RO009

Data-Driven Flood Prediction Using Drain Capacity Forecasting: A Singapore Case Study

1. Introduction

Flash floods are among the most destructive natural disasters humans can face, occurring when there is a rapid rise of water levels within 6 hours after heavy rainfall [1], usually in low-lying areas [2]. Intensified precipitation due to increased greenhouse gas concentration in the atmosphere and rising sea levels have contributed to increased flooding risks, underscoring the critical need for resilient flood detection and management systems [3]. Singapore is a heavily urbanised country that is especially at risk of flash floods, due to having a tropical forest climate (Köppen climate classification Af) [4] with abundant precipitation throughout the year (~2110 mm annually) [5] and being a low-lying coastal nation with 30% of the island being only 5m above sea level [6], increasing chances of flooding. The country's small but highly developed land area of 734.3 km² [7] and high population density of 8592 people per km^2 [8] means that severe flooding may result in major damage to both infrastructure and people.

Figure 1. Recent instances of flooding in Singapore **(A)** Flooding along Dunearn Road, 20 July 2023 **(B)** Flooding along Cavenagh Road near Orchard Plaza on 30 March 2023 **(C)** Flooding in the Jurong West area, along Boon Lay avenue, on 28 November 2023

Singapore has employed many statistical flood detection and management techniques, with drainage systems that can manage large volumes of surface runoff, integrated inland-coastal flooding models to simulate floods [9], more than 300 water level sensors to monitor the drainage system [10], and the identification of flooding hotspots [11] to quickly detect and manage floods. However, these models may be complex, involving significant time and financial investment. However, Machine Learning (ML) models are now emerging as a promising method of time-series flood forecasting, modelling the complexities of flooding with lower computation costs [12]. This research aims to test the performance of multiple ML models, including Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), and Decision Tree (DT) models such as Classification and Regression Trees (CART), Random Forest (RF), and Gradient-Boosted Decision Trees/XGBoost (XGB) in predicting floods with various short-term lead times of less than 2 hours, to determine whether flood prediction in Singapore with ML methods is viable.

2. Methodology

2.1 Data and preprocessing

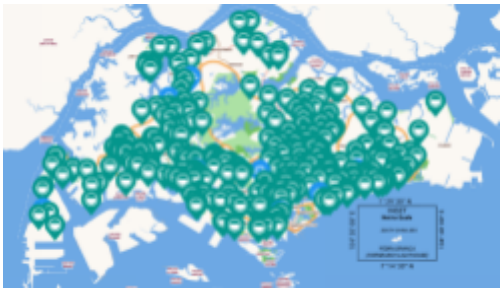

Figure 2. Singapore's network of water level sensors from which drain fill percentage data was derived

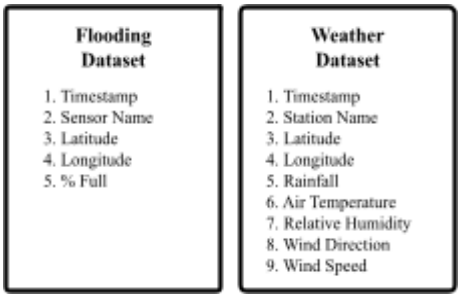

Figure 3. Data contained in the raw flooding and weather datasets.

According to the Public Utilities Board (PUB), the drain fill percentage (percentage of a drain that is filled with stormwater) is an indicator of flooding, with drain fill percentages of above 75% indicating moderate or high flood risk. The flooding dataset utilised in this research was collected over November 2023, derived from the system of water level sensors set up in drains throughout Singapore, and contains data as detailed in Figure 3. This flooding data was supplemented by seasonal weather data from weather stations around Singapore, obtained via the Application Programming Interfaces (APIs) for real-time weather data provided by the National Environment Agency (NEA). To generate the flood dataset, flooding data was filtered to only include that from November 26th to November 30th, with a known flood incident in Jurong on November 28. Each water level sensor was matched with its closest weather station and weather data from the hours prior was incorporated into the flooding dataset. The data was filtered to only include observations where the nearest weather station is within 3.5 km of the weather sensor, to ensure that the weather readings are representative of the location. To generate a model with a lead time of n hours, weather readings were taken n, n+0.5, and n+1.0 hours prior to each water level reading. For weather readings at specific time points, averages of air temperature, humidity, wind direction, and speed were computed over 30-minute intervals and were taken as the weather for that point. Precipitation was summed within these intervals to obtain total precipitation. Training datasets for lead times of 0.5 h, 1 h and 2 h were generated.

2.2 Computation and Evaluation of ML Models

Input features included prior weather data and the geographic coordinates of the water level sensor. The drain fill percentage was set as the target variable. Python's scikit-learn library was used for the computation of the SVM, MLP, CART and RF models. Python's XGBoost library was used for the computation of the XGB model. All input features were scaled with MinMaxScaler. Hyperparameters for each model were tuned with GridSearchCV with a prediction lead time of 1 h. Training data was randomly shuffled and split into 5 folds. 5-fold cross-validation was then done, where parameters for the models were computed by fitting to the training folds, before evaluating against the testing fold. Metrics such as training time, testing time, mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and R^2, also known as the coefficient of determination, were recorded for each evaluation. In the below formulae for the metrics, X_i is the i-th value predicted by the model, Y_i is the actual i-th value, n is the number of observations in the testing dataset, and epsilon ε is some small value added to Y_i to prevent division by 0 for MAPE. For this research, the value was set to $\varepsilon=10^{-10}$. $\bar{Y}$ is the mean of the actual values.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - X_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \qquad MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - X_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)^2} \qquad MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - X_i}{Y_i + \varepsilon}\right|$$

R^2 measures the proportion of the variance in the target variable explained by the independent variables, and should ideally be maximised. RMSE and MAE measure the numerical difference between the model's predicted values and the actual values, with RMSE emphasising larger errors, while MAPE represents the average percentage difference between predicted and actual values. All error values should ideally be minimised. Convergence data of models such as SVM, MLP, and XGB were also tested. While DTs such as CART, RF and XGB do not exhibit convergence where loss drops below an error range, evaluating against testing data over many boosting iterations provides insight into XGB convergence behaviour.

3. Results and Discussion

3.1 General Results

Tuned parameters used for each model can be found in the Appendix, Supplementary Table 1. All other hyperparameters were left as default values. All computations were done on a PC with Intel

Core i7-1255U, 16.0-GB memory. Codes can be found at this link: github.com/Niyaz-Mohamed/Overflow.sg.

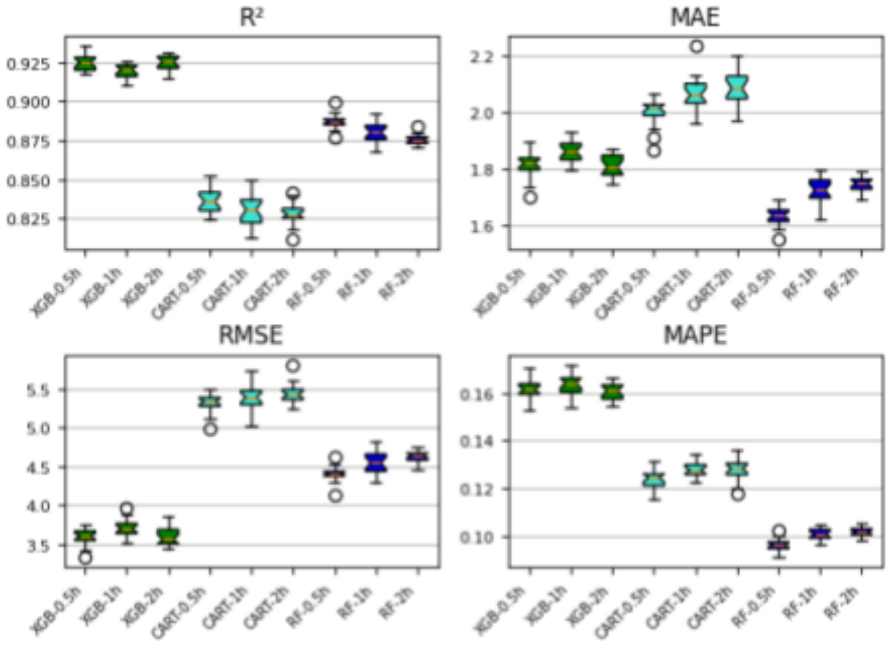

Figure 4. Boxplots of evaluation metrics (R^2, MAE, RMSE, and MAPE) obtained from multiple runs of model computation, changing the prediction lead time.

A single 5-fold cross-validation run with GridSearchCV yielded mean values for R^2, MAE, RMSE, MAPE, training time, and testing time, detailed in the Appendix, Supplementary Table 2. Generally, DT algorithms such as XGB, CART and RF performed very well, with significantly higher R^2 scores and lower errors. Unexpectedly, SVM and MLP performed very badly despite other papers reporting good performance from these models [13], recording low R^2 values of around ~0 and ~0.3 respectively and significantly higher error values than the other models, indicating inferior performance. This might be a consequence of using drain fill percentage, which may be affected by the depth of the drain, as the target variable rather than flow rate above ground, which is a direct indicator of flooding. Noise is generated in the dataset due to each water sensor having a unique relationship between its readings and the weather data, reducing SVM and MLP accuracy. However, decision trees account for differences in the patterns of each water sensor, resulting in greater accuracy. Due to their unsatisfactory performance, SVM and MLP were excluded from further analysis. The remaining DT models were further tested, each fitted 20 times with 5-fold cross-validation repeated 4 times on the training dataset. Evaluation results were collected for each fold and a boxplot was drawn.

In Figure 5, XGB exhibited the highest R^2 values and lowest RMSE values for all lead times, while RF had the lowest MAE and MAPE values across all lead times. XGB is hence the model which best fits the data and explains most of the variability in the target variable with the input variables, and is also very robust against large deviations in predicted values from the actual values as indicated by its low RMSE. RMSE was considered to be more important than MAE in this context as large errors in predicted values are very undesirable and may affect predicted flood risk levels. MAPE values were not considered to be as important as the other metrics. The large number of small or near-zero values of the target variable make MAPE an unreliable indicator of error for this use case. Meanwhile, CART is inferior to both XGB and RF across all evaluation metrics, which was expected as ensemble models such as XGB and RF generally report better performance for flood prediction [14]. Models also displayed slightly increased predictive ability with shorter lead times, with higher R^2 values and lower errors in the 0.5 h lead time models, which is to be expected as flash floods become easier to predict closer to the time of the event [15].

3.2 Comparing sensor-station distances

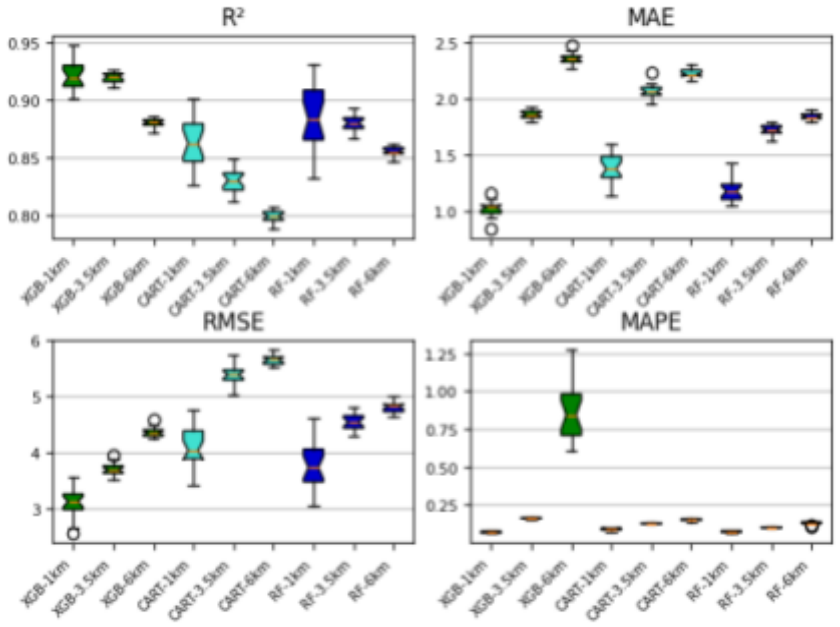

Figure 6. Boxplots of evaluation metrics (R^2, MAE, RMSE, and MAPE) obtained from multiple runs of model computation, changing the maximum sensor-station distance allowed.

20 values were collected for each evaluation metric by limiting the maximum distance between water sensors and weather stations in the dataset. When limiting the sensor-station distance to 6km, ε for MAPE calculation was adjusted to $\varepsilon=10^{-2}$ to account for numerous near-zero values of the target variable. Generally, with decreasing sensor-station distance, there was an enhancement in

model performance. Due to diminishing sensor-station distance, the weather data linked to each water level observation becomes more representative of the specific weather conditions at the corresponding water level sensor. However, models associated with shorter sensor-station distances demonstrated lower consistency, evident from the larger size of their boxplots. This inconsistency may be attributed to their relatively smaller dataset size, which significantly increases the degree to which the shuffling and splitting of the dataset before training impacts the subsequent performance of the models.

Table 1. Number of observations in the dataset when maximum sensor-station distance is limited.

Maximum Sensor-Station Distance/ km	**1.0**	**3.5**	**6.0**
Number of observations	4052	45,566	67,066

Moreover, with shorter sensor-station distances, the model can only predict floods at the few water level sensors close to weather stations. The limited number of positive flooding observations in the smaller training datasets makes it such that while the model may excel in predicting lower drain fill percentage values, it may face challenges in forecasting higher drain fill percentage values, affecting its capability to predict floods effectively and reducing its usefulness.

3.3 Convergence of XGB

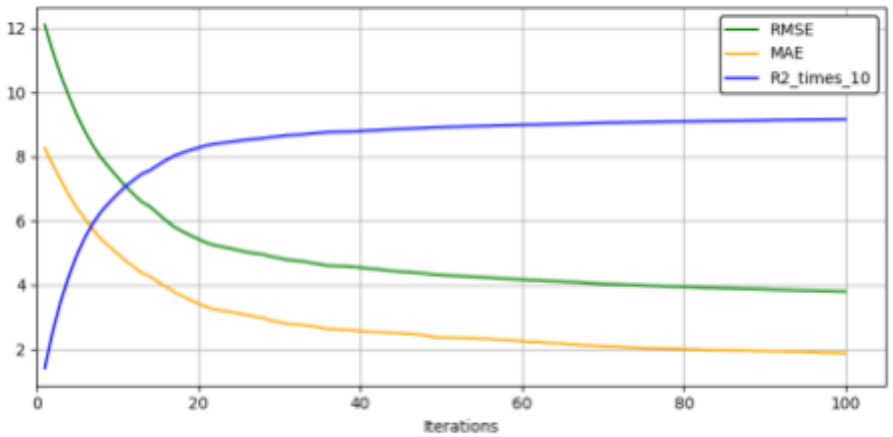

Figure 7. Variation of RMSE, MAE, and R2 over 100 boosting iterations of XGB model training

The dataset was split in a 4:1 train-test ratio and the XGB model was trained on the training data, evaluating against the testing data after every boosting iteration to obtain data on the convergence of R^2, MAE, and RMSE. The model converges smoothly to a satisfactory extent, with much slower changes in the metrics, at about 80 iterations, but continues slowly improving in further iterations.

Increasing the number of boosting iterations much beyond 100 may lead to overfitting to the training dataset, reducing the ability of the model to generalise to unseen data. A high improvement rate at first may also indicate that patterns in the training dataset were quickly identified by the XGB model, indicating that it is a suitable model for flood prediction with drain fill percentages.

3.4 General Discussion

Overall, XGB was the most performant model with a high R^2 value, low RMSE, and short training times, as recorded in the Appendix, Supplementary Table 2. Further analysis on training times, testing times, as well as performance when changing lead time or station-sensor distance, can be accessed at this link: bit.ly/flood-pred-model-data. While RF recorded the lowest MAPE and MAE values, its training time is a concern. While the prediction time for all DT models was low, training times were significantly higher for RF, which usually needed more than 200 seconds for training. In contrast, the other DT models completed training within a few seconds. This may result in scaling issues and extremely long training times if significantly larger amounts of training data are introduced in an attempt to create a more performant model, making RF less practical and leading to XGB ultimately being chosen as the most performant model overall. However, this research was severely limited due to the lack of access to flow data, which is a more direct indicator of flooding than drain fill percentage. Moreover, other relevant data which might have further improved the models, such as soil infiltration rate, were also not collected and used for training.

4. Conclusion

Overall, the aim of assessing the feasibility of ML models as a method of flood prediction in Singapore was achieved. Of the models tested, XGB was the most performant with the highest accuracy and high training speeds. Decision-tree-based models like XGB, CART, and RF were much more performant than SVM and MLP, which is likely due to noise in the training dataset disrupting SVM and MLP performance. DT models in general were the most promising in flood prediction. Possible areas for future studies include 1) inclusion of certain topographical and hydrological data like soil infiltration rate and soil moisture as input features for the model (since existing models from other countries already include this data to ensure higher prediction accuracy), 2) further research into XGB due to its promising performance, 3) integration of ML prediction with existing statistical models and flood prediction methods so that ML can be used on the ground as fast as possible.

References

1. Knocke, E. T., & Kolivras, K. N. (2007). Flash Flood Awareness in southwest virginia. *Risk Analysis*, *27*(1), 155–169. https://doi.org/10.1111/j.1539-6924.2006.00866.x

2. Lai, O. (2023, October 2). *What are the main causes and effects of floods around the world?*. Earth.Org. https://earth.org/what-are-the-main-causes-and-effects-of-floods/

3. Perera, D., Seidou, O., Agnihotri, J., Rasmy, M., Smakhtin, V., Coulibaly, P., & Mehmood, H. (2019). Flood early warning systems: A review of benefits, challenges and prospects. *UNU-INWEH Report Series*, (8), 7–11. https://doi.org/10.53328/mjfq3791

4. Winston T. L. Chow, Brendan D. Cheong, Beatrice H. Ho, "A Multimethod Approach towards Assessing Urban Flood Patterns and Its Associated Vulnerabilities in Singapore", Advances in Meteorology, vol. 2016, Article ID 7159132, 11 pages, 2016. https://doi.org/10.1155/2016/7159132

5. Meteorological Service Singapore (MSS). (n.d.). Climate of Singapore. http://www.weather.gov.sg/climate-climate-of-singapore/

6. National Climate Change Secretariat (NCCS). (2012). (publication). *Climate Change in Singapore: Challenges. Opportunities. Partnerships* (pp. 81–81). Singapore, Singapore. https://www.nccs.gov.sg/files/docs/default-source/default-document-library/national-climate-change-strategy.pdf

7. *Population and Vital Statistics*. Ministry of Health (MOH). (2023). https://www.moh.gov.sg/resources-statistics/singapore-health-facts/population-and-vital-statistics

8. Department of Statistics (DOS). (2023). *Population*. Department of Statistics (DOS). https://www.singstat.gov.sg/publications/reference/ebook/population/population

9. National University of Singapore (NUS) College of Design and Engineering (CDE). (n.d.). *Integrated Coastal-Inland Flood Model for Climate Change*. Cde.nus.edu.sg.

https://cde.nus.edu.sg/cee/research/research-clusters/ccm/integrated-coastal-inland-flood-model-for-climate-change/

10. Public Utilities Board (PUB). (n.d.). *Water Level Sensors & CCTVs*. PUB Singapore's National Water Agency. https://app.pub.gov.sg/waterlevel/pages/WaterLevelSensors.aspx

11. Poon, Y. X. (2020, May 14). *Exclusive: How Singapore is predicting floods*. GovInsider. https://govinsider.asia/intl-en/article/pub-yeo-keng-soon-hazel-khoo-exclusive-how-singapore-is-predicting-floods

12. Mosavi A, Ozturk P, Chau K-w. Flood Prediction Using Machine Learning Models: Literature Review. *Water*. 2018; 10(11):1536. https://doi.org/10.3390/w10111536

13. Maspo, N.-A., Bin Harun, A. N., Goto, M., Cheros, F., Haron, N. A., & Mohd Nawi, M. N. (2020). Evaluation of machine learning approach in flood prediction scenarios and its input parameters: A systematic review. *IOP Conference Series: Earth and Environmental Science*, *479*(1). https://doi.org/10.1088/1755-1315/479/1/012038

14. Seydi, S. T., Kanani-Sadat, Y., Hasanlou, M., Sahraei, R., Chanussot, J., & Amani, M. (2022). Comparison of machine learning algorithms for flood susceptibility mapping. *Remote Sensing*, *15*(1). https://doi.org/10.3390/rs15010192

15. Song, T., Ding, W., Wu, J., Liu, H., Zhou, H., & Chu, J. (2019). Flash flood forecasting based on long short-term memory networks. *Water*, *12*(1), 109. https://doi.org/10.3390/w12010109

16. Straits Times. (2023). *Flash Flood on Dunearn Road*. 4 hours of rain in western S'pore on Thursday nearly equals July monthly average. Retrieved December 4, 2023, from https://www.straitstimes.com/singapore/environment/4-hours-of-rain-in-western-s-pore-on-thursday-nearly-equals-july-monthly-average.

17. Straits Times. (2023). *Flooding along Boon Lay Avenue*. Heavy afternoon downpour causes flash floods in Jurong West. Retrieved December 4, 2023, from https://www.straitstimes.com/singapore/flash-floods-in-jurong-west-amid-heavy-afternoon-downpour.

18. Straits Times. (2023). *Flooding along Cavenagh Road*. Burst of heavy rain leads to flash flood warnings in several parts of Singapore. Retrieved December 4, 2023, from https://www.straitstimes.com/singapore/burst-of-heavy-rain-leads-to-flash-floods-in-several-parts-of-s-pore-including-orchard-road-area-and-bukit-timah.

Appendix

Supplementary Table 1. Tuned hyperparameters for each model

Model	Hyperparameter	Tuned value
SVM	C	0.9
MLP	activation	“tanh”
	alpha	0.001
	learning_rate _init	0.01
XGB	n_estimators	100
	max_depth	8
	learning_rate	0.1
	subsample	0.7
	gamma	2
	scale_pos_weight	0
	alpha	10
	lambda	0
CART	min_samples_split	200
	min_samples_leaf	60
RF	n_estimators	200
	min_samples_split	5
	min_samples_leaf	20

Supplementary Table 2. Mean values of evaluation results obtained from a single run of model computation across each fold of 5-fold cross-validation, rounded to 2 decimal places

Model	Prediction Lead Time / h	R^2	MAE	RMSE	MAPE	Training Time / s	Testing Time / s
SVM	0.5	-0.022	7.736	13.241	0.766	142.549	53.529
	1	-0.026	7.747	13.248	0.766	132.328	57.317
	2	-0.022	7.749	13.243	0.765	145.037	54.828
MLP	0.5	0.28	7.739	11.117	0.974	67.893	0.024
	1	0.302	7.581	10.924	0.941	71.039	0.023
	2	0.299	7.663	10.964	0.975	64.131	0.025
XGB	0.5	0.925	1.796	3.583	0.158	4.186	0.085
	1	0.919	1.872	3.713	0.165	3.532	0.052
	2	0.916	1.903	3.784	0.166	3.656	0.051
CART	0.5	0.837	1.992	5.289	0.124	1.232	0.005
	1	0.83	2.069	5.399	0.128	1.256	0.005
	2	0.829	2.076	5.409	0.128	1.248	0.006
RF	0.5	0.886	1.639	4.427	0.097	344.15	0.327
	1	0.879	1.729	4.543	0.101	286.434	0.425
	2	0.875	1.749	4.62	0.102	311.465	0.337