Water Research 232 (2023) 119710

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

Short-term Lake Erie algal bloom prediction by classification and regression models

Haiping Ai, Kai Zhang, Jiachun Sun, Huichun Zhang *

Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, OH 44106, United States

ARTICLE INFO

Keywords:
Bloom forecast
Feature selection
Long-short term memory
Machine learning
Random forest
Time series modeling

ABSTRACT

The recent outbreaks of harmful algal blooms in the western Lake Erie Basin (WLEB) have drawn tremendous attention to bloom prediction for better control and management. Many weekly to annual bloom prediction models have been reported, but they only employ small datasets, have limited types of input features, build linear regression or probabilistic models, or require complex process-based computations. To address these limitations, we conducted a comprehensive literature review, complied a large dataset containing chlorophyll-a index (from 2002 to 2019) as the output and a novel combination of riverine (the Maumee & Detroit Rivers) and meteorological (WLEB) features as the input, and built machine learning-based classification and regression models for 10-d scale bloom predictions. By analyzing the feature importance, we identified 8 most important features for the HAB control, including nitrogen loads, time, water levels, soluble reactive phosphorus load, and solar irradiance. Here, both long- and short-term nitrogen loads were for the first time considered in HAB models for Lake Erie. Based on these features, the 2-, 3-, and 4-level random forest classification models achieved an accuracy of 89.6%, 77.0%, and 66.7%, respectively, and the regression model achieved an R^2 value of 0.69. In addition, long-short term memory (LSTM) was implemented to predict temporal trends of four short-term features (N, solar irradiance, and two water levels) and achieved the Nash-Sutcliffe efficiency of 0.12–0.97. Feeding the LSTM model predictions for these features into the 2-level classification model reached an accuracy of 86.0% for predicting the HABs in 2017–2018, suggesting that we can provide short-term HAB forecasts even when the feature values are not available.

1. Introduction

In the 1960s, harmful algal blooms (HABs) frequently occurred in the western Lake Erie basin (WLEB), but disappeared in the late 1970s due to effective phosphorus control measures (Commission, 1972; DePinto et al., 1986). However, following occasional occurrences in the 1990s, HABs have started to occur annually since 2002 and reached record levels in recent years (Stumpf et al., 2016). HABs can not only release excess toxins, which can poison aquatic lives and endanger human health, but also cause oxygen depletion and turn the water body into a dead zone (USEPA, 2021). Because the wellness or deterioration of Lake Erie affects millions of residents in the whole Lake Erie watershed, timely and accurate HAB forecasts will prepare the stakeholders for upcoming HABs and prevent the public from close contact with the poisoned water.

So far, there are a number of reported predictive models that can provide HAB forecasts for the WLEB. Most of these models are statistical models that link the extent of monitored bloom biomass to various physicochemical, meteorological, and hydrodynamic variables. Briefly, the employed statistical functions range from single- (Stumpf et al., 2016, 2012) or multiple-linear regression (Ho and Michalak, 2017) and probabilistic models (Bertani et al., 2016; Obenour et al., 2014) to a spatio-temporal geostatistical model (Fang et al., 2019). The involved variables include day of year, long-term cumulative soluble reactive phosphorus (SRP), Maumee River discharge, spatial coordinates, total spring phosphorus load (TP), total spring bioavailable phosphorus (TBP) load, water column depth, and/or wind speed (Bertani et al., 2016; Fang et al., 2019; Ho and Michalak, 2017; Obenour et al., 2014; Stumpf et al., 2016). In addition to those data-driven statistical models, there are several process-based hydrodynamic models (Verhamme et al., 2016; Wynne et al., 2013, 2011), with the latest one being three dimensional at a fine scale and incorporating hydrodynamics, nutrient, sediment dynamics, and vertical mixing to track the HAB occurrence (Verhamme et al., 2016). In all the above studies, the bloom biomass is quantified by

* Corresponding author.
E-mail address: hjz13@case.edu (H. Zhang).

https://doi.org/10.1016/j.watres.2023.119710
Received 27 September 2022; Received in revised form 31 January 2023; Accepted 4 February 2023
Available online 5 February 2023

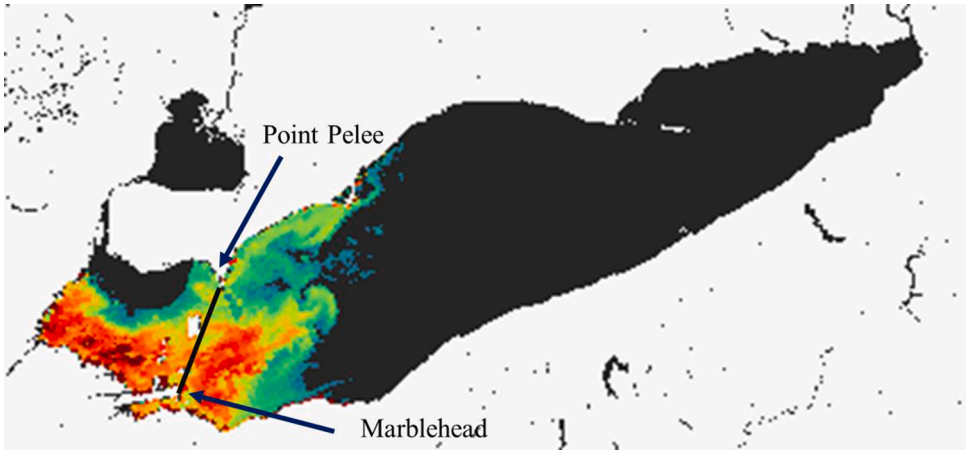

Fig. 1. A satellite image of Lake Erie with a snapshot of the maximum bloom extent monitored by NOAA between Aug 10–19, 2015. The lake area west to the Point Pelee-Marblehead line was denoted as the WLEB. The cool to warm color represents the severity of HABs from light to severe.

either remote sensing (Sayers et al., 2019; Stumpf et al., 2012) or in situ sampling (Bridgeman et al., 2013; Fang et al., 2019; Millie et al., 2014). Please see Table S1 in the supplementary material (SM) for a list of predictive models reported for the WLEB since 2010.

Despite the significant progress in the HAB modeling for Lake Erie, there are seven major limitations in the reported models. First, although nutrient—particularly spring P load—are known to be key parameters influencing HABs in Lake Erie (Stumpf et al., 2016, 2012), other temporal meteorological data, such as solar irradiance, wind speed, water temperature, and high flows which carry low levels of nutrient from the Detroit River into the WLEB, are also reportedly affecting the bloom growth in summer (Bertani et al., 2017; Michalak et al., 2013). However, only partial variables are considered in the available models for the WLEB. For example, some of these models only rely on the P load from the Maumee River to predict the annual HAB levels (Ho and Michalak, 2017; Obenour et al., 2014; Stumpf et al., 2016). The hydrodynamic model used by the National Oceanic and Atmospheric Administration (NOAA) (NOAA, 2021) is mainly based on physical transport processes rather than biological mechanisms to forecast the location and intensity of the bloom twice a week (Rowe et al., 2016; Wynne et al., 2011). Moreover, N is known to be limited in summer in the WLEB (Chaffin et al., 2013, 2014), but the amount of regenerated N in the water column can be thousand times more than the corresponding external N load in late August (Hampel et al., 2019). Therefore, legacy N or regenerated N may be an important nutrient source for HABs. Yet, no available models have considered any N-related features in the models. Models that do not consider all potentially influential variables may not accurately predict the bloom extent or reflect the temporal trend in the HABs, even though many parameters have been considered.

Second, physical transport models are computationally expensive, as they require satellite imaging, in-situ bloom biomass sampling, well defined biochemical/physiological parameters, and extensive waterflow and wind information, which are not easy to use for most users (Li et al., 2021; Wynne et al., 2011). Third, models that not only have good prediction accuracy but also provide a deep understanding of the operating HAB mechanisms are desirable for proper HAB control and management. Forth, the statistical functions in the data-driven models may not capture the nonlinear, interactive, and dynamic (even stochastic) nature of the HABs (Millie et al., 2014), like linear regression models, which can only predict the annual maximum HAB extent in Lake Erie (Stumpf et al., 2016) (additional examples in Table S1). Fifth, running process-based models often requires expert knowledge across multiple disciplines because the models involve a large number of parameters that need to be calibrated (Verhamme et al., 2016). Sixth, many of these models are based on very small sample sizes, with the data volume between 10 and 67 (Bertani et al., 2017), see Table S1. Finally, most of the reported statistical models are annual models that provide seasonal forecasts, such as the annual NOAA forecast (NOAA, 2021), but do not issue short-term forecasts to the public (Ho and Michalak, 2017; Obenour et al., 2014).

Apart from the above models for the WLEB, machine learning (ML) models have been successfully employed in bloom predictions for many lakes and rivers (Rousso et al., 2020). For example, a gradient boosting binary classification model integrated hydrological parameters, water temperature and water quality at the 10-d scale for HAB prediction for the Han River, and achieved a high accuracy of 0.9 based on the Kappa coefficient (Xia et al., 2020). Pyo et al. combined riverine, atmospheric, and environmental variables with a convolutional neural network for bloom prediction for the Nakdong River, and achieved a NSE of 0.87 (Pyo et al., 2020). When random forest (RF) was employed together with 27 environmental variables at a time scale of 4 h, the model achieved an R^2 of 0.89 for the prediction of the phytoplankton community in Lake Greifensee (Thomas et al., 2018). However, these models also did not consider all potentially influential features and did not screen for the most essential features, that is, comprehensive feature engineering. Given the powerful predictive ability of ML models, it is important to improve the feature engineering and develop ML models for bloom predictions for the WLEB. Yet, only a couple of ML models have been employed to model HABs in the WLEB and achieved some success, that is, a boosted regression tree model to compare the HAB estimates based on 8 different monitoring programs but the R^2 was only 33.1%−49.8% (Bertani et al., 2017), and artificial neural network (ANN) models for total phytoplankton and *Microcystis* biomass but the training data were only from 2009 to 2011 (Millie et al., 2014).

In this study, we first conducted a comprehensive literature review and collected data for all influencing features that had been used/discussed in previous studies on Lake Erie since 2010 (Michalak et al., 2013; Rousso et al., 2020; Sayers et al., 2019; Stumpf et al., 2016), as summarized in Table S1. Based on the available data, we then developed RF classification and regression models for short-term (10-d scale) forecast of HABs in the WELB. During the model development, careful feature selection was conducted to exclude redundant information and to identify the most important features, which is necessary for efficient and accurate HAB predictions and targeted controls. We next evaluated the effects of the identified important features, such as total nitrogen (TN) load and water levels, which have not been investigated before, on the bloom predictions. In addition to the P loads that have been widely applied in previous modeling, the TN load was recently discovered to be significantly related to HABs in the WLEB (Newell et al., 2019; Paerl et al., 2020). Therefore, short- and long-term N-based features were for the first time added into the model input. Finally, ML models were developed for the prediction of the important riverine and

Table 1
Classification of the CI values into three different levels and the number of data points for each level. SI = severity index (NOAA, 2021).

Bloom severity	SI	CI	2 levels	3 levels	4 levels	# Data points
Light	0–2	0–4.0	1	1	1	148
Mild	2–4	4–6.8		2	2	37
Significant	4–7	6.8–15.2	2	3	3	31
Extreme	>7	>15.2			4	24
					Total	240

Table 2
Initial input features for the RF models. The eight most important features identified in Section 3.3 are marked in bold.

Feature type	Feature [a]	Description
Hydrodynamic	Q10, Q20, Q30	Flowrate, ft^3/s
	WLM10, WLM20, WLM30	Water level near the Maumee River at Toledo, m
	WLD10, **WLD20**, WLD30	Water level near the Detroit River at Gibraltar, m
Meteorological	ATM10, ATM20, ATM30	Air temperature from buoy station 45005, °C
	EWD10, EWD20, EWD30	Easterly (positive value) and westerly (negative value) wind speed, m/s
	NWD10, NWD20, NWD30	Northerly (positive value) and southerly (negative value) wind speed, m/s
	PRE10, PRE20, PRE30	Precipitation, in
	SOL10, SOL20, **SOL30**	Solar irradiance, W/m^2
	SIC	Seasonal average of daily ice cover (daily ice cover >5% in area),%
	WIN10, WIN20, WIN30	Wind speed without the direction specified, m/s
Physicochemical	CON10, CON20, CON30	Conductivity in the Maumee River, μmho
	N10, N20, **N30**	Total nitrogen (=NO23+TKN)[b], metric ton as N
	1TN, **5TN**, 9TN	Long-term total N, metric ton: 1-, 5-, and 9-y cumulative TN loading [c]
	P10, P20, P30	Total phosphorus, metric ton as P
	1SRP, 5SRP, 9SRP	Long-term P, metric ton: 1-, 5-, and 9-y cumulative SRP loading [b]
	STP, **SSRP, STN**	Spring (March to June) nutrient, including spring TP, SRP and TN loadings, metric ton as P or N
	TEM10, TEM20, TEM30	Water temperature from NSRDB, °C
Time of year	**Time Period**	Number 1 to 15 for the 15 × 10-d time periods from June 1st to October 31st (details in Table S2)

[a] 10, 20 and 30 represent the average values for the time periods that correspond to the CI 10-d, the CI 10-*d* + previous 10-d, and the CI 10-*d* + previous 20-d periods, respectively. The 10-d time periods here are the same as the 10-d time periods for the composite CI values in Table S2.
[b] NO23 stands for nitrate and nitrite; TKN is the total Kjeldahl nitrogen.
[c] For cumulative nutrient loads, 1 year was defined as from March to February.

meteorological features because the feature values are not available real-time, which prevents us from using the RF models for short-term (10-d scale) bloom forecasts.

2. Materials and methods

2.1. Study area

As the nutrient sources and the occurrence of HABs in Lake Erie are mostly related to western Lake Erie, the study area was concentrated on the WLEB, as delineated by the area of around 3000 km^2 on the left side of the line from Point Pelee, Canada to Marblehead, Ohio (Fig. 1) (Bertani et al., 2017; Manning et al., 2019; Verhamme et al., 2016). Even though the monitored blooms sometimes crossed the boundary of the WLEB, the crossed parts were still considered when calculating the CI values. Correspondingly, meteorological data were collected for the WLEB.

2.2. HAB quantification

HABs were quantified by chlorophyll-a index (CI) values, using 10-d composite CI values (the maximum CI value in a 10-d window) extracted from satellite images. See Text S1 for more details of how to obtain the CI values. The reason that 10-d composite CI values were used was that the National Centers for Coastal Ocean Science (NCCOS) only published processed 10-d composite satellite images for the whole Lake from 2002 to 2019 (e.g., Fig. 1), yielding a total of 240 records. The 10-d window was determined based on the availability and representativeness of satellite images due to wind, cloud cover, and satellite revisiting time (Stumpf et al., 2012) (more discussion in Text S1). NCCOS also publishes the annual HAB severity levels: light, mild, significant, and extreme. The CI values were classified into 2, 3, or 4 levels based on the severity index (Table 1), and the levels were set as the corresponding classification model outputs. There are several additional HAB quantification programs for the WLEB. Please see Text S2 for a discussion of these programs and why we decided to use the CI values.

2.3. Selection of input features

To consider all influential factors in the input, we conducted a comprehensive literature review to identify all potentially relevant factors, including hydrodynamic, meteorological, physicochemical, and time of year, that could affect the bloom behavior (Table 2, Scheme S1). All features that have been used in the reported models for Lake Erie are summarized in Table S1. First, many studies have used spring P loads for HAB predictions (Ho and Michalak, 2017; Stumpf et al., 2016, 2012) because these loads are known to significantly affect HAB formation, either directly or after being deposited in lake sediments (Bertani et al., 2017; Michalak et al., 2013). Therefore, we included different loads of spring nutrient between 2002 and 2019 in the input features (Table S1 and Text S3). Meteorological features, including temperature, wind, light, ice cover, and precipitation, can also influence the bloom growth and were thus selected, too. Specifically, algal blooms are prone to grow in warm and still water that has high light availability (Paerl and Huisman, 2008; Sellner, 1997). Light availability, represented as solar irradiance, is indispensable for photosynthesis of autotrophs like the algae. Wind speed has reportedly influenced the vertical distribution of blooms in the water column (Chaffin et al., 2020; Hunter et al., 2008; Manning et al., 2019), while wind direction also matters for the bloom distribution (Bertani et al., 2017; Rowe et al., 2016). Although not occurring in the summertime, ice cover will block the sunlight and influence sediment resuspension in lakes, which then subsequently changes the ecology and economy of the lake regions, including the P cycle (Assel, 2005). Finally, precipitation is known for its influence on the nutrient loads (Richards et al., 2010), as precipitation-caused runoffs will carry nutrient to the receiving water.

The hydrodynamic and physicochemical features were mostly obtained for the Maumee River—referred to as the riverine features sometimes. The Maumee River provides about 5% of the water influx but approximately 45%−50% of the P loading to Lake Erie, which

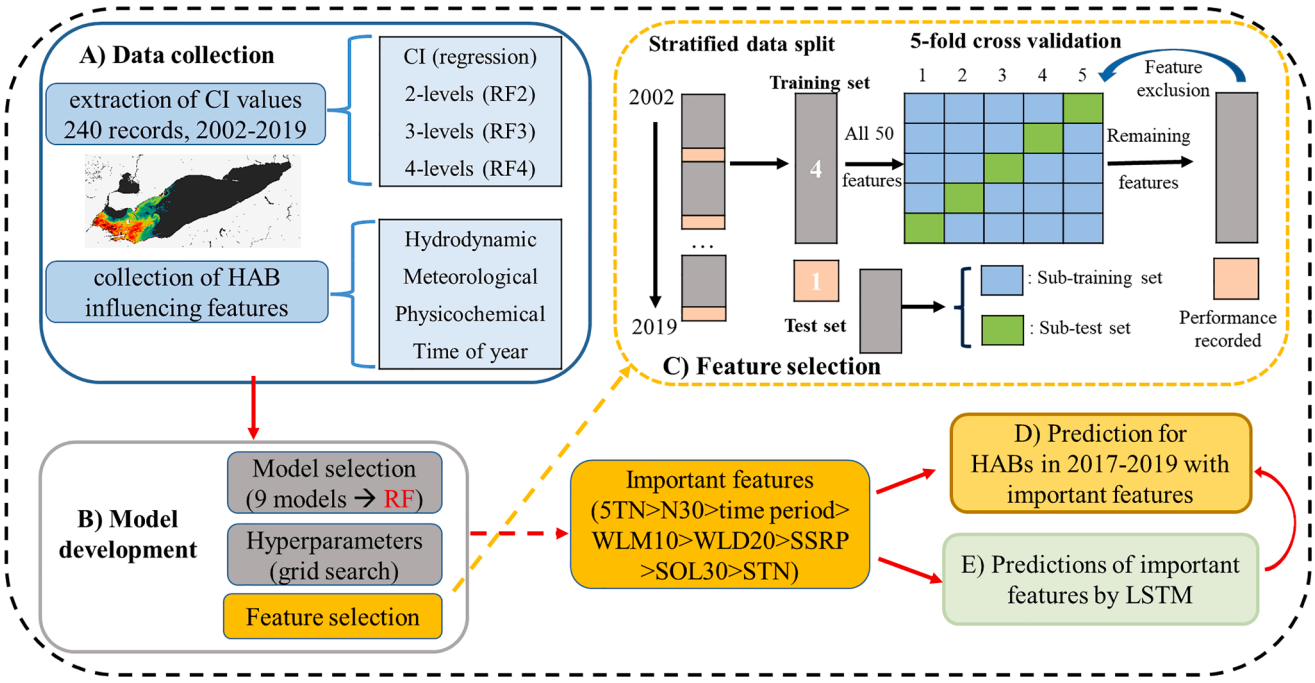

Fig. 2. Workflow of the RF model development, including A) data collection for the response and input features, B) model development to select the ML algorithm and optimize hyperparameters, C) feature selection to identify the most important features, D) model application to predict HABs in 2017–2019, and E) LSTM modeling to forecast the selected short-term features. The *feature selection* process in C) included four main steps: *i*) the 240 data points were first arranged sequentially over time (from 2002 to 2019) and then divided into the training set (in gray) and test set (in pink) in a ratio of 4:1; *ii*) the training set (in gray) was divided into sub-training sets (in light blue) and sub-test sets (in dark green), followed by 5-fold cross validation to train the model; *iii*) the performance of the optimized model was evaluated on the test set (in pink); and *iv*) based on the obtained model, the importance of the features was calculated by a permutation method; that is, for each feature, its importance score was the average of all the obtained feature importance values. The features with the lowest importance scores were excluded and the remaining features were employed in the next round of 5-fold cross validation. Steps iii) and iv) were repeated multiple times until the model performance became significantly worse. Note that for the same type of features (e.g., TEM10, TEM20, TEM30), the feature with the lowest importance score was also excluded. Please refer to Text S6 for more details of the feature selection process.

dominates the nutrient loading to the Lake (Bertani et al., 2017; Elliott, 2010; Michalak et al., 2013). Nutrients from the Maumee River are known to have a major impact on the occurrence of HABs in Lake Erie, as low flushing rates and subsequently long residence times in the lake are favorable for the algal growth. In addition, the water levels in the WLEB near the Detroit River (WLD) and the Maumee River (WLM) were selected to monitor the lake inflow (Rowe et al., 2016), and they reflect the water flow direction and water circulation conditions in the WLEB. Please see Text S3 for additional discussion on biological features.

As shown in Table S2, the CI values are usually small in early summer and in October, and reach peak in late summer. To better capture the temporal trend in the bloom occurrence, we integrated the time periods over the bloom season as part of the input, like in other models (Bertani et al., 2017; Fang et al., 2019). From June 1st to October 31st, there were fifteen 10-d time periods, and these time periods were numbered from 1 to 15 (1 for the first period of June 1–10, 2 for the second period of June 11–20, etc., Table S2).

As the CI data were 10-d composite values, most of the selected features were aggregated as 10-d averaged values so that the input and output matched in dimension. Before the onset and peak of HABs, there may be delays (time lags) in the HABs' responses to the environmental conditions (Franks, 2018; Moore et al., 2009; Valbi et al., 2019). Time-lags were thus considered by calculating 20-d and 30-d average values of the features (Table 2) (Bertani et al., 2017), to capture both the previous and current conditions. Moreover, 1-y, 5-y and 9-y cumulative SRP and total nitrogen (TN) loads were considered to represent the long-term P/N loads (Ho and Michalak, 2017). Details about the data sources are in Text S4. As some of the features were highly correlated with each other, data analysis was conducted to exclude some of the highly correlated features based on the Pearson coefficient, see Text S5 and Fig. S2. After that we obtained an initial list of 50 input features (Table 2). Compared to the small dataset (240 CI values), 50 features were still too many because using irrelevant features might worsen the model performance. So, we conducted feature selection based on the feature importance to significantly decrease the number of input features while maintaining the model performance, as details in Section 2.4.

2.4. Developing bloom prediction models

All the data were divided into training and test datasets in a ratio of 4:1 by stratified sampling, as we had compared stratified sampling with random data split and obtained better model performance with the stratified sampling (details in Text S6). The training dataset was further divided into 5 parts for a 5-fold cross validation (Hastie et al., 2009).The feature selection process involved sequentially identifying the least important features and excluding them until we obtained the smallest number of features while still maintaining a good model performance (Fig. 2C, more details in Text S6). The obtained regression models were evaluated by the R^2 and root mean square error (RMSE) values; the obtained classification models were evaluated by the prediction accuracy along with the confusion matrix and the Kappa coefficient (Text S7, Table S4). To select the best ML algorithm(s) for the modeling purposes, we compared the performance of 9 widely used ML algorithms: ANN, bootstrap aggregating (BA) (Breiman, 1996), gradient boosting (GB) (Friedman, 2001), gaussian processes (GP) (Williams and Rasmussen, 1995), k-nearest neighbors (KNN) (Taunk et al., 2019), LSTM, RF, support vector machine (SVM) (Li et al., 2021), and XGBoost (XGB) (Chen and Guestrin, 2016), based on all the 50 features and random data split (sequentially for the LSTM). Note that grid search was employed to

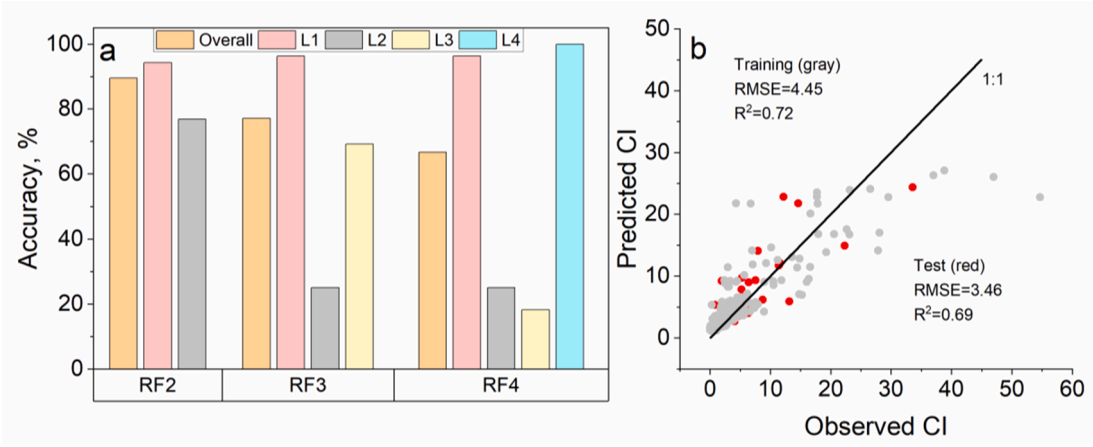

Fig. 3. (a) Prediction accuracy for the RF2, RF3, and RF4 models on the test datasets. (b) Performance of the RF regression model on the training (gray) and test (red) datasets. In the RF regression model, the following input features were used, shown in their order of importance: 5TN > time period > SSRP > SOL20 > STP.

optimize the hyperparameters for all the above algorithms during the model training process. Texts S6, S8, and S9 include more details of the RF, ANN and other model development, respectively. As the performance shown in Table S5, RF outperformed all other models in both the classification and regression models. Therefore, RF was employed to develop all the models below.

After the important features were identified following the feature selection process, RF models were developed using these features as the input. Moreover, the HABs in 2017–2019 were predicted based on the records from 2002 to 2016 to further evaluate the practical applicability of the RF classification model.

2.5. *Time series modeling with long short-term memory (LSTM)*

One major limitation in the application of the obtained models is that many of the input features are not available real time. To timely obtain these features, in-situ sampling is a helpful solution but requires high cost. In comparison, predicting these features using data-driven models would be an attractive choice.

Existing tools such as the Soil and Water Assessment Tool (SWAT), a semi-distributed hydrological and water quality model (Arnold et al., 1998; Kalcic et al., 2016), have been implemented to forecast the long term riverine nutrient data for the Maumee River (Kalcic et al., 2019; Verma et al., 2015; Yuan et al., 2020). However, little is done regarding *short-term* forecasts of the parameters. SWAT models also require many predefined parameters, such as moisture, evaporation and canopy features, which largely limits the applicability of the SWAT models. In comparison, LSTM is an advanced ML algorithm for modeling time series data (Gers et al., 2000). It can overcome the problem of stationary time series and is capable of learning long- and short-term patterns of changes (Hochreiter and Schmidhuber, 1997; Kratzert et al., 2018; Maier and Dandy, 1996), as LSTM can quickly yield fair predictions using the target itself as the input, especially when there are decades of records available. For example, based on 40 years of river flow data in Hun River, China, researchers used LSTM for river flow prediction at a 10-d scale and obtained a good performance (NSE=0.75) (Xu et al., 2020).

With decades of monitoring data, we employed LSTM to develop predictive models at the daily scale and 10-d scale for selected riverine and meteorological features, which will aid nutrient management and HAB forecast, respectively. Details of LSTM and how the LSTM models were developed are listed in Text S9 and Fig. S4. As ANN is also a common ML algorithm for water quality prediction (Chen et al., 2020), ANN models were also developed and compared with the LSTM models. The results (Table S6) showed that LSTM outperformed ANN in predicting the five riverine features, so LSTM was used for all the time series modeling.

LSTM models were optimized for four RF features—SOL30, N30, WLD20 and WLM10— at the 10-d scale. To predict the feature values for each 10-d period, we averaged the historical data into 10-d averages, to be consistent with the CI periods, and predicted 10-d averages for each feature by using all the data from the prior periods as the training data. For example, when the N10 data for the period of July 1st to July 10th, 2017 was to be predicted, it was treated as the test set and the TN data records prior to June 30th, 2017 were used as the training set. The predicted N10 is the model predicted 10-d average N values. The predicted N30 was the average of the predicted N10 and the observed TN in the previous 20 d. For the 31 × 10-d periods between 2017 and 2019 (Table S2), we obtained 31 predicted N30 values. Similarly, we obtained 31 predicted values each for SOL30, WLD20 and WLM10. The predicted values (not the observed values) for these features were then employed in the RF2 model for predicting the HAB levels in 2017–2019 and the performance was evaluated.

3. Results and discussion

3.1. *Classification and regression modeling based on RF*

According to the severity index shown in Table 1, the CI values were initially divided into four levels as the output for 4-level RF classification (RF4) models. However, as the data shown in Table 1, there are 148 data points for level 1 (RF4-L1) HABs but only between 24 and 37 data points for levels 2–4 (RF4-L2 to L4) blooms. As the whole dataset was very unbalanced, we further divided the data into 3-levels and 2-levels. For 3-level classification (RF3) model, we set light blooms as RF3-L1, mild blooms as RF3-L2, and extreme and severe blooms as RF3-L3. For 2-level classification (RF2) model, we set light and mild blooms as RF2-L1, and extreme/severe blooms as RF2-L2. For each classification model, we employed the feature selection process to eliminate insignificant features while maintaining the model performance and ended up with only eight to nine most important features (details in Section 3.2), which were employed to build the final models.

Fig. 3a and Tables S7–9 summarize the performance of the RF2, RF3, and RF4 models on the test set using the respectively identified top 8, 9, and 9 features. The RF2 model has the best performance, with the overall accuracy of 89.6% and the Kappa coefficient of 0.73. Note that a Kappa value of > 0.4 is deemed weak, that of > 0.6 is moderate and that of 0.8 is strong (McHugh, 2012). The prediction accuracy is higher for RF2-L1 blooms than for RF2-L2 blooms (94.3% vs. 76.9%). In comparison, SVM and a Naïve Bayes classifier were recently applied to predict *Karenia brevis* bloom events in the West Florida Gulf for two-level predictions (Li et al., 2021). With 318 bloom events and 765 non-bloom

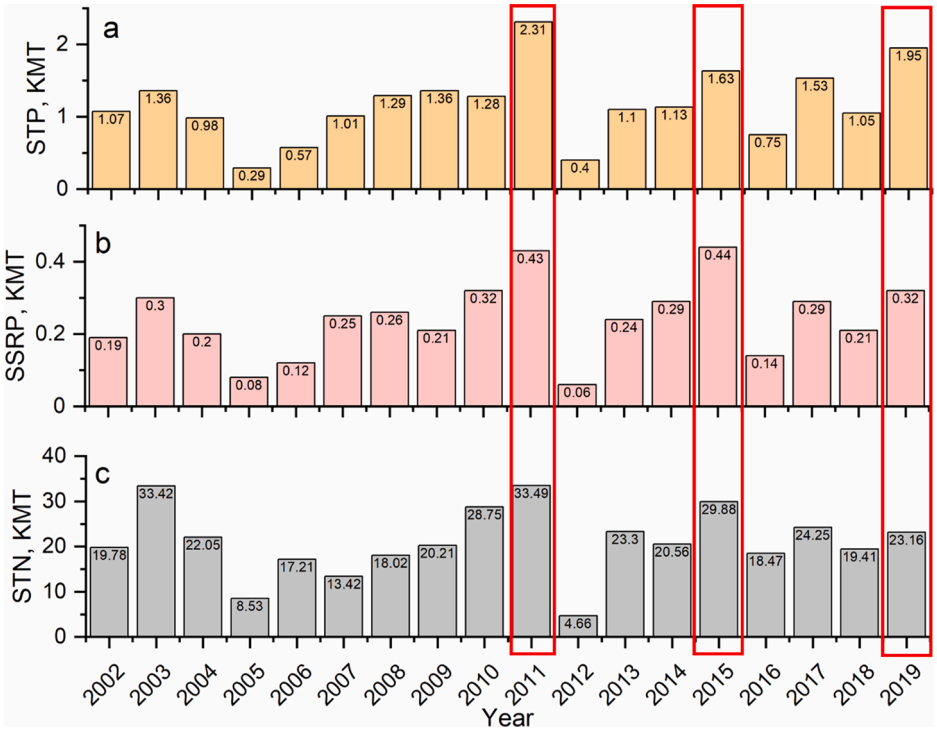

Fig. 4. (a) Spring total phosphorus (STP), (b) spring soluble reactive phosphorus (SSRP), and (c) spring total nitrogen (STN) from the Maumee River to Lake Erie in spring (March to June) between 2002 and 2019. KMT=1000 t. Years 2011, 2015 and 2019 are boxed for their high nutrient loads.

events in the dataset, they obtained an accuracy of 79% by SVM as the best performance, using riverine TP, TN, Q, temperature, wind speed/direction, and sea surface height as the features. Given that the above study had more data points (1083 vs. 240) and a more balanced data distribution between the two levels (L1: L2 = 2.4: 1 vs. 3.4: 1), the performance of the RF2 model developed in this study is quite satisfactory. We further tried to address the data imbalance issue by either synthesizing additional data or removing a portion of the RF2-L1 data, but did not observe improved model performance (details in Text S11).

The RF3 model had an overall accuracy of 77.1% and a moderate Kappa coefficient of 0.59, but it had a low accuracy of 25.0% for RF3-L2 HABs (Fig. 3a and Table S8). The RF4 model had an overall accuracy of 66.7% and a weak Kappa coefficient of 0.43. It however had a low accuracy of 25.0% and 18.2% for RF4-L2 and RF4-L3 blooms, respectively (100% for RF4-L4 blooms). The imbalanced dataset and small sample size are most likely the reasons for the poor performance because accurate ML models require both high data quality and large sample sizes (Tao et al., 2017). Nevertheless, both the RF3 and RF4 models still have a high accuracy of 96.3% for RF4-L1 blooms and good accuracy for the most severe blooms, so both models can accurately issue a safety notice to the public.

The RF regression model achieved an R^2 of 0.72 and 0.69 for the training set and test set, respectively (Fig. 3b). The model performance on the training set is similar to that of the reported linear regression or probabilistic models (R^2 from 0.71 to 0.97, Table S1), but those models are based on much smaller sampler sizes (10—28). The performance on the test set is better than that of reported ML models—R^2 of 0.181 to 0.497 for boosting regression tree models (Bertani et al., 2017), or 0.76 and 0.50 for ANN models based on only 3-years of biomass data (Millie et al., 2014). The performance of the regression model is also better than that of the reported phytoplankton models whose typical R^2 values range from 0.4 to 0.6 (Arhonditsis and Brett, 2004).

To further examine the ability of the RF models to predict HABs, we trained another RF2 model based on the data from 2002 to 2016 to predict the HABs between 2017 and 2019. As shown in Table S11, the developed model showed an overall accuracy of 74.2% and a moderate Kappa coefficient of 0.45, with the annual accuracy of 91%, 82% and 44% for the year of 2017, 2018 and 2019, respectively. Four of the 9 predictions for the year of 2019 were overestimation, probably due to the high STP and SSRP levels in 2019 (Fig. 4). From 2002 to 2019, the year of 2019 had the 2nd highest STP, 3rd highest SSRP and 7th highest STN. The high nutrient levels might have contributed to the overestimation. Additional predictions for the year of 2020 were also made based on the whole dataset covering 2002–2019, to simulate real practices, and the predictions matched the reported values (Text S12, Table S12). Overall, the above results demonstrated that the obtained RF models can be employed to successfully predict HABs in the WLEB. Given that the RF2 model performed the best and might be a more practical way, e.g., only those extreme and severe HABs are of more concern and more informative to the public, we focus on the RF2 model below to predict HAB levels for the WLEB.

3.2. *Interpretation of the feature importance*

It is essential to interpret the obtained ML models and examine if they are trustworthy—following the ground truth. This was done by ranking the feature importance, which can also help 1) achieve parsimonious models so that only the minimum number of features are needed, to facilitate the application of the models because all these feature values should be obtained through intensive data collection, monitoring, or modeling efforts. Fewer features also introduce less uncertainties; and 2) design effective HAB management plans based on the most relevant features. Following the feature selection process in Section 2.4, we identified eight most important features for the RF2 model, in the order of 5TN > N30 > time period > WLM10 > WLD20 > SSRP > SOL30 > STN, as discussed below. This feature importance rank also agreed well with the Shapley values obtained from the Shapley feature interpretation method, see Text S13 and Fig. S7. The relevance or non-relevance of all other features is discussed in Text S14.

It was initially surprising that 5TN, N30, and STN were among the identified 8 important features as they revealed the significant role of N in the HAB modeling; yet, N has not been incorporated in any of the

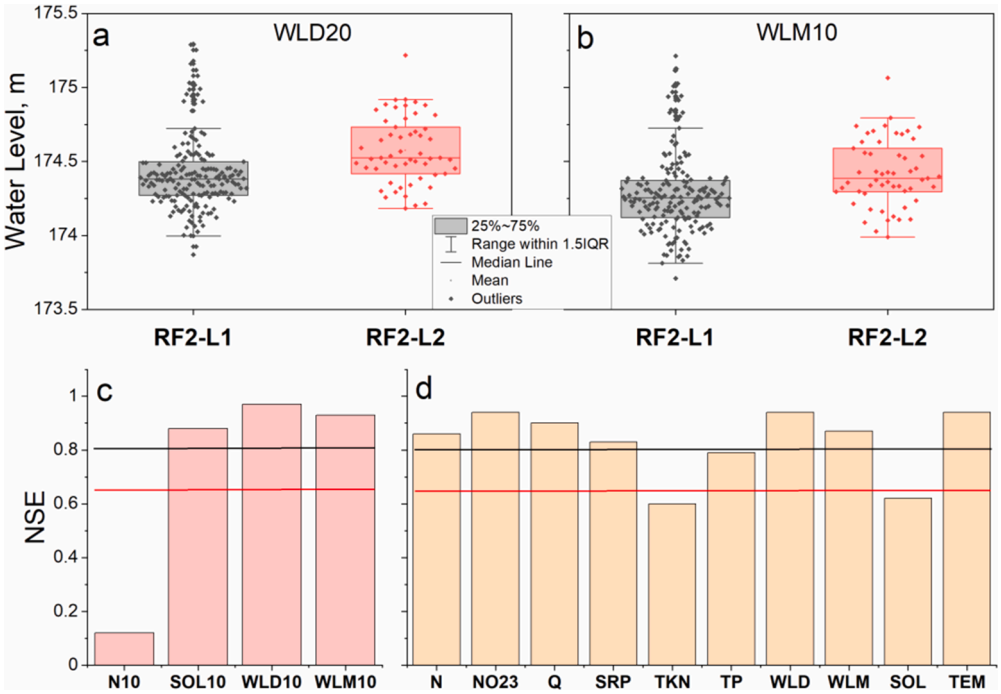

Fig. 5. Boxplots of the corresponding (a) WLD20 and (b) WLM10 for RF2-L1 versus RF2-L2 HABs. LSTM model performance (NSE) based on the parameter itself as the input for (c) 4 parameters based on 10-d averages and (d) 10 parameters based on daily averages. Note that N (in metric ton) is the daily N load based on the product of (NO23+TKN) × Q. Reference lines: red (NSE = 0.65), black (NSE = 0.8).

reported HAB prediction models for Lake Erie. According to the monthly nutrient loadings into the Lake, there is usually no N limitation except in late summer (Chaffin et al., 2013, 2014). In some extreme years such as 2011, there is no N limitation even in the late summer (Michalak et al., 2013). As a result, researchers have been focusing on P when developing models for the WLEB (Stumpf et al., 2012). However, an abundant spring TN load is necessary for the late summer bloom growth. Moreover, a recent study on the Sandusky Bay, Lake Erie, discovered that the amount of internal ammonia regenerated in the water column over the three summer months accounted for 77% of the annual external TN load, and the amount of regenerated N in late August was even 1000 times of the corresponding river N input into the Bay (Hampel et al., 2019). Similar findings have also been reported for Lake Taihu where 9% of the external N load was stored in the lake and the regenerated ammonia in the water column provided 38%−58% of the NH_4^+ needed for summer-fall blooms (Xu et al., 2021). Furthermore, recent studies have proposed both N and P reduction plans for HABs' management since N and P reportedly jointly spur the growth of HABs (Paerl et al., 2016, 2020). In agreement with the above findings, the worst HABs in Lake Erie occurred in 2011 and 2015 which had the highest spring N/P loads (Fig. 4). These findings suggest a potentially important role of N in the HAB formation in the WLEB, and the necessity to integrate N-based features, such as 5TN and N30, in the models to capture the influence of long-term and short-term N loadings, respectively. Text S15 shows our additional effort to improve the performance of the RF2 model by considering different contributions of regenerated N, but the model performance did not improve further.

Time of the year (time period) ranked the third because, as mentioned previously, HABs in the WLEB follow the temporal trend of starting in early June, peaking around August, and disappearing in October. Therefore, time is an essential feature to capture the dynamics in HAB formation. Similarly, day of year and month of year have been employed as input features in recent space-time geostatistical and machine learning models (Table S1) (Bertani et al., 2017; Fang et al., 2019).

The water levels near the Maumee River (WLM10) and the Detroit River (WLD20) were identified as two of the most important features in the RF2 models. In the literature, the hydraulic load from the Detroit River was also treated as a significant parameter affecting the HAB formation in the WLEB (Kane et al., 2014; Scavia et al., 2016). This is because high flowrates along with low nutrient concentrations from the Detroit River can dilute the lake water and the river flow will affect the lake circulation (Wynne et al., 2011). A recent study linked the HAB occurrence to low water levels because a low water level leads to water stabilization and nutrient accumulation (Rousso et al., 2020). However, we observed significantly higher water levels for RF2-L2 HABs than for RF2-L1 HABs ($p<0.001$, Fig. 5a-b) for the WLEB (Rowe et al., 2016). Based on the recorded 10-d average water levels from 2002 to 2020 (Fig. S8), there are clear trends that the water levels near the Maumee and Detroit Rivers increased from the beginning of the year to the midyear and then kept decreasing for the remaining of the year. The increasing water levels in the first half of the year means that the inflows are greater than the outflows. This can lead to the accumulation of nutrient-rich water in the WLEB over summer, which may provide a better condition for HAB formation.

Generally, P is the known limiting factor for the HAB formation in the WLEB (Chaffin et al., 2014; Joosse and Baker, 2011; Matisoff and Ciborowski, 2005). Therefore, the model correctly identified SSRP as a major feature. In addition, cumulative P load has been demonstrated to be influential to Lake Erie HABs (Del Giudice et al., 2018; Ho and Michalak, 2017). The 1SRP (1-year cumulative soluble reactive phosphorus) was the 9th important feature identified by the model, which presented a significant importance among other features. But it was not included in the final 8 features because including it did not further improve the model performance. Better features to represent long-term P effects might be needed.

Solar irradiance (SOL30) was the seventh most important feature. It is related to the clearness of the sky, as a cloudless sky is associated with higher water temperature and calm water conditions. This finding agrees with previously reported HAB prediction models for the Lake (Table S1) (Bertani et al., 2017; Millie et al., 2014; Rowe et al., 2016) where solar irradiance is often found to be positively related to HABs (Rousso et al., 2020).

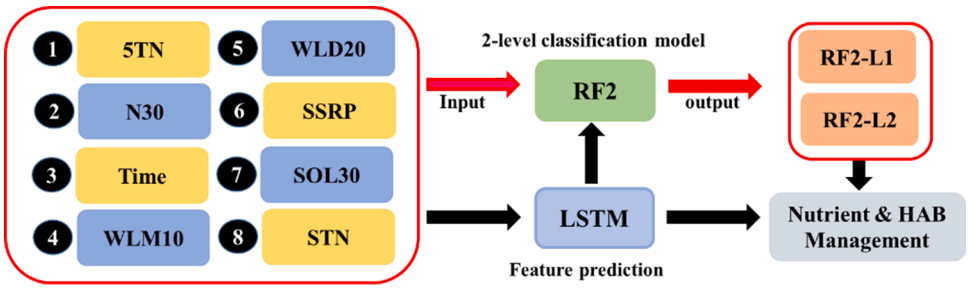

Fig. 6. Overview of the final RF2 model, including the model inputs and outputs from left to right. The features shown in blue were also modeled by the LSTM, which were then fed into the RF2 model for bloom forecast purposes. The black circled numbers represent the rank of the importance of the final 8 features, with 1 being the most important.

Similar types of 5–9 input features were also identified for the 3- and 4-level classification models and the regression model, with slight differences in the time lags or the feature importance, see the captions of Tables S8–9 and Fig. 3. In comparison, the reported ML models for the WLEB used a total of 29–31 input features (Bertani et al., 2017; Millie et al., 2014) (Table S1). The identified small number of features in this study significantly simplified the model input while still achieving very good model performance. This comparison also demonstrates the need to incorporate the most influential input features in ML modeling.

3.3. LSTM time series modeling of input features

Among the top eight input features for the RF2 model, four are short-term variables—SOL30, N30, WLD20 and WLM10—while the other four are not—5TN, time period, SSRP, and STN. Since the short-term features are not readily available real-time, we'd like to build predictive models for them so the RF models can be applied for short-term HAB forecasts. Fig. 5c shows the performance of the LSTM models for the four short-term features at the 10-d scale. SOL10, WLD10, and WLM10 all had good NSE values of 0.88–0.97. N10 showed a low NSE of 0.12, which might be due to the large fluctuation in the N values.

The optimized LSTM models were then employed to predict values for the four features (31 predictions per feature for 2017–2019). The predicted and observed data for these features agreed well, with the R^2 values of 0.75, 0.92, and 0.89 for SOL10, WLD10, and WLM10, respectively, and that for N10 being 0.22. The lower R^2 value for the N10 still might be due to the higher data fluctuation. Later the predicted 10-d average values were combined with the historical data to calculate SOL30, WLD20, and N30, and the predicted and observed data had a much better agreement with the R^2 of 0.94, 0.9, and 0.9, respectively.

To examine how well the LSTM predictions can be used for forecasting HABs, we employed the data from 2002 to 2016 to forecast the HABs in 2017–2019 and achieved an overall accuracy of 77.4% (86% for 2017–2018), with the annual accuracy of 91%, 82% and 56% for 2017, 2018 and 2019, respectively. Compared to the RF2 hindcast model in Section 3.1, this model performs almost the same (even slightly better for 2019), demonstrating the feasibility of combining the LSTM and RF models for forecasting HABs for the WLEB.

In addition, we built predictive models for the common riverine parameters (Fig. 5d) because an accurate prediction of Q would help forecast potential flooding events and facilitate subsequent response plans, and timely nutrient predictions can guide nutrient control measures for the Maumee River watershed. The performance of the optimized LSTM models for the 10 parameters at the daily scale is summarized in Fig. 5d and Table S14. All the models presented good NSE values of higher than 0.8, except for TKN (0.60) and SOL (0.62). Although TKN has a lower NSE of 0.60, the good NSE of 0.87 for N on the daily scale is more important as N is used in the RF models.

4. Conclusion

Many features are known to considerably influence the extent of HABs in the WLEB. In this work, we identified a novel combination of only eight input features—some were for the first time considered in one model— from a long list of 50+ initial features and employed them to build ML models for predicting HAB levels in the WLEB (final model summarized in Fig. 6). Overall, we have three major findings:

1. We developed effective classification and regression RF models to predict short-term (10-d scale) HAB levels in the WLEB and achieved an accuracy of up to 89.6% (2-level classification-RF2) on the test set. These models filled the gap between twice-a-week and annual bloom predictions and will benefit the public by providing short-term bloom alerts and aiding HAB management.

2. Based on the comprehensive understanding of the operating mechanisms of HABs in the WLEB and careful feature selection, we identified eight influencing features in the HAB occurrence in the order of 5TN>N30>time period>WLM10>WLD20>SSRP>SOL30>STN, some of which (5TN, STN, and N30) were for the first considered in the HAB modeling for the WLEB. These features contributed to more accurate models for better control of future HABs, including new plans to control N in addition to P. The findings also emphasized the need to include the most relevant features, not just a long list of partial features, in ML modeling to achieve good model performance (Zhong et al., 2022).

3. This work is not just restricted to retrospective historical data modeling. For the identified time-sensitive features, including total nitrogen loads, solar irradiance, and two water levels, we developed robust LSTM models to forecast them in a short-term (10-d in advance), which in turn can help forecast HABs for the WLEB. Nutrient management will also benefit from the LSTM predictions at both daily and 10-d scales for the N/P features to allow for more strategically adjusting fertilizer applications, controlling runoffs, and adapting to climate change.

Despite the above results, future work is warranted to further improve the model performance, including 1) significantly increasing the data volume, such as more satellite images at a finer time scale (more discussion in Text S2); 2) making biological features available (more discussion in Text S3); and 3) examining detailed N regeneration in the WLEB.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was funded by the Ohio Water Development Authority and NSF grant #2133576. The authors acknowledge Dr. Richard Stumpf at NOAA for providing the source of CI images, and Yili Gao and Lei Zan for their help with the image extraction.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2023.119710.

References

Arhonditsis, G.B., Brett, M.T., 2004. Evaluation of the current state of mechanistic aquatic biogeochemical modeling. Mar. Ecol. Prog. Ser. 271, 13–26.
Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large area hydrologic modeling and assessment part I: model development 1. J. Am. Water Resour. Assoc. 34 (1), 73–89.
Assel, R.A., 2005. Classification of annual Great Lakes ice cycles: winters of 1973–2002. J. Clim. 18 (22), 4895–4905.
Bertani, I., Obenour, D.R., Steger, C.E., Stow, C.A., Gronewold, A.D., Scavia, D., 2016. Probabilistically assessing the role of nutrient loading in harmful algal bloom formation in western Lake Erie. J. Great Lakes Res. 42 (6), 1184–1192.
Bertani, I., Steger, C.E., Obenour, D.R., Fahnenstiel, G.L., Bridgeman, T.B., Johengen, T. H., Sayers, M.J., Shuchman, R.A., Scavia, D., 2017. Tracking cyanobacteria blooms: do different monitoring approaches tell the same story? Sci. Total Environ. 575, 294–308.
Breiman, L., 1996. Bagging predictors. Int. J. Mach. Learn. Cybern. 24, 123–140.
Bridgeman, T.B., Chaffin, J.D., Filbrun, J.E., 2013. A novel method for tracking western Lake Erie Microcystis blooms, 2002–2011. J. Great Lakes Res. 39 (1), 83–89.
Chaffin, J.D., Bridgeman, T.B., Bade, D.L., 2013. Nitrogen constrains the growth of late summer cyanobacterial blooms in Lake Erie. Adv. Microbiol. 3 (06), 16–26.
Chaffin, J.D., Bridgeman, T.B., Bade, D.L., Mobilian, C.N., 2014. Summer phytoplankton nutrient limitation in Maumee Bay of Lake Erie during high-flow and low-flow years. J. Great Lakes Res. 40 (3), 524–531.
Chaffin, J.D., Kane, D.D., Johnson, A., 2020. Effectiveness of a fixed-depth sensor deployed from a buoy to estimate water-column cyanobacterial biomass depends on wind speed. J. Environ. Sci. 93, 23–29.
Chen, T. and Guestrin, C. 2016 Xgboost: a scalable tree boosting system, pp. 785–794.
Chen, Y., Song, L., Liu, Y., Yang, L., Li, D., 2020. A review of the artificial neural network models for water quality prediction. Appl. Sci. 10 (17), 5776.
Commission, I.J., 1972. The IJC and the Great Lakes Water Quality Agreement.
Del Giudice, D., Zhou, Y., Sinha, E., Michalak, A.M., 2018. Long-term phosphorus loading and springtime temperatures explain interannual variability of hypoxia in a large temperate lake. Environ. Sci. Technol. 52 (4), 2046–2054.
DePinto, J., Young, T., Salisbury, D., 1986. Impact of phosphorus availability on modelling phytoplankton dynamics. Hydrobiol. Bull. 20 (1), 225–243.
Elliott, J.A., 2010. The seasonal sensitivity of cyanobacteria and other phytoplankton to changes in flushing rate and water temperature. Glob. Chang. Biol. 16 (2), 864–876.
Fang, S., Del Giudice, D., Scavia, D., Binding, C.E., Bridgeman, T.B., Chaffin, J.D., Evans, M.A., Guinness, J., Johengen, T.H., Obenour, D.R., 2019. A space-time geostatistical model for probabilistic estimation of harmful algal bloom biomass and areal extent. Sci. Total Environ. 695, 133776.
Franks, P.J., 2018. Recent advances in modelling of harmful algal blooms. Glob. Ecol. Oceanogr. Harmful Algal Blooms 359–377.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
Gers, F.A., Schmidhuber, J., Cummins, F., 2000. Learning to forget: continual prediction with LSTM. Neural Comput. 12 (10), 2451–2471.
Hampel, J.J., McCarthy, M.J., Neudeck, M., Bullerjahn, G.S., McKay, R.M.L., Newell, S. E., 2019. Ammonium recycling supports toxic Planktothrix blooms in Sandusky Bay, Lake Erie: evidence from stable isotope and metatranscriptome data. Harmful Algae 81, 42–52.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical learning: Data mining, inference, and Prediction. Springer Science & Business Media.
Ho, J.C., Michalak, A.M., 2017. Phytoplankton blooms in Lake Erie impacted by both long-term and springtime phosphorus loading. J. Great Lakes Res. 43 (3), 221–228.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Hunter, P., Tyler, A., Willby, N., Gilvear, D., 2008. The spatial dynamics of vertical migration by Microcystis aeruginosa in a eutrophic shallow lake: a case study using high spatial resolution time-series airborne remote sensing. Limnol. Oceanogr. 53 (6), 2391–2406.
Joosse, P., Baker, D., 2011. Context for re-evaluating agricultural source phosphorus loadings to the Great Lakes. Can. J. Soil Sci. 91 (3), 317–327.
Kalcic, M.M., Kirchhoff, C., Bosch, N., Muenich, R.L., Murray, M., Griffith Gardner, J., Scavia, D., 2016. Engaging stakeholders to define feasible and desirable agricultural conservation in western Lake Erie watersheds. Environ. Sci. Technol. 50 (15), 8135–8145.
Kalcic, M.M., Muenich, R.L., Basile, S., Steiner, A.L., Kirchhoff, C., Scavia, D., 2019. Climate change and nutrient loading in the western Lake Erie basin: warming can counteract a wetter future. Environ. Sci. Technol. 53 (13), 7543–7550.
Kane, D.D., Conroy, J.D., Richards, R.P., Baker, D.B., Culver, D.A., 2014. *Re*-eutrophication of Lake Erie: correlations between tributary nutrient loads and phytoplankton biomass. J. Great Lakes Res. 40 (3), 496–501.
Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using long short-term memory (LSTM) networks. Hydrol. Earth Syst. Sci. 22 (11), 6005–6022.
Li, M.F., Glibert, P.M., Lyubchich, V., 2021. Machine learning classification algorithms for predicting Karenia brevis blooms on the West Florida shelf. J. Mar. Sci. Eng. 9 (9), 999.
Maier, H.R., Dandy, G.C., 1996. The use of artificial neural networks for the prediction of water quality parameters. Water Resour. Res. 32 (4), 1013–1022.
Manning, N.F., Wang, Y.-C., Long, C.M., Bertani, I., Sayers, M.J., Bosse, K.R., Shuchman, R.A., Scavia, D., 2019. Extending the forecast model: predicting Western Lake Erie harmful algal blooms at multiple spatial scales. J. Great Lakes Res. 45 (3), 587–595.
Matisoff, G., Ciborowski, J.J., 2005. Lake Erie trophic status collaborative study. J. Great Lakes Res. 31, 1–10.
McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochem. Med. (Zagreb) 22 (3), 276–282.
Michalak, A.M., Anderson, E.J., Beletsky, D., Boland, S., Bosch, N.S., Bridgeman, T.B., Chaffin, J.D., Cho, K., Confesor, R., Daloğlu, I., 2013. Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions. Proc. Natl. Acad. Sci. 110 (16), 6448–6452.
Millie, D.F., Weckman, G.R., Fahnenstiel, G.L., Carrick, H.J., Ardjmand, E., Young, W.A., Sayers, M.J., Shuchman, R.A., 2014. Using artificial intelligence for CyanoHAB niche modeling: discovery and visualization of Microcystis–environmental associations within western Lake Erie. Can. J. Fish. Aquat.Sci. 71 (11), 1642–1654.
Moore, S.K., Mantua, N.J., Hickey, B.M., Trainer, V.L., 2009. Recent trends in paralytic shellfish toxins in Puget Sound, relationships to climate, and capacity for prediction of toxic events. Harmful Algae 8 (3), 463–477.
Newell, S.E., Davis, T.W., Johengen, T.H., Gossiaux, D., Burtner, A., Palladino, D., McCarthy, M.J., 2019. Reduced forms of nitrogen are a driver of non-nitrogen-fixing harmful cyanobacterial blooms and toxicity in Lake Erie. Harmful Algae 81, 86–93.
NOAA, 2021. Harmful Algal Bloom Monitoring System. https://coastalscience.noaa.gov/research/-stressor-impacts-mitigation/hab-monitoring-system/.NOV4TH. , 2021.
Obenour, D.R., Gronewold, A.D., Stow, C.A., Scavia, D., 2014. Using a B ayesian hierarchical model to improve L ake E rie cyanobacteria bloom forecasts. Water Resour. Res. 50 (10), 7847–7860.
Paerl, H.W., Gardner, W.S., Havens, K.E., Joyner, A.R., McCarthy, M.J., Newell, S.E., Qin, B., Scott, J.T., 2016. Mitigating cyanobacterial harmful algal blooms in aquatic ecosystems impacted by climate change and anthropogenic nutrients. Harmful Algae 54, 213–222.
Paerl, H.W., Havens, K.E., Xu, H., Zhu, G., McCarthy, M.J., Newell, S.E., Scott, J.T., Hall, N.S., Otten, T.G., Qin, B., 2020. Mitigating eutrophication and toxic cyanobacterial blooms in large lakes: the evolution of a dual nutrient (N and P) reduction paradigm. Hydrobiologia 847 (21), 4359–4375.
Paerl, H.W., Huisman, J., 2008. Blooms like it hot. Science 320 (5872), 57–58.
Pyo, J., Park, L.J., Pachepsky, Y., Baek, S.-S., Kim, K., Cho, K.H., 2020. Using convolutional neural network for predicting cyanobacteria concentrations in river water. Water Res. 186, 116349.
Richards, R., Baker, D., Crumrine, J., Stearns, A., 2010. Unusually large loads in 2007 from the Maumee and Sandusky Rivers, tributaries to Lake Erie. J. Soil Water Conserv. 65 (6), 450–462.
Rousso, B.Z., Bertone, E., Stewart, R., Hamilton, D.P., 2020. A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. Water Res., 115959
Rowe, M., Anderson, E., Wynne, T.T., Stumpf, R., Fanslow, D., Kijanka, K., Vanderploeg, H., Strickler, J., Davis, T., 2016. Vertical distribution of buoyant Microcystis blooms in a Lagrangian particle tracking model for short-term forecasts in Lake Erie. J. Geophys. Res.: Oceans 121 (7), 5296–5314.
Sayers, M.J., Grimm, A.G., Shuchman, R.A., Bosse, K.R., Fahnenstiel, G.L., Ruberg, S.A., Leshkevich, G.A., 2019. Satellite monitoring of harmful algal blooms in the Western Basin of Lake Erie: a 20-year time-series. J. Great Lakes Res. 45 (3), 508–521.
Scavia, D., DePinto, J.V., Bertani, I., 2016. A multi-model approach to evaluating target phosphorus loads for Lake Erie. J. Great Lakes Res. 42 (6), 1139–1150.
Sellner, K.G., 1997. Physiology, ecology, and toxic properties of marine cyanobacteria blooms. Limnol. Oceanogr. 42 (5part2), 1089–1104.
Stumpf, R.P., Johnson, L.T., Wynne, T.T., Baker, D.B., 2016. Forecasting annual cyanobacterial bloom biomass to inform management decisions in Lake Erie. J. Great Lakes Res. 42 (6), 1174–1183.
Stumpf, R.P., Wynne, T.T., Baker, D.B., Fahnenstiel, G.L., 2012. Interannual variability of cyanobacterial blooms in Lake Erie. PLoS One 7 (8), e42444.
Tao, M., Duan, H., Cao, Z., Loiselle, S.A., Ma, R., 2017. A hybrid EOF algorithm to improve MODIS cyanobacteria phycocyanin data quality in a highly turbid lake: bloom and nonbloom condition. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10 (10), 4430–4444.
Taunk, K., De, S., Verma, S., Swetapadma, A., 2019. A Brief Review of Nearest Neighbor Algorithm For Learning and Classification. IEEE, pp. 1255–1260.
Thomas, M.K., Fontana, S., Reyes, M., Kehoe, M., Pomati, F., 2018. The predictability of a lake phytoplankton community, over time-scales of hours to years. Ecol Lett 21 (5), 619–628.
USEPA, 2021. Cyanobacterial Harmful Algal Blooms (CyanoHABs) in Water Bodies.https://www.epa.gov/cyanohabs#:~:text=Environmental%20Topics-,Cyanobacterial

%20Harmful%20Algal%20Blooms%20(CyanoHABs)%20in%20Water%20Bodies, harmful%20algal%20blooms%20or%20HABs. April 2nd 2021.
Valbi, E., Ricci, F., Capellacci, S., Casabianca, S., Scardi, M., Penna, A., 2019. A model predicting the PSP toxic dinoflagellate Alexandrium minutum occurrence in the coastal waters of the NW Adriatic Sea. Sci. Rep. 9 (1), 1–9.
Verhamme, E.M., Redder, T.M., Schlea, D.A., Grush, J., Bratton, J.F., DePinto, J.V., 2016. Development of the Western Lake Erie Ecosystem Model (WLEEM): application to connect phosphorus loads to cyanobacteria biomass. J. Great Lakes Res. 42 (6), 1193–1205.
Verma, S., Bhattarai, R., Bosch, N.S., Cooke, R.C., Kalita, P.K., Markus, M., 2015. Climate change impacts on flow, sediment and nutrient export in a Great Lakes watershed using SWAT. CLEAN–Soil, Air, Water 43 (11), 1464–1474.
Williams, C., Rasmussen, C., 1995. Gaussian processes for regression. Adv. Neural Inf. Process. Syst. 8, 514–520.
Wynne, T.T., Stumpf, R.P., Tomlinson, M.C., Fahnenstiel, G.L., Dyble, J., Schwab, D.J., Joshi, S.J., 2013. Evolution of a cyanobacterial bloom forecast system in western Lake Erie: development and initial evaluation. J. Great Lakes Res. 39, 90–99.
Wynne, T.T., Stumpf, R.P., Tomlinson, M.C., Schwab, D.J., Watabayashi, G.Y., Christensen, J.D., 2011. Estimating cyanobacterial bloom transport by coupling remotely sensed imagery and a hydrodynamic model. Ecol. Appl. 21 (7), 2709–2721.
Xia, R., Wang, G., Zhang, Y., Yang, P., Yang, Z., Ding, S., Jia, X., Yang, C., Liu, C., Ma, S., 2020. River algal blooms are well predicted by antecedent environmental conditions. Water Res. 185, 116221.
Xu, H., McCarthy, M.J., Paerl, H.W., Brookes, J.D., Zhu, G., Hall, N.S., Qin, B., Zhang, Y., Zhu, M., Hampel, J.J., 2021. Contributions of external nutrient loading and internal cycling to cyanobacterial bloom dynamics in Lake Taihu, China: implications for nutrient management. Limnol. Oceanogr. 66 (4), 1492–1509.
Xu, W., Jiang, Y., Zhang, X., Li, Y., Zhang, R., Fu, G., 2020. Using long short-term memory networks for river flow prediction. Hydrol. Res. 51 (6), 1358–1376.
Yuan, S., Quiring, S.M., Kalcic, M.M., Apostel, A.M., Evenson, G.R., Kujawa, H.A., 2020. Optimizing climate model selection for hydrological modeling: a case study in the Maumee River basin using the SWAT. J. Hydrol. (Amst.) 588, 125064.
Zhong, S., Zhang, Y., Zhang, H., 2022. Machine learning-assisted QSAR models on contaminant reactivity toward four oxidants: combining small data sets and knowledge transfer. Environ. Sci. Technol. 56 (1), 681–692.