From Remote Sensing to Multiple Time Horizons Forecasts: Transformers Model for CyanoHAB Intensity in Lake Champlain

Muhammad Adil, Patrick J. Clemins, Andrew W. Schroth, Panagiotis D. Oikonomou, Donna M. Rizzo, Peter D. F. Isles, Xiaohan Zhang, Kareem I. Hannoun, Scott Turnbull, Noah B. Beckage, Asim Zia, and Safwan Wshah

***Abstract*—Cyanobacterial Harmful Algal Blooms (CyanoHABs) pose significant threats to aquatic ecosystems and public health globally. Lake Champlain is particularly vulnerable to recurring CyanoHAB events, especially in its northern segment: Missisquoi Bay, St. Albans Bay, and Northeast Arm, due to nutrient enrichment and climatic variability. Remote sensing provides a scalable solution for monitoring and forecasting these events, offering continuous coverage where in-situ observations are sparse or unavailable. In this study, we present a remote sensing-only forecasting framework that combines Transformers and BiLSTM to predict CyanoHAB intensities up to 14 days in advance. The system utilizes Cyanobacterial Index data from the Cyanobacterial Assessment Network and temperature data from Moderate Resolution Imaging Spectroradiometer satellites to capture long-range dependencies and sequential dynamics in satellite time series. The dataset is very sparse, missing more than 30% of the Cyanobacterial Index data and 90% of the temperature data. A two-stage preprocessing pipeline addressed data gaps by applying forward fill and weighted temporal imputation at the pixel level, followed by smoothing to reduce the discontinuities of CyanoHAB events. The raw dataset is transformed into meaningful features through equal-frequency binning for the Cyanobacterial Index values and extracted temperature statistics. Transformer-BiLSTM model demonstrates strong forecasting performance across multiple horizons, achieving F1 scores of 89.5%, 86.4%, and 85.5% at one-, two-, and three-day forecasts, respectively, and maintaining an F1 score of 78.9% with an AUC of 82.6% at the 14-day horizon. These results confirm the model's ability to capture complex spatiotemporal dynamics from sparse satellite data and to provide reliable early warning for CyanoHABs management. The dataset and trained model weights are publicly available, facilitating further research in CyanoHABs forecasting.**

***Index Terms*—Transformers Based Modeling, Short and long range CyanoHABs Forecasting, CyanoHABs Intensity Prediction, Remote Sensing, Water Quality Monitoring, Early Warning System**

I. Introduction

Cyanobacterial harmful algal blooms (CyanoHABs) pose a significant environmental challenge that affects water quality, aquatic ecosystems, and public health in inland waters worldwide. CyanoHABs are characterized by the excessive growth of cyanobacteria in inland water systems, which produce toxins harmful to humans, ecological health, and the broader environment. Their formation is primarily attributed to eutrophication [1], and is exacerbated by anthropogenic activities such as intensive agriculture [2], [3] and insufficient policy actions to preserve the water quality of lake bodies [4]. The combined effect of eutrophication and anthropogenic activities leads to an increase in nutrient fluxes, accompanied by lighting conditions, and water temperature favorable to cyanobacteria that can lead to more severe and extensive CyanoHABs [4]–[6]. CyanoHABs can drastically degrade inland-water quality [7], [8], threaten public health [9]–[11], decrease biodiversity, contribute to habitat loss [12], which in turn causes damage to socioeconomic and ecological aspects of aquatic ecosystems [13]. In freshwater systems, CyanoHABs species that produce biotoxins are of particular concern as these toxins can accumulate in aquatic food webs, when consumed by humans, leading to severe illness, paralysis, and even death [14]. According to a scientific research summary from [15], CyanoHABs cost $4.6B per year in estimated damages within the United States alone. Early and reliable CyanoHABs forecasting systems can help mitigate adverse effects or even prevent them.

Numerous studies have identified temperature as a factor in the proliferation of CyanoHABs [16]–[25]. Other environmental factors include precipitation [19], [26]–[28], nutrient concentration [29]–[31], wind conditions [32], [33], and hydraulic retention time [34]–[36]. Although these CyanoHAB drivers are well understood, early CyanoHAB forecasting remains challenging, as environmental factors vary spatially and temporally [37].

Given these growing concerns, numerous CyanoHABs forecasting models have been developed in recent years. However,

M. Adil, X. Zhang, and S. Wshah are with Vermont Artificial Intelligence Laboratory (VaiL), Department of Computer Science, University of Vermont, Burlington, VT 05405 USA.

A. Zia is with Department of Community Development and Applied Economics & Department of Computer Science, University of Vermont, Burlington, VT 05405 USA.

A. W. Schroth is with Department of Geology, University of Vermont, Burlington, VT 05401 USA.

P. D. Oikonomou and D. M. Rizzo are with Department of Civil and Environmental Engineering, University of Vermont, Burlington, VT 05405 USA.

P. J. Clemins, S. Turnbull, and N. B. Beckage are with Vermont EPSCoR (Established Program to Stimulate Competitive Research), University of Vermont, Burlington, VT 05405 USA.

P. D. F. Isles is with Vermont Department of Environmental Conservation, Montpelier, VT 05602 USA.

K. I. Hannoun is with Water Quality Solutions, Inc., McGaheysville, VA 22840 USA.

Corresponding Author: Safwan Wshah, e-mail: safwan.wshah@uvm.edu.

Manuscript received Month DD, YYYY; revised Month DD, YYYY.

the majority of these algorithms rely on in-situ sensors [37]–[40] or manual water sampling [41]–[43], both of which are costly and labor-intensive. These data sources are also limited in their ability to capture large spatial areas, thereby constraining the detection of temporal and spatial changes in CyanoHABs, which are crucial for understanding their formation.

Remote sensing offers a promising solution to the limitations of sensor-based and manual sampling methods. Recent technological advances in remote sensing can transform the CyanoHABs monitoring landscape by providing data at higher temporal and spatial resolutions, enabling comprehensive and sophisticated modeling of the dynamic nature of CyanoHABs.

Many researchers have used remote sensing to forecast and predict harmful algal blooms across aquatic systems over the past decade [44]–[48]. Early demonstrations [44] focused on marine red tides using data from the Coastal Zone Color Scanner aboard Nimbus-7 in 1981. Although biologically distinct from freshwater cyanobacterial blooms, this pioneering work established the remote-sensing foundations later adapted for inland water monitoring and influenced the development of sensors such as Moderate Resolution Imaging Spectroradiometer (MODIS), Medium Resolution Imaging Spectrometer (MERIS), and the Sentinel satellites. Subsequent work applied these sensors to freshwater CyanoHAB forecasting using a range of machine-learning and statistical methods. [45] utilized the MODIS data with XGBoost, Random Forest, and Support Vector Machines to predict CyanoHAB occurrences with lead times of 1 to 11 days. A sub-monthly CyanoHAB forecasting model was developed by incorporating remote sensing and environmental factors to predict CyanoHAB cell counts [47]. Satellite-derived chlorophyll-*a* data were combined with univariate time-series models, enabling 1 to 4 weeks of CyanoHABs forecast across 15 lakes worldwide [48]. A regression-based model was developed to predict the occurrence, spatial distribution, and intensity of CyanoHABs using chlorophyll-*a* concentration imagery from MODIS satellites [46].

Although these studies offer significant improvements for CyanoHABs detection and forecasting, they have at least one of the following drawbacks: 1) Limited transferability due to a reliance on ground sensor data in addition to remote sensing data, restricting the model's scalability to regions without ground sensing data; and 2) Most of these studies do not include CyanoHABs intensities. Predicting the intensity of CyanoHABs has a significant role in effective environmental management and public health protection, and early awareness helps inform public policy and protect human health.

To address the aforementioned limitations, we present a Transformer-BiLSTM deep learning framework that combines a multi-head, self-attention Transformer [49] and Bidirectional Long Short-Term Memory (BiLSTM) architectures for extended-horizon CyanoHAB intensity forecasting, utilizing exclusively remote sensing data. The proposed Transformer-BiLSTM model combines the global context modeling capability of Transformers with BiLSTM, further refining Transformers' contextual understanding by maintaining temporal continuity and smoothing short-term variations. Unlike prior approaches that are constrained to in situ data, the proposed framework relies solely on satellite-derived inputs. To overcome the inherent data sparsity challenges in satellite observations, the framework employs a two-stage preprocessing pipeline that combines temporal imputation and smoothing techniques. The model generates forecasts across five distinct intensity classes (Low, Medium, High, Very High, and Extreme Intensity CyanoHABs) with forecast horizons of 14 days. This approach effectively predicts the likelihood and severity of CyanoHAB events across multiple time scales. Fine-grained intensity forecasting with an extended lead time provides environmental agencies and public health authorities with timely and actionable insights for proactive risk assessment and response planning.

We summarize our contributions in six principal ways:

- We introduce a hybrid Transformer-BiLSTM architecture that integrates a multi-head, self-attention transformer and BiLSTM architectures for predicting the occurrence and intensity of CyanoHABs, enabling accurate forecasting up to 14 days, bridging short and medium-range early-warning needs.
- We formulate the task as a bin-wise, five-level intensity prediction, allowing managers to assess not only whether a bloom will occur but also its expected severity.
- Our approach exclusively utilizes remote sensing data for CyanoHAB forecasting, overcoming sensor-based limitations. To our knowledge, this is the first model to rely entirely on remote sensing for multi-horizon CyanoHAB incidence and intensity prediction.
- This framework represents the first CyanoHAB early-warning system for Lake Champlain and is transferable to other Contiguous United States freshwater systems using freely available remote sensing data with minimal retraining.
- We make our processed remote sensing dataset and preprocessing pipeline publicly available to facilitate future CyanoHABs research.
- The effectiveness of the Transformer-BiLSTM model is evaluated against Advanced Deep Learning and Machine Learning models. The proposed model demonstrates the superiority of CyanoHABs' early forecasting and intensity prediction.

II. Related Work

A. In-Situ Based CyanoHABs Forecasting

In-situ based CyanoHABs prediction relies on in-situ field monitoring and is sometimes coupled with automated sensors. These approaches often use parameters such as chlorophyll-*a* concentrations, dissolved oxygen, nutrient fluxes, water temperature, pH, and cyanobacteria cell counts to predict CyanoHAB occurrence or intensity. The research investigating the detrimental effects of excessive nutrients, specifically nitrogen and phosphorus, related increased CyanoHAB activity to the exceedance of a critical N:P molar ratio [30]. A 7-day forecasting model utilizing microcystin concentration as the primary indicator and incorporating input features such as pH, precipitation, temperature, dew point, and wind speed was developed to predict CyanoHABs in Iowa lakes [39]. A hybrid deep learning approach combining wavelet analysis with LSTM was developed to forecast the CyanoHABs in Lake Mendota [40]. The research used hourly, daily, and monthly chlorophyll concentrations and cyanobacterial cell biomass counts to forecast CyanoHABs at hourly, daily, and monthly resolutions. Boosted regression trees and artificial neural networks are used to investigate the impact of high-temporal-resolution physicochemical and meteorological data features on the growth and decline of CyanoHABs in Upper Klamath Lake, Oregon. The authors concluded that decreases in inflows and lake-surface elevation, and increases in temperature and phosphorus concentration, play a significant role in CyanoHAB growth. Machine learning models for forecasting CyanoHABs in coastal areas were studied using physical, geochemical, and climate data to predict chlorophyll concentration as a proxy for CyanoHABs in Biscayne Bay [50].

Recent works have begun incorporating attention mechanisms into CyanoHABs forecasting [51]–[53]. Reverse-time and dual-stage attention mechanisms are combined with recurrent imputation to achieve 1, 7, and 14-day CyanoHABs forecasts at Nakdong River in South Korea [52]. Stacked gradient-boosting trees with an attention-based CNN-LSTM demonstrated that attention consistently boosted ensemble skill over purely boosting or purely deep learning baselines for CyanoHABs forecasting [53]. Most recently, a Temporal Convolution Network fused with multi-head attention and BiLSTM was developed to predict hour-ahead chlorophyll-*a* from high-frequency buoy measurements [51]. While these works confirm the value of attention for capturing abrupt fluctuations, they remain sensor-bound and focus on either binary CyanoHABs status or chlorophyll proxies at a single location.

Given our focus on the Lake Champlain study area, reviewing previous CyanoHABs monitoring and forecasting studies conducted in this region is essential. A modified self-organizing map was used to analyze relationships between cyanobacteria blooms and environmental conditions in Lake Champlain's Missisquoi Bay [54]. The authors concluded that low dissolved nitrogen-to-soluble reactive phosphorus ratios and sediment anoxia strongly correlate with bloom conditions. Bowling et al. A 9-year phytoplankton dataset from Missisquoi Bay was analyzed using water samples collected from four sites through manual field sampling methods, laboratory analysis, and meteorological data, showing weak correlations between environmental factors and cyanobacterial distributions, with the heterogeneity of bloom patterns remaining largely unexplained [55]. High-frequency sensor data, field sampling, and long-term monitoring data were utilized to investigate the internal drivers of cyanobacteria blooms in Missisquoi Bay [56]. The authors found that internal phosphorus loading, thermal stratification, and shifting nutrient limitations play a vital role in bloom progression.

B. Remote Sensing CyanoHABs Detection

Remote sensing can capture the temporal and spatial variability of CyanoHABs. Extensive literature [3], [45]–[48], [57]–[60] describes the monitoring, detection, and forecasting of CyanoHABs using remote sensing. MERIS, Sentinel-2, Sentinel-3, and satellites equipped with moderate-resolution imaging spectroradiometers (MODIS-Aqua and MODIS-Terra) are transforming the CyanoHABs monitoring and modeling landscape by providing higher-frequency, open-source data. They can capture detailed images and information about water bodies, enabling researchers to effectively detect and track the development of CyanoHABs.

Four machine learning models were compared for forecasting CyanoHABs in Lake Erie at 10-, 20-, and 30-day lead times [47]. The author used satellite-derived CI data alongside meteorological, hydrodynamic, and nutrient data as inputs to predict future CI values as a proxy for CyanoHABs intensity. A. Gupta [47] compared four machine learning models for forecasting harmful algal blooms in Lake Erie at 10-, 20-, and 30-day lead times. A simple, generalizable univariate forecasting model was developed to predict cyanobacterial blooms using satellite-derived chlorophyll-*a* data [48]. Forecasts were generated at 1-, 2-, and 4-week horizons for 15 lakes worldwide. The 1-week model accurately predicted high-risk blooms with 80% accuracy, comparable to that of complex models. Landsat-8 and Sentinel-2 satellite imagery were used to estimate chlorophyll-*a*, turbidity, and phycocyanin concentrations using support vector and random forest regression models [3]. The impact of spatial resolution on the detection of CyanoHABs was investigated using data from Landsat-8, Sentinel-2, and PlanetScope imagery [59]. The authors evaluated twenty bio-optical algorithms from earlier studies to predict chlorophyll concentration, biomass, phyco-cyanin concentration, etc. A spatiotemporal modeling approach utilizing Bayesian hierarchical models was presented to predict the likelihood of CyanoHABs in Florida's freshwater systems [60]. The authors used weekly composite images of maximum cyanobacteria abundance from Sentinel-3 imagery, along with ambient temperature, surface water temperature, and precipitation data. A collaborative framework for monitoring and predicting harmful algal bloom accumulation in nearshore areas of Lake Chaohu was developed using a coupled hydrodynamic-water quality-algae model [61]. The authors integrated satellite remote sensing data, 42 land-based video monitoring devices providing hourly coverage, and in-situ buoy stations measuring chlorophyll-*a* and lake currents. The framework provided real-time quantitative monitoring and daily predictions of harmful algal bloom accumulation risks in nearshore zones to support emergency lake management. An extreme gradient boosting algorithm was used to estimate chlorophyll concentrations as an indicator of CyanoHABs by training on in-situ chlorophyll measurements and using reflectance indices from Landsat-8 Operational Land Imager (OLI) data along with other environmental inputs [62]. Recently, [63] presented a multi-source monitoring and prediction framework for CyanoHABs, integrating satellite, UAV, ground-based, and in-situ systems to enable high-frequency, nearshore-focused assessment in Lake Chaohu. By simulating the growth-drift-accumulation dynamics, the presented method significantly improves prediction accuracy and practical early warning capabilities in eutrophic lakes.

Lake Champlain has been the subject of numerous

CyanoHAB studies. Empirical regression models were used to map in-situ data with corresponding Landsat bands and blue, green, and red band ratios [64]. Pigment concentration retrieval methods were evaluated in Missisquoi Bay, Lake Champlain, demonstrating that empirically calibrated QuickBird data using NIR/Red band ratios could explain approximately 80% of chlorophyll-*a* variability while achieving an R^2 of 0.68 for Phycocyanin concentration mapping [65]. CyanoHABs in Lake Champlain were mapped using Landsat-8 Operational Land Imager, Rapid Eye, and Proba Compact High-Resolution Imaging Spectrometer [43]. CyanoHABs were detected using band ratios to identify spatial patterns within the lake and problematic bays with high phycocyanin concentrations. MERIS and CI are used to estimate the cyanobacteria cell counts (cells/mL) across eastern United States (US) lakes [66]. The authors identified good performance at low (10,000 to 109,000) and high ($>$ 1,000,000) cell counts; however, the approach performed suboptimally at intermediate concentrations. An ensemble-based system using down-scaled MODIS imagery was developed to estimate chlorophyll-*a* concentrations in inland lakes [67]. The system combines adaptive modeling and Gaussian quadrature to minimize classification and estimation errors, thereby enhancing accuracy across bloom stages. Recently, the variability of CyanoHABs in Missisquoi Bay, Lake Champlain, was analyzed using satellite imagery from OLCI and MODIS with the CI algorithm to assess temporal and spatial occurrences of CyanoHABs [58].

While previous studies have enhanced our understanding of CyanoHABs' spatial and temporal distribution in Lake Champlain, forecasting is limited due to reliance on in situ sensor data and the limited temporal resolution of site-specific remote sensing data. This dependence on location-specific measurements restricts the transferability of the model to other water bodies. Developing advanced forecasting systems based exclusively on widely available remote sensing data is crucial for creating models that can be deployed across diverse freshwater ecosystems with minimal adaptation.

III. Study Area

Lake Champlain is a glacially-formed lake that consists of more than 70 islands and 54 beaches with annual economic revenues of $580 million [68]–[70]. Lake Champlain is the 13th-largest lake in the US, situated between New York and Vermont, and extending into Canada to the North. The lake is 120 miles long with a surface area of 1127 km^2 and a volume of 25.8 km^3 [71]. It is 19km at its widest point, with an estimated maximum and mean depth of 122m and 23m, respectively [70]. With more than 800km of shoreline and 21000 km^2 in watershed area, it supports an approximate population of 571,000 [68]. Despite its relatively small size, the lake drainage basin is quite significant in comparison, with a drainage ratio of 19:1 [72]. The large drainage basin makes Lake Champlain susceptible to severe precipitation events and snowmelt, with runoff from agriculture, industry, and urbanization all contributing to eutrophication, which subsequently leads to the growth of CyanoHABs. Figure 1 shows the approximate number of CyanoHABs events, at each of the 12 Lake Champlain monitoring segments from January 2016 to November 2023.

This study focuses on three lake segments: Missisquoi Bay, St. Albans Bay, and the Northeast Arm. The Northeast Arm is categorized as mesotrophic, indicative of relatively low nutrient levels and biological productivity. It makes up about one-quarter of Lake Champlain's surface area and has a mean depth of 13m. Missisquoi Bay and St. Albans Bay are eutrophic zones with elevated nutrient concentrations and biological productivity. These areas, consequently, experience the highest annual incidences of CyanoHABs, as reported in Figure 1. Both bays - Missisquoi Bay and St. Albans Bay - are shallow and warm, providing favorable conditions for CyanoHABs growth. Missisquoi Bay has a mean depth of 2.5m, while St. Albans Bay has a mean depth of 8m [73].

IV. Data Collection & Processing

We use satellite remote sensing time series from two sources: the Cyanobacterial Index from the Cyanobacterial Assessment Network (CyAN) [74], derived from ocean-color satellite imagery at 300 m spatial resolution (daily and 7-day maximum composites), and Moderate Resolution Imaging Spectroradiometer (MODIS) [75] land surface temperature (LST) at 1 km with separate day/night daily retrievals. The study period spans January 2016 to November 2024 for three Lake Champlain segments - Missisquoi Bay, St. Albans Bay, and the Northeast Arm. All inputs are in the form of GeoTiff, where shallow water pixels are masked, and aggregated to segment–day records with statistical features extracted from LST imagery. Further details are provided in the next four subsections:1) Data Collection - downloading CI data from CyAN and temperature data from MODIS, 2) Data sparsity handling - addressing substantial data sparsity through temporal imputation methods, 3) Data Extraction - extracting model ready features with five CI intensity categories and ten temperature statistics, 4) Class Balancing - temporal sampling for class balancing to focus on high activity months with later addition of negative samples from inactive months. Figure 2 illustrates the complete data pipeline steps visually.

A. Data Collection

1) Cyanobacterial Index Values: The Cyanobacterial Index (CI), was initially developed by Wynne et al. [76] and refined in [77] to detect large, monospecific cyanobacteria blooms in Lake Erie. The CI exploits the unique pigment composition of cyanobacteria, particularly their phycocyanin accessory pigments and reduced chlorophyll-*a* fluorescence, which distinguishes them from eukaryotic phytoplankton. Chlorophyll absorption is calculated by spectral analysis around 681 nm, leveraging top-of-atmosphere reflectance measured at specific wavelengths to determine the spectral shape and, subsequently, the CI value. However, limitations were encountered when estimating CI values, as highlighted first by the estimation of positive CI values without cyanobacteria measurements in Chesapeake Bay, Green Bay, and several New England lakes [78]. This was addressed by [66], who introduced a conditional equation adjustment, resulting in the CI$_{Cyano}$ algorithm.

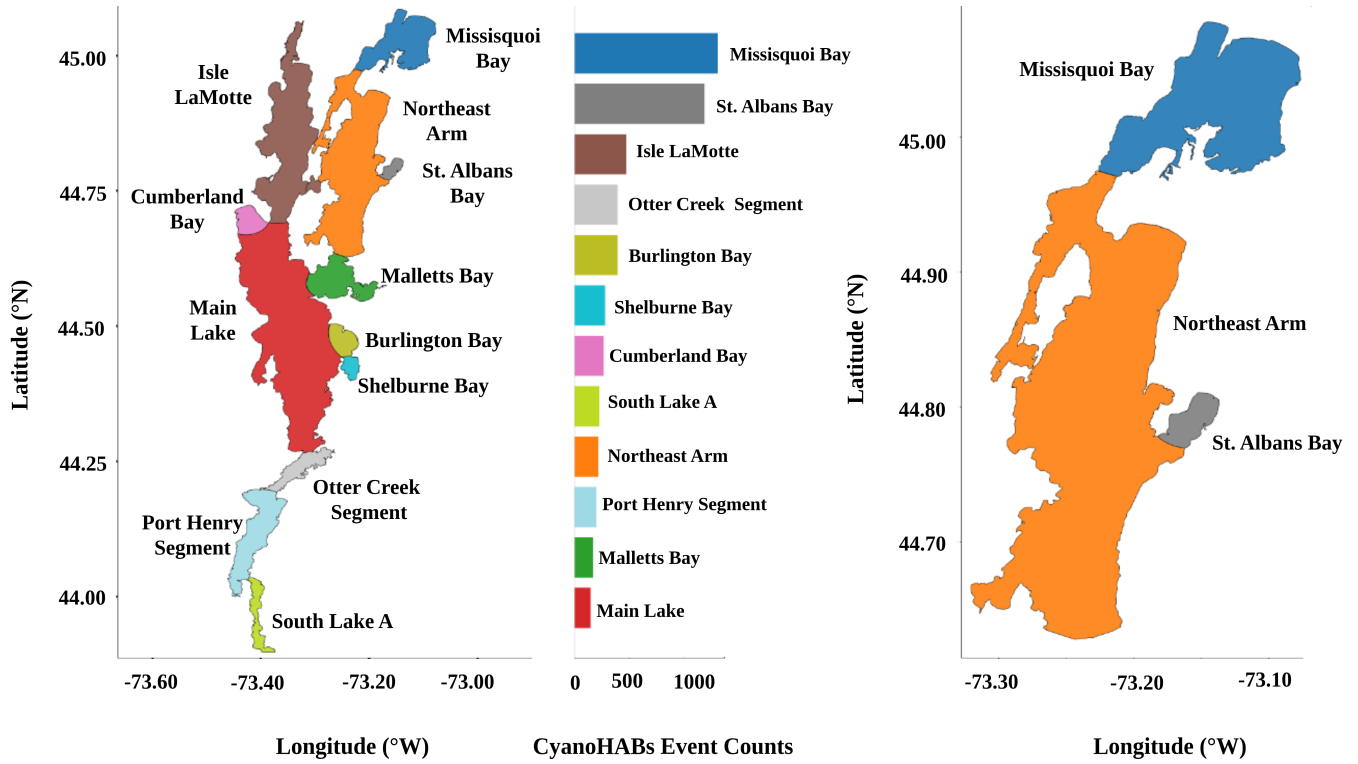

Fig. 1. The visual illustration of Lake Champlain shows 12 segments and their CyanoHABs events as a Bar Chart on the right side from 2016 to 2023. We can see that Missisquoi Bay and St. Albans Bay experienced an extensive number of CyanoHABs during that duration. The right side map shows the three stations - Missisquoi Bay, St. Albans Bay, and Northeast Arm - used in the study.

The CI$_{Cyano}$ algorithm is widely validated for CyanoHABs monitoring in freshwater systems. Mishra [79] validated CI$_{Cyano}$ for detecting toxin-producing cyanobacterial blooms across 30 U.S. lakes using MERIS and Sentinel-3 OLCI satellite data matched with field-measured microcystin concentrations. Using over 280 matchups from 2005-2019 and a microcystin threshold of $0.2\,\mu$g/L, the algorithm achieved 84% accuracy with 90% recall and 87% precision when validated against both microcystin and cyanobacteria cell density. Coffer [80] conducted a national-scale assessment of cyanobacterial blooms across 2,321 U.S. lakes, revealing strong seasonal and regional variability with typical summer-fall peaks. Wynne [58] applied CI$_{Cyano}$ to monitor CyanoHAB variability in Lake Champlain's Missisquoi Bay, finding strong interannual variability linked to temperature and atmospheric instability. The algorithm has been operationally deployed across multiple U.S. states for bloom monitoring [66], [81], [82] and has informed state health advisories in California, Oregon, New York, Idaho, New Jersey, Utah, Vermont, and Wyoming [83]–[86].

The CyAN employs the CI$_{Cyano}$ algorithm to estimate cyanobacterial concentrations within US lakes. CyAN is a collaborative effort between the United States Environmental Protection Agency, National Aeronautics and Space Administration, National Oceanic and Atmospheric Administration, and United States Geological Survey that leverages European Space Agency sensors: Medium Resolution Imaging Spectrometer (2002-2012) and Ocean and Land Color Instrument on Sentinel-3A/B (2016-present) to detect and quantify cyanobacterial algal blooms. Data from this effort are available as GeoTiff files containing either daily outputs or 7-day maximum value composites, both with a spatial resolution of 300m. We downloaded the daily GeoTiff files from January 2016 to November 2024 for our study. [1] The broad time range provides sufficient temporal data to capture the dynamic nature of the CyanoHABs.

2) Temperature: Previous studies [17]–[19], [26], [27] suggested a strong correlation between temperature and CyanoHABs growth. JC Ho [19] concluded that increased temperatures and longer summers increase chlorophyll concentrations, implying cyanobacterial growth. Therefore, we employed temperature data for CyanoHABs modeling.

The temperature data are downloaded from the Moderate Resolution Imaging Spectroradiometer (MODIS) [75], specifically from MODIS Land Surface Temperature and Emissivity. The data are in the form of GeoTiff files with a daily temporal resolution and spatial resolution of 1km and contain day and night temperatures.[2]

[1] Dataset available at https://oceancolor.gsfc.nasa.gov/about/projects/cyan/

[2] Dataset available at Google Earth Engine https://developers.google.com/earth-engine/datasets/catalog/MODIS_061_MOD11A1

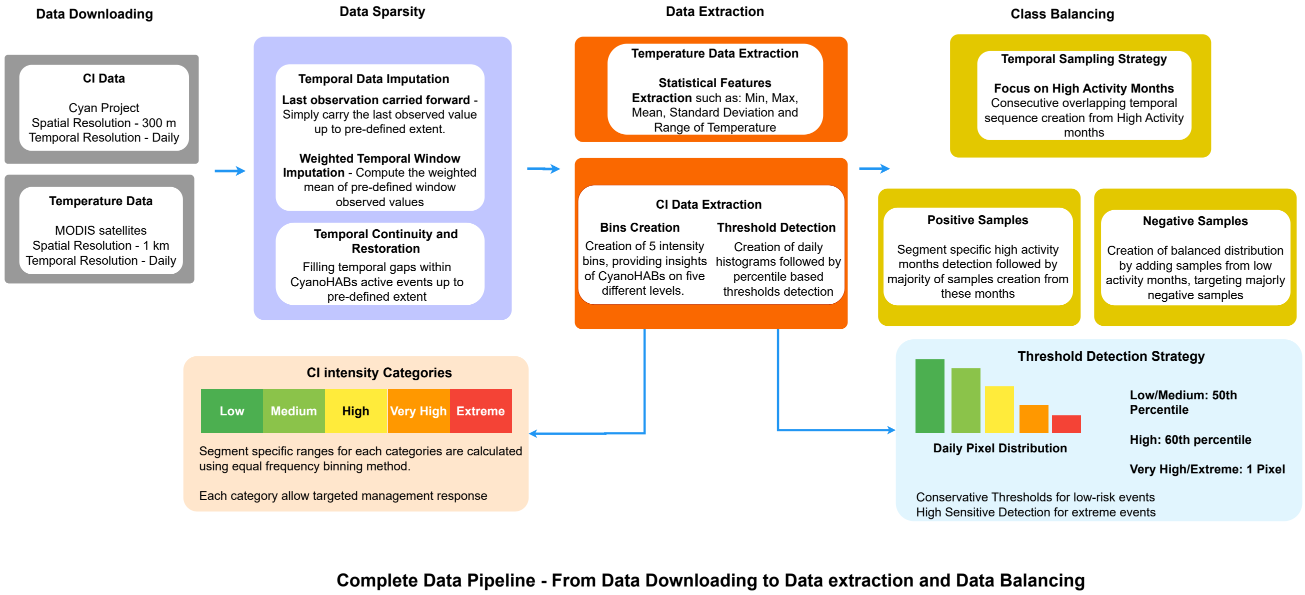
Data Downloading
CI Data
Cyan Project
Spatial Resolution - 300 m
Temporal Resolution - Daily
Temperature Data
MODIS satellites
Spatial Resolution - 1 km
Temporal Resolution - Daily
Data Sparsity
Temporal Data Imputation
Last observation carried forward - Simply carry the last observed value up to pre-defined extent.
Weighted Temporal Window Imputation - Compute the weighted mean of pre-defined window observed values
Temporal Continuity and Restoration
Filling temporal gaps within CyanoHABs active events up to pre-defined extent
Data Extraction
Temperature Data Extraction
Statistical Features Extraction such as: Min, Max, Mean, Standard Deviation and Range of Temperature
CI Data Extraction
Bins Creation
Creation of 5 intensity bins, providing insights of CyanoHABs on five different levels.
Threshold Detection
Creation of daily histograms followed by percentile based thresholds detection
Class Balancing
Temporal Sampling Strategy
Focus on High Activity Months
Consecutive overlapping temporal sequence creation from High Activity months
Positive Samples
Segment specific high activity months detection followed by majority of samples creation from these months
Negative Samples
Creation of balanced distribution by adding samples from low activity months, targeting majorly negative samples
CI intensity Categories
Low
Medium
High
Very High
Extreme
Segment specific ranges for each categories are calculated using equal frequency binning method.
Each category allow targeted management response
Threshold Detection Strategy
Daily Pixel Distribution
Low/Medium: 50th Percentile
High: 60th percentile
Very High/Extreme: 1 Pixel
Conservative Thresholds for low-risk events
High Sensitive Detection for extreme events
Complete Data Pipeline - From Data Downloading to Data extraction and Data Balancing

Fig. 2. Complete data processing pipeline for CyanoHABs prediction, from data acquisition through class balancing. The workflow consists of four stages: data downloading (CI and temperature), data sparsity management through imputation, data extraction for feature engineering, and class balancing using temporal sampling strategies.

B. Data Sparsity

Remote sensing data for CyanoHAB detection face significant challenges due to missing values caused by cloud cover and sensor limitations, with approximately 35% of the CI data and 87.88% of the temperature data missing in Missisquoi Bay. Such sparsity disrupts the temporal continuity needed for effective time-series modeling. While sophisticated spatial reconstruction methods such as Data Interpolating Empirical Orthogonal Functions (DINEOF) [87] have proven effective for satellite data gap-filling [88]–[90], we adopted a temporally-focused imputation strategy optimized for sequential forecasting. EOF-based approaches leverage spatial covariance but can attenuate onset/peak/decline transitions that are critical for learning CyanoHABs progression. Furthermore, our temperature dataset's extreme sparsity approaches the practical limits of iterative matrix decomposition, while our pixel-level approach with strict criteria (requiring ≥ 2 valid observations within 3-day windows) and linear complexity supports efficient near-real-time operational forecasting.

We present a two-stage imputation strategy to reconstruct bloom dynamics while preserving natural transitions between bloom states: (1) temporal data imputation and (2) temporal continuity restoration for CyanoHAB events. Imputation is performed on a segment-specific basis, ensuring that data from one segment does not affect others. Before processing, shallow pixels with a depth of less than 3m were filtered out using the official Lake Champlain bathymetry dataset provided by the Vermont Center for Geographic Information [91], as these pixels often display elevated CI values without actual blooms, allowing us to focus our analysis on pixels where CyanoHABs can be more reliably detected.

1) Temporal Data Imputation: The imputation strategy utilizes two complementary approaches: Last Observation Carried Forward and Weighted Temporal Window Imputation, applied at the pixel level for both the CI and temperature datasets. While Weighted Temporal Window Imputation could be used alone, we first apply Last Observation Carried Forward to preserve the day-to-day progression of CyanoHABs dynamics better. This ordering ensures that when recent values suggest persistence, the imputation reflects that continuity rather than diluting intensity through averaging.

Last Observation Carried Forward (LOCF): We implemented LOCF at the pixel level, examining each spatial coordinate sequentially across time. For pixels with missing values at time t, we checked if the exact spatial location contained a valid, non-imputed observation at time $t-1$. This value is propagated forward when available to fill the current missing value. This constraint maintains data integrity by balancing the need for completeness against the risk of introducing artificial patterns. LOCF is particularly important given the gradual day-to-day evolution typical in environmental systems. As shown in Figure 3, the LOCF application reduced the missing values of CI data at Missisquoi Bay from 35% to 27%, and missing temperature data from 87.88% to 82.89%.

Weighted Temporal Window Imputation: After LOCF, we applied a weighted temporal window approach utilizing a 3-day historical window. For each missing pixel value at time t, we computed a weighted mean of the available observations over times $t-1, t-2$, and $t-3$, assigning more importance to recent observations. Specifically, we used weights of $3, 2$, and 1 for $t-1, t-2$, and $t-3$, respectively. The imputed value $\hat{x}_t$ is calculated as:

$$\hat{x}_t = \frac{3x_{t-1} + 2x_{t-2} + x_{t-3}}{3+2+1} \qquad (1)$$

This imputation is conditionally applied only when at least two of the three observations within the window are valid (non-missing), ensuring sufficient information density for es-

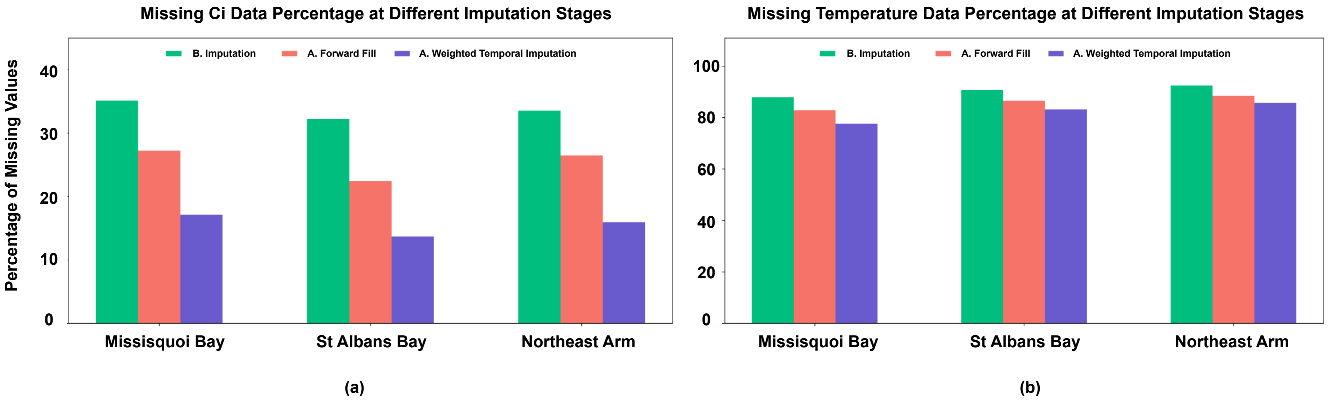

Fig. 3. Visualization of data sparsity. B. represents before while A. represents after. (a) Shows Cyanobacterial Index Values missing percentage at different imputation stages, while (b) shows Temperature data missing percentage. The green bars illustrate around 30% of the original Cyanobacterial Index Values and around 90% of the Temperature data are missing. The orange bar shows the dataset after performing Forward Fill, indicating a significant addition of data points. The purple represents the dataset after Weighted Temporal Window Imputation, indicating further reduction in data missing percentage. We constrained further imputation to avoid adding too much noise.

timation. In cases where $t-1$ is missing, the imputation is based on the available values at $t-2$ and $t-3$, reducing the influence of recent trends but still capturing nearby temporal structure. We limited weighted temporal window imputation to a maximum of two consecutive days per pixel to prevent the propagation of error. As illustrated in Figure 3, weighted temporal imputation further decreased the percentage of missing CI data at Missisquoi Bay from 27% to 17%, while the percentage of missing temperature data values was reduced from 82.89% to 77.69%.

2) Temporal Continuity Restoration in CyanoHABs Events: Following data imputation, we observed discontinuities in CyanoHAB event signatures where gaps of 3-7 days created artificial interruptions in bloom progression. We implemented a targeted continuity restoration procedure specifically for CI data to address this limitation, as these values directly indicate the presence and intensity of CyanoHABs. Temporal discontinuities were systematically identified through visual inspection of daily CI data across all segments. When these discontinuities aligned with periods of missing data during active bloom phases, a simplified version of the weighted temporal imputation was applied. The missing pixel values were estimated using the weighted mean of valid observations from the previous three days, with all missing values excluded from the computation. All continuity restoration was performed at the pixel level before any spatial summarization, ensuring that temporal coherence is restored directly in the raw CI rasters rather than in the aggregated time series

The restoration preserved the natural temporal evolution of CyanoHABs events and assisted the model to learn accurate transition patterns between CyanoHABs states. Unlike the initial data imputation phase, which uniformly addressed all missing values, this restoration step was selectively applied only to positive CyanoHABs periods. Gaps longer than 7 days were excluded to avoid introducing excessive noise, and no restoration was performed for gaps between non-bloom days, as maintaining continuity in bloom-free intervals was deemed less critical for model performance.

C. Data Extraction

The proposed data imputation strategies effectively restore temporal continuity and preserve the evolution of CyanoHABs at the pixel level. However, extending imputation beyond our constrained approach to fill all remaining gaps would introduce large-scale artificial patterns, as statistical reconstruction methods cannot capture true environmental variability during unobserved periods. The combination of remaining data sparsity and relatively coarse spatial resolution (300m for CI and 1km for temperature) limits the reliability of fine-scale spatial forecasting. To address these constraints while preserving dominant CyanoHABs dynamics, we adopt a segment-level feature-extraction that aggregates pixel information into intensity bins and statistical descriptors. This spatial summarization mitigates remaining missing values while amplifying coherent bloom signals over pixel-level noise.

1) Cyanobacterial Index Data Extraction: While forecasting the occurrence of CyanoHABs provides valuable information for policymakers and water resource managers, the practical utility of these forecasts increases substantially when intensity predictions are included. Traditional binary classification approaches predict simple presence or absence of CyanoHABs, while useful, fail to capture the critical nuance of bloom severity that drives management decisions. A low-intensity bloom may require only monitoring and public advisories, while an extreme bloom demands immediate interventions such as water treatment facility adjustments or recreational area closures.

Motivated by this operational need, we formulated the CyanoHABs forecasting problem as a multi-class classification task to predict both occurrence and intensity levels, providing actionable intelligence about the expected severity range of CyanoHAB events. The CI values, which range from 0 to 253, directly correspond to cyanobacterial concentrations, where 0 indicates CyanoHABs absence and positive values indicate CyanoHABs presence, with larger values indicating higher bloom intensity. We established five distinct intensity categories to capture the full spectrum of bloom severity:

Low Intensity, Medium Intensity, High Intensity, Very High Intensity, and Extreme Intensity CyanoHABs. The specific ranges for each intensity category are determined using equal-frequency binning applied separately to each segment's historical CI data. These categories allow stakeholders to anticipate the severity of upcoming events and implement proportional response measures.

After establishing the five intensity categories, we developed a threshold detection mechanism to determine whether each category will be active (positive) or inactive (negative) for a given forecast period. A positive classification for any intensity category indicates that CyanoHABs events are expected to occur within that category's CI value range, enabling targeted management responses. For example, a positive prediction for the Extreme Intensity category signals the anticipated presence of a severe CyanoHAB event requiring immediate intervention, while a negative prediction suggests that extreme-level blooms are unlikely to develop during the forecast period.

To determine these thresholds, we implemented a data-driven approach that analyzes the historical distribution of pixel counts within each intensity category. Rather than applying arbitrary cutoffs, past CI values are examined, and for each day, the histogram of pixel counts is created for each intensity bin. For Low and Medium intensity categories, we applied the 50th percentile as the threshold, requiring a substantial number of pixels to trigger a positive classification. This conservative approach is used to minimize false positives for these early forming CyanoHAB events. For the High-intensity category, we used the 60th percentile, acknowledging that while these events are more concerning, they still occur frequently enough to warrant moderate thresholds. However, for Very High and Extreme intensity categories, we adopted a highly sensitive approach with a threshold of one pixel. This ensures that even minimal evidence of severe bloom development triggers alerts, as missing these high-risk events could have serious public health and ecological consequences. Each segment's thresholds are calibrated independently using its specific historical data patterns.

To the best of our knowledge, no standardized ecological risk thresholds exist for CI-based CyanoHAB intensities. Therefore, we adopt a novel segment-specific binning strategy based on historical percentiles, enabling our model to reflect relative bloom severity and support actionable intensity-aware forecasting.

2) Temperature Data Extraction: Applying the same binning method used for CI data extraction might have resulted in unbalanced bin distributions due to missing temperature measurements and the dataset's relatively coarser spatial resolution. Instead, we implemented a statistical approach to extract temperature statistics for each segment of interest and each day. We extracted five distinct statistical features from both daytime and nighttime temperature measurements: minimum, maximum, mean, standard deviation, and range. The range is the difference between the maximum and minimum temperature, while standard deviation provides the spread of the temperature from the mean temperature. This approach yields ten temperature-related features per day per segment, providing a richer representation of temperature conditions than simple binning would allow.

D. Class Balancing

Our dataset exhibits substantial class imbalance in target labels with a negative-to-positive ratio of approximately 5:1. This imbalance biases model predictions toward the majority class. To address this challenge, a targeted temporal sampling strategy is implemented. For each segment, months with consistent CyanoHABs formations are identified (e.g., June through October for Missisquoi Bay), and sequence creation is extended to include the month preceding these high-activity periods. Sequences and their multi-label targets are created using overlapping windows throughout this time frame (see section V for sequence and target creation), resulting in most samples representing positive CyanoHABs events. Negative samples from inactive or low-activity months (e.g., April, May, and November for Missisquoi Bay) are selected to balance the distribution. This approach preserves the critical temporal context needed for sequence modeling while achieving a more balanced class distribution.

V. Sequential Data Formation

CyanoHABs do not emerge from a single day's conditions, and they result from the cumulative effects of environmental and hydrological conditions that develop over days to weeks. As a result, a single daily observation lacks the temporal context required to distinguish transient favorable conditions from sustained environmental drivers that lead to the initiation and intensification of CyanoHABs. To capture these dynamics, we structured the dataset sequentially by feeding the model temporally ordered input sequences rather than isolated daily observations. In addition to the CI value and temperature data, temporal features such as the day of the year, season, and month are added using the timestamp data. The input sequence and targets are created segment-wise, then concatenated into a single dataset to train the model on all segments combined. A rolling-window approach is used for sequence and target creation, where each input sequence comprises 15 consecutive days, with corresponding targets derived for a 14-day forecast horizon. The sequence window advances by one day for each new sample to ensure the capture of all temporal patterns. The segment signatures are omitted to force the model to learn a general CyanoHABs trend rather than segment-specific ones.

Given the daily data collection D for each segment s from a set of three monitoring segments S: Missisquoi Bay, St. Albans Bay, and Northeast Arm, each segment $s \in S$ contains data spanning T_s days. Let $L = 15$ represent the number of days used as input sequence length, and H is the forecast horizon.

For each starting day t such that $1 \leq t \leq T_s - (L + H)$, we create an input sequence X_t^s as:

$$X_t^s = \left(x_t^s, x_{t+1}^s, \ldots, x_{t+L-1}^s\right)$$

Here $X_t^s \in \mathbb{R}^{L \times F}$ represents an input sequence for segment s starting on day t, where F represents the dimensionality of the daily feature vector. Using a rolling window approach, we advance one day forward for each new sequence, continuing until reaching the maximum valid starting day that allows for both a complete 15-day input sequence and the full H-day forecast horizon.

For target creation, we classify each intensity bin for each day h within the H-day forecast horizon. Let $b_{d,i}$ denote the pixel count in the i^{th} intensity bin on day d. For each forecast day $h \in 1, 2, 3, \ldots, H$ and each bin $i \in 1, 2, 3, 4, 5$, the target is calculated as:

$$y^s_{t,h,i} = \begin{Bmatrix} 1 & \text{if } b^s_{d=t+L-1+h,i} \geq \tau^s_i \\ 0 & \text{Otherwise} \end{Bmatrix} \tag{2}$$

where τ^s_i represents the threshold for i^th bin for s.

These thresholds are determined through statistical analysis of historical pixel distributions. For each segment and intensity bin, we extracted historical pixel counts and calculated the 50th percentile to establish thresholds $\tau = (\tau_1, \tau_2, \tau_3, \tau_4, \tau_5)$. This approach ensures a balanced distribution between positive and negative classes rather than using arbitrary threshold values.

The target creation results in a multidimensional target tensor $Y^s_t \in 0, 1^{H\times 5}$, where each element indicates whether a specific intensity class will exceed its threshold on a particular forecast day. For a 14-day forecast horizon, this creates 70 binary classification targets per input sequence, providing detailed information about both the timing and intensity of potential bloom events.

VI. Methodology

Given that the forecasting objective is to predict the occurrence and intensity of CyanoHAB as a function of temporally evolving environmental conditions, the problem is naturally formulated as a sequence modeling task. Following data extraction, the inputs consist of structured scalar time series rather than dense spatial fields, making temporal dependency modeling the primary requirement. Accordingly, we adopt a sequence-based Transformer architecture, which is well-suited for capturing long-range temporal relationships.

The proposed Transformer-BiLSTM architecture addresses the inherent complexity of CyanoHAB dynamics through a Temporal Sequence Forecasting Attention Network (TS-FAN) designed to capture temporal patterns in sequential features. TS-FAN combines an embedding layer to process the data sparsity, followed by sinusoidal positional encodings that provide temporal context. A multi-head self-attention block processes different aspects of the sequence in parallel, while layer normalization and residual skip connections ensure stable training and improved gradient flow. A deep feed-forward network, structured as sequential neural blocks, refines feature representations through hierarchical transformations. A bidirectional LSTM further enhances sequential modeling by incorporating both past and future contexts. The forecasting task predicts the occurrence of future CyanoHABs events along with their intensities across five distinct levels (Low, Medium, High, Very High, and Extreme) for each day in a 14-day forecast horizon. Given an input sequence X_t, where $X_t \in \mathbb{R}^{L\times F}$ (see Section V) with $L = 15$ representing the sequence length and F the number of features, we define an Transformer-BiLSTM function f as:

$$Y_t = f(X_t) \quad \text{where} \quad f : \mathbb{R}^{L\times F} \rightarrow 0, 1^{H\times 5} \tag{3}$$

The function f maps the input time series with 15 time steps and F features to an $H \times 5$ dimensional binary output matrix, where $H = 14$ is the forecast horizon and 5 represents the intensity levels. For each forecast day $h \in 1, 2, ..., 14$ and intensity level $i \in 1, 2, 3, 4, 5$, the corresponding output element is either 1 to indicate the presence of CyanoHABs event within the specific intensity level or 0 to indicate the absence of the CyanoHABs event. This results in 70 binary classification outputs per input sequence, providing detailed information about both the timing and intensity of potential CyanoHABs events. The overall network architecture is illustrated in Figure 4.

A. *Transformer-BiLSTM Architecture*

1) TS-FAN: Given the critical role of temporal dependencies in the CyanoHAB's life cycle, it is essential to effectively capture the temporal relationships within the input sequence data. TS-FAN uses temporal feature extraction components to effectively model the complex temporal patterns and interactions inherent in the sequence data.

Given an input sequence X_t, it is first processed by an embedding layer. The layer takes the input sequence $X_t = (x_1, x_2, x_3, ..., x_L)$, where $L = 15$, each $x_t \in \mathbb{R}^F$ dimensional sparse feature vector, and it maps to a dense dimension D, producing an embedded sequence $E \in \mathbb{R}^{L\times D}$. The projection unifies heterogeneous numerical features into a common representation and ensures compatibility with the Transformer attention mechanism, which operates in a fixed-dimensional embedding space. Following [49], we add positional encoding to the processed input sequence, allowing the model to understand the temporal structure. The positional encoding (PE) adds a unique vector at each position p within a sequence and is defined.

$$\begin{aligned} PE(p, 2i) &= \sin\left(\frac{p}{10000^{2i/D}}\right), \\ PE(p, 2i+1) &= \cos\left(\frac{p}{10000^{2i/D}}\right). \end{aligned} \tag{4}$$

Here, p denotes the temporal position in the input sequence, and i indexes the embedding dimensions, where D is the embedding dimensionality. The division by $10000^{2i/D}$ creates wavelengths that form a geometric progression from 2π to $10000 \cdot 2\pi$, allowing the model to learn to attend to relative positions with varying periods. This scaling ensures that positional encodings across dimensions have different frequencies, making each position uniquely identifiable. Once the embeddings have been augmented with PE, the resultant positionally encoded vectors E^{PE} are forwarded to the encoder layer. Importantly, positional encoding is applied once after the embedding layer, and E^{PE} serves as the residual input to the self-attention sublayer, which adds it to the output of multi-head attention to stabilize optimization and ensure effective gradient flow. We use a single encoder layer, which

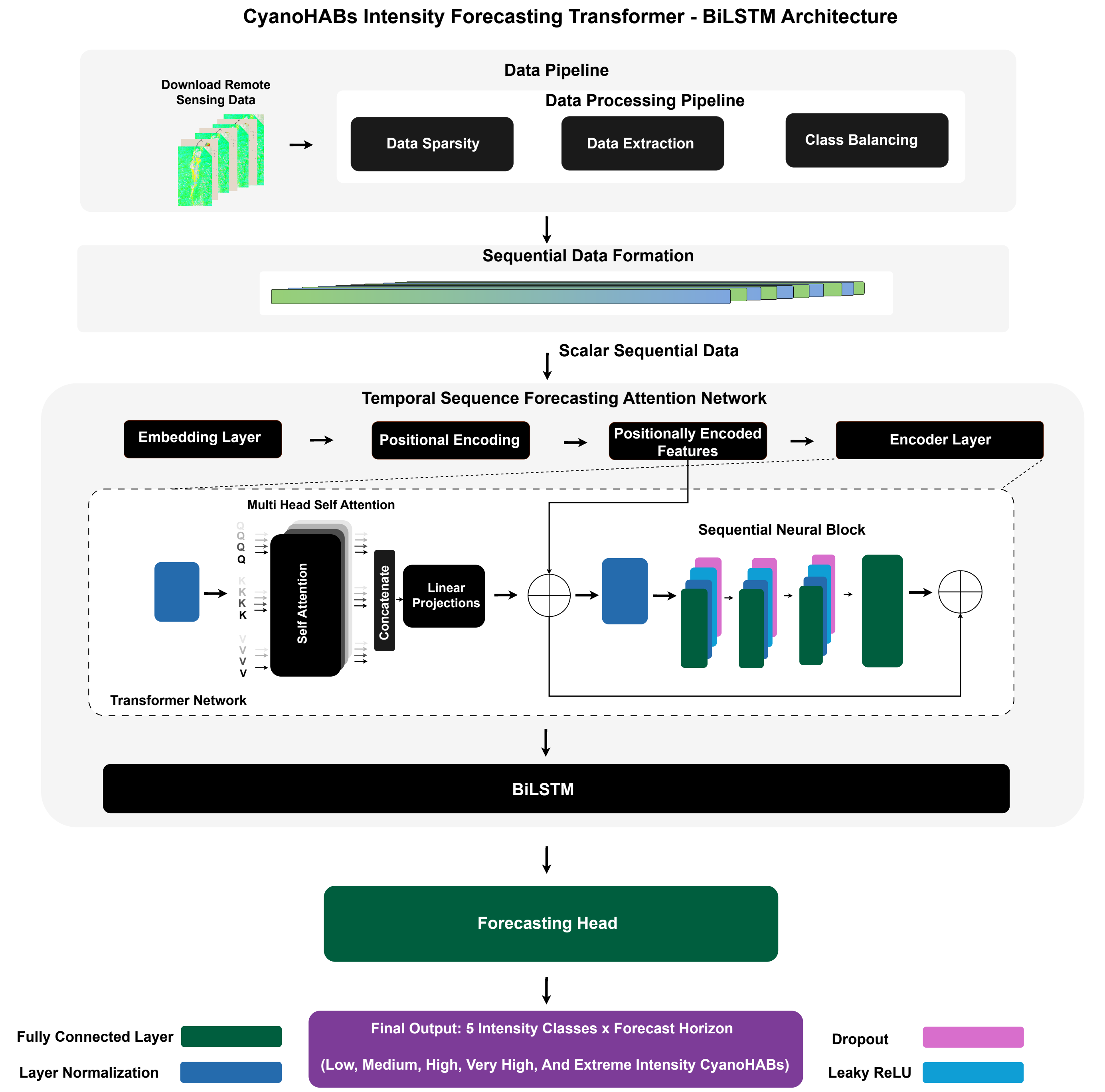
CyanoHABs Intensity Forecasting Transformer - BiLSTM Architecture
Data Pipeline
Download Remote Sensing Data
Data Processing Pipeline
Data Sparsity
Data Extraction
Class Balancing
Sequential Data Formation
Scalar Sequential Data
Temporal Sequence Forecasting Attention Network
Embedding Layer
Positional Encoding
Positionally Encoded Features
Encoder Layer
Multi Head Self Attention
Q
K
V
Self Attention
Concatenate
Linear Projections
Sequential Neural Block
Transformer Network
BiLSTM
Forecasting Head
Final Output: 5 Intensity Classes x Forecast Horizon
(Low, Medium, High, Very High, And Extreme Intensity CyanoHABs)
Fully Connected Layer
Layer Normalization
Dropout
Leaky ReLU

Fig. 4. The diagram illustrates the Transformer-BiLSTM model for CyanoHABs Intensity forecasting. The system processes 15 days of remote sensing inputs through an embedding layer, positional encoding, a Transformer encoder with multi-head self-attention, and a bidirectional LSTM block. The final output layer predicts five CyanoHAB intensity classes (Low to Extreme) across a 14-day forecast horizon. The pipeline incorporates preprocessing steps to address data sparsity and ensure class balance.

contains normalization layers, eight attention heads, and a sequential neural block. We replace the simple feed-forward network in the encoder layer with a deeper neural network, which we named sequential neural blocks. The input sequence E^{PE} is passed to the encoder layer, which first applies layer normalization to stabilize gradients.

$$E_{norm} = \text{LayerNorm}(E^{PE}) \tag{5}$$

The normalized embeddings E_{norm} are then processed by the multi-head attention mechanism within each head, where each head h computes attention scores based on queries Q, keys K, and values V. Specifically, for each h, we have:

$$Q_h, K_h, V_h = E_{norm}W_h^Q, E_{norm}W_h^K, E_{norm}W_h^V \tag{6}$$

Where, W_h^Q, W_h^K, W_h^V are learnable weighted metrics of each head h, with assigned weights that are optimized during training. Here, the network takes the representation Q_h and compares every other representation K_h of the sub-sequence to determine the attention scores. Simply, it represents the relationship of each Q_h with every other K_h, and the results are represented by values V_h. Subsequently, the softmax function is applied to normalize the attention scores between the range of 0 to 1, where 1 represents maximum correspondence, and it follows:

$$Attention(Q_h, K_h, V_h) = softmax\left(\frac{Q_h K_h^{Transpose}}{\sqrt{d_k}}\right) V_h \tag{7}$$

Here, $Transpose$ represents the transpose operation, and d_k represents the dimensionality of the keys. The latter is used to scale down the dot product so that it does not become too large, facilitating stable gradient flow during training.

Next, the outputs of all attention heads are concatenated and projected linearly. The resulting self-attention output is added to the positionally encoded input E^{PE} via a residual connection to stabilize learning. The output of the self-attention sublayer is denoted by ϑ and is computed as:

$$\vartheta = E^{PE} + Dropout(MHA(E_{norm})) \tag{8}$$

The multi-head self-attention module captures complex temporal dependencies by processing sequential data patterns. To further refine these temporal representations, we replaced the traditional feed-forward network in multi-head self-attention with a deeper and customized feed-forward network named Sequential Neural Blocks (SNB). SNB consists of three sequential blocks of fully connected layers, layer normalization, LeakyReLU, and a dropout layer, followed by a fully connected layer. These components refine complex interactions among features, stabilize learning, introduce non-linearity, and prevent overfitting.

Given resultant output sequence ϑ from multi-head self-attention, it is passed to SNB, where it is processed by the first sequential block, mapping it to a new feature space:

$$\zeta_1 = \text{Dropout}(\text{L.ReLU}(\text{L.Norm}(Z_1 = \vartheta W_1 + b_1))) \tag{9}$$

Here, W_1 & b_1 are weight metrics and bias term of a fully connected layer of the first block, respectively, and Z_1 represents the intermediate representation. This is passed to the layer norm to stabilize the Z_1 distribution, which is passed to the non-linear activation function to introduce the linearity. The subsequent dropout layer reduces the risk of overfitting.

The output of the first block is represented as ζ_1 and is passed to the second block to refine the abstractions further, which outputs ζ_2. The third block further refines those representations and passes to the fully connected layer, which transforms the representations to a requisite feature dimension ζ_4, ensuring compatibility for a residual skip connection. The ζ_4 is added to the multi-head self-attention ϑ output before passing to the BiLSTM. This is shown below:

$$En_{out} = \vartheta + \zeta_4 \tag{10}$$

The Encoder layer effectively captures temporal dependencies. We incorporate a BiLSTM layer to further enhance the model's ability to process temporal information. It processes the encoded representations sequentially in both forward and backward directions, providing temporal context that strengthens the model's predictive capabilities. Given En_{out} as input from the encoder, BiLSTM computes the hidden states $\overrightarrow{h}$ and $\overleftarrow{h}$ in forward and backward directions, respectively. Afterward, the forward and backward outputs are concatenated to form a representation that encompasses information from past and future contexts.

$$H_t = [\overrightarrow{h}; \overleftarrow{h}] \tag{11}$$

Here, H_t denotes the final aggregated BiLSTM representation.

2) Forecasting Head: The Forecasting Head consists of a fully connected layer and convert the model-specific temporal representation H_t into a multi-output binary prediction over a 14-day forecast horizon. The output consists of 70 binary values representing the presence or absence of CyanoHABs across five intensity levels for each of the next 14 days. Formally, the forecasting head performs the following operation:

$$\text{Fout}_{t,h,i} = \begin{cases} 1 & \text{if } \sigma\left(\left(H_t W_{fh} + b_{fh}\right)_{h,i}\right) > \theta_{h,i}, \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

$$\forall\, h \in \{1, \ldots, 14\}, \quad i \in \{1, \ldots, 5\}$$

Here, $\text{Fout}_{t,h,i}$ is the predicted binary label for day h and intensity level i for the input sequence at time t. The representation $H_t \in \mathbb{R}^d$ is the output of BiLSTM, and terms $Wfh \in \mathbb{R}^{d\times 70}$ and $b_{fh} \in \mathbb{R}^{70}$ are the weights and bias of the fully connected output layer. The sigmoid function σ maps each output to the range $[0, 1]$, and the threshold $\theta_{h,i}$ (typically set to 0.5) is used to determine the binary class assignment.

3) Persistence Model (Baseline): To establish a comparison baseline, we include a Persistence Model that assumes bloom intensity classes remain unchanged over the forecast horizon. This approach is motivated by the temporal continuity observed in CyanoHAB events, where bloom conditions on day t often persist to subsequent days. Formally, the persistence model assumes that the bloom intensity classes observed at time t remain unchanged for each of the next 14 days:

$$\hat{y}_{t+h} = y_t, \quad \text{for} \quad h = 1, 2, \ldots, 14 \tag{13}$$

Here, $y_t \in 0, 1^5$ is the binary vector representing the presence or absence of CyanoHABs across five intensity levels (Low, Medium, High, Very High, Extreme) on day t. The prediction $\hat{y}_{t+h}$ at future horizon h is identical to y_t, implying no evolution in bloom conditions.

While the persistence model does not involve any learning or parameter optimization, it serves as a strong short-term benchmark. Its relevance is especially notable in short-range forecasts (e.g., Day 1), where it occasionally outperforms learning-based models due to the slow-changing nature of bloom patterns. However, its inability to adapt to evolving environmental conditions leads to rapid performance degradation over longer horizons, thereby emphasizing the need for dynamic, data-driven approaches.

B. Loss Functions

The Transformer-BiLSTM outputs a binary prediction for the presence or absence of CyanoHABs across five intensity levels for each of the 14 forecast days, resulting in a total of 70 binary classification tasks per input sequence. To optimize model performance, we apply the Binary Cross-Entropy (BCE) loss to these outputs. The BCE loss is defined as:

$$\text{BCE}\left(y_i, \hat{y}_i\right) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \, log(\hat{y}_i) + (1 - y_i) \, log(1 - \hat{y}_i)\right] \tag{14}$$

Here, y_i and $\hat{y}_i$ denote the ground truth and predicted probability for the i^{th} binary output, respectively, and $N = 70$ is the total number of outputs per sequence.

VII. Temporal Data Augmentation

Transformer-based models typically require substantial training data to learn complex patterns and relationships effectively. However, the available dataset comprises only a few thousand sequences, which presents a significant challenge for model training. To address this limitation and enhance model generalization, the research implements a sophisticated temporal data augmentation strategy that respects the physical constraints and seasonality patterns of CyanoHABs.

The augmentation strategy is applied only during training to prevent data leakage into validation and test sets. It modifies two key aspects of the dataset: CyanoHABs intensity bins, and temperature statistics.

1) CyanoHABs Intensity Bin Augmentation: As introduced earlier (see Section IV-C), the CI values were discretized into five segment-specific intensity categories - Low, Medium, High, Very High, and Extreme - using an equal-frequency binning strategy. Each of these categories corresponds to a CI intensity bin, representing a distinct range of bloom severity. In this section, we refer to these five intensity classes as bins for brevity. Since CyanoHAB events exhibit strong seasonality, a month-dependent augmentation strategy introduces variability into bin values. Importantly, augmentation is applied only when a bin has at least three observed pixels, ensuring that artificial bloom events are not introduced during naturally inactive months. The magnitude and direction of adjustments vary seasonally: This is explained below:

Peak bloom months:

- If the bin count exceeds its segment-specific threshold by more than 10 pixels, it is considered a significant deviation. In this case, we randomly perturb the value by up to ±8 pixels to simulate realistic fluctuations in bloom intensity.
- If the count is above the threshold but with a smaller margin (1–10 pixels), we apply minor adjustments of up to ±3 pixels.
- If the count is below the threshold but above the minimum pixel threshold of 3, we apply small positive or negative changes up to ±3 to maintain variability near bloom onset.

Non Peak bloom months:

- Adjustments are more conservative, with a maximum perturbation of ±3 pixels to reflect the naturally lower CyanoHABs activity.
- If the bin count exceeds the threshold by more than 3 pixels, it is perturbed by up to ±3 pixels.
- If the count is just slightly above the threshold of minimum pixels of 3, we apply only positive perturbations of 0 to +2 to reflect very low cell counts.

2) Temperature Data Augmentation: Temperature serves a significant role in CyanoHAB development, where significant fluctuations can influence bloom dynamics. Controlled perturbations alter day and night temperature features while the system maintains physical consistency. This is explained below:

- Day and night minimum and maximum temperatures are perturbed within a narrow range of –0.1°C to +0.16°C to simulate natural variability without altering ecological regimes. Larger perturbations (e.g., +5°C) could shift environmental conditions from moderate to warm, misleading the model. The selected range ensures stability of bloom-relevant conditions.
- Dependent features such as mean, range, and standard deviation are recalculated following perturbation to maintain internal consistency.
- Constraints enforce that day and night minimum temperatures remain lower than their respective maximum values to preserve physically valid relationships.

This augmentation is applied to 50% of the training sequences. The objective is to expose the model to plausible thermal variations while avoiding excessive artificial noise. Despite 90% spatial missingness, the use of aggregated statistics across interpolated days provides a sufficient signal to benefit from this augmentation strategy.

VIII. Results

A. Implementation Details

The Transformer-BiLSTM is implemented using PyTorch [92], a widely adopted deep learning framework. Several strategies are used to ensure robust training despite the challenges of sparse satellite data.

For optimization, we employ the AdamW optimizer with a learning rate of 0.0005 and a weight decay of 0.09. The weight decay term helps reduce overfitting by discouraging excessively large weight values, promoting simpler and more generalizable models. A Cosine Annealing scheduler with Warm Restarts enhances the learning trajectory and cyclically resets the learning rate every five epochs. This approach allows the model to escape local minima and achieve better convergence. Early stopping prevents over-training by monitoring the validation's F1 performance and preserving the model's best state. Label smoothing [93] enhances generalization by introducing a small amount of uncertainty in the target distribution,

TABLE I
PERFORMANCE METRICS FOR INDIVIDUAL MODELS AND THE TRANSFORMER-BILSTM FOR THE 14TH DAY ACROSS A 14-DAY FORECAST HORIZON. THE TRANSFORMER-BILSTM CONSISTENTLY OUTPERFORMED ALL MODELS ACROSS ALL METRICS FOR LONG-TERM CYANOHABS FORECASTING.

Model Category	Model	Acc (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
Baseline	Persistence Model	71.34	71.68	71.34	71.54	70.42
Machine Learning	Support Vector Machines	43.29	62.20	51.91	55.29	80.12
	XGBoost	41.77	61.22	53.67	56.68	80.16
	Random Forest	45.45	63.38	57.89	59.36	82.11
Deep Learning	Transformer	74.95	74.35	74.95	74.08	77.18
	Transformer-Custom Attention	75.08	74.48	75.08	74.30	77.04
	Transformer-BiGRU	77.13	76.97	77.13	77.04	81.01
	Transformer-LSTM	77.47	77.39	77.47	77.43	80.85
	BiLSTM	77.71	77.38	77.71	77.46	80.32
	Transformer-GRU	77.95	78.43	77.95	78.13	82.53
	Transformer-BiLSTM	**78.94**	**78.80**	**78.94**	**78.86**	**82.55**

which prevents overconfidence in predictions. Gradient clipping with a threshold of 3.0 stabilizes training by preventing exploding gradients, which is particularly important for the recurrent components. The model is trained for 70 epochs with parameter updates governed by the minimization of BCE Loss. During evaluation, model outputs are thresholded using a default cutoff of 0.5. Any predicted confidence score $\hat{y} \geq 0.5$ or above is considered a positive prediction for the corresponding intensity class, while scores below this threshold are treated as negative.

The proposed model is trained on a personal computing setup equipped with a 13th Gen Intel Core i7-13700H processor, 32GB RAM, and an NVIDIA GeForce RTX 4070 GPU. This shows the model does not take much resources, which allows deployment in resource-constrained environments, making it practical for operational CyanoHAB intensity forecasting systems.

B. Evaluation Metrics

The forecasting model is evaluated using accuracy, precision, recall, and F1 scores. These are calculated as:

$$\text{Accuracy} = \frac{\text{TP + TN}}{\text{TP + TN + FP + FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}}$$

$$\text{F1 Scores} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision + Recall}}$$

Here, TP, TN, FP, and FN are True Positives, True Negatives, False Positives, and False Negatives, respectively. In addition, we report the Area Under the Curve (AUC) to evaluate each model's performance. AUC assesses the model's ability to discriminate between classes across a range of classification thresholds.

C. Intensity Forecasting Evaluation

1) Quantitative Evaluation: To evaluate the proposed Transformer-BiLSTM for CyanoHAB intensity forecasting, we compared it with two sets of architectures designed explicitly for this study. The first set covers three commonly used machine-learning algorithms, which include Random Forest (RF) [94], Support Vector Machine (SVM) [95], and XGBoost [96]. The second set consists of six custom deep-learning architectures. These include standalone Transformer-only and BiLSTM-only models, as well as four hybrid variants that pair a Transformer encoder with LSTM, BiLSTM, GRU, or BiGRU modules. In addition, we include a Transformer variant equipped with a custom temporal attention pooling mechanism, referred to as Transformer-Custom Attention. This benchmark provides a comprehensive performance evaluation against both traditional and state-of-the-art approaches.

For standalone architectures, the Transformer-only model processes input directly through the encoder layer to the forecasting head, while the BiLSTM-only model employs an embedding layer before BiLSTM processing and subsequent prediction. The hybrid configurations combine transformers with Long Short Term Memory (LSTM) [97], Gated Recurrent Unit (GRU) [98], and Bidirectional Gated Recurrent Unit (BiGRU), alongside a Custom Attention layer. In these hybrid models, encoder outputs serve as inputs to their respective recurrent or attention blocks, e.g., Transformer-BiGRU forwards the encoded sequence to a BiGRU layer for bidirectional processing. Besides utilizing established recurrent units, we incorporated a custom attention layer in Transformer-Custom Attention to provide additional emphasis on the most important time steps essential for accurate prediction. It uses a custom temporal attention pooling layer applied to the Transformer encoder outputs. Specifically, given the encoded sequence representations, a learnable linear projection assigns a scalar importance score to each time step, which is normalized using a softmax function to produce temporal attention weights. Mathematically, it is formulated as:

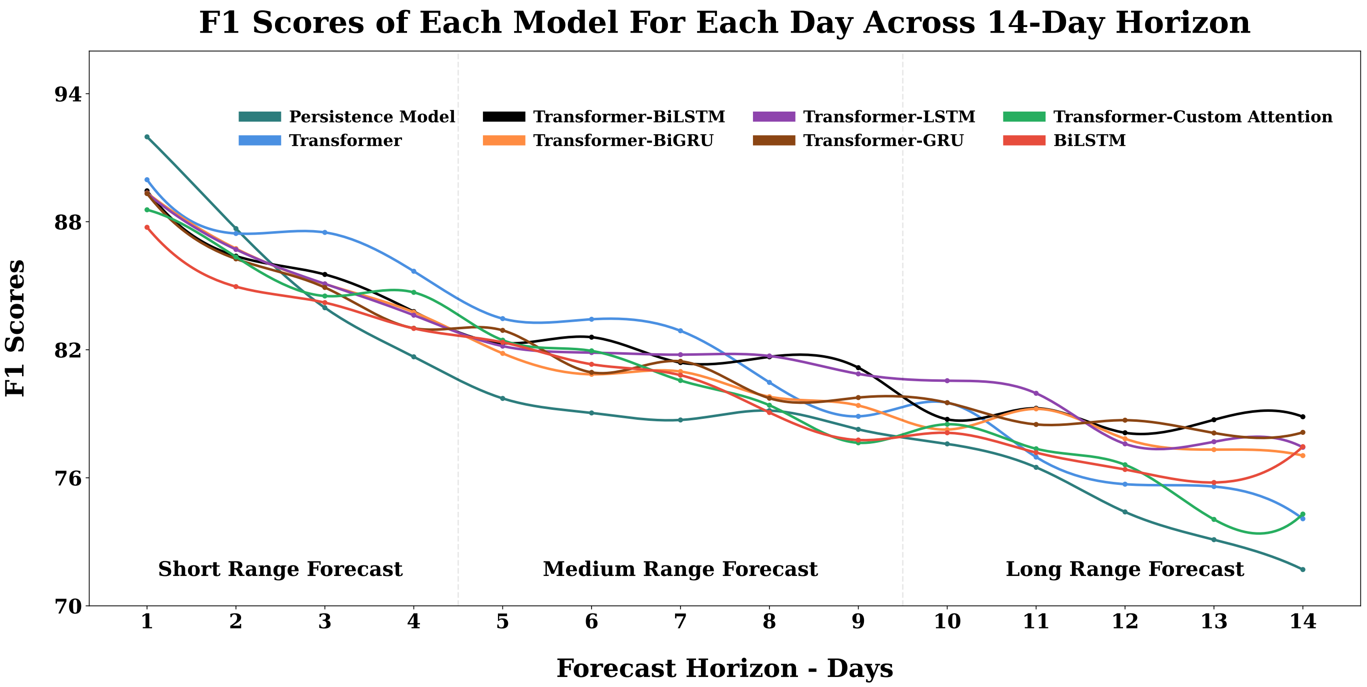

Fig. 5. Visualization F1 scores of each model for each day across the 14-day forecast horizon for all segments. F1 score trends across a 14-day forecast horizon for all individual models and the Transformer-BiLSTM model. Forecast performance generally declines with increasing lead time. The Transformer-BiLSTM outperforms individual model across longer-forecast ranges and perform competitively for short and medium forecast. Vertical dashed lines demarcate short-range (Days 1-4), medium-range (Days 5-9), and long-range (Days 10–14) forecasts. Circular markers denote discrete evaluation points at each forecast day.

$$\alpha_t = \text{softmax}(w^T h_t), Z = \sum_t \alpha_t h_t$$

Here, h_t denotes the transformer encoder output at time step t. The vector $w \in R^D$ is a trainable parameter that assigns a scalar importance score $w^T h_t$ to each time step. A softmax operation is applied across all time steps to produce normalized temporal attention weights α_t. The final sequence representation Z is computed as a weighted sum of the encoder outputs, emphasizing temporally critical observations before forecasting.

All deep learning based models were trained under identical conditions as detailed in Section VIII-A, using data from 2016 to 2021 for training, 2022 for validation, and 2023-2024 for testing, maintaining consistent data preprocessing pipelines, loss functions, and optimizer configurations with the proposed Transformer-BiLSTM architecture. Training was standardized at 70 epochs across all models, with the exception of the custom attention variant, which required extended training of 120 epochs due to its additional learnable parameters and increased computational complexity. Performance evaluation utilized standard metrics, including accuracy, precision, recall, and F1 score.

The machine learning models were trained on identical data splits but required different preprocessing since Random Forest, XGBoost, and SVM operate on fixed-length vectors rather than sequences. Input data was flattened before processing. The Random Forest comprised 600 trees grown to full depth with a minimum of two samples per leaf, wrapped in a MultiOutputClassifier from Scikit-learn [99] to predict 70 binary labels simultaneously (14 days × 5 classes). The SVM used RBF-kernel with probability estimates enabled, embedded in a One-Vs-Rest wrapper from Scikit-learn to train one classifier per label for consistent AUC computation. XGBoost followed the same preprocessing pipeline and One-Vs-Rest ensemble approach, learning one boosted tree model per output label.

Table I summarizes the forecasting performance of each model over a 14-day prediction horizon. Notably, the metrics presented in Table I represent the performance specifically for the 14th day forecast. The results show deep learning models significantly outperform traditional machine learning approaches.

The Transformer-BiLSTM architecture achieves the highest performance across all metrics, with 78.94% accuracy, 78.80% precision, 78.94% recall, 78.86% F1 score, and an AUC of 82.65%. This highlights the effectiveness of combining self-attention mechanisms with bidirectional sequence modeling to capture both contextual information and temporal dependencies in CyanoHAB dynamics. The Transformer-GRU and standalone BiLSTM models also performed strongly, achieving F1 scores of 78.13% and 77.46%, and AUC of 82.53% and 80.32%, respectively. Notably, all hybrid architectures consistently outperformed the baseline Transformer model over a more extended forecast period. While the Transformer-Custom Attention Layer showed modest improvement over the baseline, the integration of recurrent components yielded substantially better results. This highlights the importance of integrating both attention-based architectures and temporal modeling for extended forecasting tasks, as neither mechanism alone has proven sufficient to maintain performance at longer prediction intervals.

The machine-learning models trail deep learning based models in point metrics; however, their AUC scores are comparatively high. Random Forest reaches an AUC of 82.1 %, and both SVM and XGBoost exceed 80 %, despite accuracy in the low to mid 40% range. This disparity indicates that tree and kernel-based methods rank bloom risk reasonably well but struggle to convert those rankings into correct hard labels once a decision threshold is set. In other words, they discern separability yet lack calibrated decision boundaries for the multi-label, highly imbalanced setting

The Persistence Model, used as a naive temporal baseline for deep learning based models, achieved an F1 score of 71.70%. While it performed competitively in short-term forecasts (see Figure 5), its performance significantly deteriorated over longer horizons, as exemplified by its last-place ranking on Day 14. This trend reflects its inability to adapt to dynamic bloom transitions and further underscores the value of learning-based models for medium to long-range forecasting.

One can observe that Transformer-BiLSTM consistently delivered the most balanced performance across all metrics. This suggests that the combination of bidirectional sequential modeling and self-attention leads to more temporally coherent and context-rich representations. The Transformer's global attention captures long-range dependencies while BiLSTM's dual-directional processing models local temporal patterns, enabling the architecture to effectively identify subtle CyanoHAB dynamics such as bloom progression and decay phases. This combined approach proves particularly valuable for extended forecast horizons, where neither attention mechanisms nor recurrent processing alone maintains optimal prediction accuracy.

Further F1 scores analysis was conducted to examine how model performance changes across the forecast horizon. Since the Machine Learning based model showed lower F1 scores, this comparison focused on the deep-learning architectures and the Persistence model only. Figure 5 illustrates the F1 scores of the persistence model and each deep learning model for each day across the 14-day forecasting period. In the short-range forecasts (Days 1–4), all models demonstrate strong performance, with the standalone Transformer performing comparably to the Transformer-BiLSTM during the initial days. The Persistence Model, due to the temporal continuity of CyanoHAB events, performs competitively on Day 1, even marginally outperforming learning-based models. However, this advantage is short-lived; its static assumption becomes increasingly inaccurate with time, and it quickly deteriorates, reaching the lowest F1 score by Day 14.

A notable transition occurs as the forecast horizon extends into the long-range period (Days 10-14). While the standalone Transformer shows a steep decline in performance beyond Day 9, the Transformer-BiLSTM model takes control and consistently outperforms all other models in the long-range period. This superior long-range performance reflects the hybrid architecture's ability to effectively combine short-term contextual understanding via attention mechanisms with robust long-term dependency modeling through bidirectional recurrence. Other hybrid models like Transformer-GRU also show better stability than the standalone Transformer, but the Transformer-BiLSTM consistently achieves the highest F1 scores in long-range CyanoHAB intensity forecasting.

2) Qualitative Evaluation: To evaluate the qualitative performance of our forecasting models, we employed the Probability of Detection (POD), also known as the Hit Rate. POD quantifies the proportion of observed events that are correctly predicted by the model, providing insight into the model's ability to identify true positive occurrences.

The use of POD in harmful algal bloom (HAB) forecasting has been well-documented in previous studies. For instance, the National Oceanic and Atmospheric Administration (NOAA) utilized POD [100] to assess the performance of its HAB Operational Forecast System in the western Gulf of Mexico. This system provided forecasts of bloom transport and associated respiratory irritation, with POD serving as a key metric for evaluating forecast accuracy over multiple years. Similarly, the California Harmful Algae Risk Mapping system was designed to predict the spatial likelihood of Pseudo-nitzschia blooms and associated domoic acid events along the California coast [101]. In its initial skill assessments, the system incorporated POD by comparing model-generated probabilities with observed bloom occurrences to evaluate forecasting reliability.

Mathematically, the Probability of Detection is defined as:

$$\text{POD/Hit Rate} = \frac{\alpha}{\alpha + c}$$

Here, α represents the number of hits (instances where an event was both observed and correctly forecasted) and c denotes the number of misses (instances where an event was observed but not forecasted).

The POD value ranges from 0 to 1, with a value of 1 indicating perfect detection (all observed events were correctly forecasted), and a value of 0 indicating no detection (none of the observed events were forecasted). Intermediate values reflect the proportion of observed events that were successfully predicted.

While POD is formally equivalent to Recall, we used it as a concise visual summary to interpret multihorizon predictions. Given the 14-day forecast window, directly comparing daily predictions can be visually cluttered. The POD thus provides a readable behavior of detection consistency across the forecast horizon, supporting qualitative model assessment. We calculated the POD for each station-day by comparing the model's predictions with the actual observations. The comprehensive analysis for each station's hit rate is presented below.

a) ***Missisquoi Bay:*** The time series visualization for Missisquoi Bay (Figure 6) illustrates the hit rate analysis during 2023 and 2024. The visualization demonstrates the model's ability to capture CyanoHAB seasonal patterns across different CyanoHABs phases. During non-CyanoHAB periods, the model achieves a higher hit rate, as seen in June 2023, accurately identifying the absence of CyanoHABs across all intensity categories. During transition periods when CyanoHABs begin forming or dissipating, hit rates naturally decrease, reflecting the inherent challenge of predicting these

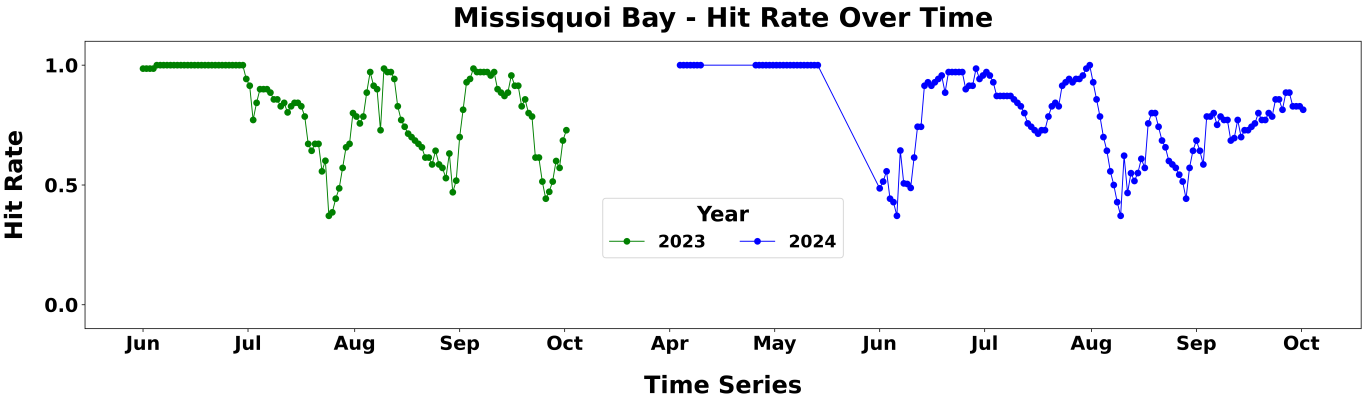

Fig. 6. Temporal hit rate analysis for Missisquoi Bay across the 2023-2024 test period. The visualization reveals seasonal patterns in CyanoHAB prediction accuracy, with higher hit rates during established CyanoHAB phases and variation during transitions. The patterns between 2023 and 2024 demonstrate the model's adaptability to different CyanoHAB dynamics while maintaining robust seasonal forecasting capabilities.

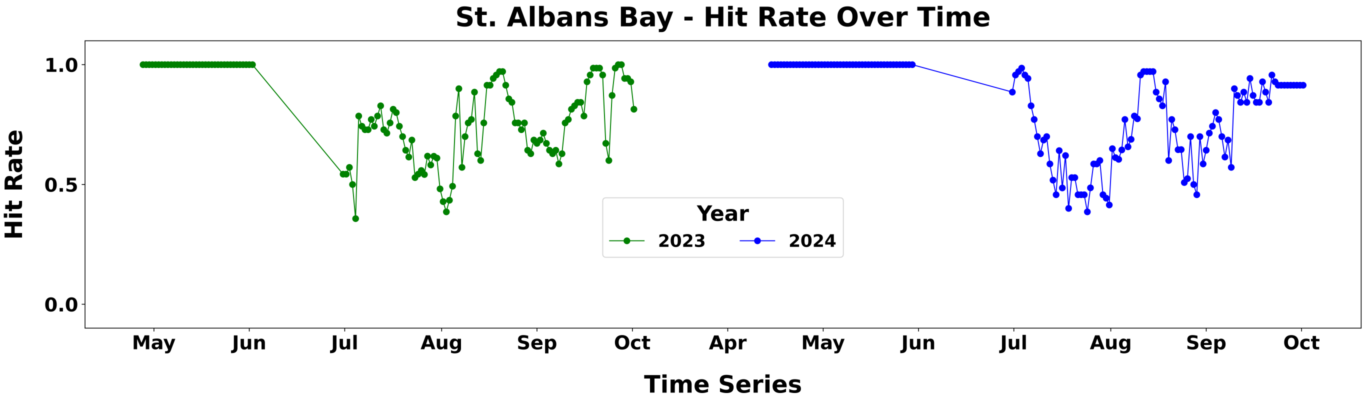

Fig. 7. Temporal hit rate analysis for St. Albans Bay across the 2023-2024 test period. The visualization shows later CyanoHABs formations compared to Missisquoi Bay. The fragmented CyanoHABs patterns create more frequent hit rate fluctuations during the season.

dynamic phases. This pattern is visible in early August 2023, where the hit rate transitions from 1.0 to approximately 0.5. The model maintains strong performance during established CyanoHABs periods, with hit rates stabilizing between 0.8 and 1.0, demonstrating its capability to forecast stable CyanoHAB states accurately.

The comparison between 2023 and 2024 reveals the model's adaptability to varying CyanoHAB dynamics. The 2023 pattern shows a clear seasonal progression: non-CyanoHABs period (June), transition phase (July), peak CyanoHABs (August-September), and gradual decline (October). The 2024 data presents a more complex pattern with earlier CyanoHABs initiation in June and multiple formation-dissipation cycles throughout the season. Despite these variations, the model successfully tracks the overall seasonal trends in both years, with hit rates recovering during peak CyanoHABs periods and maintaining high accuracy during stable conditions.

These temporal patterns illustrate the model's robust performance in capturing CyanoHAB seasonal dynamics. While hit rates fluctuate during transition periods - a natural characteristic of environmental forecasting - the model consistently identifies the broader seasonal patterns and maintains high accuracy during both CyanoHABs and non-CyanoHABs stable states.

*b) **St. Albans Bay**:* The time series visualization for St. Albans Bay (Figure 7) reveals different CyanoHABs characteristics compared to Missisquoi Bay. The visualization highlights distinct temporal and ecological patterns in CyanoHAB dynamics compared to Missisquoi Bay. CyanoHABs development in St. Albans tends to occur later in the season, typically beginning around middle or late July, as indicated by the initial drop in hit rate during that period. This temporal shift likely reflects different environmental conditions, possibly related to varying warming patterns or nutrient dynamics between the two locations.

The CyanoHABs phase in St. Albans is characterized by fragmented and intermittent activity, reflected in the oscillating hit rate values throughout August and September. These fluctuations imply multiple short-lived or low-intensity CyanoHABs events, which challenge prediction continuity. Despite this, the model demonstrates consistent recovery in hit rate during sustained CyanoHABs phases, such as late September and early October for both years, where performance stabilizes with a higher hit rate.

Year-to-year differences are also notable. While 2023 exhibits more variable hit rates during the CyanoHABs season, 2024 shows a smoother and stable trajectory, indicating longer CyanoHABs events with more substantial alignment between

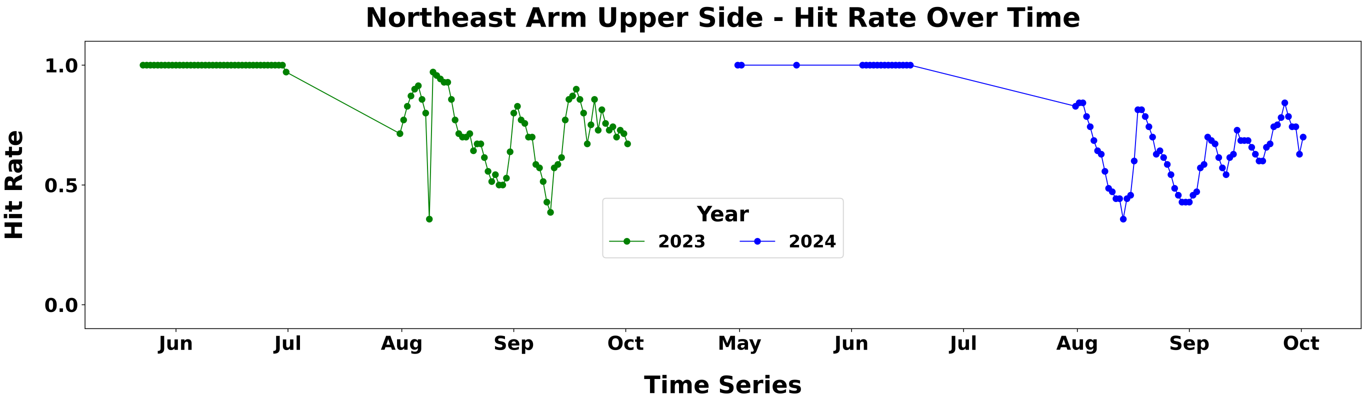

Fig. 8. Hit rate time series for the Northeast Arm Upper Side segment across 2023 and 2024. The diagram illustrates consistent fluctuations in hit rate during August to October, corresponding to fragmented and short-lived CyanoHAB events.

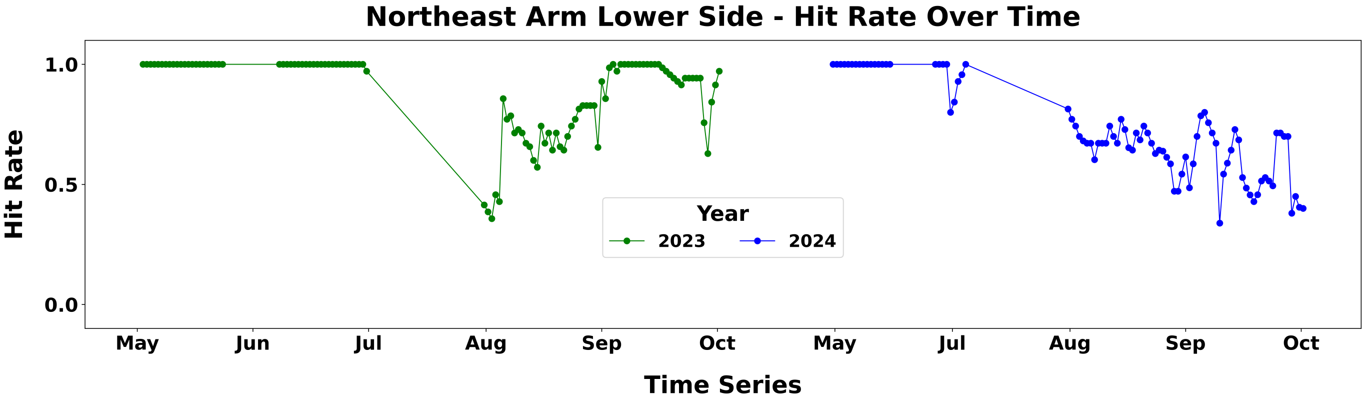

Fig. 9. Hit rate time series for the Northeast Arm Lower Side segment across 2023 and 2024. In 2023, the model achieves stable performance during sustained CyanoHABs periods in late September, reflecting more persistent CyanoHAB activity. In contrast, 2024 exhibits increased variability in hit rates, indicating fragmented and intermittent CyanoHABs patterns.

model forecasts and observations. This pattern suggests that CyanoHABs drivers in 2024 may have been more consistent, enhancing forecast reliability at this site.

*c) **Northeast Arm**:* The Northeast Arm, an oligotrophic zone characterized by relatively low nutrient levels and biological productivity, is divided into upper and lower sections for analysis due to their distinct CyanoHAB characteristics. The upper side is shown in Figure 8, exhibits CyanoHABs that start forming in late August and persist intermittently through October. The consistent fluctuations in hit rate during these months for both 2023 and 2024 suggest frequent, short-duration, and fragmented CyanoHABs events, which present challenges for sustained predictive accuracy.

The lower side is depicted in Figure 9, demonstrating a more stable CyanoHABs structure, particularly in late September 2023, as evidenced by the extended period of high hit rate. This suggests longer-lasting and more predictable CyanoHAB conditions during that period. However, 2024 presents a different pattern: the hit rate exhibits frequent fluctuations, indicating that CyanoHABs events in the lower region became more intermittent and fragmented, resembling the instability seen in the Upper Side. These variations highlight the interannual variability in CyanoHABs dynamics and environmental forcing.

The contrast between the two sub-regions, especially in 2023, demonstrates the benefit of spatial subdivision. While both sides are geographically adjacent, their temporal patterns, CyanoHABs persistence, and ecological responses differ meaningfully. This justifies the modeling decision to split the Northeast Arm, enabling more precise detection and characterization of localized CyanoHAB behavior.

IX. Discussion

Remote sensing data are used to forecast the occurrence of CyanoHABs in Lake Champlain's three key segments: Missisquoi Bay, St. Albans Bay, and Northeast Arm, over a 14-day forecast horizon. The CI Values from the Cyanobacterial Assessment network indicate the intensity of CyanoHABs. Other data features include Temperature and temporal data features. The proposed research addresses two fundamental questions: 1) Given the nature of sparse remote sensing data, can we precisely forecast the CyanoHABs? 2) If we can forecast CyanoHABs early, can we also predict their intensity?

The experimental results affirm both questions. The Transformer-BiLSTM can reliably predict the early occurrences of CyanoHABs, along with their intensities, across all five classes within a 14-day forecast horizon. Despite the inherent complexity of a 14-day forecasting window,

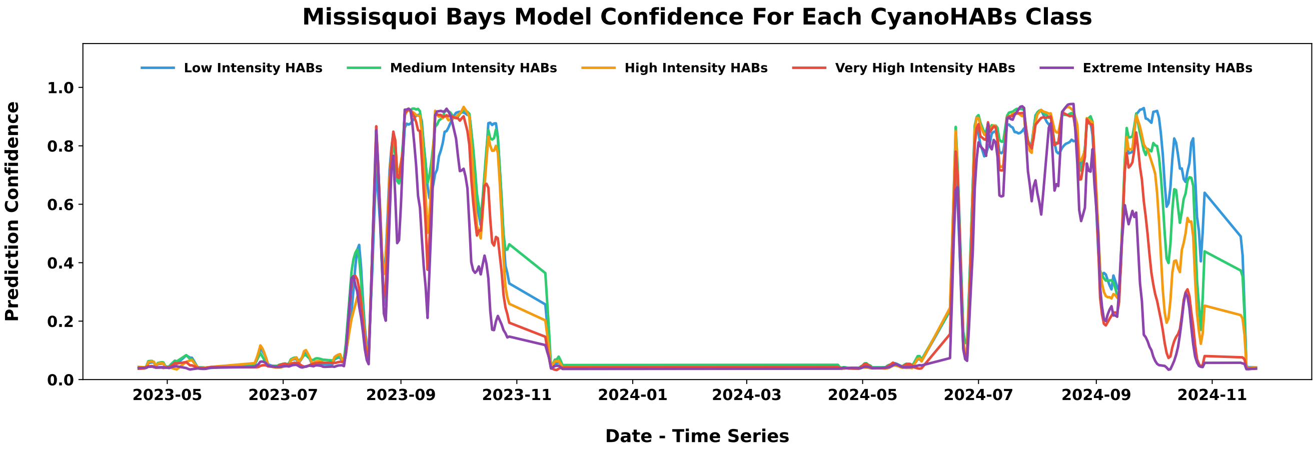

Fig. 10. Model prediction confidence scores for five CyanoHAB intensity classes in Missisquoi Bay. The graph displays confidence values for Low (blue), Medium (green), High (orange), Very High (red), and Extreme (purple) intensity classes. High confidence scores align with known bloom periods in August to October 2023 and June to September 2024, while near-zero confidence appears during non-bloom seasons, demonstrating the model's ability to recognize seasonal patterns.

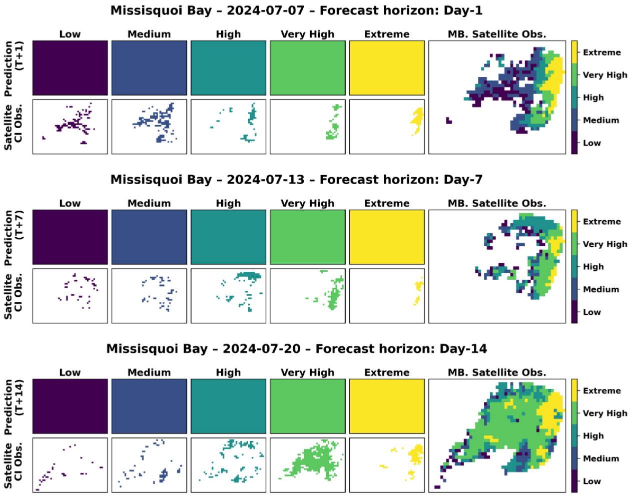
Missisquoi Bay – 2024-07-07 – Forecast horizon: Day-1
Missisquoi Bay – 2024-07-13 – Forecast horizon: Day-7
Missisquoi Bay – 2024-07-20 – Forecast horizon: Day-14
Low
Medium
High
Very High
Extreme
MB. Satellite Obs.
Prediction (T+1)
Prediction (T+7)
Prediction (T+14)
Satellite CI Obs.
Extreme
Very High
High
Medium
Low

Fig. 11. Spatial comparison between model forecasts and CyAN satellite observations of Missisquoi Bay during July 2024. Each row gives a different lead time—Day-1 (7 July), Day-7 (13 July), and Day-14 (20 July). The top shows the Transformer-BiLSTM prediction for each CI intensity class (Low, Medium, High, Very High, Extreme); the bottom shows the corresponding satellite mask, followed by the composite satellite map on the far right. One can observe the missing pixels in the complete composite of the Missisquoi Bay station. The model correctly preserves the correct class mix over the full 14-day horizon.

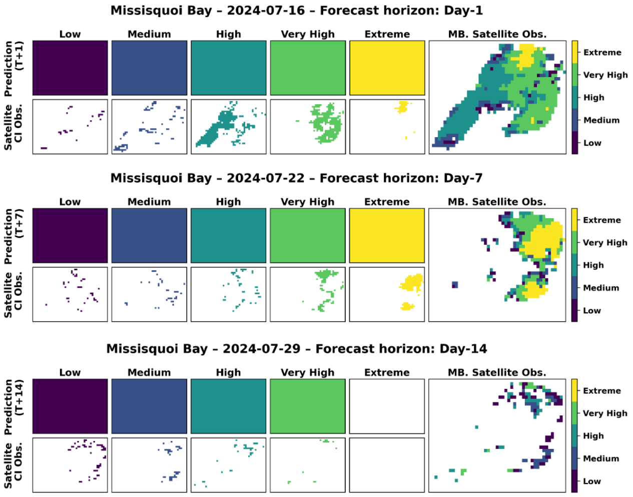
Missisquoi Bay – 2024-07-16 – Forecast horizon: Day-1
Missisquoi Bay – 2024-07-22 – Forecast horizon: Day-7
Missisquoi Bay – 2024-07-29 – Forecast horizon: Day-14
Low
Medium
High
Very High
Extreme
MB. Satellite Obs.
Prediction (T+1)
Prediction (T+7)
Prediction (T+14)
Satellite CI Obs.
Extreme
Very High
High
Medium
Low

Fig. 12. Spatial comparison between model forecasts and Cyan satellite observations of Missisquoi Bay from 16 July to 29 July, 2024. One can observe the alignment between predictions and targets. We can also see that on July 29, the targeted image is sparser compared to Day-7 and Day-1 images, which adds an additional layer of complexity due to not having enough pixels for predictions.

the Transformer-BiLSTM consistently demonstrates seasonal CyanoHABs events and predictions alignments across different stations and years. The model captures both the onset and duration of CyanoHAB activity and maintains strong forecasting performance even under varying seasonal patterns and inter-annual conditions. Across all regions, the hit rate time series reflects the seasonal nature of CyanoHABs, with prediction accuracy adapting to site-specific CyanoHABs dynamics. For instance, the model detects early CyanoHAB formation in Missisquoi Bay during June 2024, captures intermittent events in St. Albans Bay, and differentiates between the fragmented CyanoHAB behavior of the Upper Northeast Arm and the more sustained patterns seen in the Lower Side. These results also validate the spatial subdivision of the Northeast Arm, where distinct hit rate profiles reveal different ecological responses within geographically proximate zones.

Analysis of prediction confidence scores provides additional insights into the system's performance, further validating the model's forecasting reliability. The model provides confidence scores that align with known CyanoHAB

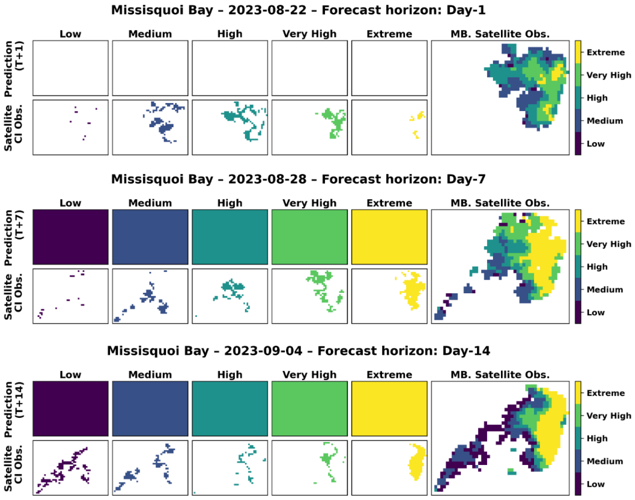
Missisquoi Bay - 2023-08-22 - Forecast horizon: Day-1
Missisquoi Bay - 2023-08-28 - Forecast horizon: Day-7
Missisquoi Bay - 2023-09-04 - Forecast horizon: Day-14
Low
Medium
High
Very High
Extreme
MB. Satellite Obs.
Prediction (T+1)
Prediction (T+7)
Prediction (T+14)
Satellite CI Obs.

Fig. 13. Same layout as earlier diagrams. The Day-1 forecast underestimates the initial outbreak, yet the Day-7 (28 August) and Day-14 (4 September) predictions catch up to the observed magnitude and intensity distribution. This rapid convergence shows that while abrupt bloom onsets can be missed, the model quickly adjusts once the larger-scale environmental drivers dominate.

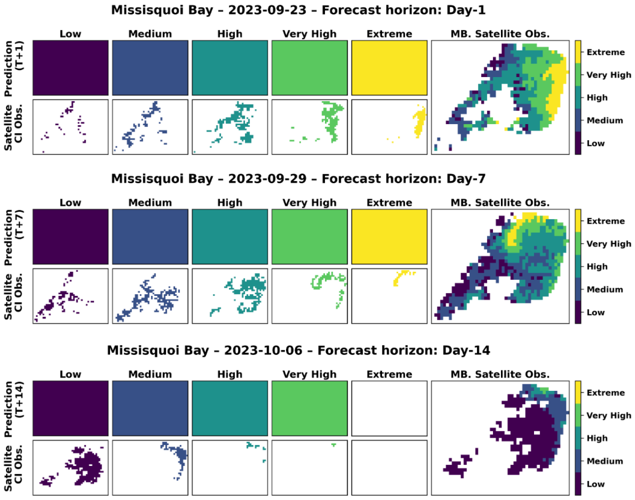
Missisquoi Bay - 2023-09-23 - Forecast horizon: Day-1
Missisquoi Bay - 2023-09-29 - Forecast horizon: Day-7
Missisquoi Bay - 2023-10-06 - Forecast horizon: Day-14
Low
Medium
High
Very High
Extreme
MB. Satellite Obs.
Prediction (T+1)
Prediction (T+7)
Prediction (T+14)
Satellite CI Obs.

Fig. 14. Forecast skill during a late-season bloom that extends from 23 September to 10 October 2023 in Missisquoi Bay. Both the model and the observations maintain High and Extreme-class pixels well past the usual summer window, confirming that the bloom persists into early October. The model mirrors this prolonged event, capturing its intensity on Day-1, retaining the full class spectrum on Day-7, and reflecting the gradual decline in Extreme pixels by Day-14.

patterns, as shown in Figure 10 for Missisquoi Bay. High confidence is observed during peak CyanoHAB periods across all intensity classes, whereas confidence levels are near zero during non-CyanoHAB seasons. These patterns demonstrate that the model has correctly learned the seasonal dynamics of CyanoHABs without overfitting. The apparent differences between 2023 and 2024 confidence patterns indicate that the model accurately captures year-to-year variations. During the CyanoHAB start and end periods, confidence scores differ between intensity classes, with low-intensity HABs exhibiting higher confidence than extreme events, reflecting the natural progression of CyanoHABs. These confidence patterns confirm that the model effectively forecasts CyanoHAB presence and intensity levels despite data sparsity.

Figures 11–14 further validate this behavior by setting predictions following satellite observations for four representative CyanoHABs events. These diagrams illustrate the bin-classification framework used in this study (See Section IV-C). Each row corresponds to a forecast lead time (Day-1, 7, 14). The five left-hand tiles show the predicted presence of CyanoHABs for the discrete CI-intensity bins, while the matching tiles beneath them display the satellite-derived masks for the same day. The composite on the far right shows the complete Missisquoi Bay station CI file after performing the data imputation (See Section IV-B). Since the problem is formulated at the bin level rather than at the true pixel scale, the model predicts each intensity bin and does not attempt sub-pixel localization. Figures 11–12 show two July 2024 cases, the forecast assigns pixels to all five intensity classes at the Day-1 horizon and maintains a similar class composition through Day-14, aligning with the observed CyanoHABs events. The Figure 13 illustrates a different behavior: the Day-1 forecast underestimates an abrupt flare-up, yet by Day-7 it reproduces the observed bloom magnitude and by Day-14 it represents the bay-wide distribution, indicating rapid model adjustment once larger-scale drivers prevail. The Figure 14 shows a late-September to early-October sequence that depicts a high-intensity bloom that persists beyond the typical season; both forecast and satellite data retain High- and Extreme-class pixels into early October, implying that warmer surface waters and extended stratification are lengthening bloom lifetimes.

Since six additional models were trained for performance evaluation against the Transformer-BiLSTM model (see Section VIII-C1), we evaluated an ensemble approach to assess potential performance gains. The ensemble method averaged predictions from all seven models to generate final forecasts. While the ensemble approach achieved marginally better performance with an F1 score of 79.09 compared to the Transformer-BiLSTM's F1 score of 78.86, the improvement was minimal at only 0.23 percentage points. However, ensemble implementation requires independent training and testing of each constituent model, significantly increasing computational complexity and deployment overhead. Given the marginal performance enhancement, the single Transformer-BiLSTM model presents a more practical solution for operational deployment, offering faster inference times and reduced computational requirements.

The model also demonstrated generalization to unusual CyanoHAB conditions not present in the training data, particularly evident during the June and July 2024 period. As illustrated in Figure 15, 2024 exhibited unprecedented early-season CyanoHAB events with exceptional intensity levels not observed during the training period. While training years showed CyanoHAB events predominantly in late July with moderate pixel counts, 2024 presented high-intensity events beginning in early June, with pixel counts higher than in any previous training year. This represents a significant departure from historical patterns, with hundreds of pixels across all intensity classes, including the extreme category. Despite never encountering such conditions during training, the model

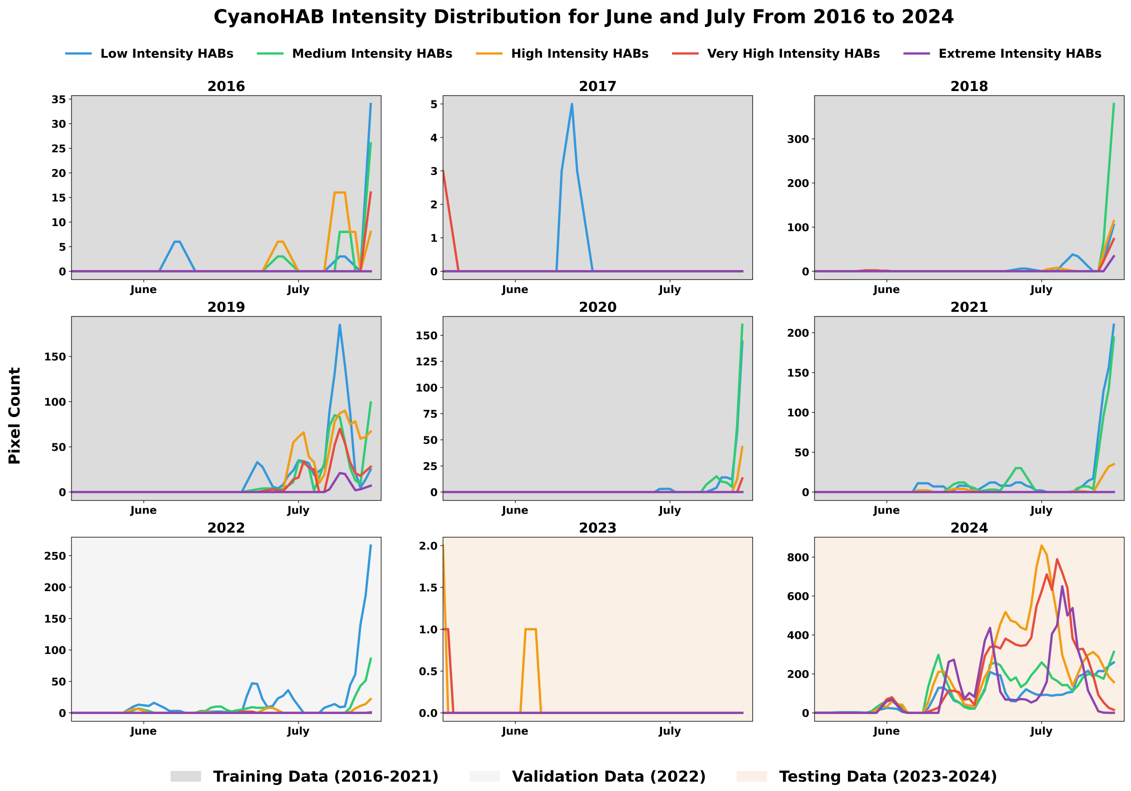

Fig. 15. CyanoHAB intensity distribution for June and July across all years from 2016 to 2024 in Missisquoi Bay, showing pixel counts for five intensity classes. Low is shown in blue, Medium in green, High in orange, Very High in red, and purple indicates Extreme Intensity CyanoHABs. Background colors indicate data partition: gray represents training data years, light gray represents validation years, and light orange represents testing years. One can observe the presence of CyanoHABs events in late July, with moderate or minimal pixel counts in June. The 2024 testing data shows a dramatic shift with unprecedented early-season blooms beginning in early June and exceptionally high pixel counts across all intensity classes.

accurately forecast these anomalous events as confirmed in Figure 6. This capacity to predict CyanoHAB patterns, rather than merely reproducing historical observations, validates the model's ability to learn fundamental environmental relationships governing CyanoHAB formation, rather than simply memorizing seasonal templates.

These predictions of unusual events reveal spatial and temporal heterogeneity across Lake Champlain that presents substantial forecasting challenges. The unprecedented early-season CyanoHAB events observed in Missisquoi Bay in June 2024, with pixel counts exceeding historical patterns, suggest potential shifts in CyanoHAB dynamics that may be attributable to changing environmental conditions. Each monitoring segment shows distinct CyanoHAB patterns: persistent high-intensity events in Missisquoi Bay, episodic patterns in St. Albans Bay, and complex regional variations in Northeast Arm. These differences highlight the localized nature of CyanoHAB formation processes. Notably, the counterintuitive finding is that the lower side of the Northeast Arm, despite its oligotrophic designation and distance from the nutrient-rich zone of Missisquoi Bay, exhibited more stable events than the upper side in 2024. This suggests that complex hydrodynamic processes, beyond simple nutrient gradients, influence the development of CyanoHABs. The significant inter-annual variability observed between 2023 and 2024 across all segments further complicates forecasting efforts, with some locations showing earlier CyanoHAB onset, others displaying more substantial fragmentation, and all exhibiting distinct intensity distribution patterns. These findings demonstrate that effective CyanoHAB forecasting requires sophisticated algorithms and consideration of site-specific ecological dynamics.

X. Conclusion

This work utilizes remote sensing data to present a Transformer-BiLSTM framework for forecasting CyanoHAB occurrences and intensity levels at three key segments in Lake Champlain: Missisquoi Bay, St. Albans Bay, and Northeast Arm. The system is designed to provide day-by-day forecasts across a 14-day horizon, offering fine-grained temporal insights into the dynamics of CyanoHABs. The research uses two remote sensing data features: Cyanobacterial Index values and Temperature. The system addresses data sparsity through a two-stage preprocessing pipeline combining forward fill and weighted temporal imputation, enabling effective forecasting despite significant missing data in both CI values and temperature measurements. A temporal data augmentation strategy is also presented to address the limited training dataset while preserving physically plausible relationships.

The model demonstrated generalization to novel CyanoHAB patterns, including unprecedented early-season high-intensity events in 2024 that were not represented in the training data. Analysis across different monitoring segments reveals distinct spatial characteristics: persistent high-intensity CyanoHABs in Missisquoi Bay, episodic patterns in St. Albans Bay, and complex regional variations in Northeast Arm. These segment-specific patterns and significant inter-annual variability between 2023 and 2024 highlight the importance of localized approaches for effective CyanoHAB management. Compared to a persistence baseline, which performed better on day 1, the Transformer-BiLSTM consistently outperformed across the 14-day horizon, with a 7.62 percentage point advantage by day 14, confirming its robustness in capturing the evolving behavior of CyanoHABs. Comparative evaluation against seven advanced deep learning architectures further validated the Transformer-BiLSTM's performance for extended forecasting horizons. While Transformer-only models demonstrated competitive performance in short-term predictions, the Transformer-BiLSTM architecture outperformed them for longer-duration forecasting tasks. Prediction confidence scores provided further validation of the model's interpretability. High confidence aligned with peak CyanoHAB periods across all intensity classes, while confidence diminished in off-season intervals. Inter-annual variations in confidence patterns between 2023 and 2024 reflected the model's sensitivity to year-specific dynamics without overfitting.

The proposed framework effectively forecasts CyanoHAB occurrences and intensity levels over a 14-day horizon. However, we acknowledge potential areas for future enhancement. Current predictions provide segment-level resolution; moving toward pixel-level localization will improve spatial specificity. The reliance on MODIS temperature data poses a challenge due to its planned decommissioning, which necessitates transitioning to alternative sources like Visible Infrared Imaging Radiometer Suite, or Sentinel-3 Sea and Land Surface Temperature Radiometer. Future work addresses these limitations by developing a pixel-level forecasting system and integrating additional satellite-based input features to ensure continuity.

FUNDING

Funding for this project was provided by the National Oceanic and Atmospheric Administration, which was awarded to the Cooperative Institute for Research on Hydrology through the NOAA Cooperative Agreement with The University of Alabama, NA22NWS4320003.

CONFLICT OF INTEREST STATEMENT

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

REFERENCES

[1] S. B. A. Rolim, B. K. Veettil, A. P. Vieiro, A. B. Kessler, and C. Gonzatti, "Remote sensing for mapping algal blooms in freshwater lakes: A review," *Environmental Science and Pollution Research*, vol. 30, no. 8, pp. 19 602–19 616, 2023.

[2] V. Klemas, "Remote sensing of algal blooms: an overview with case studies," *Journal of coastal research*, vol. 28, no. 1A, pp. 34–43, 2012.

[3] A. S. Pamula, H. Gholizadeh, M. J. Krzmarzick, W. E. Mausbach, and D. J. Lampert, "A remote sensing tool for near real-time monitoring of harmful algal blooms and turbidity in reservoirs," *JAWRA Journal of the American Water Resources Association*, 2023.

[4] D. L. Roelke, J. P. Grover, B. W. Brooks, J. Glass, D. Buzan, G. M. Southard, L. Fries, G. M. Gable, L. Schwierzke-Wade, M. Byrd *et al.*, "A decade of fish-killing prymnesium parvum blooms in texas: roles of inflow and salinity," *Journal of Plankton Research*, vol. 33, no. 2, pp. 243–253, 2011.

[5] A. C. Thomas, D. W. Townsend, and R. Weatherbee, "Satellite-measured phytoplankton variability in the gulf of maine," *Continental Shelf Research*, vol. 23, no. 10, pp. 971–989, 2003.

[6] 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.

[7] M. G. Antoniou, A. A. De La Cruz, and D. D. Dionysiou, "Cyanotoxins: New generation of water contaminants," pp. 1239–1243, 2005.

[8] Y. Deng, M. Wu, H. Zhang, L. Zheng, Y. Acosta, and T.-T. D. Hsu, "Addressing harmful algal blooms (habs) impacts with ferrate (vi): Simultaneous removal of algal cells and toxins for drinking water treatment," *Chemosphere*, vol. 186, pp. 757–761, 2017.

[9] L. C. Backer, "Cyanobacterial harmful algal blooms (cyanohabs): Developing a public health response," *Lake and reservoir Management*, vol. 18, no. 1, pp. 20–31, 2002.

[10] L. E. Fleming, B. Kirkpatrick, L. C. Backer, C. J. Walsh, K. Nierenberg, J. Clark, A. Reich, J. Hollenbeck, J. Benson, Y. S. Cheng *et al.*, "Review of florida red tide and human health effects," *Harmful algae*, vol. 10, no. 2, pp. 224–233, 2011.

[11] B. W. Brooks, J. M. Lazorchak, M. D. Howard, M.-V. V. Johnson, S. L. Morton, D. A. Perkins, E. D. Reavie, G. I. Scott, S. A. Smith, and J. A. Steevens, "Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems?" *Environmental toxicology and chemistry*, vol. 35, no. 1, pp. 6–13, 2016.

[12] J. H. Landsberg, "The effects of harmful algal blooms on aquatic organisms," *Reviews in fisheries science*, vol. 10, no. 2, pp. 113–390, 2002.

[13] I. Sanseverino, A. D. S. CONDUTO, L. Pozzoli, S. Dobricic, T. Lettieri *et al.*, "Algal bloom and its economic impact," 2016.

[14] E. Berdalet, L. E. Fleming, R. Gowen, K. Davidson, P. Hess, L. C. Backer, S. K. Moore, P. Hoagland, and H. Enevoldsen, "Marine harmful algal blooms, human health and wellbeing: challenges and opportunities in the 21st century," *Journal of the Marine Biological Association of the United Kingdom*, vol. 96, no. 1, pp. 61–91, 2016.

[15] R. Kudela, E. Berdalet, S. Bernard, M. Burford, L. Fernand, S. Lu, S. Roy, P. Tester, G. Usup, R. Magnien *et al.*, "Harmful algal blooms. a scientific summary for policy makers." 2015.

[16] J. S. Hecht, A. Zia, P. J. Clemins, A. W. Schroth, J. M. Winter, P. D. Oikonomou, and D. M. Rizzo, "Modeling the sensitivity of cyanobacteria blooms to plausible changes in precipitation and air temperature variability," *Science of The Total Environment*, vol. 812, p. 151586, 2022.

[17] C. J. Gobler, O. M. Doherty, T. K. Hattenrath-Lehmann, A. W. Griffith, Y. Kang, and R. W. Litaker, "Ocean warming since 1982 has expanded the niche of toxic algal blooms in the north atlantic and north pacific oceans," *Proceedings of the National Academy of Sciences*, vol. 114, no. 19, pp. 4975–4980, 2017.

[18] C. J. Gobler, "Climate change and harmful algal blooms: insights and perspective," *Harmful algae*, vol. 91, p. 101731, 2020.

[19] J. C. Ho and A. M. Michalak, "Exploring temperature and precipitation impacts on harmful algal blooms across continental us lakes," *Limnology and Oceanography*, vol. 65, no. 5, pp. 992–1009, 2020.

[20] H. W. Paerl and T. G. Otten, "Duelling 'cyanohabs': unravelling the environmental drivers controlling dominance and succession among diazotrophic and non-n2-fixing harmful cyanobacteria," *Environmental microbiology*, vol. 18, no. 2, pp. 316–324, 2016.

[21] Y. Cha, K. H. Cho, H. Lee, T. Kang, and J. H. Kim, "The relative importance of water temperature and residence time in predicting cyanobacteria abundance in regulated rivers," *Water research*, vol. 124, pp. 11–19, 2017.

[22] Y. Tang, Y. Feng, S. Fung, V. R. Xomchuk, M. Jiang, T. Moore, and J. Beckler, "Spatiotemporal deep-learning-based algal bloom prediction for lake okeechobee using multisource data fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 8318–8331, 2022.

[23] C. Lapucci, F. Maselli, G. Chini Zittelli, G. Betti, V. Vannucchi, M. Perna, S. Taddei, B. Gozzini, A. Ortolani, and C. Brandini, "Towards the prediction of favourable conditions for the harmful algal bloom onset of ostreopsis ovata in the ligurian sea based on satellite and model data," *Journal of Marine Science and Engineering*, vol. 10, no. 4, p. 461, 2022.

[24] A. Igwaran, A. J. Kayode, K. M. Moloantoa, Z. P. Khetsha, and J. O. Unuofin, "Cyanobacteria harmful algae blooms: causes, impacts, and risk management," *Water, Air, & Soil Pollution*, vol. 235, no. 1, p. 71, 2024.

[25] B. L. Marrone, S. Banerjee, A. Talapatra, C. R. Gonzalez-Esquer, and G. Pilania, "Toward a predictive understanding of cyanobacterial harmful algal blooms through ai integration of physical, chemical, and biological data," *ACS Es&t Water*, vol. 4, no. 3, pp. 844–858, 2023.

[26] Y. Zhou, W. Yan, and W. Wei, "Effect of sea surface temperature and precipitation on annual frequency of harmful algal blooms in the east china sea over the past decades," *Environmental Pollution*, vol. 270, p. 116224, 2021.

[27] M. L. Larsen, H. M. Baulch, S. L. Schiff, D. F. Simon, S. Sauvé, and J. J. Venkiteswaran, "Extreme rainfall drives early onset cyanobacterial bloom," *Facets*, vol. 5, no. 1, pp. 899–920, 2020.

[28] S. Haakonsson, L. Rodríguez-Gallego, A. Somma, and S. Bonilla, "Temperature and precipitation shape the distribution of harmful cyanobacteria in subtropical lotic and lentic ecosystems," *Science of the Total Environment*, vol. 609, pp. 1132–1139, 2017.

[29] D. A. Lemley, J. B. Adams, and J. L. Largier, "Harmful algal blooms as a sink for inorganic nutrients in a eutrophic estuary," *Marine Ecology Progress Series*, vol. 663, pp. 63–76, 2021.

[30] J. Wang, A. F. Bouwman, X. Liu, A. H. Beusen, R. Van Dingenen, F. Dentener, Y. Yao, P. M. Glibert, X. Ran, Q. Yao *et al.*, "Harmful algal blooms in chinese coastal waters will persist due to perturbed nutrient ratios," *Environmental Science & Technology Letters*, vol. 8, no. 3, pp. 276–284, 2021.

[31] P. M. Glibert, R. Maranger, D. J. Sobota, and L. Bouwman, "Further evidence of the haber-bosch—harmful algal bloom (hb-hab) link and the risk of suggesting hab control through phosphorus reductions only," *Just Enough Nitrogen: Perspectives on how to get there for regions with too much and too little nitrogen*, pp. 255–282, 2020.

[32] Y. Zhang, C. Chen, P. Xue, R. C. Beardsley, and P. J. Franks, "A view of physical mechanisms for transporting harmful algal blooms to massachusetts bay," *Marine Pollution Bulletin*, vol. 154, p. 111048, 2020.

[33] E. Silva, F. Counillon, J. Brajard, A. Korosov, L. H. Pettersson, A. Samuelsen, and N. Keenlyside, "Twenty-one years of phytoplankton bloom phenology in the barents, norwegian, and north seas," *Frontiers in Marine Science*, vol. 8, p. 746327, 2021.

[34] B. L. Keeler, S. Polasky, K. A. Brauman, K. A. Johnson, J. C. Finlay, A. O'Neill, K. Kovacs, and B. Dalzell, "Linking water quality and well-being for improved assessment and valuation of ecosystem services," *Proceedings of the National Academy of Sciences*, vol. 109, no. 45, pp. 18 619–18 624, 2012.

[35] 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.

[36] A. Giani, Z. E. Taranu, G. von Rückert, and I. Gregory-Eaves, "Comparing key drivers of cyanobacteria biomass in temperate and tropical systems," *Harmful Algae*, vol. 97, p. 101859, 2020.

[37] T. Kim, J. Shin, D. Lee, Y. Kim, E. Na, J.-H. Park, C. Lim, and Y. Cha, "Simultaneous feature engineering and interpretation: Forecasting harmful algal blooms using a deep learning approach," *Water Research*, vol. 215, p. 118289, 2022.

[38] D. Lee, M. Kim, B. Lee, S. Chae, S. Kwon, and S. Kang, "Integrated explainable deep learning prediction of harmful algal blooms," *Technological Forecasting and Social Change*, vol. 185, p. 122046, 2022.

[39] P. Villanueva, J. Yang, L. Radmer, X. Liang, T. Leung, K. Ikuma, E. D. Swanner, A. Howe, and J. Lee, "One-week-ahead prediction of cyanobacterial harmful algal blooms in iowa lakes," *Environmental Science & Technology*, vol. 57, no. 49, pp. 20 636–20 646, 2023.

[40] M. Liu, J. He, Y. Huang, T. Tang, J. Hu, and X. Xiao, "Algal bloom forecasting with time-frequency analysis: a hybrid deep learning approach," *Water Research*, vol. 219, p. 118591, 2022.

[41] E. Silva, F. Counillon, J. Brajard, L. H. Pettersson, and L. Naustvoll, "Forecasting harmful algae blooms: Application to dinophysis acuminata in northern norway," *Harmful Algae*, vol. 126, p. 102442, 2023.

[42] A. Molares-Ulloa, D. Rivero, J. G. Ruiz, E. Fernandez-Blanco, and L. de-la Fuente-Valentín, "Hybrid machine learning techniques in the management of harmful algal blooms impact," *Computers and Electronics in Agriculture*, vol. 211, p. 107988, 2023.

[43] N. Torbick and M. Corbiere, "A multiscale mapping assessment of lake champlain cyanobacterial harmful algal blooms," *International journal of environmental research and public health*, vol. 12, no. 9, pp. 11 560–11 578, 2015.

[44] K. A. Steidinger and K. Haddad, "Biologic and hydrographic aspects of red tides," *Bioscience*, vol. 31, no. 11, pp. 814–819, 1981.

[45] M. Izadi, M. Sultan, R. E. Kadiri, A. Ghannadi, and K. Abdelmohsen, "A remote sensing and machine learning-based approach to forecast the onset of harmful algal bloom," *Remote Sensing*, vol. 13, no. 19, p. 3863, 2021.

[46] N. F. Manning, Y.-C. Wang, C. M. Long, I. Bertani, M. J. Sayers, K. R. Bosse, R. A. Shuchman, and D. Scavia, "Extending the forecast model: Predicting western lake erie harmful algal blooms at multiple spatial scales," *Journal of Great Lakes Research*, vol. 45, no. 3, pp. 587–595, 2019.

[47] A. Gupta, M. M. Hantush, and R. S. Govindaraju, "Sub-monthly time scale forecasting of harmful algal blooms intensity in lake erie using remote sensing and machine learning," *Science of the Total Environment*, vol. 900, p. 165781, 2023.

[48] M. W. Matthews, "Near-term forecasting of cyanobacteria and harmful algal blooms in lakes using simple univariate methods with satellite remote sensing data," *Inland Waters*, vol. 13, no. 1, pp. 62–73, 2023.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[50] Z. Yan, S. Kamanmalek, and N. Alamdari, "Predicting coastal harmful algal blooms using integrated data-driven analysis of environmental factors," *Science of The Total Environment*, vol. 912, p. 169253, 2024.

[51] Y. Zhang, Y. Wang, J. Chen, L. Lin, W. Xiao, and B. Huang, "Enhancing short-term algal bloom forecasting through an anti-mimicking hybrid deep learning method," *Journal of Environmental Management*, vol. 379, p. 124832, 2025.

[52] T. Kim, J. Shin, and Y. Cha, "Incorporation of feature engineering and attention mechanisms into deep learning models to develop an early warning system for harmful algal blooms," *Journal of Cleaner Production*, vol. 414, p. 137564, 2023.

[53] J. M. Ahn, J. Kim, and K. Kim, "Ensemble machine learning of gradient boosting (xgboost, lightgbm, catboost) and attention-based cnn-lstm for harmful algal blooms forecasting," *Toxins*, vol. 15, no. 10, p. 608, 2023.

[54] A. R. Pearce, D. M. Rizzo, M. C. Watzin, and G. K. Druschel, "Unraveling associations between cyanobacteria blooms and in-lake environmental conditions in missisquoi bay, lake champlain, usa, using a modified self-organizing map," *Environmental science & technology*, vol. 47, no. 24, pp. 14 267–14 274, 2013.

[55] L. C. Bowling, S. Blais, and M. Sinotte, "Heterogeneous spatial and temporal cyanobacterial distributions in missisquoi bay, lake champlain: An analysis of a 9 year data set," *Journal of Great Lakes Research*, vol. 41, no. 1, pp. 164–179, 2015.

[56] P. D. Isles, C. D. Giles, T. A. Gearhart, Y. Xu, G. K. Druschel, and A. W. Schroth, "Dynamic internal drivers of a historically severe cyanobacteria bloom in lake champlain revealed through comprehensive monitoring," *Journal of Great Lakes Research*, vol. 41, no. 3, pp. 818–829, 2015.

[57] P. R. Hill, A. Kumar, M. Temimi, and D. R. Bull, "Habnet: Machine learning, remote sensing based detection and prediction of harmful algal blooms," *arXiv preprint arXiv:1912.02305*, 2019.

[58] T. T. Wynne, "Remote sensing of cyanobacterial blooms in lake champlain with a focus on missisquoi bay," *Journal of Great Lakes Research*, p. 102293, 2024.

[59] S. Liu, W. Glamore, B. Tamburic, A. Morrow, and F. Johnson, "Remote sensing to detect harmful algal blooms in inland waterbodies," *Science of The Total Environment*, vol. 851, p. 158096, 2022.

[60] M. H. Myer, E. Urquhart, B. A. Schaeffer, and J. M. Johnston, "Spatio-temporal modeling for forecasting high-risk freshwater cyanobacterial harmful algal blooms in florida," *Frontiers in environmental science*, vol. 8, p. 581091, 2020.

[61] Y. Qiu, H. Liu, F. Liu, D. Li, C. Liu, W. Liu, J. Huang, Q. Xiao, J. Luo, and H. Duan, "Development of a collaborative framework for quantitative monitoring and accumulation prediction of harmful algal blooms in nearshore areas of lakes," *Ecological Indicators*, vol. 156, p. 111154, 2023.

[62] Z. Cao, R. Ma, H. Duan, N. Pahlevan, J. Melack, M. Shen, and K. Xue, "A machine learning approach to estimate chlorophyll-a from landsat-8

measurements in inland lakes," *Remote Sensing of Environment*, vol. 248, p. 111974, 2020.

[63] Y. Qiu, J. Huang, J. Luo, Q. Xiao, M. Shen, P. Xiao, Z. Peng, Y. Jiao, and H. Duan, "Monitoring, simulation and early warning of cyanobacterial harmful algal blooms: An upgraded framework for eutrophic lakes," *Environmental Research*, vol. 264, p. 120296, 2025.

[64] A. Trescott, "Remote sensing models of algal blooms and cyanobacteria in lake champlain," 2012.

[65] S. M. Wheeler, L. A. Morrissey, S. N. Levine, G. P. Livingston, and W. F. Vincent, "Mapping cyanobacterial blooms in lake champlain's missisquoi bay using quickbird and meris satellite data," *Journal of Great Lakes Research*, vol. 38, pp. 68–75, 2012.

[66] R. S. Lunetta, B. A. Schaeffer, R. P. Stumpf, D. Keith, S. A. Jacobs, and M. S. Murphy, "Evaluation of cyanobacteria cell count detection derived from meris imagery across the eastern usa," *Remote Sensing of Environment*, vol. 157, pp. 24–34, 2015.

[67] A. El-Alem, K. Chokmani, I. Laurion, S. E. El-Adlouni, S. Raymond, and C. Ratté-Fortin, "Ensemble-based systems to monitor algal bloom with remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7955–7971, 2019.

[68] M. C. Vaughan, M. K. Campbell, L. Fisher, B. O'Brien, R. M. Gorney, A. Shambaugh, L. S. Sopher, O. Pierson, and E. A. Howe, "Lake champlain community scientist volunteer network communicates critical cyanobacteria information to region-wide stakeholders," *Journal of Contemporary Water Research & Education*, vol. 174, no. 1, pp. 6–20, 2021.

[69] J. Decerega, O. Guerra, and D. Tramonte, "The value of lake champlain," *Dartmouth College, PRS Policy Brief*, pp. 1516–01, 2016.

[70] Lake Champlain Basin Program, "Lake champlain basin – facts," 2023, accessed: 2024-03-08. [Online]. Available: https://www.lcbp.org/about-the-basin/facts/

[71] United States Environmental Protection Agency. (2025) Lake champlain. Accessed: Feb. 24, 2025. [Online]. Available: https://www.epa.gov

[72] D. E. Facey, J. E. Marsden, T. B. Mihuc, and E. A. Howe, "Lake champlain 2010: A summary of recent research and monitoring initiatives," *Journal of Great Lakes Research*, vol. 38, pp. 1–5, 2012.

[73] S. N. Levine, A. Lini, M. L. Ostrofsky, L. Bunting, H. Burgess, P. R. Leavitt, D. Reuter, A. Lami, P. Guilizzoni, and E. Gilles, "The eutrophication of lake champlain's northeastern arm: Insights from paleolimnological analyses," *Journal of Great Lakes Research*, vol. 38, pp. 35–48, 2012.

[74] NASA Ocean Biology Processing Group (OBPG), "Cyanobacteria assessment network (cyan), merged sentinel-3a and sentinel-3b olci regional mapped cyanobacteria index (ci) data, version 6.0," NASA Ocean Biology Distributed Active Archive Center (OB.DAAC), 2024, accessed on 2024/12/15. [Online]. Available: https://oceandata.sci.gsfc.nasa.gov/api/cyan_file_search

[75] G. Hulley and S. Hook, "Viirs/npp land surface temperature and emissivity daily l3 global 1km sin grid day v001," https://doi.org/10.5067/VIIRS/VNP21A1D.001, 2018, data set.

[76] 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.

[77] T. T. Wynne, R. P. Stumpf, M. C. Tomlinson, and J. Dyble, "Characterizing a cyanobacterial bloom in western lake erie using satellite imagery and meteorological data," *Limnology and Oceanography*, vol. 55, no. 5, pp. 2025–2036, 2010.

[78] T. Wynne, A. Meredith, T. Briggs, and W. Litaker, "Harmful algal bloom forecasting branch ocean color satellite imagery processing guidelines," NOAA Technical Memorandum NOS NCCOS, Tech. Rep. 296, 2018.

[79] S. Mishra, R. P. Stumpf, B. Schaeffer, P. J. Werdell, K. A. Loftin, and A. Meredith, "Evaluation of a satellite-based cyanobacteria bloom detection algorithm using field-measured microcystin data," *Science of The Total Environment*, vol. 774, p. 145462, 2021.

[80] M. M. Coffer, B. A. Schaeffer, J. A. Darling, E. A. Urquhart, and W. B. Salls, "Quantifying national and regional cyanobacterial occurrence in us lakes using satellite remote sensing," *Ecological Indicators*, vol. 111, p. 105976, 2020.

[81] J. M. Clark, B. A. Schaeffer, J. A. Darling, E. A. Urquhart, J. M. Johnston, A. R. Ignatius, M. H. Myer, K. A. Loftin, P. J. Werdell, and R. P. Stumpf, "Satellite monitoring of cyanobacterial harmful algal bloom frequency in recreational waters and drinking water sources," *Ecological indicators*, vol. 80, pp. 84–95, 2017.

[82] R. P. Stumpf, T. W. Davis, T. T. Wynne, J. L. Graham, K. A. Loftin, T. H. Johengen, D. Gossiaux, D. Palladino, and A. Burtner, "Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria," *Harmful algae*, vol. 54, pp. 160–173, 2016.

[83] B. A. Schaeffer, S. W. Bailey, R. N. Conmy, M. Galvin, A. R. Ignatius, J. M. Johnston, D. J. Keith, R. S. Lunetta, R. Parmar, R. P. Stumpf *et al.*, "Mobile device application for monitoring cyanobacteria harmful algal blooms using sentinel-3 satellite ocean and land colour instruments," *Environmental modelling & software*, vol. 109, pp. 93–103, 2018.

[84] Wyoming Department of Environmental Quality, "Big Sandy Reservoir Harmful Cyanobacterial Investigation 2018," Wyoming Department of Environmental Quality, Cheyenne, WY, USA, Tech. Rep., 2018, Technical Report. [Online]. Available: http://deq.wyoming.gov/media/attachments/Water

[85] ——, "Eden Reservoir Harmful Cyanobacterial Investigation 2018," Wyoming Department of Environmental Quality, Cheyenne, WY, USA, Tech. Rep., 2018, Technical Report. [Online]. Available: http://deq.wyoming.gov/media/attachments/Water

[86] ——, "Pathfinder Reservoir Harmful Cyanobacterial Bloom Investigation 2018," Wyoming Department of Environmental Quality, Cheyenne, WY, USA, Tech. Rep., 2018, Technical Report. [Online]. Available: http://deq.wyoming.gov/media/attachments/Water

[87] J.-M. Beckers and M. Rixen, "Eof calculations and data filling from incomplete oceanographic datasets," *Journal of Atmospheric and oceanic technology*, vol. 20, no. 12, pp. 1839–1856, 2003.

[88] Y. Qiu, C. Wei, F. Zhang, Y. Ge, H. Wang, Y. Jiao, Q. Xiao, and J. Luo, "Gap filling for satellite-derived products of lake aquatic environment using historical big data," *Geo-spatial Information Science*, pp. 1–11, 2025.

[89] W. Shi and M. Wang, "Ocean responses to hurricane ian from daily gap-free satellite observations," *Remote Sensing Letters*, vol. 14, no. 8, pp. 877–889, 2023.

[90] X. Ye, M. Lin, B. Zou, X. Wang, and Z. Lin, "Reconstructing global chlorophyll-a concentration for the cocts aboard chinese ocean color satellites via the dineof method," *Remote Sensing*, vol. 17, no. 20, p. 3433, 2025.

[91] Vermont Center for Geographic Information, "VT Lake Champlain Bathymetry," 2024, lake Champlain bathymetric data derived from NOAA nautical charts. Vermont Open Geodata Portal. [Online]. Available: https://geodata.vermont.gov/datasets/vt-lake-champlain-bathymetry/about

[92] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[93] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[94] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[95] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[96] T. Chen, "Xgboost: A scalable tree boosting system," *Cornell University*, 2016.

[97] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[98] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[99] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[100] K. E. Kavanaugh, K. Derner, K. M. Fisher, E. E. P. David, C. Urizar, and R. Merlini, "Assessment of the eastern gulf of mexico harmful algal bloom operational forecast system (gomx hab-ofs): A comparative analysis of forecast skill and utilization from october 1, 2004 to april 30, 2008," 2013.

[101] C. R. Anderson, R. M. Kudela, M. Kahru, Y. Chao, L. K. Rosenfeld, F. L. Bahr, D. M. Anderson, and T. A. Norris, "Initial skill assessment of the california harmful algae risk mapping (c-harm) system," *Harmful Algae*, vol. 59, pp. 1–18, 2016.