Journal of Hydrology 677 (2026) 135819

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

AI-assisted framework using physically informed rainfall–drainage features for real-time urban flood risk forecasting

Farzad Piadeh [a], Vahid Bakhtiari [b], Kourosh Behzadian [c,*], Farshad Piadeh [d]

[a] Centre for Engineering Research, School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield AL10 9AB, UK
[b] School of Architecture & Built Environment, Faculty of Science Engineering & Built Environment, Deakin University, Geelong, VIC 3220, Australia
[c] School of Computing and Engineering, University of West London, St Mary's Rd, London W5 5RF, UK
[d] School of Computer Engineering, Islamic Azad University of Mashhad, Ostad Yousefi Blvd., Mashhad 91871-47578, Iran

ARTICLE INFO

This manuscript was handled by Dan Lu, Editor-in-Chief, with the assistance of Zhiyong Liu, Associate Editor

Keywords:
Drainage systems
Dynamic mixture of expert
Ensemble modelling
Multi-class prediction
Real-time modelling
Urban flood forecasting

ABSTRACT

Urban flash flooding is becoming more severe as urban population density increases, and severe and frequent floods occur. As a result, early warning systems are required that translate raw sensor data into clear and actionable hazard situations, rather than simply providing ongoing forecasts. For this purpose, this study develops a hydrological- and hydraulics-informed framework for real-time multi-class AI-based flood warning across evaporation (low risk), drained (medium risk), and flooding (high risk) states. The methodology firstly derives physically guided rainfall features using a rule-based back-propagation neural network to estimate return-period signals alongside seasonality and antecedent-rainfall cues, while hydraulics-informed "water-level memory" (current class and class duration) captures system dynamics. These inputs feed seven weak-learner families whose outputs are fused by a time-series mixture-of-experts. The model performance is evaluated for lead times of up to 5 h using both multi-step metrics and event-based analyses. The framework is applied to the Ruislip urban drainage system, UK, using long-term IoT rainfall and water-level records (2011–2024). The results show discrimination improved compared with voting and averaging ensembles. It also increases hit rates for flood and non-flood decisions at 1–3 h. At 4–5 h, it reduces false alarms and late detections. Event-based results show more on-time hits and shorter timing lags. Feature analysis shows that rainfall intensity and duration are the main drivers. Seasonal effects and antecedent occurrence also provide added value. Residual errors concentrate at longer horizons e.g. 5 h later where transitions between adjacent states are intrinsically difficult.

1. Introduction

Flooding is one of the most severe natural disasters globally, with significant impacts on infrastructure, ecosystems, and communities (Bakhtiari et al., 2023). The frequency and intensity of urban flooding have increased significantly in recent years, primarily due to factors such as rapid urbanisation and climate change (Ferdowsi et al., 2024). Urban areas are especially vulnerable due to the proliferation of impermeable surfaces, which reduce water infiltration and exacerbate surface runoff during heavy rainfall (Tota-Maharaj et al., 2024). As a result, urban drainage systems (UDSs) are often overwhelmed, leading to more frequent and severe flood events that disrupt daily life, damage properties, and cause public health risks (Piadeh et al., 2023a). Across the last three decades, flooding has impacted nearly 2.6 billion people and led to more than 215,000 deaths, with global economic damage estimated at US$1.184 trillion (CRED, 2025). The increasing frequency and severity of such events emphasise the need for effective flood risk management strategies to protect urban areas.

Early warning systems (EWSs) are effective tools in flood risk management because they provide timely alerts before flood impacts become severe. They work by collecting monitoring data, processing that data through predictive models, and translating the outputs into actionable warnings for decision makers and communities (Bakhtiari et al., 2024). In general, their prediction mechanism follows one of two approaches: flood forecasting or flood risk classification (Byaruhanga et al., 2024). Flood forecasting predicts future water levels at specific lead times and compares them with critical thresholds to determine whether flooding is likely to occur (Almikaeel et al., 2025). Flood risk classification, in contrast, does not estimate the exact future water level. Instead, it predicts the likely flood state or risk category in the next time steps (Mishra

* Corresponding author.
E-mail address: kourosh.behzadian@uwl.ac.uk (K. Behzadian).

https://doi.org/10.1016/j.jhydrol.2026.135819
Received 2 January 2026; Received in revised form 8 May 2026; Accepted 1 June 2026
Available online 3 June 2026

et al., 2022). Although flood forecasting provides detailed information about water levels, flood risk classification offers several advantages. By simplifying predictions into risk categories, classification models can process data more quickly, allowing for faster response times, which is particularly valuable in real-time scenarios. Moreover, flood risk classification models are less sensitive to errors in water level predictions, making them more robust under uncertain conditions (Antwi-Agyakwa et al., 2023).

Flood classification has been developed using three main modelling approaches: physical models, artificial intelligence (AI) models, and hybrid models such as physics-informed neural networks (PINNs). Physical models formed the original basis of many flood EWSs. They represent flood behaviour using hydrological and hydraulic equations. This allows them to describe the physical processes of runoff generation and water movement in river or drainage networks. However, these models often need detailed input data, careful calibration, and high computational time. Their performance can also decline when data are limited or when the system is highly complex (Pandi et al., 2021). AI models were later introduced as a faster alternative. These models learn patterns directly from historical data and can produce rapid predictions (Raissi et al., 2019). They are often effective for real-time applications, but they may have limited physical interpretability and may perform poorly outside the conditions seen in training data. More recently, hybrid models such as PINNs have been proposed to combine the strengths of both approaches. These models connect neural networks with physical knowledge (Bentivoglio et al., 2022). They can do this by using physically meaningful input variables, embedding governing equations into the training process, or constraining the model outputs to remain physically consistent. In this way, PINNs can improve robustness, reduce overfitting, and provide more reliable flood prediction in complex and data-limited environments (Chew et al., 2025).

Recent studies have shown the growing value of PINN models in flood EWSs. However, the reviewed studies have mainly focused on continuous flood forecasting and using the outputs for flood routing or inundation mapping, rather than flood risk classification into future risk states. To begin with, Qian et al. (2019) integrated PINN with the shallow water equations for real-time urban flood forecasting. Bojović et al. (2022) applied PINN to one-dimensional flood wave propagation in open channels. Vongkusolkit (2022) proposed a weakly supervised physics-informed framework for near real-time flood mapping from remote sensing data. Feng et al. (2023) used PINN to improve flood wave simulation and large-scale river forecasting. Yang et al. (2024) coupled hydrodynamic modelling with deep learning to rapidly predict urban inundation maps. Donnelly et al. (2024) developed a physics-informed surrogate model to improve forecasting efficiency and accuracy. Taghizadeh et al. (2025) embedded mass conservation into a graph-based physics-informed model for river flood forecasting. Taken together, these studies confirm the strong potential of physics-informed models for flood prediction. Yet they also show that the literature has largely concentrated on forecasting continuous hydraulic variables, not on directly classifying future flood states for early warning decisions. Despite this progress, developing a PINN model for flood risk classification remains more challenging than predicting a continuous variable such as water level. Classification requires the model to distinguish between adjacent risk states, handle imbalanced class distributions, and identify the correct transition timing between non-flood, moderate-risk, and flood conditions. These difficulties become greater at longer lead times, where class overlap and transition uncertainty increase.

To address this gap, the present study introduces a hydrology- and hydraulics-informed multi-class flood risk classification framework for real-time early warning. The framework is innovative because it combines physically guided rainfall features, hydraulics-informed water-level memory, and a time-series mixture-of-experts structure to classify future flood states rather than only forecast water levels. It also contributes a dynamic lead-aware decision strategy that improves class discrimination across different forecast horizons and an event-based evaluation perspective that better reflects operational warning performance.

In the scientific domain, the study advances the application of PINNs from flood forecasting to flood risk classification and extends data representation from conventional AI modelling to physics-informed data integration. In addition, a combination of expert and agent models is developed to enable non-binary and multi-class classification. From a practical and engineering perspective, the main objective is to support faster and more reliable early warning decisions for flood emergency managers by improving detection timing, reducing false alarms, and strengthening real-time risk interpretation in urban drainage systems.

2. Methodology

The proposed framework follows a three-phase five-step methodology as shown in Fig. 1a. Phase 1 comprises two steps including four hydrology-informed features along with two conventional rainfall features (Step 1) and two hydraulics-informed features (Step 2) to feed the pre-trained models. Data aging techniques are combined with a novel rule-based back propagation neural network (BPNN) model to generate these features. In this study, the BPNN differs from a conventional BPNN by integrating a rule-based component derived from intensity–duration–frequency (IDF) relationships, and retrainable error correction rules. While the neural network learns rainfall patterns from historical data, the rule-based module constrains the predicted return-period classes according to physically defined rainfall thresholds.

Phase 2 applies agent-based modelling to develop multi-class, time-series-based models, which are then tested on unseen data (Step 3). The results are collected in a performance-based data tesseract warehouse that enables the integration of mixture-of-experts models for enhancing prediction accuracy (Step 4). Phase 3 (Step 5) acquires real-time data from sensors through an application programming interface (API) and transformed into the selected feature format. These features are dynamically fed into the flood forecasting system to generate time-series flood risk assessments. The system produces flood forecasting for different time leads with associated flood risk factors by using event-based performance assessment that is served as a key component of the EWS.

The proposed framework follows a three-phase, five-step methodology, as shown in Fig. 1a. Each phase has a clear role in the flood risk classification workflow. Phase 1 focuses on feature generation. It prepares the hydrology- and hydraulics-informed inputs required for the classification models. Phase 2 focuses on model development and integration. It builds the multi-class time-series flood risk classification models, evaluates them on unseen data, and combines them through a mixture-of-experts strategy. Phase 3 focuses on real-time deployment. It receives live sensor data through an application programming interface (API), converts them into the required feature format, and delivers flood risk classifications for different lead times as part of the EWS.

In Phase 1, Step 1 generates six rainfall-related inputs, including two conventional rainfall features and four hydrology-informed features. Two of these new features are derived using data aging techniques, and two are generated by a rule-based back propagation neural network (BPNN). Step 2 then adds two hydraulics-informed features that describe water-level memory. In Phase 2, Step 3 develops the multi-class time-series flood risk classification models and tests them on unseen data. Step 4 stores their class-wise performance in a data tesseract warehouse and uses this information to build the mixture-of-experts model. In Phase 3, Step 5 applies the selected framework to real-time sensor data and produces multi-lead flood risk classifications with event-based performance assessment for operational early warning.

PINNs are typically divided into three main types, depending on how they incorporate physical principles into the neural network structure. The first type of PINN affects the input, where physical variables such as rainfall intensity, flow velocity, and runoff parameters are incorporated directly into the model. This integration ensures that the network is

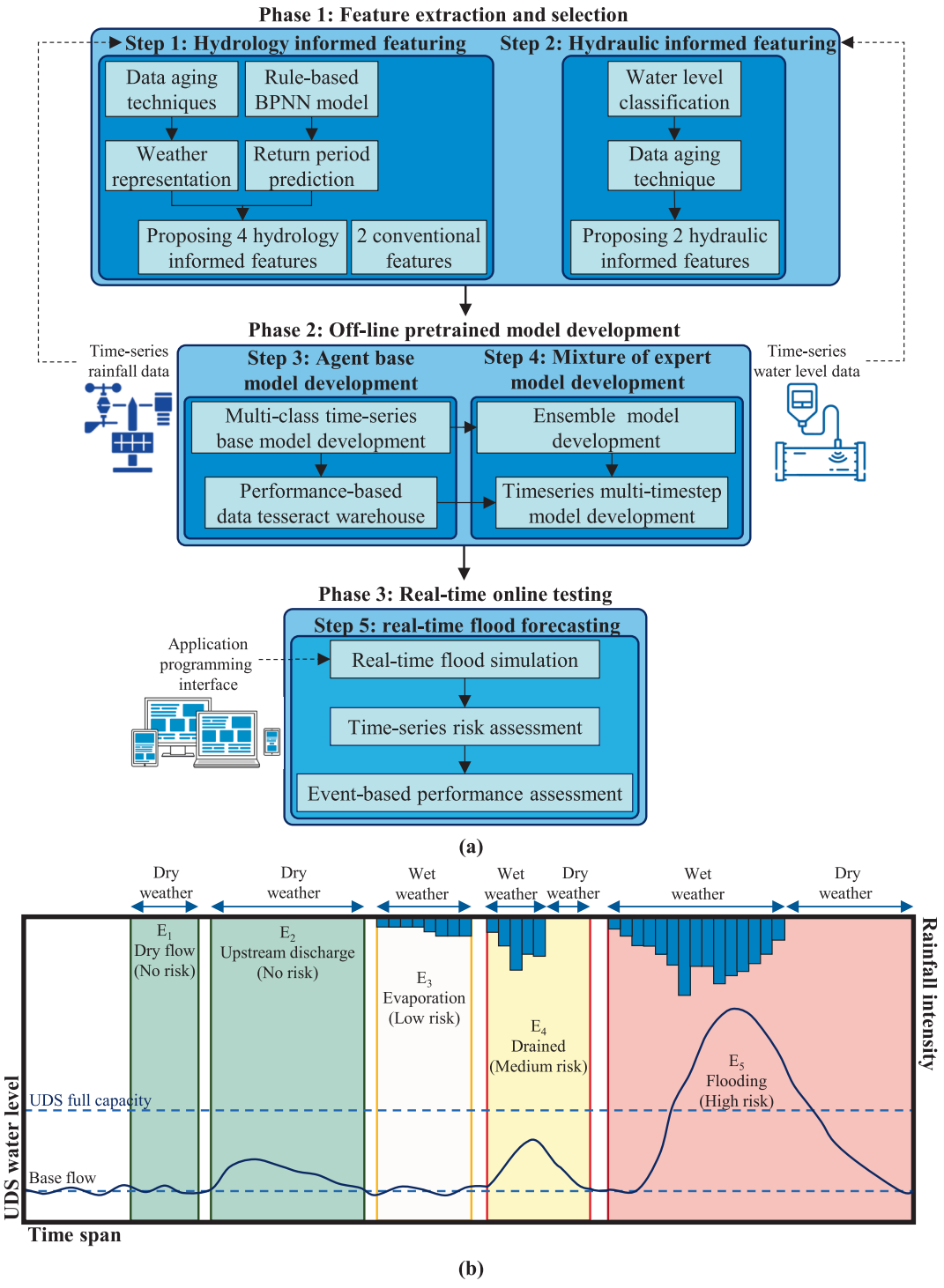
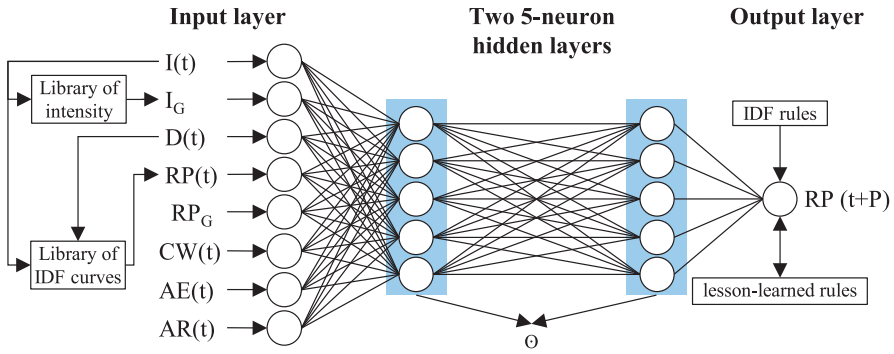


Fig. 1. Schematic structure of proposed framework for (a) multi-class real-time AI-based flood forecasting, (b) risk level classes E1-E5 in multi-class data mining.

guided by real-world data and hydrological insights, enhancing its predictive capabilities for flood events (Xia and Meng, 2024). The second type incorporates governing equations – such as the Navier-Stokes equations for fluid dynamics – into the network's structure. This allows the network to learn within the constraints of physical laws, improving accuracy by ensuring that predictions adhere to established principles of fluid behaviour during flood events (Qian et al., 2019). By embedding these laws into the network's architecture, the model adjusts its internal weights and biases according to physical dynamics, providing a more realistic simulation of flood propagation (Lin et al., 2020). The third type of PINN affects the output, making it rule-based by applying constraints based on physical laws such as mass and momentum conservation (Wang *et al.*, 2023). This approach ensures that predictions remain physically feasible, particularly when data is incomplete or sparse. This rule-based approach helps in real-time flood forecasting by maintaining consistency with physical laws, ensuring more reliable flood risk classifications.

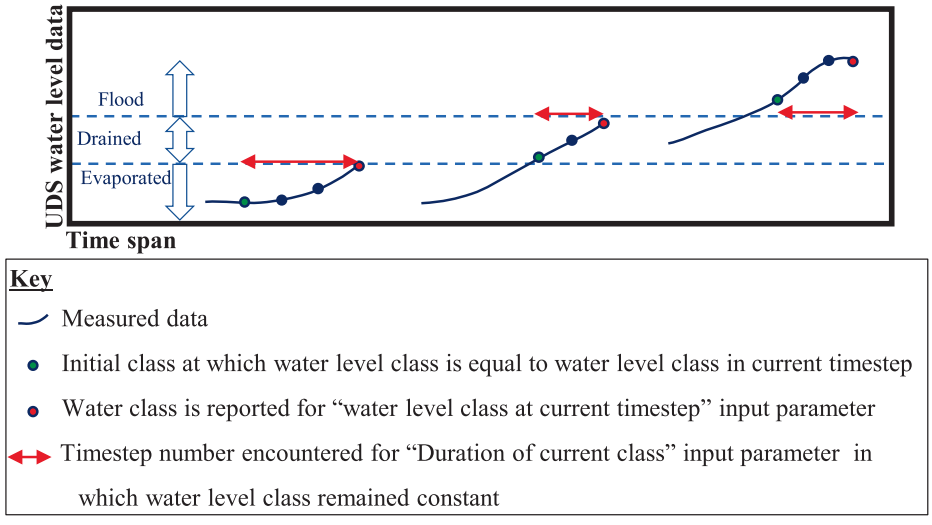
The model adopts a four-class flood risk classification scheme (Fig. 1b), adapted from Piadeh et al. (2023b), because these classes reflect the main operational states of the drainage system and support graduated early warning decisions. Class 0 represents dry weather conditions, in which no rainfall is detected, and the framework remains

inactive (E1 and E2 in Fig. 1b). This class prevents unnecessary model activation when no response is required. The system is triggered only when initial rainfall intensity is recorded by rain gauge sensors. Once activated, the framework classifies the target lead time into three operational flood risk classes. Class 1 represents the evaporation stage, where rainfall occurs but has no impact on water level in the UDS, and therefore no water-level change is observed (E3). This class indicates that rainfall has begun, but no intervention is yet needed. Class 2 denotes moderate risk, where rainfall affects the system, but the water level remains below full UDS capacity (E4). This class provides an early warning stage that can support closer monitoring and preparatory action. Class 3 indicates high risk, where the UDS exceeds full capacity and



Code	Parameter	Description	Transformation key	Unit/class
I	Intensity	The ratio of total depth to the duration	Numerical	mm/hr
IG	Intensity gradient	$I_t / I_1 \begin{cases} -1 \; if \; I(t) \geq I_{ave} \\ 1 \; f \; I(t) < I_{ave} \end{cases}$	Numerical	mm/mm
D	Duration	Time period of between the onset and end of the precipitation	Numerical	min
RP	Current RP	Class of RP for timestep t	Class	1-7
RPG	RP gradient	RP_t / RP_1	Numerical	-
CW	continuous wavelet transform	$\frac{\sum_{i=2}^{t}(R_i - R_{i-1})^2}{\overline{R}}$		mm
AE	Absolute energy	$\frac{\sum_i R_i^2}{R}$	Numerical	mm
AR	Anthropic	$\sum_i P(R_i) \times \log_2 R_i$	Numerical	-
RP	Predicted return period	RP of the rainfall for timestep of t	Class	No
Θ	Back propagation Bias	Biases of nodes in hidden layers related to backpropagation step		
t	Current timestep	-	-	-
P	Timestep ahead	Number of timestep ahead for prediction in 15min intervals	Numerical	-

(a)



(b)

Fig. 2. Structure of proposed new features for (a) rule-based BPNN model for rainfall RP prediction, (b) proposed water class data featuring.

flooding occurs (E5). This class supports immediate response and warning activation. This four-class structure is operationally meaningful because it separates inactive, low-impact, pre-critical, and critical system states rather than treating all wet conditions in the same way. As a result, the framework can support stepwise decision-making, from remaining idle in dry periods, to monitoring early rainfall response, to preparing for capacity exceedance, and finally to issuing urgent flood warnings.

The system is designed to operate using numerical data generated by IoT sensors and ground-based remote sensing. Image processing could be incorporated in Phase 1 if satellite imagery were employed; however, the current scope is restricted to numerical data sources, reflecting the widespread availability of API-based numerical data from ground monitoring systems (Girottoa et al., 2024; Bakhtiari et al., 2025). The present application also focuses on fluvial flooding, but the framework could be readily extended to other types, including pluvial, flash, and coastal flooding.

2.1. Step 1: hydrology-informed featuring

Here, the three different types of hydrological inputs used to base on their role in representing rainfall dynamics: (1) conventional rainfall features, (2) temporal features, and (3) return-period-based features. Conventional rainfall features describe the immediate characteristics of rainfall and provide the baseline information for model input. Temporal features capture the influence of time-dependent rainfall behaviour, including seasonal variability and antecedent precipitation conditions, which help represent both long-term and short-term rainfall memory in the system. Return-period-based features provide physically informed descriptors of rainfall severity by linking observed rainfall patterns to IDF relationships.

Hydrological features include two conventional rainfall features (i.e. rainfall intensity in mm (I) and rainfall duration classified in 15-minute timesteps as interval of model prediction (D)) and four new features as outlined here. Two of these new features are derived from temporal variability using data-aging techniques. The first represents long-term temporal effects by incorporating the seasonal occurrence of events (S), classified into three categories (dry, mild, and rainy) based on the Köppen climate classification (DEFRA, 2025). This feature applies seasonal relevance filtering (adapted from Chen et al., 2015) combined with a monthly period forgetting factor (adapted from Prasanthi et al., 2025). The second represents short-term temporal effects by accounting for antecedent precipitation history (A), defined as the average intensity of prior rainfall events up to a maximum look-back period equal to the time of concentration of the catchment. This feature follows an adaptive ('refreshing') feature approach adapted from Kithulgoda et al. (2018) outlined below.

Additional two new features are derived by developing a new rule-based BPNN model (as shown in Fig. 2a) to obtain the return period of current rainfall (RP) and the potential return period for future timesteps across multiple future time steps (RPt). BPNN is selected here for its proven ability to model continuous phenomena such as rainfall, capture complex input–output relationships, and automatically extract features e.g. IDF curves (Lillicrap et al., 2020). The RP is encoded as an integer from 1 to n, where 1 denotes the shortest RP and n the longest; a value of 0 is assigned when rainfall has ceased and RP is no longer applicable. RP values provide the model with information on the position of a rainfall event on the IDF curve, reflecting its relative frequency and magnitude at time t.

The BPNN model comprises two layers, each with five nodes, employing the back propagation process across all hidden-layer nodes. This architecture follows the recommendations of Lillicrap et al. (2020). The IDF rules, which define the maximum boundaries (i.e., timesteps) for each IDF curve, are converted into an if–then structure, as suggested by Park et al. (2023). This structure enables the comparison of predicted RP classes against the maximum measured IDF curves and allows the model's response to be adjusted accordingly. To generate IDF curves for various rainfall levels, the curve fitting toolbox in MATLAB 2025a is applied, fitting a general power equation as described by Hamil (2011) ($Y = aX^b$). A coefficient of determination (R^2) threshold of 0.9 or higher is used as a stopping criterion during optimisation to ensure the curves accurately represent the intensity − duration relationship.

The rule toolbox including a "lesson-learned" routine is considered to update the rules when a higher maximum timestep is observed for a specific IDF curve during real-time testing the model. In such cases, both the model and the associated rules are redefined. The rule-update mechanism is activated only when new rainfall observations exceed the previously defined maximum timestep boundaries of the IDF curves. In such cases, the rule base is updated by extending the corresponding IDF limit, and the BPNN model is retrained to incorporate the new rainfall-duration relationship.

This update process occurs outside the main model evaluation phase to ensure that performance metrics are not influenced by adaptive rule modification during testing. Therefore, all reported performance results are obtained using a fixed rule set and model configuration, ensuring that the evaluation reflects the true predictive capability of the framework rather than improvements introduced through online retraining. This adaptability is particularly relevant under changing climatic conditions, where an increase in high-intensity, short-duration rainfall events can shift RP values from longer to shorter ranges.

2.2. Step 2: hydraulics informed featuring

Two new hydraulic features are considered to provide complementary insights into system dynamics, thereby enhancing the model's ability to capture both short- and long-term memory. The first feature, the "water level class at the current timestep" (Wc) enables the model to identify the precise risk category at any given moment, improving the accuracy of immediate condition assessments. At each timestep, the water level class is assigned a value of 0, 1, 2, or 3, corresponding to the risk levels defined in the early warning stages (Fig. 1a). The second feature, the "duration of the current class" (Dw), records the number of consecutive timesteps during which the water level remains within the same risk category (Fig. 2b). This parameter allows the model to track the persistence or stability of risk levels, thereby revealing patterns of consistency or variability over time.

2.3. Step 3: agent based model development

Seven weak learner data mining models (WLDMs) (detailed provided in Table A1 in the Appendix A) were selected for their demonstrated potential and widespread use in previous hydrological classification studies (Piadeh et al., 2023b): Decision Tree (DT), K-Nearest Neighbourhood (KNN), Naive Bayes (NB), Neural Network Pattern Recognition (NPR), Support Vector Machine with Error-Correcting Output (SVM), Discriminant Analysis (DA), and Gaussian Process Regression (GPR). All models were developed and optimised in MATLAB 2025a using Phase 1 features for flood forecasting in UDS for lead times of interest. Model parameters were tuned using automated hyperparameter optimisation, minimising the 5-fold cross-validation loss over 30 iterations (demonstrated for the one-timestep-ahead forecasts in Fig. A1 in the Appendix A). Identified events were randomly distributed across training and validation datasets to ensure balanced representation. Each WLDM was optimised individually for every lead time, resulting in 140 trained models − seven WLDMs for each of the 20 lead times. The code components required for building the proposed ensemble model are stored in a dedicated library. This library enables easy training and testing of the ensemble model using the base models and the additional input parameters mentioned earlier.

To evaluate the performance of these models and to establish the mixture-of-experts model, a multi-class confusion matrix was used. Fig. 3a illustrates the transformation of forecasted data into a confusion

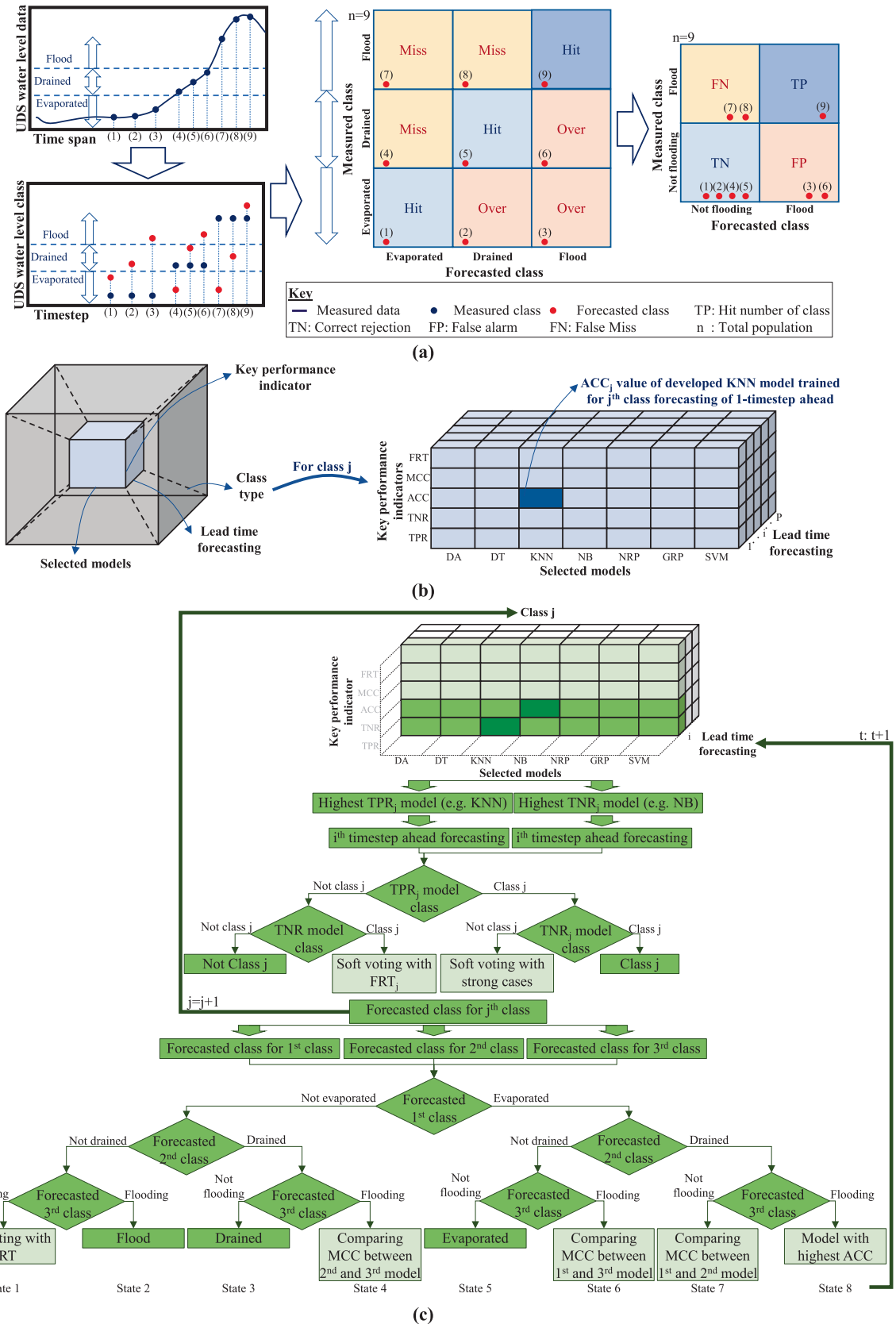


Fig. 3. Schematic illustration of (a) turning forecasted data into the confusion matrix demonstrating for flooding class, (b) constructed data tesseract warehouse, demonstrated for jth class, (c) flowchart of proposed mixture of expert model demonstrating for forecasting i-timestep ahead.

matrix, with a focus on the flooding class. Unlike binary confusion matrices, multiclass versions present additional challenges due to the need to evaluate multiple classes simultaneously. Assessing model performance requires calculating and interpreting multiple metrics for each class individually, making concise overall performance summaries more difficult (Heydarian et al., 2022). Furthermore, imbalanced class distributions – where some classes have significantly more instances than others – are common in multiclass classification problems. Such imbalances can bias performance metrics toward the majority class, potentially overstating accuracy while masking poor performance in minority classes (Markoulidakis et al., 2021).

To address these challenges, a one-versus-rest interpretation of the confusion matrix is applied when calculating class-specific performance metrics. In this approach, each class is evaluated individually by temporarily grouping the remaining classes into a single "non-target" category (Fig. 3a). For example, when evaluating the flooding class, true positives correspond to correctly detected flooding events, true negatives represent correctly detected non-flooding conditions (evaporation and drained), false negatives represent missed flooding events, and false positives correspond to overestimated flooding predictions. This dimensional reduction is used solely to define and calculate the performance indicators (as part of input feeding to the model only) and does not alter the multi-class structure of the proposed framework and multi-class evaluation of test data.The evaluated metrics include the true positive ratio (TPR), true negative ratio (TNR), accuracy (ACC), Mathew's correlation coefficient (MCC), and the non-parametric Friedman ranking test (FRT), as detailed in Equations A1-A5 in the Appendix A. These metrics are computed for each WLDM and for each class, then stored in a data tesseract warehouse (Fig. 3b) for subsequent integration into the mixture-of-experts model.

2.4. Step 4: mixture of expert model development

To develop the ensemble model, three forecasted classes are defined for each lead time. For each class and lead time, the two models with the highest levels of expertise – demonstrated by the best TPR and TNR – are selected. These models are then used to predict the class, with a decision-making process (upper decision tree in Fig. 3c) determining which of the two forecasts to accept for class j (where j = 1, 2, 3 corresponds to low, medium, and high water-level risk, respectively). This process is repeated for all three classes, and the resulting forecasts are passed to the mixture-of-experts model (lower decision tree in Fig. 3c) to determine the final prediction for each lead time.

The framework follows a set of rules to select the appropriate class in which eight possible decision pathways has resulted: (1) If one forecasted class aligns with its paired model, that class is chosen (states 2, 3

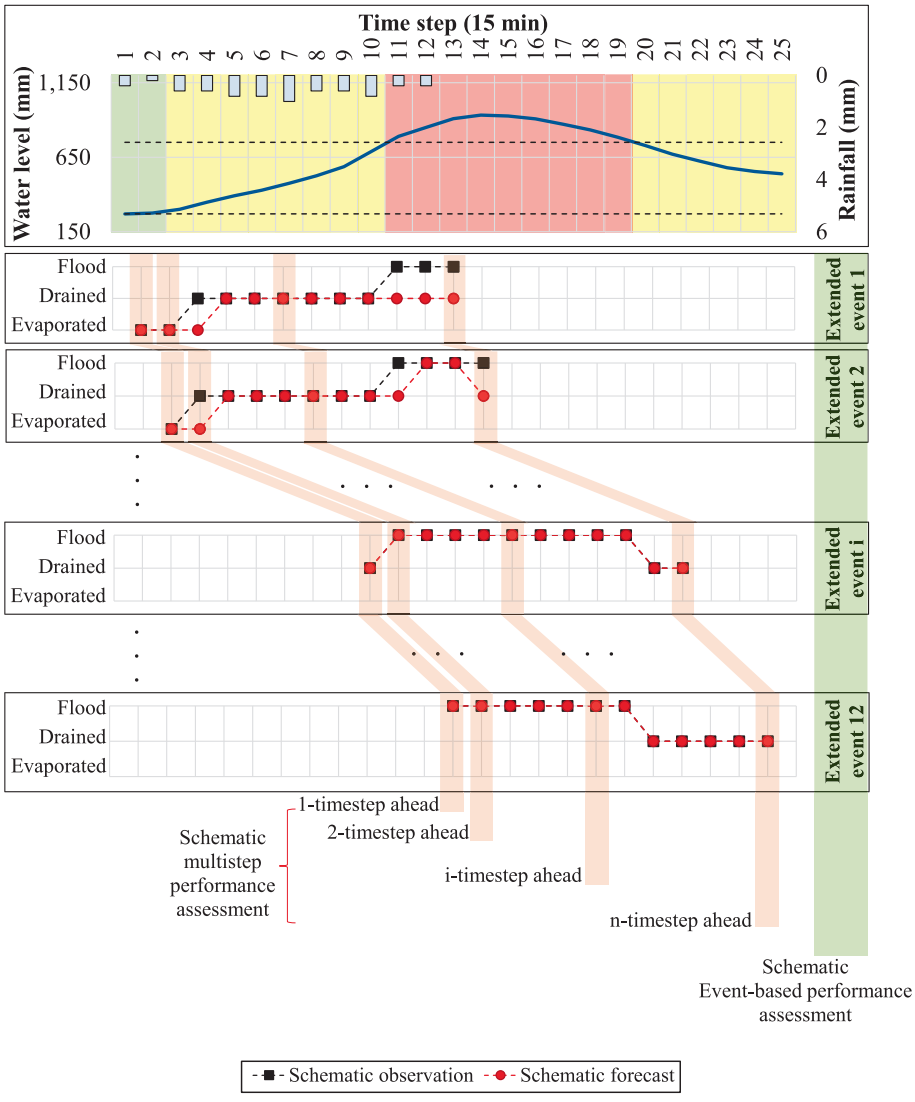


Fig. 4. Schematic illustration for comparison between event-based and multi-step performance assessment used for performance evaluation.

and 5 in Fig. 3c). For example, if the first forecast predicts evaporation, the second predicts not drained, and the third predicts flooding (state 5 in Fig. 3c), the selected class is evaporation; (2) If all models predict their respective classes, the class from the model with the highest ACC is selected (state 8 in Fig. 4); (3) If none of the models correctly predict their classes, the final decision is made through soft voting weighted by the FRT metric (state 1 in Fig. 3c); (4) If two models strongly support their respective classes without consensus, the class from the model with the highest MCC is chosen, as MCC provides a robust measure of agreement between predicted and actual classes (States 4,6,7, and 8 in Fig. 3c).

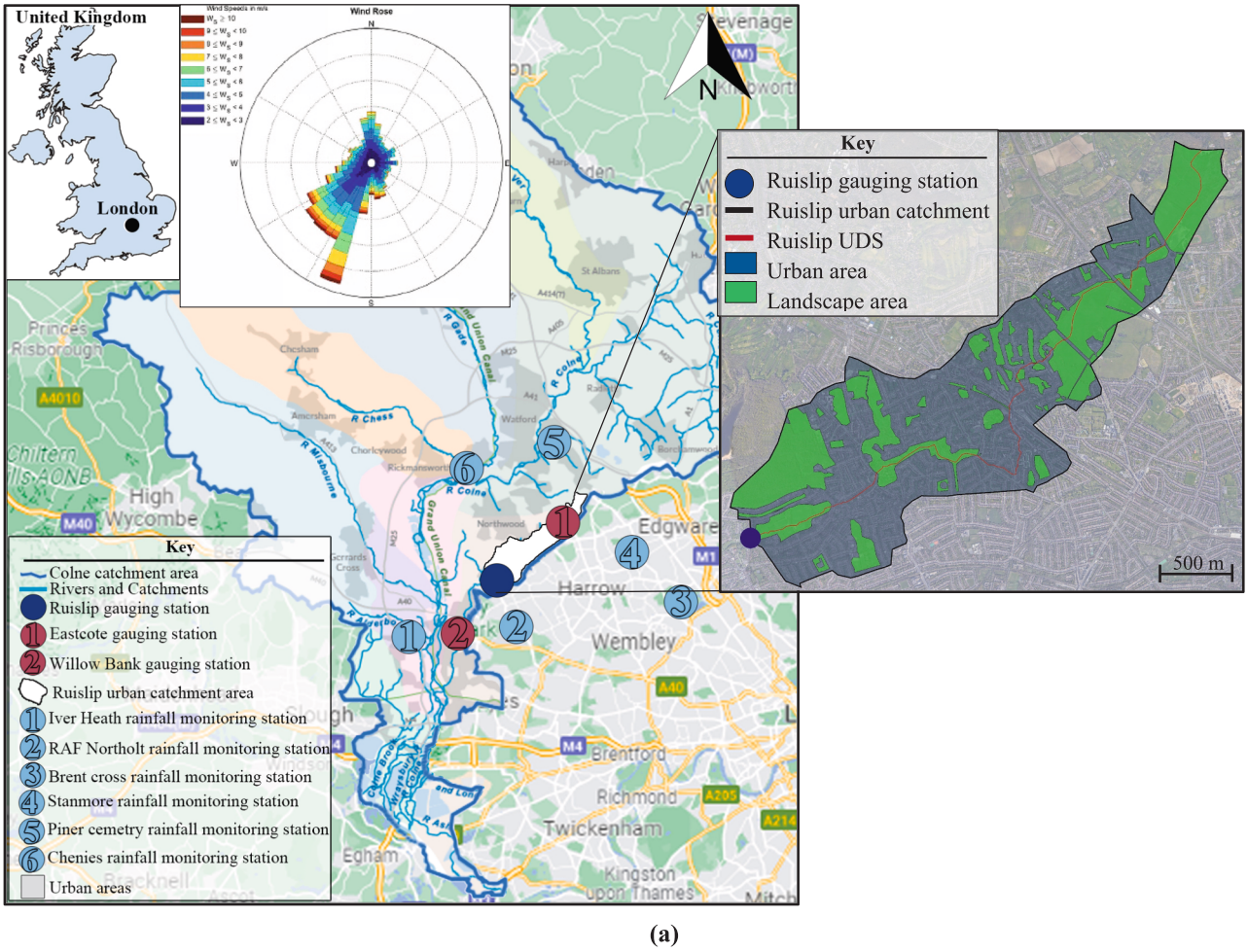


Fig. 5. Case study characteristics: (a) geographical location of the pilot study including location of case study catchment and monitoring stations, and layout of Ruislip UDS and catchment, (b) Flood event assessment between Ruislip water level rising and RAF Northolt rainfall events, and (c) IDF curve of RAF Northolt rainfall station.

2.5. Step 5: event-based and multi-step performance evaluation

The proposed framework performs multi-class flood risk classification, predicting the future flood state of the system at specified lead times. The lead-time intervals correspond to the temporal resolution of the input data but started form one step ahead to n-step ahead. The model is evaluated using continuous time-step sequences with temporal progression (2021–2024) to simulate real-time operational conditions, as illustrated in Fig. 4.

In this regard, two different performance evaluation concepts are applied here as schematically shown in Fig. 4. The first is the conventional multi-step prediction, where the model's performance is assessed separately for each forecast lead time across all events. The receiver operating characteristic (ROC) curve is also employed for validation, using its corresponding area under the curve (AUC) and optimal threshold. The second is the proposed event-based performance assessment, inspired by Piadeh et al. (2023b), which evaluates the model's performance over the entirety of individual extended events. In multi-step assessment, the accuracy performance of flood forecasting is measured for the same lead time across multiple extended events derived from a single flood event. In contrast, an event-based assessment classifies each extended event into one of the situations (See Table A2 and Figs. A3 and A4 in the Appendix A for more details), with the total count of these situations representing the overall model performance.

Finally, to provide a more detailed and balanced evaluation, class-wise and event-based performance analyses are included. Specifically, per-class accuracy along with total hit rate, miss-overestimation, and miss-underestimation, as well as AUC values for each class are provided to allow for a more comprehensive assessment of model performance across different flood-risk categories and help to account for the effects of class imbalance and asymmetric prediction errors.

3. Real case study demonstration

The Ruislip urban catchment (Fig. 5a), located in the London Borough of Hillingdon, was selected as the pilot study due to its high frequency of fluvial flooding. It conveys surface runoff from the Colne catchment in south Hertfordshire to a tributary of the River Thames, covering approximately 13 km^2 with predominantly open channels passing through parkland (Fig. 5b). Water levels are monitored every 15 min at the Ruislip gauging station on the River Pinn, one of 55 stations in the Colne catchment, using an ultrasonic IoT-based depth sensor installed in 2009 (DEFRA, 2025). Urban flooding is assumed to occur when levels exceed 850 mm, based on DEFRA's 2025 threshold. Rainfall is measured at six IoT-based tipping-bucket gauges (Fig. 5a), with stations chosen for proximity to the UDS.

The dataset contains 365,233 paired rainfall and water-level observations from 2011 to 2024 at 15-minute intervals, accessed via the UK Environment Agency API. Cross-correlation and cross-covariance analyses (Inspired by Paideh *et al.,* 2023a) identified RAF Northolt (See Fig. B1 in the Appendix B) as the representative rainfall station due to its correlation results and prevailing south-westerly wind direction. Event identification to convert the water level to flood risk classification (as shown in Fig. 5b and more details are provided in Fig. B2 in the Appendix B) revealed 5,679 wet-weather events (29,123 timesteps), comprising 595 low-risk evaporation events (8,151 timesteps), 1,218 medium-risk drained events (11,496 timesteps), and 505 high-risk flood events (5,404 timesteps). These analyses were applied only to the training and validation datasets to prevent overfitting. Initially, RP were estimated from the fitted IDF relationships for the available duration intervals derived from the 15-minute rainfall data. However, due to the limited number of extreme observations in the historical dataset, rainfall events with estimated return periods greater than 5 years were grouped into the 5-year RP category. Similarly, rainfall events associated with very short recurrence levels were consolidated into the 1-week RP class to maintain statistically meaningful bins for the classification model. This bin merging was applied to avoid unstable RP estimates in the extreme tails of the distribution while preserving representative rainfall intensity–duration patterns for model training. The fitted IDF curves used to derive the RP values are presented in Fig. B3, and the corresponding parameters are summarised in Table B1 in Appendix B.

4. Results and discussion

The selected test cases were chosen to ensure representative coverage of typical urban flooding conditions within the study area. Specifically, the events include a range of rainfall characteristics, such as short-duration high-intensity storms and longer-duration moderate rainfall events, capturing variability in both intensity and temporal patterns. In addition, the selected cases reflect different antecedent conditions and seasonal variations, which are known to significantly influence urban drainage response and flood generation. By including events that lead to both rapid surface runoff and progressive system saturation, the test set represents a spectrum of flood-risk scenarios from minor disturbances to severe flooding conditions.

4.1. Benchmark models

To ensure a robust comparison, additional ensemble methods were tested. While the proposed approach uses stacking, other stacking configurations were considered, as shown in Table B4 in the Appendix B. To further evaluate generalisation capability, Bootstrap Aggregating (Bagging) and gradient boosting regression (GBR) methods were also tested. Model selection at each timestep was optimised using MATLAB 2025a's classification and optimisation toolboxes. All benchmark models are trained and validated using the same original database and the introduced features used for training and validation of the proposed methodology (More details are provided in Tables B5-6 and Fig. B2 in the Appendix B).

All benchmark models were configured to perform the same prediction task as the proposed framework, namely multi-class flood risk classification, predicting the future system state (evaporation, drained, or flooding) at each forecast timestep rather than estimating continuous water-level values. The models use the identical input features generated in Phase 1 of the proposed framework, including conventional rainfall descriptors (rainfall intensity and duration), temporal rainfall features, and return-period-based features derived from the IDF relationships. To ensure fair benchmarking, class imbalance was addressed through ensemble configurations that incorporate sampling strategies, such as RUSBoost, which applies random undersampling to improve the representation of minority classes during training. This consistent experimental setup ensures that differences in performance arise from the learning algorithms rather than differences in data inputs or preprocessing procedures.

4.2. Performance of rule-based BPNN model

The accuracy of the BPNN model is depicted in Fig. 6, illustrating its performance across selected timesteps. The results indicate that the model achieves a high accuracy of 90% or above for all types of rainfall in the 15 min ahead prediction. Notably, it exhibits slightly superior performance in identifying dry weather and 1-year RP rainfall, achieving a remarkable accuracy of 94% (Fig. 6a). When specifically examining the model's ability to predict dry conditions, its accuracy for this class ranges from 94% in the 15 min ahead prediction to 78% in the 4 h ahead prediction, which is considered quite substantial and acceptable. However, it is important to note that the accuracy drops further to 60% in the 5 h ahead prediction. In contrast, the accuracy for predicting RP rainfall diminishes significantly, ranging from 13% to 35% at this timestep, which is deemed unsatisfactory.

Overall, the accuracy of predictions remains acceptable until 3 h ahead, with reported accuracies exceeding 75% for all time steps within

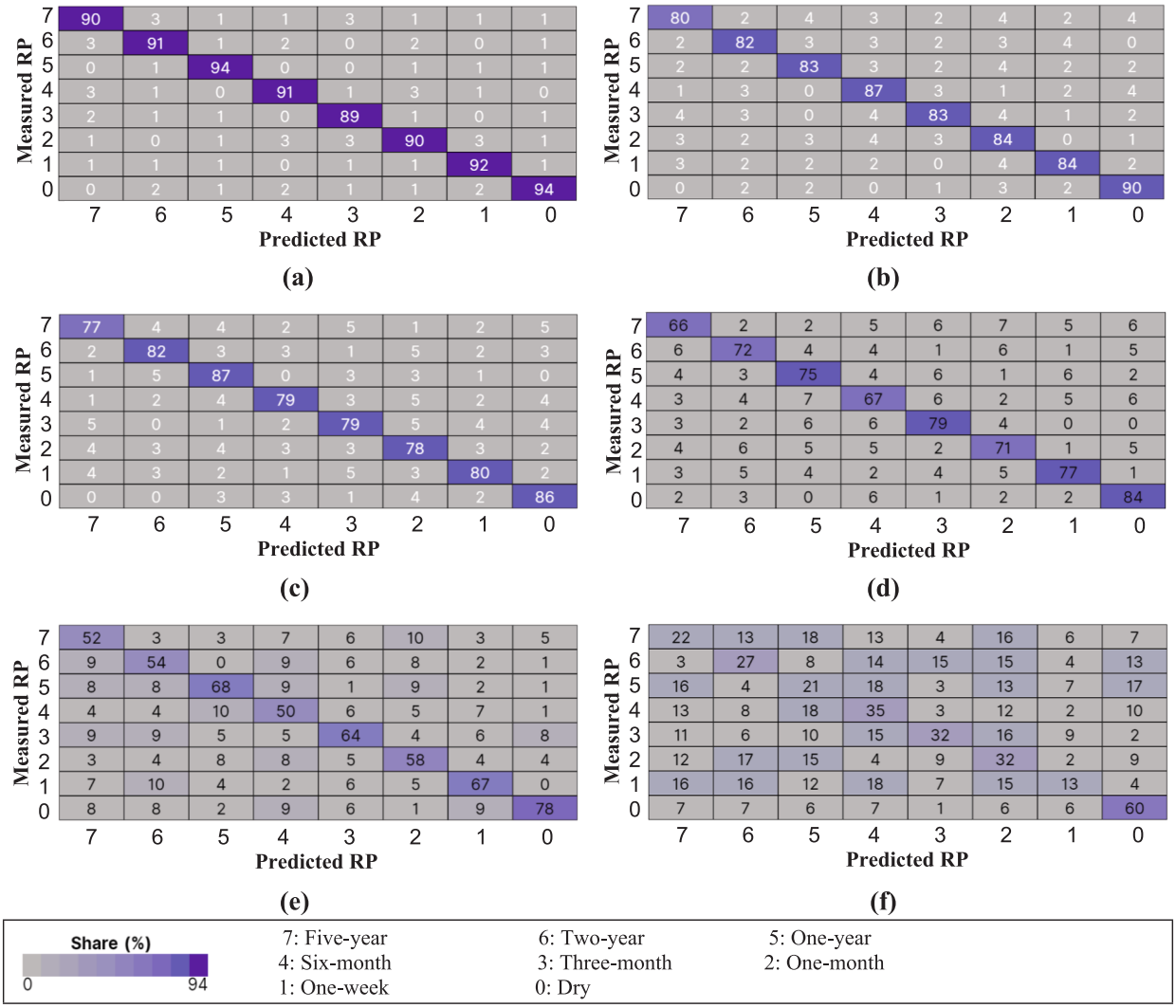


Fig. 6. Confusion matrix of BPNN model performance for: (a) 15 min, (b) 1 h, (c) 2 h, (d) 3 h, (e) 4 h, and (f) 5 h ahead of RP prediction.

this range. However, the accuracy declines to 50–68% and 13–35% for the 4 h and 5 h ahead predictions, respectively. Therefore, it becomes evident that the input parameter for multi-class stacking plays a significant role in prediction accuracy, particularly up to 3 h ahead. Analysing the accuracy distribution across different RP levels of rainfall reveals that the model performs better for higher classes, representing rainfall events with RP periods ranging from 1 to 5-year. However, its performance decreases when dealing with lower classes, i.e., rainfall with RP periods ranging from 1 to 6 months. This can be attributed to the fact that rainfall with shorter RP periods displays less fluctuation in intensity-duration patterns, resulting in enhanced predictability.

To further evaluate the model performance, Fig. 7 presents two rainfall events with a 5-year RP, illustrating the model's ability to track changes in rainfall intensity under different temporal patterns. The results show that the model performs well for short prediction horizons, particularly for 1- and 4-timestep-ahead predictions, where it successfully captures sudden and gradual changes in rainfall intensity (overall hit rate more than 80%). However, prediction errors increase as the forecast horizon extends to 8 and 20 timesteps (30–50%), indicating reduced model responsiveness and delayed adaptation to rapid intensity fluctuations.

The results also reveal that the model tends to overpredict rainfall events associated with higher RPs when abrupt intensity changes occur. In contrast, when rainfall intensity follows patterns that align with the IDF curves, the model more accurately identifies the correct RP class, especially for shorter prediction horizons. This suggests that the integration of IDF rules supports the model's ability to interpret physically consistent rainfall patterns, but its predictive reliability decreases for longer forecast horizons. A similar behaviour is observed in the second rainfall event (Fig. 7b), where rainfall intensity evolves more gradually. In this case, the model better captures the overall trend of intensity change, resulting in improved accuracy for short- and medium-range predictions. However, the performance deteriorates for longer prediction horizons, particularly when the IDF rules become less informative for distinguishing between RP classes.

4.3. *Performance of ensemble multi-class model*

Here, the performance of the EWS model is evaluated, while the detailed performance of the agent-based base models is provided completely in Appendix D for brevity. Fig. 8 demonstrates multistep performance of the time-series ensemble models for each lead time. Each model displays a distinctive pattern in forecasting each class. In the ensemble-based model, the rates of different class forecasting remain relatively similar at each time lead, as indicated in Fig. 8a. Conversely, the other two models exhibit better accuracy in predicting the low-risk (evaporation) classes. For 3 h class forecasting ahead, both models achieve a 75% accuracy in high-risk (flood forecasting). However, the soft voting model attains an accuracy of approximately 85% for medium-risk, whereas the voting-based model falls short, achieving less

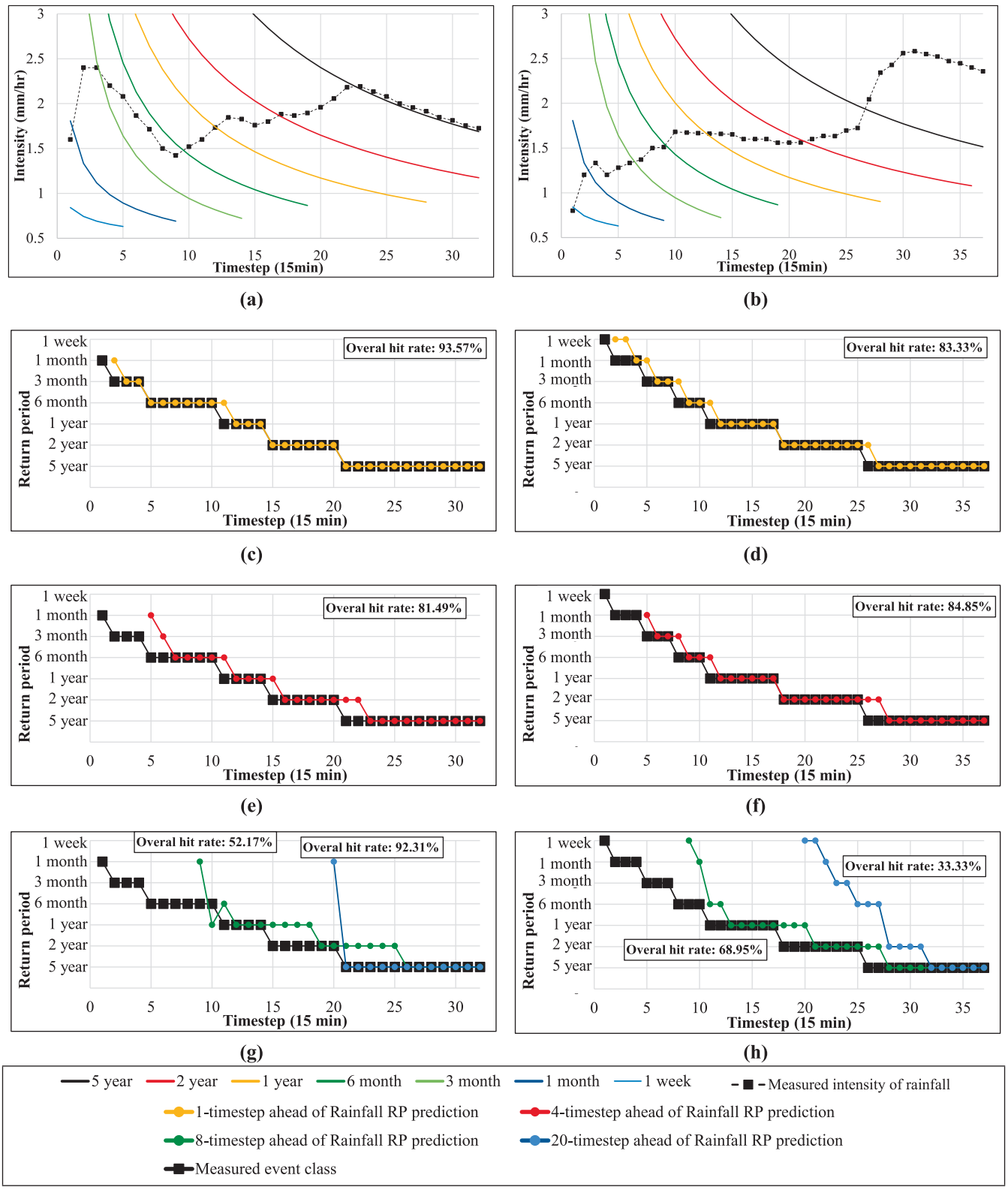


Fig. 7. BPNN Model performance on forecasting of the rainfall with 5-year RP: (left) event example 1, (right) event example 2, (a-b) selected rainfall events, (c-d) 1-timestep ahead of prediction, (e-f) 4-timestep ahead of prediction, and (g-h) 8- and 20-timestep ahead of prediction.

than 70% accuracy (compare Fig. 8b to c). This lower accuracy rate in the voting-based model may result in underestimation, suggesting a tendency to classify draining events as belonging to the evaporation class. While this may initially seem acceptable, the probability of sudden shifts from draining to flooding conditions, along with the complexities associated with longer lead times, may pose challenges in the context of EWSs.

To address this concern, the findings reveal that the developed multi-class stacking model effectively bridges these gaps by employing the proposed decision framework, resulting in a significantly improved hit

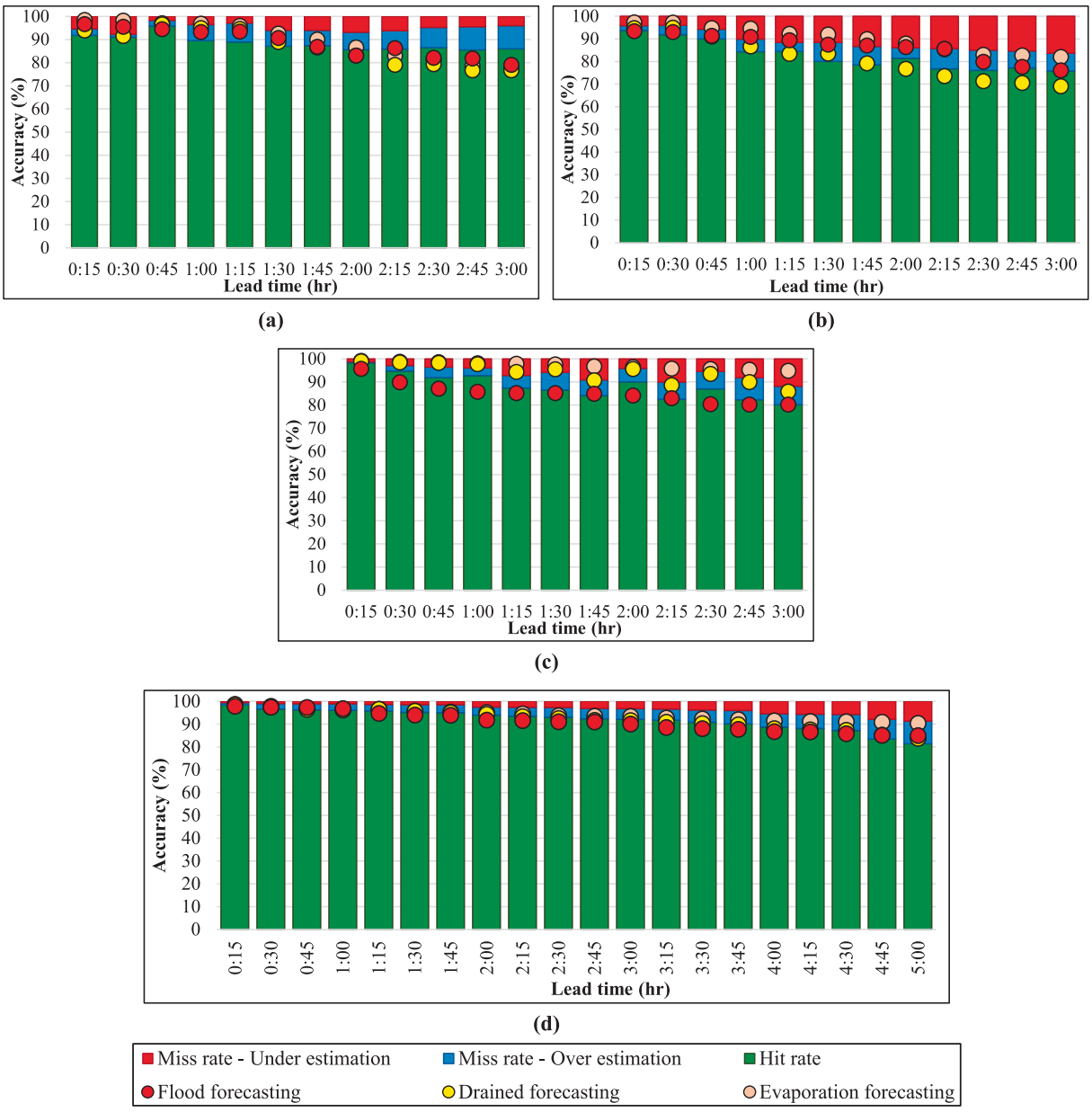


Fig. 8. Multistep performance of the time-series ensemble models for each lead time: (a) ensemble, (b) hard voting, (c) soft voting, and (d) proposed.

rate accuracy. For a 3-hours ahead, the model achieves an impressive accuracy of up to 92%, as illustrated in Fig. 8d (refer to Tables E1-2 in the Appendix E for detailed data). Notably, during this period, the model exhibits a mere 3% underestimation and 4.5% overestimation, indicating substantial improvement and an overall accuracy of approximately 92% across all flood risk classes. Furthermore, the proposed model maintains its accuracy above 80% for 5 h forecasting ahead, with specific accuracies of 85% for the flood high-risk class, 94% for the flood medium-risk class, and 90% for the flood low-risk class. However, it is worth noting that the overestimation and underestimation rates increase to 8% and 10%, respectively, within this time frame.

The results obtained from the ROC curve analysis (refer to Fig. 9a-c and Table E3 in Appendix E) provide compelling evidence of the superiority of the benchmark multiclass model. The AUC significantly improves, with values increasing from 0.77, 0.73, and 0.66 to 0.81, 0.81, and 0.85 for the ensemble-based, voting-based, and weighting-based models, respectively. Additionally, Fig. 9d presents the ROC curves for each class in the multi-class stacking model, showcasing the performance over 5 h forecast (see Table E4 in Appendix E). It is notable that all classes exhibit an AUC of over 0.75, which can be considered acceptable. Particularly noteworthy are the AUC values of 0.87 for class 3 and 0.81 for class 2, indicating strong predictive capabilities for these classes.

Moreover, examining the FPR at the optimum threshold (0.2) reveals an interesting insight. Despite the lower accuracy in forecasting class 3, there is a better balance between true positive forecasting and correct rejection within this class. While accuracy alone measures the overall correctness of predictions, but it does not account for the trade-off between true positive and false positive rates. This suggests that even though the model's accuracy in predicting class 3 may be slightly lower, it maintains a higher level of precision in distinguishing true positives from false positives, leading to more reliable predictions by correctly classifying a higher proportion of true positive cases compared to false positive cases.

Finally, in terms of class imbalance, it can be observed from Figs. 8 and 9 that the AUC values consistently range between 0.92 and 0.96 for all classes, while the difference in accuracy between classes across all time steps remains below 4%. This consistency indicates that the model

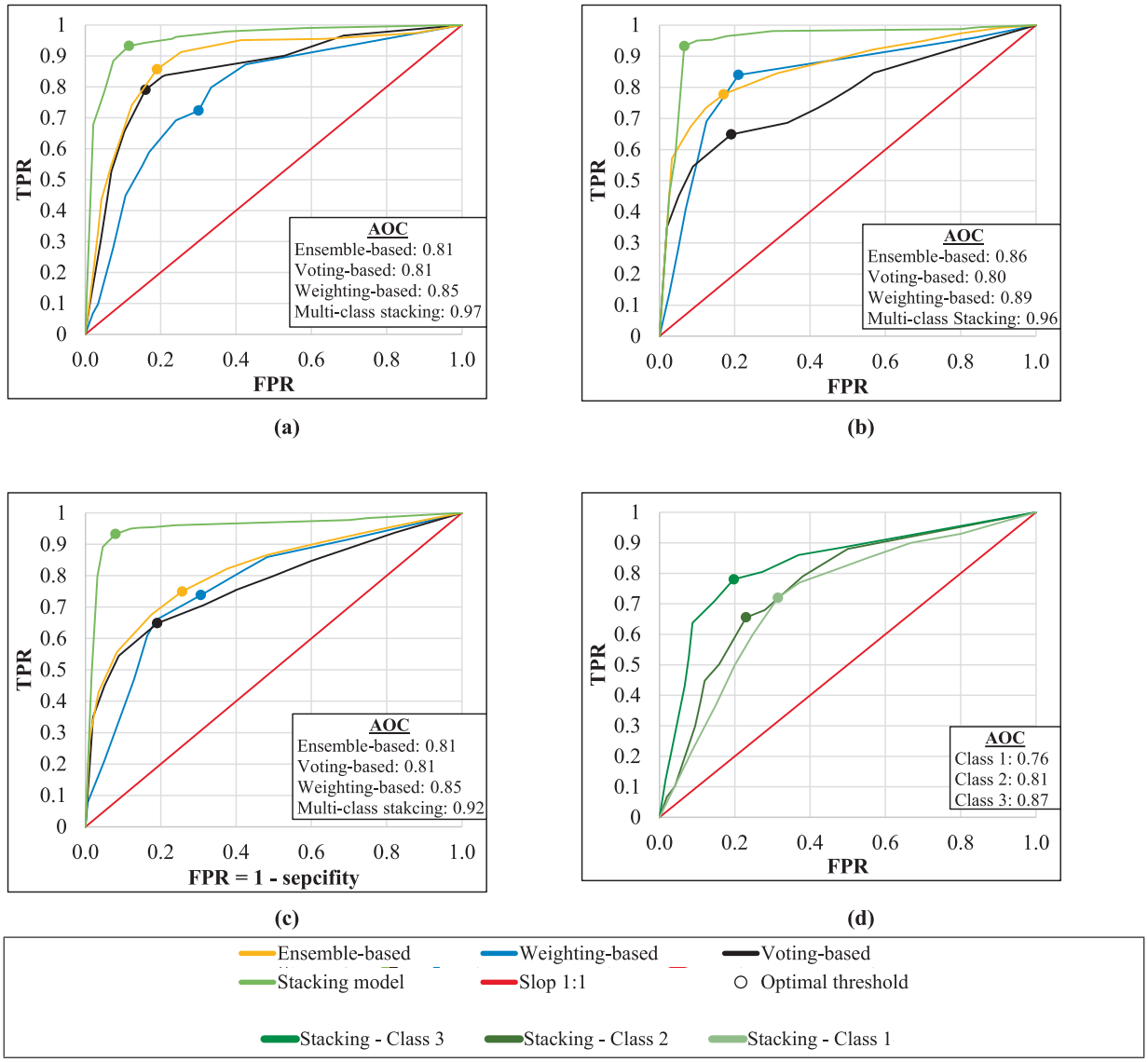


Fig. 9. ROC and AOC of developed models: (a) flood risk class in 3 h ahead, (b) drained class in 3 h ahead, (c) evaporation class in 3 h ahead, and (d) different classes forecasted by multi-class stacking model in 5 h ahead.

performs comparably well across different flood-risk categories, without a noticeable bias toward the majority class. Moreover, the relatively small variation between class-wise metrics suggests that the model maintains stable discrimination capability and balanced predictive performance, even under imbalanced data conditions. These results demonstrate that the proposed framework is robust to class imbalance and can reliably capture both low-risk and high-risk flood states.

Fig. 10 provide valuable insights into the impact of the miss rate on the accuracy of each class. The analysis highlights that the misclassification of the flood high-risk class, i.e. flooding, is primarily caused by forecasting the flood medium-risk class instead, as illustrated in Fig. 10a. More specifically, out of the total 15% miss rate in flood forecasting accuracy for a 5hr ahead, a significant portion of 13.5% can be attributed to misclassifying the flood medium-risk class. Additionally, it is noteworthy that the misclassification rate of flooding into the evaporation/dry class remains relatively low, accounting for less than 1.5% of the total for this period. This observation suggests that the model's accuracy in this specific class is not significantly affected by two-level misclassifications or significant performance gaps.

In contrast, a detailed analysis of the flood medium-risk class, i.e. drained events illustrated in Fig. 10b, demonstrates an initial miss rate stemming from overestimation. The model erroneously predicts flood risk classes, contributing to the initial inaccuracies. However, as the forecasting time steps progress, there is a gradual escalation in the misclassification into the flood low-risk classes i.e. evaporation/dry classes. This suggests that with longer forecasting time steps, both the misclassification of drained events into other classes and the model's inclination towards overestimation exhibit a notable increase. Finally, Fig. 10c reveals a distinct pattern where the model consistently misclassifies evaporation/dry events as flooding situations. This misclassification significantly contributes to the overall miss rate, with 9.56% out of the total 9.6% being attributed to this specific error in 5 h ahead. These findings highlight that despite the implemented improvements, the model continues to face challenges in accurately anticipating the end of rainfall, especially in scenarios that lead to flooding. Although a 10% error rate may be considered acceptable during the model development phase, the issue of false alarming raises concerns. Inaccurate predictions can result in unnecessary costs and necessitate additional mobilisation of staff, placing a considerable burden on the relevant authorities.

Fig. 11 presents an event-based evaluation of four ensemble multi-class early-warning models over 5-hour events, bringing together both "how often" the models are right or wrong (miss/hit rates) and "how

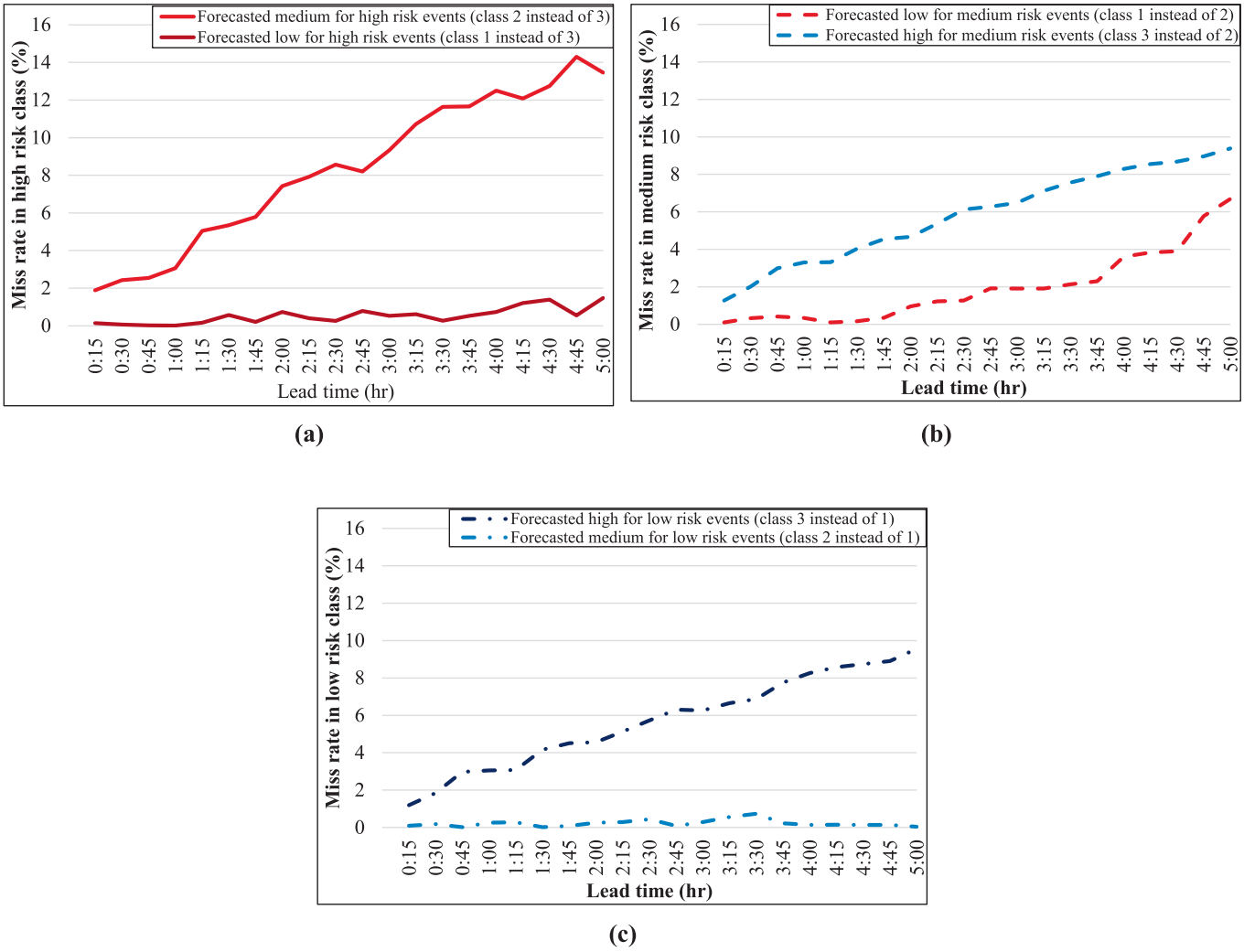


Fig. 10. Share of miss rate in forecasted class of (a) high risk i.e. flooding, (b) medium risk i.e. drained, (c) low risk i.e. evaporation.

well-timed" their correct/incorrect calls are (average lag times). The figure is read alongside Tables D5-D6, which provide the underlying percentages and average lags for each panel. Fig. 11a compares total miss rates and their composition (underestimation versus overestimation). The voting- and weighting-based ensembles have the largest overall miss, with voting strongly dominated by underestimation and weighting showing a sizable overestimation component. The simple ensemble already trims both errors, but the multi-class stacking model is the clear out-performer: by adding a short "water-class memory" (a rule-assisted BPNN input), it reduces underestimation to ~ 2% and overestimation to ~ 1% at the 5-hour horizon, giving the smallest total miss in Table D5. Fig. 11b decomposes hit rates by timing (earlier/exact/later). Here the advantage of the multi-class stacking model is most visible: exact event hits increase markedly – from ~ 67% in the predecessor to ~ 88% – while the share of early "under" hits drops to ~ 4%. Voting and weighting, in contrast, place much of their correct detections in the earlier/later bins rather than at the exact time, and the plain ensemble sits between these extremes.

Fig. 11c examines the quality of underestimation hits via average lag times (15-min units). Relative to the multi-stacking baseline, the multi-class stacking model shortens both components that matter operationally: the delay in recognising water dropping when it was forecast too early ("Earlier-Under") and the delay when detection comes late ("Later-Under") – both improve from about 1.6–1.8 to ~ 1.1-time steps. The residual underestimation that remains for the new model mainly comes from "jump" events, where sudden class shifts are intrinsically hard to anticipate. Fig. 11d repeats the timing analysis for overestimation. Again, the multi-class stacking model is the least laggy, cutting the average early false-alarm lag ("Earlier-Over") from ~ 2.1 to ~ 1.2-time steps and the delayed alarm-clear lag ("Later-Over") from ~ 1.8 to ~ 1.1. Although jump situations still exhibit noticeable early-over lag, these reductions materially lessen the operational burden of unnecessary or long-running alarms. Overall, the figure demonstrates that the proposed multi-class stacking strategy not only lowers how often events are missed but also moves correct detections closer to their true timing, which is exactly the kind of quantity-and-quality improvement sought by the event-based assessment framework.

4.4. *Sensitivity and uncertainty analysis*

In this section, a sensitivity and uncertainty analysis is carried out. The aim is to identify the relative importance of the input variables. It also examines how strongly each parameter affects model accuracy. In addition, it assesses whether the selected physically informed features provide stable and meaningful information for flood-risk classification. Accordingly, Fig. 12 summarises this analysis by showing the contribution of the input features from complementary perspectives, including variance-based ranking, sequential sensitivity, uncertainty due to input reduction, and lead-time-wise changes in feature importance. More specifically, Fig. 12a and b show that all parameters contribute to the

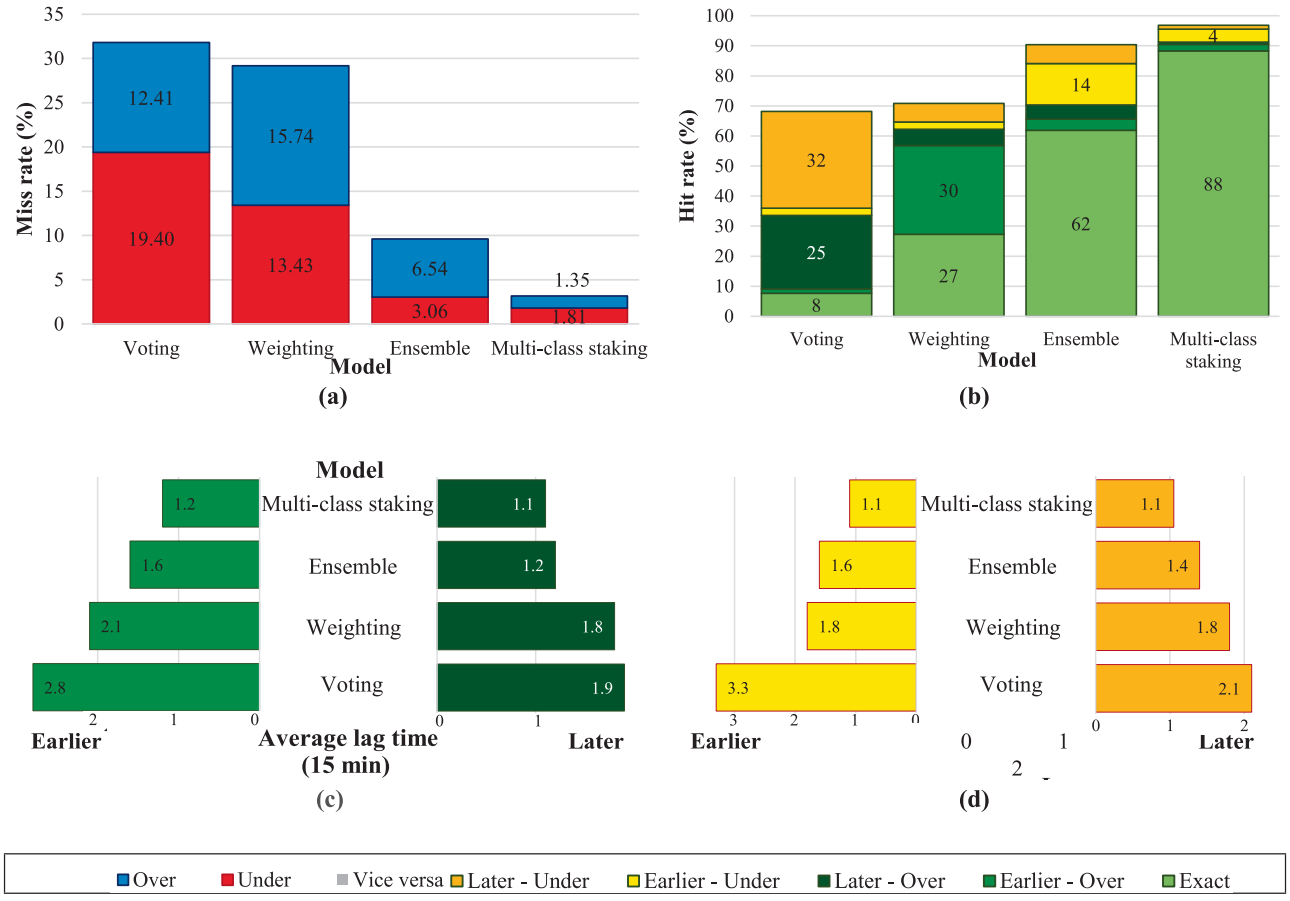


Fig. 11. Event-based performance of the ensemble multi-class models: (a) event miss rate, (b) event hit rate, (c) average underestimation time lags, and (d) average overestimation time lags.

model and should not be treated as negligible, which confirms that urban flood response is governed by multiple interacting rainfall and hydraulic controls rather than a single predictor. However, the results also indicate that the strongest influence is associated with the intensity-related variables, especially IG, followed closely by IR, while AR and RP also make a clear and meaningful contribution. This pattern is physically reasonable as IG and IR directly reflect the magnitude and temporal change of rainfall forcing, which control how quickly runoff is generated, how rapidly the drainage system is loaded, and how soon the system approaches surcharge or flooding. AR remains important because antecedent rainfall represents pre-event wetness and residual catchment storage conditions, which strongly affect infiltration losses and runoff conversion. RP is also influential because it provides a physically informed description of event severity through the IDF relationship, helping the model distinguish between ordinary rainfall and rainfall with greater flood-producing potential.

Fig. 12c shows how model performance changes when the size of the training dataset is reduced. Three clear response zones can be identified. In the first wave, the model shows strong resistance to data reduction. The error increases only slightly while the training size is reduced from the full dataset to about 70%. This means the framework can still be retrained with a smaller dataset without a major loss of performance. The likely reason is that the reduced dataset still contains the main rainfall-runoff patterns and the dominant event behaviours needed for learning. In the second wave, between about 70% and 50% of the training data, model degradation and data reduction move more in harmony. The error increases more steadily in this range. This suggests that the model is adapting to the loss of information, but its predictive strength is gradually weakening as event diversity becomes smaller. In the third wave, below about 50%, the model starts to yield. The error rises rapidly, and the deterioration becomes much more severe at very low data volumes. This indicates that the remaining dataset is no longer sufficient to represent the full variability of the system, especially nonlinear responses, transition states, and higher-risk events.

For Fig. 12d, the lead-time-wise analysis makes the role of each parameter in the final model clearer. The figure shows that the effect of each input is not fixed and changes with forecast horizon. At short lead times, rainfall duration (D) and rainfall intensity (I) have the strongest effect on model accuracy. This is physically expected because the near-future response of an UDS is mainly controlled by the ongoing storm characteristics. Seasonal condition (S) and class duration (Dw) are also more useful in the early stages. They help the model interpret the current hydrological setting and the persistence of the present water-level state. As lead time increases, the role of some variables becomes stronger. In particular, RP rises markedly and becomes one of the most influential parameters at longer horizons. This suggests that event severity becomes more important when the model looks further ahead. Antecedent rainfall (A), current water-level class (Wc), and future return-period information (RPt) also increase with lead time. This indicates that catchment memory, present hydraulic state, and future rainfall potential all support longer-horizon classification. In contrast, S and Dw gradually lose influence with time. This is reasonable because seasonal context and current-state persistence are more helpful for short-term interpretation than for extended prediction. Overall, Fig. 12d confirms that the final model benefits from a balanced structure. Immediate rainfall descriptors dominate early forecasts, while severity, memory, and hydraulic-state variables become more important at longer lead times.

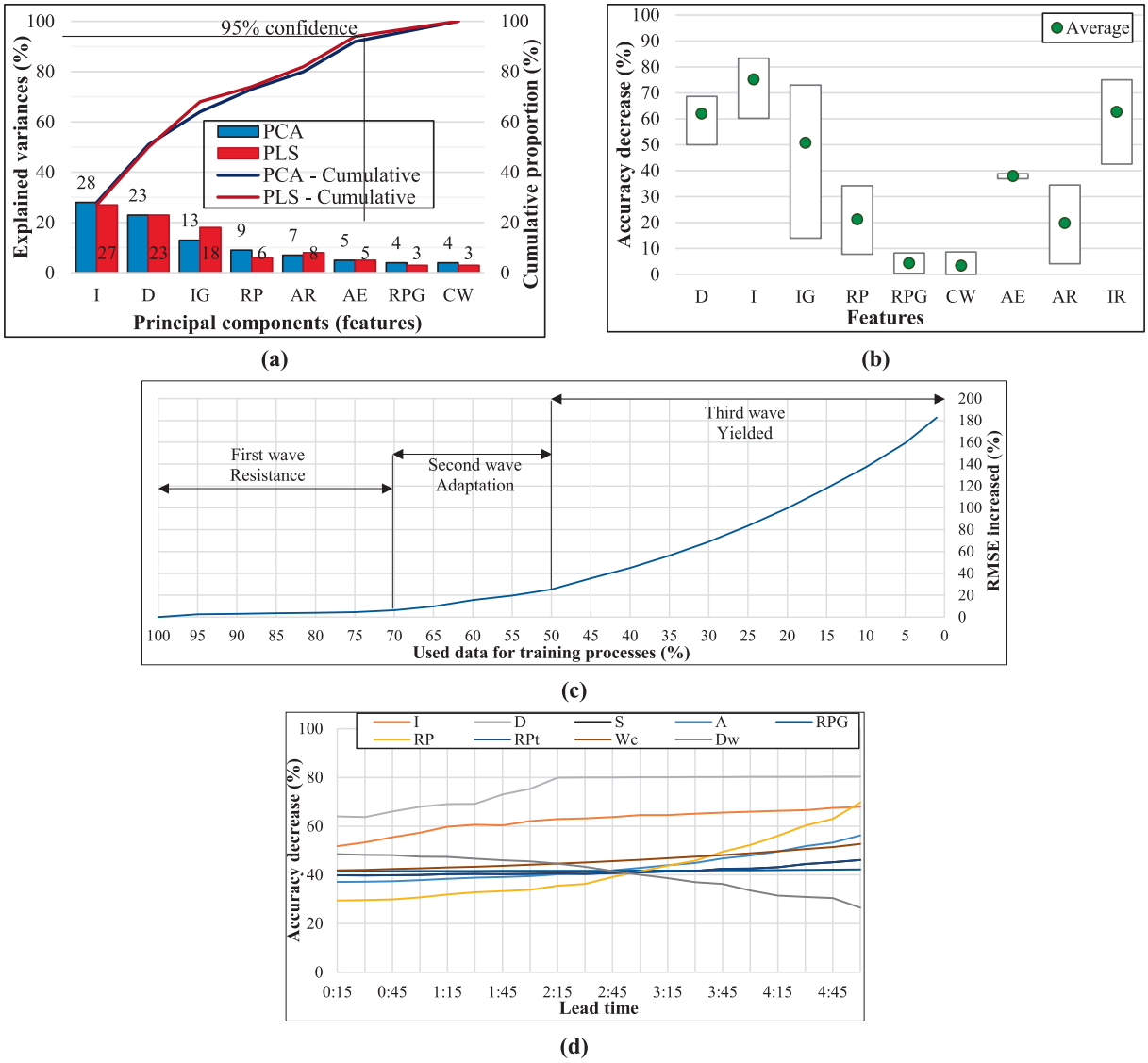


Fig. 12. Further analysis on the model developed: (a) PCA and PLS analysis on BPNN, (b) Sequential analysis on BPNN, (c) Uncertainty analysis on input reduction for PINN model, (d) Timestep wise sequential analysis on PINN model.

5. Practical benefits, current limitations and future research direction

The proposed framework offers several practical benefits for real-time urban flood early warning. First, it combines hydrological rainfall descriptors with hydraulic water-level memory. This gives the model a stronger physical basis than a purely data-driven classifier. Second, it supports multi-class flood-state classification rather than a simple binary warning. This is useful for operational decision-making because it can distinguish between non-flood, transitional, and flood conditions. Third, the lead-aware mixture-of-experts design improves model flexibility across forecast horizons. This is important because flood behaviour and model performance change with lead time. In addition, the framework is designed for real-time IoT data streams and is evaluated using both multistep and event-based criteria. This increases its practical relevance for online warning applications.

Although the proposed framework performed better than the benchmark models and showed clear value for real-time flood-state classification, several points should be considered for its next stage of development. The current results were obtained from a real-world case study and provide a strong foundation for application in practice. At the same time, further testing across different catchments, climatic conditions, and drainage settings would help confirm its broader transferability. The framework also showed its strongest performance at short to medium lead times, while longer forecast horizons and transition periods remained more challenging. This suggests an opportunity to further enrich the rainfall representation for longer-lead classification. In addition, because the framework is designed for real-time IoT-based operation, its practical performance is naturally linked to the quality, continuity, and timeliness of incoming sensor data.

Rainfall forecast uncertainty remains a major unresolved challenge in real-time flood forecasting, particularly with respect to temporal evolution, spatial heterogeneity, and extreme intensity estimation. The present study does not aim to resolve that upstream problem fully. Instead, it focuses on a downstream but operationally important task: improving the transformation of available real-time rainfall and water-level observations into physically informed multi-class flood-risk warnings. Therefore, future research should focus on integrating more advanced rainfall forecasting and nowcasting approaches, including high-resolution spatio-temporal rainfall products, radar-based estimates, and ensemble prediction systems, to better capture rainfall uncertainty and its propagation into flood predictions.

To make the framework more usable in online flood risk forecasting, future work should also extend the current classification model into a broader operational system. This includes automatic data quality control, real-time handling of missing sensor values, and integration with forecasting dashboards or digital twin platforms. The next development stage could also link flood-state classification with rapid inundation mapping and dynamic vulnerability assessment. Such integration would move the framework from warning classification toward real-time flood risk assessment and decision support. This would increase its value for emergency managers, local councils, and other flood-risk stakeholders.

Finally, in this study the use of multiple-source data with different temporal and spatial resolutions is not explicitly discussed. However, in practical applications the framework may receive inputs from multiple data sources with varying temporal and spatial resolutions (e.g., rainfall gauges, water-level sensors, and external environmental datasets). To ensure data compatibility, a harmonisation process should be applied before model training and prediction. First, all datasets should be resampled to a common temporal resolution corresponding to the model prediction interval. When higher-frequency data are available, they can be aggregated using averaging or summation depending on the variable type, whereas lower-frequency data can be interpolated using linear interpolation to match the target timestep. Second, spatial discrepancies can be addressed by selecting representative monitoring stations based on correlation analysis and catchment characteristics, ensuring that rainfall and hydraulic observations correspond to the same hydrological response area. All harmonised variables are then synchronised within a unified time-series structure prior to feature extraction. This preprocessing step ensures that heterogeneous data sources can be integrated consistently while preserving their physical and temporal relevance for flood risk classification modelling.

6. Conclusions

This study introduced a hydrology- and hydraulics-informed, multiclass early-warning framework for real-time urban flood risk classification. The framework integrates three main components: (i) rainfall descriptors derived from a rule-based back-propagation neural network, including seasonality, antecedent conditions, intensity-duration characteristics, and return-period signals; (ii) water-level memory variables, represented by the current class and class duration; and (iii) a time-series mixture-of-experts ensemble built from seven weak-learner families. Using real-time IoT sensor data, the framework produces multi-lead flood-state classifications at 15-min intervals for evaporation, drained, and flooding conditions. While the framework applied to a real-world case study, the following key findings are noted:

- The study shows that combining hydrological rainfall information with hydraulic water-level memory provides a stronger basis for flood-state classification than using a single source of information alone. This integration improves the physical relevance of the framework and supports more robust early warning.
- The rule-based BPNN is effective in extracting rainfall return-period information for short forecasting horizons. This confirms its value as a practical feature-generation tool within the proposed classification framework.
- The results show that rainfall intensity and duration are the most important descriptors for detecting flood states, while previous rainfall occurrence and seasonal conditions provide useful complementary information. This highlights the need for physically meaningful rainfall features in flood risk classification tasks.
- No single weak learner performs best across all forecast leads and classes. This confirms that flood risk classification is strongly lead-dependent and supports the need for an adaptive ensemble rather than a fixed single-model solution.
- The mixture-of-experts ensemble improved flood-state discrimination and reliability over the baseline voting and averaging ensembles. Its key advantage is adaptive, lead-aware expert selection, which allows the framework to choose the most suitable model for each class and forecast horizon. As a result, it reduced missed flood events and false alarms compared with the benchmark methods.
- Event-based evaluation shows that the proposed framework is useful not only in terms of classification accuracy, but also in terms of warning timeliness. This is important for real-time flood management, where the timing of detection is as critical as the detection itself.

Overall, the proposed framework offers a practical and physically informed approach for real-time flood-state classification in urban areas. It provides a foundation for more reliable EWSs and creates a pathway for future improvements in longer-horizon prediction, class-specific threshold design, and broader deployment in real-world flood risk management.

CRediT authorship contribution statement

Farzad Piadeh: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Vahid Bakhtiari:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Kourosh Behzadian:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Conceptualization. **Farshad Piadeh:** Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2026.135819.

Data availability

Data will be made available on request.

References

Almikaeel, W., Šoltész, A., Čubanová, L., Baroková, D., 2025. Hydro-informer: a deep learning model for accurate water level and flood predictions. Nat. Hazards 121 (4), 3959–3979.

Antwi-Agyakwa, K.T., Afenyo, M.K., Angnuureng, D.B., 2023. Know to predict, forecast to warn: a review of flood risk prediction tools. Water 15 (3), 427.

Bakhtiari, V., Kerchi, H., Piadeh, F., Behzadian, K., Nasirzadeh, F., 2025. Role of the internet of things in flood risk management: a critical review on current practices and future directions. Nat. Hazards 7589.

Bakhtiari, V., Piadeh, F., Chen, A., Behzadian, K., 2024. Stakeholder analysis in the application of cutting-edge digital visualisation technologies for urban flood risk management: a critical review. Expert Syst. Appl., 121426

Bakhtiari, V., Piadeh, F., Behzadian, K., Kapelan, Z., 2023. A critical review for the application of cutting-edge digital visualisation technologies for effective urban flood risk management. Sustain. Cities Soc. 99, 104958.

Bentivoglio, R., Isufi, E., Jonkman, S.N., Taormina, R., 2022. Deep learning methods for flood mapping: a review of existing applications and future research directions. Hydrol. Earth Syst. Sci. Discuss. 1–50.

Bojović, F., Milašinović, M., Jovanović, B., Krstić, L., Stojanović, B., Ivanović, M., Prodanović, D., Milivojević, N., 2022. Physics informed neural networks for 1D flood routing. 1st Serbian International Conference on Applied Artificial Intelligence (SICAAI).

Byaruhanga, N., Kibirige, D., Gokool, S., Mkhonta, G., 2024. Evolution of flood prediction and forecasting models for flood early warning systems: a scoping review. Water 16 (13), 1763.

Centre for Research on the Epidemiology of Disasters (CRED), 2025. Emergency events database [Online] www.emdat.be <retrieved 09/22/ 2025>.

Chen, R., Liang, C., Hong, W., Gu, D., 2015. Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm. Appl. Soft Comput. 26, 435–443.
Chew, A.W.Z., He, R., Zhang, L., 2025. Physics informed machine learning (PIML) for design, management and resilience-development of urban infrastructures: a review. Arch. Comput. Meth. Eng. 32 (1), 399–439.
Department for Environment, Food and Rural Affairs (DEFRA), 2025. DEFRA Official Website [Online] Available at Environment.data.gov.uk, <Retrieved 15/09/2025>.
Donnelly, J., Daneshkhah, A., Abolfathi, S., 2024. Physics-informed neural networks as surrogate models of hydrodynamic simulators. Sci. Total Environ. 912, 168814.
Feng, D., Tan, Z., He, Q., 2023. Physics-informed neural networks of the saint-venant equations for downscaling a large-scale river model. Water Resour. Res. 59 (2), 33168.
Ferdowsi, A., Piadeh, F., Behzadian, K., Mousavi, S., Ehteram, M., 2024. Urban water infrastructure: a critical review on climate change impacts and adaptation strategies. Urban Clim. 58, 102132.
Girottoa, C., Piadeh, F., Bakhtiari, V., Behzadian, K., Chen, A., Campos, L., Zolgharni, M., 2024. A critical review of digital technology innovations for early warning of water-related disease outbreaks associated with climatic hazards. Int. J. Disaster Risk Reduct. 100, 104151.
Hamil, L., 2011. Understanding hydraulics, 3rd Ed. Macmillan Education, London, pp. 507–585.
Heydarian, M., Doyle, T., Samavi, R., 2022. MLCM: multi-label confusion matrix. IEEE Access 10, 19083–19095.
Kithulgoda, C., Pears, R., Naeem, M., 2018. The incremental Fourier classifier: leveraging the discrete Fourier transform for classifying high speed data streams. Expert Syst. Appl. 97, 1–17.
Lillicrap, T., Santoro, A., Marris, L., Akerman, C., Hinton, G., 2020. Backpropagation and the brain. Nat. Rev. Neurosci. 21, 335–346.
Lin, Q., Leandro, J., Wu, W., Bhola, P., Disse, M., 2020. Prediction of maximum flood inundation extents with resilient backpropagation neural network: case study of Kulmbach. Front. Earth Sci. 8, 332.
Markoulidakis, I., Rallis, I., Georgoulas, I., Kopsiaftis, G., Doulamis, A., Doulamis, N., 2021. Multiclass confusion matrix reduction method and its application on net promoter score classification problem. Technologies 9 (4), 81–103.
Mishra, A., Mukherjee, S., Merz, B., Singh, V.P., Wright, D.B., Villarini, G., Paul, S., Kumar, D.N., Khedun, C.P., Niyogi, D., Schumann, G., 2022. An overview of flood concepts, challenges, and future directions. J. Hydrol. Eng. 27 (6), 03122001.
Pandi, D., Kothandaraman, S., Kuppusamy, M., 2021. Hydrological models: a review. Int. J. Hydrol. Sci. Technol. 12 (3), 223–242.
Park, S., Oh, S., Kim, E., Pedrycz, W., 2023. Rule-based fuzzy neural networks realized with the aid of linear function prototype-driven fuzzy clustering and layer reconstruction-based network design strategy. Expert Syst. Appl. 219, 119655.
Piadeh, F., Behzadian, K., Chen, A., Campos, L., Kapelan, Z., 2023a. Event-based decision support algorithm for real-time flood forecasting in urban drainage systems using machine learning modelling. J. Environ. Model. Softw. 167, 105772.
Piadeh, F., Behzadian, K., Chen, A., Kapelan, Z., Rizzuto, J., Campos, L., 2023b. Enhancing urban flood forecasting in drainage systems using dynamic ensemble-based data mining. Water Res. 247, 120791.
Prasanthi, A., Shareef, H., Khalid, S., Selvaraj, J., 2025. Optimized forgetting factor recursive least square method for equivalent circuit model parameter extraction of battery and ultracapacitor. J. Storage Mater. 119, 116298.
Qian, K., Mohamed, A., Claudel, C., 2019. Physics informed data driven model for flood prediction: application of deep learning in prediction of urban flood development. p. 1908-10312.
Raissi, M., Perdikaris, P., Karniadakis, G.E., 2019. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. Comput. Phys. 378, 686–707.
Taghizadeh, M., Zandsalimi, Z., Nabian, M.A., Shafiee-Jood, M., Alemazkoor, N., 2025. Interpretable physics-informed graph neural networks for flood forecasting. Comput. Aided Civ. Inf. Eng.
Tota-Maharaj, K., Karunanayake, C., Kunwar, K., Chadee, A.A., Azamathulla, H.M., Rathnayake, U., 2024. Evaluation of permeable pavement systems (PPS) as best management practices for stormwater runoff control: a review. Water Conserv. Sci. Eng. 9 (1), 32.
Vongkusolkit, J.J.N., 2022. Physics-informed weakly supervised learning for near real-time flood mapping. University of Wisconsin-Madison. MSc.
Xia, Y., Meng, Y., 2024. Physics-informed neural network (PINN) for solving frictional contact temperature and inversely evaluating relevant input parameters. Lubricants 12 (2), 62.
Yang, F., Ding, W., Zhao, J., Song, L., Yang, D., Li, X., 2024. Rapid urban flood inundation forecasting using a physics-informed deep learning approach. J. Hydrol. 643, 131998.