Contents lists available at https://citedness.com/index.php/jdsi

Data Science Insights

Journal Page is available to https://citedness.com/index.php/jdsi

Research article

The Impact of Rainfall Pattern Dataset Construction on Neural Network Performance for Reservoir Water Level Forecasting

Wan Hussain Wan Ishak[1*], *Raja Nurul Mardhiah Raja Mohamad*[2]

[1] *School of Computing, Universiti Utara Malaysia, Sintok, Kedah, Malaysia*
[2] *School of Computing, Universiti Utara Malaysia, Sintok, Kedah, Malaysia*
email: [1] *hussain@uum.edu,my*
** Correspondence*

ARTICLE INFO

Article history:
Received 07 January 2026
Revised 11January, 2026
Accepted 12 January, 2026
Available online 28 February 2026

Keywords:
Reservoir water level forecasting
artificial neural network
rainfall pattern representation
dataset construction
hydrological modelling

Please cite this article in IEEE style as:
W. H. Wan Ishak and R. N. M. Raja Mohamad, "The Impact of Rainfall Pattern Dataset Construction on Neural Network Performance for Reservoir Water Level Forecasting", Data Science Insights, vol. 4, no. 1, pp. 1–6, Feb. 2026.

ABSTRACT

Reservoir water level forecasting is a critical component of effective water resources management, supporting flood mitigation, water supply planning, and sustainable reservoir operation, particularly under increasingly variable rainfall conditions. During periods of heavy rainfall, inaccurate or delayed water level prediction may increase flood risk, while during low rainfall seasons, poor forecasting can compromise water storage and operational efficiency. Artificial Neural Networks (ANNs) have been widely adopted for reservoir water level forecasting due to their capability to model nonlinear rainfall–reservoir relationships. However, existing studies largely focus on algorithm selection or architectural enhancement, with limited attention given to how rainfall data representation and dataset construction influence neural network performance. This study addresses this gap by analysing the impact of rainfall pattern dataset construction on ANN performance for reservoir water level forecasting. The primary aim is to evaluate how different rainfall representations affect predictive accuracy when the learning algorithm and training configuration are held constant. Two rainfall pattern datasets were constructed using the same raw rainfall and reservoir water level data from the Timah Tasoh Reservoir, Malaysia. The first dataset represents a compact abstraction of rainfall behaviour using rainfall change indicators derived from day-to-day observations. The second dataset enriches the feature space by incorporating both rainfall change and rainfall intensity categories for each upstream station. In both datasets, the reservoir water level category serves as the prediction target. Prior to model training, redundancy and conflicting data instances were removed to ensure data consistency. A consistent ANN architecture was employed for both datasets and evaluated using 10-fold cross-validation. Model performance was assessed using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The experimental results demonstrate that the enriched rainfall pattern dataset achieved significantly lower RMSE and MAE values compared to the compact rainfall change dataset, indicating improved learning capability and generalisation performance. Although the enriched dataset required higher computational effort, the improvement in forecasting accuracy was substantial. The findings highlight that dataset construction plays a decisive role in neural-network-based reservoir water level forecasting.

Data Science Insights is an open access under the with CC By 4.0 license.

Correspondence:
Wan Hussain Wan Ishak
School of Computing, Universiti Utara Malaysia, Malaysia

1. Introduction

Reservoirs play a critical role in water resources management by regulating water storage and release to support flood control, water supply, hydropower generation, and environmental sustainability. A key operational variable in reservoir management is the reservoir water level, as it directly reflects the balance between inflow, storage capacity, and release decisions. Accurate monitoring and forecasting of reservoir water levels are therefore essential to ensure safe and efficient operation [1].

During heavy rainfall seasons, reservoirs are required to absorb excess inflow while maintaining sufficient buffer capacity to mitigate downstream flooding [1]. Inadequate anticipation of rapid water level rise may lead to delayed gate operations, increasing flood risk and potential damage to downstream communities. Conversely, during low rainfall or dry seasons, improper release decisions may result in insufficient water storage, affecting

domestic supply, agriculture, and hydropower production. These contrasting operational demands highlight the importance of reliable reservoir water level forecasting under varying rainfall conditions.

Traditional reservoir operation often relies on standard operating procedures and operator experience. However, such approaches may be insufficient when rainfall patterns exhibit high variability, nonlinearity, or abrupt changes. Recent studies have shown that rainfall-driven inflow–storage relationships are complex and difficult to capture using conventional statistical or rule-based methods. Consequently, machine learning and artificial intelligence approaches, particularly Artificial Neural Networks (ANNs), have been increasingly adopted for reservoir water level forecasting due to their ability to learn nonlinear relationships from data [2,3].

While numerous studies report promising forecasting accuracy using ANN and deep learning models [4,5,6], model performance is not determined by algorithm selection alone. An often-overlooked factor is how rainfall information is prepared, represented, and combined into datasets prior to model training. Rainfall data can be represented as raw measurements, categorical intensities, temporal changes, or hybrid combinations, each influencing the learning behaviour of neural networks. Differences in dataset construction may significantly affect prediction accuracy, model stability, and computational efficiency.

Despite growing interest in neural-network-based reservoir forecasting, limited attention has been given to systematic comparison of neural network performance across multiple rainfall pattern datasets derived from the same data source. Existing studies typically evaluate a single dataset or focus on algorithmic comparisons, leaving a gap in understanding how data representation itself influences neural network performance.

To address this gap, this article presents a comparative analysis of neural network performance across three rainfall pattern datasets, constructed using progressively enriched feature combinations. The aim of this study is to analyse how different dataset preparation strategies which are based on rainfall change, reservoir water level category, and rainfall intensity affect neural network learning and forecasting accuracy.

2. Literature Review

Numerous studies have demonstrated the effectiveness of ANN-based models for reservoir water level forecasting. Ünes et al. [7] reported improved prediction accuracy using neural networks compared to conventional models, highlighting their suitability for capturing complex hydrological responses. Subsequent studies further confirmed these findings across different geographical and hydrological contexts. For example, Valizadeh et al. [8] showed that enhancing input pattern representation significantly improved forecasting accuracy for reservoir and river water levels. More recent works have extended ANN architectures by integrating deep learning components, such as convolutional and recurrent layers, to improve temporal feature extraction and long-term dependency modelling [4,5]. In a related comparative study, Aquil and Ishak [9] evaluated several machine learning models for reservoir water level forecasting and demonstrated the overall effectiveness of data-driven approaches in capturing nonlinear reservoir behaviour, although their analysis primarily focused on algorithm comparison rather than dataset representation.

Advances in deep learning have led to the development of hybrid and attention-based models for reservoir water level prediction. Li et al. [10] demonstrated that CNN–LSTM models with improved attention mechanisms can outperform conventional ANN structures by explicitly learning spatial and temporal dependencies. Similarly, Stefenon et al. [6] employed wavelet-based sequence-to-sequence LSTM models to enhance time-series forecasting performance in hydroelectric dam applications. Alongside direct water level forecasting, intelligent systems have also been applied to support reservoir operation decisions. For instance, Mokhtar et al. [11] applied an Adaptive Neuro-Fuzzy Inference System (ANFIS) to model reservoir water release decisions, illustrating the capability of hybrid intelligent models to handle uncertainty and complex operational rules. Although the study focuses on decision modelling rather than forecasting accuracy, it reinforces the value of intelligent techniques in reservoir management.

In parallel, there has been increasing interest in incorporating alternative data sources and representations to improve forecasting accuracy. Jin et al. [12] explored the use of remote sensing data combined with deep learning techniques for reservoir water level estimation, indicating the potential of non-traditional data inputs. Other studies have focused on rainfall representation strategies, including the use of rainfall intensity categories, temporal differences, and hybrid feature sets. Sapitang et al. [2] highlighted that representing rainfall patterns rather than raw rainfall values can improve model robustness and interpretability, particularly in operational decision-support contexts.

Despite these advances, several review studies have noted that neural network performance is highly sensitive to data preparation choices. Azad et al. [3] and Rehamnia and Mahdavi-Meymand [13] emphasised that differences in input feature selection, data preprocessing, and dataset construction can lead to substantial variation in forecasting performance, even when similar machine learning models are used. However, most existing studies focus primarily on comparing different algorithms or model architectures, often treating dataset preparation as a preliminary step rather than a core research variable.

As a result, there remains a clear research gap in systematically analysing how different rainfall pattern datasets influence neural network performance in reservoir water level forecasting. While some studies implicitly modify input features or dataset composition, few explicitly compare multiple datasets derived from the same raw data source using different rainfall and reservoir-related representations. Understanding this relationship is crucial,

as it directly affects prediction accuracy, model stability, and computational efficiency, all of which are critical for practical reservoir operation.

3. Research Methods

The data preparation strategy adopted in this study follows the staged approach reported in earlier work on rainfall-pattern-based reservoir water level forecasting. The central objective is to examine how progressively enriching the dataset with additional contextual information affects neural network learning and prediction accuracy.

a) Data Source

Rainfall and reservoir water level data were collected for Tasik Timah Tasoh, Perlis, Malaysia, covering the period 1999–2012. Rainfall observations were obtained from five upstream stations: Padang Besar, Kaki Bukit, Tasoh, Lubuk Sireh, and Wang Kelian. Reservoir water level records were used to define the prediction target. Rather than relying on absolute rainfall values, the study focuses on rainfall change patterns, which better reflect short-term hydrological dynamics relevant to reservoir inflow and gate operation decisions.

b) Feature Representation

Daily rainfall values at each station were transformed into rainfall change indicators by comparing the current day's rainfall with that of the previous day, where 1 represents an increase, 0 represents no change, and −1 represents a decrease. This transformation captures implicit temporal information, as each data instance reflects short-term rainfall dynamics without explicitly introducing time-lagged variables or sequential modelling. In addition, rainfall magnitude was discretised into five intensity categories, as shown in Table 1, to represent varying levels of rainfall severity. This categorical representation provides additional contextual information on rainfall intensity, complementing the rainfall change indicators and enhancing the descriptive power of the dataset.

Table 1. Rainfall Data Representation

Rainfall (mm)	Category	Nominal Value
0	None	0
1-10	Light	1
11-30	Moderate	2
31-60	Heavy	3
>60	Very Heavy	4

The reservoir water level, which serves as the target variable, was discretised into four operational categories as shown in Table 2. This categorisation aligns the forecasting task with practical reservoir operation thresholds and reflects the decision-making levels commonly used in reservoir management.

Table 2. Reservoir Water Level Representation

Water Level (m)	Category	Nominal Value
≤ 28.0	Normal	0
29.0–29.3	Alert	1
29.4–29.5	Warning	2
≥ 29.6	Danger	3

c) Construction of the Datasets

The datasets were constructed by progressively combining rainfall change information with additional contextual attributes. Prior to model training, all datasets underwent a preprocessing stage that involved removal of redundant records and resolution of conflicting data instances. Redundant instances with identical input features and target values were reduced to a single record, while conflicting instances with identical inputs but different target categories were resolved by retaining the instance with the higher reservoir risk level. The instance counts reported for each dataset therefore represent the final number of records after preprocessing.

All datasets are treated as non-sequential instances by the Artificial Neural Network (ANN); however, temporal behaviour is implicitly encoded through rainfall change variables derived from day-to-day rainfall observations.

Dataset 1: Rainfall Change with Reservoir Water Level as Target

The first dataset was constructed by converting rainfall change into numerical form and directly associating it with the prediction target (Table 3). The input features consist of rainfall change indicators (−1 for decrease, 0 for no change, and 1 for increase) obtained from the five upstream stations. The target variable is the reservoir water level category, classified as Normal (0), Alert (1), Warning (2), or Danger (3). Temporal information is incorporated implicitly through daily rainfall changes, without the inclusion of explicit time steps or lagged variables. After redundancy removal and conflict resolution, this dataset comprises 272 instances and represents the most compact abstraction of rainfall behaviour, focusing solely on the influence of rainfall trends on reservoir water level.

Table 3. Dataset 1 - Examples of Data

Rainfall change indicator					Reservoir Level (Target)
Padang Besar	Kaki Bukit	Tasoh	Lubuk Sireh	Wang Kelian	
1	0	1	0	1	1
0	−1	0	0	−1	0
1	1	0	1	0	2

Dataset 2: Rainfall Change, Rainfall Intensity Category, and Reservoir Water Level

The second dataset was constructed by further enriching the feature set through the integration of rainfall change and rainfall intensity category information. In this dataset, rainfall change was transformed into numerical representation and combined with rainfall intensity categories ranging from None (0) to Very Heavy (4) for the upstream stations (Table 4). The prediction target remains the reservoir water level category. Temporal information is implicitly encoded through rainfall change, without explicit time-series modelling or lagged inputs. After redundancy removal and conflict resolution, this dataset comprises 2047 instances and offers the richest feature representation, preserving the highest variability in rainfall and reservoir conditions and closely reflecting real operational scenarios.

Table 4. Dataset 2 - Examples of Data

Rainfall change indicator					Rainfall Category					Reservoir Level (Target)
Padang Besar	Kaki Bukit	Tasoh	Lubuk Sireh	Wang Kelian	Padang Besar	Kaki Bukit	Tasoh	Lubuk Sireh	Wang Kelian	
1	0	1	0	1	3	1	2	1	3	2
0	−1	0	0	−1	1	0	1	0	1	0
1	1	1	0	1	4	3	3	2	3	3

4. Results and Discussion

The neural network performance comparison demonstrates clear differences between the two datasets, primarily influenced by dataset size while controlling for model configuration (Table 5). All datasets were trained using the same artificial neural network (ANN) architecture and evaluated consistently using 10-fold cross-validation, with a learning rate of 0.3 and momentum of 0.2. This controlled experimental setup ensures that any observed performance variation is attributable to data characteristics rather than differences in model design or training parameters. Dataset 1 consists of 272 instances, whereas Dataset 2 is substantially larger with 2,047 instances, which directly affects both training time and predictive accuracy.

Table 5. Neural Network Experimental Setting and Performance Comparison

Metric	Dataset 1	Dataset 2
Learning rate	0.3	0.3
Momentum	0.2	0.2
Total number of instances	272	2,047
Time taken to build model	0.30 seconds	4.31 seconds
Test option	10-fold CV	10-fold CV
Average predicted value	90.0073	90.5227
Root Mean Square Error (RMSE)	1.0073	0.6448
Mean Absolute Error (MAE)	0.7812	0.5191

In terms of computational cost, Dataset 2 required a longer model construction time (4.31 seconds) compared to Dataset 1 (0.30 seconds). This increase is expected due to the higher volume of training samples processed during each cross-validation fold. Despite the difference in size, both datasets produced very similar average predicted values (approximately 90), indicating that the ANN maintained stable output behaviour across datasets and that scaling the dataset did not distort the prediction range.

Prediction accuracy was evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Dataset 2 achieved lower error values, with an RMSE of 0.6448 and an MAE of 0.5191, compared to Dataset 1,

which recorded an RMSE of 1.0073 and an MAE of 0.7812. These results indicate that the ANN model was able to learn more representative patterns from the larger dataset, resulting in predictions that were closer to the actual values. The reduction in both RMSE and MAE suggests improvements in overall error magnitude as well as average absolute deviation.

The findings from the neural network performance comparison provide clear evidence of the influence of dataset size on predictive accuracy when model configuration and evaluation settings are held constant. In this study, all datasets were trained using the same ANN architecture and evaluated using 10-fold cross-validation with a learning rate of 0.3 and momentum of 0.2. This controlled experimental design ensures that the observed differences in performance are primarily driven by data characteristics rather than variations in network structure or training parameters, which is essential for fair and reliable model comparison. Similar controlled evaluation strategies have been adopted in previous reservoir water level forecasting studies to ensure objective performance assessment across models and datasets [7], [9].

A key outcome of the analysis is the superior predictive performance achieved using the larger dataset. Dataset 2, which contains 2,047 instances, recorded substantially lower RMSE and MAE values compared to Dataset 1 with only 272 instances. The reduction in RMSE indicates improved handling of larger prediction errors, while the lower MAE reflects a smaller average deviation between predicted and actual values. The combined use of RMSE and MAE has been widely adopted in reservoir water level forecasting studies as a complementary error evaluation strategy, as both metrics capture different aspects of prediction accuracy [2], [4], [8]. The improved results for Dataset 2 suggest that the ANN benefited from a richer and more diverse training set, enabling it to learn more representative rainfall–reservoir interaction patterns.

These findings are consistent with prior machine-learning-based reservoir studies, which report that models trained on larger and more informative datasets generally demonstrate improved generalisation capability. Comparative analyses of machine learning models for reservoir water level forecasting have shown that increased data availability supports more stable learning and improved predictive reliability, even when similar model architectures are employed [3], [9], [13]. In the context of deep and neural network models, enriched datasets help stabilise weight updates and reduce sensitivity to noise, leading to more consistent performance across validation folds [4], [5], [10]. The consistent average predicted values observed across both datasets in this study further indicate that the ANN maintained stable output behaviour, while accuracy improvements were achieved through enhanced pattern learning rather than shifts in prediction scale.

The increase in training time observed for Dataset 2 is an expected consequence of its larger size. With more instances processed during each fold of cross-validation, computational cost naturally increases. However, this additional training time represents an acceptable trade-off given the substantial improvement in predictive accuracy. Similar trade-offs between computational cost and forecasting performance have been reported in previous reservoir modelling studies, where higher computational demands are justified by gains in model robustness, accuracy, and operational reliability [5], [6], [10]. From a practical perspective, the results indicate that prioritising richer rainfall pattern representation and larger datasets can significantly enhance ANN-based reservoir water level forecasting, particularly for operational decision-support applications.

5. Conclusion

The findings clearly indicate that dataset size plays a significant role in influencing neural network performance. While the smaller dataset enabled faster model training, the larger dataset produced more accurate predictions, as evidenced by lower RMSE and MAE values. This demonstrates that increased data availability enhances the network's ability to learn representative patterns and generalise more effectively to unseen data. Overall, the results confirm that there is a trade-off between computational cost and predictive accuracy. Although training time increased with dataset size, the improvement in model performance justifies the additional computational effort. These findings are particularly relevant for applications involving complex and variable input patterns, such as rainfall-related or time-series prediction tasks, where sufficient training data is essential for achieving reliable outcomes. Future studies could explore the impact of imbalanced or heterogeneous datasets and assess model robustness using external or real-world validation datasets. These extensions would provide deeper insights into neural network behaviour and support the development of more accurate and scalable predictive models.

Acknowledgment

The authors would like to express their sincere appreciation to the Perlis Department of Irrigation and Drainage for providing the rainfall and reservoir water level data used in this study.

References

[1] W. H. W. Ishak, K. R. Ku-Mahamud, and N. M. Norwawi, "Conceptual model of intelligent decision support system based on naturalistic decision theory for reservoir operation during emergency situation," *IJCEE: International Journal of Civil & Environmental Engineering*, vol. 11, no. 2, pp. 6–11, 2011, ISSN: 2077-1258.

[2] M. Sapitang, W. M. Ridwan, K. F. Kushiar, A. N. Ahmed, and A. El-Shafie, "Machine learning application

in reservoir water level forecasting for sustainable hydropower generation strategy," *Sustainability*, vol. 12, no. 15, Art. no. 6121, 2020, doi: 10.3390/su12156121.

[3] A. S. Azad, N. Islam, M. N. Nabi, H. Khurshid, and M. A. Siddique, "Developments and trends in water level forecasting using machine learning models—A review," *IEEE Access*, vol. 13, pp. 63048–63065, 2025, doi: 10.1109/ACCESS.2025.3557910.

[4] S. C. Ibañez, C. V. G. Dajac, M. P. Liponhay, E. F. T. Legara, J. M. H. Esteban, and C. P. Monterola, "Forecasting reservoir water levels using deep neural networks: A case study of Angat Dam in the Philippines," *Water*, vol. 14, no. 1, Art. no. 34, 2022, doi: 10.3390/w14010034.

[5] H. Li, L. Zhang, Y. Yao, and Y. Zhang, "Prediction of water levels in large reservoirs based on optimization of deep learning algorithms," *Earth Science Informatics*, vol. 18, Art. no. 121, 2025, doi: 10.1007/s12145-024-01670-3.

[6] S. F. Stefenon, L. O. Seman, L. S. Aquino, and L. dos Santos Coelho, "Wavelet-Seq2Seq-LSTM with attention for time series forecasting of dam levels in hydroelectric power plants," *Energy*, vol. 274, Art. no. 127350, 2023, doi: 10.1016/j.energy.2023.127350.

[7] F. Ünes, M. Demirci, and O. Kisi, "Prediction of Millers Ferry Dam reservoir level in the USA using artificial neural networks," *Periodica Polytechnica Civil Engineering*, vol. 59, no. 3, pp. 309–318, 2015.

[8] N. Valizadeh, A. El-Shafie, M. Mirzaei, H. Galavi, M. Mukhlisin, and O. Jaafar, "Accuracy enhancement for forecasting water levels of reservoirs and river streams using a multiple-input-pattern fuzzification approach," *The Scientific World Journal*, vol. 2014, Art. no. 432976, 2014, doi: 10.1155/2014/432976.

[9] M. A. I. Aquil and W. H. W. Ishak, "Comparison of machine learning models in forecasting reservoir water level," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 31, no. 3, pp. 137–144, 2023.

[10] H. Li, L. Zhang, Y. Yao, and Y. Zhang, "Prediction of reservoir water levels via an improved attention mechanism based on CNN–LSTM," *Applied Intelligence*, vol. 55, Art. no. 506, 2025, doi: 10.1007/s10489-025-06393-6.

[11] S. A. Mokhtar, W. H. W. Ishak, and N. M. Norwawi, "Modeling reservoir water release decision using adaptive neuro-fuzzy inference system," *Journal of Information and Communication Technology*, vol. 15, no. 2, pp. 141–152, 2016.

[12] Y. Jin, D. Liu, and J. Huang, "Forecasting of reservoir water level by remote sensing and deep learning," *Research Square*, preprint, Feb. 2024, doi: 10.21203/rs.3.rs-3984208/v1.

[13] I. Rehamnia and A. Mahdavi-Meymand, "Advancing reservoir water level predictions: Evaluating conventional, ensemble, and integrated swarm machine learning approaches," *Water Resources Management*, vol. 39, pp. 779–794, 2025, doi: 10.1007/s11269-024-03990-x.