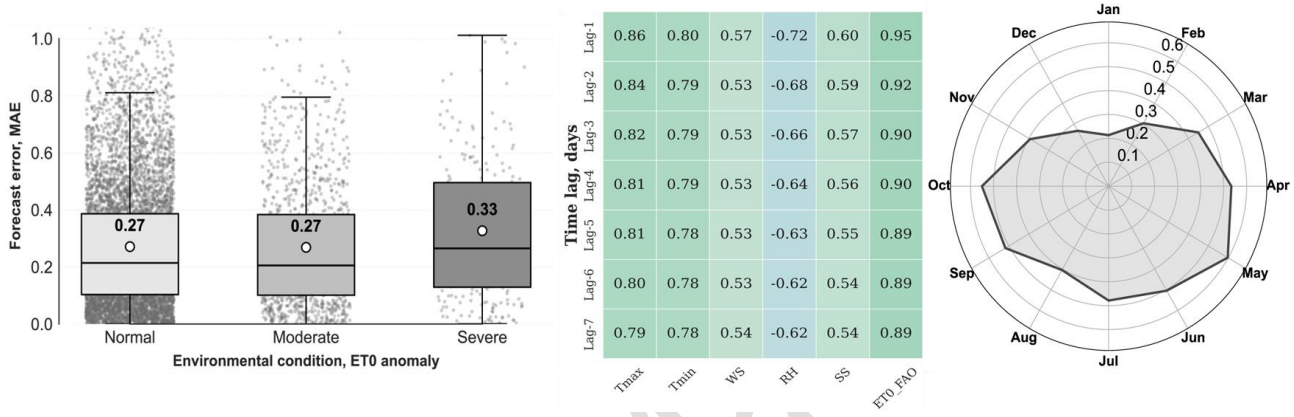
Journal of Applied Research in Water and Wastewater
E-ISSN: 2476-6283
Journal homepage: https://arww.razi.ac.ir

Explaining evapotranspiration dynamics via CNN-LSTM and temporal SHAP: a multi-step forecasting approach across diverse climates

Moein Tosan[1,*], Afshin Shayeghi[2], Javad Teymouri[3], Aydin Bakhtar[3]

[1]*Department of Irrigation and Reclamation Engineering, University of Tehran, Karaj, Iran.*
[2]*Department of Geography & Environmental Sustainability, University of Oklahoma, Norman, Oklahoma, USA.*
[3]*Department of Civil Engineering, University of Texas at Arlington, Arlington, Texas, USA.*

GRAPHICAL ABSTRACT



ARTICLE INFO

Article type:
Research Article

Article history:
Received xx Month xxx
Received in revised form xx Month xxx
Accepted xx Month xxx
Available online x Month xx

Keywords:
Climatic memory
Explainable artificial intelligence
Multi-step forecasting
SHAP

ABSTRACT

Reliable multi-step-ahead forecasting of reference evapotranspiration (ET_o) is critical for proactive water resource management, yet understanding the temporal memory of hydrological systems remains a challenge for black-box deep learning models. This study presents a novel, interpretable forecasting framework integrating temporal SHapley additive explanations (SHAP) with advanced recurrent neural networks to predict daily ET_o up to 7 days in advance across three contrasting climatic zones in Iran; Birjand (arid), Mashhad (semi-arid), and Gorgan (humid). By benchmarking long short-term memory (LSTM), bidirectional LSTM (BiLSTM), and CNN-LSTM architectures, it is demonstrated that model complexity does not always guarantee superiority; the standard LSTM proved remarkably robust, achieving high short-term accuracy ($R^2 > 0.93$ for 1-day forecast) in arid regions. However, a distinct humid-climate penalty was observed, with forecast accuracy degrading more rapidly in Gorgan due to stochastic cloud dynamics. The application of temporal SHAP revealed climate-specific memory effects: in arid zones, wind speed exhibited a persistent influence extending back 5 days, acting as a long-term driver of evaporative demand, whereas humid regions were governed by short-term radiative pulses. Furthermore, analysis of extreme events and drought propagation showed that while the model successfully captures heatwave-driven peaks, its reliability decreases under severe evaporative stress (standardized ET_o anomaly > 2). Cross-spatial generalization tests confirmed that models trained on arid data transfer effectively to humid regions ($R^2 = 0.95$), but the reverse transfer fails to capture extreme advective forcing. This study provides a transferable, physically interpretable blueprint for developing early warning systems in data-scarce regions.

1. Introduction

The increasing occurrence of hydrological extremes, due to global climate change, has intensified water shortages in arid and semi-arid areas, jeopardizing food security and ecological stability (Abbasi, 2025; Akbarpour *et al.*, 2024). In dry regions, precise management of water supplies is essential. The essential component of this management is the precise prediction of reference evapotranspiration (ET_o), a vital factor influencing the energy-water equilibrium at the terrestrial surface (Pang and Zhang, 2023; Tosan *et al.*, 2026a). Conventional estimation methods, such as the FAO-56 Penman-Monteith equation, provide accurate assessments of current water demand based on real-time meteorological data (Chang *et al.*, 2025); however, they inherently lack the ability to forecast future situations (Umutoni and Samadi, 2024). Consequently, there is a shift towards multi-step-ahead forecasting, which involves projecting agricultural water requirements multiple days

*Corresponding author Email: moein69tosan@alumni.ut.ac.ir

Cite this article: M. Tosan, A. Shayeghi, J. Teymouri, A. Bakhtar, Explaining evapotranspiration dynamics via CNN-LSTM and temporal SHAP: a multi-step forecasting approach across diverse climates, *Journal of Applied Research in Water and Wastewater*, x (x), 202x, xx-xx. doi: https://doi.org/10.22126/arww.2026.13312.1456

ahead, essential for proactive irrigation management, reservoir oversight, and the early identification of flash droughts (Dikshit and Davis, 2025; Sreeparvathy, Debdut and Mishra, 2025). The complexity of ET_o forecasting arises from its non-linear dependence on multivariate atmospheric interactions that exhibit significant temporal inertia or memory effects (Ali *et al.*, 2025; Necesito *et al.*, 2025). As a result, data-driven methodologies, especially deep learning (DL) frameworks (Qian, Wang and Chen, 2024), have become more prominent than conventional physical and statistical models. Recurrent neural networks (RNNs) (Goodarzi *et al.*, 2025), especially long short-term memory (LSTM) networks (M. Li *et al.*, 2024), have revolutionized hydrological time-series forecasting through the use of gating mechanisms that proficiently capture long-term relationships and mitigate the vanishing gradient problem (Talebi, Samadianfard and Valizadeh Kamran, 2025). Recent advancements include bidirectional LSTM networks (BiLSTM) and hybrid convolutional architectures (CNN-LSTM) that concurrently capture temporal sequences and extract local feature patterns, exhibiting enhanced performance in complex basins (Chen, Chen and Zhu, 2026; Yin *et al.*, 2025).

Notwithstanding the remarkable prediction capabilities of these sophisticated deep learning models, their practical use is often obstructed by the black-box issue (Wang *et al.*, 2026). While a model may accurately forecast an increase in evapotranspiration, it typically fails to clarify the underlying causes of this prediction (Zhang *et al.*, 2025); whether due to a prolonged heatwave that began three days earlier or a sudden wind anomaly (Ozupek *et al.*, 2025). Addressing interpretability is crucial, since comprehending the physical elements influencing hydrological dynamics is vital for fostering trust among water managers. Despite the successful implementation of explainable artificial intelligence (XAI) methodologies, such as SHapley additive explanations (SHAP), in static regression models (Elbeltagi *et al.*, 2025), their utilization in temporal deep learning architectures to elucidate the time-lagged effects of meteorological variables is still nascent (Gao *et al.*, 2025; Nourani *et al.*, 2025a). Understanding this temporal black box is crucial for grasping the dissemination of drought signals throughout the hydrological cycle across many climate regimes (Long *et al.*, 2026).

Moreover, the current literature frequently neglects the climate variability of forecast accuracy (Lakhiar *et al.*, 2025; Nkiaka *et al.*, 2024). Many studies focus on single-site applications or short-term timeframes (e.g., t+1), overlooking the decline in model performance during medium-term periods (e.g., t+7) and the applicability of acquired physical laws in diverse contexts (Ye *et al.*, 2025). The variation in memory depth of the hydrological system between energy-limited (humid) and water-limited (arid) zones is uncertain (Jiao *et al.*, 2024), as is the influence of these variances on the necessary structure of forecasting models (Rezvani Moghaddam *et al.*, 2016; X. Li *et al.*, 2024). Mitigating these uncertainties is essential for creating resilient, geographically transferable early warning systems capable of functioning in data-deficient areas (Ye *et al.*, 2025).

This paper introduces a complete and interpretable DL framework for multi-step-ahead ET_o forecasting in three distinct hydro-climatic zones of Iran: Birjand (arid), Mashhad (semi-arid), and Gorgan (humid). This study extends beyond basic accuracy evaluation by incorporating temporal SHAP with sophisticated recurrent networks to analyze the model's decision-making process. The explicit objectives are to: (1) quantify the memory effect and persistence of meteorological drivers through temporal autocorrelation analysis; (2) evaluate the efficacy of LSTM, BiLSTM, and CNN-LSTM architectures for short- to medium-term forecasting horizons; (3) utilize temporal interpretability to elucidate climate-specific lag mechanisms, differentiating between immediate radiative drivers and enduring aerodynamic forcing; and (4) examine the model's robustness to extreme events and its spatial generalization ability across varied climates. This study presents an innovative framework for physics-aware data-driven modeling, delivering practical insights for sustainable water management in a shifting environment.

2. Materials and methods

2.1. Study area and data acquisition

This study evaluates the proposed forecasting framework's robustness across diverse hydro-climatic regimes by analyzing three synoptic stations in Iran (Fig. 1); Birjand, which exemplifies an arid climate with significant advective forcing (Khalili, Fayaz and Zolfaghari, 2022; Tosan *et al.*, 2024); Mashhad, which represents a semi-arid transitional zone (Zamanipoor and Rahnama, 2025); and Gorgan, which is defined by a humid, energy-constrained coastal environment (Borna *et al.*, 2023). The unique climatic traits of different sites offer a thorough evaluation of model generalizability under diverse atmospheric boundary conditions. Daily meteorological data for a 24-year period (2000–2023) were acquired from the Iran Meteorological Organization (IRIMO). The data includes maximum and minimum air temperatures (Tmax, Tmin), average relative humidity (RH), wind speed (WS) at a height of 2 meters (u2), and duration of sunlight (n) (Table 1). ET_o was computed utilizing the standard FAO-56 Penman-Monteith equation (Allen, 2000), which functioned as the benchmark for supervised learning. Missing values, comprising less than 2% of the dataset, were imputed by linear interpolation to maintain the temporal continuity essential for RNNs.

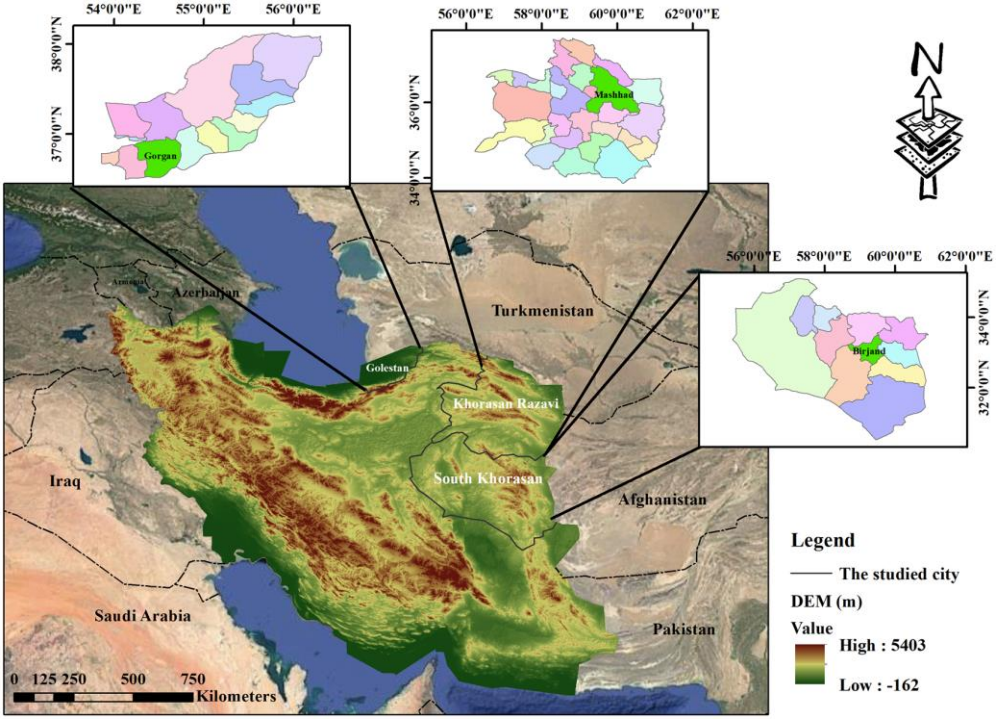


Fig. 1. Geographical location of the study area.

Table 1. Descriptive statistics of daily meteorological variables and ET_o across the study stations. (Note: SD = Standard deviation, Tmax/Tmin=Max/Min temperature, AT = Average temp, WS = Wind speed, RH = Relative humidity, SS = Sunshine hours).

Station	Variable	Unit	Mean	SD	Min	Max
Birjand	T_{max}	°C	25.36	9.36	-3.8	43.0
	T_{min}	°C	8.86	8.71	-19.6	28.8
	WS	m/s	2.99	1.52	0.0	11.4
	RH	%	31.99	18.48	3.0	97.8
	SS	hr	9.39	3.07	0.0	13.7
	ET_o	mm/d	5.65	3.29	0.4	16.5
Mashhad	T_{max}	°C	20.89	10.02	-12.4	43.0
	T_{min}	°C	7.91	7.63	-22.0	28.0
	WS	m/s	1.89	1.09	0.0	10.5
	RH	%	49.32	21.36	5.0	100.0
	SS	hr	7.97	3.86	0.0	13.5
	ET_o	mm/d	3.82	2.45	0.2	12.3
Gorgan	T_{max}	°C	22.42	7.96	0.4	46.0
	T_{min}	°C	11.96	6.81	-8.4	30.0
	WS	m/s	1.76	1.25	0.0	16.5
	RH	%	73.16	13.20	16.0	100.0
	SS	hr	6.03	3.99	0.0	12.9
	ET_o	mm/d	3.09	1.73	0.3	10.6

2.2. Data preprocessing and temporal sequencing

To ensure the robustness of the forecasting framework and prevent data leakage, the dataset was split into training (80%) and testing (20%) sets chronologically. The Min-Max scaling technique was fitted only on the training partition to estimate the minimum and maximum parameters:

$$X_{norm} = \frac{X_i - X_{\min(train)}}{X_{\max(train)} - X_{\min(train)}} \quad (1)$$

These parameters were then used to transform the testing set. This protocol ensures that future information from the test set does not influence the model training process. After normalization, a sliding window method was used to convert the supervised regression problem into a time series forecasting task (Tang *et al.*, 2025). The input matrix was reorganized into 3D sequences with a 7-day lookback window (t-7 to t-1) to capture the system's short-term climatic memory. The forecasting objectives were set for three different time horizons: one day (t+1), three days (t+3), and seven days (t+7) ahead, allowing for the measurement of model performance degradation with time.

2.3. DL architectures

Three advanced recurrent neural network architectures were developed and rigorously benchmarked to capture the non-linear temporal dependencies of ET_o The models were implemented using the Keras framework with TensorFlow backend, and training was conducted using the Adam optimizer with a learning rate of 0.001 and MSE loss function. To prevent overfitting, an early stopping mechanism was employed with a patience of 10 epochs. The hyperparameters, including the number of LSTM units (64), dropout rate (0.2), and lookback window size (7), were determined through a systematic grid search procedure. A validation subset (the last 20% of the training set) was used to evaluate various combinations, and the configuration that minimized the Root Mean Square Error (RMSE) on this validation data was selected for the final models.

2.3.1. LSTM

To overcome the vanishing gradient problem inherent in traditional RNNs, LSTM networks utilize a gating mechanism to regulate the information flow. The proposed LSTM architecture consists of an input layer accepting sequences of shape (lookback=7, features=6), followed by a hidden LSTM layer with 64 neurons and tanh activation function. A dropout layer with a rate of 0.2 was inserted to enhance regularization, followed by a dense output layer with a single neuron for regression. The mathematical operations of the LSTM cell are governed by the following equations (Hosseini et al., 2025):

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f) \quad (2)$$
$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i) \quad (3)$$
$$\tilde{C}_t = \tanh(W_c . [h_{t-1}, x_t] + b_c) \quad (4)$$
$$C_t = f_t . C_{t-1} + i_t . \tilde{C}_t \quad (5)$$
$$o_t = \sigma(W_0 . [h_{t-1}, x_t] + b_0) \quad (6)$$
$$h_t = o_t . \tanh(C_t) \quad (7)$$

where, σ denotes the sigmoid activation function, and W and brepresent the learnable weight matrices and bias vectors, respectively.

2.3.2. Bidirectional LSTM (BiLSTM)

Unlike standard LSTM which processes data only in the forward direction, BiLSTM processes the input sequence in both forward and backward directions, enabling the network to capture future context relative to past states within the training window. This architecture has demonstrated superior performance in complex hydrological modeling (Huang and Zhang, 2024). Our implementation features a bidirectional layer with 64 units (effectively 128 internal states), followed by a dropout layer (0.2) and a dense output layer. The final hidden state h_t is generated by concatenating the forward (h_t) and backward ($\overleftarrow{h_t}$) states:

$$h_t = [h_t \oplus \overrightarrow{h_t}] \quad (8)$$

2.3.3. CNN-LSTM hybrid

This hybrid architecture leverages the strengths of both convolutional neural networks (CNN) and LSTMs. A 1D convolutional layer (64 filters, kernel size=2, ReLU activation) is first employed to automatically extract high-level local features and filter high-frequency noise from raw meteorological sequences; a capability particularly beneficial for stochastic variables like wind speed in arid regions (Gao *et al.*, 2025). The output is then downsampled via a MaxPooling1D layer (pool size=2) before being fed into an LSTM layer with 50 units. Finally, a dense layer generates the multi-step forecast.

2.4. Explainable AI framework: Temporal SHAP

To unravel the black-box nature of the DL models and identify the time-lagged drivers of ET_o, SHAP were employed. Specifically, the KernelExplainer (a model-agnostic method) was utilized to approximate the contribution of each feature at each time step (Lundberg and Lee, 2017; Tosan *et al.*, 2026b). The SHAP value $\phi_{i,t}$ represents the marginal contribution of feature i at time lag t to the model's prediction, relative to a baseline background dataset. This allows for the construction of temporal heatmaps, which visualize the memory depth of the model across different climates. For the reproducibility of the XAI results, a representative background dataset of 100 samples was selected from the training partition using a k-means clustering summarizer. This ensures that the SHAP values are calculated against a baseline that captures the median climatic state of each station.

2.5. Drought propagation analysis

To investigate the reliability of the forecasting models under extreme conditions, a drought propagation analysis was conducted. The standardized anomaly (SETA) was calculated to categorize the evaporative stress levels into normal (SET A < 1), moderate (1 ≤ SET A < 2), and severe (SET A ≥ 2). The distribution of forecast errors was then analyzed across these categories to determine if the model's performance degrades during high-stress anomaly events, providing insights into the hydrological resilience of the forecasting framework (Long *et al.*, 2026; Nikoo *et al.*, 2026).

2.6. Model evaluation metrics

The predictive performance of the models was rigorously evaluated using three standard statistical metrics: the coefficient of determination (R^2), RMSE, and mean absolute error (MAE). These metrics quantify

the goodness-of-fit, the magnitude of error, and the average deviation, respectively:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (10)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (11)$$

where, y_i is the observed ET_o, is the forecasted value, $\bar{y}$ is the mean of observed values, and is the number of samples. Additionally, a spatial generalization test was performed by training the model on data from arid regions and testing it on the humid region (and vice versa) to assess transferability.

3. Results and discussion

3.1. Temporal memory and lagged correlations

Before deploying DL models, it is critical to understand the temporal persistence or memory of the hydrological system. As illustrated in Fig. 2, the temporal autocorrelation heatmaps for the three studied cities reveal how meteorological conditions from the past 1 to 7 days (lag-1 to lag-7) correlate with the current ET_o. Across all climates, the auto-correlation of ET_o itself is remarkably high at lag-1 (> 0.94), confirming that evapotranspiration is a persistent process with strong temporal continuity. This finding justifies the use of RNNs that explicitly model these sequential dependencies, consistent with recent studies emphasizing the importance of memory in hydrological forecasting (Hosseini, Prieto and Álvarez, 2025). However, a distinct climatic divergence is observed in the persistence of drivers. In the arid zone of Birjand, a unique feature is the persistence of WS, where the correlation remains strong even at lag-7 (r ≈ 0.54). This contrasts sharply with the semi-arid climate of Mashhad, where the wind correlation drops significantly to r ≈ 0.38 at lag-1. This implies that in arid regions, advective energy follows a more structured, multi-day pattern, such as prolonged wind storms, whereas in semi-arid zones, wind events are more stochastic and short-lived. Conversely, in the humid climate of Gorgan, the memory of maximum temperature (Tmax) is weaker compared to arid zones (r ≈ 0.73 at lag-1 vs. 0.86 in Birjand). This suggests that in humid climates, ET_o is more responsive to rapid, high-frequency changes in cloud cover and radiation rather than stable thermal inertia. These lag-response patterns suggest that a one-size-fits-all lookback window might be inefficient; arid regions may benefit from longer input sequences to capture wind persistence, while humid regions require models sensitive to rapid fluctuations (Khan *et al.*, 2025).

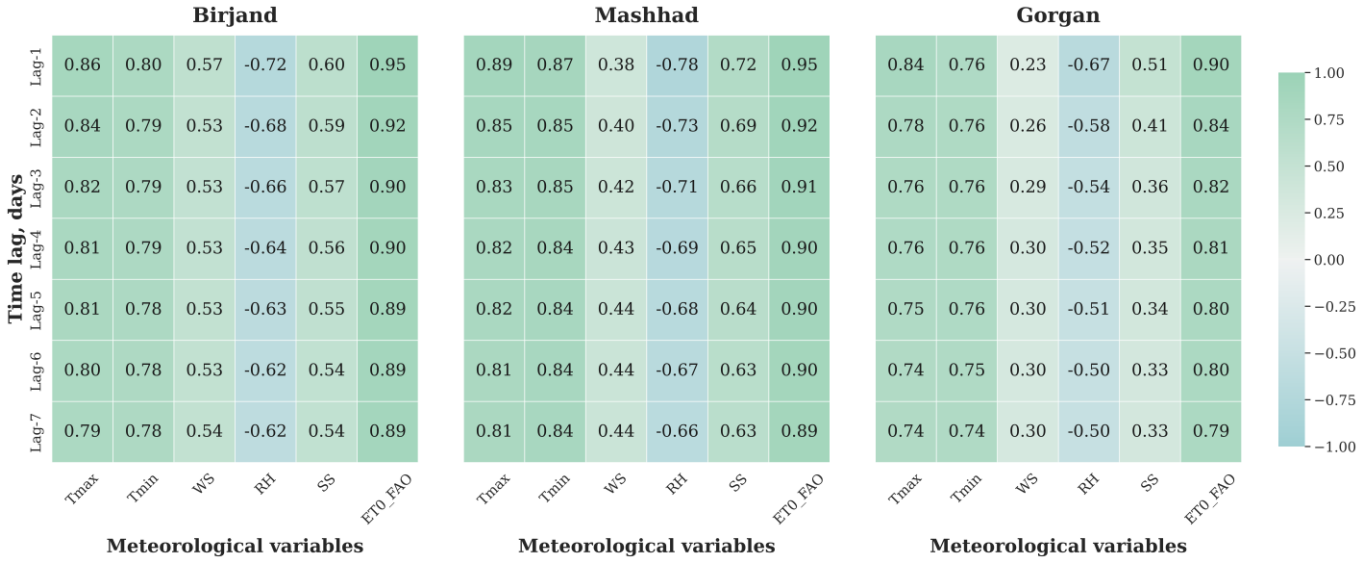


Fig. 2. Time-lag correlation heatmaps illustrating the memory effect of meteorological variables on ET_o across three climatic zones. The color intensity represents the Pearson correlation coefficient between the variable at t-k days and ET_o at day t. Note the high persistence of WS in Birjand compared to Mashhad.

3.2. Multi-step forecasting performance and model intercomparison

The predictive performance of DL architectures was assessed for forecasting horizons ranging from short-term (1-day) to medium-term (7-day) across the three climatic zones (Table 2). Fig. 3 depicts the decline in model accuracy (R^2) as the forecast horizon lengthens. Contrary to the widespread belief that increased model complexity enhances performance, the algorithmic benchmarking demonstrated that the standard LSTM model demonstrated remarkable robustness, often outperforming or equaling the performance of more intricate BiLSTM and CNN-LSTM architectures, particularly in short-term forecasting. For instance, in Mashhad (1-day), LSTM achieved the highest accuracy (R^2 = 0.934), indicating that unidirectional temporal dependencies are sufficient for effectively modeling the key semi-arid variables. However, in the arid region of Birjand over a seven-day period, the CNN-LSTM hybrid model demonstrated a slight superiority (R^2 = 0.848 compared to 0.830 for BiLSTM), suggesting that the convolutional layers effectively capture local patterns within the persistent wind series characteristic of desert climates. This advantage is also supported by Gao *et al.* (2025) in comparable hybrid modeling scenarios. A particular consequence linked to humid climates was identified during the assessment of the model's climatic resilience. Although models in arid (Birjand) and semi-arid (Mashhad) regions demonstrated high accuracy (R^2 > 0.92 for 1-day), their performance in Gorgan (humid) was comparatively less robust (R^2 approximately 0.86). This discrepancy in performance becomes increasingly evident over longer durations; after 7 days, Gorgan's accuracy decreases to approximately 0.75, whereas Birjand sustains an R^2 value of around 0.85. This phenomenon is ascribed to the stochastic variability of cloud cover and precipitation in humid regions, which generates high-frequency noise that is more challenging to forecast than the stable, inertia-driven radiative and aerodynamic cycles typical of arid zones (Yang *et al.*, 2025; Mardani *et al.*, 2025). Furthermore, the decline in accuracy demonstrates a non-linear trend, with the most significant decrease observed between the 1-day and 3-day horizons (e.g., approximately 5-7% reduction in R^2), reflecting a swift deterioration of initial condition information. Beyond a duration of three days, the decay rate stabilizes, indicating that the models accurately reflect the inherent climatic memory or seasonal baseline, thus facilitating reliable forecasts extending up to one week ahead (Tang *et al.*, 2025; Nourani *et al.*, 2025b).

3.3. Spatiotemporal interpretability: Unveiling the model's memory

To elucidate the opaque decision-making mechanism of the LSTM network, we used SHAP values tailored for time-series data, a method progressively acknowledged for improving transparency in hydrological deep learning (Gao *et al.*, 2025). In contrast to conventional feature significance that consolidates effect over time, the suggested temporal SHAP analysis (Fig. 4) delineates the contribution of each meteorological variable at distinct time lags (t-1 to t-7). In Gorgan, the SHAP heatmaps reveal a significant recency bias, with the model predominantly concentrating on the recent past (t-1 and t-2). Meteorological conditions from three to seven days prior exert minimal influence on current forecasts. This demonstrates the transient nature

of weather systems in humid or littoral environments, where atmospheric stability is limited and system memory is short-lived, a characteristic also observed in other energy-limited contexts (Elbeltagi *et al.*, 2025).

In contrast, the attention mechanism in Birjand is allocated more uniformly across the lookback window. WS demonstrates a consistent SHAP signature dating back to t-5. This validates the hypothesis that arid advective events are not instantaneous occurrences but sustained patterns; the model has effectively discerned that a windy trend initiated four days prior serves as a significant precursor for elevated evaporation today, corroborating the findings of Achite *et al.* (2025) concerning the essential role of aerodynamic factors in arid regions. Feature-specific dynamics also arise; in all cities, temperature (Tmax) consistently exhibits the most significant initial effect at t-1. RH exhibits a distinctive cumulative effect in Mashhad, with significant SHAP values persisting across multiple lag steps, indicating that the soil-atmosphere moisture equilibrium in semi-arid regions has a prolonged relaxation time compared to the energy-limited regime of Gorgan (Yonaba *et al.*, 2024). Physically, this persistent influence of wind speed in arid zones underscores the critical role of advective energy. In desert environments, prolonged wind events continuously replace the saturated air layer above the surface with dry air, maintaining a high vapor pressure deficit and sustaining elevated evaporative demand long after the initial atmospheric disturbance.

3.4. Seasonal reliability and error distribution

The operational reliability of the forecasting models at different phenological stages was assessed by examining the temporal distribution of prediction errors using polar plots (Fig. 5). The MAE exhibits distinct seasonal hysteresis that varies by climate. The error profile in Birjand is significantly asymmetric, indicating instability during the transitional season. Contrary to expectations that errors would peak at the summer maximum of ET_o, the highest MAE values are seen in the transitional months of April (spring) and October–November (autumn). This anomaly is ascribed to the synoptic instability inherent in transitional phases of arid continental climates, when abrupt wind gusts and rapid heat variations impair the climatic memory upon which the LSTM depends. This observation is consistent with the seasonal error patterns reported by Khan *et al.* (2025) in the Yellow River basin, where transitional seasons posed the greatest challenge for time-series models. In contrast, the model exhibits strong performance during the peak summer months (June–August), capitalizing on the steady, high-pressure systems that prevail in the desert environment. Conversely, Gorgan has a more concentric error distribution, albeit with a minor expansion during the winter months. This illustrates the difficulty of predicting ET_o under cloud-dominated conditions, because incoming solar radiation, the principal factor in humid areas, is markedly stochastic and disconnected from the more consistent temperature patterns (Wang and Zha, 2024). These findings indicate that adaptive management systems must integrate greater safety margins for irrigation planning in transition seasons of dry regions, whereas in humid areas, uncertainty is more evenly distributed across the year.

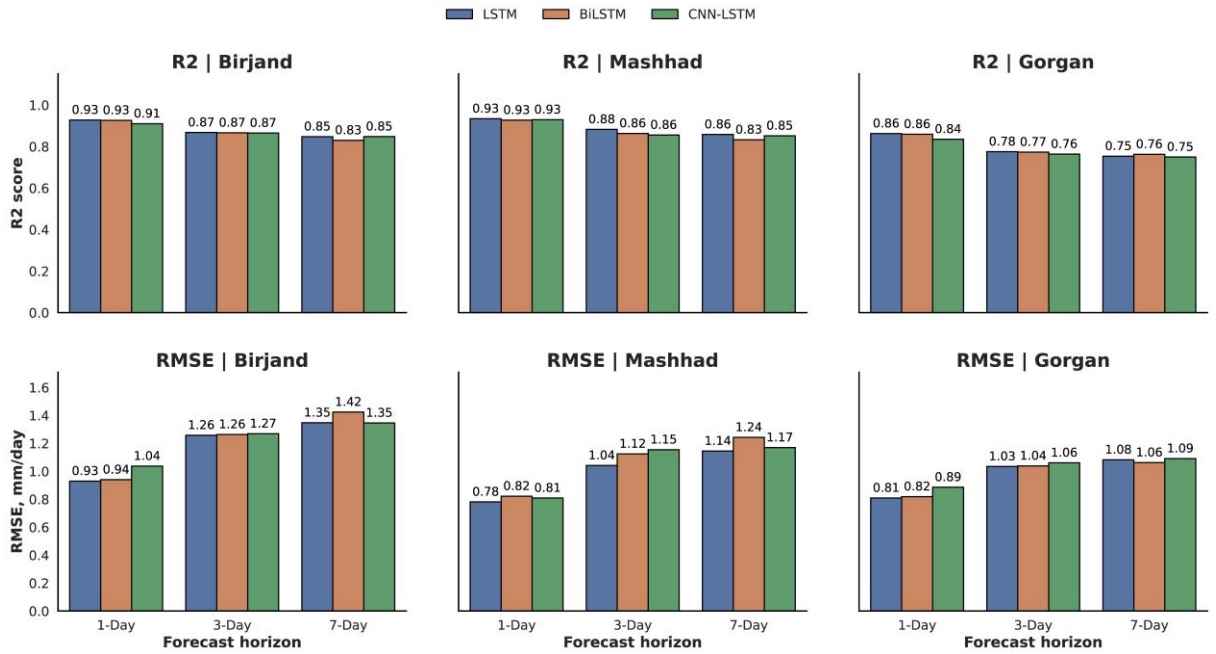


Fig. 3. Degradation of forecast accuracy (R^2) over time horizons (1, 3, and 7 days) for LSTM, Bi-LSTM, and CNN-LSTM models across three climatic zones. The steeper slope in Gorgan highlights the humid-climate penalty in medium-term forecasting.

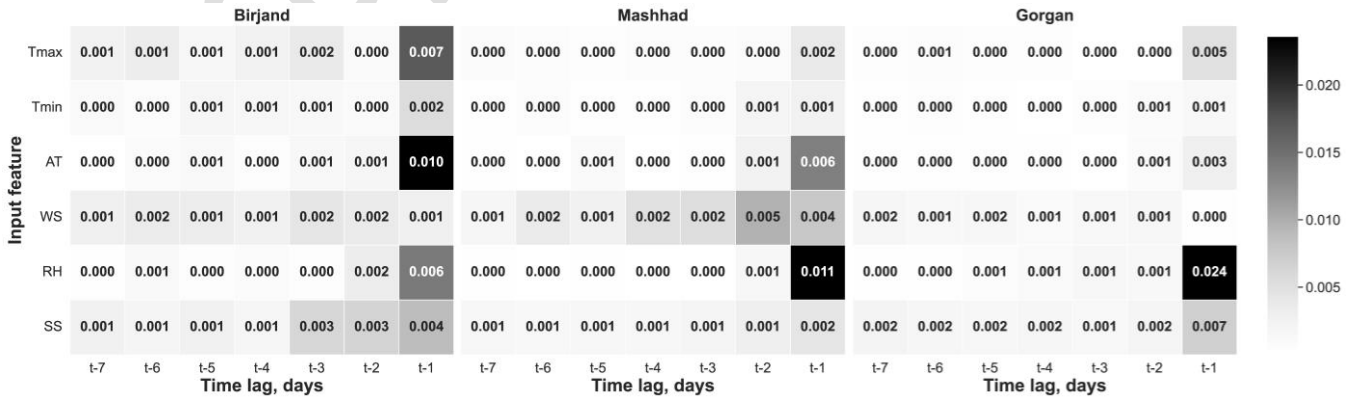


Fig. 4. Temporal SHAP heatmaps visualizing the model's attention mechanism. The x-axis represents the time lag (days past), and the y-axis represents input features. Brighter colors indicate a higher contribution to the forecast.

3.5. Anatomy of extreme evapotranspiration events

To demonstrate the operational utility of the forecasting framework, specific high-magnitude ET_o events were dissected (heatwaves/windstorms) to understand the model's decision-making process during critical periods. Fig. 6 visualizes the time-series response and the corresponding SHAP force contributions for peak events in Birjand and Gorgan. In Birjand (Fig. 6a-b), the model successfully anticipates a sharp rise in ET_o (>12 mm/day). The feature contribution analysis reveals that WS is the primary positive driver (red bar), accounting for nearly 40% of the anomaly magnitude. This confirms that the LSTM network correctly identifies advective

enhancement as the trigger for extreme evaporation in arid climates. Conversely, for a peak event in Gorgan (Fig. 6c-d), the model attributes the surge almost entirely to sunshine hours (SS) and maximum temperature, while RH exerts a negative (suppressive) influence (blue bar). This distinction highlights the model's ability to dynamically switch its internal logic based on the climatic context, moving beyond static regression coefficients to capture event-specific physics, a capability critical for managing extreme events as discussed by Maity *et al.* (2024).

3.6. Spatial generalization and cross-climatic transferability

A critical question for operational hydrology is whether a model trained in data-rich regions can be transferred to data-scarce regions with different hydro-climatic characteristics. To address this, spatial generalization tests were conducted (Fig. 7). The model trained on arid/semi-arid data (Birjand + Mashhad) and tested on the unseen humid climate of Gorgan demonstrated remarkable generalization capacity (R^2 = 0.95, bias = +0.26 mm/day). This suggests that the physical laws governing evapotranspiration in water-limited environments are inclusive enough to cover the dynamics of energy-limited environments. The slight positive bias indicates that the model, accustomed to high advective demand, tends to marginally overestimate ET_o in the humid north. Conversely, the reverse transfer (train: Gorgan → test: Birjand) yielded lower accuracy (R^2 = 0.92) and a systematic underestimation of peak events (bias = +0.33 mm/day). As evident in Fig. 7b, the scatter plot deviates from the 1:1 line at high ET_o values (> 8 mm/day). This domain shift failure occurs because the humid-trained model has never encountered the extreme aerodynamic forcing typical of arid zones, highlighting the asymmetry of transfer learning in hydrology: it is easier to downscale from harsh to mild climates than vice versa (Ye *et al.*, 2025).

Table 2. Performance comparison of deep learning models across different forecasting horizons.

City	Horizon	Model	R^2	RMSE (mm/day)
Birjand	1-Day	LSTM	**0.923**	**0.960**
		BiLSTM	0.913	1.018
		CNN-LSTM	0.884	1.179
	3-Day	LSTM	0.857	1.305
		BiLSTM	0.857	1.306
		CNN-LSTM	0.820	1.466
	7-Day	LSTM	0.844	1.365
		BiLSTM	**0.852**	**1.329**
		CNN-LSTM	0.824	1.448
Mashhad	1-Day	LSTM	0.931	0.797
		BiLSTM	**0.933**	**0.787**
		CNN-LSTM	0.895	0.985
	3-Day	LSTM	0.867	1.108
		BiLSTM	**0.877**	**1.064**
		CNN-LSTM	0.870	1.096
	7-Day	LSTM	0.829	1.259
		BiLSTM	**0.835**	**1.237**
		CNN-LSTM	0.806	1.336
Gorgan	1-Day	LSTM	**0.844**	**0.739**
		BiLSTM	0.838	0.751
		CNN-LSTM	0.817	0.800
	3-Day	LSTM	**0.757**	**0.916**
		BiLSTM	0.753	0.923
		CNN-LSTM	0.725	0.972
	7-Day	LSTM	0.697	1.026
		BiLSTM	**0.702**	**1.015**
		CNN-LSTM	0.696	1.026

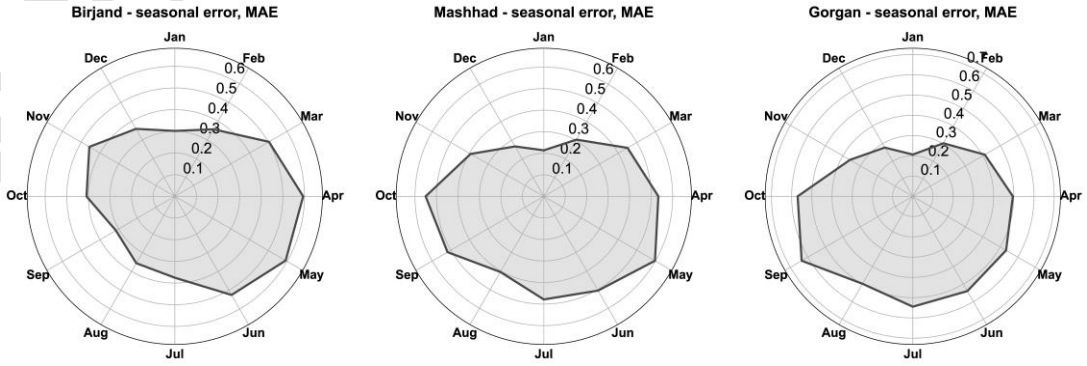


Fig. 5. Polar plots of MAE by month. The radial axis represents the error magnitude, illustrating seasonal variations in forecast reliability.

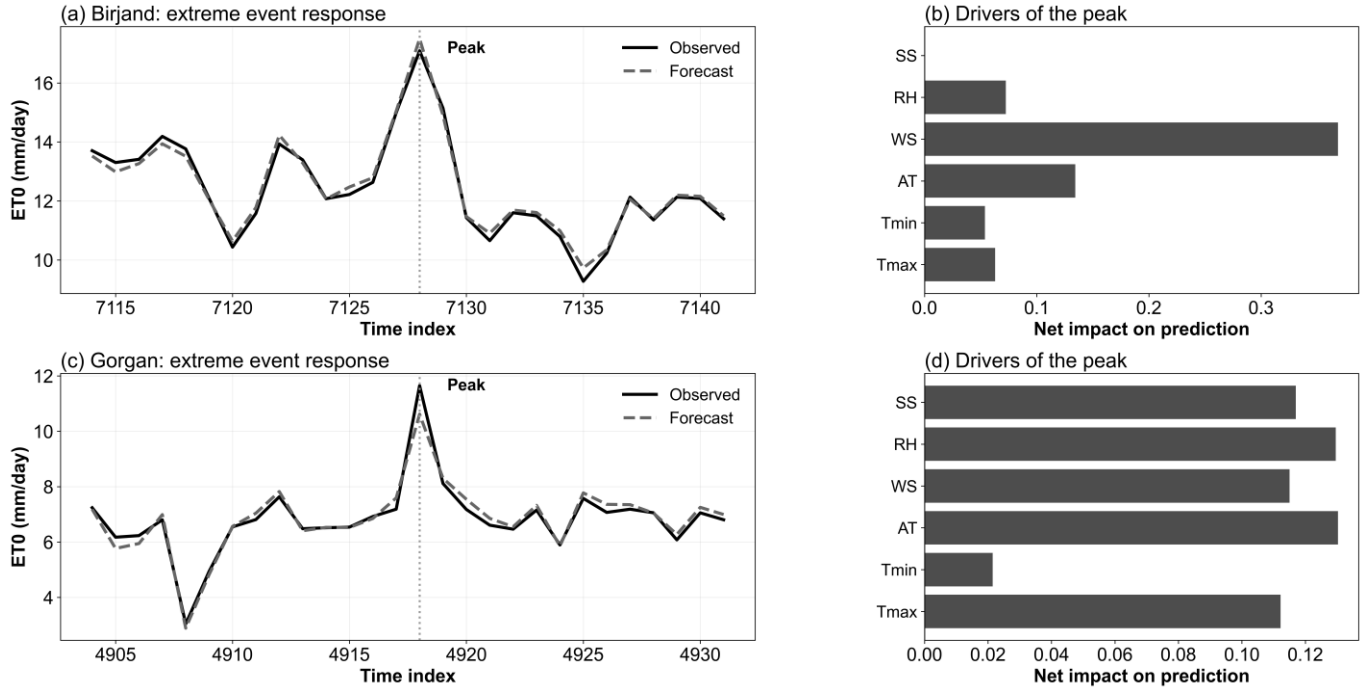


Fig. 6. Anatomy of extreme ET_o events. Left panels show the time-series forecast vs. observation during a peak event. Right panels show the SHAP contribution of each feature to the peak prediction, highlighting wind dominance in Birjand and radiative dominance in Gorgan.

3.7. Impact of drought stress on model reliability

To assess the hydrological resilience of the forecasting framework, we analyzed the distribution of prediction errors under varying levels of evaporative stress, quantified by the SETA. Fig. 8 presents the comparative boxplots of MAE across three stress categories: normal/wet (SET A < 1), moderate stress (1 ≤ SET A < 2), and severe stress (SET A ≥ 2).

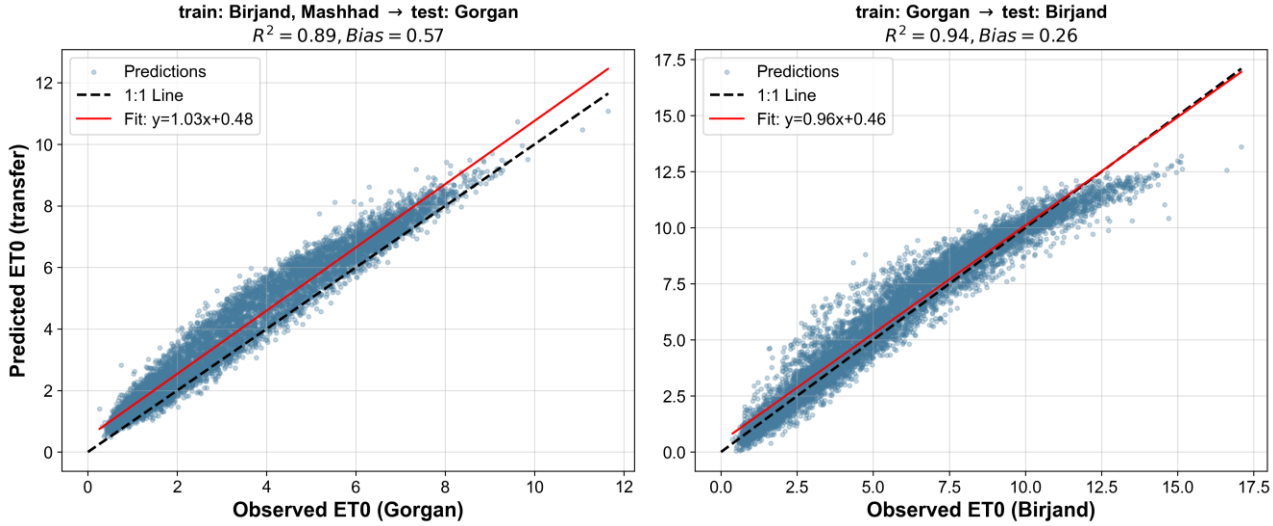


Fig. 7. Spatial generalization test results. (a) Performance of the arid-trained model tested on humid Gorgan. (b) Performance of the humid-trained model tested on arid Birjand. Note the underestimation of high values in the latter.

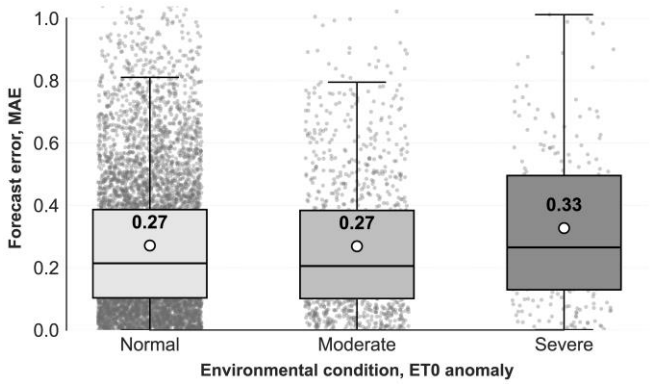


Fig. 8. Boxplots of forecast error (MAE) categorized by evaporative stress levels. The increasing spread and median of errors in the severe stress category highlight the challenge of predicting during extreme drought anomalies.

A clear degradation in model performance is observed as the system transitions into high-stress regimes. In the normal/wet phase, the forecast errors are minimal and tightly clustered (mean MAE ≈ 0.45 mm/day), indicating the model's proficiency in handling baseline conditions. However, under severe stress, the error distribution shifts significantly upward and expands (mean MAE rises to > 0.75 mm/day),

with numerous outliers extending beyond 1.5 mm/day. This phenomenon suggests a non-linear failure mode during extreme drought propagation, where the coupling between atmospheric demand and actual evaporation potential may decouple or exhibit hysteresis effects not fully captured by the 7-day lookback window. This finding aligns with the drought propagation mechanisms described by Long *et al.* (2026), who noted that hydrological response times become increasingly erratic during severe deficit periods. Consequently, while the proposed LSTM framework is robust for operational planning under typical conditions, its predictions should be interpreted with wider confidence intervals during identified severe drought anomalies.

4. Conclusions

This research developed an interpretable DL framework for multi-step-ahead forecasting of ET_o, advancing from static estimate to dynamic, time-sensitive prediction. By combining sophisticated recurrent architectures with temporal explainability methods across diverse hydro-climatic regions, the following principal results are derived:

1. Contrary to the prevailing trend of increasing model complexity, the standard LSTM network demonstrated superior stability and accuracy compared to hybrid CNN-LSTM and BiLSTM architectures, particularly for short-term forecasts (R^2 > 0.92 in arid/semi-arid zones). This suggests that for daily ET_o forecasting, capturing uni-directional temporal dependencies is often sufficient.
2. Temporal SHAP study objectively elucidated the memory depth across several climates. Arid locations (Birjand) demonstrate a prolonged memory mostly governed by WS, with advective conditions from 3-5 days earlier substantially affecting present evaporation rates. Conversely, humid areas (Gorgan) have a recency bias, predominantly influenced by immediate (1-2 day) radiative and thermal conditions, indicative of a more ephemeral atmospheric regime.
3. A non-linear decay in predictive performance was observed, with the most significant accuracy loss occurring between the 1-day and 3-day horizons. Humid climates suffered a steeper performance drop (humid-climate penalty) due to the stochastic nature of cloud cover, whereas arid zones maintained higher accuracy over longer horizons due to the persistence of high-pressure systems.

Spatial generalization tests proved that models trained on arid datasets possess a broader physical understanding and can be successfully transferred to humid regions (R^2 = 0.95). However, the reverse transfer (humid → arid) fails to predict extreme advective events, highlighting the necessity of including extreme-climate data in training corpora for universal applicability.

Conflict of Interest

The authors declare that they have no competing interests.

Author Contribution

Moein Tosan: Writing – review & editing, Methodology, Investigation, Conceptualization, Project administration, Supervision, Visualization, Writing – original draft, Resources.
Afshin Shayeghi: Visualization, Writing – original draft.
Javad Teymouri: Review & editing, Methodology.
Aydin Bakhtar: Writing – review & editing, Methodology.

Acknowledgment

The authors would like to express their sincere gratitude to the Iran Meteorological Organization (IRIMO) for providing the ground-based rain gauge data used in this study.

Data Availability Statement

Data will be available on request from the authors.

References

Abbasi, E. (2025) 'The impact of climate change on Aedes aegypti distribution and dengue fever prevalence in semi-arid regions: A case study of Tehran Province, Iran', *Environmental Research*, 275, p. 121441. doi: https://doi.org/10.1016/j.envres.2025.121441

Achite, M. *et al.* (2025) 'A coupled extreme gradient boosting-MPA approach for estimating daily reference evapotranspiration', *Journal of Theoretical and Applied Climatology*, 156(2), 113. doi: https://doi.org/10.1007/s00704-024-05313-x

Akbarpour, A. *et al.* (2024) 'Performance analysis of finite element method in groundwater studies based on Web of Science using R Biblioshiny', *Journal of Aquifer and Qanat*, 4(2), pp. 131-148. doi: https://doi.org/10.22077/jaaq.2024.7481.1071

Ali, M. *et al.* (2025) 'Improving daily reference evapotranspiration forecasts: Designing AI-enabled recurrent neural networks based long short-term memory', *Ecological Informatics*, 85, p. 102995. doi: https://doi.org/10.1016/j.ecoinf.2025.102995

Allen, R. G. (2000) 'Using the FAO-56 dual crop coefficient method over an irrigated region as part of an evapotranspiration intercomparison study', *Journal of Hydrology*, 229(1), pp. 27-41. doi: https://doi.org/10.1016/S0022-1694(99)00194-8

Borna, R. *et al.* (2023) 'Mitigation of Climate Change Impact on Bioclimatic Conditions Using Different Green Space Scenarios: The Case of a Hospital in Gorgan Subtropical Climates', *Forests*, 14(10), p. 1978. doi: https://doi.org/10.3390/f14101978

Chang, Y. *et al.* (2025) 'Machine Learning for Reference Crop Evapotranspiration Modeling: A State-of-the-Art Review and Future Directions', *Agronomy*, 15(9). p. 2038. doi: https://doi.org/10.3390/agronomy15092038

Chen, J., Chen, K. and Zhu, Y. (2026) 'Hybrid CNN-LSTM-attention model for enhanced reference crop evapotranspiration estimation toward optimized irrigation scheduling', *Journal of Agriculture and Food Research*, 25, p. 102545. doi: https://doi.org/10.1016/j.jafr.2025.102545

Dikshit, A. and Davis, C. (2025) 'Finding the link between flash drought and bushfires', *Nature Reviews Earth & Environment*, 6(5), pp. 322-322. doi: https://doi.org/10.1038/s43017-025-00675-w

Elbeltagi, A. *et al.* (2025) 'An interpretable machine learning approach based on SHAP, Sobol and LIME values for precise estimation of daily soybean crop coefficients', *Scientific Reports*, 15(1), p. 36594. doi: https://doi.org/10.1038/s41598-025-20386-y

Gao, Z. *et al.* (2025) 'An Interpretable Hybrid TCN-BiLSTM Model for Reference Evapotranspiration Prediction', *Water Resources Management*, 39(11), pp. 5481-5503. doi: https://doi.org/10.1007/s11269-025-04213-7

Goodarzi, A. *et al.* (2025) 'Prediction of pan evaporation across diverse climates and scenarios using temporal attention clockwork recurrent neural networks coupled with long short-term memory', *Water Cycle*, 6, pp. 241-253. doi: https://doi.org/10.1016/j.watcyc.2025.03.002

Hosseini, F., Prieto, C. and Álvarez, C. (2025) 'An explainable AI approach for interpreting regionally optimized deep neural networks in hydrological prediction', *Journal of Hydrology*, 661, p. 133689. doi: https://doi.org/10.1016/j.jhydrol.2025.133689

Huang, F. and Zhang, X. (2024) 'A new interpretable streamflow prediction approach based on SWAT-BiLSTM and SHAP', *Environmental Science and Pollution Research*, 31(16), pp. 23896-23908. doi: https://doi.org/10.1007/s11356-024-32725-z

Jiao, P. *et al.* (2024) 'Uncertain effect of component differences on land evapotranspiration', *Journal of Hydrology: Regional Studies*, 55, p. 101904. doi: https://doi.org/10.1016/j.ejrh.2024.101904

Khalili, S., Fayaz, R. and Zolfaghari, S. A. (2022) 'Analyzing outdoor thermal comfort conditions in a university campus in hot-arid climate: A case study in Birjand, Iran', *Urban Climate*, 43, p. 101128. doi: https://doi.org/10.1016/j.uclim.2022.101128

Khan, S. *et al.* (2025) 'Climate Impact on Evapotranspiration in the Yellow River Basin: Interpretable Forecasting with Advanced Time Series Models and Explainable AI', *Remote Sensing*, 17(1), p. 115. doi: https://doi.org/10.3390/rs17010115

Lakhiar, I. A. *et al.* (2025) 'A review of evapotranspiration estimation methods for climate-smart agriculture tools under a changing climate: vulnerabilities, consequences, and implications', *Journal of Water and Climate Change*, 16(2), pp. 249-288. doi: https://doi.org/10.2166/wcc.2024.048

Li, M. *et al.* (2024) 'Prediction of reference crop evapotranspiration based on improved convolutional neural network (CNN) and long short-term memory network (LSTM) models in Northeast China', *Journal of Hydrology*, 645, p. 132223. doi: https://doi.org/10.1016/j.jhydrol.2024.132223

Li, X. *et al.* (2024) 'A Framework for Quantifying the Uncertainty in Upscaling Evapotranspiration From Homogeneous to

Heterogeneous Underlying Surface', *IEEE Transactions on Geoscience and Remote Sensing*, 62, pp. 1-24. doi: https://doi.org/10.1109/TGRS.2024.3453589

Long, J. *et al.* (2026) 'Reconstruction of drought propagation pathways: A global analysis of multitype propagation chains and nonlinear mechanisms', *Global and Planetary Change*, 256, p. 105144. doi: https://doi.org/10.1016/j.gloplacha.2025.105144

Lundberg, S. M. and Lee, S.-I. (2017) 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems*, 30, pp. 4765-4774. doi: https://doi.org/10.48550/arXiv.1705.07874

Maity, R. *et al.* (2024) 'Revolutionizing the future of hydrological science: Impact of machine learning and deep learning amidst emerging explainable AI and transfer learning', *Applied Computing and Geosciences*, 24, p. 100206. doi: https://doi.org/10.1016/j.acags.2024.100206

Mardani, M. *et al.* (2025) 'A bibliometric analysis of research trends on the application of remote sensing in precipitation estimation with an emphasis on spatio-temporal analysis in Iran', *Iranian Journal of Rainwater Catchment Systems*, 13(2), pp. 101-118. dor: https://dor.isc.ac/dor/20.1001.1.24235970.1404.13.2.1.3

Necesito, I. V. *et al.* (2025) 'Modeling daily evapotranspiration time series based on Non-Linear Autoregressive Exogenous (NARX) method and climate variables for a data-deficient region', *PloS One*, 20(2), p. e0318675. doi: https://doi.org/10.1371/journal.pone.0318675

Nikoo, M. R. *et al.* (2026) 'Assessing the fidelity of multi-satellite precipitation estimates for drought monitoring in a mountain water tower to arid basin system', *Journal of Arid Environments*, 232, p. 105519. doi: https://doi.org/10.1016/j.jaridenv.2025.105519

Nkiaka, E. *et al.* (2024) 'Quantifying the effects of climate and environmental changes on evapotranspiration variability in the Sahel', *Journal of Hydrology*, 642, p. 131874. doi: https://doi.org/10.1016/j.jhydrol.2024.131874

Nourani, V. *et al.* (2025a) 'Ensemble machine learning-based extrapolation of Penman-Monteith-Leuning evapotranspiration data', *Ecological Indicators*, 170, p. 113012. doi: https://doi.org/10.1016/j.ecolind.2024.113012

Nourani, V. *et al.* (2025) 'Advances in multi-source data fusion for precipitation estimation: remote sensing and machine learning perspectives', *Earth-Science Reviews*, 270, p. 105253. doi: https://doi.org/10.1016/j.earscirev.2025.105253

Ozupek, E. *et al.* (2025) 'Explainable artificial intelligence to explore the intrinsic characteristics of climatic parameters governing meteorological drought forecasting: opening the black box', *Stochastic Environmental Research and Risk Assessment*, 39(8), pp. 3201-3222. doi: https://doi.org/10.1007/s00477-025-03007-y

Pang, J. and Zhang, H. (2023) 'Global map of a comprehensive drought/flood index and analysis of controlling environmental factors', *Natural Hazards*, 116(1), pp. 267-293. doi: https://doi.org/10.1007/s11069-022-05673-5

Qian, H., Wang, W. and Chen, G. (2024) 'Assessing forecast performance of daily reference evapotranspiration: A comparison of equations, machine and deep learning using weather forecasts', *Journal of Hydrology*, 644, p. 132101. doi: https://doi.org/10.1016/j.jhydrol.2024.132101

Rezvani Moghaddam, P. *et al.* (2016) 'Saffron agronomy and technology (Book of Abstracts: 2013-2016)', *Saffron Agronomy and echnology*, 4(SUPPLEMENT), pp. 1-78. doi: https://doi.org/10.22048/jsat.2016.39250

Sreeparvathy, V., Debdut, S. and Mishra, A. (2025) 'A review of advances in flash drought research: Challenges and future directions', *Earth's Future*, 13(8), p. e2025EF006603. doi: https://doi.org/10.1029/2025EF006603

Talebi, H., Samadianfard, S. and Valizadeh Kamran, K. (2025) 'Estimation of daily reference evapotranspiration implementing satellite image data and strategy of ensemble optimization algorithm of stochastic gradient descent with multilayer perceptron', *Environment, Development and Sustainability*, 27(2), pp. 3707-3729. doi: https://doi.org/10.1007/s10668-023-04037-8

Tang, M. *et al.* (2025) 'Multi-step-ahead forecasting of daily reference evapotranspiration using hybrid deep learning models for the Taklamakan Desert oasis', *Journal of Hydrology: Regional Studies*, 61, p. 102663. doi: https://doi.org/10.1016/j.ejrh.2025.102663

Tosan, M. *et al.* (2024) 'Analysis of the global research trend of saffron (Crocus sativus L.) between 2000-2023', *Saffron Agronomy and Technology*, 12(2), pp. 115-138. doi: https://doi.org/10.22048/jsat.2024.443037.1524

Tosan, M. *et al.* (2026a) 'Spatiotemporal performance and error analysis of satellite precipitation products over a topographically complex semi-arid region in Iran', *Journal of Mountain Science*, 23, pp. 118-138. doi: https://doi.org/10.1007/s11629-025-9984-6

Tosan, M. *et al.* (2026b) 'The Transparency Revolution in Geohazard Science: A Systematic Review and Research Roadmap for Explainable Artificial Intelligence', *CMES - Computer Modeling in Engineering and Sciences*, 146(1), pp. 1-41. doi: http://dx.doi.org/10.32604/cmes.2025.074768

Umutoni, L. and Samadi, V. (2024) 'Application of machine learning approaches in supporting irrigation decision making: A review', *Agricultural Water Management*, 294, p. 108710. doi: https://doi.org/10.1016/j.agwat.2024.108710

Wang, J. *et al.* (2026) 'Estimation and mechanism analysis of global evapotranspiration based on a physics-informed deep-learning model', *Journal of Hydrology*, 664, p. 134351. doi: https://doi.org/10.1016/j.jhydrol.2025.134351

Wang, Y. and Zha, Y. (2024) 'Comparison of transformer, LSTM and coupled algorithms for soil moisture prediction in shallow-groundwater-level areas with interpretability analysis', *Agricultural Water Management*, 305, p. 109120. doi: https://doi.org/10.1016/j.agwat.2024.109120

Yang, F. *et al.* (2025) 'Enhancing groundwater predictions by incorporating response lag effects in machine learning models', *Journal of Hydroinformatics*, 27(2), pp. 338-356. doi: https://doi.org/10.2166/hydro.2025.295

Ye, S. *et al.* (2025) 'Explainable transfer learning for subsurface soil moisture prediction', *Journal of Hydrology*, 661, p. 133473. doi: https://doi.org/10.1016/j.jhydrol.2025.133473

Yin, X. *et al.* (2025) 'Quantifying the time-varying period and time lag features of groundwater response: Dynamic impacts of precipitation-fed groundwater recharge', *Ecological Indicators*, 176, p. 113648. doi: https://doi.org/10.1016/j.ecolind.2025.113648

Yonaba, R. *et al.* (2024) 'Accuracy and interpretability of machine learning-based approaches for daily ET_o estimation under semi-arid climate in the West African Sahel', *Earth Science Informatics*, 18(1), p. 87. doi: https://doi.org/10.1007/s12145-024-01591-1

Zamanipoor, M. and Rahnama, M. R. (2025) 'Assessing the Role of Local Climate Zones in Shaping Land Surface Temperature in Cold Semiarid Metropolises: Evidence from Mashhad, Iran', *Weather, Climate, and Society*, 17(3), pp. 501-516. doi: https://doi.org/10.1175/WCAS-D-24-0124.1

Zhang, H. *et al.* (2025) 'Understanding Evapotranspiration Driving Mechanisms in China with Explainable Machine Learning Algorithms', *International Journal of Climatology*, 45(6), p. e8774. doi: https://doi.org/10.1002/joc.8774