Data-driven approaches for urban stormwater studies – lessons learnt from four application cases

K. Zhang[1,*], David McCarthy[2], Anna Lintern[2] & Ana Deletic[3]

[1]*Water Research Centre (WRC), School of Civil and Environmental Engineering, UNSW Sydney, High Street, Kensington, NSW 2052, Australia*
[2]*Environmental and Public Health Microbiology Laboratory (EPHM Lab), Department of Civil Engineering, Monash University, Wellington Rd, Clayton, Victoria, 3810, Australia*
[3] *School of Civil and Environmental Engineering, Queensland University of Technology, Queensland 4001, Australia*
**Corresponding author email: kefeng.zhang@unsw.edu.au*

Highlights

- Various data-driven modelling approaches applied in different stormwater studies.
- Promising performance was observed when using data-driven methods.
- Lessons learnt presented in the aspects of data-preparation and model testing.

Introduction

Effective urban stormwater management requires a good understanding of various contaminants levels from different catchments, as well as their removal processes in nature-based systems (NBS). Previously, extensive monitoring of stormwater quality (Francey *et al.*, 2010), as well as the treatment performance of various NBS have been conducted. The monitoring is, however, labour intensive and costly, and therefore, mathematical models based on physical processes are regarded as a useful approach that can assist in estimating stormwater quality, as well as designing NBS to achieve removal targets.

There are, however, many limitations of the physical-based models. For example, the processes with regards to stormwater build-up and wash-off processes are highly complex and impacted by various factors (e.g., rainfall, catchment) as well as unexpected sources (Zhang *et al.*, 2019). These are very hard to be sufficiently represented in a physically based model, and therefore their reliability is often questioned at some cases (e.g., at urban catchment scale, Bonhomme and Petrucci, 2017). The natural processes occurring in the NBS are also complicated, which could vary depending on different design and operational conditions. Another limitation of physical-based models is the demand for the high-quality data to be used for simulations, that are not easily available, as well difficulties in the model calibration. Recently, data-driven approaches, or machine learning models, have been increasingly used, in various aspects of stormwater studies (Fang *et al.*, 2021). In contract to the physical based models, the data-driven models are not reliant on the complex physical processes but learn from available data to understand the system. They are able to establish strong relationships between a high dimension of input variables (e.g., various design and operational conditions of NBS) and target output variable (e.g., effluent concentrations).

This paper presents four application cases where we applied different data-driven approaches in various aspects of urban water studies, including stormwater quality, stormwater biofilter performance modelling (and risk assessment), as well as detection of pathogen in urban waters. Through these applications, the usefulness of the data-driven methods to inform physical processes, the key input variables, as well as its further application, is demonstrated. Lessons learnt through these cases are also presented.

Methodology

Case 1: Genetic Programming to generate new pollution build-up algorithm

Water quality (and flow) monitoring data for total suspend solids (TSS), total nitrogen (TN) and total phosphorus (TP) across two catchments in Melbourne Australian were collected for 18-19 stormwater events. The relevant climate variability (e.g., ADWP, max and min Temperature, previous rainfall event, solar exposure) were also collected. GP was used to generate algorithms that best describe the relationships between C_0 (as the starting concentration for each event that reflects the pollution build-up process), and all rainfall and climate variables after initial correlation analysis. Two best generated build-up

algorithms were connected with typical first order wash-off models and then compared against traditional build-up and wash-off models. For full details of the methods please refer to Zhang *et al.* (2019). This case allowed us to learn whether further improvement of the build-up algorithm could be informed by using GP.

Table 1. Summary of the four cases employing different data-driven approaches for various purposes.

Application		Aim	Data source	Specific data-driven methods
1	Stormwater quality	Inform new build-up algorithm	Monitoring data	Genetic programming
2	Stormwater biofilter	Model system performance; identify key factors	Monitoring data	Various statistical models
3	Stormwater biofilter	Model system performance; identify key factors; risk assessment	Literature data	Multilinear Regression (MLR), Neural Network (NN) and Random Forest (RF)
4	Urban waterways	Early prediction of Campylobacter using spectrophotometer data	Monitoring data	Supported Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF)

Case 2: Modelling treatment performance of stormwater biofilters using statistical models
In this Case, stormwater biofilter treatment performance data (TP and TN), including key design and operational variables, were collected from four large laboratory column experiments. The data were fitted into three innovative statistical models of different structures with varying number of design / operational variables. The best models with the most important design and operational variables were identified. Full details of the models and testing methods could be found in Zhang *et al.* (2021). This case allowed us not only to learn the feasibility of using data-driven models in predicting biofilter performance, but also to investigate the key design and operational variables.

Case 3: **Prediction of heavy metal removal and risk mitigation by stormwater biofilters using ML**
Literature data were gathered on the performance of heavy metal removal in biofilters, alongside the essential design and operational parameters of stormwater biofilters (same as case 2). Different from Case 2, this case used data compiled from open literature, and three common machine learning (ML) models, namely multilinear regression (MLR), neural network (NN) and random forest (RF). In addition, different feature selection methods of the ML algorithms were used in this study to investigate the most important design or operational variables. Furthermore, the best models were used to predict biofilter outflow concentrations, which were then assessed for its risks to human health. Full methods are available in Fang *et al.* (2021). This case further allowed us to test different feature selection approaches by the ML models and demonstrate the how the validated ML models can be applied in risk assessment.

Case 4: Application of machine learning models in early prediction of Campylobacter in urban waters
This case aimed to investigate the potential of using spectrophotometry in predicting the presence of Campylobacter in water samples collected from various waterways of Melbourne, Australia. Three ML models, *i.e.,* supported vector machine (SVM), logistic regression (LR) and random forest (RF) were used to and predict the presence of campylobacter presence in samples by using spectrophotometry data.

Results and discussion

The GP suggested that specific build-up algorithms shall be used for different pollutants, instead of having a unique one that only considers ADWP as the key predictor of the pollution build-up. For example, the best performing build-up models informed by GP for TSS, TP and TN are always different, and inclusion of temperature as a predictor was able to improve the prediction of TP and TN. The new GP informed build-up and wash-off models were found to have better performance (e.g., Nash-Sutcliffe efficiency, NSE > 0.45 in both calibration and validation for TP in one catchment) comparing to traditional models (NSE mostly < 0).

The best statistical models for predicting TP outflow concentrations (C_{out}) from biofilter were found to include infiltration rate (IR) and inflow concentration (C_{in}) as operational variables, and filter media (FM) and filter media depth (FMD) as design variables ($NSE_{calibration}$ =0.54) and $NSE_{validation}$=0.55). As for TN, the best models included vegetation (Veg) and presence of submerged zone (SZ) as the key design variables, and IR and C_{in} as the key operational variables. The key design variables found through the statistical

models have very good agreement with literature, e.g., filter media was found to be important for TP (Hatt *et al.*, 2009), while vegetation and submerged zone presence for TN (Payne *et al.*, 2014). The statistical modelling approach reveals that the operational variables (*i.e.,* IR and C_{in}), which are relatively less studied, also have an important impact on biofilter performance. This indicates that by controlling operational conditions, the biofilter performance could be improved – in fact, the results from Case 3 suggested that by controlling IR, the risks caused by heavy metals can be maintained at low-risk.

Using the spectrophotometer data, the three ML models were able to predict the presence of campylobacter at varying accuracies (ACC), at 0.763 on average, and the false negative rate (FNR) was 13.8% across three methods (Figure 1). Although the accuracies could be improved, the models show promising ability in provisioning an early indicator to the potential risk of *Campylobacter* in water, which would then require secondary confirmations.

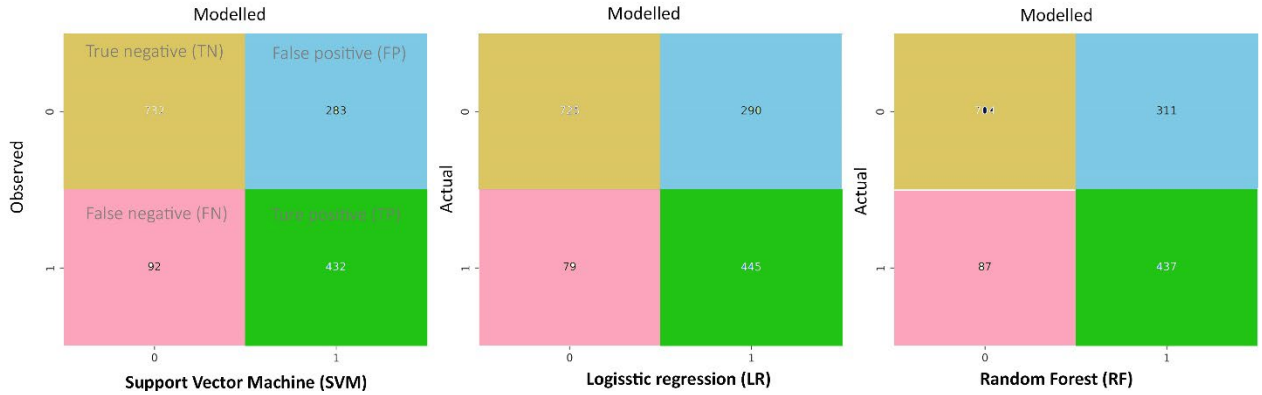

Figure 1. Confusion matrix of the model testing results, based on which ACC = (TN+TP)/(TN+TP+FN+FP), and FNR=FN/(FN+TP)

Conclusions and future work

Four different cases of applying data-driven models are presented in this paper, to showcase how they can be applied in various urban water studies. Promising results were shown in in terms of the modelling performance, and additionally the approaches were able to help identifying key factors influencing the processes (e.g., build-up, pollutant removal). The reliability of data-driven models relies on the source of data, and thus the lessons we learnt including ensuring the quality of the data used and conducting of robust testings (*e.g.,* N-fold calibration and validation). The future applications of data-driven models can be in the area of real time control technologies where the models can be used for real time predictions.

Acknowledgement

The research was funded (partially) by the Research Council Discovery Early Career Award – ARC DECRA (project number DE210101155). Dr Zhang is the recipient of the ARC DECRA.

References

Bonhomme, C. and Petrucci, G., 2017. Should we trust build-up/wash-off water quality models at the scale of urban catchments? Water Research 108, 422-431.

Fang, H., Jamali, B., Deletic, A. and Zhang, K., 2021. Machine learning approaches for predicting the performance of stormwater biofilters in heavy metal removal and risk mitigation. Water Research, 117273.

Francey, M., Fletcher, T.D., Deletic, A. and Duncan, H., 2010. New Insights into the Quality of Urban Storm Water in South Eastern Australia. Journal of Environmental Engineering 136(4), 381-390.

Hatt, B.E., Fletcher, T.D. and Deletic, A., 2009. Pollutant removal performance of field-scale stormwater biofiltration systems. Water science and technology 59(8), 1567-1576.

Payne, E.G., Pham, T., Cook, P.L., Fletcher, T.D., Hatt, B.E. and Deletic, A., 2014. Biofilter design for effective nitrogen removal from stormwater–influence of plant species, inflow hydrology and use of a saturated zone. Water science and technology 69(6), 1312-1319.

Zhang, K., Deletic, A., Bach, P.M., Shi, B., Hathaway, J.M. and McCarthy, D.T., 2019. Testing of new stormwater pollution build-up algorithms informed by a genetic programming approach. Journal of Environmental Management 241, 12-21.

Zhang, K., Liu, Y., Deletic, A., McCarthy, D.T., Hatt, B.E., Payne, E.G.I., Chandrasena, G., Li, Y., Pham, T., Jamali, B., Daly, E., Fletcher, T.D. and Lintern, A., 2021. The impact of stormwater biofilter design and operational variables on nutrient removal - a statistical modelling approach. Water Research 188, 116486.