Predicting Event Mean Concentrations (EMCs) of Nutrients and Sediments in Urban Runoff Using A Random Forest Approach

M. Shahed Behrouz, Ph.D. candidate[1,2], M. Nayeb Yazdi, Ph.D.[3], D. Sample, Ph.D.[1,2*]

[1]*Departmtnet of biological Systems Engineering, Virginia Tech, Blacksburg, Virginia*

[2]*Hampton Roads Agricultural Research and Extension Center (HR AREC), Virginia Beach, Virginia*

[3]*School of Environment and Natural Resources*, Ohio State University, Wooster, *Ohio*

**Corresponding author email: dsample@vt.edu*

Highlights

- Random Forest (RF) was used to predict water quality (WQ) event mean concentrations (EMCs)
- Land use (*LU*) and antecedent dry period (*ADP*) are not the only factors affecting WQ EMCs.
- Precipitation and site characteristics should also be considered for estimating WQ EMCs.

Introduction

Excessive nutrients and sediments in urban runoff cause degraded water quality downstream (Lintern et al., 2020). Event mean concentration (EMC) is a common method used by practitioners to estimate pollutant loads in urban runoff (Tuomela et al., 2019). Land use and antecedent dry period (*ADP*) are known factors affecting water quality (WQ) EMCs (Rossman, 2015). This study aims to investigate the effect of other precipitation (i.e. depth, *P*, and duration, *D*) and site characteristics (i.e. slope, *S*, and area, *A*) on WQ EMCs using a U.S.A dataset, the National Stormwater Quality Database (NSQD) (Figure 1).

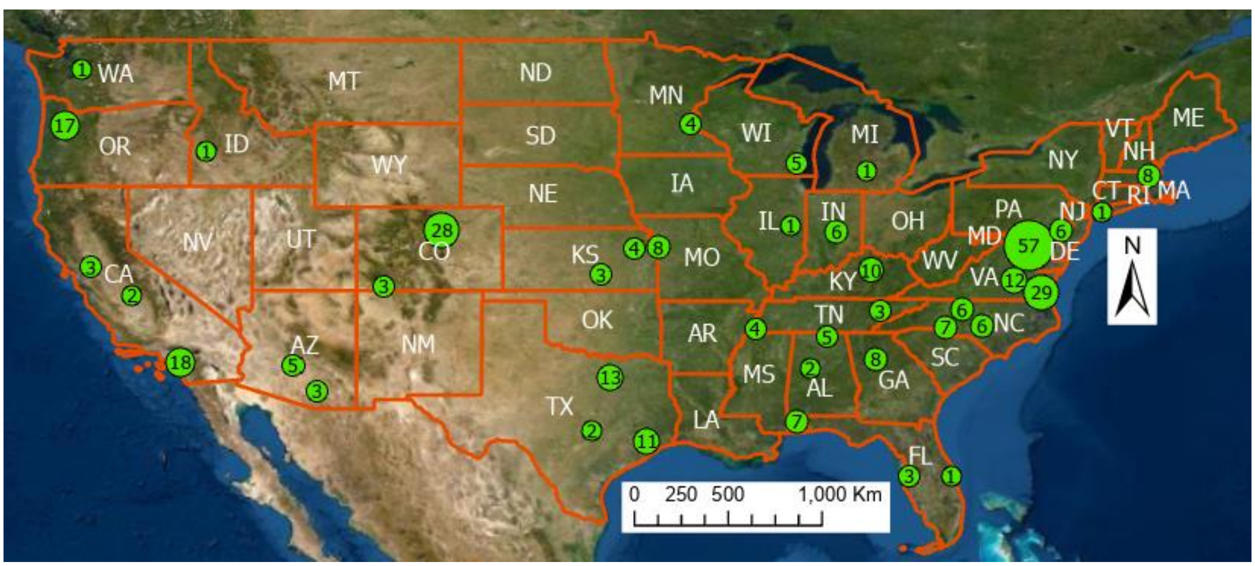

Figure 1. The number of monitoring locations with homogenous catchments (with respect to land use) in the National Stormwater Quality Database (NSQD).

Methodology

Dataset

The NSQD stores monitoring results from over 5000 storm events, collected from approximately 300 urban catchments with homogenous land uses including: Commercial, Freeways, Industrial, Institutional, Open Space, and Residential. (Figure 1). Common WQ constituents included in NSQD are total nitrogen (TN), total phosphorous (TP), and total suspended solids (TSS) (Pitt et al., 2018).

Random Forest (RF)

Random Forest (RF) is a supervised machine learning approach based on ensemble learning. RF trains and constructs several decision trees in parallel with bootstrapped aggregation (Fang et al., 2021). In this study, RF was applied to over 5000 storm events related to aforementioned land uses across the U.S.A. using a free open-source package in R, 'randomForest' (RColorBrewer and Liaw, 2018). The dataset was split randomly into training (80%) and testing periods (20%). The parameters of the RF models were held constant during the training and testing periods.

Scenarios and model evaluation criteria

Nine parameters were selected to represent precipitation and site characteristics including *P*, *D*, *ADP*, intensity (*I*), imperviousness (*Imp*), saturated hydraulic conductivity (K_{sat}), available water capacity (*AWC*), (*S*), and (*A*). Five scenarios: M1-M5, each considering different combinations of nine selected parameters, were imported in RF models (Table 1). Each RF model was iterated 100 times, by randomly selecting 80% of data as training and remaining as testing, to better evaluate the performance of the RF models.

Table 1. Scenario names and selected parameters in each scenario.

Scenario	Selected Parameters
M1	$ADP + LU$
M2	$ADP + LU + P + D + I$
M3	$ADP + LU + P + D + I + Imp$
M4	$ADP + LU + P + D + I + Imp + K_{sat} + AWC$
M5	$ADP + LU + P + D + I + Imp + K_{sat} + AWC + S + A$

Nash-Sutcliffe efficiency (NSE) was selected to evaluate the performance of the RF models. NSE ranges from -∞ to 1, with 1 being the optimal value, and positive NSE being the acceptable performance of the RF models (Moriasi et al., 2015).

Results and discussion

Model performance

The performance of the RF Models (NSE values) was approximately the same across all WQ EMCs: TN, TP, and TSS (Figure 2. a, b, c). In training period, median NSE values with M2-M5 as inputs, were above 0.7 for TN, TP, and TSS indicating excellent model fits (Figure 2.1). In testing period, median NSE values with M5 as input were 0.25, 0.25 and 0.2 for TN, TP, and TSS, respectively (Figure 2.2). All RF models have higher variability of testing NSE values (wider interquartile range) across 100 iterations relative to training NSE values (narrower interquartile range). This may be attributed to the large variances in the NSQD dataset since data have been gathered from 300 case studies conducted under different conditions, i.e. different anthropogenic activities (Fang et al., 2021).

Key variables on WQ EMCs

The findings in this study indicate considering *P, D,* and *I* (from M1 to M2) heavily improved median training NSE values for all WQ EMCs (median NSE improved from 0.05-0.21 to 0.65-0.72, Figure 2.1). For the testing periods, adding more parameters to the RF models (from M1 to M5) gradually improved testing NSE values for all WQ EMCs (median NSE improved from -0.05-0.01 to 0.2-0.25, Figure 2.2). This important finding indicated that parameters other than *LU* and *ADP* should be considered when estimating WQ EMCs.

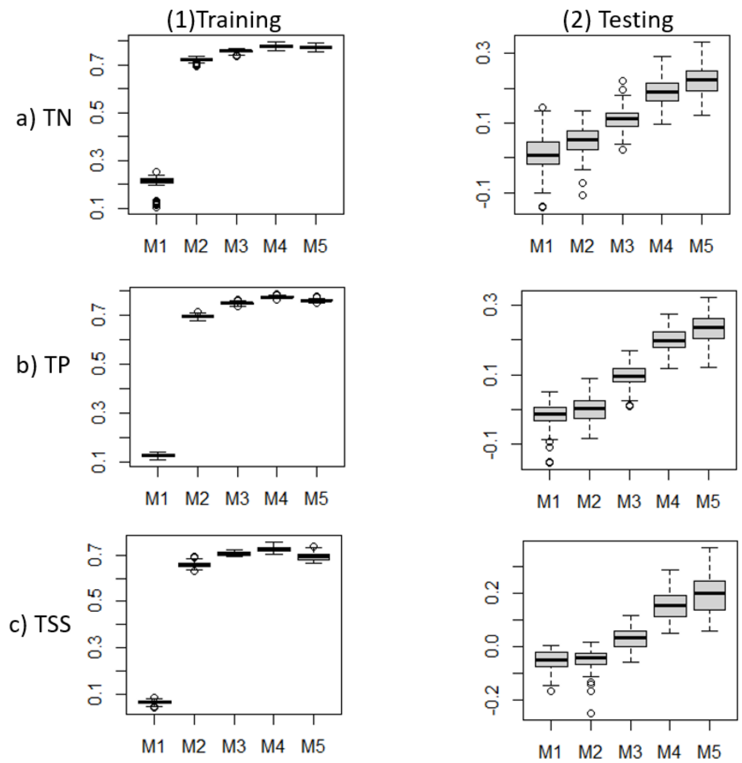

Figure 2. Boxplots of NSE values for (1) training and (2) testing generated by RF models for a) TN, b) TP, and c) TSS.

Conclusions and future work

Future Work:

RF models with different scenarios, M1-M5, will be applied on each land use type to assess estimating WQ EMCs. Sensitivity and uncertainty analysis will be performed to quantify the effect of each variable on estimated WQ EMCs and uncertainty in RF model predictions stemmed from various variables.

Lessons learned from this study:

- RF is a powerful approach for estimating WQ EMCs from homogenous land use catchments.
- Various precipitation and site characteristics should be considered for estimating WQ EMCs.

References

Fang, H., Jamali, B., Deletic, A., Zhang, K., 2021. Machine learning approaches for predicting the performance of stormwater biofilters in heavy metal removal and risk mitigation. Water Res. 200, 117273.

Lintern, A., McPhillips, L., Winfrey, B., Duncan, J., Grady, C., 2020. Best Management Practices for Diffuse Nutrient Pollution: Wicked Problems Across Urban and Agricultural Watersheds. Environ. Sci. Technol. 54, 9159–9174.

Moriasi, D.N., Gitau, M.W., Pai, N., Daggupati, P., 2015. Hydrologic and water quality models: Performance measures and evaluation criteria. Trans. ASABE 58, 1763–1785.

Pitt, R., Maestre, A., Clary, J., 2018. The national stormwater quality database (NSQD, version 4.02) 1–65.

RColorBrewer, S., Liaw, M.A., 2018. Package ‘randomForest’. URL https://www.stat.berkeley.edu/~breiman/RandomForests/

Rossman, L.A., 2015. Storm Water Management Model User’s Manual Version 5.1. United States Environ. Prot. Agency 353.

Tuomela, C., Sillanpää, N., Koivusalo, H., 2019. Assessment of stormwater pollutant loads and source area contributions with storm water management model (SWMM). J. Environ. Manage. 233, 719–727.