JSCE (Journal of System and Computer Engineering)
Vol.7, No.3, July 2026, pp. 292-310
ISSN (online): 2723-1240
DOI : https://doi.org/10.61628/jsce.v7i3.2741

Explainable Machine Learning for Long-Term Monthly Hydroclimatic Forecasting and Extreme-Event Detection

[1,a,*]Arif Fadillah, [2,b]Markani Pato, [3,c]Nuraida Latif, [3,d] Muhajirin, [4,e]Benny Leornard Encrico Panggabean, [3,f]Muhammad Rizal H, [3,g]Mursalim

[1] *Sistem Informasi, Universitas Muhammadiyah Ahmad Dahlan, Palembang, Indonesia*
[2] *Sistem Informasi, Universitas Teknologi Akba Makassar, Makassar, Indonesia*
[3] *Teknik Informatika, Universitas Teknologi Akba Makassar, Makassar, Indonesia*
[4] *Ilmu Komputer, Universitas Pancasakti, Makassar, Indonesia*

[a]aafifah170114@gmail.com; [b]markani@akba.ac.id; [c]nuraida@akba.ac.id; [d]aji@akba.ac.id; [e]blep@unpacti.ac.id; [f]rizal@unitama.ac.id, [g]mursalim@unitama.ac.id
* Corresponding author

Abstract

Long-term hydroclimatic prediction in arid urban environments remains methodologically demanding because monthly records are often intermittent, highly seasonal, zero-inflated, and dominated by rare but consequential extreme events. Using a 121-year monthly hydroclimatic record for Makkah, Saudi Arabia, spanning January 1901 to December 2021, this study develops an explainable hybrid machine-learning framework for monthly forecasting, seasonal diagnostics, and extreme-event detection. The dataset contains 1,452 monthly observations with a mean value of 6.19, median of 3.00, standard deviation of 8.05, and maximum of 52.00, indicating a strongly skewed distribution. Exploratory analysis reveals pronounced seasonality: November, December, and January exhibit the highest hydroclimatic values, whereas June is consistently dry across the full record. A temporal feature set was constructed using lag variables, rolling statistics, annual seasonal memory, cyclical month encodings, and trend indicators. Several predictive models were evaluated, including Random Forest, Extra Trees, Histogram Gradient Boosting, XGBoost, and a hybrid SARIMA–Random Forest residual-correction model. Extra Trees achieved the best forecasting performance on the holdout period, with MAE = 2.997, RMSE = 5.603, sMAPE = 57.669%, and R^2 = 0.518. Extreme-event detection was performed using a 90th-percentile threshold of 17.68, identifying 146 extreme months over the full record. The best classification trade-off was obtained by Histogram Gradient Boosting, while Random Forest produced the highest ROC-AUC. SHAP-based interpretation demonstrates that seasonal phase variables and annual memory features dominate model behaviour, especially month_cos, month_sin, same_month_last_year, and lag_12. The findings show that interpretable ensemble learning can provide a more transparent and operationally relevant framework than accuracy-only forecasting for arid-region hydroclimatic risk assessment.

Keywords—Hydroclimatic forecasting; explainable artificial intelligence; machine learning; extreme-event detection; SHAP

1. Introduction

Hydroclimatic variability has become a central concern in climate-risk research because water-related extremes are no longer peripheral anomalies but recurrent stressors for urban infrastructure, water-resource planning, and disaster preparedness. The scientific literature has repeatedly shown that changes in precipitation extremes are not simply proportional to changes in mean precipitation; heavy precipitation can intensify even where annual totals remain low or uncertain. Evidence from global observational and modelling studies indicates robust increases in extreme precipitation across both dry and wet regions, while reviews of short-duration rainfall extremes emphasize the growing relevance of high-intensity events for flash-flood risk and urban drainage design (Donat et al., 2016; Westra et al., 2014; Fischer & Knutti, 2016; Min et al., 2011).

The Arabian Peninsula represents a particularly demanding context for hydroclimatic modelling. It combines aridity, water scarcity, high spatial heterogeneity, limited observational density, and rapidly expanding urban systems. Recent regional syntheses identify the Eastern Mediterranean and Middle East as a climate-change hotspot, with warming and weather extremes posing multi-sectoral risks, while new Saudi rainfall products explicitly characterize the Arabian Peninsula as one of the most arid, water-stressed, and data-sparse regions on Earth (Zittis et al., 2022; Wang et al., 2025). Within Saudi Arabia, earlier climatological work has shown that rainfall behaviour is strongly seasonal and regionally variable, making local-scale forecasting and extreme-event screening especially important for cities such as Makkah, where episodic rainfall may have disproportionate hydrological and infrastructural consequences (Almazroui et al., 2012).

Machine learning has increasingly reshaped hydroclimatic modelling by allowing nonlinear temporal relationships to be inferred from long observational records. In hydrology, deep learning and tree-based learning have shown strong predictive value, particularly when conventional process models are constrained by sparse inputs, uncertain parameterization, or nonlinear catchment response. Yet the literature also warns against treating predictive accuracy as a sufficient criterion for scientific usefulness. Reviews in water resources and Earth-system science argue that data-driven models become more valuable when they are coupled with process understanding, uncertainty reasoning, and interpretable diagnostics (Shen, 2018; Reichstein et al., 2019; Nearing et al., 2021; Tripathy & Mishra, 2024).

Three research gaps motivate the present work. First, many hydroclimatic forecasting studies remain accuracy-centred, reporting regression metrics without explaining why a model predicts a wet, dry, or extreme month. Second, forecasting and extreme-event detection are often treated as separate tasks, even though decision-makers need both expected monthly values and warning signals for unusually high events. Third, hybrid models are frequently presented as inherently superior, whereas their performance depends on the structure of the series, the residual signal, the degree of zero inflation, and the match between model assumptions and hydroclimatic behaviour. Recent reviews of hybrid hydroclimatic prediction explicitly call for physically explainable and operationally usable hybrid systems, not merely algorithmic combinations (Slater et al., 2023).

The novelty of this study lies in integrating long-term monthly forecasting, hybrid statistical–machine learning modelling, explainable AI, and extreme-event detection within a single reproducible workflow. Rather than asking only which model minimizes error, the analysis asks how Makkah's monthly hydroclimatic behaviour is structured, whether nonlinear ensemble models can forecast it more effectively than a seasonal hybrid baseline, which temporal features drive predictions, and how well the same feature system supports extreme-event identification. This framing moves the analysis beyond model ranking toward a more decision-relevant understanding of seasonal memory, event rarity, and interpretable risk signals.

1.1 Related Works

1.1.1 Hydroclimatic forecasting between statistical structure and machine learning flexibility

Time-series forecasting has historically relied on statistical models that impose interpretable structure on temporal dependence, seasonality, and residual behaviour. ARIMA, exponential smoothing, and state-space models remain attractive because they are parsimonious, computationally efficient, and transparent. Automated ARIMA and exponential-smoothing frameworks have been widely adopted precisely because they provide robust baselines for seasonal and nonseasonal series (Hyndman & Khandakar, 2008; De Livera et al., 2011). However, empirical forecasting competitions have complicated the assumption that one methodological family is universally superior. The M4 competition showed that hybrid and ensemble strategies can outperform single-model approaches when they combine complementary strengths across diverse time-series structures (Makridakis et al., 2020).

Deep learning has expanded the forecasting landscape by enabling nonlinear sequence modelling, probabilistic prediction, and multi-horizon forecasting. DeepAR introduced autoregressive recurrent probabilistic forecasting for related time series, while Temporal Fusion Transformers explicitly combine multi-horizon accuracy with interpretability through attention and feature-selection mechanisms (Salinas et al., 2020; Lim et al., 2021). Yet recurrent-neural-network reviews caution that deep models are not automatic replacements for statistical baselines, particularly where individual series are short, noisy, or strongly seasonal (Hewamalage et al., 2021). This caution is especially relevant for a single monthly hydroclimatic series with 1,452 observations: the sample is long climatologically, but modest for high-capacity deep architectures.

1.1.2 Machine learning in hydrology and water-resources science

Hydrological machine learning has advanced rapidly because water systems often display nonlinear, threshold-like, and memory-dependent behaviour. LSTM-based rainfall–runoff modelling demonstrated the capacity of recurrent networks to learn long-term hydrological dependencies, while subsequent studies showed that large-sample learning can improve prediction in ungauged basins and across multiple temporal scales (Kratzert et al., 2018, 2019; Gauch et al., 2021). At the same time, tree-based ensemble methods such as Random Forest and Extra Trees remain attractive in applied hydroclimatology because they perform well with engineered lag features, handle nonlinear interactions, and are comparatively robust under moderate sample sizes (Breiman, 2001; Geurts et al., 2006; Pham et al., 2021).

Recent reviews stress that hydrological modelling is moving from a dichotomy between physical and data-driven models toward a continuum of hybrid, physics-guided, and explainable systems. Deep learning is increasingly used for drought and flood forecasting, hydroclimate downscaling, groundwater modelling, and water-quality assessment, but its scientific value depends on whether it can generate physically plausible and interpretable outputs (Tripathy & Mishra, 2024). For univariate monthly hydroclimatic forecasting, therefore, the relevant methodological question is not whether a deep model is fashionable, but whether the data support a model complex enough to improve prediction without sacrificing interpretability.

1.1.3 Hybrid forecasting and residual correction

Hybrid forecasting has emerged as a promising response to the limitations of both purely statistical and purely machine-learning approaches. Statistical models are effective at capturing

DOI : https://doi.org/10.61628/jsce.v7i3.2741

regular seasonal structure, whereas nonlinear learners can exploit residual patterns that remain after trend and seasonality have been modelled. Slater et al. (2023) define hybrid hydroclimatic forecasting as the deliberate blending of data-driven methods with dynamical, statistical, or physics-based prediction systems to improve skill while acknowledging multiple sources of predictability.

Nevertheless, the literature does not support a simplistic assumption that hybridization is always beneficial. A residual-correction model can improve forecasts only when residuals contain learnable information rather than noise, structural breaks, or unobserved covariates. This is particularly important in arid hydroclimatic series, where extreme months may be generated by synoptic-scale atmospheric disturbances that are not encoded in a univariate historical record. Hybrid modelling must therefore be evaluated empirically and interpreted diagnostically rather than treated as a methodological guarantee.

1.1.4 Explainable AI and hydroclimatic interpretability

The growing use of black-box models has intensified demand for explainability. XAI reviews argue that interpretability is not a cosmetic add-on but a condition for accountability, trust, and scientific learning, especially when model outputs support consequential decisions (Arrieta et al., 2020). In tree-based modelling, SHAP provides a principled way to translate complex model behaviour into feature-level contributions, connecting local predictions with global interpretability (Lundberg et al., 2020).

In hydroclimatic studies, explainability has a particular scientific role: it can reveal whether models rely on plausible temporal signals such as seasonality, annual memory, antecedent wetness, or rolling variability. If an accurate model depends primarily on spurious trend indicators or unstable short-term noise, its operational credibility weakens. Conversely, when SHAP diagnostics show dominance of physically meaningful seasonal and lag-based features, the model becomes more defensible for risk-informed decision-making.

1.1.5 Extreme-event detection and anomaly reasoning

Extreme hydroclimatic events require different evaluation logic from average-value forecasting. A model may achieve low RMSE by predicting seasonal means while failing to detect rare high-impact months. Anomaly-detection literature distinguishes between ordinary predictive error and structurally unusual observations, while comparative evaluations show that unsupervised anomaly methods can complement supervised classification when labelled extremes are scarce (Chandola et al., 2009; Goldstein & Uchida, 2016).

For arid-region hydroclimatology, this distinction is crucial. The same monthly series may contain many dry or near-dry months and a small number of large peaks. Consequently, a forecast system should be assessed not only by continuous-value metrics but also by its ability to flag months exceeding a decision-relevant threshold. This study therefore combines regression forecasting, threshold-based extreme classification, and unsupervised anomaly detection.

2. Method

2.1 Research design

The research design follows a sequential and reproducible workflow: data acquisition, preprocessing, exploratory analysis, feature engineering, model development, hybrid forecasting, explainability analysis, extreme-event detection, and evaluation. The workflow is summarized in

Figure 1. Its central logic is that long-term hydroclimatic forecasting should not be reduced to a single predictive algorithm. A scientifically useful framework must expose the temporal structure of the series, benchmark multiple model classes, explain model behaviour, and translate predictions into risk-relevant outputs.

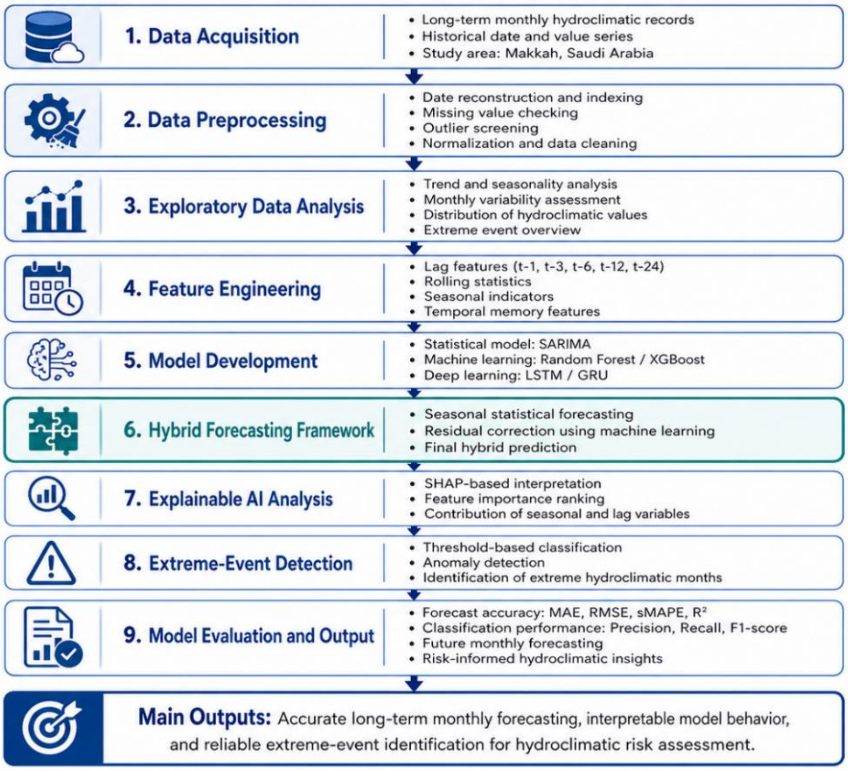


Figure 1. Research Framework

2.2 Data source and temporal reconstruction

The empirical dataset consists of a monthly hydroclimatic dataset for Makkah, Saudi Arabia, stored in a file named Mecca.csv. The dataset contains two variables, namely Date and Value, representing monthly hydroclimatic observations collected from January 1901 to December 2021., containing two variables: Date and Value. The file includes 1,452 monthly observations. Because the date strings use two-digit years, the chronological index was reconstructed as a continuous monthly sequence from January 1901 to December 2021. This reconstruction is justified by the exact correspondence between the number of records and a complete 121-year monthly sequence.

The variable Value is treated as a monthly hydroclimatic amount. Since the uploaded file does not provide physical units or metadata, the analysis avoids assigning a definitive measurement unit. However, the zero-inflated distribution, strong seasonality, and extreme high values are consistent with rainfall-like hydroclimatic behaviour in an arid environment.

2.3 Preprocessing and exploratory diagnostics

Preprocessing included numeric conversion of Value, missing-value checking, chronological sorting, and outlier screening through descriptive statistics and graphical inspection. No missing values were found. The descriptive analysis used count, mean, standard deviation, quartiles, and extrema. Seasonality was diagnosed using month-wise boxplots and monthly summary statistics. Extreme months were initially explored through distributional inspection and subsequently formalized through a percentile-based threshold.

2.4 Feature engineering

The feature set was designed to encode short-term persistence, annual memory, seasonal phase, and local variability. After applying the maximum lag of 24 months, 1,428 usable observations remained. Thirty predictors were constructed:

- calendar and trend indicators: year, month, time_idx;
- cyclic seasonal encodings: month_sin, month_cos;
- lag features: lag_1, lag_2, lag_3, lag_6, lag_12, lag_24;
- rolling statistics over 3, 6, 12, and 24 months: rolling mean, standard deviation, minimum, and maximum;
- seasonal-memory variables: same_month_last_year, same_month_2y_ago;
- annual seasonal contrast: seasonal_diff_12.

All lagged and rolling features were shifted so that only information available prior to the prediction month entered the model. This avoids look-ahead bias.

2.5 Train-test partitioning

The engineered dataset was split chronologically rather than randomly. The first 80% of observations formed the training set, and the final 20% formed the holdout set. The training period extended from January 1903 to February 1998, while the holdout period extended from March 1998 to December 2021, comprising 286 months. This design respects temporal dependence and evaluates out-of-sample forecasting performance under a realistic forward-prediction setting.

2.6 Forecasting models

Five forecasting approaches were evaluated to model the nonlinear, seasonal, and intermittent structure of the monthly hydroclimatic series: Random Forest Regressor, Extra Trees Regressor, Histogram Gradient Boosting Regressor, XGBoost Regressor, and a hybrid SARIMA–Random Forest residual-correction model. These models were selected to represent complementary forecasting paradigms, including bagging-based ensembles, randomized tree ensembles, boosting-based learners, regularized gradient boosting, and hybrid statistical–machine learning forecasting. The purpose of this comparison was not only to identify the model with the lowest prediction error, but also to examine whether different algorithmic assumptions were suitable for a long, zero-inflated, and strongly seasonal hydroclimatic time series. The model comparison follows the manuscript's original experimental design, where these five approaches were evaluated on the chronological holdout period.

Let y_t denote the observed monthly hydroclimatic value at time t, and let $\mathbf{x}_t = (x_{t1}, x_{t2}, \dots, x_{tp})$ denote the vector of engineered predictors, including calendar variables, cyclic seasonal encodings, lagged values, rolling statistics, and annual-memory features. The general supervised forecasting problem can be expressed as:

$$\hat{y}_t = f(\mathbf{x}_t) + \varepsilon_t \tag{1}$$

where $\hat{y}_t$ is the predicted monthly value, $f(\cdot)$ is the forecasting function learned from the training data, and ε_t is the residual error.

2.6.1 Random Forest Regressor

The Random Forest Regressor was used as a bagging-based ensemble model capable of capturing nonlinear relationships and high-order interactions among temporal predictors. Random Forest constructs an ensemble of B regression trees, where each tree is trained on a bootstrap sample of the training data and each split is selected from a random subset of predictors. The final prediction is obtained by averaging the outputs of all individual trees:

$$\hat{y}_t^{RF} = \frac{1}{B}\sum_{b=1}^{B} T_b\,(\mathbf{x}_t) \tag{2}$$

where $T_b(\mathbf{x}_t)$ is the prediction of the b-th regression tree and B is the total number of trees. This averaging mechanism reduces variance and improves generalization, particularly in feature-engineered hydroclimatic data where seasonal and lag-based predictors may interact nonlinearly.

2.6.2 Extra Trees Regressor

The Extra Trees Regressor, or Extremely Randomized Trees, extends the ensemble-tree principle by introducing stronger randomization during tree construction. Unlike Random Forest, which typically searches for the best split among candidate thresholds, Extra Trees selects split thresholds more randomly. This additional randomization can reduce variance and improve robustness when the dataset contains noisy, intermittent, or highly skewed observations.

The prediction of the Extra Trees model is similarly defined as the average of B randomized regression trees:

$$\hat{y}_t^{ET} = \frac{1}{B}\sum_{b=1}^{B} T_b^*(\mathbf{x}_t) \tag{3}$$

where $T_b^*(\mathbf{x}_t)$ represents the b-th extremely randomized tree. In this study, Extra Trees was particularly suitable because the monthly hydroclimatic series contains strong seasonality, repeated dry months, and abrupt extreme peaks. The manuscript reports that Extra Trees achieved the best forecasting performance among all evaluated models.

2.6.3 Histogram Gradient Boosting Regressor

The Histogram Gradient Boosting Regressor was included as a boosting-based model that sequentially improves prediction by fitting new trees to the residual errors of previous trees. Gradient boosting estimates the forecasting function as an additive expansion of weak learners:

$$\hat{y}_t^{HGB} = F_M(\mathbf{x}_t) = F_0(\mathbf{x}_t) + \sum_{m=1}^{M} \nu\, h_m(\mathbf{x}_t) \tag{4}$$

where $F_0(\mathbf{x}_t)$ is the initial prediction, $h_m(\mathbf{x}_t)$ is the regression tree fitted at boosting iteration m, M is the total number of boosting iterations, and ν is the learning rate. At each iteration, the model minimizes a differentiable loss function $L(y_t, F(\mathbf{x}_t))$ by fitting the next tree to the negative gradient:

$$r_{tm} = -\left[\frac{\partial L(y_t, F(\mathbf{x}_t))}{\partial F(\mathbf{x}_t)}\right]_{F=F_{m-1}} \tag{5}$$

where r_{tm} denotes the pseudo-residual at iteration m. The histogram-based implementation improves computational efficiency by binning continuous predictors, which is useful when many lagged and rolling features are included.

2.6.4 XGBoost Regressor

XGBoost was used as a regularized gradient-boosting framework designed to improve prediction accuracy while controlling model complexity. The XGBoost prediction is expressed as an additive ensemble of regression trees:

$$\hat{y}_t^{XGB} = \sum_{k=1}^{K} f_k(\mathbf{x}_t), \quad f_k \in \mathcal{F} \tag{6}$$

where K is the number of trees and $\mathcal{F}$ is the space of regression-tree functions. The model is trained by minimizing a regularized objective function:

$$\mathcal{L}^{(K)} = \sum_{t=1}^{n} l\,(y_t, \hat{y}_t) + \sum_{k=1}^{K} \Omega\,(f_k) \tag{7}$$

where $l(y_t, \hat{y}_t)$ is the prediction loss and $\Omega(f_k)$ is a regularization term that penalizes tree complexity. A common form of the regularization function is:

$$\Omega(f_k) = \gamma J + \frac{1}{2}\lambda \sum_{j=1}^{J} w_j^2 \tag{8}$$

where J is the number of terminal leaves, w_j is the score assigned to leaf j, γ controls the penalty for adding leaves, and λ is the L_2 regularization coefficient. This formulation allows XGBoost to balance model fit and complexity. However, in the manuscript's empirical results, XGBoost performed less effectively than the tree-bagging approaches, indicating that stronger algorithmic regularization and boosting depth did not necessarily improve performance for this intermittent monthly hydroclimatic series.

2.6.5 Hybrid SARIMA-Random Forest residual-correction model

The hybrid SARIMA–Random Forest model was designed to combine statistical seasonal modelling with nonlinear machine-learning residual correction. The SARIMA component first models the linear autocorrelation and seasonal structure of the monthly series. A general seasonal ARIMA model can be written as:

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D y_t = \Theta_Q(B^s)\theta_q(B)\varepsilon_t \tag{9}$$

where B is the backshift operator, s is the seasonal period, $\phi_p(B)$ and $\theta_q(B)$ are the nonseasonal autoregressive and moving-average polynomials, $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ are the seasonal autoregressive and moving-average polynomials, and d and D are the nonseasonal and seasonal differencing orders.

The SARIMA forecast is denoted as:

$$\hat{y}_t^{SARIMA} \tag{10}$$

The residual from the SARIMA model is then computed as:

$$e_t = y_t - \hat{y}_t^{SARIMA} \tag{11}$$

A Random Forest model is subsequently trained to predict the residual component using the engineered feature vector $\mathbf{x}_t$:

$$\hat{e}_t^{RF} = \frac{1}{B} \sum_{b=1}^{B} T_b\,(\mathbf{x}_t) \tag{12}$$

The final hybrid forecast is obtained by adding the SARIMA forecast and the machine-learning residual correction:

$$\hat{y}_t^{Hybrid} = \hat{y}_t^{SARIMA} + \hat{e}_t^{RF} \tag{13}$$

This formulation assumes that the SARIMA model captures regular seasonal and autocorrelative structure, while the Random Forest component captures nonlinear residual patterns that remain unexplained by the statistical model. However, the empirical results showed that the hybrid SARIMA–Random Forest model did not outperform Extra Trees or Random Forest. This suggests that the residual component may not have contained sufficiently learnable structure from the available univariate historical predictors, especially for rare and abrupt extreme events.

2.7 Explainability analysis

SHAP analysis was applied to the best-performing forecasting model. SHAP values quantify the marginal contribution of each feature to the model output and allow interpretation of both global feature importance and local prediction behaviour. Where SHAP is used with tree ensembles, it provides a computationally efficient and theoretically grounded explanation framework for nonlinear models (Lundberg et al., 2020).

2.8 Evaluation metrics

Continuous forecasts were evaluated using MAE, RMSE, sMAPE, and R^2. sMAPE was used because conventional MAPE becomes unstable when observed hydroclimatic values are zero or near zero. Extreme-event classifiers were evaluated using precision, recall, F1-score, and ROC-AUC. This metric combination distinguishes average forecasting accuracy from rare-event discrimination.

3. Results And Discussion

3.1 Long-term hydroclimatic distribution

Table 1 summarizes the descriptive statistics of the full 1901–2021 record.

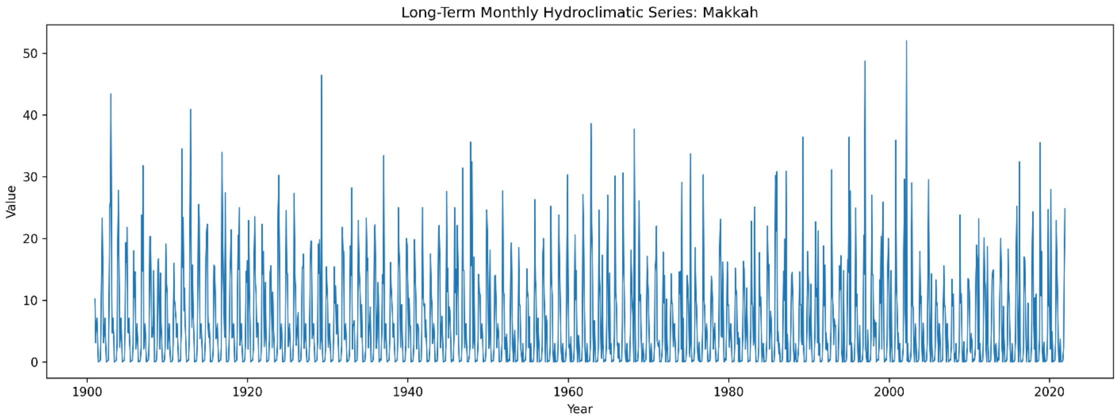
Table 1. Descriptive statistics of monthly hydroclimatic values

Statistic	Value
Count	1,452
Mean	6.188
Standard deviation	8.052
Minimum	0.000
25th percentile	0.200
Median	3.000
75th percentile	9.600
Maximum	52.000

The distribution is strongly right-skewed. The mean exceeds the median by more than twofold, indicating that the long-term average is pulled upward by relatively infrequent high-value months. A total of 174 months, equivalent to approximately 12.0% of the record, have values equal to zero. This confirms that the series combines dry-month intermittency with episodic high events, a structure typical of arid and semi-arid hydroclimatic regimes.

Figure 2 shows the full monthly series from 1901 to 2021. The temporal pattern is dominated by recurring low values interrupted by sharp peaks. The largest observed value is 52.0 in March 2002. Other notable peaks occur in January 1997, April 1929, January 1903, and December 1912. The absence of a visually smooth long-term trend suggests that month-to-month variability and seasonality dominate the record more strongly than monotonic secular change. This does not imply climatic stationarity, but it indicates that any trend analysis would require formal testing rather than visual inference alone.

Figure 2. The full monthly series from 1901 to 2021



3.2 Monthly seasonality and dry-season structure

The seasonal boxplot in Figure 3 reveals a highly structured annual cycle. November, December, and January exhibit the highest median and interquartile values, with November showing the highest monthly mean. June is uniquely dry: all 121 June observations are equal to zero. July, August, and September remain near zero, although they are not entirely dry.

Figure 3. Monthly Seasonality Distribution

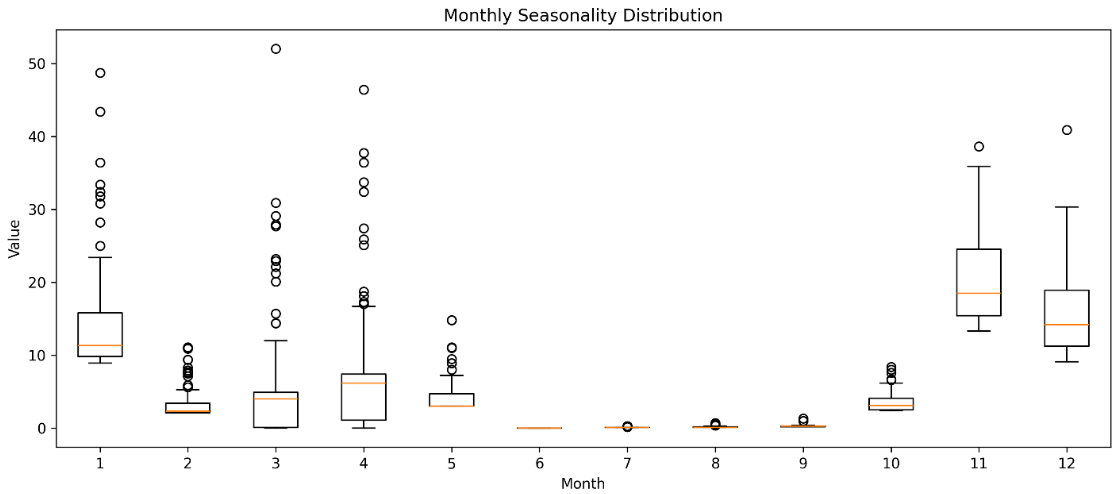


Table 2. Month-wise hydroclimatic summary

Month	Mean	Median	Maximum	Zero months
Jan	14.309	11.3	48.7	0
Feb	3.156	2.3	11.1	0
Mar	5.488	4.0	52.0	30
Apr	7.306	6.2	46.4	23
May	4.174	3.0	14.8	0
Jun	0.000	0.0	0.0	121
Jul	0.113	0.1	0.3	0
Aug	0.188	0.1	0.7	0
Sep	0.288	0.3	1.3	0

Oct	3.554	3.1	8.4	0
Nov	20.212	18.5	38.6	0
Dec	15.470	14.2	40.9	0

The seasonal structure carries two implications. First, the predictive problem is not purely autoregressive; calendar phase contains substantial information. Second, extreme months are not evenly distributed across the year. Threshold exceedances are concentrated in November, December, January, March, and April, while the June–September period contributes almost no extreme events. This explains why seasonal encodings later emerge as dominant predictors in the SHAP analysis.

3.3 Forecasting performance on the holdout period

Table 3 reports forecasting performance for the 286-month holdout period from March 1998 to December 2021.

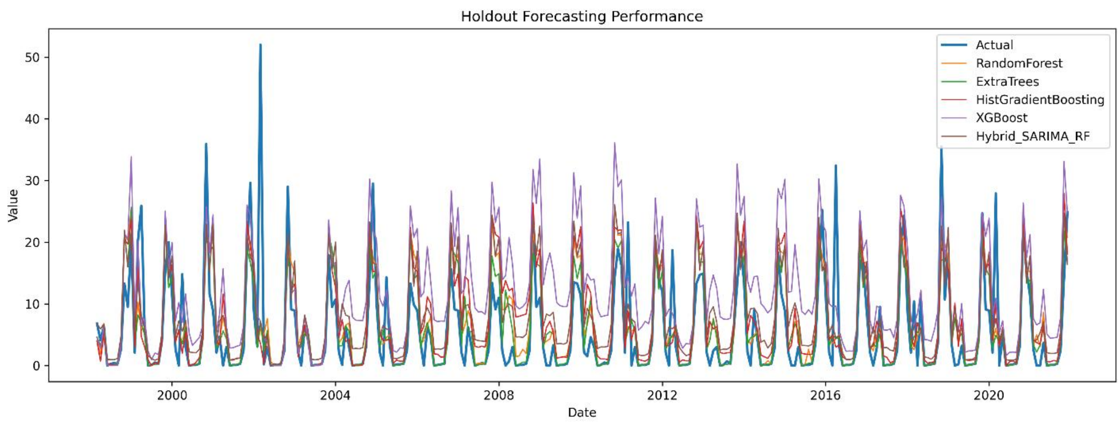
Table 3. Holdout forecasting performance

Model	MAE	RMSE	sMAPE (%)	R^2
Extra Trees	2.997	5.603	57.669	0.518
Random Forest	3.345	5.796	67.833	0.484
Hybrid SARIMA–RF	4.463	6.260	104.671	0.398
Histogram Gradient Boosting	3.981	6.329	95.434	0.385
XGBoost	7.431	9.183	120.815	-0.295

Extra Trees outperforms all other models across every continuous metric. Its RMSE of 5.603 and R^2 of 0.518 indicate moderate explanatory power in a highly intermittent monthly series. Random Forest ranks second, suggesting that bagged or randomized tree ensembles are well matched to the engineered feature structure. Histogram Gradient Boosting and XGBoost perform less favourably, with XGBoost producing negative R^2, implying worse performance than a mean baseline on this holdout period.

Figure 4 illustrates the same result visually. The ensemble models reproduce the seasonal cycle reasonably well, especially during recurring wet-season months, but all models struggle with the most abrupt peaks. The March 2002 event, where the observed value reaches 52.0, remains severely underestimated. This is not merely a model weakness; it reflects the difficulty of predicting rare, synoptic-scale extremes from a univariate monthly historical series without atmospheric covariates.

Figure 4. Holdout Forecasting Performance



The hybrid SARIMA–Random Forest model performs worse than Extra Trees and Random Forest. This outcome is methodologically important. It shows that hybridization does not automatically improve forecasting. In this dataset, the SARIMA component captures broad seasonal structure but appears insufficient for the zero-inflated and sharply peaked distribution. The residual-correction stage cannot fully recover extremes that are not predictable from the available lagged information.

3.4 Extreme-event detection

Using the 90th-percentile training threshold of 17.68, the full record contains 146 extreme months, equivalent to approximately 10.1% of all observations. The largest detected events are shown in Table 4.

Table 4. Highest detected extreme hydroclimatic events

Rank	Date	Value
1	March 2002	52.0
2	January 1997	48.7
3	April 1929	46.4
4	January 1903	43.4
5	December 1912	40.9
6	November 1962	38.6
7	April 1968	37.7
8	January 1995	36.4
9	April 1989	36.4
10	November 2000	35.9

Figure 5 confirms the episodic character of these extremes. The dashed threshold line separates ordinary seasonal variability from unusually high monthly events. Extreme events are not uniformly scattered across the calendar; they cluster strongly in the wet season. Of the 146 extreme months, 69 occur in November, 34 in December, 23 in January, 10 in March, and 10 in April. No extreme events occur from May to October under the selected threshold.

Figure 5. Detected extreme hydroclimatic events

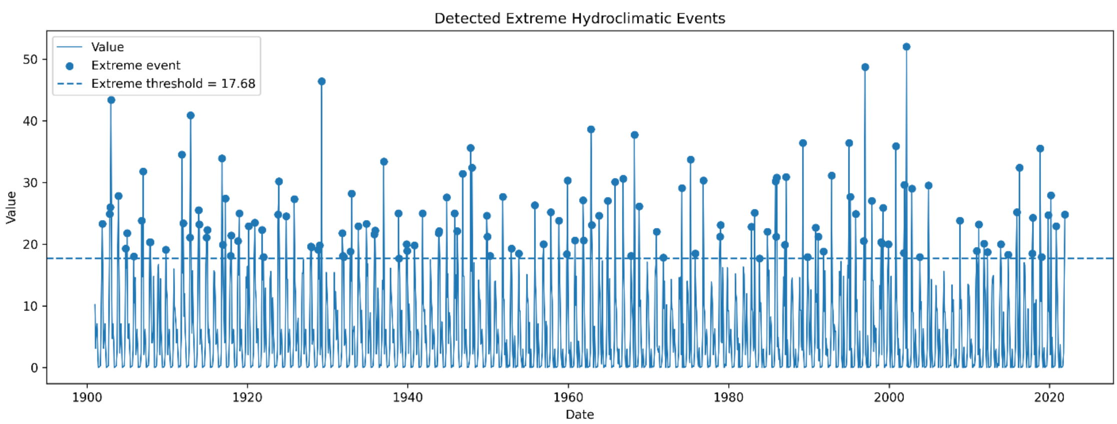


Table 5. Extreme-event classification performance

Model	Threshold	Precision	Recall	F1	ROC-AUC
Histogram Gradient Boosting	17.68	0.440	0.393	0.415	0.858
XGBoost Classifier	17.68	0.364	0.429	0.393	0.847
Random Forest Classifier	17.68	0.471	0.286	0.356	0.868

Table 5 reports the holdout classification performance. The classifiers reveal a familiar rare-event trade-off. Random Forest produces the highest precision and ROC-AUC, suggesting strong ranking ability and fewer false alarms. XGBoost achieves the highest recall, detecting more actual extreme events at the cost of additional false positives. Histogram Gradient Boosting offers the best F1-score, balancing precision and recall more effectively. From a risk-management perspective, the preferred classifier would depend on whether false alarms or missed events are more costly.

3.5 Anomaly detection

The Isolation Forest model identifies 29 anomalous months in the holdout period. These anomalies cluster notably around 2002–2003, immediately following the record maximum in March 2002. The most anomalous month is April 2002, followed by May and June 2002. This pattern indicates that anomaly detection is not simply flagging large values; it is identifying unusual temporal configurations in the feature space. A low or moderate monthly value can be anomalous if it occurs in an unexpected sequence following a highly unusual preceding event.

This distinction matters scientifically. Extreme-event classification is threshold-based and value-centred, whereas anomaly detection is context-sensitive. The two approaches therefore provide complementary information: one identifies high-impact magnitude exceedances, while the other detects unusual temporal behaviour.

3.6 Recursive 12-month future forecast

The best forecasting model, Extra Trees, was used to generate a recursive 12-month forecast for 2022. The predicted values are reported in Table 6 and visualized in Figure 6.

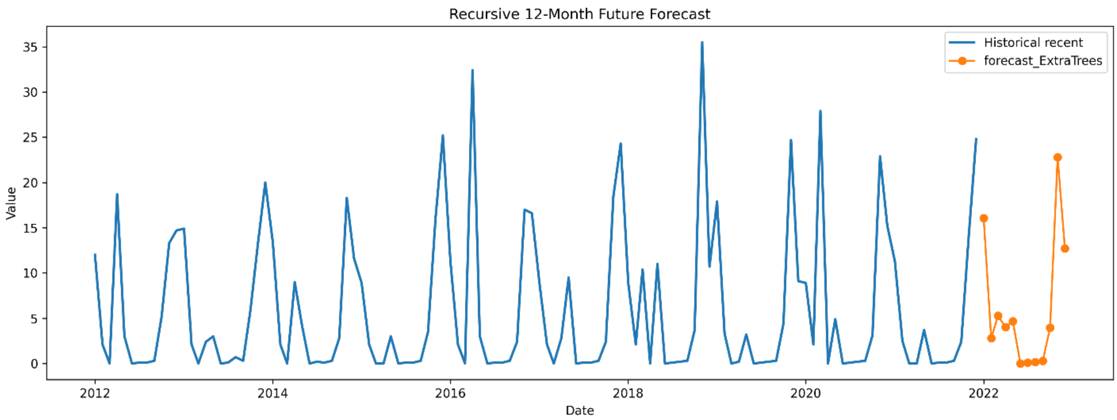


Figure 6. Recursive 12-month future forecast

Table 6. Recursive 12-month forecast for 2022 using Extra Trees

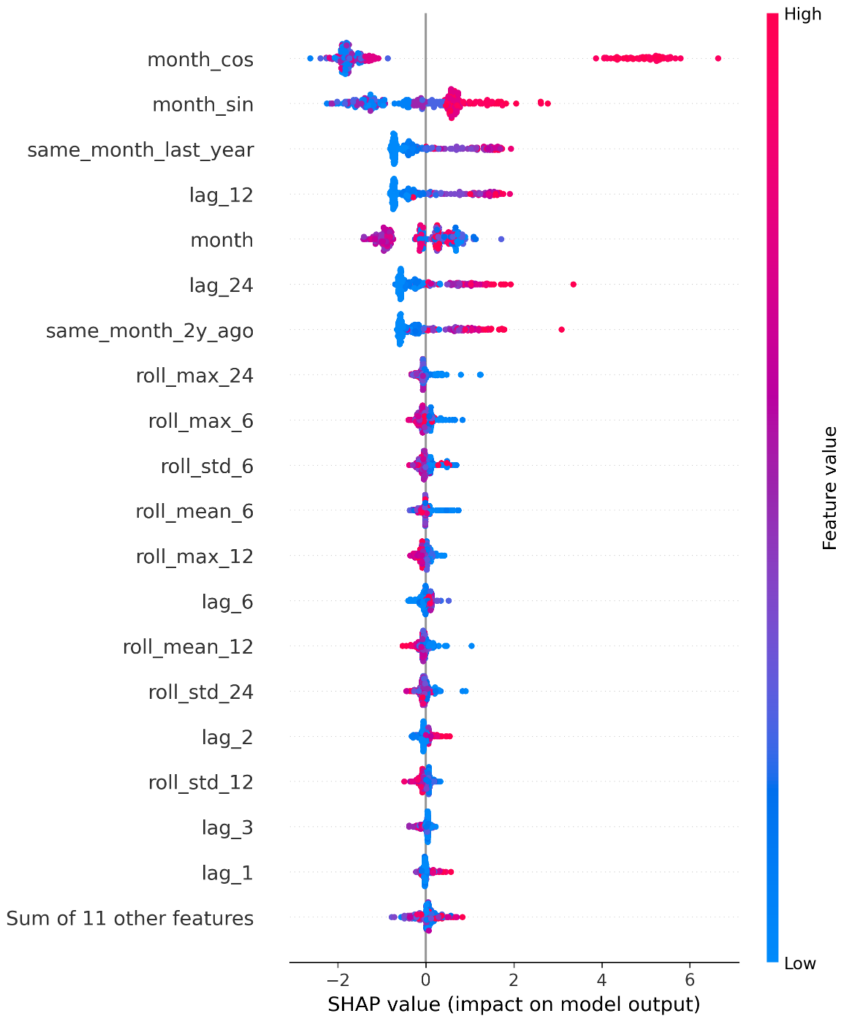
Month	Forecast
Jan 2022	16.069
Feb 2022	2.793
Mar 2022	5.275
Apr 2022	4.009
May 2022	4.660
Jun 2022	0.000
Jul 2022	0.108
Aug 2022	0.150
Sep 2022	0.298
Oct 2022	3.963
Nov 2022	22.782
Dec 2022	12.710

The forecast preserves the historical seasonal regime: near-zero values dominate June–September, moderate values occur in February–May and October, and the wet-season peak appears in November–January. The largest forecasted value occurs in November 2022, followed by January and December. This seasonal coherence supports the SHAP finding that the model relies heavily on calendar phase and annual memory.

3.7 Explainable AI results

Figure 7 presents the SHAP summary plot for the Extra Trees model. The most influential variables are month_cos, month_sin, same_month_last_year, lag_12, month, lag_24, and same_month_2y_ago. This ranking confirms that the model's predictive behaviour is dominated by seasonality and annual recurrence rather than short-term autoregression alone.

Figure 7. The SHAP summary plot for the Extra Trees model



The interpretation is hydrologically plausible. month_cos and month_sin encode the annual cycle continuously, allowing the model to differentiate wet-season and dry-season months without treating month numbers as purely ordinal. The importance of lag_12 and same_month_last_year indicates that the same calendar month in the preceding year provides useful information, while lag_24 and same_month_2y_ago suggest a weaker but still relevant two-year seasonal memory. Shorter lags such as lag_1, lag_2, and lag_3 contribute less, implying that immediate month-to-month persistence is secondary to seasonal positioning.

This explainability result strengthens confidence in the model. The best-performing algorithm is not exploiting arbitrary numerical artefacts; it is learning the dominant seasonal architecture of the Makkah hydroclimatic record.

3.8 Discussions

The empirical results show that Makkah's long-term monthly hydroclimatic behaviour is governed by a strong seasonal regime with pronounced intermittency. The wet-season concentration in November, December, and January, coupled with the consistently dry June, suggests that any model ignoring calendar structure would be misspecified. This finding aligns with regional climatological evidence that rainfall across Saudi Arabia and the Arabian Peninsula is highly seasonal and spatially heterogeneous (Almazroui et al., 2012; Wang et al., 2025).

The dominance of Extra Trees over boosting and hybrid SARIMA–RF has several methodological implications. Randomized tree ensembles are especially effective when predictive information is embedded in nonlinear interactions among seasonal indicators, lagged values, and rolling statistics. Extra Trees introduces additional split randomization, which may reduce variance and improve generalization under a moderately sized feature-engineered dataset. By contrast, XGBoost performed poorly in this empirical configuration, likely because the series combines long dry intervals, abrupt peaks, and limited covariates. This reinforces the forecasting literature's warning that algorithmic sophistication does not guarantee superiority when the data-generating process is sparse, intermittent, or only partially observed (Makridakis et al., 2020; Hewamalage et al., 2021).

The underperformance of the hybrid SARIMA–RF model is equally instructive. Hybrid forecasting is often advocated because statistical models can capture seasonality while machine learning models correct nonlinear residuals. Yet the residual-correction strategy depends on residuals being learnable from available predictors. In this case, the most extreme months appear to reflect event-scale atmospheric drivers not contained in the univariate monthly record. The result does not invalidate hybrid forecasting; rather, it clarifies the boundary conditions under which hybridization is useful. Hybrid models require meaningful residual structure, not simply a statistical baseline plus a machine-learning correction layer.

The SHAP analysis provides a critical safeguard against purely accuracy-driven modelling. It shows that the Extra Trees model bases its predictions primarily on seasonal phase and annual memory. This is consistent with XAI literature, which argues that interpretability can reveal whether a model's internal logic is plausible and trustworthy rather than merely numerically accurate (Arrieta et al., 2020; Lundberg et al., 2020). In the present case, explainability confirms that the model has learned a climatologically meaningful structure: high predictions are associated with wet-season calendar positions and prior annual values, while dry-season months suppress predicted values.

The extreme-event results reveal a different performance landscape from continuous forecasting. While Extra Trees is the best regression model, Histogram Gradient Boosting produces the best F1-score for extreme classification, and Random Forest produces the highest ROC-AUC. This divergence shows that average-value prediction and rare-event detection are not identical tasks. A model optimized for RMSE may reproduce seasonal means but still miss rare extremes. This is particularly important in arid urban environments, where risk is often concentrated in infrequent high-intensity events rather than in the long-term mean.

Finally, the anomaly-detection results broaden the interpretation of risk. The clustering of anomalies around 2002–2003 indicates that unusual temporal sequences may persist beyond the single month of maximum magnitude. This matters for post-event assessment: after an exceptional event, subsequent months may remain anomalous even if their values are not themselves extreme. Integrating threshold-based classification with anomaly detection therefore provides a richer picture of hydroclimatic disturbance.

4. Conclusions

This study develops and evaluates an explainable hybrid machine-learning framework for long-term monthly hydroclimatic forecasting and extreme-event detection in Makkah, Saudi Arabia. The dataset reveals a strongly seasonal, zero-inflated, and right-skewed monthly hydroclimatic regime. Wet-season months, particularly November, December, and January, dominate both central tendency and extreme-event occurrence, while June is consistently dry across the 121-year record.

Among the evaluated models, Extra Trees provides the best continuous forecasting performance, outperforming Random Forest, Histogram Gradient Boosting, XGBoost, and the hybrid SARIMA–Random Forest residual-correction model. This result demonstrates that randomized tree ensembles can effectively exploit engineered seasonal and lag-based predictors in a long monthly arid-region series. However, the limited ability of all models to reproduce the largest peaks indicates that univariate historical data alone cannot fully resolve rare hydroclimatic extremes.

The main intellectual contribution is not simply that one model performs best. Rather, the study demonstrates that a scientifically useful hydroclimatic forecasting framework should integrate prediction, explanation, and event detection. SHAP analysis shows that the model's behaviour is governed by seasonal phase and annual memory, lending interpretive credibility to the forecasting results. Extreme-event classification and anomaly detection add a risk-oriented layer that is not captured by regression metrics alone. In this sense, the study moves beyond accuracy-centred forecasting toward interpretable hydroclimatic intelligence for arid urban risk assessment.

Acknowledgements

The researcher would like to thank the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia for funding this research through the 2026 Fundamental Research Scheme with Grand Number No. 453/LL9/PPIP.PL-MULTI/2026

References

Almazroui, M. (2011). Calibration of TRMM rainfall climatology over Saudi Arabia during 1998–2009. Atmospheric Research, 99(3–4), 400–414. doi:10.1016/j.atmosres.2010.11.006

Almazroui, M., Nazrul Islam, M., Athar, H., Jones, P. D., & Rahman, M. A. (2012). Recent climate change in the Arabian Peninsula: Annual rainfall and temperature analysis of Saudi Arabia for 1978–2009. International Journal of Climatology, 32(6), 953–966. doi:10.1002/joc.3446

Alexander, L. V., Zhang, X., Peterson, T. C., Caesar, J., Gleason, B., Klein Tank, A. M. G., et al. (2006). Global observed changes in daily climate extremes of temperature and precipitation. Journal of Geophysical Research: Atmospheres, 111, D05109. doi:10.1029/2005JD006290

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence. IEEE Access, 6, 52138–52160. doi:10.1109/ACCESS.2018.2870052

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence: Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. doi:10.1016/j.inffus.2019.12.012

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis,

120, 70–83. doi:10.1016/j.csda.2017.11.003
Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. Nature, 619, 533–538. doi:10.1038/s41586-023-06185-3
Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. doi:10.1023/A:1010933404324
Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), Article 15. doi:10.1145/1541880.1541882
De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. Journal of the American Statistical Association, 106(496), 1513–1527. doi:10.1198/jasa.2011.tm09771
Donat, M. G., Lowry, A. L., Alexander, L. V., O'Gorman, P. A., & Maher, N. (2016). More extreme precipitation in the world's dry and wet regions. Nature Climate Change, 6, 508–513. doi:10.1038/nclimate2941
Dueben, P. D., & Bauer, P. (2018). Challenges and design choices for global weather and climate models based on machine learning. Geoscientific Model Development, 11, 3999–4009. doi:10.5194/gmd-11-3999-2018
Fischer, E. M., & Knutti, R. (2016). Observed heavy precipitation increase confirms theory and early models. Nature Climate Change, 6, 986–991. doi:10.1038/nclimate3110
Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. doi:10.1214/aos/1013203451
Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. Hydrology and Earth System Sciences, 25, 2045–2062. doi:10.5194/hess-25-2045-2021
Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63, 3–42. doi:10.1007/s10994-006-6226-1
Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLOS ONE, 11(4), e0152173. doi:10.1371/journal.pone.0152173
Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. Nature, 573, 568–572. doi:10.1038/s41586-019-1559-7
Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. International Journal of Forecasting, 37(1), 388–427. doi:10.1016/j.ijforecast.2020.06.008
Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, 27(3), 1–22. doi:10.18637/jss.v027.i03
Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679–688. doi:10.1016/j.ijforecast.2006.03.001
Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., & Nearing, G. (2022). Uncertainty estimation with deep learning for rainfall–runoff modeling. Hydrology and Earth System Sciences, 26, 1673–1693. doi:10.5194/hess-26-1673-2022
Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using Long Short-Term Memory networks. Hydrology and Earth System Sciences, 22, 6005–6022. doi:10.5194/hess-22-6005-2018
Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. Water Resources Research, 55(12), 11344–11354. doi:10.1029/2019WR026065
Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., et al. (2023). Learning skillful medium-range global weather forecasting. Science, 382, 1416–1421. doi:10.1126/science.adi2336
Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4),

1748–1764. doi:10.1016/j.ijforecast.2021.03.012
Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2, 56–67. doi:10.1038/s42256-019-0138-9
Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting, 36(1), 54–74. doi:10.1016/j.ijforecast.2019.04.014
Min, S.-K., Zhang, X., Zwiers, F. W., & Hegerl, G. C. (2011). Human contribution to more-intense precipitation extremes. Nature, 470, 378–381. doi:10.1038/nature09763
Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. International Journal of Forecasting, 36(1), 86–92. doi:10.1016/j.ijforecast.2019.02.011
Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., & Gupta, H. V. (2021). What role does hydrological science play in the age of machine learning? Water Resources Research, 57(3), e2020WR028091. doi:10.1029/2020WR028091
Pham, L. T., Luo, L., & Finley, A. (2021). Evaluation of random forests for short-term daily streamflow forecasting in rainfall- and snowmelt-driven watersheds. Hydrology and Earth System Sciences, 25, 2997–3015. doi:10.5194/hess-25-2997-2021
Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A benchmark dataset for data-driven weather forecasting. Journal of Advances in Modeling Earth Systems, 12, e2020MS002203. doi:10.1029/2020MS002203
Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., et al. (2021). Skilful precipitation nowcasting using deep generative models of radar. Nature, 597, 672–677. doi:10.1038/s41586-021-03854-z
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. Nature, 566, 195–204. doi:10.1038/s41586-019-0912-1
Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., et al. (2022). Tackling climate change with machine learning. ACM Computing Surveys, 55(2), Article 42. doi:10.1145/3485128
Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting, 36(3), 1181–1191. doi:10.1016/j.ijforecast.2019.07.001
Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resources Research, 54(11), 8558–8593. doi:10.1029/2018WR022643
Slater, L. J., Arnal, L., Boucher, M.-A., Chang, A. Y.-Y., Moulds, S., Murphy, C., Nearing, G., et al. (2023). Hybrid forecasting: Blending climate predictions with AI models. Hydrology and Earth System Sciences, 27, 1865–1889. doi:10.5194/hess-27-1865-2023
Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. International Journal of Forecasting, 36(1), 75–85. doi:10.1016/j.ijforecast.2019.03.017
Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence: Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813. doi:10.1109/TNNLS.2020.3027314
Tripathy, K. P., & Mishra, A. K. (2024). Deep learning in hydrology and water resources disciplines: Concepts, methods, applications, and research directions. Journal of Hydrology, 628, 130458. doi:10.1016/j.jhydrol.2023.130458
Wang, X., Alharbi, R. S., Baez-Villanueva, O. M., Green, A., McCabe, M. F., Wada, Y., Van Dijk, A. I. J. M., Abid, M. A., & Beck, H. E. (2025). Saudi Rainfall (SaRa): Hourly 0.1° gridded rainfall for Saudi Arabia via machine learning fusion of satellite and model data. Hydrology and Earth System Sciences, 29, 4983–5003. doi:10.5194/hess-29-4983-2025
Westra, S., Fowler, H. J., Evans, J. P., Alexander, L. V., Berg, P., Johnson, F., et al. (2014). Future

changes to the intensity and frequency of short-duration extreme rainfall. Reviews of Geophysics, 52, 522–555. doi:10.1002/2014RG000464

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2022). Integrating scientific knowledge with machine learning for engineering and environmental systems. ACM Computing Surveys, 55(4), Article 66. doi:10.1145/3514228

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159–175. doi:10.1016/S0925-2312(01)00702-0

Zittis, G., Almazroui, M., Alpert, P., Ciais, P., Cramer, W., Dahdal, Y., et al. (2022). Climate change and weather extremes in the Eastern Mediterranean and Middle East. Reviews of Geophysics, 60(3), e2021RG000762. doi:10.1029/2021RG000762