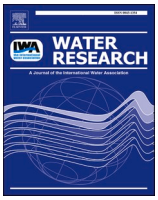
Water Research 247 (2023) 120791

Contents lists available at ScienceDirect

Water Research

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



Enhancing urban flood forecasting in drainage systems using dynamic ensemble-based data mining

Farzad Piadeh [a,b], Kourosh Behzadian [a,e,*], Albert S. Chen [c], Zoran Kapelan [d], Joseph P. Rizzuto [a], Luiza C. Campos [e]

[a] School of Computing and Engineering, University of West London, St Mary's Rd, London W5 5RF, UK
[b] School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield AL10 9AB, UK
[c] Centre for Water Systems, Faculty of Environment, Science and Economy, University of Exeter, Exeter EX4 4QF, UK
[d] Department of Water Management, Faculty of Civil Engineering and Geoscience, Delft University of Technology (TU Delf), Delft, Netherlands
[e] Centre for Urban Sustainability and Resilience, Department of Civil, Environmental and Geomatic Engineering, University College London, Gower St, London WC1E 6BT, UK

ARTICLE INFO

Keywords:
Data mining
Drainage systems
Dynamic ensemble modelling
Real-time modelling
Urban flood forecasting

ABSTRACT

This study presents a novel approach for urban flood forecasting in drainage systems using a dynamic ensemble-based data mining model which has yet to be utilised properly in this context. The proposed method incorporates an event identification technique and rainfall feature extraction to develop weak learner data mining models. These models are then stacked to create a time-series ensemble model using a decision tree algorithm and confusion matrix-based blending method. The proposed model was compared to other commonly used ensemble models in a real-world urban drainage system in the UK. The results show that the proposed model achieves a higher hit rate compared to other benchmark models, with a hit rate of around 85% vs 70 % for the next 3 h of forecasting. Additionally, the proposed smart model can accurately classify various timesteps of flood or non-flood events without significant lag times, resulting in fewer false alarms, reduced unnecessary risk management actions, and lower costs in real-time early warning applications. The findings also demonstrate that two features, "antecedent precipitation history" and "seasonal time occurrence of rainfall," significantly enhance the accuracy of flood forecasting with a hit rate accuracy ranging from 60 % to 10 % for a lead time of 15 min to 3 h.

1. Introduction

The early warning of floods in urban drainage systems (UDS) is critical to mitigate potential social, economic, and environmental losses caused by the growing concern of urban flooding worldwide (Zounemat-Kermani et al., 2020). The highly complex and temporally-restricted nature of UDS flooding, along with spatial limitations (Piadeh et al., 2022a), underscores the importance of early warning systems for decision-makers and communities. Such advance notice enables the reduction of financial and human losses/fatalities, minimisation of infrastructure damage, and better preparation of safety services (Mobini et al., 2022). Historical long-term and large-scale databases that are continuously updated with real-time rainfall and water level data in UDS can be integrated with data-driven models for simulating dynamic and continuous flood events, particularly for flood forecasting. Prominently, single or ensemble weak learner data mining models (WLDMs) have been developed for this purpose (Munawar et al., 2021).

Several predictive data mining models (DMs) aim to estimate future flood events by flood regression, time series analysing of hydrological characteristics, flood risk, flood classification, and forecasting flood events (Zounemat-Kermani et al., 2021). Among these wide-ranging applications, classification and forecast of floods have attracted more attention, especially for water level rise or flooding in UDS. Earlier studies have demonstrated the application of various weak learner models, such as support vector machine (SVM), k-nearest neighbourhood (KNN), Naïve Bayes (NB), and neural network pattern recognition (NNPR), for water level forecasting (Mosavi et al., 2018). These models have also been used for flood forecasting when the flow associated with water level rise exceeds the full capacity of the UDS conduits. Other strong DMs, particularly various feedback forward and recurrent neural network (RNN) models, show more capability in the accuracy of flood

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

https://doi.org/10.1016/j.watres.2023.120791
Received 1 March 2023; Received in revised form 31 August 2023; Accepted 27 October 2023
Available online 27 October 2023

forecasting, handling big data and high-speed computation (Rasheed et al., 2022). Despite the advantages of these DMs, the accuracy of overflow predictions made for periods longer than one hour has been significantly reduced (Piadeh et al., 2022a).

To overcome this limitation, hybrid models have been developed recently, in which WLDMs are combined with RNN, such as nonlinear autoregressive network with exogenous inputs and long short-term memory (Piadeh et al., 2022b; Zhang et al., 2023). In addition to water level forecasting, WLDMs have been used to classify urban flooding, specifically for flood susceptibility, overflow probability, and flood risk index (Shahzad et al., 2022). However, only a few studies have investigated the classification of floods in UDS, which involves identifying flood or non-flood conditions due to rainfall occurrence. For instance, recent studies have employed a combination of decision tree (DT) and random forest (RF) to predict multivariate flood status in the UK (Aswad et al., 2022) and to forecast flash floods in cities in Malaysia and China (Zang et al., 2022).

Ensemble modelling is another popular approach in flood forecasting that involves combining multiple forecasts from different models to improve the accuracy of predictions. The resulting ensemble forecast provides a range of possible outcomes and their probabilities, allowing forecasters to make more informed decisions (Bui et al., 2019). Ensemble modelling has been applied to various flood forecasting problems, including riverine, flash floods, and storm surges. However, its application in real-time urban flood forecasting is a promising area for further investigation, particularly for longer periods of forecasting (Zounemat-Kermani et al., 2021) or using dynamic platforms in which ensemble models are adapted based on timesteps of forecasting (Zhou et al., 2023). These studies demonstrate that flood forecasting methods can achieve higher accuracy and faster processing with less computational efforts compared to other conventional data mining models. However, these methods may fail to provide accurate predictions for longer lead times, such as projections of more than two hours ahead (Wagenaar et al., 2020).

Besides, the recent progress in flood forecasting has primarily concentrated on improving the accuracy of projection for a single or multiple timesteps ahead, but dynamic forecasting of almost all timesteps of flood events, known as event-based modelling still requires more attention. This approach can enable a more detailed about forecasting the timing, magnitude, and duration of a specific flood event, in comparison to just providing an overall performance of the model across single or multiple timesteps of the events (Yaseen et al., 2019). This is particularly crucial in the context of flood risk management and emergency response, where accurate forecasts of specific flood events are vital to minimise potential impacts on society and the environment (Orellana-Alvear et al., 2021). Such accurate forecasts of specific events can facilitate informed decision-making and support effective risk management during flood events. However, development of time-series dynamic DMs that can accurately forecast nature of flood events commonly known as a challenging task and further research is required to develop reliable and effective models for this purpose (Munawar et al., 2021).

To overcome the above challenges, this study introduces several innovative contributions to the field of urban flood forecasting. Firstly, this study utilises a novel time-series ensemble-based data mining model, which has yet to be extensively applied. This approach provides a fresh perspective on addressing the existing knowledge gaps in current flood forecasting methods by enhancing the accuracy of flood forecasting with longer lead times, which is crucial for effective flood risk management. Secondly, this study presents the concept of dynamic modelling for ensemble-based data mining to move beyond this focus on single-timestep forecasting and investigates the performance of the proposed model for event forecasting. Finally, the study couples the decision tree algorithm with the concept of ensemble modelling to blend the performance of WLDM models.

2. Methodology

Two general approaches are documented for UDS flood forecasting: (1) direct relationship approach that goals to establish a direct relationship between rainfall data and the UDS water. This means that the model attempts to predict the water level directly based on the incoming rainfall data. This approach is simpler and more straightforward, as it involves no intermediate step of forecasting rainfall, (2) rainfall forecasting approach that involves forecasting the intensity of rainfall in the area. The predicted rainfall data is then used as an input to a model that simulates the behaviour of the UDS to predict the resulting water levels. This approach can be more accurate in capturing the dynamics of the system but involves more complexity and uncertainty due to the two-step forecasting process (Zounemat-Kermani et al., 2021; Piadeh et al., 2022a). This study adopts the first approach mainly due to its more simplicity and response time, data availability and finally less UDS complexity in comparison to large catchments of river basins or reservoirs dams (Zhou et al., 2023).

The proposed modelling framework, as illustrated in Fig. 1, involves two key stages: (1) an offline data mining framework for pre-training, and (2) a real-time online platform for implementation. To ensure the framework's applicability to scenarios with limited data availability, only time-series data from a single rainfall station and a single water level station, which are commonly used in hydrological practices, are utilised and other time-series data resources are excluded to maintain simplicity. While there might be various rainfall stations to choose from, the selection process is guided by two parameters, drawing inspiration from the research of Piadeh et al. (2023): (1) dominant wind direction determined based on wind rose data aligned with the geographical positioning of both the rainfall station and the UDS station, (2) highest cross-correlation coefficient between the data from the rainfall stations and the UDS water levels (See Figure C1 in the Appendix C for further description).

The continuous time-series data of rainfall and water level are partitioned into dry and wet weather events using the methodology outlined in Section 2.1. These events are subsequently transformed into rainfall features to develop WLDMs, as explained in Section 2.2. The commonly used WLDMs are developed and stored in a data warehouse, and their performance indicators on forecasting unseen data is also stored in a data cube structure for ensemble model development. Further details are presented in Section 2.3. Lastly, a dynamic platform utilising time-series ensemble models is developed using the methodology proposed in Section 2.4.

2.1. Event identification

This study follows an event identification procedure, as described in Piadeh et al. (2021) and Piadeh et al. (2022b), to classify time-series data into flood and non-flood events. The classification is based on typical event states as illustrated in Fig. 2a, and the event identification procedure outlined in Fig. 2b. The time-series data is initially divided into dry (R_1, R_4, R_6) and wet (R_2, R_3, R_5) weather classifications based on rainfall data. Then, the water level is classified into six categories (S_1-S_6): (S1) dry flow event, which is characterised by no rainfall and trivial/no change in water level; (S_2) upstream discharge event, which denotes an increase in water level without rainfall, caused by leakage/exfiltration, infiltration, or discharging diurnal wastewater into combined sewerage; (S_3) evaporation event, which indicates rainfall without an increase in water level due to either evaporation or infiltration into the soil; (S_4) overflow event, which occurs when rainfall and water level increase with a time lag, ultimately leading to water level exceeding the full UDS capacity; (S_5) depletion event, which describes a scenario of no rainfall but a decrease in water level back to the dry flow state in the falling limb of the hydrograph; and (S_6) drained event, which occurs when rainfall and water level rise but less than the full UDS capacity, allowing for the safe drainage of excessive water.

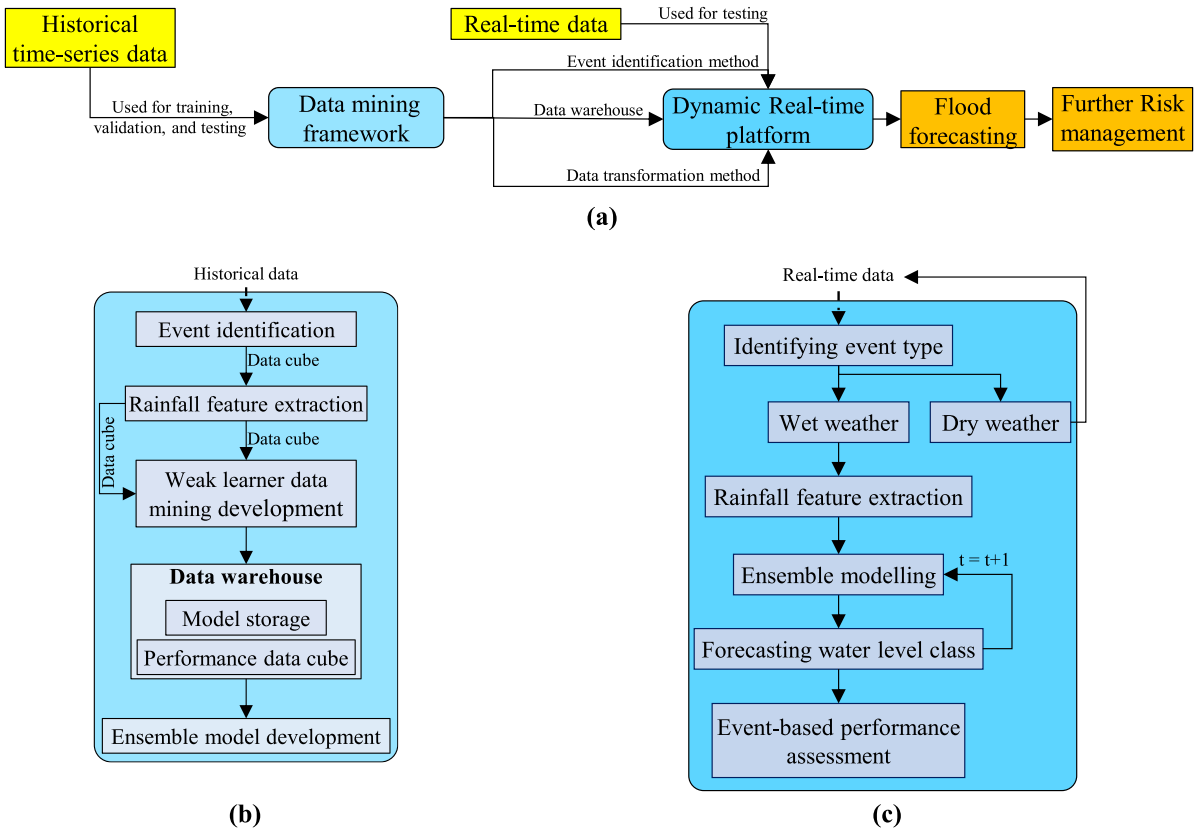


Fig. 1. Schematic structure of the: (a) proposed framework for real-time dynamic flood forecasting, (b) pre-training offline data mining platform, (c) online dynamic flood forecasting platform.

This study has been developed for real-time applications, and since the duration of a real-time rainfall events are unknown in this period of time, it is necessary to break down the identified events into multiple events with different rainfall durations, known as extended events. This breakdown is then used for WLDM and ensemble model training. For instance, for a rainfall event with n timesteps schematically represented in Fig. 3a, at real-time timestep i, only the first i timesteps of rainfall are illustrated for forecasting water level at any given lead time from timestep i onwards (e.g., 1, ..., i*, ..., p). Therefore, this event can provide maximum n states of extended events in practice in which p water level data are collected. A rainfall data cubic structure is then created for all identified extended rainfall events, as shown in Fig. 3b. One dimension represents the ID number of the original rainfall events (ranging from 1 to k in Fig. 3b). The rainfall intensity of the associated extended events is then organised ranging from 1 to n for the number of extended events and t_1 to t_n for the timesteps. Accordingly, water level data is considered up to the maximum interested lead time (Here p) is stored in another data cube shown in Fig. 3c. This data cube is structured similarly to the rainfall data cube, but instead of rainfall intensity, it stores water level data ranging from 1 to p. It should be noted that stored data here are numerical and time-series rainfall and water level data.

2.2. Event classification and key rainfall feature extraction

Stored numerical data are transformed into categorised orders, referred to as rainfall features and water level class which can be utilised for developing and testing WLDMs. Various extracted potential features are listed in Table 1, including (1) current rainfall characteristics such as duration, depth, average intensity, and peak depth, as demonstrated in previous studies by Bui et al. (2019) and Hosseini et al. (2020), (2) antecedent precipitation history, which reflects the short-term effects of soil moisture and surface temperature on model development, and is represented by two forms of recent rainfall occurrence and average intensity, and (3) time occurrence, which represents the long-term effects of average air temperature on model development, and is indicated by the season code based on the Köppen climate classification (DEFRA, 2022) and the average intensity of the long-term history of similar rainfall. Furthermore, the forecasted water level data for each lead time is classified into a binary structure, referred to as "class 1" and "class 2," for the purpose of flood forecasting. Class 1 corresponds to the situation where the water level remains under the full capacity level of the UDS, while class 2 denotes the occurrence of flooding when the water level exceeds the UDS full capacity level. These binary classes are defined for each timestep ahead, ranging from 1 to p, and stored in the data cube of features for water level class (as illustrated in Fig. 5a).

The features extracted from the extended rainfall events (Table 1) are then refined using three established techniques: principal component analysis (PCA), partial least squares (PLS), and sequential sensitivity analysis. These techniques are widely accepted as prerequisite steps to identify key variables that enhance classification performance and reduce computation times (Masahiko et al., 2019). PCA is a process that extracts principal components representing the original variables and explains the majority of the variance in the dataset. PLS is used to estimate linear relationships between dependant and independent variables, showing the direct effect of independent variables on the dependant variables. Further sensitivity analysis is conducted by removing one feature at a time and measuring the accuracy difference of the developed WLDM models discussed in the next section. Key rainfall features are finally selected based on the results of these three methods and determined individually for each extended event. The values of these features (1 to r in Fig. 4b) are stored in the data cube for each extended event, ranging from 1 to n.

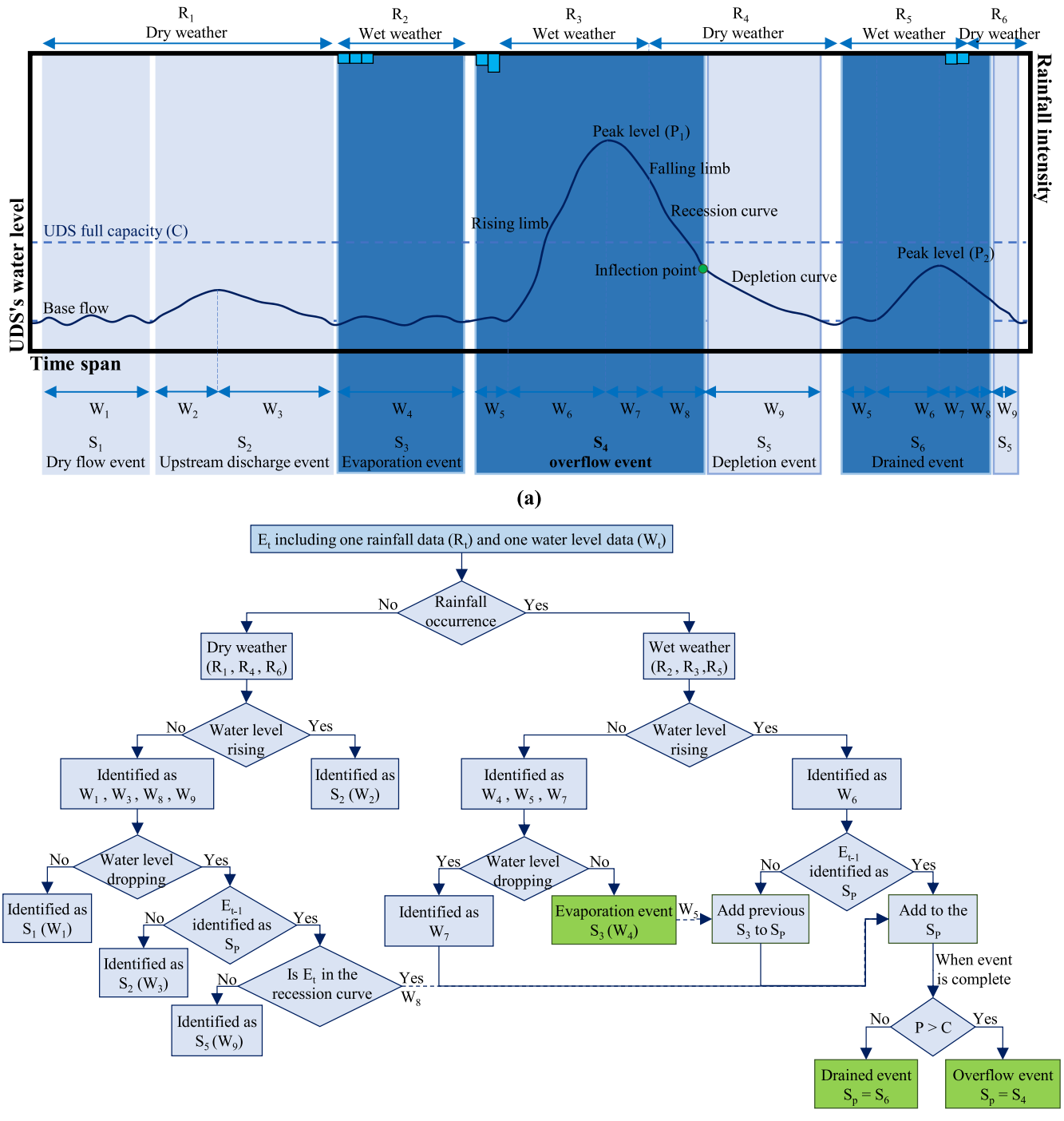


Fig. 2. Event classification method used in this study: (a) schematic representation of typical events, and (b) event identification procedure.

2.3. Development of weak learner data mining models

This study employed seven WLDMs, namely DT, KNN, NB, NPR, SVM, discriminant analysis (DA), and Gaussian process regression (GPR), which were chosen based on their widespread use and potential in previous hydrological classification studies (Zounemat-Kermani et al., 2020). These models were developed in MATLAB 2021a using rainfall features and water level class (data cubes in Fig. 4) to forecast flooding in UDS for various lead times up (limited here up to 12-timesteps). Additionally, the 5-fold cross-validation method was employed to reduce error bias, which is commonly used in practical data mining applications (Gharib and Davies, 2021). All identified events were randomly distributed across the training, validation, and testing databases to ensure equal representation. All models were optimised using automatic hyperparameter optimisation in MATLAB 2021a by minimising the five-fold cross-validation loss over 30 iterations, as detailed in Table A1 in Appendix A. Each model was optimised individually for all lead times, ranging from 1 to 12 timesteps. The optimisation process for flood forecasting for one timestep ahead are presented in Figure A1 in Appendix A. As a result, 84 models were developed, consisting of

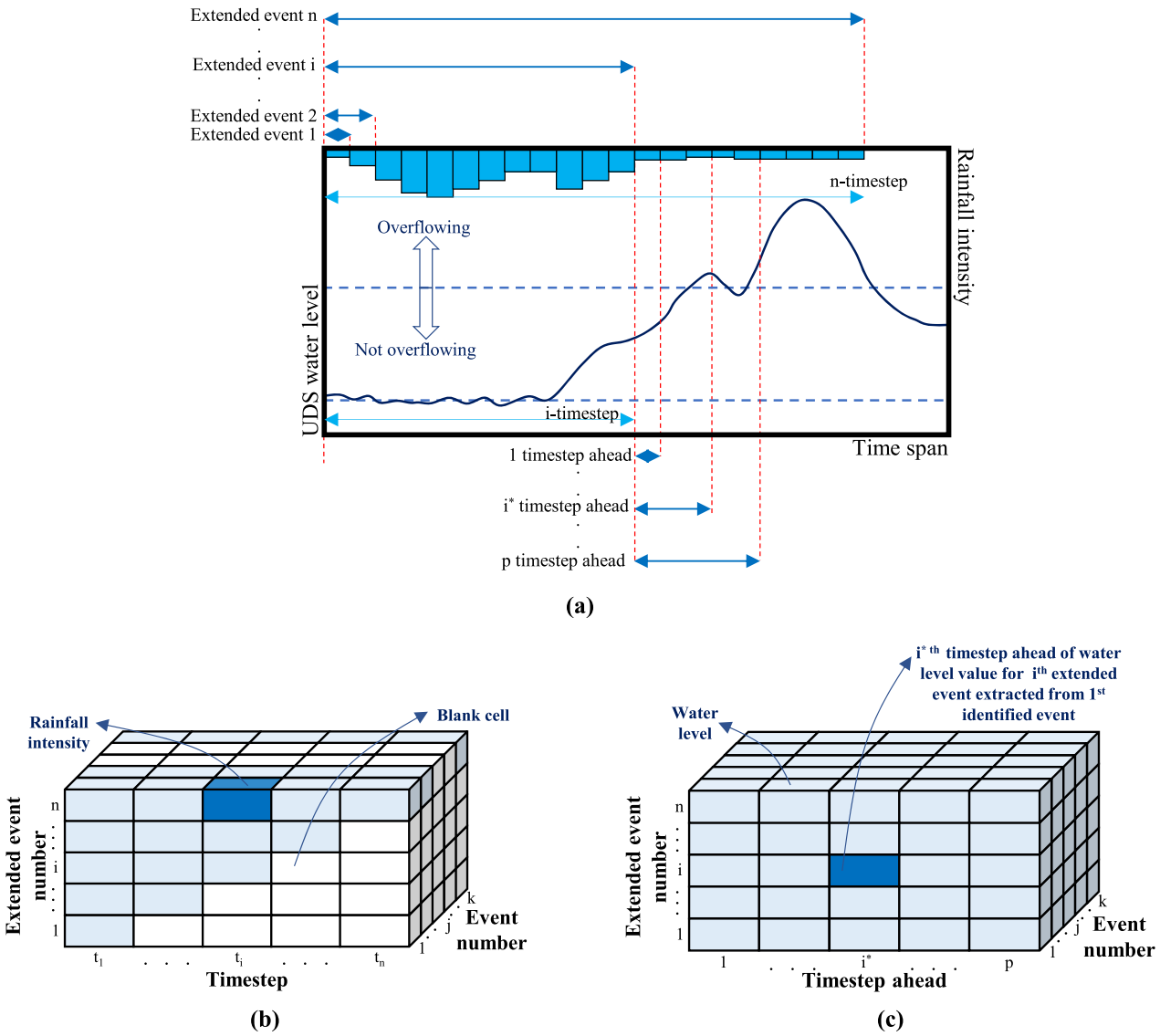


Fig. 3. Conceptual development of extended events based on the identified events: (a) Schematic hyetograph of extracting extended events from one identified overflow (flood) event (S_4 in event identification method) and hydrograph used for extracting required water level data for the extended *i*th event, (b) structure of data cube for rainfall data, (c) structure of data cube for water level data.

Table 1
Potential rainfall features extracted for developing data mining models.

Group feature	Extracted rainfall feature	Code	Description	Transformation key	Unit/ class
Current rainfall characteristics	Duration	F_1	Time period of between the onset and end of the precipitation	Numerical	min
	Depth	F_2	Maximum water depth if all rainfall cumulated in saturated impervious surface	Numerical	mm
	Intensity	F_3	The ratio of total depth to the duration	Numerical	mm/h
	Peak depth	F_4	Maximum rainfall intensity	Numerical	mm
Antecedent precipitation history	Occurrence	F_5	Previous rainfall occurred until maximum previous period equalled to time of concentration	Binary	0:No 1:Yes
	Average intensity	F_6	The average rainfall intensity of previous rainfall occurred until maximum previous period equalled to time of concentration	Numerical	mm/h
Time occurrence	Season	F_7	A different class of humid temperate climate	Class	1:Dry 2:Mild 3:Rainy
	Long-term similarity	F_8	Average of past 10 years' rainfall intensity for a similar duration of current event	Numerical	mm/h

seven WLDMs for each of the 12 different lead times and stored in library shown in Fig. 6a.

To evaluate the performance of these models in flood forecasting, this study employed the confusion matrix concept as a statistical classification technique (Tharwat, 2021). The process involved mapping the forecasted water level classes (i.e., overflow or not overflow) onto the

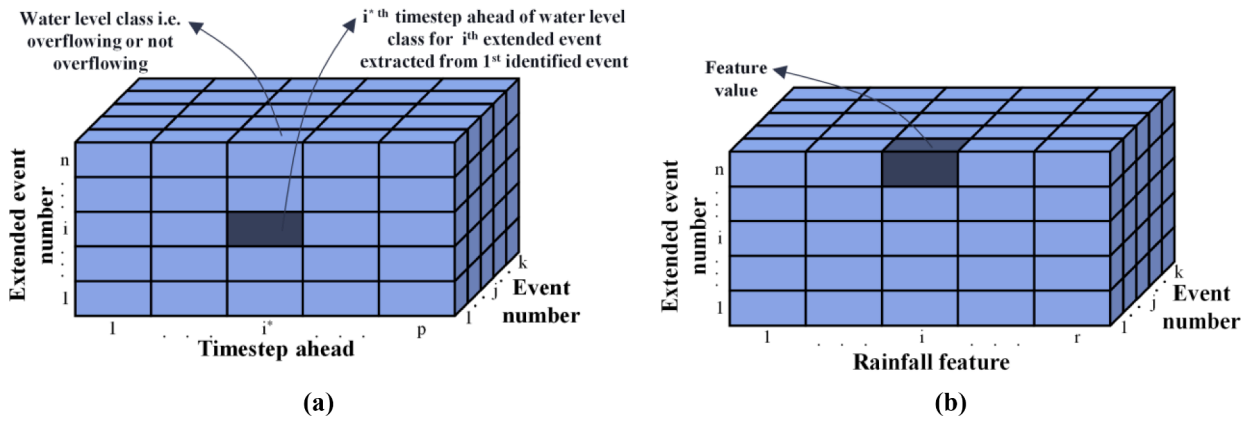


Fig. 4. Schematic structure of data cube used for developing data mining models: (a) water level class, (b) rainfall features.

confusion matrix, with a schematic representation shown in Fig. 5. The comparison of observed and forecasted water levels yielded four possible states for evaluating the model's performance in flood forecasting: True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP). Using this mapping of the confusion matrix, this study defined seven key performance indicators (KPIs) for assessing the performance of the developed models. These KPIs were selected based on their wide application and are presented in Table 2 (Grandini et al., 2020; Chicco et al., 2021). As each KPI represents an independent aspect of model performance, the non-parametric Friedman ranking test (FRT) was employed to compare the average rank of the selected KPIs for all models (Ghosh et al., 2022).

To facilitate analysis and minimising ambiguity in the individual models' KPIs, their associated KPIs and Friedman rank of all 84 models were stored in a data warehouse, as depicted in Fig. 6b in a data cubic structure for 1–12-timesteps. Subsequently, the built data warehouse was utilised to develop time-series ensemble models for further analysis.

2.4. Development of dynamic ensemble model

This study developed ensemble models by combining multiple WLDMs to create a more robust and accurate forecasting model. Due to the homogeneous nature and high variance and bias of flood forecasting, the stacking method was selected as it is well-suited for creating ensemble models (Prasad et al., 2021). In this method, all WLDMs are trained on the same set of training data and then blended using a variety of techniques such as weighted averages, decision trees, voting, Bayesian averaging, or machine learning algorithms (Yao et al., 2022). Two ensemble models are developed here for flood forecasting, as schematically shown in Fig. 7 for k-timestep of forecasting. The first model is a hybrid model, which takes into account the different abilities of WLDMs. The second model is a smart model, which uses an initiative decision framework for selecting the best classification. Relevant pre-trained WLDMs are recalled from the data warehouse and used for flood forecasting in a specific lead time, such as k-timestep ahead.

The hybrid model selects individual models that outperform others with respect to TPR, F_1-score, MCC, DP, and CKR indicators for flood forecasting in k-timestep ahead. It employs a simple but effective voting approach to determine the forecasting class. Similarly, the smart model selects two individual models with a higher rate of TPR and TNR indicators for flood forecasting in k-timestep ahead. The smart model uses a decision tree framework, as shown in Fig. 7b, to determine the water level class by using FRT and ACC ratios of the selected models and considering the purpose of the selected model. The model prioritises the highest TPR and TNR rates for flooding and non-flooding, respectively. Selected WLDMs are dynamically varied in each timestep of forecasting which equips the ensemble models to utilise the full potential of all 84 developed WLDMs.

In addition to assessing the multistep forecasting performance, a new event-based performance assessment is introduced in this study. It evaluates the accuracy of ensemble models in predicting each individual event, whether it is a flood or non-flood event. The forecasted events are divided into two classes: missed classes and hit classes. Missed classes occur when the model is unable to forecast the event at any timestep, while hit classes occur when the model predicts the event correctly, with or without a time lag. Missed classes can result in complete incorrect

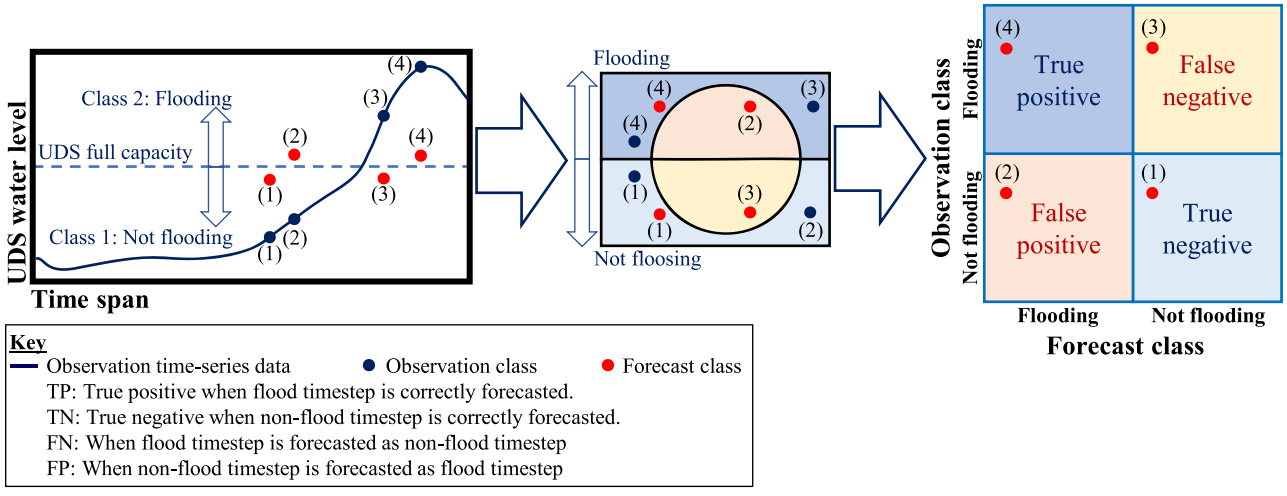


Fig. 5. Schematic visualisation of mapping classification of forecasted data onto the confusion matrix concept.

Table 2
Selected key performance indicators used for performance assessment of WLDMs.

Code	Description	Formula
TPR	Model sensitivity in recalling actual flood condition, i.e., accuracy of flood class	$\frac{TP}{TP+FN} \times 100$
TNR	Model specificity in selecting actual non-flood condition, i.e., accuracy of non-flood class	$\frac{TN}{TN+FP} \times 100$
ACC	Probability in that the model forecasting is correct, i.e., interested in forecasting the right classes without caring about the type of the class or class distribution	$\frac{TP+TN}{n} \times 100$
MCC	Highlighting correlation and agreement between observed and predicted classes	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$
DP	Determining the likelihood of correct flood and non-flood conditions	$\frac{\sqrt{3}}{\pi} \times \left[\log\left(\frac{TPR}{1-TNR}\right) + \log\left(\frac{TNR}{1-TPR}\right)\right]$
F_1 score	Revealing the best trade-off between overflow and not-flood forecasting by interpretation as a weighted average between PPV and TPR	$\frac{2 \times TPR \times PPV}{TPR+PPV}$
CKR	Measuring the concordance between ACC, TPR, and TNR	$\frac{ACC - [TPR \times (1-TPR) + TNR \times (1-TNR)]}{1 - [TPR \times (1-TPR) + TNR \times (1-TPR)]}$

ACC: ACCuracy of true classification.
CKR: Cohen's Kappa Rate.
DP: Discriminant Power.
F_1-score: Harmonic mean.
MCC: Matthews Correlation Coefficient.
PPV: Positive Predictive Value.
TNR: True Negative Rate.
TPR: Total Positive Rate.

forecasting, as shown in Fig. 8b,i,j, or partial incorrect forecasting, as shown in Fig. 8f,l (underestimation) and b–d (overestimation). On the other hand, correct forecasting of each event (hit class) can occur in various situations, which are illustrated in Fig. 8 and listed in Table 3 for brevity.

3. Results and discussion

The proposed methodology's results for real-time forecasting of water level class at the gauging station of a real-world case study are presented using unseen original test data. To facilitate discussion, the proposed methodology is compared to three commonly used stacked conventional methods: selecting the best WLDMs in each timestep, voting, and Bayesian averaging (Rahman et al., 2021; Zounemat-Kermani et al., 2021; Piadeh et al., 2022a). 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. The performance assessment of flood forecasting is carried out for up to 12 timesteps.

This study reports the results of a proposed methodology for real-time forecasting of water level class at a gauging station in a real-world case study described in Section 3.1. The methodology is compared with three commonly-used stacked ensemble models, namely choosing the best TPR-performed WLDMs in each timestep (Piadeh et al., 2022a), voting, and Bayesian averaging (Zounemat-Kermani et al., 2021; Aswad et al., 2022). These benchmark models are trained and validated using the same original database and features as those used for training and validation of the proposed methodology. The performance of flood forecasting is assessed for up to 12-timestep of flood forecasting on unseen original test data.

3.1. Study area and time-series data acquisition

Fig. 9a displays the location of the Ruislip gauging station, RAF Northolt rain gauge, and urban catchment area. The Ruislip UDS, an open channel situated in the northwest of London, Borough of Hillingdon, collects surface runoff from a catchment area of 9.3 km^2 and transfers it through the river Pinn to a tributary of the River Thames in England. Rainfall events, as illustrated in Fig. 9b, were mainly recorded throughout the year, with a duration and depth typically less than 600 min (i.e., 10 h) and 10 mm, respectively. These rainfall events resulted in several fluvial floods and water escaping into Ruislip's urban neighbourhoods, leading to road traffic, significant water puddles on pavements, and damage to properties and infrastructure. The Ruislip gauging station measures real-time UDS water levels every 15 min, and the rain gauge station, as shown in Fig. 9a, was selected based on the prevailing wind direction in the pilot study (i.e., southwest) to obtain 15-min rainfall data. The entire database comprises 365,233 data samples for both rainfall and water level, spanning a continuous 12-year period (2009–2021), which can be accessed through the application programming interface of the UK Environment Agency (DEFRA, 2022). Missing values were infilled using the copula-based regression method recommended by Ben Aissia et al. (2017). The input data was first divided into two subsets, with 80 % and 20 % of the data allocated for model development of WLDMs and time-series ensemble models, respectively. Specifically, the WLDMs were trained and validated on the 60 % subset and tested on the 20 % subset, which served as unseen data.

The time of concentration (ToC) for the urban catchment area as depicted in Fig. 9c is estimated to be approximately 79 min, based on the longest length of 9.3 km, an average slope of 18.3 %. Consequently, ToC is assumed to be approximately 6-timesteps long (90 min) in the feature extraction analysis. It should be noted, however, that using this parameter to exclude durations of historic rainfalls larger than ToC is challenging due to the non-uniformity of historic rainfall patterns observed through time-series analysis of rainfall hyetograph data, and the lack of historic hydrograph data for water level in which the water

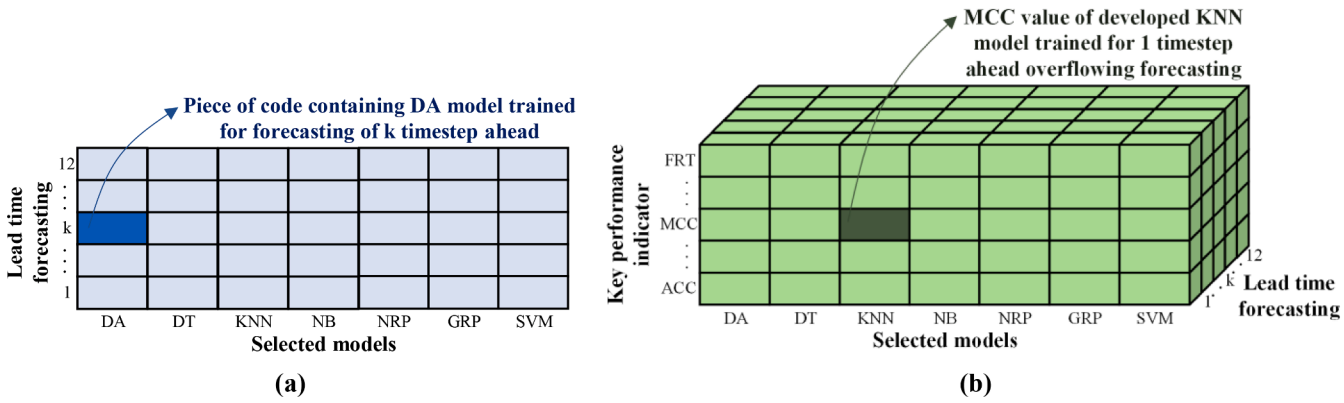


Fig. 6. Schematic illustration of constructed data warehouse: (a) library of developed WLDM models, (b) structure of performance data cube.

level began to stabilise during the rainfall.

3.2. Feature selection analysis

Fig. 10a presents the results of PCA and PLS analysis for all eight rainfall features in the database (refer to Table 1). The analysis indicates that rainfall duration (F_1) and intensity (F_3) are the most influential features, followed by antecedent precipitation occurrence (F_5) and season type (F_7). On the other hand, rainfall depth (F_2) does not have a strong correlation compared to these four rainfall characteristics, especially when rainfall duration and intensity are used. In addition, the occurrence of antecedent rainfall is determined to be more important than its intensity (6 times more in PCA and 4 times in PLS). Furthermore, long-term similarity of rainfall occurrence is excluded, possibly due to the uncertainty in the rainfall distribution that may occur on the same date in different years.

The results of sensitivity analysis for all 84 WLDMs are also presented in Fig. 10b with detailed data available in Table A2 in appendix A. The results confirm the features identified by PCA/PLS analysis and demonstrate consistency between these three analytical methods. However, the results indicate that rainfall intensity is the most sensitive feature affecting the accuracy performance of the models. While the range of differences is almost identical for the season, rainfall duration, and intensity, different models exhibit varying degrees of sensitivity for antecedent rainfall occurrence.

3.3. Performance of WLDM models

The results of the best-performing WLDMs for three KPIs, namely CKR, TPR, and TNR, across different lead times are shown in Table 4 (also see Tables A3 and A4 and Figure A2 in Appendix A). Findings reveal that none of the WLDMs exhibited superior performance in the majority of KPIs, and thus, no ideal WLDM can be chosen from Table 4. Moreover, the trend of the decreasing rate of the best model performance varies across different lead times concerning each KPI. Specifically, TNR is insignificantly impacted by various lead times, with a minor drop from 98 % for 15 min to 94 % for 3 hrs, compared to TPR (which shows a significant drop from 87 % for 15 min to 66 % for 3 hrs) and CKR (with a moderate drop from 92 % for 15 min to 78 % for3 h). This implies that the accuracy of true flood forecasting (i.e., TPR) is highly sensitive to lead times, while this has a minor impact on correct forecasting non-flood events (i.e., TNR). Such results may stem from the large correct forecasting of non-flood events relative to the small number of forecasted floods during the test period. Comparing TPR with TNR confirms this result that WLDMs can identify non-flood states more accurately than flood events, indicating a high miss rate of WLDMs for early flood warning.

Table 5 presents the results of the Friedman test performance metric for the WLDMs across various lead times, as well as the rankings of the models in the last row based on the Friedman average metrics of all lead times. The analysis reveals that the DA model is ranked first, followed by the GPR model in second place, and the NB model in third place. Specifically, the DA model achieved the best metric for six different lead times, which represents 50 % of all lead times. While analysis identified the top-performing models based on various KPIs and lead times, none of the models consistently outperformed the others across all KPIs and lead times. Therefore, to achieve accurate time-series multistep flood forecasting, it is recommended to use an dynamic approach that selects or combines the best-performing models based on the specific lead time and KPI of interest.

3.4. Performance of ensemble models

The accuracy performance of the ensemble models for each lead time is depicted in Fig. 11. The accuracy of the forecast for the hit rate, which denotes the correct detection of flood or non-flood events, ranges from 80 % to 90 % for a 15 min lead time. Specifically, the voting-based model achieved an 80 % hit rate, the TPR-based model achieved an 90 % hit rate, and other models achieved over 90 % hit rate. However, the hit rate of the voting-based model significantly decreased to nearly 65 % for 12-timestep, whereas the TPR-based and weighting-based models maintained an 75 % hit rate. Moreover, the miss rate, which refers to the incorrect detection of events, varied significantly across the models. For instance, the TPR-based model tends to classify non-flood events as flood events, while the voting-based model has a greater tendency towards underestimation of forecasts, i.e., forecasting non-flood events as flood events. This difference in miss rate is particularly evident for longer lead times, as illustrated in Fig. 11b and c.

The analysis of the accuracy performance of the weighting-based model in terms of the balance between the rate of underestimation and overestimation shows similar results to the TPR-based model for a 3-hour lead time (refer to Fig. 11a vs 11b and c). This suggests that the weighting-based model has a greater ability to predict non-flood events as compared to accurately forecasting flood events. The lower rate of flood forecasting and a significant share of underestimations observed in the voting-based model is reasonable, given the inherent ability of the WLDMs to better forecast non-overflow events. It is also expected that the weighting-based model may have higher overestimation and underestimation rates due to the use of all WLDMs, including models with lower TPR and TNR scores. However, the flexible use and application of a weighted average of the WLDMs in different timesteps can help overcome this limitation and improve the accuracy performance of the model by reducing the range of both overestimated and underestimated predictions.

Comparison of the performance of all ensemble models reveals that the hybrid and smart models outperform the other three ensemble models with respect to the four metrics shown in Fig. 11. This superiority may be attributed to the development of ensemble models based on

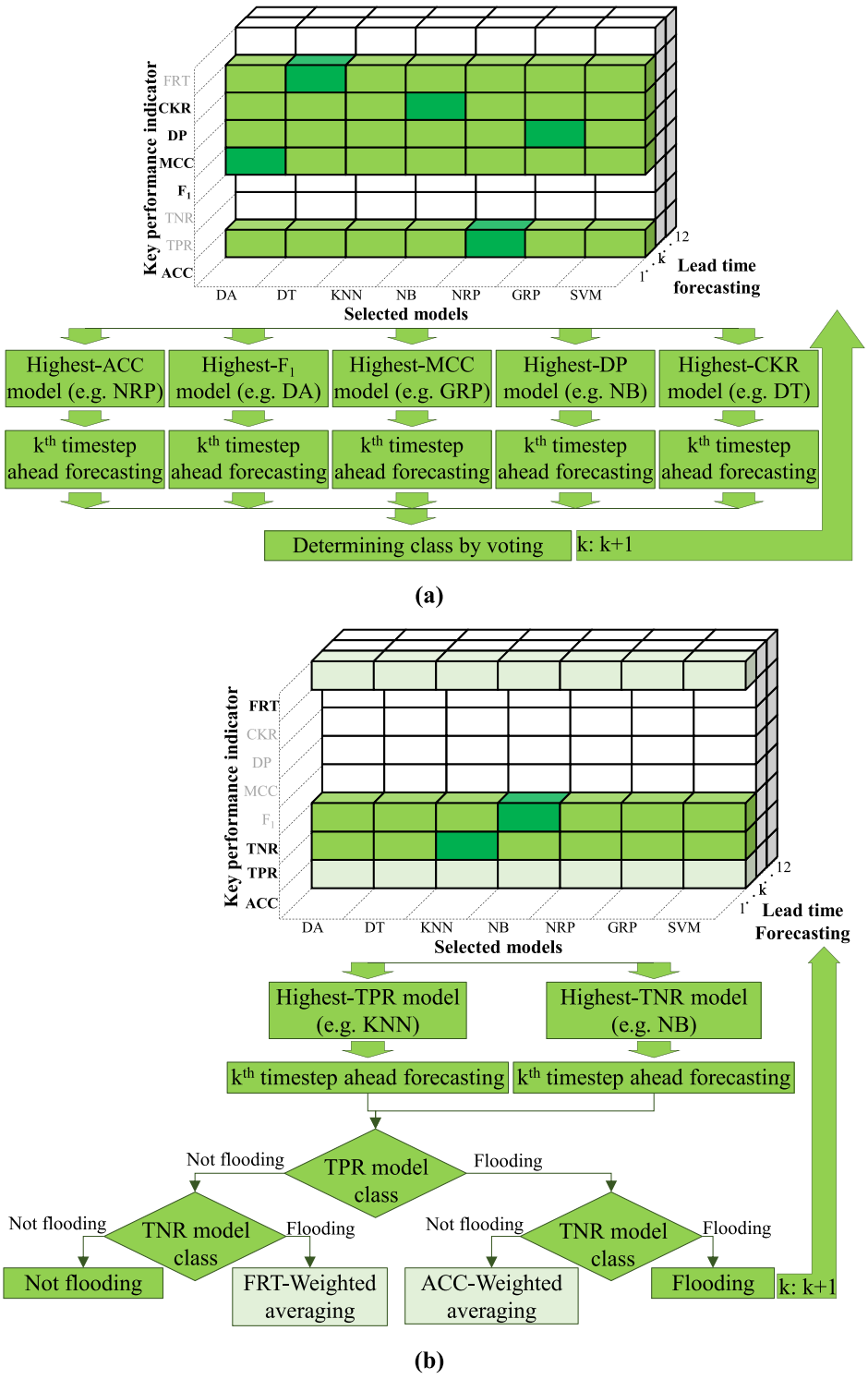


Fig. 7. Visual representation of sacked generalisation of developed dynamic ensemble models: (a) hybrid model and (b) smart model.

combined key performance indicators (KPIs) such as CKR and MCC, leading to the emergence of the hybrid model (Fig. 11d), which exhibits a remarkable improvement in forecasting non-overflow states. Consequently, overestimated and underestimated rates are improved not only in the 15 min but also in the 3 h lead time, by more than 90 % and 85 %, respectively. However, rate of the flood forecasting in the hybrid model drops to 70 % for the 3 h lead time. On the other hand, the smart model, which selects the best TPR and TNR models through a smart decision framework as shown in Fig. 7b, outperforms the hybrid model, especially in longer lead times, such as 3 h. Although the underestimated rate in the smart model seems to be higher than that in the hybrid model for some specific lead times, such as 1:15 h or 1:30 h, the flood forecasting rate remains above 80 % for all forecasted timesteps. The superior flood forecasting performance achieved by the smart model can also be compared with other RNN models developed in previous studies for urban flood forecasting, such as those reported by Noymanee et al. (2017), Mosavi et al. (2018), and Wagenaar et al. (2020), which determined accuracies of around 70 % for 2 h lead times.

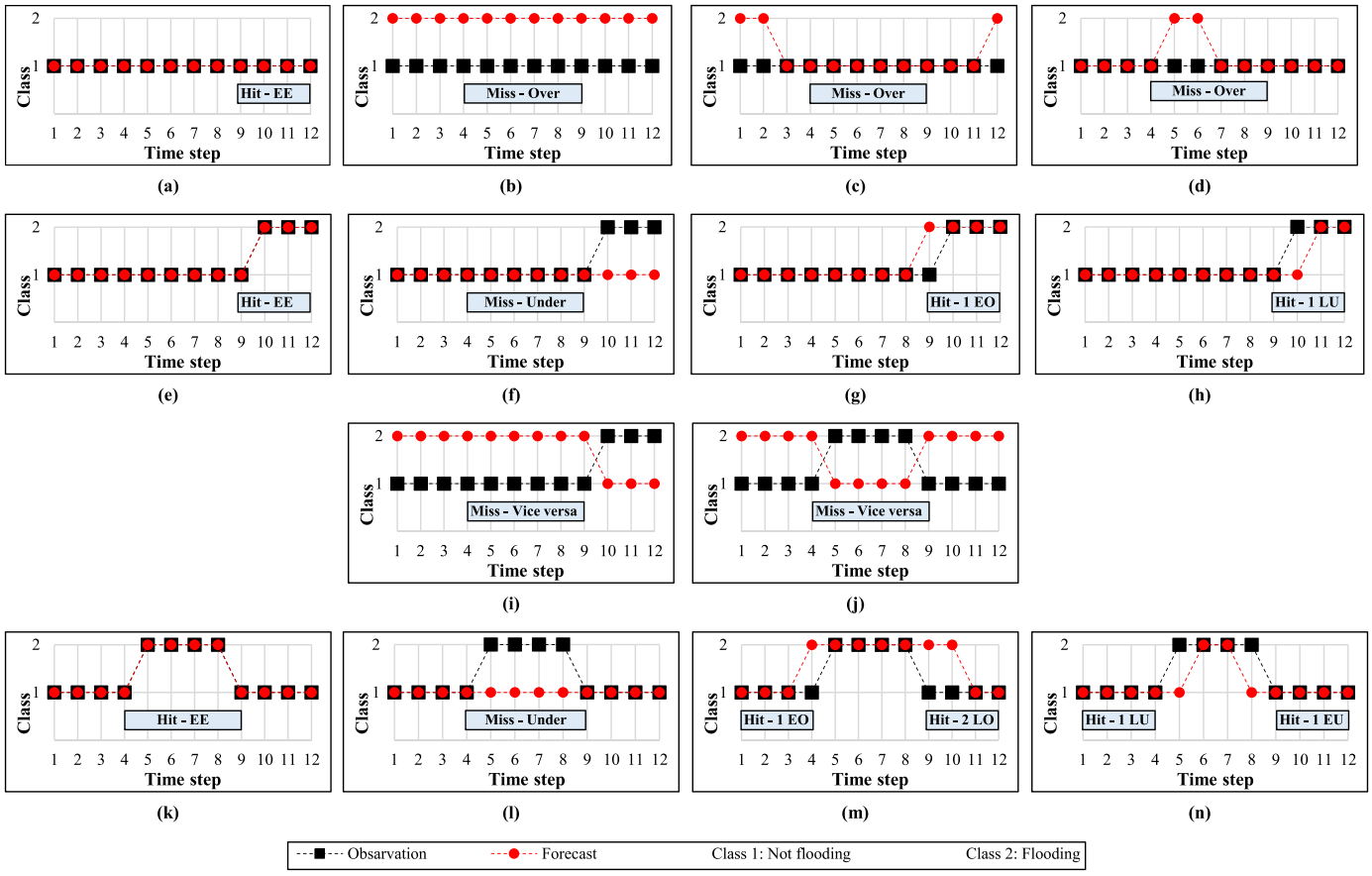


Fig. 8. Different examples of event-base performance assessment: (a–d) Non overflow events, (e–j) non-ended Overflow events, (k–n) ended Overflow events.

Table 3
Event-based performance assessment proposed for dynamic ensemble models.

Miss class[1]			Hit class[2]		
Under	Over	Vice versa			
(A)	(B)	(C)	**Earlier**	**Exact**	**Late**
Average lag		**Under**	(D)	–	(E)
		Exact	–	(F)	–
		Over	(G)	–	(H)

1: When the model cannot forecast the true conditions at any event such as non-flood for flood events and flood for non-flood events.
2: When the model can forecast the event correctly regardless of any time lag.
(A): When the model forecasts non-flood condition for all timesteps of flood event (see Figure F8f and F8l).
(B): When the model forecasts the flood condition for all timesteps of non-flood event (see Fig. 8b, c, and d).
(C): When the model forecasts non-flood for flood condition and flood for non-flood condition (see Fig. 8j and k).
(D): When the model forecasts non-flood for flood condition earlier than the exact timestep (see Fig. 8i and o).
(E): When the model forecasts non-flood for flood condition in timesteps later than the exact timestep (see Fig. 8h and o).
(F): When the model forecasts all timesteps precisely (see Fig. 8a, e and l).
(G): When the model forecasts flood for non-flood conditions earlier than the exact timestep (see Fig. 8g and m).
(H): When the model forecasts flood for non-flood conditions later than the exact timestep (see Fig. 8n).

The performance of five ensemble models is presented in Fig. 12, where three real events with a 3-h time window (12-timesteps) were simulated. While the majority of ensemble models accurately predicted the non-flood event (as evidenced by predictions in the left graphs of Fig. 12), the TPR-based and voting-based models exhibited some misses and tended to overestimate this event within the last timesteps. For the flood event shown in the middle graphs, all models could forecast the flood occurrence, but the TPR-based and voting-based models indicated flood classes for the non-flood situation in earlier timesteps (i.e., 3 to 6-timesteps). This suggests that these models may estimate flood occurrence earlier than usual, which seems initially be useful for an early warning system, but the considerable time lag may necessitate excessive costs and efforts for further risk management actions. Conversely, the weighting-based model missed two timesteps of flooding (see Fig. 12h) and announced flood occurrence later than the measured situation. Alternatively, the hybrid and smart models could forecast this event with minor time lags. The privilege of these models was highlighted in the third event example (right column in Fig. 12), where the TPR-based model in Fig. 12c or the weighting-based model in Fig. 12i missed the flood condition or the voting-based model falsely alarmed an overflow condition (see Fig. 12f). However, against the hybrid model, the smart model could forecast the classes of this event correctly, with only one missing timestep at the 10th timestep.

For a comprehensive assessment, the model performance of all developed ensemble models is presented in Fig. 13, based on the introduced event-based sheet in Table 3. Further details are provided in Table B1 and Table B2 in appendix B. Overall, the miss rate of TPR-based and weighting-based models is significantly higher compared to other models, with rates of 35 % and 26 %, respectively, in contrast to 12 % for TPR-based, and only 5 % for both hybrid and smart models. Despite better accuracy of multistep forecasting achieved by the weighting-based model compared to TPR and voting-based models, this model shows a higher miss rate of event forecasting.

Comparing the results in Fig. 13 with those in Fig. 11, it can be seen that although the TPR-based model has better accuracy in multi-

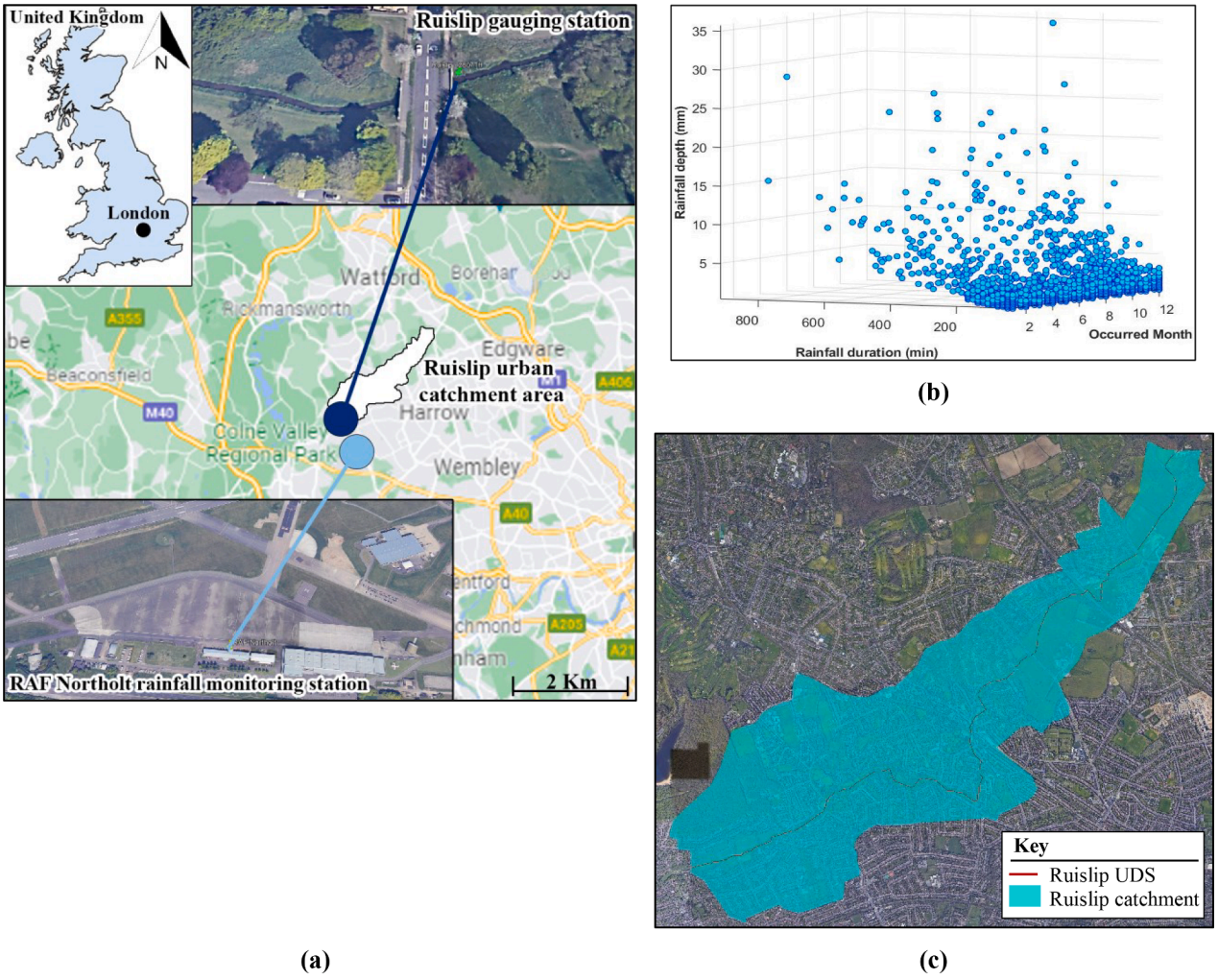


Fig. 9. Geographical map and hydrological data of the pilot study: (a) location of stations and layout of catchment, (b) Characteristics of recorded rainfalls and (c) layout of Ruislip UDS and catchment.

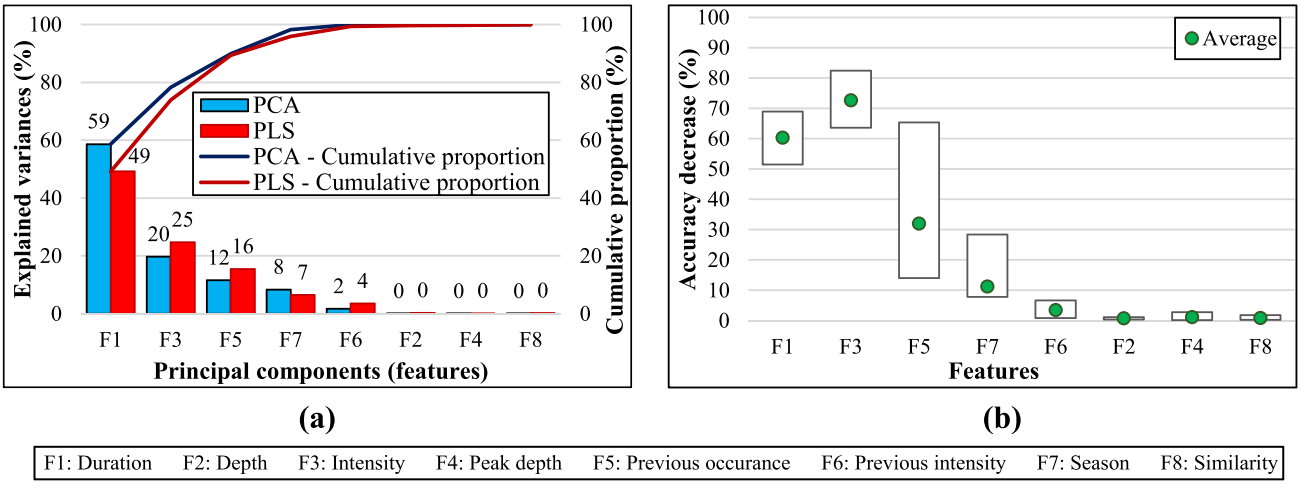


Fig. 10. Performance of the potential rainfall features as input data of the WLDM models based on (a) PCA and PLS methods, (b) sensitivity analysis.

forecasting than the voting-based models, it shows a higher miss rate in events forecasting. This suggests this model is likely to forecast many non-flood events as flood conditions in real-time applications. On the other hand, a high rate of underestimation obtained by the weighting-based model can result in missing flood conditions and consequently a lack of proper risk action in the concept of early warning systems.

The findings presented in Fig. 13b indicate that the voting-based model is unable to accurately forecast the behaviour of events, i.e., correctly predicting all timesteps in one event. Only 17 % of the total events are precisely forecasted by this model, whereas the TPR-based

Table 4
Best WLDMs based on the selected three key performance indicators.

	KPIs		
Lead time	CKR	TPR	TNR
15 min	DA (0.92)	NB (87 %)	DT (98 %)
30 min	DA (0.92)	NB (88 %)	DT (97 %)
45 min	KNN (0.90)	NB (87 %)	KNN (98 %)
1 h	GPR (0.90)	NB (84 %)	GPR (98 %)
1:15 h:min	DA (0.87)	DA (83 %)	DT (95 %)
1:30 h:min	KNN (0.87)	DA (81 %)	DT (95 %)
1:45 h:min	DA (0.85)	NB (75 %)	DA (95 %)
2 h	DA (0.85)	DA (76 %)	KNN (94 %)
2:15 h:min	KNN (0.80)	NB (73 %)	SVM (94 %)
2:30 h:min	NRP (0.80)	NRP (70 %)	SVM (93 %)
2:45 h:min	DA (0.79)	NRP (67 %)	SVM (94 %)
3 h	SVM (0.78)	NB (66 %)	SVM (94 %)

Table 5
Friedman test ranking of the WLDMs in each lead time.

Timestep (Lead time)	Average ranking of WLDM						
	DA	DT	GPR	KNN	NB	NRP	SVM
15 min	**(1.7)**	5.9	3.3	2.1	5.3	4.6	5.1
30 min	**(1.4)**	6.1	4.0	2.4	5.0	3.7	5.4
45 min	4.0	6.3	3.3	**(1.7)**	5.0	2.4	5.3
1 h	3.0	6.0	**(1.7)**	3.3	5.4	4.3	4.3
1:15 h:min	**(1.4)**	5.9	4.3	2.1	5.6	3.1	5.6
1:30 h:min	2.0	5.4	4.0	**(1.3)**	6.0	4.3	5.0
1:45 h:min	**(1.1)**	5.6	4.9	2.3	6.0	3.6	4.6
2 h	**(1.3)**	5.9	4.6	2.3	6.0	3.4	4.6
2:15 h:min	3.0	6.4	4.9	**(1.7)**	5.3	2.4	4.3
2:30 h:min	3.0	6.6	4.4	2.2	5.7	**(1.9)**	4.1
2:45 h:min	**(1.7)**	6.7	4.6	2.9	5.6	3.0	3.6
3 h	3.1	6.9	4.3	3.3	5.0	**(2.0)**	3.4
Average	**2.23**	6.13	2.30	5.49	3.23	4.01	4.61
Rank	**1**	7	2	6	3	4	5

and weighting-based models demonstrate superior performance with 30 % and 48 %, respectively. This implies that, while the miss rate of the voting-based model is significantly lower than that of the other two models (12 %), it primarily relies on overestimating flood conditions earlier or later than the actual event (32 % for later-under and 25 % for later-over conditions, respectively), potentially resulting in additional costs and unnecessary alert procedures in real-time systems.

Conversely, the hybrid and smart models exhibit a miss rate of approximately 5 %, wherein flood events are primarily underestimated, as illustrated in Fig. 13a. Despite the hit rate of these models (Fig. 13b) being more than 90 %, the hybrid model could only forecast the event entirely in approximately half of the total events (52 % of the exact rate). Further scrutiny of events reveals that it is susceptible to overestimation in the latest timesteps (refer to Fig. 8g). Consequently, the hybrid model is best suited for the nearest lead time. On the other hand, it appears that the smart model is appropriate for real-time early warning systems due to its highest rate of exact forecasting (71 %), relatively lower rate of false alarms (13 % for later-over), and a lower rate of underestimation (7 % for earlier-under) in later timesteps.

However, it is worth noting that these rates only reflect the quantity of forecasted events, and the quality of forecasting necessitates further analysis through lag time specification. The distribution of lag times for different overestimation and underestimation forecasting is presented in Fig. 13c and d. Although benchmark models exhibit high ranges of earlier and later lags, hybrid and smart models can reduce lag time by less than 1.5-timesteps (approximately 22 min). More importantly, the smart model's low rate of underestimated lag time for forecasted events, whether earlier or later (1.2 and 1.1, respectively), alleviates concerns of missing long timesteps of overflowing.

3.5. Further analysis and recommendation

This study presents a time-series ensemble model that utilises water level and rainfall data as inputs to improve the accuracy of flood forecasting. While this novel approach offers potential benefits in urban flood forecasting, further validations by other time-series models are necessary to assess its efficacy in diverse contexts. In order to generalise the findings, the proposed methodology underwent testing on two additional UDS stations: Eastcote and Willow Bank (Detailed geographical information can be found in Part 1 of Appendix C). This extension aimed to enhance the applicability of the approach to more complex and diverse geographic regions. The results clearly illustrate the superior performance of the proposed hybrid and smart model when compared to the benchmark models, as visualised in Fig. 14. However, the outcomes also reveal a noteworthy trend: as the distance between the rainfall station and the UDS increases, the accuracy of the models tends to decrease, as one would expect. For instance, in the case of Eastcote, where the rainfall station is situated at twice the distance from Willow Bank (5.09 km compared to 2.08 km), the accuracy of flood forecasting and hit accuracy both experience a reduction of approximately 1.5 % and 2.5 %, respectively, for a 3 h ahead of forecasting. While this decline might not be drastic, it highlights the influence of the spatial separation between rainfall and UDS stations. Consequently, it is recommended that future research include sensitivity analysis focusing on the impact of the distance between the data source of rainfall and the UDS station. This becomes especially pertinent when dealing with alternative forms of rainfall data, such as satellite or radar data. Moreover, despite the relatively consistent outcomes across different UDS stations, the applicability of this model should be further explored within real-time forecasting scenarios in other applications such as reservoir or river basins. This expanded investigation could lead to a broader range of practical applications for the developed model.

The assessment delved into the influence of alternative available rainfall stations, encompassing scenarios where data from various stations were combined. Diverse rainfall merging techniques, as outlined in Piadeh et al. (2022a), were evaluated and finally resulted in the selection of two optimal techniques: KED (Kriging with external drift) for interpolation and MSF (multiquadric surface fitting) for bias adjustment. Additional details are provided in Part 2 of Appendix C. The findings, illustrated in Table 6, shows a compelling trend using a single rainfall station that exhibits stronger correlation with the UDS (refer to Part 1 in Appendix C) yields better results in comparison to scenarios where less correlated stations are integrated into the data mining models. At first glance, this outcome might seem to challenge recommendations from other research works, such as Zounemat-Kermani et al. (2020), Zounemat-Kermani et al. (2021), and Paideh et al. (2022a). However, it is important to recognise that the methodology employed here revolves around identifying flood events, encompassing both rainfall occurrences and rising water levels. The manipulation of more correlated rainfall data, coupled with the adjustment or interpolation of less correlated stations, can potentially alter the duration and intensity of rainfall events. This, in turn, exerts a significant influence on the proposed models, as previously elucidated in the feature analysis depicted in Fig. 10. This conclusion gains further support from the observation that the KED techniques are notably more sensitive to the data from other rainfall stations. This heightened sensitivity translates into a considerable increase in the underestimation rate, particularly evident when employing the KED merging technique. This results in an upward spike of miss rates, exceeding 30 % for KED merging, in contrast to approximately 20 % for MSF merging.

Finally, data from two other rainfall stations, exhibiting stronger correlation, were integrated as eight additional features to construct WLDM models. This step aimed to assess the influence of these supplementary resources on the model's accuracy (Further details can be found in Table C2 in Appendix C). The outcomes of the sequential sensitivity analysis conducted on these newly introduced features yield a notable

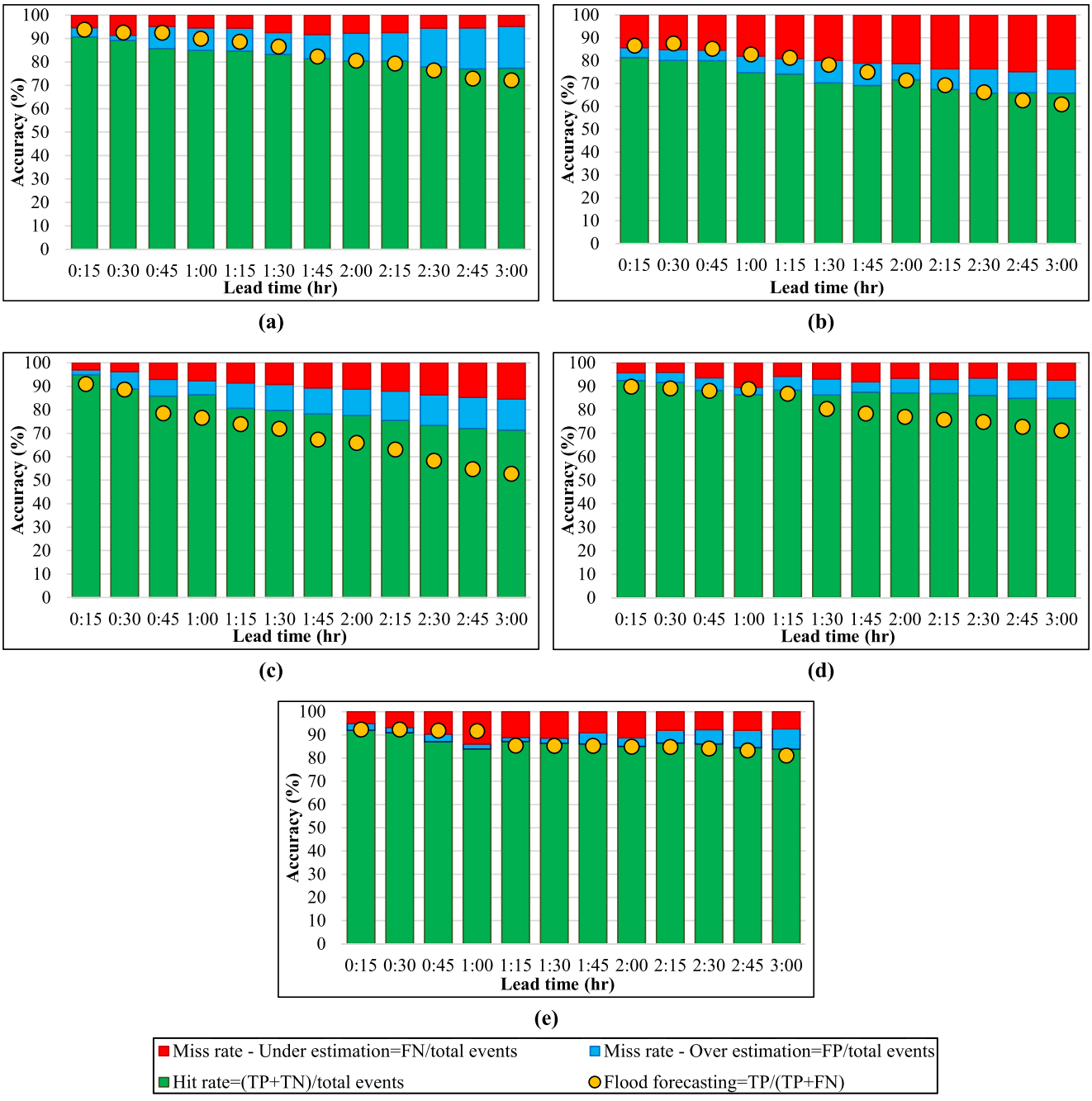


Fig. 11. Performance of the time-series ensemble models for each lead time: (a) TPR -based, (b) voting-based, (c) weighting-based, (d) hybrid, and (e) smart models.

revelation: none of the newly incorporated features succeeded in enhancing the accuracy of the WLDM models. Specifically, the parameters of depth and peak depth displayed minimal impact on model accuracy, effectively registering near-zero improvements. Moreover, when considering extended lead times, the inclusion of duration from other rainfall stations as an additional rainfall feature shows a significant 57 % decrease in accuracy for forecasting flood classes in next 12-timestep ahead. This outcome strongly indicates that the model misconstrues the primary rainfall characteristics, causing it to lose its ability to discern the prevailing conditions upon the addition of other rainfall events. This result suggests that the model's performance can be affected when confronting with the incorporation of rainfall events from secondary sources.

Besides, as mentioned earlier, this model follows a direct relationship between rainfall data and UDS water level. The other approach i.e., providing rainfall forecasting and using its output for forecasting water level should also be tested and compared with this approach. Additionally, the selected KPIs for the latest lead times require improvement, which can be achieved by partitioning water level data into more classes or incorporating classified rainfall data to develop models. The proposed model also serves as a comparative analysis and requires further evaluation and enhancements to ensure its effectiveness in diverse settings.

4. Conclusions

This paper presented a dynamic ensemble data mining modelling approach for real-time flood forecasting, with a focus on its application in urban drainage systems. The study analysed the main rationales behind the event identification method, rainfall feature extraction, the development of different WLDMs, and the blended ensemble

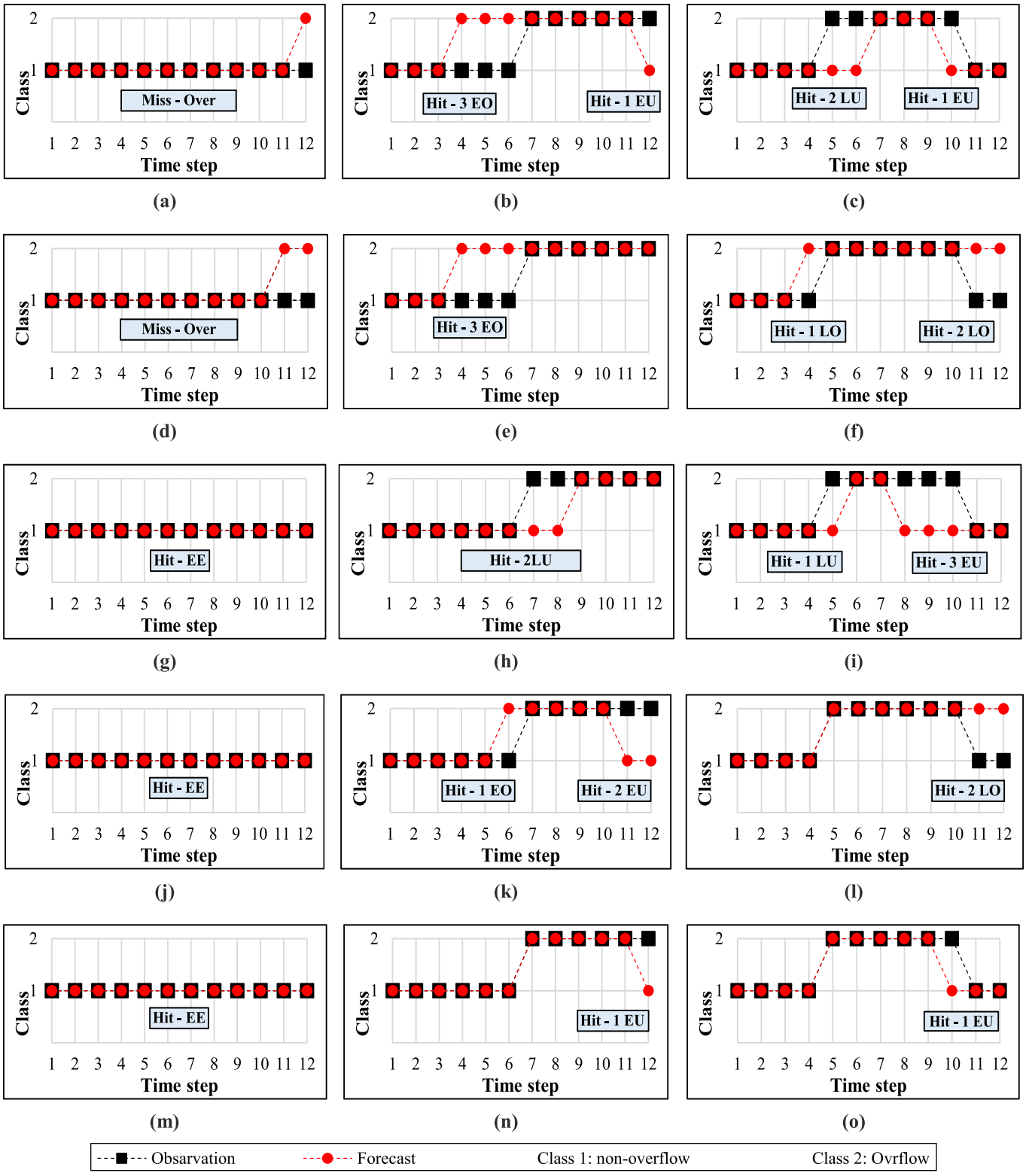


Fig. 12. Performance of the dynamic ensemble models for three event examples: (Left): non-flood event, (Middle): flood event occurring in the middle and continuing after 3 h, (Right) flood occurring and finishing within 3 h, (a–c): TPR-based, (d–f): voting-based, (g–i): weighting-based, (j–l): hybrid, (m–o): smart models.

approaches. The approach combined time-series forecasting with an ensemble modelling technique to improve the accuracy of non-flood and flood detection, particularly for longer timesteps, which has yet to be properly addressed in previous attempts. The use of applied data cubes and data warehouse significantly aids in constructing the real-time platform, and the KPI performance of WLDMs for blended ensemble models shows advances in developing an efficient early warning system. While the framework is applied to a case study, the main research findings are summarised as follows:

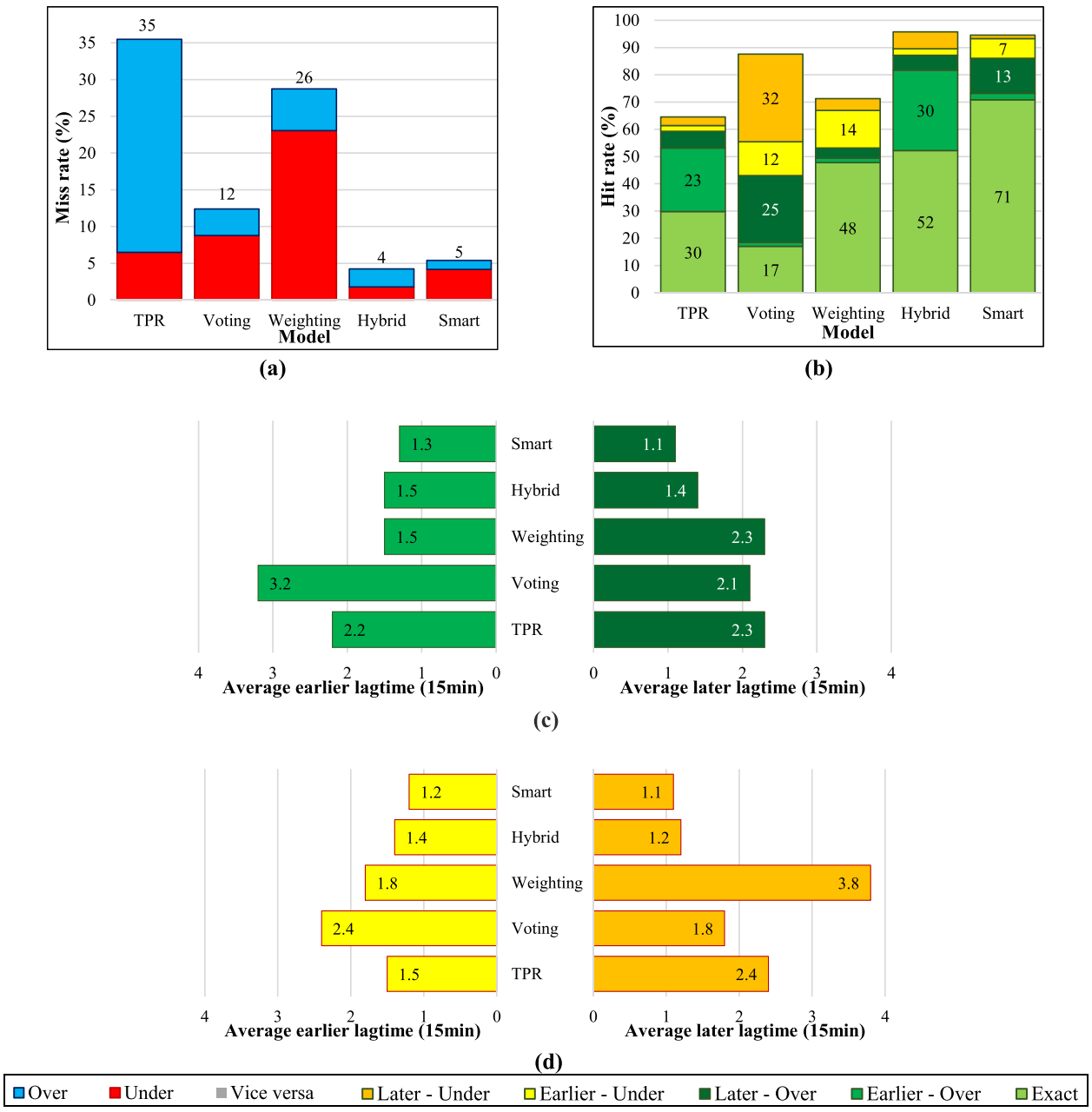


Fig. 13. Event-based performance of the ensemble models: (a) event miss rate, (b) event hit rate, (c) average overestimation time lags, (d) average underestimation time lags.

- The smart model achieved a higher accuracy in forecasting events (71 %), with a correct classification of all timesteps in each event without any lag time. This was higher than the hybrid model's accuracy of 52 % and the weighting-based model's accuracy of 48 %. This results in reducing false alarms, unnecessary risk management actions, and extra costs in real-time early warning applications.
- Although both the voting-based hybrid model and decision-based smart model achieved a hit rate near 90 % for 3 hrs ahead of forecasting, the higher rate of the hybrid model mainly relied on successful forecasting of non-flood events. This finding was supported by a significant reduction (20 %) between 15 min and 3 hrs flood forecasting rates.
- The event-based miss rate of the smart model was limited to only 5 %. 13 % of the total event-based hit rate was forecasted with an average of 1.1 overestimation lag time (16 min earlier) and 7 % was forecasted with an average of 1.3 underestimation lag time (20 min later). These lag times were significantly better than the at least 30 min delay or 40 min earlier flood forecasting reported by other developed models.
- Comparison of different event-based hit rates, i.e., with or without lag time flood forecasting, shows that there is no correlation between the distribution of lag time and their values. For example, the rate of flood forecasting with earlier false alarm (EO) is very small for the voting-based model (1.43 %) in comparison to the rate of overflow forecasting with later overestimation (LO) (25 %). However, the value of the EO lag time is documented more than the LO lag time (3.2 vs. 2.1). This finding highlights the importance of both the quantity (rate of different hits) and quality (average lag times) of event-based performance assessment to provide a comprehensive evaluation of models in forecasting different events.
- The proposed features named "short time history of past rainfall" and "seasonal time occurrence of current rainfall" significantly improved

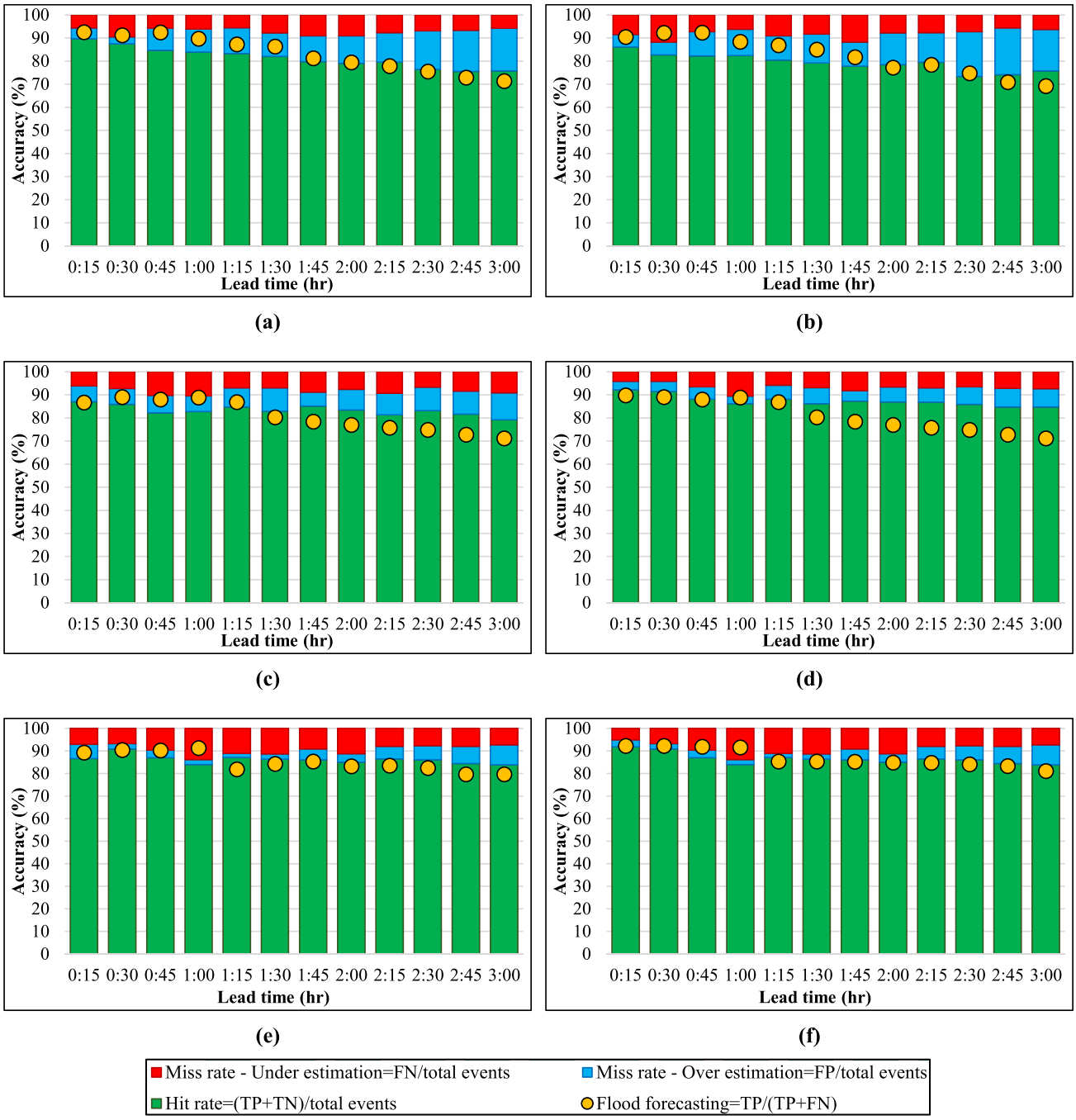


Fig. 14. Performance of the time-series ensemble models for each lead time: (left): Eastcote UDS station, (right) Willow Bank UDS station, (a and b) best performed benchmark model, (c and d) hybrid model, and (e and f) smart model.

Table 6
Miss rate of the developed ensemble models for other merging models.

Model	Miss class rate (%)			Total rate (%)
	Under	Over	Vice versa	
Hybrid model				
Best correlated Single station	1.79	2.44	0	4.23
KED merging	25.47	4.76	0	**30.22**
MSF merging	11.46	7.97	0	19.43
Smart model				
Best correlated Single station	4.18	1.23	0	5.41
KED merging	26.87	10.68	0	**37.54**
MSF merging	10.94	8.09	0	19.03

KED: Kriging with external draft MSF: Multiquadric surface fitting.

the accuracy of the WLDMs. Rainfall duration and rainfall intensity had a greater impact on WLDM accuracy than other rainfall characteristics such as total depth or peak depth.

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.

Data availability

Data will be made available on request.

Acknolwedgment

Chen work in the manuscript is supported by the Devon Resilience Innovation Project (DRIP), a part of the Defra/EA's Flood and Coastal Resilience Innovation Programme.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2023.120791.

References

Aswad, F., Kareem, A., Khudhur, A., Khalaf, B., Mostafa, S., 2022. Tree-based machine learning algorithms in the Internet of Things environment for multivariate flood status prediction. J. Intell. Syst. 31 (1), 1–14.

Bui, D., Tsangaratos, P., Thi Ngo, P., Dat Pham, T., Thai Pham, B., 2019. Flash flood susceptibility modeling using an optimized fuzzy rule based feature selection technique and tree based ensemble methods. Sci. Total Environ. 668, 1038–1054.

Chicco, D., Tötsch, N., Jurman, G., 2021. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min. 14 (13), 1–22.

Gharib, A., Davies, E., 2021. A workflow to address pitfalls and challenges in applying machine learning models to hydrology. Adv. Water Resour. 152, 103920.

Ghosh, S., Saha, S., Bera, B., 2022. Flood susceptibility zonation using advanced ensemble machine learning models within Himalayan foreland basin. Nat. Hazard. Res. 2 (4), 363–374.

Grandini, M., Bagli, E., Visani, G., 2020. Metrics for multi-class classification: an overview. Comput. Sci. 1–17. ArXiv: 2008.05756.

Hosseini, F., Choubin, B., Mosavi, A., Nabipour, N., Shamshirband, S., Darabi, H., Haghighi, A., 2020. Flash-flood hazard assessment using ensembles and Bayesian-based machine learning models: application of the simulated annealing feature selection method. Sci. Total Environ. 711, 135161.

Masahiko, H., Cian, F., Lall, U., 2019. Leveraging Global and Local Data Sources for Flood Hazard Assessment and Mitigation: an Application of Machine Learning to Manila. United Nations Office for Disaster Risk Reduction. Contributing Paper to GAR 2019.

Mobini, S., Pirzamanbein, B., Berndtsson, R., Larsson, R., 2022. Urban flood damage claim analyses for improved flood damage assessment. Int. J. Disaster Risk Reduct., 103099

Mosavi, A., Ozturk, P., Chau, K., 2018. Flood prediction using machine learning models: literature review. Water 10 (11), 1536.

Munawar, H., Hammad, A., Waller, S., 2021. A review on flood management technologies related to image processing and machine learning. Autom. Constr. 132, 103916.

Noymanee, J., Nikitin, N., Kalyuzhnaya, A., 2017. Urban pluvial flood forecasting using open data with machine learning techniques in Pattani Basin. Proc. Comput. Sci. 119, 288–297.

Orellana-Alvear, J., Veldhuis, J., Clemens, F., 2021. Event-based evaluation of a hydrodynamic flood forecasting system: application to a complex urban area. Water (Basel) 13 (5), 665.

Piadeh, F., Behzadian, K., Alani, A.M., 2021. The role of event identification in translating performance assessment of time-series urban flood forecasting. In: 15th *UWL Annual Conference*. London: UK. Available at. repository.uwl.ac.uk/id/eprint/9113 [Accessed 27/02/2023].

Piadeh, F., Behzadian, K., Alani, A.M., 2022a. A critical review of real-time modelling of flood forecasting in urban drainage systems. J. Hydrol. 607, 127476.

Piadeh, F., Behzadian, K., Alani, A.M., 2022b. Multi-step flood forecasting in urban drainage systems using time-series data mining techniques. In: Water Efficiency Conference. West Indies: Trinidad and Tobago.

Piadeh, F., Behzadian, K., Chen, A., Campos, L., Kapelan, Z., 2023. Event-based decision support algorithm for real-time flood forecasting in urban drainage systems using machine learning modelling. J. Environ. Modell. Softw. 167, 105772.

Prasad, P., Loveson, V., Das, B., Kotha, M., 2021. Novel ensemble machine learning models in flood susceptibility mapping. Geocarto Int. 37 (16), 4571–4593.

Rahman, M., Chen, N., Elbeltagi, A., Islam, M., Alam, M., Pourghasemi, H., Tao, W., Zhang, J., Shufeng, T., Faiz, H., Baig, M., Dewan, A., 2021. Application of stacking hybrid machine learning algorithms in delineating multi-type flooding in Bangladesh. J. Environ. Manag. 295, 113086.

Rasheed, Z., Aravamudan, A., Sefidmazgi, A., Anagnostopoulos, G., Nikolopoulos, E., 2022. Advancing flood warning procedures in ungauged basins with machine learning. J. Hydrol. 609, 127736.

Shahzad, H., Myers, B., Boland, J., Hewa, G., Johnson, T., 2022. Stormwater runoff reduction benefits of distributed curbside infiltration devices in an urban catchment. Water Res. 215, 118273.

Tharwat, A., 2021. Classification assessment methods. Appl. Comput. Inform. 17 (1), 168–192.

Wagenaar, D., Curran, A., Balbi, M., Bhardwaj, A., Soden, R., Hartato, E., Mestav Sarica, G., Ruangpan, L., Molinario, G., Lallemant, D, 2020. Invited perspectives: how machine learning will change flood risk and impact assessment. Natl. Hazards Earth Syst. Sci. 20, 1149–1161.

Yaseen, Z., Sulaiman, S., Deo, R., Chau, K., 2019. An enhanced extreme learning machine model for river flow forecasting: state-of-the-art, practical applications in water resource engineering area and future research direction. J. Hydrol. 569, 387–408.

Yao, J., Zhang, X., Luo, W., Liu, C., Ren, L., 2022. Applications of Stacking/Blending ensemble learning approaches for evaluating flash flood susceptibility. Int. J. Appl. Earth Observ. Geoinform. 112, 102932.

Zang, Y., Meng, Y., Guan, X., Lv, H., Yan, D., 2022. Study on urban flood early warning system considering flood loss. Int. J. Dis. Risk Reduct. 77, 103042..

Zhang, Z., Tian, W., Liao, Z., 2023. Towards coordinated and robust real-time control: a decentralized approach for combined sewer overflow and urban flooding reduction based on multi-agent reinforcement learning. Water Res. 229, 119498.

Zhou, Y., Wu, Z., Xu, H., Wang, H., Ma, B., Lv, H., 2023. Integrated dynamic framework for predicting urban flooding and providing early warning. J. Hydrol. 618, 129205.

Zounemat-Kermani, M., Matta, E., Cominola, A., Xia, X., Zhang, Q., Liang, Q., Hinkelmann, R., 2020. Neurocomputing in surface water hydrology and hydraulics: a review of two decades retrospective current status and future prospects. J. Hydrol. 588, 125085.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., Hinkelmann, R., 2021. Ensemble machine learning paradigms in hydrology: a review. J. Hydrol. 598, 126266.