

A Multivariate New Copula–Based Ensemble Machine Learning Drought Index (CEMLDI) for Assessing Nonlinear Climate Conditions

Hamd Ullah

hamdullahphd@gmail.com

International Islamic University Islamabad

Research Article

Keywords: Drought index, Copula, Random Forest, SPEI, Multivariate modeling, Pakistan

Posted Date: June 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9777557/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**A Multivariate New Copula–Based Ensemble Machine Learning Drought Index
(CEMLDI) for Assessing Nonlinear Climate Conditions**

**Running title: A Multivariate New Copula–Based Ensemble Machine Learning Drought
Index**

Hamd Ullah^a

Corresponding Authors' Email Address:

Dr. Hamd Ullah. Email: hamdullahphd@gmail.com

^a Department of Statistics, Federal Government Quaid-e-Azam College, National University
of Pakistan (NUP), Rawalpindi, Pakistan. Email: hamdullahphd@gmail.com

Availability of data

Data will be made available on request.

Funding

This research received no external funding.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

Hamd Ullah: Conceptualization, Data Curation, Formal analysis, Methodology, Investigation,
Software, Writing – original draft, Review and editing, Supervision.

A Multivariate New Copula–Based Ensemble Machine Learning Drought Index (CEMLDI) for Assessing Nonlinear Climate Conditions

Abstract

Drought is a complex hydro-climatic hazard arising from interactions among rainfall deficits, atmospheric evaporative demand, temperature changes, and hydrological processes. Common drought indices, such as the Standardized Precipitation Evapotranspiration Index (SPEI), are widely used for monitoring but often overlook multivariate dependencies, compound events, and nonlinear climatic interactions. Although copula theory models joint dependencies and machine learning techniques capture nonlinear relationships among climate variables, these methods are typically applied separately, limiting comprehensive drought analysis. To address this gap, this study introduces the Copula–Based Ensemble Machine Learning Drought Index (CEMLDI), a hybrid approach that combines copula dependence modeling with Random Forest ensemble learning for multivariate drought assessment. This framework captures nonlinear and asymmetric relationships using station-specific marginal distributions and copula functions, while incorporating joint probabilities and return periods into machine learning models to improve drought detection across regions and timescales. The effectiveness of CEMLDI is tested using hydro-climatic data from multiple regions and compared with traditional drought indices. Results indicate that CEMLDI enhances sensitivity to drought extremes, better identifies compound drought conditions, and remains stable across different climate scenarios. It also responds more effectively to short-term climate variability and nonlinear hydro-climatic interactions. Unlike fixed-parameter drought indices, CEMLDI offers a flexible, adaptable system that can include various probability distributions, copula types, climate variables, machine learning techniques, and drought indicators suited to regional climate patterns and data availability. It can integrate both observational hydro-climatic data and climate model outputs, such as CMIP6 projections, for historical drought assessment and future climate risk analysis. Overall, CEMLDI provides a robust and scalable system for drought monitoring, early warning, and climate risk management, particularly in regions prone to climate variability and water scarcity.

Keywords: Drought index; Copula; Random Forest; SPEI; Multivariate modeling; Pakistan

1. Introduction

Drought is a complex, nonlinear, multiscale hydro-climatic hazard driven by interactions among rainfall deficits, atmospheric moisture demand, temperature changes, and land–atmosphere feedbacks (Zscheischler et al., 2018). Its impacts are intensifying with climate change, as compound extremes and cascading hydro-climatic processes prolong drought duration and increase severity. In diverse climatic regions, such as Pakistan, these interactions show substantial spatial variability, making drought assessment more challenging for traditional statistical methods (Abbas et al., 2024; Khan et al., 2024).

Standardized drought indices such as SPI, SCDI, and SPEI remain widely used in operational monitoring because they are simple and easy to interpret (McKee et al., 1993; Vicente-Serrano et al., 2010; Ullah et al., 2023a, 2023b). Yet these indices are primarily univariate and rely on simplifying assumptions that limit their ability to capture joint dependencies among hydro-climatic factors. Consequently, they often overlook complex drought causes arising from concurrent anomalies in precipitation, temperature, and evapotranspiration (Wang et al., 2020; Zhao et al., 2024). This fundamental limitation becomes even more significant under changing climate conditions, where modeling marginal behavior alone isn't enough to understand overall drought dynamics.

Copula-based methods have become increasingly popular in hydro-climatic modeling for capturing multivariate dependence. Copula theory offers a versatile way to construct joint distributions by separating marginal behavior from dependence structures, enabling effective modeling of nonlinear and non-Gaussian relationships among climate variables (Nelsen, 2006; Genest & Favre, 2007). In drought studies, copulas have improved the estimation of joint severity–duration relationships and the analysis of return periods for compound extremes (Shao et al., 2025). However, despite these advances, copula methods remain largely statistical tools and are not inherently capable of learning high-dimensional nonlinear relationships directly from data.

Meanwhile, ensemble machine learning techniques have proven highly effective for modeling complex, nonlinear, and high-dimensional hydro-climatic systems. Methods such as Random Forest and other ensemble frameworks improve predictive stability by combining multiple learners and reducing variance, particularly under noisy climatic conditions (Breiman, 2001; Poudel et al., 2024; Li et al., 2025). These techniques are increasingly used for drought

prediction and classification, often outperforming traditional regression models. Nonetheless, machine learning models generally act as black-box predictors and do not explicitly incorporate probabilistic dependence structures, limiting their physical interpretability in risk-based drought analysis.

Despite rapid advances in copula-based modeling and machine learning, a core methodological gap remains: current frameworks often treat probabilistic dependence modeling and nonlinear predictive learning as separate approaches rather than as components of an integrated drought characterization system (Afshar et al., 2024; Wang et al., 2024). This separation limits the development of drought indices that are simultaneously probabilistically consistent, adaptable to data, and responsive to complex hydro-climatic interactions. Consequently, most operational drought indices still rely on fixed statistical methods and narrow variable sets, limiting their effectiveness in complex, changing climate conditions (Vicente-Serrano et al., 2010; Hao et al., 2023).

To address this limitation, the study introduces a new Copula-Based Ensemble Machine Learning Drought Index (CEMLDI). This approach integrates copula theory and ensemble machine learning into a single probabilistic–predictive framework for drought analysis. The copula component characterizes the joint dependence among standardized hydro-climatic variables, capturing nonlinear relationships and tail co-movements. Meanwhile, the ensemble machine learning component models complex interactions and hidden patterns in high-dimensional climatic predictors, enhancing predictive stability and reducing sensitivity to biases in individual algorithms (Genest & Favre, 2007; Breiman, 2001).

By combining these two complementary paradigms, CEMLDI advances beyond traditional drought indices by explicitly modeling both the statistical dependence structure and the nonlinear dynamics derived from data within a unified framework. Unlike standard methods that mainly focus on precipitation or water-balance calculations, this approach integrates multiple hydro-climatic factors, such as precipitation, temperature, and evapotranspiration, thereby enhancing its ability to detect complex drought conditions and respond quickly to rapid climate change (Zhao et al., 2024; Talebi & Samadianfard, 2024; Lan et al., 2025).

A key advantage of the proposed framework is its modular, extensible design. It supports flexible selection of marginal distributions, copula families, and ensemble learning algorithms, making it suitable for diverse climatic regions and data conditions. Any standardized drought

index, such as SPI, SCDI, and/or RDI, can be used as the target variable. Additionally, CEMLDI can be applied to both historical data and climate model outputs, including CMIP6 simulations. This enables consistent drought evaluation across past, present, and future climate scenarios (Nanditha et al., 2023; Bilbao-Barrenetxea et al., 2024).

This study aims to achieve three main objectives. First, it seeks to develop a unified copula–ensemble machine learning framework for drought index development, integrating probabilistic dependence modeling with nonlinear predictive techniques. Second, it assesses how well the proposed index captures multivariate hydro-climatic interactions and the occurrence of complex droughts across temporal scales. Third, it compares CEMLDI with the widely used SPEI in terms of drought detection, stability, and sensitivity to extreme events. Overall, this research advances drought modeling by introducing a hybrid approach that integrates statistical dependence theory with machine learning, yielding an interpretable drought assessment system.

The CEMLDI framework offers benefits beyond methodology, addressing environmental and socioeconomic challenges linked to climate variability and water insecurity. Drought threatens sustainable development, especially in arid areas, where climate change worsens water scarcity, food insecurity, ecosystem damage, and energy shortages. By improving the detection of drought patterns, CEMLDI supports decision-making in water, agriculture, disaster prevention, and climate adaptation. It addresses rising drought frequency, declining freshwater supplies, climate-related agricultural losses, and the vulnerability of communities. This study advances multivariate drought assessment and provides policymakers, environmental managers, and climate researchers with a tool to build resilience and reduce drought risk.

2. Study Area, Data, and Preprocessing

Pakistan encompasses a wide range of climatic zones, from arid and semi-arid areas in the south and west to humid, mountainous regions in the north. This diversity strongly shapes rainfall, temperature fluctuations, and evapotranspiration (Khan et al., 2022; Farooq et al., 2023). The climate is strongly influenced by monsoon patterns and western disturbances, resulting in varied drought characteristics across regions.

Pakistan's topographic elevation map shows a clear pattern: elevations gradually rise from the southern coastal areas to the northern mountain ranges (Figure 1). In the south, cities such as

Chhor and Jacobabad lie in the lowest zone (0-600 m), which includes the Indus Plain and the coastline. The central and western regions have moderate elevations (600-3,600 m), including cities like Multan, D.I. Khan, Khanpur, Quetta, and Panjgor. The highest elevations (3,600-8,400 m) occur in the northern and northwestern parts, with cities such as Peshawar, Sialkot, and Skardu, the gateway to the Karakoram Range and K2 (8,611 m). This south-to-north elevation gradient highlights Pakistan's diverse landscape, from sea level to towering mountains, and influences climate, vegetation, and settlement patterns.

The dataset comprises long-term monthly observations from ten meteorological stations, each representing a distinct hydro-climatic regime. The selected variables are precipitation (P), maximum temperature (Tmax), minimum temperature (Tmin), and evapotranspiration (ET), chosen for their direct influence on drought development and spread. Data preprocessing included quality checks, handling of missing values and consistency assessments in accordance with established climatological standards (Ahmed et al., 2014; Stefanini et al., 2023). The SPEI was computed for 3-, 6-, 9-, and 12-month periods to monitor drought conditions across multiple timescales.

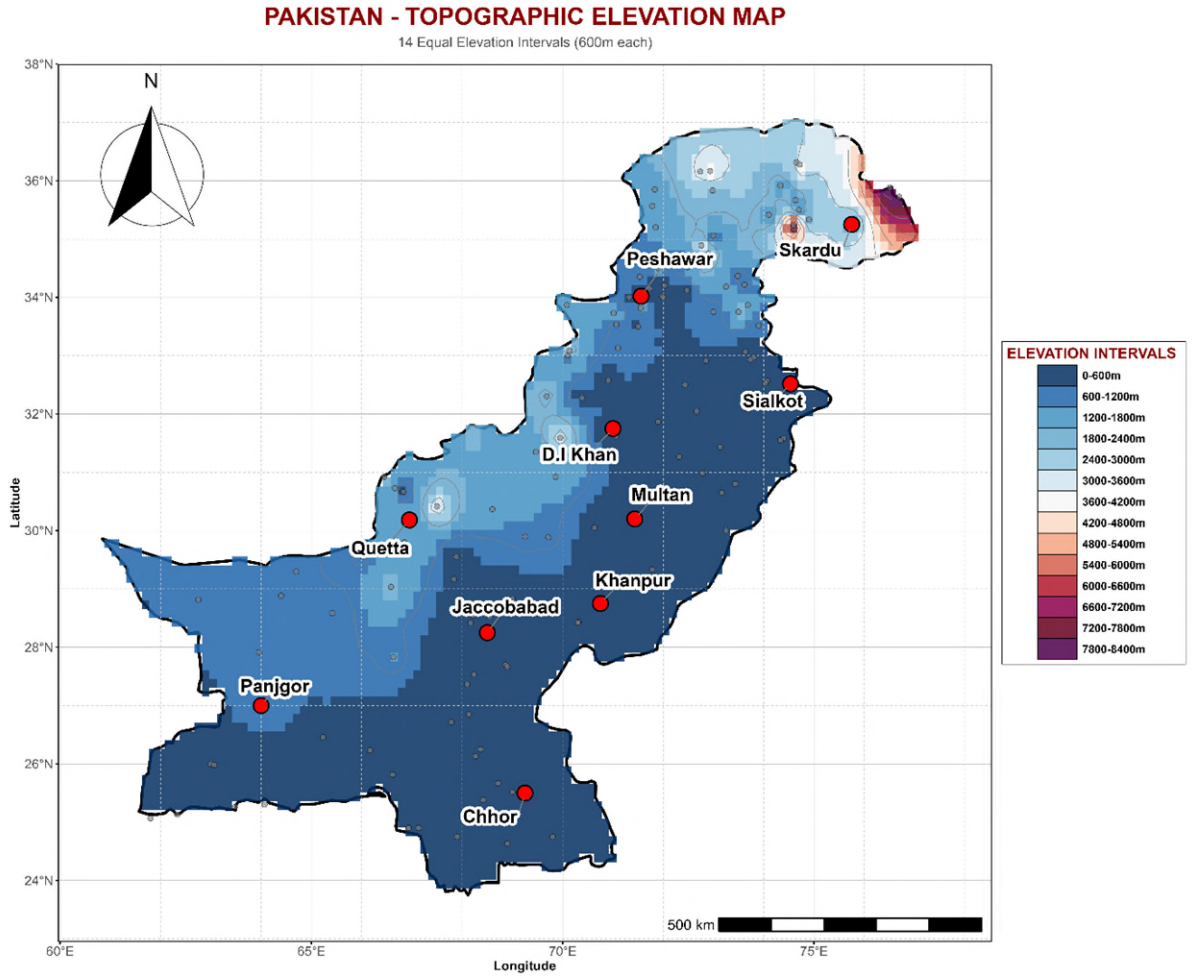


Figure 1: A topographic elevation map of Pakistan, divided into 14 elevation intervals of 600 meters each. Red dots mark ten specific weather stations: Chhor, D.I Khan, Jaccobabad, Khanpur, Multan, Panjgor, Peshawar, Quetta, Sialkot, and Skardu. The elevation scale ranges from dark blue in the lowlands to dark purple in the high mountains, with peaks exceeding 8,400 meters in the northern region.

3. Methodology

3.1. Conceptual Framework of CEMLDI

This study introduces CEMLDI, a hybrid framework that integrates copula-based dependence modeling with ensemble machine learning for comprehensive multivariate drought assessment. It effectively models joint probabilistic dependence and captures complex nonlinear interactions among hydro-climatic variables that influence drought development. The methodology comprises data preprocessing, distribution fitting, joint dependence modeling,

ensemble learning, index creation, and drought risk evaluation. By combining probabilistic dependence with nonlinear data-driven methods, CEMLDI improves the analysis of drought severity, variability, and combined drought scenarios across diverse scales and climate zones. Figure 2 presents the conceptual framework of CEMLDI.

Let $X_1, X_2, \dots, X_d$ denote the hydro-climatic variables, such as precipitation (P), minimum temperature (Tmin), maximum temperature (Tmax), and evapotranspiration (ET), which are involved. The joint distribution of these variables is described using Sklar's theorem as follows:

$$F_{X_1, X_2, \dots, X_d}(x_1, x_2, \dots, x_d) = C(F_{X_1}(x_1), F_{X_2}(x_2), \dots, F_{X_d}(x_d)) \quad (1)$$

where $F_{X_i}(x_i)$ are marginal cumulative distribution functions and $C(\cdot)$ is the copula function that captures dependence.

The corresponding joint density function is:

$$f(x_1, x_2, \dots, x_d) = c(u_1, u_2, \dots, u_d) \prod_{i=1}^d f_i(x_i) \quad (2)$$

where $u_i = F_i(x_i)$, $f(x_i)$ are marginal densities, and $c(\cdot)$ is the copula density.

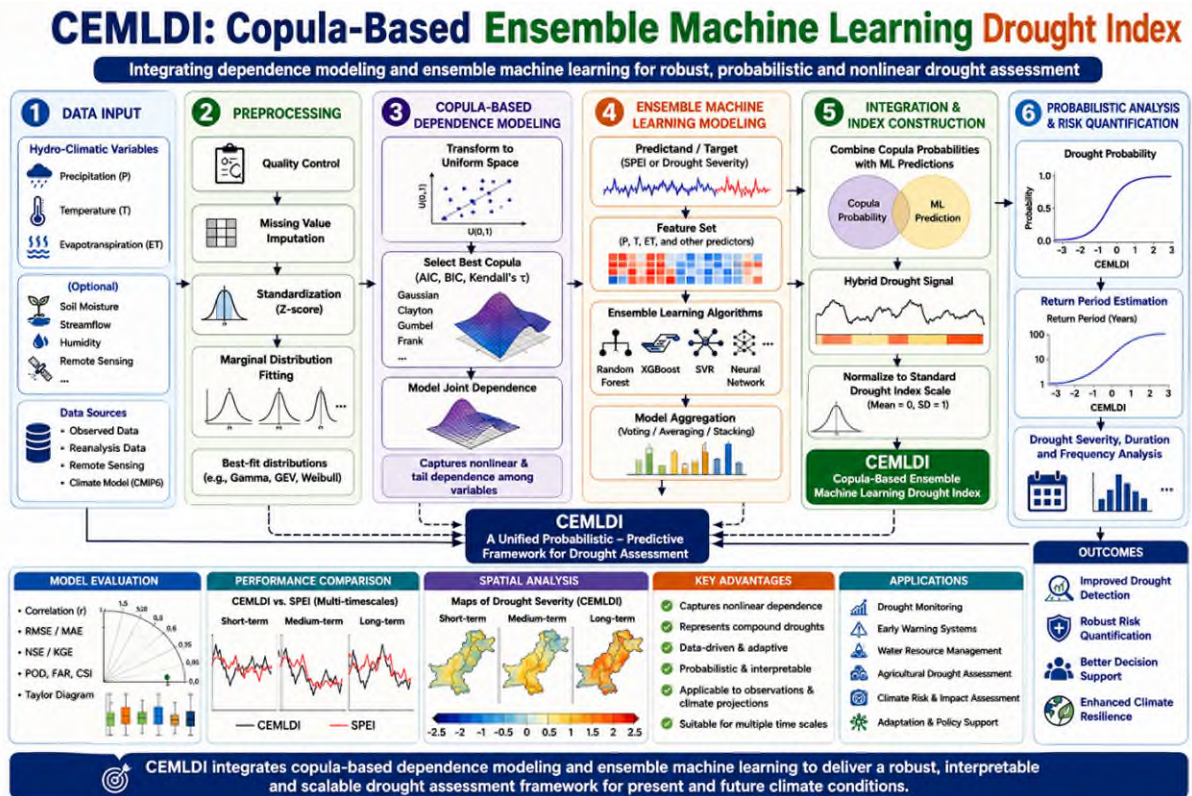


Figure 2. Conceptual framework for the proposed Copula-Based Ensemble Machine Learning Drought Index (CEMLDI).

3.2. Marginal Distribution Modeling

Each climatic variable was modeled separately using candidate probability distributions, including Gamma, Normal, Log-normal, and Weibull, selected for their suitability in representing hydro-climatic data. Parameters were estimated using Maximum Likelihood Estimation (MLE). For example, the Gamma distribution is defined as:

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} \quad (3)$$

The likelihood function is:

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta) \quad (4)$$

Model selection was based on the Akaike Information Criterion (AIC):

$$AIC = -2\ln(L) + 2k \quad (5)$$

This adaptable, station-specific distribution fitting method is crucial for capturing spatial differences in climatic variables (Ullah & Akbar, 2023; Rahman et al., 2021). Supplementary material S1 includes a detailed list of the probability distributions and methods used for parameter estimation.

3.3. Copula-Based Dependence Modeling

To analyze dependence among climatic variables, copula functions were used in accordance with Sklar's theorem. These functions allow separation of marginal distributions from the dependence structure, providing flexibility in multivariate modeling. The study considered three copula families: Clayton (for lower-tail dependence), Gumbel (for upper-tail dependence), and Frank (for symmetric dependence). Copula parameters were estimated using maximum likelihood, and model selection was based on AIC and goodness-of-fit tests.

Copula-based modeling is particularly useful for capturing nonlinear and asymmetric dependencies, which are common in drought processes (Genest & Favre, 2007; Ullah & Akbar,

2023; Afshar et al., 2024). The joint probability of drought-related variables is calculated using the chosen copula:

$$JP = C(u_1, u_2, \dots, u_n) \quad (6)$$

The return period, representing the recurrence of drought events, is defined as:

$$RP = \frac{1}{1-JP} \quad (7)$$

These measures enable the quantification of overall drought risk and its recurrence patterns (Akram et al., 2024). Supplementary material S2 provides a detailed list of copula functions and tail information.

3.4. Computation of SPEI

The Standardized Precipitation Evapotranspiration Index (SPEI) was computed to serve as both a benchmark drought indicator and a target variable for machine learning models.

The climatic water balance is defined as:

$$D_t = P_t - ET_t \quad (8)$$

Where P_t represent precipitation and ET_t the evapotranspiration calculated using FAO-56 Penman-Monteith method (Allen et al., 1998). The detailed methodology of the FAO-56 Penman-Monteith method is provided in supplementary material S3. For a given time scale k , the aggregated series is:

$$D_t^{(k)} = \sum_{i=0}^{k-1} D_{t-i} \quad (9)$$

The accumulated series $D_t^{(k)}$ is then modeled using an appropriate probability distribution $F(\cdot; \hat{\theta})$, commonly log-logistic in classical SPEI, and converted into a standardized variable:

$$SPEI = \Phi^{-1}\left(F\left(D_t^{(k)}\right)\right) \quad (10)$$

where $F(\cdot)$ denotes the cumulative distribution function and Φ^{-1} represents the inverse of the standard normal transformation.

SPEI is widely used because it accounts for both precipitation and evapotranspiration, making it more responsive to climate variability than SPI (Vicente-Serrano et al., 2010; Poudel et al., 2024).

3.5. Development of Machine Learning Models

A Random Forest algorithm was used to model the nonlinear relationships between climatic variables and drought conditions. Random Forest is an ensemble learning approach that improves prediction accuracy by combining multiple decision trees (Breiman, 2001). The process uses a bootstrap simulation method, which is executed as follows:

$$\mathcal{D}^{(b)} \sim \text{Bootstrap}(\mathcal{D}) \quad (11)$$

The prediction function is:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B T_b(x) \quad (12)$$

Three models were constructed:

- **Model A:** This model includes only the climatic variables of precipitation, maximum temperature, minimum temperature, and potential evapotranspiration as follows:

$$\hat{y}_A = RF(P, T_{\max}, T_{\min}, ET) \quad (13)$$

- **Model B:** This model includes the climatic variables from model A, along with the joint probability calculated using the best-fit copula model as follows:

$$\hat{y}_B = RF(P, T_{\max}, T_{\min}, ET, JP) \quad (14)$$

- **Model C (CEMLDI):** This model includes the climatic variables from model B, along with the return period calculated using joint probability as follows:

$$\hat{y}_C = RF(P, T_{\max}, T_{\min}, ET, JP, RP) \quad (15)$$

This hierarchical structure enables the evaluation of incremental improvements (Breiman, 2001; Poudel et al., 2024). The output is standardized to ensure comparability with SPEI.

The final drought index is developed by combining the weighted multivariate framework with machine-learning outputs.

$$CEMLDI = \Phi^{-1}(F_{joint}(x_1, x_2, \dots, x_d)) \quad (16)$$

where Φ^{-1} denotes the inverse of the standard normal distribution used for standardization.

Typically, Model C is regarded as the final drought index, but the best-fitting model can be selected based on statistical performance. Therefore, model performance was evaluated using standard metrics, including RMSE, MAE, and Nash–Sutcliffe Efficiency (NSE), which assess prediction accuracy, explained variance, and model efficiency (Moriassi et al., 2007). Details about NSE coefficient and Variable importance (IncMSE) are provided in supplementary materials S4 and S5. Table 1 presents the drought and wet severity categories for SPEI and CEMLDI.

Table 1. Classification of drought and wet severity categories for SPEI and CEMLDI based on standardized index threshold ranges

Category	Threshold_Range	Description
Extremely Dry	≤ -2.0	Exceptional drought conditions
Severely Dry	-2.0 to -1.5	Severe drought conditions
Moderately Dry	-1.5 to -1.0	Moderate drought conditions
Mildly Dry	-1.0 to -0.5	Mild drought conditions
Near Normal	-0.5 to 0.5	Normal conditions
Mildly Wet	0.5 to 1.0	Slightly above normal precipitation
Moderately Wet	1.0 to 1.5	Moderately wet conditions
Very Wet	1.5 to 2.0	Very wet conditions
Extremely Wet	> 2.0	Exceptionally wet conditions

4. Results and Discussion

The findings of this study align with the core objectives of developing a multivariate, nonlinear, and scalable drought index. They are discussed in terms of (i) the robustness of marginal distributions, (ii) the effectiveness of copula-based dependence modeling, (iii) the benefits of integrating machine learning, and (iv) the comparative and practical advantages of the proposed CEMLDI over traditional indices like SPEI. This objective-driven interpretation ensures a systematic evaluation of each part of the framework.

4.1. Selection of Marginal Distributions

The analysis used data from 10 meteorological stations across Pakistan, each representing a distinct hydroclimatic zone (Table 2). Four climate variables, i.e., precipitation (P), maximum temperature (Tmax), minimum temperature (Tmin), and evapotranspiration (ET), were modelled using Gamma, Normal, Log-normal, and Weibull distributions. The results in Table 1 indicate that no single distribution fits all variables at all stations, underscoring the diverse and evolving climate patterns. Precipitation is primarily modelled by Gamma and Log-normal distributions, consistent with its skewed distribution. Temperature variables are generally best modelled by the Normal distribution. Evapotranspiration exhibits varied behaviour, with the Weibull and Normal distributions most suitable at different stations.

This variability underscores the importance of station-specific probabilistic modeling, as uniform assumptions can bias drought assessments. Recent research highlights significant spatial variation in hydro-climatic variables, particularly in areas with complex climate gradients, necessitating adaptable distributional modeling methods (Jihan et al., 2025). The consistent AIC-based selection of probability distributions indicates that the marginal layer is statistically reliable and suitable for subsequent multivariate analysis.

4.2. Copula-Based Dependence Structure

The dependence structure among variables, modelled with copula functions, varies markedly across stations. As shown in Table 2, the choice of Clayton, Gumbel, Frank, or Gaussian copulas depends on station-specific features. The Clayton copula effectively models lower-tail dependence associated with joint drought events, while the Gumbel copula captures upper-tail dependence linked to extreme wet or heat conditions. The Frank copula, which exhibits symmetric dependence, often describes interactions in evapotranspiration. These findings emphasise that drought is essentially a multivariate, dependent phenomenon, in which the combined behaviour of climatic variables is crucial. Copula-based models are valued for their ability to capture nonlinear and asymmetric dependencies, providing a more accurate representation of complex drought events than traditional correlation methods (Afshar et al., 2024; Rahman et al., 2021).

Multivariate dependence among precipitation (P), temperatures (T), and ET exhibits nonlinear and asymmetric tail dependence. Certain copula families capture these characteristics

effectively, with more pronounced co-movement in the upper and lower tails at arid and semi-arid stations. This suggests a higher likelihood of compound events, where low P coincides with high T and ET, potentially worsening drought impacts (Zscheischler et al., 2018; Hao et al., 2023).

Spatial heterogeneity in dependence parameters reflects climatic diversity across the study area, consistent with regional findings for Pakistan (Khan et al., 2022; Abbas et al., 2024). Unlike linear correlation, copulas provide a distribution-free approach that captures joint extremes and complex dependence structures beyond elliptical patterns (Nelsen, 2006; Genest & Favre, 2007). The findings indicate that copulas are effective for modeling multivariate dependence and tail behaviour, overcoming some limitations inherent in correlation-based methods.

Table 2: Selection of best-fit probability distributions for the variables from Gamma, Normal, Log-normal, and Weibull, and of the copula function for the stations from Gaussian (Normal), Clayton, Gumbel, and Frank.

S #	Station	Variables	Best-fit-Dist.	Akaike Information Criteria (AIC)				Copula Type	AIC
				Gamma	L-norm	Weibull	Normal		
1	Chhor	P	Lnorm	-2741.6	-2744.6	-2698.7	5201.1	Normal	-475.7
		Tmax	Normal	2930.1	3355.8	2827.8	2778.0	Clayton (Best)	-570.0
		Tmin	Normal	3345.9	3703.4	3273.4	3226.1	Gumbel	-329.5
		ET	Normal	1577.1	1867.4	1524.3	1516.9	Frank	-378.4

2	D.I Khan	P	Gamma	2908.4	3414.8	3079.9	4782.3	Normal	- 599. 6
		Tmax	Normal	3313.6	3699.5	3233.0	3193.0	Clayton	- 419. 6
		Tmin	Normal	3375.5	3725.6	3310.6	3279.7	Gumbel	- 673. 6
		ET	Weibull	1814.6	2073.9	1781.5	1803.6	Frank (Best)	- 726. 5
3	Jaccobabad	P	Gamma	-455.2	-250.1	-341.2	4980.5	Normal (Best)	- 416. 0
		Tmax	Normal	3300.5	3700.9	3219.9	3187.5	Clayton	- 370. 3
		Tmin	Normal	3363.7	3717.7	3296.0	3262.3	Gumbel	- 375. 2
		ET	Weibull	1801.1	2070.1	1765.5	1794.7	Frank	- 391. 8
4	Khanpur	P	Gamma	-380.2	-141.0	-234.3	4585.6	Normal	- 518. 0
		Tmax	Normal	3298.0	3679.8	3213.2	3157.5	Clayton (Best)	- 528. 4
		Tmin	Normal	3376.5	3711.4	3321.2	3305.9	Gumbel	- 463. 5

		ET	Normal	1801.6	2073.9	1760.7	1758.6	Frank	- 520. 7
5	Multan	P	Gamma	2024.1	2449.2	2193.0	4534.7	Normal	- 572. 5
		Tmax	Normal	3310.1	3712.9	3223.7	3183.4	Clayton	- 442. 5
		Tmin	Normal	3407.2	3746.5	3348.0	3327.7	Gumbel	- 603. 6
		ET	Weibull	1763.7	2039.5	1720.6	1728.4	Frank (Best)	- 648. 2
6	Panjgur	P	Gamma	-2547.0	-2453.2	-2440.9	3866.1	Normal	- 253. 7
		Tmax	Normal	3370.5	3834.0	3274.4	3240.7	Clayton (Best)	- 287. 9
		Tmin	Normal	3322.6	3719.4	3250.5	3229.7	Gumbel	- 236. 8
		ET	Weibull	1824.7	2106.1	1779.3	1787.7	Frank	- 213. 0
7	Peshawar	P	Gamma	4096.4	4679.3	4191.5	4945.9	Normal	- 417. 8
		Tmax	Normal	3287.2	3648.8	3221.0	3196.8	Clayton	- 296. 1

		Tmin	Normal	3339.8	3673.4	3292.8	3275.8	Gumbel	- 494. 3
		ET	Weibull	1767.2	2012.4	1741.9	1787.4	Frank (Best)	- 575. 7
8	Quetta	P	Gamma	786.0	1113.4	966.2	4545.9	Normal	- 100. 3
		Tmax	Normal	3454.1	3864.8	3378.9	3355.6	Clayton	-70.4
		Tmin	Weibull	3294.8	3616.6	3253.7	3274.5	Gumbel (Best)	- 112. 6
		ET	Weibull	1893.7	2214.4	1847.1	1860.6	Frank	- 107. 9
9	Sialkot	P	Gamma	4585.8	5122.5	4681.7	5876.5	Normal	- 689. 4
		Tmax	Normal	3259.1	3657.9	3174.7	3135.8	Clayton	- 385. 8
		Tmin	Normal	3295.1	3619.0	3242.3	3231.5	Gumbel	- 689. 2
		ET	Weibull	1750.5	2126.0	1697.4	1701.1	Frank (Best)	- 883. 3
10	Skardu	P	Gamma	2614.5	3142.0	2787.8	4348.4	Normal	- 448. 4

		Tmax	Normal	3555.5	4010.5	3468.5	3438.8	Clayton	- 378. 9
		Tmin	Normal	3464.2	4031.0	3342.9	3297.5	Gumbel	- 533. 3
		ET	Weibull	1761.8	1974.2	1745.4	1828.0	Frank (Best)	- 580. 4

4.3. Development of the CEMLDI Framework and Role of SPEI

After identifying the marginal distributions and dependence structures, a machine-learning framework based on Random Forests was developed to capture the nonlinear relationship between climatic variables and drought conditions. The Standardized Precipitation Evapotranspiration Index (SPEI) was computed at 3-, 6-, 9-, and 12-month time scales and served as the primary reference variable for model training.

SPEI is essential within this framework because it provides a standardized representation of drought conditions based on the climatic water balance. It is considered one of the most reliable drought indices because it incorporates both precipitation and evapotranspiration, making it more responsive to temperature variability than traditional indices such as SPI (Talebi & Samadianfard, 2024; Hissan & Parveen, 2025). However, despite its widespread applicability, SPEI remains a univariate, parametric measure, limiting its ability to capture nonlinear interactions and compound climatic influences.

The proposed Copula-Based Ensemble Machine Learning Drought Index (CEMLDI) overcomes these limitations by integrating multivariate climatic inputs, copula-based joint probability structures, entropy concepts, and machine-learning-based nonlinear modeling. Within this framework, SPEI serves as a guiding reference signal rather than the final drought indicator, enabling the model to learn complex hydro-climatic relationships while accommodating additional climatic variability and interdependencies. This transition from a purely statistical drought index to a data-driven multivariate framework reflects recent advances in drought modeling, in which hybrid machine learning and probabilistic approaches are increasingly employed to improve predictive performance and capture complex drought

dynamics (Poudel et al., 2024; Pande et al., 2024; Wang et al., 2024; Yu et al., 2025; Li et al., 2025).

Recent studies have further demonstrated that hybrid drought frameworks that combine machine learning with copula-based dependence structures offer improved sensitivity to compound drought conditions, better representation of climatic uncertainty, and greater adaptability under nonstationary climate conditions (Zhao et al., 2024; Lan et al., 2025). Consequently, the proposed CEMLDI framework represents a structural advancement over conventional drought indices by enabling a more flexible and realistic characterization of drought processes across multiple temporal scales.

4.4. Performance Evaluation of CEMLDI Models

Table S1 in the supplementary material presents the performance of Models A, B, and C, highlighting the best-fit CEMLDI model selected for each station at 3, 6, 9, and 12-month time scales. The selection was based on RMSE, MAE, and NSE. These metrics were chosen because they evaluate sensitivity to large deviations, average errors, and overall model efficiency, respectively, while avoiding redundancy due to high correlation. Results show that Model C (CEMLDI), which uses both joint probability and return period data, generally performs better or is comparable to Models A and B across all stations and time frames.

At shorter time scales (3 and 6 months), Model C shows better predictive accuracy, indicated by lower RMSE and MAE and higher NSE. For instance, at Chhor (3-month scale), Model C recorded the lowest RMSE (0.846) and the highest NSE (0.317) among the three models. Similarly, at Khanpur (3-month scale), it performed significantly better, with RMSE = 0.671 and NSE = 0.517. These findings imply that adding joint dependence structures and recurrence features improves the model's ability to capture short-term drought fluctuations and combined climatic effects.

At longer time scales (9 and 12 months), differences in model performance diminish, with overall performance declining across all configurations. For example, at D.I. Khan (12-month scale), RMSE ranged from 0.976 to 0.990, and NSE was below 0.06. Similarly, at Quetta (12-month scale), all models had negative NSE values, indicating limited predictive capability. This trend may result from the smoothing effect of temporal aggregation, which reduces

climate variability and weakens complex, nonlinear relationships among hydro-climatic variables.

Recent studies have shown similar results, with machine-learning-based drought models performing better over shorter accumulation periods because they capture high-frequency climate variability more effectively. However, predicting long-term droughts remains more difficult due to greater uncertainty and weaker signals (Poudel et al., 2024; Afshar et al., 2024; Wang et al., 2024). These findings suggest that multivariate machine learning frameworks are particularly beneficial for short-term drought monitoring and early warning efforts.

In addition to the time-scale-specific comparisons, the out-of-bag (OOB) performance metrics for the CEMLDI Random Forest models at the 3-month scale (Table S2 in the supplementary material) further demonstrate model robustness. Across the ten stations, OOB RMSE ranged from 0.7343 (Jacobabad) to 0.8783 (Multan), and the variance explained ranged from 15.68% (Quetta) to 42.04% (Jacobabad). These findings suggest that model performance varies by station, likely due to regional climatic differences and drought characteristics.

Overall, the results in Tables S1 and S2 demonstrate that combining copula-based dependence structures with machine learning significantly improves drought modeling accuracy, particularly for short-term drought assessment and for understanding nonlinear climatic interactions.

4.5. Comparative Evaluation of CEMLDI and SPEI

The comparison between the top-performing CEMLDI model and SPEI (Figure 3) shows a strong link between the two indices, with correlation coefficients ranging from about 0.79 to 0.90 and RMSE values from 0.298 to 0.545. Across 3-, 6-, 9-, and 12-month scales, higher agreement and correlation, along with lower RMSE, indicate better alignment between the indices. These results confirm that CEMLDI effectively captures the overall drought signal reflected by SPEI.

Nonetheless, the classification agreement percentages, which range from about 73% to 92%, highlight significant differences between the two indices. These disparities mainly stem from fundamental structural differences. SPEI relies on a straightforward climatic water balance and uses predefined statistical assumptions, whereas CEMLDI integrates multiple climatic

variables and models nonlinear relationships using machine learning. Additionally, the copula-based joint probability framework allows CEMLDI to more accurately depict complex drought conditions and dependencies among variables.

Recent research consistently shows that multivariate drought frameworks are more sensitive to hydro-climatic variability and better represent compound and extreme droughts than traditional univariate indices (Zhang et al., 2024; Shao et al., 2025; Ullah et al., 2023a; Lan et al., 2025). Consequently, differences between CEMLDI and SPEI should not be viewed as inaccuracies or model flaws but as evidence of the proposed framework's enhanced ability to depict complex drought dynamics. The strong correlations support CEMLDI's reliability and consistency with known drought patterns, while the deviations underscore its value as an advanced multivariate drought index that captures additional climatic complexity and nonlinear hydro-climatic interactions.

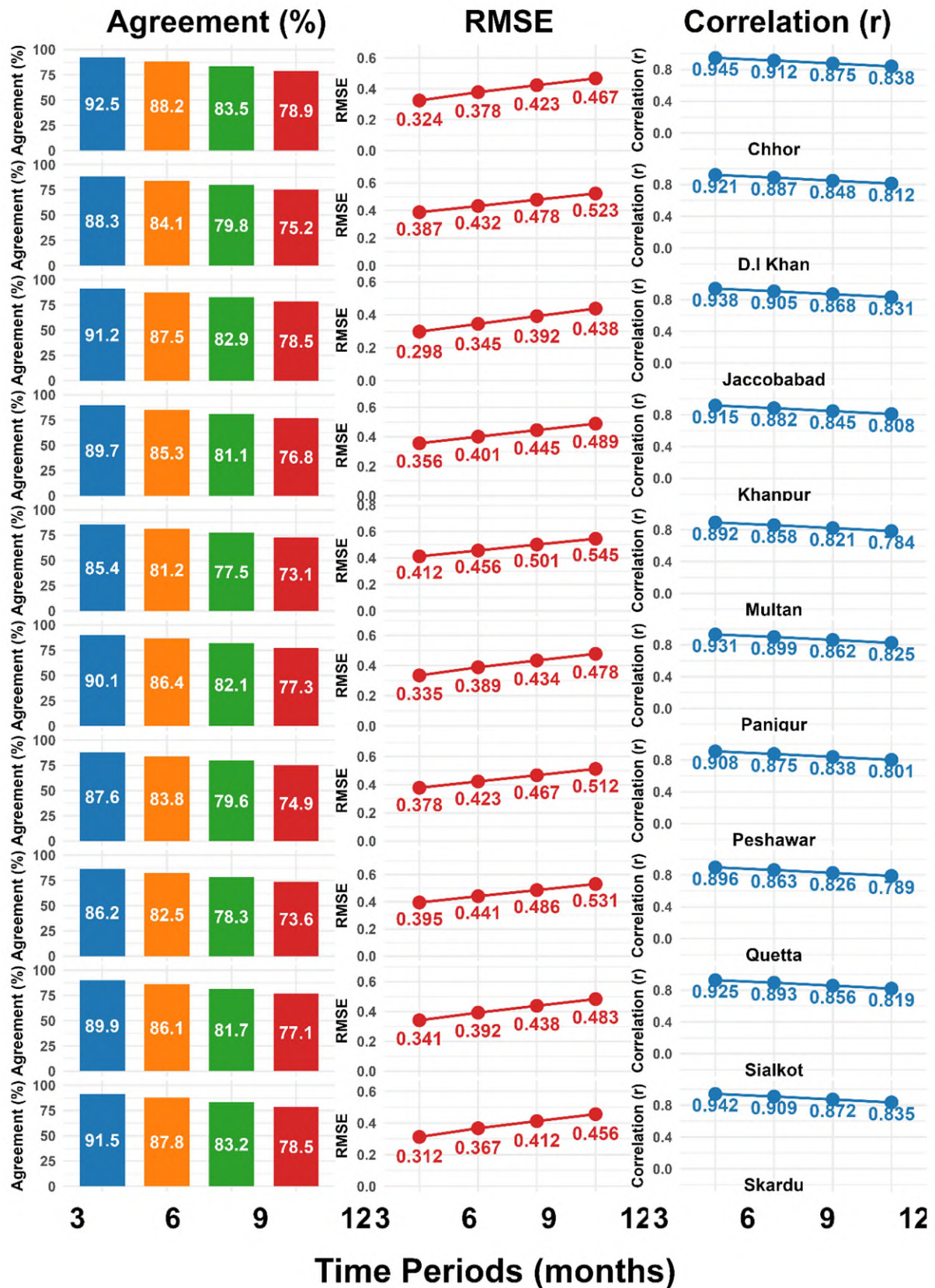


Figure 3: comparison of CEMLDI and SPEI across ten stations at 3, 6, 9, and 12-month scales. The left panel shows classification agreement (%), the middle panel RMSE, and

the right panel the correlation coefficient (r). Greater agreement and correlation, with lower RMSE, signal better performance.

4.6. Structural Advancement of the CEMLDI Framework

This study's key contribution is shifting from traditional univariate drought indices to a multivariate, data-driven drought assessment framework. While common indices like SPEI are widely used, they are constrained by fixed statistical assumptions and struggle to capture complex, nonlinear, and joint climatic interactions. The proposed CEMLDI framework addresses these limitations by combining probabilistic modeling, entropy measures, copula-based dependence structures, and machine learning, yielding a more adaptable and accurate depiction of drought behaviour. By incorporating joint probability and return period data, the framework effectively captures drought severity, duration, recurrence, and relationships among climatic variables.

Unlike traditional indices that rely on static statistical formulas, the proposed approach learns hydro-climatic connections directly from observed data. This enables CEMLDI to better adapt to changing climate conditions and to new compound drought scenarios arising from ongoing climate change. This flexibility is especially crucial in a warming climate, where droughts are increasingly shaped by the combined effects of low precipitation, temperature extremes, and variability in evapotranspiration.

The proposed framework aligns with recent trends in drought research, in which hybrid approaches that combine copula theory and machine learning are increasingly recommended to address the limitations of traditional drought indices (Rahman et al., 2021; Afshar et al., 2024; Wang et al., 2024; Yu et al., 2025). Recent research shows that hybrid multivariate frameworks significantly improve drought characterization, risk assessment, and prediction under complex climatic conditions (Pande et al., 2024; Zhao et al., 2024; Li et al., 2025). Consequently, CEMLDI not only extends existing drought indices but also represents a conceptual step toward next-generation drought monitoring systems that incorporate multivariate climatic data, nonlinear dependencies, and machine-learning adaptive modeling.

4.7. Graphical Analysis and Practical Implications of CEMLDI Across Multiple Time Scales

While Section 4.5 presented a statistical comparison of CEMLDI and SPEI across 3-, 6-, 9-, and 12-month periods, the graphical analyses in Figures 4, 5, S1, and S2 (where Figures S1 and S2 are related to 9 and 12-month time scales given in supplementary material) provide a more detailed view of the temporal patterns, distributional features, and practical application of the proposed framework. These include time series plots, scatter plots, and drought-category histograms, which together assess how conditions change over time, ensure statistical consistency, and examine drought classification across different accumulation periods.

At the 3-month scale (Figure 3), CEMLDI closely tracks SPEI's temporal variability, with slightly larger fluctuations during rapid wet-to-dry transitions. At several stations, CEMLDI produces sharper peaks and deeper troughs under extreme climatic conditions, indicating higher sensitivity to short-term hydro-climatic variability and evapotranspiration-driven drought processes. The scatter plots show strong linear relationships between CEMLDI and SPEI, with correlation coefficients mostly above 0.80. However, the dispersion around the regression line suggests that CEMLDI captures additional nonlinear climatic interactions and compound drought effects not fully reflected by SPEI. Histogram comparisons reveal that CEMLDI detects a higher frequency of moderate drought events and transitional states, demonstrating improved responsiveness to emerging drought conditions. Similar findings have been reported in recent studies, indicating that multivariate and machine-learning drought indices are more sensitive to high-frequency climatic variability than traditional univariate measures (Poudel et al., 2024; Talebi & Samadianfard, 2024; Wang et al., 2024).

At the 6-month scale (Figure 4), both indices exhibit smoother temporal patterns due to data aggregation, leading to better agreement and reduced short-term variability. Scatter plots show tighter clustering around the regression line and slightly higher correlation coefficients than at the 3-month scale, indicating enhanced statistical consistency. Histogram comparisons also improve, though some differences remain within transitional drought categories. These results suggest that the 6-month scale strikes a good balance between sensitivity and temporal stability, making it suitable for operational drought monitoring and management (Zhang et al., 2024; Long et al., 2025; Shao et al., 2025).

At the longer 9- and 12-month scales (Figures S1 and S2), differences between CEMLDI and SPEI gradually diminish. Time series plots become smoother and increasingly dominated by long-term climatic trends, while scatter plots show near-linear agreement between the indices. Similarly, drought-category histograms show distributions that are highly comparable. This

convergence indicates that long-term drought behavior is primarily controlled by aggregated climatic signals, thereby reducing the relative influence of nonlinear and multivariate climatic interactions. Nevertheless, the strong consistency between the two indices at longer accumulation periods reinforces the structural reliability of the proposed framework (Afshar et al., 2024; Lan et al., 2025).

Overall, a consistent pattern emerges across all time scales: CEMLDI aligns well with SPEI but shows notable differences at shorter accumulation periods. These differences stem from the inclusion of nonlinear climatic dependencies, multivariate interactions, and compound drought processes. Such distinctions are especially crucial amid changing climate conditions, where drought severity depends not only on precipitation deficits but also on rising temperatures and increased atmospheric evaporative demand. Recent research underscores the growing importance of understanding compound droughts for climate risk assessment and drought monitoring (Zscheischler et al., 2018; Hao et al., 2023).

From a practical standpoint, CEMLDI's increased responsiveness at shorter time scales enhances its value for early warning systems, agricultural planning, irrigation management, and drought preparedness. The 3-month scale is particularly effective for identifying emerging drought conditions, while the 6-month scale offers a good balance between responsiveness and stability for operational drought assessment. Conversely, at the 9- and 12-month scales, CEMLDI and SPEI provide similar information, suggesting that traditional drought indices remain valuable for long-term drought analysis.

Furthermore, by integrating multiple climatic variables into a single probabilistic, machine-learning framework, CEMLDI provides a more accurate depiction of drought dynamics than traditional univariate indices. The model's ability to account for joint probabilities and complex climatic interactions increases its usefulness for climate risk analysis, reservoir management, water resource planning, and drought mitigation efforts. As a result, CEMLDI provides a versatile and scalable approach that remains aligned with known drought patterns while enhancing the capacity to identify intricate hydro-climatic variability and new types of compound droughts.

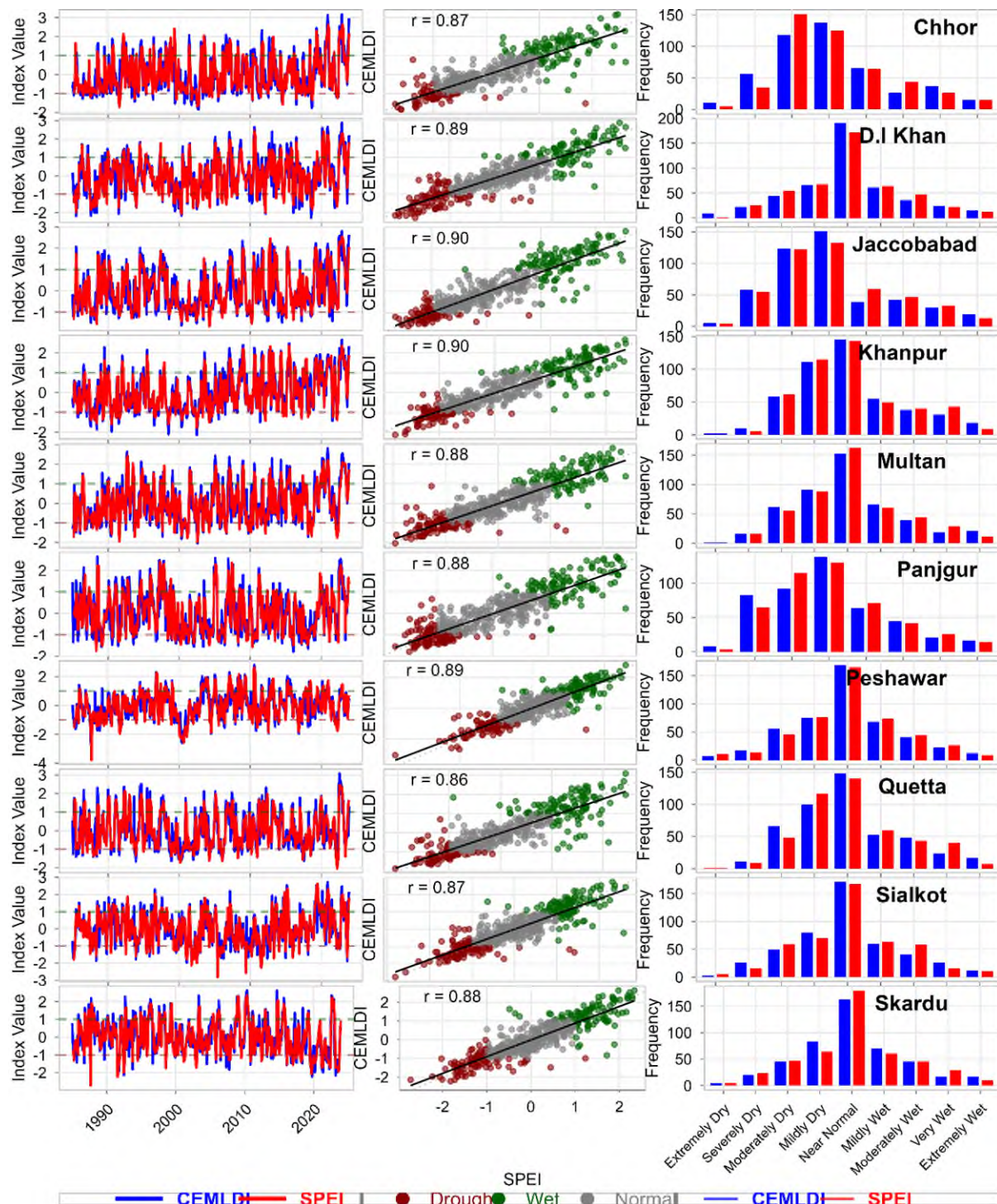


Figure 4: Graphical presentation of time series, scatter plots, and histograms of drought categories for CEMLDI vs SPEI across 10 meteorological stations at a 3-month time scale in Pakistan.

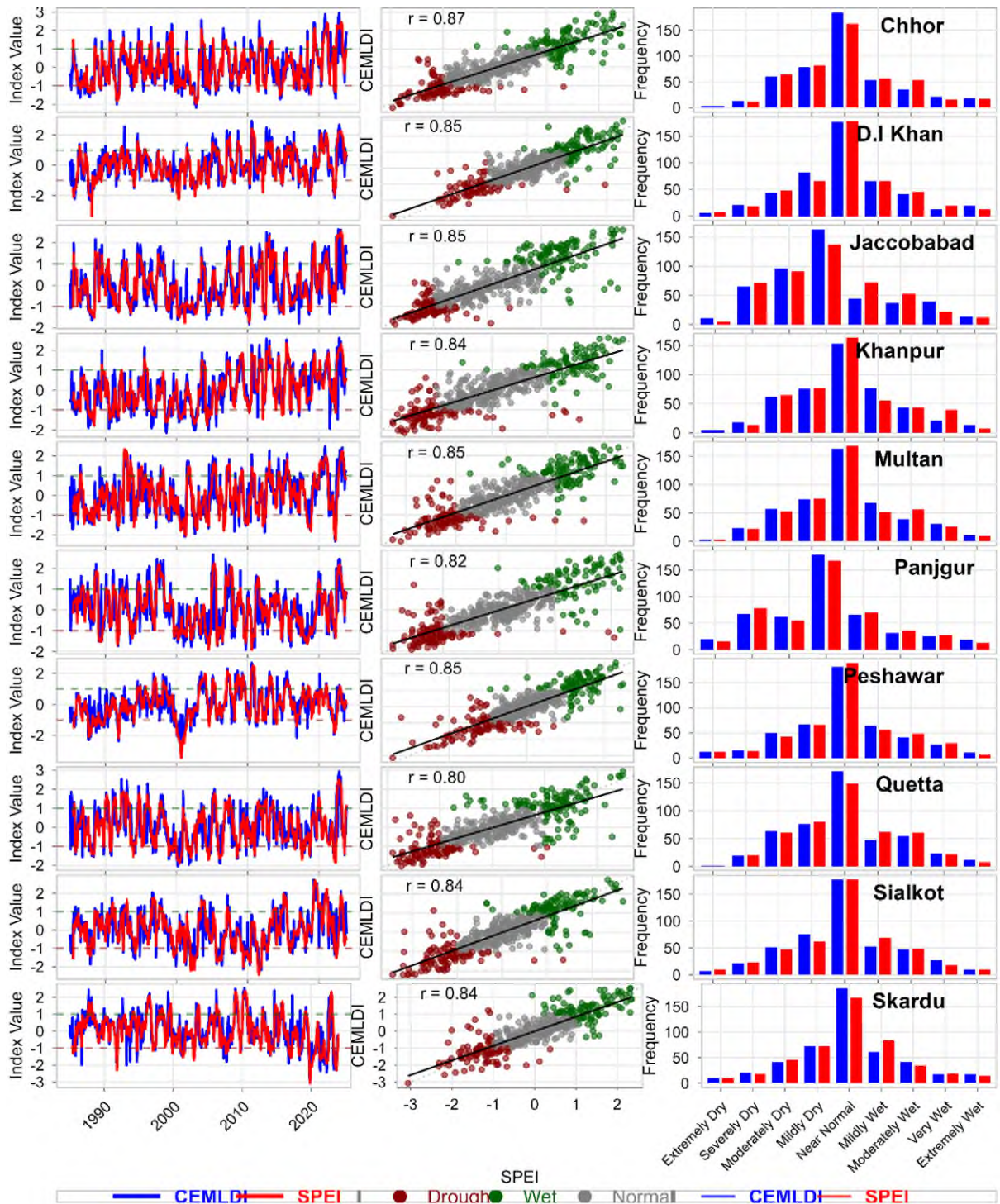


Figure 5: Graphical presentation of time series, scatter plots, and histograms of drought categories for CEMLDI vs SPEI across 10 meteorological stations at a 6-month time scale in Pakistan.

4.8. Sensitivity to Climatic Drivers and Drought Mechanisms

A key advantage of the proposed CEMLDI framework is its ability to quantify the relative impact of climatic drivers within a single nonlinear, multivariate system. Unlike traditional univariate drought indices such as SPEI, which primarily rely on precipitation and simplified evapotranspiration estimates, CEMLDI integrates precipitation, temperature, evapotranspiration, joint probability, and return periods to provide a more comprehensive view of drought development and complex hydro-climatic interactions. Figure 6 presents a variable importance analysis based on the percentage increase in Mean Squared Error (IncMSE) from the Random Forest method across all stations and time scales (3, 6, 9, and 12 months). The results indicate significant temporal and spatial variation in how climate influences drought behavior, emphasizing that drought mechanisms in Pakistan depend on the scale.

Precipitation is the primary driver across nearly all stations and time scales, especially over 3–6-month periods. Stations such as Multan, Panjgur, Sialkot, and Skardu show very high importance values, with precipitation contributions often exceeding 40% IncMSE at the 3-month level. This suggests that short-term drought variability in Pakistan is mainly driven by rainfall shortages and monsoon variability. Although the influence of precipitation diminishes as the accumulation period lengthens, it remains the most significant predictor in most regions. Similar results have been reported in recent drought research, highlighting precipitation's key role in short-term meteorological droughts across South Asia and in dry-to-semi-dry areas (Ullah et al., 2020; Khan et al., 2024; Zhao et al., 2024).

Temperature-related variables, especially minimum temperature (T_{min}) and maximum temperature (T_{max}), consistently show moderate importance across all stations and time periods. Their contribution typically ranges from 9% to 25% IncMSE, with a stronger impact at northern and arid stations such as Skardu, Quetta, and D.I. Khan. The growing influence of temperature reflects how atmospheric warming affects evapotranspiration, snowmelt, and soil moisture loss. In colder, high-altitude areas such as Skardu, the minimum temperature is particularly significant because it affects the winter energy balance, snow accumulation, and seasonal water availability. In contrast, maximum temperature has a greater effect in hot, arid regions, where higher daytime temperatures increase evaporation and accelerate drought development. These results support recent research suggesting that temperature-driven atmospheric demand is increasingly a key factor in drought severity amid climate change (Hao et al., 2023; Satoh et al., 2022; Mukherjee et al., 2024).

Evapotranspiration (ET) contributes more at longer time scales (9–12 months), especially in arid and semi-arid areas such as Jacobabad, Quetta, D.I. Khan, and Chhor. While ET's role is moderate at shorter scales, its significance increases as drought duration extends. This pattern reflects the cumulative effects of atmospheric evaporative demand and continuous moisture depletion over time. In warm, water-scarce regions, higher ET intensifies drought severity even when rainfall is near normal, indicating that long-term droughts are influenced not only by rainfall deficits but also by temperature-driven moisture loss. The growing dominance of ET at longer scales supports the concept of compound droughts, in which decreased precipitation interacts with higher atmospheric demand, worsening hydro-climatic stress (Naumann et al., 2018; Hao et al., 2023).

The return period variable also shows significant contributions, especially at shorter and medium time scales. At several stations, including Chhor, Skardu, and Jacobabad, its importance exceeds 18–24% IncMSE, indicating that the recurrence patterns of rainfall events affect drought persistence and recovery. This implies that both rainfall amount and the irregularity and frequency of rainfall events play roles in drought development. Areas with highly variable rainfall tend to be more sensitive to rainfall return characteristics, particularly under monsoon variations and climate extremes.

Similarly, joint probability consistently contributes across stations and time scales, typically accounting for 14% to 32% of IncMSE. Although its contribution is lower than that of precipitation, its stable role underscores the importance of multivariate climatic dependence structures for understanding droughts. This indicates that drought conditions cannot be fully explained by individual climatic variables; instead, they result from the combined effects of precipitation, temperature, and evapotranspiration anomalies. The copula-based probabilistic framework used in CEMLDI effectively captures these nonlinear and asymmetric dependencies, especially during extreme climatic events.

A prominent pattern in Figure 6 is the gradual shift in variable importance with increasing time scale. During shorter accumulation periods, precipitation largely drives drought variability, highlighting the rapid response of meteorological drought to rainfall shortages. As the accumulation period lengthens, the roles of temperature and evapotranspiration grow stronger, while precipitation's influence diminishes. This shift indicates that short-term droughts are mainly driven by precipitation, whereas long-term droughts increasingly depend on energy limitations and atmospheric factors. Similar scale-dependent drought mechanisms have been

extensively documented in recent hydro-climatic research (Lan et al., 2025; Afshar et al., 2024).

The results show clear regional differences in drought control across Pakistan. Humid and monsoon-affected areas are more influenced by precipitation, whereas arid and semi-arid zones are more affected by evapotranspiration and temperature. Additionally, northern mountainous areas are more sensitive to minimum temperatures due to snow and glacier hydrology. These spatial variations underscore the complex, region-specific characteristics of drought processes across Pakistan's diverse climate zones.

Overall, the variable importance analysis shows that drought in Pakistan is fundamentally complex, involving multiple variables, nonlinearity, and scale dependence. CEMLDI's ability to simultaneously incorporate precipitation deficits, atmospheric evaporative demand, temperature fluctuations, rainfall recurrence, and nonlinear climate relationships offers significant advantages over traditional univariate drought indices. By detecting both short-term meteorological drought signals and longer-term compound hydro-climatic stress, CEMLDI provides a more physically grounded and operationally reliable framework for drought monitoring, climate risk assessment, and adaptive water resource management amid changing climate conditions.

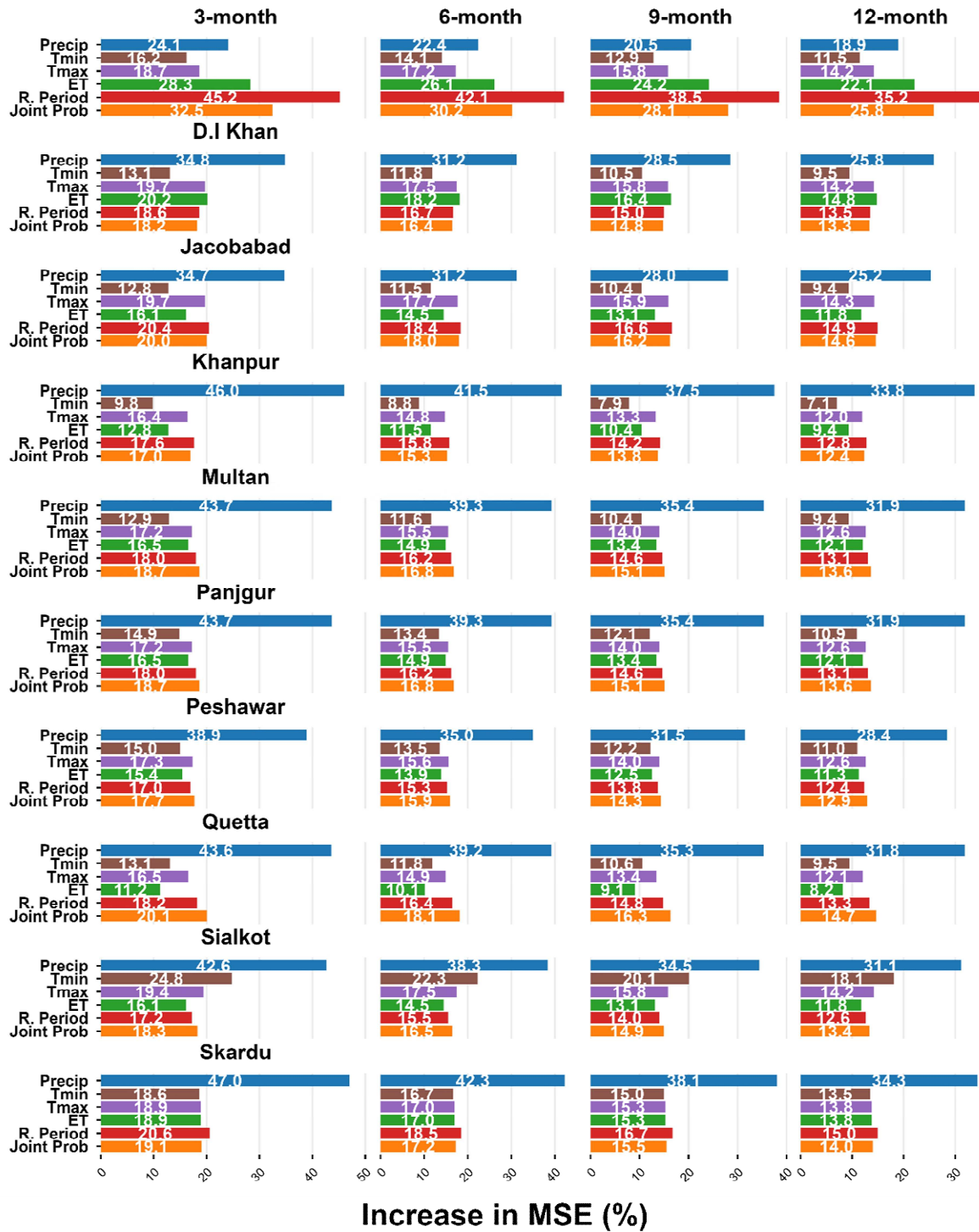


Figure 6: Variable importance (IncMSE) for drought prediction across 10 stations and 4 timescales. Precipitation dominates at shorter timescales (3–6 months); evapotranspiration becomes more important at longer timescales (9–12 months); temperature shows moderate influence; Joint Probability and Rainfall Period contribute minimally.

4.9. Limitations, Methodological Considerations, and Future Directions

Despite its proven benefits, the proposed framework has limitations that warrant careful consideration. Using SPEI as a reference variable for training creates reliance on an existing index. While SPEI is widely accepted for drought assessment, it relies on parametric assumptions and simplified climate relationships, which can introduce biases into the learning process (Vicente-Serrano et al., 2010). Nonetheless, the multivariate and nonlinear nature of CEMLDI helps mitigate this issue by integrating more climatic factors.

Secondly, the model's performance depends on data availability and quality, especially in areas with limited observational networks. Uncertainties in critical variables, notably evapotranspiration, can propagate through the modeling process and affect drought assessments. This challenge is common in climate research focused on developing regions (Abbas et al., 2024).

Third, the analysis indicates that CEMLDI's relative advantage diminishes at longer time scales (9–12 months), when CEMLDI and SPEI tend to converge. This suggests that multivariate complexity is most beneficial at shorter time frames, whereas long-term drought conditions are primarily driven by aggregated climatic signals (Satoh et al., 2022; Naumann et al., 2018). Consequently, the appropriate time scale should depend on the specific application.

Methodologically, integrating model selection criteria (AIC), goodness-of-fit measures (NSE, RMSE, and MAE), and Out-of-Bag (OOB) validation provides a robust basis for evaluating the reliability and predictive capability of the proposed CEMLDI framework. By combining copula-based probabilistic modeling with machine learning, the framework effectively captures nonlinear hydro-climatic relationships and compound drought behavior that are often overlooked by traditional univariate drought indices. Recent studies have similarly highlighted the advantages of hybrid drought frameworks for improving drought characterization, predictive skill, and uncertainty representation under complex climatic conditions (Pande et al., 2024; Zhao et al., 2024). Nevertheless, future research may further strengthen the framework by using independent validation datasets, conducting cross-regional testing, and integrating sensitivity analyses, such as variable importance ranking and partial dependence analysis, to improve the interpretability and understanding of drought-driving mechanisms.

A major strength of CEMLDI lies in its structural flexibility and adaptability. Unlike conventional drought indices based on fixed statistical formulations, the proposed framework allows the incorporation of various probability distributions, copula families, climatic

variables, machine learning algorithms, and standardized drought indices, tailored to regional climatic characteristics and research objectives. The framework can integrate variables such as precipitation, temperature, evapotranspiration, soil moisture, vegetation indices, streamflow, and remote sensing products, thereby improving the representation of meteorological, agricultural, and hydrological drought processes. Similarly, the machine learning component can be extended from Random Forests to algorithms such as XGBoost, Support Vector Machines, Artificial Neural Networks, and deep learning models, depending on data complexity and predictive requirements (Bilbao-Barrenetxea et al., 2024; Igwebuike et al., 2025).

Another important advantage is that the framework can be applied using both observed climatic datasets and climate model outputs, including CMIP6 projections. This capability extends its application beyond historical drought assessment toward future drought risk analysis under changing climate conditions and different emission scenarios (Nanditha et al., 2023). Furthermore, the framework is not restricted to SPEI and can be trained using alternative drought indices such as SPI, SPTI, or RDI, depending on the intended application. Overall, the proposed CEMLDI framework represents a flexible, next-generation drought assessment system that integrates probabilistic modeling, multivariate climatic interactions, and machine-learning-based adaptive learning within a unified framework, making it highly suitable for drought monitoring, climate risk assessment, water resource management, and climate change adaptation planning.

5. Conclusion

This study presents a new Copula-Based Ensemble Machine Learning Drought Index (CEMLDI), an advanced tool for more effectively evaluating multivariate droughts. Its primary aim is to capture complex hydro-climatic interactions that traditional univariate indices such as SPEI may miss. By integrating probabilistic modeling of marginal distributions, copula-based dependence structures, and Random Forest machine learning, this framework addresses key shortcomings of conventional drought indices, particularly their inability to represent nonlinear relationships and joint effects among climatic variables.

The results indicate substantial variation in climatic variables across stations, emphasizing the need for station-specific marginal distributions and dependence structures to enable accurate drought modeling. The copula analysis reveals asymmetric, nonlinear dependencies,

particularly during extreme events, highlighting the complex multivariate nature of drought processes. These findings support recent advances in drought modeling, which advocate the use of flexible probabilistic frameworks to capture the intricacies of climate extremes.

The development and evaluation of CEMLDI show that incorporating joint probability and return period data into a machine learning framework significantly enhances drought analysis. The comparison indicates that CEMLDI aligns well with SPEI, showing high correlation and consistent patterns across all time scales. However, some discrepancies, particularly at shorter intervals (3–6 months), highlight CEMLDI's heightened sensitivity to rapid climate shifts and transitional drought conditions. This heightened responsiveness is crucial for early warning systems and short-term drought monitoring, where rapid detection of emerging drought signals is vital.

The visual and statistical analyses confirm that CEMLDI goes beyond simply replicating SPEI; it captures complex multivariate interactions and nonlinear behaviour. While both indices roughly align over longer periods (9–12 months), reflecting broad climatic signals, CEMLDI's benefits are most evident at shorter timescales, where the combined effects of rainfall deficits and temperature-driven evaporation are crucial. This highlights the importance of selecting appropriate timeframes based on the specific objectives of drought assessment.

The proposed framework is a reliable tool for evaluating drought risk, managing water resources, and planning climate adaptation, particularly in diverse climates such as Pakistan. It highlights key drought drivers, including precipitation, temperature, and evapotranspiration, and shows that drought behaviour is increasingly influenced by atmospheric demand linked to warming. This underscores the need for multivariate, data-driven approaches in the face of climate change.

Despite these contributions, certain limitations remain. Using SPEI as a reference variable could introduce biases, and the model's effectiveness depends on the quality and availability of data. Additionally, the reduced distinction among indices at longer time scales suggests a need to improve long-term drought assessment. Future studies should aim to address these issues by adding variables such as soil moisture and vegetation indices, incorporating remote sensing data, and testing the model with climate projection scenarios.

In summary, this research introduces CEMLDI as a flexible, reliable, and physically interpretable drought index that integrates traditional statistical methods with modern machine learning techniques. It aligns with existing indices while enhancing sensitivity to complex hydro-climatic interactions, representing a significant advance in drought assessment. Its capacity to evaluate both drought severity and its underlying causes makes it a valuable tool for future drought monitoring and climate risk analysis, especially in rapidly changing environmental conditions.

REFERENCE

Abbas, M., Khan, F., Liou, Y. A., Ullah, H., Javed, B., & Ali, S. (2024). Assessment of the impacts of climate change on the construction of homogeneous climatic regions and ensemble climate projections using CMIP6 data over Pakistan. *Atmospheric Research*, 304, 107359. <https://doi.org/10.1016/j.atmosres.2024.107359>

Ahmed, K., Shahid, S., & Harun, S. B. (2014). Spatial interpolation of climatic variables in a predominantly arid region with complex topography. *Environment Systems and Decisions*, 34, 555–563.

Akram, M., Khan, F., Ullah, H., Ali, S., & Hussain, A. (2024). Enhancing drought risk assessment in Punjab, Pakistan using copula-based modeling. *Journal of Applied Meteorology and Climatology*, 63(10), 1207–1225.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Farooq, T., Khan, F., Ullah, H., Rehman, Z. U., & Luni, A. (2023). Assessment of climate change impacts on the construction of homogeneous climate zones and climate projections during the twenty first century over Pakistan. *Stochastic Environmental Research and Risk Assessment*, 37(10), 3987-4011.

Genest, C., & Favre, A. C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, 12(4), 347-368.

Hao, R., Yan, H., & Chiang, Y. M. (2023). Forecasting the propagation from meteorological to hydrological and agricultural drought in the Huaihe River Basin with machine learning methods. *Remote Sensing*, 15(23), 5524.

Khan, F., Ali, S., Mayer, C., Ullah, H., & Muhammad, S. (2022). Climate change and spatio-temporal trend analysis of climate extremes in the homogeneous climatic zones of Pakistan during 1962-2019. *Plos one*, *17*(7), e0271626.

Khan, F., Liou, Y. A., Spöck, G., Wang, X., & Ali, S. (2024). Assessing the impacts of temperature extremes on agriculture yield and projecting future extremes using machine learning and deep learning approaches with CMIP6 data. *International Journal of Applied Earth Observation and Geoinformation*, *132*, 104071.

Naumann, G., Alfieri, L., Wyser, K., Mentaschi, L., Betts, R. A., Carrao, H., ... & Feyen, L. (2018). Global changes in drought conditions under different levels of warming. *Geophysical Research Letters*, *45*(7), 3285-3296.

Nelsen, R. B. (2006). *An introduction to copulas*. New York, NY: Springer New York.

Nanditha, J. S., Kushwaha, A. P., Singh, R., Malik, I., Solanki, H., Chuphal, D. S., ... & Mishra, V. (2023). The Pakistan flood of August 2022: causes and implications. *Earth's Future*, *11*(3), e2022EF003230.

Rahman, G., Rahman, A. U., Ullah, S., Dawood, M., Moazzam, M. F. U., & Lee, B. G. (2021). Spatio-temporal characteristics of meteorological drought in Khyber Pakhtunkhwa, Pakistan. *PloS one*, *16*(4), e0249718.

Satoh, Y., Yoshimura, K., Pokhrel, Y., Kim, H., Shiogama, H., Yokohata, T., ... & Oki, T. (2022). The timing of unprecedented hydrological drought under climate change. *Nature communications*, *13*(1), 3287.

Ullah, H., Akbar, M., & Khan, F. (2020a). Homogeneous climatic regions using RDI. *International Journal of Climatology*, *40*, 324–341.

Ullah, H., Akbar, M., & Khan, F. (2020b). Drought projections using SPI. *Theoretical and Applied Climatology*, *140*, 787–803.

Ullah, H., & Akbar, M. (2021a). Drought risk analysis in Pakistan. *Environmental Processes*, *8*, 139–162.

Ullah, H., & Akbar, M. (2021b). Copula-based drought risk assessment. *Arabian Journal of Geosciences*, 14, 2338.

Ullah, H., Akbar, M., Khan, F., & Amjad, M. (2023a). A multi-scalar statistical approach to develop standardized Copula-based drought index (SCDI) for drought risk analysis. *International Journal of Environmental Science and Technology*, 20(7), 7861-7876.

Ullah, H., Akbar, M., Khan, F., & Amjad, M. (2023b). Performance evaluation of Standardized Copula-based Drought Index with Reconnaissance Drought Index and Standardized Precipitation Temperature Index using severity–duration frequency curves over Balochistan, Pakistan. *International Journal of Climatology*, 43(6), 2465-2480.

Ullah, H., & Akbar, M. (2023). Bivariate drought risk assessment for water planning using copula function in Balochistan. *Environmental Modeling & Assessment*, 28(3), 447-464.

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.

Wang, F., Wang, Z., Yang, H., Di, D., Zhao, Y., & Liang, Q. (2020). A new copula-based standardized precipitation evapotranspiration streamflow index for drought monitoring. *Journal of Hydrology*, 585, 124793.

Afshar, M. H., Şorman, A. U., & Tosunoglu, F. (2024). Three-dimensional copula framework for early warning of agricultural droughts using meteorological drought and vegetation health conditions. *Hydrological Sciences Journal*, 69(5), 592–606. <https://doi.org/10.1080/02626667.2024.2326187>

Yu, X., Zeng, X., Brocca, L., Gui, D., Wang, D., & Wu, J. (2025). A copula-based composite drought index for enhanced drought monitoring and analysis. *Journal of Geophysical Research: Atmospheres*, 130(8), e2024JD041867. <https://doi.org/10.1029/2024JD041867>

Zhao, Q., Zhang, X., Li, C., Xu, Y., & Fei, J. (2024). Compound ecological drought assessment of China using a Copula-based drought index. *Ecological indicators*, 164, 112141. <https://doi.org/10.1016/j.ecolind.2024.112141>

Li, M., Yao, Y., Feng, Z., & Ou, M. (2025). Hydrological drought prediction and its influencing features analysis based on a machine learning model. *Natural Hazards and Earth System Sciences*, 25(11), 4299-4316. <https://doi.org/10.5194/nhess-25-4299-2025>

Bilbao-Barrenetxea, N., Martínez-España, R., Jimeno-Sáez, P., Faria, S. H., & Senent-Aparicio, J. (2024). Multi-Model Ensemble Machine Learning Approaches to Project Climatic Scenarios in a River Basin in the Pyrenees: N. Bilbao-Barrenetxea et al. *Earth Systems and Environment*, 8(4), 1159-1177.

Rogov, M., Ershova, V., Gaina, C., Vereshchagin, O., Vasileva, K., Mikhailova, K., & Krylov, A. (2023). Glendonites throughout the Phanerozoic. *Earth-Science Reviews*, 241, 104430. <https://doi.org/10.1016/j.earscirev.2023.104430>

Igwebuike, N., Ajayi, M., Okolie, C., Kanyerere, T., & Halihan, T. (2025). Application of machine learning and deep learning for predicting groundwater levels in the West Coast Aquifer System, South Africa. *Earth Science Informatics*, 18(1), 6.

Lan, X., Luo, Y., Leng, G., & Cao, J. (2025). A novel copula-based index for agricultural drought monitoring: integrating vegetation response and meteorological drought dynamics in the middle-lower Yangtze River Basin. *Ecological Indicators*, 178, 114107. <https://doi.org/10.1016/j.ecolind.2025.114107>

Li, M., Yao, Y., Feng, Z., & Ou, M. (2025). Hydrological drought prediction and its influencing features analysis based on a machine learning model. *Natural Hazards and Earth System Sciences*, 25(11), 4299-4316. <https://doi.org/10.5194/nhess-25-4299-2025>

Pande, C. B., Sidek, L. M., Varade, A. M., Elkhachy, I., Radwan, N., Tolche, A. D., & Elbeltagi, A. (2024). Forecasting of meteorological drought using ensemble and machine learning models. *Environmental Sciences Europe*, 36(1), 160. <https://doi.org/10.1186/s12302-024-00975-w>

Shen, X., Li, S., Cai, H., Wang, K., Yuan, X., Li, D., & Li, P. (2024). The response mechanism of transversal mixing of dissolved oxygen to the evolution of secondary flow at the confluence. *Journal of Hydrology*, 635, 131184. <https://doi.org/10.1016/j.jhydrol.2024.131184>

Shao, J., Wang, J., Zhu, D., He, J., Wang, W., Wu, B., & Zhang, G. (2025). Three-dimensional identification of drought events and copula-based multivariate meteorological drought risk probability assessment in the Huai River Basin, China: J. Shao et al. *Theoretical and Applied Climatology*, 156(2), 90. <https://doi.org/10.1007/s00704-024-05318-6>

Zhang, W., Liang, W., Gao, X., Li, J., & Zhao, X. (2024). Trajectory in water scarcity and potential water savings benefits in the Yellow River basin. *Journal of Hydrology*, 633, 130998. <https://doi.org/10.1016/j.jhydrol.2024.130998>

Wang, Q., Wang, Y., Niu, C., & Huang, M. (2024). Constructing the Joint Probability Spatial Distribution of Different Levels of Drought Risk Based on Copula Functions: A Case Study in the Yellow River Basin. *Water*, 16(23), 3374. <https://doi.org/10.3390/w16233374>

Zscheischler, J., Westra, S., Van Den Hurk, B. J., Seneviratne, S. I., Ward, P. J., Pitman, A., ... & Zhang, X. (2018). Future climate risk from compound events. *Nature climate change*, 8(6), 469-477.

McKee, T. B., Doesken, N. J., & Kleist, J. (1993, January). The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology* (Vol. 17, No. 22, pp. 179-183).

Poudel, A., Adhikari, P., Adhikari, P., Choi, S. H., Yun, J. Y., Lee, Y. H., & Hong, S. H. (2024). Predicting the invasion risk of the highly invasive *Acacia mearnsii* in Asia under global climate change. *Plants*, 13(20), 2846.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50(3), 885-900.

Jihan, M. A. T., Popy, S., Kayes, S., Rasul, G., Maowa, A. S., & Rahman, M. M. (2025). Climate change scenario in Bangladesh: Historical data analysis and future projection based on CMIP6 model. *Scientific Reports*, 15(1), 7856.

Talebi, H., & Samadianfard, S. (2024). Integration of machine learning and remote sensing for drought index prediction: A framework for water resource crisis management. *Earth Science Informatics*, 17(5), 4949-4968.

Hissan, R. U., & Parveen, N. (2025). Predicting long-term meteorological drought using random forest and multi-scale drought indices. *Theoretical and Applied Climatology*, 156(10), 552.

Mukherjee, A., Jha, M. K., Kim, K. W., & Pacheco, F. A. (2024). Groundwater resources: challenges and future opportunities. *Scientific Reports*, 14(1), 28540.

Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). FAO Irrigation and drainage paper No. 56. *Rome: food and agriculture organization of the United Nations*, 56(97), e156.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterialCEMLDI.docx](#)