Scientific Reports

https://doi.org/10.1038/s41598-026-40713-1

Article in Press

# Multi-resolution adaptive channel fusion transformer encoder LSTM for accurate streamflow prediction

Received: 2 November 2025

Accepted: 16 February 2026

Published online: 21 February 2026

Cite this article as: Apak S., Kilinc H.C., Yurtsever A. *et al.* Multi-resolution adaptive channel fusion transformer encoder LSTM for accurate streamflow prediction. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-40713-1

Sina Apak, Huseyin Cagan Kilinc, Adem Yurtsever, Hilal Haznedar & Furkan Ozkan

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

# Multi-Resolution Adaptive Channel Fusion Transformer Encoder LSTM for Accurate Streamflow Prediction

Sina Apak[1], Huseyin Cagan Kilinc*[2], Adem Yurtsever[3], Hilal Haznedar[4], Furkan Ozkan[5]

[1]Department of Management Information Technology, Istanbul Aydın University, Istanbul, Turkey
[2]Department of Civil Engineering, Istanbul Aydın University, Istanbul, Turkey
[3]Department of Environmental Engineering, Istanbul University-Cerrahpaşa, Turkey.
[4]Department of Computer Engineering, Hasan Kalyoncu University, Gaziantep, Turkey
[5]Department of Computer Engineering, Çukurova University Adana, Turkey.

*Corresponding author: huseyincagankilinc@aydin.edu.tr

**Abstract**

Accurate streamflow prediction plays a vital role in water management and flood mitigation. However, conventional deep learning models often fail to simultaneously capture short-term variability and long-term dependencies, particularly in univariate time series common in hydrological settings. To address these challenges, we propose the Multi-Resolution Adaptive Channel Fusion Transformer Encoder LSTM (MR-ACF-TE-LSTM)—hybrid architecture designed to improve predictive accuracy and interpretability by modeling temporal patterns at multiple scales. The model constructs pseudo-multivariate inputs from lagged observations, statistical summaries, and seasonal indicators, which are dynamically fused through an adaptive attention-based mechanism. Before making a prediction with LSTM, this fused representation are encoded by a transformer and then refined over time. Extensive experiments on three benchmark streamflow datasets demonstrate that MR-ACF-TE-LSTM consistently outperforms baseline models, including Transformer, Transformer-LSTM, and Bayesian CNN, achieving lower RMSE and higher $R^2$ scores. Ablation and comparative analyses demonstrate that MR-ACF-TE-LSTM attains the minimal RMSE across all FMSs, with multi-resolution yielding enhancements of up to 13% and overall RMSE reductions varying from 28% to 48% in comparison to baseline and state-of-the-art models. In comparison to baseline models, the RMSE at Dereevi FMS improved by 39%, from 1.436 to 0.874. Cross-dataset evaluations further highlight the model's robustness and generalization capabilities across heterogeneous catchments. Ablation studies confirm the critical contributions of multi-resolution inputs and adaptive fusion, while attention weight visualizations reveal the model's ability to selectively focus on temporally inputs. These findings establish MR-ACF-TE-LSTM as a powerful and interpretable framework for univariate hydrological forecasting.

**Keywords**: Deep learning, adaptive fusion, streamflow, LSTM, multi-resolution

## 1. Introduction

Accurate streamflow forecasting plays a crucial role in sustainable water resource management, flood control, and early warning systems, enabling proactive decision-making that can mitigate the impacts of extreme hydrological events[1]. However, traditional statistical and process-based hydrological models often struggle to capture the nonlinear dynamics and long-term temporal

ARTICLE IN PRESS

dependencies embedded in hydrological processes[2,3]. The inability of these models to explicitly model multiple temporal resolutions and dynamically attend to relevant historical patterns limits their interpretability and generalization, especially in complex river basins.

In particular, while standard LSTM/GRU models can miss multi-scale hydrological dependencies, CNN-based temporal models are often limited by receptive field for long-range dynamics, and Transformer-only architectures can be computationally expensive for long sequences and data-hungry—motivating our multi-resolution and adaptive fusion design.

Recent studies on time series analysis reveal the limitations of deterministic approaches based on single-resolution data, demonstrating their tendency to overlook uncertainties. Inspired by the constraints of existing models in capturing complex, nested, and multi-periodic patterns in long-term time series forecasting, Wang et al.[4] proposed a multi-resolution architecture. Their proposed model explicitly represents both the intrinsic features and the periodicity of the time series. In multivariate forecasting, it notably adapts to the unique temporal patterns of each variable. Experimental findings indicate that such methods provide more reliable feature representations by mitigating the influence of spurious signals and achieve superior accuracy compared to single-resolution models. Similarly, Multi-Res-Former[5] utilizes adaptive multi-resolution modeling to capture varying periodicities in time series. By establishing multi-resolution branches within each block, the model processes both short-term and long-term dependencies using shared Transformer architectures. This approach allows it to generate data-specific and meaningful representations without being constrained by fixed hierarchical temporal scales. In modeling time series data, hybrid approaches that can adapt to different data characteristics are also noteworthy. For instance, Cheng and Liu[6] used a hybrid surrogate model based on meta-features to overcome the challenges of small and diverse datasets. This approach successfully adapted to the complex nature of time series data by dynamically selecting the most suitable sub-models.

Furthermore, the integration of multi-resolution inputs in distribution-based forecasting models was highlighted as playing a critical role in uncertainty reduction[7,8]. Specifically, by enabling high-resolution data to capture fluctuations in greater detail, the probability distributions of future variability can be modeled more reliably. Using data from diverse temporal scales in the model's distribution forecasting component provides more granular information than single-

ARTICLE IN PRESS

resolution inputs, thereby mitigating uncertainty and enhancing forecasting accuracy. In addition, Zhang et al.[9] proposed the multi-resolution time series transformer framework to simultaneously model diverse temporal patterns at different resolutions in long-term forecasting problems. Likewise, Li et al.[10] proposed a multi-scale temporal fusion transformer architecture to overcome the challenges of complex nonlinear correlations. Both models significantly enhance prediction accuracy by deeply extracting long-range trend features across different temporal scales. Integrating information from multiple sources in time series data, especially combining data at different temporal resolutions, presents a critical challenge for improving prediction accuracy. In this context, the MGSF former model[11] introduces a Transformer architecture that eliminates information redundancy between data with different sampling intervals. The model dynamically assesses the importance of data at various granularities to fuse this information effectively. Consequently, multi-resolution data integration is a critical strategy in streamflow forecasting, as it reduces uncertainties while simultaneously modeling temporal and cross-channel relationships[12].

Adaptive channel fusion techniques have emerged as an effective means of integrating heterogeneous feature channels, allowing models to dynamically assign importance weights and capture complementary information across multiple temporal or spatial channels. For instance, Olorunnimbe and Viktor[13] emphasized the effectiveness of adaptive fusion across temporal contexts, integrating multiple temporal transformer models trained on different historical windows. Their model adaptively fuses information from distinct timeframes, enhancing robustness and prediction accuracy under both volatile and non-volatile conditions. In this context, Tao et al.[14] highlighted the potential of adaptive fusion strategies for balancing physical realism and data-driven precision in hydrological forecasting. Their model utilizes the outputs and their residual errors as inputs to Transformer, enabling the deep model to correct systematic biases and improve the overall accuracy. Additionally, transformer-based encoder architectures have recently gained significant attention in time series forecasting due to their superior ability to model long-range temporal dependencies and capture complex non-linear relationships between sequential inputs. The Rel-Informer[15], developed in this context, has improved performance in multi-step flow forecasting by utilizing a relative position encoding approach. Specifically, encoder-decoder based models were found to be significantly more successful in basins with a high volume of data. However, conventional full-attention mechanisms suffer from quadratic

ARTICLE IN PRESS

complexity and inefficient use of the entire historical sequence. To address these issues, Zhang et al.[16] proposed Dozerformer, a sparse Transformer framework with an attention mechanism that enables adaptive modeling of local, seasonal, and global dependencies while maintaining high computational efficiency. Their results highlighted the potential of sparsity-driven encoding for efficient temporal representation learning. To overcome challenges such as non-stationarity and limited sample sizes in time series data, attention mechanisms enriched with frequency information and meta-learning approaches have also been utilized. For instance, MetaGNSD former model[17], employs a meta-learning strategy to rapidly adapt to new tasks with few samples, demonstrating high performance in both point and interval predictions even under limited data conditions. Furthermore, Kang H. and Kang P.[18] addressed the problem of anomaly detection in multivariate time series by focusing on the challenge of simultaneously modeling complex inter-variable correlations and temporal dependencies. Their proposed approach adapts the Transformer's attention mechanism for dual use, separately capturing temporal dependencies within sequences and inter-variable relationships across sensors. The incorporation of temporal attention mechanisms enhances interpretability by allowing models to explicitly focus on critical time steps that contribute most to streamflow variations and peak discharge events[19]. Furthermore, Lee and Ahn[20] proposed Dualformer, a dual-encoder Transformer that separately encodes historical hydro-climatological records and meteorological forecast data before fusing them in a shared decoder. Similarly, Liu et al.[21] proposed a Transformer-based Double-Encoder model to overcome the limitations of conventional time series forecasting models that tend to support only univariate predictions. The model effectively addresses the multivariate forecasting problem by employing a cross-attention mechanism designed to capture the complex dynamic relationships between two distinct time series. These integrated approaches demonstrated that multi-source information fusion can significantly enhance both the accuracy and generalizability of time series forecasting models. A significant challenge in multivariate time series forecasting is the presence of missing data, which can lead to the incorrect modeling of spatiotemporal correlations. To address this, models like GinAR+[22] offer an end-to-end framework that combines interpolation attention mechanisms and adaptive graph convolution. Moreover, it enhances model robustness against varying missing data rates by employing contrastive learning.

Sequential decoding using Long Short-Term Memory (LSTM) layers remains a fundamental approach for generating temporally coherent predictions, as it effectively captures sequential

ARTICLE IN PRESS

dependencies and ensures consistency between input and output sequences. For instance, Liu et al.[23], integrated their model with an encoder–decoder LSTM structure to enhance medium- and long-term streamflow predictions. The dual-LSTM configuration improves the consistency between input and output sequences, leading to superior accuracy in predicting peak flow events, an essential aspect for early flood warning systems. Barzegar et al.[24] proposed a multi-time resolution model to overcome the limitations of traditional physically based modeling tools in problems characterized by high uncertainty. The model employs an ensemble architecture composed of short-sequence LSTM units, each utilizing distinct input spaces derived from local dynamics. Through the application of a multi-rate sampler, the approach enables more effective feature extraction from non-stationary time series, leading to a significant improvement in prediction accuracy. Xu et al.[25] proposed a multi-resolution LSTM approach to account for complex and multi-scale temporal variations. A two-layer LSTM structure benefits from multiple time resolutions to construct an integrated historical representation, while effectively capturing both long- and short-term temporal dependencies within it.

The hybrid method developed in this study aims to enhance the accuracy of river flow predictions. The irregular nature of river flows (i.e., non-linear structure and hydrological behavior) necessitates a multifaceted modeling approaches. To overcome these challenges, our study developed dense Transformer encoders based on a multi-resolution approach, combining adaptive channel fusion to enhance feature interactions, multi-resolution learning for scale-aware representation, and LSTM units to maintain sequential consistency. This synergistic design results in a physically consistent and highly accurate model capable of capturing complex hydrological dynamics and establishing a new benchmark for robust flow prediction.

The literature supports the fact that deep learning modeling has made significant progress in flow prediction; however, it is also observed that current models need better prediction targets in single-scale approaches. This is thought to be due to limitations caused by insufficient fit to hydrological process characteristics and hydrological parameters. While traditional CNN and LSTM-based structures struggle to simultaneously capture rapid flow responses and long-term watershed memory, Transformer-based models, despite their strengths in modeling long-term analyses, cannot maintain temporal continuity and physical consistency when used alone. Furthermore, most hybrid methods limit their adaptability across hydrological regimes by

ARTICLE IN PRESS

treating feature interactions as fixed. To overcome these limitations, this study presents a hydrology-sensitive hybrid architecture integrating multi-resolution learning, adaptive channel coupling, dense Transformer encoders, and LSTM units. This proposed hybrid method addresses fundamental hydrological conditions with more robust prediction targets. For example, long-term (time interval) hydro-meteorological predictions are captured via Transformers, temporal continuity and basin memory via LSTM units, and regime-sensitive predictive relevance via adaptive channel coupling. All these contributions form a physically consistent, robust, and state-of-the-art deep learning framework for flow velocity estimation.

This paper is organized as follows: Section 2 provides methodology and data collection, results are illustrated in Section 3, discussion is given in Section 4, finally, Section 5 presents the conclusion.

## 2. Material and Method

### 2.1. Study Region

The Western Black Sea Basin lies between the north latitudes of 40°34′42″ – 41°27′52″ and the east longitudes of 30°52′33″ – 35°12′12″, and constitutes 3.7% of Turkey's total land area (Fig. 1). The average yearly precipitation per square meter is 335 mm, and its precipitation area is approximately 28,855 $km^2$[26]. The average basin yield is 10.6 $L/s/km^2$, and the average annual flow is 9.93 $km^3$[27]. The basin was chosen as the study area due to several characteristics, including rapid surface runoff, abrupt flow changes, high flood risks, and intense rainfall that fluctuates over short periods. The practical application of the models in this study is further enhanced by the region's critical need for irrigation for agricultural crops and hydroelectric power generation. As a result, the Western Black Sea Basin can be regarded as an appropriate natural laboratory for evaluating the precision and adaptability of hydrological time-series forecasting models, especially when it comes to short-term streamflow predictions.

ARTICLE IN PRESS

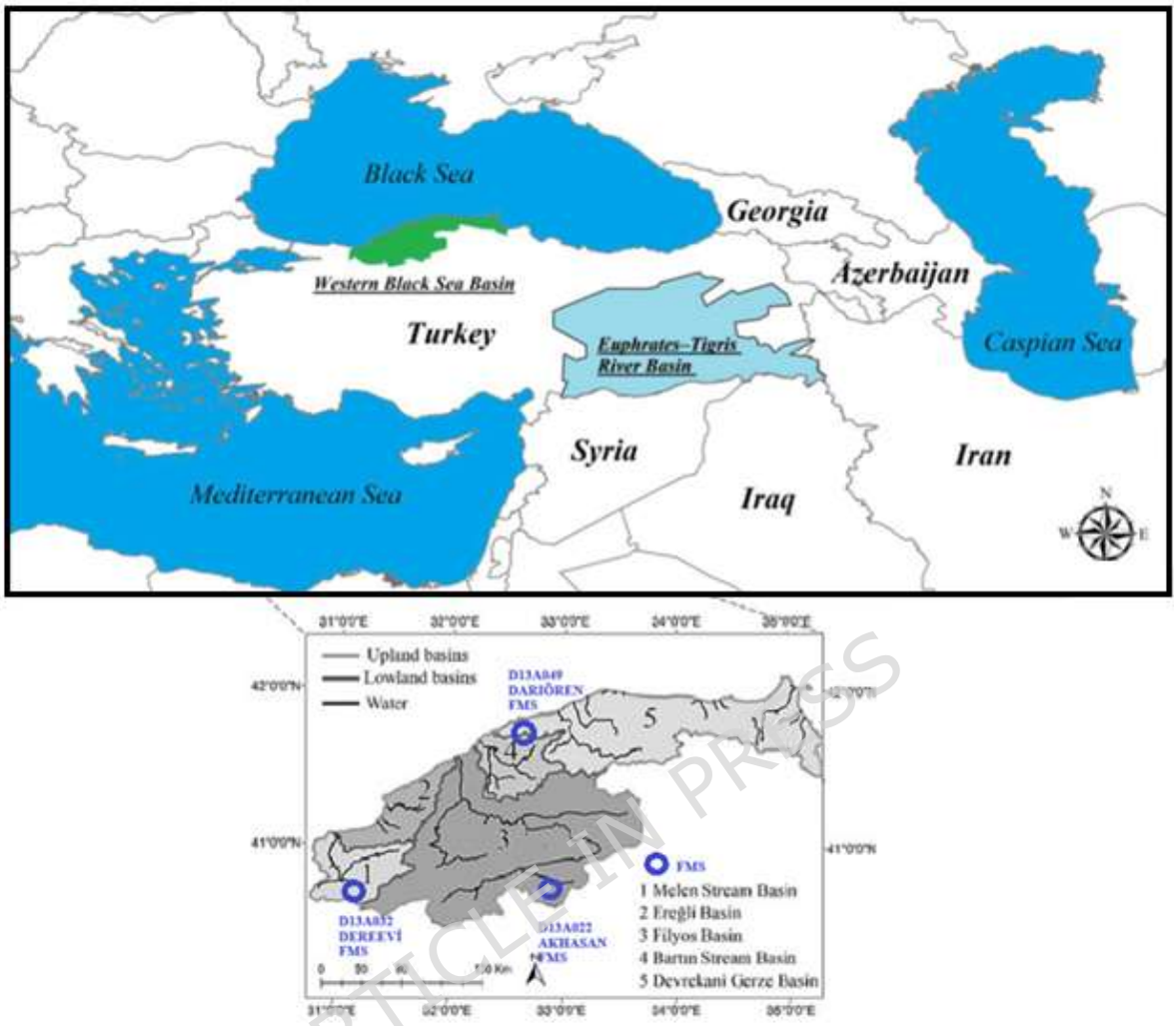



**Fig. 1.** Location of the Basin and FMSs[28,29]

This study used 10-year flow data (2001–2011) from the Darıören (D13A049), Dereevi (D13A032), and Akhasan (D13A022) Flow Measurement Stations (FMS) in the Western Black Sea Basin. The characteristics of FMSs are given in Table 1 and the daily flow rates observed in these FMSs are presented in Fig. 2.

ARTICLE IN PRESS

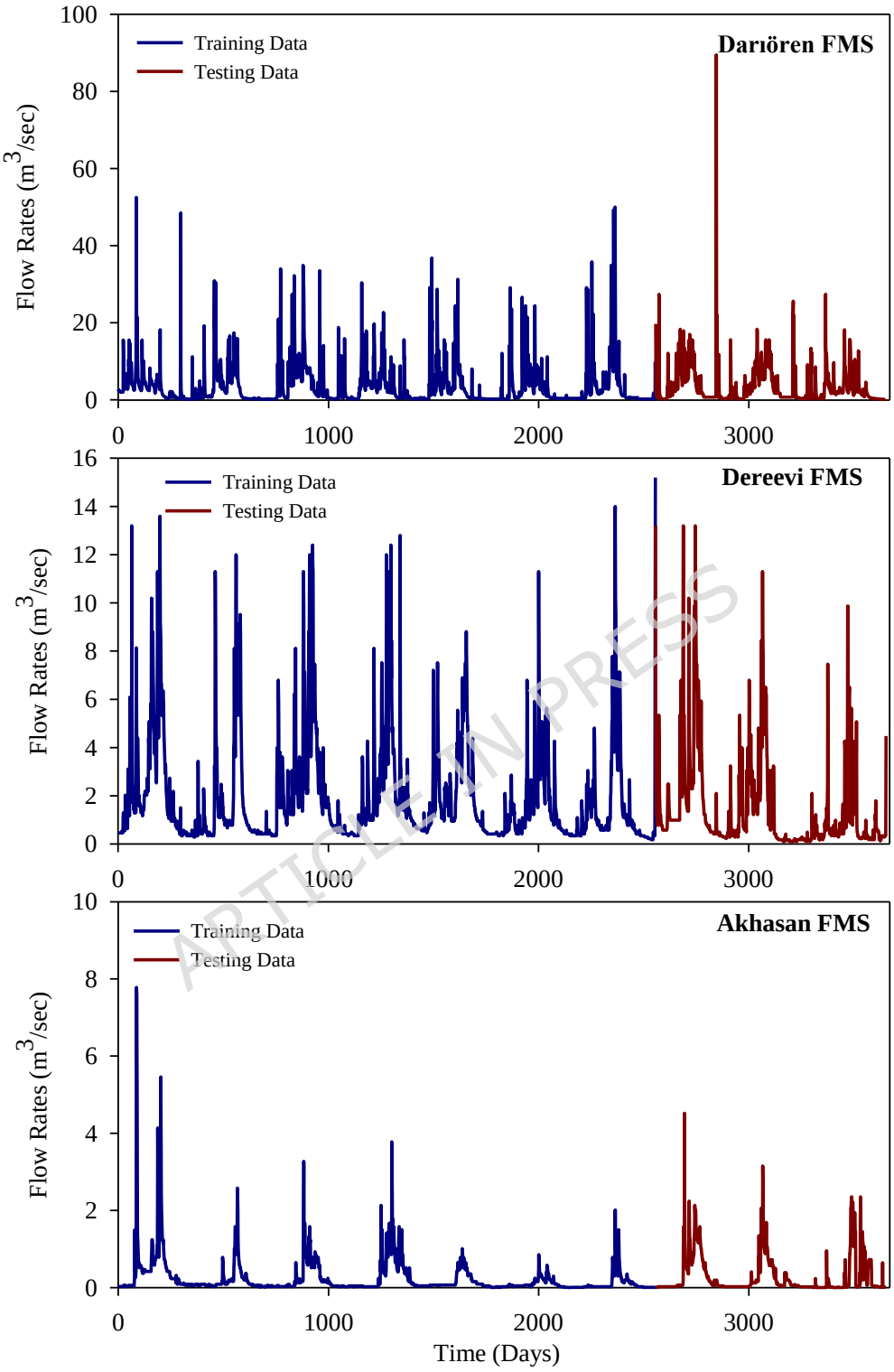



**Fig. 2.** Daily flow rates for studied FMSs

ARTICLE IN PRESS

**Table 1.** The characteristics of studied FMSs

| FMS | Darıören FMS | Dereevi FMS | Akhasan FMS |
|---|---|---|---|
| **FMS No** | D13A049 | D13A032 | D13A022 |
| **Stream** | Arıt Stream | Aksu Stream | Akhasan Stream |
| **Coordinates** | 32°31'57" E - 41°39'34" N | 30°59'50" E – 40°42'34" N | 32°48'40" E - 40°46'14" N |
| **Elevation (m)** | 230 | 873 | 1180 |
| **Catchment Area ($km^2$)** | 137 | 71.50 | 76.50 |
| **Maximum Flow Rates ($m^3/s$)** | 89.50 | 15.20 | 7.78 |
| **Minimum Flow Rates ($m^3/s$)** | 0.12 | 0.10 | 0.001 |
| **Average Flow Rates ($m^3/s$)** | 3.42±5.24 | 1.77±2.10 | 0.26±0.50 |

For model training and evaluation, the dataset for each station was split into training (70%) and testing (30%) sets. This 70:30 ratio is a common practice in similar hydrological forecasting studies and was determined to be optimal for our models after preliminary assessments. For the prediction phase, a time delay of five days was used across all models. Similarly, the daily flow values ninety-six days prior to the forecast day are used as input in all models, i.e., the input window begins 96 days in advance.

ARTICLE IN PRESS

### 2.2.Methodology

#### 2.2.1. Multi-Resolution Input Construction

Hydrological time series often exhibit complex dynamics arising from the interplay of short-term fluctuations, seasonal cycles, and long-term trends. To capture these diverse temporal patterns, we construct a multi-resolution input representation that transforms the univariate streamflow sequence into a pseudo-multivariate format. This design enables the model to jointly learn from multiple temporal granularities, thereby improving its ability to capture both transient events and persistent seasonal dependencies[30,31].

The input sequence is decomposed into resolution-specific channels, each designed to extract complementary information:

- Lagged Observations: A fixed-length sliding window of recent historical streamflow values is extracted to provide short-term temporal context. These raw lagged values

ARTICLE IN PRESS

constitute the base-resolution channel, preserving fine-grained temporal signals, as widely employed in hydrological forecasting studies[32].

- Statistical Features: For each window, statistical descriptors such as mean, standard deviation, minimum, and maximum are computed. These features summarize the distributional properties of the window, providing a smoothed, mid-resolution abstraction that is less sensitive to local noise, a practice consistent with statistical feature engineering techniques in hydrology and machine learning literature[33,34].

- Seasonal and Temporal Indicators: To account for recurring seasonal behavior, additional channels encode temporal position using categorical variables (e.g., month, day-of-year) and continuous cyclical representations (sine–cosine encodings). Previous studies have shown that such temporal indicators improve model performance in capturing seasonal and periodic hydrological phenomena[35,36].

The resulting channels are concatenated along the feature dimension to form a composite multi-resolution tensor for each time step. This tensor serves as the input to the Adaptive Channel Fusion (ACF) module, which learns to dynamically weigh the contributions of each temporal resolution during training. By explicitly incorporating multi-resolution information, this approach overcomes a major limitation of conventional univariate models, which often struggle to represent heterogeneous temporal dynamics[37]. The resulting enriched input representation provides a more robust foundation for downstream feature extraction and forecasting.

ARTICLE IN PRESS

### 2.2.2. Adaptive Channel Fusion Module

The Adaptive Channel Fusion (ACF) module is designed to integrate information from multiple temporal resolutions in a data-driven, learnable manner. Since streamflow patterns are influenced by short-, medium-, and long-term variability, naive concatenation of features may not effectively capture their relative importance. The ACF module addresses this by dynamically learning fusion weights, enabling the model to emphasize the most relevant resolution at each time step, consistent with previous findings on multi-resolution learning in hydrological forecasting[38].

*1. Input Representation*

ARTICLE IN PRESS

The multi-resolution input space is formalized as $X = \{X^1, X^2, \ldots, X^{(R)}\}$, where each component $X^{(r)} \in R^{T \times d_r}$ denotes a feature matrix at the $r^{th}$ temporal resolution, $T$ is the sequence length, and $d_r$ is the feature dimensionality at resolution $r$. Each matrix consists of a temporal sequence of length $T$ with $d_r$-dimensional features at resolution $r$. These resolutions may correspond to lag-based raw values, statistical descriptors, or seasonal indicators, similar to feature engineering techniques used in hydrological time series modeling[34].

*2. Learnable Fusion Weights*

To adaptively integrate information across different temporal resolutions, the Adaptive Channel Fusion module employs a shared attention-like gating network to compute normalized importance weights for each resolution channel[39]. This mechanism enables the model to dynamically emphasize the most informative temporal resolution depending on the input sequence, rather than relying on static or hand-crafted weighting schemes. Formally, each resolution-specific feature matrix $X^{(r)}$ is first summarized over time to obtain a compact representation:

$$\overline{\mathbf{X}}^{(r)} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t^{(r)}, \qquad r = 1, \ldots, R \tag{1}$$

where $\overline{\mathbf{X}}^{(r)}$ denotes the temporal average of $X^{(r)} \in R^{T \times d_r}$ over $T$ time steps at resolution $r$. These summary vectors are then passed through a small feed-forward network to produce scalar fusion logits:

$$z^{(r)} = \mathbf{w}^{\cdot}\, \sigma\left(\mathbf{W}_f \overline{\mathbf{X}}^{(r)} + \mathbf{b}_f\right) \tag{2}$$

where $\mathbf{W}_f \in R^{d \times d_r}$, $b_f \in R^d$, and $w \in R^d$ are learnable parameters and here, $d$ denotes the shared embedding dimension. $\sigma(.)$ is a non-linear activation function ReLU, and $w$ is a trainable projection vector mapping the hidden representation to a scalar importance score. To ensure comparability across resolutions, these logits are passed through a softmax function to obtain normalized fusion weights:

ARTICLE IN PRESS

$$\alpha^{(r)} = \frac{\exp\left(z^{(r)}\right)}{\sum_{j=1}^{R} \exp\left(z^{(j)}\right)}, \qquad r = 1, \ldots, R \tag{3}$$

such that $\alpha(r) \geq 0$ and $\sum \alpha^{(r)} = 1$[40,41]. This formulation is analogous to the attention scoring function proposed by Bahdanau et al.[42] and widely adopted in Transformer-based self-attention[39]. It can also be interpreted as a gating network, a concept well established in mixture-of-experts models and in dynamic multimodal fusion networks that learn to balance the contribution of different modalities[40,41].

The resulting normalized weights $\alpha^{(r)}$ are subsequently used to compute the fused multi-resolution representation as a convex combination of the resolution-specific channels. This data-driven weighting not only improves predictive performance by emphasizing the most relevant temporal scales but also provides a measure of interpretability, as the learned weights indicate which resolution contributed most to the model's decision for each input sequence.

*3. Fused Representation*

As shown in Eq. (4), the final fused feature vector $X_{fused}$ is obtained by calculating a weighted average of the individual feature channels $x^{(r)}$. This process uses the normalized attention scores $a^{(r)}$ as weights to emphasize the more salient channels.

$$\mathbf{X}_{\text{fused}} = \sum_{r=1}^{R} \alpha^{(r)} \cdot \mathbf{X}^{(r)} \tag{4}$$

This weighted summation yields a unified sequence $X_{fused} \in R^{T \times d}$ where $d$ is the shared embedding dimension. This adaptively emphasizes the most informative resolutions, a strategy that improves forecasting performance[43].

### 2.2.3. Transformer Encoder Module

The Transformer encoder is employed to capture long-range temporal dependencies in the fused representation $\mathbf{X}_{\text{fused}}$. Unlike recurrent neural networks, which process sequences step-by-step, the Transformer leverages self-attention to model interactions between all-time steps in parallel,

ARTICLE IN PRESS

improving both computational efficiency and the ability to capture global temporal patterns[39]. The fused input sequence is first projected into a $d_{model}$-dimensional embedding space as $\mathbf{H}_0 = \mathbf{X}_{\text{fused}}\mathbf{W}_e + \mathbf{b}_e$, and then augmented with sinusoidal positional encodings to incorporate temporal order information. This yields $Z_0 = H_0 + P$, where represents the positional encodings[39]. Each encoder layer applies multi-head self-attention (MHSA) to learn contextual relationships across all time steps:

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^{\bullet}}{\sqrt{d_k}}\right)V \tag{5}$$

where $Q = Z_0W_Q$, $K = Z_0W_K$ and $V = Z_0W_V$ are the query ($Q$), key ($K$), and value ($V$) matrices derived from the fused input. Multi-head attention computes *h* parallel attention maps:

$$\text{MHSA}(\mathbf{Z}_0) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)\mathbf{W}_O \tag{6}$$

where each head computes attention independently and $\mathbf{W}_O$ is the output projection[43]. This mechanism allows the model to jointly attend to patterns at different temporal scales and lags. The attention outputs are then refined through a position-wise feed-forward network defined as:

$$\text{FFN}(x) = \max(0, x\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \tag{7}$$

where $\mathbf{W}_1$, $\mathbf{W}_2$ are weight matrices and $\max(0,.)$ denotes the ReLU activation.

Each sub-layer (MHSA and FFN) is wrapped with residual connections and layer normalization for stable training[39]. In the MR-ACF-TE-LSTM framework, the Transformer encoder produces a context-rich representation of the input sequence that captures both local and global dependencies. This representation is subsequently refined by the temporal attention mechanism before being decoded by the LSTM [25].

$$\mathbf{H}_l = \text{LayerNorm}\left(\mathbf{H}_{l-1} + \text{MHSA}(\mathbf{H}_{l-1})\right) \tag{8}$$

$$\mathbf{Z}_l = \text{LayerNorm}\left(\mathbf{H}_l + \text{FFN}(\mathbf{H}_l)\right) \tag{9}$$

where $l$ indexes the encoder layer. Stacking $L$ such layers yield the final encoder output $Z_L$.

ARTICLE IN PRESS

### 2.2.4. Temporal Attention Mechanism

While the Transformer encoder captures contextual information across the entire sequence, not all-time steps contribute equally to the final prediction. To enhance interpretability and performance, we integrate a temporal attention mechanism that assigns a learnable importance score to each time step in the encoder output[44]. This approach follows the additive attention formulation proposed by Bahdanau et al.[42], which has proven effective in time-series forecasting and sequence-to-sequence models[45,46].

*1. Attention Score Computation*

Let $Z_L = \{z_1, z_2, \ldots, z_T\}$ denote the final output where $\mathbf{z}_t \in R^d$ represents the encoder output at time step $t$. The attention score for each time step is computed as:

$$e_t = \mathbf{v}_a^{\cdot} \tanh(\mathbf{W}_a \mathbf{z}_t + \mathbf{b}_a) \tag{10}$$

where $W_a \in R^{d_a x d_{\text{model}}}$ is a learnable weight matrix, $b_a \in R^{d_a}$ is a bias term, $v_a \in R^{d_a}$ is a trainable projection vector that converts the hidden representation into a scalar relevance score.

*2. Normalized Attention Weights*

The attention scores are normalized via a Softmax function to produce a probability distribution over time steps:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)} \tag{11}$$

where $\alpha_t(r) \geq 0$ and $\sum_{t=1}^{T} a_t = 1$, ensuring interpretability as relative importance weights.

*3. Context Vector and Weighted Summation*

The final context vector is computed as a weighted sum of the Transformer Encoder's hidden states:

ARTICLE IN PRESS

$$\mathbf{c} = \sum_{t=1}^{T} \alpha_t \mathbf{z}_t \tag{12}$$

The context vector $\mathbf{c}$ highlights the most influential time steps and serves as a compact, attention-weighted representation of the input sequence. It is subsequently fed into the LSTM decoder for streamflow prediction.

### 2.2.5. LSTM-Based Sequential Decoder

The final component of the proposed MR-ACF-TE-LSTM architecture is the LSTM-based decoder, which transforms the attention-weighted sequence into the final streamflow prediction. While the Transformer encoder effectively captures global temporal dependencies, the LSTM decoder is crucial for modeling local temporal continuity, a characteristic that is particularly important in hydrological processes where neighboring observations exhibit strong autocorrelation. The attended feature representation $c$, produced by the temporal attention mechanism, is fed into the LSTM. The LSTM then leverages its recurrent memory cell dynamics to capture residual temporal dependencies and preserve local continuity. The LSTM decoder computes its hidden and cell states iteratively according to the standard LSTM recurrence relations[47]. At each step $t$, the LSTM updates its hidden state $h_t$ and cell state $s_t$ using the standard gating mechanism:

$$\mathbf{i}_t = \sigma(\mathbf{W}_i \mathbf{c} + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i) \tag{13}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_f \mathbf{c} + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f) \tag{14}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o \mathbf{c} + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o) \tag{15}$$

$$\tilde{\mathbf{s}}_t = \tanh(\mathbf{W}_c \mathbf{c} + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c) \tag{16}$$

$$\mathbf{s}_t = \mathbf{f}_t \odot \mathbf{s}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{s}}_t \tag{17}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{s}_t) \tag{18}$$

where $\mathbf{i}_t$, $f_t$, $o_t$ represent the input, forget, and output gates, respectively, $\sigma(.)$ is the sigmoid activation, and $\odot$ denotes element-wise multiplication. After processing the full sequence, the final hidden state $h_T$ is passed through a fully connected layer to yield the forecasted streamflow value:

ARTICLE IN PRESS

$$\hat{y}_T = \mathbf{W}_y \mathbf{h}_T + b_y \tag{19}$$

This decoder stage refines the global context captured by the Transformer encoder. It achieves this by modeling fine-grained sequential dependencies, filtering noise through its gating mechanisms, and aligning the model output with the continuity and memory properties of hydrological processes. Consequently, this improves the model's prediction stability and generalization.

### 2.2.6. Proposed Approach

The proposed MR-ACF-TE-LSTM architecture is designed to enhance interpretability and cross-domain generalization in streamflow forecasting. It integrates a temporal attention mechanism that operates after multi-resolution fusion and Transformer-based encoding. After constructing pseudo-multivariate inputs through resolution-specific channels and adaptively fusing them, the model employs a Transformer encoder to capture long-range dependencies, which are then processed by a temporal attention module. This attention module learns to assign dynamic importance scores to each step within the input sequence, enabling the model to prioritize the most relevant temporal regions for prediction. Unlike fixed-weight architectures, the temporal attention mechanism can flexibly concentrate on recent observations or distribute focus across multiple lagged intervals depending on the temporal dynamics of the dataset.

ARTICLE IN PRESS

As illustrated in Fig. 3, attention weights vary significantly across samples and training datasets. In some cases, the model highlights earlier and mid-range input steps (e.g., steps 10–25 and 55–75), suggesting reliance on long-term dependencies. In others, it emphasizes recent data (steps 60–85), aligning with hydrological principles where short-term antecedent conditions strongly influence streamflow. This adaptive temporal focusing contributes not only to improved accuracy but also to interpretability, as it reveals the internal decision-making rationale of the model under both within-domain and cross-domain settings. Table 2 displays the pseudocode for the suggested method. By combining resolution-aware feature extraction, Transformer-based sequence encoding, and a learnable attention mechanism, the proposed model strikes a robust balance between predictive power and explanatory clarity. This balance is particularly critical in heterogeneous and data-scarce hydrological environments.

ARTICLE IN PRESS

**Table 2:** Multi-Resolution Attentive Channel Fusion Transformer-Encoder LSTM (MR-ACF-TE-LSTM)

| |
|---|
| Input: Time series data $X=\{x_1,x_2,\ldots, x_T\}$, Window size W.<br>Output: Forecasted value $\hat{x}_{T+1}$. |
| 1. Multi-Resolution Input Construction |
| $X_{Multi}$=[month, day-of-year, sin(2πt/365), cos(2πt/365)]. |
| 2. Adaptive Channel Fusion |
| - Attention Score Computation |
| - Score Normalization |
| - Fused Representation Computation |
| 3. Transformer Encoder Block |
| H=[h1,h2,…,hN] = TransformerEncoder ($X_{fused}$) |
| 4. Temporal Attention Mechanism |
| - Compute attention scores |
| - Normalize via SoftMax |
| - Compute context vector |
| 5. LSTM Decoder and Output Generation |

ARTICLE IN PRESS

ARTICLE IN PRESS

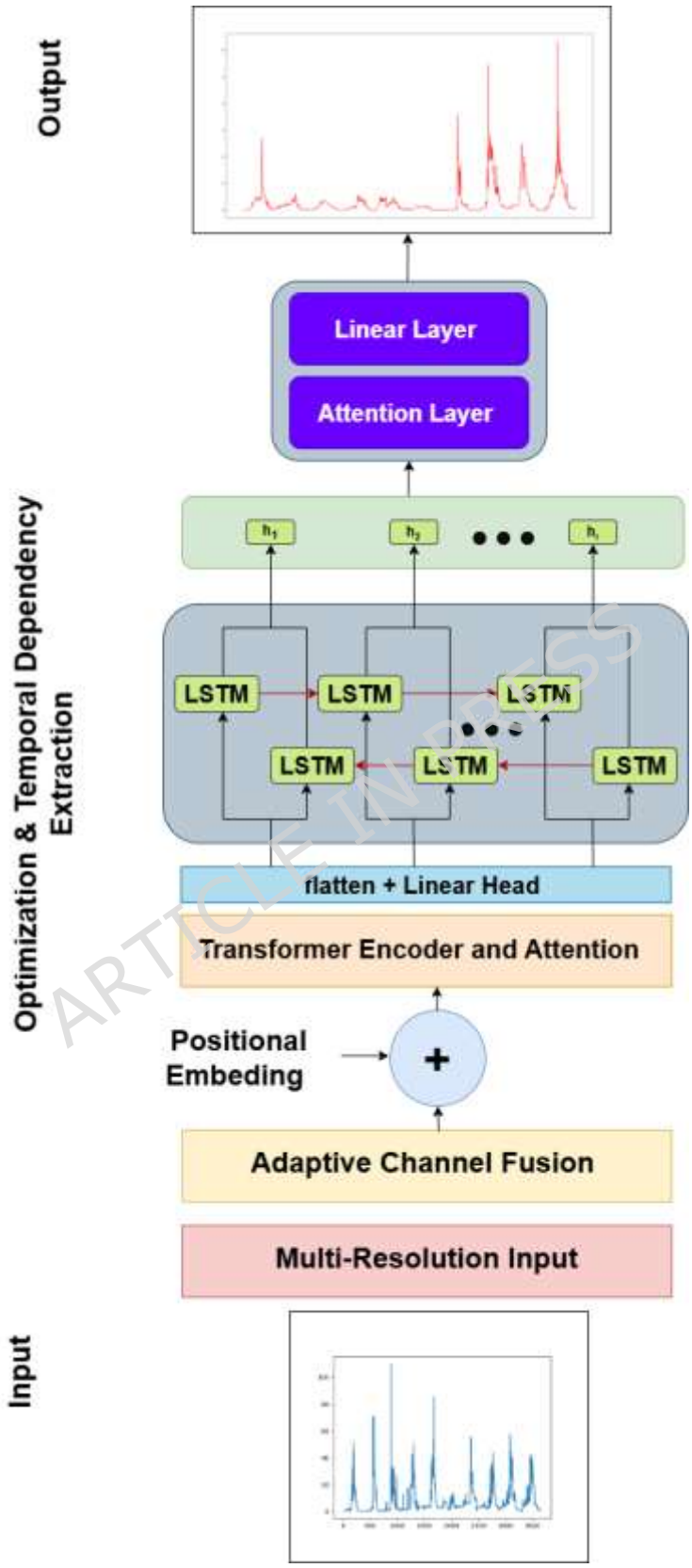



**Fig. 3.** Proposed Approach architecture

ARTICLE IN PRESS

## 3. Experimental Results

All experiments were performed on a personal workstation featuring an NVIDIA RTX 4060 GPU and an Intel Core i5-2800 CPU, which provided adequate computational resources for training deep learning models. Each dataset was processed using a sliding window approach with a window size of 32, where every input sample included 96 successive daily time steps to predict the subsequent value. Before model training, the data was normalized using Min-Max scaling and divided into training and testing sets in a 70:30 ratio, reserving 30% of the data for evaluating the model's generalization ability. Models were trained using a batch size of 32 for up to 1000 epochs per run, taking advantage of GPU acceleration to ensure efficient learning. This experimental setup was consistently applied across all datasets to maintain fairness and ensure reproducibility. In addition, to ensure a fair comparison, we tuned the hyperparameters of each baseline model using a validation split from the training set. Specifically, we conducted a small grid/random search over key hyperparameters (learning rate, hidden/embedding dimension, dropout rate, and—where applicable—the number of layers/attention heads), and selected the configuration that achieved the best validation performance. The final results reported in the paper correspond to these tuned baseline configurations.

ARTICLE IN PRESS

### 3.1. Ablation Study

Table 3 presents the ablation study results conducted on three benchmark streamflow datasets—Dereevi FMS, Darıören FMS, and Akhasan FMS—to assess the contribution of key components in the proposed Multi-Resolution Adaptive Fusion Transformer-LSTM model. The baseline configuration, which includes all components (multi-resolution processing, adaptive fusion, and Transformer-based attention), consistently outperforms its ablated variants across all datasets. Notably, removing the adaptive fusion module resulted in noticeable performance degradation, especially in the Darıören and Dereevi datasets, with RMSE increases of approximately 8–10% and corresponding drops in $R^2$. Similarly, excluding multi-resolution features led to increased prediction errors and reduced explanatory power, underscoring the importance of capturing both short- and long-term temporal patterns.

The most substantial decline in performance was observed when both the adaptive fusion and multi-resolution features were disabled simultaneously. For instance, in the Dereevi dataset, this

ARTICLE IN PRESS

combination yielded an RMSE of 1.0157 and an $R^2$ of 0.6119, compared to 0.8748 and 0.7008 in the baseline. Although removing the Transformer encoder alone caused less dramatic declines, it still affected the model's ability to capture long-range dependencies, reflected in slight drops in $R^2$ across datasets. These findings confirm that each architectural component—multi-resolution decomposition, adaptive fusion, and Transformer-enhanced temporal modeling—contributes meaningfully to the model's overall performance, and their combined use leads to the most accurate and robust streamflow predictions.

**Table 3.** Ablation study of proposed approach based on benchmark dataset.

| Dataset | **Model Variant** | **RMSE** | **MAE** | **$R^2$** |
|---|---|---|---|---|
| | **Baseline (Full Model)** | **0.8748** | **0.4943** | **0.7008** |
| | Without Adaptive Fusion | 0.9451 | 0.5328 | 0.6623 |
| Dereevi | Without Multi-Resolution Features | 0.9803 | 0.5586 | 0.6417 |
| | Without Adaptive Fusion + Multi-Resolution | 1.0157 | 0.5892 | 0.6119 |
| | Without Transformer Encoder | 0.9014 | 0.5070 | 0.6835 |
| | **Baseline (Full Model)** | **2.8542** | **2.0171** | **0.4920** |
| | Without Adaptive Fusion | 3.1090 | 2.1890 | 0.4920 |
| Darıören | Without Multi-Resolution Features | 3.2480 | 2.2660 | 0.4100 |
| | Without Adaptive Fusion + Multi-Resolution | 3.4170 | 2.3910 | 0.3710 |
| | Without Transformer Encoder | 2.9470 | 2.0730 | 0.4710 |
| | **Baseline (Full Model)** | **0.2935** | **0.1566** | **0.6589** |
| | Without Adaptive Fusion | 0.3190 | 0.1720 | 0.6230 |
| Akhasan | Without Multi-Resolution Features | 0.3370 | 0.1830 | 0.6010 |
| | Without Adaptive Fusion + Multi-Resolution | 0.3550 | 0.1950 | 0.5760 |
| | Without Transformer Encoder | 0.3060 | 0.1630 | 0.6400 |

ARTICLE IN PRESS

### 3.2. Baseline and State-of-the-Art Comparisons

To comprehensively evaluate the effectiveness of the proposed MR-ACF-TE-LSTM model, we compared its performance against several recently introduced baseline and state-of-the-art architectures in the field of streamflow prediction. The architectural specifications of the basic models are listed in Table 4. These models, representing contemporary advancements in literature, serve as strong benchmarks for assessing the predictive accuracy, temporal modeling capabilities, and robustness of the proposed approach. These include: (i) a conventional Transformer model with standard positional encoding, (ii) a hybrid Transformer Encoder–LSTM model designed to capture both long-term attention and short-term sequential dynamics, (iii) a Fusion LSTM-GRU[48] model representing a robust state-of-the-art approach known for

ARTICLE IN PRESS

balancing temporal learning capacity and model complexity, and (iv) a Bayesian Convolutional Neural Network (BCNN)[49], which introduces uncertainty estimation into the forecasting process. This comparative analysis aims to highlight the unique strengths of our model in capturing multi-resolution temporal features and long-range dependencies, while also assessing its generalization ability and robustness against established methods. The baseline architecture and key hyperparameter settings were selected based on commonly adopted configurations in prior streamflow forecasting studies to ensure realistic and reproducible comparisons; the final specifications used in our experiments are summarized in Table 4. In addition, we verified these literature-informed settings using a validation split by checking a small set of plausible alternatives (e.g., learning rate, hidden/embedding size, dropout, and—where applicable—layers/heads) and retained the best validation-performing configuration.

**Table 4.** Architectural Specifications of the Baseline Models

| **Layer Category** | **Layer Type** | **Transformer** | **Transformer-LSTM** |
|---|---|---|---|
| Input Processing | Initial Feature Embedding | Dense layer (64 nodes) | Dense layer (64 nodes) |
| Core Transformer Block | Multi-Head Attention | Number of heads 2, Key dimension 32. | Number of heads 2, Key dimension 32. |
| | Residual Connection & Norm | Concatenate and layer normalization | Concatenate and layer normalization |
| | Feed-Forward Network | Dense layer with [128,64] nodes and ReLU activation function | - |
| Sequential Processing | LSTM Layer | - | LSTM with 64 unites |
| Regularization & Pooling | Dropout | Rate = 0.2 | Rate = 0.2 |
| | Pooling | Global Average pooling | Global Average pooling |
| Output Generation | Output Layer | Dence layer | Dence layer |

ARTICLE IN PRESS

Table 5 summarizes the comparative performance results of the proposed MR-ACF-TE-LSTM model against the selected baseline and state-of-the-art architectures across three benchmark datasets: Dereevi, Darıören, and Akhasan. The evaluation metrics—RMSE, MAE, and $R^2$—demonstrate that MR-ACF-TE-LSTM consistently outperforms all other models, achieving the lowest prediction errors and highest explanatory power across datasets. Notably, while models such as Fusion LSTM-GRU and Bayesian CNN offer reasonable performance, they fall short in capturing the full spectrum of temporal dynamics and resolution-specific patterns. The results validate the effectiveness of combining multi-resolution processing, adaptive channel fusion, and Transformer-enhanced temporal modeling, thereby establishing MR-ACF-TE-LSTM as a robust and accurate framework for univariate streamflow forecasting.

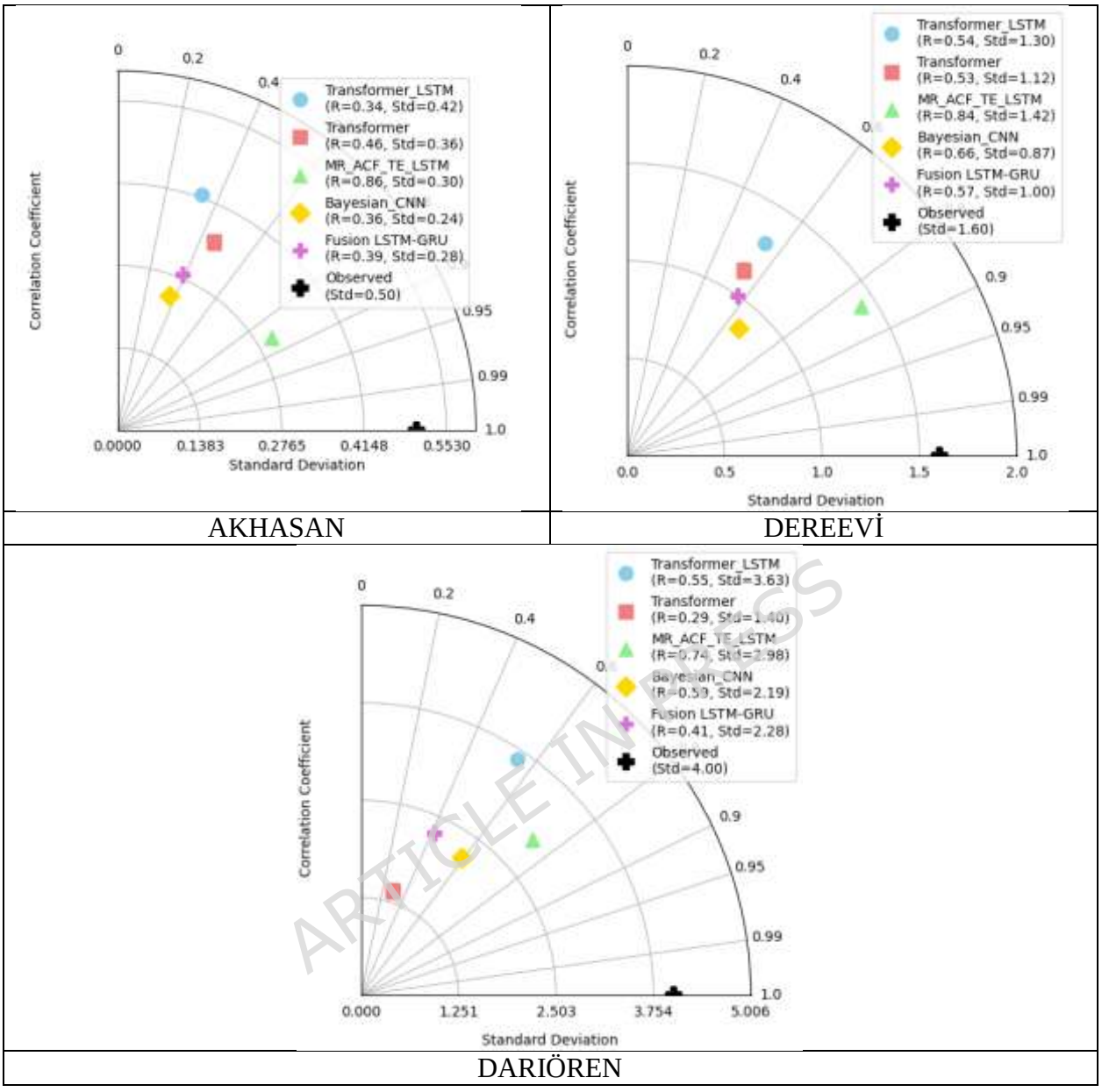
ARTICLE IN PRESS

**Table 5.** Comparative performance of baseline models across benchmark streamflow datasets.

| Dataset | Model | RMSE | MAE | $R^2$ | KGE | NSE |
|---|---|---|---|---|---|---|
| Dereevi | Transformer | 1.436 | 0.919 | 0.281 | 0.420 | 0.193 |
| | Transformer LSTM | 1.435 | 0.935 | 0.296 | 0.379 | 0.194 |
| | Fusion LSTM-GRU | 1.317 | 0.850 | 0.324 | 0.411 | 0.32 |
| | Bayesian CNN | 1.356 | 0.795 | 0.434 | 0.365 | 0.281 |
| | **MR-ACF-TE-LSTM** | **0.874** | **0.494** | **0.700** | **0.808** | **0.705** |
| Darıören | Transformer | 3.841 | 2.668 | 0.079 | 0.420 | 0.193 |
| | Transformer LSTM | 3.991 | 2.932 | 0.006 | 0.379 | 0.194 |
| | Fusion LSTM-GRU | 3.706 | 2.452 | 0.143 | 0.411 | 0.321 |
| | Bayesian CNN | 3.236 | 2.110 | 0.346 | 0.365 | 0.281 |
| | **MR-ACF-TE-LSTM** | **2.854** | **2.017** | **0.492** | **0.808** | **0.705** |
| Akhasan | Transformer | 0.468 | 0.271 | 0.133 | 0.341 | 0.142 |
| | Transformer LSTM | 0.569 | 0.394 | 0.281 | 0.075 | 0.281 |
| | Fusion LSTM-GRU | 0.472 | 0.284 | 0.115 | 0.229 | 0.117 |
| | Bayesian CNN | 0.472 | 0.282 | 0.116 | 0.184 | 0.116 |
| | **MR-ACF-TE-LSTM** | **0.293** | **0.156** | **0.658** | **0.645** | **0.663** |

For better visualization of model performance, Taylor diagrams were employed to simultaneously depict correlation, standard deviation, and centered RMSE in a unified polar plot, as presented in Fig. 4. This approach allows for a compact yet comprehensive comparison of predictive accuracy. For instance, on the Akhasan dataset, MR-ACF-TE-LSTM achieved a high correlation coefficient of 0.86 and a standard deviation of 0.30, closely matching the observed standard deviation of 0.50, while the Transformer model reached only 0.46 correlation. Similarly, in the Dereevi dataset, MR-ACF-TE-LSTM attained R = 0.84 and Std = 1.42, performing better than the Bayesian CNN (R = 0.66, Std = 0.87). These visual and numerical results emphasize the superior alignment of MR-ACF-TE-LSTM predictions with observed data across various temporal and statistical characteristics.

ARTICLE IN PRESS




AKHASAN DEREEVİ

DARIÖREN

**Fig. 4.** Taylor diagram of proposed approach compared to baseline and state of art methods.

Fig. 5 presents boxplots of residuals (Observed – Predicted) for five models on the benchmark datasets. Across all cases, MR-ACF-TE-LSTM exhibits the most compact and centered residual distribution, with interquartile ranges approximately between -0.5 and 0.5 for Akhasan, -1.0 and 1.0 for Dereevi, and -2.0 and 2.0 for Darıören. Its residuals also show fewer extreme outliers, with maxima generally under ±6, while Transformer and Bayesian CNN models display broader spreads with residuals extending to ±10 in Dereevi and even beyond ±25 in Darıören. Transformer LSTM and Fusion LSTM-GRU demonstrate skewed distributions, often

ARTICLE IN PRESS

underestimating flow values. These patterns highlight the superior accuracy, consistency, and generalization ability of the MR-ACF-TE-LSTM model across varied hydrological conditions.

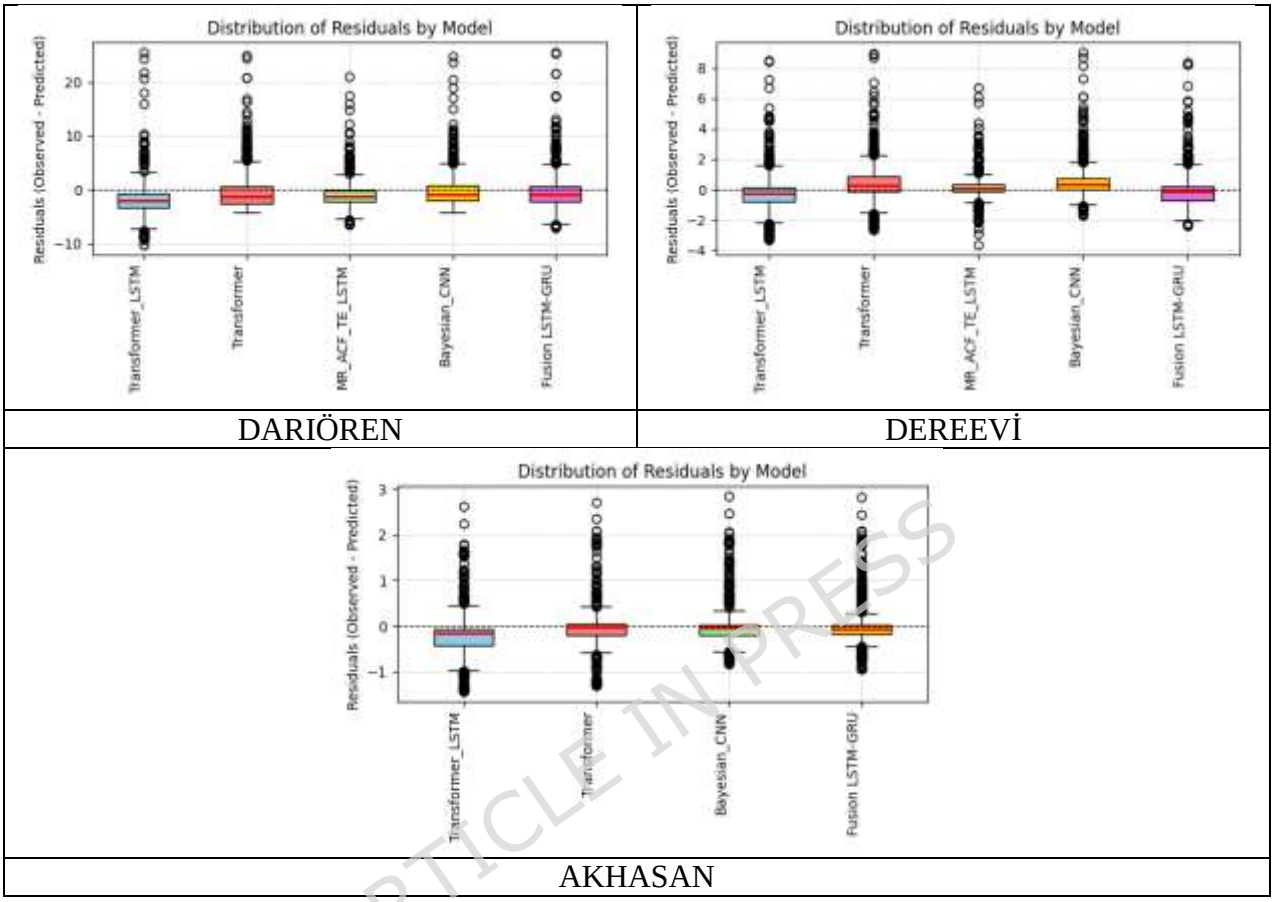



**Fig.5.** Boxplot of residuals across datasets based on Baseline methods.

### 3.3. Generalization and Robustness Analysis

To assess the generalization capability and robustness of the proposed MR-ACF-TE-LSTM model across heterogeneous hydrological datasets, we conducted cross-dataset evaluations where the model was trained on one dataset and tested on the remaining two. The results are summarized in Table 6. When trained on the Akhasan dataset, the MR-ACF-TE-LSTM achieved an RMSE/MAE of 2.50/1.50 on Dereevi and 5.80/3.01 on Darıören, outperforming both Transformer and Transformer-LSTM in generalization to the Dereevi dataset but slightly underperforming compared to Transformer-LSTM on Darıören. When trained on Dereevi, the model maintained strong performance on Akhasan (1.01/0.93) and competitive results on Darıören (5.57/2.94). Finally, training on Darıören showed limited generalizability, particularly

ARTICLE IN PRESS

towards Akhasan and Dereevi, yet MR-ACF-TE-LSTM still outperformed baseline Transformer in most settings. These findings demonstrate that the proposed architecture is more robust to domain shifts than baseline models, particularly when trained on mid-complexity datasets (e.g., Dereevi).

**Table 6.** Cross-Dataset Evaluation (Train → Test) Performance (RMSE / MAE).

| Train \ Test | Model | Akhasan | | Dereevi | | Darıören | |
|---|---|---|---|---|---|---|---|
| | | RMSE | MAE | RMSE | MAE | RMSE | MAE |
| Akhasan | Transformer | - | - | 2.99 | 1.92 | 6.37 | 3.82 |
| | Transformer - LSTM | - | - | 2.28 | 1.242 | 5.85 | 3.13 |
| | MR-ACF-TE-LSTM | - | - | 2.50 | 1.50 | 5.80 | 3.01 |
| Dereevi | Transformer | 0.72 | 0.58 | - | - | 6.08 | 3.75 |
| | Transformer - LSTM | 1.39 | 1.26 | - | - | 5.35 | 2.6 |
| | MR-ACF-TE-LSTM | 1.01 | 0.93 | - | - | 5.57 | 2.94 |
| Darıören | Transformer | 1.27 | 1.08 | 3.09 | 2.79 | - | - |
| | Transformer - LSTM | 3.06 | 2.9 | 2.45 | 2.15 | - | - |
| | MR-ACF-TE-LSTM | 2.90 | 2.83 | 2.15 | 1.78 | - | - |

Fig. 6 illustrates the temporal attention weights learned by the proposed model. These weights are shown for three representative test samples across six distinct cross-dataset evaluation scenarios. The attention weight patterns shown in Fig. 6 exhibit clear physical interpretability when examined from a hydrological perspective. Higher attention values assigned to recent time steps during high-flow periods reflect the dominance of fast rainfall–runoff responses, whereas the broader temporal attention span observed during low-flow conditions is consistent with groundwater-dominated baseflow and catchment storage processes. Similarly, the varying temporal focus across seasons aligns with known hydrological memory effects, where antecedent moisture conditions exert stronger influence during wet seasons compared to dry periods. These results indicate that the attention mechanism is not merely a technical artifact but rather captures physically meaningful hydrological dependencies embedded within the streamflow generation process. These visualizations highlight how the model dynamically allocates attention over the 96-step input window when generalizing from one hydrological catchment to another.

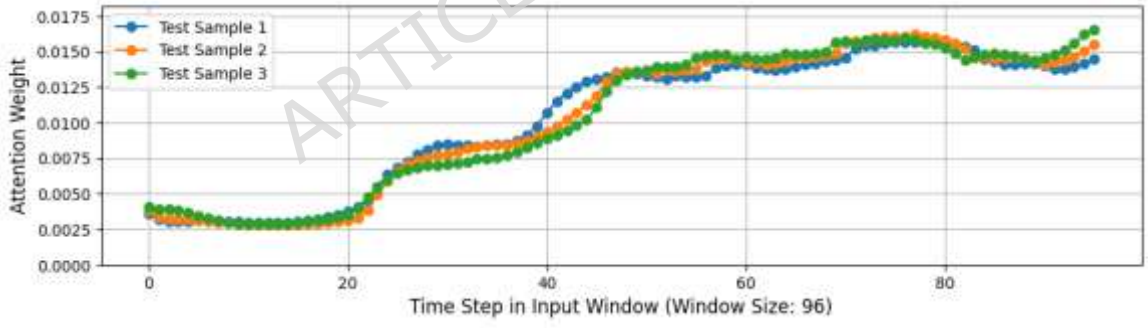
When trained on Akhasan, the model shows moderate to widely distributed attention (first two plots), especially for Sample 1, where focus spans both earlier (steps 10–25) and mid-to-late (steps 55–75) input periods—demonstrating the model's ability to incorporate information from

ARTICLE IN PRESS

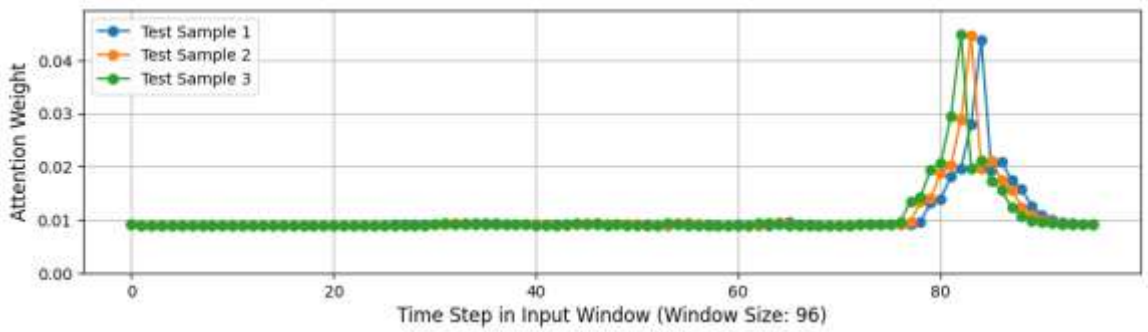
multiple temporal segments. In contrast, Samples 2 and 3 from the same plots concentrate attention more narrowly within the 60–85 range, emphasizing the importance of recent input patterns in prediction. In the scenarios where the model is trained on Darıören (third and fourth plots), attention weights become more peaked and localized—most prominently around steps 78–88—indicating the model's reliance on immediate historical data for forecasting. This concentrated focus reflects a learned dependency on recent streamflow conditions, especially visible in Samples 2 and 3. For Dereevi-trained models (last two plots), attention maps show smoother progression. While Sample 1 still presents multi-period dependencies, Samples 2 and 3 retain a strong focus on the latest segments (around time steps 75–90), aligning with the typical hydrological intuition that recent rainfall-runoff interactions play a dominant role in short-term streamflow forecasting.

Overall, these attention patterns confirm the temporal adaptability of the MR-ACF-TE-LSTM model across diverse data domains. The model can either distribute its attention across multiple influential regions or narrow its focus to immediate precursors. Depending on the spatiotemporal characteristics of the data, such flexibility enhances the model's interpretability and robustness in cross-catchment generalization tasks.

ARTICLE IN PRESS

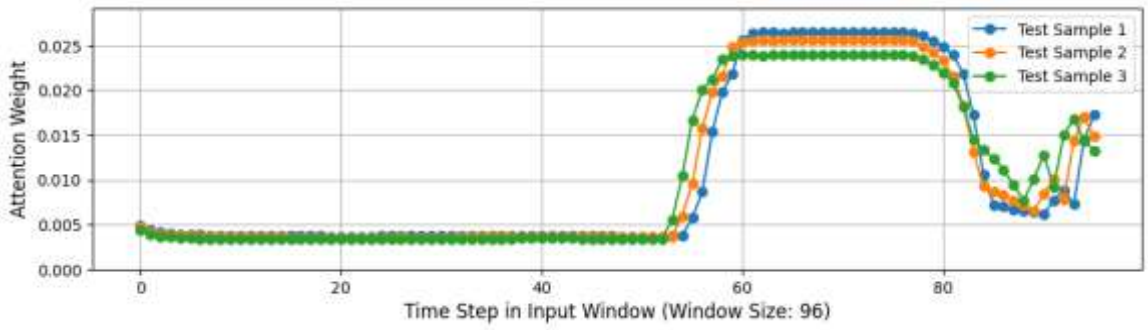



AKHASAN (Train) - DARIÖREN (Test)

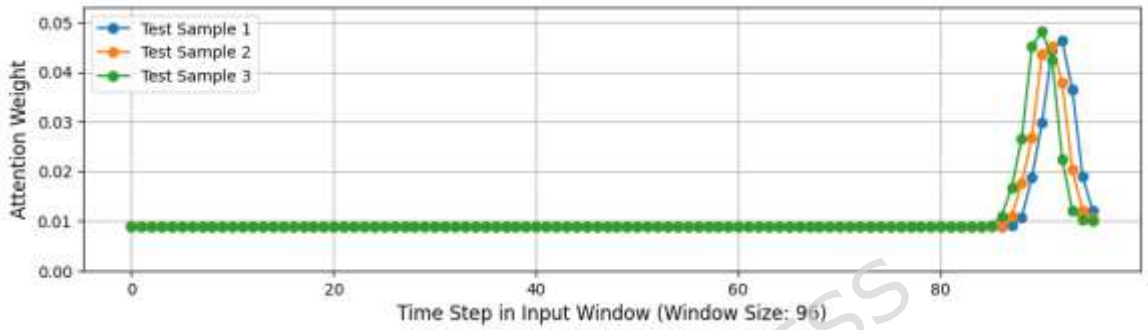



AKHASAN (Train) - DEREEVİ (Test)

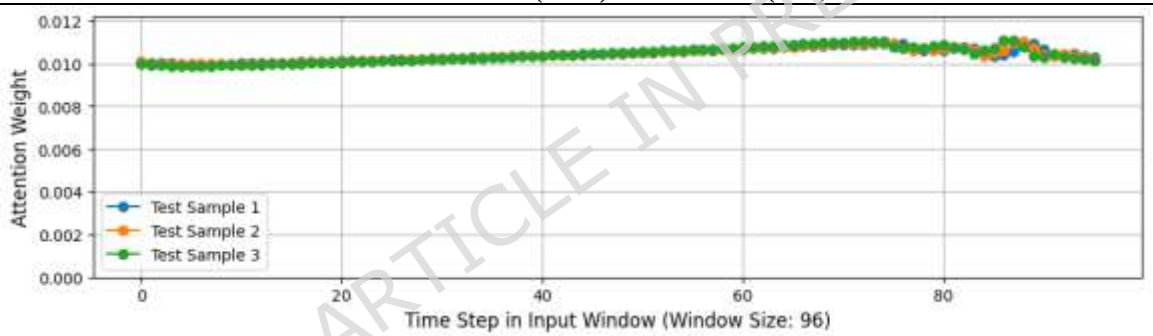
ARTICLE IN PRESS



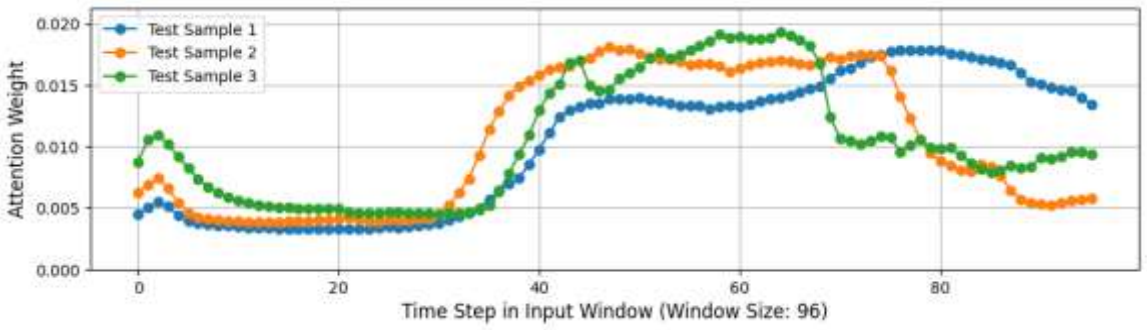

DARIÖREN (Train) - DEREEVİ (Test)




DARIÖREN (Train) - AKHASAN (Test)




DEREEVİ (Train) - AKHASAN (Test)




DEREEVİ (Train) - DARIÖREN (Test)

**Fig. 6.** Temporal Attention Weights across Time Steps for Test Sample

ARTICLE IN PRESS

## 4. Discussion

In this study, it has been suggested a novel hybrid model named MR-ACF-TE-LSTM which combines adaptive channel fusion and multi-resolution temporal analysis to improve interpretability and prediction accuracy. The model efficiently captures short- and long-term dependencies for accurate time-series forecasting by combining lagged data, statistical features, and seasonal indicators using an attention-driven Transformer encoder and fine-tuning them with an LSTM. The use of multi-resolution in the hybrid model is one of the study's most crucial components. Multi-resolution architecture has greatly improved the capacity to record long-term time series data by allowing models to analyze temporal phenomena at different scales and successfully decipher intricate patterns. This approach is particularly useful for applications where time series data—like daily flow data—are entangled with seasonal fluctuations and long-term trend data. Recent research has shown that hybrid frameworks that combine sophisticated neural network architectures, like Transformers and Temporal Convolutional Networks (TCNs), with multi-resolution techniques can improve predictive accuracy and robustness across a range of forecasting tasks[50]. By utilizing hierarchical representations of the data, multi-resolution strategies enable a more dynamic capture of long-term dependencies. To accurately model high-dimensional time series, for example, the Sparse Multi-scale Temporal Convolutional Network (SM-TCN) uses different kernel sizes to extract features at different temporal resolutions[51]. Furthermore, by breaking down the time series into different scales, multi-resolution techniques allow for better management of the temporal dynamics of streamflow data, enabling the capture of both long-term trends and short-term fluctuations. Deep learning can be used to create multi-resolution global hydrological model proxies, according to a study by Droppers et al.[52]. The study highlights that the performance of the multi-resolution surrogate was comparable to that of multiple single-resolution surrogate variants, suggesting that there were few trade-offs between the surrogate's performance and its wide spatial applicability. In this study, it was used a multi-resolution approach to more accurately forecast streamflow data with irregular structures. In ablation studies, the proposed MR-ACF-TE-LSTM model produced the lowest RMSE values among all FMSs. The RMSE was 0.3370 without multi-resolution, but it decreased to 0.2935 with multi-resolution, resulting in a 13% improvement particularly in Akhasan FMS. It was also obtained enhancements of 11% and 12% in the Darıören FMS and Dereevi FMSs, respectively. Badrzadeh et al.[53] demonstrated that the model becomes simultaneously sensitive to both short-

ARTICLE IN PRESS

term fluctuations and long-term trends when streamflow data are broken down into distinct frequency components using multi-resolution analysis. For instance, compared to a single-scale model, improvements in NSE and RMSE of roughly 10–15% and 12%, respectively, were reported when wavelet-based multi-resolution inputs were used. Consequently, it can be concluded that the findings of this investigation are consistent with those of other studies in the literature.

Fig. 2 shows that the stations have different peak flow characteristics. At Darıören FMS, there are very high peak values during seasonal changes, with a maximum of about 55 $m^3$/sec. This means that the flow conditions are very dynamic. Station Dereevi, on the other hand, has moderate peak flows of about 14 $m^3$/sec, and Akhasan FMS has relatively stable hydrological behavior, with peak values usually close to 4 $m^3$/sec. Further analysis concentrating on high-flow periods illustrates that the proposed model maintains robust predictive accuracy across diverse peak magnitudes, effectively encompassing both sudden extreme events and lower-amplitude peak flows. This confirms that the proposed framework is strong enough to work in a variety of hydrological settings.

Long-range dependencies in noisy and irregular data, like hydrological time series, are also modeled using the Transformer Encoder in addition to the Multi-Resolution Approach. Consequently, the hybrid model in this study was developed using this module, which enhanced the model's computational effectiveness and capacity to represent global temporal patterns[39]. When compared to traditional models such as LSTM, Zhang et al.[54] show how effective the Patch Transformer is by showcasing its superior predictive ability across a range of lead times. According to their research, conventional physics-based hydrological models frequently have limited prediction horizons and only offer accurate forecasts for brief periods of time. In the study of Liu et al.[21], it was proposed a transformer-based double-encoder-enabled model known as the double-encoder Transformer. Its unique feature is the "cross-attention" mechanism, which uses river flow observation to capture the relationship between two time series sequences. The study found that in the 21 years of predictions, the double-encoder Transformer's average $R^2$ was roughly 0.91 which is 0.1 higher than the other model, and its RMSE was 2579 $m^3$/month, much lower than the compared models. Similar to the literature, the use of Transformer encoder

ARTICLE IN PRESS

module in the current study was increased the prediction accuracy ($R^2$) and reduced the error rates (RMSE and MAE) (Table 3).

In the Western Black Sea Basin, catchment heterogeneity, which includes differences in drainage area, topography, and how water flows, is very important for how well cross-dataset generalization works. FMSs with greater variability and intricate runoff dynamics generally present more significant challenges for transferability, while stations with more uniform flow regimes facilitate more stable generalization. The proposed model shows that it can work well in a wide range of hydrological conditions, which shows that it is strong enough to handle basin- and station-specific features. Although the study was conducted within a single basin, the selected stations represent different hydroclimatic regimes and flow formation mechanisms, including differences in seasonal precipitation patterns, altitude-dependent runoff responses, and discharge accumulation scales. This intra-basin heterogeneity allows for a meaningful assessment of the model's robustness in the face of varying hydrological conditions, rather than simply through spatial interpolation. Therefore, in this study, the process was primarily analyzed through the generalization of the model across contrasting hydrometeorological regimes and flow dynamics within a basin.

Overall, hydrological data, especially stream flow data, have an irregular and complicated structure and show notable changes over a range of time scales. This is because hydrological data fluctuate greatly due to annual and seasonal variability. Even though the FMSs employed in this study had these characteristics as well, the novel and hybrid MR-ACF-TE-LSTM model produced notable improvements in every metric. When examining 10-year flow rates, the Darıören FMS, which has a higher standard deviation (3.42±5.24 $m^3/s$)—that is, more fluctuations and a more complex structure—had lower $R^2$ values and higher RMSE values and the suggested model produced superior results compared to other models. The Dereevi FMS, which has a more homogeneous pattern and a lower standard deviation (1.77±2.10 $m^3/s$), had the highest $R^2$ (0.700). These outcomes show how effective the suggested model is, even when dealing with intricate data structures. While our results demonstrate strong performance using discharge-only inputs, incorporating exogenous forcings such as precipitation and temperature is expected to improve physical interpretability and will be explored in future work when consistent forcing records are available.

ARTICLE IN PRESS

## 5. Conclusion

In this study, a hybrid model called MR-ACF-TE-LSTM was proposed to improve both prediction accuracy and interpretability in streamflow forecasting. The model integrates multi-resolution temporal analysis and adaptive channel fusion with a Transformer encoder and an LSTM-based decoder, enabling the effective modeling of short- and long-term dependencies for precise time-series forecasting. The Transformer encoder captures long-range temporal dependencies and global contextual patterns, while the temporal attention mechanism improves interpretability by highlighting the time steps that contribute most to the prediction. The LSTM-based decoder complements this structure by preserving local temporal continuity and ensuring sequential consistency in the forecasted outputs. The proposed MR-ACF-TE-LSTM model was evaluated using ten years of daily streamflow data from three flow measurement stations and consistently outperformed baseline and state-of-the-art models. Ablation analyses further demonstrated that each architectural component plays a critical role in the overall performance, confirming the effectiveness of the proposed hybrid design. Beyond predictive performance, the learned fusion weights and attention distributions provide valuable insights into the temporal dynamics of streamflow processes, enhancing the model's applicability in practical hydrological analysis. Despite its strengths, the multi-layered structure of the proposed architecture introduces increased computational and structural complexity, which may limit scalability and generalization when applied to short, noisy, or data-scarce time series. Future research will focus on developing lightweight and parameter-efficient model variants, validating the proposed model across diverse climatic and hydrological regimes, and incorporating domain adaptation to further enhance robustness and operational reliability in streamflow forecasting.

ARTICLE IN PRESS

## Data Availability

The data will be available upon reasonable request.

## References

ARTICLE IN PRESS

1. Granata, F. & Di Nunno, F. Pathways for Hydrological Resilience: Strategies for Adaptation in a Changing Climate. *Earth Syst. Environ.* https://doi.org/10.1007/s41748-024-00567-x (2025) doi:10.1007/s41748-024-00567-x.

2. Gao, S. *et al.* A hydrological process-based neural network model for hourly runoff forecasting. *Environ. Model. Softw.* **176**, 106029 (2024).

3. Khazaeiathar, M. & Schmalz, B. Addressing Volatility and Nonlinearity in Discharge Modeling: ARIMA-iGARCH for Short-Term Hydrological Time Series Simulation. *Hydrology* **12**, 197 (2025).

4. Wang, X. *et al.* MPPN: Multi-Resolution Periodic Pattern Network For Long-Term Time Series Forecasting. Preprint at https://doi.org/10.48550/arXiv.2306.06895 (2023).

5. Du, L. *et al.* MultiResFormer: Transformer with Adaptive Multi-Resolution Modeling for General Time Series Forecasting. Preprint at https://doi.org/10.48550/arXiv.2311.18780 (2024).

6. Cheng, F. & Liu, H. Charging strategies optimization for lithium-ion battery: Heterogeneous ensemble surrogate model-assisted advanced multi-objective optimization algorithm. *Energy Convers. Manag.* **342**, 120170 (2025).

7. Feng, H., Qian, W. & Laima, S. A machine learning-based probabilistic wind speed prediction model with multi-resolution data, quantile regression and bound estimation. *Eng. Struct.* **322**, 119098 (2025).

8. Wang, L., Li, Z., Ye, F. & Liu, T. A Probability Model for Short-Term Streamflow Prediction Based on Multi-Resolution Data. *Water Resour. Manag.* **37**, 5601–5618 (2023).

9. Zhang, Y., Ma, L., Pal, S., Zhang, Y. & Coates, M. Multi-resolution Time-Series Transformer for Long-term Forecasting. in *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics* 4222–4230 (PMLR, 2024).

10. Li, S. *et al.* A Novel Multi-Scale Time Fusion Transformer for Long-Range Spectrum Occupancy Prediction. *IEEE Trans. Veh. Technol.* **74**, 9299–9312 (2025).

ARTICLE IN PRESS

ARTICLE IN PRESS

11. Yu, C. *et al.* MGSFformer: A Multi-Granularity Spatiotemporal Fusion Transformer for air quality prediction. *Inf. Fusion* **113**, 102607 (2025).

12. Han, W. *et al.* MCformer: Multivariate Time Series Forecasting with Mixed-Channels Transformer. Preprint at https://doi.org/10.48550/arXiv.2403.09223 (2024).

13. Olorunnimbe, K. & Viktor, H. Ensemble of temporal Transformers for financial time series. *J. Intell. Inf. Syst.* **62**, 1087–1111 (2024).

14. Tao, J. *et al.* Coupling SWAT and Transformer Models for Enhanced Monthly Streamflow Prediction. *Sustainability* **16**, 8699 (2024).

15. Multi-step ahead forecasting of daily streamflow based on the transform-based deep learning model under different scenarios | Scientific Reports. https://www.nature.com/articles/s41598-025-89837-w.

16. Zhang, Y., Wu, R., Dascalu, S. M. & Harris, F. C. Sparse transformer with local and seasonal adaptation for multivariate time series forecasting. *Sci. Rep.* **14**, 15909 (2024).

17. Cheng, F., Liu, H. & Lv, X. MetaGNSDformer: Meta-learning enhanced gated non-stationary informer with frequency-aware attention for point-interval remaining useful life prediction of lithium-ion batteries. *Adv. Eng. Inform.* **69**, 103798 (2026).

18. Kang, H. & Kang, P. Transformer-based multivariate time series anomaly detection using inter-variable attention mechanism. *Knowl.-Based Syst.* **290**, 111507 (2024).

19. Zhou, R., Wang, Q., Jin, A., Shi, W. & Liu, S. Interpretable multi-step hybrid deep learning model for karst spring discharge prediction: Integrating temporal fusion transformers with ensemble empirical mode decomposition. *J. Hydrol.* **645**, 132235 (2024).

20. Lee, D.-G. & Ahn, K.-H. Improving medium-range streamflow forecasts over South Korea with a dual-encoder transformer model. *J. Environ. Manage.* **368**, 122114 (2024).

21. Liu, C., Liu, D. & Mu, L. Improved Transformer Model for Enhanced Monthly Streamflow Predictions of the Yangtze River. *IEEE Access* **10**, 58240–58253 (2022).

ARTICLE IN PRESS

ARTICLE IN PRESS

22. Yu, C. *et al.* GinAR+: A Robust End-to-End Framework for Multivariate Time Series Forecasting With Missing Values. *IEEE Trans. Knowl. Data Eng.* **37**, 4635–4648 (2025).

23. Liu, J., Xu, T., Lu, C., Yang, J. & Xie, Y. Variational mode decomposition coupled LSTM with encoder-decoder framework: an efficient method for daily streamflow forecasting. *Earth Sci. Inform.* **18**, 38 (2024).

24. Barzegar, V., Laflamme, S., Hu, C. & Dodson, J. Multi-Time Resolution Ensemble LSTMs for Enhanced Feature Extraction in High-Rate Time Series. *Sensors* **21**, 1954 (2021).

25. Xu, T., Han, G., Zhu, H., Taleb, T. & Peng, J. Multi-Resolution LSTM-Based Prediction Model for Remaining Useful Life of Aero-Engine. *IEEE Trans. Veh. Technol.* **73**, 1931–1941 (2024).

26. Aydin, H., Yenigun, K., Isinkaralar, O. & Isinkaralar, K. Hydrological low flow and overlapped trend analysis for drought assessment in Western Black Sea Basin. *Nat. Hazards* **121**, 1833–1863 (2025).

27. Çeribaşı, G. Batı Karadeniz Havzasının Yağış Verilerinin Yenilikçi Şen Yöntemi İle Analizi. *Acad. Platf. - J. Eng. Sci.* **6**, 168–173 (2018).

28. Isinkaralar, O., Isinkaralar, K., Sevik, H. & Küçük, Ö. Spatial modeling the climate change risk of river basins via climate classification: a scenario-based prediction approach for Türkiye. *Nat. Hazards* **120**, 511–528 (2024).

29. Nuri Balov, M. & Altunkaynak, A. Spatio-temporal evaluation of various global circulation models in terms of projection of different meteorological drought indices. *Environ. Earth Sci.* **79**, 126 (2020).

30. Kratzert, F. *et al.* Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrol. Earth Syst. Sci.* **23**, 5089–5110 (2019).

31. Lees, T. *et al.* Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. *Hydrol. Earth Syst. Sci.* **25**, 5517–5534 (2021).

ARTICLE IN PRESS

ARTICLE IN PRESS

32. Fang, K., Kifer, D., Lawson, K., Feng, D. & Shen, C. The Data Synergy Effects of Time-Series Deep Learning Models in Hydrology. *Water Resour. Res.* **58**, e2021WR029583 (2022).

33. Grismer, M. E. Estimating agricultural deep drainage lag times to groundwater: application to Antelope Valley, California, USA. *Hydrol. Process.* **27**, 378–393 (2013).

34. Mosavi, A., Ozturk, P. & Chau, K. Flood Prediction Using Machine Learning Models: Literature Review. *Water* **10**, 1536 (2018).

35. Papacharalampous, G., Tyralis, H., Koutsoyiannis, D. & Montanari, A. Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale. *Adv. Water Resour.* **136**, 103470 (2020).

36. Taylor, S. J. & Letham, B. Forecasting at Scale. *Am. Stat.* **72**, 37–45 (2018).

37. Addor, N. *et al.* Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges. *Hydrol. Sci. J.* **65**, 712–725 (2020).

38. Gauch, M. *et al.* Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. *Hydrol. Earth Syst. Sci.* **25**, 2045–2062 (2021).

39. Vaswani, A. *et al.* Attention is All you Need. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).

40. Tsai, Y.-H. H. *et al.* Multimodal Transformer for Unaligned Multimodal Language Sequences. *Proc. Conf. Assoc. Comput. Linguist. Meet.* **2019**, 6558–6569 (2019).

41. Gao, P. *et al.* Dynamic Fusion With Intra- and Inter-Modality Attention Flow for Visual Question Answering. in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 6632–6641 (IEEE, Long Beach, CA, USA, 2019). doi:10.1109/CVPR.2019.00680.

42. Bahdanau, D., Cho, K. & Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. Preprint at https://doi.org/10.48550/arXiv.1409.0473 (2016).

43. Lim, B., Arık, S. Ö., Loeff, N. & Pfister, T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **37**, 1748–1764 (2021).

ARTICLE IN PRESS

ARTICLE IN PRESS

44. Sheng, Z. *et al.* Residual Temporal Convolutional Network With Dual Attention Mechanism for Multilead-Time Interpretable Runoff Forecasting. *IEEE Trans. Neural Netw. Learn. Syst.* **36**, 8757–8771 (2025).

45. Qin, Y. *et al.* A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction. Preprint at https://doi.org/10.48550/arXiv.1704.02971 (2017).

46. Xu, K. *et al.* Show, Attend and Tell: Neural Image CaptionGeneration with Visual Attention.

47. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).

48. Kilinc, H. C., Apak, S., Ozkan, F., Ergin, M. E. & Yurtsever, A. Multimodal Fusion of Optimized GRU–LSTM with Self-Attention Layer for Hydrological Time Series Forecasting. *Water Resour. Manag.* **38**, 6045–6062 (2024).

49. Kilinc, H. C. *et al.* Enhancing hydrological time series forecasting with a hybrid Bayesian-ConvLSTM model optimized by particle swarm optimization. *Acta Geophys.* **73**, 3549–3566 (2025).

50. Wu, J. *et al.* A Hybrid Framework Integrating Past Decomposable Mixing and Inverted Transformer for GNSS-Based Landslide Displacement Prediction. *Sensors* **25**, 6041 (2025).

51. Guo, Z., Sun, Y. & Wu, T. SM-TCN: Multi-Resolution Sparse Convolution Network for Efficient High-Dimensional Time Series Forecast. *Sensors* **25**, 6013 (2025).

52. Droppers, B., Bierkens, M. F. P. & Wanders, N. A Multi-Resolution Deep-Learning Surrogate Framework for Global Hydrological Models. *Water Resour. Res.* **61**, e2024WR037736 (2025).

53. Badrzadeh, H., Sarukkalige, R. & Jayawardena, A. W. Impact of multi-resolution analysis of artificial intelligence models inputs on multi-step ahead river flow forecasting. *J. Hydrol.* **507**, 75–85 (2013).

ARTICLE IN PRESS

ARTICLE IN PRESS

54. Zhang, B., Xu, Q. & Ouyang, C. *Runoff Forecasting in Unmeasured Catchments and Rapid Flash Flood Prediction Based on Deep Learning.* https://meetingorganizer.copernicus.org/EGU25/EGU25-4387.html (2025) doi:10.5194/egusphere-egu25-4387.

**Author contributions**

H.C.K: Supervision, Conceptualization, Methodology, Software, Formal analysis, Writing – review & editing. S.A: Methodology, Conceptualization, Investigation, Software, Writing – original draft. A.Y: Data curation, Methodology, Investigation, Writing – original draft. H.H: Methodology, Writing – review & editing F.O: Data curation, Writing – review & editing. All authors give their full consent for the publication of this manuscript.

**Funding**

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

**Declarations**

**Competing interests**

The authors declare no competing interests.

**Additional information**

Correspondence and requests for materials should be addressed to H.C.K

ARTICLE IN PRESS