Spatio-Temporal Deep Learning-based Algal Bloom Prediction for Lake Okeechobee Using Multi-Source Data Fusion

Yufei Tang, *Senior Member, IEEE,* Yingqi Feng, Sasha Fung, Veronica Ruiz Xomchuk, Mingshun Jiang, Tim Moore, and Jordon Beckler

***Abstract*—This study focuses on predicting harmful algal bloom (HAB) events in Lake Okeechobee, a shallow lake in Florida. A spatio-temporal deep learning model is employed to predict the levels of cyanobacteria *Microcystis aeruginosa* (M. aeruginosa) present in the lake for a single-day and a 14-day prediction horizon. Datasets collected from remote sensing (i.e., satellite images from Jan. 2018 to Dec. 2020) and from a physics-based simulation model (i.e., daily simulation from Jan. 2018 to Dec. 2020) are available. Due to the low quality of remote sensing data caused by various environmental and technical issues, the two available datasets are fused together to create a multi-source hybrid dataset for deep learning model training. A convolutional long-short term memory (ConvLSTM) deep neural model is trained on the datasets, and the results of the predictions are compared to the true Cyanobacterial Index (CI) for that time period. Findings include 1) the deep learning model, ConvLSTM, shows promising performance for short- and mid-term HAB forecasting; and 2) the hybrid dataset that fuses remote sensing with physics-based modeling (a.k.a. modeling based on fundamental physical and biogeochemical principles) speeds up the model learning and improves its performance significantly. The proposed methodologies are reliable, and cost-effective, and could be used to forecast algal bloom occurrences in shallow lakes with limited sparse observations.**

***Index Terms*—Harmful Algal Blooms (HABs), Multi-Source Data Fusion, Spatio-Temporal Prediction, Deep Learning Modeling, Convolutional Long-Short Term Memory (ConvLSTM)**

I. Introduction

A. Research Motivation

FOR the past decades, many major lakes have experienced an increased occurrence of algal blooms, including Lake Victoria in Africa [1], Lake Taihu in China [2], and Lake Okeechobee in the United States [3]. Specifically, Florida's Lake Okeechobee is the second-largest lake within the contiguous United States. A huge number of microorganisms reside in the waters of Lake Okeechobee. Some microorganisms are often the cause of algal blooms, which include cyanobacteria, also known as blue-green algae. Algal blooms are high concentrations of phytoplankton and harmful algal blooms (HABs) are problematic algal blooms that produce toxic or harmful effects on the ocean environment and human health. The dominant HABs in the Lake Okeechobee are the toxic cyanobacteria blooms, such as the *Microcystis aeruginosa (M. aeruginosa)*, which not only produces a dense surface mat that blocks waterways and emits a foul smell that cause hypoxia events but also produces microcystin, which is a potent hepatotoxin related to skin diseases, respiratory distress, and liver damage in animals and humans [4], [5]. This common freshwater species is quickly becoming a global health threat, with reported increases in both the frequency and intensity of blooms around the world [6], [7].

Improved understandings of the ecology and persistence of *M. aeruginosa* blooms and the distributions and bio-accumulation of their toxins represent a key challenge for scientists and water managers [8], [9]. To combat freshwater eutrophication and protect human and ecosystem health, this paper develops a HABs prediction model based on state-of-the-art deep learning algorithms, historical remote sensing data, and physics-based numerical simulation. This enhances the existing HAB monitoring program by predicting the bloom distributions and timing and helps management efforts to control the spread of HABs. Moreover, the proposed methodology in this research is applicable to other shallow water or lake algal bloom forecasting.

B. Machine Learning for HAB Modeling and Prediction

Deep learning methods and non-deep learning methods have been developed for HAB detection and prediction.

Classical Machine Learning Models: HAB monitoring is divided into detecting the occurrence and tracking motion caused by wind. Currently, available HAB detection systems are mainly based on empirical relationships obtained from previous observations. These methods have a high false alarm rate because they do not consider any temporal aspect or spatiotemporal dependencies. It is critical to understand the spatio-temporal dynamic behavior of HABs, especially for data-driven modeling approaches. Previous work shows that the spatio-temporal hybrid model can improve the predictive ability of traditional neural network (NN) and multivariate regression (MR) models [10]. The combination of kernel principal component analysis (KPCA) and support vector machine (SVM) has been proposed for generalized and improved

This work was supported in part by the National Science Foundation under Grant No. CMMI-2145571 and the U.S. State of Florida Department of Environmental Protection under Grant No. MN016.

Y. Tang, Y. Feng, and S. Fung are with the Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431 USA. {tangy, yfeng2016, sfung2017}@fau.edu.

V. R. Xomchuk, M. Jiang, T. Moore, and J. Beckler are with the Harbor Branch Oceanographic Institute, Florida Atlantic University, Vero Beach, FL 32963 USA. {vruizxomchuk, jiangm, mooret, jbeckler}@fau.edu.

HAB detection [11]. Back-propagation (BP) neural network, generalized regression neural network (GRNN), and SVM are then compared to demonstrate that the improved BP algorithm and SVM are better than the GRNN for HABs detection [12]. Random forest model using MODIS and MERIS satellite data while applying a threshold filter to balance the training inputs and labels has also been proposed, showing significantly better performance [13].

Deep Learning Models: Deep learning models have been recently employed for HAB detection and prediction. For example, multilayer perceptron (MLP), recurrent neural network (RNN), and long short-term memory (LSTM) have been proposed for modeling the HABs [7]. Results demonstrated that the LSTM model has the highest prediction rate, which reveals the potential for predicting algal blooms using LSTM and deep learning. Following this work, a variety of deep learning architectures that utilized the state-of-the-art spatio-temporal analysis methods based on convolutional neural networks (CNN), LSTM components together with random forest, and SVM classification methods have been investigated [14]. Results were favorable with about 91% accuracy for detection and 86% accuracy for prediction. From the analysis of the above articles, the prediction of HAB must take into account its spatio-temporal dependencies. In similar research, a convolutional LSTM (ConvLSTM) is used to build a trainable model for spatio-temporal sequence forecasting problems and applied to end-to-end precipitation nowcasting [15] and Chlorophyll-a concentration prediction [16], [17]. Results show that the ConvLSTM network can capture spatio-temporal correlations well, and this inspires the use of ConvLSTM for HAB prediction in this paper.

C. Challenges and Contributions

The following scientific challenges are identified to develop a HAB prediction model for the Lake Okeechobee:

- **Complex and fast lake phytoplankton dynamics:** Lake Okeechobee is a shallow lake in southern Florida that is highly eutrophic. Phytoplankton lives in a seemingly chaotic environment where winds, waves, tides, and convection act to mix up the water that surrounds them. Phytoplankton in the lake often forms patchy structures on a wide range of scales, in spite of the turbulent mixing.
- **Spatial and temporal dependencies:** HABs growing in lakes are spatially and temporally correlated yet highly variable. Its spatio-temporal footprint varies, ranging from weeks to months, from a few square kilometers to thousands of square kilometers, with patterns that vary daily, seasonally, and yearly [18]. It is critical to consider those factors when developing a spatio-temporal deep learning model that can precisely and timely pinpoint the areas that are or will be affected.
- **Sparse and noisy remote sensing data:** Due to the required working conditions of satellite sensors and the influence of the atmospheric environment, remote sensing images often suffer from missing information and low quality, such as dead pixels and thick clouds [19]. As a result, satellite images only represent snapshots of the lake blooms, and there is no useful information for relatively long periods of time. Therefore, it is inadequate for effectively training a deep learning model alone.

To address these challenges, this paper develops a novel deep learning-based HAB forecasting model using multi-source datasets. The major contribution is two-fold:

- **Creating a hybrid dataset from multi-source:** Two datasets are available to train a deep learning model to predict the Cyanobacterial Index (CI) levels in the lake. The remote sensing dataset is collected from images taken by the satellites, which represents critical yet discontinuous information about the lake blooms. The second dataset is generated by a coupled hydrodynamic-biological model that simulates chlorophyll levels in the lake based on the environmental conditions of the lake. This dataset is clean and continuous. However, because it is based on a numerical model, it is a simplification of the real blooms. The remote sensing images and physics-based model data are fused together to form a hybrid dataset, enabling physics-informed data-driven modeling. The resulted dataset has the clean and continuous properties of the simulation along with the close to true-value information from the remote sensing.
- **Developing a spatio-temporal prediction model using deep learning:** HABs have an inherent aspect entrenched in space and time, therefore a combination of spatial and temporal analysis is required for the most effective prediction [14]. Utilizing the hybrid dataset as a continuous sequence of images, we develop a HAB prediction model based on convolutional long-short term memory (ConvLSTM) [15], that heritages the advantage of long-short term memory for capturing the temporal correlations in the sequence of daily lake condition and the advantage of convolutional neural network for extracting spatial feature in the images.

The rest of the paper is organized as follows. Section II first discusses the preliminaries of datasets, and then presents the hybrid dataset generated from remote sensing and physics-based simulation. Section III presents the proposed HABs prediction model based on ConvLSTM. Section IV presents the comparative results and Section V draws conclusions and future works. Finally, the mathematical background of the deep learning model is presented in the Appendix.

II. Multi-Source HABs Data Fusion

A. Background of the Datasets

Remote Sensing: Satellite remote sensing technology enables real-time, large-scale, routine monitoring and prediction of HABs, significantly reducing unnecessarily wasted expense and time [20]. Although detection of HABs in thin water surface layers is challenging, remote sensing still provides an effective tool for identifying high biomass HABs such as red tides [21] and *M. aeruginosa* [22]. *M. aeruginosa* is the HAB of most concern in Lake Okeechobee. The organism can outcompete other phytoplankton for light by regulating its own buoyancy, allowing it to float at the surface in turbid waters. This fortuitously allows detection via satellite remote and thus

Fig. 1: True color image from Sentinel-2A on July 4, 2021, where algal bloom can be clearly identified around the center of the Lake Okeechobee.

inclusion of this data for a predictive model. For example, Fig. 1 is a typical remote sensing image from satellites, where the greenish streaks and patches at the surface across different parts of the lake are the cyanobacteria. During bloom peaks, thick green mats of highly concentrated cyanobacteria will aggregate at the surface.

Physics-based Simulation: The concept of the physical model is to combine a hydrodynamic model with a biological model to simulate all of the major processes including hydrodynamics and nutrient cycling to predict phytoplankton blooms. The Regional Ocean Model System (ROMS) is a finite difference and terrain-following ocean model that has been widely applied to coastal and regional ocean modeling [23]. The modeling system is flexible, allowing modular implementation of many important oceanic components such as tides, circulation, biology, and sediment transport. However, its applications to inland waters such as lakes and reservoirs remain limited.

In the physical model, each of these variables interacts with several variables at once. Each variable has a dynamic equation with several terms describing how interactions take place. For example, the phytoplankton equation is as follows:

$$\frac{dP}{dt} = \overbrace{P(\mu - m)}^{P\text{ growth}} - \overbrace{gZ}^{Z\text{ graze}} - \overbrace{\tau(DON + P)P}^{P\text{ coag}} - \overbrace{w_p \frac{P \sin k}{dz}} \quad (1)$$

where on the right-hand side of this equation describes, in order of the terms, the net growth rate of phytoplankton (P), where μ is dependent on light, temperature and nutrient concentration, grazing by zooplankton (Z), coagulation of phytoplankton (PON) and sinking. One can calculate each of these terms to find which of them dominates.

In order to simulate phytoplankton blooms, a biogeochemical model is usually co-simulated. For example, one of the widely used model is the 8-component Fennel model [24], which includes 7 variables for simulating nitrogen cycle and phytoplankton blooms, and one variable for dissolved oxygen (DO). Dissolved inorganic nitrogen is represented by nitrate (NO_3), which also includes nitrite (NO_2), and ammonia (NH_4). There is only one group of phytoplankton and one group of zooplankton, but chlorophyll is also directly simulated. Chlorophyll (Chl) concentration is treated as an independent variable, but it is closely related to phytoplankton biomass. The biogeochemical model is coupled with the ROMS hydrodynamic module to simulate the physical-biological coupled dynamics. Dynamics of phytoplankton blooms is complex and highly nonlinear. While the model is designed to capture all of the major processes, its predictive capability is limited due to the limited knowledge of the biology and biogeochemistry and imperfect numerical solutions.

B. Remote Sensing Dataset

Following the idea of Hill *et al.* [14], the remote sensing data are initially used as the input data for our model. The used satellite products were imagery from the Ocean Land Color Imagery (OLCI) sensor on board the twin satellite pair Sentinel-3A/B operated by the European Space Agency. Sentinel-3 provides near real-time basic information for oceans and weather forecasts. The mission is based on two identical satellites operating in a constellation for optimal global coverage and data transfer. The 1,270-kilometer-wide OLCI provides global coverage every two days. For nominal orbit, at sub-satellite point, OLCI full resolution is about 300 m above ground [25]. Acquired images were then processed through NASA's SeaDAS image processing package, producing two ocean color products for every image – Turbidity and the Cyanobacteria Index (CI). For our research object, we mainly focus on the CI product. The CI is an optical product based on a line height form using three bands in the red/NIR region [26], specifically the reflectance at 665 nm, 681 nm and 709 nm. The CI captures an optical feature that is related to particles in suspension near the surface of the water with a combination of spectral absorbing and scattering properties [27]. Here, the CI is calculated as:

$$\begin{aligned} \mathrm{CI} = -[(\rho_{681} - \rho_{665}) - (\rho_{709} - \rho_{665}) \\ \times (681 - 665)/(709 - 665)] \end{aligned} \quad (2)$$

where ρ_{xxx} is the partial reflectance at that wavelength. A positive CI indicates presence of cyanobacteria, which we used as a flag in data plots. Note that the negative sign for CI asserts a positive value when there is a trough detected at 681 nm. The CI product uses wavelengths in the red and near-infrared spectrum, shown in Fig. 2, and is insensitive to aerosol loads in the atmosphere. To derive this product, images were produced from partially atmospheric correction where only the molecular scattering is removed, leaving aerosols. These partially corrected images, which resulted in spectral reflectance products, were then subsequently used in the CI algorithm developed in [26]. The cloud and turbidity masks for the CI product were applied using the NOAA operational scheme [28]. These image products were saved into a NetCDF file. Owing to the type of algorithm applied for the surface cyanobacteria detection, the partial atmospheric correction

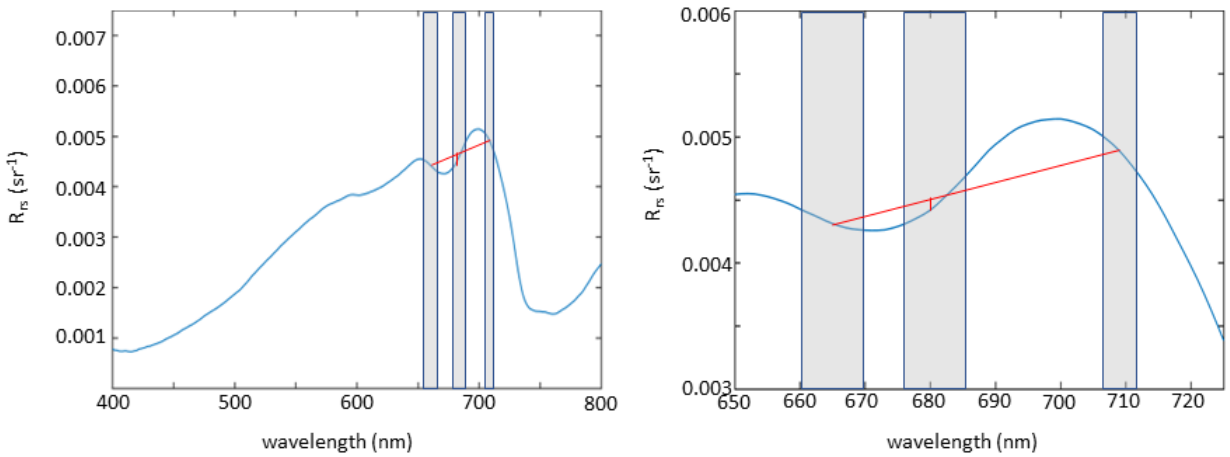


Fig. 2: A reflectance spectrum measured in Lake Okeechobee, with the spectral bands used in the CI highlighted.

could be applied for the CI product, avoiding some limitations associated with the full atmospheric correction. However, only cyanobacteria immediately at the surface could be detected with any degree of accuracy with this method. Nonetheless, the patterns in the images were useful for observing the spatial scale of events and were very useful in detecting surface cyanobacteria populations. While these are more qualitative, the expression of CI indicating surface cyanobacteria in and of itself has quantitative information.

For turbidity, a full atmospheric correction is required. We used a variation of the NASA standard scheme called the Management Unit of the North Sea Mathematical Models (herein MUMM) algorithm. This scheme was developed to account for near infrared (NIR) backscattering using a different water model and assumes a constant shape in the NIR and was derived from waters where particles are dominated by inorganic types [29], [30]. The MUMM was used for image processing and generating spectral remote sensing (Rrs) products. A turbidity product was generated from a semi-empirical single band turbidity retrieval algorithm [31], [32] for turbid waters with Rrs as input. The algorithm relates turbidity (T) and water reflectance $\rho_w(\lambda)$ at wavelength λ using:

$$T = \frac{A_T^{\lambda}\rho_w(\lambda)}{\left(1-\rho_w(\lambda)/C^{\lambda}\right)}[FNU] \tag{3}$$

where A_T and C are two wavelength-dependent calibration coefficients. In our study, we used the λ = 865nm wavelength. The CI product was generated for the same image but with the partial atmospheric correction using rho (reflectance) as input.

This dataset can be considered large, containing 191*216 pixel remote sensing images from October 2016 to the present. Although the dataset is large, there are a few issues that render some samples inoperable. On the one hand, although Sentinel-3 consists of two satellites, Sentinel-3A and Sentinel-3B, the two satellites in orbit make the revisit time of Lake Okeechobee only less than two days, there is no way to guarantee that we will get remote sensing images of the lake every day. There are also environmental sources that degrade image quality such as clouds, sun glint, and turbidity. It is often cloudy over Lake Okeechobee, and there is additional haze from the atmosphere, e.g., the sugar cane farms surrounding the lake routinely light fires as part of the cultivation process. The types of satellites used to detect cyanobacteria cannot see through the clouds, as they are passive sensors that detect the visible light field (much like our eyes). The near-total reflection of the sun off the water directly into the viewing sensors of the satellite is known as sun glint, and it is particularly acute between April and September. The extreme level of turbidity in Lake Okeechobee further limits the capabilities of the satellites, in particular, detecting cyanobacteria beneath the surface of the water. All these factors cause random discontinuities in the dataset. Fig. 3 shows some remote sensing data from June 19 to June 28, 2021. It can be seen that these samples contain missing and bad data, which compromises the integrity of the entire dataset.

Because HABs are related to space and time, their prediction requires high temporal continuity. Lack of critical information and discontinuous datasets will greatly affect the performance of deep learning models and lead to inaccurate predictions. A remedy to this problem involves the reconstruction of the remote sensing data. J. Li *et al.* proposed a method to determine total algal biomass in shallow lakes by combining remote sensing image data and hydrological data under non-algal bloom conditions [33]. Q. Zhang *et al.* [19] used a framework called STS-CNN to do the reconstruction. However, the amount of missing data in our dataset is too influential and would cause significant discrepancies between the reconstructed remote sensing data and the actual data. The deviation from the real data would negatively impact the prediction results or even render the model not trainable.

C. Physics-based Modeling Dataset

The model is driven by rivers' inputs and outputs of water, phytoplankton, nutrients, and organic matter, and surface forces including winds, heat fluxes, and water fluxes.

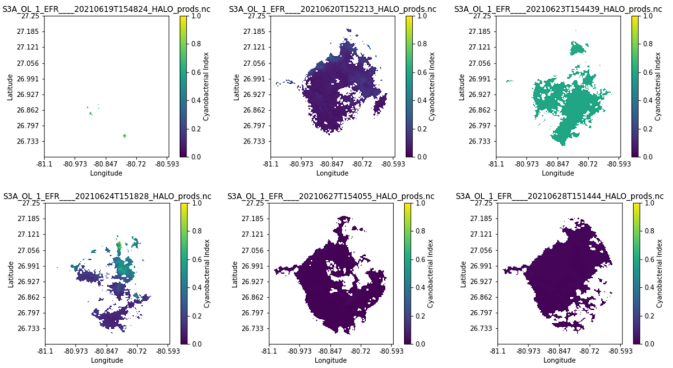


Fig. 3: Remote sensing data images for the Lake Okeechobee from the Sentinel-3A and Sentinel-3B satellites.

Considering the input and output of the rivers surrounding the lake, our physical model domain needs to cover the entire lake and part of the watershed. Therefore, our physical model uses coastlines to delineate lakes, and the model domain includes western wetlands. We choose to start with the ROMS model where the Lake Okeechobee ROMS domain covers the entire lake with a 386*386 horizontal grid (resolution ∽150 m), and 10 vertical layers. River flow and water quality data are derived from the observations by South Florida Water Management District (SFWMD), which include daily averaged water flow and biweekly or monthly water quality data. Atmospheric forcing is derived from the NCEP North American Regional Reanalysis (NARR) products, which has a spatial resolution about 30 km. Temporally, the NARR model output has a 3 hour interval, which resolves most of the diurnal variations.

Sediment fluxes are likely an important source of nutrients and a sink for dissolved oxygen. Sediment processes, however, are not directly simulated. Rather, particulate organic matter is allowed to sink out of the bottom, and, at the same time, sediment input of nutrients and sediment oxygen demand (SOD) are specified based on simple empirical formulations. Fisher *et al.* [34] estimates the DIN sediment flux to be 4500 metric tons annually with some spatial variation noted. This roughly equates to 0.5 mmol/m2/day. Therefore we use a simple empirical model to specify the sediment DIN flux as spatially variable depending on water depth as follows:

$$F_{DIN} = \frac{1 \times 10^{-5}}{\max(1, H)} \text{mmol} m^{-2}\,\text{sec}^{-1} \tag{4}$$

where H is the water depth. Note that this is constant throughout the year.

Several modifications have been made to ROMS codes to accommodate specific needs for modeling inland water such as Lake Okeechobee because ROMS was designed for coastal and regional oceans. These include: 1) allowing outflows of water from the modeling domain for canals or rivers; 2) including surface water fluxes (precipitation minus evaporation) in the water volume budget; and 3) using model tracer (e.g. temperature or NO3) concentration at the river mouth for tracer export (for outgoing flow only). In addition, ROMS evaporation algorithm under-estimates evaporation in the lake. An empirical evaporation formula is adopted instead.

The values of biological parameters, such as phytoplankton growth rate, zooplankton grazing rate, are adapted from [35]. Concentrations of both particles and organic matter is very high (e.g. DON≥50 umol/L), which includes a significant amount of CDOM. Therefore the light attenuation is high. Model results are calibrated against available historical water quality data collected from Lake Okeechobee by the SFWMD, which include monthly measurements of NO3, NH4, Chl, DO, and organic matter at a number of monitoring stations around the lake. We chose 8 representative stations for model calibration purpose. Model results at the same locations as the monitoring stations are directly compared with observations.

Once a basic calibration has been undertaken for the year 2018, a 3-year simulation is ran from Jan. 2018 to Dec. 2020. The result of this 3-year simulation makes up the physics-based model dataset, which includes daily averaged output of key physical and biogeochemical variables including currents (u, v) and chlorophyll. The physics-based model Chl values are empirically converted to CI through the relationship $CI = (Chl - 24.2)/(3.083 \times 10^3)$. Note that this empirical liner relationship was developed using summer data because *Microcystis* blooms in the lake usually take place during the summer season. As an example, Fig. 4 shows a comparison of model Chl and remote sensing CI for a few days when relatively complete remote sensing data images were available. It can be seen that there are differences between the physics-based model results and the remote sensing data. Although the physics-based model dataset is less reflective of the real HAB conditions than the remote sensing dataset, it provides a continuous and clean dataset as additional information that can be used to create a hybrid dataset for training the deep learning model. Below, these converted physics-based daily CI will be fused with remote sensing CI to derive the hybrid training dataset.

D. Fusion of Multi-Source Data as a Hybrid Dataset

Here we attempt to fuse the remote sensing and physics-based model data as a hybrid dataset. We have a total of 1093 physics-based model data from Jan. 2018 to Dec. 2020, therefore we selected remote sensing data of the same time period for data fusion. Three types of data samples are combined: remote sensing Satellite A denoted as X_A, remote sensing Satellite B denoted as X_B, and physics-based model denoted as X_P. After our manual statistics, there are 310 days of remote sensing data that are completely missing and 178 days of bad remote sensing data. So we total have 605 days of remote sensing data available. Among them, there are 100 days when both satellites have data. The remote sensing samples are manually scrutinized for the removal of corrupt data. On days that include remote sensing data from both satellites, the samples X_A and X_B are combined such that the resulting sample X_C are:

$$X_C(a,b) = \begin{cases} X_A(a,b), & \text{if } X_A(a,b) \in \mathbb{R}, X_B(a,b) \notin \mathbb{R} \\ X_B(a,b), & \text{if } X_B(a,b) \in \mathbb{R}, X_A(a,b) \notin \mathbb{R} \\ (X_A + X_B)/2, & \text{otherwise} \end{cases} \tag{5}$$

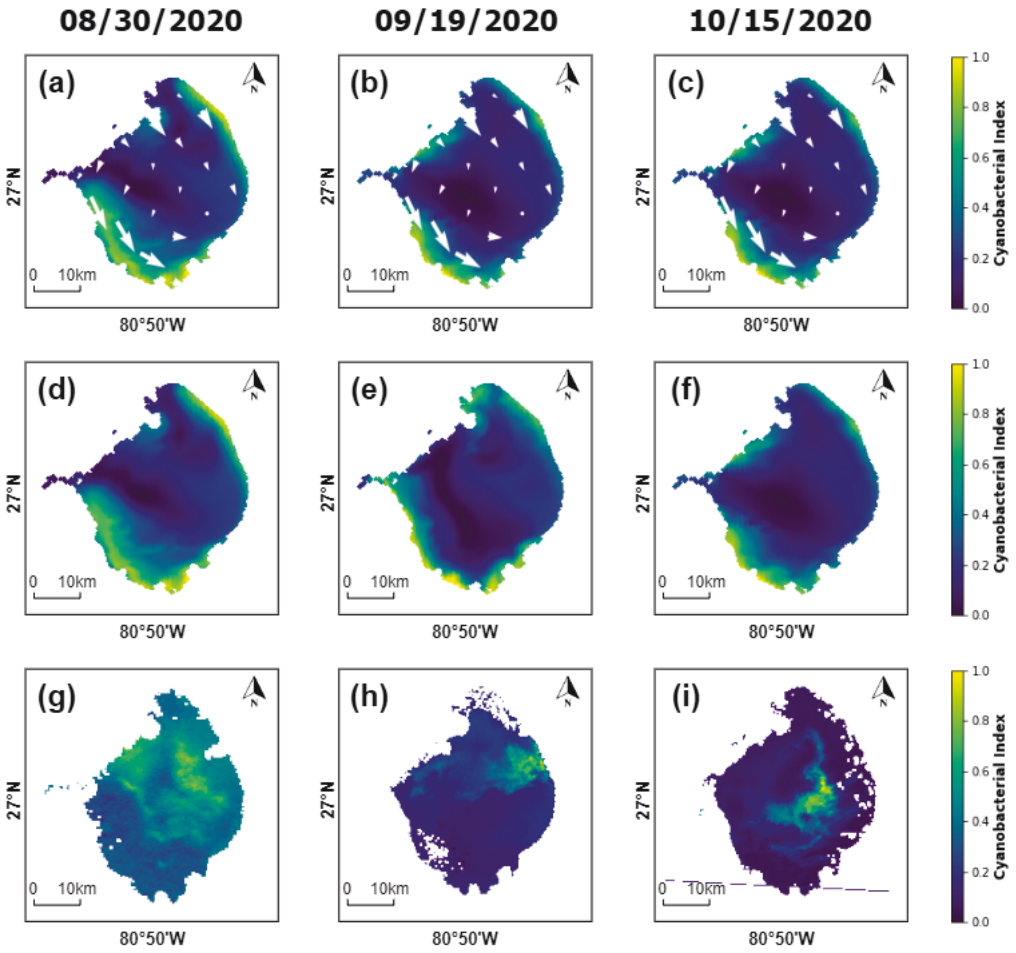


Fig. 4: Comparison of physics-based model simulated images and remote sensing data images of the same date. (a)-(c): Physics-based model data with u and v. (d)-(f): Physics-based model data without u and v. (g)-(i): Remote sensing data.

where a and b are the horizontal and vertical coordinates of the remote sensing images.

The final remote sensing sample X_{RS} is either X_A or X_B if only one exists, X_C if both exist, or missing if neither exist. For each day, a hybrid data sample X_H is then assembled by the imputation of X_P (the physics-based model generated data) onto X_{RS}. On days with missing remote sensing data, $X_H = X_P$, otherwise imputation is applied such that:

$$X_H(a,b) = \begin{cases} \alpha * X_{RS}(a,b) + \beta * X_P(a,b), & \text{if } X_{RS}(a,b) \in \mathbb{R} \\ X_P(a,b), & \text{otherwise} \end{cases} \tag{6}$$

where α and β are weights to balance the importance of different data sources. In this paper, we empirically set $\alpha = 0.8$ and $\beta = 0.2$ to count more on remote sensing when quality images can be obtained. This results in a hybrid dataset that utilizes the physics-based model information while considering quality information from remote sensing. Fig. 5 shows an example of the data fusion process and result.

As mentioned above, when remote sensing data are missing, the physics-based model data are used for the hybrid dataset. The spread and growth of algal blooms have certain rules, but due to the differences between physics-based model data and remote sensing data, this causes some parts of our data appear irregular. In order to improve this aspect, in the hybrid dataset, if the physics-based model data is used on a certain day d, that is, $X_H(d) = X_P$, and the data of the day before and the day after this certain day d are all hybrid data $X_H(d \pm 1) = \alpha * X_{RS} + \beta * X_P$ and, we then take the average of the data before and after of this certain day d as $X_H(d) = (X_H(d-1) + X_H(d+1))/2$.

The size of the physics-based model image is 386*386, while the remote sensing images have a lower resolution of 191*216. In order to fuse the data, the resolution of the physics-based model data needs to be reduced to match the remote sensing data. The image size of our hybrid dataset is 191*216. Moreover, both the physics and hybrid datasets contain a large amount of information that requires downsampling before it can be used as input to the machine learning models. Furthermore, all data undergo min-max normalization. The physics and hybrid data samples are individually resized to 112*112*1 and 83*88*1 respectively. Although the sharpness of the pictures is reduced, it is beneficial for speeding up the deep learning model training and online deployment when computational resources are limited.

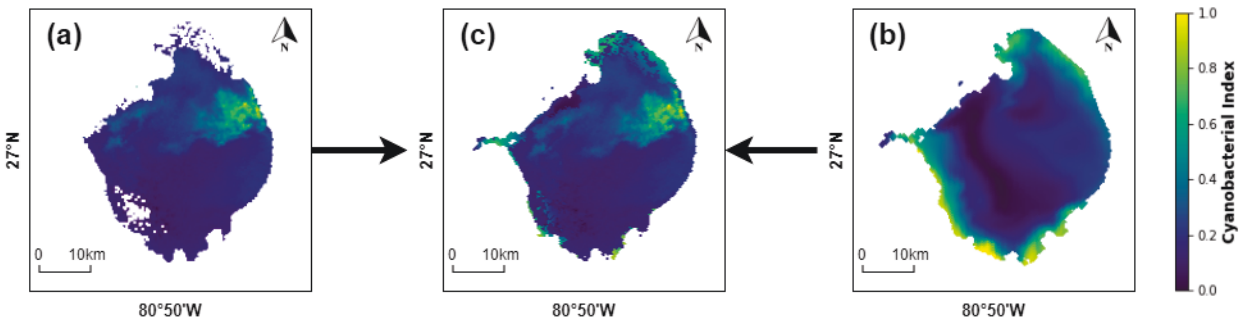


Fig. 5: Data fusion using data on September 19, 2020. (a): Remote sensing data. (b): Physics-based data. (c): Hybrid data.

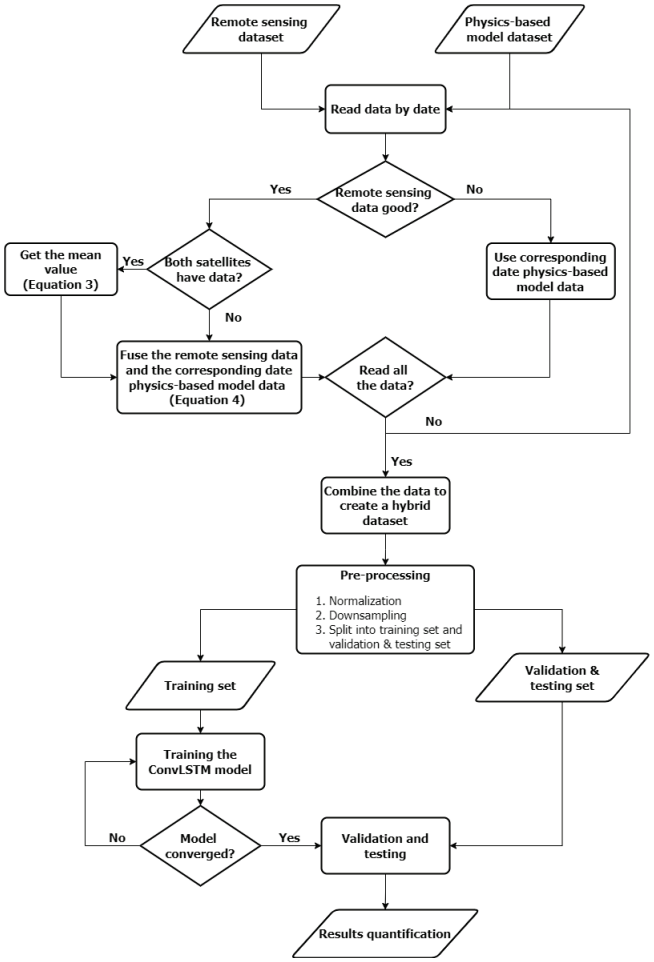


Fig. 6: Flowchart of our proposed approach, including data fusion, pre-processing, deep learning model training and testing, and results reporting.

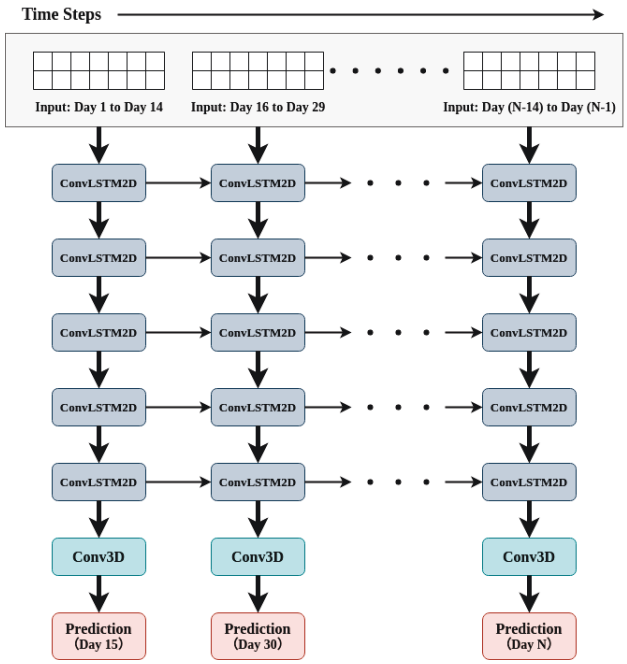


Fig. 7: Implementation diagram of the HAB prediction model based on ConvLSTM. For each training sample, the input to the model consists of 14-days of consecutive daily images which are processed to output a 1-day image prediction. To ensure a sequence learning and prediction, a rolling window mechanism is used.

III. Deep Learning-based HAB Prediction Model

A. Mathematical Background of ConvLSTM

We develop our HAB prediction model based on convolutional long-short term memory (ConvLSTM), a spatio-temporal deep learning model that heritages the advantage of long-short term memory for capturing the temporal correlations in the daily lake condition sequence and the advantage of convolutional neural network for extracting spatial feature in the images. The foundation and principles of ConvLSTM are presented in the Appendix.

B. Model Implementation Details

An overview of the deep learning-based HAB prediction model development is shown in Fig. 6. ConvLSTM implementation for HAB prediction in this paper is based on [36], [37], where the model is built using 5 ConvLSTM2D layers with batch normalization and is then followed by a Conv3D layer for spatio-temporal outputs. Fig. 7 shows the diagram of the model training and rolling window prediction. Since the ConvLSTM2D layer only accepts the inputs that have a specific shape (batch size, sequence, width, height, channels),

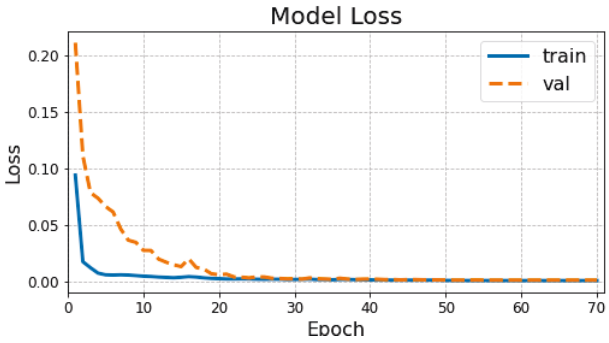


Fig. 8: Training loss and validation loss, where the losses are decreased to small values proving that the model learns from the data.

our hybrid dataset needs to be pre-processed before using for deep learning model training. The dataset contains 1093 samples (i.e., 1093 days of lake images), but in order to reshape the data more easily, only 1080 were used. For each training sample, the input to the model consists of 14-days of consecutive daily images which are processed to output a 1-day image prediction. Specifically, the hybrid dataset images have been reshaped to (72, 15, 83, 88, 1) corresponding to the numbers of batch size, sequence, width, height, and channels. This produces 72 batches of data, each batch has a sequence of 15 data images, and each image has a size of 83*88*1. The overall dataset is divided into a training set and a validation/testing set, where 80% are for training and 20% are for validation/testing.

We do the same processing for the physics-based model dataset, and the only difference is the width and height of the images. Note that the size of the images in the hybrid dataset is 83*88, while the size of the images in the physics-based model dataset is 112*112. The loss function used for model training is the mean squared error (MSE) (i.e., this will be further discussed in the following section), and the optimizer used is the Adam optimizer. It can be seen from Fig. 8 that the loss values (both training and validation) decrease to 0, which the model convergences demonstrate successful learning and prediction.

IV. Result and Discussion

A. Evaluation Metrics

The following image similarity measures are used [38]:

- **Root Mean Square Error (RMSE)** is a common comparison metric for measuring pixel-wise differences. A value of 0 indicates that the ground truth image and the predicted image are the same. The formula is as follows:

$$RMSE = \sqrt{\frac{1}{M*N}\sum_{i=0,j=0}^{M-1,N-1}[P(i,j)-G(i,j)]^2} \quad (7)$$

where G is the ground truth image and P is the predicted image. M represents the numbers of rows of pixels of the images and i represents the index of that row. N represents the number of columns of pixels of the image and j represents the index of that column.

- **Peak Signal-to-Noise Ratio (PSNR)** is often used to quantify the reconstructed quality of images and videos affected by compression. It measures the ratio between the maximum possible power of a signal and the destructive noise power that affects the fidelity of its representation, usually expressed on a logarithmic decibel scale. The higher the PSNR, the better the quality of the compressed or reconstructed image. To calculate PSNR, first calculate the mean squared error (MSE):

$$MSE = \frac{\sum_{M,N}[G(i,j)-P(i,j)]^2}{M*N} \quad (8)$$

$$PSNR = 10\log_{10}\left(\frac{R^2}{MSE}\right) \quad (9)$$

where R is the maximum pixel value of the image.

- **Structural Similar Index Measure (SSIM)** quantifies image quality degradation caused by processing or data transfer losses [39]. SSIM measures the perceptual difference between two similar images. The value of SSIM is between -1 and 1, where a value of 1 means that the two given images are very similar or identical and a value of -1 means that the two given images are very different. The SSIM index is calculated on various windows of an image. The measure between two windows x and y of common size N x N is:

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (10)$$

where μ_x is the average of x, μ_y is the average of y. σ_x^2 is the variance of x, σ_y^2 is the variance of y, σ_{xy} is the covariance of x and y. $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$ are two variables to stabilize the division with weak denominator. L is the dynamic range of the pixel-values. $k_1 = 0.01$ and $k_2 = 0.03$ by default.

- **Feature Similarity Indexing (FSIM)** is mainly used to compare the structural and feature similarity measures between the recovered object and the original object [40]. Phase Congruency (PC) is used as the primary feature in FSIM, and Image Gradient Magnitude (GM) as a secondary feature. The FSIM value is between 0 and 1, where 1 is perfect feature similarity. If the similarity between images f1 and f2 is to be calculated, the following formula is needed:

$$S_{PC}(x) = \frac{2PC_1(x)\cdot PC_2(x) + T_1}{PC_1^2(x) + PC_2^2(x) + T_1} \quad (11)$$

$$S_G(x) = \frac{2G_1(x)\cdot G_2(x) + T_2}{G_1^2(x) + G_2^2(x) + T_2} \quad (12)$$

$$S_L(x) = [S_{PC}(x)]^\alpha \cdot [S_G(x)]^\beta \quad (13)$$

$$PC_m(x) = \max(PC_1(x), PC_2(x)) \quad (14)$$

$$\text{FSIM} = \frac{\sum_{x\in\Omega} S_L(x)\cdot PC_m(x)}{\sum_{x\in\Omega} PC_m(x)} \quad (15)$$

where PC_1 and PC_2 represent the PC maps extracted from f_1 and f_2, respectively, and G_1 and G_2 represent the GM maps extracted. T_1 is a constant that increases the stability

TABLE I: Prediction performance on the datasets using different metrics.

Testing Data	Testing Dates	Training Data	RMSE	PSNR	SSIM	FSIM	SRE
Physics Simulation Data	Single-Day (July 03)	Physics	0.0008	62.2858	0.9992	0.8820	67.1597
	14-Day (April 9 to April 22)	Physics	0.0028	50.9297	0.9926	0.8150	65.2296
Remote Sensing Data	Single-Day (July 03)	Hybrid	**0.0079**	**48.4578**	**0.9692**	**0.4515**	**59.6935**
		Physics	0.0154	36.2355	0.9167	0.3880	54.6491
	4 Days of 14-Day (June 22 - June 25)	Hybrid	**0.0097**	**42.6706**	**0.9599**	**0.3985**	**59.8354**
		Physics	0.0142	36.9117	0.9308	0.3750	57.9144

of S_{PC} and T_2 is a constant that depends on the dynamic range of the GM value. $S_{PC}(x)$ and $S_G(x)$ are combined to obtain the similarity $S_L(x)$ of $f_1(x)$ and $f_2(x)$.

- **Signal to Reconstruction Error ratio (SRE)** measures the error related to signal power [41]. Using SRE is more suitable for making errors comparable between images of different brightness. SRE is calculated as:

$$SRE = 10\log_{10}\frac{\mu_x^2}{|\hat{x}-x|^2/n} \tag{16}$$

where μ_x is the average value of x.

RMSE, PSNR, and SRE are measures of how different two images are, which can help us judge whether the predicted image is similar to the "ground truth" image, but they do not take into account the quality of the image itself. This is solved by considering image structure (SSIM) and display features (FSIM) [38]. To sum up, the closer the RMSE is to 0, the higher the PSNR and SRE, and the closer the SSIM and FSIM are to 1, indicating that the ground truth image and the predicted image are more similar.

B. Comparative Results

Physics Data for HAB Prediction Model Validation: TABLE I shows the prediction performance on the benchmark datasets using the five similarity measures. There are two trials of experiments using two different testing data. In the first trial, we use the physics-based simulation data for the testing of the deep learning-based HAB prediction model. This can help to judge whether the ConvLSTM model is able to capture the spatio-temporal information of HABs in the Lake and make predictions. We carry out a single-day prediction using July 3, 2020, since the remote sensing data on this day is complete with high quality. The ultimate goal is to make a two-week (14-day) prediction, so a rolling window is created and the prediction result is used as input to predict the next day. More specific, when using the trained deep learning model for two-week predictions (i.e., days d to d+13), the first day d will be predicted using the data in days d-1 to d-14, then this predicted data in day d will be used together with data in days d-1 to d-13 to predict the second day d+1. This will continue until all 14 days have been predicted.

For ease of display and comparison, Fig. 9 shows the first 4 days of the 14-day prediction results for the model using the physics-based model dataset only. Since rolling predictions inevitably have information superposition, these results in the predicted image being blurry compared to the original image. Downsampling is used, which further causes the image resolution to decrease and become blurry. This "blur" can accumulate and further degrade the performance of the prediction. We observe that the model can successfully capture the spatio-temporal "evloving" patterns in the whole lake. In terms of quantitative metrics in TABLE I, the RMSE is very close to 0, and the SSIM and FSIM are also fairly close to 1, which further demonstrates the effectiveness of our proposed prediction method.

Remote Sensing Data for Hybrid Data Validation: Physics-based model simulation data may be different from the real distribution of the HABs since it is a simplification of the real mechanism. Although it suffers from sparse observation issues due to large quantities of missing data, using remote sensing data is still the best choice to reflect the real situation. To validate the effectiveness of our proposed hybrid data for deep learning model training, we use sparse but complete remote sensing data as the testing dataset in the second trial. In other words, we use the hybrid dataset to train our deep learning model and use the remote sensing data for testing. We also compare it with using the physics-based model simulation as training and remote sensing as testing.

As mentioned above, some relatively complete remote sensing data are selected. Fig. 10 shows the remote sensing data on July 3, 2020, and the prediction results using the physics-based model dataset as training or the hybrid dataset as training. Fig. 11 shows the remote sensing data from June 22 to 25, 2020, and 4 of the 14-day model prediction results for both datasets. June 22 to July 5 are chosen as the dates to be predicted since relatively complete and continuous remote sensing data can be found in this date range. From the figures, the prediction results of the model using the hybrid dataset are closer to the remote sensing data, whether it is a single-day prediction or a 14-day prediction. In the comparison metrics, since there are only 4 days of complete and continuous remote sensing data, the first 4 days from the 14-day prediction results of the two dataset prediction models are compared. In terms of metrics in TABLE I, all metrics using the hybrid dataset outperform those using the physics-based model dataset. This result also shows that although the model using the physics-based model dataset learns better, the prediction results deviate from the real situation due to the discrepancy between the dataset and

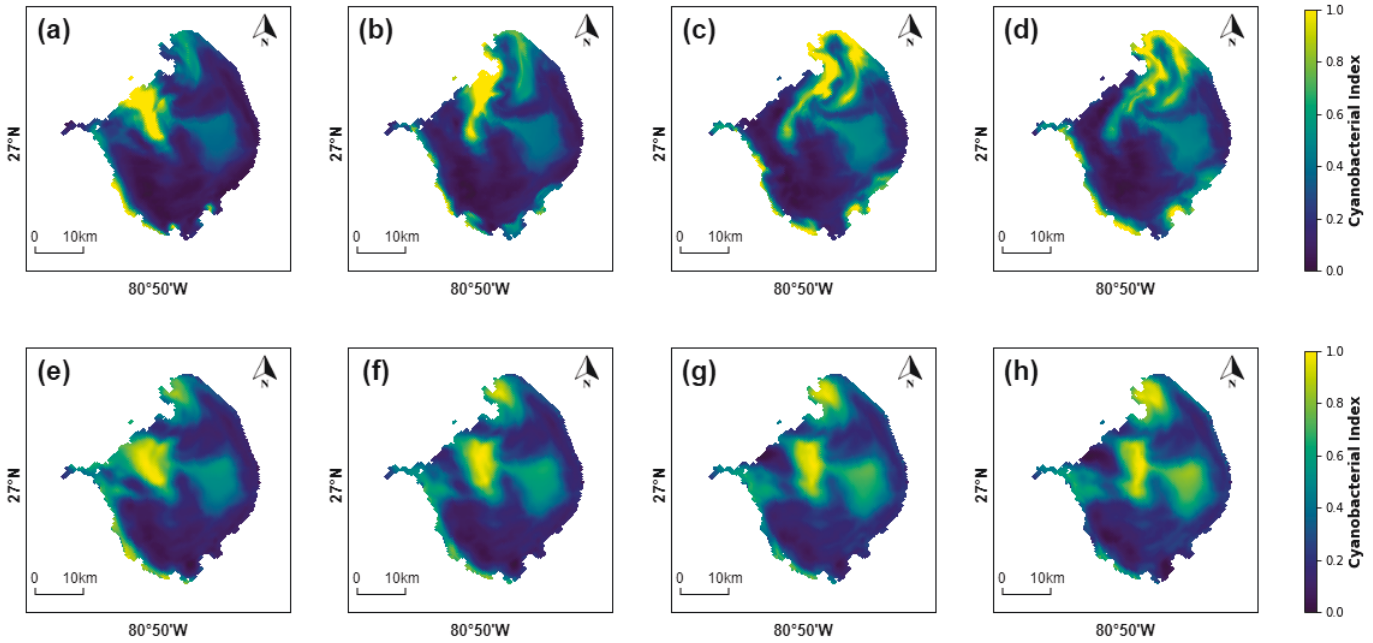


Fig. 9: Using physics-based model dataset for training and testing. Data from March 26 to April 8, 2019 are input to the model for predicting the HAB of the lake from April 9 to April 22, 2019. The first 4 days (April 9 to April 12) of the 14-day predictions are shown. (a-d): Physics-based model simulation results. (e-h): Proposed model prediction results.

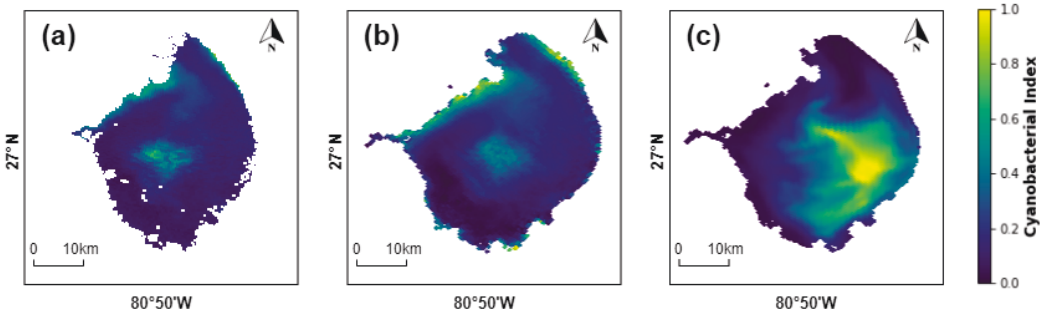


Fig. 10: Single-day prediction. Using data from June 19 to July 2, 2020 as input to predict the result for July 3, 2020. (a): Remote sensing data as the ground truth. (b): Prediction result using the hybrid dataset for training. (c): Prediction result using the physics-based model dataset for training.

the real situation, while the model predictions using the hybrid dataset are closer to the ground truth (i.e., the remote sensing data).

V. Conclusion and Future Work

Our study shows that the HAB prediction model built by ConvLSTM was effective and that ConvLSTM can capture the spatio-temporal dependencies for prediction. Moreover, multi-source data fusion as a new solution to the problems of a large number of missing, discontinuous, and low-quality remote sensing datasets in prediction was proposed. The hydrodynamic model was combined with the biological model to create a physical model to simulate the growth and spread of algal blooms, generating a physics-based model dataset that is continuous and clean. Due to the discrepancy between the simulation results of the physics-based model and the real situation, the physics-based model dataset was combined with the remote sensing dataset to generate a hybrid dataset for physics-informed data-driven modeling. It turns out that the prediction results of the model trained using the hybrid dataset are close to the remote sensing data, proving that data fusion is a promising approach.

There are also plans to improve the quality of the prediction, such as fixing the blurred prediction images caused by the model rolling prediction, improving the image blur caused by downsampling, and increasing the quantity of usable remote sensing data by patching and reconstructing the corrupt images. In terms of data fusion, a simple interpolation was used for this experiment, and the edges of some data are not as smooth as the original images. This issue will be addressed in future work.

Wind and temperature play a key role in driving the phy-

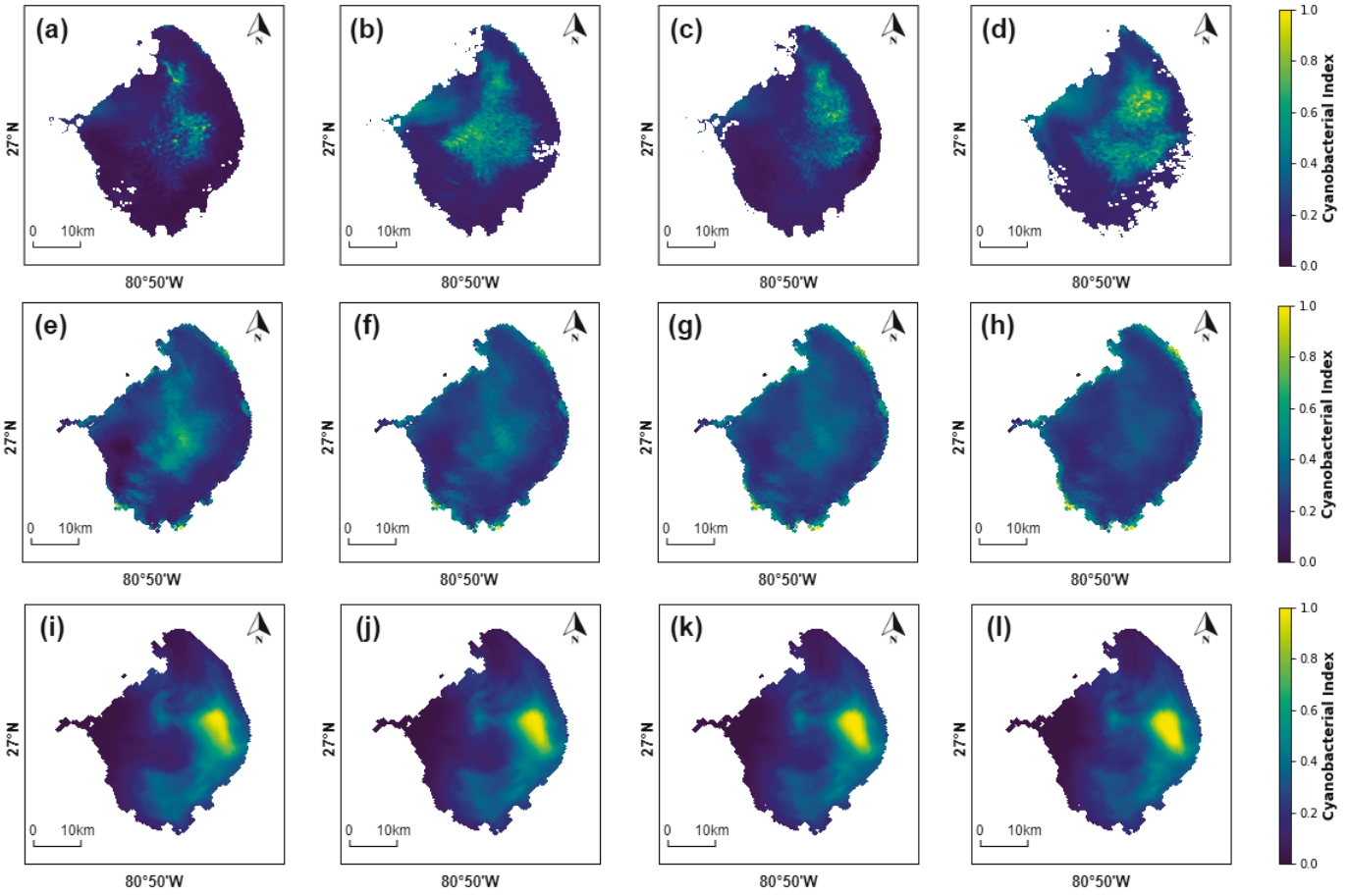


Fig. 11: 14-day prediction. Using data from June 8 to June 21, 2020 as input to predict the 14-day results from June 22 to July 5, 2020. The first 4 days (June 22 to June 25) of the 14-day predictions are shown. (a-d): Remote sensing data as Ground Truth. (e-h): Prediction results using the hybrid dataset for training. (i-l): Prediction results using the physics-based model dataset for training.

toplankton blooms, mainly M. aeruginosa. It appears that M. aeruginosa migration, while important, is not strong enough to overcome strong wind mixing. Field observations suggest that most of the strongest bloom events occur under low wind conditions. This is particularly true for the central lake where water is deeper and strong winds should disrupt any potential blooms. Temperature is also important in affecting the growth directly. One combined effect of these two factors is that strong blooms mostly occur in the summer and early fall when water is warm and winds are typically weak. In other word, the growing patterns of algal are seasonal dependent, and this factor will also be considered when developing the prediction model in the future.

ACKNOWLEDGEMENT

We thank Dr. Ashley Brereton for assistance with physic-based model simulation, and Lindsay Steis for help with the dataset cleaning and pre-processing.

APPENDIX

Mathematical Background of ConvLSTM: ConvLSTM is a deep learning model built from the long-short term memory network and convolutional neural network. The idea is the same as LSTM, which utilizes the previous layer's output as the input for the following layer (i.e., recurrent operation). The most significant change is that each layer's weight computations are a convolutional operation. The inner structure of ConvLSTM is shown in Fig. 12. With the addition of the convolution operation, ConvLSTM can not only establish a timing relationship similar to LSTM but also has a spatial feature extraction capability similar to CNN. The key equations of ConvLSTM are shown below [15]:

$$\begin{aligned} i_t &= \sigma\left(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i\right) \\ f_t &= \sigma\left(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f\right) \\ C_t &= f_t \circ C_{t-1} + i_t \circ \tanh\left(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c\right) \\ o_t &= \sigma\left(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o\right) \\ H_t &= o_t \circ \tanh\left(C_t\right) \end{aligned} \quad (17)$$

where σ is the logistic sigmoid function, t represents the time step, i, f, and o are respectively the input gate, forget gate, output gate, X is the input, H is the hidden state, C is the cell output, and b is the bias. W is the weight matrix, where different subscripts have different meanings, for example, W_{hi}

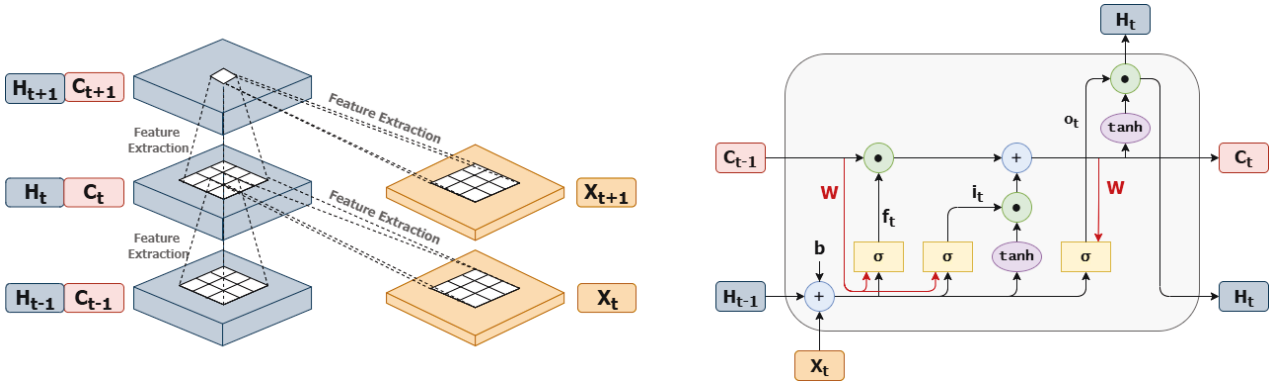


Fig. 12: Inner structure of the ConvLSTM. Left shows the convolutional operation and right shows the recurrent network.

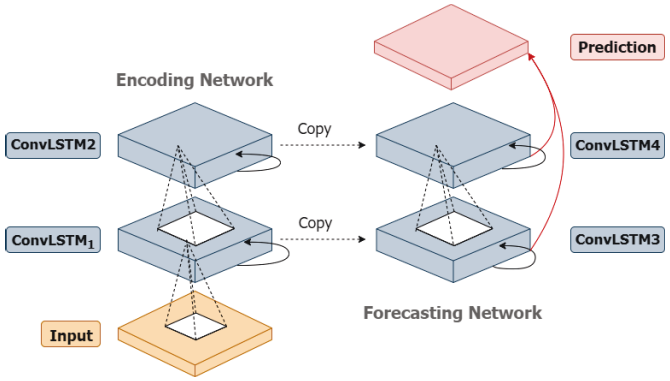


Fig. 13: Encoding-forecasting ConvLSTM structure for capturing the spatio-temporal dependencies.

is the hidden-input gate matrix. The $*$ is the convolution operator and $\circ$ is the Hadamard product.

The encoding-forecasting ConvLSTM structure illustrated in Fig. 13 is used to solve the spatio-temporal prediction. It is made up of two networks: an encoding network and a forecasting network. The forecasting network's initial state and cell output are replicated from the encoding network's final state. Both networks are created by stacking several ConvLSTM layers. All states are connected in the forecasting network and fed to the 1*1 convolutional layer to construct the final prediction since the target and input of spatio-temporal prediction generally have the same dimensions.

REFERENCES

[1] M. Olokotum, V. Mitroi, M. Troussellier, R. Semyalo, C. Bernard, B. Montuelle, W. Okello, C. Quiblier, and J.-F. Humbert, "A review of the socioecological causes and consequences of cyanobacterial blooms in lake victoria," *Harmful Algae*, vol. 96, p. 101829, 2020.

[2] Y. Zhang, R. Ma, H. Duan, S. A. Loiselle, J. Xu, and M. Ma, "A novel algorithm to estimate algal bloom coverage to subpixel resolution in lake taihu," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 7, pp. 3060–3068, 2014.

[3] E. J. Phlips, S. Badylak, N. G. Nelson, and K. E. Havens, "Hurricanes, el niño and harmful algal blooms in two sub-tropical florida estuaries: Direct and indirect impacts," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[4] W. W. Carmichael, "Health effects of toxin-producing cyanobacteria:"the cyanohabs"," *Human and ecological risk assessment: An International Journal*, vol. 7, no. 5, pp. 1393–1407, 2001.

[5] R. Dawson, "The toxicology of microcystins," *Toxicon*, vol. 36, no. 7, pp. 953–962, 1998.

[6] C. C. Carey, B. W. Ibelings, E. P. Hoffmann, D. P. Hamilton, and J. D. Brookes, "Eco-physiological adaptations that favour freshwater cyanobacteria in a changing climate," *Water research*, vol. 46, no. 5, pp. 1394–1407, 2012.

[7] S. Lee and D. Lee, "Improved prediction of harmful algal blooms in four major south korea's rivers using deep learning models," *International journal of environmental research and public health*, vol. 15, no. 7, p. 1322, 2018.

[8] A. W. Griffith and C. J. Gobler, "Harmful algal blooms: a climate change co-stressor in marine and freshwater ecosystems," *Harmful Algae*, vol. 91, p. 101590, 2020.

[9] H. B. Kim, S. Cho, Y. Lee, W. Wu, and U.-H. Ha, "Weigela florida inhibits the expression of inflammatory mediators induced by pseudomonas aeruginosa and staphylococcus aureus infection," *Journal of Microbiology*, pp. 1–8, 2022.

[10] R. Elkadiri, C. Manche, M. Sultan, A. Al-Dousari, S. Uddin, K. Chouinard, and A. Z. Abotalib, "Development of a coupled spatiotemporal algal bloom model for coastal areas: A remote sensing and data mining-based approach," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 11, pp. 5159–5171, 2016.

[11] B. Gokaraju, S. S. Durbha, R. L. King, and N. H. Younan, "A machine learning based spatio-temporal data mining approach for detection of harmful algal blooms in the gulf of mexico," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 4, no. 3, pp. 710–720, 2011.

[12] X. Li, J. Yu, Z. Jia, and J. Song, "Harmful algal blooms prediction with machine learning models in tolo harbour," in *2014 International Conference on Smart Computing*. IEEE, 2014, pp. 245–250.

[13] W. Song, J. M. Dolan, D. Cline, and G. Xiong, "Learning-based algal bloom event recognition for oceanographic decision support system using remote sensing data," *Remote Sensing*, vol. 7, no. 10, pp. 13 564–13 585, 2015.

[14] P. R. Hill, A. Kumar, M. Temimi, and D. R. Bull, "Habnet: Machine learning, remote sensing-based detection of harmful algal blooms," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3229–3239, 2020.

[15] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[16] T. Lee, J.-H. Choi, M. Jang, J. Won, and J. Kim, "Enhancing prediction of chlorophyll-a concentration with feature extraction using higher-order partial least squares," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2020, pp. 1666–1668.

[17] Y. Choi, Y. Park, W.-A. Lim, S.-H. Min, and J.-S. Lee, "Convolution neural network for the prediction of cochlodinium polykrikoides bloom in the south sea of korea," *Journal of Marine Science and Engineering*, vol. 10, no. 1, p. 31, 2021.

[18] D. Blondeau-Patissier, J. F. Gower, A. G. Dekker, S. R. Phinn, and V. E. Brando, "A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans," *Progress in oceanography*, vol. 123, pp. 123–144, 2014.

[19] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei, "Missing data reconstruction in remote sensing image with a unified spatial–temporal–

spectral deep convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4274–4288, 2018.
[20] D. Xu, Y. Pu, M. Zhu, Z. Luan, and K. Shi, "Automatic detection of algal blooms using sentinel-2 msi and landsat oli images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8497–8511, 2021.
[21] L. Shen, H. Xu, and X. Guo, "Satellite remote sensing of harmful algal blooms (habs) and a potential synthesized framework," *Sensors*, vol. 12, no. 6, pp. 7778–7803, 2012.
[22] N.-B. Chang, B. Vannah, and Y. J. Yang, "Comparative sensor fusion between hyperspectral and multispectral satellite sensors for monitoring microcystin distribution in lake erie," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2426–2442, 2014.
[23] A. F. Shchepetkin and J. C. McWilliams, "The regional oceanic modeling system (roms): a split-explicit, free-surface, topography-following-coordinate oceanic model," *Ocean modelling*, vol. 9, no. 4, pp. 347–404, 2005.
[24] K. Fennel, J. Wilkin, J. Levin, J. Moisan, J. O'Reilly, and D. Haidvogel, "Nitrogen cycling in the middle atlantic bight: Results from a three-dimensional model and implications for the north atlantic nitrogen budget," *Global Biogeochemical Cycles*, vol. 20, no. 3, 2006.
[25] C. Donlon, B. Berruti, A. Buongiorno, M.-H. Ferreira, P. Féménias, J. Frerick, P. Goryl, U. Klein, H. Laur, C. Mavrocordatos *et al.*, "The global monitoring for environment and security (gmes) sentinel-3 mission," *Remote sensing of Environment*, vol. 120, pp. 37–57, 2012.
[26] T. Wynne, R. Stumpf, M. Tomlinson, R. Warner, P. Tester, J. Dyble, and G. Fahnenstiel, "Relating spectral shape to cyanobacterial blooms in the laurentian great lakes," *International Journal of Remote Sensing*, vol. 29, no. 12, pp. 3665–3672, 2008.
[27] T. S. Moore, C. B. Mouw, J. M. Sullivan, M. S. Twardowski, A. M. Burtner, A. B. Ciochetto, M. N. McFarland, A. R. Nayak, D. Paladino, N. D. Stockley *et al.*, "Bio-optical properties of cyanobacteria blooms in western lake erie," *Frontiers in Marine Science*, vol. 4, p. 300, 2017.
[28] T. Wynne, A. Meredith, R. Stumpf, T. Briggs, and W. Litaker, "Harmful algal bloom forecasting branch ocean color satellite imagery processing guidelines, 2020 update," 2020.
[29] K. G. Ruddick, F. Ovidio, and M. Rijkeboer, "Atmospheric correction of seawifs imagery for turbid coastal and inland waters," *Applied optics*, vol. 39, no. 6, pp. 897–912, 2000.
[30] K. G. Ruddick, V. De Cauwer, Y.-J. Park, and G. Moore, "Seaborne measurements of near infrared water-leaving reflectance: The similarity spectrum for turbid waters," *Limnology and Oceanography*, vol. 51, no. 2, pp. 1167–1179, 2006.
[31] B. Nechad, K. Ruddick, and G. Neukermans, "Calibration and validation of a generic multisensor algorithm for mapping of turbidity in coastal waters," in *Remote Sensing of the Ocean, Sea Ice, and Large Water Regions 2009*, vol. 7473. SPIE, 2009, pp. 161–171.
[32] A. I. Dogliotti, K. Ruddick, B. Nechad, D. Doxaran, and E. Knaeps, "A single algorithm to retrieve turbidity from remotely-sensed data in all coastal and estuarine waters," *Remote sensing of environment*, vol. 156, pp. 157–168, 2015.
[33] J. Li, Y. Zhang, R. Ma, H. Duan, S. Loiselle, K. Xue, and Q. Liang, "Satellite-based estimation of column-integrated algal biomass in non-algae bloom conditions: A case study of lake chaohu, china," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 2, pp. 450–462, 2016.
[34] M. Fisher, K. Reddy, and R. T. James, "Internal nutrient loads from sediments in a shallow, subtropical lake," *Lake and Reservoir Management*, vol. 21, no. 3, pp. 338–349, 2005.
[35] K. Fennel, R. Hetland, Y. Feng, and S. DiMarco, "A coupled physical-biological model of the northern gulf of mexico shelf: model description, validation and analysis of phytoplankton variability," *Biogeosciences*, vol. 8, no. 7, pp. 1881–1899, 2011.
[36] K. Team, "Keras documentation: Next-frame video prediction with convolutional lstms," keras.io. [Online]. Available: https://keras.io/examples/vision/conv_lstm/
[37] C. W. Zhang, "Spatial-temporal convlstm for crash prediction," Medium, 11 2021. [Online]. Available: https://towardsdatascience.com/spatial-temporal-convlstm-for-crash-prediction-411909ed2cfa
[38] N. Ekhtiari, "Comparing ground truth with predictions using image similarity measures." [Online]. Available: https://up42.com/blog/tech/image-similarity-measures?utm_content=151794785&utm_medium=social&utm_source=facebook&hss_channel=fbp-381810672402775
[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
[40] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.
[41] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler, "Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 146, pp. 305–319, 2018.

Yufei Tang (SM'22) received the Ph.D. degree in Electrical Engineering from the University of Rhode Island, Kingston, RI, USA, in 2016. He is currently an Associate Professor with the Department of Electrical Engineering and Computer Science, and a Faculty Fellow with the Institute for Sensing and Embedded Network Systems Engineering, Florida Atlantic University, Boca Raton, FL, USA. His research interests include machine learning, data mining, dynamical systems, and renewable energy.

Dr. Tang was a recipient of the IEEE International Conference on Communications Best Paper Award in 2014, the National Academies Gulf Research Program Early-Career Research Fellowship in 2019, and the U.S. National Science Foundation CAREER Award in 2022.

Yingqi Feng received the B.S. degree in Computer Science from Florida Atlantic University, Boca Raton, FL, USA, in 2021.

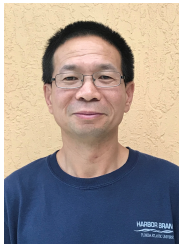
She is currently working toward the Ph.D. degree in Computer Science, and is a Graduate Research Assistant with the Institute for Sensing and Embedded Network Systems Engineering, Florida Atlantic University, Boca Raton, FL, USA. Her research interests include machine learning, graph learning, data mining, and applications in remote sensing and medical images.

Sasha Fung received the B.S. degree in Computer Engineering from Florida Atlantic University, Boca Raton, FL, USA, in 2020.

She is currently working toward the Ph.D. degree in Electrical Engineering, and is a Graduate Research Assistant with the Institute for Sensing and Embedded Network Systems Engineering, Florida Atlantic University, Boca Raton, FL, USA. Her research interests include machine learning, smart grid security, and marine renewable energy.

Veronica Ruiz Xomchuk received the Ph.D. degree in Oceanography at Texas A&M University, College Station, TX, USA, in 2020. She is now a Senior Research Fellow in Oceanographic and Environmental Data Science at FAU's Harbor Branch Oceanographic Institute, Fort Pierce, FL, USA.

Her research interest is in physical drivers of biogeochemical processes in coastal environments, including HAB's, hypoxia development and the carbon cycle. She is also interested in computational oceanography developing python tools for analysis of oceanographic and environmental data from models, and fixed and mobile continuous recording platforms.



Mingshun Jiang received his Ph.D. degree in Physical Oceanography from Ocean University of Qingdao (now Ocean University of China) in 1994. He is currently an Associate Research Professor with Harbor Branch Oceanographic Institute, Florida Atlantic University.

His research interests focus on modeling coastal and regional ocean dynamics, biogeochemical cycles including nutrients, carbon and Fe cycle, and phytoplankton blooms. His recent focus has been developing numerical models for modeling water quality and harmful algal blooms (HABs) in south Florida freshwater and estuaries including Lake Okeechobee, Indian River Lagoon, and Florida Bay.

Tim Moore is a Research Associate Professor at FAU's Harbor Branch Oceanographic Institute, Fort Pierce, FL, USA. He has been working with ocean color satellite imagery for over 20 years. He has been part of the NASA MODIS Science Team, participated in work groups on reports for the International Ocean Color Coordinating Group (IOCCG), has been an instructor at many IOCCG and NASA workshops on using ocean color remote sensing in research.

His research has focused on bio-optical algorithms for freshwater and marine environments, developed an optical water type framework as a way to integrate multiple algorithm tuned for specific waters, and developed mapped uncertainties for a variety of ocean color products, including NASA's MEaSUREs program. His software has been embedded in NASA's SEADAS software and ESA's BEAM software which are satellite processing systems for ocean color data.

Jordon Beckler received a Ph.D. degree in Earth and Atmospheric Sciences with a minor in Inorganic Chemistry from the Georgia Institute of Technology in 2014. He is currently an Assistant Research Professor and holds a joint appointment with Florida Atlantic University's Harbor Branch Oceanographic Institute and the FAU Institute for Sensing and Embedded Network Systems Engineering (I-SENSE).

As the Principal Investigator of the Geochemistry and Geochemical Sensing Lab, his research interests primarily pertain to the development and application of autonomous sensors for lacustrine, estuarine, or marine monitoring. Between 2020 and 2022, he was the lead Principal Investigator of the Harmful Algal Bloom Assessment of Lake Okeechobee (HALO), multi-institution effort to understand and develop a predictive capacity for these detrimental events in a critically-sensitive Florida ecosystem.