Unsupervised Identification of Post-Rainfall Behavior in a Medium Reservoir Using Feature-Engineered Clustering

Sakchan Luangmaneerote
Department of Computer Technology, Faculty of Agriculture and Technology, Rajamangala University of Technology Isan, Surin Campus, Surin, Thailand
tsakchan@hotmail.com

Anyawee Chaiwachirakhampon
Department of Computer Technology, Faculty of Agriculture and Technology, Rajamangala University of Technology Isan, Surin Campus, Surin, Thailand
anyawee_chai@hotmail.com (corresponding author)

Jeeranun Tasuntia
Department of Computer Technology, Faculty of Agriculture and Technology, Rajamangala University of Technology Isan, Surin Campus, Surin, Thailand
jeeranunt.ta@gmail.com

Received: 6 October 2025 | Revised: 31 October 2025 | Accepted: 15 November 2025

 DOI: https://doi.org/10.48084/etasr.15335

ABSTRACT

Understanding reservoir behavior after rainfall is essential for effective water management, especially in medium-sized reservoirs with limited data. This study uses a daily time-series dataset comprising 2,922 days of observations from 2014 to 2023 and proposes a novel unsupervised learning approach to identify operational patterns during post-rainfall periods, which are critical for deciding whether to store or release water. The method integrates rainfall-regime segmentation with feature engineering to emphasize hydrologically significant periods. Real-world data from a medium-sized reservoir were analyzed, focusing on inflow, storage, outflow, and usable water variables. A post-rainfall subset was derived to capture the system's recovery behavior following rainfall. Clustering of these post-rainfall days revealed three distinct operational modes: (1) flood response with high inflow and release, (2) recharge with moderate inflow and retention, and (3) idle or drought with minimal activity. Internal validation demonstrated strong performance (Silhouette coefficient = 0.565, Davies–Bouldin Index (DBI) = 0.550), outperforming traditional approaches. These findings highlight the value of context-aware clustering in revealing interpretable, operation-relevant patterns, providing a practical decision-support tool for medium reservoirs under data limitations.

Keywords-reservoir operation; unsupervised learning; rainfall regime; clustering analysis; feature engineering

I. INTRODUCTION

Reservoirs play a central role in water resource management by balancing water supply, flood control, and ecological sustainability [1]. While large reservoirs are often supported by robust monitoring systems and automated decision tools [2], medium-sized reservoirs, which form the backbone of local and regional water systems, frequently operate with limited instrumentation and imperfect data records [3]. This lack of high-quality data presents a significant challenge for understanding reservoir behavior and optimizing operations, particularly during hydrologically sensitive periods [4].

In recent years, unsupervised learning techniques, particularly clustering algorithms, have been increasingly used in hydrology to detect patterns in drought occurrence, precipitation variability, and reservoir operation regimes. These methods are especially attractive in cases where labeled data are unavailable, allowing for data-driven discovery of hidden structures [5]. However, most existing studies apply clustering to entire datasets without accounting for the hydrological context [6], such as whether the system is responding to rain,

recovering from dry spells, or operating in normal conditions [7]. This "one-size-fits-all" approach risks blending dissimilar operational modes, leading to clusters that are less interpretable or operationally meaningful.

Several studies have emphasized the importance of preprocessing in hydrological modeling, particularly in handling irregular time-series data such as missing rainfall or outflow measurements. Several preprocessing approaches have been explored, including general techniques for improving rainfall forecasts [8], statistical methods for imputing missing streamflow data [9], and machine learning models for imputing hydrological records [10]. Comparable approaches have also been demonstrated in smart irrigation systems, where sensor-based preprocessing and classification techniques improved water management accuracy [11]. For instance, authors in [12] reduced the computational cost of flood simulations by clustering hydrologic model responses in the frequency domain, though the focus remained on ensemble outputs rather than on detailed feature-level transformations. Authors in [13] reviewed core preprocessing methods in water-related machine learning but did not explore how these integrate with domain-specific feature engineering. Similarly, time series alignment techniques have been used to reconcile inter-station lags prior to clustering or similarity analysis [14]. Together, these studies underscore the foundational role of preprocessing while revealing a recurring gap: limited linkage to meaningful, task-specific feature extraction.

Other research has emphasized domain-specific feature engineering to capture temporal and spatial patterns [15, 16]. Techniques such as lag-based inflow features [17], cumulative rainfall [18], and data decomposition [19] have shown promise in representing system dynamics. Authors in [20] combined the Standardized Precipitation Index (SPI) with Principal Component Analysis (PCA) to identify drought zones in Serbia, though their clustering focused on static metrics. Authors in [21] applied multi-timescale SPI clustering to classify drought types in China but used only a single indicator.

In time-series analytics, authors in [22] proposed characteristic-based clustering using global features such as trend, seasonality, and autocorrelation, offering a more informative representation than raw sequences. Authors in [23] employed tensor decomposition and flow-based dissimilarity to reduce simulation loads, although their approach was limited to simplified 2D models. These studies highlight the value of engineered features while underscoring the need for richer, context-aware representations to enhance model generalizability.

Although preprocessing and feature engineering are recognized as essential components for effective clustering, their integration into unified frameworks specific to hydrological applications remains limited. Authors in [24] developed a clustering-based preprocessing framework for subannual hydrological model calibration, showing that structured preprocessing improves interpretability but may still miss complex interactions. Authors in [25] provided a foundational survey of clustering methods, but its scope is general and not tailored to reservoir-specific challenges. More recently, authors in [26] developed GRDL, a machine learning-based global reservoir area-storage-depth dataset, whereas authors in [27] applied a physics-guided neural network to simulate lake temperature profiles, showing how physical constraints can enhance data-driven modeling. Despite these advances, such methods focus on large-scale generalization rather than localized operational interpretability. In contrast, this study presents a rainfall-regime-aware clustering approach that segments reservoir behavior based on engineered features and contextual hydrological rules, offering improved interpretability in data-limited environments.

Accordingly, this study proposes a new approach for analyzing reservoir behavior specifically during post-rainfall periods, a time when reservoir inflows often surge and critical operational decisions must be made. Instead of clustering all days together, the method first segments the dataset into rainfall regimes using a rule-based classification (e.g., Dry, Heavy Rain, Post-Rain). Clustering is then applied exclusively to Post-Rain days, supported by carefully selected features and engineered variables such as cumulative three-day rainfall. This context-aware approach aims to isolate distinct behavioral patterns, such as flood response, passive recharge, or drought conservation modes.

The core contribution of this research lies in demonstrating that even in data-limited, medium-scale reservoirs, meaningful behavioral regimes can be discovered through feature-engineered, regime-based clustering. By improving cluster interpretability and aligning results with real operational scenarios, the approach offers practical value for reservoir managers seeking insight from imperfect datasets.

II. METHOD

This study proposes a context-aware clustering framework that combines rainfall-regime segmentation with feature engineering to identify operational patterns in medium-scale, data-limited reservoirs. The methodology consists of eight key steps, beginning with a description of the study area and data sources, followed by a rigorous preprocessing pipeline to ensure data quality and alignment. A rule-based segmentation approach is then used to classify hydrological regimes, after which clustering is applied specifically to post-rainfall days using task-specific, hydrologically meaningful features. The clustering results are validated both statistically and operationally, and additional sensitivity tests are performed to evaluate the robustness of the method. All procedures are implemented in a reproducible computational environment, with open-source tools and data-sharing provisions to support transparency and replicability.

A. Study Area and Data Sources

This study uses data from the Huai Saneng reservoir, located approximately 7 km southeast of Surin City in northeastern Thailand. Named after the local Khmer word "Saneng", the reservoir was initially developed as part of Surin Province's first irrigation project, launched by the Royal Irrigation Department in 1952 to enhance water retention in the region. Serving as the core of the Surin Irrigation Project since 1984, Huai Saneng features an earthen dam completed in 1978, which stands 20 m high and supports a total storage capacity of 20.8 Mm^3. The reservoir is hydraulically connected to the

nearby Ampil reservoir, forming an integrated water management system that plays a vital role in regional flood control and agricultural water supply. Huai Saneng spans a surface area of 10.4 km^2 and lies at an elevation of 137 m above mean sea level. In addition to its role in irrigation, the reservoir supports fisheries, provides a local water supply, and contributes to tourism development in the area.

This study uses a daily time-series dataset consisting of 2,922 valid records from the Huai Saneng reservoir, covering the period from January 1, 2014, to December 31, 2023. The dataset initially contained 3,652 days, but 730 incomplete or inconsistent records were removed to ensure data quality and feature consistency.

The dataset comprises 13 variables: Date_times (observation date), Capacity (total storage capacity in Mm^3), Minimum_amount_of_RNOL (Minimum Required Normal Operating Level, RNOL), Reservoir_Storage_Volume (daily stored volume), Percent_of_RNOL (percentage of minimum level achieved), inflow_volume and outflow (daily inflow and outflow in Mm^3), cumulative_inflow_volume and cumulative_outflow (running totals), usable_water_volume (available water for use), Rain (daily rainfall in mm), Evaporation (daily evaporation loss), and wind_current (daily wind speed in km/h). All data were obtained from the Royal Irrigation Department and verified for reliability before analysis.

To enhance hydrological interpretation, the derived feature Rain_3day_sum(t) = Rain(t−1) + Rain(t−2) + Rain(t−3) was introduced, excluding the current day (t), to capture the cumulative rainfall effect preceding dry days. This feature highlights post-rainfall conditions influencing reservoir recovery, supporting more meaningful clustering.

B. Preprocessing and Feature Engineering

The Date_times column was converted to a proper datetime format and set as the index to support time-based operations, enabling temporal slicing, rolling calculations, and lag-based feature derivation. To capture hydrological context, a rainfall-regime classification was introduced. A three-day rolling sum of rainfall (Rain_3day_sum) was computed, and each day was categorized into one of four regimes based on rainfall conditions: Dry (no rain today or in the past three days), Heavy Rain (daily rainfall > 10 mm), Light Rain (daily rainfall > 0 mm but ≤ 10 mm), and Post-Rain (no rainfall today but more than 10 mm accumulated in the past three days). Days not fitting any of these rainfall-regime rules were automatically labeled as "Other," representing residual or borderline rainfall cases retained for completeness. These labels were assigned using a rule-based function applied row-by-row.

For clustering, meaningful hydrological features were selected to characterize reservoir behavior, as illustrated in Figure 1, including inflow_volume, Reservoir_Storage_Volume, outflow, and usable_water_volume. These variables represent the input, system state, management response, and available supply, respectively. Together, they encapsulate the cause-state-response dynamics of reservoir operations, allowing the identification of operational regimes even with limited data. Records with missing values were excluded to ensure data integrity, and all numerical features were standardized using z-score normalization prior to clustering.

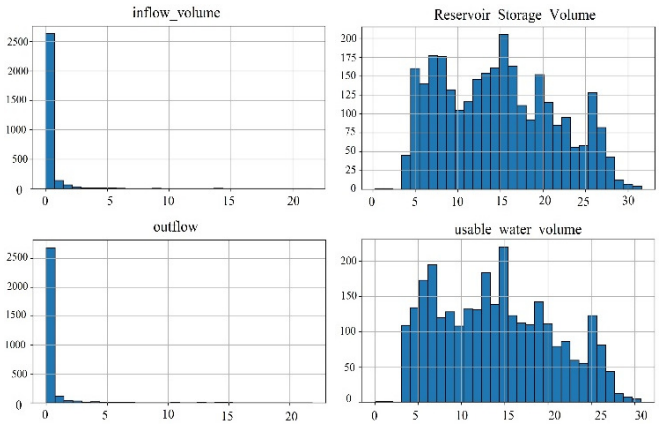

Fig. 1. Histogram of selected features.

Figure 2 shows the frequency of rainfall regimes over the ten-year period (2014–2023), with Dry days dominating the dataset (approximately 1,600 out of 2,922 valid records), followed by Light Rain, Heavy Rain, and Post-Rain, each occurring in roughly 300 instances. Although Post-Rain days account for only about 10% of the total records, they represent the most critical period for reservoir decision-making. By focusing clustering specifically on this subset, the analysis avoids being dominated by the more numerous Dry days, which could dilute meaningful signals. This targeted approach improves cluster interpretability and helps capture key operational patterns.

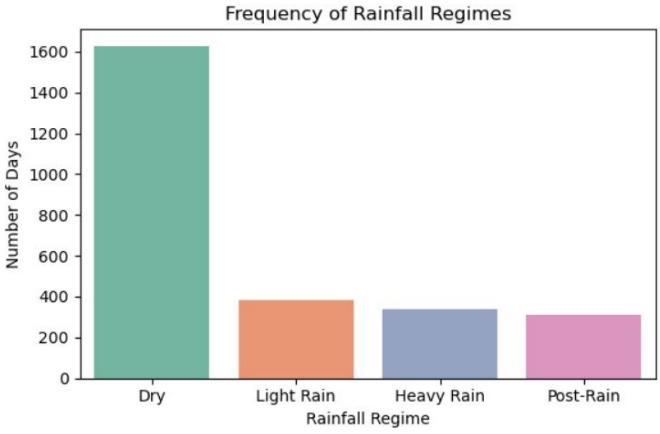

Fig. 2. Frequency of rainfall regimes.

Figure 3 shows the three-day rolling rainfall sum from 2014 to 2023, with each day classified into four rainfall regimes. The plot reveals clear seasonal variability, with pronounced peaks corresponding to the main rainy seasons and extended flat periods indicating prolonged dry spells. Post-Rain days (orange) follow heavy rainfall, marking key inflow response periods and confirming that post-rain events are distinct and operationally significant for reservoir management.

III. RESULTS AND DISCUSSION

A. Regime-Specific Clustering Patterns in Inflow–Storage Space

Figure 4 visualizes K-means [28] clustering results for each rainfall regime in the inflow-versus-storage plane (z-score scale). In the Dry panel, clusters separate almost exclusively by storage, reflecting routine draw-down and refill cycles under negligible inflow. Light Rain days begin to spread along the inflow axis, producing clusters that capture modest runoff and incremental recharge. Under Heavy Rain conditions, inflow dominates: one cluster combines high inflow with high storage (initial flood capture), a second shows high inflow with lower storage (rapid release), and a third contains low-inflow residuals—evidence of diverse operator responses during intense events.

In the Post-Rain panel, extreme inflow points disappear and clusters compact into three clear modes: recharge (moderate inflow, rising storage), idle/drought (minimal inflow, low storage), and active-release (moderate storage with ongoing outflow, not visible in this 2-D slice). The contrast across panels confirms that whole-series clustering would merge hydrologically distinct behaviors, whereas regime-specific clustering yields interpretable, operation-relevant patterns.

B. Evaluation of Clustering Quality Using Internal Metrics

To evaluate the quality of clustering results, this study uses three internal validation metrics: Silhouette score [29], Davies–Bouldin Index (DBI) [30] and Calinski–Harabasz Index (CHI) [31]. These measures assess cluster cohesion, separation, and variance structure. Silhouette indicates how well each point fits within its assigned cluster. DBI reflects the compactness and distinctness of clusters, where lower values indicate better quality. CHI evaluates the ratio of between-cluster to within-cluster dispersion, with higher values representing stronger separation. Together, these metrics provide a robust assessment of clustering performance and support the selection of an optimal number of clusters.

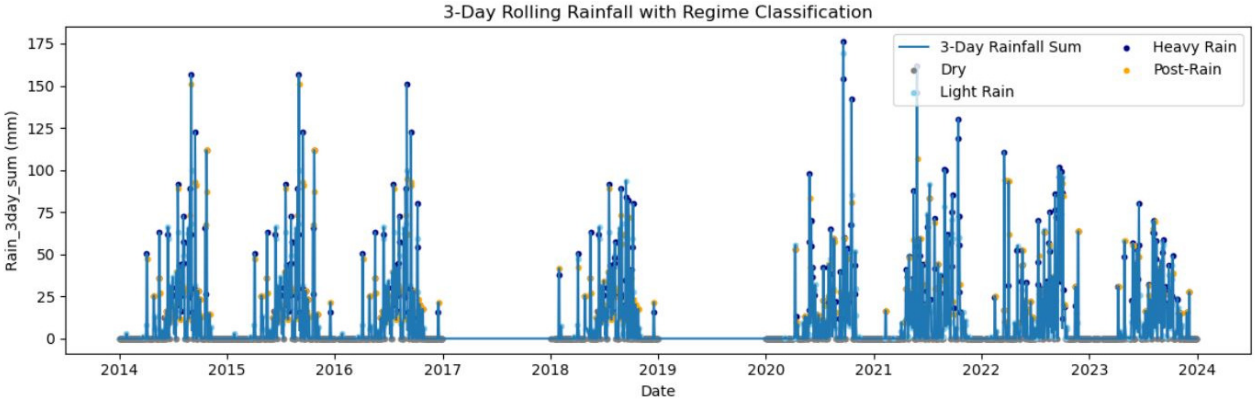

Fig. 3. Three-day rolling rainfall with regime classification.

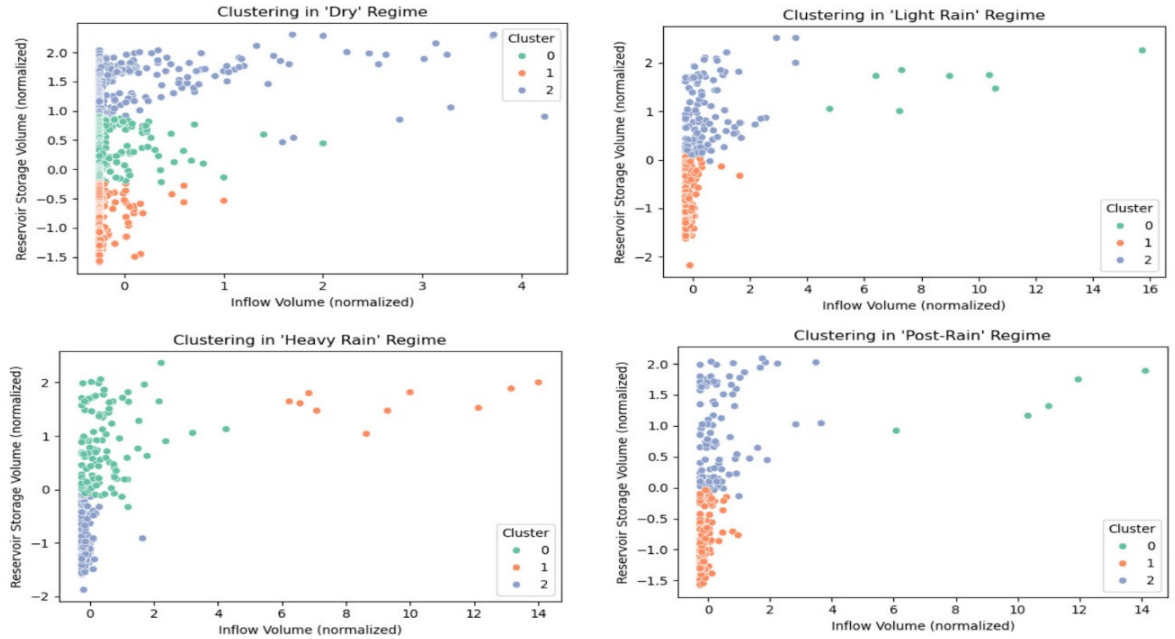

Fig. 4. K-means clustering results across rainfall regimes in the inflow–storage plane (normalized).

Table I compares internal validation metrics between the proposed regime-specific clustering (applied only to post-rainfall days) and a traditional clustering approach (applied to the full dataset without rainfall segmentation).

TABLE I. COMPARISON OF CLUSTERING PERFORMANCE BETWEEN REGIME-SPECIFIC AND TRADITIONAL APPROACHES

Metric	Post-Rain clustering	Traditional clustering
Silhouette score	0.565	0.517
DBI	0.550	0.627
CHI	638.50	3,971.06

The comparison of clustering performance shows that the regime-specific approach, focused solely on post-rainfall days, performs better than traditional all-day clustering in terms of cluster cohesion and separation. The Silhouette score for the post-rainfall model is 0.565, higher than the traditional model's score of 0.517, indicating that the regime-specific clusters are more cohesive and well separated. The DBI is also lower for the post-rain model, at 0.550 compared to 0.627, suggesting more compact and distinct clusters.

Although the CHI is substantially higher for the traditional model, with a value of 3,971.06 compared to 638.50 for the post-rain model, this metric is highly sensitive to sample size. The traditional model uses the full dataset, which likely inflates this score without necessarily indicating better clustering quality. Overall, the results support rainfall-regime segmentation, especially during post-rainfall periods, as a more effective strategy for identifying meaningful and interpretable patterns in reservoir behavior. Typical benchmarks suggest Silhouette > 0.5 and DBI < 1 indicate good cluster cohesion and separation.

Figure 5 illustrates the relationship between the number of clusters (k) and the Silhouette score. While k = 2 yields the highest score (0.90), it may oversimplify reservoir behavior. At k = 3, the score drops to 0.58 but still indicates well-separated clusters, balancing performance and interpretability. Scores decline and plateau beyond k = 3, suggesting limited benefit from added complexity. Thus, k = 3 is optimal, aligning both statistically and with known operational modes: flood response, recharge, and idle. Figure 6 shows the histogram of Silhouette coefficients for the post-rainfall subset, revealing a strong clustering structure. Most values lie between 0.40 and 0.85, with a prominent peak near 0.80, indicating that data points are, on average, much closer to their own centroids than to those of other clusters. Only a few observations fall below 0.20, and negative coefficients are rare, implying minimal overlap or misclassification and well-defined cluster boundaries. The right-skewed shape, characterized by many high values and a thin left tail, suggests tightly packed cluster cores with only a small number of ambiguous edge cases, a pattern consistent with real reservoir behavior. Collectively, these features confirm that the regime-based K-means model with k=3 achieves high cohesion and separation, validating the focus on post-rainfall days; observations with low or negative coefficients warrant closer inspection as potential transitional cases in operational analyses.

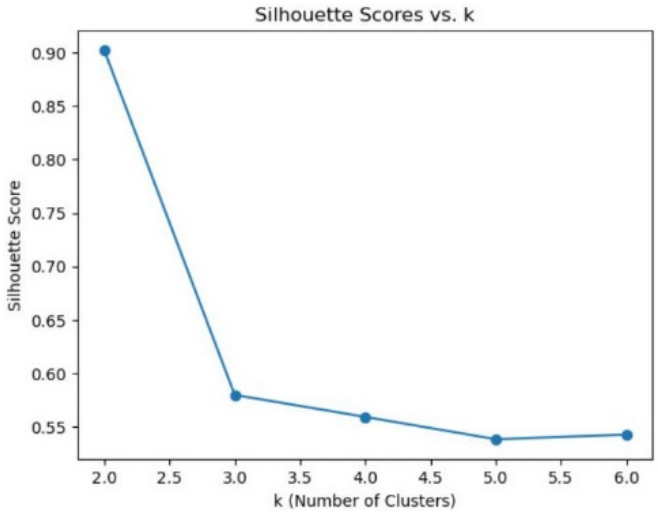

Fig. 5. Silhouette scores for varying cluster counts (k) in post-rainfall clustering.

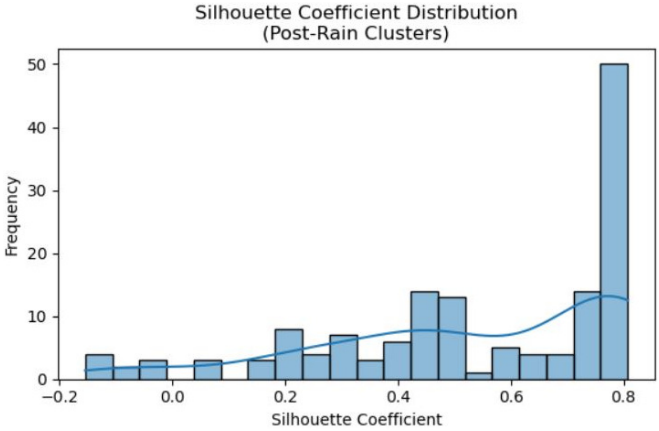

Fig. 6. Silhouette coefficient distribution for Post-Rain clusters.

C. Characteristics of Each Cluster

Table II presents the operational cluster centroids for the Post-Rain regime, derived from the real reservoir dataset. It provides a summary of the mean values (centroids) for inflow, storage, outflow, and usable water volume across each of the three clusters identified during post-rainfall conditions, with all values expressed in Mm³. These centroids reveal distinct operational modes of the reservoir: Cluster 0 corresponds to flood-response conditions, characterized by high inflow and active water release; Cluster 1 represents idle or drought conditions, with minimal inflow and low storage levels; and Cluster 2 indicates recharge behavior with moderate inflow and storage recovery.

TABLE II. OPERATIONAL CLUSTER CENTROIDS FOR POST-RAINFALL REGIME

Regime	Cluster	n_days	Inflow (Mm³)	Storage (Mm³)	Outflow (Mm³)	Usable (Mm³)
Post-Rain	0	5	14.77	24.17	15.46	23.35
Post-Rain	1	189	0.13	8.35	0.06	7.53
Post-Rain	2	106	0.81	20.40	0.60	19.58

In Table III, the rows list every combination of rainfall regime and K-means cluster, reporting (i) cluster size in days, (ii) centroids in z-score space, and (iii) the same centroids back-transformed to the reservoir's native units in Mm^3. Reading across the columns reveals several data-driven insights:

- Dry regime (variance driven by storage, not inflow): All three Dry clusters have near-zero inflow_z and outflow_z, confirming negligible hydrologic forcing. The only separation axis is storage_z, which ranges from −0.80 (low drawdown) to +1.44 (high conservation). In operational terms, the reservoir cycles between low-level drought and high-level refill even when no rainfall occurs. This information can support the refinement of rule curves for dry season operations.
- Light Rain regime (two stable modes plus a rare outlier group): Cluster 1, with 259 days, mirrors the Dry regime's low-storage group. Cluster 2, with 116 days, captures modest recharge, with storage_z around 0.94. Cluster 0 contains only eight days and shows extreme z-scores across all axes. These rare, high-inflow cases may indicate outliers or edge conditions and should be reviewed carefully before being used in decision-making processes.
- Heavy Rain regime (inflow becomes the dominant separator): Cluster 1 exhibits inflow_z around 9.4 and storage_z near 1.6, representing a clear flood-surge response, though this cluster includes only ten days. Cluster 0 reflects a moderate flood-capture scenario, whereas Cluster 2 represents the recession limb where storage has already been drawn down. The presence of both high and low storage conditions within the same rainfall category highlights the variability in operational responses during extreme events.
- Post-Rain regime (three interpretable operational modes): Cluster 0, with five days, retains the tail of the flood pulse, featuring the highest inflow and outflow. Cluster 2, with 106 days, shows moderate inflow and elevated storage, consistent with a recharge strategy. Cluster 1, with 189 days, shows negative z-scores on all variables, indicating an idle or drought condition once the storm event has passed. These three behavioral patterns align closely with typical post-rainfall decisions, such as releasing, storing, or conserving water. This supports the effectiveness of rainfall segmentation and the selection of three clusters.
- Other regime (heterogeneous catch-all class): The Other category includes low-storage idle days and a small high-storage cluster with 49 days. These records are mixed and may reflect incomplete or misclassified rainfall conditions. This regime may need to be redefined or excluded in future iterations to improve clustering clarity.
- Sample-size cues for model trust: Clusters with fewer than ten days, such as Light Rain 0, Heavy Rain 1, and Post-Rain 0, are rare and should be treated cautiously. Their extreme centroids could distort statistical summaries or influence model interpretation if not separated or clearly flagged as exceptional cases.

TABLE III. CLUSTER SUMMARY FOR EACH RAINFALL REGIME

Regime	Cluster	n_days	Behavioral signature (Mm^3)	Interpretation
Dry	0	702	inflow = 0.07, storage = 17	Moderate storage, negligible inflow/outflow; routine conservation during long dry spells
	1	559	inflow = 0.04, storage = 9	Low storage and minimal releases; drought or draw-down phase
	2	384	inflow = 0.42, storage = 24	Above-average storage sustained by small but steady baseflow; passive refill
Light Rain	0	8	inflow = 12.4, storage = 25	Rare, high-runoff bursts after light storms; active release
	1	259	inflow = 0.11, storage = 9	Light showers with little runoff; remains in drought-like state
	2	116	inflow = 0.96, storage = 21	Incremental recharge without the need for release
Other	0	104	inflow = 0.08, storage = 7	Miscellaneous very-low events; essentially idle
	1	49	inflow = 1.29, storage = 25	Short unclassified surges leading to near-full storage
	2	103	inflow = 0.19, storage = 16	Intermediate behavior between idle and recharge
Heavy Rain	0	119	inflow = 0.84, storage = 20	Moderate flood capture with limited release
	1	10	inflow = 13.0, storage = 26	Extreme inflow and high storage; emergency release expected
	2	209	inflow = 0.09, storage = 8	Tail end of storms; storage already drawn down
Post-Rain	0	5	inflow = 14.8, storage = 24	Immediate flood-response: high inflow matched by high outflow
	1	189	inflow = 0.13, storage = 8	Idle/drought mode after rain subsides; reservoir returns to low level
	2	106	inflow = 0.81, storage = 20	Recharge mode: moderate inflow retained to rebuild supply

The centroid patterns confirm that rainfall-segmented clustering captures distinct operational behaviors that would be masked in a full-series clustering approach. These modes can be linked to specific gate operations. For example, rapid releases may be triggered when inflow_z exceeds 9, or recharge strategies may apply when inflow_z is moderate and storage_z remains below 1. At the same time, small clusters should be interpreted as rare conditions, requiring careful treatment in planning and modeling.

D. Discussion of Results

Regime-specific clustering uncovered operational patterns that were hidden in whole-series analysis. In the inflow–storage plane shown in Figure 4, Dry days separated primarily by storage. Light Rain days introduced modest inflow variation, whereas Heavy Rain days produced flood-capture, rapid-release, and recession clusters. Post-Rain days formed three distinct modes: flood response, recharge, and idle. This demonstrates that rainfall segmentation effectively isolates key hydrologic drivers.

Validation supports these findings. The regime-based model achieved a higher Silhouette score and a lower DBI compared to the traditional approach. Its Silhouette coefficient distribution, as illustrated in Figure 6, indicates tight cluster cohesion. Centroid analysis in Table II confirms operational relevance: Cluster 0 corresponds to emergency release, Cluster 2 reflects recharge behavior, and Cluster 1 represents idle or drought conditions. Table III extends this view across all rainfall regimes, revealing that some small but high-inflow clusters mark rare, critical events. In practice, managers can link gate operations to these behavioral modes; however, small clusters should be interpreted cautiously as edge cases. Despite data limitations, the study demonstrates that rainfall-segmented clustering offers statistically robust and operationally meaningful insights.

These results are consistent with existing literature. The finding that segmenting the dataset by rainfall regime prior to clustering significantly improves cluster cohesion and interpretability aligns with authors in [24], who emphasized structured preprocessing for hydrological model interpretability. The methodology of creating a hydrologically relevant feature (Rain_3day_sum) aligns with the principle of characteristic-based clustering proposed by authors in [22], which advocates using meaningful extracted features rather than raw time series. The clear identification of a flood-response cluster corroborates the focus of authors in [2], who developed advanced models for real-time reservoir flood control. Similarly, the isolation of an idle or drought cluster echoes clustering-based drought classification studies by authors in [20] and [21].

IV. CONCLUSION

This research introduced a novel, context-aware approach for identifying reservoir behavior during post-rainfall periods using unsupervised learning. By combining data preprocessing, feature engineering, and rainfall-regime segmentation, the study addressed key limitations in traditional clustering methods, which often analyze all days together without considering hydrological context. Focusing specifically on the post-rainfall regime, a critical yet often overlooked operational period, allowed the identification of three distinct reservoir behavior modes: flood response, recharge holding, and drought idle. Despite relying on imperfect and incomplete data from a medium-scale reservoir, the proposed method achieved strong internal clustering performance, with a Silhouette score of 0.565 and a Davies–Bouldin Index (DBI) of 0.550. These results outperformed traditional full-dataset clustering in terms of both clarity and interpretability. The use of engineered features such as recent rainfall accumulation (Rain_3day_sum), enhanced the model's ability to uncover patterns consistent with real-world operational practices. Overall, the findings show that unsupervised learning can be both practical and effective in data-limited environments when hydrological context is carefully incorporated into the analysis. The proposed framework provides a useful tool for decision-making in medium reservoirs, especially in systems that lack formal operating rules but require data-driven guidance.

ACKNOWLEDGMENT

I sincerely thank the Surin Royal Irrigation Project Office 8, Huai Saneng, and the Royal Irrigation Department for providing essential data. I am also grateful to Boonlueo Nabumroong and the Department of Computer Technology, Faculty of Agriculture and Technology, RMUTI Surin Campus, for their support, as well as to colleagues, family, and friends for their invaluable encouragement.

REFERENCES

[1] J. W. Labadie, "Optimal Operation of Multireservoir Systems: State-of-the-Art Review," *Journal of Water Resources Planning and Management*, vol. 130, no. 2, pp. 93–111, Mar. 2004, https://doi.org/10.1061/(ASCE)0733-9496(2004)130:2(93).

[2] J. Zhang, X. Cai, X. Lei, P. Liu, and H. Wang, "Real-time reservoir flood control operation enhanced by data assimilation," *Journal of Hydrology*, vol. 598, July 2021, Art. no. 126426, https://doi.org/10.1016/j.jhydrol.2021.126426.

[3] D. Li, Y. Chen, L. Lyu, and X. Cai, "Uncovering Historical Reservoir Operation Rules and Patterns: Insights From 452 Large Reservoirs in the Contiguous United States," *Water Resources Research*, vol. 60, no. 8, Aug. 2024, Art. no. e2023WR036686, https://doi.org/10.1029/2023WR036686.

[4] W. W.-G. Yeh, "Reservoir Management and Operations Models: A State-of-the-Art Review," *Water Resources Research*, vol. 21, no. 12, pp. 1797–1818, Dec. 1985, https://doi.org/10.1029/WR021i012p01797.

[5] E. Aytaç, "Unsupervised learning approach in defining the similarity of catchments: Hydrological response unit based k-means clustering, a demonstration on Western Black Sea Region of Turkey," *International Soil and Water Conservation Research*, vol. 8, no. 3, pp. 321–331, Sept. 2020, https://doi.org/10.1016/j.iswcr.2020.05.002.

[6] A. Wunsch, T. Liesch, and S. Broda, "Feature-based Groundwater Hydrograph Clustering Using Unsupervised Self-Organizing Map-Ensembles," *Water Resources Management*, vol. 36, no. 1, pp. 39–54, Jan. 2022, https://doi.org/10.1007/s11269-021-03006-y.

[7] D. Li, G. Sang, W. Liu, Y. Liu, and M. Zhao, "Cluster analysis-based hydrological similarity assessment in small watersheds of Shandong Province's hilly areas," *Hydrology Research*, vol. 56, no. 2, pp. 93–107, Dec. 2024, https://doi.org/10.2166/nh.2024.017.

[8] I. Ebtehaj, H. Bonakdari, M. Zeynoddin, B. Gharabaghi, and A. Azari, "Evaluation of preprocessing techniques for improving the accuracy of stochastic rainfall forecast models," *International Journal of Environmental Science and Technology*, vol. 17, no. 1, pp. 505–524, Jan. 2020, https://doi.org/10.1007/s13762-019-02361-z.

[9] Y. Gao, M. Taie Semiromi, and C. Merz, "Efficacy of statistical algorithms in imputing missing data of streamflow discharge imparted with variegated variances and seasonalities," *Environmental Earth Sciences*, vol. 82, no. 20, Sept. 2023, Art. no. 476, https://doi.org/10.1007/s12665-023-11139-z.

[10] V. Sharma and K. Yuden, "Imputing Missing Data in Hydrology using Machine Learning Models," *International Journal of Engineering Research & Technology*, vol. 10, no. 1, pp. 78–82, Jan. 2021, https://doi.org/10.17577/IJERTV10IS010011.

[11] A. H. Blasi, M. A. Abbadi, and R. Al-Huweimel, "Machine Learning Approach for an Automatic Irrigation System in Southern Jordan Valley," *Engineering, Technology & Applied Science Research*, vol. 11, no. 1, pp. 6609–6613, Feb. 2021, https://doi.org/10.48084/etasr.3944.

[12] A. E. Sikorska-Senoner, "Clustering model responses in the frequency space for improved simulation-based flood risk studies: The role of a cluster number," *Journal of Flood Risk Management*, vol. 15, no. 1, Mar. 2022, Art. no. e12772, https://doi.org/10.1111/jfr3.12772.

[13] F. Ghobadi and D. Kang, "Application of Machine Learning in Water Resources Management: A Systematic Literature Review," *Water*, vol. 15, no. 4, Feb. 2023, Art. no. 620, https://doi.org/10.3390/w15040620.

[14] S. Lee *et al.*, "Clustering of Time Series Water Quality Data Using Dynamic Time Warping: A Case Study from the Bukhan River Water

Quality Monitoring Network," *Water*, vol. 12, no. 9, Sept. 2020, Art. no. 2411, https://doi.org/10.3390/w12092411.

[15] C.-M. Forke and M. Tropmann-Frick, "Feature Engineering Techniques and Spatio-Temporal Data Processing," *Datenbank-Spektrum*, vol. 21, no. 3, pp. 237–244, Nov. 2021, https://doi.org/10.1007/s13222-021-00391-x.

[16] G. Papacharalampous *et al.*, "Global-scale massive feature extraction from monthly hydroclimatic time series: Statistical characterizations, spatial patterns and hydrological similarity," *Science of The Total Environment*, vol. 767, May 2021, Art. no. 144612, https://doi.org/10.1016/j.scitotenv.2020.144612.

[17] A. M. Rushworth, A. W. Bowman, M. J. Brewer, and S. J. Langan, "Distributed Lag Models for Hydrological Data," *Biometrics*, vol. 69, no. 2, pp. 537–544, June 2013, https://doi.org/10.1111/biom.12008.

[18] G. Papacharalampous, H. Tyralis, Y. Markonis, and M. Hanel, "Hydroclimatic time series features at multiple time scales," *Journal of Hydrology*, vol. 618, Mar. 2023, Art. no. 129160, https://doi.org/10.1016/j.jhydrol.2023.129160.

[19] M. Sit, B. Z. Demiray, Z. Xiang, G. J. Ewing, Y. Sermet, and I. Demir, "A comprehensive review of deep learning applications in hydrology and water resources," *Water Science and Technology*, vol. 82, no. 12, pp. 2635–2670, Aug. 2020, https://doi.org/10.2166/wst.2020.369.

[20] S. Shamshirband *et al.*, "Clustering project management for drought regions determination: A case study in Serbia," *Agricultural and Forest Meteorology*, vol. 200, pp. 57–65, Jan. 2015, https://doi.org/10.1016/j.agrformet.2014.09.020.

[21] P. Yang, Z. Xiao, J. Yang, and H. Liu, "Characteristics of clustering extreme drought events in China during 1961–2010," *Acta Meteorologica Sinica*, vol. 27, no. 2, pp. 186–198, Apr. 2013, https://doi.org/10.1007/s13351-013-0204-x.

[22] X. Wang, K. Smith, and R. Hyndman, "Characteristic-Based Clustering for Time Series Data," *Data Mining and Knowledge Discovery*, vol. 13, no. 3, pp. 335–364, Nov. 2006, https://doi.org/10.1007/s10618-005-0039-x.

[23] E. G. D. Barros, F. K. Yap, E. Insuasty, P. M. J. van den Hof, and J. D. Jansen, "Clustering Techniques for Value-of-information Assessment in Closed-loop Reservoir Management," in *ECMOR XV - 15th European Conference on the Mathematics of Oil Recovery*, Amsterdam, Netherlands, 2016, https://doi.org/10.3997/2214-4609.201601858.

[24] T. Lan *et al.*, "A Clustering Preprocessing Framework for the Subannual Calibration of a Hydrological Model Considering Climate-Land Surface Variations," *Water Resources Research*, vol. 54, no. 12, pp. 10034-10052, Dec. 2018, https://doi.org/10.1029/2018WR023160.

[25] P. Berkhin, "A Survey of Clustering Data Mining Techniques," in *Grouping Multidimensional Data: Recent Advances in Clustering*, J. Kogan, C. Nicholas, and M. Teboulle, Eds. Berlin, Heidelberg, Germany: Springer, 2006, pp. 25–71, https://doi.org/10.1007/3-540-28349-8_2.

[26] Z. Hao *et al.*, "GRDL: A New Global Reservoir Area-Storage-Depth Data Set Derived Through Deep Learning-Based Bathymetry Reconstruction," *Water Resources Research*, vol. 60, no. 1, Jan. 2024, Art. no. e2023WR035781, https://doi.org/10.1029/2023WR035781.

[27] X. Jia *et al.*, "Physics-Guided Machine Learning for Scientific Discovery: An Application in Simulating Lake Temperature Profiles," *ACM/IMS Transactions on Data Science*, vol. 2, no. 3, May 2021, Art. no. 20, https://doi.org/10.1145/3447814.

[28] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, Berkeley, CA, USA, 1965, pp. 281–297.

[29] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987, https://doi.org/10.1016/0377-0427(87)90125-7.

[30] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, https://doi.org/10.1109/TPAMI.1979.4766909.

[31] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, Jan. 1974, https://doi.org/10.1080/03610927408827101.