Evaluating River-Level Rate of Change as a Training-Data Selection for LSTM Flash Flood Forecasting in UK Catchments

Thitipoom Chailert[1], Mark A. Trigg[1], Abdulrahman Altahhan[2], Evangelos Pournaras[2]

[1] School of Civil Engineering, University of Leeds, Leeds LS2 9JT, UK
[2] School of Computing, University of Leeds, Leeds LS2 9JT, UK

Corresponding author: cntch@leeds.ac.uk

Abstract.
Flash floods are challenging to forecast with data-driven models, as their rarity in hydrological records produces severe class imbalance during training. To mitigate this problem, we propose an event-based training data selection method that uses the rate of change (RoC) of river levels as a filtering criterion and assess how the choice of RoC threshold affects forecasting performance across three UK catchments. For each catchment, a cascaded LSTM architecture, which is a rainfall model whose forecasts feed into a river-level model, is trained on event datasets filtered at a range of RoC thresholds and evaluated against the 20 highest-magnitude testing events per site. The results reveal a trade-off in the use of RoC thresholds. Higher thresholds, which filter the training data toward rapid-rise events, improve the forecasting of high-magnitude events at the two smaller, faster-responding catchments, where the median RMSE improves from 0.18 m at the optimum low-magnitude RoC threshold to 0.05 – 0.09 m at the optimum high-magnitude threshold. Lower thresholds produce more accurate forecasts of low-magnitude events at all three sites; however, no single threshold is optimal across all ranges of event magnitudes. Cross-catchment comparison shows that the optimal threshold depends on both catchment characteristics and the target event magnitude. The improvement in high-magnitude forecasting at higher thresholds is substantial in the smaller, flashier catchments but weak in the larger, slower-responding areas. Higher RoC-based event filtering is therefore most effective for high-magnitude forecasting in small, fast-responding catchments.

Keywords: Deep learning, Flash flood forecast, rate of river-level change, Long Short-Term Memory (LSTM), FEH Catchment descriptors, Imbalanced data.

1 Introduction

Floods are the most frequent natural disasters globally, resulting in devastating economic losses and impacts on human life [1, 2]. Among these, flash floods are particularly dangerous due to their rapid onset and fast-moving water. Flash floods are characterized by a short lag time, often fewer than six to twelve hours between the initial rainfall and the peak discharge [3]. Although factors such as reservoir failures, rapid snow melt, and ice jams can cause flash floods, they often arise from prolonged high-intensity rainfall in small, steep and impervious catchments [3]. The sudden nature of flash floods creates a short window time for flood warning and emergency response, making accurate and timely forecasting a critical and challenging task.

Deep learning has gained popularity recently in flood forecasting due to its ability to extract complex non-linear patterns from raw data [4]. Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) designed to capture long-range dependencies in sequential time-series data [5]. A recent review by Santos et al. [6] highlighted that LSTM is the most used deep learning method for flash floods forecast models, accounting for 60% of the surveyed literature. However, the study also noted that no single approach consistently outperforms others because the model performance is influenced by the length of the forecast horizon, the quality of the data, and the regional hydrological characteristics.

In typical hydrological time series, extreme events, such as flash floods, represent only a negligible fraction of the total data, whereas the majority of observations consist of typical river levels and no-rain period conditions. Consequently, the application of deep learning to flash flood forecasting is challenging due to a severe class imbalance. LSTMs as data-driven architectures rely on the quality and composition of the training data. While precipitation and flow have been identified as the most used variables for flash flood forecast [7], models trained on full time-series data are prone to predictive bias because hydrological data are dominated by low-flow observations [8]. This training approach often fails to capture the non-linear dynamics of rare, high-magnitude events, leading to the underestimation of peak levels, as the model learns that the river level is most likely to remain low [9, 10].

Despite these challenges, many studies in the flood forecast continue to utilize unfiltered datasets to train their flood forecast models [11-13]. Akinsoji et. al. [14] conducted a review of data strategies that have been applied in flood forecast using machine learning. They confirmed the importance of data preprocessing for developing flood forecasting models. Among the various techniques that have been studied, feature engineering is widely used. However, adapting a feature engineering technique to select flood relevant data in hydrological data setting has not been so far explored.

Several recent studies have utilized specific discharge return intervals or thresholds for flood event selection, yet they did not explore the impact of alternative thresholds on model performance. For instance, Zhao et al. [15] used a 1 in 2-year discharge return interval to identify 161 events for a deep learning probabilistic flash flood warning model. While their model outperformed a deterministic hydrological model in forecasting flash flood occurrence probabilities, the effect of varying the return intervals remains unexamined. Similarly, Heidari et al. [7] utilized a 1-year return period threshold but did not test alternative thresholds against models trained with full time-series data. Other approaches, such as Chen et al. [16], filtered hourly and daily data by a discharge threshold and demonstrated reduced errors in peak time forecasts. However, their study did not explore the performance of the model to alternative thresholds or compare the results against models trained with the full time series. Although flood events have been filtered from continuous data and used in flood forecasting models, a fair comparison of how different flood thresholds impact overall model performance is still missing.

When selecting flood events, researchers have so far used a return period of discharge or flow as flood thresholds [7, 15-18]; however, measured river flow or discharge data and the required accurate rating curve equation for a specific location are often unavailable [19], especially for small catchment areas. This scarcity of hydrological information therefore creates a limitation for flood event selection. Using river levels which are easier to measure and obtain [6], can be a potential solution. River level data are typically recorded at fixed time-steps, capturing the rise and fall of water levels over time. From these measurements, the rate of change of river level can be calculated, describing how rapidly the level changes over a given interval. RoC reflects the dynamic response of the river making it an alternative criterion for flood event selection. However, the optimum RoC threshold for this purpose has yet to be explored.

This study addresses this gap by investigating the impact of different event-based datasets on the performance of LSTM-based river level forecasting models. We propose a framework for event selection using a range of selected rate of change of river level thresholds and perform a comparative analysis to determine how these thresholds influence model robustness. The predictive accuracy of each configuration is evaluated using standard hydrological metrics, including Root Mean Square

Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Nash-Sutcliffe Efficiency (NSE).

2 Study area

2.1 Location and Data

Three catchments located in different parts of the United Kingdom (UK) (Figure 1.A) were selected for this study which are Walsden, Boscastle, and Hawnby (Figure 1.B-D). All three are in active flash flood areas, but each has different physical characteristics that allow an inter-comparison of model performance across diverse hydrological conditions. A summary of their flood estimation handbook (FEH) catchment descriptors [20] is given in Table 1.

Walsden (Figure 1.B) is the headwater of the River Calder in West Yorkshire. The catchment is small (AREA 13.60 km², Table 1) and steep (DPSBAR 158.4 m/km), with the highest mean altitude of the three sites (ALTBAR 319 m). It is also the wettest, receiving 1434 mm of standard-period average annual rainfall (SAAR) and with soil that is wet for 57% of the time (PROPWET 0.57). Combined with a moderate baseflow index (BFIHOST19 0.39), the highest standard percentage runoff of the three sites (SPRHOST 44.09%) and the shortest drainage path (LDP 6.44 km, DPLBAR 2.8 km) among the three areas, these conditions make the area prone to rapid runoff and flash flooding. Consistent with its small size, steep gradient and high runoff potential, the catchment response time is short, estimated at 1 hour using Detrending Moving-Average Cross-Correlation Analysis (DMCA) [21]. According to Calder Rivers Trust [22], notable flash flood events occurred in 2012, 2013, and 2015.

Boscastle (Figure 1.C) is a village port on the north coast of Cornwall, situated on a steep slope (ALTBAR 199 m) at the confluence of three rivers which are the Valency, Jordan, and Paradise. The catchment is small and steep (AREA 19.99 km², DPSBAR 124.3 m/km), but differs from Walsden in having more permeable soils, reflected in the highest baseflow index (BFIHOST19 0.50) and the lowest standard percentage runoff (SPRHOST 34.48%) of the three. The very small mean floodplain depth (FPDBAR 0.12 m) and the short drainage path (LDP 8.44 km, DPLBAR 4.62 km) indicate limited natural floodplain storage, so when rainfall is intense the steep channel network can flow rapidly at the village outlet. The DMCA-derived catchment response time is 2 hours [21], slightly longer than Walsden owing to the more permeable soils and longer drainage path, but still indicative of a rapidly responding catchment. Boscastle experienced severe flash floods in August 2004 and June 2007 [23 - 25].

Hawnby (Figure 1.D) is a small crossroads village in the North York Moors National Park, North Yorkshire, with the River Rye flowing through it. In contrast to the other two sites, Hawnby drains a larger catchment (AREA 131.28 km²) with the longest drainage path (LDP 22.07 km, DPLBAR 11.7 km), implying a slower overall hydrological response. The terrain is nevertheless steep (DPSBAR 144 m/km), and the catchment is the driest of the three (SAAR 882 mm, PROPWET 0.34), so flash flooding here is driven primarily by intense convective rainfall acting on relatively impermeable soils (BFIHOST19 0.38, SPRHOST 43.2%). This longer travel distance is reflected in a DMCA-derived catchment response time of 5.25 hours [21], the slowest of the three sites. A flash flood occurred in June 2005 following heavy rainfall [26-27]. Together, the three catchments span a useful range of conditions: a smallest, steepest, wettest, fastest resepnse, high-runoff upland (Walsden); a small, steep, but more permeable coastal catchment (Boscastle); and a much larger, drier, slower-responding moorland catchment (Hawnby). This contrast underpins the inter-comparison of model performance.

All three catchments have a river level gauging station at the catchment outlet, and each has nearby rainfall stations which are three at Walsden, two at Boscastle, and one at Hawnby. All data are recorded at 15-minute intervals, with river level in metres (m) and rainfall in millimetres (mm). The period of analysis was determined by the availability of overlapping rainfall and river-level measurements. The data were divided into training and testing sets as summarised in Tables 2 and 3. Training periods covered nine years at each site (315,552 records each), comprising 2006 - 2015 for Walsden, 2013 - 2022 for Boscastle, and 2014 - 2023 for Hawnby. Testing data spanned 2015 -2016 for Walsden (35,136 records), 2011 - 2013 for Boscastle, and 2023 - 2025 for Hawnby (70,176 records each). For Walsden and Hawnby the test set followed the training period, whereas for Boscastle an earlier interval was withheld because the most extreme observed data. All datasets are publicly available via the Department for Environment, Food & Rural Affairs (Defra) [17].

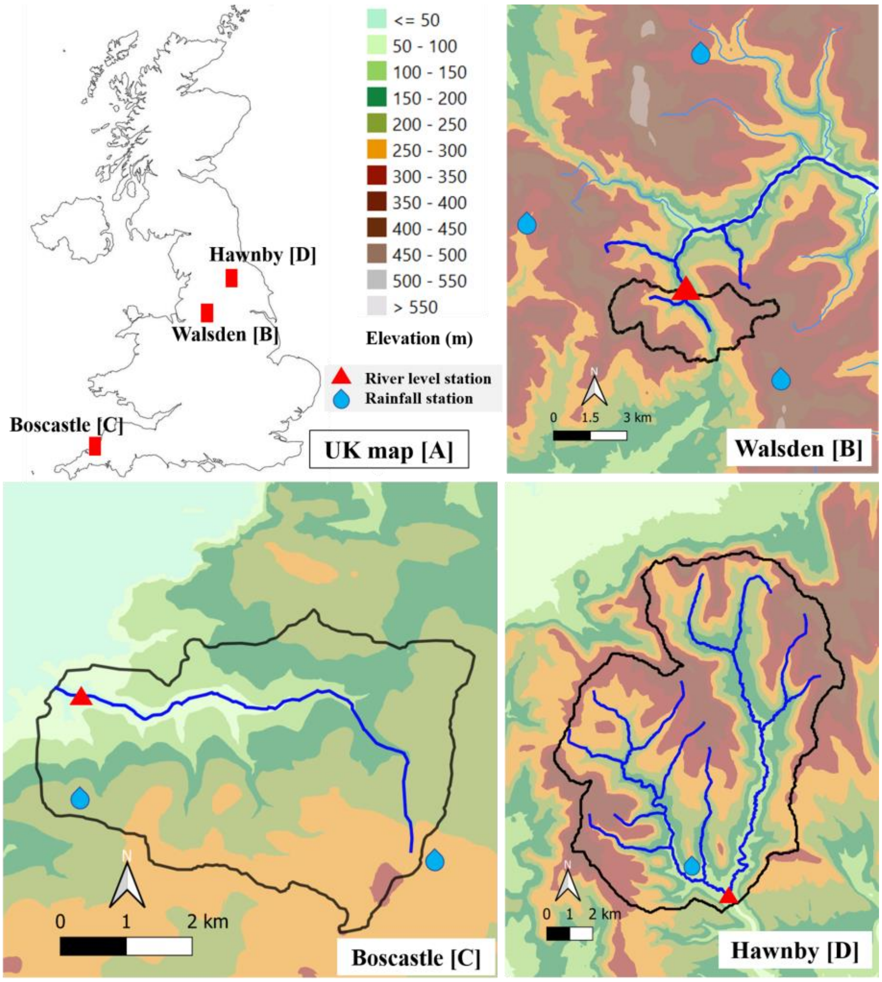

Figure 1 Approximate locations of the three study areas are shown on the UK location map (A): Walsden (B), Boscastle (C), and Hawnby (D). Each panel shows the main rivers (blue lines), river level stations (red triangles), rainfall stations (blue droplet symbols), and elevation (m, colour ramp). Data source: Department for Environment, Food & Rural Affairs (Defra) [28]

Table 1 FEH catchment descriptors [20] of Walsden, Boscastle and Hawnby.

FEH catchment descriptors	Walsden	Boscastle	Hawnby	catchment description
AREA	13.60	19.99	131.28	Catchment drainage area (km2)
ALTBAR	319.00	199.00	268.00	Mean catchment altitude (m above sea level).
BFIHOST19	0.39	0.50	0.38	Base flow index
DPLBAR	2.80	4.62	11.73	Mean of distances between each node on the IHDTM grid and the catchment outlet (km)
DPSBAR	158.40	124.30	144.00	Index of overall catchment steepness (m/km), >300 in mountainous terrain to <25 in the flattest parts.
FPDBAR	0.66	0.12	0.20	The mean depth of water on floodplains in a 100-year event.
LDP	6.44	8.44	22.07	Longest drainage path (km)
PROPWET	0.57	0.46	0.34	Catchment wetness index, over 80% in the wettest catchments to less than 20% in the driest parts.
SAAR	1434.00	1257.00	882.00	Standard-period Average Annual Rainfall
SPRHOST	44.09	34.48	43.18	Standard percentage runoff (%)
Catchment response time (hour)	1.00	2.00	5.25	Calculated based on the Detrending Moving-Average Cross-Correlation Analysis (DMCA) [21]

2.2 River flood indicators

The Environment Agency (EA) defines the flood risk for Walsden based on river levels [28]. The first indicator is set at 0.38 m, representing the upper limit of the normal range. At this stage, low-lying areas are prone to flooding, and a flood alert is issued. The second indicator is set at 0.60 m, marking the stage where property flooding becomes probable, triggering a flood warning. This study adopts the 0.38 m indicator as the primary indicator of flood level at Walsden. For Boscastle, the EA sets two levels at the River Valency level at Boscastle County Bridge which are 1.00 m and 2.10 m [29]. The flood level at 1.00 m will be used in this study. For Hawnby, only one flood level has been set by the EA which is 0.40 m [30].

The statistical data of full time-series training and testing datasets at Walsden, Boscastle, and Hawnby are shown in Tables 2 and 3. Overall, three catchments are mainly in long no-rain period, and low-magnitude river levels. That can be observed from their 80th percentiles of rainfall are zero, and their 95th percentiles of river levels are below the EA flood levels. Moreover, the river levels that are above the EA flood level show a very low proportion compared to their entire data, particularly Boscastle. Their full time-series training data contains only 19 of 315,552 (0.006%) records that are above flood level. These low proportions of flood data in the entire datasets confirm the severe class imbalance in the full time-series data.

Table 2 Rainfall and river level percentiles for the full time-series training data of the Walsden, Boscastle, and Hawnby catchments, along with the proportion of river level records exceeding the EA flood indicator.

Full time-series Training data	**Walsden**	**Boscastle**	**Hawnby**
Period (year)	2006 - 2015	2013 - 2022	2014 - 2023
Total records	315552	315552	315552
Rain (mm/15-min)			
80th percentile	0.0	0.0	0.0
95th percentile	0.6	0.4	0.2
Level (m)			
80th percentile	0.12	0.50	0.19
95th percentile	0.21	0.57	0.33
The EA flood indicator (m)	0.38	1.00	0.40
River level above the EA flood indicator (records)	1341	19	10740
River level above the EA flood indicator (% of total records)	0.425%	0.006%	3.403%

Table 3 Rainfall and river level percentiles for the full time-series testing data of the Walsden, Boscastle, and Hawnby catchments, along with the proportion of river level records exceeding the EA flood indicator.

Continuous Testing data	**Walsden**	**Boscastle**	**Hawnby**
Period (year)	2015 - 2016	2011 - 2013	2023 - 2025
Total records	35136	70176	70176
Rain (mm/15-min)			
80th percentile	0.0	0.0	0.0
95th percentile	0.8	0.4	0.2
Level (m)			
80th percentile	0.15	0.34	0.21
95th percentile	0.24	0.43	0.37
The EA flood indicator (m)	0.38	1.00	0.40
River level above the EA flood indicator (records)	383	4	3092
River level above the EA flood indicator (% of total records)	1.090%	0.005%	4.406%

2.3 Data preprocessing

To ensure data integrity, a preprocessing workflow was implemented to handle missing values. For rainfall records, missing values were infilled using the average of the nearest known observations. For river level data, missing values were filled using linear interpolation.

In addition to the measured data, two additional variables were calculated, which are cumulative rainfall over 120 minutes (mm/120-min) and the rate of change in river level over 180 minutes (m/180-min). The cumulative rainfall of 120-minutes is the total precipitation recorded over a rolling two-hour window. This feature provides a representation of heavy rainfall over a specific period, allowing the model to learn the intensity of rain over time. The 180-minute RoC was calculated as the difference between the current river level and its previous level recorded 180 minutes (three hours).

The input features used for the forecasting models are 15-minute rainfall, 120-minute cumulative rainfall, 15-minute river level, and 180-minute RoC of river level. These features were held constant across all three study catchments. This was an experimental-design choice, as the objective of the study is to evaluate whether a consistent RoC threshold can be used to filter relevant flood events for training. This question requires that the input features are standardised across catchments so that any differences in optimal threshold can be attributed to catchment characteristics rather than to differences in feature engineering. Changing the features individually per catchment (for example, matching the cumulative-rainfall window to each catchment's response time) might yield better forecasting performance for that catchment, but would prevent cross-catchment comparison of the threshold's behaviour. The cost of this design, particularly for catchments whose response time differs from the standardised feature windows, is acknowledged and considered in the discussion.

3 Methodology

3.1 Event data selection process

Hydrological data are typically characterized by prolonged periods with no rainfall and low magnitude of river level (Tables 2 and 3). Training data-driven models on such raw time series data often results in performance biased toward these dominant low river level conditions [20, 21]. Flash floods are rare events, making them a very small proportion compared to the entire data record. Forecasting models trained on full time-series tend to optimize for global error metrics by prioritizing the dominant low magnitude of river level conditions, resulting in underprediction of the magnitude of rare peaks. To minimize the dominance of uninformative data and mitigate this class imbalance, the continuous hydrological time series data were filtered into discrete potential flash flood data, called ‘event data’, each event having the same window time. This selective training scheme is similar to calibration processes in traditional hydrological forecasting; for example, the conceptual rainfall-runoff Probability-Distributed Model (PDM) calibrates parameters against a subset of the highest observed flow events [33].

Potential flash flood events were identified using the rate of change in river level as a selection criterion. RoC characterises flood events by capturing the rapid dynamic response of the river rise. The RoC measures the rise in river level over a defined interval. It responds only to the rising limb of the hydrograph and avoids the over-counting that can occur with fixed river level thresholds when slow recessions cross flood level. The effectiveness of RoC depends on the interval over which it is computed. It controls the balance between sensitivity to rapid rises and filtering out minor fluctuations that do not develop into a flood.

To determine an appropriate RoC interval, a sensitivity analysis was carried out comparing four RoC intervals, which are 15, 60, 180, and 360 minutes, against a reference flood-event classification defined by exceedance of river levels that are above 0.38 m as per the EA flood indicator at Walsden (Appendix A). Across 90th, 95th, and 99th percentile settings tested, precision against this reference increased with interval length, while the number of selected events decreased. The 15- and 60-minute intervals were over-sensitive to short-duration fluctuations, producing flagged events of which 38 - 72% did not reach flood stage. The 360-minute interval achieved high precision but selected too few events (5 - 39 events across the three percentiles tested) for model training. The 180-minute interval was the candidate that satisfied three operational requirements that are precision above 80% against the river level threshold reference, a sample size sufficient for model training, and a temporal window long enough to capture the rise characteristic of flash flood events while ignoring minor rises.

To select events, once an assigned threshold was met, an event window was extracted comprising data from a fixed period before and after the threshold crossing (Figure 2). The window lengths were chosen to span each catchment's hydrological response time identified by the DMCA method (Table 1). For Walsden and Boscastle (response times of 1- 2 hours), an 8-hour window was applied, which is from the six hours before and two hours after the threshold, equivalent to 32-time steps at the 15-minute records. For Hawnby, with a longer 5.25-hour response time, the window was extended to 16 hours (11 hours before and 5 hours after, or 64-time steps). The asymmetric window, which is longer before the threshold than after, captures the rising limb in its event hydrograph, as it is the important part of the flood event and the focus of the forecasting models.

The threshold values tested at each catchment were chosen to span the operational range from frequent low-magnitude rises to rare high-magnitude events, with the upper limit determined by the point at which the resulting event number became too small for model training (Table 3). At Walsden and Boscastle, eight thresholds between 0.01 and 0.35 m/180-min were applied. Above this range, the number of events is low considered insufficient for training (with 7 and 11 events at the upper limit, respectively). At Hawnby, where the RoC distribution extends to higher values, eleven thresholds between 0.01 and 0.60 m/180-min were tested, again to the point of a low number of events (8 events at the upper limit). Threshold values were spaced at intervals of 0.05 m/180-min, with the upper end at Hawnby extending in steps of 0.10 m/180-min, so that each threshold produced a different training set from the next. At higher thresholds, the number of events drops with each step, so setting finer intervals between each threshold would have produced training sets differing by only few events.

The event-selection results in Table 4 reveal an inverse relationship between the RoC threshold and the number of selected events. Lower thresholds yielded a higher number of events, while higher thresholds captured only steep-rise events, resulting in a lower number of selected events. As the threshold acts as a minimum cutoff, these steep-rise events are therefore subsets of those generated by lower thresholds (Figure 3). The proportion of flood-level observations within the selected events increases with the threshold, addressing the severe class imbalance present in the continuous records, where flood-level observations account for only 0.425%, 0.006%, and 3.403% of the training data at Walsden, Boscastle, and Hawnby, respectively (Table 2). At Walsden, for example, this proportion rises from 0.425% in the continuous record to 40.400% in the event dataset selected at a 0.30 m/180-min threshold (Table 4). However, this increase varies between sites. Hawnby exceeds 50% flood representation from RoC 0.25 onwards, while Boscastle remains below 5% from the highest tested threshold, reflecting the rarity of flood-exceeding events in the record (only 19 records in a decade of 15-minute observations, Table 2). Overall, the results indicate that RoC-based event selection is an effective way to reduce class imbalance in flood-prediction training data without requiring synthetic data augmentation examples.

The 20 testing events (Table 3, and Figure 4) were selected from the peak river level events of each catchment during its testing period. This approach is similar to the calibration process of traditional hydrological flood forecasting models, such as PDM, in which model parameters are calibrated by comparing outputs against the 20 highest observed river flow events [33]. Even though the peak river levels have been selected for testing, not all data are above the EA flood indicator of each site. The proportion of data points within these selected events that exceed the EA flood indicator are Walsden (18.33%) and Boscastle (0.62%), except for Hawnby (57.42%).

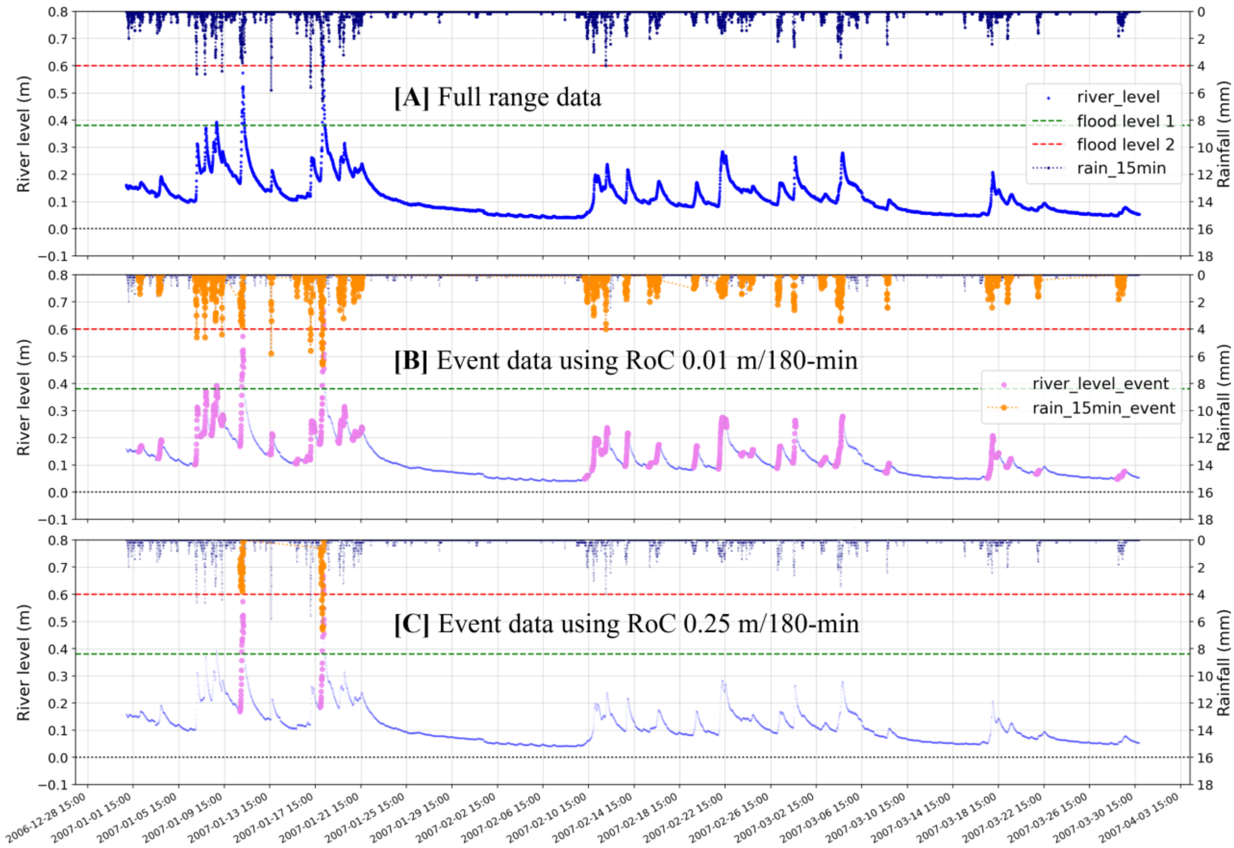

Figure 2 Illustration of the event-selection process applied to 15-minute hydrological time-series data from Walsden, 1 January to 31 March 2007. (A) Full time-series of river level (blue, left axis) and rainfall (navy, right axis, inverted). Events extracted using rate-of-change thresholds of 0.01(B) and 0.25 (C) m/180-min, respectively, with selected river-level (pink) and rainfall (orange) overlaid on the full time-series (faded) for comparison. The lower threshold (B) has many events ranging from minor fluctuations to peak levels, whereas the higher threshold (C) isolates only the two largest events in this period. Horizontal dashed lines mark the Environment Agency flood levels 1 (green) and 2 (red) for Walsden. This comparison illustrates the sensitivity of event extraction to the choice of RoC threshold.

As illustrated in Figure 4, this scarcity of high-magnitude river-level observations is also evident in the testing events, even though these events were selected based on their highest river levels. The number of testing events exceeding the EA flood indicator differs between catchments. All 20 test events at Hawnby exceed 0.40 m, the EA flood line, 9 of 20 exceed Walsden's flood line 1 (0.38 m), but only 2 of 20 exceed Boscastle's 1.00 m flood line.

These differences reflect the catchment characteristics summarised in Table 1. Hawnby is the largest catchment (AREA 131.28 km²) with a low BFIHOST (0.38) and a flood threshold set close to its routine flow regime, so test events readily exceed it. Walsden is small and steep (AREA 13.60 km², DPSBAR 158.4 m/km) with the highest SAAR of the three sites (1434 mm) and a 1-hour response time, producing frequent rapid-rise events. Boscastle, by contrast, has a higher BFIHOST (0.50) than the other two sites (0.38–0.39), meaning a larger proportion of rainfall is absorbed and released slowly rather than reaching the river as rapid runoff; combined with a flood threshold (1.00 m) above its typical flow regime, this means river levels rarely reach the EA flood indicator.

Table 3 The number of training events based on RoC thresholds and the amount of flood data percentage in training and testing sets from three catchments.

Dataset	RoC threshold (m/180-min)	Walsden			Boscastle			Hawnby		
		Selected training data (events)	The EA flood level (m)	Data above flood level (%)	Selected training data (events)	The EA flood level (m)	Data above flood level (%)	Selected training data (events)	The EA flood level (m)	Data above flood level (%)
Train	RoC 0.01	1633		2.08%	2151		0.02%	732		13.94%
	RoC 0.05	552		5.05%	339		0.17%	281		30.96%
	RoC 0.10	271		7.84%	153		0.38%	183		41.87%
	RoC 0.15	123		13.44%	92		0.64%	133		48.38%
	RoC 0.20	56		22.46%	62		0.95%	109		48.68%
	RoC 0.25	33	0.38	28.93%	33	1.00	1.79%	73	0.40	52.40%
	RoC 0.30	15		40.40%	19		2.79%	52		53.46%
	RoC 0.35	7		37.66%	10		5.31%	40		52.46%
	RoC 0.40							25		51.38%
	RoC 0.50	n/a		n/a	n/a		n/a	15		51.25%
	RoC 0.60							8		48.63%
Test		20	0.38	18.33%	20	1.00	0.62%	20	0.40	57.42%

These contrasting hydrological characters produce contrasting event distributions in both the training and testing sets, which have direct consequences for how training data should be selected. Figure 4 also plots the maximum 180-minute RoC of each testing event against the number of training events available at each threshold. As the RoC threshold rises, the number of training events drops sharply, while testing events, particularly those above the EA flood indicators, have high RoC values where comparable training data is scarce. This train–test mismatch motivates the need to optimise the training event selection.

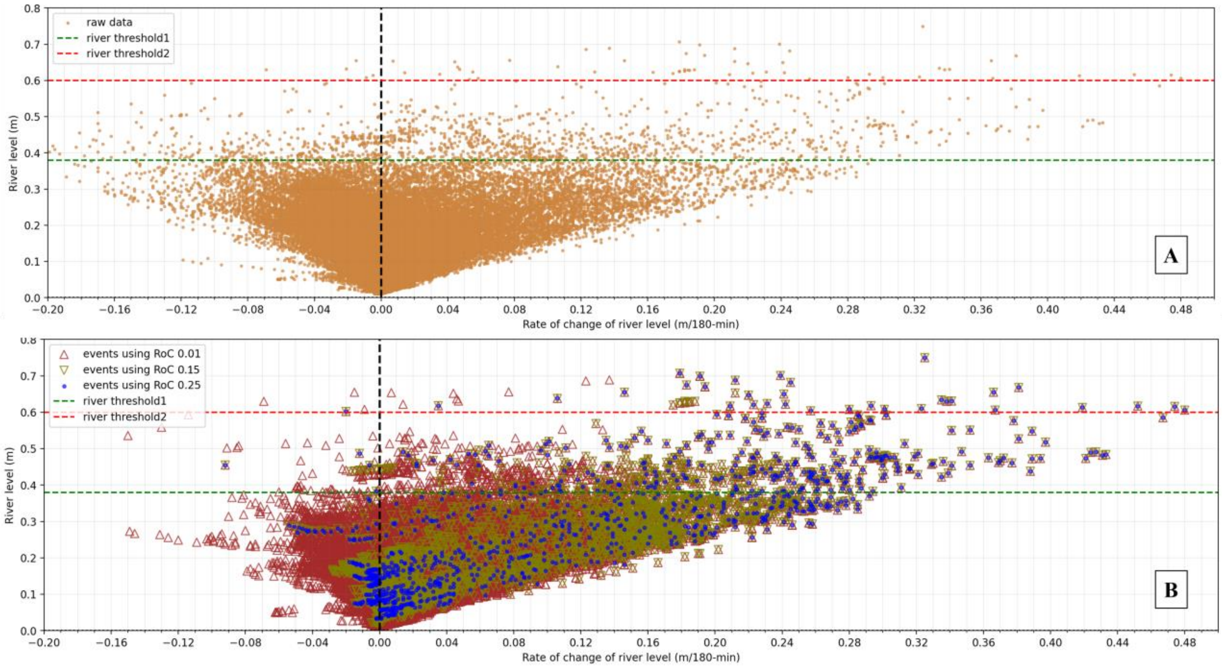

Figure 3 Scatter plots comparing full time-series data (orange, A) with extracted events from Walsden catchment. The selected event data using 180-minute rate-of-change thresholds of 0.01 m/180-min (red triangles), 0.15 m/180-min (olive reverse triangles), and 0.25 m/180-min (blue dots) (B). The EA flood indicators at Walsden are 0.38 m (green dashed line) and 0.60m (red dashed line). The vertical dashed line at zero RoC separates rising-limb (positive) from falling-limb (negative) observations. Events extracted using higher RoC thresholds form a subset of those captured at lower thresholds.

3.2 Forecasting structures

The rainfall and river level models adopt a multi-step forecasting approach with a fixed-length sliding window. The lengths of the input window (the historical observations available to the model at each prediction step) and the forecast horizon (the number of future steps predicted) were chosen based on the catchment response times derived from the DMCA technique [21].

For Walsden and Boscastle, with response times of 1.00 and 2.00 hours respectively, the input window was set to 12-time steps and the forecast horizon to 12 future time steps, corresponding to 3.00 hours of historical context and a 3-hour forecast at 15-minute resolution. The configuration was sized to accommodate the longer of the two response times (Boscastle's 2.00 hours); consequently, the input and forecast lengths for Walsden are several times its response time. A common configuration was used across the two catchments to support cross-catchment comparison, consistent with the standardised input design described in the Data preprocessing Section. For Hawnby, with a longer response time of 5.25 hours, the input window and forecast horizon were both extended to 24-time steps (6.00 hours each), maintaining the principle of an input window long enough to capture the slower hydrological build-up characteristic of this catchment. The sliding window advances by one time step across each event, so that each successive prediction window overlaps the previous one by all but one time step, producing a dense sequence of multi-step forecasts across the duration of the event.

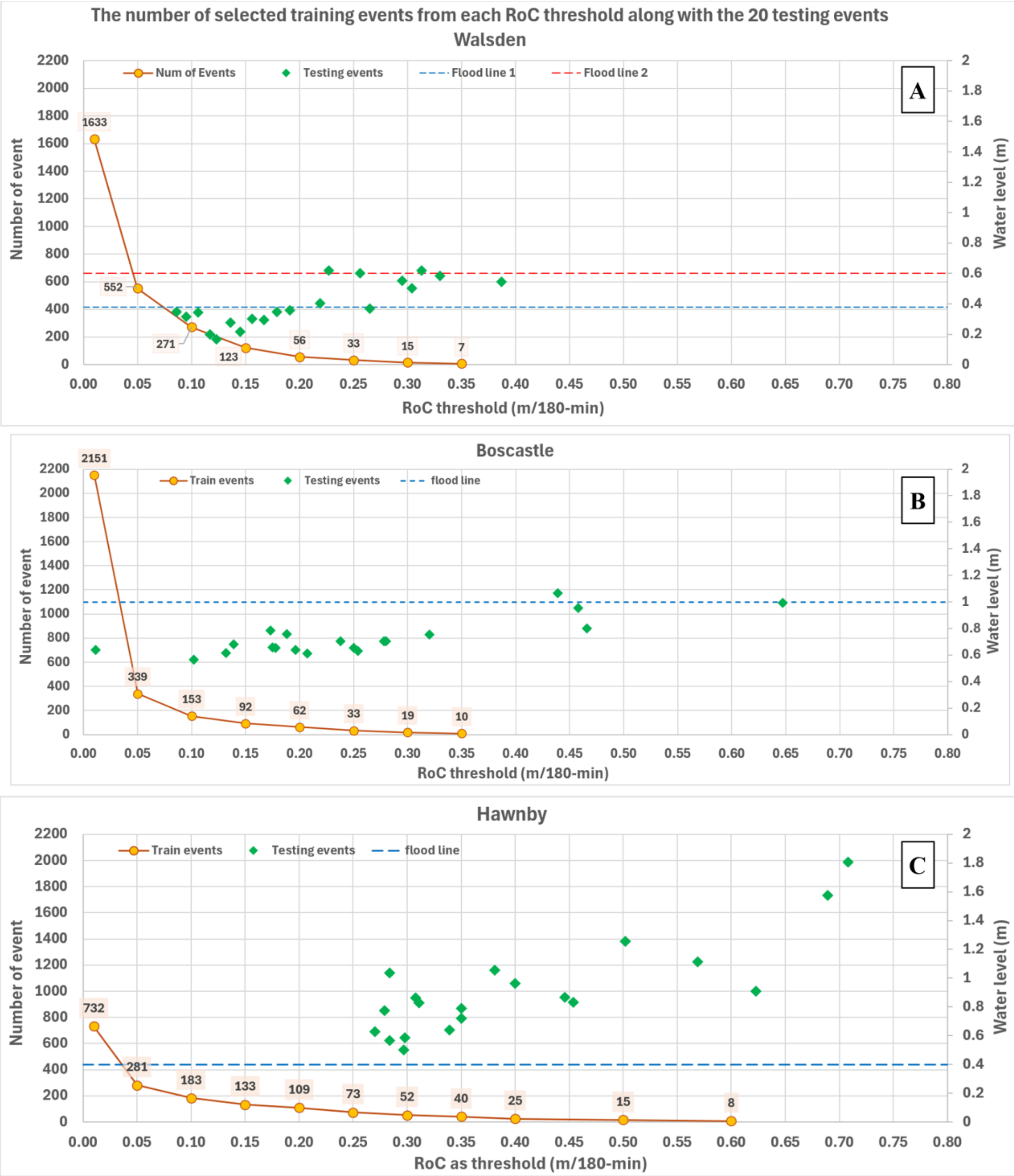

Figure 4 Distribution of testing events and the number of available training events across rate of change (RoC) thresholds for the three catchments: (A) Walsden, (B) Boscastle, and (C) Hawnby. Orange line (left Y-axis, number of events) show the number of selected training events at each RoC threshold, with its values. Green diamonds show the 20 testing events at each catchment, plotted by their maximum 180-minute RoC on the x-axis and by their peak water level on the right Y-axis. Horizontal dashed lines are the Environment Agency (EA) flood indicators: flood line 1 (blue, 0.38 m) and flood line 2 (red, 0.60 m) at Walsden, and the single flood line (blue) at 1.00 m for Boscastle and 0.40 m for Hawnby. All panels used common axis ranges for cross-catchment comparison (RoC threshold: 0 - 0.80 m/180-min (X-axis); number of events: 0 - 2200 (left Y-axis); water level: 0 - 2.0 m (right Y-axis)).

A common feature of all three configurations is that the forecast horizon was set to exceed the catchment response time. This is a methodologically demanding choice. For lead times shorter than the catchment response time, river behaviour is largely determined by rainfall that has already fallen, and forecasts benefit from the catchment's own hydrological memory of observed recent inputs. Beyond the response time, this advantage diminishes. The river response depends on rainfall that will fall during the forecast period, which is not observed and must instead be supplied by a coupled rainfall forecasting model. The river level forecasts therefore depend not only on the river level model's ability to translate rainfall into runoff, but also on the accuracy of the rainfall forecast that drives it which is a more challenging configuration that reflects the operational requirements of flash flood early warning, where useful lead times must exceed the catchment's natural response time so that warnings can be issued and acted upon.

3.3 Forecasting models

The discrete event data extracted in the previous step (Event data selection process section) were used as the primary datasets for model development, rather than the continuous raw time series. Two forecasting models, which are rainfall and river level, were developed using LSTM networks. Both employ an encoder–decoder sequence-to-sequence architecture, in which stacked LSTM layers encode the input history into hidden states and a separate stack of LSTM cells then generates the forecast autoregressively, feeding each prediction back as input to the next forecast step. The autoregressive decoding strategy was chosen for its alignment with the temporal structure of the forecasting task. The prediction errors accumulating across the forecast horizon addresses the limitations.

The forecasting structures have been set at 12 for 12 steps at Waslden and Boscastle, while 24 for 24 steps are for Hawnby where it has a longer catchment response time. At each historical time step, the input feature set comprises four variables, which are 15-minute rainfall, 120-minute cumulative rainfall, the river level at the 15-minute interval, and the 180-minute rate of change of river level. The rainfall model produces multi-step forecasts of 120-minute cumulative-rainfall features, which are then supplied to the river level model as additional inputs alongside the historical features. To ensure that training shares the same input distribution, the river level model is trained using the rainfall model's predicted future rainfall rather than observed future rainfall. The implications of using the rainfall model's predictions on its own training data are revisited in the limitations.

To enable a controlled evaluation of how event-based dataset size, which varies with the chosen RoC threshold, affects model performance, hyperparameters were held constant across both models and across all RoC thresholds. Each model used two stacked LSTM layers with 64 hidden units. Dropout regularisation was applied at a rate of 0.1 on the input and recurrent connections of the encoder and on the decoder cell output. Training used the Adam optimiser with a learning rate of 0.001, a batch size of 32, and Mean Squared Error as the loss function. All inputs and outputs were normalised to the range (0, 1) using Min–Max statistics fitted on the training portion of each dataset.

Due to dataset size constraints, training and testing data splitting was utilised. Training was capped at 50 epochs, and early stopping was implemented by monitoring the loss on the testing set. Training will be stopped if the loss failed to improve for five consecutive epochs following an initial five-epoch warm-up period. The model weights from the epoch with the lowest testing error were restored for final evaluation. While this mitigates overfitting to the training data, it necessitates that the stopping criterion was informed by the testing data. Therefore, absolute error values represent the best-epoch performance on the testing data. The implications of this approach are addressed in the limitations section. Finally, to quantify variance introduced by network initialisation, each configuration was trained with five random seeds.

4 Results

To evaluate how the choice of RoC threshold affects forecasting performance, river level models with identical architecture, hyperparameters, and training schedule were trained on event subsets defined by each of the RoC thresholds (Table 4). For Walsden and Boscastle, eight thresholds were tested, yielding eight trained models per catchment; for Hawnby, eleven thresholds were tested, yielding eleven trained models. Because the model configuration was held strictly constant within each catchment, any differences in performance can be confidently attributed to the composition and size of the training set induced by the varying thresholds, rather than to differences in modelling choices.

Each trained model was evaluated against the testing set of 20 events for its catchment, selected as the 20 events with the highest river level in the testing data. The peak RoC values within these test events varied by catchment (Figure 4), which are from 0.09 to 0.39 m/180-min at Walsden, 0.01 to 0.65 at Boscastle, and 0.27 to 0.71 at Hawnby. Because the RoC thresholds used to filter the training data act as minimum cutoffs (selecting events whose peak RoC equals or exceeds the threshold), higher-threshold training sets are nested subsets of lower-threshold sets, and a model trained at any threshold is exposed to the full range of peak RoC values present in its training data, not only those near the threshold value itself.

Forecasting performance was quantified using four error metrics, which are RMSE, MAE, MAPE and NSE. The first three metrics have an optimal value of zero; NSE has an optimal value of one. These metrics provide views of forecast accuracy in which RMSE emphasises large errors, MAE treats all errors equally, MAPE expresses error in relative terms, and NSE places performance on a normalised scale that allows comparison across catchments. The forecasting results are presented in Figures 5, 6, and 7 for Walsden, Boscastle, and Hawnby, respectively.

Inspection of the results reveals that forecasting performance depends not only on the RoC threshold used to filter the training data but also on the magnitude of the test event itself. To examine this dependence, the 20 test events at each catchment have been divided into two groups based on the maximum RoC observed within each event. For Walsden (Figure 5), a split point of 0.20 m/180-min produces 11 lower-RoC and 9 higher-RoC events. The split point was selected based on the result value that clearly separated the two response patterns. For Boscastle and Hawnby, these points are at 0.30 m/180-min and 0.35 m/180-min, respectively. Each box plot in Figures 5, 6, and 7 summarises the distribution of a metric across the testing events in one group, for a single training threshold. Each trained model is evaluated on the same 20 test events which are divided into the two groups that allows the dependence of forecast performance on test-event magnitude to be examined.

The two groups show contrasting responses to the training threshold. For higher-RoC test events at Walsden (Figure 5, right column), the median RMSE (0.11 m) is high at the low training threshold (0.05 m/180-min) and decreases as the training threshold rises, reaching the lowest of median RMSE 0.05 m (A lower RMSE means the model's predictions are more accurate) at thresholds 0.25 – 0.30 m/180-min, where median NSE also peaks at 0.75 - 0.78. The pattern indicates that filtering the training set to retain only higher-RoC improves forecasting of the higher-RoC test events. For lower-RoC test events (left column), the pattern is inverted. The median RMSE (0.10 - 0.14 m) is high when the model is trained at the optimal threshold of higher-RoC test events (0.25 – 0.30 m/180-min), but the median RMSE reached the lowest value (0.03 m) at threshold 0.05 m/180-min. The NSE values show the same pattern in inverted form. The highest median NSE (0.61) occurs at threshold 0.05 m/180-min and declines as the threshold increases, becoming negative for higher thresholds. MAE and MAPE confirm the same patterns observed in RMSE for both groups.

The Boscastle results (Figure 6) show the same patterns observed at Walsden but with higher threshold values. A split point of 0.30 m/180-min separates the 20 test events into 15 lower-RoC and 5 higher-RoC events. For the 5 higher-RoC test events at Boscastle (Figure 6, right column), median RMSE decreases from 0.15 - 0.17 m at training threshold 0.05 - 0.15 m/180-min to the lowest median RMSE 0.07 - 0.09 m at training thresholds 0.20 – 0.30 m/180-min, with median NSE rising from around -0.80 to 0.50 across the same threshold ranges. The direction of the pattern matches Walsden's finding that higher-RoC events benefit from higher training thresholds; however, it is worth noting that there is a small sample of 5 events under this high-RoC range.

For the 15 lower-RoC test events at Boscastle (Figure 6, left column), the median RMSE (0.12 - 0.13 m) is high when the model is trained at the optimal threshold of higher-RoC test events (0.20 – 0.30 m/180-min). The median RMSE reached the lowest value (0.05 m) at threshold 0.05 - 0.15 m/180-min. Median NSE shows the corresponding pattern, peaking at 0.13 - 0.20 in the 0.05 - 0.15 m/180-min range and declining at higher thresholds. MAE and MAPE follow the same trends across both groups. The optimal training thresholds at Boscastle 0.20 - 0.30 m/180-min for higher-RoC events and 0.05 – 0.15 m/180-min for lower-RoC events which are shifted upward relative to Walsden's optima, consistent with Boscastle's broader RoC distribution.

The Hawnby results (Figure 7) show the similar pattern of lower-RoC events observed at Walsden and Boscastle but a weaker version of the higher-RoC pattern. A split point of 0.35 m/180-min produces 11 lower-RoC and 9 higher-RoC events. For the 9 higher-RoC test events at Hawnby (Figure 7, right column), median RMSE varies modestly across training thresholds, from approximately 0.12 m to 0.19 m, with overlap between adjacent thresholds. The lowest median RMSE is 0.12 m at 0.35m/180-min threshold, but the difference from neighbouring thresholds (0.25 – 0.50 m/180-min) is small relative to the within-threshold spread of metric values (0.12 - 0.14 m). For higher-RoC events at Hawnby, the choice of training threshold therefore appears less consequential than at the faster-responding catchments.

For the 11 lower-RoC test events at Hawnby (Figure 7, left column), the results show similar pattern that were observed at Walsden and Boscastle for the low-RoC events. Median RMSE is the lowest values (0.05 – 0.06 m) at training thresholds 0.15 – 0.30 m/180-min and rises at higher thresholds (up to 0.25 m at threshold 0.60 m/180-min). Median NSE confirms the pattern, peaking at 0.76 - 0.85 in the 0.15 – 0.30 m/180-min range. This is the highest NSE values observed across the three studied catchments. MAE and MAPE again follow the same patterns.

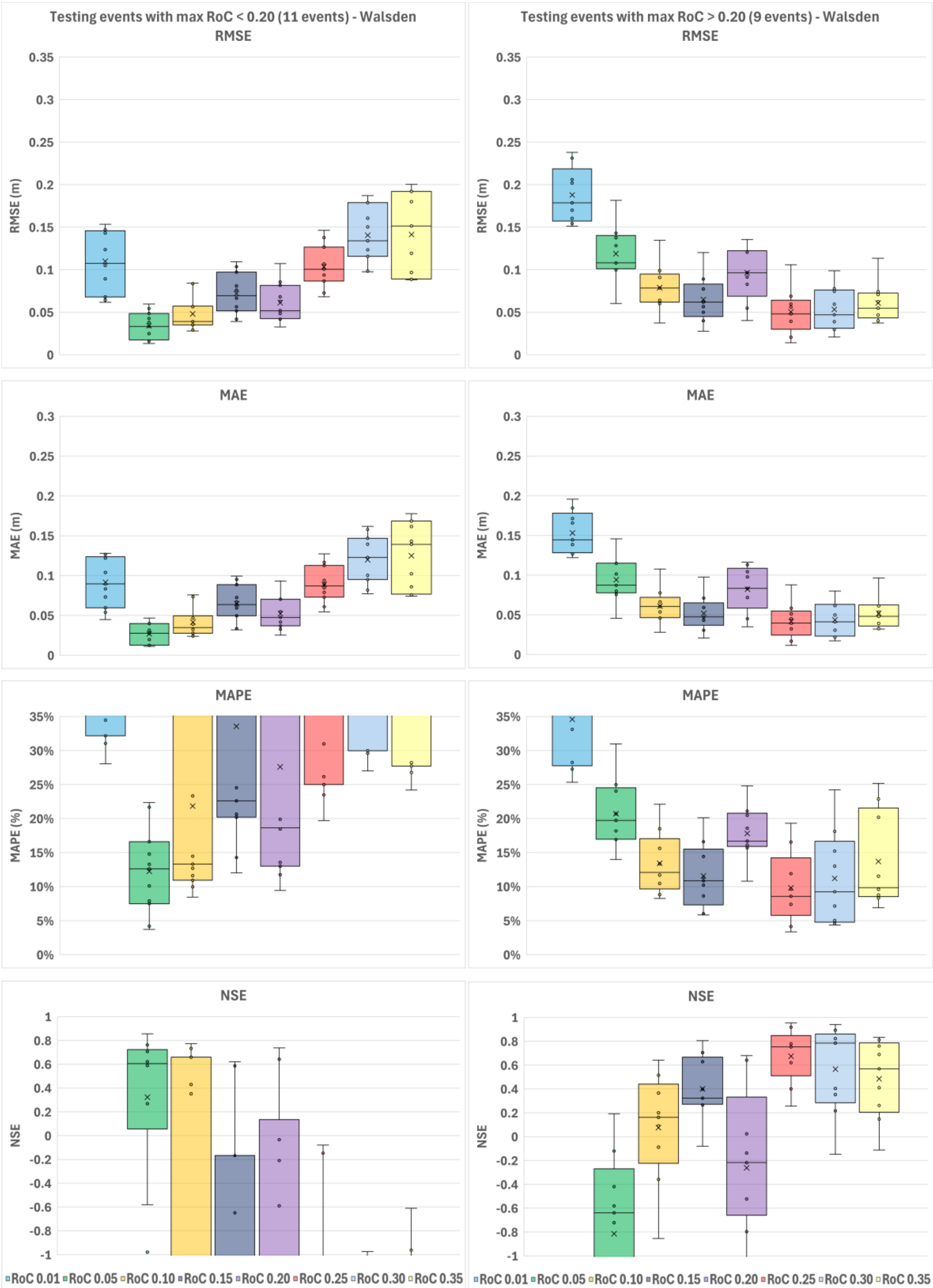

Figure 5 Comparative box plots of four evaluation metrics (RMSE, MAE, MAPE, and NSE) for the Walsden catchment. The results are divided into low-magnitude events (left column) and extreme flood events (right column). Box colours represent the specific RoC threshold used to filter the model's training data. The y-axis scales are standardized across both columns to enable direct comparison; consequently, some MAPE and NSE distributions are truncated.

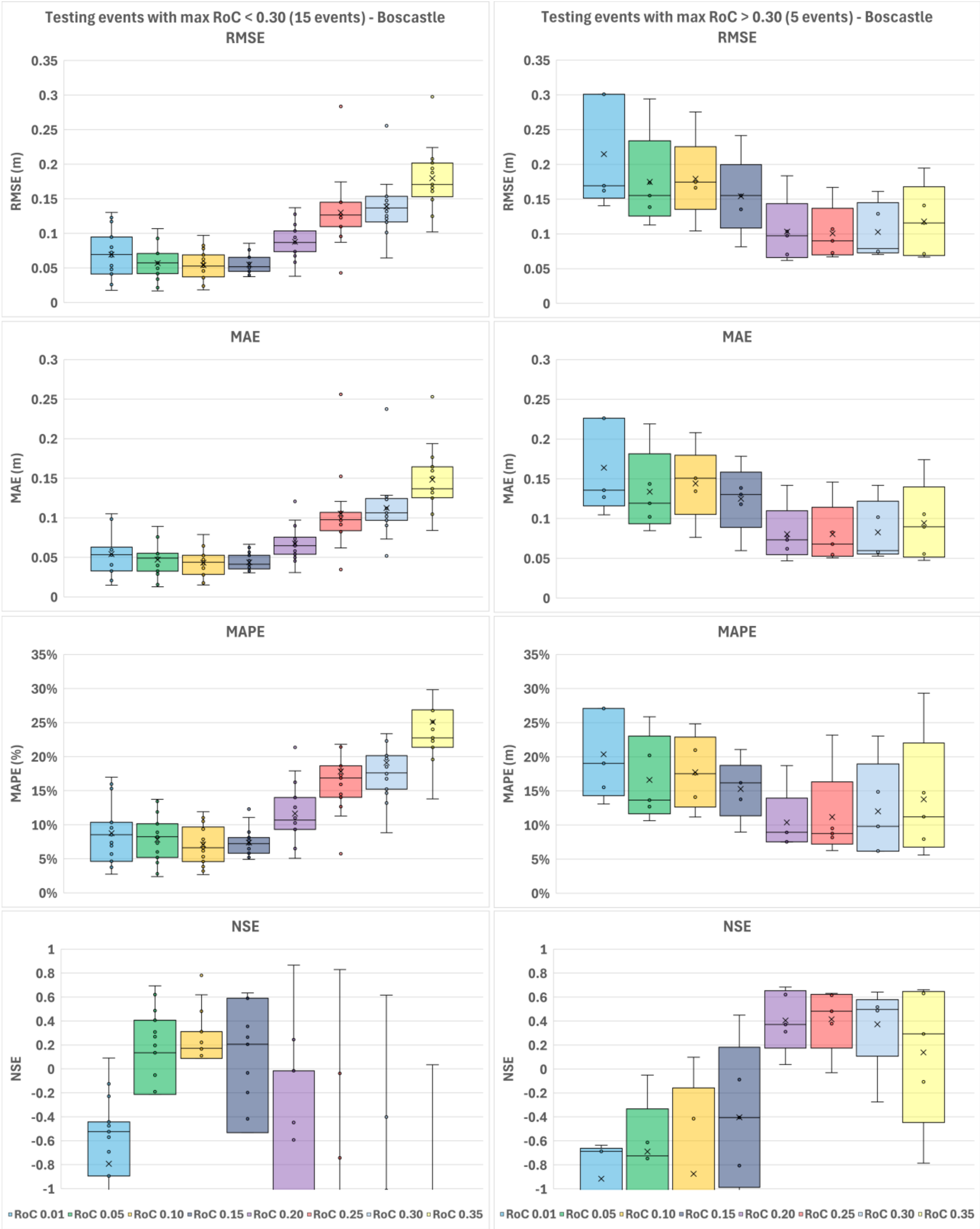

Figure 6 Comparative box plots of four evaluation metrics (RMSE, MAE, MAPE, and NSE) for the Boscastle catchment. The results are divided into low-magnitude events (left column) and extreme flood events (right column). Box colours represent the specific RoC threshold used to filter the model's training data. The y-axis scales are standardized across both columns to enable direct comparison; consequently, some NSE distributions are truncated.

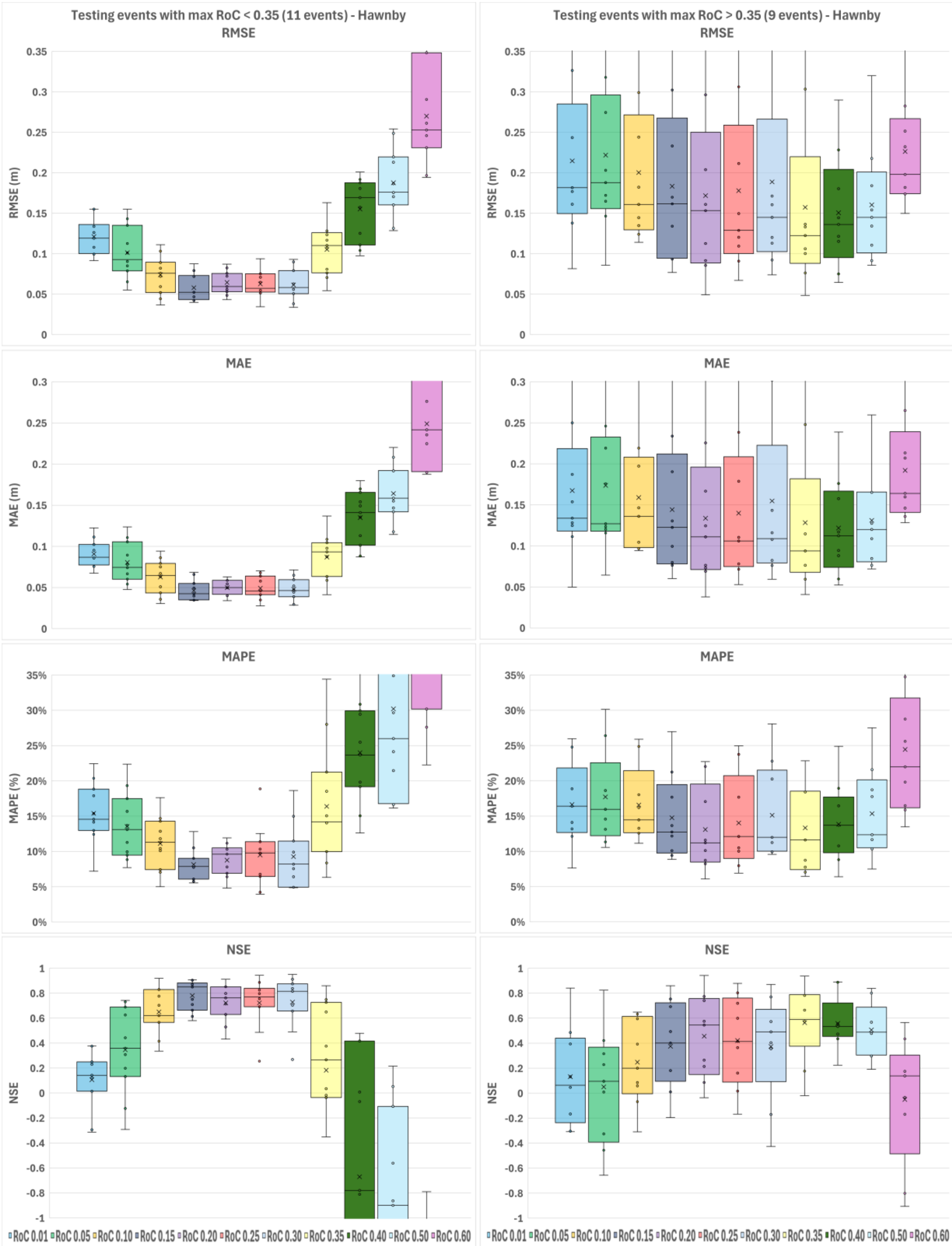

Figure 7 Comparative box plots of four evaluation metrics (RMSE, MAE, MAPE, and NSE) for the Hawnby catchment. The results are divided into low-magnitude events (left column) and extreme flood events (right column). Box colours represent the specific RoC threshold used to filter the model's training data. The y-axis scales are standardized across both columns to enable direct comparison; consequently, some RMSE, MAE, MAPE, and NSE distributions are truncated.

5 Discussion

Across the three study catchments, the evaluation results reveal a relationship between forecasting accuracy and the rate of change threshold used to select the training data, although this relationship varies between catchments. Two contrasting trends can be observed depending on the magnitude of the events being forecast. Models trained with lower RoC thresholds achieve the lowest errors (RMSE, MAE, MAPE) and highest NSE when forecasting lower-magnitude river levels and slower rates of rise events. Conversely, accurate forecasting of higher-magnitude river levels and rapid rates of rise events requires models trained on data filtered with higher RoC thresholds. Based on these relations, there is no single training threshold that can optimize across the full range of event magnitudes. The threshold acts as a filter on event magnitude in the training data, and its optimal setting depends on the magnitude of the events the model is required to forecast.

These results indicate that that the selection of relevant training data influences the performance of deep learning models for flood forecasting. This study uses the rate of change of river level as the criterion for filtering relevant events from the full hydrological record. RoC offers a practical alternative to the flow return periods relied upon in previous studies [7, 15]. RoC can be computed directly from river level, a standard and widely available measurement [6], and does not require the longer records or flow-frequency analysis that return-period approaches depend on.

The RoC threshold is filtering the composition of the training set. As the RoC threshold increases, the proportion of flood-specific, rapid-rise conditions in the training data rises and the proportion of low-magnitude events falls (Table 4), allowing the model to train on rapid-rise dynamics. Conversely, models trained at low thresholds are dominated by low-magnitude data (Figure 3), although their training sets still contain some extreme events. This class imbalance limits the model's ability to learn and reproduce extreme flood behaviour, which is reflected in the higher errors these models produce on higher-magnitude test events. The same mechanism explains the inverse pattern for lower-magnitude events in which models trained at high thresholds rarely encounter low-magnitude rises during training and consequently forecast them less accurately.

The catchment characteristics which vary among the three study areas, strongly influence this mechanism. The trade-off was observed clearly at Walsden and Boscastle, where higher training thresholds improved forecasting of higher-magnitude events. At Hawnby, the trade-off was only partially observed. The lower-magnitude events at Hawnby showed the same threshold dependence at the other two catchments, but forecasting of higher-magnitude events was insensitive to the training threshold across a broad RoC range. The FEH catchment descriptors (Table 1) offer a physically plausible explanation grounded in the concept of river flashiness which are the rapidity and frequency of fluctuations in river [34]. The notable differences between the three catchments are size and wetness. Walsden (13.60 km^2) and Boscastle (19.99 km^2) are small, wet, fast-responding catchments (SAAR 1434 mm and 1257 mm respectively). Both have hydrological response times of 1.00 – 2.00 hours (based on the DMCA method [21]) and would be classified as high flashiness. Hawnby (131.28 km^2) is six to ten times larger across all size-related descriptors (AREA, LDP), is in the drier north-east area (SAAR 882 mm), and has a response time of 5.25 hours indicating low flashiness.

In a large and less flashy catchment like Hawnby, the outlet hydrograph integrates runoff arriving from across a wide area at different times, producing a more gradual and smoothed response. This integration is likely to make event shape more uniform across the magnitude range, so that filtering the training data by rate of change performs less effectively between event types. In flashier catchments like Walsden and Boscastle, RoC thresholds are different between event types, so filtering training data by RoC segregates effectively between rapid-onset events and more gradual rises. This is consistent with the observation at Hawnby that the model trained on high-RoC threshold made a slight improvement compared with the models trained on adjacent RoC thresholds for forecasting high-magnitude events. The finding is aligned with the study from Santos et al. [6], in which they suggested that the regional hydrological characteristics have an impact on forecast performance.

The consistency between the observed forecasting behaviour and the catchment physical descriptors suggests that the applicability of RoC-based training filtering is greatest in small, fast-responding catchments, which are prone to flash flooding, making the technique relevant for operational use. While limited to a three-catchment design, this study provides an initial assessment of how these findings might transfer across different environments. The RoC threshold effect for lower-magnitude events transferred consistently across all three sites, while the corresponding effect for higher-magnitude events transferred to the two flashier catchments but not to the slower-responding Hawnby. This indicates that while RoC-based training filtering is a generalisable technique, its specific benefit for predicting extreme events is conditional on catchment responsiveness.

For operational applications, these findings suggest that flash flood forecasting systems intended to specialise in high-magnitude events should be trained on filtered subsets of higher RoC, particularly in fast-responding catchments where this provides the best performance. Conversely, systems intended to forecast across the full range of event magnitudes require a lower threshold that contains lower-magnitude events, with the optimal thresholds tuned to each catchment's response characteristics. In larger, slower-responding catchments such as Hawnby, the choice of training threshold for high-magnitude events appears less critical, and dataset size considerations may take precedence over magnitude filtering. Nonetheless, because this study is restricted to three sites, future research utilising a larger sample of catchments spanning diverse sizes, response times, and flashiness indices is required to verify these operational guidelines.

6 Limitations

This study evaluates the performance of RoC as a threshold for training data filtering, some limitations in methodology and datasets must be acknowledged.

First, the generalisability of the findings is limited by the small sample size of three UK catchments. Although the study of Walsden, Boscastle, and Hawnby allowed for a comparative analysis between fast-responding and slower-responding systems, a three-site study is still too small to be applied in all other areas. Verifying the broader applicability of this technique will require a larger number and diverse areas to potentially derive transfer functions based on hydrologically physical descriptors.

Second, a constraint of the model architecture is the cascading uncertainty introduced by the forecast rainfall inputs. The river level model was trained using the forecast rainfall from the rainfall model, rather than the observed future rainfall. Although this approach is a realistic operational simulation, it limits the river model's performance because any forecast errors generated by the rainfall model accumulate and propagate through the autoregressive forecasting steps.

Lastly, due to the low number of selected training events, particularly using higher-RoC thresholds, the testing dataset was used to monitor validation loss for early stopping. While this approach was necessary to prevent overfitting on small datasets, it implies that the testing data informed the stopping criterion, resulting in slightly optimistic absolute error metrics (e.g., RMSE). However, because identical architectures, hyperparameters, and early stopping protocols were applied across all RoC threshold configurations, this limitation does not invalidate the relative performance comparisons. The central finding, which is the trade-off in predictive accuracy between low- and high-magnitude events, is still valid.

7 Conclusion

Flash flood forecasting with deep learning is impacted by the composition of the training data. Extreme events are rare in continuous hydrological records, and conventional training on the full record allows low-magnitude observations to dominate the training data. This study introduces the rate of change of river level as a physically motivated criterion for selecting flood-relevant training events from continuous time-series and evaluates its effect on cascaded LSTM forecasts across three hydrologically distinct UK catchments.

The central finding is a trade-off in the choice of RoC threshold. There is no a single threshold that can optimise forecasting for both low- and high-magnitude events. Higher thresholds, which filter the training data toward rapid-rise events, improve forecast accuracy on higher-magnitude test events. At Walsden and Boscastle, the same set of high-magnitude test events for each catchment was forecast with median RMSE of 0.11 - 0.17 m when the model was trained at the threshold that optimised error on low-magnitude events. The median RMSE dropped to 0.05 - 0.09 m when the threshold was raised to the optimum RoC for high-magnitude events (0.25 - 0.30 m/180-min at Walsden and 0.20 - 0.30 m/180-min at Boscastle). The improvement on high-magnitude events came at the cost of accuracy on low-magnitude events, whose optimal training thresholds (0.05 - 0.10, 0.10 - 0.15, and 0.15 - 0.30 m/180-min at Walsden, Boscastle, and Hawnby, respectively) are below those optimal RoC for high-magnitude events. Therefore, the optimal choice depends on which part of the operation the model is intended to serve.

The strength of this trade-off, however, depends on catchment characteristics. At Walsden and Boscastle, which are smaller catchments with fast response times of 1.00 – 2.00 hours, higher RoC-based filtering improved higher-magnitude forecasting. At Hawnby, which is a larger catchment (131 km² versus 14 - 20 km²) with a slower 5.25-hour response time, the high-magnitude forecasting was insensitive to threshold choice across a broad range. This is consistent with low flashiness of larger catchments producing more uniform runoff at the outlet, reducing the performance of the rate of change to filter the event data between each event types. RoC-based event filtering is therefore most effective in small, fast-responding catchments which are also prone to flash flooding and where the technique is most operationally relevant.

The catchment dependence is both an insight and a limitation. The insight is that the technique's domain of value coincides with the operational need it is intended to serve. The limitation is that optimal thresholds cannot be transferred between sites without calibration, and the present study is restricted to three UK catchments. Extending the methodology to a larger and more hydrologically diverse areas would test whether the dependence on response time generalises and could open the possibility of estimating the optimal RoC threshold for unseen catchments directly from their physical descriptors. Further methodological work could also examine alternative metrics that respond to the timing or shape of the rising limb rather than its rate alone, which may capture aspects of event character that pure rate of change misses, particularly in larger catchments.

Nonetheless, the results establish RoC-based event extraction as a simple, interpretable, and physically grounded technique for addressing training-data composition in flash flood forecasting. Rate of change can be computed directly from a standard hydrological measurement and provides a tuneable filter that allows the training set to be shaped according to the operational target. In the small, flashy catchments where flash flooding is most consequential, RoC-based filtering offers a concrete path toward more reliable forecasts of the high-magnitude events whose accurate prediction has the greatest operational value.

Reference

[1] J. Buszta, K. W´ojcik, C. A. Guimar˜aes Santos, K. Koziol , and K. Maciuk, “Historical analysis and prediction of the magnitude and scale of natural disasters globally,” Resources, vol. 12, no. 9, p. 106, 2023.

[2] D. Teh and T. Khan, “Types, definition and classification of natural disasters and threat level,” in Handbook of disaster risk reduction for resilience: new frameworks for building resilience to disasters, Springer, 2021, pp. 27–56.

[3] K. Sene, “Flash floods,” in Hydrometeorology: forecasting and applications, Springer, 2016, pp. 273–312.

[4] V. Kumar, H. M. Azamathulla, K. V. Sharma, D. J. Mehta, and K. T. Maharaj, “The state of the art in deep learning applications, challenges, and future prospects: A comprehensive review of flood forecasting and management,” Sustainability, vol. 15, no. 13, p. 10543, 2023.

[5] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] L. B. Santos, E. V. Escobar-Silva, L. F. Satolo, Oyarzabal, R.S., Diniz, M.M., Negri, R.G., Lima, G.R., Stephany, S., Soares, J.A., Duque, J.S. and Saraiva Filho, F.L., “Machine learning-based hydrological models for flash floods: A systematic literature review,” Smart Construction and Sustainable Cities, vol. 3, no. 1, p. 21, 2025.

[7] E. Heidari, V. Samadi, and A. A. Khan, “Leveraging recurrent neural networks for flood prediction and assessment,” Hydrology, vol. 12, no. 4, p. 90, 2025.

[8] Z. Liu, N. Coleman, F. I. Patrascu, K. Yin, X. Li, and A. Mostafavi, “Artificial intelligence for flood risk management: A comprehensive state-of-the-art review and future directions,” International Journal of Disaster Risk Reduction, vol. 117, p. 105 110, 2025.

[9] M. Albared, H.-P. Beise, and M. Stu¨ber, “How confident and reliable are deep learning models for streamflow prediction under flood conditions,” Natural Hazards, vol. 121, no. 16, pp. 18529–18549, 2025.

[10] M. Luppichini, M. Barsanti, R. Giannecchini, and M. Bini, “Deep learning models to predict flood events in fast-flowing watersheds,” Science of the total environment, vol. 813, p. 151885, 2022.

[11] W. Li, A. Kiaghadi, and C. Dawson, “Exploring the best sequence lstm modeling architecture for flood prediction,” Neural Computing and Applications, vol. 33, no. 11, pp. 5571– 5580, 2021.

[12] M. Fu, T. Fan, Z. Ding, S. Q. Salih, N. Al-Ansari, and Z. M. Yaseen, “Deep learning data-intelligence model based on adjusted forecasting window scale: Application in daily streamflow simulation,” Ieee Access, vol. 8, pp. 32632–32651, 2020.

[13] X.-H. Le, H. V. Ho, G. Lee, and S. Jung, “Application of long short-term memory (lstm) neural network for flood forecasting,” Water, vol. 11, no. 7, p. 1387, 2019.

[14] A. H. Akinsoji, B. Adelodun, Q. Adeyi, R. A. Salau, G. Odey, and K. S. Choi, “Integrating machine learning models with comprehensive data strategies and optimization techniques to enhance flood prediction accuracy: A review,” Water Resources Management, vol. 38, no. 12, pp. 4735–4761, 2024.

[15] Y. Zhao, X. Wu, W. Zhang, Lan, P., Qin, G., Li, X. and Li, H., “A deep learning-based probabilistic approach to flash flood warnings in mountainous catchments,” Journal of Hydrology, vol. 652, p. 132 677, 2025.

[16] C. Chen, B. Li, H. Zhang, H., Zhao, M., Liang, Z., Li, K. and An, X., “Performance enhancement of deep learning model with attention mechanism and fcn model in flood forecasting,” Journal of Hydrology, vol. 658, p. 133 221, 2025.

[17] M. Jeong, M. Kwon, J.-H. Cha, and D.-H. Kim, “High flow prediction model integrating physically and deep learning based approaches with quasi real-time watershed data assimilation,” Journal of Hydrology, vol. 636, p. 131 304, 2024.

[18] T. Song, W. Ding, J. Wu, H. Liu, H. Zhou, and J. Chu, “Flash flood forecasting based on long short-term memory networks,” Water, vol. 12, no. 1, p. 109, 2019.

[19] S. Nevo, E. Morin, A. Gerzi Rosenthal, Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G. and Begelman, G., “Flood forecasting with machine learning models in an operational framework,” Hydrology and Earth System Sciences, vol. 26, no. 15, pp. 4013–4032, 2022.

[20] A. Bayliss,Catchment Descriptors. “Volume 5 of the Flood Estimation Handbook,” Centre for Ecology & Hydrology, 1999.

[21] 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, e2020WR028201, 2021.

[22] Calder Rivers Trust, River history, [Online]. Available: https://calderandcolneriverstrust. org/site/river-history/, [Accessed 03-02-2026].

[23] Internet Geography, "Flood Management Scheme – Boscastle.", [Online]. Available: https://www.internetgeography.net/geotopics/flood-management-scheme-boscastle/, [Accessed 02-06-2026].

[24] National Trust, "History of Boscastle.", https://www.nationaltrust.org.uk/visit/cornwall/boscastle/history-of-boscastle#rt-flood-alleviation, [Accessed 02-06-2026].

[25] Howden, N. J. K., and L. K. Deeks. "Soil Examination, Rainfall Simulation and Soil Runoff and Infiltration Experiments following a flood event in the Boscastle area.", 2007.

[26] BBC News, Flood Damage to Yorkshire Image, [Online]. Available: http://news.bbc.co.uk/1/hi/england/north_yorkshire/4646031.stm, [Accessed 02-06-2026].

[27] Met Office, Flooding in North Yorkshire, [Online]. Available: https://weather.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/weather/learn-about/uk-past-events/interesting/2005/flooding-in-north-yorkshire---19-june-2005---met-office.pdf, [Accessed 02-06-2026].

[28] Department for Environment, Food & Rural Affairs (Defra), Data services platform: Catchment data explorer, [Online]. Available: https://environment.data.gov.uk/ catchment-planning/WaterBody/GB104027062600, [Accessed 05-02-2026].

[28] Environment Agency. Walsden Water level at Walsden. GOV.UK, [Online]. Available: https://check-for-flooding.service.gov.uk/station/8211, [Accessed 02-06-2026].

[29] Environment Agency. River Valency level at Boscastle County Bridge. GOV.UK, [Online]. Available: https://check-for-flooding.service.gov.uk/station/3402, [Accessed 02-06-2026].

[30] Environment Agency. River Rye level at Broadway Foot. GOV.UK, [Online]. Available: https://check-for-flooding.service.gov.uk/station/8033, [Accessed 02-06-2026].

[31] A. Shirzadi, H. Shahabi, A. Salvati, Nodoushan, E.J., Hoseini, S.M., Tahan, M.H. and Clague, J.J., “Leveraging imbalanced dataset in urban flood susceptibility prediction: A case study of sanandaj city,” Journal of Hydrology, p. 134727, 2025.

[32] H. Wang, Y. Meng, H. Xu, Wang, H., Guan, X., Liu, Y., Liu, M. and Wu, Z., “Prediction of flood risk levels of urban flooded points though using machine learning with unbalanced data,” Journal of Hydrology, vol. 630, p. 130742, 2024.

[33] Moore, R.J. The PDM rainfall-runoff model. Hydrology and Earth System Sciences, 11(1), pp.483-499, 2007.

[34] Olin, B. and Beevers, L., River flashiness in Great Britain: A spatio-temporal analysis. Atmosphere, 15(9), p.1025, 2024.

8 Appendix

A. Selection of the differencing interval for RoC computation

The rate of change of river level is used in this study as a criterion for selecting flood events to train flood forecasting models. The choice of differencing interval (the time window over which the change in water level is computed) affects both the character of the events selected and the size of the training sample availabilities. To justify the potential RoC interval, a sensitivity analysis was carried out comparing four RoC intervals: 15, 60, 180, and 360 minutes (Table A1). These four intervals were selected to cover from instantaneous localized spikes to slower responses.

Table A1 Comparison of differencing time intervals and percentile thresholds for RoC-based event selection, validated against the 0.38 m the EA flood indicator. This testing performs on Walsden catchment. For each combination of differencing interval (15, 60, 180, 360 minutes) and percentile threshold (90th, 95th, 99th), the table reports the RoC threshold value (in m per interval and equivalent rate in m/h), the total number of events flagged in the study period, the number of these events that exceeded 0.38 m the EA flood indicator (classified as flood events), and the precision and recall against the 118 flood events using the EA flood indicator.

Interval	Percentile	RoC at that percentile (m/interval)	Equivalent RoC (m/h)	Total flagged events	Flood events (flagge events above 0.38 m)	Precision	Recall
15 min	90th	0.024	0.1	242	68	28%	58%
	95th	0.034	0.14	118	48	41%	41%
	99th	0.056	0.22	31	18	58%	15%
60 min	90th	0.092	0.09	133	53	40%	45%
	95th	0.125	0.13	63	38	60%	32%
	99th	0.185	0.18	21	13	62%	11%
180 min	90th	0.229	0.08	55	40	73%	34%
	95th	0.280	0.09	30	27	90%	23%
	99th	0.343	0.11	8	8	100%	7%
360 min	90th	0.313	0.05	39	39	100%	33%
	95th	0.367	0.06	16	16	100%	14%
	99th	0.446	0.07	5	5	100%	4%

For each time interval (15, 60, 180, and 360 minutes), the RoC time series was computed over the full study period for the Walsden catchment. Thresholds were established at the 90th, 95th, and 99th percentiles of each interval; this quantile-matched approach ensures consistent comparisons across the varying RoC intervals. Events that exceeded a given RoC threshold were designated as 'flagged events.' To evaluate their accuracy, these flagged events were cross-referenced against actual flood events, which were defined as instances where the river level reached or exceeded the EA flood indicator of 0.38 m. Across the entire historical record, this 0.38 m baseline identified 118 total flood events.The classification performance of each RoC threshold was evaluated using precision and recall metrics. Precision represents the proportion of flagged events that were actual floods (i.e., where river levels equal or above 0.38 m). Recall represents the proportion of the 118 actual flood events that were successfully detected by the given RoC threshold.

It should be noted that this cross-referencing is not an independent validation of RoC as a flood detector. Both methods operate on the same river-level record but with a different measurement (the RoC-based metric evaluates the rate of change, whereas the EA indicator evaluates absolute depth). RoC is preferred in the analysis because, unlike a river level flood indicator, the positive RoC responds only to the rising limb of the hydrograph and is therefore not subject to the over-counting that can occur from slow recessions crossing the river level threshold.

As the events identified by RoC intervals will be used as training data for forecasting models in the main analysis, a sufficient number of events to support model fitting and evaluation is necessary. The 360-minute interval yields only 16 events at the 95th percentile and 5 at the 99th (Table A1), which are low numbers of events for model training. Moreover, training data should be informative. The flagged events should be flood events rather than short-duration fluctuations. The 15-minute and 60-minute intervals can detect many events (21 - 242 events); however, their maximum precision did not exceed 62% at any percentile tested, meaning at least 38% of training samples would not correspond to flood events. Lastly, the optimum RoC interval should be long enough to ignore minor fluctuations of river levels that do not develop into a flood while remaining short enough to capture rapid-onset events. The 180-minute interval satisfies most of these requirements. At the 90th percentile it yields 55 flagged events with 73% precision (capturing 40 true flood events), and at the 99th percentile achieves 100% precision while still selecting a non-trivial number of events. Therefore, the 180-minute interval was adopted for the main analysis.