Supplementary Material

Recent Changes in Cyanobacteria Algal Bloom Magnitude in Lakes across the Contiguous United States

Sachidananda Mishra[1, 2], Richard P. Stumpf[2], Blake A. Schaeffer[3], P. Jeremy Werdell[4]*

[1]Consolidated Safety Services Inc., Fairfax, VA 22030, USA
[2]National Oceanic and Atmospheric Administration, National Centers for Coastal Ocean Science, Silver Spring, MD 20910, USA
[3]U.S. Environmental Protection Agency, Office of Research and Development, Durham, NC 27709, USA
[4]Ocean Ecology Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA

[*]**Corresponding author**: Phone: +1 (240) 621-1680, Email: sachi.mishra@noaa.gov

Content Summary: 16 pages; 7 figures; 3 Tables, SM References

1. Lake outline data

In order to calculate bloom magnitude within lake boundaries, lake polygon boundaries were selected from the National Hydrography Dataset Plus version 2.0 (NHDPlusV2) lake polygons dataset (McKay et al., 2012), with the condition that a satellite image should resolve each selected water body with a 300x300 m pixel resolution. Details of the selection method are available elsewhere (Mishra et al., 2019). Thus, 2357 lakes can be assessed with MERIS/OLCI observations (300x300m spatial resolution). Of that, 26 lakes in the Northern Gulf of Mexico (GOM) Louisiana marsh, nine lakes in the southern tip of the Florida panhandle, and one on the GOM coast in Texas were brackish water lakes. Thus, we removed 36 saline/brackish lakes bringing the total lake count to 2,321. In addition, 440 lakes did not have all nine years of observation (2008-2011, 2016-2020). As we wanted to keep the years of observation consistent to determine directly comparable statistics (e.g., Sen slope and Kendal's τ), we analyzed 1881 lakes with all nine years of data (Fig. S1, Fig. 4). The surface area of the selected lakes varied from 0.94 km^2 to 4,310 km^2 with a median value of 8.15 km^2. However, for independent satellite-based assessment and monitoring of lakes, all 2357 lakes can be used for estimating bloom magnitudes.

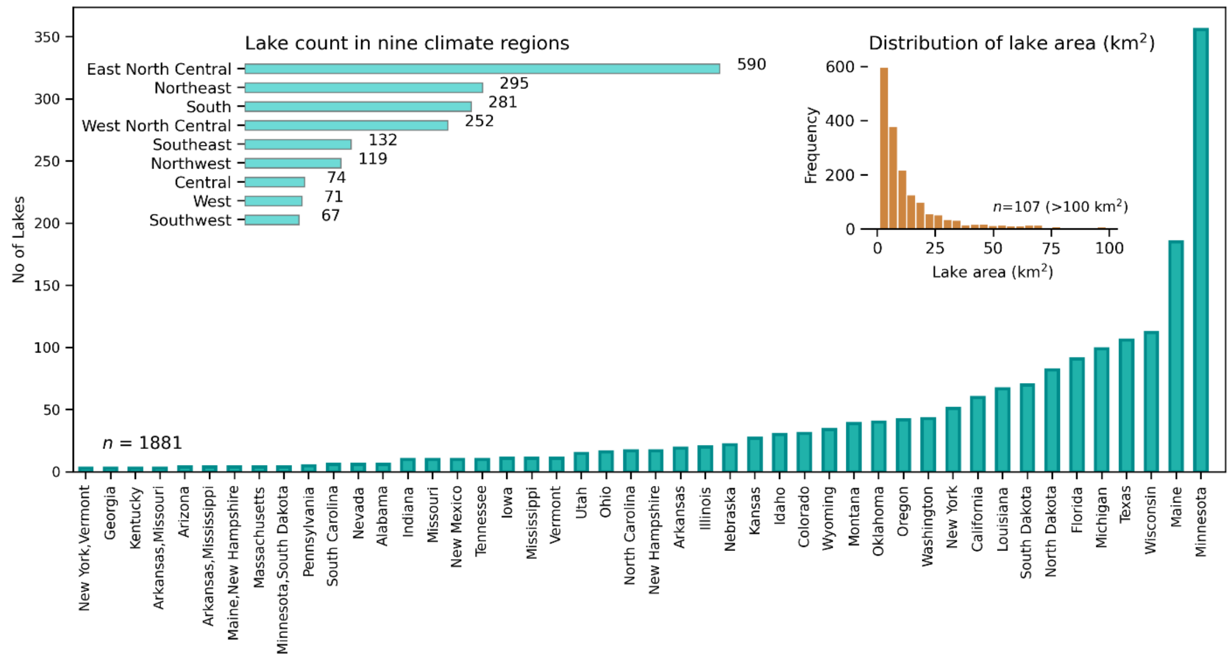

Figure S1. Distribution of lakes in the contiguous United States grouped by state and nine climate regions (left inset, light turquoise bars). Distribution of lake area is provided (right inset, orange bars) (range 0.94 km^2 to 4,310 km^2; median area: 8.15 km^2) . For better visualization of the majority of the lakes, lakes larger than 100 km^2 (n=107) surface area are not shown.

2. World Health Organization (WHO) Alert Levels

WHO recently recommended an update to the cyanobacteria harmful algal bloom (cyanoHAB) monitoring strategy in recreational water considering water sports and other activities in water are likely to be a major route of cyanotoxin exposure (Chorus and Welker, 2021). Thus the updated Alert Level Framework (ALF), which is a monitoring and management action sequence, replaced the old WHO guidelines of low, moderate, and high risk categories (Chorus and Bartram, 1999). Based on the new recommendation, there are three levels based on biovolumes or chlorophyll-a (chl-a) concentrations in recreational water bodies that triggers an action. Here we list the concentration-based action levels when cyanobacteria are the dominant algal-type in the water bodies:

i. Vigilance Level: chl-a concentration is within 3-12 mg m^{-3} (0.00045 – 0.0018 CI_{cyano}) (assess for toxin-producing cyanobacteria)

ii. Alert Level 1: chl-a biomass is within 12-24 mg m^{-3} (0.0018 – 0.0036 CI_{cyano}) (watch for scums and if possible, conduct toxin analysis; inform site users to avoid recreational activities)

iii. Alert Level 2: chl-a biomass is > 24 mg m^{-3} (0.0036 CI_{cyano}) chl-a with presence of toxins.

Our satellite-based biomass detection primarily focuses on cyanobacteria. Therefore, the updated ALF is applicable for cyanoHAB monitoring and assessment. However, note that no toxin analysis was carried out to determine the Alert Level-2 in this study. It was solely determined based on satellite-derived chl-a biomass. For detailed description on the ALF refer to (Chorus and Welker, 2021).

3. Random forest model and feature selection

Random Forest (RF) model grows *n* number of trees by randomly selecting a subset of features and splitting them following the Classification and Regression Tree (CART) methodology. RF regression model measures the importance of each feature based on the reduction in the model accuracy when the feature in question is excluded from a subset of features within a tree (Breiman, 2001). Thus, decision trees with subsets of features excluding highly informative features will lead to higher model error or reduced prediction accuracy, highlighting the feature's importance to the decision tree. The model accuracy averaged across decision trees with and without the feature in question provides the feature's importance and ranks them based on their importance. Based on the drastic change in feature rank and their importance, we selected eight LULC and climate features for modeling.

Selected LULC features

- *All_crops_acr_pct_hu12*: is the percentage of the total acreage of all croplands in the HUC 12, representing the agricultural activity in the hydrologic unit surrounding a lake under study. Therefore, that would serve as a proxy of nutrient loading to a lake in the form of excess nutrients transferred from surrounding agricultural land to the lake through surface runoff.
- *Forest_shrub_acr_pct_hu8:* is the percent area of the HU with code eight surrounding a lake covered by forest and shrubland. Lakes in hydrologic units with higher forest and shrubland cover would be expected to be in pristine condition with less anthropogenic disturbance.
- *Grassland_pasture_acr_pct_hu10*: is the percent area of the HU with code ten surrounding a lake covered by grassland and pasture. Grasslands and pastures can act as sources by working as a nonpoint source of excessive fertilizer. It can also serve as a sink by absorbing nutrients from the surface runoff by taking the role of cover crops.
- *Wetland_acr_pct_hu12:* is the percent area of the HU with code 12 surrounding a lake covered by wetlands. Wetlands can serve as nutrient sources or sinks, influencing the bloom condition in a lake.

Selected climate features

- PDSI above normal ($PDSI_{AN}$): The Palmer Drought Severity Index (PDSI) is a standardized index computed from temperature and precipitation data to estimate relative dryness. Generally, it varies from -10 (dry) to +10 (wet), although operation maps typically vary from -4 to +4. $PDSI_{AN}$ represents the percentage area of the climate region with severe moisture surplus (equivalent to the highest tenth percentile of the local period of record) based on the PDSI. It varies from 0 (extreme condition was nowhere recorded) to 100% (extreme condition was recorded everywhere) in the climate region.
- T_{max} (Mar-Oct) (°C) is the maximum temperature observed from March to October.
- Cumulative precipitation (Jun - July) is the accumulation of precipitation over June to July measured in mm.
- Cooling Degree Days (CDD)(°F) represents how much warmer the mean air temperature is compared to a baseline temperature (E.g., 65 °F in this study). For example, if the daily mean temperature for a day were 78 °F, the CDD for the day would be 13 °F (78°F - 65°F). Thus, the accumulation of such CDDs over a time period would mean the prevalence of warmer air conditions in a region.

4. Geographically Weighted Regression (GWR)

Geographically weighted regression (GWR) extends ordinary least-square (OLS) regression. Using a spatial weight matrix allows models to vary over space, addressing the non-stationary effect of independent variables on the response variable (Brunsdon et al., 1996; Fotheringham et al., 1997; Fotheringham et al., 2001).

$$y_i = \beta_{i0} + \sum_{k=1}^{m} \beta_{ik} x_{ik} + \varepsilon_i \quad \text{(S1)}$$

Where y_i is the dependent variable at lake year i; β_{i0} refers to the regression intercept; β_{ik} refers to the independent parameter; X_{ik} is the value of the k^{th} regression parameter; ε_i refers to the model residuals at lake year location i.

$$\hat{\beta}_i = (X^T W_i X)^{-1} X^T W_i y \quad \text{(S2)}$$

$$w_{ij} = \left[-\frac{1}{2} \left(\frac{d_{ij}}{b} \right)^2 \right] \quad \text{(S3)}$$

where d_{ij} is the Euclidian distance between observation point j and regression point i with planar coordinates, and b is the kernel bandwidth.

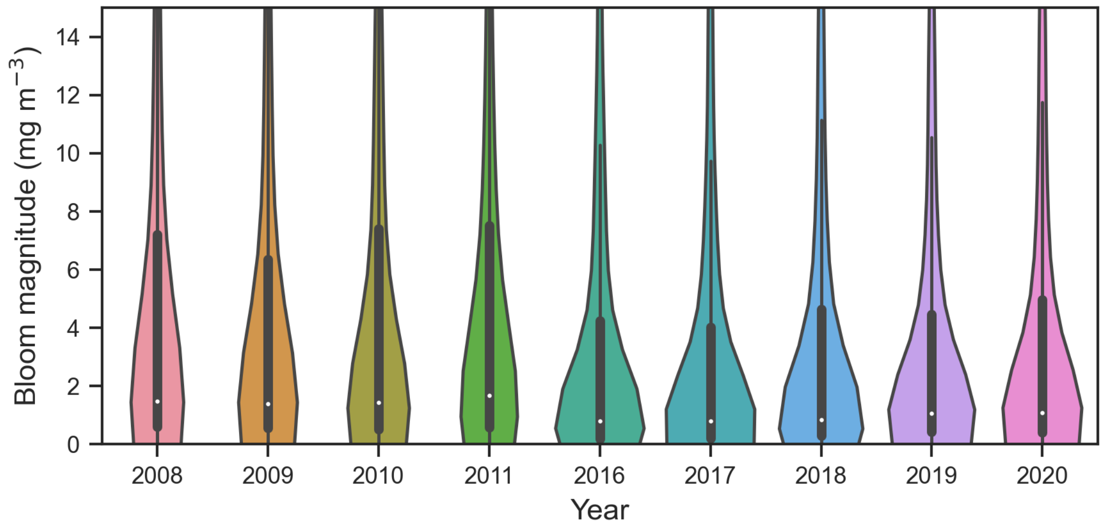

Figure S2. Distribution of bloom magnitude in the CONUS lakes over the 2008-2011 and 2016-2020 time period. Violin-like shapes show the distribution of bloom magnitude data by year (color-coded). Thus, the width of the violin represents the distribution shape (density) of the data (or number of lakes with a certain bloom magnitude) in a given year. The top and bottom bound of the black boxes inside the violin shapes represents the interquartile range. The whiskers show 1.5 times of the interquartile range. The white dot in the middle is the median. For better visualization, we trimmed the Y-axis to focus on the majority of the data, thus losing the extreme values (outliers). See Table S1 for the summary of the entire dataset.

5. Change analysis with extended MERIS time series

Although spatial coverage of MERIS full resolution (300×300m) data prior to 2008 was patchy across CONUS, for comparison, we used the entire MERIS time series with annual data (2003-2011) to calculate the cyanoHAB magnitude trends (Fig. S3). With the constraint of lakes requiring 14 years of data (2003-2011, 2016-2020), the total lake count came down to 1651 as 230 lakes lacked observations of at least one year from 2003-2008. With the extended MERIS time series data, ~2.5% and 25.6% of the lakes showed an increase and decrease in CyanoHAB magnitude trend, respectively. The same numbers derived from the 2008-2020 time series are 4.7% and 22% (Fig. 3). Similarly, 72% of the lakes showed no trend (at $|\tau| > 0.3$) matching the

2008-2020 trend analysis number (73%). Extension of the MERIS time series data back to 2003 shows the status of bloom magnitude from 2003-2008 and underlines the decreasing change rate compared to the lakes showing cyanoHAB intensification over the same time in a broader spatial scale across the CONUS.

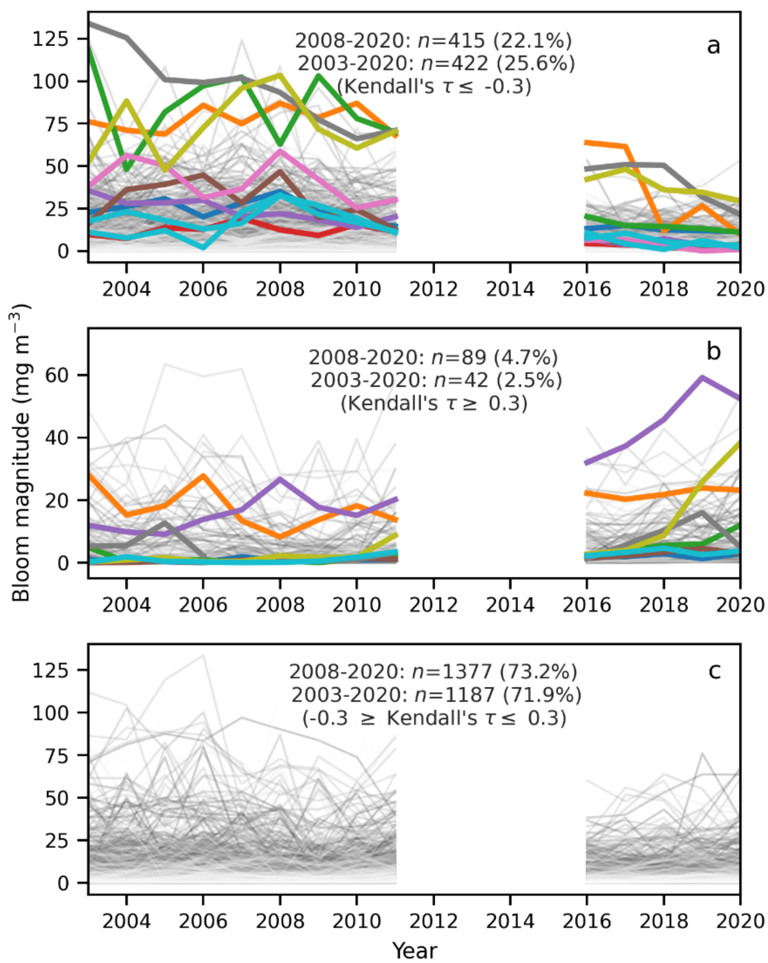

Figure S3. Cyanobacterial chl-a time series in lakes in contiguous United States as observed from the satellite-derived data (2003-2020). a) lakes where the bloom magnitudes have moderately or strongly decreased; b) Lakes where bloom magnitudes have moderately or strongly increased; c) lakes with weak decreasing or increasing trends over the observation period. Gray lines indicate change over time with moderate (Kendall’s $|\tau| > 0.3$), and colored lines indicate strong (Kendall’s $|\tau| > 0.5$). Note satellite observation gap from 2012 through 2015.

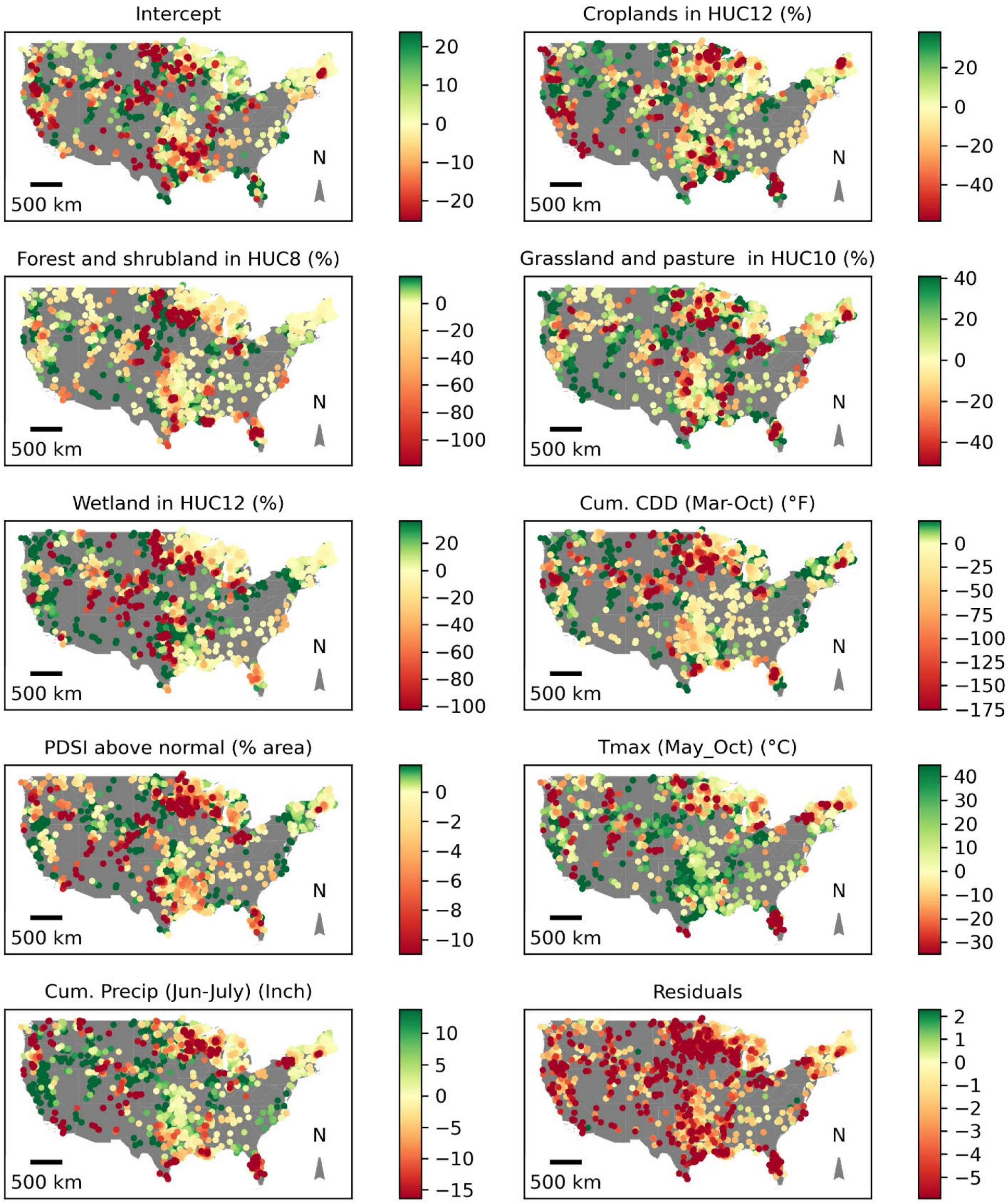

Figure S4. Surface maps of the model coefficients from the Geographically Weighted Model (GWR). Map of model residuals is also provided. Units are dimensionless.

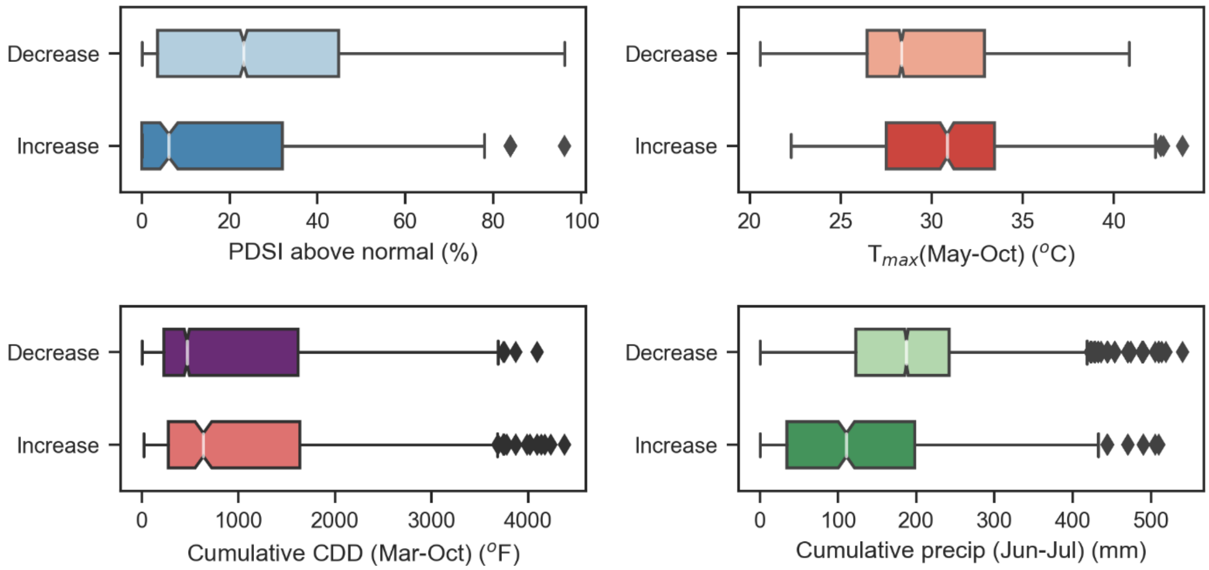

Figure S5. The distribution of climate variables used in the model with in groups ('*Increase*,' '*Decrease*') of lakes where bloom magnitude increased or decreased. The left and right bounds of the boxes represent the first and third quartiles, respectively. The whiskers show 1.5 times of the interquartile range. The white bar in the middle is the median, and the diamonds are detected as outliers.

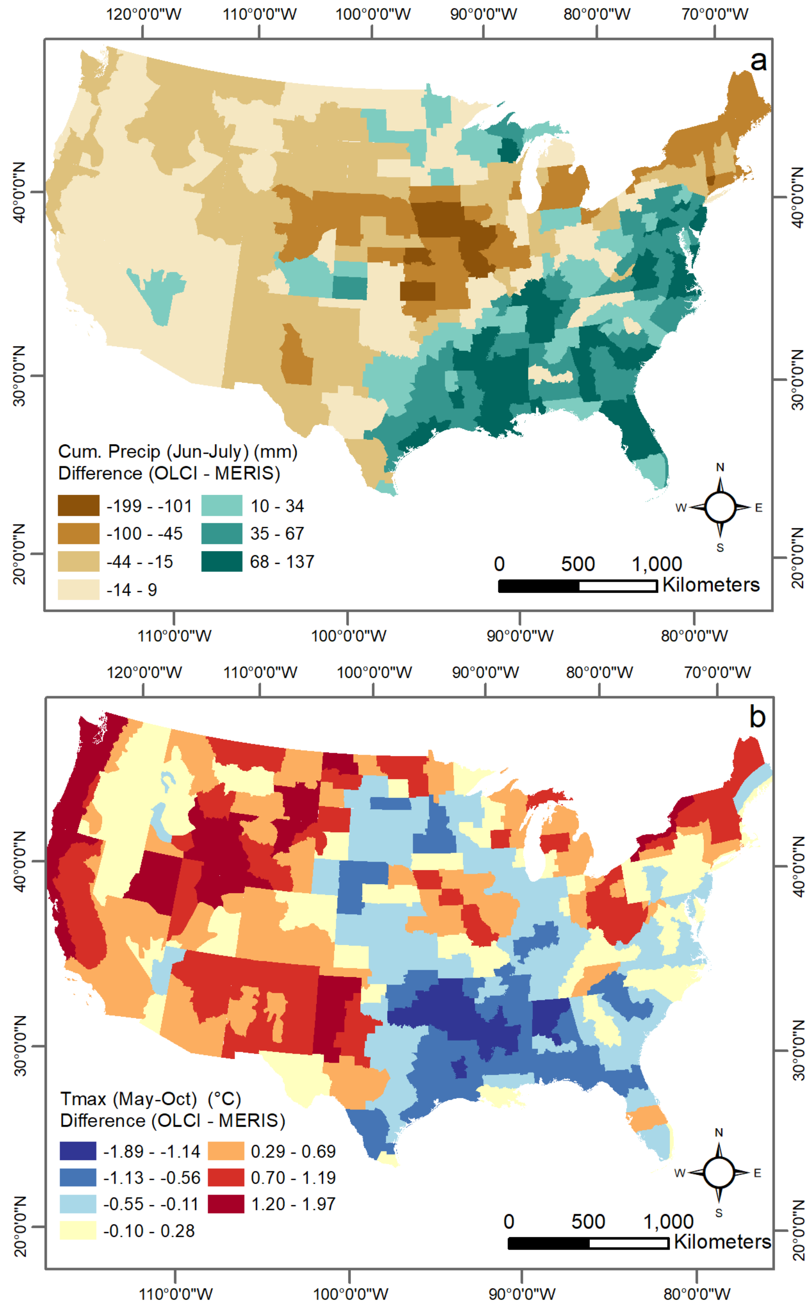

Figure S6. a) Difference in Cum. precipitation (Jun-July) (mm) at climate divisions level over 2008-2011 (MERIS) and 2016-2020 (OLCI) observation periods. A positive difference indicates the median over the OLCI period to be larger; b) difference between climate division-level median T_{max} (May-Oct) (°C) over the same observation period across the CONUS.

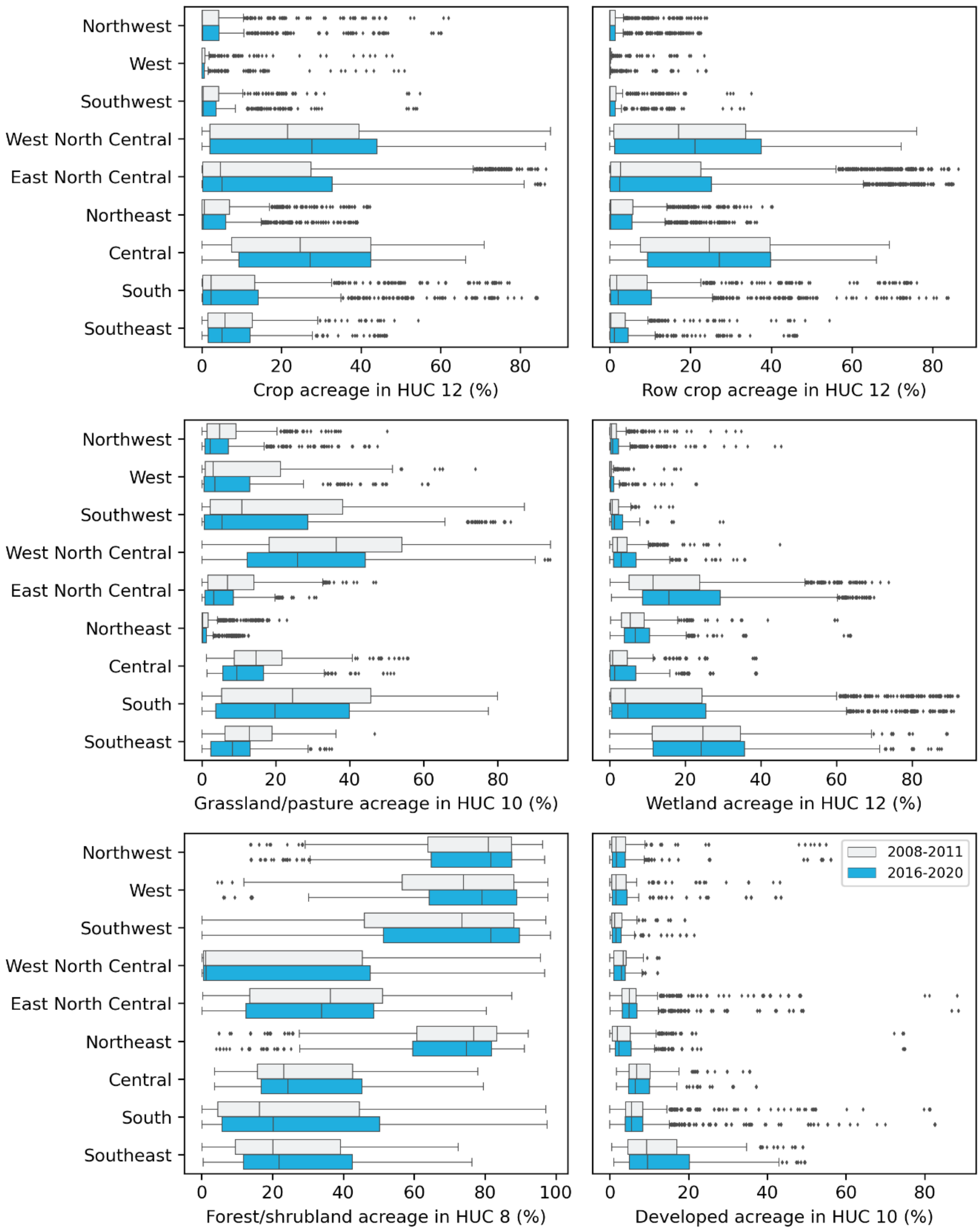

Figure S7. The distribution of land use land cover types in the corresponding hydrologic units in each climate region in the CONUS over 2008-2011 (MERIS, gray boxes) and 2016-2020 (OLCI, blue boxes). Left and right bound of the boxes represent the first and third quartiles, respectively. The whiskers show 1.5 times of the interquartile range. The vertical bars in the middle of the boxes are the median, and the diamond markers are outliers.

Table S1. Descriptive statistics of bloom magnitude (mg m^{-3}) in CONUS over 2008-2011 and 2016-2020 time period. Std is standard deviation.

Year	mean	std	min	25%	50%	75%	99%	max
2008	7.06	13.61	0.01	0.59	1.48	7.22	63.44	172.69
2009	6.21	12.13	0.02	0.56	1.39	6.35	58.97	157.35
2010	6.30	11.05	0.01	0.52	1.44	7.40	53.04	141.86
2011	6.62	12.03	0.02	0.57	1.68	7.52	54.89	144.27
2016	3.98	7.97	0.00	0.19	0.79	4.25	34.25	127.23
2017	3.84	7.77	0.00	0.22	0.79	4.03	37.56	107.84
2018	4.33	8.50	0.00	0.30	0.84	4.64	37.33	134.14
2019	4.33	8.17	0.00	0.42	1.07	4.47	36.94	111.65
2020	4.68	8.88	0.00	0.40	1.08	4.97	40.91	120.33

Table S2. Summary of model coefficients from the geographically weighted regression model with Land use/Land Cover (LULC) and climate variables as the explanatory variables.

	5th percentile	1^{st} quantile	Mean	Median	3^{rd} quantile	95th percentile
Intercept	-40.80	-8.67	11.75	1.64	15.21	83.88
All croplands fraction (%) in HUC12	-94.63	-15.63	21.14	3.03	28.28	126.83
Forest and shrubland fraction (%) in HUC8	-431.34	-27.10	197.40	-2.05	10.71	499.94
Grassland and pasture fraction (%) in HUC10	-76.39	-14.62	13.01	7.49	33.31	132.77
Wetland fraction (%) in HUC12	-182.34	-23.66	-3.85	0.31	23.26	168.17
Cum. CDD (Mar-Oct) (°F)	-262.26	-80.30	-42.15	-16.66	14.02	133.71
PDSI above normal (% area)	-16.74	-4.74	-1.88	-1.21	1.04	10.35
T_{max} (May_Oct) (°C)	-78.62	-6.67	11.02	10.31	37.49	107.05
Cum. Precip (Jun-July) (Inch)	-27.45	-4.78	4.43	1.01	10.32	51.57
Residuals	-8.67	-2.48	-0.05	-0.35	1.40	10.45
Local R^2	0.17	0.35	0.42	0.46	0.58	0.74

Table S3. Summary statistics of key model covariates associated with the groups ('*Increase,*' '*Decrease*') of lakes where bloom magnitude has increased or decreased (see methods: Bloom magnitude ratio). Sample size (number of lakes × number of years) in increase and decrease categories are 6,444 and 621, respectively. Differences between the means were computed using Cohen's *d* metric(Cohen, 1988; Sawilowsky, 2009).

		Decrease group	**Increase group**	**Mean and Median difference (Cohen's *d*)**
PDSI above normal (%)	mean	26.33	16.62	-9.71 (*d = -0.6*)
	std	24.17	21.54	
	1%	0.00	0.00	
	25%	3.60	0.00	
	50%	23.20	6.20	-17.00
	75%	44.80	32.00	
	99%	83.90	83.90	
T_{max} (May-Oct) °C	mean	29.32	30.69	1.37 (*d = 0.49*)
	std	3.85	4.05	
	1%	22.28	23.24	
	25%	26.44	27.50	
	50%	28.33	30.83	2.50
	75%	32.89	33.44	
	99%	38.84	41.70	
Cum. CDD (Mar-Oct) °F	mean	958.52	1117.12	158.6 (*d = 0.213*)
	std	1005.77	1098.13	
	1%	16.00	42.40	
	25%	229.00	275.00	
	50%	467.50	639.00	171.50
	75%	1619.25	1638.00	
	99%	3377.00	4021.60	
Cum. precip (Jun-Jul) (mm)	mean	187.55	127.83	-59.72 (*d = -0.806*)
	std	101.04	108.44	
	1%	3.81	0.00	
	25%	122.68	34.54	
	50%	187.45	110.74	-76.71
	75%	242.06	198.12	
	99%	470.92	444.80	

SI Reference

Breiman L. Random forests. Machine learning 2001; 45: 5-32.
Brunsdon C, Fotheringham AS, Charlton ME. Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical analysis 1996; 28: 281-298.
Chorus I, Bartram J. Toxic cyanobacteria in water: a guide to their public health consequences, monitoring and management.© 1999. London: E & FN Spoon, on behalf of WHO, 1999.
Chorus I, Welker M. Exposure to cyanotoxins: Understanding it and short-term interventions to prevent it. Toxic Cyanobacteria in Water. CRC Press, 2021, pp. 295-400.
Cohen J. Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates. Hillsdale, NJ 1988: 20-26.
Fotheringham AS, Charlton M, Brunsdon C. Two techniques for exploring non-stationarity in geographical data. Geographical Systems 1997; 4: 59-82.
Fotheringham AS, Charlton ME, Brunsdon C. Spatial variations in school performance: a local analysis using geographically weighted regression. Geographical and environmental Modelling 2001; 5: 43-66.
McKay L, Bondelid T, Dewald T, Johnston J, Moore R, Rea A. NHDPlus Version 2: User Guide. 2012.
Mishra S, Stumpf RP, Schaeffer BA, Werdell PJ, Loftin KA, Meredith A. Measurement of Cyanobacterial Bloom Magnitude using Satellite Remote Sensing. Scientific Reports 2019; 9: 1-17.
Sawilowsky SS. New effect size rules of thumb. Journal of modern applied statistical methods 2009; 8: 26.