

Multi-Scale SPEI Drought Characterisation and Forecasting across Eleven Contrasting Climates Using ARIMA, Support Vector Regression, XGBoost and LSTM

Ahmad Pahlavanravi¹✉

¹ Associate Professor, Faculty of Water and Soil, Department of Range and Watershed Management, University of Zabol, Zabol, Iran.

*Corresponding Authors: Email: (✉ pahlavanravi@uoz.ac.i)

Received: 14 November 2025

Accepted: 07 December 2025

Available online: 08 January 2026

Abstract: Drought early warning requires both a sound description of past droughts and an honest estimate of how far ahead they can be predicted. This study used 108 years (1900–2007) of monthly climatic water balance (precipitation minus potential evapotranspiration) at eleven locations, from arid Lahore and Kimberley to humid Abashiri and São Paulo. The Standardized Precipitation Evapotranspiration Index (SPEI) was computed at 1-, 3- and 12-month scales with a Python implementation that reproduced the reference R package to within 10^{-12} . Run theory, the modified Mann–Kendall test and the Pettitt test were applied, and one- and three-month-ahead forecasts of SPEI-3 and SPEI-12 from ARIMA, support vector regression (SVR), XGBoost and a long short-term memory (LSTM) network were compared with persistence and climatology benchmarks for 1980–2007, with the SPEI distribution fitted to 1900–1979 only. Annual SPEI-12 decreased significantly at ten of the eleven sites (Sen slopes -0.08 to -0.17 per decade), with change points between 1932 and 1977. The number of SPEI-3 drought events rose from 414 in 1900–1953 to 687 in 1954–2007. Relative to the reference climate, the test period was markedly drier: SPEI-12 was below -1 in 18–67% of months, against 14–19% before 1980. At a one-month lead, SVR forecast SPEI-3 best (mean NSE = 0.65, critical success index 0.55), followed by LSTM (0.62) and ARIMA (0.52), and was significantly better than ARIMA at ten sites. For SPEI-12, the overlap of the accumulation windows made persistence hard to beat (NSE = 0.84 at a one-month lead), and LSTM gave the largest gain at a three-month lead (NSE = 0.59 vs. 0.51; skill score +17%). No model forecast SPEI-3 three months ahead (NSE < 0). Forecast skill at monthly scales therefore comes mainly from the memory built into the index, and ML models add modest but consistent value only when this memory is partial.

Keywords: Drought, SPEI, Run theory, LSTM, Support vector regression, Drought forecasting

1. Introduction

Drought is a slow-onset hazard that affects water supply, agriculture and ecosystems more widely than any other natural disaster, and its frequency and severity are expected to increase as warming raises atmospheric evaporative demand (Dai, 2011; Mishra and Singh, 2010). Standardized drought indices convert a hydro-climatic variable into a probability-based anomaly that can be compared across climates and time scales. The Standardized Precipitation Index (SPI; McKee et al., 1993) uses precipitation only. The Standardized Precipitation Evapotranspiration Index (SPEI; Vicente-Serrano et al., 2010) uses the climatic water balance, precipitation minus potential evapotranspiration (PET), and therefore also responds to temperature-driven changes in water demand. The SPEI is usually fitted with a log-logistic or generalized logistic distribution estimated by probability-weighted moments or L-moments (Hosking, 1990; Beguería et al., 2014; Stagge et al., 2015).

Global and regional analyses have reported increases in drought frequency and severity during the twentieth century, particularly where rising temperatures have increased evaporative demand (Spinoni et al., 2014; Vicente-Serrano et al., 2014). Drought events are commonly described with run theory, which defines an event as an uninterrupted period below a threshold and

characterises it by duration, severity and intensity (Yevjevich, 1967). Change-point and trend tests help to separate gradual changes from shifts (Pettitt, 1979).

Drought forecasting has been attempted with stochastic models such as ARIMA (Mishra and Desai, 2005) and with data-driven models including artificial neural networks, support vector regression (SVR) and extreme learning machines (Mishra and Desai, 2006; Belayneh et al., 2014; Deo and Şahin, 2015; Mokhtarzad et al., 2017). More recently, long short-term memory (LSTM) networks have been used for longer lead times (Dikshit et al., 2021). Reviews show that reported skill varies widely and depends strongly on the index time scale, the lead time and the validation design (Anshuka et al., 2019). Two issues deserve particular attention. First, an SPEI- k value shares $k - 1$ months of data with the value one month earlier, so part of the apparent skill at long scales is inherited from persistence. Second, fitting the SPEI distribution to the whole record, including the test period, leaks information about the test climate into the target. Neither issue is always addressed, and naive benchmarks are rarely reported.

This study aimed to (i) characterise drought frequency, duration, severity and long-term change at eleven climatically diverse locations over 1900–2007; (ii) compare ARIMA, SVR, XGBoost and LSTM for one- and three-month-ahead forecasting of SPEI-3 and

SPEI-12 under a strictly out-of-sample design with persistence and climatology benchmarks; and (iii) evaluate both continuous accuracy and the detection of drought conditions.

2. Materials and Methods

2.1. Data and Study Sites

The data were the monthly climatic water balance $D = P - PET$ (mm) for January 1900 to December 2007 (1296 months) at eleven locations: Indore (India), Kimberley (South Africa), Albuquerque (USA), Valencia (Spain), Vienna (Austria), Abashiri (Japan), Tampa (USA), São Paulo (Brazil), Lahore (Pakistan),

Punta Arenas (Chile) and Helsinki (Finland) (Figure 1). The series are distributed with the SPEI package (Beguería and Vicente-Serrano, 2023) and were derived from the CRU TS 3.1 gridded climate data (Harris et al., 2014) and the 20th Century Reanalysis (Compo et al., 2011). They are the example data of that package, and the series were complete. Site coordinates in Figure 1 are approximate city locations. The sites cover a gradient of mean annual water balance from -1205 mm (Lahore) to +432 mm (Abashiri) (Table 1), and in the text they are ordered from driest to wettest.

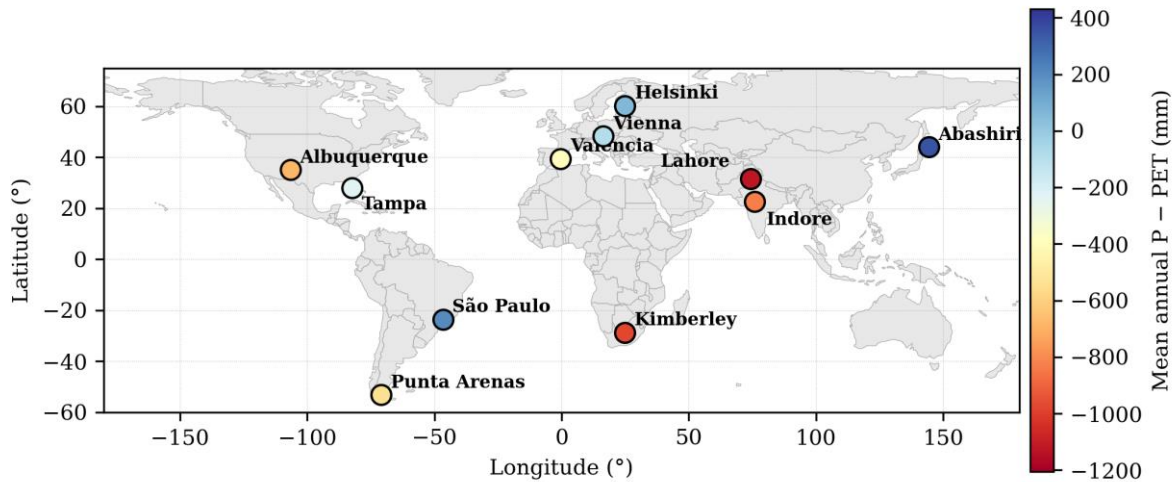


Figure 1. Locations of the eleven sites coloured by mean annual climatic water balance ($P - PET$), 1900–2007

Table 1. Water-balance characteristics of the study sites (1900–2007)

Site	Country	Mean annual D (mm)	SD annual D (mm)	Driest month (mean D, mm)	Wettest month (mean D, mm)	Months with D > 0	Skewness (monthly D)
Lahore	Pakistan	-1205	260	Jun (-288)	Jan (+13)	2	-0.32
Kimberley	South Africa	-644	180	Jan (-95)	Jun (-21)	0	-0.02
Indore	India	-607	311	May (-273)	Jul (+147)	3	0.94
Albuquerque	USA	-414	99	Jun (-86)	Dec (+7)	2	-0.34
Punta Arenas	Chile	-295	120	Feb (-59)	Jul (+9)	4	0.58
Valencia	Spain	-256	196	Jul (-92)	Dec (+14)	5	1.19
Tampa	USA	+46	277	May (-47)	Aug (+39)	8	1.05
Vienna	Austria	+208	134	Aug (-6)	Dec (+43)	9	0.39
Helsinki	Finland	+322	123	Jun (-9)	Nov (+60)	9	0.16
São Paulo	Brazil	+399	256	Aug (-23)	Jan (+125)	7	1.03
Abashiri	Japan	+432	144	Jun (+10)	Jan (+63)	12	0.70

D = precipitation – potential evapotranspiration.

2.2. Standardized Precipitation Evapotranspiration Index

For a time scale k , the water balance was accumulated as

$$D_k(t) = \sum_{j=0}^{k-1} D(t-j) \quad (1)$$

and, separately for each calendar month, a three-parameter generalized logistic (GLO) distribution was fitted to $D_{\{k\}}$ by unbiased probability-weighted moments converted to L-moments (Hosking, 1990). Its cumulative distribution function is

$$F(x) = [1 + \exp(-y)]^{-1}, \quad y = -\frac{1}{\kappa} \ln \left[1 - \frac{\kappa(x-\xi)}{\alpha} \right] \quad (2)$$

where ξ , α and κ are the location, scale and shape parameters. The SPEI is the standard normal quantile of F :

$$\text{SPEI}_k(t) = \Phi^{-1}[F(D_k(t))] \quad (3)$$

This is the default procedure of the SPEI package version 1.8.1 (Beguería et al., 2014). It was coded in Python and checked against the reference outputs distributed with the package for all eleven series. The maximum absolute differences were 1.0×10^{-13} for SPEI-1 and 1.6×10^{-13} for SPEI-12. For drought characterisation the distribution was fitted to the full 1900–2007 record. For forecasting it was fitted to 1900–1979 only, so that the test period (1980–2007) was not

used in the standardisation. Values outside the fitted support (0.42% of all values) were bounded at ± 3.09 .

2.3. Drought Characteristics and Trend Tests

Drought events were identified by run theory (Yevjevich, 1967) as consecutive months with $\text{SPEI} < -1$. For each event e , duration L_e (months), severity and intensity were computed as

$$S_e = -\sum_{t \in e} \text{SPEI}(t), \quad I_e = S_e/L_e \quad (4)$$

Events were counted separately for 1900–1953 and 1954–2007. Trends in December SPEI_{12} , i.e. the annual water-balance anomaly, were tested with the Mann–Kendall test using the Hamed and Rao (1998) correction for serial correlation, and slopes were estimated by the Sen (1968) method. Abrupt shifts were located with the Pettitt (1979) test, whose statistic is

$$K = \max_{1 \leq \tau < n} \left| \sum_{i=1}^{\tau} \sum_{j=\tau+1}^n \text{sgn}(x_j - x_i) \right|, \quad p \approx \frac{1}{2 \exp\left(\frac{-6K^2}{n^3 + n^2}\right)} \quad (5)$$

2.4. Forecasting Models

Forecasts of SPEI_{-3} and SPEI_{-12} were made at leads of $h = 1$ and 3 months from each origin t , using only information available at t . Samples whose target fell in 1900–1979 were used for training and those whose target fell in 1980–2007 for testing (336 months). Two benchmarks were used: **persistence**, $\text{SPEI}_k(t + h) = \text{SPEI}_k(t)$, and **climatology**, $\text{SPEI}_k(t + h) = 0$, the expected value under the reference climate.

ARIMA($p, 0, q$) models with a constant (Box et al., 2015) were fitted to the training SPEI_k series, with $p, q \leq 3$ selected by the Akaike information criterion:

$$\phi(B)(x_t - \mu) = \theta(B)\varepsilon_t \quad (6)$$

where B is the backshift operator. With the parameters fixed, the Kalman filter was run through the full series, and dynamic h -step forecasts were issued from every test origin. The machine learning models used the same predictors: SPEI_k at lags 0–5, SPEI_{-1} at lags 0–2, and the sine and cosine of the target month. **SVR** used a radial basis function kernel (Vapnik, 1995; grid $C \in \{0.1, 1, 10, 100\}$, $\gamma \in \{0.001, 0.01, 0.05, 0.1\}$, $\varepsilon \in \{0.01, 0.1, 0.2\}$). **XGBoost** (Chen and Guestrin, 2016) used subsampling of 0.8 (grid: 200 or 500 trees, learning rate 0.02 or 0.05, depth 2, 3 or 5, minimum child weight 1 or 5). The hyperparameters of both were tuned by five-fold forward-chaining cross-validation on the training period. The **LSTM** network (Hochreiter and

Schmidhuber, 1997) received 12-month sequences of SPEI_k , SPEI_{-1} and the seasonal terms. It had one LSTM layer with 32 units, dropout of 0.2, a 16-unit ReLU layer and a linear output, and was trained with Adam (learning rate 0.001, batch size 32, maximum 300 epochs) with early stopping (patience 20) on the last 20% of the training period. The forecast was the mean of three networks with different random seeds. The software used was statsmodels, scikit-learn (Pedregosa et al., 2011), xgboost and Keras.

2.5. Evaluation

Continuous skill was measured by the Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970), RMSE and the Pearson correlation R . The mean-squared-error skill score relative to persistence was also computed:

$$SS = 1 - \frac{MSE_{model}}{MSE_{persistence}} \quad (7)$$

Drought detection at the threshold $\text{SPEI} < -1$ was assessed with the probability of detection (POD), false alarm ratio (FAR) and critical success index (CSI) (Wilks, 2011):

$$POD = \frac{a}{a+c}, \quad FAR = \frac{b}{a+b}, \quad CSI = \frac{a}{a+b+c} \quad (8)$$

where a, b and c are the numbers of hits, false alarms and misses. The best ML or ARIMA model at each site was compared with ARIMA and with persistence by the Diebold–Mariano test (Diebold and Mariano, 1995) with the Harvey et al. (1997) correction and a variance estimator allowing $h - 1$ autocovariances.

3. Results

3.1. Water Balance Regimes

The eleven sites span very different water-balance regimes (Table 1; Figure 2). At Lahore and Kimberley, D was negative in almost every month, with the largest deficits in the pre-monsoon (Lahore, June: -288 mm) and austral summer (Kimberley, January: -95 mm) seasons. At Indore the monsoon produced a strong surplus in July ($+147$ mm) between long dry seasons. Valencia, Albuquerque and Punta Arenas had moderate annual deficits (-256 to -414 mm), with a summer deficit and small winter surpluses. Tampa was close to balance, and Vienna, Helsinki, São Paulo and Abashiri had annual surpluses ($+208$ to $+432$ mm). Abashiri was the only site with a positive mean D in all twelve months.

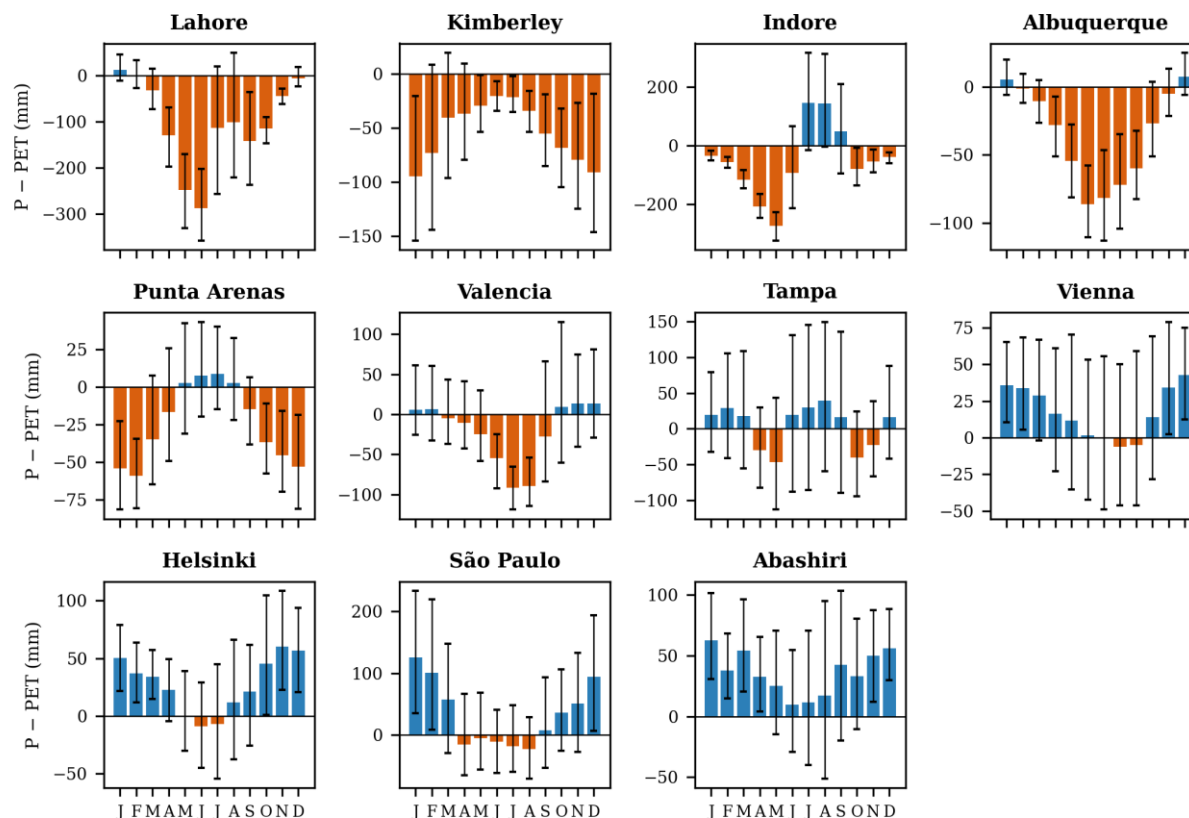


Figure 2. Mean monthly climatic water balance (bars) with 10th–90th percentile range (whiskers), 1900–2007

3.2. Drought Characteristics and Long-Term Change

With the SPEI-3 threshold of -1 , each site experienced 87–113 drought events in 108 years (Table 2; Figure 3). Mean durations were 1.9–2.7 months and maximum durations 7–14 months. Events were shortest and most frequent in humid Abashiri, São Paulo and Vienna, and longest on average at Kimberley, Indore and Albuquerque. At the 12-month scale, events were fewer (29–50) and longer (mean 4.4–7.6 months). The

longest SPEI-12 droughts occurred at Valencia (35 months from November 1997), Indore (34 months from September 2000), Albuquerque (33 months from October 2001) and Tampa (32 months from February 1999). The total number of SPEI-3 events increased from 414 in 1900–1953 to 687 in 1954–2007, and the number of drought months doubled from 825 to 1658. Abashiri (58 $\rightarrow$ 55 events) was the only site with fewer events in the second half.

Table 2. Run-theory drought characteristics for SPEI-3 and SPEI-12 (threshold -1 , 1900–2007)

Site	SPEI-3 events	Mean dur. (mo)	Max dur. (mo)	Mean sev.	Events 1900–53/1954–2007	SPEI-12 events	Mean dur. (mo)	Max dur. (mo)	Max sev.	Most severe SPEI-12 event: start (months)
Lahore	106	2.1	11	3.1	44/62	44	5.0	22	40.9	1946-09 (22)
Kimberley	93	2.7	9	3.8	35/58	34	6.9	25	40.1	1983-01 (25)
Indore	87	2.5	8	3.7	24/63	31	7.4	34	61.9	2000-09 (34)
Albuquerque	88	2.5	11	3.7	31/57	29	6.8	33	59.4	2001-10 (33)
Punta Arenas	101	2.3	14	3.3	38/63	42	5.3	25	44.6	1936-11 (25)
Valencia	95	2.3	9	3.4	27/68	31	7.6	35	60.3	1997-11 (35)
Tampa	97	2.3	11	3.3	38/59	42	6.0	32	58.8	1999-02 (32)
Vienna	109	2.1	7	3.0	39/70	46	4.6	26	43.7	2000-06 (26)
Helsinki	101	2.2	9	3.2	40/61	46	5.1	24	41.0	2002-08 (23)
São Paulo	111	2.1	13	2.9	40/71	49	4.4	17	35.8	1963-07 (17)
Abashiri	113	1.9	7	2.7	58/55	50	4.4	19	27.1	1987-09 (19)

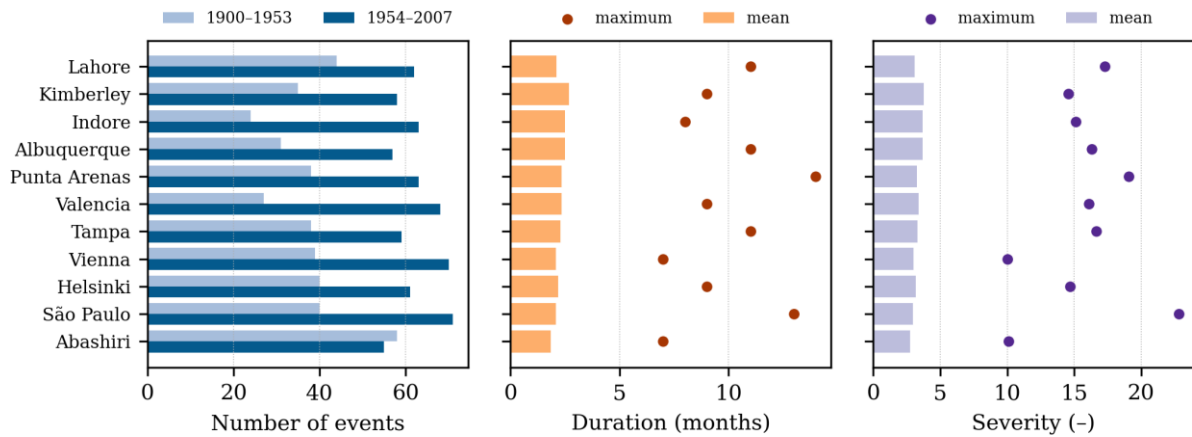


Figure 3. SPEI-3 drought events: number in the two half-periods, mean and maximum duration, and mean and maximum severity

Annual SPEI-12 decreased significantly at 10 of the 11 sites (Table 3; Figure 4). At these sites, Sen slopes ranged from -0.08 to -0.17 standardized units per decade, and the strongest declines were at Vienna, Valencia, Tampa and Albuquerque. Lahore was the only site without a significant trend ($Z = -1.30$, $p = 0.20$). The Pettitt test found significant change points at the same ten sites. Most of them fell in 1944–1967, with the earliest at Helsinki (1932) and the latest at Valencia (1977). After the change point, mean SPEI-12 was negative at every site. The heat map in Figure 4 shows that the decades after 1980 were predominantly dry at most sites.

Table 3. Trend (modified Mann–Kendall, Sen) and change-point (Pettitt) tests of annual SPEI-12, 1900–2007

Site	Z	p	Sen (decade ⁻¹)	Change year	p (Pettitt)	Mean before	Mean after
Lahore	-1.29	0.195	-0.038	1931	0.396	+0.31	-0.13
Kimberley	-7.73	0.000	-0.137	1963	0.003	+0.28	-0.41
Indore	-3.43	0.001	-0.139	1964	0.000	+0.39	-0.59
Albuquerque	-3.59	0.000	-0.157	1944	0.000	+0.69	-0.49
Punta Arenas	-4.44	0.000	-0.140	1958	0.000	+0.37	-0.44
Valencia	-5.72	0.000	-0.167	1977	0.000	+0.31	-0.79
Tampa	-6.19	0.000	-0.162	1960	0.000	+0.44	-0.57
Vienna	-5.22	0.000	-0.168	1944	0.000	+0.58	-0.41
Helsinki	-2.87	0.004	-0.080	1932	0.016	+0.47	-0.21
São Paulo	-3.17	0.002	-0.093	1950	0.037	+0.28	-0.25
Abashiri	-3.09	0.002	-0.090	1967	0.022	+0.24	-0.40

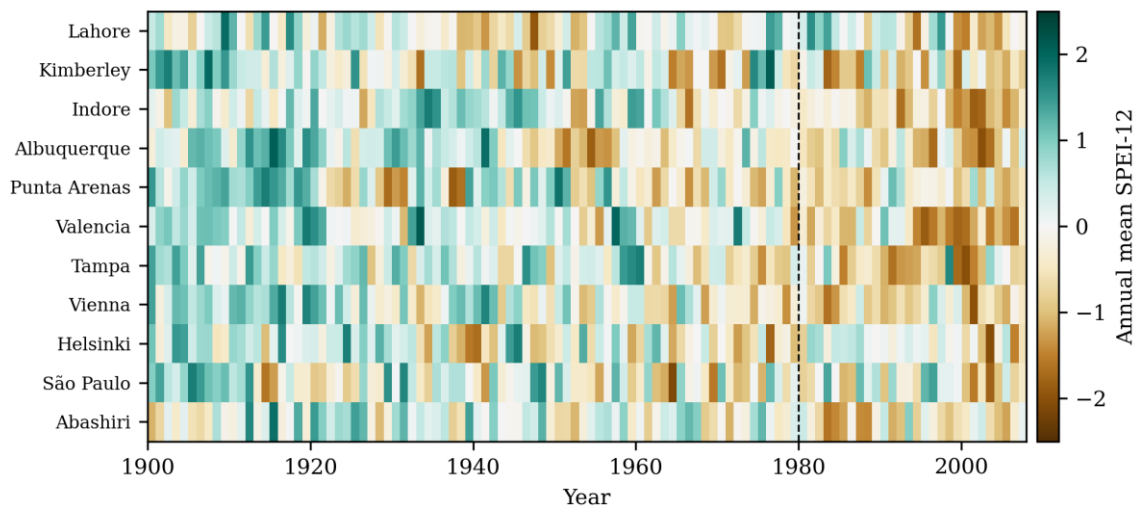


Figure 4. Annual mean SPEI-12 at the eleven sites (driest at top); the dashed line marks the start of the forecast test period

Because the forecast SPEI was standardised on 1900–1979, the test period appears as a strong departure from the reference climate. Mean test-period SPEI-12 ranged from -0.10 (Helsinki) to -1.35 (Valencia), and the proportion of months with $\text{SPEI-12} < -1$ was 18–67%, compared with 14–19% in the reference period. This non-stationarity is the reason why the climatology benchmark had negative NSE at all sites (Figure 5).

3.3. Forecast Performance

Table 4 summarises the mean test-period skill, and Figure 5 shows NSE for every site. At a one-month lead, SPEI-3 was forecast best by SVR (mean NSE = 0.65, RMSE = 0.65), followed by LSTM (0.62), XGBoost (0.60) and ARIMA (0.52). All models were far above persistence (0.34). SVR was the best model at eight sites and LSTM at three (Table 5). The MSE skill score

relative to persistence averaged 0.47 for SVR and 0.43 for LSTM (Figure 6). The best model was significantly more accurate than ARIMA at ten of the eleven sites (Table 6).

Table 4. Mean test-period (1980–2007) skill over the eleven sites

Case	Model	NSE	RMSE	R	POD	FAR	CSI	SS vs. persistence
SPEI-3, h = 1	Persistence	0.34	0.89	0.67	0.65	0.34	0.49	–
	Climatology	-0.23	1.22	–	–	–	–	–
	ARIMA	0.52	0.76	0.74	0.52	0.24	0.44	+0.28
	SVR	0.65	0.65	0.81	0.65	0.23	0.55	+0.47
	XGBoost	0.60	0.70	0.79	0.52	0.22	0.45	+0.39
	LSTM	0.62	0.68	0.80	0.59	0.22	0.51	+0.43
SPEI-3, h = 3	Persistence	-0.94	1.53	0.03	0.35	0.64	0.22	–
	Climatology	-0.23	1.22	–	–	–	–	–
	ARIMA	-0.15	1.18	0.04	0.00	0.06	0.00	+0.40
	SVR	-0.16	1.19	0.03	0.01	0.13	0.01	+0.40
	XGBoost	-0.19	1.20	0.07	0.00	0.11	0.00	+0.38
	LSTM	-0.19	1.20	0.04	0.00	0.00	0.00	+0.38
SPEI-12, h = 1	Persistence	0.84	0.37	0.92	0.87	0.13	0.77	–
	Climatology	-0.70	1.20	–	–	–	–	–
	ARIMA	0.84	0.36	0.92	0.82	0.11	0.75	+0.02
	SVR	0.85	0.36	0.92	0.84	0.12	0.75	+0.04
	XGBoost	0.82	0.39	0.91	0.81	0.11	0.73	-0.12
	LSTM	0.85	0.36	0.93	0.84	0.11	0.76	+0.05
SPEI-12, h = 3	Persistence	0.51	0.65	0.75	0.73	0.27	0.58	–
	Climatology	-0.70	1.20	–	–	–	–	–
	ARIMA	0.53	0.63	0.76	0.58	0.20	0.52	+0.06
	SVR	0.57	0.61	0.77	0.65	0.20	0.56	+0.13
	XGBoost	0.50	0.65	0.75	0.53	0.18	0.48	-0.00
	LSTM	0.59	0.59	0.80	0.66	0.19	0.57	+0.17

POD, FAR and CSI for SPEI < -1. The climatology forecast (SPEI = 0) never predicts drought.

At a three-month lead, SPEI-3 could not be forecast. All models had negative mean NSE (-0.15 to -0.19) and near-zero correlation with the observations. They converged towards the reference mean and therefore detected almost no droughts (CSI ≤ 0.01). This is expected, because a three-month accumulation window three months ahead shares no months with the information available at the forecast origin. The formally significant improvement over persistence (Table 6) is trivial, because persistence itself had an NSE of -0.94.

SPEI-12 at a one-month lead was dominated by persistence (0.84), since eleven of the twelve accumulated months are already known. LSTM (0.85), SVR (0.85) and ARIMA (0.84) improved on it only marginally, and XGBoost was worse (0.82). Persistence even had the highest POD and CSI. At a three-month lead the models added clearer value. LSTM had the highest mean NSE (0.59), was the best model at seven sites and had a mean skill score of +0.17 relative to persistence, followed by SVR (0.57; +0.13) and ARIMA (0.53). The best model was significantly better than persistence at eight sites and than ARIMA at five. Figure 7 illustrates these SPEI-12 forecasts for an arid, an intermediate and a humid site. The forecasts follow the major drought episodes, such as the prolonged 1990s drought at Valencia, but lag their onset and termination.

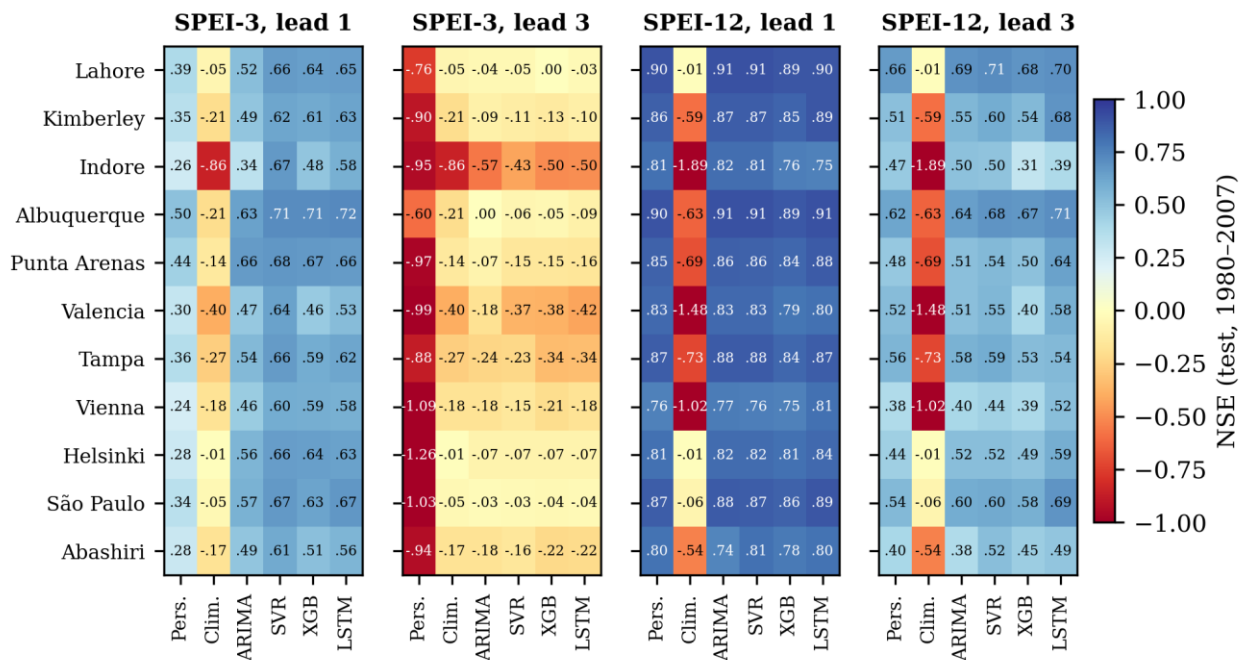
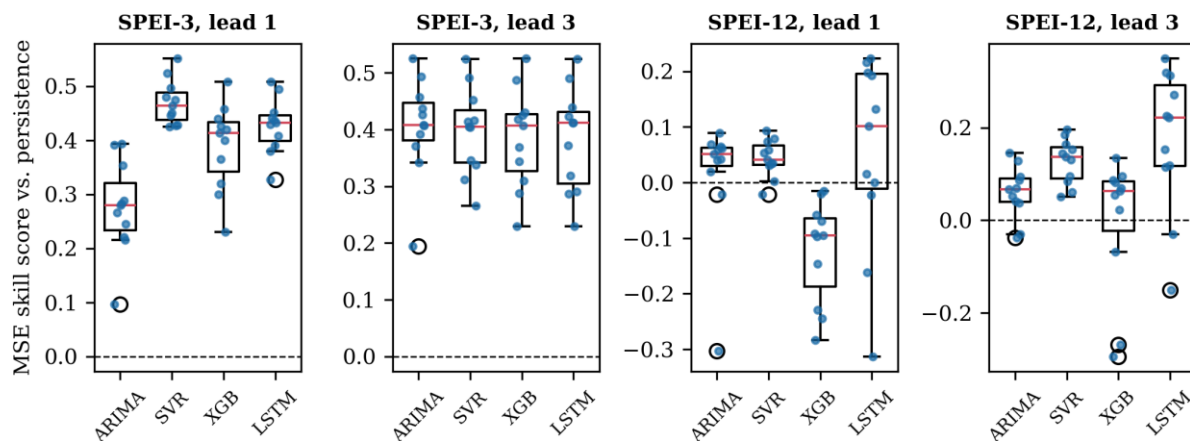
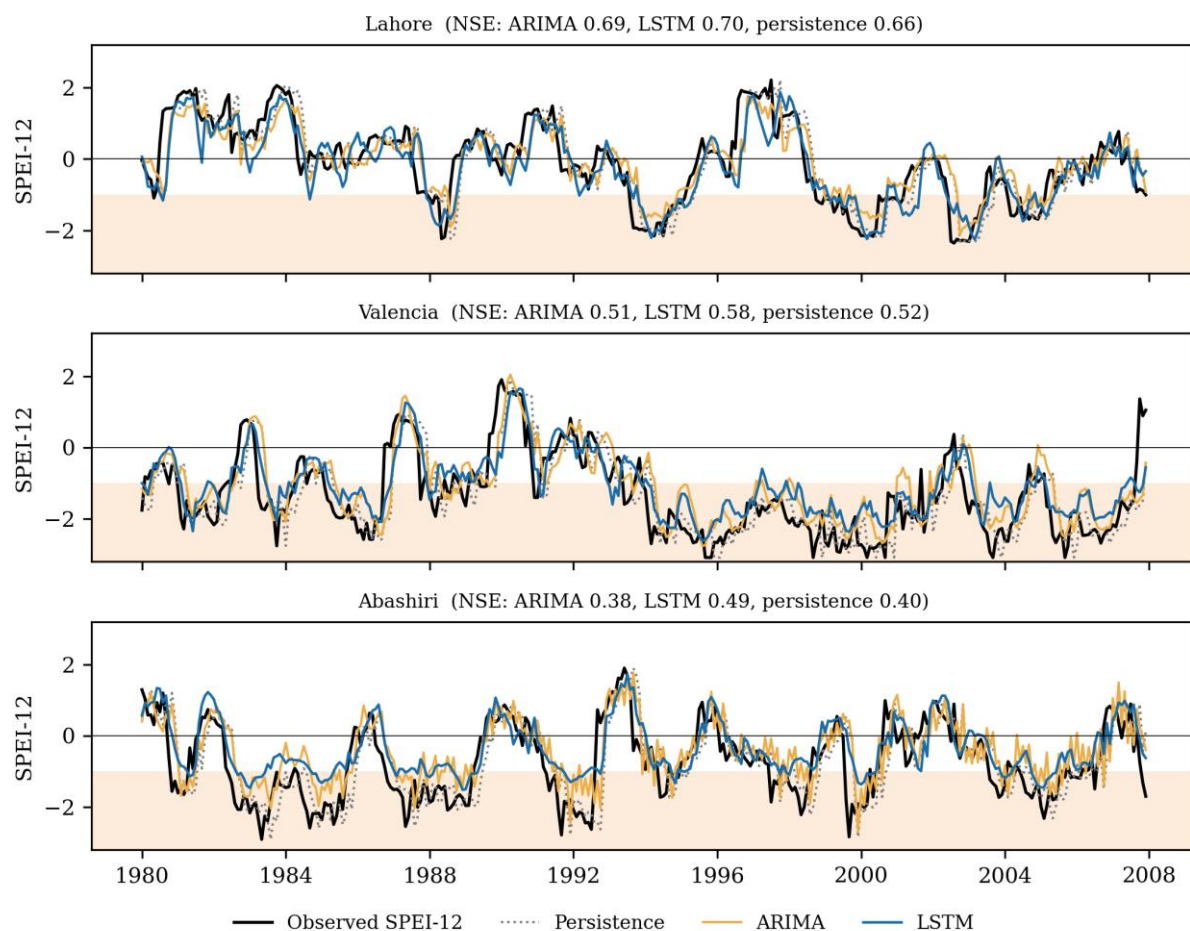


Figure 5. Test-period NSE of the benchmarks and forecast models for SPEI-3 and SPEI-12 at leads of 1 and 3 months

Table 5. Number of sites where each model was best (highest NSE), mean RMSE rank among six forecasts, and Diebold–Mariano results for the best model

Case	Best: ARIMA	Best: SVR	Best: XGB	Best: LSTM	Rank SVR	Rank LSTM	Rank ARIMA	Rank Pers.	Sig. better than ARIMA	Sig. better than persistence
SPEI-3, h = 1	0	8	0	3	1.27	2.09	3.82	5.00	10/11	11/11
SPEI-3, h = 3	5	4	2	0	2.36	3.45	2.18	6.00	3/6	11/11
SPEI-12, h = 1	2	3	0	6	2.18	2.18	2.36	3.45	4/9	5/11
SPEI-12, h = 3	1	3	0	7	1.82	1.73	3.00	4.36	5/10	8/11

Rank 1 = lowest RMSE among persistence, climatology, ARIMA, SVR, XGBoost and LSTM. Significance at $p < 0.05$; the denominator is the number of sites where the best model was not the reference.

**Figure 6. MSE skill scores of the forecast models relative to persistence (boxes: eleven sites)****Figure 7. Observed and three-month-ahead forecast SPEI-12 at Lahore (arid), Valencia and Abashiri (humid); shading marks drought (SPEI < -1)**

Drought detection (Figure 8) followed the continuous skill. For SPEI-3 at a one-month lead, SVR combined a high POD (0.65) with a low FAR (0.23), which gave the best CSI (0.55). Persistence detected as many droughts (POD 0.65) but with many more false alarms (FAR 0.34). ARIMA and XGBoost were more

conservative and missed about half of the drought months (POD ≈ 0.52). For SPEI-12 at a three-month lead, LSTM had the best balance (POD 0.66, FAR 0.19). Persistence reached a similar CSI (0.58) with a higher FAR (0.27).

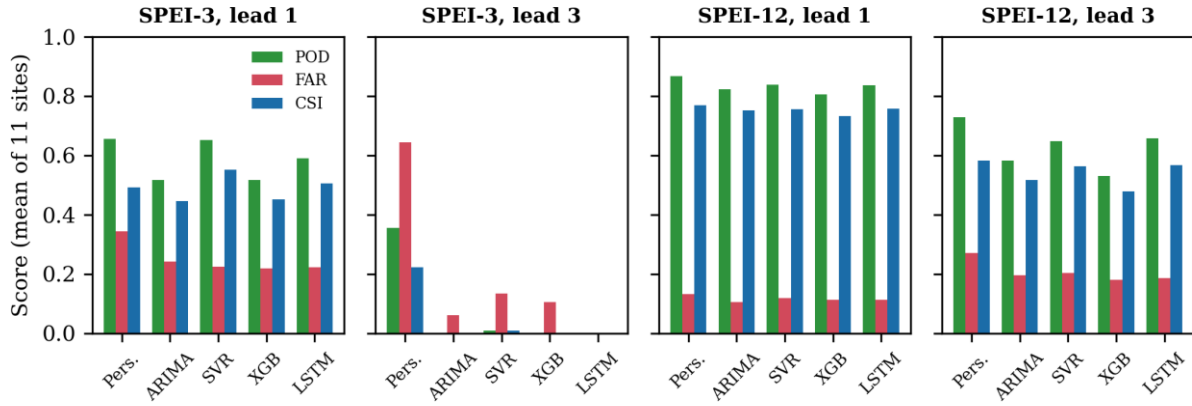


Figure 8. Mean probability of detection (POD), false alarm ratio (FAR) and critical success index (CSI) for drought (SPEI < -1)

4. Discussion

The widespread decline of SPEI-12 over the twentieth century agrees with global assessments that link increasing drought severity to higher evaporative demand (Dai, 2011; Spinoni et al., 2014). The declines at Valencia and Vienna are consistent with the drying reported for southern and central Europe (Vicente-Serrano et al., 2014). The change points around the mid-century, however, should be interpreted with care. The early decades of gridded climate data rest on sparse station networks, and the PET component of the series combines two data sources. Part of the apparent shift may therefore reflect changes in data quality rather than in climate (Harris et al., 2014). The results are nonetheless consistent in direction across ten of the eleven sites and across very different climates.

For forecasting, the decisive factor was the overlap between the accumulation window and the lead time. When overlap was complete except for one month (SPEI-12, $h = 1$), persistence was nearly unbeatable, and the reported NSE of about 0.85 should not be taken as evidence of model skill. When there was no overlap (SPEI-3, $h = 3$), no model had skill, which agrees with the rapid loss of skill with lead time reported in reviews (Anshuka et al., 2019). The models added real value in the intermediate cases (SPEI-3, $h = 1$; SPEI-12, $h = 3$), where they had to predict the unknown new months that enter the window. SVR and LSTM were the most consistent here, in line with earlier comparisons (Belayneh et al., 2014; Dikshit et al., 2021). Their advantage over the linear ARIMA model suggests that the seasonal and non-linear structure of the water balance carries some predictability. XGBoost was the weakest learner, probably because shallow trees cannot extrapolate into the drier test climate.

The non-stationarity of the test period is an important practical lesson. A drought index standardised on an earlier reference period will classify many more months as drought if the climate has shifted. Forecast models trained on the reference period must then extrapolate beyond their training distribution. Operational systems should therefore update the reference period regularly or use non-stationary index formulations. The study used only past values of the index as predictors. Adding large-scale climate indices or seasonal forecasts could extend useful lead times, particularly for short accumulation scales. Other limitations are the use of eleven single grid-point series and a fixed LSTM architecture.

5. Conclusions

SPEI drought characteristics and forecasts were analysed for eleven contrasting climates over 1900–2007. The main conclusions are:

(1) SPEI-12 decreased significantly at ten of the eleven sites (-0.08 to -0.17 per decade), with change points between 1932 and 1977, and the number of SPEI-3 drought events increased by 66% between the first and second halves of the record.

(2) For SPEI-3 one month ahead, SVR was the most accurate model (mean NSE 0.65, CSI 0.55) and was significantly better than ARIMA at ten sites. LSTM was a close second.

(3) SPEI-12 one month ahead was dominated by persistence (NSE 0.84). At three months ahead, LSTM added the most value (NSE 0.59; skill +17% over persistence).

(4) SPEI-3 three months ahead was not predictable from the index history (NSE < 0 for all models).

(5) Drought-forecasting studies should report persistence and climatology benchmarks, fit the index distribution without using the test period, and consider

the overlap between accumulation window and lead time when interpreting skill.

Data Availability Statement

The water-balance series are distributed with the open-source SPEI R package (<https://CRAN.R-project.org/package=SPEI>). Scripts, including the validated Python SPEI implementation, are available from the corresponding author on request.

Conflict of Interest

The authors declare no conflict of interest.

References

- Anshuka, A., van Ogtrop, F.F., & Vervoort, R.W. (2019). Drought forecasting through statistical models using standardised precipitation index: a systematic review and meta-regression analysis. *Natural Hazards*, 97, 955–977. <https://doi.org/10.1007/s11069-019-03665-6>
- Beguería, S., & Vicente-Serrano, S.M. (2023). SPEI: Calculation of the Standardized Precipitation-Evapotranspiration Index. R package version 1.8.1. <https://CRAN.R-project.org/package=SPEI>
- Beguería, S., Vicente-Serrano, S.M., Reig, F., & Latorre, B. (2014). Standardized precipitation evapotranspiration index (SPEI) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. *International Journal of Climatology*, 34, 3001–3023. <https://doi.org/10.1002/joc.3887>
- 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 network and wavelet support vector regression models. *Journal of Hydrology*, 508, 418–429. <https://doi.org/10.1016/j.jhydrol.2013.10.052>
- Box, G.E.P., Jenkins, G.M., Reinsel, G.C., & Ljung, G.M. (2015). *Time Series Analysis: Forecasting and Control* (5th ed.). Wiley, Hoboken, NJ.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Compo, G.P., Whitaker, J.S., Sardeshmukh, P.D., Matsui, N., Allan, R.J., Yin, X., et al. (2011). The Twentieth Century Reanalysis Project. *Quarterly Journal of the Royal Meteorological Society*, 137, 1–28. <https://doi.org/10.1002/qj.776>
- Dai, A. (2011). Drought under global warming: a review. *Wiley Interdisciplinary Reviews: Climate Change*, 2, 45–65. <https://doi.org/10.1002/wcc.81>
- Deo, R.C., & Şahin, M. (2015). Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia. *Atmospheric Research*, 153, 512–525. <https://doi.org/10.1016/j.atmosres.2014.10.016>
- Diebold, F.X., & Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Dikshit, A., Pradhan, B., & Alamri, A.M. (2021). Long lead time drought forecasting using lagged climate variables and a stacked long short-term memory model. *Science of the Total Environment*, 755, 142638. <https://doi.org/10.1016/j.scitotenv.2020.142638>
- Hamed, K.H., & Rao, A.R. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, 204, 182–196.
- Harris, I., Jones, P.D., Osborn, T.J., & Lister, D.H. (2014). Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset. *International Journal of Climatology*, 34, 623–642. <https://doi.org/10.1002/joc.3711>
- Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2), 281–291.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Hosking, J.R.M. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society, Series B*, 52(1), 105–124.
- 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*, American Meteorological Society, Anaheim, CA, 179–184.
- Mishra, A.K., & Desai, V.R. (2005). Drought forecasting using stochastic models. *Stochastic Environmental Research and Risk Assessment*, 19, 326–339. <https://doi.org/10.1007/s00477-005-0238-4>
- Mishra, A.K., & Desai, V.R. (2006). Drought forecasting using feed-forward recursive neural network. *Ecological Modelling*, 198, 127–138. <https://doi.org/10.1016/j.ecolmodel.2006.04.017>
- Mishra, A.K., & Singh, V.P. (2010). A review of drought concepts. *Journal of Hydrology*, 391, 202–216. <https://doi.org/10.1016/j.jhydrol.2010.07.012>
- Mokhtarzad, M., Eskandari, F., Jamshidi Vanjani, N., & Arabasadi, A. (2017). Drought forecasting by ANN, ANFIS, and SVM and comparison of the models. *Environmental Earth Sciences*, 76, 729. <https://doi.org/10.1007/s12665-017-7064-0>
- Nash, J.E., & Sutcliffe, J.V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3), 282–290.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pettitt, A.N. (1979). A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 28(2), 126–135.
- Sen, P.K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, 63, 1379–1389.
- Spinoni, J., Naumann, G., Carrao, H., Barbosa, P., & Vogt, J. (2014). World drought frequency, duration, and severity for 1951–2010. *International Journal of Climatology*, 34, 2792–2804. <https://doi.org/10.1002/joc.3875>

- 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, 4027–4040. <https://doi.org/10.1002/joc.4267>
- Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.
- 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, 1696–1718. <https://doi.org/10.1175/2009JCLI2909.1>
- Vicente-Serrano, S.M., Lopez-Moreno, J.-I., Beguería, S., Lorenzo-Lacruz, J., Sanchez-Lorenzo, A., García-Ruiz, J.M., et al. (2014). Evidence of increasing drought severity caused by temperature rise in southern Europe. *Environmental Research Letters*, 9, 044001. <https://doi.org/10.1088/1748-9326/9/4/044001>
- Wilks, D.S. (2011). *Statistical Methods in the Atmospheric Sciences* (3rd ed.). Academic Press, Oxford.
- Yevjevich, V. (1967). An objective approach to definitions and investigations of continental hydrologic droughts. *Hydrology Papers* 23, Colorado State University, Fort Collins, CO.