5S–5N, 120E–80W	https://www.pmel.noaa.gov/tao/wwv/data/wwv.dat
13	Precipitation including Accumulated Non-Monsoon precipitation (ANMP)	Monthly and Daily in Basin/Sub-Basin	https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGM_06/summary
14	South Indian Ocean SST Index (IOD)		https://psl.noaa.gov/gcos_wgsp/Timeseries/Data/dmi.had.long.data
15	North Atlantic Oscillation (NAO)		https://psl.noaa.gov/data/correlation/nao.data
16	Pacific Decadal Oscillation (PDO)		https://psl.noaa.gov/data/correlation/pdo.data
17	Atlantic Multidecadal Oscillation (AMO)		https://psl.noaa.gov/data/correlation/amon.us.data

monthly time scale using Equation (1):

$$a(t) = \frac{(x(t) - mean(x))}{std(x)} \tag{1}$$

where $a(t)$ is the anomaly of the predictor, $x(t)$ is the monthly value of the predictor and mean(x) and std(x) are the mean and standard deviation in the respective month of the time series of predictors. The base period for the calculation of monthly anomalies of predictors has been taken as 1991–2020.

4.3. Correlation analysis

The prevailing conditions of climate predictors in the non-monsoon period (October–May) are considered as the key indicators for early forecast of seasonal precipitation over the basin during the monsoon period (June–September). To assess the strength of the relationship between each predictor belonging to non-monsoon months with the aggregated seasonal precipitation of the monsoon season (June–September), the Pearson correlation coefficient for all predictors comprising monthly anomalies, means of monthly anomalies over 2 consecutive months and means of monthly anomalies over 3 consecutive months were calculated using the respective time series from 2000 to 2020.

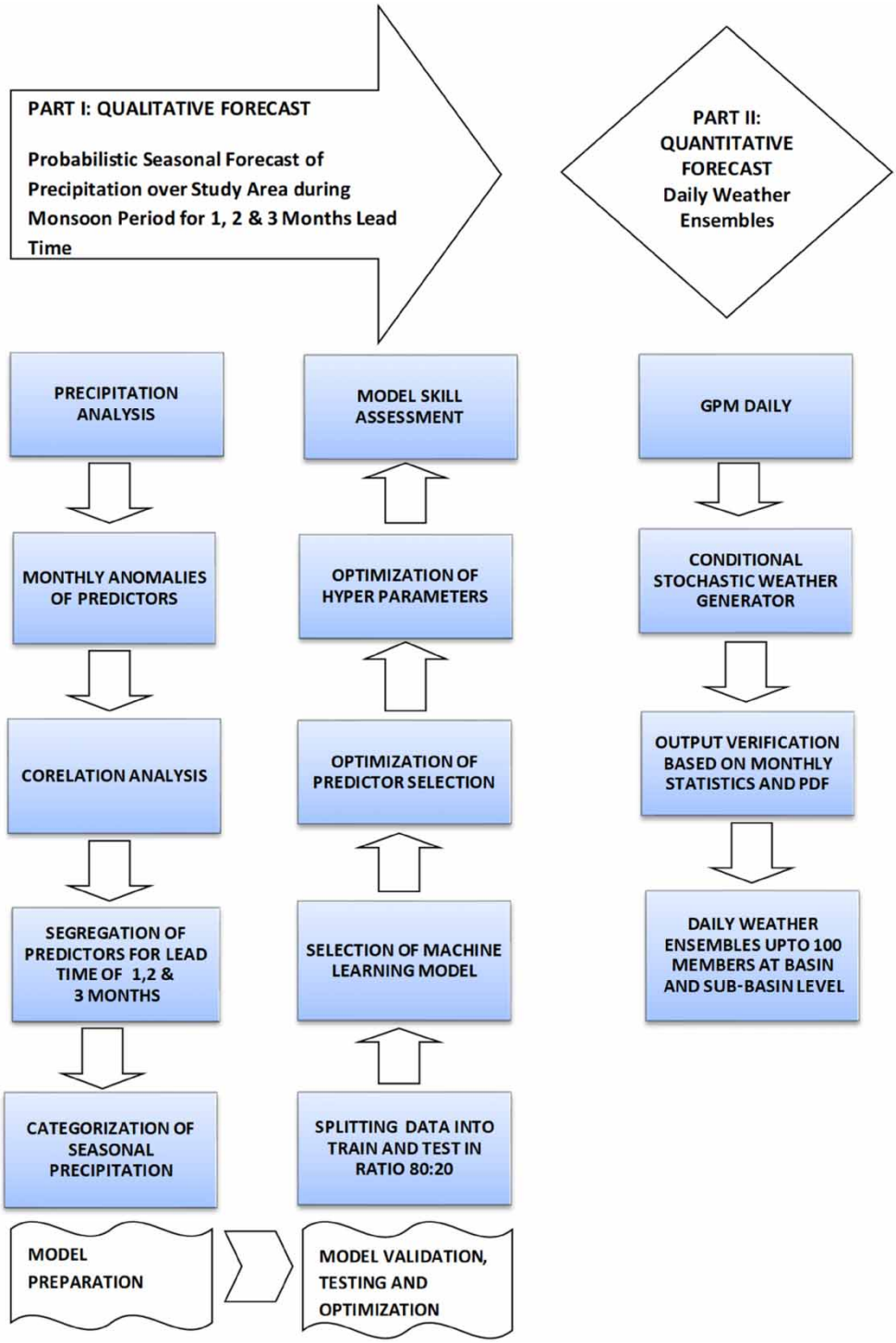

Figure 2 | Flow diagram of the methodology.

The associated significance level (p-value) between the variables used in the correlation analysis was assessed using scipy.stats.pearsonr function in python. A threshold significance level of 0.15 has been adopted in view of the conditional relationship between predictors, limited availability of data, and relatively small size of the basin. The predictors having p-values lower than 0.15 has been identified as statistically significant LSCPs.

4.4. Classification-based ML modelling

The precipitation totals in the monsoon season for the period from 2000 to 2020 were categorized into three equal groups as AN, N, and BN based on quantile analysis. Seven ML-based classification models that are logistic regression (LR), naïve Bayes (NB), KNN, decision tree (DT), random forest (RF), gradient boosting (GB), and support vector machine (SVM) were selected initially based on their ability to classify either using linear or non-linear techniques including ensemble approach. The scikit-learn package in python has been used to carry out the analysis for the above seven ML classification algorithms (Pedregosa *et al.* 2011).

The dataset containing the potential predictors for a lead time of 1 month was split randomly in a ratio of 80:20 to obtain training and test datasets. The training set was fitted to each ML model and the test dataset was used to evaluate each model's performance based on the accuracy score as defined in Equation (2). This process was repeated 2,000 times, which is about 10% of the total split sets possible. The model with the highest mean test score based on accuracy was selected as the most optimal for further analysis. The same ML model was adopted for lead time of 2 and 3 months as many predictors for lead time of 1 month are common with those for lead time of 2 and 3 months. The ML model with the highest accuracy score for 1 month lead time was then run on all possible combinations of potential predictors for lead time of 1, 2 and 3 months using a Leave-One-Out CV scheme on the complete dataset of 21 years. The number of combinations run for lead time of 1 month was 32,767 and for lead time of 2 and 3 months, it was 1,023. The combination for which the CV score based on mean accuracy was highest was adopted as the most optimal combination of predictors for each lead time.

The choice of hyperparameters for the ML model greatly affects the separability of the classes and performance of the algorithm. The hyperparameters of the ML model with highest accuracy score were fine tuned using the random search optimization technique (Pedregosa *et al.*) The mean accuracy score using a Leave-One-Out CV scheme was evaluated over 2,000 iterations using different sets of hyperparameters. The hyperparameter set which provided the highest mean accuracy score was adopted for each lead time.

4.5. Model skill assessment

The performance of the seasonal forecast models was evaluated by calculating the Accuracy, Precision, Recall, F1 score, Multiclass Brier Score (MBS), and Brier Skill Score (BSS) for the period from 2000 to 2020. Accuracy is defined as the fraction of forecasts that the model got right. Precision is the proportion of positive identifications, which was actually correct, whereas Recall is the proportion of actual positives, which was identified correctly. The F1score is the Harmonic mean of the Precision and Recall. The value of 1 for Precision, Recall, Accuracy and F1 score indicates the best possible fit, while 0 indicates poor fit. The scores are defined in Equation (2):

$$\begin{aligned}
\text{Precision(P)} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\
\text{Recall(R)} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \\
\text{Accuracy(A)} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \\
\text{F1 Score} &= \frac{2*\text{R}*\text{P}}{\text{R}+\text{P}}
\end{aligned} \tag{2}$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, and FN is the number of false negatives.

The MBS is a metric that measures the https://www.statisticshowto.com/accuracy-and-precision/ accuracy of a probabilistic forecast. The best possible MBS is 0 for total accuracy and the lowest possible score is 1, which means the forecast was completely inaccurate. Smaller scores that are closer to zero indicate better forecasts. Scores in the middle can be hard to interpret as 'good' or 'bad', so these are sometimes converted to BSS. MBS can indicate how accurate a forecast was, but it does not indicate how accurate it is compared to other forecasts. BSS relates MBS to a benchmark forecast based on climatology. It measures the relative skill of a probabilistic forecast over that of climatology.

Table 2 | Segregation of predictors for lead time of 1, 2, and 3 months

Lead time		
1-month (May)	**2-month (April)**	**3-month (March)**
NINO3.4: Monthly Anomaly in May	NCPZW: Mean of monthly anomaly over three consecutive months from Feb to April	NCPZW: Mean of monthly anomaly over two consecutive months from Feb to March
NCPZW: Mean of monthly anomaly over three consecutive months from April to May	WWV-Monthly Anomaly in April	WWV-Monthly Anomaly in March
WWV-Monthly Anomaly in May		
EASLP – Monthly Anomaly in May		
NINO1 + 2: Monthly Anomaly in May		
NINO3: Monthly Anomaly in May		
NINO4: Monthly Anomaly in May		
NWESLP : Mean of monthly anomaly over 3 consecutive months from Jan to March		
NASST-Monthly Anomaly in Feb		
NASLP:Monthly anomaly in Dec (Previous year)		
NAO-Monthly Anomaly in Jan		
AMO- Mean of monthly anomaly over 2 consecutive months from Feb to March		
ASST- Mean of monthly anomaly over 3 consecutive months from Nov (Previous year) to Jan		
IOD- Monthly Anomaly in March		
ANMP from Oct (Previous year) to Feb		

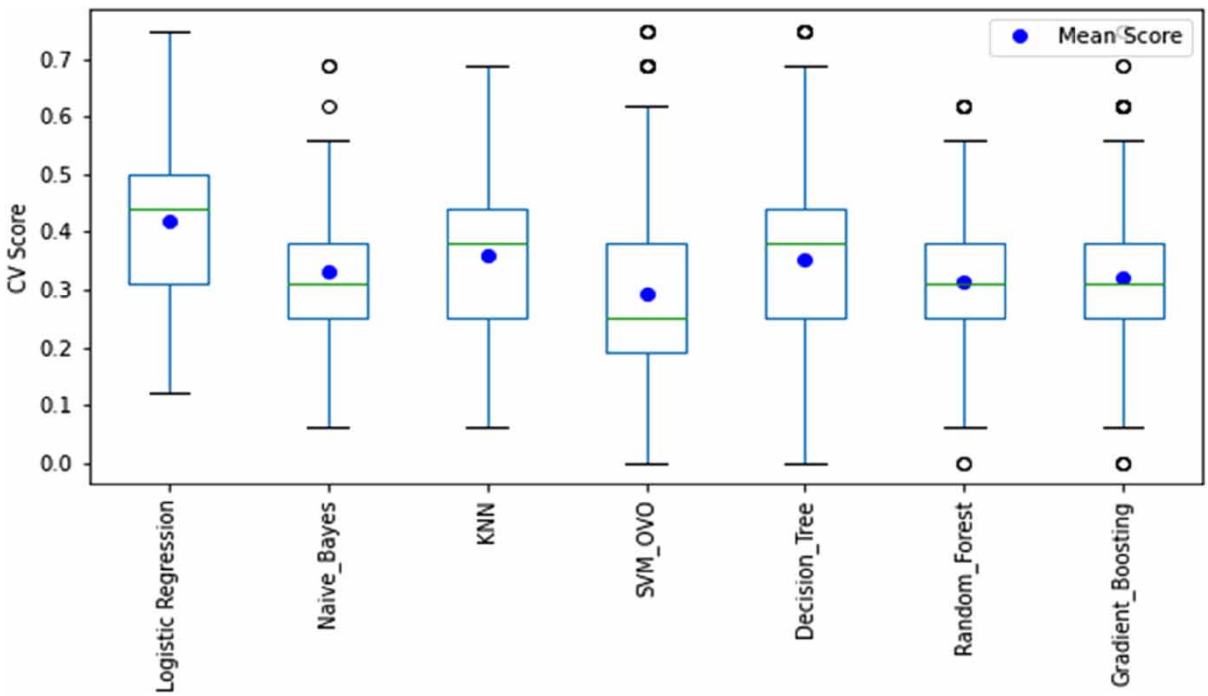

Figure 3 | Performance of ML models based on random split train/test analysis.

Table 3 | Probabilistic model output of SPM vis-a-vis actual category

Correct forecast False forecast Highest AN, N and BN years

Year	Probabilistic Seasonal Forecast from Model						Actual Category
	Lead Time of 1 Month			Lead Time of 2 and 3 Months			
	BN	N	AN	BN	N	AN	
2000	6.4	16.1	77.5	4.2	16.7	79	AN
2001	25	56.3	18.7	44.1	51.4	4.5	N
2002	7.8	4	88.2	5.8	22.2	72	AN
2003	13.6	4.7	81.8	6.8	17.1	76.1	AN
2004	10.6	20.2	69.2	14.6	47.6	37.8	AN
2005	37.9	36.6	25.5	67.5	13.2	19.3	BN
2006	90.7	4.8	4.5	76.2	22.5	1.3	BN
2007	0.8	6.1	93.2	12.5	34.1	53.4	AN
2008	28.7	50.8	20.5	45.6	40	14.4	N
2009	76.7	21.7	1.6	21.9	51.3	26.8	BN
2010	7.2	68.5	24.3	3.9	49.9	46.2	N
2011	28.8	40.2	31	1.3	26.9	71.8	AN
2012	24.4	65.7	9.9	70.7	28	1.3	N
2013	18.4	20.8	60.8	52.1	47.4	0.5	BN
2014	43.5	24.1	32.3	41.7	49.8	8.5	N
2015	49.6	47.9	2.6	66	12.2	21.8	BN
2016	20.6	72.6	6.7	20	45.5	34.4	N
2017	73.5	23.6	2.9	41.2	55.5	3.3	BN
2018	90.9	4.6	4.6	71.1	28.8	0.1	BN
2019	31.1	62.6	6.3	14.7	29.7	55.5	N
2020	13.7	48.2	38.1	20.5	18.3	61.1	AN

MBS and BSS for probabilistic seasonal precipitation forecasts were determined using Equations (3) and (4):

$$\text{MultiClass Brier Score (MBS)} = \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{R}(f_{ti} - o_{ti})^2 \tag{3}$$

$$\text{Brier Skill Score (BSS)} = 1 - \frac{\text{MBS}}{\text{MBS}_{\text{reference}}} \tag{4}$$

where N is the number of years, R is the number of categories, f_{ti} are predicted probabilities of each category for the given monsoon season, o_{ti} are actual probabilities of each category for the given monsoon season, and $MBS_{reference}$ is based on the climatological forecast that is 33.33% for each category.

Table 4 | Evaluation metrics of SPM

Category	Precision	Recall	F1 score	Overall acuracy
For a lead time of 1 month				
BN	0.86	0.86	0.86	0.81
N	0.75	0.86	0.80	
AN	0.83	0.71	0.77	
For a lead time of 2 and 3 months				
BN	0.71	0.71	0.71	0.71
N	0.57	0.57	0.57	
AN	0.86	0.86	0.86	

5. RESULTS AND DISCUSSION

Out of 15 adopted significant predictors, 11 predictors showed a negative correlation ranging from −0.35 to −0.6, whereas four predictors showed a positive correlation ranging from 0.35 to 0.4. Based on the correlation between the anomalies of predictors and aggregated precipitation during monsoon season14 potential predictors out of 16 affecting the ISMR were found to be statistically significant. The *P*-values for nine potential predictors were found to be less than 0.05. Five potential predictors had a p-value 'between' 0.05 to 0.15. As further segregation based on lead time, the number of potential predictors is 15 for lead time of 1 month and 10 each for lead time of 2 and 3 months. Here, eight potential predictors are common to lead time of 1, 2 and 3 months. This includes one local predictor that is ANMP. The segregated potential predictors with a lead time of 1, 2 and 3 months for the seasonal forecast are shown in Table 2.

From the seven classification ML algorithms used with the dataset, LR indicated in highest mean accuracy score as compared to other ML models, as is shown in Figure 3. The LR model was optimized with respect to potential predictors, which were converged for the lead time of 1, 2, and 3 months based on the mean accuracy score of Leave-One-Out CV scheme. For lead time of 1 month, NINO3 and NINO1 + 2 SST over the Pacific, NCPZW, ANMP, ASST, and EASSP were found to be the most optimal combination of potential predictors from the 32,767 combinations tested on the 15 potential predictors with a mean accuracy score of 0.714. For lead time of 2 and 3 months, both the ML models produced similar results. NAO, NWESLP and NASST were found to be the most optimal combination of potential predictors from the 1,023 combinations tested on the 10 potential predictors with a mean accuracy score of 0.57.

For lead time of 2 and 3 months, the mean accuracy score further improved from 0.57 to 0.62 through optimization of hyperparameters of LR- ML model using a random search optimization technique. However, no improvement could be achieved in the mean CV score of 0.714 for lead time of 1-month.

The seasonal forecasts of precipitation in qualitative terms (AN, N, and BN) for lead times of 1, 2, and 3 months for the period 2000–2020 have been shown in Table 3. The actual category of observed precipitation has also been shown in Table 3. A comparison of the seasonal forecasts with the observed precipitation indicates a good agreement between the two. The overall variation in Precision, Recall and F1 Score is more (0.57–0.86) in 2- and 3-month lead time as compared to (0.71–0.86) for 1 month lead time. The Accuracy is more (0.81) for lead time of 1 month as compared to (0.71) for lead time of 2 and 3 months. For 2- and 3-month lead times, AN category has greater (0.86) precision, recall and F1 score as compared to 0.71 for BN category, while N category shows low value of 0.57. The result for lead time of 2 and 3 months indicates SPM model performance for the AN category is better replicated than the other two categories. For lead time of 1 month, performance of SPM is better for all three categories. The results are shown in Table 4.

6. CONCLUSIONS

Probabilistic forecast of seasonal precipitation for a basin using the information of large-scale circulation is an important issue that has been adequately addressed in the present work. In the context of lead time of 1 month, five climate predictors, namely NINO3, NINO1 + 2, SST over the Pacific, NCPZW, ASST and EASSP were found to impact precipitation over the Kosi Basin. In particular, the ENSO conditions in the month of May captured using SST in the region of NINO3 and NINO1 + 2 along with the intensity of the trade winds over the region indicated by NCPZW appeared to play a crucial

role in enhancing or attenuation of precipitation over the Kosi Basin during the monsoon season. The relationship between the SST predictors and the seasonal precipitation over the basin was found to be inversely proportional. Additionally, it could also be inferred that the conditions in the Arabian Sea and East Asia Sea act as an important catalyst in concentrating the monsoon winds over the basin. The results of the research carried out in this work would prove to be useful in addressing water management issues in the Kosi Basin, especially during the monsoon season.

There are several potential advantages of the model developed in the present research. Based on seasonal forecasts, it would be possible to undertake crop planning and estimate irrigation demands. In case the forecast indicates AN rainfall, farmers can plant crops that require more water without worrying about water shortages. Water resource managers can use seasonal precipitation forecasts to develop strategies for meeting water demands in case of BN precipitation forecasts. Seasonal precipitation forecasting can aid communities and emergency services to prepare for potential disasters such as floods and droughts. With the model proposed herein, it would be possible for the businesses to plan for seasonal fluctuations in demand for their products and services that are related to precipitation. An important potential benefit of the research presented herein is that it could assist water resource managers in developing climate change mitigation and adaptation strategies based on seasonal precipitation forecasts. LSCPs can provide prognostic information for seasonal forecasting. Each predictor is likely to have its own affected zones and seasons. The significance of the creation of basin and sub-basin climate indices based on SST, SSP has been shown by Lamb (2010) on the Colorado River basin, where a new SST region is identified as a Hondo region and compared to other established LSCPs like SOI, PDO, NAO and AMO. The test result demonstrated that Hondo performs better at longer lead times than the existing LSCPs. For future research, a statistically-based seasonal precipitation forecast model that automatically identifies suitable predictors from globally gridded SST and climate variables could be developed. A statistical modelling approach could that utilizes ML techniques for categorical forecast of precipitation in terciles (BN, N, and AN) in combination with a conditional stochastic weather generator (CSWG) could also be developed.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories listed in Table 1.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Abid, M. A., Almazroui, M., Kucharski, F., O'Brien, E. & Yousef, A. E. 2018 ENSO relationship to summer rainfall variability and its potential predictability over Arabian Peninsula region. *Climate and Atmospheric Science* **1**, 20171.

Anderson, J., Dool, H., Barnston, A., Chen, W., Stern, W. & Ploshay, J. 1999 Present-day capabilities of numerical and statistical models for atmospheric extratropical seasonal simulation and prediction. *Bulletin of the American Meteorological Society* **80**, 1349–1362.

Apel, H., Abdykerimova, Z., Agalhanova, M., Baimaganbetov, A., Gavrilenko, N., Gerlitz, L., Kalashnikova, O., Unger-Shayesteh, K., Vorogushyn, S. & Gafurov, A. 2018 Statistical forecast of seasonal discharge in Central Asia using observational records: development of a generic linear modelling tool for operational water resource management. *Hydrology and Earth System Sciences* **22**, 2225–2254.

Archer, D. R. & Fowler, H. J. 2008 Seasonal forecasting of runoff on the River Jhelum, Pakistan, using meteorological data. *Journal of Hydrology* **361**, 10–23.

Arnal, L., Wood, A. W., Stephens, E., Cloke, H. L. & Pappenberger, F. 2017 An efficient approach for estimating stream flow forecast skill elasticity. *Computational Hydrology* **18**, 1715–1729. *American Meteorological Society*.

Barlow, M. A. & Tippett, M. K. 2008 Variability and predictability of Central Asia river flows: antecedent winter precipitation and large-scale tele connections. *Journal of Hydrometeorology* **9**, 1334–1349.

Barnston, A. G., Dool, H. M., Zebiak, S. E., Barnett, T. P., Ji, M., Rodenhuis, D. R., Cane, M. A., Leetmaa, A., Graham, N. E., Ropelewski, C. R., Kousky, V. E., O'Lenic, E. A. & Livezey, R. E. 1994 Long-lead seasonal forecasts – where do we stand? *Bulletin of the American Meteorological Society* **75**, 2097–2114.

Behera, S. K., Krishnan, S. & Yamagata, T. 1999 Anomalous air–sea coupling in the southern tropical Indian Ocean during the boreal summer of 1994. *Geophysical Research Letters* **26**, 3001–3004.

Bharath, R. & Srinivas, V. V. 2015 Delineation of homogeneous hydrometeorological regions using wavelet-based global fuzzy cluster analysis. *International Journal of Climatology* **35**, 4707–4727. Wiley Online Library.

Bharati, L., Gurung, P., Priyantha, J., Vladimir, S. & Bhattarai, U. 2014 The projected impact of climate change on water availability and development in the Koshi Basin. *Nepal Mountain Research and Development* **34** (2), 118–130.

Bierkens, M. F. P. & Van Beek, L. P. H. 2009 Seasonal predictability of European discharge: NAO and hydrological response time. *Journal of Hydrometeorology* **10**, 953–968.

Chinnasamy, P., Bharati, L., Bhattarai, U., Khadka, A., Dahal, V. & Wahid, S. 2015 Impact of planned water resource development on current and future water demand in the Koshi River basin, Nepal. *Water International* **40**, 1004–1020.

Colomo, R. A., Nieves, D. C. & Méndez, M. 2019 Comparative analysis of rainfall forecast models using machine learning in islands with complex orography: Tenerife Island. *Applied Sciences* **9**, 4931.

Dong, X. 2016 Influences of the Pacific decadal oscillation on the East Asian summer Monsoon in non-ENSO years. *Atmospheric Science Letters* **17**, 115–120.

Eldaw, A. K., Salas, J. D. & Garcia, L. A. 2003 Long-range forecasting of the Nile River flows using climatic forcing. *Journal of Applied Meteorology and Climatology* **42**, 890–904.

Feliks, Y., Groth, A., Robertson, A. W. & Ghil, M. 2013 Oscillatory climate modes in the Indian Monsoon, North Atlantic, and tropical Pacific. *Journal of Climate* **26**, 9528–9544.

Feng, P., Wang, B., Liu, D. L., Ji, F., Niu, X., Ruan, H., Shi, L. & Yu, Q. 2020 Machine learning-based integration of large-scale climate drivers can improve the forecast of seasonal rainfall probability in Australia. *Environmental Research Letters* **15**, 084051.

Gámiz-Fortis, S. R., Esteban-Parra, M. J., Trigo, R. M. & Castro-Diez, Y. 2010 Potential predictability of an Iberian river flow based on its relationship with previous winter global SST. *Journal of Hydrology* **385**, 143–149.

Goswami, B. N., Madhusoodan, M. S., Neema, C. P. & Sengupta, D. 2006 A physical mechanism for North Atlantic SST influence on the Indian summer Monsoon. *Geophysical Research Letters* **33**, L02706.

Hannachi, A., Jolliffe, I. & Stephensenson, D. B. 2007 Empirical orthogonal functions and related techniques in atmospheric science: a review. *International Journal of Climatology*. **27**, 1119–1152.

Kashid, S. S. & Maity, R. 2012 Prediction of monthly rainfall on homogeneous Monsoon regions of India based on large scale circulation patterns using Genetic Programming. *Journal of Hydrology* **454**, 26–41.

Kirono Dewi, G. C., Chiew Francis, H. S. & Kent David, M. 2010 Identification of best predictors for forecasting seasonal rainfall and runoff in Australia. *Hydrological Processes* **24**, 1237–1247.

Krishnamurthy, L. & Krishnamurthy, V. 2016 Teleconnections of Indian Monsoon rainfall with AMO and Atlantic tripole. *Climate Dynamics* **46**, 2269–2285.

Krishnan, R. & Sugi, M. 2003 Pacific decadal oscillation and variability of the Indian summer Monsoon rainfall. *Climate Dynamics* **21**, 233–242.

Kumar, K., Rajagopalan, B., Hoerling, M., Bates, G. & Cane, M. 2006 Unraveling the mystery of Indian Monsoon failure during El Niño. *Science* **314**, 115–119.

Kumar, A., Pai, D. S., Singh, J. V., Singh, R. & Sikka, D. R. 2012 Statistical models for long-range forecasting of southwest monsoon rainfall over India using step wise regression and neural network. *Atmospheric and Climate Sciences* **2** (3), 322–336.

Kurths, J., Agarwal, A., Shukla, R., Marwan, N., Rathinasamy, M., Caesar, L., Krishnan, R. & Bruno, M. 2019 Unravelling the spatial diversity of Indian precipitation teleconnections via a non-linear multi-scale approach. *Nonlinear Processes in Geophysics* **26**, 251–266.

Lamb, K. W. 2010 *Improving Ensemble Streamflow Prediction Using Interdecadal/Interannual Climate Variability*. Thesis Dissertation.

Lavers, D. A. 2011 *Seasonal Hydrological Prediction in Great Britain–an Assessment*. *PhD dissertation*, University of Birmingham.

Mishra, N., Soni, H. K., Sharma, S. & Upadhyay, A. K. 2018 Development and analysis of artificial neural network models for rainfall prediction by using time-series data. *International Journal of Intelligent Systems and Applications* **10**, 16–23.

Mokhov, I., Smirnov, D. A., Nakonechny, P. I., Kozlenko, S. S., Seleznev, E. P. & Kurths, J. 2012 Relationship between El-Nino/Southern Oscillation and the Indian Monsoon Izvestiya. *Atmospheric and Oceanic Physics: SP MAIK Nauka/Interperiodica* **48**, 47–56.

Pedregosa, F., Varoquaux, G., Gramfort, A., Vincent, M., Bertrand, T., Olivier, G., Mathieu, B., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. 2011 Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830.

Praveen, B., Talukdar, S., Shahfahad, Mahato, S., Mondal, J., Sharma, P., Islam, A. R. T. & Rahman, A. 2020 Analyzing trend and forecasting of rainfall changes in India using nonparametrical and machine learning approaches. *Scientific Reports* **10**, 10342.

Purdie, J. M. & Bardsley, W. E. 2010 Seasonal prediction of lake inflows and rainfall in a hydro-electricity catchment, Waitaki river, New Zealand. *International Journal of Climatology: A Journal of the Royal Meteorological Society* **30**, 372–389.

Rajeevan, M., Pai, D. S., Kumar, A. R. & Lal, B. 2007 New statistical models for long-range forecasting of southwest Monsoon rainfall over India. *Climate Dynamics* **28**, 813–828.

Rasouli, K. R., Hsieh, W. W. & Cannon, A. J. 2012 Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology* **414**, 284–293.

Saha, M., Santara, A., Mitra, P., Arun, C. & Nanjundiah, R. S. 2021 Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. *International Journal of Forecasting* **37**, 158–157.

Shalev-Shwartz, S. & Ben-David, S. 2014 *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, USA.

Sharma, A. & Goyal, M. K. 2015 Bayesian network model for monthly rainfall forecast. IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, India, 2015, 241–246, doi:10.1109/ICRCICN.2015.7434243.

Subrahmanyam, K. V., Ramsenthil, C., Imran, A., Chakravorty, A., Sreedhar, R., Ezhilrajan, E., Bala Subrahamanyam, D., Ramachandran, R., Kumar, K. K., Rajasekhar, M. & Jha, C. S. 2021 Prediction of heavy rainfall days over a peninsular Indian station using the machine learning algorithms. *Journal of Earth System Science* **130**, 240.

Vathsala, H. & Koolagudi, S. G. 2017 Prediction model for peninsular Indian summer monsoon rainfall using data mining and statistical approaches. *Computers & Geosciences* **98**, 55–63.

First received 10 December 2022; accepted in revised form 16 May 2023. Available online 26 May 2023