Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multi-Horizon SPI Drought Forecast Skill under Alternative Accumulation Protocols: A Rolling-Origin Benchmark for Bangladesh

Md. Shafiqul Islam
International University of Business Agriculture and Technology
Md Abu Jubayer Bijoy
2026-2-83-007@std.ewubd.edu

East West University

Research Article

Keywords: Standardized Precipitation Index, Meteorological drought, Forecast lead time, Accumulation protocol, Rolling-origin evaluation, Bangladesh, Accumulated precipitation

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11353622/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multi-Horizon SPI Drought Forecast Skill under Alternative Accumulation Protocols: A Rolling-Origin Benchmark for Bangladesh

Authors

Md. Shafiqul Islam

Associate Professor

Department of Environmental Science

International University of Business Agriculture and Technology

4 Embankment Drive Road, Uttara Model Town

Sector No. 10, Dhaka-1230, Bangladesh

Email: dislam.env@iubat.edu

Md Abu Jubayer Bijoy

Department of Mathematics and Data Science

East West University, Dhaka, Bangladesh

Email: 2026-2-83-007@std.ewubd.edu

ORCID: 0009-0000-4947-4037

Corresponding Author

Md Abu Jubayer Bijoy

Department of Mathematics and Data Science

East West University, Dhaka, Bangladesh

Email: 2026-2-83-007@std.ewubd.edu

Multi-Horizon SPI Drought Forecast Skill under Alternative Accumulation Protocols: A Rolling-Origin Benchmark for Bangladesh

Abstract

The Standardized Precipitation Index (SPI) is based on accumulated precipitation, but there is a difference between the scale of accumulation and the forecast lead time. When a forecast is issued at month t and verified at $t + h - 1$, each post-issue month during the target accumulation period must either be forecast or be provided via a specified alternative. Although the method for constructing the SPI from observed and forecast precipitation is well established in operational practice, its impact on comparisons of lead times across different models has rarely been quantified. We assess three accumulation methods; a full-trajectory forecast, a climatology-gap operational baseline, and a perfect-intermediate-information (PII) diagnostic. This study uses a dataset comprising 10 stations with complete records from Bangladesh, six expanding rolling-origin folds, ten forecasting techniques, SPI values of 3, 6 and 12 months, and lead times of 1, 3, 6, 9 and 12 months. Sensitivity analyses were carried out on a 34-station dataset with variable records and an empirical SPI. For the Gamma SPI-12, the difference between the PII and full-trajectory approaches increased with lead time and reaches +0.760 F1, +0.732 PR-AUC, and +1.873 R^2. No method performed best in all criteria and simple baselines of climatology and residual-trees were highly competitive. This shows that the accumulation protocol can qualitatively affect the conclusions about lead-time skill and model performance.

Keywords: Standardized Precipitation Index, Meteorological drought, Forecast lead time, Accumulation protocol, Rolling-origin evaluation, Bangladesh, Accumulated precipitation.

1 Introduction

Drought is one of the most expensive and hardest to predict of all natural hazards, with a chain of effects on agriculture, water security, food systems, and people's livelihoods (Wilhite et al., 2007; Mishra & Singh, 2010). In South Asia, Bangladesh is especially vulnerable due

to its agriculture which depends on the monsoon and the country's high population density. According to Monir et al. (2026) there are increasing trends in short-, medium-, and long-duration dry periods at 82.35%, 73.53%, and 50% of the stations examined, respectively. While Shahid (2010) found considerable spatial and seasonal variation in rainfall and SPI during wet and dry periods.

The Standardized Precipitation Index (SPI) (McKee et al., 1993) is the primary tool used worldwide for monitoring meteorological drought. It presents accumulated precipitation anomalies in terms of a standard-normal distribution. This standardization enables comparisons to be made across different accumulation periods and locations. The World Meteorological Organization (2012) and Hayes et al. (2011) have accepted the SPI as the global standard for monitoring meteorological drought because of its simplicity, its ability to be compared across regions, and its flexibility with respect to the accumulation period. According to McKee et al. (1993) and the World Meteorological Organization (2012), SPI values below -1.0 show moderate drought, those below -1.5 indicate severe drought, and values below -2.0 mean extreme drought.

In recent years there has been an increase in the use of machine learning and deep learning techniques for forecasting droughts. Aghelpour and others (2021) compared hybrid optimization methods with standalone drought-prediction methods and found the hybrid approach to be the best performing. Gyaneshwar et al. (2023) looked at deep learning drought models but did not concentrate specifically on the construction of accumulated drought indices over time. Osman et al. (2025) examined machine learning models used for drought monitoring and observed that SPI and SPEI are the most widely used indices and precipitation is the main input variable. Márquez-Grajales and others (2024) discovered that LSTM is the deep learning algorithm most frequently used for this purpose.

Many previous studies have involved a variety of different methods of constructing forecasts. In the case of operational studies, the forecast precipitation is usually combined with the observed antecedent data to produce the future SPI values (Cancelliere et al., 2007; Quan et al., 2012; Yoon et al., 2012; Yuan et al., 2013; Dutra et al., 2013, 2014; Mo and Lyon, 2015; Sutanto et al., 2020; Shin et al., 2020; Torres-Vázquez et al., 2024; Brands et al., 2025). On the other hand, some dynamical studies compute the SPI from the forecast precipitation for the target season rather than attaching the forecast to the observed antecedent trajectory (Mwangi et al., 2014; Ma et al., 2015; Lavaysse et al., 2020). Statistical and machine-learning studies can alternatively predict the precomputed SPI/SPEI value, the drought class, or the probability of

a class transition directly (Lyon et al., 2012; Belayneh et al., 2014; Bonaccorso et al., 2015; Behrangi et al., 2015; Liu et al., 2018; Ali et al., 2019; Park et al., 2020; Felsche and Ludwig, 2021; Karbasi et al., 2023; Yu et al., 2024; Lee et al., 2025). Since these represent different types of forecasting tasks, the skill reported should not be considered directly comparable unless the forecast construction, the definition of lead time, and the index-fitting procedure are taken into account. The current study does not introduce the observed–forecast partition; instead, it quantifies how alternative accumulation protocols affect reported skill across ten models under identical folds, targets, and metrics.

The SPI is a measure based on accumulated precipitation; in order to calculate a 12-month SPI, one needs precipitation data for 12 months. If a forecast is produced at time t for a period ending at time $t+h-1$, then the forecaster has only observed precipitation up to $t-1$; therefore all the months from t through $t+h-1$ must either be predicted or be replaced by some operational estimate. The level of detail given in published reports regarding the accumulation scale, the forecast lead time and the method used for the intermediate months differs from report to report. This variation in the way the reports present this information makes it difficult to carry out consistent comparisons of operational skill.

Within the wider body of literature on machine learning, considerable attention has been given to temporal information boundaries. Mammas and Lekkas (2021) showed that there is a bias in the characterization of droughts and contamination of the evaluation when SPI formulations are fitted using the entire record prior to out-of-sample validation, which constitutes a type of parameter-estimation bias. Kapoor and Narayanan (2023) also stressed the need for systematic consideration of leakage and the reproducibility crisis in machine-learning based prediction with regard to information boundaries. However, these studies mainly concerned themselves with parameter estimation in the computation of the SPI rather than with the evaluation-protocol sensitivity that is addressed in this paper.

The study tackles this gap by systematically assessing multi-horizon SPI forecasts under three different protocols, each of which deals with the unknown part of the SPI accumulation window in a distinct way. The full-trajectory protocol makes use of the observed rainfall up to the date of issuance and then predicts the rest with the model. The climatology-gap operational baseline fills in the missing months with climatological averages and thus provides a baseline that is valid for operational use. The perfect-intermediate-information (PII) diagnostic applies the actual rainfall to the intermediate future months but still keeps the model forecast for the

target month. The PII diagnostic is a deliberately non-operational counterfactual designed to show how much the reported skill can change when realized intermediate information is taken into account. In the current experiment it acts as an optimistic diagnostic for the various leads and metrics, but it is not a guaranteed upper limit for every metric, station, or fold.

This paper makes four contributions. The first is to offer a rolling-origin benchmark for SPI-w forecasts produced at lead h, the evaluation being carried out using the same set of folds, targets and metrics. The second is to measure protocol sensitivity by means of a controlled comparison between the full-trajectory, climatology-gap and PII constructions, all using the same predictions, targets, folds and metrics. Thirdly, it presents a baseline-aware assessment of the models through a pairwise comparison of neural, tree-based, naïve, fixed-climatology and adaptive-climatology forecasts across continuous, ranking, threshold and event-level metrics. Fourthly, it supplies a structured audit and checklist concerning the way in which relevant studies report on forecast origin, target date, accumulation construction, treatment of the intermediate month and the SPI fitting period.

2 Related Work

2.1 The monitoring of SPI and the preparation of the forecast

The Standardized Precipitation Index (SPI) was developed by McKee et al. (1993) as a means of providing a standardized method for measuring precipitation anomalies over a variety of time scales and at different locations. It is obtained by fitting a probability distribution to the historical precipitation data and then transforming the cumulative precipitation amount for a given accumulation period into a standard normal distribution. The World Meteorological Organization (2012) adopted the SPI as the primary index for monitoring meteorological drought on a global basis and recommended that it should be used in conjunction with other indices if a full assessment of drought is to be made. Guttman (1999) gave a detailed algorithm for calculating the SPI based on the different methods of distribution fitting and parameter estimation.

The SPI has a number of advantages; one of them is that it requires only precipitation data, which are generally available and reliable. Another advantage is that it can be used to calculate drought for any length of accumulation period, from short-term (agricultural) droughts lasting one to three months, through medium-term droughts lasting six to nine months, to long-term (hydrological) droughts lasting twelve to twenty-four months (McKee et al., 1993; World

Meteorological Organization, 2012). A third advantage is that the SPI is standardized, so that it can be compared between regions having different climatic conditions. Lastly, the SPI's standardized probabilistic scale enables the identification and comparison of the time at which drought begins, its duration and how severe it is.

By means of the SPI drought classification technique (McKee et al., 1993), the various drought categories are defined on the basis of the SPI values: moderate drought corresponds to the range from -1.0 to -1.49, severe drought to -1.5 to -1.99, and extreme drought to -2.0 or below. A drought event is said to begin when the SPI falls below -1.0 and ends when the SPI becomes positive once again. This method has been extensively applied in drought monitoring and early warning systems.

Although it is widely used, conventional SPI has some limitations. It frequently attempts to fit a gamma distribution to the accumulated precipitation for each month, a method which may not be valid for all regions and time scales (Stagge et al., 2015). Furthermore, the SPI does not consider temperature or evapotranspiration factors, both of which can be important when evaluating agricultural drought (Vicente-Serrano et al., 2010). To remedy this drawback, the Standardized Precipitation Evapotranspiration Index (SPEI) was developed by incorporating potential evapotranspiration (Vicente-Serrano et al., 2010). Yet, the SPI remains the most commonly used index for monitoring meteorological drought due to its simplicity and effectiveness (Osman et al., 2025; Gyaneshwar et al., 2023).

2.2 The use of machine learning in predicting SPI events and droughts

In recent years there has been an increasing use of machine learning techniques in the field of drought forecasting. According to a thorough review of machine learning models for drought monitoring and forecasting carried out by Osman et al. (2025), the most frequently used algorithms are artificial neural networks (ANN), support vector machines (SVM), random forest (RF), and long short-term memory (LSTM) networks. The review also mentioned that SPI and SPEI are the most commonly predicted metrics.

Márquez-Grajales and colleagues in 2024 looked at drought prediction through the use of deep learning models and discovered that the LSTM model was the one most frequently used. They noted that hybrid models, which involve combining wavelet transforms with LSTM (W-LSTM) and other deep learning techniques, have shown promising results in the field of drought forecasting. Gyaneshwar and co-authors also carried out a study on deep learning models applied

to drought prediction. Furthermore, Hao and others conducted a review of seasonal drought prediction and looked at the advances, the challenges, and the future prospects.

In Bangladesh several studies have applied machine learning to drought forecasting. Awal and Al Mamun (2022) developed a wavelet artificial neural network to predict long-term meteorological droughts in the north-western part of Bangladesh. Hossain and others (2025) used an ensemble learning method for the analysis and forecasting of droughts in central Bangladesh. Al Mamun and colleagues (2024) identified the weather parameters that are important for seasonal drought prediction in Bangladesh by means of machine learning. Sarkar and co-authors (2025) applied SPI-gamma random forest modelling for the description and prediction of meteorological drought in the Bengal Delta area. Sadiq and others (2023) carried out an assessment of meteorological drought in northern Bangladesh using a machine learning-based approach that considered remote sensing metrics.

Recent studies have placed a greater emphasis on multi-output forecasting and on forecasts that are operationally explicit. He et al. (2024) created a direct multi-step method for forecasting monthly rainfall and supplied a methodological procedure for forecasting complete future rainfall trajectories. Brust et al. (2021) developed DroughtCast to enable multi-lead predictions of the categories employed in the United States Drought Monitor; their work highlighted the need to evaluate drought forecasts at specifically defined issue dates and lead times. More recently, Magallanes-Quintanar et al. (2026) have compared different transformer architectures when carrying out long-horizon SPI-12 forecasting, indicating that interest in transformer-based drought prediction is increasing.

2.3 The boundaries of temporal information and retrospective preprocessing

It is necessary to assess forecasts if we are to judge their operational value. Tashman (2000) carried out a comprehensive review of measures of forecasting accuracy and of the different evaluation techniques, one of which is rolling-origin evaluation. Rolling-origin evaluation involves sequentially updating the forecast origin and then preparing forecasts based on each origin, thus providing a more reliable assessment of forecast performance than simple holdout validation (Tashman, 2000). Cerqueira et al. (2020) likewise carried out a review of the various methods for evaluating the performance of time series forecasting models.

The importance of information boundaries when assessing statistical and machine learning forecasts should not be overlooked. In the case of time-series forecasting, look-ahead contami-

nation may result from using future data for scaling or normalisation, from employing statistics calculated on the evaluation dataset for feature engineering, or from including predictors that would not be accessible at the time a forecast is issued (Kaufman et al., 2012). Mammas and Lekkas (2021) showed that estimating the SPI distribution parameters from the full record prior to out-of-sample validation can lead to a bias in the characterisation of droughts and cause contamination of the evaluation. Kapoor and Narayanan (2023) stressed that reproducibility in machine-learning-based science relies on giving systematic attention to information boundaries.

When it comes to drought forecasting based on the SPI, there are two different retrospective options. One of them is to estimate the parameters of the SPI distribution using all the available data—not just that from the training subset—rather than merely the training data. This represents a form of retrospective fitting and the extent of its impact will vary according to the length of the sample, the extent of station coverage, and the distributional assumptions; it cannot be assumed to cause a large bias in all cases. The other aspect is that the way in which the target accumulation window is constructed at the time the forecast is issued can differ from study to study, and this difference has not yet been systematically measured against a common benchmark. It is this second issue that is the focus of the present study.

2.4 Evaluation Gaps

The way that the published SPI forecasting reports explain the difference between accumulation scale and forecast lead differs from report to report, as well as the way they handle the unknown intermediate months. Because of this, it is not always simple to compare the amount of operational information contained in the SPI forecasts reported in different studies.

The study addresses these gaps by establishing a common standard for evaluating multi-horizon SPI forecasts that considers the information available at the time the forecasts are issued. By comparing three different evaluation methods, it assesses the degree of skill differences caused by the protocols and the way in which these differences depend on the fraction of the SPI window. It also looks at how various models perform when evaluated using the same protocol in order to determine whether greater architectural complexity provides any advantage over simpler causal and statistical baseline models.

3 Study Area and Data

3.1 Study Area

Bangladesh is located in the Bengal Delta and has a tropical monsoon climate characterised by a wet season (from June to September) and a dry season (from November to February). About two-thirds of its land area is situated at a height of less than 10 metres above sea level. There is a large variation in rainfall both in space and in time, the annual amount ranging from less than 1,500 mm in the west to more than 3,000 mm in the northeast (Shahid, 2010).

Drought is a common hazard in Bangladesh and has a major effect on agriculture, water resources and food security; it mostly takes place during the pre-monsoon period (March–May) and the post-monsoon period (October–November) since the rainfall is low and the temperature is high (Habiba & Shaw, 2013). In the wider Indo-Bangladesh Bengal Delta, Sarkar et al. (2025) found a great deal of variation in the severity of drought, there being prolonged periods of extreme dryness in western West Bengal and significant historical SPI anomalies at stations in Bangladesh.

The meteorological stations extend across the main climatic areas of Bangladesh, including the dry area in the north-west, the transitional area in the centre, the wet area in the north-east, the coastal area in the south-west, and the hilly region in the south-east. There were 34 meteorological stations altogether for which data were adequate for analysis.

3.2 Data

We used a publicly distributed compilation of monthly Bangladesh meteorological-station records attributed to the Bangladesh Meteorological Department. The dataset covers 1948–2013 (66 years) of monthly observations and includes the following variables: maximum temperature (Tmax, °C), minimum temperature (Tmin, °C), relative humidity (RH, %), wind speed (m/s), cloud coverage (oktas), bright sunshine (hours), and rainfall (mm).

At the beginning there were 35 stations. The Chittagong City Ambagan station was removed since it had less than 200 valid months, which left 34 stations for analysis. Each station has its coordinates (latitude, longitude) and elevation.

We conducted a coverage and consistency audit, including station record length, missingness, temporal completeness, and repeated-value patterns. Only 10 stations had complete records for the whole of the 1948–2013 period (792 months) and the other 24 stations had between 276

and 780 valid months. The coverage statistics for the stations are given in Table 1. Because source-level quality-control metadata were unavailable, this audit should not be interpreted as a full climatological homogenization assessment.

Table 1: Station coverage statistics

Category	Number of stations	Median months	Min months	Max months
Complete (1948–2013)	10	792	792	792
Eligible but non-complete	24	576	276	780
Total eligible	34	660	276	792

The public compilation does not provide complete historical metadata on station homogenization, quality-control interventions, source-level imputation, or station relocation. Repeated early values at some stations suggest that upstream preprocessing may have occurred. Consequently, the present analysis cannot independently reconstruct all source-level quality-control decisions.

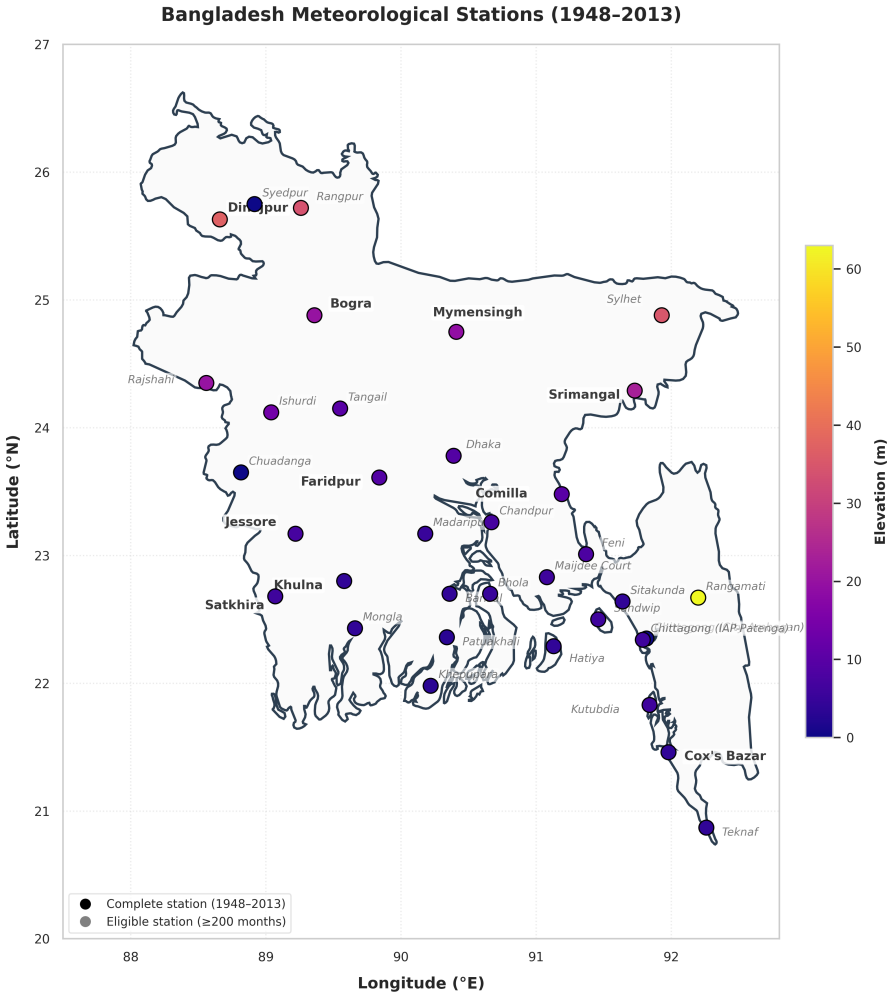


Figure 1: Study area map and station coverage.

3.3 Preprocessing

Each sample used in the supervised learning process included the previous 24 months' data for seven meteorological variables along with sine and cosine transformations of the calendar month. The aim is to predict the entire future rainfall pattern over the following 12 months.

The fixed-network primary panel consisted of the 10 complete-record stations and the 34-station variable-record panel was employed as a means of carrying out a sensitivity analysis. In each sensitivity fold a station was accepted only if its 24-month input window and 12-month target trajectory were fully available within the fold. Backward filling of any values did not take place before the station started to operate.

Within each outer fold, the seven input variables, including the untransformed rainfall history, were standardized using means and standard deviations estimated by pooling station-month observations from the training partition only. The 12-month rainfall target was transformed using $\log(1 + P)$ and then standardized using the training-partition mean and standard deviation of log-rainfall. The fitted input and target transformations were applied unchanged to validation and test samples. Forecasts were transformed back to rainfall units before accumulation and SPI calculation. Since the direct trajectory output was available, it was possible to assess the SPI-3, SPI-6 and SPI-12 values at lead times of 1, 3, 6, 9 and 12 months without a need to insert observations from a future unknown period.

The data were split using six expanding-window folds:

- Fold 1: Train (1948–1992), Val (1993–1995), Test (1996–1998)
- Fold 2: Train (1948–1995), Val (1996–1998), Test (1999–2001)
- Fold 3: Train (1948–1998), Val (1999–2001), Test (2002–2004)
- Fold 4: Train (1948–2001), Val (2002–2004), Test (2005–2007)
- Fold 5: Train (1948–2004), Val (2005–2007), Test (2008–2010)
- Fold 6: Train (1948–2007), Val (2008–2010), Test (2011–2013)

A sample was only used if its full 12-month forecast path was contained within the relevant three-year test period. As a result, each station with a complete record provided 25 issue dates per fold. The input history was allowed to extend into the previous training or validation period since the observations from those periods would be available on the issue date.

The long-term monthly averages were calculated for each station and month using only the training period, and this climatology was then used as a benchmark as well as forming part of the climatology gap protocol.

4 Methodology

4.1 Problem Formulation

Let $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$ be the set of stations, where $N = 10$ for the fixed complete-record primary panel and $N \leq 34$ for the availability-filtered sensitivity panel. For station s and month t, we obtained a vector $\mathbf{x}_t^{(s)} \in \mathbb{R}^7$ of seven meteorological variables. Each sample was associated with a station identifier. Neural models represented this identifier through an eight-dimensional learned station embedding, while Residual Extra Trees used one-hot station indicators. Geographic coordinates and elevation were retained for station description but were not used in the final forecasting models for prediction.

Neural models were global station-conditioned forecasters: model weights were shared across stations, while learned station embeddings captured station identity. No inter-station graph or spatial message-passing operation was used.

The input consists of the 24 months immediately preceding the issue month t:

$$\mathbf{x}_{t-24:t-1}^{(s)} = \left\{\mathbf{x}_{t-24}^{(s)}, \mathbf{x}_{t-23}^{(s)}, \ldots, \mathbf{x}_{t-1}^{(s)}\right\} \tag{1}$$

The output is the complete 12-month future rainfall trajectory:

$$\widehat{\mathbf{P}}_{t:t+11}^{(s)} = F_\theta\left(\mathbf{x}_{t-24:t-1}^{(s)}, \mathbf{e}_s, \mathbf{c}_t\right) \tag{2}$$

where $\mathbf{e}_s$ is the station embedding, $\mathbf{c}_t$ contains calendar sine/cosine features, and t is the first unknown month. This direct 12-month trajectory makes evaluation at leads $h \in \{1, 3, 6, 9, 12\}$ possible without re-estimating for each lead.

4.2 SPI Computation

This paper computes SPI following the procedure from McKee et al. (1993) and Guttman (1999). For each station, calendar month, and accumulation window $w \in \{3, 6, 12\}$, the procedure is:

1. Accumulate rainfall over the current month and the previous $w-1$ months.

2. For each station and calendar month, fit a mixed discrete-continuous distribution to the accumulated training-period values. The discrete component is estimated as

$$q_{s,m,w} = \Pr\left(A_{s,m,w} \leq 0.1 \text{ mm}\right) \tag{3}$$

and a Gamma distribution is fitted to values exceeding 0.1 mm.

3. Transform the accumulated value to a standard normal quantile using the piecewise function

$$H_{s,m,w}(A) = \begin{cases} \dfrac{q_{s,m,w}}{2}, & A \leq 0.1 \text{ mm}, \\ q_{s,m,w} + (1 - q_{s,m,w}) G(A; \alpha_{s,m,w}, \beta_{s,m,w}), & A > 0.1 \text{ mm}, \end{cases} \tag{4}$$

where G is the fitted Gamma CDF with shape α and scale β. Then

$$\text{SPI}_{s,m,w} = \Phi^{-1}\Big[\text{clip}\Big(H_{s,m,w}(A), 10^{-6}, 1 - 10^{-6}\Big)\Big]. \tag{5}$$

The available sample size varied by fold and station. A station–calendar-month SPI fit was retained only when at least 20 valid accumulated precipitation observations were available in the corresponding training partition. Station–calendar-month combinations that did not meet this threshold were excluded from the SPI evaluation for that fold and station rather than substituted by an alternative distribution. This rule was applied identically to the primary and sensitivity panels.

For the empirical sensitivity analysis, let $R_{\leq}(A) = \#\{A_i \leq A\}$ denote the right-sided insertion rank of the out-of-sample accumulated value A relative to the sorted training sample $\{A_i\}_{i=1}^{n}$. The empirical plotting position used in the implementation was

$$\widehat{F}(A) = \frac{R_{\leq}(A) - 0.44}{n + 0.12}, \tag{6}$$

followed by

$$\text{SPI}_{\text{emp}} = \Phi^{-1}\Big[\text{clip}\Big(\widehat{F}(A), 10^{-6}, 1 - 10^{-6}\Big)\Big]. \tag{7}$$

Ties were therefore handled using the right-sided insertion rank. All reference samples were estimated from the corresponding outer-fold training partition only.

4.3 Evaluation Protocols

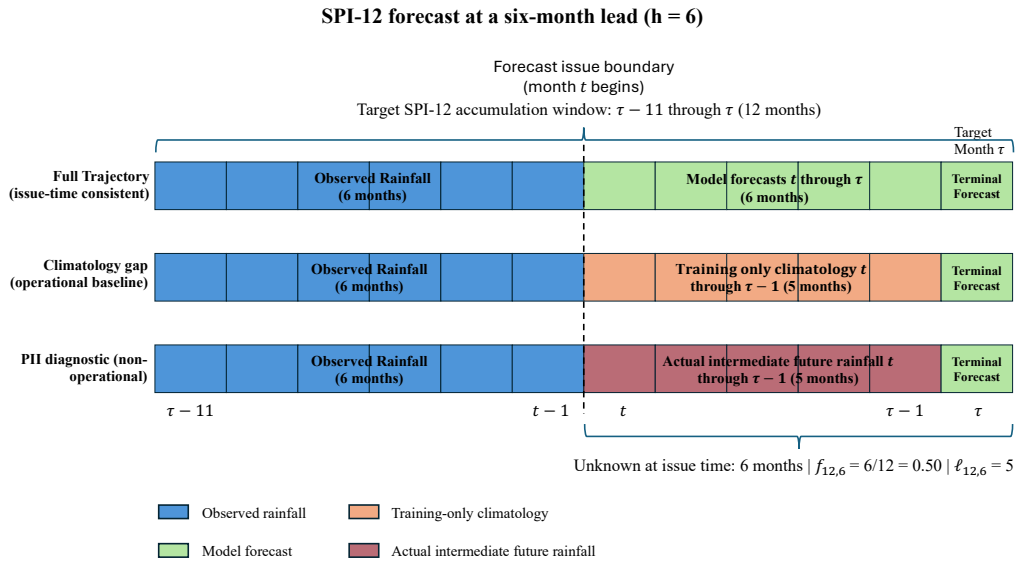


Figure 2: Information boundary for the forecast-issue time and the three evaluation protocols.

Let t be the forecast issue month, $\tau = t+h-1$ the target month, w the SPI accumulation scale, and $a = \tau - w + 1$ the beginning of the target SPI window. The target accumulation window is

$$W_{w,\tau} = \{a, \ldots, \tau\}. \tag{8}$$

The portion genuinely observed at forecast issuance is

$$O_{w,h,t} = \{i : a \le i \le \min(t-1, \tau)\}, \tag{9}$$

and the unknown portion is

$$U_{w,h,t} = \{i : \max(t, a) \le i \le \tau\}. \tag{10}$$

Sums over empty index sets are treated as zero throughout.

Full-trajectory forecast. The accumulated rainfall is the sum of observed rainfall for all months up to the issue date plus the forecast for all months in the unknown portion:

$$\widehat{A}^{\text{traj}}_{w,\tau|t} = \sum_{i=a}^{\min(t-1,\tau)} P_i + \sum_{i=\max(t,a)}^{\tau} \widehat{P}_{i|t}. \tag{11}$$

Climatology-gap operational baseline. Unknown months before the target month are replaced by climatological averages, and only the terminal month is forecast:

$$\widehat{A}^{\text{gap}}_{w,\tau|t} = \sum_{i=a}^{\min(t-1,\tau)} P_i + \sum_{i=\max(t,a)}^{\tau-1} C_{s,m(i)} + \widehat{P}_{\tau|t}. \tag{12}$$

Perfect-intermediate-information (PII) diagnostic. The diagnostic uses actual rainfall for intermediate future months but retains the model forecast for the terminal target month:

$$\widehat{A}^{\text{PII}}_{w,\tau|t} = \sum_{i=a}^{\min(t-1,\tau)} P_i + \sum_{i=\max(t,a)}^{\tau-1} P_i + \widehat{P}_{\tau|t}. \tag{13}$$

The terminal month is always forecast in all protocols. The PII diagnostic is a deliberately non-operational counterfactual that provides a diagnostic of how strongly realized intermediate information can alter reported skill. It is not proposed as a deployable forecasting method.

4.4 Forecasted Fraction

For a given accumulation window w and forecast lead h, we define the forecasted fraction as

$$f_{w,h} = \frac{\min(h, w)}{w} = \min\left(1, \frac{h}{w}\right). \tag{14}$$

The number of intermediate months that would be inserted under the PII diagnostic is

$$\ell_{w,h} = \min(h - 1, w - 1). \tag{15}$$

At short leads relative to the accumulation window (e.g., $h = 1$, $w = 12$), f is small and most of the window is already observed. At long leads (e.g., $h = 12$, $w = 12$), $f = 1$ and the entire window must be forecast.

We use $f_{w,h}$ as a descriptive measure of how much of the target accumulation window depends on post-issue information. We examine empirically whether the PII–full-trajectory contrast varies with $f_{w,h}$; monotonicity is not assumed because forecast-error magnitude and covariance can vary by lead, season, and model.

4.5 Evaluation Metrics

The principal summary ranking criterion was mean PR-AUC under the complete panel Gamma-SPI full-trajectory protocol, equally averaged across SPI-3, SPI-6 and SPI-12 and leads of 1, 3, 6, 9 and 12 months.

For continuous SPI prediction, we report R^2 (coefficient of determination), MAE (mean absolute error) and RMSE (root mean squared error). R^2 is the coefficient of determination $1 - \mathrm{SSE}/\mathrm{SST}$, not the squared Pearson correlation. Negative values are retained and indicate performance worse than predicting the evaluation-set mean.

For binary drought detection at SPI < -1, we report precision, recall, F1, PR-AUC (Saito & Rehmsmeier, 2015) and ROC-AUC. PR-AUC and ROC-AUC were computed from a monotonic drought-risk score derived from predicted SPI:

$$r(\widehat{\mathrm{SPI}}) = \frac{1}{1 + \exp\left\{4\left(\widehat{\mathrm{SPI}} + 1\right)\right\}}. \tag{16}$$

Lower predicted SPI produces a higher risk score, as required for a drought-risk ranking.

For each combination of SPI scale and lead time, the metrics were calculated by combining the predictions from eligible station-month entries across different stations and issue dates. The model-level summaries were then obtained by taking the equal average of the 15 combinations of SPI scale and lead time. These summaries are macro-averages over the tasks and not averages of the pooled metrics across all tasks.

4.6 Event-Level Evaluation

An episode of drought threshold was defined as one or more consecutive evaluated target months during which the SPI was less than -1. The episode ended when the SPI had returned to being greater than or equal to -1. A predicted episode was regarded as detected if it shared at least one month with a true episode. The onset error was calculated by subtracting the observed onset month from the predicted onset month. Positive values meant that the detection had occurred later than it should have. False warning episodes were those predicted episodes that did not overlap with any true episode. Event-level results are reported for AC-LSTM as a case study rather than for all ten models; they should not be generalized to the full benchmark. False-warning counts are normalized per 10 station-years.

4.7 Compared Models

The primary 10-station panel comparison includes the following models.

Fixed Climatology predicts the training-period station- and calendar-month mean for each future month and serves as a simple causal seasonal baseline.

Adaptive Climatology takes the average of the most recent k observations from the same month, with $k \in \{10, 20, 30, 40, 50\}$ chosen independently in each fold based on the validation trajectory error; the validation trajectory criterion is a combination of the normalized monthly rainfall MAE and the normalized accumulated errors at 3, 6 and 12 months.

Seasonal Naive uses the value observed in the same month of the previous year.

Residual Extra Trees is an ensemble of decision trees designed to predict the residuals with respect to the climatological baseline. The model used the flattened 24-month standardized meteorological history, the calendar encodings and the one-hot encoded station identity to predict the log-rainfall residuals relative to a fixed climatology. The minimum leaf size was chosen from $\{5, 10, 20\}$; residual shrinkage was selected from $\{0, 0.1, 0.25, 0.5\}$. The choice was based on the validation trajectory error; the final model was then refitted on the combined training and validation data. During validation, 250 trees were used; in the final refit 350 trees were used with max_features $= 0.7$, random_state $= 123$ and n_jobs $= -1$.

DLinear is a linear model with seasonal and trend decomposition (Zeng et al., 2023); it decomposes the input into trend and seasonal components by average pooling and then applies linear layers to each.

TCN is a temporal convolutional network using dilated convolutions to capture long-range dependencies (Bai et al., 2018); it has four layers with dilations $1, 2, 4$ and 8.

iTransformer-Lite is a lightweight architecture inspired by iTransformer, which applies attention to the feature dimensions rather than to the time steps (Liu et al., 2024).

TFT-Lite is a lightweight model inspired by TFT that combines LSTM, self-attention, and gating mechanisms (Lim et al., 2021).

SMA-TFT-Trajectory is a version of TFT-Lite with a learnable set of 12 seasonal prototypes—one for each calendar month—retrieved using a gated mechanism.

AC-LSTM is a two-layer LSTM with an accumulation loss added to encourage the model to match both monthly and SPI-accumulated rainfall. The loss for monthly rainfall is

$$\mathcal{L}_{\text{monthly}} = \text{SmoothL1}(\hat{\mathbf{y}}, \mathbf{y}), \tag{17}$$

and the accumulation loss for window w is

$$\mathcal{L}_{\text{acc},w} = \text{SmoothL1}\left(\frac{\sum_{j=1}^{w} \widehat{P}_{t+j-1}}{\widetilde{\sigma}_{A,w}}, \frac{\sum_{j=1}^{w} P_{t+j-1}}{\widetilde{\sigma}_{A,w}}\right), \tag{18}$$

where $\widetilde{\sigma}_{A,w} = \max(\sigma_{A,w}, 1)$ and $\sigma_{A,w}$ is the batch standard deviation of true w-month accumulations. The total loss is

$$\mathcal{L}_{\text{AC}} = \mathcal{L}_{\text{monthly}} + \frac{0.10}{3} \sum_{w \in \{3,6,12\}} \mathcal{L}_{\text{acc},w}. \tag{19}$$

For neural architectures, the latent dimension was 64. TCN and selected feed-forward components used dropout 0.1, whereas the two-layer LSTM backbone used dropout 0.15. All neural models were optimized with AdamW. Common training parameters included learning rate 1×10^{-3}, weight decay 1×10^{-4}, maximum of 40 epochs, early stopping with patience 7, and batch size 128. Five seeds (11, 22, 33, 44, 55) were used for neural models. Neural models retained the best training-only checkpoint selected by validation loss; they were not refitted on training plus validation.

Table 2: Model architecture summary

Model	Temporal mechanism	Parameters	Output	Loss
DLinear	Moving-average decomposition	5,396	12-month rainfall	Smooth L1
TCN	Dilations 1/2/4/8	44,380	12-month rainfall	Smooth L1
iTransformer-Lite	Feature-token attention	125,948	12-month rainfall	Smooth L1
TFT-Lite	LSTM and multi-head attention	60,380	12-month rainfall	Smooth L1
SMA-TFT-Trajectory	TFT-Lite + 12 prototypes	77,696	12-month rainfall	Smooth L1
AC-LSTM	Two-layer LSTM	58,012	12-month rainfall	Monthly and accumulation

4.8 Paired Uncertainty Analysis

To summarize the uncertainty between folds and assess the stability of paired comparisons, we used fold-level bootstrap resampling with 10,000 iterations. For paired comparisons (for

example, the PII diagnostic versus the full-trajectory protocol), we resampled the folds with replacement and calculated the mean difference for each bootstrap sample. The 95% fold-resampling interval (FRI) is the range between the 2.5th and 97.5th percentiles of the bootstrap distribution. A paired difference was considered stable over the evaluated folds if the 95% FRI did not include zero.

For neural models, the metrics were first averaged across seeds for each model, fold, SPI, and horizon. These fold-level values were used in the model summaries and the paired uncertainty analysis, so neural models were not given greater weight merely because they had more than one seed. Protocol differences were averaged across seeds for each model and fold, and then across models for each fold.

Fold-resampling intervals summarize between-fold stability but should be interpreted cautiously because only six test folds are available and their expanding training periods overlap.

4.9 Structured Reporting Audit

We carried out a structured methodological audit of how the accumulation protocol and the SPI fitting period are presented in the SPI/SPEI forecasting literature. This audit is a focused methodological evaluation, not a systematic review, and does not follow a PRISMA protocol. Papers were chosen on grounds of topical relevance to SPI or SPEI forecasting, and the full text was read where the PDF was available.

For each paper we coded: (i) whether the forecast issue date was given; (ii) whether a target date or target period was given; (iii) whether the method used to construct the accumulation window was described; (iv) how intermediate unknown months were handled; (v) the causality status of the SPI/SPEI fitting procedure; and (vi) whether the study recognized any information-boundary concern. Papers were divided into five categories: observed-plus-forecast trajectory, forecast-only target-period precipitation, direct future index, drought-class or transition, and non-SPI contextual studies. SPI fitting causality was coded as one of: training-only historical; rolling/expanding causal; leave-one-year-out; hindcast/model climatology; full-record retrospective; unstated; N/A. Coding was carried out by a single reviewer in two passes. Counts are reported with denominators rather than percentages. Seven papers appeared in the audit corpus but their PDFs could not be obtained; these are stated separately and excluded from the prevalence counts.

5 Results

5.1 Protocol-Induced Skill Differences

Table 3 shows the protocol-induced differences for SPI-12 at various forecast leads.

Table 3: SPI-12 protocol-induced differences by forecast lead

Lead h	$f_{w,h}$	ℓ	F1 diff.	95% FRI	R^2 diff.	95% FRI	PR-AUC diff.
1	0.083	0	0.000	[0.000, 0.000]	0.000	[0.000, 0.000]	0.000
3	0.250	2	+0.273	[0.194, 0.360]	+0.269	[0.233, 0.306]	+0.258
6	0.500	5	+0.416	[0.370, 0.473]	+0.707	[0.619, 0.819]	+0.413
9	0.750	8	+0.581	[0.527, 0.643]	+1.254	[1.080, 1.449]	+0.576
12	1.000	11	+0.760	[0.681, 0.824]	+1.873	[1.631, 2.147]	+0.732

At $h = 1$ there is no difference since the month being forecast is also the target month. At $h = 12$, when the whole window has to be forecast, the F1 difference reaches 0.760, the R^2 difference 1.873, and the PR-AUC difference 0.732. For all $h \geq 3$, the fold-resampling intervals excluded zero, indicating a stable contrast direction across the evaluated folds.

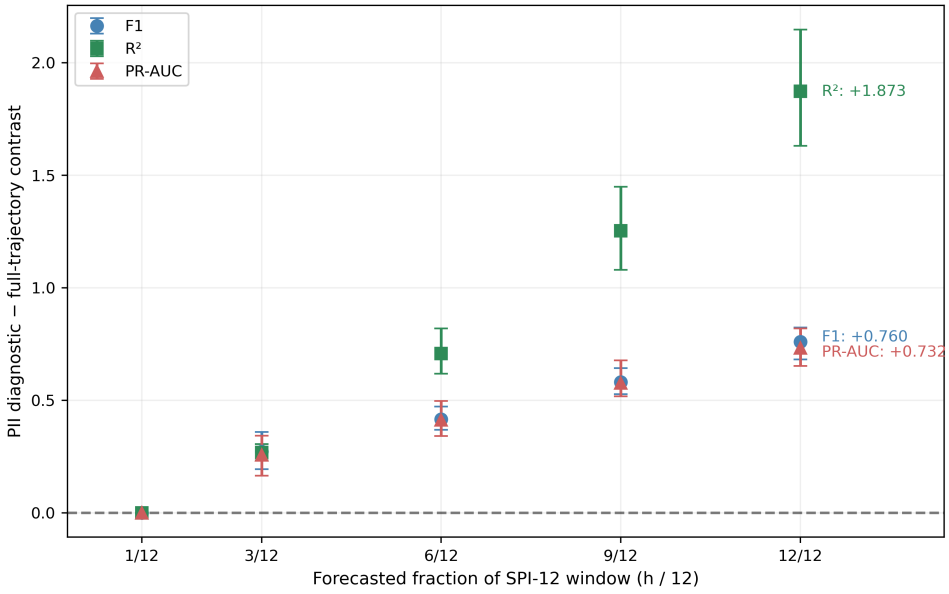


Figure 3: Protocol-induced differences for SPI-12 across forecast leads.

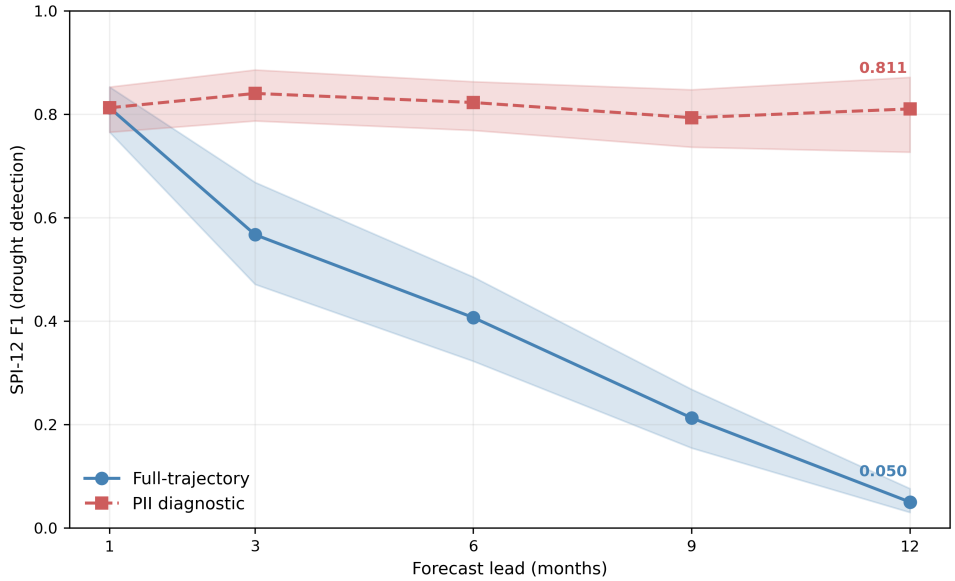


Figure 4: Full-trajectory versus PII-diagnostic SPI-12 F1 across forecast leads. Shaded bands are 95% fold-resampling intervals.

At a one-month lead, both protocols produce identical mean F1 of 0.813 because no intermediate future month is available for insertion under the PII diagnostic. The F1 subsequently decreases to 0.567, 0.407, 0.213 and 0.050 at leads of 3, 6, 9 and 12 months, whereas the PII diagnostic maintains values of 0.841, 0.823, 0.794 and 0.811.

5.2 Comparison of the Evaluation Protocols

Table 4: SPI-12 performance by protocol (cross-model mean)

Lead h	Protocol	F1	PR-AUC	R^2	MAE
1	All protocols	0.813	0.891	0.890	0.153
3	Full-trajectory	0.567	0.647	0.620	0.311
3	Climatology-gap	0.621	0.721	0.712	0.277
3	PII diagnostic	0.841	0.905	0.889	0.157
6	Full-trajectory	0.407	0.456	0.169	0.507
6	Climatology-gap	0.453	0.548	0.455	0.423
6	PII diagnostic	0.823	0.869	0.876	0.168
9	Full-trajectory	0.213	0.280	−0.389	0.690
9	Climatology-gap	0.164	0.351	0.142	0.553
9	PII diagnostic	0.794	0.856	0.865	0.169
12	Full-trajectory	0.050	0.132	−0.994	0.851
12	Climatology-gap	0.000	0.161	−0.183	0.667
12	PII diagnostic	0.811	0.864	0.879	0.160

Climatology-gap generally improved continuous SPI R^2 and MAE and increased drought-risk PR-AUC relative to full-trajectory forecasting, but it did not consistently improve fixed-threshold F1. At $h = 9$ and $h = 12$, its F1 was lower than the full-trajectory protocol. Both valid protocols remained far below the PII diagnostic at long leads.

5.3 Skill Degradation under the Full-Trajectory Protocol

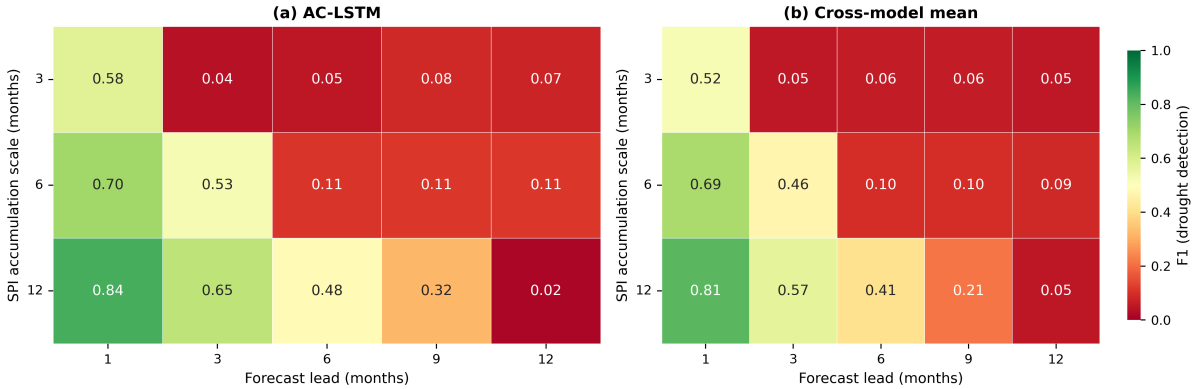


Figure 5: Skill degradation across SPI scales and leads.

Table 5: Full-trajectory F1 by SPI scale and lead (cross-model mean)

Lead h	SPI-3 F1	SPI-6 F1	SPI-12 F1
1	0.523	0.689	0.813
3	0.053	0.465	0.567
6	0.062	0.095	0.407
9	0.060	0.095	0.213
12	0.051	0.086	0.050

Drought detection skill decreases with lead time, with a pattern that varies by SPI scale. With SPI-3, skill is moderate at short lead times (F1 $\approx$ 0.52) but drops considerably by $h = 3$ (F1 $\approx$ 0.05). For SPI-6, skill is moderate at short lead times (F1 $\approx$ 0.69) but falls to about 0.10 by $h = 6$. For SPI-12, skill is high at short lead times (F1 $\approx$ 0.81), decreases to about 0.41 by $h = 6$, and becomes very weak by $h = 12$ (F1 $\approx$ 0.05).

5.4 Ranking of Models

Table 6: Model ranking (PR-AUC, full-trajectory, complete panel)

Rank	Model	Mean PR-AUC
1	AC-LSTM	0.4036
2	Adaptive Climatology	0.3983
3	Fixed Climatology	0.3935
4	Residual Extra Trees	0.3901
5	TCN	0.3687
6	SMA-TFT-Trajectory	0.3563
7	iTransformer-Lite	0.3542
8	TFT-Lite	0.3542
9	DLinear	0.3449
10	Seasonal Naive	0.3031

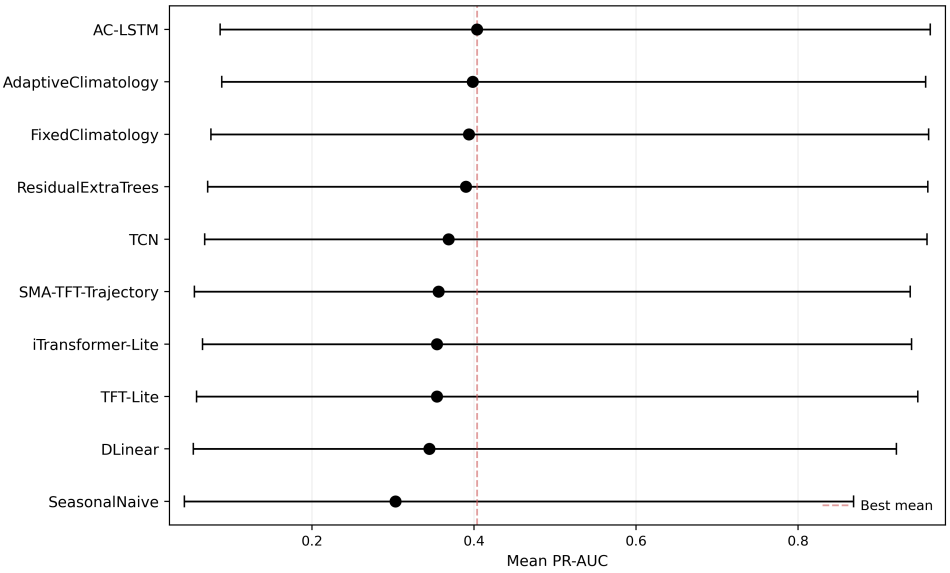


Figure 6: Model ranking with fold-resampling intervals.

The paired comparisons were as follows:

$$\Delta\text{PR-AUC}_{\text{AC}-\text{Adaptive}} = +0.0053, \quad 95\%\ \text{FRI} = [-0.0280, 0.0327],$$

$$\Delta\text{F1}_{\text{AC}-\text{Adaptive}} = +0.0727, \quad 95\%\ \text{FRI} = [0.0499, 0.0944],$$

$$\Delta R^2_{\text{AC}-\text{Adaptive}} = -0.2034, \quad 95\%\ \text{FRI} = [-0.3892, -0.0056].$$

AC-LSTM achieved the highest mean PR-AUC, but its PR-AUC advantage over Adaptive Climatology was inconclusive, and its mean continuous SPI R^2 was lower with the paired fold-resampling interval excluding zero. The top four consisted of three non-neural methods and one

neural method.

Table 7: Model ranking (SPI R^2, full-trajectory, complete panel)

Rank	Model	Mean SPI R^2
1	Adaptive Climatology	0.2138
2	Residual Extra Trees	0.2119
3	Fixed Climatology	0.1992
4	AC-LSTM	0.0104
5	SMA-TFT-Trajectory	−0.2214
6	TFT-Lite	−0.2292
7	iTransformer-Lite	−0.2673
8	TCN	−0.3753
9	DLinear	−0.4413
10	Seasonal Naive	−0.5471

The discrepancy between the PR-AUC and R^2 rankings highlights the importance of using multiple evaluation metrics. Adaptive Climatology provided stronger continuous SPI fidelity but lower fixed-threshold F1 than AC-LSTM, and Residual Extra Trees achieved the lowest mean MAE (0.5565) among the top-performing methods.

5.5 Sensitivity Analysis: Variable-Record Panel

We repeated the analysis on the eligible 34-station panel (all stations with $\geq$ 200 months of data). Table 8 presents the PR-AUC ranking.

Table 8: Model ranking (PR-AUC, full-trajectory, eligible panel)

Rank	Model	Mean PR-AUC
1	Residual Extra Trees	0.3814
2	Adaptive Climatology	0.3781
3	Fixed Climatology	0.3724
4	TFT-Lite	0.3468
5	SMA-TFT-Trajectory	0.3453
6	iTransformer-Lite	0.3445
7	TCN	0.3424

The ranking change indicates sensitivity to station composition and record length.

5.6 Empirical versus Gamma SPI Distribution

Table 9: PR-AUC by SPI distribution (full-trajectory, complete panel)

Model	Gamma SPI	Empirical SPI
AC-LSTM	0.4036	0.3845
Adaptive Climatology	0.3983	0.3888
Fixed Climatology	0.3935	0.3774
Residual Extra Trees	0.3901	0.3792

Under empirical SPI, simple climatological and residual-tree baselines remained competitive, although absolute PR-AUC values and the leading model changed. The full three-protocol contrast was evaluated only under Gamma SPI.

5.7 Event-Level Analysis

For SPI-12 we carried out an event-level analysis under the Gamma-SPI full-trajectory protocol. This analysis is reported for AC-LSTM as a case study rather than for all ten models. False-warning counts are normalized per 10 station-years using the complete-panel test exposure of 125 station-years.

Table 10: SPI-12 event-level metrics by lead (AC-LSTM, full-trajectory, complete panel)

Lead h	True events	Detected events	Detection rate	False warnings	Median onset error (months)
1	26	23.2	0.892	0.512	0.00
3	26	15.8	0.608	1.184	0.00
6	31	14.8	0.477	2.368	2.25
9	35	19.4	0.554	2.416	2.00
12	35	1.0	0.029	0.400	1.00

False warnings are expressed per 10 station-years, computed from the mean count across neural seeds divided by 125 station-years of complete-panel exposure and multiplied by 10. Exposure = 10 stations × 25 issue months × 6 folds.

The event detection rate decreased from 0.892 at $h = 1$ to 0.029 at $h = 12$, though it rose from 0.477 at $h = 6$ to 0.554 at $h = 9$. The median onset error stayed low at leads 1 and 3 before increasing slightly thereafter. The false-warning rate peaked at leads 6 and 9 before falling at lead 12.

5.8 Reporting Practices in Prior Studies

Of the 30 fully audited studies, 11 used an observed-plus-forecast trajectory, 4 calculated the target index from forecast-only target-period precipitation, 10 forecast a future drought index

directly, 1 forecast drought-class transitions, and 4 were non-SPI contextual studies. Seven papers could not be retrieved in full and are excluded from the counts.

Table 11: Reporting practices in SPI/SPEI forecasting studies ($N = 30$ full-PDF audited)

Category	N	Intermediate treatment	Causal-bounded fitting	Hindcast climatology	Full-record retrospective	Unstated/N/A
Observed + forecast	11	11 explicit	4	2	3	2
Forecast-only precipitation	4	4 explicit	0	3	1	0
Direct future index	10	1 explicit; 9 not required	0	0	7	3
Class / transition	1	Not required	0	0	1	0
Non-SPI contextual	4	Mixed/N/A	0	0	1	3

"Causally bounded fitting" = training-only historical, rolling/expanding causal, or leave-one-year-out. The seven papers whose PDFs could not be obtained are excluded from the totals.

All observed-plus-forecast studies explicitly reported forecast origin, target date, accumulation construction, and intermediate-month treatment. Their SPI fitting procedures nevertheless differed: 4 used training-only, rolling, or leave-one-year-out fitting; 2 used model or hindcast climatologies; 3 used full-record retrospective fitting; and 2 did not state the fitting sample. Of the 10 direct-index studies, 9 did not require intermediate-month treatment because the future index was predicted directly; the one analytical persistence study (Lyon et al., 2012) explicitly represented future precipitation when constructing the future SPI distribution. Seven direct-index studies constructed the index retrospectively from the full record, while 3 did not state the fitting period.

6 Discussion

6.1 Protocol Sensitivity and Accumulation Memory

In the setting under evaluation, the contrasts between protocols were greater than the differences in the average model values, especially for long-lead SPI-12. When most of the SPI window has already been observed, there is little or no difference between the full-trajectory protocol and the PII diagnostic; when most of the window has to be predicted, the difference becomes substantial. The magnitude of the protocol-induced differences shows that, when looking at the skill of SPI-12 over long lead times, it is necessary to state clearly how the intermediate months were produced.

The operational implications are equally important. The cross-model average SPI-12 F1 at lead 12 was about 0.05 under the full-trajectory design, indicating that the evaluated models displayed very weak fixed-threshold performance at a 12-month lead. Reporting the same skill under the PII diagnostic without qualification could create an unjustified sense of confidence in

the forecasts.

6.2 Model Performance under the Full-Trajectory Protocol

The model rankings indicate that increased architectural complexity does not necessarily lead to better performance in forecasting SPI values. Under the full-trajectory protocol, both adaptive climatology and residual tree models were consistently competitive or superior to the neural architectures. Adaptive Climatology had the highest mean R^2 and Residual Extra Trees the lowest mean MAE.

There are several factors that could account for the small advantage enjoyed by neural models. One is that the temporal structure of SPI can be adequately handled by straightforward statistical methods. Adaptive climatology can detect the seasonal patterns particular to each station by using data from the same month in recent years, without needing the complexity of neural networks. Residual Extra Trees bases its predictions on residuals with respect to climatology and provides a flexible nonparametric approach that can capture nonlinear relationships.

In the dataset and experimental setup used here, SMA-TFT did not show a practically significant overall improvement over TFT-Lite: its advantage in F1 was minor (+0.0028) and that in PR-AUC was not clear (+0.0021, with the fold-resampling interval containing zero).

6.3 The Forecasted Fraction as a Reporting Diagnostic

The value $f = \min(1, h/w)$ provides a simple way of determining how much of the target accumulation period is actually forecast. We recommend that studies evaluating SPI over multiple horizons state the fraction of the target accumulation period that is forecast rather than the fraction that is observed. The lower the value of f, the greater the portion of the target index that comes from the observations rather than from the forecast model.

6.4 Connection with Previous Work

The present study concentrates on the sensitivity of reported skill to the treatment of intermediate future months within the target accumulation period. The study is complementary to operational trajectory studies. Cancelliere et al. (2007) expressed future SPI analytically in terms of observed and unknown precipitation; Quan et al. (2012), Yoon et al. (2012), Yuan et al. (2013), Dutra et al. (2013, 2014), Mo and Lyon (2015), Sutanto et al. (2020) and Shin et al. (2020) explicitly appended forecast precipitation to observed analyses; Torres-Vázquez et al.

(2024) have provided an open-source implementation with explicit issue date and accumulation partition; and Brands et al. (2025) compared reanalysis and model-based backfilling for SPEI-3.

The fact that adaptive climatology remained as competitive as neural models is consistent with earlier work showing that simple baselines can match complex architectures (Zeng et al., 2023). For the station-only predictors and models examined here, full-trajectory SPI-12 skill became marginal only after six months and was weak at longer lead times. This decline is in agreement with the statement that drought predictability depends strongly on lead time, predictor information, initial conditions and slowly varying climate drivers (Hao et al., 2018). Lee et al. (2026) have analysed the deterministic autocorrelation structure of overlapping SPI windows and cautioned that apparent SPI skill may reflect accumulation memory rather than forecast information. Since we were not able to check the full paper at the time of writing, we regard it only as related work.

6.5 Limitations and Future Study

This study has a number of limitations. Since the analysis focuses only on Bangladesh, a country with a tropical monsoon climate, the results cannot be extended to other types of precipitation patterns. The compilation of data from public stations does not include complete metadata regarding homogenization, quality control, and station relocation, which means that the original decisions behind the data cannot be fully reconstructed. Although the set of models is varied, it is by no means complete, and the calculation of the SPI is based on a mixed Gamma distribution which might not be the best choice for all station–month combinations; the contrasts between protocols were only assessed under this approach. A fixed window of 24 months was used as the input, and the fold-resampling intervals, which are based on six overlapping test folds, indicate the stability between folds but are not conventional confidence intervals. Lastly, large-scale climate factors such as ENSO and IOD have not been incorporated.

The framework should be extended in future studies to other areas and climate types, include large-scale climate indices, and the three-protocol comparison should be carried out under different SPI distributions. A systematic review of the literature accompanied by independent double-coding would reinforce the recommendations regarding reporting, and the forecasted fraction diagnostic should be applied to other accumulated indices (for example, SPEI, SSI, snowpack, streamflow) in order to examine its generality.

6.6 Implications for Operational Drought Forecasting

Operational systems should make clear what information is available at the time a forecast is issued and then evaluate the forecasts on that basis. The full-trajectory protocol provides an assessment consistent with the issue time because it explicitly includes all unknown months. The climatology-gap protocol provides a baseline suitable for operational use that can be put into practice in real time.

The forecasted fraction should be given with the evaluation results as a reporting diagnostic. Since the forecasted fraction is not a calibrated measure of forecast reliability and should not be used as an operational probability, it is a descriptive reporting diagnostic showing the proportion of the target accumulation window that depends on post-issue information.

Forecasts with short to medium lead times offered the most reasonable basis for early-warning purposes, while those with longer lead times would require further predictive information, calibration and prospective validation. The fact that adaptive climatology and residual tree models give similar results in a competitive environment shows that operational systems do not always have to use complex neural networks.

Studies reporting multi-horizon SPI skill should give the forecast issue date, the target date, how the accumulation window is constructed, the way the unknown intermediate months are treated, the period used to fit the SPI, and the forecasted fraction.

7 Conclusion

The study has systematically evaluated multi-horizon SPI drought forecasts using three different evaluation protocols, each of which handles the months after the issue date within the target accumulation period in a distinct way. The main conclusion is that the accumulation protocol has a greater effect on the reported multi-horizon SPI skill than the model's architecture. For SPI-12 at a 12-month lead, the perfect-intermediate-information (PII) diagnostic raised the F1 score by 0.760 (95% FRI: 0.681–0.824), the PR-AUC by 0.732 (95% FRI: 0.653–0.820), and the R^2 by 1.873 (95% FRI: 1.631–2.147) when compared to the full-trajectory protocol. When the forecasts were assessed using the full-trajectory method, the mean F1 score for SPI-12 dropped from 0.813 at a one-month lead to 0.050 at a twelve-month lead. No individual model performed best in all the evaluation criteria: AC-LSTM achieved a higher fixed-threshold F1 than Adaptive Climatology by 0.0727 (95% FRI: 0.0499–0.0944), but it decreased the continuous SPI R^2 by

0.2034 (95% FRI: −0.3892 to −0.0056), while the simple climatological and residual-tree baseline models stayed competitive under both the Gamma and empirical SPI methods.

The conclusion is that the multi-horizon SPI skill reported should be understood only when there is a clear statement of the forecast-time information boundary. It is therefore suggested that multi-horizon SPI studies include information on the date of the forecast issuance, the target date, the way in which each unknown accumulation month is treated, the SPI fitting period, and the fraction of the forecast window. Further research should extend the framework to other regions and climate regimes, to other accumulated indices such as SPEI and SSI, and to situations that make use of large-scale climate predictors or initialized seasonal forecasts, in order to determine whether the collapse of long-lead skill described here indicates a fundamental limit of SPI predictability or is instead a result of the use of a station-only predictor set.

CRediT authorship contribution statement

Md. Shafiqul Islam: Supervision, Project administration, Writing – review & editing, Validation.

Md Abu Jubayer Bijoy: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

Data Availability: The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Generative AI and AI-Assisted Technologies Declaration: During the preparation of this manuscript, the authors used the following AI-assisted technologies:

- Semantic Scholar and SciSpace were used to assist with literature review and identification of relevant sources. All sources identified were independently verified by the authors for relevance and accuracy.

- DeepSeek-V3 was used for minor grammar correction, language polishing, and formatting suggestions. This included improving sentence clarity and ensuring consistent formatting of references and tables.

All AI-assisted outputs were independently reviewed and verified by the authors. The authors take full responsibility for the accuracy, integrity, and final content of the manuscript. No AI tools were used for data analysis, generation of results, or creation of tables and figures.

References

Aghelpour, P., Mohammadi, B., Mehdizadeh, S., Bahrami-Pichaghchi, H., & Duan, Z. (2021). A novel hybrid dragonfly optimization algorithm for agricultural drought prediction. *Stochastic Environmental Research and Risk Assessment*, 35, 2459–2477. https://doi.org/10.1007/s00477-021-02011-2

Ahmed Osman, A. I., 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 forecasting. *Climate Risk Management*, 50, 100758. https://doi.org/10.1016/j.crm.2025.100758

Al Mamun, M. A., Sarker, M. R., Sarkar, M. A. R., Roy, S. K., Nihad, S. A. I., McKenzie, A. M., Hossain, M. I., & Kabir, M. S. (2024). Identification of influential weather parameters and seasonal drought prediction in Bangladesh using machine learning algorithm. *Scientific Reports*, 14, 566. https://doi.org/10.1038/s41598-023-51111-2

Ali, M., Deo, R. C., Maraseni, T., & Downs, N. J. (2019). Improving SPI-derived drought forecasts incorporating synoptic-scale climate indices in multi-phase multivariate empirical mode decomposition model hybridized with simulated annealing and kernel ridge regression algorithms. *Journal of Hydrology*, 576, 366–384. https://doi.org/10.1016/j.jhydrol.2019.06.032

Awal, M. A., & Al Mamun, M. M. A. (2022). Long term meteorological drought forecasting for northwestern region of Bangladesh using wavelet artificial neural network. *Revista Brasileira de Meteorologia*, 37(4), 453–465. https://doi.org/10.1590/0102-77863740058

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint*, 1803.01271. https://doi.org/10.48550/arXiv.1803.01271

Behrangi, A., Nguyen, H., & Granger, S. (2015). Probabilistic seasonal prediction of meteorological drought using the bootstrap and multivariate information. *Journal of Applied Meteorology and Climatology*, 54(7), 1512–1526. https://doi.org/10.1175/JAMC-D-14-0162.1

Belayneh, A., Adamowski, J., Khalil, B., & Ozga-Zielinski, B. (2014). Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural net-

work and wavelet support vector regression models. *Journal of Hydrology*, 508, 418–429. https://doi.org/10.1016/j.jhydrol.2013.10.052

Bonaccorso, B., Cancelliere, A., & Rossi, G. (2015). Probabilistic forecasting of drought class transitions in Sicily (Italy) using Standardized Precipitation Index and North Atlantic Oscillation Index. *Journal of Hydrology*, 526, 136–150. https://doi.org/10.1016/j.jhydrol.2015.01.070

Brands, S., Iturbide, M., Díez González-Pardo, J., Herrera, S., Bedia, J., Manzanas, R., Rodríguez-Guisado, E., Beguería, S., Vicente-Serrano, S. M., & Gutiérrez, J. M. (2025). Seasonal drought predictions in the Mediterranean using the SPEI index: Paving the way for their operational applicability in climate services. *Climate Services*, 38, 100555. https://doi.org/10.1016/j.cliser.2025.100555

Brust, C., Kimball, J. S., Maneta, M. P., Jencso, K., & Reichle, R. H. (2021). DroughtCast: A machine learning forecast of the United States Drought Monitor. *Frontiers in Big Data*, 4, 773478. https://doi.org/10.3389/fdata.2021.773478

Cancelliere, A., Di Mauro, G., Bonaccorso, B., & Rossi, G. (2007). Drought forecasting using the Standardized Precipitation Index. *Water Resources Management*, 21(5), 801–819. https://doi.org/10.1007/s11269-006-9062-y

Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109, 1997–2028. https://doi.org/10.1007/s10994-020-05910-7

Dutra, E., Di Giuseppe, F., Wetterhall, F., & Pappenberger, F. (2013). Seasonal forecasts of droughts in African basins using the Standardized Precipitation Index. *Hydrology and Earth System Sciences*, 17(6), 2359–2373. https://doi.org/10.5194/hess-17-2359-2013

Dutra, E., Pozzi, W., Wetterhall, F., Di Giuseppe, F., Magnusson, L., Naumann, G., Barbosa, P., Vogt, J., & Pappenberger, F. (2014). Global meteorological drought – Part 2: Seasonal forecasts. *Hydrology and Earth System Sciences*, 18(7), 2669–2678. https://doi.org/10.5194/hess-18-2669-2014

Felsche, E., & Ludwig, R. (2021). Applying machine learning for drought prediction in a perfect model framework using data from a large ensemble of climate simulations. *Natural Hazards and Earth System Sciences*, 21(12), 3679–3691. https://doi.org/10.5194/nhess-21-3679-2021

Guttman, N. B. (1999). Accepting the standardized precipitation index: A calculation algorithm. *Journal of the American Water Resources Association*, 35(2), 311–322. https://doi.org/10.1111/j.1752-1688.1999.tb03592.x

Gyaneshwar, A., Mishra, A., Chadha, U., Raj Vincent, P. M. D., Rajinikanth, V., Pattukandan Ganapathy, G., & Srinivasan, K. (2023). A contemporary review on deep learning models for drought prediction. *Sustainability*, 15(7), 6160. https://doi.org/10.3390/su15076160

Habiba, U., & Shaw, R. (2013). Farmers' response to drought in northwestern Bangladesh. In R.

Shaw (Ed.), *Risks and Conflicts: Local Responses to Natural Disasters* (Vol. 14, pp. 131–161). Emerald Group Publishing Ltd. https://doi.org/10.1108/S2040-7262(2013)0000014012

Hao, Z., Singh, V. P., & Xia, Y. (2018). Seasonal drought prediction: Advances, challenges, and future prospects. *Reviews of Geophysics*, 56(1), 108–141. https://doi.org/10.1002/2016RG000549

Hayes, M., Svoboda, M., Wall, N., & Widhalm, M. (2011). The Lincoln declaration on drought indices: Universal meteorological drought index recommended. *Bulletin of the American Meteorological Society*, 92(4), 485–488. https://doi.org/10.1175/2010BAMS3103.1

He, R., Zhang, L., & Chew, A. W. Z. (2024). Data-driven multi-step prediction and analysis of monthly rainfall using explainable deep learning. *Expert Systems with Applications*, 235, 121160. https://doi.org/10.1016/j.eswa.2023.121160

Hossain, M. A., Begum, M., Akhtar, M. N., Talukder, M. A., Rahman, N., & Rahman, M. (2025). An ensemble learning approach for drought analysis and forecasting in central Bangladesh. *Advances in Meteorology*, 2025(1), Article 4904248. https://doi.org/10.1155/adme/4904248

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4, 100804. https://doi.org/10.1016/j.patter.2023.100804

Karbasi, M., Jamei, M., Malik, A., Kisi, O., & Yaseen, Z. M. (2023). Multi-steps drought forecasting in arid and humid climate environments: Development of integrative machine learning model. *Agricultural Water Management*, 281, 108210. https://doi.org/10.1016/j.agwat.2023.108210

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), 1–21. https://doi.org/10.1145/2382577.2382579

Lavaysse, C., Stockdale, T., McCormick, N., & Vogt, J. (2020). Evaluation of a new precipitation-based index for global seasonal forecasting of unusually wet and dry periods. *Weather and Forecasting*, 35(4), 1345–1361. https://doi.org/10.1175/WAF-D-19-0196.1

Lavaysse, C., Vogt, J., & Pappenberger, F. (2015). Early warning of drought in Europe using the monthly ensemble system from ECMWF. *Hydrology and Earth System Sciences*, 19(7), 3273–3286. https://doi.org/10.5194/hess-19-3273-2015

Lee, S., Danandeh Mehr, A., Moriasi, D., & Mirchi, A. (2025). Large-scale drought forecasting in the U.S. Southern Plains through a hybrid cluster-based wavelet-machine learning approach. *Water Resources Research*, 61(4), e2024WR039744. https://doi.org/10.1029/2024WR039744

Lee, T., Kong, Y., Singh, V. P., & Yoon, H.-C. (2026). Autocorrelation structure of SPI and its implication for drought forecasting. *International Journal of Climatology*. Advance online publication. https://doi.org/10.1002/joc.70240

Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

Liu, Z., Lu, G., He, H., Wu, Z., & He, J. (2018). A conceptual prediction model for seasonal drought processes using atmospheric and oceanic standardized anomalies: Application to regional drought processes in China. *Hydrology and Earth System Sciences*, 22(1), 529–546. https://doi.org/10.5194/hess-22-529-2018

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2024). iTransformer: Inverted transformers are effective for time series forecasting. *The Twelfth International Conference on Learning Representations*. https://arxiv.org/abs/2310.06625

Lyon, B., Bell, M. A., Tippett, M. K., Kumar, A., Hoerling, M. P., Quan, X.-W., & Wang, H. (2012). Baseline probabilities for the seasonal prediction of meteorological drought. *Journal of Applied Meteorology and Climatology*, 51(7), 1222–1237. https://doi.org/10.1175/JAMC-D-11-0132.1

Ma, F., Yuan, X., & Ye, A. (2015). Seasonal drought predictability and forecast skill over China. *Journal of Geophysical Research: Atmospheres*, 120(16), 8264–8275. https://doi.org/10.1002/2015JD023185

Magallanes-Quintanar, R., Galván-Tejada, C. E., Galván-Tejada, J. I., Méndez-Gallegos, S. de J., & García-Domínguez, A. (2026). Standardized precipitation index forecasting comparison using transformer models. *Forecasting*, 8(3), 44. https://doi.org/10.3390/forecast8030044

Mammas, K., & Lekkas, D. F. (2021). Characterization of bias during meteorological drought calculation in time series out-of-sample validation. *Water*, 13(18), 2531. https://doi.org/10.3390/w13182531

Márquez-Grajales, A., Villegas-Vega, R., Salas-Martínez, F., Acosta-Mesa, H. G., & Mezura-Montes, E. (2024). Characterizing drought prediction with deep learning: A literature review. *MethodsX*, 13, 102800. https://doi.org/10.1016/j.mex.2024.102800

McKee, T. B., Doesken, N. J., & Kleist, J. (1993). The relationship of drought frequency and duration to time scales. *Proceedings of the 8th Conference on Applied Climatology*, 179–184.

Mishra, A. K., & Singh, V. P. (2010). A review of drought concepts. *Journal of Hydrology*, 391(1–2), 202–216. https://doi.org/10.1016/j.jhydrol.2010.07.012

Mo, K. C., & Lyon, B. (2015). Global meteorological drought prediction using the North American Multi-Model Ensemble. *Journal of Hydrometeorology*, 16(3), 1409–1424. https://doi.org/10.1175/JHM-D-14-0192.1

Monir, M. M., Sarker, S. C., Rahman, M. M., & Islam, M. N. (2026). Analysis of the trend of dry spells and how ocean factors affect their patterns during the summer monsoon in Bangladesh using the Mann–Kendall and frontier atmospheric general circulation model. *Geosystems and Geoenvironment*, 5(2), 100472. https://doi.org/10.1016/j.geogeo.2025.100472

Mwangi, E., Wetterhall, F., Dutra, E., Di Giuseppe, F., & Pappenberger, F. (2014). Forecasting droughts in East Africa. *Hydrology and Earth System Sciences*, 18(2), 611–620. https://doi.org/10.5194/hess-18-611-2014

Park, S., Im, J., Han, D., & Rhee, J. (2020). Short-term forecasting of satellite-based drought indices

using their temporal patterns and numerical model output. *Remote Sensing*, 12(21), 3499. https://doi.org/10.3390/rs12213499

Quan, X.-W., Hoerling, M. P., Lyon, B., Kumar, A., Bell, M. A., Tippett, M. K., & Wang, H. (2012). Prospects for dynamical prediction of meteorological drought. *Journal of Applied Meteorology and Climatology*, 51(7), 1238–1252. https://doi.org/10.1175/JAMC-D-11-0194.1

Sadiq, M. A., Sarkar, S. K., & Raisa, S. S. (2023). Meteorological drought assessment in northern Bangladesh: A machine learning-based approach considering remote sensing indices. *Ecological Indicators*, 157, 111233. https://doi.org/10.1016/j.ecolind.2023.111233

Saito, T., & Rehmsmeier, M. (2015). The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Sarkar, B., Mohinuddin, S., Islam, A., Islam, A. R. M. T., Saha, U. D., Sengupta, S., Pal, S. C., Chu, H.-J., & Huang, J.-C. (2025). SPI-gamma random forest modelling for meteorological drought characterization and prediction in the Bengal Delta, Indo-Bangladesh region. *Theoretical and Applied Climatology*, 156(1), Article 51. https://doi.org/10.1007/s00704-024-05293-y

Shahid, S. (2010). Rainfall variability and the trends of wet and dry periods in Bangladesh. *International Journal of Climatology*, 30(15), 2299–2313. https://doi.org/10.1002/joc.2053

Shin, C.-S., Huang, B., Dirmeyer, P. A., Halder, S., & Kumar, A. (2020). Sensitivity of U.S. drought prediction skill to land initial states. *Journal of Hydrometeorology*, 21(12), 2793–2811. https://doi.org/10.1175/JHM-D-20-0025.1

Stagge, J. H., Tallaksen, L. M., Gudmundsson, L., Van Loon, A. F., & Stahl, K. (2015). Candidate distributions for climatological drought indices (SPI and SPEI). *International Journal of Climatology*, 35(13), 4027–4040. https://doi.org/10.1002/joc.4267

Sutanto, S. J., Van Lanen, H. A. J., Wetterhall, F., & Llort, X. (2020). Potential of Pan-European seasonal hydrometeorological drought forecasts obtained from a multihazard early warning system. *Bulletin of the American Meteorological Society*, 101(4), E368–E393. https://doi.org/10.1175/BAMS-D-18-0196.1

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0

Torres-Vázquez, M. Á., Di Giuseppe, F., Dutra, E., Halifa-Marín, A., Jerez, S., Ramón, J., Montávez, J. P., Doblas-Reyes, F. J., & Turco, M. (2024). Probabilistic predictions for meteorological droughts based on multi-initial conditions. *Journal of Hydrology*, 640, 131662. https://doi.org/10.1016/j.jhydrol.2024.131662

Vicente-Serrano, S. M., Beguería, S., & López-Moreno, J. I. (2010). A multiscalar drought index sensitive to global warming: The standardized precipitation evapotranspiration index. *Journal of Climate*, 23(7), 1696–1718. https://doi.org/10.1175/2009JCLI2909.1

Wilhite, D. A., Svoboda, M. D., & Hayes, M. J. (2007). Understanding the complex impacts of drought: A key to enhancing drought mitigation and preparedness. *Water Resources Management*, 21(5), 763–774. https://doi.org/10.1007/s11269-006-9076-5

World Meteorological Organization. (2012). *Standardized Precipitation Index user guide* (WMO-No. 1090). Geneva.

Yoon, J.-H., Mo, K., & Wood, E. F. (2012). Dynamic-model-based seasonal prediction of meteorological drought over the contiguous United States. *Journal of Hydrometeorology*, 13(2), 463–482. https://doi.org/10.1175/JHM-D-11-038.1

Yu, J., Ma, T., Jia, L., Rong, H., Su, Y., & Wahab, M. M. A. (2024). Multivariate spatio-temporal modeling of drought prediction using graph neural network. *Journal of Hydroinformatics*, 26(1), 107–124. https://doi.org/10.2166/hydro.2023.134

Yuan, X., Wood, E. F., Chaney, N. W., Sheffield, J., Kam, J., Liang, M., & Guan, K. (2013). Probabilistic seasonal forecasting of African drought by dynamical models. *Journal of Hydrometeorology*, 14(6), 1706–1720. https://doi.org/10.1175/JHM-D-13-054.1

Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2023). Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9), 11121–11128. https://doi.org/10.1609/aaai.v37i9.26317