Journal of Hydroinformatics

Journal of Hydroinformatics Vol 26 No 12, 3266 doi: 10.2166/hydro.2024.263

Improved daily streamflow forecasting for semi-arid environments using hybrid machine learning and multi-scale analysis techniques

Salah Difi [a], Salim Heddam [b], Bilel Zerouali [c,d], Sungwon Kim [e], Yamina Elmeddahi [a], Nadjem Bailek [f,g,*], Celso Augusto Guimarães Santos [h] and Habib Abida [i]

[a] Department of Hydraulic, Civil Engineering and Architecture Faculty, Vegetal Chemistry-Water-Energy Laboratory (LCV2E), University of Hassiba Benbouali, Chlef, Algeria
[b] Faculty of Science, Agronomy Department, Hydraulics Division, Laboratory of Research in Biodiversity Interaction Ecosystem and Biotechnology, University 20 Août 1955, Route El Hadaik, BP 26, Skikda, Algeria
[c] Laboratory of Architecture, Cities and Environment, Department of Hydraulic, Civil Engineering and Architecture Faculty, University of Hassiba Benbouali, Chlef, Algeria
[d] Vegetal Chemistry-Water-Energy Laboratory, Faculty of Civil Engineering and Architecture, Department of Hydraulic, Hassiba Benbouali, University of Chlef, B.P. 78C, Ouled Fares, Chlef 02180, Algeria
[e] Department of Railroad Construction and Safety Engineering, Dongyang University, Yeongju, Republic of Korea
[f] Laboratory of Mathematics Modeling and Applications, Department of Mathematics and Computer Science, Faculty of Sciences and Technology, Ahmed Draia University of Adrar, Adrar 01000, Algeria
[g] Jadara University Research Center, Jadara University, P.O. Box 733, Irbid, P.C. 21110, Jordan
[h] Department of Civil and Environmental Engineering, Federal University of Paraíba, 58051-900 João Pessoa, Paraíba, Brazil
[i] Laboratory GEOMODELE, Faculty of Sciences, University of Sfax, 3000 Sfax, Tunisia
*Corresponding author. E-mail: bailek.nadjem@univ-adrar.edu.dz

SD, 0000-0003-2935-0277; SH, 0000-0002-8055-8463; BZ, 0000-0003-4735-9750; SK, 0000-0002-9371-8884; YE, 0000-0002-2561-1524; NB, 0000-0001-9051-8548; CAGS, 0000-0001-7927-9718

ABSTRACT

This study aimed to improve daily streamflow forecasting by combining machine learning (ML) models with signal decomposition techniques. Four ML models were hybridized with five families of maximum overlap discrete wavelet transforms (MODWTs). The hybrid models were applied to predict daily streamflow at the Bir Ouled Taher station in northern Algeria. Model performance was evaluated using multiple statistical metrics and compared to standalone ML models. The hybrid MODWT-Gaussian process regression (GPR) model using Symlet wavelets (MODWT-GPR3 sym4) achieved the best performance, with $R = 0.99$ and NSE = 0.98 during validation. This significantly outperformed the standalone models tested and other hybrid combinations. The MODWT-GPR3 sym4 model demonstrated a superior ability to capture nonlinearities and predict peak flows. Hybridization of ML models with wavelet transforms, particularly the MODWT-GPR approach, can substantially improve daily streamflow prediction accuracy compared to standalone models. However, model performance may vary between watersheds due to differences in hydrological characteristics. Consideration of catchment concentration time when selecting model inputs could further enhance forecasting capabilities.

Key words: hybrid models, hydrological modeling, signal decomposition techniques, streamflow prediction, watershed

HIGHLIGHTS

- Novel hybrid models combining MODWT and machine learning improve streamflow forecasting in semi-arid environments.
- Multi-scale analysis enhances the capture of complex streamflow patterns in semi-arid watersheds.
- Findings contribute to improved water management strategies under climate variability.

1. INTRODUCTION

Water management remains a significant challenge, particularly in arid and semi-arid regions where water resources are scarce and rainfall patterns are unpredictable (Zerouali *et al.* 2024b). In response to this challenge, numerous efforts have been made to enhance the accuracy and reliability of rainfall-runoff modeling. These efforts encompass a range of approaches, including conceptual models, which simplify complex processes into a set of mathematical equations (Sugawara 1979). Physically based models, such as that assessed by Santos *et al.* (2003), attempt to represent the physical processes

governing rainfall-runoff relationships in a more detailed manner. More recently, machine learning (ML) algorithms have emerged as powerful in this domain, as highlighted by do Nascimento *et al.* (2022).

ML algorithms are increasingly favored over conventional models because they excel at handling the complex, nonlinear relationships between rainfall and runoff data – relationships that are notoriously difficult to capture with traditional physical equations. This capability has led to their widespread adoption in practical engineering applications, as noted by Saraiva *et al.* (2021). However, the performance of ML models is highly dependent on the quality of the input data, necessitating the use of various preprocessing techniques to enhance data quality and model performance (El-kenawy *et al.* 2022; Gomaa *et al.* 2023; Zerouali *et al.* 2023a, b). Preprocessing techniques are essential for preparing data for ML algorithms, especially in complex domains such as hydrology and water resource management (Farajpanah *et al.* 2024). One advanced preprocessing method that has shown significant promise in enhancing the accuracy of rainfall-runoff models is the maximal overlap discrete wavelet transform (MODWT) (Küllahcı & Altunkaynak 2024). The MODWT is a powerful tool for decomposing input data into various frequency components, facilitating more nuanced analysis and superior feature extraction. This method improves upon the traditional discrete wavelet transform (DWT) by retaining time alignment and better handling nonstationary data, making it particularly suitable for hydrological applications where data characteristics can vary over time (Amini *et al.* 2024).

The integration of MODWT in rainfall-runoff modeling allows ML algorithms to leverage both temporal and frequency information, leading to more accurate and reliable predictions. Studies have demonstrated that the use of MODWT in preprocessing significantly enhances the model's ability to detect patterns and trends that may not be apparent in raw data, resulting in improved predictive performance (Daif & Hebal 2024). This technique's ability to capture multi-scale features of hydrological processes makes it a valuable addition to the suite of preprocessing tools available for hydrological modeling (Zerouali *et al.* 2023a, b). In addition to MODWT, other preprocessing techniques play crucial roles in preparing data for ML. Normalization and standardization are fundamental steps that ensure that each feature contributes equally to the model's performance. Typically, normalization rescales data to a specific range of 0 to 1, which is essential for algorithms sensitive to the scale of input features (Habib & Okayli 2024). Standardization adjusts the data to have a mean of zero and a standard deviation of one, which is beneficial for algorithms that assume normally distributed data. Principal component analysis (PCA) is another widely used technique that reduces the dimensionality of data. By transforming the original features into a set of uncorrelated principal components, PCA helps eliminate redundant information and focus on the most significant features (El-Rawy *et al.* 2024). This reduction in dimensionality not only simplifies the model but also enhances its efficiency and accuracy, particularly when dealing with high-dimensional datasets. Feature selection methods, such as recursive feature elimination (RFE) and mutual information, are employed to identify and retain the most relevant features (Zheng *et al.* 2024). RFE iteratively fits the model and removes the least important features, while mutual information measures the dependency between variables to select the most predictive features. By focusing on the most informative features, these methods improve model performance and reduce the risk of overfitting.

In scenarios where data are limited, data augmentation techniques such as synthetic data generation can be employed to increase the size and diversity of the training dataset. Techniques such as synthetic minority over-sampling technique (SMOTE) generate synthetic samples by interpolating between existing data points, which is particularly useful for addressing class imbalances in classification problems (Ni *et al.* 2024). Handling missing data is another critical preprocessing step. Techniques such as the mean imputation, k-nearest neighbors (KNN) imputation, and regression imputation are used to fill in gaps in the dataset, ensuring that the ML model has a complete and reliable set of inputs (Abnane *et al.* 2023). Mean imputation replaces missing values with the mean of the available data, KNN imputation uses values from the nearest neighbors, and regression imputation predicts missing values based on other features in the dataset (Li *et al.* 2024). Finally, noise reduction methods such as smoothing and filtering help to remove unwanted fluctuations and outliers from the data (Cloez *et al.* 2024). Smoothing techniques, such as moving averages, reduce short-term fluctuations and highlight longer-term trends, while filtering methods, including low-pass filters and wavelet-based denoising, remove high-frequency noise while preserving important signal characteristics (Dodig *et al.* 2024). These techniques lead to more stable and accurate models by ensuring that the input data are clean and reliable. By integrating these advanced preprocessing techniques, including MODWT, researchers and engineers can significantly enhance the performance of ML algorithms in rainfall-runoff modeling. The robust framework provided by these preprocessing methods, combined with powerful ML tools, offers a comprehensive approach to addressing the complexities of water management in challenging environments.

For example, Roushangar *et al.* (2017) presented different strategies to explore the spatiotemporal variation in the rainfall-runoff process for a watershed in northwest Iran using an extreme learning machine (ELM), and DWT preprocessed the temporal features. Quilty *et al.* (2019) proposed a stochastic data-driven ensemble forecasting framework for urban water demand in Montreal, Canada, using wavelet-based forecasts as input data. Alizadeh *et al.* (2021) simulated the precipitation and runoff data of the Shaharchay River basin, one of the most important basins of Lake Urmia in northwestern Iran, using a combined ELM, differential evolution, and DWT. Alizadeh *et al.* (2020) integrated a new learning machine with DWT to predict runoff-precipitation amounts in the same river basin. They tested several mother wavelets to identify the best family member. Roy *et al.* (2021) proposed an integrated model, combining an equilibrium optimizer with an ELM, and a deep neural network for one-day-ahead rainfall-runoff modeling. They tested the proposed models in two different catchments in the UK. They also tested six other well-known ML models. The proposed models were combined with the DWT preprocessing technique to improve their performance. Khan *et al.* (2021) compared the performances of single decision tree (SDT), tree boost (TB), decision tree forest (DTF), multi-layer perceptron (MLP), and gene expression programming (GEP) methods in rainfall-runoff modeling of a Pakistanian river basin. Additionally, they assessed the impact of wavelet preprocessing through MODWT on the model performance.

Furthermore, Alizadeh *et al.* (2018) presented an integrated artificial neural network (IANN) model that incorporates observed and predicted time series as input variables combined with wavelet transform to predict flow discharge at multiple lead times. Gomes & Blanco (2021) developed a hybrid MODWT-ANN model for daily rainfall estimation, considering the seasonality of rainfall data. The study demonstrated that this hybrid model performed well in forecasting daily rainfall using both satellite and national water agency data, indicating its potential utility in similar applications for rainfall estimation in other regions.

As noted by Freire *et al.* (2019), Freire & Santos (2020), and Abda *et al.* (2020), the selection of the mother wavelet may influence the results. Thus, this study aims to enhance the accuracy and reliability of rainfall-runoff modeling in the Oued Rouina Zeddine watershed by leveraging both ML techniques and signal decomposition methods. The focus is on evaluating the performance of standalone ML models, such as Gaussian process regression (GPR), long short-term memory (LSTM), general regression neural network (GRNN), and multi-layer perceptron neural network (MLPNN), in predicting daily streamflow at the Bir Ould Taher station. Additionally, the study explores hybrid models that integrate ML with different MODWT wavelet families to enhance prediction accuracy, aiming to identify the most effective configurations for capturing streamflow nonlinearities and improving water resource management in arid and semi-arid regions.

In arid and semi-arid regions, water management faces significant challenges due to streamflow variability and data scarcity, exacerbated by sporadic rainfall and prolonged dry periods (Freire *et al.* 2019; Abda *et al.* 2020; Freire & Santos 2020). Given that traditional rainfall-runoff models often struggle to capture the nonlinear and irregular hydrological patterns in such settings, this challenge manifests particularly in the Oued Rouina Zeddine watershed, where these issues are prevalent. To address these limitations, this study improves streamflow forecasts using varied ML and sophisticated signal decomposition approaches. In this analytical framework, our study uses GPR, LSTM, GRNN, and MLPNN models to address semi-arid hydrological complexities, in contrast to many humid studies. These models, selected for their ability to capture nonlinear relationships, are further enhanced by integrating MODWT to decompose input data into multiple frequency components, revealing underlying patterns that raw data may miss. Additionally, this study analyzes how mother wavelet families affect prediction accuracy, thereby improving streamflow forecasting.

2. MATERIALS AND METHODS

2.1. Study area and data used

This paper utilizes data from the National Agency of Hydraulic-Resources (ANRH). The hydrometric station of Bir Ouled Tahar (code 011905) was selected as a case study. This station is situated in the Oued Rouina Zeddine watershed. The Oued Rouina Zeddine watershed covers an area of 891.46 km² and is part of the northern section of the larger Cheliff basin (Supplementary Figure A1). It is located between longitudes 1°40′ and 2°10′ E and latitudes 35°50′ and 36°10′ N. Oued Rouina Zeddine is a minor tributary of the Oued Cheliff. This watershed is monitored by both a rain gauge station and a hydrometric station. The elevations in this watershed are moderate, rarely exceeding 1,700 m. Due to its geographical location, it experiences a temperate semi-arid climate, with an average annual temperature of approximately 16.6 °C. The average annual precipitation is 487 mm. Streamflow (Q) and precipitation (P) data are available at daily time scales (01

September 2000 to 31 August 2010). The data were divided into training (70%) and validation (30%) sets. Therefore, the training and validation subsets were 2,555 and 1,094, respectively, for the Bir Ouled Tahar station. In Supplementary Table A1, in terms of the statistical descriptions of (Q) and (P), the mean, maximum, minimum, standard deviation, and coefficient of variation were reported. According to the results of the statistical parameters in Supplementary Table A1, the table provides a comprehensive overview of the streamflow and precipitation data, highlighting the variability and correlation of these parameters across different subsets. The streamflow data show a maximum value of 25.84 m³/s in the training and all data subsets, with a lower maximum of 14.62 m³/s in the validation subset. The mean streamflow values are relatively low, indicating that high streamflow events are infrequent. The standard deviation of 1.22 m³/s suggests moderate variability in the streamflow data. The coefficient of variation (C_v) values indicate high variability in the data, with the highest C_v observed in the training subset.

For precipitation, the maximum value is 42.10 mm in both the validation and all the data subsets, with a lower maximum of 27.30 mm in the training subset. The mean precipitation values are low, similar to the streamflow data, indicating that high precipitation events are rare. The standard deviation values suggest greater variability in precipitation than in streamflow. The C_v values for precipitation also indicate high variability, with the highest C_v observed in the validation subset. The coefficient of correlation (R) between streamflow subsets is 1.00 for all streamflow subsets, indicating a perfect linear relationship. In contrast, the correlation values for precipitation are lower, approximately 0.30, indicating a weaker relationship between precipitation and streamflow.

2.2. Maximum overlap discrete wavelet transforms

The DWT was first introduced in the late 1980s by Daubechies (1988) and Mallat (1989). DWT is implemented using a discrete set of scales and wavelet translations obeying certain rules. This transform decomposes the signal into a set of mutually orthogonal wavelets. DWT analysis consists of performing a local comparison of a signal with wavelet patterns, such as a mathematical microscope, allowing zooming in on the signal at different scales. Wavelet orthonormal bases allow a multiresolution analysis based on very fast decomposition and reconstruction algorithms for a finite discrete signal (Daubechies 1992). These are functions produced by the process of dilation and translation of a mother wavelet function $\psi_{a,\tau}(t)$, which is given as follows:

$$\psi_{a,\tau}(t) = \psi\left(\frac{t-\tau}{a}\right) \tag{1}$$

One of the advantages of DWT is the flexibility in the selection of the mother wavelet, depending on the experimenter's use or the time series characteristics. The wavelet transform is generally written as follows:

$$C_x(a,\ \tau) = \int_{-\infty}^{+\infty} x(t)\overline{\psi_{a,\tau}(t)}\, d(t) \tag{2}$$

This research employs the MODWT, which was applied as a combined approach with the various ML methods mentioned above. The MODWT is a modified version of the DWT. The MODWT does not use the subsampling process during the filtering and decomposition stage, which provides more information about the resulting wavelet coefficients than does the DWT, which makes the MODWT more robust to boundary effects.

In its mathematical framework, the MODWT decomposes the time series X_t into an approximation component ($A_{j,\ t}$) using a low-pass filter ($\tilde{g}_{j,\ 1} = g_{j,1}/2^{j/2}$) and into a detail component ($D_{j,\ t}$) using a high-pass filter ($\tilde{h}_{j,\ 1} = h_{j,1}/2^{j/2}$), where $\tilde{g}_{j,\ 1}$ and $\tilde{h}_{j,\ 1}$ are the jth of the MODWT (Seo *et al.* 2017). Following Percival & Walden (2000), the MODWT is expressed through the following equations:

$$X = \sum_{j=1}^{L} D_j + A_{J0'} \tag{3}$$

$$D_{j,t} = \sum_{l=1}^{n-1} \tilde{h}_{j,l}^{0} W_{j,t+1 \bmod n'} \tag{4}$$

$$A_{j,t} = \sum_{l=1}^{n-1} \tilde{g}_{j,l}^{0} V_{j,t+1 \bmod n'} \tag{5}$$

Despite its advantages, critical analysis of the literature reveals that a key limitation of both DWT and MODWT lies in selecting the appropriate mother wavelet. Therefore, this research employed the most commonly used mother wavelets, including Haar, Debauchies, Symlet, Coiflets, and Fejer-Korovkin (Supplementary Figure A2). For comprehensive details regarding the mathematical implementation of MODWT, readers are directed to the significant contributions of Seo *et al.* (2017) and Barzegar *et al.* (2021).

2.3. Development methodology

To forecast daily streamflow (Q) at time (t), we selected specific time lags for both streamflow and precipitation (P) based on autocorrelation and cross-correlation analyses. The ACF, PACF, and XCF plots for the Bir Ouled Tahar station are presented in Figure 1. Based on the result, the optimal lags for streamflow were identified with the help of the ACF, whereas the PACF revealed significant precipitation lags.

By using the XCF, the lagged correlation in precipitation and streamflow was analyzed to show the effect of past precipitation representing future variation in streamflow. The procedure of this methodology enables us to select those combinations of lags that have the highest predictive power and are suitable for this watershed, where current and previous precipitation events affect streamflow.

This research focused on the daily streamflow forecast using precipitation and streamflow data only because precipitation records are, even when incomplete, consistently more available than others. We then selected, as illustrated in Figure 1, two streamflow lags, namely, $(t-1)$ and $(t-2)$, together with three precipitation lags, namely, (t), $(t-1)$, and $(t-2)$, as input variables, while streamflow at the time (t) was the output variable. Thus, four combinations of five components were considered in this study, as listed in Supplementary Table A2. Two modeling scenarios have been considered:

The first was standalone modeling, for which four different ML models had been applied: MLPNN, GPR, GRNN, and LSTM. Each model used the selected precipitation and streamflow lags independently without pretreatment. These models were chosen based on their application strengths regarding streamflow prediction:

- MLPNN models the nonlinear relationships quite realistically, which is so significant in rainfall-runoff processes.
- GPR offers probabilistic predictions and accounts for uncertainty, enhancing its usefulness in streamflow forecasting.
- GRNN also adapts well and fast to new data, making it well-suited for dynamic and variable hydrological conditions.
- LSTM is used to capture the long-term dependency, which is essential in time series data for accurate predictions based on historical rainfall.

The second scenario proposed a hybrid model to overcome the problem of nonstationarity in the streamflow data. All the ML models from scenario 1 were coupled with MODWT. For this hybrid model, the sub-series produced by decomposing the original time series was used as input for the ML models for further predictions. The MODWT decomposition stabilized such sub-series signals and thus enabled a more in-depth look into the periodicity and structure of the data. Some of the major points of the process are as follows:

1. This includes using PACF and XCF to decompose the selected precipitation and streamflow lags MRAs and residual components by MODWT.
2. The application of various mother wavelets, such as Haar, Daubechies (db3), Symlet (sym4), Coiflets (coif1), and Fejer-Korovkin (fk8) in analyzing streamflow and precipitation at t, $t-1$, and $t-2$ produced seven MRAs, namely MRA1 to MRA7, and one residual signal denoted as RSD. Each wavelet has some unique strengths: Haar detects jumps or discontinuities; Daubechies and Symlet provide a very good tradeoff between regularity and computational efficiency; Coiflets symmetrically preserve the trends of data; and Fejer-Korovkin introduces the minimum phase distortion.
3. Split the decomposed signals further into training and validation in order to optimize model learning.
4. The decomposed signals, specifically from db3, sym4, coif1, and fk8, were then used to train ML models for streamflow forecasting at time t.

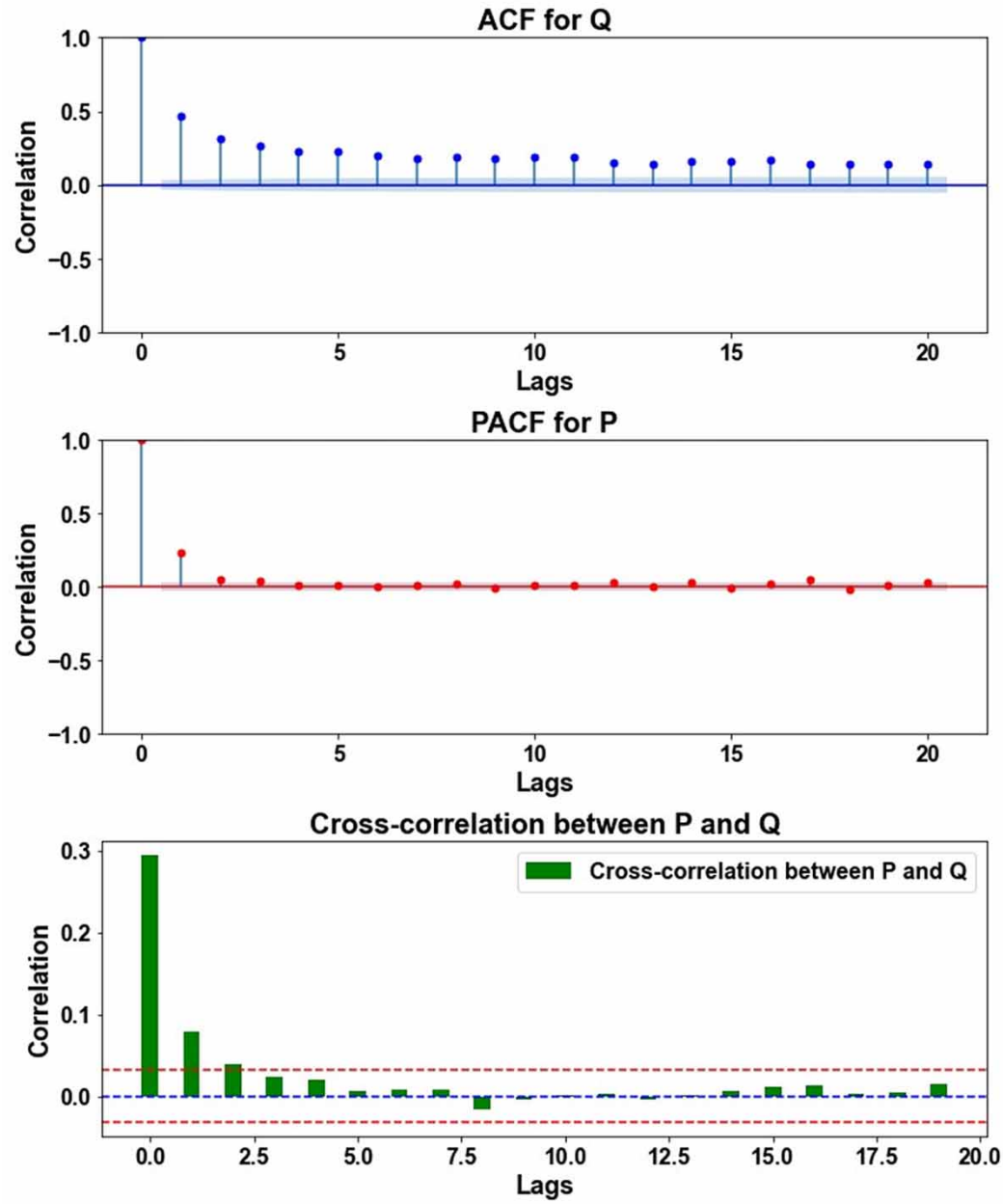

Figure 1 | ACF for streamflow (*Q*), PACF for precipitation (*P*), and cross-correlation between precipitation (*P*) and streamflow (*Q*).

This integration of the MODWT with ML improved the performance of the models by capturing the short-term fluctuations along with long-term patterns exhibited in streamflow variation. Supplementary Figure A2 describes the methodology in detail, together with a flowchart for daily streamflow prediction by the MODWT-ML algorithm.

2.4. Performance assessment of the models

This study evaluated model accuracy in predicting daily streamflow through metrics including root mean square error (RMSE), mean absolute error (MAE), correlation coefficient (*R*), and Nash Sutcliffe efficiency (NSE) (Ali *et al.* 2024; Belletreche *et al.* 2024; El-kenawy *et al.* 2024; Ferkous *et al.* 2024; Ibrahim *et al.* 2024; Oulimar *et al.* 2024; Zerouali *et al.*

2024a, b), as defined in Equations (6)–(9):

$$R = \left[\frac{\frac{1}{N}\sum (Q_{iO} - \overline{Q}_{iO})(Q_{iP} - \overline{Q}_{iP})}{\sqrt{\frac{1}{N}\sum_{i=1}^{n} (Q_{iO} - \overline{Q}_{iO})^2} \sqrt{\frac{1}{N}\sum_{i=1}^{n} (Q_{iP} - \overline{Q}_{iP})^2}} \right] \tag{6}$$

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{N} [Q_{iO} - Q_{iP}]^2}{\sum_{i=1}^{N} [Q_{iO} - \overline{Q}_{iO}]^2} \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (Q_{iO} - Q_{iP})^2} \tag{8}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |Q_{iO} - Q_{iP}| \tag{9}$$

where N represents the total number of data points, $\overline{Q}_{iO}$ is the average of the measured daily streamflow, $\overline{Q}_{iP}$ is the average of the forecasted daily streamflow, Q_{iP} is the forecasted daily streamflow, and Q_{iO} is the measured daily streamflow.

3. RESULTS

The performance metrics of the standalone ML models throughout the training and validation stages are shown in Table 1. Initial analysis reveals the GPR3 model outperformed the LSTM1, GRNN1, and MLPNN1 models during training, obtaining the lowest RMSE and MAE values of ≈0.032m³/s, ≈0.143 m³/s, and ≈0.993, respectively, as well as the greatest R and NSE values. The best-performing models utilized either the first or third input variable combinations from Supplementary Table A2, suggesting that incorporating both recent precipitation and streamflow data enhances model performance.

The validation phase results present a different outcome, as shown in Table 1, where the LSTM1 model achieved the highest level of accuracy, with $R \approx 0.805$, NSE ≈ 0.642, RMSE ≈ 0.729 m³/s, and MAE ≈ 0.225 m³/s, closely followed by the MLPNN1 model. The GPR1 model also performed well, achieving minimum values for the error metrics (RMSE and MAE) and higher values for R and NSE during the validation period. This performance distinction between phases is significant, as while GPR3 excelled during training, LSTM1 demonstrated superior generalization ability in validation, emphasizing the importance of assessing models across both phases. This discrepancy suggests that GPR3 may be overfitting the training data, while LSTM1 exhibits better generalization to unseen data.

The visualization in Figure 2(a) and 2(b) represents the performance metrics from Table 1 for the training and validation phases, respectively. The figures clearly illustrate the superior performance of the GPR3 model during training and the LSTM1 model during validation. These comparative results underscore the importance of selecting models based on validation results to ensure better generalization.

The integration of wavelet analysis with GPR models yielded compelling results. Table 2 presents the outcomes of hybridizing the GPR model with different MODWT wavelet families. The hybrid MODWT-GPR models demonstrate excellent predictive accuracy, with considerably reduced error measures (RMSE and MAE) and notably improved fit indices (R and NSE) compared to the standalone GPR model. This systematic improvement across all wavelet families suggests that MODWT preprocessing effectively captures underlying patterns in the streamflow data.

A detailed comparison of the best-performing hybrid MODWT-GPR model with the best standalone GPR model appears in Figure 3(a) and 3(b) for both phases. The training phase results show the MODWT-GPR3 (haar) model achieved the maximum improvement, reducing the GPR RMSE and MAE to approximately ≈0.004 and ≈0.001 m³/s, respectively. This significant improvement suggests that the Haar wavelet is particularly effective at capturing the underlying structure of the streamflow data. The quantitative gains, as visualized in Figure 3(a), indicate that MODWT preprocessing addresses some of the standalone GPR model's limitations identified earlier.

Analysis of the validation phase revealed the MODWT-GPR3 (sym4) model as superior, with an RMSE of ≈0.171 m³/s and an MAE of ≈0.117 m³/s, representing a significant improvement over the standalone GPR models. The differential

Table 1 | Results of streamflow prediction obtained by various standalone ML models

	Training				Validation			
Models	***R***	**NSE**	**RMSE**	**MAE**	***R***	**NSE**	**RMSE**	**MAE**
GPR1	0.990	0.979	0.175	0.043	**0.743**	**0.551**	**0.816**	**0.261**
GPR2	0.959	0.920	0.345	0.160	0.571	0.317	1.007	0.575
GPR3	**0.993**	**0.986**	**0.143**	**0.032**	0.551	0.293	1.025	0.487
GPR4	0.902	0.811	0.529	0.224	0.539	0.267	1.043	0.614
GPR5	0.922	0.849	0.473	0.126	0.363	0.080	1.266	0.861
GPR6	0.645	0.413	0.932	0.378	0.125	0.190	1.329	0.557
LSTM1	**0.829**	**0.676**	**0.692**	**0.190**	**0.805**	**0.642**	**0.729**	**0.225**
LSTM2	0.764	0.580	0.788	0.248	0.741	0.485	0.875	0.474
LSTM3	0.784	0.610	0.759	0.229	0.714	0.507	0.856	0.311
LSTM4	0.705	0.496	0.864	0.329	0.695	0.477	0.881	0.411
LSTM5	0.603	0.363	0.971	0.269	0.615	0.378	0.961	0.304
LSTM6	0.689	0.468	0.887	0.265	0.729	0.502	0.860	0.333
GRNN1	**0.884**	**0.773**	**0.579**	**0.220**	**0.685**	**0.460**	**0.896**	**0.378**
GRNN2	0.719	0.505	0.856	0.341	0.580	0.313	1.010	0.470
GRNN3	0.706	0.470	0.886	0.292	0.583	0.323	1.003	0.410
GRNN4	0.641	0.400	0.942	0.370	0.592	0.307	1.014	0.480
GRNN5	0.567	0.297	1.020	0.331	0.486	0.233	1.067	0.432
GRNN6	0.291	0.082	1.166	0.438	0.310	0.094	1.159	0.559
MLPNN1	**0.690**	**0.475**	**0.882**	**0.258**	**0.749**	**0.557**	**0.811**	**0.290**
MLPNN2	0.624	0.389	0.951	0.376	0.666	0.441	0.911	0.464
MLPNN3	0.611	0.371	0.965	0.269	0.640	0.409	0.937	0.282
MLPNN4	0.589	0.346	0.984	0.384	0.637	0.393	0.949	0.470
MLPNN5	0.559	0.312	1.009	0.278	0.616	0.378	0.960	0.301
MLPNN6	0.319	0.101	1.153	0.444	0.309	0.094	1.159	0.575

The bold values indicate the best performance metric achieved within each category of models (standalone or hybrid) during either the training or validation phase.

performance of wavelets – MODWT-GPR3 (haar) in training versus MODWT-GPR3 (sym4) in validation – reinforces the importance of prioritizing validation performance for model selection.

The investigation of LSTM hybridization presents additional insights. Table 3 demonstrates the results of combining LSTM models with various MODWT families. Consistent with previous observations, the hybrid MODWT-LSTM models exhibit improvements in terms of the numerical performance criteria *R*, NSE, RMSE, and MAE in both the training and validation phases compared to the standalone LSTM model.

Performance visualization in Figure 4 illustrates the comparative results of the MODWT-LSTM hybrid models against the standalone LSTM model. The analysis identifies the MODWT-LSTM3 (sym4) model's superiority in both phases with respect to RMSE and MAE. This consistency across both training and validation phases contrasts with the variable performance of the standalone models observed earlier. The stability of MODWT-LSTM3 (sym4) indicates effective resolution of the previously identified generalization issues.

The application of MODWT to GRNN models yielded significant results, as shown in Table 4. The experimental data demonstrates strong performance of the GRNN-MODWT hybrid models in terms of the numerical performance criteria *R*, NSE, RMSE, and MAE during training phases. Among the hybrid models, four of the five best performers exceeded the performance of the standalone model.

Comparative analysis in Figure 5 illustrates the performance metrics of the MODWT-GRNN hybrid models against the standalone GRNN model. The validation results show three of the five best hybrid models outperformed the GRNN3 model in terms of *R*, NSE, RMSE, and MAE. This widespread improvement across different wavelet families indicates

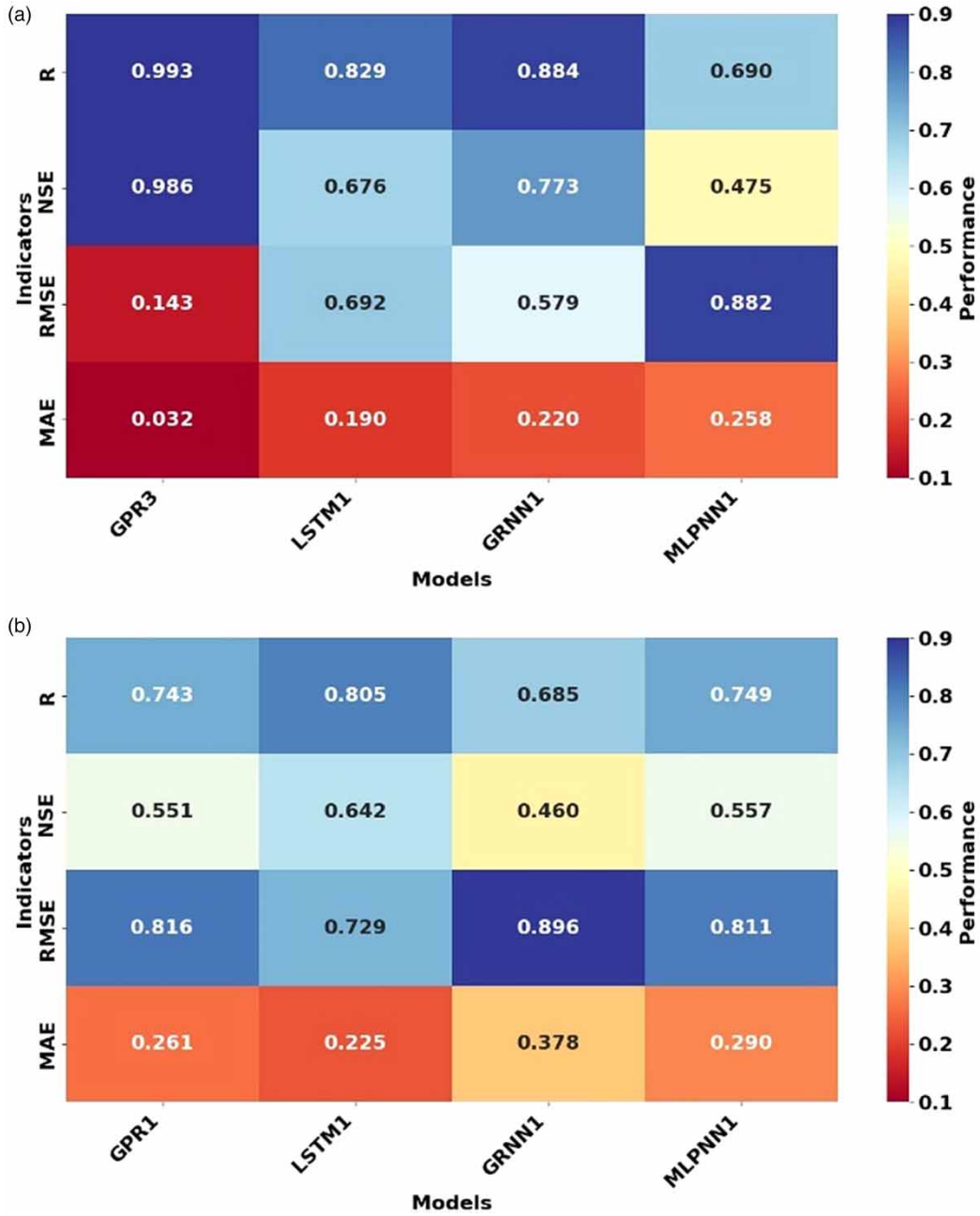

Figure 2 | The best performance criteria obtained by the standalone ML models. (a) Training and (b) validation.

that GRNN models benefit significantly from the multi-resolution analysis provided by MODWT. These results parallel the improvements observed in the MODWT-GPR and MODWT-LSTM models, suggesting a general trend of enhanced performance through MODWT preprocessing.

The MLPNN hybridization results reveal similar enhancements. Table 5 presents the outcomes of combining the MLPNN model with various MODWT algorithms. The data indicate particularly strong performance from the MODWT-MLPNN5 model (sym4), which excelled in both the training and validation phases, significantly outperforming the standalone MLPNN model.

The performance visualization in Figure 6 demonstrates the comparative metrics of the MODWT-MLPNN hybrid models against the standalone MLPNN model. The exceptional performance of MODWT-MLPNN5 (sym4) in both phases is clearly

Table 2 | Results of streamflow prediction obtained by hybrid MODWT-GPR models on a daily time scale for the Bir Ouled Tahar station

		Training				Validation			
Mother wavelet	**Models**	***R***	**NSE**	**RMSE**	**MAE**	***R***	**NSE**	**RMSE**	**MAE**
Coiflets wavelet (coif1)	MODWT-GPR1	0.999	0.999	0.005	0.002	**0.933**	**0.870**	**0.439**	**0.285**
	MODWT-GPR2	0.997	0.993	0.099	0.040	0.549	0.286	1.029	0.560
	MODWT-GPR3	**0.999**	**0.999**	**0.005**	**0.002**	0.926	0.854	0.465	0.326
	MODWT-GPR4	0.996	0.993	0.103	0.042	0.516	0.247	1.057	0.570
	MODWT-GPR5	0.999	0.999	0.004	0.002	0.883	0.767	0.588	0.404
	MODWT-GPR6	0.997	0.994	0.095	0.035	0.426	0.143	1.128	0.596
Daubechies wavelet (db3)	MODWT-GPR1	0.999	0.999	0.008	0.003	0.886	0.781	0.570	0.384
	MODWT-GPR2	0.997	0.994	0.093	0.035	0.518	0.226	1.072	0.590
	MODWT-GPR3	**0.999**	**0.999**	**0.006**	**0.003**	**0.898**	**0.803**	**0.540**	**0.334**
	MODWT-GPR4	0.997	0.994	0.094	0.034	0.530	0.242	1.061	0.602
	MODWT-GPR5	0.999	0.999	0.005	0.002	0.812	0.649	0.722	0.454
	MODWT-GPR6	0.997	0.994	0.097	0.038	0.471	0.140	1.130	0.671
Symlet wavelet (sym4)	MODWT-GPR1	0.999	0.999	0.009	0.004	0.990	0.980	0.174	0.118
	MODWT-GPR2	0.997	0.994	0.098	0.036	0.593	0.334	0.994	0.479
	MODWT-GPR3	**0.999**	**0.999**	**0.008**	**0.003**	**0.990**	**0.980**	**0.171**	**0.117**
	MODWT-GPR4	0.997	0.993	0.101	0.037	0.578	0.311	1.011	0.488
	MODWT-GPR5	0.999	0.999	0.007	0.003	0.988	0.976	0.187	0.118
	MODWT-GPR6	0.997	0.994	0.096	0.034	0.390	0.076	1.171	0.607
Haar wavelet (haar)	MODWT-GPR1	0.999	0.999	0.004	0.001	0.605	0.313	1.010	0.645
	MODWT-GPR2	0.999	0.999	0.022	0.006	0.173	−0.027	1.235	0.640
	MODWT-GPR3	**0.999**	**0.999**	**0.004**	**0.001**	**0.646**	**0.402**	**0.942**	**0.599**
	MODWT-GPR4	0.999	0.999	0.028	0.008	0.154	−0.053	1.250	0.644
	MODWT-GPR5	0.999	0.999	0.004	0.001	0.623	0.362	0.973	0.643
	MODWT-GPR6	0.999	0.999	0.040	0.010	0.140	−0.009	1.224	0.633
Fejer-Korovkin wavelet (fk8)	MODWT-GPR1	0.998	0.995	0.086	0.030	**0.891**	**0.793**	**0.554**	**0.331**
	MODWT-GPR2	0.996	0.991	0.114	0.039	0.556	0.297	1.021	0.485
	MODWT-GPR3	0.999	0.999	0.011	0.003	0.792	0.627	0.744	0.373
	MODWT-GPR4	0.995	0.991	0.117	0.037	0.535	0.274	1.038	0.497
	MODWT-GPR5	**0.999**	**0.999**	**0.005**	**0.002**	0.718	0.515	0.848	0.457
	MODWT-GPR6	0.996	0.993	0.105	0.037	0.329	0.019	1.207	0.596

The bold values indicate the best performance metric achieved within each category of models (standalone or hybrid) during either the training or validation phase.

evident. These findings align with our earlier observations about the sym4 wavelet's effectiveness, as demonstrated in the MODWT-LSTM3 (sym4) model.

A comprehensive evaluation appears in Supplementary Figure A3 through Taylor diagrams of the best standalone and hybrid models during the validation phase. These diagrams provide a concise visual summary of model observation matches in terms of correlation, root mean square difference, and variance ratios. The distribution pattern of hybrid models in the optimal regions of the Taylor diagrams reinforces the consistent improvement achieved through MODWT hybridization.

Further validation appears in Figure 7 through scatterplots comparing predicted and observed daily flow values for the Bir Ouled Tahar station during validation. The results confirm the positive effect of hybridization, as these models exhibit less dispersion and linear trends closer to the *yx* line. This visual representation aligns with the numerical improvements observed earlier and demonstrates enhanced prediction accuracy across all observation ranges.

The computational efficiency analysis in Supplementary Figure A8 presents processing times for various models with and without MODWT decompositions. Baseline results indicate that the GRNN and MLPNN models achieve calculation speeds of 8–11 s, making them suitable for rapid prediction applications. The hybrid implementations of MODWT-MLPNN5 (sym4) and MODWT-MLPNN5 (fk8) maintain efficiency with calculation times of 8 s, while also improving predictive capabilities. In contrast, the more complex GPR and LSTM networks require 48 and 46 s, respectively. Computational demands peak with MODWT-GPR3 (haar) at 99 s, whereas MODWT-GPR3 (sym4) achieves an optimal balance, providing superior predictions within 54 s of processing time.

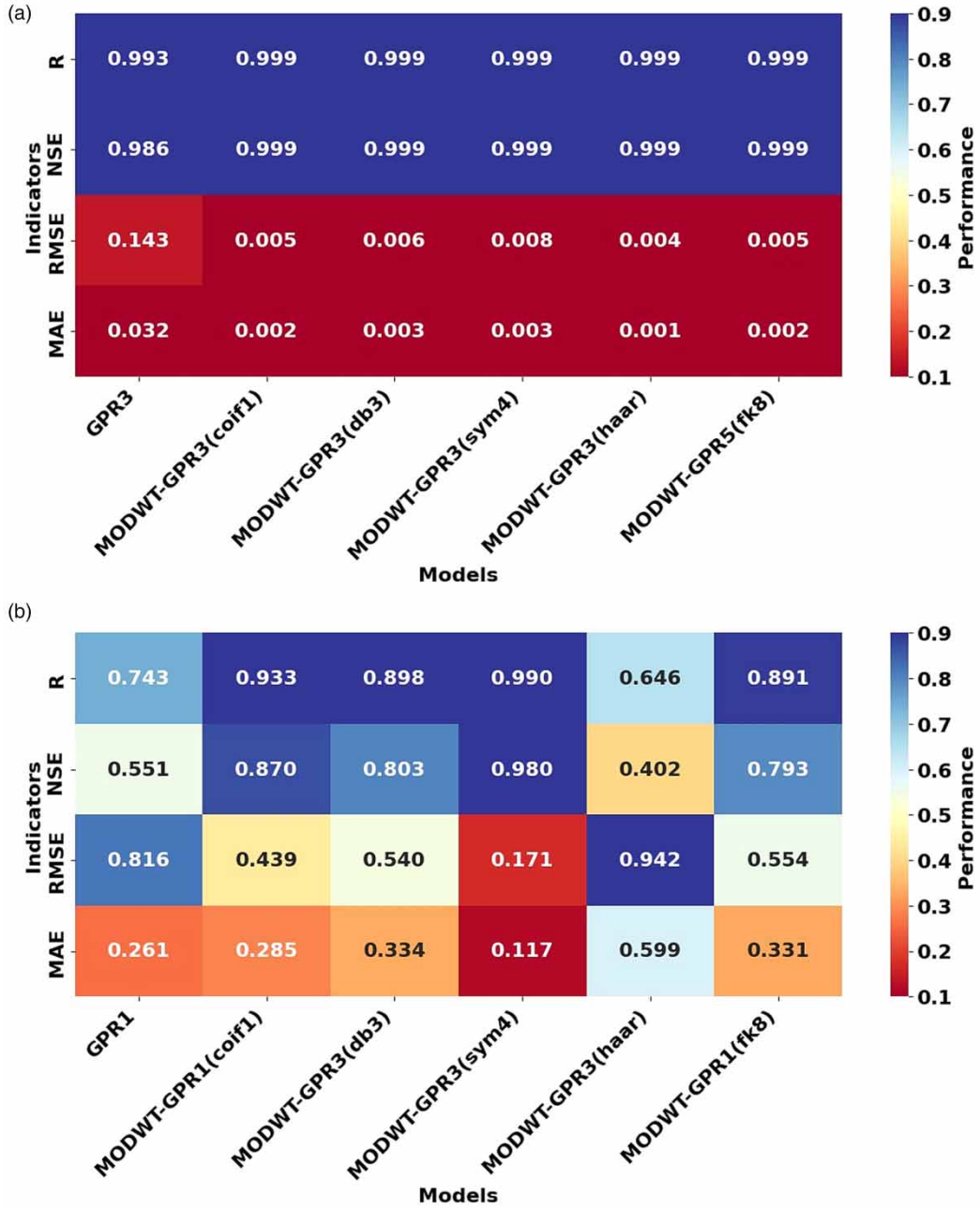

Figure 3 | The best performance criteria obtained by the hybrid MODWT-GPR models for daily streamflow prediction at the Bir Ouled Tahar station. (a) Training and (b) validation.

The temporal analysis in Figure 8 examines measured versus calculated streamflow variations during the validation period. Results demonstrate that the MODWT-GPR3 (sym4) model successfully captures significant nonlinearities and accurately predicts maximum values. This exceptional performance in capturing both trends and extreme events aligns with the superior metrics observed earlier and has substantial implications for flood prediction and water resource management.

Finally, to underscore the significance of these findings, Supplementary Table A3 provides a comprehensive comparison of data-driven and hybrid models applied in Algeria for streamflow forecasting. The proposed GPR-MODWT hybrid demonstrates remarkable accuracy, with an R value of 0.990 for daily forecasts and an RMSE of approximately 0.174 m³/s. These metrics surpass previous methods, including the neuro-fuzzy approach (RMSE ≈ 3.61 m³/s, $R \approx 0.90$) and the wavelet-support vector regression model (RMSE ≈ 0.15 m³/s, $R \approx 0.97$).

Table 3 | Results of streamflow prediction obtained by hybrid MODWT-LSTM models on a daily time scale for the Bir Ouled Tahar station

		Training				Validation			
Mother wavelet	**Models**	***R***	**NSE**	**RMSE**	**MAE**	***R***	**NSE**	**RMSE**	**MAE**
Coiflets wavelet (coif1)	MODWT-LSTM1	**0.950**	**0.898**	**0.388**	**0.188**	**0.896**	**0.790**	**0.559**	**0.365**
	MODWT-LSTM2	0.817	0.660	0.710	0.279	0.650	0.415	0.932	0.451
	MODWT-LSTM3	0.932	0.861	0.453	0.179	0.852	0.725	0.638	0.361
	MODWT-LSTM4	0.780	0.603	0.767	0.307	0.622	0.379	0.960	0.440
	MODWT-LSTM5	0.945	0.885	0.412	0.170	0.848	0.717	0.648	0.374
	MODWT-LSTM6	0.750	0.556	0.811	0.331	0.630	0.378	0.961	0.459
Daubechies wavelet (db3)	MODWT-LSTM1	0.885	0.772	0.581	0.231	0.825	0.676	0.694	0.418
	MODWT-LSTM2	0.737	0.537	0.828	0.327	0.546	0.290	1.027	0.535
	MODWT-LSTM3	0.910	0.821	0.515	0.209	0.837	0.699	0.669	0.392
	MODWT-LSTM4	0.758	0.570	0.798	0.317	0.363	0.082	1.267	0.642
	MODWT-LSTM5	**0.926**	**0.849**	**0.473**	**0.182**	**0.868**	**0.734**	**0.628**	**0.364**
	MODWT-LSTM6	0.782	0.607	0.762	0.303	0.600	0.348	0.984	0.461
Symlet wavelet (sym4)	MODWT-LSTM1	0.949	0.892	0.399	0.137	0.951	0.901	0.383	0.220
	MODWT-LSTM2	0.837	0.696	0.671	0.284	0.658	0.367	0.969	0.497
	MODWT-LSTM3	**0.959**	**0.914**	**0.356**	**0.113**	**0.960**	**0.919**	**0.347**	**0.181**
	MODWT-LSTM4	0.822	0.672	0.697	0.290	0.758	0.575	0.795	0.389
	MODWT-LSTM5	0.946	0.883	0.417	0.111	0.966	0.929	0.325	0.182
	MODWT-LSTM6	0.748	0.554	0.812	0.325	0.688	0.473	0.885	0.443
Haar wavelet (haar)	MODWT-LSTM1	0.888	0.780	0.570	0.259	0.784	0.596	0.774	0.479
	MODWT-LSTM2	0.672	0.451	0.902	0.362	0.233	−0.154	1.309	0.747
	MODWT-LSTM3	0.898	0.800	0.544	0.264	**0.792**	**0.623**	**0.748**	**0.489**
	MODWT-LSTM4	0.589	0.346	0.984	0.385	0.363	0.120	1.143	0.581
	MODWT-LSTM5	**0.904**	**0.805**	**0.537**	**0.220**	0.724	0.412	0.934	0.567
	MODWT-LSTM6	0.633	0.399	0.943	0.355	0.247	0.037	1.196	0.575
Fejer-Korovkin wavelet (fk8)	MODWT-LSTM1	**0.858**	**0.731**	**0.631**	**0.266**	0.757	0.571	0.798	0.403
	MODWT-LSTM2	0.755	0.564	0.803	0.304	0.644	0.410	0.936	0.462
	MODWT-LSTM3	0.852	0.719	0.645	0.255	0.810	0.654	0.717	0.367
	MODWT-LSTM4	0.753	0.563	0.805	0.324	0.537	0.271	1.040	0.486
	MODWT-LSTM5	0.858	0.727	0.635	0.227	**0.861**	**0.737**	**0.625**	**0.316**
	MODWT-LSTM6	0.726	0.526	0.838	0.333	0.621	0.383	0.957	0.482

The bold values indicate the best performance metric achieved within each category of models (standalone or hybrid) during either the training or validation phase.

The model's adaptability is evident across diverse datasets. For example, while wavelet-ANN models applied to Algeria's semi-arid and humid regions show higher RMSE (=2.46 mm) but stronger correlation ($R \approx 0.994$), the current approach maintains consistent performance across varying conditions. The integration of MODWT with advanced ML techniques enhances hydrological forecast accuracy, demonstrating broad applicability across Algerian watersheds.

4. DISCUSSION

The results presented in this study underscore the effectiveness of hybrid models, particularly the MODWT-GPR algorithm (sym4), in predicting daily streamflows at the Bir Ouled Taher station. The hybrid model's superior performance is evident when compared to the standalone models, as it consistently yields lower error metrics and higher fit indices during both the learning and validation phases. This performance aligns with previous research that highlights the advantages of hybrid models in handling complex hydrological data. The MODWT-GPR (sym4) model outperformed several advanced ML models reported in the literature. For instance, Gomaa *et al.* (2023) introduced a hybrid EMD-MLP-PSO model, achieving an R value of 0.982 and an NSE of 0.961. However, the MODWT-GPR (sym4) model in this study achieved even higher R and NSE values of 0.99 and 0.98, respectively. This suggests that the wavelet transform, when combined with GPR, can enhance model performance beyond what is possible with empirical mode decomposition (EMD) and other optimization techniques like PSO.

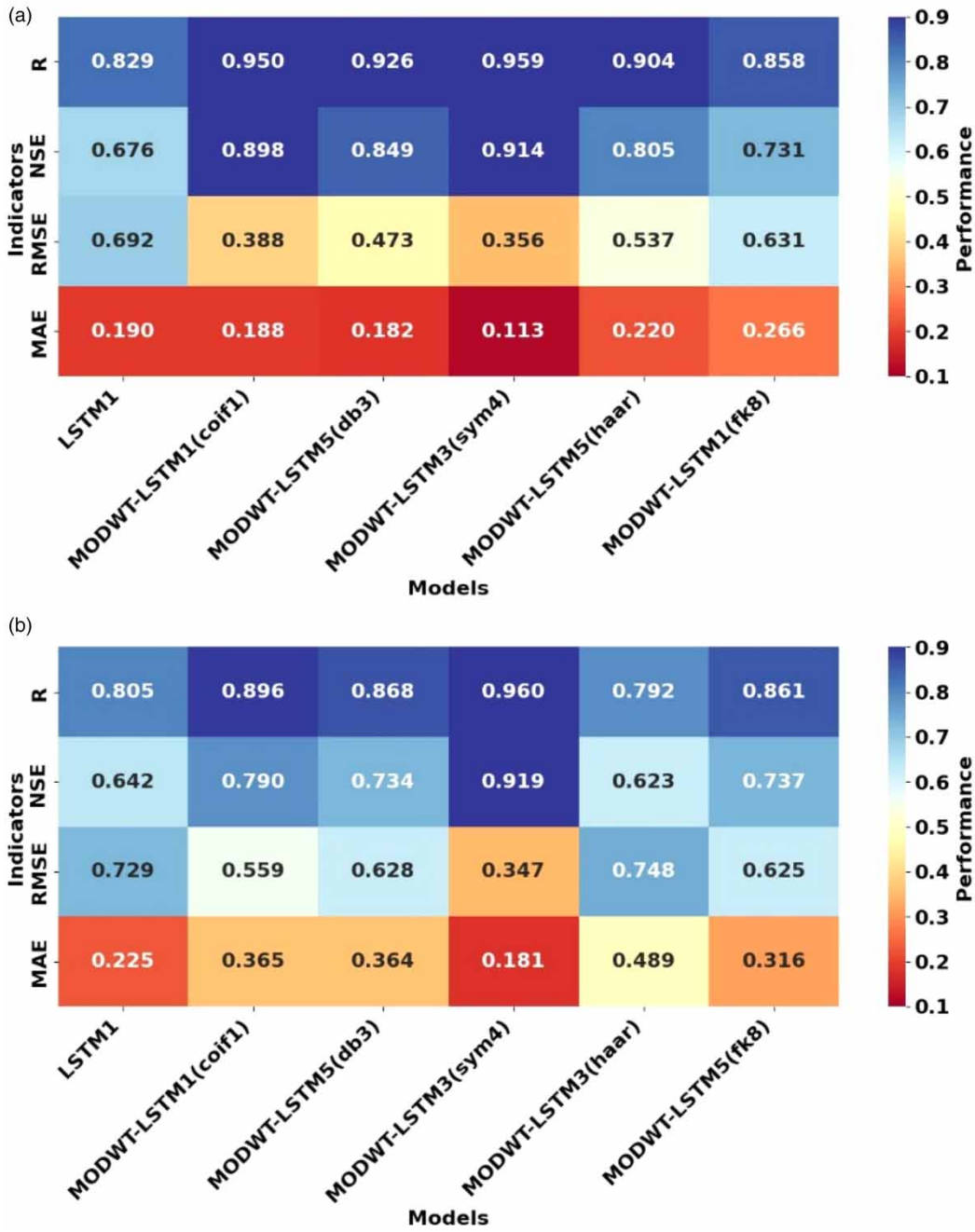

Figure 4 | Thematic maps showing the best performance criteria obtained by the hybrid MODWT-LSTM models for daily streamflow prediction at the Bir Ouled Tahar station. (a) Training and (b) validation.

One possible reason for the superior performance of the MODWT-GPR (sym4) model in the validation phase is its ability to capture multi-scale hydrological patterns. The sym4 wavelet, in particular, excels at approximating both the low- and high-frequency components of the time series, providing a better balance in capturing both short-term fluctuations and long-term trends in streamflow data. The sym4 wavelet's capacity to separate high- and low-frequency components allows the GPR model to perform better by effectively predicting streamflow under various hydrological conditions.

Similarly, Chakraborty & Biswas (2023) developed wavelet-based models, showing that hybridization with wavelet transforms significantly improved predictive accuracy. Their models achieved high NSE values, such as 0.9985 at the Teesta

Table 4 | Results of streamflow prediction obtained by hybrid MODWT-GRNN models on a daily time scale for the Bir Ouled Tahar station

		Training				Validation			
Mother wavelet	**Models**	***R***	**NSE**	**RMSE**	**MAE**	***R***	**NSE**	**RMSE**	**MAE**
Coiflets wavelet (coif1)	MODWT-GRNN1	**0.996**	**0.991**	**0.115**	**0.034**	0.767	0.583	0.787	0.377
	MODWT-GRNN2	0.982	0.964	0.229	0.086	0.382	0.121	1.142	0.464
	MODWT-GRNN3	0.988	0.975	0.191	0.071	**0.803**	**0.640**	**0.731**	**0.368**
	MODWT-GRNN4	0.966	0.929	0.324	0.135	0.406	0.155	1.120	0.460
	MODWT-GRNN5	0.964	0.927	0.329	0.141	0.803	0.635	0.736	0.388
	MODWT-GRNN6	0.796	0.606	0.764	0.280	0.391	0.107	1.151	0.489
Daubechies wavelet (db3)	MODWT-GRNN1	**0.996**	**0.992**	**0.109**	**0.030**	0.678	0.446	0.907	0.421
	MODWT-GRNN2	0.984	0.967	0.222	0.088	0.295	−0.005	1.221	0.511
	MODWT-GRNN3	0.989	0.978	0.182	0.063	**0.782**	**0.607**	**0.764**	**0.381**
	MODWT-GRNN4	0.968	0.933	0.314	0.137	0.310	0.055	1.184	0.495
	MODWT-GRNN5	0.968	0.935	0.310	0.129	0.772	0.591	0.779	0.398
	MODWT-GRNN6	0.857	0.702	0.664	0.270	0.480	0.221	1.075	0.488
Symlet wavelet (sym4)	MODWT-GRNN1	**0.997**	**0.994**	**0.093**	**0.017**	0.866	0.707	0.660	0.275
	MODWT-GRNN2	0.988	0.976	0.190	0.069	0.541	0.284	1.031	0.433
	MODWT-GRNN3	0.995	0.990	0.121	0.034	0.910	0.805	0.538	0.257
	MODWT-GRNN4	0.970	0.939	0.300	0.114	0.574	0.325	1.001	0.422
	MODWT-GRNN5	0.981	0.962	0.239	0.076	**0.920**	**0.821**	**0.515**	**0.267**
	MODWT-GRNN6	0.825	0.656	0.713	0.264	0.343	0.004	1.216	0.468
Haar wavelet (haar)	MODWT-GRNN1	**0.993**	**0.985**	**0.148**	**0.041**	0.664	0.433	0.918	0.442
	MODWT-GRNN2	0.934	0.859	0.456	0.127	0.282	0.068	1.176	0.500
	MODWT-GRNN3	0.971	0.941	0.295	0.103	0.641	0.404	0.941	0.454
	MODWT-GRNN4	0.867	0.716	0.649	0.203	0.331	0.097	1.158	0.496
	MODWT-GRNN5	0.921	0.844	0.481	0.189	**0.704**	**0.484**	**0.876**	**0.455**
	MODWT-GRNN6	0.471	0.139	1.129	0.406	0.104	0.010	1.212	0.562
Fejer-Korovkin wavelet (fk8)	MODWT-GRNN1	**0.995**	**0.991**	**0.118**	**0.022**	0.647	0.406	0.939	0.357
	MODWT-GRNN2	0.983	0.965	0.229	0.071	0.425	0.156	1.119	0.437
	MODWT-GRNN3	0.989	0.977	0.183	0.048	**0.685**	**0.446**	**0.907**	**0.355**
	MODWT-GRNN4	0.948	0.892	0.400	0.133	0.481	0.195	1.093	0.434
	MODWT-GRNN5	0.946	0.891	0.401	0.120	0.633	0.375	0.963	0.378
	MODWT-GRNN6	0.833	0.631	0.739	0.257	0.500	0.197	1.092	0.451

The bold values indicate the best performance metric achieved within each category of models (standalone or hybrid) during either the training or validation phase.

Bazaar station. The current study's MODWT-GPR model, with an NSE of 0.98, shows comparable effectiveness, further validating the utility of wavelet-based hybrid models in streamflow prediction. Shabbir *et al.* (2023) proposed a hybrid method using HD-SVR, HD-KNN, and HD-ARIMA models, reporting RMSE values as low as 7.9314 m^3/s in the Indus River basin. While the RMSE values from the MODWT-GPR (sym4) model in this study are much smaller, especially during the validation phase (≈0.171 m^3/s), it is clear that the proposed model's ability to reduce error metrics is superior. This advantage can be attributed to the effectiveness of MODWT in capturing the multi-scale characteristics of hydrological time series, which might not be fully exploited by decomposition techniques like EMD. Moreover, Wang *et al.* (2021) developed the VMD-LSTM-PSO model and demonstrated its high accuracy and stability. Although this model showed strong predictive performance, particularly in the Yellow River basin, the MODWT-GPR (sym4) model presented in this study achieved even lower RMSE and higher NSE values, highlighting its robustness across different hydrological contexts.

This study marks a significant advancement in the application of hybrid models for streamflow prediction. By integrating MODWT with GPR, the study introduces a novel approach that outperforms both traditional ML models and other hybrid models previously documented. The superior performance of the MODWT-GPR (sym4) model suggests that it can effectively capture the complex, nonlinear relationships inherent in hydrological data, making it a valuable tool for accurate streamflow prediction.

The findings align with the growing body of research advocating for the use of hybrid models in hydrology. For instance, Hu *et al.* (2020) and He *et al.* (2019) both emphasized the importance of combining decomposition techniques like VMD with

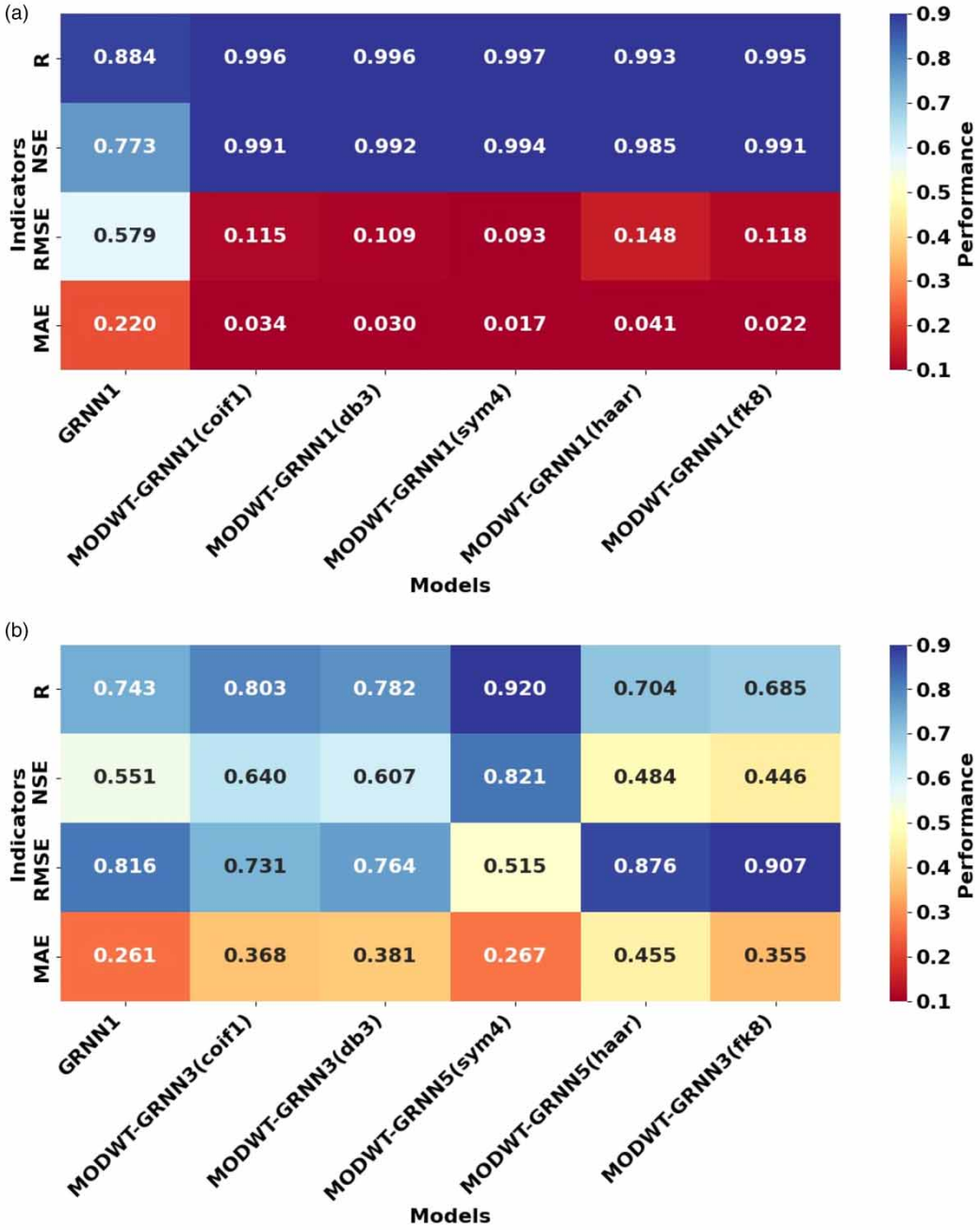

Figure 5 | Thematic maps showing the best performance criteria obtained by the hybrid MODWT-GRNN models for daily streamflow prediction at the Bir Ouled Tahar station. (a) Training and (b) validation.

advanced ML models to improve forecasting accuracy. The current study supports this view, demonstrating that the combination of MODWT and GPR offers a powerful approach to improving predictive accuracy. In their paper, Xie *et al.* (2019) used a new hybrid model, VMD-DBN-IPSO, to improve the accuracy of runoff forecasting at the Yangxian and Ankang hydrological stations in the Han River basin, China. Variable mode analysis (VMD) is used to analyze the original daily runoff series, and then, using the hybrid model combining the improved particle swarm optimization (IPSO) algorithm and the deep belief network (DBN), runoff is predicted. The results show that the VMD-DBN-IPSO model can still achieve the best performance in the training and testing phases and has good stability and representation; moreover, the NSE coefficient remains above 0.8, and the peak flow prediction error is less than 20%.

Table 5 | Results of streamflow prediction obtained by hybrid MODWT-MLPNN models on a daily time scale for the Bir Ouled Tahar station

		Training				Validation			
Mother wavelet	**Models**	***R***	**NSE**	**RMSE**	**MAE**	***R***	**NSE**	**RMSE**	**MAE**
Coiflets wavelet (coif1)	MODWT-MLPNN1	0.923	0.817	0.521	0.235	**0.890**	**0.787**	**0.562**	**0.344**
	MODWT-MLPNN2	0.664	0.436	0.914	0.399	0.656	0.424	0.925	0.512
	MODWT-MLPNN3	0.931	0.851	0.469	0.220	0.892	0.794	0.553	0.354
	MODWT-MLPNN4	0.640	0.408	0.936	0.389	0.690	0.461	0.894	0.444
	MODWT-MLPNN5	**0.939**	**0.878**	**0.425**	**0.226**	0.887	0.786	0.563	0.348
	MODWT-MLPNN6	0.513	0.262	1.045	0.409	0.543	0.292	1.025	0.491
Daubechies wavelet (db3)	MODWT-MLPNN1	0.884	0.775	0.577	0.251	**0.879**	**0.771**	**0.583**	**0.360**
	MODWT-MLPNN2	0.636	0.401	0.942	0.402	0.597	0.355	0.978	0.524
	MODWT-MLPNN3	0.837	0.686	0.681	0.297	0.817	0.644	0.727	0.463
	MODWT-MLPNN4	0.653	0.423	0.924	0.379	0.511	0.254	1.052	0.525
	MODWT-MLPNN5	**0.926**	**0.857**	**0.461**	**0.212**	0.863	0.701	0.666	0.388
	MODWT-MLPNN6	0.515	0.238	1.062	0.392	0.531	0.251	1.054	0.494
Symlet wavelet (sym4)	MODWT-MLPNN1	0.984	0.960	0.245	0.068	0.984	0.967	0.221	0.138
	MODWT-MLPNN2	0.600	0.350	0.981	0.370	0.639	0.389	0.952	0.451
	MODWT-MLPNN3	0.952	0.904	0.377	0.100	0.972	0.945	0.285	0.147
	MODWT-MLPNN4	0.657	0.431	0.918	0.352	0.630	0.393	0.949	0.458
	MODWT-MLPNN5	**0.991**	**0.980**	**0.171**	**0.056**	**0.985**	**0.968**	**0.218**	**0.128**
	MODWT-MLPNN6	0.491	0.231	1.067	0.374	0.523	0.256	1.051	0.487
Haar wavelet (haar)	MODWT-MLPNN1	**0.894**	**0.783**	**0.567**	**0.263**	0.794	0.628	0.743	0.444
	MODWT-MLPNN2	0.562	0.315	1.007	0.458	0.347	0.106	1.152	0.612
	MODWT-MLPNN3	0.888	0.776	0.576	0.268	**0.824**	**0.676**	**0.693**	**0.405**
	MODWT-MLPNN4	0.520	0.267	1.042	0.445	0.341	0.097	1.158	0.609
	MODWT-MLPNN5	0.873	0.750	0.608	0.267	0.821	0.674	0.695	0.388
	MODWT-MLPNN6	0.305	0.091	1.160	0.461	0.047	−0.056	1.252	0.617
Fejer-Korovkin wavelet (fk8)	MODWT-MLPNN1	0.810	0.639	0.731	0.272	0.836	0.690	0.679	0.330
	MODWT-MLPNN2	0.625	0.390	0.950	0.382	0.580	0.336	0.993	0.501
	MODWT-MLPNN3	0.775	0.598	0.772	0.273	0.827	0.677	0.692	0.339
	MODWT-MLPNN4	0.605	0.359	0.974	0.386	0.548	0.299	1.020	0.505
	MODWT-MLPNN5	**0.863**	**0.737**	**0.624**	**0.285**	**0.846**	**0.701**	**0.666**	**0.314**
	MODWT-MLPNN6	0.521	0.268	1.041	0.386	0.534	0.279	1.035	0.473

The bold values indicate the best performance metric achieved within each category of models (standalone or hybrid) during either the training or validation phase.

This study also shows that regardless of the base algorithm – whether GPR, LSTM, GRNN, or MLPNN – integrating MODWT preprocessing consistently enhances model performance. This finding aligns with earlier studies advocating wavelet-based hybridization for improving hydrological modeling accuracy.

An interesting observation emerged when comparing standalone and hybrid models. While the GPR3 model performed best during training, the LSTM1 model excelled in validation. This highlights the importance of evaluating models on independent datasets to avoid overfitting, as seen with the GPR3 model, and ensure predictions remain reliable.

The study's computational efficiency analysis reveals practical considerations. While MODWT-MLPNN5 (sym4) and MODWT-MLPNN5 (fk8) models provide superior predictions with rapid calculation speeds of approximately 8 s, more complex models like GPR and LSTM require longer processing times between 46 and 99 s. This tradeoff suggests that the MODWT-MLPNN models are ideal for real-time applications, while the MODWT-GPR (sym4) model may be better suited for in-depth offline analyses that prioritize accuracy.

The model's exceptional performance in capturing both trends and extreme events has promising implications for flood prediction and water management, as shown in the temporal analysis in Figure 8. By accurately forecasting peak and low flows, the model supports effective flood mitigation and sustainable water distribution.

Overall, the MODWT-GPR (sym4) hybrid model's accuracy in predicting daily streamflows, along with its adaptability across different regions of Algeria, marks significant progress in hydrological forecasting. The study's findings point to the future potential of integrating wavelet analysis with ML in water resource management and the continued development of data-driven hydrological models.

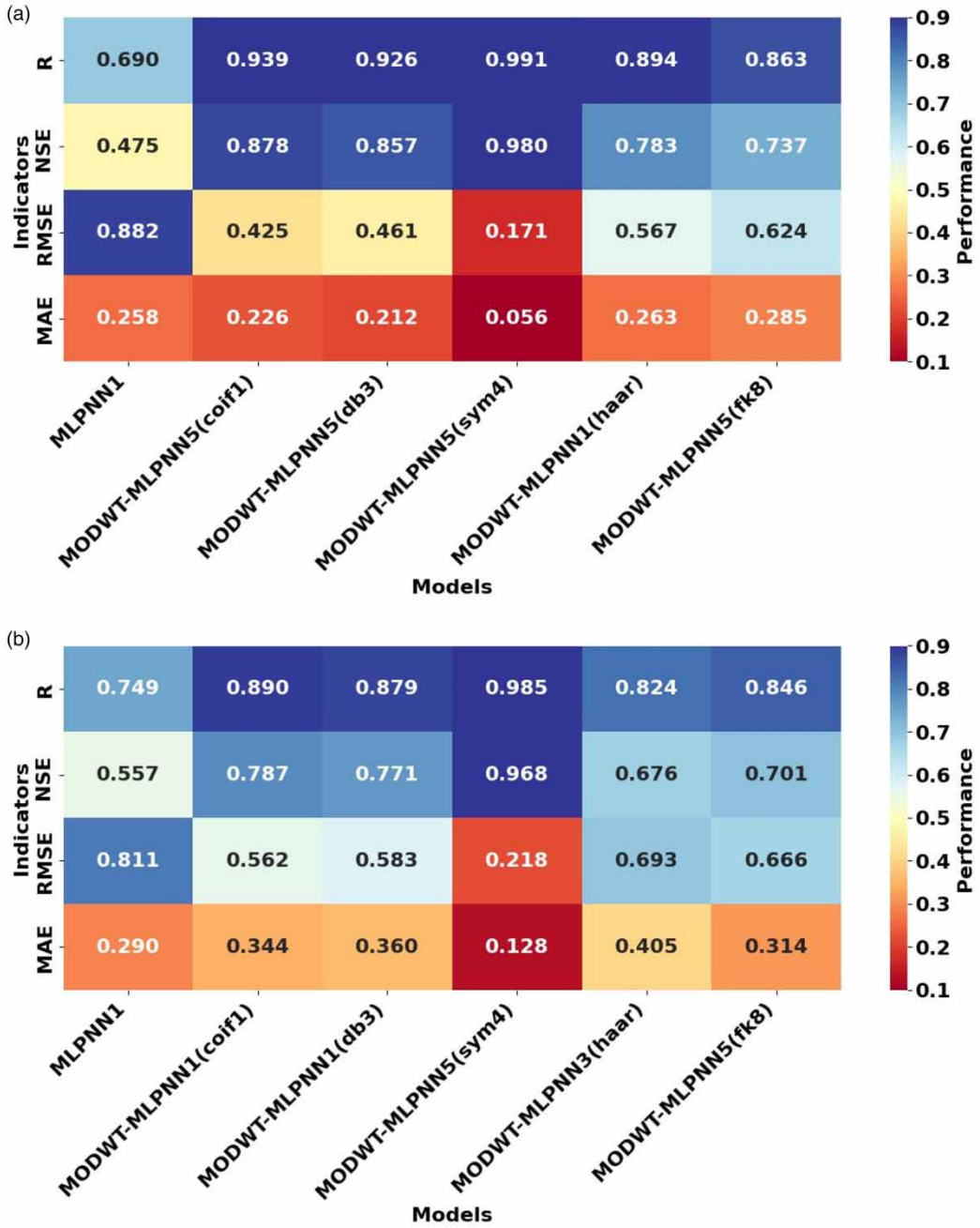

Figure 6 | Thematic maps showing the best performance criteria obtained by the hybrid MODWT-MLPNN models for daily streamflow prediction at the Bir Ouled Tahar station. (a) Training and (b) validation.

5. CONCLUSION

The goal of this study was to improve the predictability of daily streamflow in the Oued Rouina Zeddine watershed in northern Algeria, focusing on enhancing water flow predictions using hybrid models that combine signal analysis techniques with ML. The methods applied in this study were designed to explore the benefits of combining signal decomposition with ML techniques for streamflow prediction. Initially, four standalone models – GPR, LSTM, GRNN, and MLPNN – were developed and tested using historical data on streamflow and precipitation. These models were evaluated based on key performance metrics, such as *R*, NSE, RMSE, and MAE. Next, the MODWT was applied to decompose the data into various components,

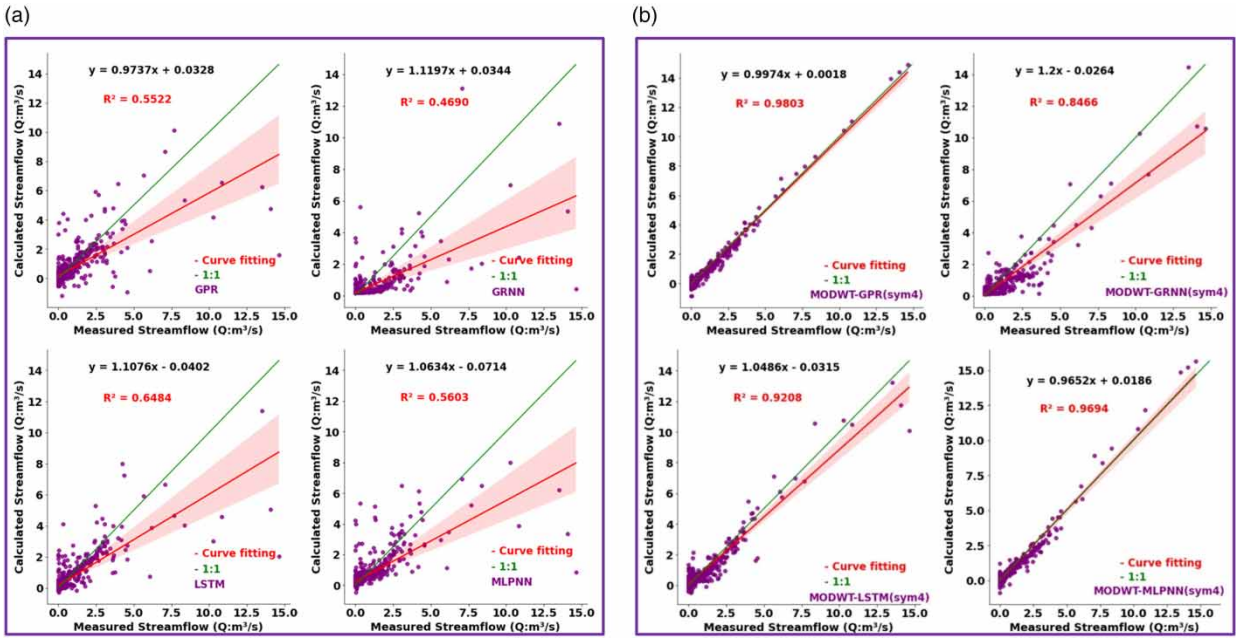

Figure 7 | Scatter plot of measured vs. calculated daily streamflows for the best (a) standalone and (b) hybrid ML models in the validation stage for the Bir Ouled Tahar station.

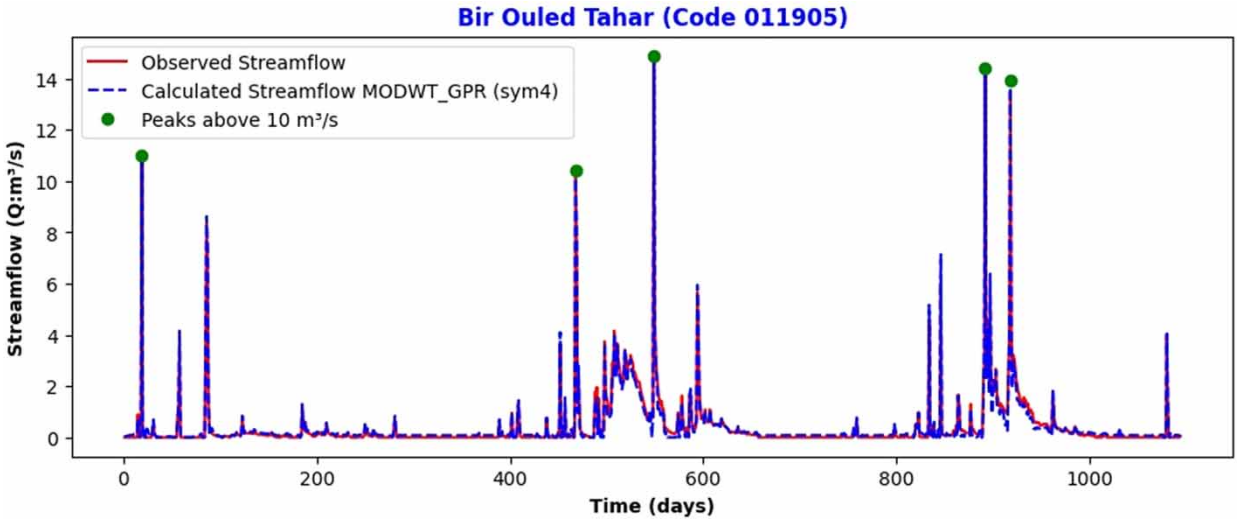

Figure 8 | Comparison between measured and predicted daily streamflow using the hybrid model MODWT-GPR3 (sym4) at the Bir Ouled Tahar station.

which were then fed into hybrid models. The study tested different wavelet families (coif1, db3, sym4, haar, and fk8) to determine which combination would yield the best results for streamflow prediction. The performance of the hybrid models was compared with the standalone models in both learning and validation phases to identify the most effective approach.

The results showed a clear improvement in prediction accuracy with the hybrid models, especially in comparison to the standalone models. Among the standalone models, the GPR3 model performed the best during the learning phase, achieving the highest correlation ($R = 0.993$) and NSE (0.986) values, along with the lowest RMSE (0.143 m³/s) and MAE (0.032 m³/s). In the validation phase, the LSTM1 model, with an R value of 0.805 and NSE $\approx$ 0.642, had the best performance among the

standalone models, though its RMSE (≈0.729 m³/s) and MAE (≈0.225 m³/s) were higher than those of the GPR3 model during training.

When combining MODWT with ML models, especially using the Symlet wavelet family (sym4), significant improvements were achieved. The hybrid model MODWT-GPR3 (sym4) emerged as the top performer, with superior accuracy in both the learning and validation phases. During validation, it reduced RMSE to 0.171 m³/s and MAE to 0.117 m³/s, outperforming the best standalone model (LSTM1). Other hybrid models, such as MODWT-LSTM3 (sym4), MODWT-GRNN5 (sym4), and MODWT-MLPNN5 (sym4), also showed notable improvements over their standalone counterparts.

The results were consistently supported by scatterplot analysis and performance graphs, which highlighted the superiority of the MODWT-GPR3 (sym4) model. This hybrid model was particularly effective in capturing nonlinear patterns in the data and accurately predicting peak flow values, as evidenced by the time series comparisons of measured and predicted streamflow.

Overall, this study provides strong evidence for the effectiveness of the MODWT-GPR (sym4) hybrid model in streamflow prediction. It highlights the potential for combining signal decomposition with ML techniques to enhance hydrological forecasts. To further enhance these findings, future research should explore other wavelet families and hybrid models, extending this approach to diverse hydrological environments.

AUTHOR CONTRIBUTIONS

All authors have read and agreed to the published version of the manuscript.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Abda, Z., Chettih, M. & Zerouali, B. (2020) Assessment of neuro-fuzzy approach based different wavelet families for daily flow rates forecasting, *Modeling Earth Systems and Environment*, **7** (3), 1523–1538. http://dx.doi.org/10.1007/s40808-020-00855-1.

Abnane, I., Idri, A. & Abran, A. (2023) Optimized fuzzy clustering-based K-nearest neighbors imputation for mixed missing data in software development effort estimation, *Journal of Software: Evolution and Process*, **36**, e2529. http://dx.doi.org/10.1002/smr.2529.

Ali, E., Zerouali, B., Tariq, A., Katipoğlu, O. M., Bailek, N., Santos, C. A. G., Ghoneim, S. S. M. & Towfiqul Islam, A. R. M. (2024) Fine-tuning inflow prediction models: integrating optimization algorithms and TRMM data for enhanced accuracy, *Water Science and Technology*, **90**(3), 844–877. https://doi.org/10.2166/wst.2024.222.

Alizadeh, M. J., Nourani, V., Mousavimehr, M. & Kavianpour, M. R. (2018) Wavelet-IANN model for predicting flow discharge up to several days and months ahead, *Journal of Hydroinformatics*, **20** (1), 134–148.

Alizadeh, A., Rajabi, A., Shabanlou, S., Yaghoubi, B. & Yosefvand, F. (2020) Simulation of rainfall and runoff time series using robust machine learning, *Irrigation and Drainage*, **70** (1), 84–102. http://dx.doi.org/10.1002/ird.2518.

Alizadeh, A. Rajabi, A., Shabanlou, S., Yaghoubi, B. & Yosefvand, F. (2021) Modeling long-term rainfall-runoff time series through wavelet-weighted regularization extreme learning machine, *Earth Science Informatics*, **14** (2), 1047–1063. http://dx.doi.org/10.1007/s12145-021-00603-8.

Amini, A., Dolatshahi, M. & Kerachian, R. (2024) Real-time rainfall and runoff prediction by integrating BC-MODWT and automatically-tuned DNNs: comparing different deep learning models, *Journal of Hydrology*, **631**, 130804. http://dx.doi.org/10.1016/j.jhydrol.2024.130804.

Barzegar, R., Aalami, M. T. & Adamowski, J. (2021) Coupling a hybrid CNN-LSTM deep learning model with a boundary corrected maximal overlap discrete wavelet transform for multiscale lake water level forecasting, *Journal of Hydrology*, **598**, 126196. http://dx.doi.org/10.1016/j.jhydrol.2021.126196.

Belletreche, M., Bailek, N., Abotaleb, M., Bouchouicha, K., Zerouali, B., Guermoui, M., Kuriqi, A., Alharbi, A. H., Khafaga, D. S., EL-Shimy, M. & El-kenawy, E.-S. M. (2024) Hybrid attention-based deep neural networks for short-term wind power forecasting using meteorological data in desert regions, *Scientific Reports*, **14** (1), 21842. http://dx.doi.org/10.1038/S41598-024-73076-6.

Chakraborty, S. & Biswas, S. (2023) River discharge prediction using wavelet-based artificial neural network and long short-term memory models: a case study of Teesta River Basin, India, *Stochastic Environmental Research and Risk Assessment*, **37** (8), 3163–3184.

Cloez, B., Fontez, B., González-García, E. & Sanchez, I. (2024) Kalman filter with impulse noised outliers: a robust sequential algorithm to filter data with a large number of outliers, *The International Journal of Biostatistics*. http://dx.doi.org/10.1515/ijb-2023-0065.

Daif, N. & Hebal, A. (2024) Enhanced daily streamflow forecasting in northeastern Algeria: integrating hybrid machine learning with advanced wavelet transformation techniques, *Modeling Earth Systems and Environment*, 1–29. http://dx.doi.org/10.1007/s40808-024-02067-3.

Daubechies, I. (1988) Time-frequency localization operators: a geometric phase space approach, *IEEE Transactions on Information Theory*, **34** (4), 605–612. http://dx.doi.org/10.1109/18.9761.

Daubechies, I. (1992) Ten lectures on wavelets. In: *Society for Industrial and Applied Mathematics*. https://doi.org/10.1137/1.9781611970104.

do Nascimento, T. V. M., Santos, C. A. G., Farias, C. A. S. d. & Silva, R. M. d. (2022) Monthly streamflow modeling based on self-organizing maps and satellite-estimated rainfall data, *Water Resources Management*, **36** (7), 2359–2377. http://dx.doi.org/10.1007/s11269-022-03147-8.

Dodig, A., Ricci, E., Kvascev, G. & Stojkovic, M. (2024) A novel machine learning-based framework for the water quality parameters prediction using hybrid long short-term memory and locally weighted scatterplot smoothing methods, *Journal of Hydroinformatics*, **26** (5), 1059–1079. http://dx.doi.org/10.2166/hydro.2024.273.

El-Kenawy, E. S. M., Zerouali, B., Bailek, N., Bouchouich, K., Hassan, M. A., Almorox, J., Kuriqi, A., Eid, M. & Ibrahim, A. (2022) Improved weighted ensemble learning for predicting the daily reference evapotranspiration under the semi-arid climate conditions. *Environmental Science and Pollution Research*, **29** (54), 81279–81299.

El-kenawy, E. S. M., Bailek, N., Bouchouicha, K., Zerouali, B., Hassan, M. A., Kuriqi, A., Jamil, B., Colak, I., Khalil, A. & Ibrahim, A. (2024) Global scale solar energy harnessing: An advanced intra-hourly diffuse solar irradiance predicting framework for solar energy projects. *Neural Computing and Applications*, **36** (18), 10585–10598

El-Rawy, M., Wahba, M., Fathi, H., Alshehri, F., Abdalla, F. & El Attar, R. M. (2024) Assessment of groundwater quality in arid regions utilizing principal component analysis, GIS, and machine learning techniques, *Marine Pollution Bulletin*, **205**, 116645. http://dx.doi.org/10.1016/j.marpolbul.2024.116645.

Farajpanah, H., Adib, A., Lotfirad, M., Esmaeili-Gisavandani, H., Riyahi, M. M. & Zaerpour, A. (2024) A novel application of waveform matching algorithm for improving monthly runoff forecasting using wavelet–ML models, *Journal of Hydroinformatics*, **26** (7), 1771–1789. http://dx.doi.org/10.2166/hydro.2024.128.

Ferkous, K., Guermoui, M., Bellaour, A., Boulmaiz, T. & Bailek, N. (2024) Enhancing photovoltaic energy forecasting: a progressive approach using wavelet packet decomposition, *Clean Energy*, **8** (3), 95–108.

Freire, P. K. d. M. M. & Santos, C. A. G. (2020) Optimal level of wavelet decomposition for daily inflow forecasting, *Earth Science Informatics*, **13** (4), 1163–1173. http://dx.doi.org/10.1007/s12145-020-00496-z.

Freire, P. K. d. M. M., Santos, C. A. G. & Silva, G. B. L. d. (2019) Analysis of the use of discrete wavelet transforms coupled with ANN for short-term streamflow forecasting, *Applied Soft Computing*, **80**, 494–505. http://dx.doi.org/10.1016/j.asoc.2019.04.024.

Gomaa, E., Zerouali, B., Difi, S., El-Nagdy, K. A., Santos, C. A. G., Abda, Z., Ghoneim, S. M., Bailek, N., da Silva, R. M. & Rajput, J. (2023) Assessment of hybrid machine learning algorithms using TRMM rainfall data for daily inflow forecasting in Três Marias Reservoir, eastern Brazil, *Heliyon*, **9**, e18819.

Gomes, E. P. & Blanco, C. J. C. (2021) Daily rainfall estimates considering seasonality from a MODWT-ANN hybrid model, *Journal of Hydrology and Hydromechanics*, **69** (1), 13–28.

Habib, M. & Okayli, M. (2024) Evaluating the sensitivity of machine learning models to data preprocessing technique in concrete compressive strength estimation, *Arabian Journal for Science and Engineering*, 1–19. http://dx.doi.org/10.1007/s13369-024-08776-2.

He, X., Luo, J., Zuo, G. & Xie, J. (2019) Daily runoff forecasting using a hybrid model based on variational mode decomposition and deep neural networks, *Water Resources Management*, **33**, 1571–1590.

Hu, H., Zhang, J. & Li, T. (2020) A comparative study of VMD-based hybrid forecasting model for nonstationary daily streamflow time series, *Complexity*, **2020**, 4064851.

Ibrahim, A., Khodadadi, Ehsan, Khodadadi, Ehsaneh, Dutta, P. K., Bailek, N. & Abdelhamid, A. A. (2024) Apple perfection: assessing apple quality with waterwheel plant algorithm for feature selection and logistic regression for classification, *Journal of Artificial Intelligence in Engineering Practice*, **1** (1), 34–48.

Khan, M. T., Shoaib, M., Hammad, M., Salahudin, H., Ahmad, F. & Ahmad, S. (2021) Application of machine learning techniques in rainfall–runoff modelling of the Soan River basin, Pakistan, *Water*, **13** (24), 3528. http://dx.doi.org/10.3390/w13243528.

Küllahcı, K. & Altunkaynak, A. (2024) Maximizing daily rainfall prediction accuracy with maximum overlap discrete wavelet transform-based machine learning models, *International Journal of Climatology*, **44** (10), 3405–3426. http://dx.doi.org/10.1002/joc.8530.

Li, J., Guo, S., Ma, R., He, J., Zhang, X., Rui, D., Ding, Y., Li, Y., Jian, L., Cheng, J. & Guo, H. (2024) Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets, *BMC Medical Research Methodology*, **24** (1), 41.

Mallat, S. G. (1989) Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$, *Transactions of the American Mathematical Society*, **315** (1), 69–87. http://dx.doi.org/10.1090/s0002-9947-1989-1008470-5.

Ni, Z., Zhu, Y., Qian, Y., Li, X., Xing, Z., Zhou, Y., Chen, Y., Huang, L., Yang, J. & Zhuge, Q. (2024) Synthetic minority over-sampling technique-enhanced machine learning models for predicting recurrence of postoperative chronic subdural hematoma, *Frontiers in Neurology*, **15**, 1305543.

Oulimar, I., Bouchouicha, K., Bailek, N. & Bellaoui, M. (2024) Statistical study of global solar radiation in the Algerian desert: a case study of Adrar Town, *Theoretical and Applied Climatology*, **155** (4), 3493–3504.

Percival, D. B. & Walden, A. T. (2000) *4 Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge, UK.

Quilty, J., Adamowski, J. & Boucher, M. (2019) A stochastic data-driven ensemble forecasting framework for water resources: a case study using ensemble members derived from a database of deterministic wavelet-based models, *Water Resources Research*, **55** (1), 175–202. http://dx.doi.org/10.1029/2018wr023205.

Roushangar, K., Alizadeh, F. & Nourani, V. (2017) Improving capability of conceptual modeling of watershed rainfall–runoff using hybrid wavelet-extreme learning machine approach, *Journal of Hydroinformatics*, **20** (1), 69–87. http://dx.doi.org/10.2166/hydro.2017.011.

Roy, B., Singh, M. P., Kaloop, M. R., Kumar, D., Hu, J.-W., Kumar, R. & Hwang, W.-S. (2021) Data-driven approach for rainfall-runoff modelling using equilibrium optimizer coupled extreme learning machine and deep neural network, *Applied Sciences*, **11** (13), 6238. http://dx.doi.org/10.3390/app11136238.

Santos, C. A. G., Srinivasan, V. S., Suzuki, K. & Watanabe, M. (2003) Application of an optimization technique to a physically based erosion model, *Hydrological Processes*, **17** (5), 989–1003. http://dx.doi.org/10.1002/hyp.1176.

Saraiva, S. V., Carvalho, F. de O., Santos, C. A. G., Barreto, L. C. & Freire, P .K. de M. M. (2021) Daily streamflow forecasting in Sobradinho Reservoir using machine learning models coupled with wavelet transform and bootstrapping, *Applied Soft Computing*, **102**, 107081. http://dx.doi.org/10.1016/j.asoc.2021.107081.

Seo, Y., Choi, Y. & Choi, J. (2017) River stage modeling by combining maximal overlap discrete wavelet transform, support vector machines and genetic algorithm, *Water*, **9** (7), 525. http://dx.doi.org/10.3390/w9070525.

Shabbir, M., Chand, S. & Iqbal, F. (2023) Prediction of river inflow of the major tributaries of Indus River basin using hybrids of EEMD and LMD methods, *Arabian Journal of Geosciences*, **16** (4), 257.

Sugawara, M. (1979) Automatic calibration of the tank model/L'étalonnage Automatique d'un Modèle à Cisterne, *Hydrological Sciences Bulletin*, **24** (3), 375–388. http://dx.doi.org/10.1080/02626667909491876.

Wang, X., Wang, Y., Yuan, P., Wang, L. & Cheng, D. (2021) An adaptive daily runoff forecast model using VMD-LSTM-PSO hybrid approach, *Hydrological Sciences Journal*, **66** (9), 1488–1502.

Xie, T., Zhang, G., Hou, J., Xie, J., Lv, M. & Liu, F. (2019) Hybrid forecasting model for non-stationary daily runoff series: a case study in the Han River Basin, China, *Journal of Hydrology*, **577**, 123915.

Zerouali, B., Santos, C. A. G., de Farias, C. A. S., Muniz, R. S., Difi, S., Abda, Z., Chettih, M., Heddam, S., Anwar, S. A. & Elbeltagi, A. (2023a) Artificial intelligent systems optimized by metaheuristic algorithms and teleconnection indices for rainfall modeling: the case of a humid region in the Mediterranean basin, *Heliyon*, **9** (4).

Zerouali, Bilel, Pawar, U. V, Elbeltagi, A., Abda, Z., Chettih, M., Santos, C. A. G. & Difi, S. (2023b) Change-point detection in monsoon rainfall of Narmada River (Central India) during 1901–2015, *Journal of Earth System Science*, **132** (3), e15355. http://dx.doi.org/10.1007/s12040-023-02140-y.

Zerouali, Bilel, Bailek, N., Tariq, A., Kuriqi, A., Guermoui, M., Alharbi, A. H., Khafaga, D. S. & El-kenawy, E.-S. M. (2024a) Enhancing deep learning-based slope stability classification using a novel metaheuristic optimization algorithm for feature selection, *Scientific Reports*, **14** (1), 1–21. http://dx.doi.org/10.1038/S41598-024-72588-5.

Zerouali, B, Bailek, N., Islam, A. R. M. T., Katipoğlu, O. M., Ayek, A. A. E., Santos, C. A. G., Rajput, J., Wong, Y. J., Abda, Z., Chettih, M. & Elbeltagi, A. (2024b) Enhancing groundwater potential zone mapping with a hybrid analytical method: the case of semiarid basin, *Groundwater for Sustainable Development*, **26**, 101261.

Zheng, Z., Jiang, Y., Zhang, Q., Zhong, Y. & Wang, L. (2024) A feature selection method based on relief feature ranking with recursive feature elimination for the inversion of urban river water quality parameters using multispectral imagery from an unmanned aerial vehicle, *Water*, **16** (7), 1029. http://dx.doi.org/10.3390/w16071029.

First received 14 August 2024; accepted in revised form 29 November 2024. Available online 9 December 2024