Soft Computing (2019) 23:8399–8412
https://doi.org/10.1007/s00500-019-04120-1

FOCUS

Drought prediction based on SPI and SPEI with varying timescales using LSTM recurrent neural network

S. Poornima[1] · M. Pushpalatha[1]

Published online: 6 June 2019

Abstract
Over years, natural calamities like drought have taken a huge toll on human life and resources. As the prediction methods increase, the effects of natural calamities can be reduced to an extent by preplanning and providing warnings to the people. Metrological drought indices like standardized precipitation index and standardized precipitation evapotranspiration index are used to identify drought and its severity level. By forecasting these indices, the occurrences of drought are predicted using the prediction models which help the society to take preventive measures due to the effect of drought. Many research works on prediction majorly focused on statistical methods such as Holt–Winters and ARIMA, but these methods lack accuracy to provide long-term forecasts. However, with advances in the area of machine learning especially artificial neural networks and deep neural networks, there seems to be a method to predict drought in the long term with a good accuracy. Long short-term memory is used in recurrent neural network to predict the drought indices which handle the real-time nonlinear data well and good that can help authorities better prepare and mitigate natural disasters. In this paper, we compare the 1-, 6- and 12-month prediction of the ARIMA statistical model with LSTM using multivariate input in hopes of bettering said performance.

Keywords Data analytics · Big data · Drought · Long short-term memory

1 Introduction

1.1 Drought

Precipitation is a field that is random in character and as such makes drought prediction a complex task. Drought materializes from a deficiency of rainfall over a period of time. Drought over short timescales (months) characterize meteorological drought, whereas long-term scales (years) showcase hydrological drought. Drought causes immense damage to the environment, economy and society. There is an abrupt increase in fires and deflation intensity loss of biodiversity and introduction of pests and diseases. Agricultural losses result in higher suicide rates among farmers, higher cost of food production, lower hydrological energy output and depleted water supply and tourism. They are also responsible for excessive heat waves, limitation of water supplies for consumption, high stress due to failed harvests, etc. Therefore, there is a vital requirement to give accurate prediction of drought occurrence especially for a longer timescale.

1.2 Overview of drought indices

The standardized precipitation index (SPI) was proposed in order to help monitor relative wetness and dryness over multiple timescales (Mckee et al. 1993). Short timescales imply that the SPI is closely related to soil moisture, while at longer timescales it indicates groundwater and reservoir storage. SPI can be used across regions with differing climatic conditions. This results from the quantification of observed precipitation as a selected probability distribution function that is modeled over raw precipitation data.

SPI can be used over 1–36-month timescale and can be interpreted as the number of standard deviations by which

Communicated by Sahul Smys.

✉ S. Poornima
poornima.se@ktr.srmuniv.ac.in

M. Pushpalatha
pushpalatha.m@ktr.srmuniv.ac.in

[1] SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

the observed anomaly deviates from long-term mean. Since SPI is not conducive to climate change associated with evapotranspiration, the standardized precipitation evapotranspiration index (SPEI) has also been proposed. Inclusion of SPEI is to ensure that the limited ability of SPI to capture the effect of increased temperatures is overcome. Other indices include palmer drought severity index (PDSI) and multivariate standardized drought index (MSDI) (Hao and AghaKouchak 2014).

SPI cannot be applied directly to standard time series methods due to two hindrances. Correlation of rainfall is not pertaining to when drought occurs. Also summing up of the rainfall levels over time on the chosen scale results in time-associated correlations (Box and Jenkins 1970). A commonly used method in statistics is the autoregressive (AR) model. Having a long precipitation time series may make it intuitive to apply AR model on it. The AR model would give its prediction as the extract of the seasonal cycle of precipitation.

Drought prediction, however, may not be feasible since it depends heavily on the departure from the seasonality of precipitation. Our proposal is to provide a more accurate long-term prediction using the indices as inputs along with other inputs that are highly related to the indices. We also compare them to the standard univariate method autoregressive integrated moving average (ARIMA).

1.3 Understanding the drought indices

1.3.1 SPI

Several indices pertaining to drought were formulated and used all over the scientific world. These were based on percentage of rainfall and percentile values and sometimes were very complex like the PDSI. The main aim of developing this index was to get an index that was easy in complexity and calculation so that they were applicable in regions all over the world.

The deficit of rainfall has varied effects on various water-related resources and phenomenon such as stream flow, groundwater and soil moisture (Mckee et al. 1993). This resulted in them formulating the SPI. The index is powerful, flexible and fairly easy to use and is simple to calculate. Precipitation of the time series is the only time series required.

Its effectiveness is broad with respect to the phenomenon of rainfall as it can analyze drought and floods in equal measures. The SPI formulation was based on the purpose to implement the precipitation values for several timescales. From the timescales, we can obtain the effects of drought on whether certain water bodies and resources are available or viable for use. The SPI calculation for the place is calculated on the long-term precipitation values for the period one desires. This long-term record is fit into a probability distribution, which is then transformed into a normal distribution. This implies that the mean SPI for the location and period desired is zero (Edwards et al. 1997).

Positive SPI values show greater than median precipitation, while negative values show less than median rainfall. Because the SPI is normalized over a standard distribution, wetter and drier climates can be represented in the same way. Wet periods, whether it is slight or heavy rainfall, can also be monitored using the SPI.

Drought classification is based on Table 1, referred from Mckee et al. (1993). A drought is set to have occurred if the SPI value is continuously negative and reaches a magnitude of -1 or less. The event is supposed to end when the SPI reaches a positive value.

SPI has flexibility and can be calculated for several time periods or timescales. Shorter time period or scale SPI is known to provide early warning of drought and also help predict very accurately the drought intensity. It provides a means of comparison of locations in different climates. Since it is probabilistic in nature, it acquires historical basis that would help in making important decisions. However, it is based on explicitly only the precipitation parameter and has no soil–water balance indicator, and no associated ratios of evapotranspiration/potential evapotranspiration could be evaluated. A new variation of the index in Vicente-Serrano et al. (2010), addressed the potential evapotranspiration (PET) issue by including a temperature component in the calculation of their index called the standardized precipitation evapotranspiration index (SPEI).

1.3.1.1 Computation of SPI The index is shown as the number of standard deviations one gets from the observed precipitation deviating from the long-term mean, for the normal distribution. This distribution of precipitation is not normally distributed, and therefore, before being fitted into probability distribution for the precipitation time series, it has to undergo a transformation such as the gamma function or normal distribution or Pearson distribution.

Initial step in calculation of index is adequate selection of the probability distribution that appropriately fits the

Table 1 SPI range for drought

2.0+	Extremely wet
1.5 to 1.99	Very wet
1.0 to 1.49	Moderately wet
− .99 to .99	Near normal
− 1.0 to − 1.49	Moderately dry
− 1.5 to − 1.99	Severely dry
− 2 and less	Extremely dry

long-term time series of the rainfall. It is known that gamma distribution fits reliably the precipitation distribution. Maximum likelihood estimators are used for fitting of parameters.

The probability density function for gamma function is:

$$g(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1}\, \mathrm{e}^{-x/\beta} \quad \text{for} \quad x > 0$$

where α is a shape parameter, β is a scale parameter and x is the amount of precipitation. $\Gamma(\alpha)$ is the gamma function, which is defined as

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}\mathrm{e}^{-y}\, \mathrm{d}y$$

This distribution to the data requires to be estimated. Edwards et al. (1997) suggest estimating these parameters using the approximation in Thom (1958) for maximum likelihood.

From the below mathematical action, estimates of and yields an expression for the cumulative probability $G(x)$ of an observed amount of precipitation occurring for a given month and timescale.

$$G(x) = \int_0^x g(x)\mathrm{d}x = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_0^x x^{\alpha}\mathrm{e}^{-x/\beta}\, \mathrm{d}x$$

$$G(x) = \frac{1}{\Gamma(\alpha)} \int_0^x t^{\alpha-1}\mathrm{e}^{-1}\, \mathrm{d}t$$

which is the incomplete gamma function. Values of this are calculated based on an algorithm taken from Press (2007). Since the gamma distribution is undefined for $x = 0$, and $q = P(x = 0) > 0$ where $P(x = 0)$ is the probability of zero precipitation, the cumulative probability becomes

$$H(x) = q + (1 - q)G(x)$$

The cumulative probability distribution is then transformed into the standard normal distribution to yield the SPI. Following Edwards et al. (1997), we employ the approximate conversion provided in Abramowitz and Stegun (1965) as an alternative:

$$Z = \mathrm{SPI} = -\left(t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 + d_2 t^2 + d_3^3}\right) \quad \text{for} \quad 0 < H(x) \le 0.5,$$

$$Z = \mathrm{SPI} = +\left(t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 + d_2 t^2 + d_3^3}\right) \quad \text{for} \quad 0.5 < H(x) < 1$$

where

$$t = \sqrt{\ln\left(\frac{1}{(H(x))^2}\right)} \quad \text{for} \quad 0 < H(x) \le 0.5$$

$$t = \sqrt{\ln\left(\frac{1}{1 - (H(x))^2}\right)} \quad \text{for} \quad 0 < H(x) \le 1$$

The design of SPI was to quantify the deficit in precipitation for several timescales. The timescales taken reflect the impacts of drought various water resources which help guide the decision-making process for mitigation of various water-associated issues. The SPI can be calculated for timescales from 1 to 72 months. Statistically, 1–24 months is the best practical range of application (Guttman 2007). This suits for dataset having 50–60-year data for 24 month or more timescales. For longer timescales, 80–100-year data are required. Considering the 33-year time series, SPI-1, SPI-6 and SPI-12 are used for this work.

1.3.1.2 Types of SPI *1-month SPI:* SPI values for one month show the 30-day period representing the normal precipitation percentage. Since the distribution is normalized, SPI is an accurate representation of the precipitation distribution. Therefore, 1-month periods compare the precipitation values for every month with the total of the value for all months of all years taken into consideration. It shows short-term effects, conditions of precipitation that is related to crop loss, etc.

6-month SPI: The 6-month SPI is used to compare the precipitation for that period with the same 6-month period over the historical record. Say, the 6-month period starts from January; then, the precipitation summation for the January–June period with all precious totals for that period is compared and used for future predictions. This type of SPI shows seasonal to medium-term trends in rainfall, considered to be more sensitive than the Palmer index (PDSI).

12-month SPI: The SPI at this timescale is capable of reflecting long-term precipitation patterns. This compares precipitation for 12 consecutive months with that recorded in the same 12 months in previous years in the dataset taken into consideration for prediction. SPIs of these timescales are usually tied to stream flows, reservoir levels and even groundwater levels at longer timescales. In some locations, the 12-month SPI is most closely related to the Palmer index and both the indices can reflect similar conditions.

1.3.2 SPEI

The standardized precipitation evapotranspiration index (SPEI) is a multi-scalar in character and therefore is capable of fulfilling the requirements of a drought index by being flexible enough to be implemented in different scientific disciplines. It too, like the SPI, can measure drought severity and also identify onset of drought. This index allows for juxtaposition of drought intensity in time and space. Moreover, Keyantash and Dracup (2002) showed

that indices indicative of drought must be robust and easy to calculate statistically.

A vital advantage of the SPEI is the inclusion of potential evapotranspiration (PET) in its calculation and would therefore be able to reflect the effect of PET on drought severity and enables identification of different drought types and impacts in the context of global warming.

1.3.2.1 Calculation of SPEI The SPEI is easy to calculate, following the same procedure used in the calculation of the aforementioned SPI. The SPEI uses weekly or monthly difference between precipitation and potential evapotranspiration. This represents a simple climatic water balance that reflects the different timescales in which SPEI can represent the water balance.

Several equations fit to the model the PET over the data such as Thornthwaite equation (Thornthwaite 1948) and Penman–Monteith equation (Allen et al. 1998). The SPEI does not show any linkage to any specific equation mentioned above. Either of the first two seems to be apt in modeling PET. With a value for PET, the difference between the precipitation P and PET for the month i is calculated

$$D_i = P_i - \mathrm{PET}_i$$

which provides a simple measure of the water surplus or deficit for the analyzed month. The calculated D_i values are aggregated at different timescales, following the same procedure as for the SPI.

Normally in time series weather data, the daily rainfall and stream flow follows log-logistic distribution which has heavier tail than gamma distribution. Hence, log-logistic was selected for the gradual decrease in rainfall values that fit the curve with respect to the other useful distributions. The probability density function of a three parameter log-logistic distributed variable is expressed as

$$f(x) = \frac{\beta}{\alpha}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1}\left(1+\left(\frac{x-\gamma^{\beta}}{\alpha}\right)\right)^{-2}$$

where α, β and γ are scale, shape and origin parameters, respectively. Therefore, D_i values ranges between $\gamma > D_{i-} < \propto$ by replacing D_i for x.

When L-moments are computed for the parameters α, β and γ to identify the suitable probability distribution based on the goodness of fit, kappa distribution is obtained by Singh et al. (1993) which is a special case of log-logistic distribution. Accordingly, the probability weighted moments (PWMs) are calculated as:

$$\beta = \frac{2W_1 - W_0}{6W_1 - W_0 - 6W_2}$$

$$\alpha = \frac{(W_0 - 2W_1)\beta}{\Gamma\left(1+\frac{1}{\beta}\right)\Gamma\left(1-\frac{1}{\beta}\right)}$$

$$\gamma = W_0 - \alpha\Gamma\left(1+\frac{1}{\beta}\right)\Gamma\left(1-\frac{1}{\beta}\right)$$

where γ is the gamma function for N number of data and F frequency estimator using the approach (Hosking and Wallis 2005). D_i is the difference between precipitation and potential evapotranspiration for the month I; then, the cumulative probability distribution function of D_i according to the log-logistic distribution is given by

$$F(x) = \left[1 + \left(\frac{\alpha}{x-\gamma}\right)^{\beta}\right]^{-1}$$

With $F(x)$, the SPEI as standardized values is obtained. For example, following the classical approximation of Abramowitz and Stegun (1965):

$$\mathrm{SPEI} = W - \left(\frac{C_0 + C_1 W + C_2 W^2}{1 + d_1 W + d_2 W^2 + d_3 W^3}\right)$$

where $W = -2\ln(P)$ for $P \leq 0.5$ where P is probability of exceeding a determined D_i value, $P = 1 - F(x)$. If $P > 0.5$, P is replaced with $1 - P$ and sign of the resultant value SPEI is changed to opposite to what it was. The average value of the SPEI is 0, and the standard deviation is 1. An SPEI of 0 indicates a value corresponding to 50% of the cumulative probability of D_i, according to a log-logistic distribution. The range of values used for SPEI is similar to the one used by SPI as both fit the standard normal distribution.

2 Literature survey

Many types of methods have been proposed for drought forecasting over the years with varied results. All these studies can be broadly classified into two major categories.

2.1 Statistical approaches

The gamma highest probability (GAHP) method (Bordi et al. 2007) overcomes the problem of extraction of just the seasonal cycle when implementing the AR Model. This limits any knowledge about large deviations that occur. The GAHP approach is focused on forecasting the precipitation value for a month in the future as the most probable precipitation value as shown by the probability density function of the precipitation for the month in consideration. Since it is a gamma distribution, it requires

the estimation of the parameters that fit the frequency histogram of the observed precipitation of the month best. Predicted precipitation for the next month is the observed mode of the fitted distribution. The method again rests on assuming that consecutive precipitation events are not correlated and precipitation is related only to seasonality.

The aforementioned assumptions are empirical and related to the characteristics of precipitation of the region under scrutiny. It is clear that the AR model estimates future precipitation as the mean value in approximation of the precipitation observed for that month. Mean value is not always the most probable value, in matters associated with precipitation, especially since precipitation does not fit the Gaussian distribution of any month. Future discussions must involve predictions for longer timescales. MSDI is a multivariate version of the SPI (Hao and AghaKouchak 2014). It is able to composite multi-index drought information: precipitation and soil moisture.

Ensemble stream flow prediction (ESP) was used to predict the MSDI values where it was shown that the persistence-based model required higher quality and accuracy of the initial conditions and could not account for fluctuations in or missing data.

2.2 Neural network approaches

Artificial neural network (ANN) has been proved to be suitable for complex time series forecasting (Zhang et al. 1998). It showed the resourcefulness of deep neural networks by proposing and unsupervised greedy layer-wise training for deep networks (Hinton and Salakhutdinov 2006). In Morid et al. (2007), the study was conducted in six different provinces of Tehran. An artificial neural network was adopted for prediction of future droughts. Of the various indices present, standard precipitation index and the effective drought index (EDI) were chosen. The study used 32 years of rainfall data for modeling the ANN and also for prediction purposes.

The monthly rainfall data were converted to the chosen indices using the Drought Index Package (DIP) software. The Southern Oscillation Index (SOI) and the North Atlantic Oscillation (NAO) were also considered to be the extra parameters along with the drought indices as input to the ANN. This study used the multilayer perceptron (MLP) as the training algorithm for the ANN. The index data were standardized to range between FMIN and FMAX (FMIN = 0.1 and FMAX < 1) using the following equation

$$X_n = \text{FMIN} + \frac{(X_u - \text{factmin})}{(\text{factmax} - \text{factmin})} \times (\text{FMAX} - \text{FMIN})$$

where $X(u)$ and $X(n)$ represent original and standardized values and the factmin and factmax indicate the minimum and the maximum values in the original dataset.

Various architectures of the ANN were tested with R square, root mean square error (RMSE), moving average error (MAE) as the error measures for validation. After training all the architectures and comparing the results, the 5–6–1 model with five input neurons, six hidden neurons and one output neuron was selected to be the well suited for the EDI. This model produced an R^2 error of 0.84 and 0.79 on training and validation for the EDI. A similar model with a lot more past information was found to be the best for the SPI giving R^2 values in the range 0.66–0.80. The study also showed that the extra parameters like the SOI and the NAO were found to have almost no impact on the performance of the network. It was concluded that ANN was well suited to medium-term forecast (6 months) and EDI outperformed the SPI because of its sluggish nature with no immediate fluctuation which leads to better results. In study Ali et al. (2017), ANN was again used for forecasting droughts in the region of Pakistan. This study used the standardized precipitation evapotranspiration index for forecasting. The ANN model was selected because of its superiority in modeling hydrological data. The data used for calculating the SPEI were from 1975 to 2012, and the SPEI was calculated for four different timescales (1, 3, 6 and 12 months), respectively. The SPEI was calculated by taking the difference between the precipitation data and the potential evapotranspiration, and this difference was later fit to various probability distributions like the gamma, generalized extreme-valued distribution, log-logistic distribution, generalized Pareto distribution and standardized to obtain the final SPEI values for the given data.

The ANN model was selected to be 30–8–1 in input, hidden and output layers, respectively, after experimenting with several other architectures. The momentum for the network was set to 0.5, it was trained for 10,000 epochs and each epoch consisted of a vector of 30 previous values as the input; it was trained on 80% of the data and validated on the rest of the 20%. Error measures like the MAE, correlation coefficient (R) and the RMSE were used to evaluate the model. The study gave the following results measured on the correlation coefficient: 0.887–0.987 for SPEI-1, 0.876–0.994 for SPEI-3, 0.876–0.994 for SPEI-6 and 0.780–0.970 for SPEI-12.

Another study Illeperuma and Sonnadara (2009) used the similar approach of ANN for prediction purpose. This study was based in the Sri Lankan region. The data were collected from 13 meteorological stations over a period of 100 years from 1870 to 1980, covering both dry and wet zones. The SPI was used by fitting the precipitation data to the frequency distribution and fitting this to a probability

distribution. This was performed separately for each month. The dataset was split into training set which consisted of data from 1870 to 1950, and the training set consisted of data from 1951 to 1980. After trying and testing various models, the following model of 30–8–1, with 30 input nodes, 8 hidden nodes and 1 output node, was selected as the best. This model was trained over 500 epochs using the Levenberg–Marquardt optimization weight updating algorithm. The momentum was set to 0.9, and the weight gradient was set to the lowest of 10–6. The error measures used to evaluate the model were MAE, RMSE, *R* and percentage error. The error rates for the SPI with smaller windows were high, and when the window size was increased, the error rate reduced considerably. The results can be summarized as shown in Table 2, on the RMSE error:

The reason for the decrease in error rate with increase in timescale is that as the time period increases, the correlation coefficient between the prediction model and actual value also increases. This happens due to the mapping of SPI values on normal distribution which leads to smooth curve that was easier to predict by the ANN, whereas if the mapping is not done, then rainfall values may generate sudden peak in the curve. When the trained model was used for forecasting the future values, it was found that the best results were obtained for 1-month ahead forecasts, and when the time period was increased to 4–5 months, the accuracy became low. Reason for accuracy degradation may be due to the computation of SPI based on rainfall parameter alone; hence, if additional parameters are used on which rainfall depends then there may be a chance of increase in the accuracy of ANN. Taking a different approach to drought prediction, Dastorani and Afkhami (2011) used ANN to predict drought in Yazd city in Iran. Here the variables like precipitation, minimum temperature, maximum temperature, evaporation, wind speed and wind direction were used directly instead of converting them to a standardized scaled like the SPI or SPEI. The data used were from April 1953 to December 2005. For training the ANN model, the data from 1975 to 2001 were used and the data from 2002 to 2007 were used to test the model. The RMSE and R were chosen as the error measure.

Table 2 RMSE error table

SPI	RMSE	Correlation coefficient
SPI-6	0.005	0.999
SPI-4	0.028	0.999
SPI-3	0.074	0.994
SPI-2	0.193	0.963
SPI-1	0.444	0.823

The lead time was selected to be 12 months in this study. The two ANNs used were the BP neural network and the time lag recurrent network (TLDR); the tangent hyperbolic function was used in the hidden layers, and the sigmoid function was used in the output layer. After running correlation experiments on the different variables, it was found that maximum temperature had the highest impact on the precipitation data; thus, it was selected to help in prediction purposes. The TLDR gave the best results with an R of 0.95 and RMSE of 0.05, this result was obtained after the model converged after training for over 22,000 iterations and the prediction time was 12 months ahead.

In another approach (Chen et al. 2012), this study touched upon the use of state-of-the-art deep learning method for short-term forecasting. The study was conducted over four regions in the eastern part of China; the data collected were over the period of 1958–2006. These data were then used to calculate the SPI-9, SPI-3, SPI-6 and SPI-12. The training set consisted of data from 1958 to 1999, and the rest was used for testing. The new approach was the use of restricted Boltzmann machine which consists of two layers: the input and output; when these restricted Boltzmann machines (RBMs) are stacked together to create many hidden layers, they are called as deep belief networks. The study used the following steps to create and train the model:

1. Compute the different timescale SPI series.
2. Normalize the SPI series.
3. Determine the optimum number of input, hidden and output nodes required to attain efficient performance by trial and error method.

After going through the above steps, the model was set to the following architecture 9–5–10–1, with 9 input nodes, 5 nodes in the first hidden layer, 10 nodes in the second hidden layer and 1 node in the output layer. The error measures used to reach the above architecture were MAE, RMSE. To show the superior performance of deep belief networks (DBNs), the same data were fed to a simple ANN trained using the backpropagation algorithm.

The results for one of the stations for SPI-3 and SPI-6 are summarized in Table 3. Thus, it was concluded that the deep belief network outperformed the BP neural network on all timescales of the SPI, hence proving their superiority in short-term forecasts.

Table 3 Results for the station Bengbu

Station	Model	SPI-3	SPI-6
Bengbu	DBN	RMSE 0.68	RMSE 0.65
		MAE 0.54	MAE 0.52
	BP neural network	RMSE 0.98	RMSE 0.69
		MAE 0.75	MAE 0.58

2.3 Other approaches

2.3.1 Atmospheric electricity

The first attempt to use atmospheric electricity for rainfall and subsequently drought prediction is given in Kulkarni (2015). Instead of using standard drought indices such as the SPI or SPEI, he came up with the atmospheric electrical columnar resistance calculated using satellite data. It is practical and easy to calculate and does not involve the use of probability. Aerosols and convection aid in drop formation, thereby allowing the use of such a parameter, i.e., atmospheric electricity, to predict such complex nonlinear phenomenon. The atmospheric electrical columnar resistance (Rc) is the resistance of a column of the unit cross-sectional area starting from the surface of the earth to the atmosphere. It was found that the anti-correlation lag between the All India Rainfall (AIRF) time series and Rc series over the region of Bay of Bengal is highly significant.

2.3.2 Machine learning methods

Most of the research work on prediction is undergone using data mining algorithms, but data mining techniques have issues in dealing with big data (Poornima and Pushpalatha 2018). Machine learning methods have become more apt and mainstream with respect to accuracy and ease of operation, mainly due to their effectiveness in handling nonlinear characteristics of hydrological data. To overcome the issue of non-stationary data, suffered by both the ANN and SVR methods, researchers have begun to us wavelet analysis to preprocess the input hydrological data (Belayneh and Adamowski 2013).

A wavelet transform is a mathematical tool that provides a time–frequency representation of a signal in the time domain (Partal and Ozgur 2007). It reduces the noise but preserves the features in the input data. Hence, input data are processed using wavelet transform first and then fed to ANN for training in wavelet analysis neural network (WANN). This is taken as the first machine learning model; then, forecasting of SPI for 3- and 6-month timescales is computed using support vector machine taken as second model; finally, SPI computation using ANN without wavelet transform is taken as the third model for comparison. They found that 6-month timescales with month lead times were the most accurate, more than even 1-month lead time predictions. SVR models performed slightly better than ANN in 1-month lead time, whereas ANN had slightly better performance in 3-month lead time. The best results were, however, yielded from the forecasts of wavelet analysis neural network. The results indicate that the use of wavelet analysis as a preprocessing tool provided good forecast results for both ANN and SVR models irrespective of forecast lead time (Table 4).

The retrieved literature used statistical methods like ARIMA on various drought indices like the SPI and SPEI. Most of the latest methods included the use of traditional ANNs for short-term and long-term forecasting. One of the methods also included the use of RBMs, a new deep learning methodology for prediction, and it gave better results compared to the traditional ANN. Other approaches were also taken like the use of Atmospheric electricity for predicting drought and also the use of simple machine learning methods like SVR. Future work may include the use of much advanced deep learning methods like the [long short-term memory (LSTM)] neural network and other variations of it.

3 Implementation

3.1 Dataset

The dataset is composed of maximum temperature, minimum temperature, maximum relative humidity, minimum relative humidity, precipitation, wind speed, sunshine and evapotranspiration of daily reading from 1958 to 2014. First 55-year (1958–2013) data are taken as training data for LSTM, and the last one-year (2014) data are taken for testing. Our aim is to give an accurate prediction of SPI and SPEI for the year 2014 using 1-, 6- and 12-month timescale. Drought index prediction for January 2014 is carried

Table 4 Model-wise results

Model type	Meisso SPI-3 RMSE	Hirna SPI-3 RMSE
ANN L1	0.106	0.108
ANN L3	0.130	0.150
SVR L1	0.086	0.100
SVR L3	0.110	0.122
WANN L1	0.029	0.023
WANN L3	0.029	0.089

out, and the results are compared with ARIMA. The dataset was obtained from CRIDA Labs, Hyderabad, for the Hyderabad region.

3.2 Preprocessing and cleaning dataset

The original dataset consists of minimum and maximum temperature. This is converted to average temperature by adding the two columns and dividing by two. Using the precipitation and evapotranspiration values, SPI and SPEI can be calculated. Evapotranspiration values were not recorded from the year 1958–1979. As a result, we consider data only from 1980. The year 2014 is to be used for comparison with the predicted value and therefore will not be used in training and testing in our proposed method. Hence, the years from 1980 to 2013 will be used for processing and prediction. That gives us a dataset of 34 years which gives us more than 12,000 data points, more than enough to allow good prediction using both autoregressive integrated moving average and long short-term memory—recursive multistep neural network.

Using the package SPEI in *R*, the values of precipitation and evapotranspiration are used to generate SPI and SPEI values. The dataset also has a date parameter that records the data for every day. This is converted to the date data type in R and then used to convert the entire dataset into a time series. Provision of this time series is tantamount to easier processing, especially while employing ARIMA. This conversion to a time series allows for easier processing in R.

3.3 Exploratory analysis

A brief analysis of the dataset is required in order to get an idea of the scope and range of the work to be accomplished.

The SPEI value seems to have a lot of extreme values, while that is not the case for SPI as shown in Figs. 1 and 2. Already it seems SPI is able to handle the drought-related features of the region better. This could be because the region is predominantly mildly wet all through the year.

3.4 Correlation tests

The proposed model involves giving two inputs to the LSTM neural network and one output. The inputs are the drought index (SPI or SPEI) and another parameter not included in the calculation of the index taken into consideration. To determine the second input, correlation test is run to see the correlation coefficient that each parameter has with the respective drought indices. SPI has a positive correlation only with precipitation, average temperature and humidity as shown in Fig. 3. Precipitation is already used in the calculation of SPI and therefore not used as second input. SPEI has a positive correlation with average temperature only. Therefore, a consensus is made as to the selection of average temperature and humidity as the second input to the model LSTM network.

This gives us four combinations of the neural network: SPI + average temperature, SPI + humidity, SPEI + average temperature and SPEI + humidity. This is applied using SPI-1, SPI-6, SPI-12, SPEI-1, SPEI-6 and SPEI-12. This gives us 12 different outputs that will be compared along with the 1-, 6- and 12-month ARIMA predictions for both the indices giving us 18 different predictions to calculate and compare.

3.4.1 SPI correlation

It was found that SPI has high correlation with both humidity and temperature as shown in Figs. 4 and 5. Correlation with other parameters such as wind speed showed negative correlation coefficient values.

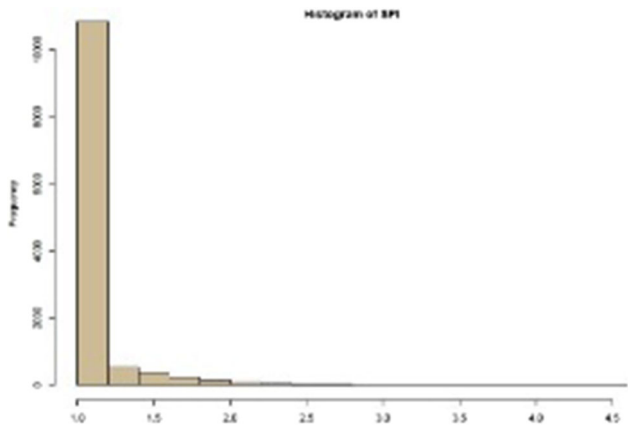

Fig. 1 SPI histogram

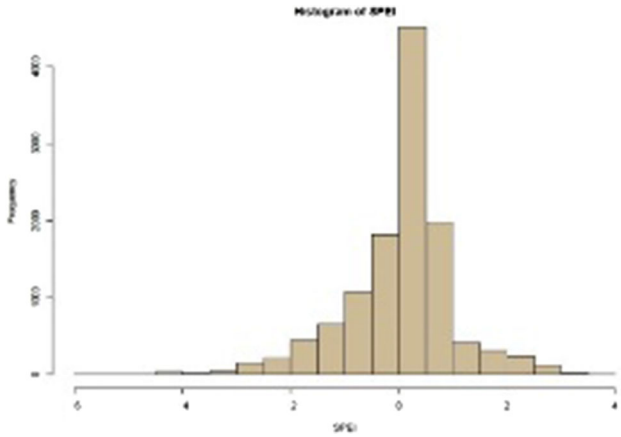

Fig. 2 SPEI histogram

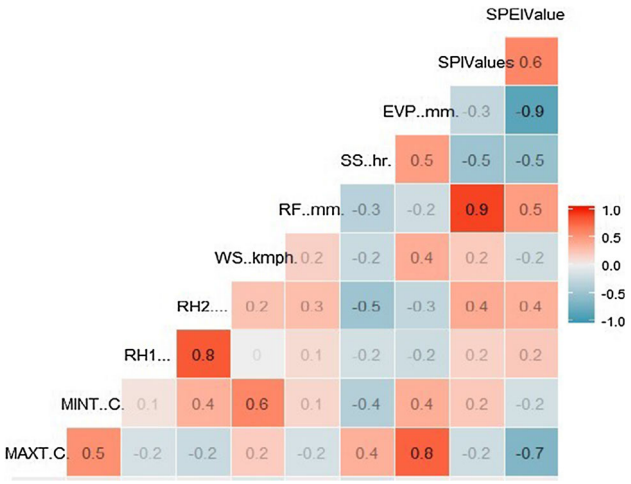

Fig. 3 Correlation map

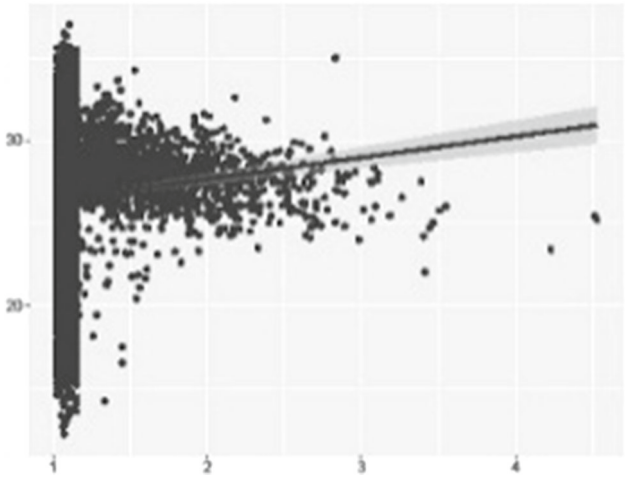

Fig. 4 SPI versus humidity

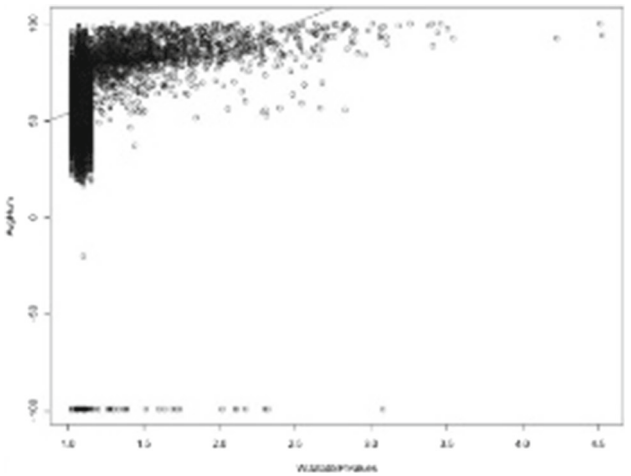

Fig. 5 SPI versus temperature

3.4.2 SPEI correlation

It was found that SPEI has high correlation with humidity but a lower correlation with temperature as shown in Figs. 6 and 7 and negative correlation with all other parameters just as SPI. For continuity's sake, SPEI is also considered with humidity and temperature.

3.5 ARIMA

ARIMA is considered to be one of the most effective prediction methods for univariate time series models. ARIMA models are generally applied where time series show non-stationarity in their data. Therefore, an initial differencing step must be applied one or more times to eliminate the non-stationarity. AR implies that the evolving variable of interest is regressed on its own lagged (prior) values. Moving average (MA) indicates that the regression error is actually a linear combination of error values that occurred contemporaneously. The I part indicates that the values have been replaced by the difference between their values and their previous values. The purpose of this is to fit the model in the data.

Given a time series of data $X(t)$ where t is an integer index and the $X(t)$ are real numbers, an ARMA (p', q) model is given by

$$X_t - \alpha_1 X_{t-1} - \alpha_{p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q'}$$

or equivalently

$$\left(1 - \sum_{i=1}^{p'} \alpha_i L^i\right) X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

where L is the lag operator, the α_i are the parameters of the autoregressive part of the model, the θ_i are the parameters

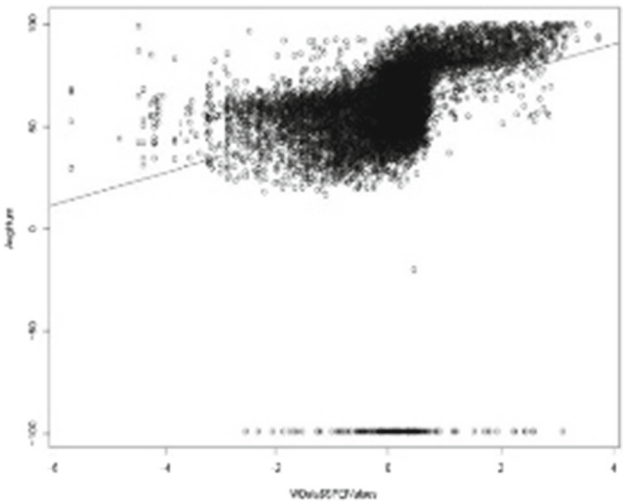

Fig. 6 SPEI versus humidity

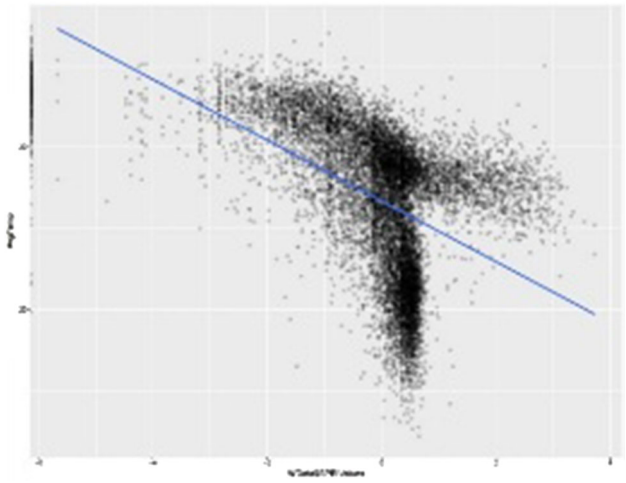

Fig. 7 SPEI versus temperature

of the moving average part and ε_t are error terms. The error terms ε_t are generally assumed to be independent, identically distributed variables sampled from a normal distribution with zero mean.

3.5.1 General ARIMA process

There are off-line and online several resources that give a detailed explanation of how ARIMA works, and are too large to explore in the context of this report, but condensed, the general process of ARIMA is as follows.

1. Plot (dataset) while (graph is non-stationary = true) smoothen graph autocorrelation function (ACF)/partial autocorrelation function (PACF) (stationary graph).
2. Estimate all possible model parameters.
3. Calculate Akaike information criterion (AIC) values of all model parameters.
4. Plot residual graph with no lag.
5. If residual graph has no lag, forecast dataset with these parameters.
6. Else choose other estimated model parameters and repeat step 3.

3.5.2 ARIMA implementation

The proposed model of work revolves around implementing the ARIMA model for SPI and SPEI for 1-, 6- and 12-month timescales.

The lag order for SPI ARIMA is obtained and also can be identified from the ACF for the differenced series as shown in Fig. 8. The MA order of 2 is chosen since using it gives higher negative AIC and BIC values. The PACF for SPI differenced series teeters off at the end as shown in

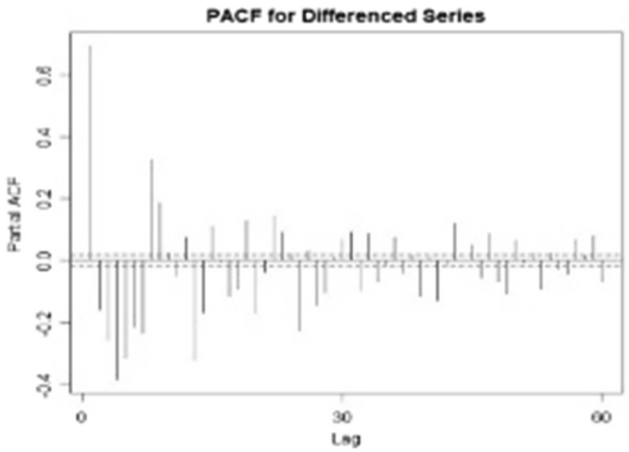

Fig. 8 SPI ACF differenced series

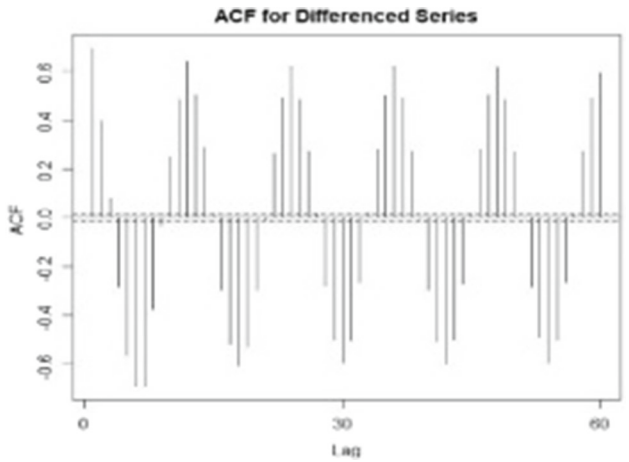

Fig. 9 SPI PACF differenced series

Fig. 9. The fit model gives AIC = − 77,524.96 and BIC = − 77,524.92.

The lag order for SPEI ARIMA is obtained and also can be identified from the ACF for the differenced series as

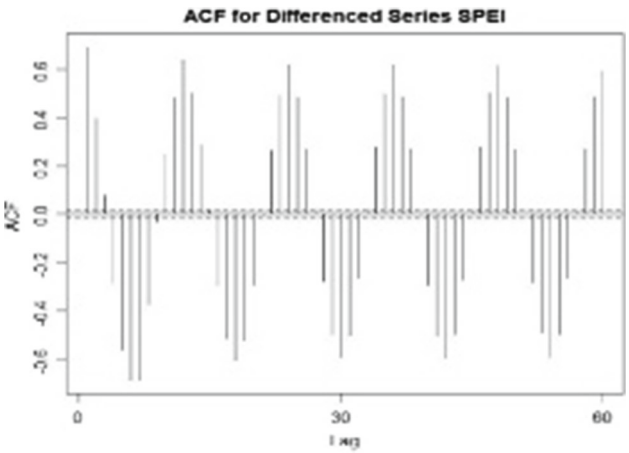

Fig. 10 SPEI ACF differenced series

shown in Fig. 10. The MA order of 2 is chosen since using it gives higher negative AIC and BIC values. The PACF for SPEI differenced series teeters off at the end just like SPI as shown in Fig. 11. The fit model gives AIC = − 96,417.95 and BIC = − 96,417.94. The ARIMA model (1, 2, 2) is fit for both SPI and SPEI.

3.6 LSTM

Long short-term memory (LSTM) is a recurrent neural network (RNN) architecture (an artificial neural network) proposed in (Hochreiter and Schmidhuber 1997). RNNs can capture the dynamics of sequences via cycles in the network. But some RNNs suffer from the vanishing and exploding gradients problem in which gradients are either squashed to zero or increase without bound during back-propagation through a large number of time steps. LSTM is introduced primarily to overcome the problem of vanishing gradients. LSTM network is well suited to learn from experience to classify process and predict time series when there are time lags of unknown size and bound between important events. LSTM is able to model long-term dependencies by using a memory unit called cell state. It has a chain like structure, having four gates which are implemented using the logistic function. All the four gates take the previous state as input along with the current input. The role of each gate is as follows: Forget gate controls the extent to which the value remains in memory, input gate allows the flow of new values in the memory, candidate gate generates the new update for the cell state, and finally, output gate allows the value in memory to compute the output activation of the block further. Forget gate, input gate and output gate use sigmoid function to perform its task, whereas the candidate gate and output gate use tanh function. All the logistic functions are computed by applying a weight and a bias to trigger the neurons and normalize the inputs. Every neuron in the hidden layer of recurrent neural network is implemented with LSTM unit and undergoes for number of states for prediction. The above-mentioned functionalities of LSTM are shown in Fig. 12.

The information given depicts the forward pass and backward pass in LSTMs. In terms of the forward pass, the LSTM can learn when to let activation into the internal state. As long as the input gate takes value zero, no activation can get in. Similarly, the output gate learns when to let the value out. When both gates are closed, the activation is trapped in the memory cell, neither growing nor shrinking, or affecting the output at intermediate time steps. In terms of the backwards pass, the constant error carousel enables the gradient to propagate back across many time steps, neither exploding nor vanishing. In this sense, the gates are learning when to let an error in and when to let it out. The minimization of LSTM total error is achieved by implementing the iterative gradient descent such as backpropagation. It changes each weight in proportion to its derivative with respect to the error.

3.6.1 LSTM implementation

The LSTM RNN model proposed in this study is split into two models. First one is the univariate case where only the SPI and SPEI values are used. In this architecture, we use one layer of LSTM which consists of 1024 cells. The backpropagation through time (BPTT) is limited to one step. A dropout layer is included between the two hidden layers for regularization. It will randomly exclude 50% of the activations of the previous layer from propagating to prevent overfitting.

The root mean square (RMS) loss is reduced using the AdaGrad algorithm which increases the learning rate for more sparse parameters and decreases the learning rate for less sparse ones. This strategy often improves convergence performance over standard stochastic gradient descent in settings where data are sparse and sparse parameters are more informative. The initial learning rate is taken to be 7 and is exponentially decreased when the RMS loss does not improve for more than 10 epochs. The training was stopped after the loss started to fluctuate despite very low learning

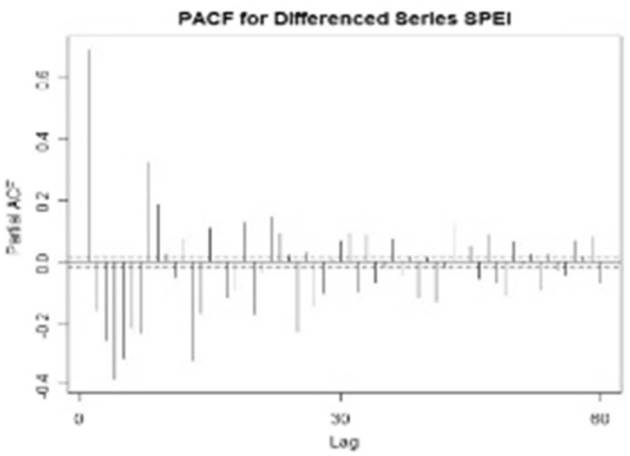

Fig. 11 SPEI PACF differenced series

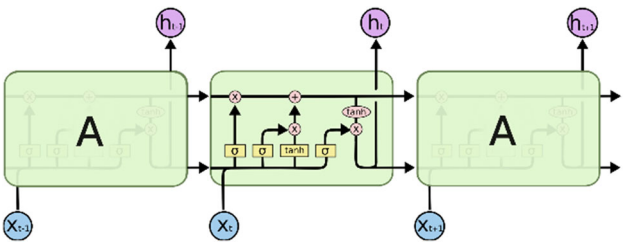

Fig. 12 LSTM network

rate. The number of epochs came to be 8000. These parameters were selected after trying out other architecture.

The second architecture is shown in Fig. 13 is complex compared to the first one to accommodate the extra parameters like temperature and relative humidity. This model uses two layers of LSTM with 512 units in each of the layers. The dropout is set to 50% and this is trained using RMSProp algorithm which was better suited for this particular case since it resolves the problem of diminishing learning rates. The training was stopped after 4000 epochs as the loss started to plateau. From the two methods of implementation mentioned above, it was understood that the increase in number of layers increases the accuracy. Due to the same reason, the number of epochs is also decreased to almost half with the least network loss. But the only constraint in adding the number of layers is execution time that too in recurrent neural network every neuron in the hidden layer undergoes for recursion to certain number of states that increases the time complexity almost to double. It is precise to use two hidden layers for the dataset used in this work for good accuracy in prediction, but as the size of the dataset increases then accordingly deeper network can be used for better accuracy and learning with minimum loss. The current focus of our research work is to design a better LSTM network that reduces the number of epochs to a minimal range by preserving good accuracy and learning rate by avoiding vanishing gradient problem that arises due to sigmoid and hyperbolic tangent function used in LSTM.

4 Results and discussions

The error metric used is root mean square error (RMSE). This gives us a good indication of model performance for each model and allows us to get a good inclination of the order of performance. We also include accuracy calculation and compare the 1-, 6- and 12-month predictions with the SPI and SPEI values of the year 2014 whose actual data are already available.

4.1 ARIMA results

The results of SPI and SPEI predictions using ARIMA are shown in Table 5. Accuracy is calculated based on the difference in the value predicted using ARIMA and the actual value available in the dataset for January 2014. For 1-month scale, ARIMA performs very well with an accuracy of 99% and RMSE of 0.01. For SPI-6 and 12, however, the performance deteriorates because of the working principle of ARIMA Model. The value stagnates because the expected value is a cumulative sum of the period considered before.

Table 5 ARIMA prediction for SPI and SPEI

	Given	Predicted	Accuracy (%)	RMSE
SPI timescale				
SPI-1	1.089221	1.083111	99	0.010
SPI-6	0.617220	1.119165	25	0.03
SPI-12	0.405042	1.1321	15	0.019
SPEI timescale				
SPEI-1	0.493370	0.502132	98	0.06
SPEI-6	0.398685	1.119164	10	0.06
SPEI-12	0.448110	1.119168	12	0.06

ARIMA (1,2,2) is a non-stationary time series with one period ahead forecast. It is just a random walk, i.e., a cumulative sum of innovations or shocks. Forecasting can be done any number of times based on one period ahead in ARIMA. But during the long-term prediction, forecast of many periods ahead has to be computed where ARIMA is not a suitable model, since if the expected value of a new innovation is zero, then the expected cumulative sum one period ahead is just equal to the current value of the cumulative sum. Therefore, the forecast is equal to the last observed value. Those applications which involve weather data may often suffer with such issue because the rainfall value may be zero for a continuous period of time even more than 3 months. Thus, the SPI and SPEI values may be normalized to 0 which leads to the above-mentioned problem during future predictions using ARIMA. Meanwhile, a forecast for a stationary time series will almost never be equal to the last observation (although there may be some special cases). Therefore, as observed the performance for SPI 6 and 12 are quite inaccurate.

SPEI follows a similar trend and shows a good performance in SPEI-1. The 6- and 12-month performances are quite poor; hence, this univariate prediction model is not conducive in long run.

4.2 LSTM results

The SPI and SPEI prediction results using LSTM are shown in Table 6. SPI + relative humidity on a 1-month timescale gives a good accuracy of 99% and an RMSE of

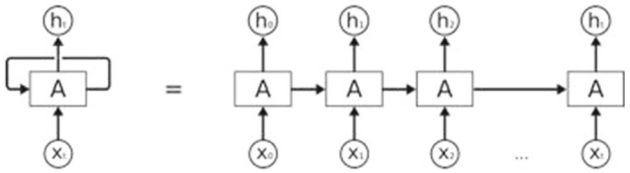

Fig. 13 LSTM RNN architecture

Table 6 LSTM prediction for SPI and SPEI with humidity and temperature

Timescale	Given	Predicted	Accuracy (%)	RMSE
SPI + humidity				
SPI-1	1.089221	1.083111	99	0.011
SPI-6	0.617220	0.600285	97	0.03
SPI-12	0.405042	0.403718	99	0.015
SPI + temperature				
SPI-1	1.089221	1.081234	94	0.03
SPI-6	0.617220	0.598785	89	0.05
SPI-12	0.405042	0.400811	95	0.03
SPEI + humidity				
SPEI-1	0.493370	0.507919	97.05	0.2
SPEI-6	0.398685	0.418762	94	0.2
SPEI-12	0.448110	0.460447	97	0.1
SPEI + temperature				
SPEI-1	0.493370	0.515212	91	0.2
SPEI-6	0.398685	0.418762	94	0.2
SPEI-12	0.448110	0.451347	96	0.1

0.01. This shows that LSTM can be useful for short-term predictions even when it is computationally expensive. On the longer timescales, LSTM has vastly outperformed ARIMA which plateaued after a period of time. The LSTM is thus a better solution for a long timescale prediction. It performs better in both 6- and 12-month timescales.

Temperature has a lower correlation with SPI compared to relative humidity and this is reflected in the results. SPI + temp on a 1-month timescale gives a good accuracy of 94% and an RMSE of 0.03. On the longer timescales, LSTM was still not very good at 6-month and 12-month timescales compared to the other models; it had a lower accuracy and a higher RMSE.

SPEI + relative humidity on a 1-month timescale gives a good accuracy of 97.05% and an RMSE of 0.2. On the longer timescales, LSTM has vastly outperformed ARIMA which plateaued after a period of time. Even though ARIMA had a better RMSE result, it failed to accurately predict for 6-month and 12-month timescales.

SPEI + temperature on a 1-month timescale gives a considerable accuracy of 91% and an RMSE of 0.2. On the longer timescales, LSTM was not very good at a 6-month and 12-month timescales compared to the other models; it had a lower accuracy and a higher RMSE.

5 Conclusion and future enhancement

ARIMA and LSTM models were applied on the given dataset of about 12,000 data points. ARIMA is a univariate approach, whereas LSTM follows a multivariate approach. Temperature and humidity were the two parameters not involved in the calculation of SPI and SPEI that had positive correlation with the two indices. Therefore, they were included in the multivariate approach as used in the LSTM neural network. It can be concluded that the long short-term memory model has showed better results compared to the ARIMA model for predictions on a longer timescale, i.e., 6 month and 12 month. It was also seen that LSTM performed better when extra variables which had positive correlation with SPI and SPEI such as relative humidity and temperature were added to the training. LSTM computation, however, is more resource intensive than the ARIMA model. It can also be observed that ARIMA provided a good solution to short-term prediction and was also computationally lighter compared to the LSTM network.

Future work in the area can be done by using a larger dataset which might yield better results and also by using ensemble deep learning methods which have a better chance at capturing the intricacies of a difficult pattern like drought. At some regions, the parameters that are correlated might be different based on the general precipitation cycle. Certain weather phenomenon such as El Nino may also need to be taken under consideration.

Funding Funding is not applicable.

Compliance with ethical standards

Conflict of interest All authors declare that there is no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Not applicable.

References

Abramowitz M, Stegun IA (1965) Handbook of mathematical formulas, graphs, and mathematical tables. In: National bureau of stand mathematics series, vol 55

Ali Z, Hussain I, Faisal M, Nazir HM, Hussain T, Shad MY, Mohamd Shoukry A, Hussain Gani S (2017) Forecasting drought using multilayer perceptron artificial neural network model. Adv Meteorol 2017:1–9

Allen RG, Pereira LS, Raes D, Smith M (1998) Crop evapotranspiration- guidelines for computing crop water requirements-FAO irrigation and drainage paper 56. FAO, Rome

Belayneh A, Adamowski J (2013) Drought forecasting using new machine learning methods. J Water Land Dev 18:3–12

Bordi I, Fraedrich K, Petitta M, Sutera A (2007) Extreme value analysis of wet and dry periods in sicily. Theoret Appl Climatol 87(1):61–71

Box G, Jenkins G (1970) Time series analysis; forecasting and control. Holden-Day, San Francisco

Chen J, Jin Q, Chao J (2012) Design of deep belief networks for short-term prediction of drought index using data in the huaihe river basin. Math Probl Eng 2012:1–16. https://doi.org/10.1155/2012/235929

Dastorani MT, Afkhami H (2011) Application of artificial neural networks on drought prediction in Yazd (Central Iran). Desert 16(1):39–48

Edwards C, McKee T, Doesken N, Kleist J (1997) Historical analysis of drought in the United States. In: 7th conference on climate variations, 77th AMS annual meeting, vol 27

Guttman NB (2007) Accepting the standardized precipitation index: a calculation algorithm. JAWRA J Am Water Resour Assoc 35(2):311–322

Hao Z, AghaKouchak A (2014) A nonparametric multivariate multi-index drought monitoring framework. J Hydrometeorol 15(1):89–101

Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Hosking JRM, Wallis JR (2005) Regional frequency analysis: an approach based on L-moments. Cambridge University Press, Cambridge

Illeperuma G, Sonnadara U (2009) Forecasting droughts using artificial neural networks. In: National symposium on disaster risk reduction and climate change adaptation, vol 1

Keyantash J, Dracup JA (2002) The quantification of drought: an evaluation of drought indices. Bull Am Meteorol Soc 83(8):1167–1180

Kulkarni MN (2015) A new tool for predicting drought: an application over India. Sci Rep 5:1–8, Article 7680

Mckee TB, Doesken NJ, Kleist J (1993) The relationship of drought frequency and duration to time scales. In: 8th Conference on applied climatology, California, 7–22 Jan 1993

Morid S, Smakhtin V, Bagherzadeh K (2007) Drought forecasting using artificial neural networks and time series of drought indices. Int J Climatol 27(15):2103–2111

Partal T, Ozgur K (2007) Wavelet and neuro-fuzzy conjunction model for precipitation forecasting. J Hydrol 342(1):199–212

Poornima S, Pushpalatha M (2018) A survey of predictive analytics using big data with Data mining. Int J Bioinform Res Appl 14(3):269–282

Press WH (2007) Numerical recipes, 3rd edition: the art of scientific computing. Cambridge University Press, Cambridge

Singh VP, Guo H, Yu FX (1993) Parameter estimation for 3-parameter log-logistic distribution (LLD_3) by pome. Stoch Hydrol Hydraul 7(3):163–177

Thom HCS (1958) A note on gamma distribution. Mon Weather Rev 86(4):117–122

Thornthwaite CW (1948) An approach toward a rational classification of climate. Geogr Rev 38(1):55–94

Vicente-Serrano SM, Begueria S, Lopez-Moreno JI (2010) A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. J Clim 23(7):1696–1718

Zhang G, Patuwo BE, Hu MY (1998) Forecasting with artificial neural networks: the state of the art. Int J Forecast 14(1):35–62