Article

Prediction of Nitrate and Phosphorus concentrations using machine learning algorithms in watersheds with different landuse

Aayush Bhattarai [1], Sandeep Dhakal [1], Yogesh Gautam [1] and Rabin Bhattarai [2,*]

[1]Department of Mechanical and Aerospace Engineering, Institute of Engineering, Pulchowk Campus, Kathmandu 44700, Nepal

[2]Department of Agricultural and Biological Engineering, University of Illinois at Urbana Champaign, 1304W Pennsylvania Ave #338, Urbana, IL, 61801, USA

[*]Correspondence: rbhatta2@illinois.edu; Tel.: +1 217-300-0001

Supporting Material

Table S1. Characteristics of studied watersheds upstream of the USGS gaging station (streamflow, total suspended solids, nitrate concentration, and total phosphorus concentration data provided by NCWQR at Heidelberg University, Ohio).

Watersheds	USGS station #	Monitoring period	Watershed size (km^2)	Data frequency	Land-use (%)					Anthropogenic impact levels
					Agriculture	Pasteur	Forest	Urban	Other	
Cuyahoga	4208000	1982-2020	1,830	Daily	9.01	11.84	33.55	39.54	6.06	High
Grand	4212100	1988-2006	1,774	Daily	40.00	*	50.10	0.90	13.10	Low
Maumee	4193500	1976-2021	16,395	Daily	73.33	6.34	6.47	10.61	3.24	Medium
Raisin	4176500	1982-2020	2,698	Daily	49.56	18.69	11.01	10.75	9.99	Medium
Sandusky	4198000	1976-2021	3,239	Daily	77.59	4.32	8.82	8.10	1.18	Medium

* Pasteur is not separate from agriculture

Table S2. Descriptive statistics of parameters for all watersheds.

Watershed	Parameter	Unit	Minimum	Median	Mean	Maximum	Standard Deviation
Cuyahoga	Flow	ft^3/s	0	650.783	1031.009	13884.000	1134.092
	Total Suspended Solids	mg/L	0	31.900	102.627	10090.000	259.145
	Total Phosphorus	mg/L	0.024	0.165	0.217	4.175	0.179
	Nitrate	mg/L	0	2.220	2.456	8.540	1.215
Grand River	Flow	ft^3/s	2.520	445.900	1045.197	14220.000	1556.223
	Total Suspended Solids	mg/L	0	19.200	64.997	3103.333	160.962
	Total Phosphorus	mg/L	0.006	0.058	0.083	1.420	0.086
	Nitrate	mg/L	0	0.387	0.449	3.700	0.346

Maumee	Flow	ft^3/s	22.000	2564.000	6366.227	110565.000	9625.053
	Total Suspended Solids	mg/L	0.200	44.566	75.407	2323.500	103.443
	Total Phosphorus	mg/L	0.030	0.192	0.236	2.808	0.153
	Nitrate	mg/L	0	4.070	4.192	26.750	3.062
Raisin	Flow	ft^3/s	44.000	480.780	892.173	22570.500	1168.412
	Total Suspended Solids	mg/L	0	24.193	45.686	1918.900	84.278
	Total Phosphorus	mg/L	0.007	0.098	0.127	1.827	0.109
	Nitrate	mg/L	0	2.420	2.896	19.460	2.232
Sandusky	Flow	ft^3/s	0	469.040	1504.826	32153.333	2752.192
	Total Suspended Solids	mg/L	0	32.168	71.809	1910.000	113.688
	Total Phosphorus	mg/L	0.003	0.133	0.196	2.440	0.183
	Nitrate	mg/L	0	3.640	3.975	22.933	3.193

Table S3. Parameter settings for each ML algorithm to predict Nitrate and Phosphorus concentration

Algorithm	Parameter	Value for Nitrate prediction	Value for Phosphorus prediction
F-SVM	Kernel function	Gaussian	Gaussian
	Kernel scale	0.35	0.43
	Box constraint	Automatic	Automatic
	Epsilon	Automatic	Automatic
	Standardize data	TRUE	TRUE
M-SVM	Kernel function	Gaussian	Gaussian
	Kernel scale	1.4	1.7
	Box constraint	Automatic	Automatic
	Epsilon	Automatic	Automatic
	Standardize data	TRUE	TRUE
kNN	k	5	5
	Distance function	Euclidean	Euclidean
RF	Number of trees	10	10
ANN	Neurons in hidden Layer	100	100
	Activation	ReLu	ReLu
	Solver	L-BFGS-B	L-BFGS-B

Table S4. Hyperparameters and search spaces of Regression Tree (RT), Ensemble, and Gaussian Process Regression (GPR) models.

Algorithm	Parameter	Cuyahoga	Grand	Maumee	Raisin	Sandusky
Nitrate						
RT	Minimum leaf size	1-4198	1-1767	1-4497	1-3234	1-4014
Ensemble	Ensemble method			Bag, LSBoost		
	Minimum leaf size	1-4198	1-1767	1-4497	1-3234	1-4014
	Number of learners			10-500		
	Learning rate			0.001-1		
	Number of predictors to sample			1-2		
GPR	Sigma	0.0001-11.7395	0.0001-3.3716	0.0001-31.9037	0.0001-23.0761	0.0001-33.2232
	Basis function	Constant, Zero, Linear				
	Kernel function	Nonisotropic Exponential, Nonisotropic Matern 3/2, Nonisotropic Matern 5/2, Nonisotrpoic Rational Quadratic, Nonisotropic Squared Exponential, Isotropic Exponential, Isotropic Matern 3/2, Isotropic Matern 5/2, Isotropic Rational Quadratic, Isotropic Squared Exponential				
	Kernel scale	12.2517-12251.6582	13.3967-13396.6667	110.543-110543	22.5265-22526.5	25.4828-25482.8333
	Standardize	TRUE, FALSE				
Phosphorus						
RT	Minimum leaf size	1-4198	1-1767	1-4497	1-3234	1-4014
Ensemble	Ensemble method			Bag, LSBoost		
	Minimum leaf size	1-4198	1-1767	1-4497	1-3234	1-4014
	Number of learners			10-500		
	Learning rate			0.001-1		
	Number of predictors to sample			1-3		
GPR	Sigma	0.0001-1.8724	0.0001-0.92927	0.0001-1.5828	0.0001-1.1633	0.0001-1.7914
	Basis function	Constant, Zero, Linear				
	Kernel function	Nonisotropic Exponential, Nonisotropic Matern 3/2, Nonisotropic Matern 5/2, Nonisotrpoic Rational Quadratic, Nonisotropic Squared Exponential, Isotropic Exponential, Isotropic Matern 3/2, Isotropic Matern 5/2, Isotropic Rational Quadratic, Isotropic Squared Exponential				
	Kernel scale	12.2517-12251.6582	13.3967-13396.6667	110.543-110543	22.5265-22526.5	25.4828-25482.8333
	Standardize	TRUE, FALSE				

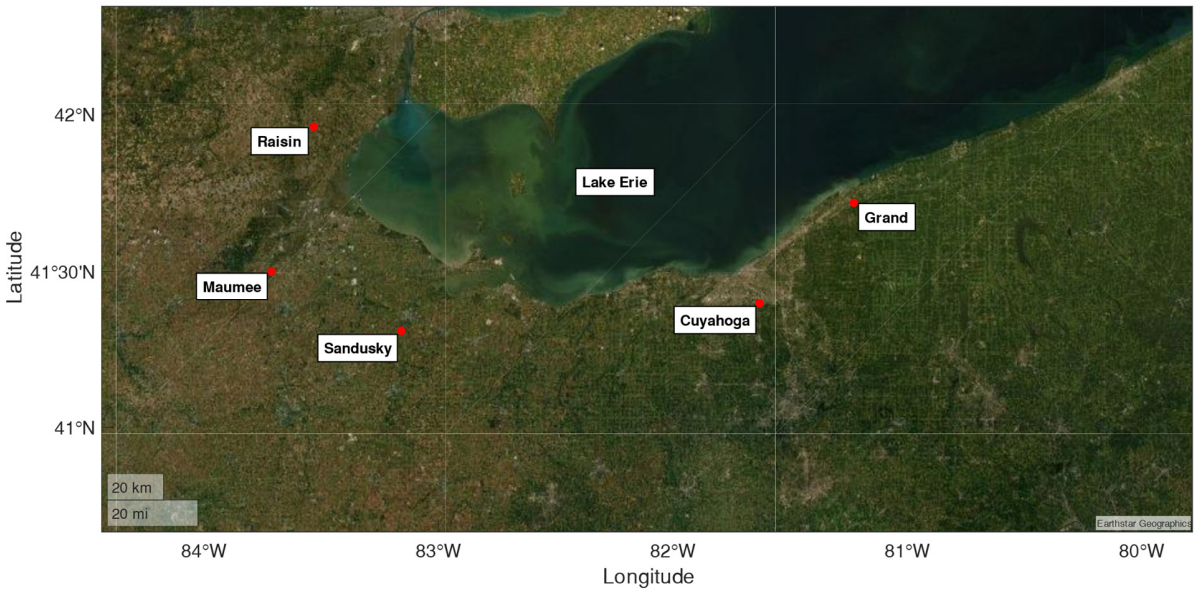

Figure S1. Locations of the Cuyahoga, Grand, Maumee, Raisin, and Sandusky watersheds and gaging stations (Source: Earthstar Geographics).

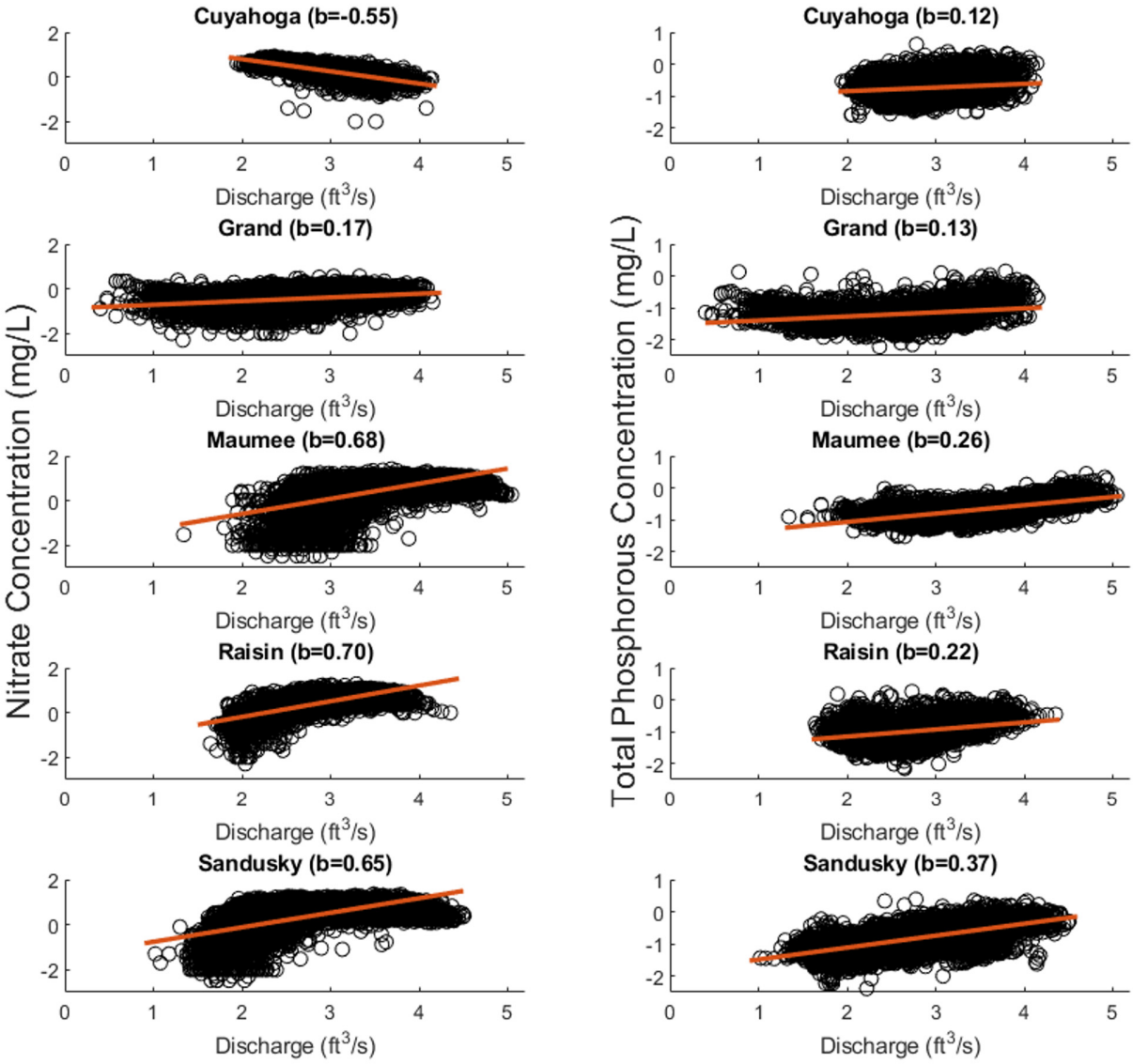

Figure S2. C-Q relation for each watershed. Slope of the C-Q relationship 'b' is given alongside the figure. Both axes are in logarithmic scale.