Multi-step flood forecasting in urban drainage systems using time-series data mining techniques

Water Efficiency Conference 2022

Multi-Step Flood Forecasting in Urban Drainage Systems Using Time-series Data Mining Techniques

Farzad Piadeh[1], Kourosh Behzadian *[1], Amir Alani[1]

[1] *School of Computing and Engineering, University of West London, St Mary's Rd, London, W5 5RF, UK*

ABSTRACT

While early warning systems are recognised as the most cost-effective solution in urban flood risk management, highly accurate flood forecasting is limited to short-term timesteps, usually less than a few hours especially for prediction of overflowing in urban drainage systems . This study aims to provide a framework for more accurate overflow predictions for longer lead times by using data mining models applied to time series data for multi-step flood forecasting. The framework including event identification, feature analysis and developing models is demonstrated by its application to a pilot study in London. All numerical rainfall data and water levels in urban drainage systems are first turned to the categorical events on which 6 common weak learner models are developed. Then, three new time-series models, including overflowing-based, non-overflowing-based, and accuracy-based, are developed based on these models to predict overflow states among all identified events. Three weak learner models, i.e. discriminant analysis, naive Bayes, and decision tree are considered as the best models based on accuracy, total overflowing detection and total non-overflowing detection. Furthermore, while the accuracy of these models is changed between 95 to 85% from 1 to 12-step ahead of prediction, these models can detect the non-overflow conditions better than overflow detection. To cover this gap, new time series developed models could significantly reduce the overestimation and underestimation of water levels, including correct predicting of 50% of the total events after 12-step ahead by overflow-based model. This result shows the potential of using time-series data-demanding models for effective and highly accurate predictions of overflow events.

Keywords: Data mining; Drainage system; Flooding classification; Multistep prediction Overflow prediction

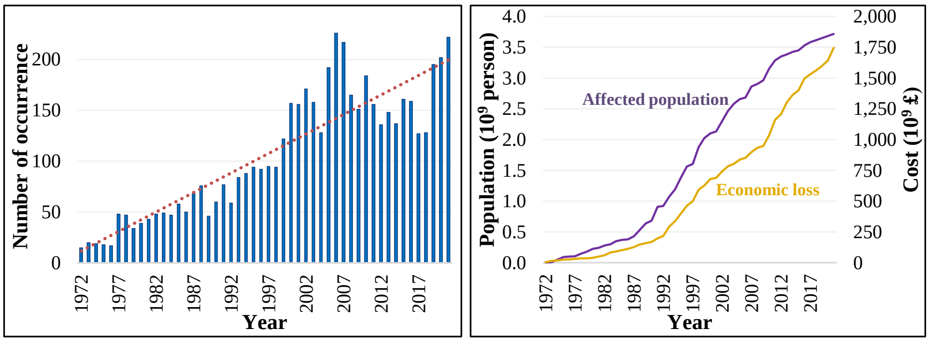
1. INTRODUCTION

Flooding is recognised as a worldwide natural hazard, which is responsible alone for over 30% of global economic loss and 60% of the total affected people by all types of natural hazards [1]. According to UNDRR [2], the number of flood occurrences increased in the recent 50 years (Figure 1a), in which more than 3.5 billion people have been affected and near 1,750 billion pounds loss is estimated so far (Figure 1b). Therefore, early warning systems can now commerce to a reliable and practical solution for predicting floods' overflowing of drainage systems by using weak learner data mining models (WLDM). However, they are unable to forecast flood for long time steps due to complicated non-periodic and chaotic mechanism of rainfall occurrence and weak correlation between flooding and drainage systems' water level [3].

Previous research works applies WLDM such as support vector machine (SVM), k-nearest neighbourhood (KNN), discriminant analysis (DA), decision tree (DT), Gaussian process regression (GPR), naive Bayes (NB), and neural network pattern recognition (NNPR) to forecasting water level in urban drainage systems (UDS) [4,5]. They were also used to determine overflow conditions in which the flow in UDS exceeds the full capacity of UDS and spills into urban areas and causes flooding. However, other data mining models (DMs), particularly feedback forward and recurrent neural network (RNN) models show more potential due to representing a significant forecasting accuracy, handling big data and high-speed computation [6]. Despite the good performance of these DMs, the accuracy of forecasting overflow in these models for periods longer than 60 minutes is reduced significantly [1]. To tackle this, hybrid models have been developed recently in which WLDM are mixed with RNN

* Tel.: +44 (0) 20 8231 2466
E-mail address: kourosh.behzadian@uwl.ac.uk

[7]. While these models could increase the accuracy of overflow prediction in near lead time, they are still unable to provide reliable and accurate estimation for lead time longer than two hours. [8,9]. Additionally applied WLDM focus heavily on increasing model accuracy in only one specific time step [10] instead of following the concept of time-series DMs for multi-step prediction [11,12].



(a) **(b)**

Figure 1. Recent 50-year recorded flood events all around the world: a) number of flood occurrences, b) associated cumulative social and economic loss

Hence, the present study aims to propose flexible time-series WLDM models to fill the above gaps, including (1) providing more accurate overflow predictions for longer lead time, and (2) using the concept of time series DMs to find the best method for multi-step prediction, (3) investigating model performance on not only different lead times, but also multistep overflow detection of flood events. This model can enable the development of a high-speed and outperformed real-time overflow classification model that can be trained based on a limited temporal set of data, i.e. only rainfall and water level in UDS. The proposed framework along with the case study in the UK will be described in the next section followed by presenting research findings, critical discussion, and finally highlighting key findings and final remarks in conclusions.

2. METHODOLOGY

The proposed framework as shown in Figure 2 comprises two main parts: (1) data collection and preparation, and (2) model development and performance assessment. The time-series of rainfall and water level data of UDS (described in section 2.1) collected from a public domain online database and their missing data are infilled by linear regression, are used to identify both flood and non-flood events [13]. Identified numerical events are then turned into categorial order, named hereafter featured events, through the method proposed in Section 2.2. Several widely used WLDMs are developed based on these featured events which are introduced in Section 2.3. The time-series data mining models for predicting overflow conditions in floods and non-floods events are developed and evaluated based on the model performance criteria introduced in section 2.4.

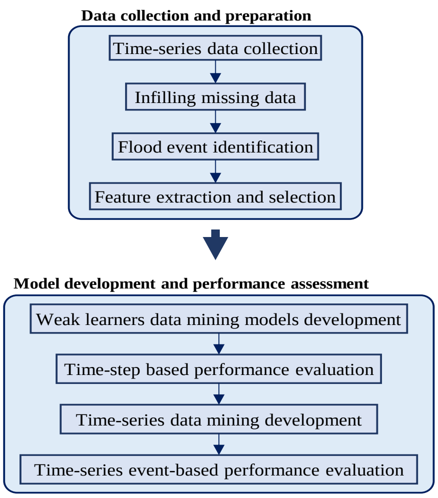


Figure 2. The proposed framework for flood overflow detection

2.1 Data collection and study area

The proposed framework is demonstrated here through its application for forecasting flood overflow in a real-world UDS pilot study in the UK. Figure 2 shows the entire catchment area located in the London Borough of Hillingdon including the Ruislip urban catchment area analysed in this pilot study. The Ruislip UDS drives the Colne catchment surface runoff from south Hertfordshire to a tributary of the River Thames in England. The UDS located in the northwest of London collects the surface runoff through the river Pinn from a catchment area of 13 km^2. The pilot study was selected due to its vulnerability to frequent fluvial flooding over the Ruislip urban neighbourhoods. Ruislip gauging station in the river Pinn located at the outlet of the Ruislip UDS is one of the 55 gauging stations installed in the Colne catchment area and is responsible for measuring and recording the water level. An ultrasonic depth monitor system is used to record the time-series of water level every 15 min at the station since 2009 [14]. Furthermore, the rainfall observed every 15 minutes at RAF Northolt rain gauge station, shown in Figure 2, is also used here [14]. The entire database includes 365,233 data for both rainfall and water level, 15-minute time intervals and a continuous duration of 11 years (2009-2020) are accessible through the application programming interface (API) of the UK Environment Agency [14].

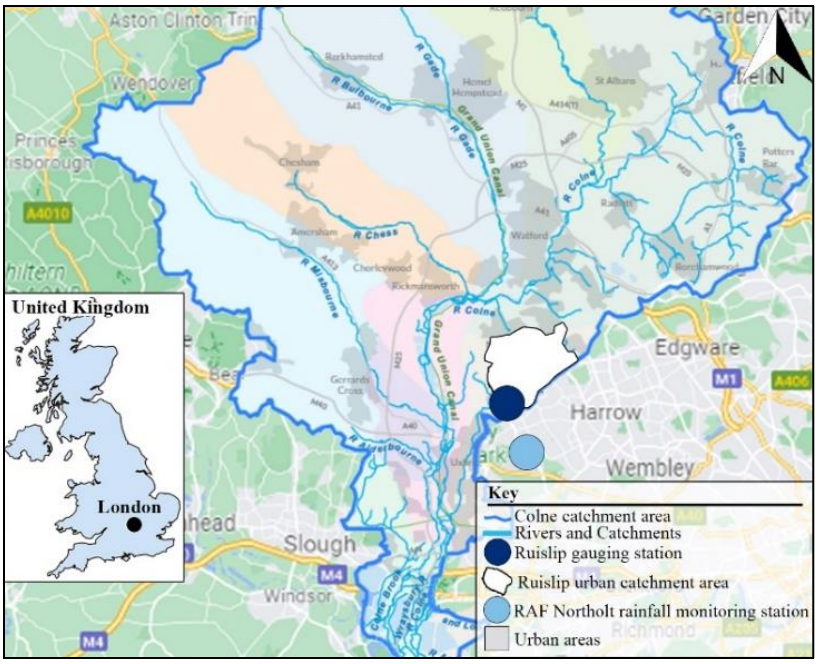


Figure 2. The layout of the case study and location of rainfall and gauging stations

2.2 Feature extraction and selection

Based on rainfall and water level records, data are firstly converted to (1) overflow events i.e., rainfall occurrence causes water level rising in the UDS and overflowing, (2) non-overflowing events, i.e. despite rainfall, there is no water level rise as outlined in the event identifications [13]. Then all identified events are converted into several features of rainfall events listed in Table 1. Principle component analysis [15] is used as criteria (See Figure 3) to determine the final parameters including (1) rainfall duration, (2) rainfall intensity, (3) intensity of previously occurred rainfall, (4) season of the event and (5) overflowing state.

Table 1. Extracted and final selected features for turning identified events to featured events

Group feature	Extracted feature	Description	Type of used data	Transformation
Rainfall data	Duration (F_1)	Duration of rainfall occurrence in the area of interest	Actual	Timestep (Every 15 mins)
	Total depth (F_2)	Liquid precipitation covering a horizontal surface area of interest	Actual	mm
	Intensity (F_3)	The average rainfall rate for a specific duration	Actual	mm/hr.
	Peak depth (F_4)	The maximum amount of rainfall Intensity	Actual	mm
Previous rainfall	Occurrence (F_5)	Evidence showing rainfall occurred before the current time	Categorised	0 (No) 1 (Yes)
	Intensity (F_6)	The average rainfall intensity of previously recorded rainfall	Actual	mm/hr.
Date of the year	Season (F_7)	Time of the year	Categorised	1(Dry) 2 (Mild) 3 (Rainy)
	Long-term history (F_8)	Average rainfall intensity of this date for the past 10 years	Actual	mm/hr.
Overflow	Existence	State of water level in comparison to full capacity of UDS	Categorised	Class 1 (No overflowing) Class 2 (Overflowing)

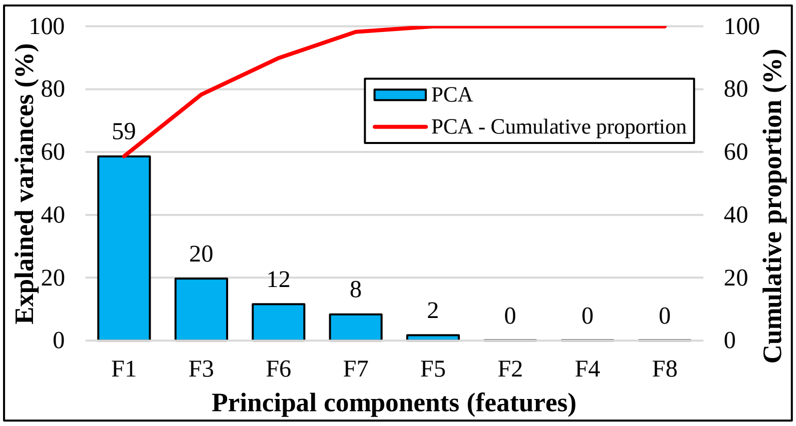


Figure 3. Principal components analysis (PCA) on extracted features

2.3 Developed data mining models

Based on the literature, 7 WLDMs are selected to develop here, including: (1) DA, (2) DT, (3) GPR, (4) KNN, (5) NB, (6) SVM, and (7) NNPR. Models are developed based on 2 classes (1) non-overflow state, and (2) overflow detection. Time-series models are also developed based on the process shown in Figure 4, in which 12 timesteps of each event are predicted by a specified developed model. Here, WLDMs are used to build three time-series models based on the best overall accuracy (called "ACC" model), the best overflow detection (called "TPR" model) and the best non-overflow detection (called "TNR" model) of previous developed WLDM. For this purpose, among all developed WLDM, the best model is used for the prediction of the event's class in first time step ahead, and then this process goes continued iteratively for further time steps.

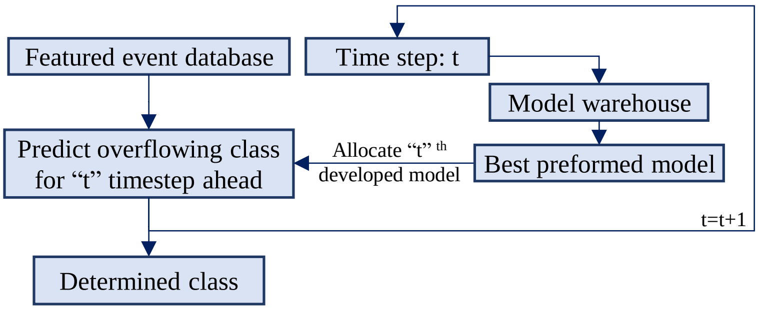


Figure 4. Iterative process of developing time-series data mining models

All models are built for prediction from 1-step to 12-step ahead using MATLAB 2021a and then individually optimised based on 30 different iterations. 75% and 25% of total events are used for building WLDMs and time-series models, respectively. All WLDM models are developed based on 70% (53% of the total database) for model training, 15% (11% of the total database) for validation and 15% of data for test. In each timestep, training, validation and test dataset are generated randomly based on characteristics of the entire database, meaning 70% of non-overflow events and 30% of overflow events. All developed WLDM models are stored in a model warehouse (library) used then for developing time-series models.

2.4 Key performance indicators

Performance of WLDM models is evaluated by using three main indicators listed in Table 2, including (1) accuracy, (2) total correct detection of overflow events, and (3) Total correct detection of non-overflow events. Besides, developed time-series models evaluated based on the confusion matrix shown in Figure 5, consisting of (1) hit rate, predicting correct event's class in correct timestep, (2) miss rate, underestimated prediction in both event class and timestep, (3) Over rate, overestimated prediction in event class, (4) acceptable rate, predicting correct event class in correct or earlier timestep.

Table 2. Key performance indicators used for performance assessment of WLDM models

Metric	Covered concern	Equation	Range
Accuracy (ACC)	Probability in that the model prediction is correct, i.e. interested in predicting the right classes without caring about the type of the class or class distribution.	$\frac{TP+TN}{n}$	[0,1]
Total positive rate (TPR)	Sensitivity of model in recalling actual overflow condition, i.e. accuracy of overflow class	$\frac{TP}{TP+FN}$	[0,1]
Total negative rate (TNR)	Specificity of the model in selecting actual non-overflow condition, i.e. accuracy of non-overflow class	$\frac{TN}{TN+FP}$	[0,1]

TP: True overflow detection TN: True non-overflow detection n: total number of events
FP: Non-overflow event is detected as an overflow condition
FN: Overflow event is detected as a non-overflow condition

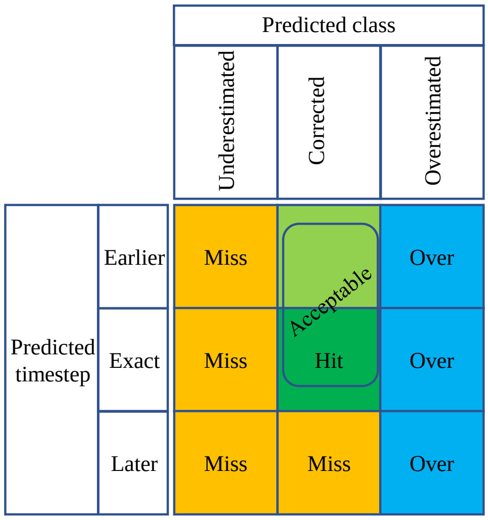


Figure 5. Structure of time-series event-based performance assessment

3. RESULTS AND DISCUSSION

Performance of developed WLDM models for different prediction timesteps (1-12 timesteps ahead) are demonstrated in Figure 6 and the best model performance is indicated in Table 3. Overall, ACC and TNR reduced from near 95% to 80% from 1 timestep to 12 timesteps for all models (Figure 6a and 6C), whereas TPR dropped up to 50% for the longest lead time (Figure 6b). While the number of the observed non-overflow event is 3 times more than overflow events, high accuracy of TNR shows that developed WLDM models are more capable to detect non-overflow events than detecting correct overflow conditions. Furthermore, Table 3 indicates that no absolute and unique WLDM model can perfectly show the best performance in comparison to other models. For example, although DA is recognised as the best ACC model, positioning the first rank in 6 out of total 12 timesteps with an average of 88.65%, it could not obtain the best TPR or TNR score, whereas NB and DT models have the best performance in these metrics, respectively.

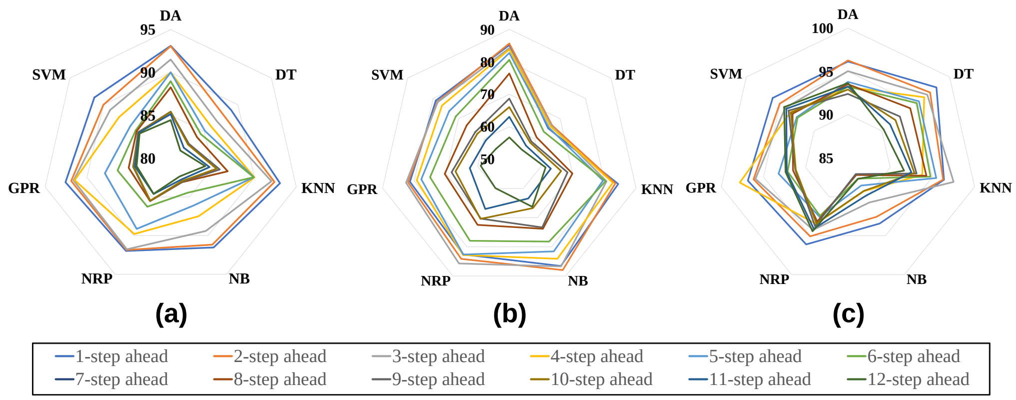


Figure 6. Performance of WLDMs in different prediction timesteps: (a) ACC, (b) TPR, (c) TNR

Based on the recognised best WLDM models for all metrics, the performance of these models is investigated in time-series modelling, which is shown in Figure 7. While the accuracy of exact prediction, i.e. correct detection of events, is generally reduced from around 85% (1 step ahead) to near 80% (12 steps ahead) for all models, the TPR-based model shows slightly better performance in which the accuracy is about 80% for all time steps. However, these models are distinguished from each other in the rate and trend of underestimation and overestimation accuracy. The ACC-based model for longer lead times tends to overestimate flood forecasting while the TNR-based model has more underestimated flood forecasting. Although the result of the TNR-based model was expected because of the ability of WLDM models to better prediction of non-overflow events in comparison to overflow events, it was expected that the TPR-based model has more underestimated forecasting because of using WLDM with the lower range of TPR rather than TNR and ACC score. However, flexible use of WLDM in time-series models could overcome this gap and shows better performance on the low range of both overestimated and underestimated predictions.

Table 3. Best WLDMs based on the key performance indicators

Time step	Best developed model		
	ACC	TPR	TNR
1	**DA** (93.10%)	**NB** (86.56%)	**DT** (98.07%)
2	**DA** (93.06%)	**NB** (87.93%)	**DT** (97.13%)
3	KNN (92.03%)	**NB** (86.68%)	KNN (97.51%)
4	GPR (91.47%)	**NB** (84.07%)	GPR (97.81%)
5	**DA** (90.01%)	DA (82.74%)	**DT** (95.49%)
6	KNN (89.92%)	DA (80.61%)	**DT** (95.16%)
7	**DA** (88.42%)	**NB** (74.63%)	DA (95.39%)
8	**DA** (88.27%)	DA (76.42%)	KNN (94.29%)
9	KNN (85.86%)	**NB** (73.35%)	SVM (93.77%)
10	NRP (85.47%)	NRP (70.44%)	SVM (93.41%)
11	**DA** (85.15%)	NRP (67.04%)	SVM (94.10%)
12	SVM (84.64%)	**NB** (66.29%)	SVM (94.42%)
Best model[1*]	**DA (6[2*], 88.65%[3*])**	**NB (7[2*], 76.94%[3*])**	**DT (4[2*], 94.47%[3*])**

1*: Best model is selected based on the Friedman test for all 12 timesteps
2*: Frequency of best model among total 12 timesteps
3*: Average value for all 12 timesteps

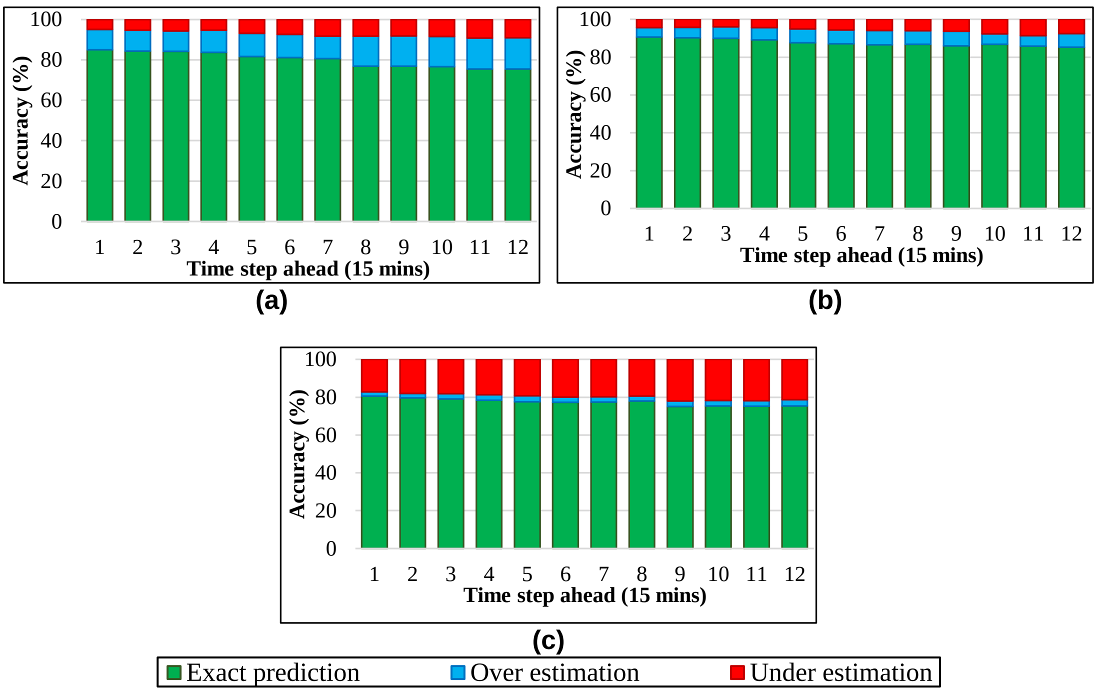


Figure 7. Performance assessment of developed time-series models in each time step: (a) ACC-based, (b) TPR-based, (C) TNR-based

Model performance should be also investigated for all duration of events, as shown in Figure 8. Results show that the miss rate of prediction is still low for both ACC-based and TPR-based models (less than 3% in Figure 8a). However, this rate suddenly increases to 25% for the TNR-based model. On the other hand, as can be seen in Figure 8b, 13% of total events have overestimation forecasting in the ACC-based model, whereas this rate is reduced to 5.68% and 2.44% for TPR-based and TNR-based, respectively. Finally, while hit rate is quite low for all models (42%, 22% and 11% for TPR, ACC, and TNR models, shown in Figure 8c, respectively), acceptable rate illustrated in Figure 8d shows these models can satisfactorily predict the correct class, meaning overflowing or not overflowing, in correct timestep or slightly earlier. More specifically, the TPR-based model could show a 91% acceptable rate with just an average of 1.13 timestep lag (average of all lagged time between actual timestep and predicted timestep of true predicted class). These promising results can show time series models, particularly the TPR-based model, can be simply but effectively applied for early warning overflow detection systems.

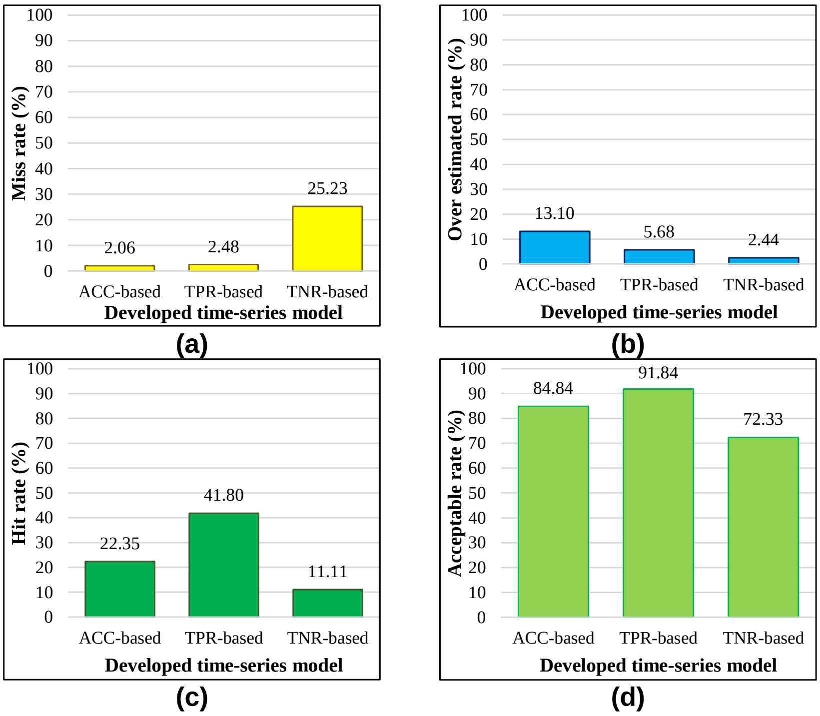


Figure 8. Event-based performance assessment of developed time-series models: (a) Miss rate, (b) Over estimated rate, (c) Hit rate, (d) Acceptable rate

4. CONCLUSIONS

The present study provides a framework for developing WLDMs followed by time-series models based on ACC, TPR and TNR metrics, for early warning overflow detection systems. Analysis of WLDMs, applied for the real case study, shows that none of the selected models could outperform each other for all metrics or for all prediction timesteps. While models indicating better ACC rather than TPR are more capable in detecting non-overflow events than overflow conditions (80-95% vs 60-90% accuracy of TNR and TPR). However, time-series models, especially the TPR-based model, could cover this accuracy by choosing the best WLDM in each time step, which result in reducing overestimation and underestimations for different timesteps of prediction as well as reflecting more than 90% of the acceptable rate. Hence, the application of time-series data mining models can enable the development of a high-speed and real-time overflow classification model that can be trained based on limited features obtained from only rainfall and water level in UDS. However, the applied concept requires further studies such as using advantages of ensemble modelling and involving more input decision variables, especially temperature, soil moisture and wind characteristics.

COMPETING INTERESTS

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.

REFERENCES

[1] Piadeh F., Behzadian K. Alani A. (2022). A Critical Review of Real-Time Modelling of Flood Forecasting in Urban Drainage Systems. Journal of Hydrology, 127476.

[2] United Nations Office for Disaster Risk Reduction (UNDRR). (2022). Annual report for the United Nations Office for Disaster Risk Reduction 2022. [Online] Available at http://www.undrr.org, [Accessed 09 Apr. 2022].

[3] 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. Journal of Hydrology, 588: 125085.

[4] Chen, S., Yu, P., (2007). Pruning of support vector networks on flood forecasting. Journal of Hydrology, 347 (1-2), pp. 67-78.

[5] Damle, C., Yalcin, A. (2007). Flood prediction using Time Series Data Mining. Journal of Hydrology, 333(2–4), pp. 305-316.

[6] Rasheed, Z., Aravamudan, A., Sefidmazgi, A., Anagnostopoulos, G., Nikolopoulos, E. (2022). Advancing flood warning procedures in ungauged basins with machine learning. Journal of Hydrology, 609: 127736.

[7] Zhao, G., Pang, B., Xu, Z., Xu, L. (2020). A hybrid machine learning framework for real-time water level prediction in high sediment load reaches. Journal of Hydrology, 581: 124422.

[8] Noymanee, J., Nikitin, N., Kalyuzhnaya, A. (2017). Urban Pluvial Flood Forecasting using Open Data with Machine Learning Techniques in Pattani Basin. Procedia Computer Science, 119, pp. 288-297.

[9] Wagenaar, D., Curran, A., Balbi, M., Bhardwaj, A., Soden, R., Hartato, E., et al. (2020). Invited perspectives: How machine learning will change flood risk and impact assessment. National Hazards Earth System Science, 20, pp. 1149–1161.

[10] Munawar, H., Hammad, A., Waller, S. (2021). A review on flood management technologies related to image processing and machine learning. Automation in Construction, 132: 103916.

[11] Mosavi, A., Ozturk, P., Chau, K. (2018). Flood Prediction Using Machine Learning Models: Literature Review. Water, 10(11):1536.

[12] 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. Journal of Hydrology, 569, pp. 387-408.

[13] Piadeh F., Behzadian K. Alani A. (2021). The Role of Event Identification in Translating Performance Assessment of Time-Series Urban Flood Forecasting. UWL annual conference, July 2021, London UK.

[14] Department for Environment, Food and Rural Affairs (DEFRA). (2022). DEFRA Official Website [Online] Available at http://Environment.data.gov.uk, [Accessed 12 Jan. 2022].

[15] Masahiko, H., Fabio, C., Upmanu, L. (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.