








Improved seasonal prediction of harmful algal blooms in Lake Erie using large-scale climate indices

Mukul Tewari ¹✉, Chandra M. Kishtawal², Vincent W. Moriarty ¹, Pallav Ray ³, Tarkeshwar Singh ⁴, Lei Zhang ^{5,7}, Lloyd Treinish ¹ & Kushagra Tewari ⁶

Harmful Algal Blooms lead to multi-billion-dollar losses in the United States due to shellfish closures, fish mortalities, and reluctance to consume seafood. Therefore, an improved early seasonal prediction of harmful algal blooms severity is important. Conventional methods for harmful algal blooms prediction using nutrient loading as the primary driver have been found to be less accurate during extreme bloom years. Here we show that a machine learning approach using observed nutrient loading, and large-scale climate indices can improve the harmful algal blooms prediction in Lake Erie. Moreover, the seasonal prediction of harmful algal blooms can be completed by early June, before the expected peak in harmful algal bloom activity from July to October. This improved early seasonal prediction can provide timely information to policymakers for adopting proper planning and mitigation strategies such as restrictions in harvesting and help in monitoring toxins in shellfish to keep contaminated products off the market.

¹IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA. ²ClimateAI, San Francisco, CA, USA. ³Florida Institute of Technology, Melbourne, FL, USA. ⁴Nansen Environmental and Remote Sensing Center and Bjerknes Centre for Climate Research, Bergen, Norway. ⁵University of Colorado, Boulder, CO, USA. ⁶University of Southern California, Los Angeles, CA, USA. ⁷Present address: South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou 510301, China. ✉email: mtewari@us.ibm.com

Phytoplankton blooms with the potential to harm aquatic ecosystems and human health are referred to as Harmful Algal Blooms (HABs). Most HAB phenomena are caused by blooms of microscopic algae or phytoplankton (including certain cyanobacteria), although the term also applies to blooms of macroalgae (seaweeds)¹. HABs are of worldwide concern due to their potential to harm the environment and human health². It also has tremendous economic impacts with over \$4.6B per annum in estimated damages within the United States alone³.

The Laurentian Great Lakes of North America contain approximately 95% of the surface freshwater supply in the United States⁴. The Great Lakes are also the world's largest and most biologically diverse freshwater resource⁵. Lake Erie is the fourth-largest by area (surface area = 25874 km²), the shallowest (mean depth = 19 m), and the smallest by water volume (484 km³). It is also the most southern among the Great Lakes with its watershed containing both urbanized environments that include large manufacturing facilities, and rural environments that include extensive areas of agriculture. Hence, Lake Erie (Fig. 1a) is severely impacted by excessive nutrient loading, especially from its western basins⁶. Over the past decades, western Lake Erie has witnessed increasingly intense algal blooms including cyanobacterial blooms^{7–9}. From analyses of long-term trends in nutrient loading and meteorological conditions (including severe springtime precipitation events coupled with long-term trends in agricultural land use and practices), it was determined that the 2011 extreme HAB (Fig. 1a) in western Lake Erie was largely driven by trends in these factors¹⁰. The bloom extended over 16 km from the shores, and in the central basin, it was observed at a depth of 60 m. The bloom density also caused damage to the navigation because it disrupted boat motors.

A seasonal severity index (SI) for HABs in Lake Erie has been made available by the National Oceanic and Atmospheric Administration (NOAA) since the 2002 season⁹, and contains estimates for the maximum bloom biomass seen for the peak 30 days of a bloom. Stumpf et al.⁹ found that a blended model incorporating springtime discharge and measured total phosphorus (TP) appeared sufficient to predict the bloom magnitude in western Lake Erie. However, the linear and non-linear regression models based on TP and water discharge alone could not successfully explain some extreme HAB events (e.g., the extreme HAB event in 2011). It is therefore hypothesized that some other unknown biological, chemical, hydrodynamical, and meteorological factors may be responsible for this unexplained variance.

Previous studies have analyzed the relationships between algal growth rates and toxin production in response to temperature and inorganic nutrients^{11–14}. However, the relationship between HABs and large-scale atmospheric conditions has rarely been addressed. Nonetheless, the identification and understanding of atmospheric conditions favoring HABs could be essential to predict these events. Baez et al.¹⁵ showed that HAB events in southwest Europe are linked to atmospheric teleconnections, which in turn are correlated with hydrographical conditions favoring HABs. Phlips et al.¹⁶ examined the roles that hurricanes and El Niño play in contributing to HAB events in two subtropical Florida estuaries. They found that an increase in rainfall associated with hurricanes and El Niño resulted in increased nutrient loads, ultimately driving HABs within the Indian River Lagoon and St. Lucie Estuary. Previous studies have also explored the relationship between interannual and multi-decadal variability of atmospheric circulation and the occurrences of HABs (e.g., refs. 17–21). Zhou et al.²² analyzed bloom extent in Lake Erie and developed a model for explaining interannual variability of Lake Erie hypoxia. Their model, based on four variables (April to June tributary discharge, May to July soluble reactive phosphorus (SRP) loading, July wind stress, and June northwesterly wind duration) can explain 82% of the interannual variability of hypoxia. They discussed the need for meteorological factors in the development of nutrient management strategies by showing that a discharge-only model explained only 39% of the variability (i.e., if the meteorological factors were not included in the model).

It is widely accepted that the occurrence of HABs in freshwater bodies results from complex non-linear interactions between chemical, biological, hydrological, and meteorological processes that take place at different spatial and temporal scales. Data-driven machine learning techniques may help understand these complex non-linear interactions and develop predictive modeling that may capture these non-linear interactions in a way that would not be possible using statistical or dynamical models. A recent review paper²³ discussed the advances made in the machine learning methods for predicting HABs and shellfish biotoxin contamination, with a particular focus on autoregressive models, support vector machines, random forest, probabilistic graphical models, and artificial neural networks (ANNs). Likewise, most measures of local conditions used by ecologists fail to capture complex associations between weather and ecological processes²⁴. As a result, large-scale, seasonal indices of climate

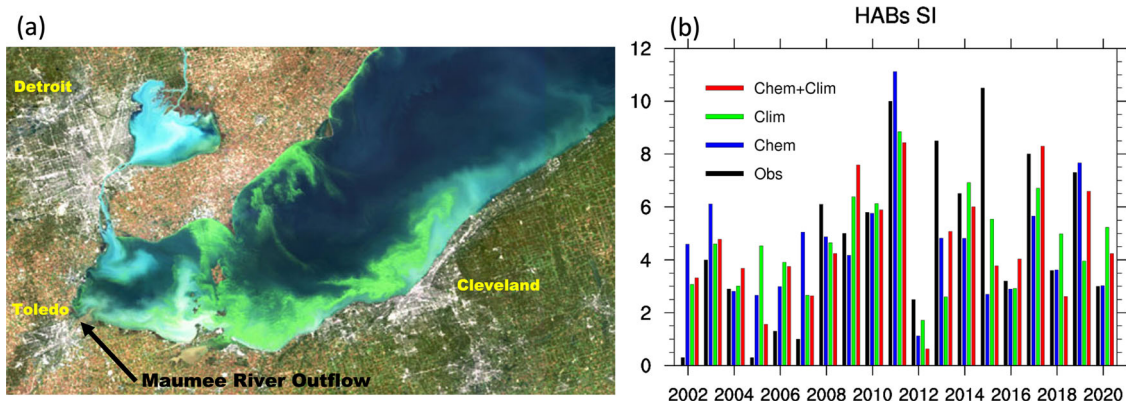


Fig. 1 Satellite image of algal bloom over Lake Erie and seasonal prediction of HABs severity index. **a** Optical three-band visible image from the Envisat satellite (MERIS_FP) at 300 m resolution showing one of the largest algae blooms from October 2011 extended from Toledo to beyond Cleveland and along the Ontario shore (courtesy of European Space Agency). Data from three visible bands were used each with a bandwidth of 10 nm (442.5, 460, and 665 nm band centers). **b** HABs severity index (SI) from observation (black bar), prediction using GA-chem model that is based on nutrient loading only (blue bar), prediction using GA-clim model that is based on climate indices only (green bar) and GA-chem-clim that is based on nutrient loading and climate indices (red bar).

Table 1 Performance (root mean square error (RMSE) and coefficient of determination (R^2)) of GA-chem-clim (using nutrient loading and climate indices) developed using Genetic Algorithm under the jackknife procedure.

Year	HABs SI function	RMSE	R^2
2002	$((6.93) \cdot TP) - ((PNA(Dec) - PDO(Nov)) - (2.29))$	1.43	0.60
2003	$((10.11) + PDO(Nov)) \cdot TP - (PNA(Dec) + (-0.42))$	1.14	0.77
2004	$(TP \cdot (7.32)) - (PNA(Dec) - (PDO(Nov) + (1.97)))$	1.52	0.60
2005	$(TP \cdot ((12.27) - (PNA(Dec) \cdot PNA(Dec)))) - PNA(Dec)$	0.96	0.72
2006	$((PDO(Nov) - PNA(Dec)) + (SOI(Apr) + TP)) + (4.29)$	1.54	0.59
2007	$(SOI(Apr) - ((-4.75) - (PDO(Nov) \cdot (1.61)))) - PNA(Dec)$	1.63	0.60
2008	$(2.07) + (PDO(Nov) - (PNA(Dec) + ((-6.85) \cdot TP)))$	1.47	0.61
2009	$(SOI(Apr) + PDO(Nov)) + ((4.65) - (PNA(Dec) \cdot (1.69)))$	1.48	0.62
2010	$(PDO(Nov) - (PNA(Dec) - PDO(Nov))) + ((4.73) + SOI(Apr))$	1.77	0.60
2011	$(SOI(Apr) + (PDO(Nov) \cdot (1.63))) - (PNA(Dec) + (-4.59))$	1.57	0.56
2012	$((8.01) \cdot TP) - (PNA(Dec) - (PDO(Nov) + (1.59)))$	1.45	0.61
2013	$(PDO(Nov) + (PDO(Nov) + ((4.55) - (PNA(Dec) - SOI(Apr))))$	1.52	0.63
2014	$((7.52) \cdot TP) - PNA(Dec) + (PDO(Nov) - (-1.80))$	1.52	0.59
2015	$((PDO(Nov) - TP) + (12.68)) \cdot TP$	1.37	0.55
2016	$(PDO(Nov) - ((-4.78) - PDO(Nov))) + (SOI(Apr) - PNA(Dec))$	1.75	0.60
2017	$((TP \cdot (7.75)) + ((1.78) - PNA(Dec))) + PDO(Nov)$	1.55	0.57
2018	$(TP \cdot ((10.55) + PDO(Nov))) + ((-1.30) \cdot PNA(Dec))$	1.20	0.76
2019	$(TP \cdot ((11.62) + PDO(Nov))) - (PNA(Dec) + SL)$	1.19	0.73
2020	$(TP \cdot ((10.13) + PDO(Nov))) + ((0.36) - PNA(Dec))$	1.10	0.76

The potential predictors selected by GA-chem-clim are: El Niño–Southern Oscillation (Southern Oscillation Index): ENSO (SOI) (April current year), Pacific Decadal Oscillation: PDO (November previous year) and Pacific–North American teleconnection pattern: PNA (December previous year) and TP (March–May). The scaling of chemical loading data is as follows: CL: 2×10^5 , TN: 4×10^4 , SD: 8×10^4 , SL: 3×10^5 , SRP: 7×10^2 , TKN: 1.2×10^4 , TP: 2.5×10^3 , and TSS: 1×10^6 . All the chemical loading data is in metric ton.

spanning several months can outperform local climatic factors indicating the importance of climatic influences on population ecology²⁵.

In this study, we develop a machine learning-based predictive model of the Lake Erie SI. Building on previous approaches (e.g., ref. 26) that ingest observed nutrient information only, we modify this approach by including a set of physically-based, easily-available, large-scale climate indices as potential predictors for the seasonal prediction of HABs. We believe that the present study is the first to use a machine learning-based approach that uses climate indices in combination with nutrient loading to predict the seasonal SI of HABs.

Results

Our predictive model makes use of the peak three 10-day periods of bloom biomass in Lake Erie to estimate the SI⁹ and is available from <https://www.weather.gov/cle/LakeErieHAB>. The SI is a non-dimensional positive number for each season (Supplementary Fig. S1, blue line). During the observed period of 2002–2020, the highest SI (10.5) was observed in 2015, and the lowest SI (0.3) was in 2002 and 2005.

For the seasonal prediction of HABs in Lake Erie, we use Genetic Algorithm (GA, see details in “Materials and methods”) models. We used a jackknife procedure to assess the uncertainty of the predictors, as well as the predictions. For this, we used 18 years of data for training the model and one leftout year for cross-validation. This process was done repeatedly. Thus, at the end of the jackknife method, we have 19 models and 19 cross-validations. In the GA-chem model that is based on nutrient loading only (Supplementary Table S1), a total of eight parameters (total phosphorus (TP), SRP, total Kjeldahl nitrogen (TKN), total nitrogen ((nitrite + nitrate), TN), chlorides (CL), total suspended solids (TSS), silica dioxide (SiO₂, referred here as SD), and sulfate (SO₄, referred here as SL)) from March to May were used as the model input. Out of these eight parameters, accumulated TP [March–May] was selected by 12 models followed by accumulated SRP [March–May] (selected by 8 models),

accumulated TSS [March–May] (selected by 6 models), accumulated SD [March–May] (selected by 3 models), accumulated SL [March–May] (selected by 2 models), and accumulated CL [March–May] (selected by 1 model). There is little to no data available before March because of low discharge due to freezing conditions.

The GA-chem-clim model that is based on chemical loading and climate indices (Table 1), selected TP in combination with three monthly climate indices, the Pacific–North American (PNA, Wallace and Gutzler²⁷) pattern, El Niño–Southern Oscillation (ENSO) (Southern Oscillation index or SOI, Ropelewski and Jones²⁸), and Pacific Decadal Oscillation (PDO, Mantua et al.²⁹) from the input set of climate indices consisted of Arctic Oscillation Index (AO), PNA, Quasi-biennial Oscillation (QBO), Atlantic Multidecadal Oscillation (AMO), ENSO SOI Index, Niño index (Niño3.4), North Atlantic Oscillation (NAO), and PDO. Out of these potential predictors, ENSO (SOI) index from April, PNA from December, PDO from November, and chemical loadings of TP from March–May turned out to be the most useful predictors.

For the GA-clim model (Supplementary Table S2), which is based on the climate indices only, the prediction is made using the three climate indices ENSO (SOI), PNA, and PDO. Qin et al.³⁰ in their study of extreme climate anomalies enhancing the cyanobacterial blooms in Lake Taihu, China found that a recorded algal bloom in 2017 was connected with the 2016–2017 El Niño, AMO and PDO. They found the persistent warmth during 2016–2017 was related to the warm phases of AMO and PDO. McKibben et al.³¹ showed evidence of climatic regulation of domoic acid in shellfish over the past 20 years in the Northern California Current regime. They found that the timing of elevated domoic acid is strongly related to warm phases of the PDO and the Oceanic Niño Index, an indicator of El Niño events.

Seasonal prediction of HABs severity index using nutrient loading. Predictive modeling for seasonal severity of HABs (HABs SI) using the GA-chem model (Supplementary Table S1)

was developed using the observations of integrated chemical loading (March to May for 2002–2020) for Lake Erie with the aim that the seasonal forecast would be made available in early June.

Using the jackknife method, TP was found to be the dominant predictor by the GA-chem models. The interannual variability of TP selected by the maximum number of GA-chem models as a potential predictor is shown in Supplementary Fig. S1. A systematic relationship between the HABs SI and TP is evidenced by a high correlation between them (0.59), which is consistent with other studies⁹. Based on the predictors selected by the GA-chem model, the monthly variation in the most dominant predictors TP and SRP was examined (Supplementary Fig. S2). We found that the correlation between TP loading and HABs SI was 0.31 in March, 0.43 in April, and 0.37 in May. The correlation between HABs SI and SRP was 0.41 in March, 0.35 in April, and 0.34 in May (Supplementary Table S3). Apart from these two dominant predictors, we found SD and TSS as the two other potential predictors. In our GA-chem experiments, finding SD as a potential predictor was a novel finding (to the best of our knowledge) given that while past studies have found a connection between SD and algal blooms^{32,33}, these were primarily diatom blooms. Previous work on diatom blooms in Lake Erie (e.g., ref. ³⁴) found that rapid sedimentation of diatoms may contribute to hypoxic zones near the lake bottom. Similar hypoxic zones have been associated with increased internal loading of phosphorus in Lake Erie^{35–37}. Given the importance of SD as a predictor by the GA-chem model, it is possible that increase in SD may lead to an indirect increase in cyanobacteria biomass via internal loading of phosphorus resulting from increased diatom sedimentation rates. TSS is the most visible indicator of water quality and can come from soil erosion, runoff, discharges, disturbed bottom sediments and algae. As such, finding TSS in some of the predictive models is not a surprise. There are two other predictors, SL and CL, which are not discussed here since they are selected by only two models (SL) and one model (CL) in the jackknife method.

Based on the two potential predictors (TP and SRP) and four other predictors (TSS, SL, CL, and SD) selected by some GA-chem predictive models, HABs SI simulation (Fig. 1b) was successful for the extreme year of 2011 (with an overprediction in HABs SI by 1.12). For another extreme year 2015, the GA-chem model missed the prediction of HABs SI by 7.8 units. It was found (<https://www.washingtonpost.com/news/capital-weather-gang/wp/2015/11/12/this-years-disgusting-green-algal-bloom-in-lake-erie-was-the-most-severe-on-record/>) that Maumee River watershed received about eight inches more rain than normal in June 2015. It was also the fourth wettest June in Toledo, Ohio, and one of the top 20 wettest months since records began in 1880. As a result, our GA-chem model that lacked nutrient loading data from June, may have performed poorly. Therefore, to test the impact of nutrient loading from June, a GA-chemplus model was developed, which is similar to GA-chem but makes use of chemical loading data from March to June (instead of March to May as in GA-chem). Using the GA-chemplus, the prediction of HABs SI was improved considerably in 2015 compared to GA-chem (Observed value: 10.5, Predicted value using GA-chemplus: 10.28, predicted values using GA-chem: 2.70). For the entire time period (2002–2020), the root mean squared error (RMSE) of predicted HABs SI using GA-chem is 2.67 and the correlation with the HABs SI is 0.53 (Table 2). However, for years with HABs SI > 7, the performance of GA-chem deteriorates as evidenced by higher RMSE (4.02). The standard deviation of the observed HABs SI is 3.17, compared to the standard deviation for the GA-chem model (2.74). To further improve the prediction of SI, we look at predictors other than nutrients and discharge data.

Table 2 Performance of GA-chem (based on chemical loading data for March–May), GA-clim (based on climate indices data for November (previous year)–May (current year)), and GA-chem-clim (based on chemical loading data for March–May, and climate indices data for November–May) against observed HABs severity index during 2002–2020.

Experiment	RMSE	Correlation	Standard deviation of bias
	All events	All events	All events
GA-chem	2.67	0.53	2.74
GA-clim	2.52	0.58	2.59
GA-chem-clim	2.26	0.67	2.32

Seasonal prediction using nutrient plus climate indices. Large-scale circulation patterns are known to have an important association with the hydrometeorological conditions for the Great Lakes, and hence, may be considered as potential predictors for HABs. ENSO is the dominant climate mode on the planet³⁸, and its associated teleconnection pattern, i.e., the PNA teleconnection^{27,39}, provides the greatest potential for seasonal prediction of hydrometeorology over the Great Lakes^{40–42}. Similar to GA-chem, a set of predictive models, using the jackknife method, was developed using climate indices only (GA-clim, Supplementary Table S2). We found that PDO, PNA and ENSO (SOI) were the most important predictors of all the climate indices included in this analysis. The performance of GA-clim for all events is comparable to GA-chem (RMSE of GA-chem: 2.67, GA-clim: 2.52) and R^2 for each predictive model is in the range 0.29–0.69 for GA-chem and 0.49–0.63 for GA-clim (Supplementary Tables S1, S2).

Notaro et al.⁴¹ examined the relationship between large-scale circulation and the regional climate of Northeast United States for early winter using regional climate simulations and observations. They found that during positive PNA, there is greater subsidence over land, lower atmospheric moisture content and higher stability, resulting in reduced total simulated cloud cover, precipitation and snowfall. Also, during positive PNA, lower surface air temperature over land resulted in higher simulated cloud cover over the Great Lakes and the Atlantic Ocean due to increased thermal contrast. The time-series of ENSO (SOI) and PNA (Supplementary Fig. S3) indicate that they were mostly in-phase in some years, but were out-of-phase in other years. Thus, no linear relationship of these indices with the severity of HABs over Lake Erie could be established. This finding emphasizes the importance of including climate indices in machine learning models to capture such non-linearity.

The improvement in prediction in GA-chem-clim compared to GA-chem is evident in Fig. 1b. The RMSE in GA-chem-clim is reduced to 2.26 from 2.67 as found in GA-chem (Table 2). At the same time, the correlation between observed and simulated SI increased from 0.53 in GA-chem to 0.67 in GA-chem-clim. Moreover, the observed standard deviation of SI was captured well by the GA-chem-clim compared to GA-chem (Table 2) indicating that GA-chem-clim can represent the observed variation in HABs better than GA-chem. Considering years with HABs SI > 7, GA-chem-clim also performs better than GA-chem (Table 2), and is particularly evident in 2017 (Fig. 1b; Obs:8, GA-chem:5.65, GA-chem-clim:8.29). This improvement is notable because one of the major problems mentioned in Stumpf et al.⁹ model was the prediction in years with high HABs SI.

We further investigated large-scale atmospheric conditions to understand the role of climate variability (PNA and ENSO) and their connection to HABs. For this purpose, we analyzed the

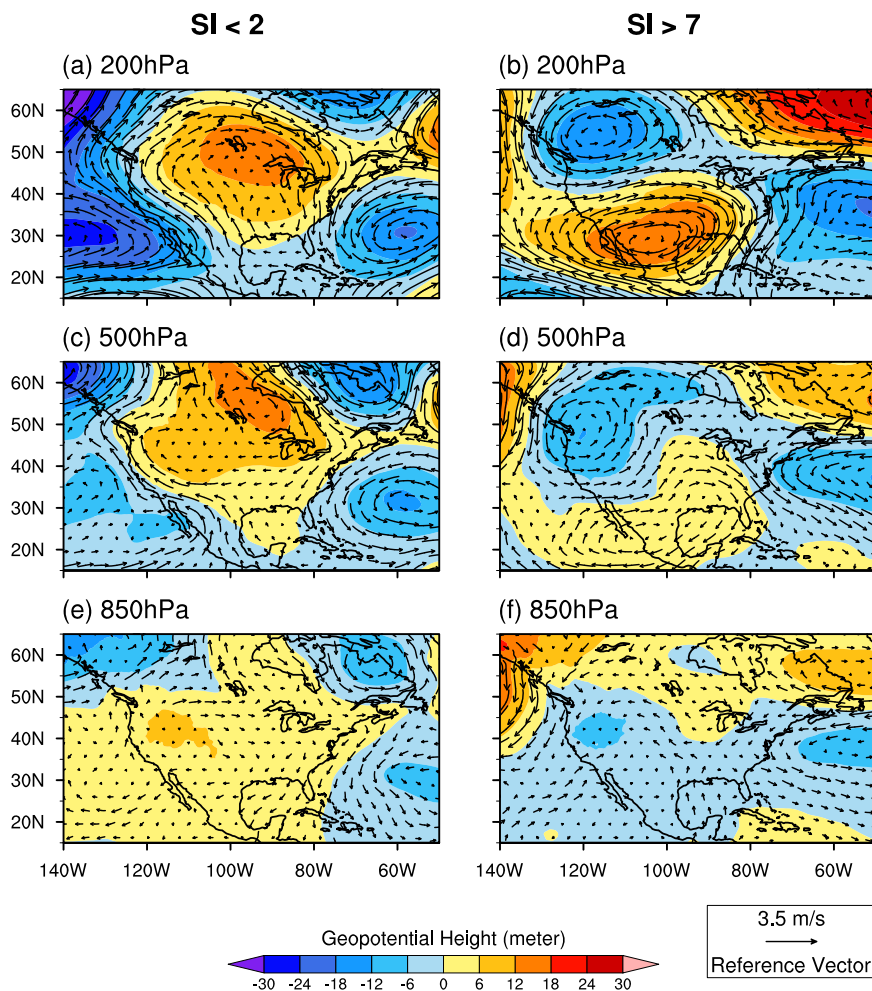


Fig. 2 Geopotential height and winds during years with mild and severe bloom activity. **a** Anomalies of geopotential height (shaded, m) and winds (vector, m s^{-1}) for December, January and February for years with HABs SI < 2 at 200 hPa. **b** Same as **(a)** but for HABs SI > 7. **c** Same as **(a)** but for 500 hPa. **d** Same as **(c)** but for HABs SI > 7. **e** Same as **(a)** but for 850 hPa. **f** Same as **(e)** but for HABs SI > 7. Anomalies were calculated after subtracting the climatology based on 2002–2020.

large-scale circulations based on data obtained from the European Center for Medium-Range Weather Forecasts (ECMWF) Reanalysis 5 (ERA5, Hersbach et al.⁴³) for years with low SI (<2; 2002, 2005, 2006, and 2007) and high SI (>7; 2011, 2013, 2015, 2017, and 2019). The reasoning behind this categorization is to focus on the role of nutrient loading and atmospheric circulations for these extreme low and high HABs SI years.

The composites of large-scale circulation patterns from the winter months show distinct differences that lead to years with low (SI < 2) and high (SI > 7) severity of blooms. Here we analyze the geopotential height—the height of a pressure surface (e.g., 500 hPa) above the mean sea level—which is closely related to wind anomalies due to the geostrophic balance and is often used to identify atmospheric teleconnection patterns. We find higher geopotential heights at 500 hPa over the northern U.S. (Fig. 2) during low severity years, which result in less cold air outbreaks in the winter months and, thus, higher temperature. Meanwhile, the high-pressure anomaly center also tends to shift the jet stream southward, which leads to less snowfall over the Great Lake areas. The higher geopotential height anomalies over the U.S. can be traced back to the tropical Pacific as the positive PNA wave train, with low-pressure anomalies found over the Aleutian Islands region and high-pressure anomalies over the subtropical North Pacific. A positive PNA is typically observed during El Niño, agreeing with our findings of a prominent positive sea surface

temperature (SST) anomaly in the central tropical Pacific (Fig. 3). In addition, the positive phase of PDO can also induce positive PNA teleconnection pattern^{44,45}. Conversely, lower geopotential heights are found during winter months prior to high severity blooms, which allowed greater intrusions of Arctic air mass during winters of preceding years with a high HAB severity index. Over the Pacific Ocean, high pressure anomalies weaken the Aleutian Low while low pressure anomalies are found over and to the west of Hawaiian Islands, accompanied by prominent negative SST anomalies in the central equatorial Pacific. These large-scale climate conditions are associated with the negative PNA and La Niña, which lead to cooler temperature and greater snowfall over the U.S. (Fig. 4) in the winter resulting in more runoff in the summer, and potentially stronger bloom seasons.

Prediction of multiple years using nutrient loading and climate indices. In the previous sections, we used maximum available data length for training and a left out year for prediction. To further test the capability of the GA models (GA-chem, GA-clim and GA-chem-clim) and the current approach, we reduced the data length for training i.e., 79%, (15 years) of the data used for training and 21% (4 years) of data for prediction and performed 4 set of experiments. In designing these experiments, at least one extrema is part of the prediction phase (see Table 3). Monthly

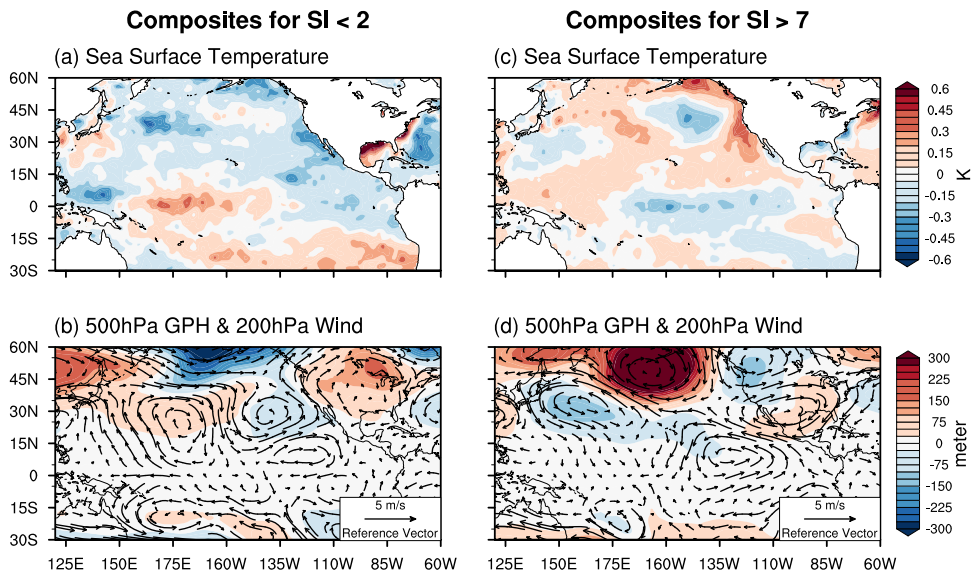


Fig. 3 Oceanic and atmospheric conditions during years with mild and severe bloom activity. **a** Anomalies of SST ($^{\circ}\text{C}$) for December, January, and February for years with HABs $\text{SI} < 2$. **b** Anomalies of 500-hPa geopotential height (shaded, m) and 200-hPa wind (vector, m s^{-1}) for December, January, and February for years with HABs $\text{SI} < 2$. **c** Same as **(a)** but for HABs $\text{SI} > 7$. **d** Same as **(b)** but for HABs $\text{SI} > 7$. SSTs were obtained from HadISST (Rayner et al.⁵⁴), and geopotential height and winds were obtained from the ERA5 reanalysis. Anomalies were calculated after subtracting the climatology based on 2002–2020.

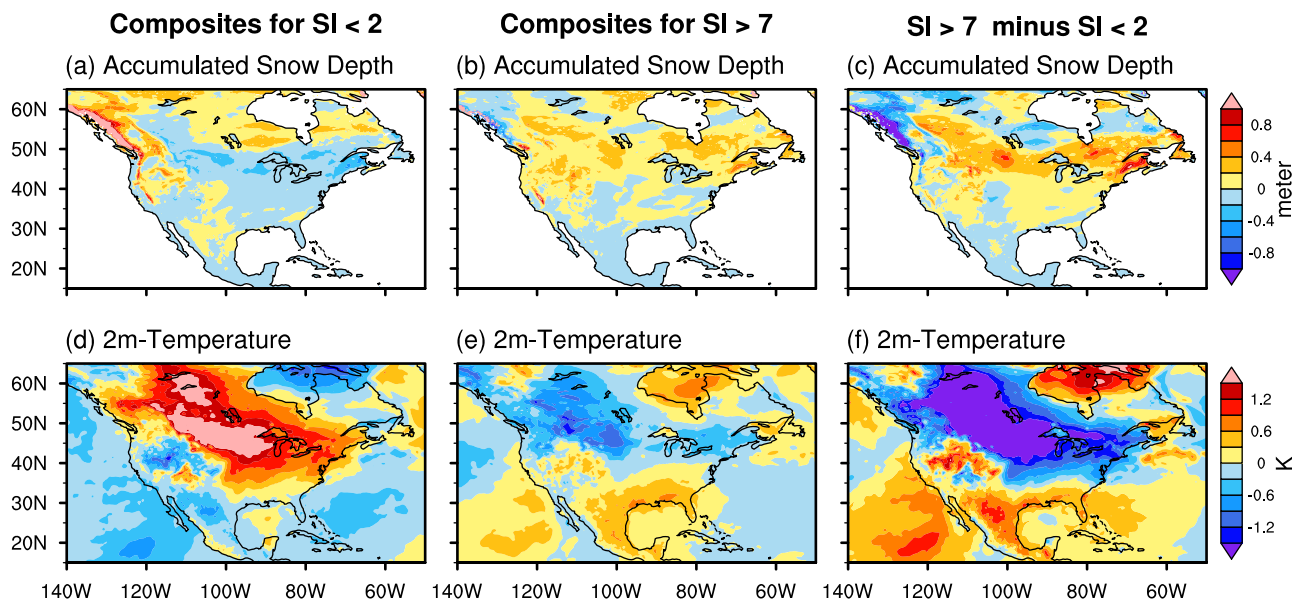


Fig. 4 Accumulated snow depth and near-surface temperature during years with mild and severe bloom activity. **a** Anomalies of accumulated snow depth (m) for years with HABs $\text{SI} < 2$. **b** Same as **(a)** but HABs $\text{SI} > 7$. **c** Anomalies of accumulated snow depth (m) for HABs $\text{SI} > 7$ minus HABs $\text{SI} < 2$. **d** Same as **(a)** but for 2 m temperature (K). **e** Same as **(b)** but for 2 m temperature (K). **f** Same as **(c)** but for 2 m temperature (K). These anomalies for December, January, and February were calculated after subtracting the climatology based on 2002–2020.

Table 3 Root mean squared error (RMSE) and mean bias between observations and HABs severity index for GA-chem and GA-chem-clim models for experiments 1 to 4.

Experiments	Years (predicted)	RMSE GA-chem/GA-clim/GA-chem-clim	Mean Bias GA-chem/GA-clim/GA-chem-clim
Experiment-1	2004, 2006, 2012, 2016	1.12/2.27/0.35	1.10/1.56/−0.20
Experiment-2	2004, 2007, 2008, 2012	2.25/2.54/1.15	1.60/2.20/−0.19
Experiment-3	2013, 2016, 2018, 2020	2.28/1.93/1.70	−0.28/1.70/−0.39
Experiment-4	2015, 2016, 2018, 2020	2.92/3.44/1.66	−1.48/0.63/−0.67

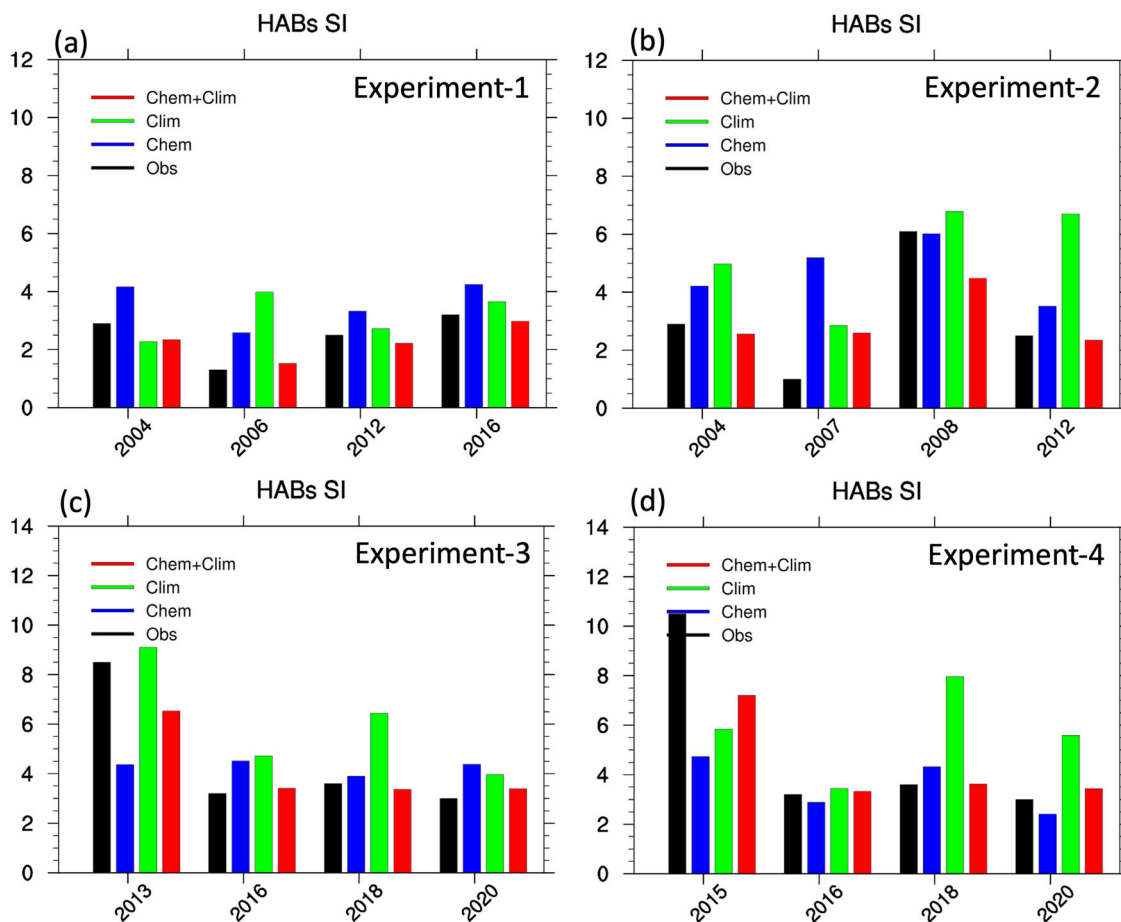


Fig. 5 Prediction of HABs severity index using different input parameters. **a** HABs severity index (SI) for Experiment-1 from observation (black bar), prediction using GA-chem model that is based on nutrient loading only (blue bar), prediction using GA-clim model that is based on climate indices only (green bar) and GA-chem-clim model that is based on nutrient loading and climate indices (red bar). **b** Same as (a) but for Experiment-2. **c** Same as (a) but for Experiment-3. **d** Same as (a) but for Experiment-4. The years on x-axis show the predicted years for each of these four experiments.

Table 4 HABs SI functions for GA-chem, GA-clim and GA-chem-clim for the 4 experiments.

Experiments	HABs SI function (GA-chem)	HABs SI function (GA-clim)	HABs SI function (GA-chem-clim)
Experiment-1	$(CL[Mar] + ((TP[May] - 1.61) * (TS[May] - CL[Mar])) * 0.23))$	$5.60 - 1.95 * PDO[Nov] * SOI[May] - PNA[Nov] * NAO[Dec]$	$((CL[Mar] - CL[Apr]) - (PNA[Dec] - (TP[Apr] + TP[May]))) - TS[May]$
Experiment-2	$(TS[May] + ((0.31) * (CL[Mar] - TS[May])) * (4.68) - TS[May]))$	$((4.54 - PNA[Dec]) + (PDO[Dec] + (SOI[Dec] - (NAO[Mar] * SOI[May])))$	$((TP[Apr] - TS[Mar]) * (0.16) * TP[May]) + (CL[Mar] - PNA[Dec])$
Experiment-3	$(CL[Mar] + (((0.23) * (CL[Mar] - TP[May])) * (CL[May] - TP[May])))$	$5.14 + PDO[Nov] - PNA[Dec] - SOI[May] * (NAO[Mar] - SOI[Nov])$	$(CL[Mar] - (PNA[Dec] + ((TS[May] - PNA[Dec]) * ((TP[Apr] - TS[Mar]) * (-0.19))))$
Experiment-4	$(TP[May] - ((-0.58) * (TP[Apr] + ((CL[Mar] - CL[May]) - TS[May])))$	$(PDO[Nov] - (PNA[Dec] - (4.80 + (SOI[Apr] - (NAO[May] * NAO[Mar])))$	$((CL[Mar] - PNA[Dec]) + ((0.12) * (TP[Apr] - TS[Mar])) * TP[May])$

The provided equations are sample equations for the prediction of 2006 (experiment-1), 2008 (experiment-2), 2013 (experiment-3), and 2014 (experiment-4).

accumulations of chemical loading data and monthly climate indices are used for GA-chem and GA-chem-clim models. In all four experiments (Fig. 5), GA-chem overpredicts the low extrema years and underpredicts the high extrema years. GA-chem-clim model improves the prediction compared to GA-chem model as shown in Table 3 by reduction of RMSE for GA-chem-clim models. In Table 4, we show the HABs SI functions for Experiments 1–4. Apart from PNA for December month, we found TP for April and May, TSS for March, and CL for the months of March and April as predictors for HABs SI. The priority of

monthly chloride loading given by the GA-chem and GA-chem-clim models in experiments 1–4 has important implications for phytoplankton dynamics in Lake Erie. There is strong evidence that large inputs of road de-icing salt (via spring runoff) can disrupt lake phytoplankton dynamics⁴⁶. Some species of freshwater cyanobacteria, including *Microcystis aeruginosa*, have demonstrated high tolerances to elevated salinity, and increased chloride concentrations have been associated with cyanobacterial blooms⁴⁷. The importance of chloride loading in March–May as a predictor may also serve as an indirect factor, as road salt

application rates may function as a proxy measure for the severity of a previous season's winter (as shown by the GA models in selecting PNA for the month of December).

Although there is an overall improvement in GA-chem-clim model compared to GA-chem (Figs. 1b, 5 and Tables 2, 3), there may be other environmental predictors which can add value to both these models. To test this hypothesis, we incorporated local data (2 m-temperature and 10 m wind speed) at five locations surrounding the Lake Erie (Toledo (83.4764°W, 41.5631°N), Cleveland (81.8528°W, 41.4050°N), Buffalo (78.7358°W, 42.9408°N), Delhi (80.5453°W, 42.8635°N), and Niagara District (79.1717°W, 43.1917°N)) in the GA models. The inclusion of wind speed (early April to mid-May) and temperature (early March to late April) in the GA models showed that they can play important roles in the prediction of HABs SI. Since the results are still preliminary and only used local data over 5 stations and two variables (temperature and wind speed), and need more experiments and further investigation, we do not report results from these experiments here. Also, in this study, due to the small size of the dataset (2002–2020), the number of predictors is limited to avoid overfitting⁴⁸. We hope to design further experiments in future to select the predictors from chemical, climate and local parameters in order to improve the prediction of HABs SI.

Concluding remarks. The Lake Erie seasonal severity index was successfully predicted using a GA model with predictors selected from two types of data, chemical data and climate indices. Despite limited availability (19 years) of the bloom severity data, our GA-chem-clim model provides a more accurate prediction than GA-chem as seen by lower RMSE and higher correlation in GA-chem-clim than GA-chem model (Tables 2, 3 and Figs. 1b, 5). Out of the five years with HABs SI > 7, the prediction for 2013 was the worst with a negative bias of ~3.5 units by both GA-chem and GA-chem-clim. The comparison of performance by GA-chem, GA-clim, and GA-chem-clim (Supplementary Table S4) show residual standard error for GA-chem-clim is lower than GA-chem and GA-clim. The Pearson product-moment correlation and *p*-value further confirms the superior performance of GA-chem-clim model. We further demonstrated that the GA-chem-clim model improves (as shown in Table 3) the performance of GA-chem model by predicting a set of four different years comprising of low and high HABs SI.

To understand the potential role of climate indices for improved HABs prediction, we further analyzed the large-scale atmospheric circulations using reanalysis data. It was found that for the mild HABs years (SI < 2), the large-scale meteorological features are distinctly different from those for the severe HABs years (SI > 7). This suggests that large-scale circulation may affect the seasonal evolution of HABs in Lake Erie. Moreover, the seasonal prediction of HABs can be completed by early June (using GA-chem, GA-clim, and GA-chem-clim), before the expected peak in HAB activity during July to October. This improved early seasonal prediction can provide timely information to policymakers for adopting proper planning and mitigation strategies such as restrictions in harvesting and help in monitoring toxins in shellfish to keep contaminated products off the market. The incorporation of climate indices in the GA-models provides the flexibility of early prediction with a greater lead time and also shows improvement over GA-model which only uses chemical loading.

Overall, the present approach shows promise for early-season prediction of HABs SI in Lake Erie, and it would be interesting to apply the same approach to other water bodies. At present, this GA-model is deterministic. It would be worthwhile to design a probability-based system that can provide the likely range of

HABs severity and their probability of occurrence. Work will be undertaken in the future to include the stochastic framework within our methodology to include these aspects.

Materials and methods

Machine learning techniques like the Genetic Algorithm (GA; refs. 49–52) are search heuristics based on the laws of natural selection, where a large number of randomly-generated mathematical functions are initially assumed to explain the relationship between predictor and predictand variables. New forms of functional relations evolve during an iterative process that involves reproduction and mutation among the most fitting individual functions. After a large number of iterations (typically in thousands), the best individual function emerges.

Predictive modeling for seasonal HABs using the GA was developed using observations of integrated chemical loading (March to May) into Lake Erie as well as the monthly large-scale climate indices seven months prior to June (i.e., from November of the previous year to May of the current year). The small number (19, one event per year for the period 2002–2020) of observations, and the corresponding predictor observations compelled us to constrain the GA process to select only a few (4 to 5) predictors to avoid the so called “curse of dimensionality” and the risk of overfitting⁴⁸. GA procedures were initiated with 2000 size-constrained random functions, which were allowed to reproduce among one another, and mutate at certain random intervals, to produce more and more fit individual functions in an iterative process. The measure of the fitness of a model is defined as how best its predictions match the observations of HABs SI. With each iteration, the GA procedure evolves and refines its functions. The mean of seasonal HABs SI in our limited data set was 4.72 with a standard deviation of 3.17 units. In our model development procedure, we considered only those models, which achieved an accuracy of 2.0 units or less in the GA procedure. The value represents the maximum RMSE from Table 1, Supplementary Tables S1, S2. The GA iterative process continued until the fitness of the best-evolved function reached saturation below the predefined threshold, which took about 4000–6000 iterations.

The paucity of observations of HAB events (only 19 seasons) does not permit the probability distribution of HAB to be represented. Further, a single predictive model may not be sufficient to explain the relationship between predictors and predictands. In order to analyze the bias and the variance of the GA-based prediction, we employed the jackknife method⁵³. Using this method, we developed 19 predictive models, using 18 data points for training, leaving one data point (one season) for prediction. Hence, the mean of the jackknife variance for the 19 models was considered the variance of the GA model. With the availability of more observations in the future, this procedure may be repeated to obtain improved predictive models.

Supplementary Tables S1, S2 show the performance, and the predictors selected by each of the 19 models developed by GA-chem and GA-clim, respectively. The mean jackknife standard error of the models was found to be 1.76, which is much smaller compared to the natural variability of the seasonal index record (3.17). The mean strength (equivalent to r^2 in regression modeling) of the models was found to be 0.37. Within the 19 different data sets, the GA procedure picked different predictors. Integrated March-to-May TP, SRP, TSS, CL and SL loading were the most common predictors selected by GA-chem.

In the GA-clim models that used only climate indices data, the major predictors selected were April ENSO (SOI) index, December PNA, and November PDO. The mean strength of the models was found to be 0.57, and the standard error of prediction was 1.61 units.

As mentioned earlier, the combination of climatic indices with chemical loading led to enhancement of model strength (0.63), and reduction in standard error (1.43). This analysis indicates that a suitable combination of ENSO, PNA and other climate indices, and bio-chemical loading may be responsible for the observed variability of HABs over Lake Erie.

Data availability

We have used the nutrient loading data, which was obtained from <https://ncwqr.org/monitoring/>. ERA5 data were downloaded from the Copernicus Data System (<https://cds.climate.copernicus.eu/cdsapp#!/home>). Climate indices datasets were downloaded from Climate Prediction Center, NOAA (<https://www.cpc.ncep.noaa.gov/>) and Physical Sciences Laboratory, NOAA (<https://psl.noaa.gov/>).

Code availability

All the analyses and figures are made using FORTRAN and NCL (<https://doi.org/10.5065/D6WD3XH5>), and the codes are available from the corresponding author on reasonable request.

Received: 26 April 2021; Accepted: 27 July 2022;

Published online: 31 August 2022

References

1. Anderson, D. M., Cembella, A. D. & Hallegraef, G. M. Progress in understanding harmful algal blooms: Paradigm shifts and new technologies for research monitoring and management. *Ann. Rev. Marine Sci.* **4**, 143–176 (2011).
2. McPartlin, D. A. et al. Biosensors for the monitoring of harmful algal blooms. *Curr. Opin. Biotechnol.* **43**, 164–169 (2017).
3. Kudela, R. M. Harmful algal blooms: A scientific summary for policy makers – UNESCO Digital Library. IOC/INF-1320 REV <https://unesdoc.unesco.org/ark:/48223/pf0000233419> (2015).
4. Grannemann, N. G. & Reeves, H. W. Great Lakes basin water availability and use. A study of the national assessment of water availability and use program, USGS, Fact Sheet 2005-3113, pp 1–4. https://pubs.usgs.gov/fs/2005/3113/pdf/FS2005_3113.pdf (2005).
5. Magnuson, J. et al. Potential effects of climate change on aquatic systems: Laurentian Great Lakes and Precambrian Shield region. *Hydrol. Processes* **11**, 825–871 (1997).
6. Leon, L. F. et al. Application of a 3D hydrodynamic-biological model for seasonal and spatial dynamics of water quality and phytoplankton in Lake Erie. *J. Great Lakes Res.* **37**, 41–53 (2011).
7. Vincent, R. K. et al. Phycocyanin detection from LANDSAT TM data for mapping cyanobacterial blooms in Lake Erie. *Remote Sens. Environ.* **89**, 381–392 (2004).
8. Rinta-Kanto, J. M. et al. Lake Erie Microcystis: Relationship between microcystin production, dynamics of genotypes and environmental parameters in a large lake. *Harmful Algae* **8**, 665–673 (2009).
9. Stumpf, R. P., Wynne, T. T., Baker, D. B. & Fahnenstiel, G. L. Interannual variability of cyanobacterial blooms in Lake Erie. *PLoS One* **7**, e24444 (2012).
10. Michalak, A. M. et al. Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions. *Proc. Natl Acad. Sci. USA* **110**, 6448–6452 (2013).
11. Paul, V. J. Global warming and cyanobacterial harmful algal blooms. In: Hudnell, H.K. (ed.), *Cyanobacterial harmful algal blooms: State of the science and research needs. Adv. Exp. Med. Biol.* **619**, 239–257 (2007).
12. Hallegraef, G. M. Ocean climate, phytoplankton community responses, and harmful algal blooms: A formidable predictive challenge. *J. Phycol.* **46**, 220–235 (2010).
13. Fu, F. X., Tatters, A. O. & Hutchins, D. A. Global change and the future of harmful algal blooms in the ocean. *Mar. Ecol. Prog. Ser.* **470**, 207–233 (2012).
14. O'Neil, J. M., Davis, T. W., Burford, M. A. & Gobler, C. J. The rise of harmful cyanobacteria blooms: The potential roles of eutrophication and climate change. *Harmful Algae* **14**, 313–334 (2012).
15. Baez, J. C. et al. The North Atlantic Oscillation and the Arctic Oscillation favour harmful algal blooms in SW Europe. *Harmful Algae* **39**, 121–126 (2014).
16. Philips, E. J. et al. Hurricanes, El Niño and harmful algal blooms in two subtropical Florida estuaries: Direct and indirect impacts. *Sci. Rep.* **10**, 1910 (2010).
17. Belgrano, A., Lindahl, O. & Hernroth, B. North Atlantic Oscillation primary productivity and toxic phytoplankton in the Gullmar Fjord, Sweden (1985–1996). *Proc. R. Soc. Lond. B* **266**, 425–430 (1999).
18. Weyhenmeyer, G. A., Blenckner, T. & Pettersson, K. Changes of the plankton spring outbreak related to the North Atlantic oscillation. *Limnol. Oceanogr.* **44**, 1788–1792 (1999).
19. Irigoien, X. et al. Feeding selectivity and egg production of *Calanus helgolandicus* in the English channel. *Limnol. Oceanogr.* **45**, 44–59 (2000).
20. Edwards, M., Reid, P. & Planque, B. Long-term and regional variability of phytoplankton biomass in the Northeast Atlantic (1960–1995). *CES J. Marine Sci.* **58**, 39–49 (2001).
21. Hinder, S. et al. Changes in marine dinoflagellate and diatom abundance under climate change. *Nat. Clim. Change* **2**, 271–275 (2012).
22. Zhou, Y., Michalak, A. M., Beletsk, D., Rao, Y. R. & Richards, R. P. Record-breaking Lake Erie hypoxia during 2012 drought. *Environ. Sci. Technol.* **49**, 800–807 (2015).
23. Cruz, R. C., Reis Costa, P., Vinga, S., Krippahl, L. & Lopes, M. B. A Review of recent machine learning advances for forecasting Harmful Algal Blooms and shellfish contamination. *J. Mar. Sci. Eng.* **9**, 283 (2021).
24. Hallett, T. et al. Why large-scale climate indices seem to predict ecological processes better than local weather. *Nature* **430**, 71–75 (2004).
25. Stenseth, N. C. et al. Studying climate effects on ecology through the use of climate indices: The North Atlantic Oscillations, El Niño Southern Oscillations and beyond. *Proc. R. Soc. B. Bio. Sci.* **270**, 2087–2096 (2003).
26. Lee, S. & Lee, D. Improved prediction of harmful algal blooms in four major South Korea's rivers using deep learning models. *Int. J. Environ. Res. Public Health* **15**, 1322 (2018).
27. Wallace, J. M. & Gutzler, D. S. Teleconnections in the geopotential height field during the northern hemisphere winter. *Mon. Weather Rev.* **109**, 784–812 (1981).
28. Ropelewski, C. F. & Jones, P. D. An extension of the Tahiti–Darwin Southern Oscillation Index. *Mon. Weather Rev.* **115**, 2161–2165 (1987).
29. Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M. & Francis, R. C. A Pacific interdecadal climate oscillation with impacts on salmon production. *Bull. Amer. Meteor. Soc.* **78**, 1069–1079 (1997).
30. Qin, B. et al. Extreme climate anomalies enhancing cyanobacterial blooms in Eutrophic Lake Taihu, China. *Water Resour. Res.* **57**, e2020WR029371 (2021).
31. McKibben, S. M. et al. Climatic regulation of the neurotoxin domoic acid. *Proc. Natl Acad. Sci. USA* **114**, 239–244 (2017).
32. Egge, J. K. & Aksnes, D. L. Silicate as regulating nutrient in phytoplankton competition. *Mar. Ecol. Prog. Ser.* **83**, 281–292 (1992).
33. Allen, J. et al. Diatom carbon export enhanced by silicate upwelling in the northeast Atlantic. *Nature* **437**, 728–732 (2005).
34. Lashaway, A. R. & Carrick, H. J. Effects of light, temperature and habitat quality on meroplanktonic diatom rejuvenation in Lake Erie: Implications for seasonal hypoxia. *J. Plankton Res.* **32**, 479–490 (2010).
35. Nürnberg, G. K. et al. Evidence for internal phosphorus loading, hypoxia and effects on phytoplankton in partially polymictic Lake Simcoe, Ontario. *J. Great Lakes Res.* **39**, 259–270 (2013).
36. Paytan, A. et al. Internal loading of phosphate in Lake Erie Central Basin. *Sci. Total Environ.* **579**, 1356–1365 (2017).
37. Nürnberg, G. K., T. Howell, T. & Palmer, M. Long-term impact of Central Basin hypoxia and internal phosphorus loading on north shore water quality in Lake Erie. *Inland Waters* **9**, 362–373 (2019).
38. McPhaden, M. J., Zebiak, S. E. & Glantz, M. H. ENSO as an integrating concept in earth science. *Science* **314**, 1740–1745 (2006).
39. Hoskins, B. J. & Karoly, D. J. The steady linear response of a spherical atmosphere to thermal and orographic forcing. *J. Atmos. Sci.* **38**, 1179–1196 (1981).
40. Rodionov, S. N. & Assel, R. Atmospheric teleconnection patterns and severity of winters in the Laurentian Great Lakes Basin. *Atmosphere-Ocean* **38**, 601–635 (2000).
41. Notaro, M., Wang, W.-C. & Gong, W. Model and observational analysis of the Northeast U.S. regional climate and its relationship to the PNA and NAO patterns during early winter. *Mon. Wea. Rev.* **134**, 3479–3505 (2006).
42. Wang, J. et al. Temporal and spatial variability of Great Lakes ice cover, 1973–2010. *J. Climate* **25**, 1318–1329 (2012).
43. Hersbach, H., Coauthors. The ERA5 Global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
44. Alexander, M. Extratropical Air-Sea Interaction, Sea Surface Temperature Variability, and the Pacific Decadal Oscillation. In *Climate Dynamics: Why Does Climate Vary?* (eds Sun, D.-Z. & Bryan, F.) 123–148 <https://doi.org/10.1029/2008GM000794> (2010).
45. Newman, M. et al. The Pacific decadal oscillation, revisited. *J. Clim.* **29**, 4399–4427 (2016).
46. Szklarek, S., Górecka, A. & Wojtal-Frankiewicz, A. The effects of road salt on freshwater ecosystems and solutions for mitigating chloride pollution—A review. *Sci Total Environ.* **805**, 150289 (2022).
47. Tonk, L., Bosch, K., Visser, P. M. & Huisman, J. Salt tolerance of the harmful cyanobacterium *Microcystis aeruginosa*. *AME* **46**, 117–123 (2007).
48. Lever, J., Krzywinski, M. & Altman, N. Model selection and overfitting. *Nat Methods* **13**, 703–704 (2016).
49. Szpiro, G. G. Forecasting chaotic time series with genetic algorithms. *Phys. Rev. E* **55**, 2557–2568 (1997).
50. Alvarez, A., Lopez, C., Riera, M., Hernandez-Garcia, E. & Tintore, J. Forecasting the SST space-time variability of the Alboran Sea with genetic algorithms. *Geophys. Res. Lett.* **27**, 2709–2712 (2000).
51. Alvarez, A., Orfila, A. & Tintore, J. DARWIN, an evolutionary program for nonlinear modeling of chaotic time series. *Comput. Phys. Commun.* **136**, 334–349 (2001).
52. Kishtawal, C. M., Basu, S., Patadia, F. & Thapliyal, P. K. Forecasting summer rainfall over India using genetic algorithm. *Geophys. Res. Lett.* **30**, 1–5 (2003).
53. Efron, B. & Gong, G. A. Leisurely look at the bootstrap, the jackknife and cross-validation. *Am. Stat.* **37**, 36–48 (1983).
54. Rayner, N. A. et al. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.* **108**, 4407 (2003).

Acknowledgements

We thank two anonymous reviewers and the Editor for constructive comments and suggestions that improved the manuscript. We would also like to thank Dr. Nathan Manning and Dr. Laura Johnson from Heidelberg University for sharing the information on nutrient loading data. We would like to thank Dr. R. Stumpf (NOAA) and Dr. K. Kavanaugh (NOAA) for sharing the HABs severity data and their research papers, which were helpful in understanding the challenges related to the seasonal prediction of HABs. We acknowledge the use of remotely-sensed imagery from the European Space Agency. This research was supported by The Jefferson Project at Lake George, a partnership

among IBM, Rensselaer Polytechnic Institute and the FUND for Lake George. There are no external funding sources for this research.

Author contributions

M.T. and C.M.K. formulated the problem, designed and conducted the experiments, and wrote the manuscript; V.M. contributed to the manuscript and analysis of results; P.R., T.S., and L.Z. contributed to the analysis and writing, K.T. contributed to the analysis, L.T. contributed towards the refinement of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s43247-022-00510-w>.

Correspondence and requests for materials should be addressed to Mukul Tewari.

Peer review information *Communications Earth & Environment* thanks Bryony L. Townhill and the other anonymous reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Edmond Sanganyado and Clare Davis.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022