remote sensing

MDPI

Supporting Document

A machine learning-based assessment of proxies and drivers of Harmful Algal Blooms in the Western Lake Erie Basin using satellite remote sensing

Neha Joshi[1], Armeen Ghoorkhanian[1,2], Jongmin Park[1,3,4], Kaiguang Zhao[3], Sami Khanal[1*]

[1] Department of Food, Agricultural and Biological Engineering, The Ohio State University, Columbus, OH 43210, USA; joshi.304@osu.edu
[2] Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, 43210, USA; ghoorkhanian.5@osu.edu
[3] School of Environment and Natural Resources, The Ohio State University, Columbus, OH 43210, USA; zhao.1423@osu.edu
[4] Department of Environmental Engineering, Korea National University of Transportation, Chungju, 27469, Korea; jmpark1@ut.ac.kr
* Correspondence: khanal.3@osu.edu

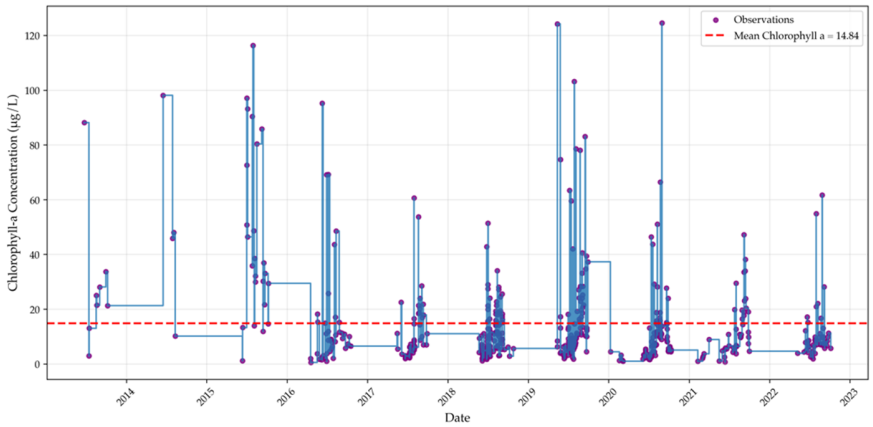

Figure S1. Temporal variation in chlorophyll-a concentration from June 2013 to October 2022.

Table S1. Minimum, maximum, mean, median and standard deviation by month for Chl-a concentration

Month	Min	Max	Mean	Median	St. Dev.	#of observations
January	4.40	4.40	4.40	4.40	N/A	1
February	0.92	3.22	1.79	1.23	1.25	3
March	0.96	3.68	1.98	1.69	1.03	6
April	0.53	8.88	3.79	1.97	4.47	3
May	1.03	124.25	17.41	6.22	30.50	19
June	0.71	98.13	9.10	4.24	17.76	93
July	1.58	116.36	13.88	6.52	20.61	164
August	3.03	124.55	19.37	14.03	17.51	135
September	4.47	85.84	18.41	13.37	14.29	91
October	2.79	29.43	8.15	5.86	6.50	20

Table S2. Minimum, maximum , mean, median, and standard deviation by year for Chl-a concentration

Year	Min	Max	Mean	Median	St. Dev.	#of observations
2013	2.91	88.18	29.19	23.22	25.62	8
2014	10.14	98.13	50.53	46.93	36.17	4
2015	1.16	116.36	46.84	36.39	31.95	24
2016	0.53	95.23	14.93	8.70	20.44	42
2017	1.90	60.60	11.09	6.37	11.64	51
2018	1.10	51.44	10.68	8.59	8.62	116
2019	1.79	124.25	18.80	10.50	22.66	93
2020	0.96	124.55	10.85	6.39	15.69	100
2021	0.71	47.18	12.47	9.60	10.77	45
2022	1.88	61.71	10.64	7.65	10.85	52

Table S3. Minimum, maximum, mean, median, and standard deviation for all variables over the study period

Variable	Min	Max	Mean	Median	St. Dev.	#of observations
Chlorophyll a (μg/L)	0.53	124.55	14.84	8.71	18.63	535
Water Temperature (ºC)	-0.82	29.89	22.52	23.95	4.43	535
Humidity (g/kg)	1.65	19.41	12.39	12.57	3.29	535
Wind Speed (m/s)	0.92	7.29	2.72	2.5	1.08	535
Precipitation (Mm)	0.0	31.57	1.38	0.0	4.79	535
TN (μg NP/L)	26.25	5984.94	935.45	736.11	684.31	477
TP (μg P/L)	1.16	177.91	31.58	22.70	28.63	535
Microcystin (μg/L)	0.0	21.53	0.98	0.25	2.00	451
SD (M)	0.04	9.15	1.79	1.58	1.22	481
DOM (μg/L)	0.0	2.55	0.58	0.47	0.43	515
Solar Radiation or Sun Light (kW-hr/m2/day)	3.74	11.6	10.20	10.74	1.50	535
Wind Direction (Degrees)	17.56	337.69	196.15	215.38	75.13	535
Blue (Nm)	0.0011	0.52	0.043	0.039	0.030	535
Green (Nm)	0.0088	0.44	0.063	0.058	0.030	535
Red (Nm)	0.0023	0.43	0.039	0.032	0.031	535
NIR (Nm)	0.0001	0.44	0.029	0.017	0.042	535
SWIR2 (Nm)	0.0	0.13	0.019	0.012	0.021	535
NDVI	-0.99	0.91	-0.29	-0.29	0.30	535
EVI	-0.20	0.64	-0.029	-0.035	0.067	535
SABI	-0.58	7.85	-0.085	-0.14	0.47	535
NDTI	-0.86	0.03	-0.29	-0.28	0.15	535

Data quality control

For satellite images, the HLS Fmask band [1] was used to discard the pixels with noise, including cloud, cloud shadow, and snow. Fmask for Landsat data is dependent on the thermal band for noise detection, Fmask for Sentinel-2 uses the cirrus band as it lacks thermal band [1]. The Fmask was handled through a quality assessment (QA) layer in both L30 and S30, encoded at a bit level. Among the seven bits, each bit corresponds to a particular state (Table 2), with Bit 0 linked to cirrus, Bit 1 to cloud, Bit 2 to adjacent cloud, and so on. A value of 0 indicated an absence, while 1 indicates a presence. Based on the HLS User guide [1], pixels with a value of 1 for any of the first three bits were masked out.

Table S4. Quality assessment to evaluate a pixel by indicating different conditions associated with different bit combinations

Bit Number	**QA description**	**Bit Combination**	**Description**
7-6	Aerosol Quality	00	Climatology
		01	Low
		10	Average
		11	High
5	Water	1	Yes
		0	No
4	Snow/Ice	1	Yes
		0	No
3	Cloud Shadow	1	Yes
		0	No
2	Adjacent Cloud	1	Yes
		0	No
1	Cloud	1	Yes
		0	No
0	Cirrus	1	Yes
		0	No

Table S5. List of hyperparameters and their values used for the models

Model	Hyperparameter	Physical Significance	Fitted Range
RF	bootstrap	decides whether sampling with replacement is used when building trees	TRUE
	max_depth	max number of levels in each decision tree	10
	max_features	number of features considered when splitting.	auto
	min_samples_leaf	min number of data points allowed in a leaf node	10
	min_samples_split	min number of data points placed in a node before the node is split	10
	n_estimators	number of decision trees in the forest	200
SVR	C	tradeoff between model complexity and training error	$[10^{-3}, 10^{3}]$ (log-uniform)
	gamma	influence of a single training point	$[10^{-6}, 10^{1}]$ (log-uniform)
	epsilon	Margin of tolerance	[0.01, 1.0] (uniform)
	degree	Degree of the polynomial kernel function	[2, 5] (integer)
	kernel	function used to map data into higher-dimensionality space	{rbf, poly, sigmoid}
XGB	colsample_bytree	Fraction of features used when constructing each tree	[0.1, 1.0] (uniform)
	gamma	Minimum loss reduction required for splitting a node	$[e^{-5}, e^{0}]$ (log-uniform)
	learning_rate	Learning rate	[0.0, 1.0] (uniform)
	max_depth	Maximum depth of each tree	[1, 10] (integer)
	min_child_weight	Minimum sum of instance weight for a child node to split further	[1, 10] (integer)
	n_estimators	Number of trees	[500, 1000] (integer step 10)
	reg_alpha	Adds sparsity to leaf weights	$[e^{-5}, e^{0}]$ (log-uniform)
	reg_lambda	Smoothens the model by shrinking leaf weights	$[e^{-5}, e^{0}]$ (log-uniform)
	subsample	Fraction of randomly sampled observations for each tree	[0.1, 1.0] (uniform)

Table S6. Distribution of R^2 across three models/variable combination and ten random seeds

Model/Variable Combination	R^2 across random seeds									
	2	**11**	**42**	**78**	**99**	**100**	**114**	**150**	**300**	**456**
RF										
ALL	**0.54**	0.62	**0.68**	**0.56**	**0.65**	0.51	0.54	**0.61**	**0.66**	**0.65**
B+BI	0.34	0.27	0.18	0.19	0.24	0.22	0.27	0.23	0.27	0.28
B+MET	0.30	0.07	0.14	0.18	0.20	0.18	0.22	0.12	0.19	0.18
B+PC	0.53	**0.63**	**0.68**	**0.56**	**0.65**	**0.52**	**0.56**	**0.61**	**0.66**	**0.65**
SVR										
ALL	0.36	0.34	**0.61**	0.65	0.61	0.51	**0.47**	0.66	0.70	**0.72**
B+BI	**0.48**	0.34	0.30	0.30	0.35	0.38	0.41	0.25	0.38	0.30
B+MET	0.34	0.12	0.12	0.14	0.21	0.21	0.18	0.10	0.16	0.12
B+PC	0.30	**0.48**	0.56	**0.73**	**0.72**	**0.76**	0.44	**0.69**	**0.76**	0.69
XGB										
ALL	0.57	0.70	0.72	**0.63**	0.60	0.58	0.66	**0.58**	**0.65**	0.53
B+BI	0.36	0.33	0.27	0.26	0.29	0.30	0.33	0.23	0.42	0.36
B+MET	0.28	0.17	0.24	0.19	0.12	0.22	0.17	0.08	0.16	0.28
B+PC	**0.65**	**0.72**	**0.73**	0.59	**0.62**	**0.64**	**0.69**	**0.58**	**0.65**	**0.61**

Note: Bolded numbers indicate the highest R^2 values across variable combinations for each of the three models, for each random seed.

Table S7. Distribution of RMSE across three models/variable combination and ten random seeds

Model/Variable Combination	RMSE across random seeds									
	2	**11**	**42**	**78**	**99**	**100**	**114**	**150**	**300**	**456**
RF										
ALL	13.46	12.38	11.52	**15.13**	**9.44**	11.35	13.56	**11.11**	**11.30**	10.94
B+BI	**11.63**	18.84	18.32	17.45	15.52	15.60	14.71	16.59	15.58	17.26
B+MET	11.97	21.17	18.85	17.60	15.88	15.94	15.19	17.67	16.45	18.40
B+PC	13.51	**12.34**	**11.45**	15.14	9.45	**11.23**	**13.30**	11.13	11.33	**10.84**
SVR										
ALL	15.80	16.33	**12.68**	13.51	9.91	11.36	14.57	10.39	10.63	**9.68**
B+BI	**10.30**	17.88	16.96	16.24	14.37	13.89	**13.25**	16.38	14.41	16.97
B+MET	11.63	20.64	19.06	18.00	15.82	15.63	15.63	17.87	16.72	19.06
B+PC	16.52	**14.48**	13.48	**11.83**	**8.41**	**8.04**	15.01	**9.91**	**9.53**	10.26
XGB										
ALL	12.90	11.07	10.68	**13.84**	10.07	10.52	11.66	**11.51**	11.53	12.56
B+BI	**11.50**	18.04	17.32	16.73	15.01	14.74	14.16	16.55	13.88	16.29
B+MET	12.20	20.03	17.70	17.51	16.70	15.55	15.76	18.15	16.75	17.31
B+PC	11.72	**10.60**	**10.59**	14.52	**9.73**	**9.78**	**11.22**	11.59	**11.49**	**11.54**

Note: Bolded numbers indicate the lowest RMSE values across variable combinations for each of the three models, for each random seed.

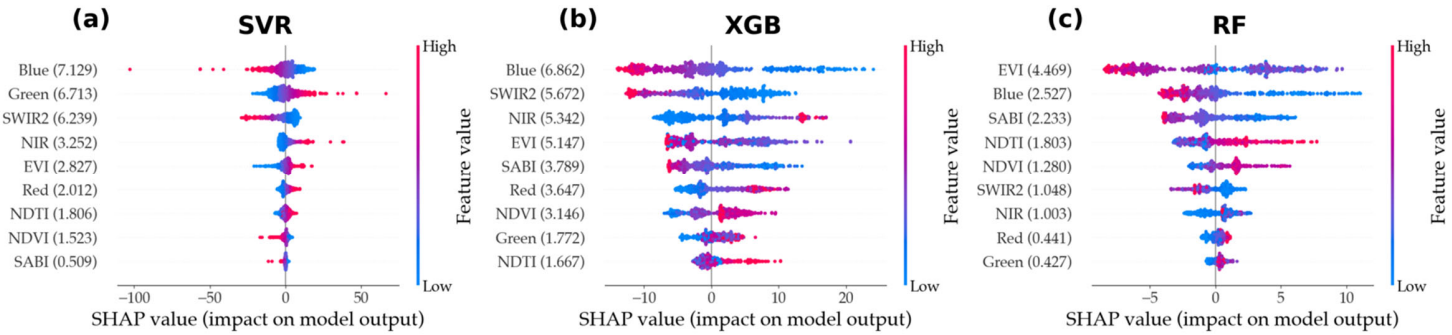

Figure S2 SHAP plots showing the influence of spectral bands and derived indices on Chl-a prediction using models (a) SVR (b) XGB, and (c) RF. Predictor variables are ranked by their influence on the model, with the mean absolute SHAP value shown alongside the corresponding variables.

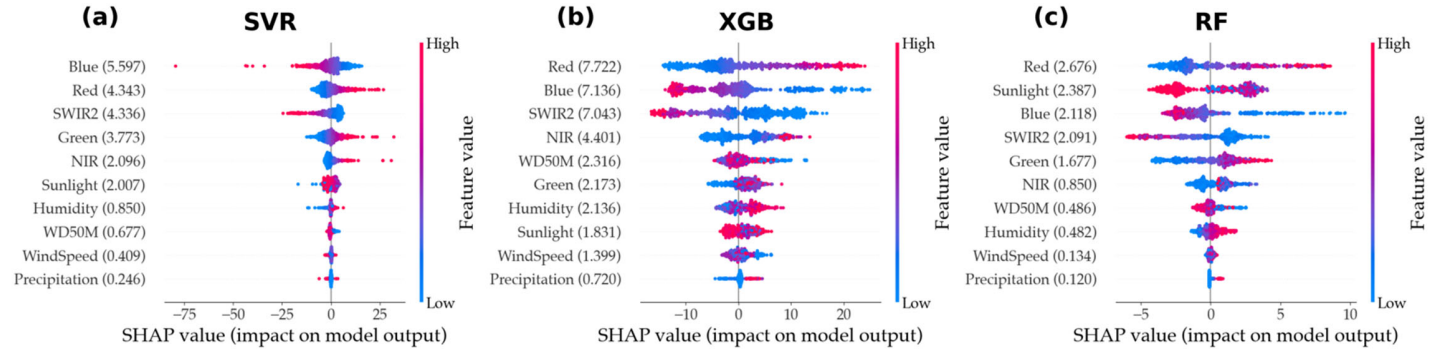

Figure S3. SHAP plots showing the influence of spectral bands and meteorological variables on Chl-a prediction using (a) SVR (b) XGB, and (c) RF models.

References

1. Claverie, M.; Ju, J.; Masek, J.G.; Dungan, J.L.; Vermote, E.F.; Roger, J.C.; Skakun, S. V.; Justice, C. The Harmonized Landsat and Sentinel-2 Surface Reflectance Data Set. *Remote Sensing of Environment* **2018**, *219*, 145–161, doi:10.1016/J.RSE.2018.09.002.