



OPEN

A support vector machine based drought index for regional drought analysis

Mohammed A Alshahrani¹, Muhammad Laiq², Muhammad Noor-ul-Amin², Uzma Yasmeen³ & Muhammad Nabi⁴✉

The increased global warming has increased the likelihood of recurrent drought hazards. Potential links between the frequency of extreme weather events and global warming have been suggested by earlier research. The spatial variability of meteorological factors over short distances can cause distortions in conclusions or limit the scope of drought analysis in a particular region when extreme values predominate. Therefore, it is challenging to make trustworthy judgments regarding the spatiotemporal characteristics of regional drought. This study aims to improve the quality and accuracy of regional drought characterization and the process of continuous monitoring. The new drought indicator presented in this study is called the Support Vector Machine based drought index (SVM-DI). It is created by adding different weights to an SVM-based X-bar chart that is displayed with regional precipitation aggregate data. The SVM-DI application site is located in Pakistan's northern area. Using the Pearson correlation coefficient for pairwise comparison, the study compares the SVM-DI and the Regional Standard Precipitation Index (RSPI). Interestingly, compared to RSPI, SVM-DI shows more pronounced regional characteristics in its correlations with other meteorological stations, with a significantly lower Coefficient of Variation. These results confirm that SVM-DI is a useful tool for regional drought analysis. The SVM-DI methodology offers a unique way to reduce the impact of extreme values and outliers when aggregating regional precipitation data.

Keywords Drought index, Machine learning, Support vector machine, Drought analysis

A natural phenomenon that occurs all over the world is drought. In many regions of the world droughts are becoming increasingly frequent and severe as mentioned by¹. Coles and Tawn² define drought as a decrease in the amount of water available over a specific period and area. Precipitation stands out as the most crucial climatological factor influencing both droughts and floods. The variability in precipitation can give rise to either of these natural hazards. Wilhite³ analyzed precipitation and drought climatologist offers valuable insights for enhancing water management strategies, environmental protection, agricultural production, and socioeconomic development in specific regions. Drought stemming from insufficient precipitation in a particular area is both a disaster and a naturally occurring hazard. Hirabayashi et al.⁴ highlights that understanding precipitation and drought patterns can significantly contribute to the effective management of water resources, environmental preservation, agricultural practices and socioeconomic progress. Paulo et al.⁵ discussed the frequency, severity and duration of drought exhibit variations across diverse climatic zones. Cai et al.⁶ recognized that drought is one of the most impactful climatic extremes affecting a larger population than any other type of natural disaster.

Mohamadi et al.⁷ discussed that drought indices play a crucial role as tools for monitoring and assessing different types of droughts, including: (a) meteorological drought, denoting a period with insufficient precipitation over a region; (b) hydrological drought, linked to insufficient surface and subsurface water during a specific timeframe; (c) agricultural drought, typically associated with reduced soil moisture leading to crop failure; and (d) socio-economic drought, which is linked to a time when water resource systems are unable to meet demand for water. However, the meanings of drought are always changing to take into account its effects on society and the ecosystem. Kaur et al.⁸ presented that multitude of indices utilizing different variables have been devised to detect and measure occurrences of drought. Included in these are the Surface Water Supply Index (SWSI), the Rainfall Anomaly Index (RAI), the Streamflow Drought Index (SDI), the Palmer Drought Severity Index (PDSI),

¹Department of Mathematics, College of Sciences and Humanities, Prince Sattam Bin Abdulaziz University, 11942 Alkharj, Saudi Arabia. ²Department of Statistics, COMSATS University Islamabad-Lahore Campus, Lahore, Pakistan. ³Department of Mathematics and Statistics, Brock University, St. Catharines, Canada. ⁴Khost Mechanics Institute, Khost, Afghanistan. ✉email: m.nabi.com1990@gmail.com

the Reconnaissance Drought Index (RDI), the Standardized Precipitation Index (SPI), the Standardized Precipitation Evapotranspiration Index (SPEI), and the Standardized Runoff Index (SRI). McKee et al.⁹ introduce the SPI which stands out as the most commonly employed indicator for evaluating meteorological drought, characterized by a brief decline in precipitation leading to reduced water resources availability and ecosystem carrying capacity. Vicente-Serrano¹⁰ mentioned exclusive consideration of precipitation simplifies the calculation process as compared to more intricate indices, enabling the comparison of diverse drought conditions across various temporal and geographical dimensions. Capra and Scicolone¹¹ discussed that SPI is particularly robust and practical as it allows assessment over diverse time spans, facilitating the exploration of various drought categories.

Barker et al.¹² discussed SPI is a non-linear method relies on conventional statistical approaches for drought prediction introducing a considerable level of uncertainty. The application of machine learning (ML) algorithms has been employed for SPI estimation. Support Vector Regression (SVR) has been employed by Borji et al.¹³ to use support vector machines for regression tasks, demonstrating effectiveness in estimating drought. Liu et al.¹⁴ utilize a single-layer feedforward neural network for swift and efficient learning, successfully applied in modeling drought. Deo et al.¹⁵ adopts regression splines to capture non-linear relationships, proving effective in estimating drought. Rhee and Im¹⁶ leverage extremely randomized decision trees for precise and robust drought modeling. Nguyen et al.¹⁷ integrate fuzzy logic and neural networks for adaptive modeling, achieving success in drought estimation. Banadkooki et al.¹⁸ use interconnected nodes to emulate the human brain's learning process, applying it to drought modeling with promising outcomes. Elbeltagi et al.¹⁹ employ an ensemble of multiple decision trees to enhance accuracy and reduce overfitting, demonstrating efficacy in drought estimation. Kushwaha et al.²⁰ discussed that the Super Vector Machine (SVM) is a novel machine-learning algorithm has been recognized as a reliable approach for addressing complex data related issues. Achirul Nanda²¹ discuss the implementation of SVM boasts essential features such as advanced validation, geometric explanation, and precise statistical tracking all achieved with a relatively low number of training data sets. Sihag et al.²² utilizes the kernel functions play a crucial role in SVM model serving as a valuable tool in optimizing the dataset for a more accurate classification method. The prediction of drought is imperative for understanding future drought intensity, enabling effective planning to mitigate the impact of drought conditions and climate changes. Sakaa et al.²³ discussed various models have been developed to forecast drought in semi-arid regions. These machine learning models exhibit the ability to predict information accurately by utilizing the correct input variables. This research clearly reveals a research gap concerning the application of machine learning algorithms in semi-arid environments and drought predictions, in contrast to earlier studies. The results of this work effectively address and overcome a sizable gap in the field of machine learning models for agricultural and meteorological drought prediction as well as drought forecasting.

This study aims to establish a statistical framework that improves the regionality and representativeness of diverse scattered observations within a specific area. As a result, the research introduces a novel tool for monitoring regional drought, grounded in an unequal weighting scheme based on the X-bar chart and SVM regression. This study specifically does: (1) It creates a new way to combine precipitation data from different stations in the region, called the SVM-DI, (2) It calculates the Standardized Precipitation Index (SPI) for each station and compares it to the new SVM-DI for the entire region, (3) It checks how well the SVM-DI compares to the regional SPI by looking at their correlations with each other. In simple terms, the study tries to make a better system for understanding drought in a specific area. It creates a new index called SVM-DI, compares it to existing methods, and checks how accurate it is by comparing it to the RSPI.

Methodology

In the proposed methodology for the SVM-based X-bar control chart the first step involves the comprehensive collection of relevant data pertaining to the targeted process variable. Support Vector Machines (SVM) are employed to train the model, requiring the definition of input features and corresponding target values based on historical process data. In the following subsections, the SVM and the statistical control chart for the mean have been discussed.

Support vector machine

The Support Vector Machines (SVM) for drought prediction/classification involves adapting the SVM framework to the specifics of monthly precipitation data. The mathematical equations would be similar to the general SVM equations, but with considerations for the features and labels related to drought.

The linear equation of the hyperplane for drought prediction is:

$$w \cdot X + b = 0 \quad (1)$$

where, w is the weights of the vector, X is the feature matrix and b is the bias.

The objective function for SVM optimization to maximize the margin while ensuring correct classification would include the term:

$$\text{minimize } \left(\frac{1}{2} \|w\|^2 \right) \quad (2)$$

Subject to the constraints:

$$x_i (w \cdot x + b) \text{ for all } i.$$

The input features for SVM include historical meteorological data. Using the monthly precipitation data, we split the data into train and test. The featured variable is considered as time and the target variables is

precipitation. By using the errors of SVM we make the X-bar control charts and detect the out-of-control points. The X-bar control chart is explained in the next section.

Shewhart X-bar control chart

In statistics, the Statistical Quality Control (SQC) presents the many of control charts for the surveillance of both industrial processes and environmental processes²⁴. In 1924, Shewhart introduced the first control chart, and since then various chart types have evolved to monitor diverse processes. Shewhart control charts are acknowledged as memory-less, signifying their lack of consideration for past information. These charts are particularly adept at identifying significant process shifts, prompting numerous studies aimed at enhancing their efficacy.

The x-bar control chart has the following mathematical structure:

$$\mu = \frac{x_1 + x_2 + \dots + x_n}{n} \quad (3)$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (4)$$

The UCL, CL and LCL are shown in equations as:

$$UCL = \mu + 3\sigma_{\bar{x}} \quad (5)$$

$$CL = \mu \quad (6)$$

$$LCL = \mu - 3\sigma_{\bar{x}} \quad (7)$$

Proposed drought index based on support vector machine

This section provides the description of support vector machine-based drought index SVM-DI. As labeled in Sect. 1, SVM has important role in drought monitoring. In SVM-DI, we predict the values of standard error using the SVM. Next, we use the standard error of SVM to create the X-bar control chart. To do this we find the Out-of-Control Point (OCP) using the X-bar control chart. Additionally, the weights for the aggregation of region data are integrated using the cumulative count of OCP (COCP). The incorporation of SVM in the X-bar control chart is pivotal for the overall effectiveness of SVM-DI. It introduces machine learning adaptability, enhances drought prediction, and contributes to the identification of OCPs. This integration enables SVM-DI to capture unique regional drought patterns with improved accuracy and reliability.

Let $M \in M_1, M_2, M_3, \dots, M_k$ represent the precipitation time series data from various meteorological stations within a designated area. The primary goal in this case is to create weights for the aggregation $M \in M_1, M_2, M_3, \dots, M_k$ in a way that places stations with high COCPs at a lower weight than stations with low COCPs. Figure 1 shows the SVM-DI flow chart.

Stage I: Incorporate of SVM based X-bar control chart.

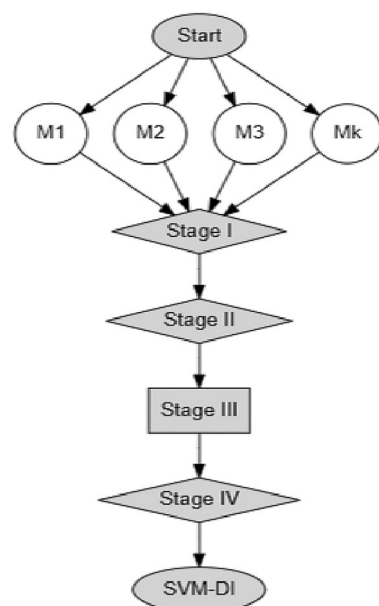


Figure 1. Flow chart of the SVM-DI.

We use X-bar control in the first stage of SVM-DI to detect the OCP. The UCL and LCL in the X-bar control chart are determined by averaging the time series data from all meteorological stations within a given geographic area.

Let Z_i represent the meteorological station mean time series data. Under the X-bar control chart, we constructed the upper and lower control limits for Z_i . The vector with total values outside of the UCL and LCL is displayed by the $(COCP_1, COCP_2, COCP_3, \dots, COCP_k)$. Here, i indicated the quantity of meteorological stations.

Stage 2: Weights Estimation.

In the second stage of SVM-DI, weights for the aggregation of time series data from multiple meteorological stations are determined. In this instance, the weight estimate is based on the COCP that is specific to each weather station. Assuring that observatories with relatively higher COCP are given lower weights and those with relatively smaller COCP are given higher weights are the following formulas.

$$v_i = 1 - \frac{COCP_i}{U} \tag{8}$$

In the above equation, $U_i = \sum_{i=1}^k i$.

$$w_i = \frac{v_i}{\sum_{i=1}^k v_i} \tag{9}$$

The condition is $\sum_{i=1}^k w_i = 1$. Here, k denotes the number of meteorological stations.

Stage 3: Fusion.

After allocating the estimated weights for the regional precipitation data aggregation is the task of this stage in our process. Mathematically, the Weighted Mean Time Series Data (WMTSD) can be computed as follows:

$$WMTSD_i = \sum_{j=i}^k M_{ij}w_{ij} \tag{10}$$

$WMTSD_i$ denotes the regional precipitation data in the equation above using the suggested weighting scheme, where M_{ij} denotes the time series data from the j^{th} meteorological station and the w_{ij} represents the estimated weights in the region.

$WMTSD$ by considering the variability in the importance of meteorological stations through the COCP-based weights contributes to a more nuanced and accurate representation of regional drought. It plays a pivotal role in enhancing the reliability and effectiveness of SVM-DI as a drought monitoring tool. $WMTSD$ is a critical component in this stage ensuring that the regional precipitation data is aggregated in a way that accounts for the significance of individual meteorological stations. Its role extends beyond aggregation, influencing the normalization process and ultimately contributing to the formulation of SVM-DI. This comprehensive approach enhances the ability of SVM-DI to accurately depict the spatiotemporal characteristics of regional drought conditions.

Stage 4: Normalization.

In SVM-DI, normalization is the final step. To model hydrological data, we have used KCGMD²⁵. In this stage, the precipitation time series data are spatiotemporally aggregated regionally, and the Cumulative Distribution Function (CDF) of the KCGMD is fitted to them to normalize them. The KCGMD's CDF has the following mathematical expression:

$$F(q) = F(WMTSD_{i1}) + F(WMTSD_{i2} + \dots + F(WMTSD_{ik}) \tag{11}$$

This paper derives SVM-DI using the normalization approximation proposed by Abramowitz and²⁶. The normalization approximation's mathematical structure is provided by the equation that follows.

$$SVMDI = -(l + \frac{d_0 + d_1l + d_2l^2}{1 + g_1l + g_2l^2 + g_3l^3}) \tag{12}$$

$$l = \sqrt{\ln(\frac{1}{\{F(q)\}^2})}$$

$$0 \leq F(q) \leq 0.5$$

$$SVMDI = +(l + \frac{d_0 + d_1l + d_2l^2}{1 + g_1l + g_2l^2 + g_3l^3})$$

$$l = \sqrt{\ln(\frac{1}{\{1 - F(q)\}^2})}$$

$$0.5 \leq F(q) \leq 1$$

$d_0 = 2.515517, d_1 = 0.802853, d_2 = 0.010328, g_1 = 1.432788, g_2 = 0.985269, g_3 = 0.001308$ are constant.

The choice of K-component Gaussian mixture distribution

The computation of the SDI involves fitting a suitable probability distribution to the time series data of diverse climatic variables. The selection of an appropriate probability model demands careful consideration as the accuracy and reliability of estimates depend significantly on the fitness of the chosen model. Numerous univariate probability distributions have been proposed in previous research for modeling precipitation data in SDI calculations. The gamma distribution has been frequently employed by various authors in computing the SPI, as demonstrated by studies such as those conducted by⁹ and²⁷.

In the last decade, researchers such as²⁸ have put forth various probability models beyond the gamma distribution. The process of selecting the appropriate probability function can be facilitated through the utilization of R packages such as 'fitdistrplus' and 'Propagate,' as suggested by²⁹. These tools contribute to the refinement of the SDI computation by aiding in the identification of a probability distribution that best fits the climatic variables' time series data, thereby enhancing the accuracy and reliability of the index estimates. The use of univariate probability models is deemed inadequate for achieving accurate inferences. Instead, employing a multi-model distribution or a mixture of probability functions is recommended to enhance computational accuracy. In line with this approach³⁰, have proposed the utilization of K-CGMM (K-component Gaussian Mixture Model) for modeling precipitation time series. This advanced modeling technique considers a combination of probability functions, offering a more nuanced and robust representation of the underlying patterns in precipitation data. By adopting such an approach, the study aims to improve the precision and reliability of computational inferences related to precipitation, acknowledging the limitations of traditional univariate probability models in capturing the complexities of climatic variables.

Application

The central focus of the research application is the five weather stations located in northern region of Pakistan (see Fig. 2). This is situated on the second-highest plateau in the world, this selected region is of great significance to the nation's water resource management system. The dataset of selected meteorological stations consist of 41 years from January 1981 to December 2021, together with their corresponding latitude and longitude coordinates, are Astor (35.36 N, 74.84E), Bunji (35.64 N, 74.63E), Gilgit (35.92 N, 74.30E), Gupis (36.22 N, 73.44E), and Skardu (35.30 N, 75.61E). The challenges posed by climate change and global warming, the water resources in this region have experienced unprecedented depletion, leading to heightened risks of drought. This study evaluates the suggested model's implications for more precise regional drought monitoring by utilizing time-series data on monthly precipitation accumulation from 1981 to 2021. The source of the data is power data access viewer of NASA website. The arithmetic mean (AM), standard deviation (SD), and coefficient of variation (CV) in the correlation coefficient between the SPIs of distinct meteorological stations of the SVM-DI and RSPI. The Simple Mean Time Series Data (SMTSD) from all meteorological stations is normalized in this article to produce the RSPI. It's crucial to remember that the normalization used for the Weighted Mean Time Series Data (WMTSD) and SMTSD is the same. By following the³⁰, the RSPI is computed by standardizing the simple average of time series (SATSD) data for all the stations. However, the standardization of SATSD is the same as that used for SVM-DI. The RSPI allows for the identification of wet and dry periods over different time scales, contributing to a nuanced understanding of the region's climatic variability. The RSPI is unable to handle the extreme values in the data where the SVM-DI is useful in this situation.

Results and discussion

The proposed methodology is applied in this section. This section is further divided into three subsections.

Detecting OCP using SVM based X-bar chart

This section presents and analyzes the outcomes concerning the identification of Out-of-Control Points (OCPs) by using residuals based X-bar control chart. The residuals are obtained by using the SVM based regression. The residual control charts are useful when the process exhibits autocorrelations³¹. The correlation in the Table 1 represents the correlation coefficient between the predicted values and the actual values of the target variable. The choice of the SVM based regression method is made on the basis of the correlation coefficient, mean absolute error, root mean square error, relative absolute error and root relative squared error results presented in the Table 1. Based on the results, SVM exhibited the most effective performance for precipitation (correlation = 0.6604, MAE = 0.5766, RMSE = 0.7280, RAE = 0.6853 and RRSE = 0.7524) when compared to the other models.

To improve the accuracy in achieving normality, we partitioned and organized the data on a monthly basis. For each month, UCLs and LCLs were estimated accordingly. The monthly data segregation and plotting for each station made it easier to identify OCP (see Fig. 3). The cumulative number of OCP recorded for each station during each month is displayed in Table 2. In January, 23 out of 41 observations from the Astor station were detected. In contrast, OCPs 23, 29, 21, and 22 have been found for Skardu, Gilgit, Gupis, and Bunji. Table 2 displays that the monthly distribution of OCPs in the precipitation time series data for each of the chosen meteorological stations.

Weights estimation

These weights are estimated based on the degree of heterogeneity or homogeneity observed between stations in different months. The conclusions related to the weights derived from the recommended weighting scheme are discussed and clarified in this section. This method assigns weights to stations based on their COCP; stations with higher COCP are given higher weights, and stations with lower COCP are given lower weights. This makes sense because it is well-known that stations with lower COCPs are more closely aligned with the regional data, while those with higher COCPs diverge from it more. Therefore, it makes sense to assign higher weights to

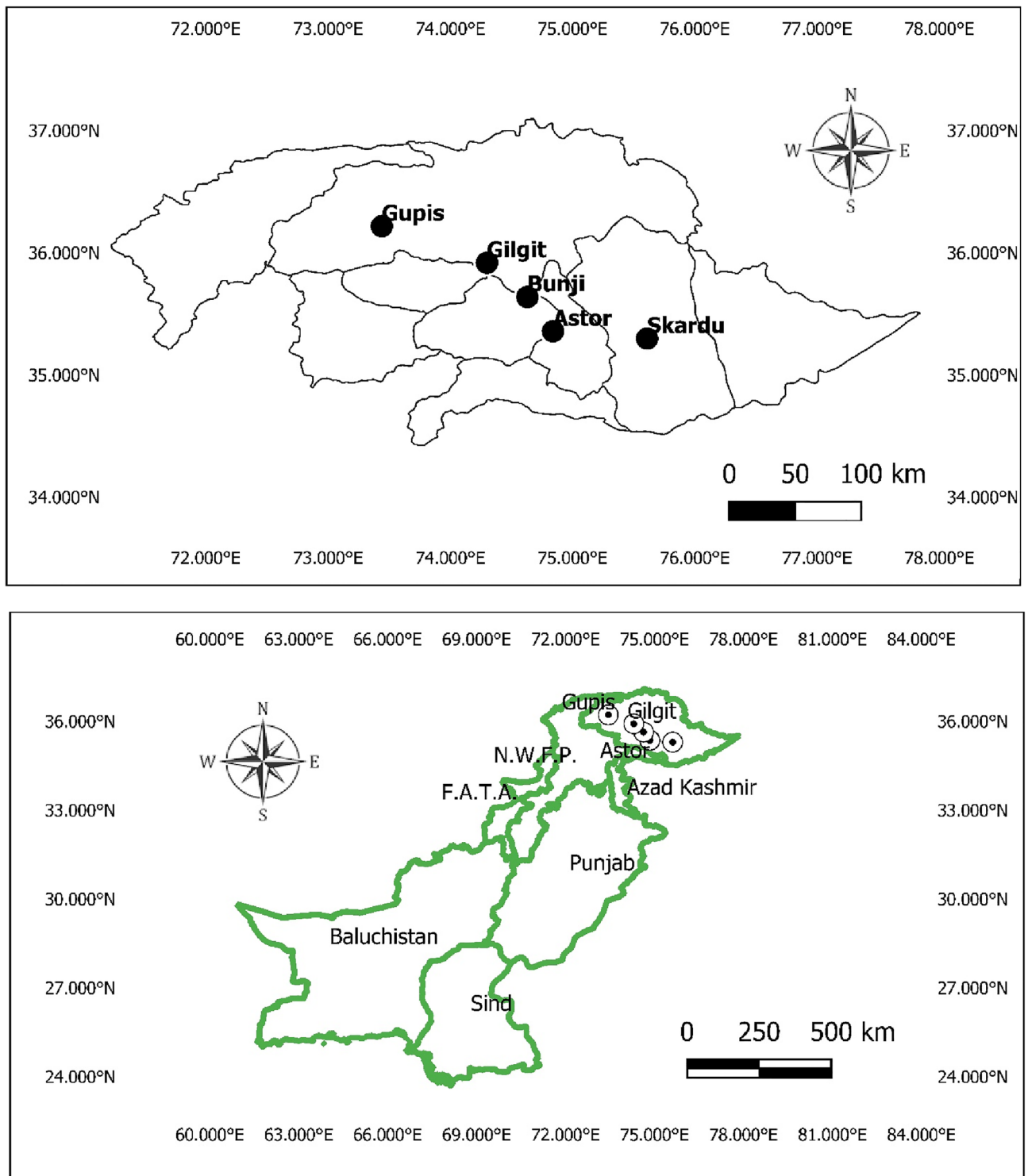


Figure 2. Meteorological stations chosen to assess the SVM-DI implications.

Model	Correlation	MAE	RMSE	RAE	RRSE
Support Vector Regression	0.6604	0.5766	0.7280	0.6853	0.7524
Decision Tree Method	0.5976	0.6631	0.8177	0.7819	0.8119
Gradient Boosting Method	0.6076	0.7106	0.8441	0.8279	0.8378

Table 1. Comparison of the machine learning models.

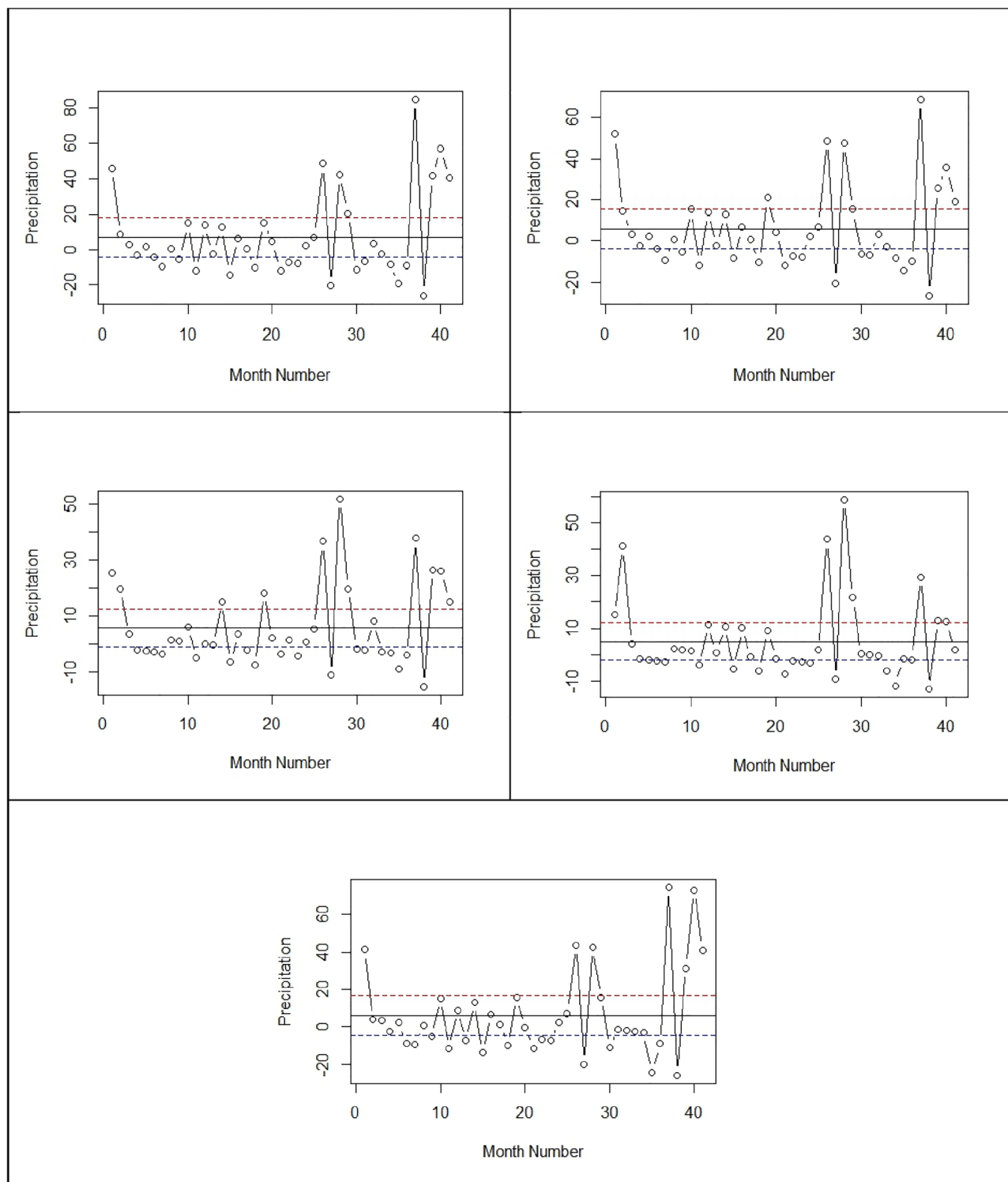


Figure 3. Control Charts Depicting OCP with UCL and LCL for each Station.

stations with lower COCP. Additionally, Table 2 displays the weight distribution for each station on a monthly basis under the suggested weighting scheme. These weights' spatiotemporal variation reflects how each station's significance varies from month to month.

Consistency and efficiency of SVM-DI and RSPI

The Weighted Mean Time Series Data (WMTSD) log-likelihood and Bayesian Information Criterion (BIC) values for each time scale are shown in Table 3.

Month	Astor	Bunji	Gilgit	Gupis	Skardu
January	23	23	29	21	22
Weights	0.2013	0.2013	0.1886	0.2205	0.2039
February	26	25	24	26	26
Weights	0.1988	0.2008	0.2027	0.1988	0.1988
March	32	32	29	24	32
Weights	0.1963	0.1963	0.2013	0.2097	0.1963
April	25	27	30	28	25
Weights	0.2037	0.2000	0.1944	0.1981	0.2037
May	32	32	29	24	32
Weights	0.1963	0.1963	0.2013	0.2097	0.1963
June	24	22	22	24	25
Weights	0.1987	0.2030	0.2030	0.1987	0.1966
July	27	26	30	22	26
Weights	0.1985	0.2004	0.1927	0.2080	0.2004
August	25	25	26	20	21
Weights	0.1966	0.1966	0.1944	0.2073	0.2051
September	19	20	26	24	22
Weights	0.2072	0.2049	0.1914	0.1959	0.2004
October	22	20	28	25	21
Weights	0.2026	0.2069	0.1896	0.1961	0.2047
November	24	21	25	24	21
Weights	0.1978	0.2043	0.1956	0.1978	0.2043
December	26	25	28	28	23
Weights	0.2000	0.2019	0.1961	0.1961	0.2058

Table 2. Monthly allocation pattern of weights and COCP.

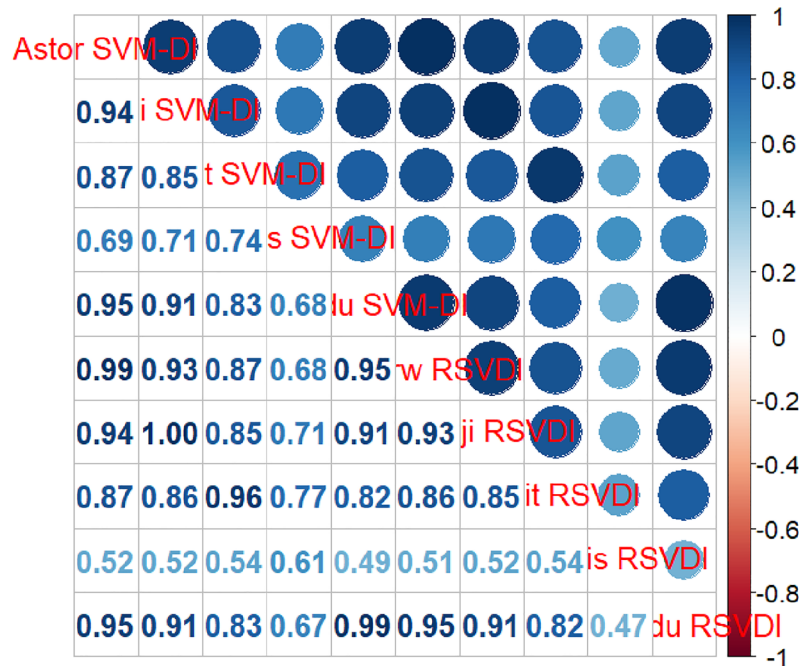
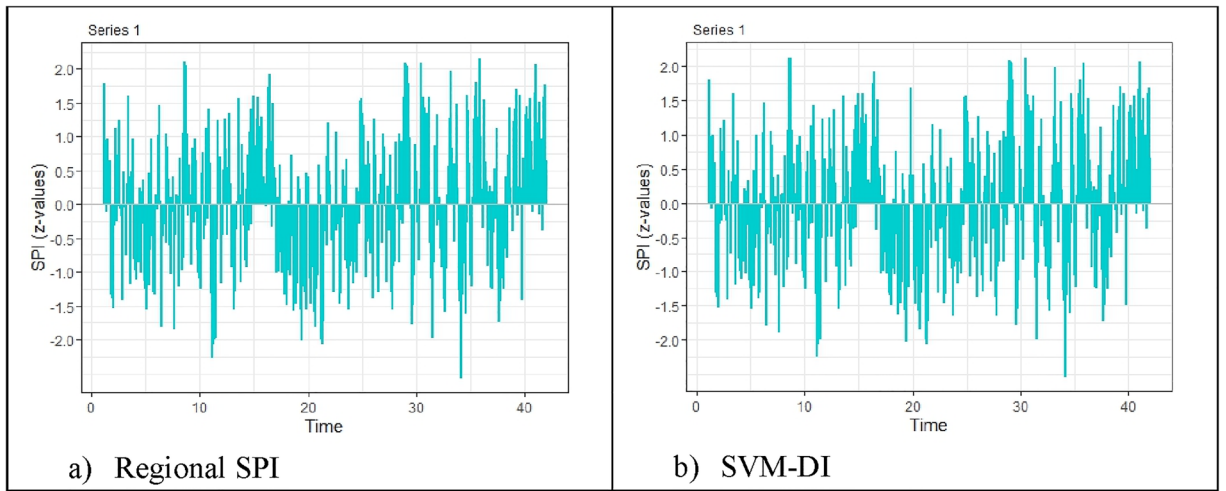
Time scale	Likelihood	BIC
1	698.3220	-345.0667
3	689.3926	-332.4133
6	8713.9890	-332.4248
9	732.7200	-329.5109
12	757.3073	-329.5215
24	842.1898	-322.8307
48	1034.0674	-320.5051

Table 3. Log likelihood and BIC for the KCGMD.

The log-likelihood values are significantly high, and the BIC values are consistently low on all time scales, confirming that the K-CGMD model is appropriate for the WMTSD data. Utilizing the Coefficient of Variation (CV) and Pearson correlation (r), one can evaluate the efficacy of the SVM-DI. We compared the SVM-DI and RSPI correlations with the SPIs of specific meteorological stations in order to assess consistency. The temporal and associative behavior of RSPI and SVM-DI are shown in Figs. 4a–c. Table 4 presents the correlations between the RSPI and SVM-DI and the data from specific meteorological stations-based SPIs over several important time intervals.

The Astor station has the highest correlation value of 0.67081 between SVM-DI and SPI on a one-month time scale, while the Gupis station has the lowest correlation value of 0.62422. Thus for Gupis and Skardu, the highest and lowest RSPI correlations with SPI are 0.80401 and 0.64328. For every other time scale, comparable ranges between the maximum and minimum correlation values are presented in Table 4. The discrepancies indicate that SVM-DI is more consistent than RSPI with regard to the SPIs of individual weather stations.

After the evaluation of consistency assessment, the effectiveness of SVM-DI in comparison to RSPI is assessed. The correlation coefficient statistics (mean, standard deviation, and coefficients of variation) for RSPI and SVM-DI are shown and contrasted in Table 5. The mean correlation between SVM-DI and individual meteorological stations is higher than RSPI at the one-month time scale. The observation that the standard deviation of individual meteorological stations is low implies that SVM-DI is more homogeneous than RSPI. Last but not least, SVM-DI is more consistent than RSPI, as indicated by the low CV in correlation values with the SPI of individual meteorological stations (see Table 5). All of these thorough results point to SVM-DI's greater regional emphasis



c) Correlation

Figure 4. Correlation plot between SVM-DI and RSPI.

than RSPI. This implies that the SVM-DI is a more appropriate tool for representing regional drought in an effective manner. The claim that SVM-DI is a more appropriate indicator for accurate and effective regional drought monitoring is firmly supported by provided results. In conclusion, water resource managers and policymakers will find the implications of SVM-DI to be beneficial, as they will provide deeper understanding of drought conditions and enable the creation of more efficient drought mitigation strategies.

Time scale	Index	SPI				
		Astor	Bunji	Gilgit	Gupis	Skardu
1	SVM-DI	0.67081	0.66115	0.62813	0.62422	0.65529
	RSPI	0.67025	0.66054	0.60996	0.80401	0.64328
3	SVM-DI	0.72413	0.72931	0.70897	0.72974	0.74398
	RSPI	0.73191	0.72773	0.69119	0.82292	0.73922
6	SVM-DI	0.74952	0.75680	0.75636	0.77760	0.78271
	RSPI	0.76309	0.75499	0.75430	0.83331	0.77608
9	SVM-DI	0.76684	0.77382	0.77691	0.79308	0.80790
	RSPI	0.78293	0.77215	0.87449	0.84219	0.80484
12	SVM-DI	0.77394	0.75687	0.78381	0.79508	0.81655
	RSPI	0.78839	0.75499	0.79937	0.85501	0.81402
24	SVM-DI	0.81491	0.83387	0.83072	0.80651	0.86458
	RSPI	0.82529	0.83185	0.86979	0.89646	0.86543
48	SVM-DI	0.78151	0.85470	0.86477	0.81963	0.85631
	RSPI	0.78851	0.82525	0.89443	0.94726	0.85964

Table 4. Correlation analysis of SVM-DI and RSPI.

Time scale	Statistics	SVM-DI	RSPI
1	Mean	0.64790	0.67761
	Standard Deviation	0.02067	0.07429
	CV	3.18977	10.96396
3	Mean	0.72719	0.74259
	Standard Deviation	0.01260	0.04858
	CV	1.73313	6.54229
6	Mean	0.76459	0.77635
	Standard Deviation	0.01464	0.03303
	CV	1.91404	4.25395
9	Mean	0.78368	0.81532
	Standard Deviation	0.01661	0.04256
	CV	2.11926	5.22013
12	Mean	0.78519	0.80235
	Standard Deviation	0.02241	0.03659
	CV	2.85429	4.56061
24	Mean	0.83009	0.85776
	Standard Deviation	0.02226	0.02927
	CV	2.68185	3.41233
48	Mean	0.83538	0.86302
	Standard Deviation	0.03473	0.06138
	CV	4.15784	7.11288

Table 5. Assessment of Correlation Coefficients between SVM-DI and RSPI Values.

Conclusion

The impacts of global warming and climate change have resulted in frequent drought occurrences, significantly affecting various facets of life. To enhance regional drought monitoring, this paper introduces an innovative weighting scheme for the SVM-based X-bar control chart. As a result, the study introduces the SVM-DI, a novel regional drought index. The study focuses on several meteorological stations in Pakistan's northern region in order to assess the efficacy of SVM-DI. CVs in the correlations between RSPI and SVM-DI with specific meteorological stations are investigated for comparative analysis. The numerical results show that, in comparison to the straight forward RSPI, SVM-DI consistently displays more homogenous correlation values across all significant time scales. The higher mean correlation values for SVM-DI highlight its overall stronger association with meteorological stations' SPIs. The lower standard deviation indicates that SVM-DI performance is more stable and less susceptible to variations. These findings imply that RSPI is less able to adequately represent the entire region than SVM-DI. The study's findings demonstrate in summary that SVM-DI based drought monitoring is a useful method for examining drought characteristics at the local level. Practitioners of drought management

can accurately define regional climatology with the help of these findings. This study is limited to the machine learning method by using precipitation data. One may enhance this study by incorporating some other covariates such as the temperature. The deep learning methods can also be utilized in multivariate case.

Data availability

All data analyzed during this study is obtained from the website of NASA power Data Access Viewers (1981–2021) and the website link is <https://power.larc.nasa.gov/data-access-viewer/>.

Received: 8 December 2023; Accepted: 25 April 2024

Published online: 29 April 2024

References

- Beran, M. & Rodier, J. *Hydrological aspects of drought, studies and reports in Hydrology* 39 (Unesco- WMO, 1985).
- Coles, S. G. & Tawn, J. A. Modelling extreme multivariate events. *J. R. Stat. Soc. Ser. B Stat Methodol.* **53**(2), 377–392 (1991).
- Willhite, D. A. Drought as a natural hazard: concepts and definitions (2000).
- Hirabayashi, Y., Kanae, S., Emori, S., Oki, T. & Kimoto, M. Global projections of changing risks of floods and droughts in a changing climate. *Hydrol. Sci. J.* **53**(4), 754–772 (2008).
- Paulo, A. A., Rosa, R. D. & Pereira, L. S. Climate trends and behaviour of drought indices based on precipitation and evapotranspiration in Portugal. *Nat. Hazard.* **12**(5), 1481–1491 (2012).
- Cai, W. *et al.* Increasing frequency of extreme El Niño events due to greenhouse warming. *Nat. Clim. Change* **4**(2), 111–116 (2014).
- Mohamadi, S. *et al.* Zoning map for drought prediction using integrated machine learning models with a nomadic people optimization algorithm. *Nat. Hazards* **104**, 537–579 (2020).
- Kaur, A. & Sood, S. K. Artificial intelligence-based model for drought prediction and forecasting. *Comput. J.* **63**(11), 1704–1712 (2020).
- McKee, T. B., Doesken, N. J., & Kleist, J. The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology* (Vol. 17, No. 22, pp. 179–183) (1993).
- Vicente-Serrano, S. M. Differences in spatial patterns of drought on different time scales: An analysis of the Iberian Peninsula. *Water Resour. Manag.* **20**, 37–60 (2006).
- Capra, A. & Scicolone, B. Spatiotemporal variability of drought on a short–medium time scale in the Calabria Region (Southern Italy). *Theor. Appl. Climatol.* **110**, 471–488 (2012).
- Barker, L. J., Hannaford, J., Chiverton, A. & Svensson, C. From meteorological to hydrological drought using standardised indicators. *Hydrol. Earth Syst. Sci.* **20**(6), 2483–2505 (2016).
- Borji, M., Malekian, A., Salajegheh, A. & Ghadimi, M. Multi-time-scale analysis of hydrological drought forecasting using support vector regression (SVR) and artificial neural networks (ANN). *Arab. J. Geosci.* **9**, 1–10 (2016).
- Liu, C., Yang, C., Yang, Q. & Wang, J. Spatiotemporal drought analysis by the standardized precipitation index (SPI) and standardized precipitation evapotranspiration index (SPEI) in Sichuan Province China. *Sci. Rep.* **11**(1), 1280 (2021).
- Deo, R. C., Kisi, O. & Singh, V. P. Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model. *Atmos. Res.* **184**, 149–175 (2017).
- Rhee, J. & Im, J. Meteorological drought forecasting for ungauged areas based on machine learning: Using long-range climate forecast and remote sensing data. *Agricult. Forest Meteorol.* **237**, 105–122 (2017).
- Nguyen, L. B., Li, Q. F., Ngoc, T. A., & Hiramatsu, K. Adaptive Neuro-Fuzzy inference system for drought forecasting in the Cai River Basin in Vietnam (2015).
- Banadkooki, F. B., Singh, V. P. & Ehteram, M. Multi-timescale drought prediction using new hybrid artificial neural network models. *Nat. Hazards* **106**, 2461–2478 (2021).
- Elbeltagi, A. *et al.* Drought indicator analysis and forecasting using data driven models: Case study in Jaisalmer, India. *Stoch. Environ. Res. Risk Assess.* **37**(1), 113–131 (2023).
- Kushwaha, N. L. *et al.* Data intelligence model and meta-heuristic algorithms-based pan evaporation modelling in two different agro-climatic zones: a case study from Northern India. *Atmosphere* **12**(12), 1654 (2021).
- Achirul Nanda, M., Boro Seminar, K., Nandika, D. & Maddu, A. A comparison study of kernel functions in the support vector machine and its application for termite detection. *Information* **9**(1), 5 (2018).
- Sihag, P., Jain, P. & Kumar, M. Modelling of impact of water quality on recharging rate of storm water filter system using various kernel function based regression. *Model. Earth Syst. Environ.* **4**, 61–68 (2018).
- Sakaa, B. *et al.* Water quality index modeling using random forest and improved SMO algorithm for support vector machine in Saf-Saf river basin. *Environ. Sci. Pollut. Res.* **29**(32), 48491–48508 (2022).
- Shamsuzzaman, M. *et al.* Effective monitoring of carbon emissions from industrial sector using statistical process control. *Appl. Energy* **300**, 117352 (2021).
- Ali, F., Li, B. Z. & Ali, Z. A new weighting scheme for diminishing the effect of extreme values in regional drought analysis. *Water Resour. Manage.* **36**, 4099–4114 (2022).
- Abramowitz, M., & Stegun, I. A. (Eds.). Handbook of mathematical functions with formulas, graphs, and mathematical tables (Vol. 55). *US Government printing office* (1948).
- Bong, C. H. J. & Richard, J. Drought and climate change assessment using standardized precipitation index (SPI) for Sarawak River Basin. *J. Water Clim. Change* **11**(4), 956–965 (2020).
- Stagge, J. H., Tallaksen, L. M., Gudmundsson, L., Van Loon, A. F. & Stahl, K. Candidate distributions for climatological drought indices (SPI and SPEI). *Int. J. Climatol.* **35**(13), 4027–4040 (2015).
- Issam, B. K. & Mohamed, L. Support vector regression based residual MCUSUM control chart for autocorrelated process. *Appl. Math. Comput.* **201**(1–2), 565–574 (2008).
- Ali, Z. *et al.* A new regional drought index under X-bar chart based weighting scheme—the quality boosted regional drought index (QBRDI). *Water Resour. Manage.* **37**, 1895–1911 (2023).
- Spieß, A. N. Propagation of uncertainty using higher-order Taylor expansion and Monte Carlo simulation. **11**, 44 (2018).

Acknowledgements

The authors are deeply thankful to the editor and reviewers for their valuable suggestions to improve the quality and presentation of the paper. This project is sponsored by Prince Sattam Bin Abdulaziz University (PSAU) as part of funding for its SDG Roadmap Research Funding Programme project number PSAU-2023-SDG-2023/SDG/65.

Author contributions

Mohammed Alshahrani: Conceptualization, development of the Support Vector Machine based drought index (SVM-DI), evaluation of machine learning methods Muhammad Laiq: Design of study, Statistical Analysis Muhammad Noor-ul-Amin: Supervision, Software, Programming Uzma Yasmeen: Writeup, Methodology Muhammad Nabi: Manuscript review, modification of methodology and conclusion sections, reorganization of computation work and results validation.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024