Flash Flood Forecasting and the Role of Catchment Response Time: Predictions via Deep Learning Models

Thitipoom Chailert*, Mark A. Trigg*, Abdulrahman Altahhan†, Evangelos Pournaras†
*School of Civil Engineering
†School of Computer Science
University of Leeds, United Kingdom

***Abstract*—River level forecasting plays a critical role in water resources management, hydropower operations, and flood risk mitigation. Short-term forecasts are especially critical in the context of flash floods, which can develop rapidly—often within six hours of intense rainfall. Although previous research on data-driven models for river level forecasting has focused on single-step daily or monthly forecasts which typically achieve high accuracy, the influence of rainfall and river level response time has received limited attention. In this study, we evaluate the performance of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models for short-term river level forecasting at 1, 3, and 6 hour lead times using 15 minute resolution data. We also investigate how forecast accuracy varies across three catchments with different response times. Results indicate that both LSTM and GRU outperform the Probability-Distributed Model (PDM) and a naïve baseline in terms of RMSE, NSE, and MAPE. Notably, LSTM demonstrates higher accuracy at shorter lead times (1 hour), while GRU performs better at longer horizons (6 hours). Furthermore, the findings suggest that catchment response time serves as a practical upper limit for effective forecasting, with accuracy decreasing when the lead time exceeds the catchment response time.**

***Index Terms*—River Level Forecasting, Flash Flood Forecasting, LSTM, GRU, Probability-Distributed Model (PDM), Catchment Response Time**

I. Introduction

Flash floods, characterized by rapid increases in river levels and heavy rainfall [1], are among the most hazardous and devastating natural hazards, posing substantial risks to urban infrastructure and human life with over 5,000 fatalities annually [2]. Approximately 85% of floods are flash floods [2]; precise and prompt flash flood forecasting is essential for efficient early warning systems, disaster planning, and mitigation strategies. Flash floods are primarily caused by heavy rainfall, which leads to a rapid rise in river levels. Predicting these events is challenging due to the nonlinear relationship between rainfall and river discharge [3]

River level forecasting uses hydrological, meteorological, and statistical models to estimate future water levels or discharges. Among statistical approaches, the seasonal autoregressive integrated moving average (SARIMA) model has been shown to outperform the autoregressive integrated moving average (ARIMA) model in forecasting annual runoff [4]. The Probability-Distributed Model (PDM) [5], a widely used hydrological rainfall-runoff model by the UK Environment Agency, is another approach. In recent years, data-driven methods have gained popularity due to their ability to capture complex, nonlinear relationships between meteorological variables, hydrological factors, and river discharge levels [6].

Deep learning models such as Long Short-Term Memory (LSTM) [7] and Gated Recurrent Units (GRU) [8] are widely used data-driven approaches. These models are suitable for capturing the temporal dependencies and nonlinear interactions in hydrometeorological data, making them effective in river level forecasting, a specific type of time series forecasting [9], [10]. Studies from [11] and [12] applied LSTM and GRU models for daily or monthly forecasting, which is useful for long-term water resource management. However, these timescales are too coarse for flash flood prediction. In contrast, our study applies LSTM and GRU models at finer temporal resolutions, specifically at minute- and hourly-level forecasting, which are appropriate for capturing the rapid dynamics of flash floods. To the best of our knowledge, a performance comparison between PDM and deep learning models is still missing in the literature. This study aims to evaluate and compare their effectiveness for short-term river level prediction.

Catchment response time, which is the duration between rainfall onset and the resulting rise in river levels at the catchment outlet [13], is another important factor in river level forecasting, particularly for flash flood prediction. Fast-responding catchments are typically characterized by steep slopes, impermeable surfaces, and shorter flow paths, while slow-responding catchments tend to have flatter terrain, permeable soils, and larger drainage areas. Wang et al. [14] found that prediction accuracy of multi-step forecast models was lower in a fast-responding watershed compared to two

ISBN 978-3-903176-72-0 © 2025 IFIP

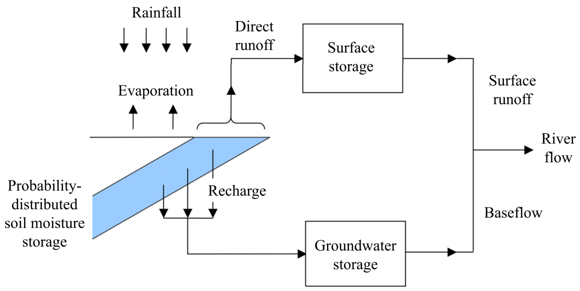

Fig. 1. Conceptual diagram of the Probability-Distributed Model. Adapted from [5]

slower basins, supporting our hypothesis that catchment response time influences the performance of forecasting models.

This study addresses two key areas: comparing the accuracy of LSTM, GRU, and PDM for short lead-time river level forecasting in different catchments, and assessing the impact of different catchment response time on single-value forecasting models.

II. Background

Flash floods typically occur within six hours of heavy rainfall [1], though the timing varies depending on catchment characteristics, with shorter response times in small and steep basins. The interval between peak rainfall and peak river level, termed the catchment response time [15], is a critical factor in flash flood dynamics. For forecasting models, two key challenges arise: accurately predicting river levels and accounting for the impact of catchment response time on forecast reliability.

PDM is a conceptual hydrological rainfall-runoff model that simulates river flow based on rainfall and evaporation [5]. PDM considers that different parts of the basin have varying capacities to store rainwater by using a probability distribution to describe this variability. The soil moisture storage is updated continuously, and runoff is generated when storage thresholds are exceeded. The PDM concept diagram illustrates in Fig. 1. More details can be found in [5]. As PDM has been adopted for flood forecasting in the UK, it is an established benchmark model.

Several studies have applied LSTM and GRU models to river level forecasting. For example, Li et al. [16] developed a hybrid LSTM model for one-hour-ahead streamflow forecasting using rainfall and streamflow as input data, achieving high accuracy with a single-step structure. Similarly, Farfán-Durán and Cea [17] employed LSTM and GRU with hourly input data, finding GRU robust but noting declining accuracy at longer lead times, especially in smaller basins. Xu et al. [18] combined particle swarm optimization with LSTM for rainfall-runoff simulation across two catchments, reporting strong performance but not considering the impact of catchment response time in their study. Wang et al. [14] applied Encoder-Decoder, feedback and attention mechanism in LSTM for multiple-step hourly streamflow prediction in fast-flowing watershed, using future precipitation and forecast streamflow as additional inputs. Their results show the prediction accuracy of the fast-flowing watershed lower than the other two study watersheds at the same lead time. These studies mention the impact of catchment characteristics on forecasting performance, which aligns with our hypothesis that the catchment response time impacts the performance of the forecasting models. However, the influence of catchment response time on model accuracy across lead times has received limited attention.

In this study, we run a PDM simulation in our study area and compare its performance with LSTM and GRU models. To clarify the role of catchment response time in forecast reliability, we evaluate how single-value, time-specific predictions perform across catchments with varying response times. Importantly, we use only historical data as inputs to isolate each model ability to learn past hydrological patterns without the influence of future rainfall or forecast river levels.

III. Methodology

A. Study Area

The River Calder, located in West Yorkshire, Northern England, extends approximately 72 kilometers from its source and flows through several catchments. The river has a long history of flooding, with notable events in 2000, 2012, 2015, and 2020 [19], making the area susceptible to inundation. This study focuses on three catchments within the River Calder: Walsden, Elland, and Wakefield [Table I]. As shown in Fig. 2, Walsden is in the upper reaches of a steep, high-altitude valley and is the smallest catchment. Elland, located midstream, serves as a transition zone between mountainous and lowland areas. Wakefield, at the downstream end, is the largest catchment and lies at the lowest elevation.

There are nine rain measurement stations in the study area; however, each rainfall station corresponds to a specific river level station based on its location (Fig. 2). Table I presents the Standard Annual Average Rainfall (SAAR) [20] which describes the total annual rainfall during 1961-1990. Walsden is likely to receive the highest yearly rainfall (1434 millimeters or mm) among the other two catchments. This is another indicator that shows high flood risk in Walsden area.

Flood threshold levels (Table I) for each catchment are defined by the Environment Agency. Threshold-1 indicates river levels above the normal range and may trigger flood warnings. Threshold-2 denotes a higher risk of flooding, where inundation of low-lying areas becomes likely, potentially resulting in flood alerts.

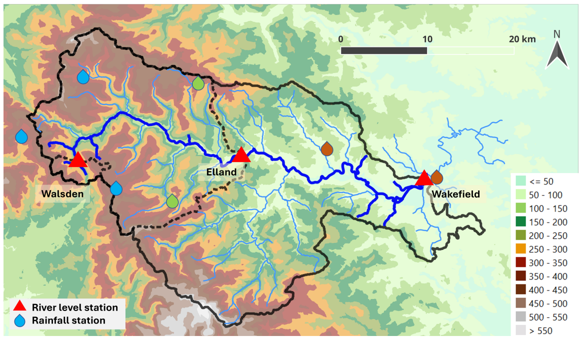

Fig. 2. Map of the three study catchments—Walsden, Elland, and Wakefield (left to right)—with river level stations (red triangles), rainfall stations (colored teardrops, one color per catchment), and elevation (metre). Nested catchment boundaries of each catchment are outlined in black. Data source: [20]

Despite the distinct physical and hydrological characteristics of each catchment, their shared position within the same river system allows for a fair comparison of river level forecasting performance and the influence of catchment-specific factors under similar hydrological conditions.

TABLE I
CATCHMENT DETAILS [20]

Locations	Size (km2)	Altitude (m)	SAAR (mm)	Rainfall stations	Threshold-1 (m)	Threshold-2 (m)
Walsden	13.6	319	1434	3	0.38	0.60
Elland	340.8	292	1257	2	1.35	1.80
Wakefield	841.8	225	1073	2	0.85	1.20

B. Dataset

The hydrological data collected from the three catchments includes rainfall and river levels [20]. The number of rainfall and river level measurement stations is shown in Table I and Fig. 2.

The dataset covers a 10-year period, from October 1, 2006, to September 30, 2016, with measurements recorded at 15-minute intervals, resulting in 350,688 data entries. The data characteristics and statistical values are presented in Table II.

TABLE II
STATISTICAL VALUES OF THE RAINFALL AND RIVER LEVEL.

Stat values	Rainfall			River level		
	Walsden	Elland	Wakefield	Walsden	Elland	Wakefield
Data entries	350,688					
Q80	0.0	0.0	0.0	0.131	0.626	0.178
Q95	0.2	0.2	0.2	0.220	0.812	0.340
Min	0.0	0.0	0.0	0.110	0.244	0.010
Mean	0.039	0.036	0.020	0.091	0.536	0.124
Max	8.6	11.6	7.0	0.750	2.758	1.615
STD	0.149	0.151	0.108	0.066	0.148	0.119

An analysis of the dataset reveals three key characteristics:

- Zero rainfall values dominate the dataset, as indicated by the 80th quantile (Q80), where all rainfall data entries across the three catchments are zero.
- Low river levels are prevalent, with 95th quantile (Q95) river levels remaining below the first threshold level in all three catchments.
- River levels exceeding the first threshold—indicating potential flood conditions—occur in only 0.49% of entries for Walsden (1,724 out of 350,688), 0.17% for Elland (598), and 0.27% for Wakefield (961). The frequency drops further for the second threshold: 0.03% for both Walsden and Elland, and 0.05% for Wakefield. These results highlight the rarity of flood-related data.

These characteristics highlight two potential challenges when applying statistical evaluation metrics to the forecasting models:

- The abundance of zero rainfall values skews the model's performance assessment.
- As floods are rare events, the limited number of river level observations that exceed the threshold may not be adequately captured in the model evaluation results.

C. Deep learning models

LSTM and GRU models were configured for single-value forecasting at T+4 (1 hour), T+12 (3 hours), and T+24 (6 hours) lead times [Fig. 3], using 48 previous steps (12 hours) as input. The input data consists of historical rainfall and river levels. Data preprocessing included removing missing values and outliers, followed by data normalization using Z-score. The dataset was chronologically split into training set (70%), validation set (20%), and testing set (10%).

For hyperparameter optimization, we first identified the optimal configuration using Walsden catchment data, then applied the same parameters to Elland and Wakefield. This approach was justified because all three catchments belong to the same River Calder system, sharing similar hydrological characteristics, and Walsden is the fastest response time, making it the most challenging catchment in this study. This methodology also enables evaluation of LSTM and GRU transferability across catchments. The optimal hyperparameters were determined as 64 neurons, 480 batch size, Adam optimizer with 0.0001 learning rate, and MSE loss function.

D. Catchment response time

According to McCuen [21], catchment response time is commonly defined by two concepts:

- The time from the end of rainfall excess to the inflection point on the falling limb of the hydrograph.

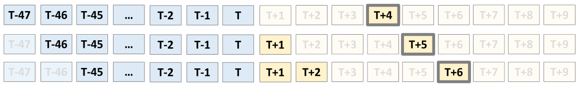
T-47 T-46 T-45 ... T-2 T-1 T T+1 T+2 T+3 T+4 T+5 T+6 T+7 T+8 T+9
T-47 T-46 T-45 ... T-2 T-1 T T+1 T+2 T+3 T+4 T+5 T+6 T+7 T+8 T+9
T-47 T-46 T-45 ... T-2 T-1 T T+1 T+2 T+3 T+4 T+5 T+6 T+7 T+8 T+9

Fig. 3. The diagram shows single-value forecasting using a fixed sliding window. In the T+4 (1 hour) example, the model uses 48 past steps (T-47 to T) to predict the value at T+4. The window then shifts forward for the next prediction. The same approach applies to T+12 (3 hours) and T+24 (6 hours) forecasts.

- The time from the center of mass of rainfall excess to the center of mass of direct runoff (also known as time lag).

Giani, Rico-Ramirez, and Woods [22] applied the first concept from McCuen and proposed the Detrending Moving-average Cross-correlation Analysis (DMCA). The DMCA provides an average estimate of catchment response time from rainfall-runoff time series containing multiple events. This technique produces results similar to the traditional method based on the Flood Estimation Handbook [23]. In this study, we use DMCA to identify catchment response times.

E. Model Benchmark

We run a PDM simulation for Walsden using the same parameters as the Environment Agency. At Elland, PDM is site-calibrated and used for forecasting. However, due to limited flow data and unavailable parameters, PDM could not be applied at Wakefield. Instead, a naïve forecast—using the most recent observed river level as the prediction—is used as a baseline. Beck, Dovern, and Vogl [24] show that naïve forecasts can outperform complex models like LSTM and ARIMA in highly volatile datasets. However, they also note that machine learning models perform well when data contains meaningful, learnable patterns. In our study, if LSTM or GRU models do not surpass the naïve forecast, it may suggest both models are not capturing relevant patterns in the rainfall and river level data.

F. Evaluation of Model Performance

The results are evaluated using several statistical metrics. Root Mean Squared Error (RMSE (1), ideal 0), assesses model accuracy. Nash-Sutcliffe Efficiency (NSE (2), ideal 1), evaluates the model's consistency across different catchments and time periods. Mean Absolute Percentage Error (MAPE (3), ideal 0), quantifies forecast error as a percentage of actual values. Finally, the cross-correlation function (CCF) is used to assess time lags between forecasted and observed values.

$$\mathrm{RMSE}(y,\hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i-\hat{y}_i)^2}{N}} \tag{1}$$

$$\mathrm{NSE}(y,\hat{y}) = 1 - \frac{\sum_{i=0}^{N-1}(y_i-\hat{y}_i)^2}{\sum_{i=0}^{N-1}(y_i-mean(y))^2} \tag{2}$$

$$\mathrm{MAPE}(y,\hat{y}) = \frac{100\%}{N}\sum_{i=0}^{N-1}\frac{\mid y_i-\hat{y}_i\mid}{\mid y_i\mid} \tag{3}$$

where y_i is the actual value for the i^{th} observation, $\hat{y}_i$ is the forecast value for the i^{th} observation, and N is the number of observations.

This study focuses on short-term river level forecasting for flash flood prediction, with particular emphasis on threshold-exceeding levels that indicate flood risk. Evaluation metrics were applied to two data categories which are the entire testing dataset, and the subset of the testing dataset containing only river level is above each catchment first threshold. This selective analysis minimizes the influence of nonflood conditions, specifically revealing model performance during high-flow events. We refer to as the threshold-based subset data.

IV. Results

The results are divided into two subsections: river level forecasting and catchment response time. The first subsection evaluates the forecasting model performance at different lead times. The second subsection focuses on catchment response time, presenting findings derived using the DMCA technique.

A. River level forecasting

LSTM and GRU models were configured to forecast river levels at three lead times: T+4 (1 hour), T+12 (3 hours), and T+24 (6 hours), using the previous 48 time steps (12 hours) as input. Forecast performance was evaluated using RMSE, NSE, and MAPE, with results visualized in Fig. 4. The figure shows model performance on both the full test dataset (34,850 entries; blue triangles) and a threshold-based subset where river levels exceeded the first threshold (orange squares) in three catchments. The subset contains 383 entries for Walsden, 218 for Elland, and 289 for Wakefield.

Across all models, RMSE and MAPE increased while NSE decreased with longer forecast lead times. Performance on the full dataset generally remained closer to ideal metric values compared to the threshold-based subset. These differences became more pronounced at the longest forecast window (T+24). When comparing threshold-based results using the dimensionless metrics NSE and MAPE, Walsden exhibited the lowest NSE values but the highest MAPE among the three catchments.

For visualization purposes, only the results for the shortest (T+4, 1 hour) and longest (T+24, 6 hours) forecast lead times are shown in the river level time series plot [Fig. 5]. The plot covers a 24-hour period from December 12–13, 2015, during which river levels exceeded the second flood threshold. Observed river levels (blue markers) are plotted alongside the model forecasts and flood threshold levels (green and red vertical lines; see the study area subsection in the Methodology for details).

At the T+4 (1 hour) forecast window [Fig. 5], both LSTM (yellow) and GRU (green) produce forecasts that closely follow the observed river levels, particularly near the peak values in Walsden and Elland. In contrast, PDM (red) underperforms in these cases and is unavailable for Wakefield. At the longer lead time (T+24, 6 hours), the gap between observed and forecast river levels increases across all models, with the most pronounced errors occurring in Walsden. However, LSTM and GRU predict higher water levels than PDM.

To assess the timing differences between observed and forecasted river levels, the cross-correlation function (CCF) method was applied. The results are presented in Table III, which shows that most models exhibit delayed predictions across all lead times, except for PDM at T+4 (1 hour) in the Elland catchment, which forecasts 1.25 hours earlier than the observed peak. The time difference generally increases with longer forecast lead times. For example, at the shortest lead time (T+4), LSTM and GRU show a time lag of 0.5 hours in Walsden, and 0.25 hours in Wakefield. However, at the longest lead time (T+24, 6 hours), their time differences increase to 3.75 hours in Walsden, while remaining relatively smaller in Wakefield at 0.75 hours for LSTM and 1.50 hours for GRU.

B. Catchment response times

Table IV presents the response times for three catchments, as determined using the DMCA method. A range of time windows was tested to ensure that the DMCA captured all possible catchment response times.

The DMCA results show that Walsden (1 hour) has the fastest response time, followed by Elland (4.5 hours) and Wakefield (6.75 hours).

V. Discussion

A. Forecasting results between using full test data and threshold-based subset data

Forecast performance metrics indicate higher RMSE and MAPE values, along with lower NSE scores, when evaluated on the threshold-based subset compared to the full test dataset [Fig. 4]. This discrepancy is attributed to the data distribution as the 95th percentile river levels (Q95; Table II) remain below the flood threshold in all catchments, with only 1.09% (Walsden), 0.62% (Elland), and 0.82% (Wakefield) of observations exceeding the first threshold. Consequently, flood conditions are rare within the dataset. These findings highlight that model evaluation using the full dataset may overestimate forecasting skill during flood events. Therefore, we advocate for threshold-based evaluation, as implemented in this study, to more rigorously assess model performance under high river level conditions.

B. LSTM, GRU, PDM, and naive performance comparison

This study evaluates LSTM and GRU models for river level forecasting, with PDM and Naïve forecasts as baselines. Using the same input data, both deep learning models outperform the baselines across all lead times, especially at the longest lead time (T+24, 6 hours; Fig. 4). For example, in Walsden, PDM shows the highest MAPE at 52%, compared to 38% for LSTM and GRU. Similarly, NSE scores are around -8 for LSTM and GRU, versus -13 for PDM. These results support the conclusion that LSTM and GRU offer better performance.

On the threshold-based subset [Fig. 4], LSTM outperforms GRU at short lead times (T+4) with lower MAPE (3.6–6.1% vs. 4.4–6.4%) and comparable NSE (0.6–0.9). At longer lead times (T+24), GRU performs slightly better, with MAPE of 16–38% (LSTM 18–38%) and NSE of –0.1 to –7.9 (LSTM –0.3 to –8.1). This suggests LSTM is more effective short-term, while GRU may suit longer forecasts, though further validation is recommended.

A limitation of using PDM in this study is the need for recalibration for each catchment, which involves numerous parameters and requires hydrological knowledge. Additionally, PDM relies on both flow and river level data, along with a rating curve (an equation to convert between level and flow). Without one of these inputs, the model cannot produce reasonable results. In contrast, the LSTM and GRU models demonstrated flexibility in handling input data while maintaining strong performance across catchments even using identical hyperparameters. Notably, both deep learning models consistently outperformed PDM despite this setting, suggesting their potential for transferable flood forecasting applications.

C. Catchment response time and model forecasting structures

The time difference between observed and forecast river levels, commonly referred to as time lag, is a critical factor in river level forecasting. The time lags could reduce the usefulness of forecasts in operational settings, where timely flood warnings are critical. Ideally, forecasts should align closely with observations across all lead times.

As shown in Table III, time lag increases with forecast lead time and is notably influenced by catchment response time. Despite using identical LSTM and GRU architectures and hyperparameters, differences in timing accuracy are observed across catchments. For instance, at T+24 (6 hours), Walsden, with a response time of 1 hour based on DMCA analysis, exhibits a 3.75-hour lag, whereas Wakefield, with a 6.75-hour response time, shows lags of only 0.75 (LSTM) and 1.50 (GRU)

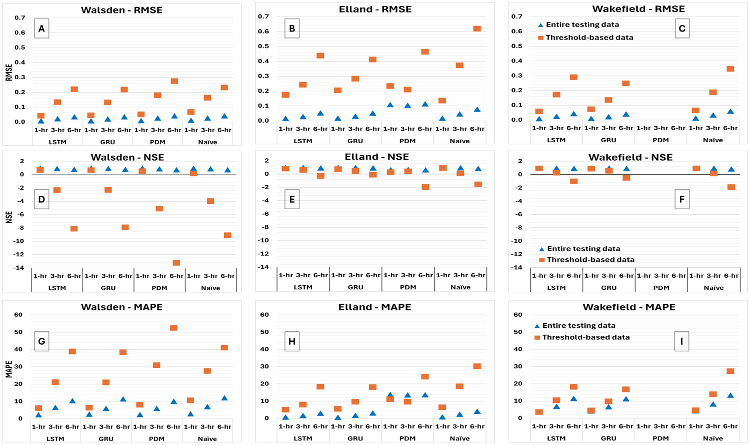

Fig. 4. Forecasting performance evaluation for the three catchments. Each catchment is shown in a vertical column: Walsden (left), Elland (middle), and Wakefield (right). Rows represent evaluation metrics — RMSE on the top (A–C), NSE in the middle (D–F), and MAPE on the bottom (G–I).

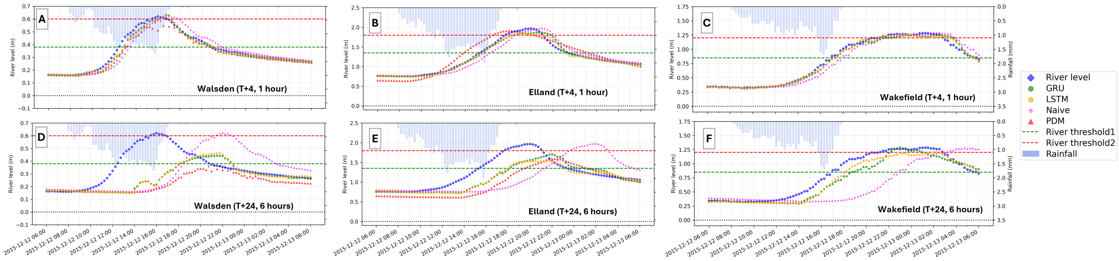

Fig. 5. The example event from December 12-13, 2015, is shown for three catchments (from left to right: Walsden, Elland, and Wakefield). Plots A–C on the top display T+4 (1 hour) forecasts, while plots D–F on the bottom show T+24 (6 hours) forecasts. Forecast results from the LSTM, GRU, PDM, and Naïve models are represented.

TABLE III
TIME DIFFERENCES (HOURS) BETWEEN OBSERVED AND FORECAST RIVER LEVELS: POSITIVE VALUES INDICATE DELAYED FORECASTS, NEGATIVE VALUES INDICATE EARLY FORECASTS.

Models	**T+4 (1 hour) forecast**				**T+12 (3 hours) forecast**				**T+24 (6 hours) forecast**			
	LSTM	**GRU**	**PDM**	**Naïve**	**LSTM**	**GRU**	**PDM**	**Naïve**	**LSTM**	**GRU**	**PDM**	**Naïve**
Walsden	0.50	0.50	0.50		1.75	1.75	1.75		3.75	3.75	4.00	
Elland	0.50	0.50	-1.25	1.00	0.50	0.75	0.00	3.00	1.00	1.25	2.75	6.00
Wakefield	0.25	0.25	N/A		0.75	1.00	N/A		0.75	1.50	N/A	

TABLE IV
THE CATCHMENT RESPONSE TIMES OF THREE AREAS USING THE DMCA TECHNIQUE.

Maximum moving average test (hours)	**DMCA results (Catchment response time, hours)**		
	Walsden	**Elland**	**Wakefield**
2	0.75	0.75	0.75
6	**1.00**	2.75	2.75
12		**4.50**	5.75
24			**6.75**
48			

hours [Tables III, IV]. These results suggest that fast-responding catchments are more challenging for data-driven models, likely because the models struggle to capture rapid hydrological changes when using only historical inputs, and because their structural configurations may operate on timescales slower than the catchment response time. As illustrated in Fig. 5, alignment between observed and forecast values is better at shorter lead times (e.g., T+4) but deteriorates at longer horizons, especially in catchments with fast response times like Walsden.

These findings highlight a key limitation in the generalizability of deep learning models for river level forecasting. Identical LSTM and GRU configurations yielded different performance across catchments, with faster-responding basins exhibiting larger time lags and reduced forecast accuracy. This underscores the importance of incorporating hydrological response characteristics into model design to improve timing reliability in different catchment response times.

VI. Conclusion

This study achieved two key objectives: comparing the performance of deep learning (LSTM, GRU) versus rainfall-runoff (PDM) models for river level forecasting, and evaluating how catchment response time influences single-value forecast accuracy. The results show that both LSTM and GRU models outperform PDM and Naïve benchmarks across all three catchments, as evidenced by superior RMSE, NSE, and MAPE metrics at multiple lead times. We also observed the constraints of the PDM approach, particularly its strict input data requirements and need for individual catchment calibration.

Our results demonstrate that both catchment response time and model structure influence forecasting performance in single-value prediction models. A finding reveals that when forecast lead times exceed a catchment response time, a time lag emerges between observed and predicted values, leading to reduced accuracy. Furthermore, our threshold-based evaluation approach offers a more precise assessment of flood risk, as it focuses on river levels that have the potential to cause flooding.

Lastly, our observations suggest that LSTM performs better at short lead times (1 hour), while GRU is more effective at longer lead times (6 hours). As a future direction, further experiments should be conducted to validate this finding.

Acknowledgment

Evangelos Pournaras supported by a UKRI Future Leaders Fellowship (MR-/W009560-/1): '*Digitally Assisted Collective Governance of Smart City Commons–ARTIO*' and the European Union, under the Grant Agreement GA101081953 attributed to the project H2OforAll—*Innovative Integrated Tools and Technologies to Protect and Treat Drinking Water from Disinfection Byproducts (DBPs)* . Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. Funding for the work carried out by UK beneficiaries has been provided by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10043071].

References

[1] L. Marchi, M. Borga, E. Preciso, and E. Gaume, "Characterisation of selected extreme flash floods in europe and implications for flood risk management," *Journal of Hydrology*, vol. 394, no. 1-2, pp. 118–133, 2010.

[2] WMO, *Sustainability Strategy for the Flash Flood Guidance System with Global Coverage*. World Meteorological Organization, 2023. [Accessed 8-04-2025].

[3] P. C. Young and K. J. Beven, "Data-based mechanistic modelling and the rainfall-flow non-linearity," *Environmetrics*, vol. 5, no. 3, pp. 335–363, 1994.

[4] M. Valipour, "Long-term runoff study using sarima and arima models in the united states," *Meteorological Applications*, vol. 22, no. 3, 2015.

[5] R. Moore, "The pdm rainfall-runoff model," *Hydrology and Earth System Sciences*, vol. 11, no. 1, pp. 483–499, 2007.

[6] M. N. A. Zakaria, M. A. Malek, M. Zolkepli, and A. N. Ahmed, "Application of artificial intelligence algorithms for hourly river level forecast: A case study of muda river, malaysia," *Alexandria Engineering Journal*, vol. 60, no. 4, pp. 4015–4028, 2021.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[9] F. Kratzert, D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing, "Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets," *Hydrology and Earth System Sciences*, vol. 23, no. 12, pp. 5089–5110, 2019.

[10] Y. Man, Q. Yang, J. Shao, G. Wang, L. Bai, and Y. Xue, "Enhanced lstm model for daily runoff prediction in the upper huai river basin, china," *Engineering*, vol. 24, pp. 229–238, 2023.

[11] S. Zhu, X. Luo, X. Yuan, and Z. Xu, "An improved long short-term memory network for streamflow forecasting in the upper yangtze river," *Stochastic Environmental Research and Risk Assessment*, vol. 34, pp. 1313–1329, 2020.

[12] A. M. Ahmed, R. C. Deo, Q. Feng, A. Ghahramani, N. Raj, Z. Yin, and L. Yang, "Deep learning hybrid model with boruta-random forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity," *Journal of Hydrology*, vol. 599, 2021.

[13] K. McGuire, J. J. McDonnell, M. Weiler, C. Kendall, B. McGlynn, J. Welker, and J. Seibert, "The role of topography on catchment-scale water residence time," *Water Resources Research*, vol. 41, no. 5, 2005.

[14] F. Wang, J. Mu, C. Zhang, W. Wang, W. Bi, W. Lin, and D. Zhang, "Deep learning model for real-time flood forecasting in fast-flowing watershed," *Journal of Flood Risk Management*, vol. 18, no. 1, p. e70036, 2025.

[15] O. J. Gericke and J. C. Smithers, "Review of methods used to estimate catchment response time for the purpose of peak discharge estimation," *Hydrological sciences journal*, vol. 59, no. 11, pp. 1935–1971, 2014.

[16] Y. Lin, D. Wang, G. Wang, J. Qiu, K. Long, Y. Du, H. Xie, Z. Wei, W. Shangguan, and Y. Dai, "A hybrid deep learning algorithm and its application to streamflow prediction," *Journal of Hydrology*, vol. 601, p. 126636, 2021.

[17] J. F. Farfán-Durán and L. Cea, "Streamflow forecasting with deep learning models: A side-by-side comparison in northwest spain," *Earth Science Informatics*, pp. 1–27, 2024.

[18] Y. Xu, C. Hu, Q. Wu, S. Jian, Z. Li, Y. Chen, G. Zhang, Z. Zhang, and S. Wang, "Research on particle swarm optimization in lstm neural networks for rainfall-runoff simulation," *Journal of hydrology*, vol. 608, p. 127553, 2022.

[19] . Calder Rivers Trust, "River history." https://calderandcolneriverstrust.org/site/river-history/. [Accessed 31-03-2025].

[20] . Data Services Platform, "Catchment data explorer." https://environment.data.gov.uk/. [Accessed 31-03-2025].

[21] R. H. McCuen, "Uncertainty analyses of watershed time parameters," *Journal of Hydrologic Engineering*, vol. 14, no. 5, pp. 490–498, 2009.

[22] G. Giani, M. A. Rico-Ramirez, and R. A. Woods, "A practical, objective, and robust technique to directly estimate catchment response time," *Water Resources Research*, vol. 57, no. 2, p. e2020WR028201, 2021.

[23] . Institute of Hydrology, *Flood Estimation Handbook (five volumes)*. Centre for Ecology Hydrology, 1999. [Accessed 8-05-2025].

[24] N. Beck, J. Dovern, and S. Vogl, "Mind the naive forecast! a rigorous evaluation of forecasting models for time series with low predictability," *Applied Intelligence*, vol. 55, no. 6, p. 395, 2025.