Natural Resources for Human Health
eISSN: 2583-1194

Original Research
Received 18 July 2026
Revised 14 Aug 2026
Accepted 28 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (19s): 1535–1548

KEYWORDS:
agricultural drought; SPI; prediction lead time; persistence baseline; long short-term memory; forecast verification.

Persistence-Relative Evaluation of an Adaptive-Weighted LSTM for Multi-Scale, Multi-Lead Agricultural Drought Prediction

Ritu Khandelwal[1], Dr. Hemlata Goyal[2]* and Dr. Rajveer Singh Shekhawat[3]

[123]School of Computer Science and Engineering, Manipal University Jaipur, Jaipur, India
[2*]hemlata.goyal@jaipur.manipal.edu (Corresponding author)

ABSTRACT: The accuracy of deep-learning drought prediction models is normally determined by the ability of the forecasted values to match actual values, which can be used to evaluate prediction performance. Standardized Precipitation Index (SPI) is the major agriculture drought index and calculated on overlapping rainfall totals, to accumulate extended periods and generate less volatile time series. It is used to predict adequate persistence forecast carrying forward the current value with increasing accumulation period to reflect model predictive ability. In this work, we forecast SPI for four accumulation periods (3, 6, 9, and 12 months) and for three lead times (1, 3, and 6 months) at 55 grid locations in the Jaisalmer district of Rajasthan, India, employing a Long Short-Term Memory network with temporal attention mechanism. The main model utilized precipitation anomalies, SPI, and seasonal calendar features, while the role of satellite-derived indices is examined in separate sensitivity analysis. . The hyperparameters are tuned using validation samples within each walk-forward sample split. All twelve scale– lead pairs are assessed relative to persistence and climatology using Diebold–Mariano tests in three walk-forward sample splits. For a one-month forecast lead, the absolute accuracy is very high for SPI-3, SPI-6, and SPI-9, but not as high for SPI-12, with the R^2 values being 0.67, 0.68, 0.65 and 0.33, respectively. With respect to persistence, there are only two configurations that outperformed it in all three walk-forward splits while maintaining the positive absolute R^2: SPI-3 at a one-month forecast lead and SPI-6 at a three-month forecast lead. Drought classification followed the same trend. The ROC-AUC is higher than 0.85 in eight configurations, while AW-LSTM outperformed persistence only in four configurations.

1. INTRODUCTION

An agricultural drought is a slow onset hazard that poses a problem to agriculture that depends entirely on rains. In the arid and semi-arid areas, crops and animals survive on rainfall that lasts for only a short duration of time. This is the case with the Thar Desert in western Rajasthan, where mean annual rainfall is only 254mm, occurring during the southwest monsoon season of May to September. Small deviations from this norm have impacts on the sowing, feeding, groundwater recharge and animal survival and hence drought forecast is of paramount importance[1].

Standardized Precipitation Index (SPI), which is formulated by McKee et al. [2], is an index of precipitation anomaly based on accumulation period. While short accumulation periods are sensitive to precipitation deficits of recent past, long periods capture accumulation of deficits over a period of months indicating dry spell. Thus, SPI forecast at different accumulation periods gives drought forecast at multiple time horizons. A few studies have investigated multi-scale SPI forecast using single unified framework analysis, ranging from SPI-1 to SPI-12 [3-5]. LSTM (Long Short-Term Memory) networks [6] have been extensively used in hydroclimatic series prediction due to their capability of capturing nonlinear temporal dependencies through sequential learning [7-10]. Another approach to finding the relative importance of elements of an input sequence is through attention

mechanisms [11]. The use of this method in the case of drought prediction and interpretation is studied in [12], where a model of national-scale drought prediction over India is built [13]. There have also been studies on decomposition and hybrid models [14-16], spatio-temporal architectures [17], multi-target prediction methods [18] and explainable sub-seasonal-to-seasonal agricultural drought forecast models [19]. These recent reviews show that artificial intelligence drought forecast methods develop very rapidly, while forecast lead time and predictor choice have been shown to affect the performance of the models described [20-21].

Despite all the advances mentioned above, assessing prediction accuracy is still a challenging task. In the case where the accumulation period increases, successive SPI values are based on the precipitation totals that include many common months. The result is smoothing out of the time series and an increase of its autocorrelation. A persistence forecast, that simply extends the most recent SPI value forward, may thus show better performance at larger accumulation scales despite learning nothing. Larger R^2 for SPI-12 compared to SPI-3 does not necessarily mean that the model learned more. Accuracy in absolute terms combines two factors at once: how predictable the target variable is and what additional information the model brings. R^2 cannot distinguish between the two. Forecast assessment in relation to some reference forecast has a long history in drought and hydrological prediction [22-23]. Operational experience also supports this approach: The International Research Institute for Climate and Society assesses its seasonal SPI forecasts against optimal persistence [24]. Evaluation design is also recognized as a drawback of artificial-intelligence drought research [21].

Accumulation scale and lead time have an interactive relationship; with increasing lead time, the more recent SPI value becomes less representative of the target month, and hence persistence reduces. However, SPI at a long accumulation scale retains considerable dependence in time, so that which scale works better could itself be conditional on the lead time. Multi-scale analysis considers various accumulation scales, while the interaction between scale, lead time and improvement in accuracy has not yet been studied directly.

This study explores the performance of an adaptive-weighted LSTM model with temporal attention mechanism (AW-LSTM) for drought index prediction in Jaisalmer district, Rajasthan, India. SPI models at four accumulation scales - SPI-3, SPI-6, SPI9, and SPI-12 – for lead times of 1, 3, and 6 months, giving rise to twelve scale-lead combinations. The performance is assessed through absolute accuracy and an increment in predictive capability using Diebold-Mariano tests [25].

This research introduces three key contributions as follows-

- First, it measures the degree of temporal dependence in each of the SPI targets and correlates it with prediction accuracy.
- Second, it performs ablation analysis with respect to different preprocessing and objective function combinations and demonstrates how some improvements and some don't.
- Finally, it compares its findings with the published research on prediction SPI and assesses whether this comparison can be justified by the documentation of lead time and reference forecasts in the literature.

The informative comparators state the lead times directly [26-28]. The WPT-ANN combination achieved NSE of 0.82 at six months, yet the decomposition of the entire time series prior to splitting may introduce the leakage of future information, Quilty and Adamowski [29] discuss the leakage-free approach. No decomposition is applied in the present study.

As such, the smaller values observed herein relate to more stringent criteria: explicit horizons, naive benchmarks, significance testing per split, and the exclusion of the filled years. The published results without these restrictions cannot be regarded as being measured on the same scale.

2. STUDY AREA AND DATA

2.1 Study area

The district of Jaisalmer lies within the hyper-arid zone of the Thar Desert in western Rajasthan, India, roughly lying between 26.9°–28.1°N and 69.0°–72.5°E. According to the Köppen-Geiger climate classification, the area belongs to the BWh category, which refers to the hot desert climate. The research area consists of 55 grid points at a resolution of 0.25° and covers approximately 32,400 km². The rainfall is highly seasonal, mainly taking place between May and September. The clear seasonality of precipitation makes this area ideal for analysing the association between SPI accumulation scale, persistence, and prediction accuracy.

2.2 Variables and calendar reconstruction

The input dataset consists of monthly precipitation, SPI on four accumulation periods, three indices estimated from satellites, and calendar month. Used in the research variables and their functions are provided in Table 1.

Satellite indices include Vegetation Condition Index (VCI), Temperature Condition Index (TCI), and Vegetation Health Index (VHI). The formula of VHI implies its value as equally weighted mean of VCI and TCI as given in eq.(1)

$$\mathrm{VHI} = 0.5\ \mathrm{VCI} + 0.5\ \mathrm{TCI} \tag{1}$$

Cyclical nature of calendar month is maintained through sine and cosine transformations. Table 1 presents the variables used in the analysis, including their time-period and their uses. The precipitation source holds data for May-September only, while calculating the SPI index for long accumulation periods requires a full monthly precipitation series. For October-April, therefore, zero precipitation months are generated. The independent station workbook records zero mm precipitation for these months, and the two sources differ by a mean absolute difference of 0.154 mm over the overlapping months. This step is designed to generate the seasonal calendar through the supplementary station records, not to fill gaps in the precipitation data.

Table 1. Variables assembled for the study and their roles.

Variable	Description	Coverage	Role
Precipitation	Monthly precipitation total (mm), expressed as an anomaly for modelling	1991–2023, all reconstructed months	Model input
SPI-3, SPI-6, SPI-9, SPI-12	SPI calculated separately for each grid point and calendar month	1991–2023, subject to index availability	Forecast target and input
VCI	Vegetation Condition Index (0–100)	May–September	Evaluated as an input in sensitivity analysis
TCI	Temperature Condition Index (0–100)	May–September	Evaluated as an input in sensitivity analysis
VHI	VHI is calculated as 0.5 * VCI + 0.5 * TCI	May–September	Evaluated as an input in sensitivity analysis
Calendar month	Sine–cosine encoding of the calendar month	All months in the reconstructed calendar	Model input

2.3 Multi-scale SPI computation

The calculation of SPI is done according to McKee et al. [2]. For each accumulation scale k from 3 to 12 months, precipitation is summed up for k months before that, for each grid-point individually and for each month of the year:

$$g(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0 \tag{2}$$

where α is the shape parameter, β the scale parameter and $\Gamma(\alpha)$ the gamma function. Months of zero rainfall are frequent in this climate, so the cumulative distribution is mixed with the empirical probability q of zero accumulation:

$$H(x) = q + (1 - q)G(x) \tag{3}$$

where q is the empirical probability of zero accumulation and $G(x)$ is the fitted gamma cumulative distribution function. SPI is subsequently obtained by transforming the adjusted probability using the inverse standard normal distribution:

$$\mathrm{SPI} = \Phi{-}1(H(x)) \tag{4}$$

where Φ^{-1} denotes the inverse cumulative distribution function of the standard normal distribution. SPI cannot be calculated for a pair of a grid-point and a month of the year if the number of non-zero observation is less than 5. As a result, the SPI-3 can be undefined during parts of the dry season due to successive months without any precipitation. Gamma distribution is estimated based on the training part of each temporal validation split, and then those parameters are fixed for the corresponding validation and test splits [30-31].

Natural Resources for Human Health
eISSN: 2583-1194

3. METHODOLOGY

The detailed methodology is discussed in subsequent section (3.1-3.5) to achieve the contributions for the proposed work.

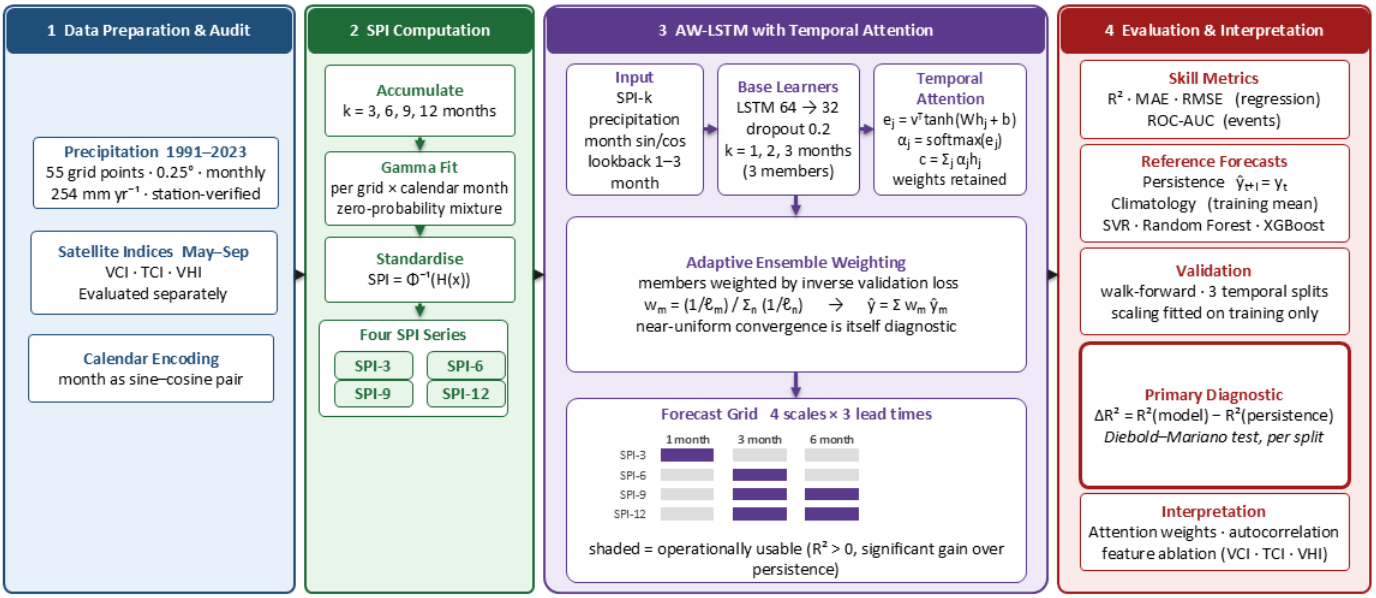


Figure 1. Methodological workflow for multi-scale SPI construction, AW-LSTM architecture, and persistence-relative evaluation.

3.1 Forecast formulation

Every SPI accumulation scale k, along with lead L in time, is considered a different prediction problem. In each grid cell, the aim is to predict SPI-k in L months from the forecast start time, giving twelve different scale–lead combinations. For each prediction time step, the primary model is fed four features: SPI-k, precipitation anomaly, and sin/cos encodings of the calendar month. The VCI, TCI, and VHI are considered separately in the sensitivity analysis due to their limited seasonality and range inconsistencies [32].

Input sliding windows are limited to the months from May to September during which vegetation indices calculated from satellites are available. This helped in achieving a longer lead forecast. For instance, a July forecast could be aimed at October (three months ahead) or January of the next year (six months ahead).

3.2 AW-LSTM architecture

The design of the AW-LSTM is made up of two LSTM layers [6] with an additive attention mechanism [11] which assigned an importance weight to each position in the input sequence window:

$$e_j = v^T \tanh(W h_j + b) \tag{5}$$

Scores are normalised by a SoftMax to weights summing to unity, and the context vector is their weighted sum of hidden states:

$$\alpha_j = \frac{\exp(e_j)}{\sum_{i=1}^{k} \exp(e_i)}, \qquad c = \sum_{j=1}^{k} \alpha_j h_j \tag{6}$$

The context vector is passed through the fully connected ReLU layer with dropout and the linear output. Three base learners are trained with lookback periods of one, two and three months to cover varying amount of temporal context and their outputs are weighted by the inverse validation loss:

$$w_m = \frac{\frac{1}{\ell_m}}{\sum_{n=1}^{M} \frac{1}{\ell_n}}, \qquad \hat{y} = \sum_{m=1}^{M} w_m \hat{y}_m \tag{7}$$

where ℓ_m is the mean squared error of member m on the validation set. Different members created for different windows of inputs give rise to different amounts of samples, and each forecast is treated according to its origin of forecast and grid point and not according to its position, with averaging being performed over samples that all members can create.

3.3 Objective function and input representation

According to the findings of ablation analysis described in Section 4.5, two decisions regarding design are made for the main experiments. The first one is that the loss function used is the Huber loss function instead of standard MSE loss function. The values of SPIs in this study range up to -8.65 which implies that very high impact can be created by such extreme data points while using squared error loss function. It is defined as

$$L_\delta(y,\hat{y}) = \begin{cases} \frac{1}{2}(y-\hat{y})^2, & |y-\hat{y}| \le \delta, \\ \delta|y-\hat{y}| - \frac{1}{2}\delta^2, & |y-\hat{y}| > \delta \end{cases} \tag{8}$$

where y and $\hat{y}$ denote the observed and predicted SPI values, respectively, and the threshold is set to $\delta = 1$.

The second difference is that the precipitation input variable is represented as the anomaly with respect to its monthly mean, instead of as its raw monthly total. The representation enabled the network to capture deviations of precipitation from the norm within the season without having to learn the entire annual precipitation cycle first.

3.4 Hyperparameter optimisation

The following seven hyperparameters are tuned with Optuna, using two LSTM layers, dropout rate, learning rate, batch size, attention dimension, and the width of the dense layer. Each scale-lead combination is optimised 15 times using the Huber loss objective function.

Optimisation of hyperparameters is done solely based on the validation set corresponding to each walk-forward split and test set data points did not enter the process and thus information leakages are avoided. Predictions for test periods are calculated by averaging outputs for five different random initialisations.

3.5 Baselines, validation and significance testing

The assessment of the model performance involved three walk-forward splits. The last year of the training split is 2000, 2003, and 2006 for the respective splits. In each case, the following three years are reserved for the validation stage while all the other observations are allocated for the test set. Parameters of the feature scaling and the SPI gamma distribution are estimated on the training split and used without changes in the validation and test splits.

Two reference forecasts are calculated using the same samples as AW-LSTM. The persistence forecast took the most recent observation of SPI-k and extended it to the target month. It is used as the primary reference forecast for measuring the added value of the learning model. The climatology reference forecast assumed predicting the mean SPI value from the training sample. Incremental performance relative to persistence is quantified as

$$\Delta R^2 = R^2_{\text{AW-LSTM}} - R^2_{\text{persistence}} \tag{9}$$

where positive values indicate that AW-LSTM achieved a higher coefficient of determination than persistence.

Drought detection is evaluated by defining drought as SPI < -1 and calculating the area under the receiver operating characteristic curve (ROC-AUC), following Wilks [33]. Model AUC is compared with persistence AUC computed on identical test samples.

Differences in squared forecast error between AW-LSTM and persistence are tested with the Diebold–Mariano test [25], using a Newey–West correction at multi-month lead times, separately on each walk-forward split. Counting the splits on which a gain is significant shows whether it holds across evaluation periods, which a single pooled test would hide.

A configuration is called regression-usable when it achieved a positive absolute R^2 and a significant improvement over persistence on all three walk-forward splits. This separates configurations with real improved accuracy from those whose strong absolute scores come mostly from the persistence of the target [34].

4. RESULTS

4.1 Temporal structure of the targets

Autocorrelation of the target SPI at forecast lead times of one, three, and six months, along with the percent of grid-months where each index is computed, is summarized in Table 2. Values are means over the 55 grid points The autocorrelation values increased with scale for all lead times. For lead time one month, this increase is from 0.709 for SPI-3 to 0.923 for SPI-12. For a three-month lead, these values varied from 0.086 to 0.748. These variations demonstrate the large differences in the inherent predictability of the target variable.

Table 2. Index coverage and autocorrelation at each forecast lag

Accumulation scale	Defined (% of grid months)	Lead time 1 month	Lead time 3 months	Lead time 6 months
SPI-3	58.3	0.709	0.086	0.096
SPI-6	82.1	0.825	0.612	0.122
SPI-9	96.8	0.838	0.633	0.367
SPI-12	95.6	0.923	0.748	0.488

4.2 Predictive accuracy across accumulation scales and lead times

Table 3 displays all twelve scale-lead combinations, averaged across three splits and five runs in which asterisk (*) indicates significant improvement over persistence and bold text rows, show configurations with reliable improvement, and Figure 2 summarizes them. One month ahead, AW-LSTM achieved R^2 of 0.669, 0.675 and 0.653 for SPI-3, SPI-6 and SPI-9, respectively, ROC-AUC being in the range of 0.906-0.931. On its own, it appears that these results represent consistent good performance at short leads for various accumulation scales.

Table 3. Accuracy scores for 12 configurations

Lead	Target	AW-LSTM R^2 ± Std. Deviation	Persistence R^2	ΔR^2 vs Persistence	ROC-AUC	Persistence AUC	ΔAUC
1	**SPI-3**	**0.669 ± 0.149**	**0.201**	**+0.468***	**0.906**	**0.809**	**+0.097**
1	SPI-6	0.675 ± 0.077	0.631	+0.044	0.931	0.945	−0.014
1	SPI-9	0.653 ± 0.065	0.620	+0.033	0.916	0.945	−0.029
1	SPI-12	0.330 ± 0.099	0.426	−0.097	0.870	0.916	−0.046
3	SPI-3	−0.360 ± 0.145	−1.584	+1.224*	0.523	0.534	−0.011
3	**SPI-6**	**0.601 ± 0.065**	**0.085**	**+0.516***	**0.884**	**0.863**	**+0.020**
3	SPI-9	0.465 ± 0.169	0.453	+0.011	0.895	0.920	−0.025
3	SPI-12	0.456 ± 0.079	0.432	+0.024	0.869	0.921	−0.053
6	SPI-3	−0.012 ± 0.026	−3.475	+3.463*	0.729	0.322	+0.407
6	SPI-6	−0.175 ± 0.170	−1.730	+1.554*	0.528	0.444	+0.084
6	SPI-9	0.281 ± 0.318	0.144	+0.137	0.820	0.840	−0.019
6	SPI-12	0.122 ± 0.131	0.369	−0.247	0.857	0.874	−0.017

Persistence changed this, in the case of one-month lead, persistence obtained R^2 of 0.631 for SPI-6, 0.620 for SPI-9 and 0.426 for SPI-12. Therefore, at SPI-6 and SPI-9 AW-LSTM improved the score by +0.044 and +0.033, correspondingly, while deteriorating the result in the case of SPI-12 ($\Delta R^2 = -0.097$). All these improvements are statistically significant only on one of three splits. In the case of SPI-3 with one-month lead, it is $\Delta R^2 = +0.468$ and significant on all three splits.

There are two combinations meeting the usability criterion: SPI-3 with one-month lead ($R^2 = 0.669$, $\Delta R^2 = +0.468$) and SPI-6 with three-month lead ($R^2 = 0.601$, $\Delta R^2 = +0.516$). They both exhibited a statistically significant improvement over persistence on all three splits.

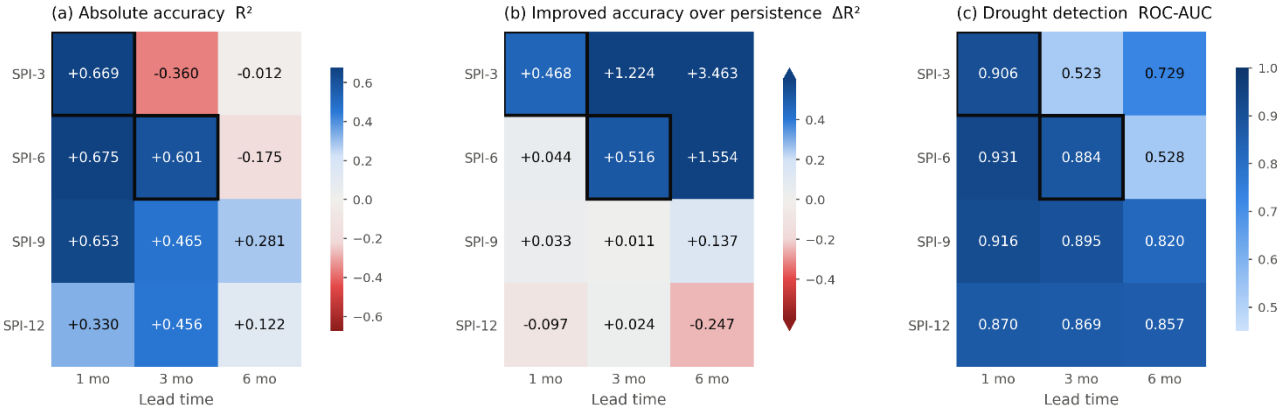


Figure 2. Accuracy performance showing absolute R^2, Improved accuracy relative to persistence, and drought-event detection. Highlighted configurations indicate those meeting the usability criterion.

Incremental ability must also be seen in relation to absolute ability. SPI-3 with six months' lead achieved $\Delta R^2 = +3.463$, which is the greatest increase in this experiment and significant in all three splits - although its absolute $R^2 = -0.012$ and ROC-AUC = 0.729. The increase is large since persistence drops to $R^2 = -3.475$, not due to good predictions by AW-LSTM.

4.3 Drought-event detection relative to persistence

Similar to accuracy, the ROC-AUC value by itself does not prove improvement in forecasting and must be compared to a persistence benchmark. Due to strong autocorrelation properties, the current value of the SPI variable can predict whether or not the target month is below the drought threshold. Hence, a high AUC does not imply that the model has contributed something to the improvement of predictions. Persistence AUC is estimated using the same test samples as AW-LSTM. It obtained 0.945 for SPI-6 and SPI-9 at one-month lead, 0.921 for SPI-12 at three-month lead, and 0.916 for SPI-12 at one-month lead. AWLSTM improved AUC in only four out of the twelve configurations. In five configurations where persistence AUC surpassed 0.90, AW-LSTM underperformed, with differences varying between -0.014 and -0.053. Figure 3 illustrates the ROC curves of the two configurations that satisfied the usability criterion. For SPI-3 at one-month lead, the test period AUC is 0.891 for AWLSTM and 0.799 for persistence, resulting in $\Delta AUC = +0.092$, and a bootstrap 95% CI of [+0.012, +0.145]. The non-inclusion of zero in the above CI shows that there is indeed an improvement in event detection. For the SPI-6 leading three months, the AUC difference is +0.033, which has a 95% confidence interval of [−0.015, +0.104]. This interval contains zero, and, consequently, cannot prove a statistically significant improvement in drought detection.

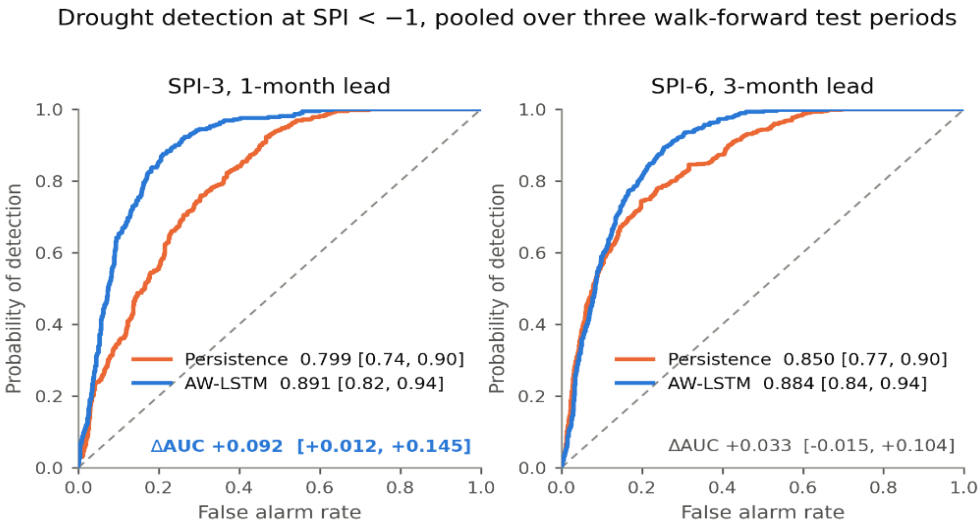


Figure 3. ROC curves for drought detection at SPI < −1 for SPI-3 at a one-month lead and SPI-6 at a three-month lead, comparing AW-LSTM with persistence

The discrepancy between the performance of regression and drought detection is critical. For instance, SPI-6 with the three-month lead improved $\Delta R^2 = +0.516$ and is statistically significant in the Diebold–Mariano test but had an insignificant gain of AUC. These two measures evaluate different aspects: Diebold–Mariano compares squared errors of the full probability distribution, while ROC-AUC estimates the ranking capability of the model – how well it separates drought months from non-drought months. The model may show an improvement in the former case without showing improvement in the latter.

The results for tables further elucidate the SPI-3 one-month configuration. There are 377 observed drought events among 2,802 testing samples. AW-LSTM flagged 188 of the drought events, while persistence recognized 243 droughts, which corresponds to probabilities of detection of 0.50 and 0.64, correspondingly. This trade-off needs to be considered consciously while evaluating its suitability in an early warning system, rather than using a simple discriminatory measure. The AUC values given in Table 4 represent averages of the split-wise values. In contrast, the AUC values in Figure 3, are computed after pooling the predictions from all three test periods. Table 4 reports the mean of the AUCs computed separately for each period. The two differ because the test periods differ in sample size and drought base rate, and pooling also compares cases across periods. For SPI-3 at a 1-month lead, Table 4 reports 0.906 for AW-LSTM and 0.809 for persistence, whereas the pooled values are 0.891 and 0.799, respectively. For SPI-6 at three months lead, they are 0.884 and 0.863, while pooled, 0.884 and 0.850, respectively. Both have been used since they serve different purposes.

4.4 Relationship between target autocorrelation and Improved accuracy

Figure 4 shows the relationship between accuracy and the autocorrelation of the target SPI at the corresponding lead time. For two configurations that passed the usability criterion, their autocorrelation ranged from about 0.61 to 0.71. On the contrary, for SPI-3 with leads of three and six months, the autocorrelation is 0.086 and 0.096, respectively, along with negative absolute R^2 value. At the other extreme, for SPI-12 at one-month lead, the autocorrelation is 0.923, while for AW-LSTM, it failed to outperform persistence.

Considering the twelve configurations examined above, the results imply that the improved accuracy lies within the intermediate range of target autocorrelation, ranging from 0.60 to 0.78. Only two configurations lay in the reported range and passed the usability criterion.

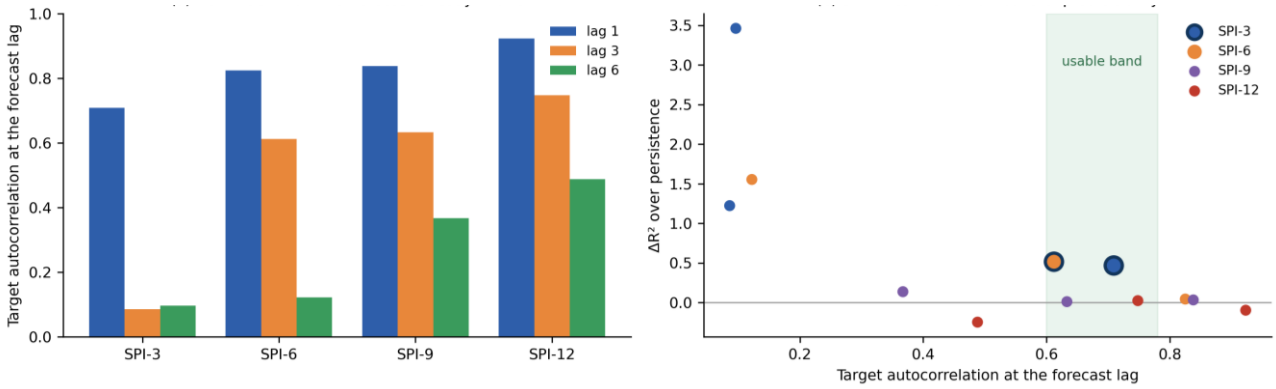


Figure 4. (a) Target autocorrelation at each forecast lag (b) Improved accuracy relative to persistence plotted against target autocorrelation (highlighted markers indicate the configurations meeting the regression-based usability criterion)

Target persistence, thus, partly accounts for the differences in improved accuracy across different configurations. This range is a pattern discovered in this dataset, not a general threshold; autocorrelation is not enough to claim that a target is predictable or to predict whether a certain model improves over persistence.

4.5 Ablation study

Nine different preprocessing and objective variations have been tested according to the same protocol with the use of identical data splits and hyperparameters except for one change each time. The assessment is concentrated on two variations that satisfied the regression-based criteria. Table 4 and Figure 5 give the details of the outcomes.

Table 4. Ablation over preprocessing and objective choices, evaluated on the two regression-usable configurations

Variant	R^2 with SPI-3 & Lead time at 1-month	ΔR^2	R^2 with SPI-6 & Lead time at 3-month	ΔR^2	Both improved (yes/no)
B Baseline	0.634	+0.433	0.494	+0.409	—
V1 log(1+precipitation)	0.601	+0.400	0.381	+0.296	no
V2 Four nearest-neighbour grids	0.600	+0.399	0.445	+0.360	no
V3 Residual target (y − persistence)	0.486	+0.285	0.363	+0.278	no
V4 Huber loss	0.644	+0.442	0.555	+0.469	yes
V5 Precipitation anomaly	0.648	+0.447	0.505	+0.420	yes
V6 V2 + V3	0.456	+0.255	0.342	+0.256	no
V7 V1 + V2 + V3	0.506	+0.305	0.492	+0.407	no
V8 V4 + V5	0.655	+0.453	0.575	+0.490	yes

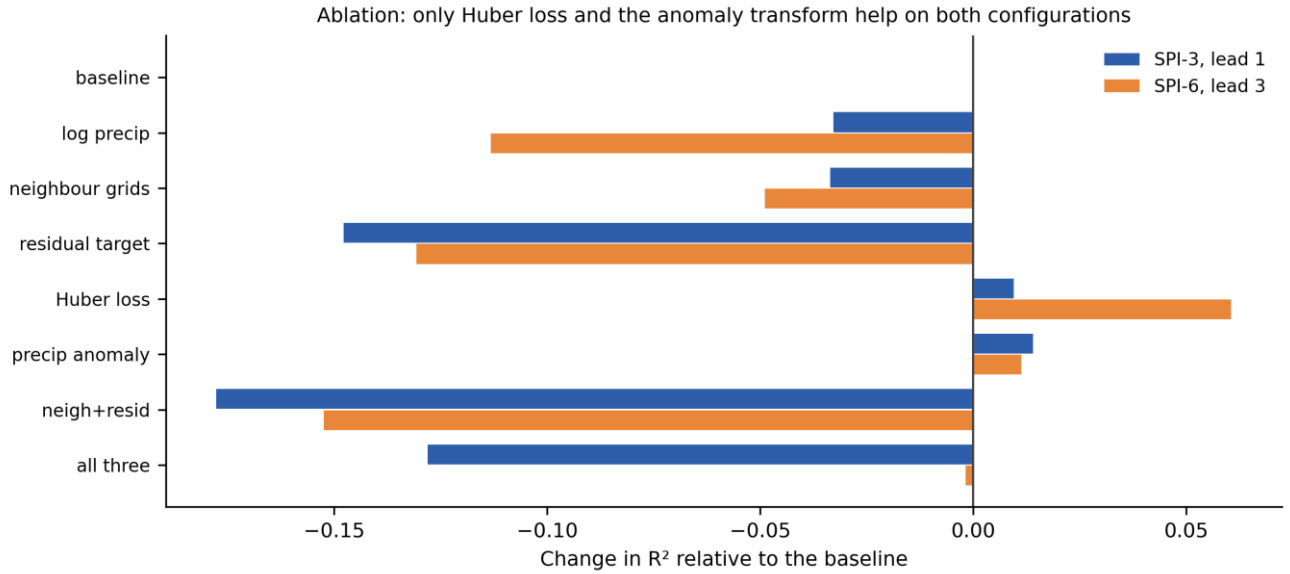


Figure 5. Change in R^2 relative to the baseline for each ablation variant, for both reference configurations.

Huber loss function and precipitation anomaly are the only individual variables that made both variations better. The Huber loss function helped increase R^2 from 0.634 to 0.644 for SPI-3 one month ahead and from 0.494 to 0.555 for SPI-6 three months ahead. Precipitation anomaly representation helped improve these indicators up to 0.648 and 0.505 correspondingly. This outcome is consistent with the purpose of both methods since the first one decreases the impact of outliers while the latter represents deviation from the seasonal average.

Adding the two together and retraining the model resulted in an R^2 of 0.669 for SPI-3 at one month out and an R^2 of 0.601 for SPI-6 at three months out. This represents absolute improvements of 0.035 and 0.107, respectively, above the baseline values. Accuracy improved for both models, and the model ultimately chosen remained statistically significant for all three splits.

Two other variations decreased performance. Including average SPI and precipitation from the four nearest-neighbour grid points brought R^2 down to 0.600 for SPI-3 at one month and to 0.445 for SPI-6 at three months out. This suggests that the formulation of the neighbour features did not improve the model. It does not mean that spatial information or neural graph architectures are necessarily useless.

Predicting the residual compared to persistence also harmed the performance, dropping R^2 to 0.486 and 0.363 for both configurations. In this experiment, learning the residual failed to enhance the baseline predictions. It appears that the residual structure used for learning is not a valuable prediction target using the present dataset and configuration of the model.

4.6 Optimised hyperparameters

The hyperparameters selected in Table 5 are presented for each configuration of scales and leads. The selected settings differed from one configuration to another and, in some instances, also between temporal splits. For instance, the learning rate selected for the predictions of SPI-12 with a one-month lead varied from 0.00013 to 0.0053 across three splits. These differences show that selected configurations are subject to the variance of validation sets.

As only a few years are observed, the validation sets are small and possibly contain variations of selected hyperparameters that could have resulted from sampling uncertainty. The configuration-specific values of hyperparameters are therefore presented rather than assuming that there is one optimum value for all configurations.

Table 5. Hyperparameters selected for each scale–lead configuration from fifteen trials scored on validation data

Target	Lead time	Window	LSTM-1	LSTM-2	Dropout	Learning rate	Batch Size	Attention	Dense layer width
SPI-3	1	3	64	16	0.091	0.0053	64	64	32
SPI-6	1	3	64	64	0.145	0.0045	256	32	16
SPI-9	1	3	64	64	0.221	0.0054	256	32	32
SPI-12	1	3	64	16	0.393	0.0077	64	64	32
SPI-3	3	3	64	16	0.091	0.0053	64	64	32
SPI-6	3	3	128	64	0.065	0.0015	128	32	16
SPI-9	3	3	64	16	0.061	0.0067	64	32	32
SPI-12	3	3	64	64	0.145	0.0045	256	32	16
SPI-3	6	1	32	16	0.389	0.0007	256	16	64
SPI-6	6	3	64	16	0.091	0.0053	64	64	32
SPI-9	6	3	64	16	0.083	0.0069	64	64	64
SPI-12	6	3	64	64	0.385	0.0028	64	32	16

5. DISCUSSION

Absolute accuracy and gradual improvement varied by configuration. While the R^2 values are high for short leads, the gains obtained through AW-LSTM over persistence are low at the higher SPI levels since the autocorrelation value at lag-one increased from 0.709 for SPI-3 to 0.923 for SPI-12. For configurations other than SPI-3 at a one-month lead and SPI-6 at a three-month lead, the usability requirement is not satisfied.

For the detection of events, AW-LSTM had ROC-AUC above 0.85 in eight configurations but gained over persistence in only four configurations. In the case of SPI-6 at a three-month lead, the regression performance improved without a marked improvement in AUC value, indicating the importance of separately evaluating continuous predictions and event detection tasks. From ablation studies, the use of the Huber loss function and precipitation anomalies proved to improve the reference configurations, while neighbour-grid and persistence residual targets worsened the results.

Advances in drought prediction using decomposition approaches, recurrent neural networks, and spatiotemporal deep learning have been made by recent works. Table 6 places the accuracy reported here beside four recent drought-prediction studies. For SPI at short lead times, the reported values cluster between roughly 0.57 and 0.73, and the present results, 0.669 and 0.601, sit inside that range. The agreement is worth noting but should not be over-read: the studies differ in index, accumulation scale, region and evaluation protocol; one forecasts the vegetation health index rather than SPI, and some report a single best station rather than a domain-wide value. The comparison is most useful for what the table does not contain. None of the four studies reports improved accuracy relative to a persistence forecast.

Table 6. Comparison with recent drought prediction studies

Ref.	Model	Reported accuracy	Lead Time
[35]	LSTM (SPI-1)	≤ 0.11 (metric not mentioned)	1 month
[36]	Gradient boosting; target is VHI, not SPI	R^2 0.6–0.75 across most grid cells at a 1-month lead; MAE 5–6%	1 month
[37]	Random Forest	R^2 0.57	1–3 months
[17]	Deformable convolution + Swin Transformer + LSTM	R^2 0.5875 (SPI-3), 0.7265 (SPI-6)	1–3 months
Proposed work	AW-LSTM with temporal attention; SPI-3, SPI-6, SPI-9 and SPI-12	R^2 = 0.669 (SPI-3, 1-month lead) and 0.601 (SPI-6, 3-month lead); ΔR^2 over persistence +0.468 and +0.516	1, 3, 6 months

The analysis is confined to Jaisalmer for 33 years, and satellite-based indices are considered only for May–September, considering the anomalies associated with the spatial extent of the indices. The analysis needs to be extended to longer time series, other locations, different neural network architecture, and error detection methods in future.

6. CONCLUSION

The present analysis of AW-LSTM included tests for four SPI accumulation scales with three lead times (1, 3, 6 months) utilizing walk-forward validation, baseline of persistence and climatology, and split-wise Diebold-Mariano tests. Only SPI-3 with a one-month lead time and SPI-6 with a three-month lead time reached the threshold of being usable for regression. While some combinations yielded an excellent ROC-AUC, only SPI-3 with a one-month lead time had a statistically significant increase in pooled AUC compared to persistence.

The Huber loss and precipitation anomaly inputs enhanced the reference combination performance, while the neighbour grid and persistence residual target variables failed to improve them. Multi-scale studies ought to include the absolute and persistence-based performance, significance, exact horizon points, temporal validation, and data screening.

Acknowledgements

The authors would like to acknowledge the support provided by their institution during this research

Author Contributions

Ritu Khandelwal: Data collection & analysis, conceptualization, methodology, modeling, writing original draft

Dr. Hemlata Goyal: Supervision, Correspondence, Review and editing

Prof. Rajveer Singh Shekhawat: Review and editing

Conflict of Interest

The authors declare no conflict of interest.

Funding

The authors state that no funds, grants, or other forms of support are received during the preparation of this manuscript.

Data Availability Statement

This study is based on previously published data and publicly available datasets, which are cited within the manuscript.

REFERENCES

[1] Osman, A.I.A., AlDahoul, N., Chong, K.L., Huang, Y.F., Ng, J.L., Elshafie, A., Sherif, M., Ahmed, A.N., 2025. A review on machine learning models for drought monitoring and prediction. *Climate Risk Management* 50, 100758. https://doi.org/10.1016/j.crm.2025.100758

[2] McKee, T.B., Doesken, N.J., Kleist, J., 1993. The relationship of drought frequency and duration to time scales, in: Proceedings of the 8th Conference on Applied Climatology. American Meteorological Society, Anaheim, CA, 17–22 January 1993, pp. 179–184.

[3] Labban, A.H., Ali, M., Diykh, M., 2026. Accurate multi-scaler SPI drought prediction by designing ORDWT-based signal decomposition and stacked LSTM. *Water* 18(14), 1686. https://doi.org/10.3390/w18141686

[4] Çitakoğlu, H., Aktürk, G., Demir, V., 2025. Hybrid machine learning for drought prediction at multiple time scales: a case study of Ağrı station, Türkiye. *Acta Geophysica* 73(2), 1643–1677. https://doi.org/10.1007/s11600-024-01501-5

[5] Yusuf, B., Shabri, A., Rezaiy, R., 2025. An innovative hybrid drought modelling approach using wavelet-long-short term memory (W-LSTM) with standardised precipitation index (SPI) data for Badeggi district, Niger State-Nigeria. *Discover Applied Sciences* 7(10), 1129. https://doi.org/10.1007/s42452-025-07649-z

[6] [6] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural Computation* 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[7] Taylan, E.D., 2024. An approach for future droughts in Northwest Türkiye: SPI and LSTM methods. *Sustainability* 16(16), 6905. https://doi.org/10.3390/su16166905

[8] Oruc, S., Hinis, M.A., Tugrul, T., 2024. Evaluating performances of LSTM, SVM, GPR, and RF for drought prediction in Norway: a wavelet decomposition approach on regional prediction. *Water* 16(23), 3465. https://doi.org/10.3390/w16233465

[9] Zhou, J., Fan, Y., Guan, Q., Feng, G., 2024. Research on drought monitoring based on deep learning: a case study of the Huang-Huai-Hai region in China. *Land* 13(5), 615. https://doi.org/10.3390/land13050615

[10] Afan, H.A., Almawla, A.S., Al-Hadeethi, B., Khaleel, F., AbdUlameer, A.H., Khan, M.M.H., Ma'arof, M.I.N., Kamel, A.H., 2024. LSTM model integrated remote sensing data for drought prediction: a study on climate change impacts on water availability in the arid region. *Water* 16(19), 2799. https://doi.org/10.3390/w16192799

[11] Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate, in 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA. arXiv:1409.0473.

[12] Dikshit, A., Pradhan, B., Assiri, M.E., Almazroui, M., Park, H.-J., 2022. Solving transparency in drought prediction using attention models. *Science of the Total Environment* 837, 155856. https://doi.org/10.1016/j.scitotenv.2022.155856

[13] Pathania, A., Gupta, V., 2025. Interpretable transformer model for national scale drought prediction: attention-driven insights across India. *Environmental Modelling & Software* 187, 106394. https://doi.org/10.1016/j.envsoft.2025.106394

[14] Talebi, H., Çitakoğlu, H., Samadianfard, S., Erol, A., 2026. Advanced hybrid machine learning for precise short-term drought prediction: a comparative study of SPI and SPEI indices in Iran's arid and semi-arid regions. *Pure and Applied Geophysics* 183(2), 793–819. https://doi.org/10.1007/s00024-025-03876-y

[15] Latifoğlu, L., Bayram, S., Aktürk, G., Çitakoğlu, H., 2024. Drought index time series prediction via three-in-one machine learning concept for the Euphrates basin. *Earth Science Informatics* 17(6), 5841–5898. https://doi.org/10.1007/s12145-02401471-8

[16] En-Nagre, K., Aqnouy, M., Ouarka, A., Naqvi, S.A.A., Bouizrou, I., Stitou El Messari, J.E., Tariq, A., Soufan, W., Li, W., El-Askary, H., 2024. Assessment and prediction of meteorological drought using machine learning algorithms and climate data. *Climate Risk Management* 45, 100630. https://doi.org/10.1016/j.crm.2024.100630

[17] Peng, H.Y., Wu, C., Du, Y., Liu, H., 2026. DCSwinLSTM for spatiotemporal meteorological drought prediction. *iScience* 29(6), 115902. https://doi.org/10.1016/j.isci.2026.115902

[18] Safari, M.J.S., Arashloo, S.R., Vaheddoost, B., 2025. Non-linear output structure learning: a novel multi-target technique for multi-station and multi-index drought modelling. *International Journal of Climatology* 45(14), e70105. https://doi.org/10.1002/joc.70105

[19] Verma, S., Karthikeyan, L., Ghosh, S., 2026. An explainable deep learning framework for sub-seasonal to seasonal agricultural drought prediction over India. *SSRN Electronic Journal* (preprint). https://doi.org/10.2139/ssrn.6303746

[20] Kogan, F.N., 1995. Application of vegetation index and brightness temperature for drought detection. *Advances in Space Research* 15(11), 91–100. https://doi.org/10.1016/0273-1177(95)00079-T

[21] Hameed, M.M., Mohd Razali, S.F., Wan Mohtar, W.H.M., Khaleel, F., AlOmar, M.K., Al-Suwaiyan, M.S., Yaseen, Z.M., 2025. Advancements in drought modeling: a comprehensive review of artificial intelligence and statistical models. *Environmental Earth Sciences* 84(16), 458. https://doi.org/10.1007/s12665-025-12432-9

[22] Sutanto, S.J., van der Weert, M., Blauhut, V., van Lanen, H.A.J., 2020. Skill of large-scale seasonal drought impact forecasts. *Natural Hazards and Earth System Sciences* 20(6), 1595–1608. https://doi.org/10.5194/nhess-20-1595-2020

[23] Belayneh, A., Adamowski, J., Khalil, B., Ozga-Zieliński, B., 2014. Long-term SPI drought prediction in the Aish River Basin in Ethiopia using wavelet neural network and wavelet support vector regression models. *Journal of Hydrology* 508, 418–429. https://doi.org/10.1016/j.jhydrol.2013.10.052

[24] International Research Institute for Climate and Society, Drought forecast and analysis tools for North America and the Globe. IRI Data Library Map Room, Columbia Climate School. https://iridl.ldeo.columbia.edu/maproom/Global/Drought (accessed 11 May 2026).

[25] Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* 13(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

[26] Lesinger, K., Tian, D., 2025. Skillful subseasonal soil moisture drought forecasts with deep learning-dynamic models. *Nature Communications* 16, 7461. https://doi.org/10.1038/s41467-025-62761-3

[27] Sibiya, S., Ramroop, S., Melesse, S., Mbatha, N., 2026. Meteorological drought trend analysis and prediction using a hybrid SG-CEEMDAN-ARIMA-LSTM model based on SPI from rain gauge data. *Natural Hazards and Earth System Sciences* 26(1), 315–342. https://doi.org/10.5194/nhess-26-315-2026

[28] Hinge, G., Piplodiya, J., Sharma, A., Hamouda, M.A., Mohamed, M.M., 2022. Evaluation of hybrid wavelet models for regional drought prediction. *Remote Sensing* 14(24), 6381. https://doi.org/10.3390/rs14246381

[29] Quilty, J., Adamowski, J., 2018. Addressing the incorrect usage of wavelet-based hydrological and water resources prediction models for real-world applications with best practices and a new prediction framework. *Journal of Hydrology* 563, 336–353. https://doi.org/10.1016/j.jhydrol.2018.05.003

[30] Khandelwal, R., Goyal, H., Shekhawat, R.S., 2023. Agricultural drought index selection using probability distribution: statistical and linear regression approach, in: 2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT). IEEE, pp. 1–6. https://doi.org/10.1109/CISCT57197.2023.10351409

[31] Khandelwal, R., Goyal, H., Shekhawat, R.S., 2023. Spatio-temporal standardized precipitation index selection using geospatial technique, in 2023 IEEE International Conference on Contemporary Computing and Communications (InC4), vol. 1. IEEE, pp. 1–5. https://doi.org/10.1109/InC457730.2023.10263047

[32] Khandelwal, R., Goyal, H., Shekhawat, R.S., 2022. Comparative analysis of machine learning techniques using predictive modeling. Recent Advances in Computer Science and Communications 15(3), 466–477. https://doi.org/10.2174/2666255813999200904164539

[33] Wilks, D.S., 2019. Statistical Methods in the Atmospheric Sciences, 4th ed. Elsevier, Amsterdam. ISBN 978-0-12-815823-4.

[34] Khandelwal, R., Goyal, H., Shekhawat, R.S., 2026. A systematic review of spatiotemporal deep learning approaches for drought monitoring and prediction using remote sensing data. *Recent Advances in Computer Science and Communications*, online ahead of print. https://doi.org/10.2174/0126662558510620260916054954

[35] Natarajan, N., Dharshini, S., & Vasudevan, M. (2026). Prediction of standardized precipitation index based on LSTM and variants of LSTM. *Physics and Chemistry of the Earth, Parts A/B/C*, 104422. https://doi.org/10.1016/j.pce.2026.104422

[36] Gallear, J.W., Galdos, M.V., Zeri, M., Hartley, A., 2025. Evaluation of machine learning approaches for large-scale agricultural drought forecasts to improve monitoring and preparedness in Brazil. *Natural Hazards and Earth System Sciences* 25, 1521–1541. https://doi.org/10.5194/nhess-25-1521-2025

[37] Jian, J., Mahmoudi, P., Jafari, P., Ghaemi, A., Yang, J., Firoozi, F., 2026. Short-term drought prediction in Iran using multisource machine learning: an assessment of autoregressive, teleconnection-driven, and hybrid paradigms. *Natural Hazards and Earth System Sciences* 26, 4407–4427. https://doi.org/10.5194/nhess-26-4407-2026