Karlsruher Institut für Technologie

Development of Big Data and Deep Learning Concepts in Water Environment Management

example of three different water bodies in China, from monitoring to prediction

Zur Erlangung des akademischen Grades eines
DOKTORS DER INGENIEURWISSENSCHAFTEN (Dr.-Ing.)
von der KIT-Fakultät für Bauingenieur-, Geo- und Umweltwissenschaften

des Karlsruher Instituts für Technologie (KIT)
genehmigte
DISSERTATION
von
M.Sc.Jing Qian
aus Chongqing, China

Tag der mündlichen Prüfung: 21.07.2023

Referent: Prof. Dr. Stefan Norra
Korreferent: Prof. Dr. Yonghong Bi
Korreferent: Prof. Dr. Stefan Hinz

Karlsruhe 2023

Erklärung

Ich versichere wahrheitsgemäß, die Arbeit selbstständig angefertigt, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde.

Karlsruhe den 09. 04. 2023

Mottos

Nicht auf Preußens Liberalismus sieht Deutschland, sondern auf seine Macht; Bayern, Württemberg, Baden mögen dem Liberalismus indulgieren, darum wird ihnen doch keiner Preußens Rolle anweisen; Preußen muß seine Kraft zusammenfassen und zusammenhalten auf den günstigen Augenblick, der schon einige Male verpaßt ist; Preußens Grenzen nach den Wiener Verträgen sind zu einem gesunden Staatsleben nicht günstig; nicht durch Reden und Majoritätsbeschlüsse werden die großen Fragen der Zeit entschieden – das ist der große Fehler von 1848 und 1849 gewesen –, sondern durch Eisen und Blut.

— Otto von Bismarck

I do not plan to come back. I have no reason to come back. I plan to do my best to help the Chinese people build up the nation to where they can live with dignity and happiness.

— Xuesen Qian

Thought and learning are of small value unless translated into action.

— Yangming Wang

Acknowledgements

I would like to take this opportunity to express my deepest gratitude to everyone who has supported, guided, and encouraged me throughout the course of my Ph.D. journey. The completion of this thesis would not have been possible without the contributions of numerous individuals who have played a vital role in my academic and personal growth.

First and foremost, I would like to extend my sincerest appreciation to my supervisor, Prof. Dr. Stefan Norra, for his unwavering support, guidance, and mentorship. His profound knowledge, insightful feedback, and dedication to my success have been invaluable in shaping my research and fostering my growth as a scholar. I am honored to have had the opportunity to work under their tutelage.

I am also deeply grateful to my co-supervisor, Prof. Dr. Yonghong Bi, for his expert advice and continuous encouragement throughout my research journey. His enthusiasm, wisdom, and commitment to my project have been essential in the development and completion of this thesis.

My heartfelt thanks go to the members of my doctoral committee, Prof. Dr. Stefan Hinz, Prof. Dr. Andreas Tiehm, Prof. Dr. Olivier Eiff, and Prof. Dr. Florian Wittmann, for their time, constructive criticism, and valuable suggestions. Their expertise and dedication have played a crucial role in refining my work and ensuring its quality.

I would like to express my appreciation to the faculty, staff, and fellow students of the Institute of Applied Geosciences at KIT and the Institute of Hydrobiology at CAS. In particular, I am grateful to Jonas Bauer for his camaraderie, insightful discussions, and continuous support, which have enriched my understanding and contributed to my personal and academic growth.

I am indebted to the numerous researchers and scholars whose work has laid the groundwork for my research. Their pioneering efforts have inspired me to delve deeper into my chosen field and strive for excellence in my work.

I would like to extend my heartfelt gratitude to my friends and family for their unwavering love, support, and understanding throughout this journey. I am particularly grateful to my parents, my wife and my little son, for their constant encouragement, unconditional love, and belief in my abilities. They have instilled in me a passion for learning and a strong work ethic, which have been the foundation of my academic pursuits.

My deepest appreciation goes to my partner, Xiaobai Xue and Nan Pu, for their patience, love, and unwavering support during the most challenging moments of this journey. Their

constant encouragement, understanding, and companionship have been my source of strength and inspiration, and I could not have achieved this milestone without them by my side.

In conclusion, I am profoundly grateful to all those who have contributed to the success of this thesis. The completion of this work would not have been possible without the support, guidance, and encouragement of numerous individuals, and I am forever grateful for their contributions.

Zusammenfassung

Diese Doktorarbeit untersucht das Potenzial von Big Data im Wasserumweltmanagement (WEM), einem kritischen Aspekt der nachhaltigen Entwicklung. Ein Big-Data-Framework wird erstellt, erstellt, um vier zentrale Ziele zu erreichen: Weiterentwicklung von Big-Data Methoden für die Überwachung und Bewertung der Wasserqualität, die Identifizierung von Wachstumsfaktoren für die Algenentwicklung und der Aufbau eine Frühwarnsystems des Algenwachstums.

Der Aufgabenknoten zur Überwachung der Wasserqualität konzentriert sich auf das Qingcaosha-Reservoir und verwendet eine Umwelt-Big-Data-Plattform (EBDP), die im Rahmen dieser Doktorarbeit aus Satellitenfernerkundungsdaten (Sentinel-2) und Kreuzfahrtüberwachungsgeräten (BIOFISH) erstellt wurde. In dieser Arbeit wurden mittels eines tiefes neuronalen Netzwerkes die Daten ausgewertet und ein Überwachungskonzpet ausgearbeitet.

Der Aufgabenknoten zur Bewertung der Wasserqualität, ebenfalls im Qingcaosha-Reservoir angesiedelt, verwendet eine EBDP, die Ergebnisse aus dem Aufgabenknoten zur Überwachung der Wasserqualität enthält. Die verbesserte Technik des Deep Embedding Clustering (IDEC) zeigt vier deutlich getrennte gemeinsame Managementzonen auf, wobei die charakteristischen Faktoren jeder Zone durch statistische Methoden bestimmt werden und eine Grundlage für regionale gemeinsame Managementstrategien bieten.

Der Aufgabenknoten zur Identifizierung der Wachstumsfaktoren von Algen untersucht die mittlere Route des Süd-Nord-Wasserumleitung-Projekts unter Verwendung einer EBDP, die hochfrequente, vierjährige manuelle Stichprobendaten enthält. Bloomformer-1, entwickelt auf Basis der Transformer-Kernstruktur, erreicht eine hohe Leistung sowohl in Einzelunterstandort- als auch in Vollliniensimulationen des Algenwachstums und identifiziert Gesamtphosphor (TP) als den kritischsten treibenden Faktor. Die Kontrolle und Reduzierung von Phosphorgehalten sind wesentliche Strategien zur Steuerung des Algenwachstums und zur Aufrechterhaltung der Stabilität der Wasserqualität.

Der Aufgabenknoten zur Frühwarnung von Algenwachstum untersucht den Taihu-See und verwendet eine EBDP, die aus Daten eines vertikalen Wasserqualitätsüberwachungssystems (BIOLIFT) erstellt wurde. Wertgewinnungswerkzeuge umfassen DeepDPM, spektrale Clusterung und Bloomformer-2, ebenfalls entwickelt auf Basis der Transformer-Kernstruktur. Die kombinierte Verwendung von DeepDPM und spektraler Clusterung gruppiert Tiefenabschnitte in Cluster und optimiert so die Systemeffizienz. Bloomformer-2 zeigt hervorragende Leistungen sowohl bei Einzelschritt- als auch bei Mehrschrittvorhersagen für alle Tiefenkombinationen,

wobei eine verbesserte Interpretierbarkeit die Zuverlässigkeit und Anwendbarkeit in realen Szenarien gewährleistet.

Zusammenfassend hebt das Ergebnis die zahlreichen Vollständigkeit von Big Data im WEM hervor, einschließlich hoher Anpassungsfähigkeit, Genauigkeit, Umfassendheit und feiner Granularität. Der durch vier Aufgabenknotenpunkte geformte Industriecluster-Effekt hat das Konzept einer präzisen Wasserumweltmanagement eingeläutet. Diese Forschung veranschaulicht das Potenzial datengesteuerter Ansätze bei der Bewältigung komplexer Herausforderungen im Wasserumweltmanagement, fördert unser Verständnis für das Wasserressourcenmanagement und bietet praktische Lösungen für das nachhaltige Wasserumweltmanagement. Wenn zukünftige Forschungen auf diesem Rahmen aufbauen und zusätzliche Aufgabenknoten integrieren, wird erwartet, dass die Vorteile und die Wirksamkeit des Big-Data-WEM-Ansatzes weiter zunehmen und letztendlich zum übergeordneten Ziel der nachhaltigen Entwicklung und verantwortungsvollen Bewirtschaftung der Wasserressourcen beitragen.

Abstract

This dissertation scrutinizes the potentiality of big data within the realm of Water Environment Management (WEM), a crucial facet of sustainable development. A big data framework is conceived and established with four task nodes: to further develop big data methodologies for monitoring and assessment of water quality; to identify algal growth driving factors; and to construct an early warning system for algal growth.

The water quality monitoring task node focuses on the Qingcaosha Reservoir, using an Environmental Big Data Platform (EBDP) built from satellite remote sensing (Sentinel-2) and cruise monitoring devices (BIOFISH) data in this Ph.D. thesis. Additionally, a deep neural network was used to analyze the data and develop a monitoring concept.

The water quality assessment task node, also centered on the Qingcaosha Reservoir, employs an EBDP comprising results from the water quality monitoring task node. The Improved Deep Embedding Clustering (IDEC) technique reveals four distinct joint management zones, with each zone's characteristic factors determined through statistical methods, providing a basis for regional joint management strategies.

The identification of algal growth driving factors task node examines the middle route of the South-to-North Water Diversion Project, using an EBDP containing high-frequency, four-year manual sampling data. Bloomformer-1, developed based on the Transformer core structure, achieves high performance in both single sub-site and full-line simulations of the algal growth, identifying total phosphorus (TP) as the most critical driving factor. Controlling and reducing phosphorus levels are essential strategies for managing algal growth and maintaining water quality stability.

The algal growth early warning task node investigates Lake Taihu, utilizing an EBDP constructed from data collected by a vertical water quality monitoring system (BIOLIFT). Value mining tools include DeepDPM, Spectral clustering, and Bloomformer-2, also developed based on the Transformer core structure. The combined use of DeepDPM and spectral clustering groups depth segments into clusters, optimizing system efficiency. Bloomformer-2 demonstrates outstanding performance in both single-step and multi-step predictions for all depth combinations, with enhanced interpretability ensuring reliability and applicability to real-world scenarios.

In summary, the result highlights the numerous advantages of big data in WEM, including high adaptability, accuracy, comprehensiveness, and fine granularity. The industry cluster effect, fashioned by four task nodes, has heralded the concept of precise WEM. This research

exemplifies the potential of data-driven approaches in addressing complex water environment management challenges, advancing our understanding of water resources management and offering practical solutions for sustainable water environment management. As future research builds upon this framework and integrates additional task nodes, the benefits and effectiveness of the big data WEM approach are expected to further increase, ultimately contributing to the overarching goal of sustainable development and responsible water resource management.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADCP	Acoustic Doppler Current Profiler
AGW	Institute of Applied Geoscience
BMBF	German Federal Ministry of Education and Research
BP	Back-scattered Particles
CDOM	Colored Dissolved Organic Matter
CH	Carinski-harabasz
CMT	Cruise Monitoring Technology
DEC	Deep Embedding Clustering
DL	Deep Learning
DNN	Deep Neural Network
DO	Dissolved Oxygen
EBDP	Environmental Big Data Platform
El.cond.	Electrical Conductivity
ETR	Extra Trees Regression
FLAASH	Fast Line of Sight Atmospheric Analysis of Hypercubes
GBRT	Gradient Boosting Regression Tree
IDEC	Improved Deep Embedding Clustering
IHB	Institute of Hydrobiology
KIT	Karlsruhe Institute of Technology
LSTM	Long Short-Term Memory
MAD	Median Absolute Deviation
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MHSA	Multi-Head-Self-Attention
MLP	Multilayer Perceptron
MLR	Multiple Linear Regression
MOST	Chinese Ministry of Science and Technology
MRP	The Middle Route of the South-to-North Water Diversion Project
MSE	Mean Squared Error
NLP	Natural Language Processing
RFR	Random Forest Regression
RMSE	Root-Mean-Square Error
RNN	Recurrent Neural Network

RST	Remote Sensing Technology
SIGN	Sino-German Water Supply Network
STD	Standard Deviation
SVR	Support Vector Regression
Temp.	Temperature
TLLER	Taihu Laboratory for Lake Ecosystem Research
TN	Total Nitrogen
TOC	Total Organic Carbon
TP	Total Phosphorus
TSP	Time Series Prediction
Turb.	Turbidity
USST	University of Shanghai for Science and Technology
WD	Wind Direction
WEM	Water Environment Management
WS	Wind Speed

1 Introduction

1.1 From small data to Big data in water environment management

Water environment management (WEM) is the practice of ensuring the sustainable use of water resources while protecting and preserving the environment [1]. It involves the management of water quality, aquatic ecosystems, and the overall health of watersheds. Effective water environment management is critical for the long-term sustainability of water resources and the preservation of aquatic ecosystems. The importance of water environment management is increasingly recognized as the world faces the challenges of population growth, urbanization, and climate change. These factors are placing unprecedented pressure on water resources and ecosystems, making it essential to implement effective management practices.

Early in their evolutionary process, humans realized that the world is not only made up of independent facts (i.e., data) but that these independent facts are interwoven by an intricate web of cause-and-effect relationships [2]. It is the exploration of these causal explanations that has shaped much of our scientific knowledge today. Historically, much of the knowledge in WEM has been gained through empirical or hypothesis-driven research, where the pace of synthesis was managed by individual researchers or research groups [3]. This is the hallmark of scientific research in the era of small data, where a hypothesis of causality is set, and then this hypothesis is verified through the collection and analysis of data in tightly controlled experiments. However, the conclusions obtained from this approach to research have been found to be insufficient (especially in the study of WEM) [4]. First, this research approach tends to consider only a single or a limited number of natural processes, while nature is a very complex system of super-many natural processes, and the individual natural processes interact with each other, leading to conclusions that often do not work well in practical applications. In other words, the natural environmental system is thought of as too simple. Second, the small amount of data used for inference and verification of causality makes the conclusions

poor in generalization, which means the conclusions are simply explainable under certain hard preconditions(e.g., temperature and air pressure).

In recent years, with the development of data science, scientists seem to have found an alternative path of using data-driven search for correlation instead of using a hypothesis-driven search for causality [5]. This is the core idea of the big data era, which is simply to know what instead of why. It has advantages. First, the interrelationship of the natural processes in the complex system is contained in the historical data, and the conclusions drawn from the study of the correlation of the historical data are not one-sided but systematic, which is much better in practical applications [6]. Secondly, the correlations are derived from an extensive dataset, thereby enhancing the generalizability of the findings. Simultaneously, as the relationships are not causal in nature, there are no any preconditions; instead, any potential preconditions are systematically incorporated as an additional dimension within the analysis [7].

In addition, due to the development of various sensors, the speed and dimensionality of data generation greatly exceed the capacity of traditional data compilation and analysis. In Earth observations, the total amount of data stored in the NASA Earth Observing System Data and Information System Archive at the beginning of 2017 was about 22 PB, and the NISAR satellite mission is expected to add up to 85 TB of data per day to the archive [8]. UAVs carrying multispectral cameras, in situ water quality monitors, and Acoustic Doppler Current Profiler (ADCP), which are widely used in WEM, are also generating large amounts of different kinds of data [9]. If we continue to use the research approach of the small data, there is bound to be a strong asymmetry between the data generation and knowledge extraction pipelines, resulting in the so-called "dark data" or "data iceberg" situation [10], i.e., the data stream cannot be parsed in a timely manner, thus losing most of the data value [11]. Therefore, in summary, the movement from small data to big data in WEM is positive and necessary.

I cannot entirely agree with the hypothesis-driven search for causality based on small data (i.e., the desperate thirst for why), nor the data-driven search for correlation founded on big data (i.e., abandoning the "why" and only pursuing the "what"). The author posits that within the realm of WEM research, it is advisable to integrate both approaches, initially focusing on identifying correlations and subsequently, through the incorporation of prior knowledge, gradually progressing towards uncovering causality.

1.2 Characteristics of big data in WEM

Embracing the power of big data has become critical to solving complex WEM challenges as the data generation rate in all water-related areas accelerates. Big data consists of a wide range of data sets that are characterized by volume, velocity, variety, veracity and value - (5Vs) - and requires scalable architectures for efficient storage, manipulation and analysis [12].

Volume: Managing Massive Water-Related Datasets

The volumetric attribute refers to the size of the data. A data volume is considered large if it is larger than the traditional on-premise IT infrastructure can handle in a reasonable amount of time. The volume of data is relative, and data can often be considered "large" if traditional

analytics are found to be helpless [13]. In addition to the usual spatial dimension, the volume of data also has a temporal dimension. This is particularly true for WEM, which often needs to store water environment observations over long periods of time (years to decades) and at high frequencies (seconds to hours). Thus, even though each individual scene data is not large in the spatial dimension, it represents a large volume of data in spatial and temporal dimensions.

Velocity: Rapid Data Processing for Timely Decision-Making

The velocity attribute refers not only to the speed of data generation but also to the speed of data analysis required for data ingestion. The speed of data generation, processing, and analysis speed plays a critical role in water environment management, especially in emergencies such as flooding or water pollution events. Processing and analyzing real or near real-time data is critical to making informed decisions and deploying appropriate actions. The speed of data generation is a direct result of increasing connectivity, widespread use of smart devices, and real-time monitoring of networks. High-speed data sets generated continuously from different sources form data streams that need to be processed in real or near-real time [14]. The generation of high-speed data streams, while having value-added information, also poses new challenges for data storage and processing algorithms, requiring high-throughput data stream servers with low latency (e.g., in-memory processing) and efficient online artificial intelligence algorithms that can identify, filter, analyze, and process data streams as they pass through [15].

Variety: Integrating Diverse Data Sources for Comprehensive Water Management

Variety, one of the five core dimensions of the 5V framework in Big Data, refers to the diversity of data types and formats generated by numerous sources. With the rapid development of digital technology, data diversity has become an essential aspect of big data analysis and management [16]. Due to the nature of observations, WEM generates a wide variety of data types and formats, and there are three main types of data based on their structure [17]. The first type is structured data. This type of data is highly organized, easily searchable, and can be processed and analyzed using traditional database systems. Structured data usually exists in relational databases or spreadsheets and consists of well-defined data types such as text, numbers, and dates. Examples include sensor readings and manual monitoring. The second type is semi-structured data. Semi-structured data is somewhere between structured and unstructured data and has some organizational elements but lacks strict structure. This type of data typically contains metadata or tags that provide context and facilitate analysis. Examples of this include XML files in satellite remote sensing, which contain data in a hierarchical format, but do not adhere to a fixed schema like structured data. The third type is unstructured data. Unstructured data does not have a specific format or organization, making it more challenging to process, store, and analyze. It forms a large part of the data generated today and includes various forms such as text files, images, videos, audio files, and social media posts. Integrating and analyzing these disparate data sets provides a holistic view of water resources and facilitates informed decision-making. Advanced analytics, such as natural language processing (NLP) and deep learning (DL), enable water resource managers to process and mine valuable content from a variety of data sources to facilitate integrated water resource management [18].

Veracity: Ensuring Data Quality and Accuracy in Water Management

Authenticity refers to the credibility, quality, and accuracy of the data. Data authenticity is most important in WEM because poor data quality can lead to poor decisions and can have serious consequences. Data authenticity includes accuracy, consistency, completeness, and timeliness [19]. Data accuracy refers to the correctness of the data collected and used for analysis. Inaccurate data can come from various sources, such as human error in the data entry process, measurement errors in sensors, or inconsistencies between different data sources. Data consistency refers to maintaining the same data values in all situations where data are stored and used. Inconsistencies occur when data are updated or changed in one place and not in another, leading to discrepancies and unreliable analysis. Data completeness means having all the information needed for accurate analysis and decision-making. Incomplete data sets can lead to biased or incomplete insights that undermine the effectiveness of data-driven decision-making. Timeliness refers to how current and relevant the data is at the time of analysis. Outdated data can result in decisions based on historical trends that are no longer applicable, negatively impacting the outcomes of data-driven initiatives [12]. A number of data management measures need to be taken, including governance, validation, cleansing, and auditing, to ensure the authenticity of data. In WEM, data cleansing [20] is the most frequently used and time-consuming. There are unpredictable factors such as complexity and variability in real natural scenarios, and the data collected cannot be guaranteed to be accurate (e.g., sensors need time to stabilize during sudden environmental changes), so identifying and correcting errors, inconsistencies, and inaccuracies in the data is a key pre-requisite step to placing it in the database.

Value: Extracting Actionable Insights for Sustainable Water Management

Value is the most central element of the 5V framework in Big Data and refers to the actionable insights and benefits that result from processing and analyzing large databases. The ways in which Big data can create value in WEM are diverse. Big data is used to collect and analyze data from a variety of sources, such as sensors, satellite imagery, and social media, to monitor and assess water quality. This allows organizations to identify trends, detect pollution events, and develop strategies to address water quality issues [21]. Big data analytics are used to optimize water allocations, predict demand, and assess the impact of climate change on water resources. These insights help organizations develop sustainable water management plans and improve water use efficiency [22]. Information on species distribution and abundance is monitored to help organizations develop effective conservation strategies, restore degraded habitats, and maintain ecosystem health. It is important to note that building a large database that conforms to the 4Vs previously described is the first step and provides material for subsequent value mining. In turn, the capability of value mining depends almost entirely on the development and use of mining tools. A value mining approach with an advanced kernel that fits the data modality, is easy to understand, and has outstanding performance can enable us to obtain more and higher value information on existing data sets to make informed decisions, optimize processes, and efficiently allocate resources to address water-related challenges.

In conclusion, Big data has tremendous potential to revolutionize water environment management. The use of large datasets and deep learning can provide insights into water

quality, consumption, distribution, and waste treatment that were previously impossible to achieve.

1.3 Big data framework in WEM

In recent WEM research, scholars such as Peters-Lidard [23] advocate for the adoption of burgeoning data science methodologies to synthesize, evaluate theories and models, and bolster data support for Earth system change mechanisms, thereby positioning data science as a novel paradigm. However, at present, numerous inquiries, ambiguities, and even skepticism surround the distinctions between emergent big data science and traditional small data science. Moreover, the implementation of nascent research modalities, such as Big Data, within the realm of WEM is primarily confined to singular tasks, sessions, and scenarios, failing to fully manifest the inherent benefits of Big Data [24]. To secure a competitive edge in an increasingly digitized and interconnected global landscape, the geoscience community necessitates a comprehensive understanding of the pertinent technologies underpinning contemporary data science and a paradigmatic research framework delineating the potential applications of these novel technologies within the context of WEM. Consequently, I propose a big data WEM research framework 1.1.

The cornerstone of this framework is the Environmental Big Data Platform (EBDP), which occupies the central position. To ensure the EBDP possesses the attributes of big data, various data collection methods are employed, such as remote sensing satellites, cruise monitoring, manual sampling, and so on. The process of extracting value from the EBDP centers on the current cutting-edge artificial intelligence technique, DL. DL enables computers to process data within the EBDP in a manner reminiscent of human cognitive processing, thereby facilitating more profound value extraction [25]. The structural differentiation of DL allows it to address diverse data processing tasks such as clustering, classification, and regression. Another fundamental aspect of this framework involves leveraging existing expertise to align WEM tasks with the appropriate DL categories and to select the optimal structures within the corresponding categories for the associated WEM tasks, such as water quality monitoring, water quality assessment, and algal bloom prediction. The aggregation of these tasks will form a comprehensive closed-loop chain encompassing WEM, with each WEM task representing a node. This closed-loop, comprising individual nodes, serves as the foundation for constructing a large-scale, multi-task, and multi-modal WEM model in the future.

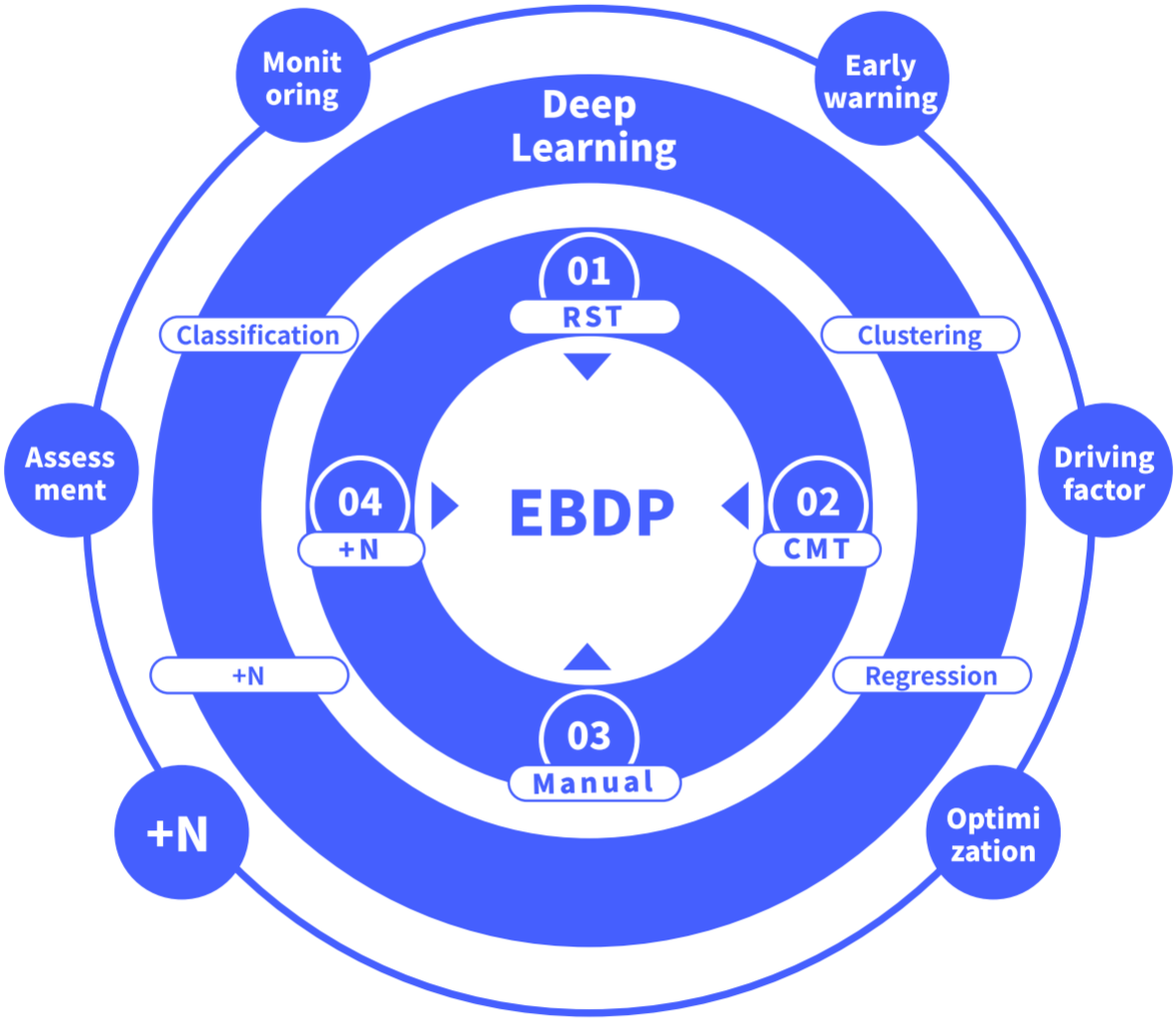


Figure 1.1: Big data framework in WEM

2

Research Objectives

The goal of this Ph.D. work is to improve water resource management through the development of a big data framework in WEM and to show examples of task node completion and their clustering effect through practical applications in various water bodies. Additionally, this research seeks to evaluate the benefits and transformative potential of big data as a new paradigm in WEM. This leads to the following objectives:

1) **Large-scale water quality monitoring and assessment in Qingcaosha Reservoir.** As one of the indispensable foundations of WEM and the node of the research framework, an economical, accurate, and practical water quality monitoring system has become essential for researchers, practitioners, and policymakers [26]. Traditional water quality monitoring methods are point-based, observing water quality for a given time series and placing a fixed number of stations at varying densities and dispersions (Section 4.1). This becomes challenging when researchers have limited resources available, such as employees, time, equipment, funding, and access to funding (Section 4.1). Cruise monitoring technology (CMT) is more effective in extracting environmentally relevant parameters on a large scale than point-based environmental monitoring. The data collected by CMT is line-based, which transforms the "point-inferred area" problem into a "line-inferred area" problem combined with space interpolation [27]. However, CMT combined with spatial interpolation still requires an extensive monitoring network to provide accurate estimates, which means there is still a heavy reliance on research resources. This reliance seems to be alleviated in a sense by the development of remote sensing technology (RST). The data collected by RST is area-based, and RST can directly scan the target area and combine it with manual in situ measurements, a traditional process, to build inverse models that present the state of the environment on a large scale (Section 4.3.1). However, the amount of data at the measured point is much smaller than the amount of RST data for the entire study area (usually differing by one to two orders of magnitude), resulting in instability and uncertainty in the inversion results. As a result, the overall environmental state of the target area cannot be presented clearly and accurately. After the appeal analysis, CMT and RST are complementary for environmental monitoring problems in large-scale

areas. Therefore, in the task node of water quality monitoring, EBDP consists of data from CMT and RST. Value mining means using the Deep Neural Network (DNN) because it can accurately find the complex nonlinear relationship between water quality parameters and RST observations (Section 4.3.3). As an important part of WEM, representative and reliable water quality assessment helps to identify potential problems, risks and areas for improvement to ensure the safety of water resources [28]. Traditional assessment methods such as the single-factor assessment method, water quality classification method, and integrated pollution index method are often difficult to make a meaningful water quality assessment of the large and complex matrix of water quality attributes [29]. Therefore, in this task node of water quality assessment, I use the deep clustering method, Improved Deep Embedding Clustering (IDEC), as a value mining tool, that is, similar elements are assigned to the same group, and different elements are assigned to different groups so as to find a quick solution to the pollution problem (Section 4.3.4). The EBDP of water quality assessment is composed of the results of the water quality monitoring task node. Qingcaosha reservoir, located in the middle of the Yangtze estuary, one of the world's largest tidal reservoirs and the new largest source of drinking water for about 12 million Shanghai residents (Section 4.2), was selected as the model study area for the water quality monitoring and assessment node.

2) **Identifying the algal growth driving factors in the South-to-North Water Diversion Project.** Water environment management is an important aspect of sustainable development because it plays a vital role in protecting natural resources, ensuring public health, and maintaining ecological balance (Section 1.1). Effective management strategies rely on a comprehensive understanding of the driving factors that affect water quality, availability, and ecosystem health (Section 5.1). Identifying these driving factors is paramount to designing targeted interventions and adaptive management plans to address pressing water-related challenges, so driving factor identification can be an important task node in the WEM big data framework. One of the most important issues for ecologists, algal growth and water quality relationships [30], was chosen by me as an example of the completion of this task node. The South-North Water Diversion Middle Route Project, a mega water project in China serving 69 million people, was selected as the implementation area for the driving factor identification node (Section 5.2). While traditional process-based ecodynamic models can systematically represent the relationship between a single output and multiple inputs, they often require extensive up-front experiments to identify variables as well as parameters and are difficult to calibrate because the physical transport and biological processes that regulate algal biomass dynamics are highly variable at a different residence and biological time scales [31]. With the development of data science, some traditional data-driven methods, such as regression analysis and multivariate analysis methods, have been proposed and applied (Section 5.1). However, due to their linear functional basis, these methods are ineffective in seeking such nonlinear relationships between environmental factors and algal biomass [30]. Therefore, machine learning has been used in recent years as a better tool for driver identification [32, 33, 34]. However, as environmental studies begin to migrate from small to large data, traditional machine learning is unable to respond effectively. To solve this problem, Bloomformer-1 (Section 5.3.2, 5.3.3), a DL model based on the Transformer core proposed by Google (Section 5.3.2), was developed as a value mining method and implemented in EBDP consisting of

high-frequency manual sampling and analysis data from the middle route of the south-north water diversion over four years.

3) **Building algal growth early warning system in Lake Taihu.** Algal growth is a significant environmental and public health concern worldwide. They occur when colonies of algae grow rapidly and produce toxic or harmful effects on aquatic life, ecosystems, and human health. Early warning systems for algal growth are essential for minimizing their impacts and facilitating timely and effective management responses. These systems involve a combination of monitoring and forecasting to detect, predict and guide stakeholders on how to prevent the coming algal growth. The occurrence and extinction of algal growth are rapid and random and fluctuate at different depths, so monitoring of algal growth based on different continuous depths and high temporal resolution is essential. Traditional monitoring tools such as manual sampling or fixed-point in situ water quality monitoring devices are difficult to do both due to their technical bottlenecks and research resource limitations. To solve this problem, a vertical aquatic monitoring system, named BIOLIFT, was applied and the generated monitoring data was used as EBDP for early warning task nodes. Chlorophyll-a future value prediction depends not only on its own previous values but also on the previous/present values of other water quality parameters. Traditional time series prediction (TSP) methods can be difficult when interpreting long background series and extending to complex variable relationships. Deep learning models overcome these challenges by utilizing large data sets. The popular Long Short-Term Memory (LSTM) DL models (Section 6.2.3) can only moderate the vanishing gradient and exploding gradient problems to some extent when dealing with long time series data; therefore, LSTM is only applicable to time series data of average length and its prediction becomes poor for longer time series data. In addition, the algal growth early warning system needs to be able to predict Chl-a accurately and also shows clearly and specifically the drivers of the predicted values so that preventive measures can be developed. The mechanism and structure of the LSTM dictate that it does not have this capability. In order to achieve high performance in processing long time series data with the function of displaying the drivers of predicted values, the DL model Bloomformer-2 (Section 6.2.4), based on the Transformer core proposed by Google (Section 5.3.2), has been developed and used as a value mining method for early warning task nodes. Taihu Laboratory for Lake Ecosystem Research, located on the southern side of Meiliang Bay (Section 6.1), was selected as the model study area for the algal growth early warning task node.

3 Projects

3.1 Qingcaosha reservoir project

The Qingcaosha reservoir project is a joint research project initiated by Karlsruhe Institute of Technology (KIT) and the University of Shanghai for Science and Technology (USST). The project aims to develop and adopt innovative sensor-based scientific and technical approaches for effective water quality monitoring to generate new knowledge about the system. The project will contribute to risk assessment and early warning concepts for water pollution and recommendations for improving water quality in the Qingcaosha Reservoir (Yangtze River). The duration of this project, originally planned for May 2017 to December 2020, has been extended to August 2021 due to the epidemic. The lead scientists from Germany and China are Prof. Dr. Stefan Norra and Prof. Dr. Hongbo Liu.

I am involved in the project from July 2019 to August 2021. My work mainly includes 1) applying a multi-sensor cruise monitoring system (BIOFISH) for water quality monitoring in Qingcaosha Reservoir; 2) collecting water samples from the reservoir and completing laboratory analysis; 3) collecting and digitizing remote sensing images of Qingcaosha Reservoir, and 4) performing high-value mining of EBDP composed of BIOFISH and remote sensing images to complete examples of water quality monitoring and assessment task nodes.

3.2 Middle route of the south-to-north water division project

The middle route of the south-to-north water division project is undertaken by the Center for Algal Biology and Application Research, Institute of Hydrobiology (IHB), Chinese Academy of Sciences, and aims to conduct in-depth research on the spatial and temporal distribution and growth factors of algae in the middle route of south-north water diversion and to contribute to the ecological management of the mega-water project. The leading scientist is Prof. Dr. Yonghong Bi. The duration of the project is from Jan 2017 to December 2025.

I am involved in one of the sub-project "Study on the mechanism of algal outbreak and its key driving factors in long-distance water transmission main canals" from June 2021 to February 2023. My work mainly includes 1) field sampling and expedition to the South-North Water Transfer Central Line, 2) collecting water samples from each monitoring station and completing laboratory analysis, and 3) high-value mining of EBDP consisting of 4 years of water sample analysis data and completing example of driving factor identification task node.

3.3 SIGN project

The Sino-German Water Supply Network (SIGN) is a collaborative initiative between China and Germany aimed at addressing water management challenges through the exchange of knowledge, technology, and best practices in the water supply sector. The partnership focuses on enhancing the quality, efficiency, and sustainability of water supply systems in both countries, leveraging their respective strengths and experiences. The SIGN research project consists of phases I and II, running from 2015 to 2021, and its research is funded bilaterally by the Chinese Ministry of Science and Technology (MOST) and the German Federal Ministry of Education and Research (BMBF).

The subproject DYNAQUA in SIGN I and the subproject AMORIS in SIGN II were designed and supervised by Prof. Dr. Stefan Norra at the Institute of Applied Geoscience (AGW), Karlsruhe Institute of Technology (KIT). The main development objective of both projects is to establish an algal early warning system and assess the ecological risk of sediment resuspension events by developing an in situ online multi-sensor device for stationary monitoring of water quality and meteorology.

I was involved in the project from January 2019 to August 2021. My work mainly includes 1) applying the vertical water quality monitor (BIOLIFT) to conduct fieldwork on Taihu Lake in different seasons and times; 2) collecting water samples and completing laboratory analysis on Taihu Lake; 3) using BIOLIFT and laboratory data to establish EBDP and complete algal growth early warning system development.

4

Water Quality Monitoring and Assessment Task Node

4.1 Published paper

Title: Water quality monitoring and assessment based on cruise monitoring, remote sensing, and deep learning: A case study of Qingcaosha Reservoir

Authors: Jing Qian, Hongbo Liu, Li Qian, Jonas Bauer, Xiaobai Xue, Gongliang Yu, Qiang He, Qi Zhou, Yonghong Bi and Stefan Norra

Journal: Frontiers in Environmental Science, Volume 10-2022, doi: 10.3389/fenvs.2022.979133

Authorship statement: This peer-reviewed scientific journal article is based on data obtained from the January 2020 field trip as well as remote sensing data from that day. I designed, implemented, and was responsible for the entire research process. Due to the confidential nature of the Qingcaosha reservoir, I started communication and preparation with my Chinese colleagues at USST starting in July 2019. the only, one-day field trip in January 2020 was conducted under the authority of Prof. Hongbo Liu. During the field trip, Jonas Bauer and I worked together to set up the BIOFISH instrument and complete the field trip activities. The pre-processing of the remote sensing data was done entirely by me alone. I completed the deep learning model building with the help of Li Qian (Ludwig Maximilian University of Munich) and discussed the results with Jonas Bauer (KIT) and Xiaobai Xue (MioTech Research) for productization and industrialization. Finally, Prof. Dr. Stefan Norra (KIT), Prof. Dr. Qiang He (Chongqing University), Dr. Yugonglaing (IHB), and Prof. Dr. Qi Zhou (Tongji University) were involved in the revision of this paper. Prof. Dr. Stefan Norra, Prof. Dr. Hongbo Liu (USST), and Prof. Dr. Yonghong Bi (IHB) supervised this project, and Prof. Dr. Stefan Norra provided funding. All co-authors critically reviewed the manuscript and gave their consent for publication.

Abstract: Accurate monitoring and assessment of the environmental state, as a prerequisite for improved action, is valuable and necessary because of the growing number of environmental problems that have harmful effects on natural systems and human society. This study developed an integrated novel framework containing three modules: remote sensing technology (RST), cruise monitoring technology (CMT), and deep learning, to achieve a robust performance for environmental monitoring and the subsequent assessment. The deep neural network (DNN), a type of deep learning, can adapt and take advantage of the big data platform effectively provided by RST and CMT to obtain more accurate and improved monitoring results. It was proved by our case study in the Qingcaosha reservoir that DNN showed a more robust performance (R^2=0.89 for pH, R^2=0.77 for DO, R^2=0.86 for conductivity, and R^2=0.95 for back-scattered particles) compared to the traditional machine learning including multiple linear regression, support vector regression and random forest regression. Based on the monitoring results, the water quality assessment of Qingcaosha reservoir was achieved by applying a deep learning algorithm called improved deep embedding clustering. Deep clustering analysis enables scientific delineation of joint control regions and determines the characteristic factors of each area. This study presents a high value of the framework with a core of big data mining for environmental monitoring and follow-up assessment in a manner of high-frequency, multi-dimensionality, and deep-hierarchy.

4.2 Study area

Located approximately 30 kilometers northeast of downtown Shanghai as Figure 4.1, the Qingcaosha reservoir was built as part of a long-term strategy to address the growing demand for fresh water in a rapidly expanding urban landscape. The reservoir's primary function is to provide a stable and secure water supply, and its development has contributed to a significant improvement in the quality of life for 12 million Shanghai residents, ensuring that access to clean water is no longer a luxury, but a fundamental right. Completed in 2010, this massive reservoir covers approximately 70 km^2 and has become synonymous with sustainable development and efficient resource management in the region [35]. One of the world's largest estuarine and tidal reservoirs, the Qingcaosha reservoir is strategically located at the mouth of the Yangtze River (31.42-31.49N, 121.55-121.71E) and takes full advantage of the abundant resources provided by Asia's longest river. With a storage capacity of 435 million cubic meters, the reservoir is designed to withstand natural disasters and seasonal fluctuations in water levels, ensuring a steady supply of fresh urban water [36].

4.3 Framework of water quality monitoring and assessment

In the task node of water quality monitoring, EBDP consists of data from CMT and RST. Deep neural network is used for value mining means (Figure 4.2. After that, the results of the water quality monitoring node consist of new EBDP for deep clustering method, IDEC, for further value mining to complete the task node of water quality assessment (Figure 4.3).

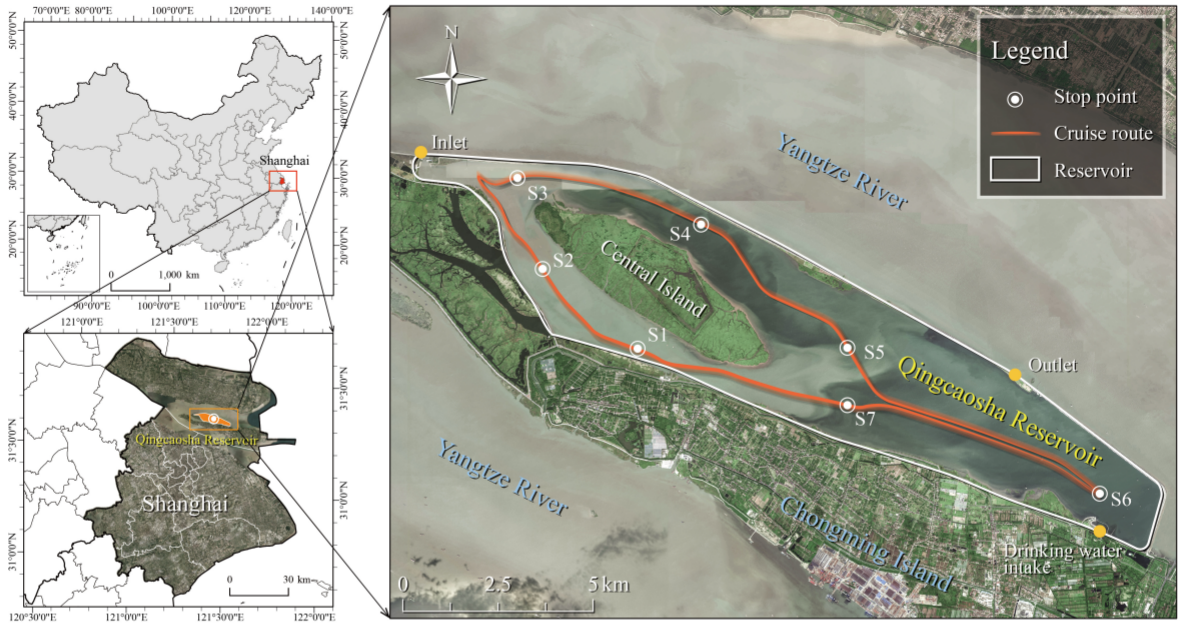


Figure 4.1: Study area and cruise route

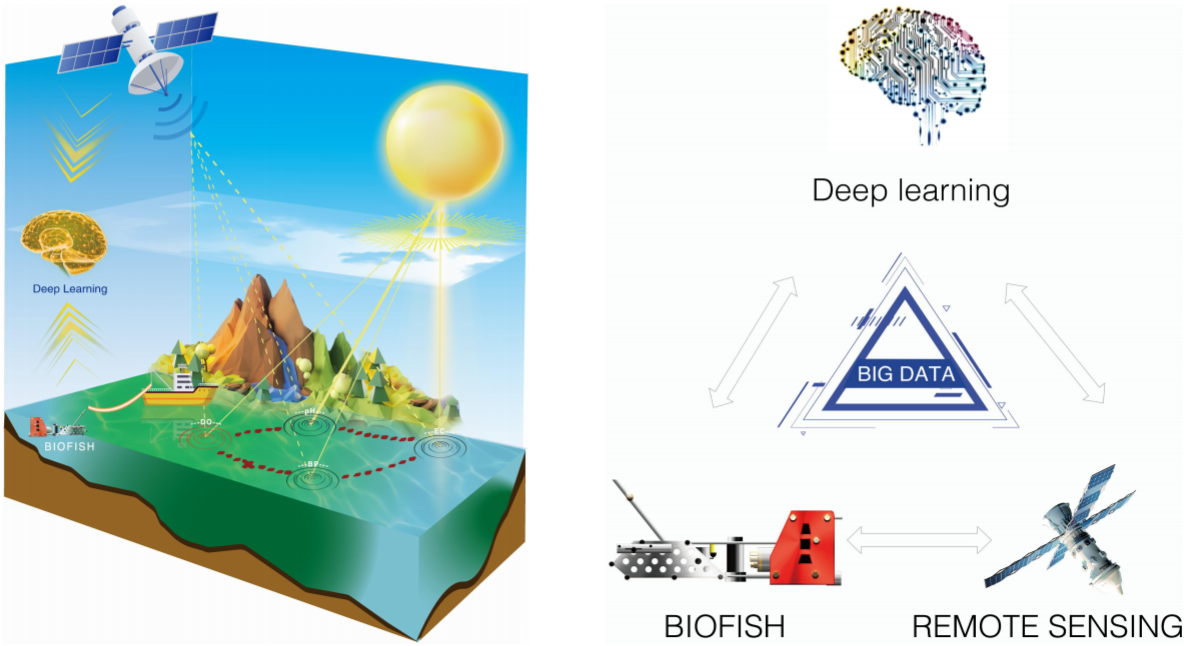


Figure 4.2: Schematic diagram of monitoring task nodes

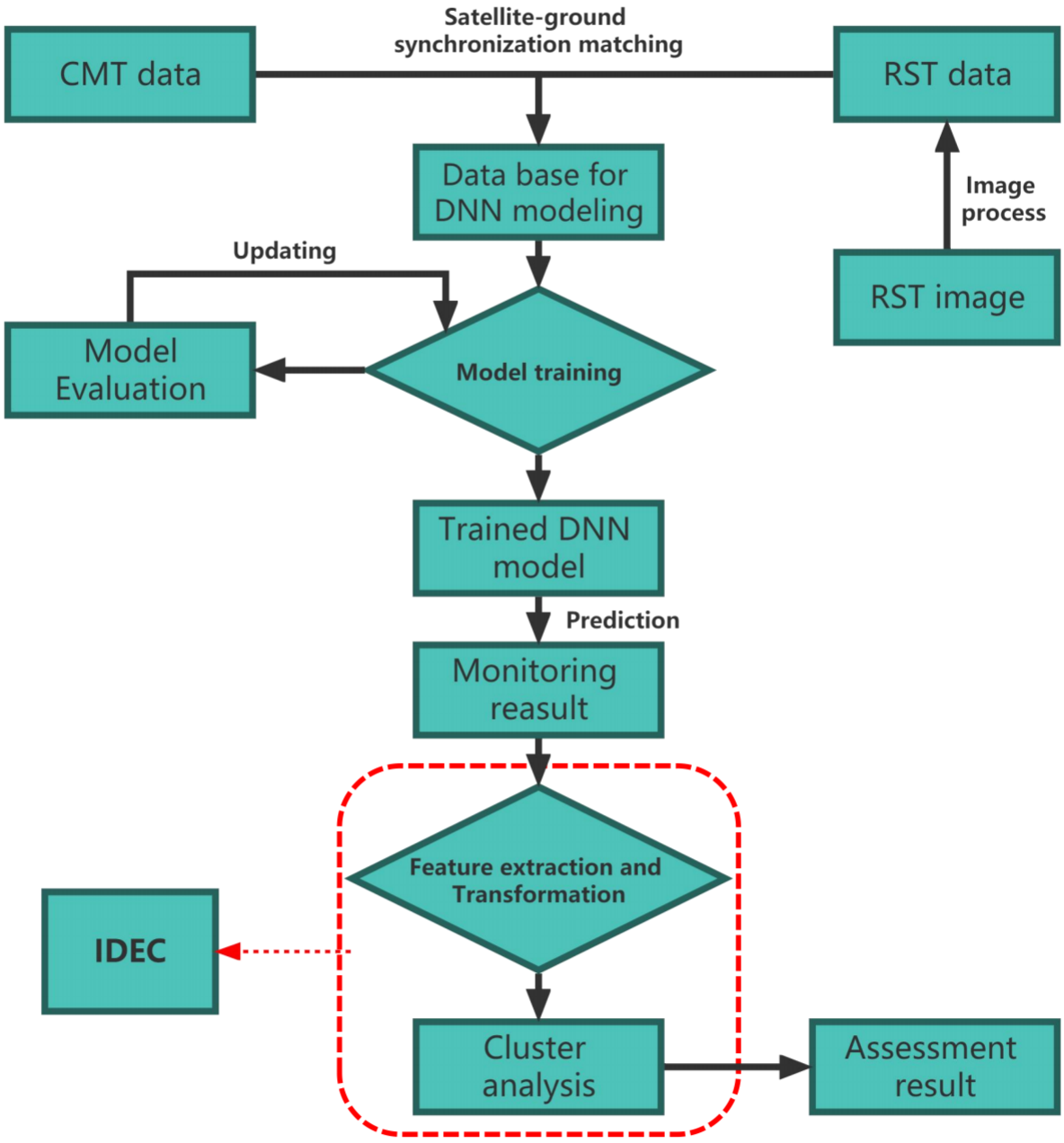


Figure 4.3: Flow chart of the monitoring and assessment task nodes

4.4 Remote sensing

The Sentinel-2 Earth observation satellite was selected as the observation satellite for this task node. Its multispectral instrument acquires 13 spectral bands from $440nm$ to $2200nm$ with spatial resolutions containing $10m$, $20m$ and $60m$ [37]. The information of Sentinel-2 is shown in Table 4.1. Sentinel 2 images were downloaded from the official website of the U.S. Geological Survey (https://earthexplorer.usgs.gov/). The Level-1C data product was selected for this task node, and this series has been radiometrically and geometrically corrected (including orthorectification). The downloaded completed Sentinel-2 remote sensing images underwent a series of pre-processing from radiometric correction, atmospheric correction, RST image fusion and study area clipping to complete the digitization from the images. In this task node, Fast Line of Sight Atmospheric Analysis of Hypercubes (FLAASH) [38] was set as the atmospheric correction algorithm with parameter settings including ground elevation, atmospheric model, aerosol retrieval and water retrieval obtained by reading the source file from the FLAASH assistant plug-in. The Gram-Schmidt Pan Sharpening method [39] was selected as the image fusion algorithm to bring the spatial resolution of all bands resampled to 10m. The remote sensing data were Z-score normalized before being input to the model according to the following equation [40]:

$$Z_i = \frac{x_i - \overline{x_i}}{\sigma_i} \tag{4.1}$$

where Z_i is the standard score of i-th data, x_i is the i-th original data, $\overline{x_i}$ is the mean of i-th data, and σ_i is the standard deviation of i-th data.

Table 4.1: Bands of Sentinel-2 and their specifications

Bands	**Specification**	**Central Wavelength** (μm)	**Resolution** (m)
Band 1	Coastal aerosol	0.443	60
Band 2	Blue	0.490	10
Band 3	Green	0.560	10
Band 4	Red	0.665	10
Band 5	Vegetation Red Edge	0.705	20
Band 6	Vegetation Red Edge	0.740	20
Band 7	Vegetation Red Edge	0.783	20
Band 8	NIR	0.842	10
Band 8A	Vegetation Red Edge	0.865	20
Band 9	Water vapor	0.945	60
Band 10	SWIR - Cirrus	1.375	60
Band 11	SWIR	1.610	20
Band 12	SWIR	2.190	20

4.5 Cruise monitoring

In this task node, multi-sensor cruise monitoring is performed by BIOFISH. It is an aquatic cruise monitoring system equipped with multiple sensors and connected to the ship by a data

Table 4.2: Sensors of BIOFISH and their specifications

Parameter	Principle	Range	Resolution	Accuracy
Pressure	piezo-resistive	0-100dBar	0.01dBar	±0.1dBar
Temperature	Pt 100	0-36°C	0.001°C	±0.01°C
pH	Potentiometric (Ag/AgCl)	0-12pH	0.01pH	±0.02pH
DO	Potentiometric (Clark electrode)	0-100%	0.01%	±0.01%
El.cond	7-pole-cell	0-60mS/cm	1uS/cm	±10uS/cm
Backscattered particles	Mie backscattering	0-100%	0.01%	±0.01%

transmission cable. Data on water quality parameters were recorded in real-time with GPS latitude and longitude positions. In this task node, BIOFISH swam between 10 cm and 20 cm below the water surface, so a swimming mode with a floating body was chosen instead of a diving mode with fins. The monitoring parameters were customized by the demand side. The study area Qingcaosha reservoir is one of the largest tidal reservoirs in the world. Of particular interest to tidal reservoirs is the introduction of freshwater and the backflow of seawater. The Yangtze River is a source of introduced freshwater and generally has a high turbidity level. And the specific value and diffusion of this turbidity is an important factor affecting the residence time. Electrical conductivity (El.cond.) and pH, as parameters reflecting the backflow of seawater, are undoubtedly also of interest. In particular, the subsequent diffusion of these two parameters and the degree of their influence can be of great help in adjusting the response of the reservoir to seawater inversion. Finally, the concentration of dissolved oxygen (DO) reflects the self-purification capacity of the reservoir. The level of self-purification capacity is closely related to the reservoir's ability to supply water. Therefore, back-scattered particles (BP), El.cond., pH and DO were selected as the monitoring parameters in this task node. The sensors of BIOFISH are made by ADM Elektronik and their specifications are shown in Table 4.2

Due to the limitation of authority, the sampling of BIOFISH needs to be completed within 5 hours. The cruise route is shown in Figure 4.1. and was designed to cover as much of the study area as possible. S1 is the start and end point of the cruise route. S1-S7 are the seven stops designed for BIOFISH calibration with YSL ProDSS to ensure data accuracy. BIOFISH data were Z-score (Section 4.3.1) normalized before being input to the model. Since the high sampling density of BIFOISH means that multiple BIOFISH sampling points can be found randomly in a pixel block of size 10m x 10m, a satellite-ground synchronization matching process is required for the BIOFISH data, i.e., to fix the BIOFISH sampling points in the same pixel block and derive their representative values. The first step is to specify the spatial information of all BIOFISH sampling points and pixel block centroids. The geodesic distances [41] between the pixel grid centroids and the BIOFISH sampling points can be calculated by the Python package Geopy, which is modeled as an ellipsoid, i.e., WGS-84. Then, by finding the shortest geodesic distance between them, the pixel grid corresponding to the BIOFISH sampling points can be extracted. The next step is to calculate the representative values of BIOFISH measurements within each pixel block by arithmetic mean (AM).

4.6 Deep neural network

DNNs are the basic form of deep learning. As the left side of Figure 4.4 shows, DNN is a connectionist system with multiple hidden layers between the input and output layers [42]. Each hidden layer contains multiple neurons, called nodes. Any nodes in lth layer must be connected to any node in $l+1$th layer, and the following equation indicates the non-linear relationship between the DNN layers shown in the right side of Figure 4.4:

$$\mathrm{a}_j^{l+1} = f\left(\sum_{i=1}^{n} a_i^l w_{ij}^l + b_j^l\right) \tag{4.2}$$

Where a_i^l is the activation value of ith node in lth layer, a_j^{l+1} is the activation value of jth node in $l+1$th layer,$w_{ij}^{l+1}$is the weight between a_i^l and a_j^{l+1}, b_j^{l+1} is the bias value of jth node in $l+1$th layer, and $f(\cdot)$ is the active function.

The process of training is depicted on the left side of Figure 4.4. Calculating and storing intermediate variables (including outputs) from the input layer to the output layer is forward propagation. Back propagation is the process of computing the gradient of neural network parameters and updating them based on the difference between the output and the actual value. In this task node, *relu* [43] was set as the active function and *adam* as the optimizer of all models for adjusting hyperparameters. Apart for the El.cond.-spectral value, the layer and neural units of models were (256,256,256,256,256) (256,256,256). In addition, batch size and learning rate were optimized within a suitable range.

Several indicators including the coefficient of determination (R^2), Root-mean-square error (RMSE), Mean absolute percentage error (MAPE) and Median absolute deviation (MAD) were used to evaluate the performace of models. The Units of RMSE and MAD are the same as the respective water quality parameter units, and the unit of MAPE is %.

The coefficient of determination [44](R^2) was calculated as:

$$R^2\left(y, \hat{y}\right) = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y_i}\right)^2} \tag{4.3}$$

RMSE [45] was calculated as:

$$RMSE(y, \hat{y}) = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2}{n}} \tag{4.4}$$

MAPE [46] was calculated as:

$$MAPE\left(y, \hat{y}\right) = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y_i}}{y_i}\right| \tag{4.5}$$

MAD [47] was calculated as:

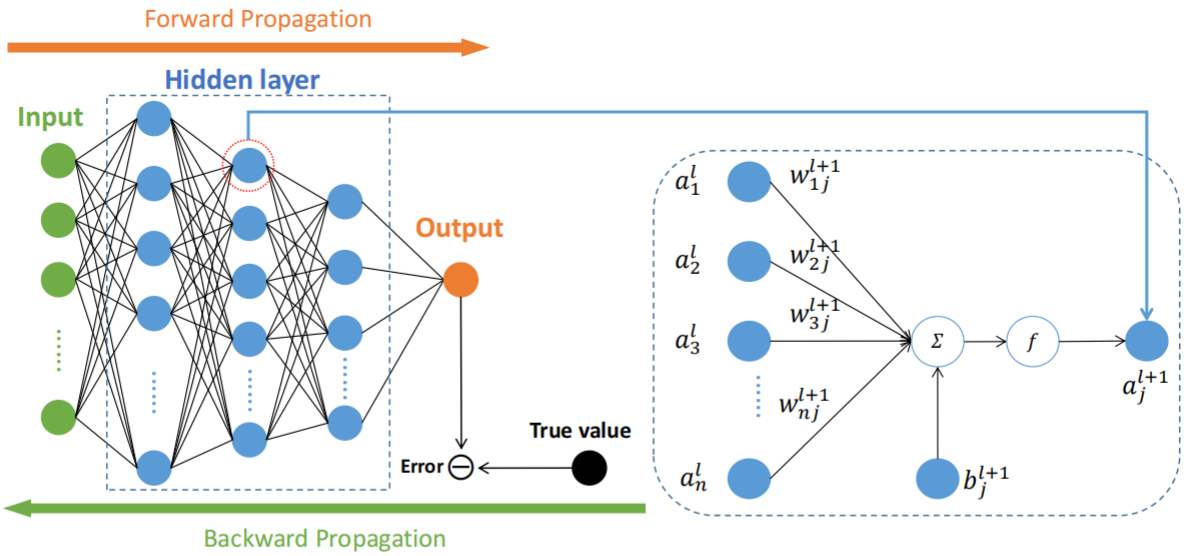


Figure 4.4: Architecture of DNN

$$MAD\,(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y_i}| \tag{4.6}$$

where $\widehat{y_i}$ is the predicted value of the ith sample, y_i is the corresponding true value of the total n samples, and $\overline{y_i}$ is the mean of true value.

4.7 Improved Deep Embedding Clustering

IDEC is an advanced clustering algorithm that combines the advantages of deep learning and unsupervised clustering techniques to improve the performance of traditional clustering methods. It is an extension of the original Deep Embedding Clustering (DEC) algorithm, which uses an autoencoder to learn meaningful feature representations from high-dimensional data and subsequently clusters the learned features [48]. The IDEC algorithm incorporates several improvements to address certain limitations and improve the clustering performance of the DEC method. The main components of the IDEC algorithm include (Figure 4.5):

1) Autoencoder and decoder: This is a deep learning neural network that learns to compress the input data to a lower dimensional representation and then reconstructs the original data from the compressed representation. Autoencoders are trained to minimize the reconstruction error between the input data and the reconstructed output. In the context of IDEC, the autoencoder is used to extract meaningful features from high-dimensional data.

2) Clustering layer: The clustering layer is added to the encoder part of the autoencoder, resulting in a new model that maps the input data directly to the clustering task. The clustering layer uses a Gaussian Mixture Model or another clustering algorithm to assign each data point to a cluster centroid based on the learned feature representation.

3) Joint optimization: The IDEC algorithm involves a joint optimization process that simultaneously refines feature representation learning and cluster assignment. The optimization objective combines the reconstruction error of the autoencoder and the clustering loss term.

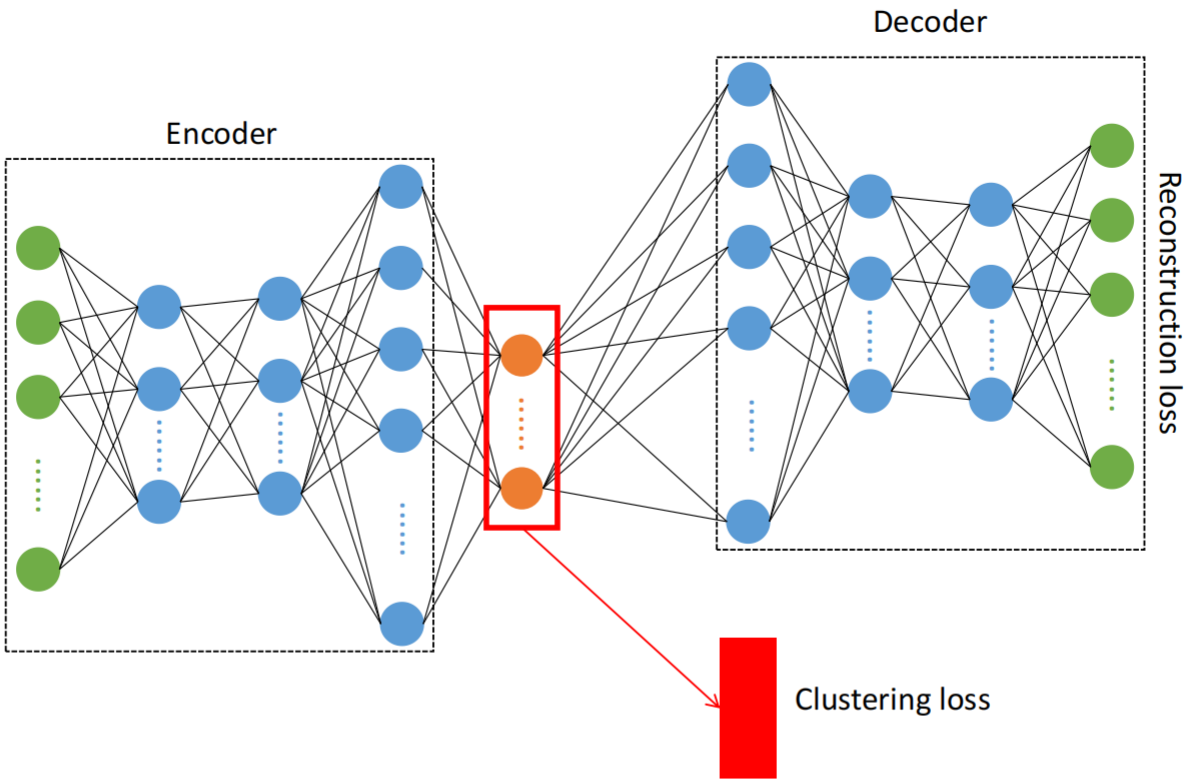


Figure 4.5: Architecture of IDEC [48]

The clustering loss term is usually based on Kullback-Leibler divergence [49]. It is defined as KL divergence between distributions P and Q, where Q is the distribution of soft labels measured by Student's distribution and P is the target distribution derived from Q.

4) Self-supervised fine-tuning: In the joint optimization process, the model adjusts its weights to minimize the overall loss, taking into account the reconstruction error and clustering loss. This self-supervised fine-tuning process helps to improve the quality of the learned feature representations and thus the clustering performance.

In this task node, the encoder network is set as a fully connected multilayer perceptron (MLP) with dimensions 4-125-125-500-10. the decoder network is the mirror image of the encoder with dimensions 10-500-125-125-4. *relu* is set as the active function and *adam* is the optimizer for all models. The coefficient of cluster loss *gamma* is set to 0.1 and the batch size is 256. the convergence threshold *delta* is set to 0.1%. The update interval T was 1 iteration. the IDCE and CH methods were performed by PyTorch.

The number of clusters was determined by the Carinski-harabasz (CH) method [50]. The Calinski-Harabasz (CH) method, also known as the Calinski-Harabasz index, is a widely used evaluation metric for determining the optimal number of clusters in a dataset. It is an internal cluster validation method that measures the ratio of the between-cluster variance to the within-cluster variance. The higher the CH index, the better the clustering performance, as it indicates that the clusters are more compact and well-separated.

The CH index is calculated using the following formula:

$$SSW = \sum_{i=1}^{N} \|x_i - C_{pi}\|^2 \tag{4.7}$$

$$SSB = \sum_{i=1}^{k} n_i \left\| c_i - \bar{X} \right\|^2 \tag{4.8}$$

$$CH = \frac{SSB/(k-1)}{SSW/(N-k)} \tag{4.9}$$

Where k is the number of clusters, P is the partitions, $X = \{x_1, x_2 ... x_N\}$ represents the data set with N-dimensional points, $\overline{X} = \sum_{i=1}^{N} x_i/N$ is the center of the entire data set, $C = \{c_1, c_2 ... c_k\}$ represents the centroids of cluster, c_i is the ith cluster.

4.8 Computational environment

The experiment was carried out on a PC with the following features: Hardware: CPU i7-6950X, RAM 64GB, dual GeForce RTX 3090, VRAM 24GB; Software: Ubuntu 20.04, Python3.6, Pytorch 1.10.0, Numpy 19.2.

4.9 Results

4.9.1 Model performance evaluation

DNN demonstrate superior performance in terms of accuracy and stability when compared to traditional machine learning methods such as Multiple Linear Regression (MLR), Support Vector Regression (SVR), and Random Forest Regression (RFR). A summary of the model performance metrics can be found in Table 4.3. Figure 4.6 presents a comparison between the predicted and measured values for the test set. Notably, the slopes of the DNN test results (0.90 for pH, 0.67 for DO, 0.82 for El.cond, and 0.92 for BP) are considerably larger than those of MLR, SVR, and RFR. Consequently, the DNN model significantly enhances inversion accuracy in comparison to MLR, SVR, and RFR methods.

4.9.2 Result of water quality monitoring and assessment in Qingcaosha reservoir

The illustrative thermal cartography, showcasing concentration gradients of individual factors such as pH, DO, El.cond., and BP, can be located within Figure 4.7, meticulously arranged in distinct subsections. Through the application of the CH technique, the vast expanse of the Qingcaosha reservoir bifurcates into four unique aquatic categories: I, II, III, and IV. The spatial distribution of these clusters, as demarcated by the IDEC approach, is delineated in Figure 4.8. An encompassing collation of the statistical evaluation, encapsulating extremities

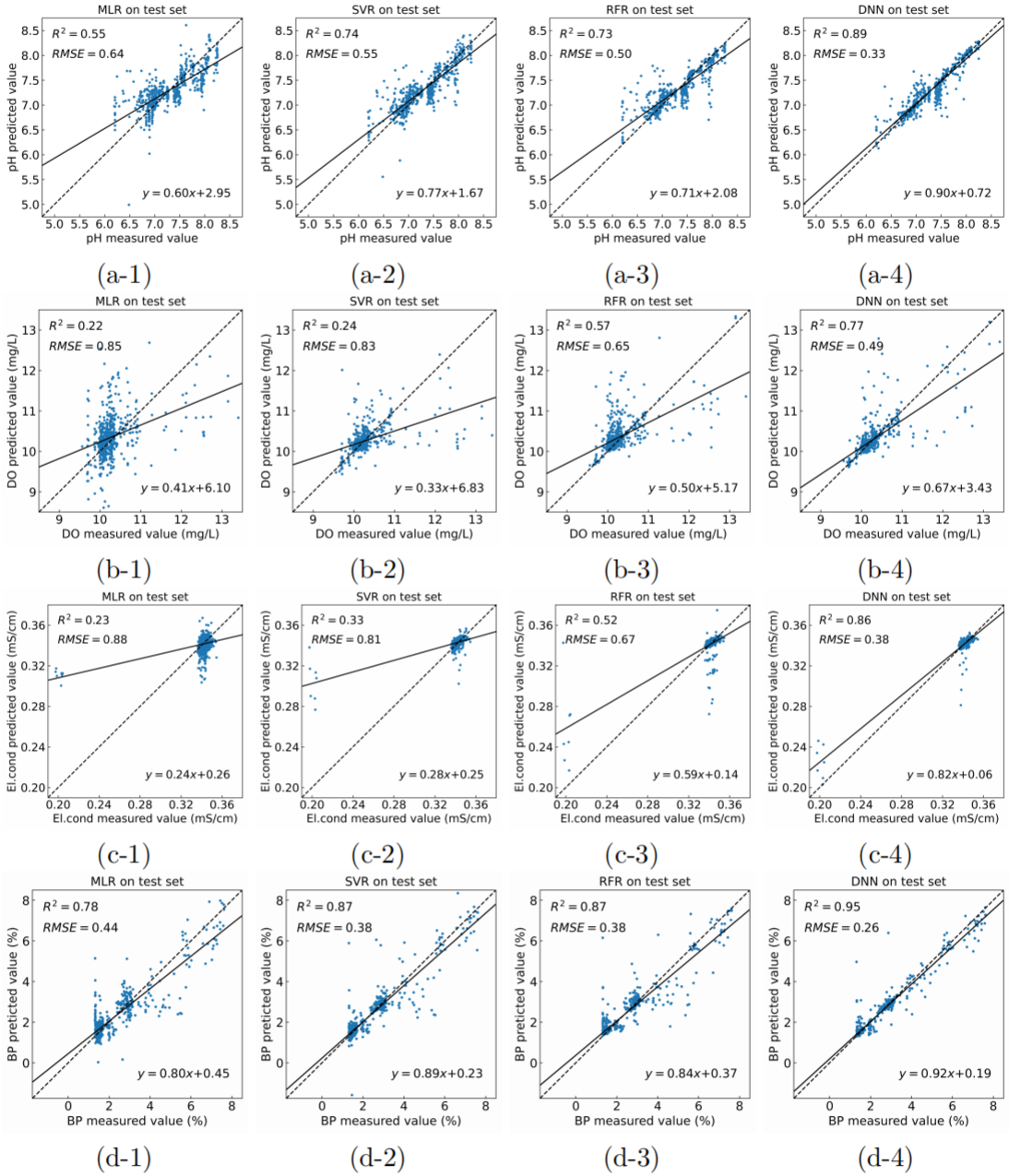


Figure 4.6: Regression model performance evaluation by comparison of the predicted data and measured data on test set, where (a), (b), (c), (d) represent the test results of the pH, DO, El.cond and BP, respectively, and (1), (2), (3), (4) represent the test results of the MLR, SVR, RFR and DNN, respectively.

Table 4.3: Results of model evaluation

Parameter	Model	R^2	RMSE	MAPE	MAD
pH	MLR	0.55	0.64	0.86	0.41
	SVR	0.74	0.55	0.69	0.21
	RFR	0.73	0.50	0.57	0.17
	DNN	**0.89**	**0.33**	**0.52**	**0.10**
DO	MLR	0.22	0.85	2.83	0.30
	SVR	0.24	0.83	**1.30**	0.12
	RFR	0.57	0.65	1.59	0.14
	DNN	**0.77**	**0.49**	1.61	**0.06**
El.cond	MLR	0.23	0.88	9.62	0.20
	SVR	0.33	0.81	**1.54**	0.08
	RFR	0.52	0.67	1.78	0.10
	DNN	**0.86**	**0.38**	1.74	**0.06**
BP	MLR	0.78	0.44	3.07	0.14
	SVR	0.87	0.38	3.34	0.07
	RFR	0.87	0.38	**2.72**	0.06
	DNN	**0.95**	**0.26**	3.10	**0.03**

(Max, Min), dispersion measure (standard deviation - STD), and central tendency (Median), is presented in Table 4.4.

Group I, encapsulating a commanding 73.79% of the entire Qingcaosha reservoir's expanse, displays median values for each constituent element relative to its counterparts. This collective finds its predominance in the northeastern periphery of the central isle, extending its influence to the terminating zone. Group II and III, although minor, claim ownership of 4.99% and 3.23% of the water proportions respectively. Group II, distinguished by significantly amplified BP indices in comparison to the other cohorts, is primarily nestled at the reservoir's apex and along the southern flank. Contrarily, Group III, marked by superior dissolved oxygen concentrations vis-a-vis its peers, situates itself in the vicinity of the potable water outlet. Concluding with Group IV, enveloping 17.99%, it is characterized by heightened pH metrics, insinuating a subtly alkaline aquatic body. This cluster's presence is predominantly felt in the southwestern precincts of the central island, with minor footprints at the tail area and proximate to the drinking water inlet.

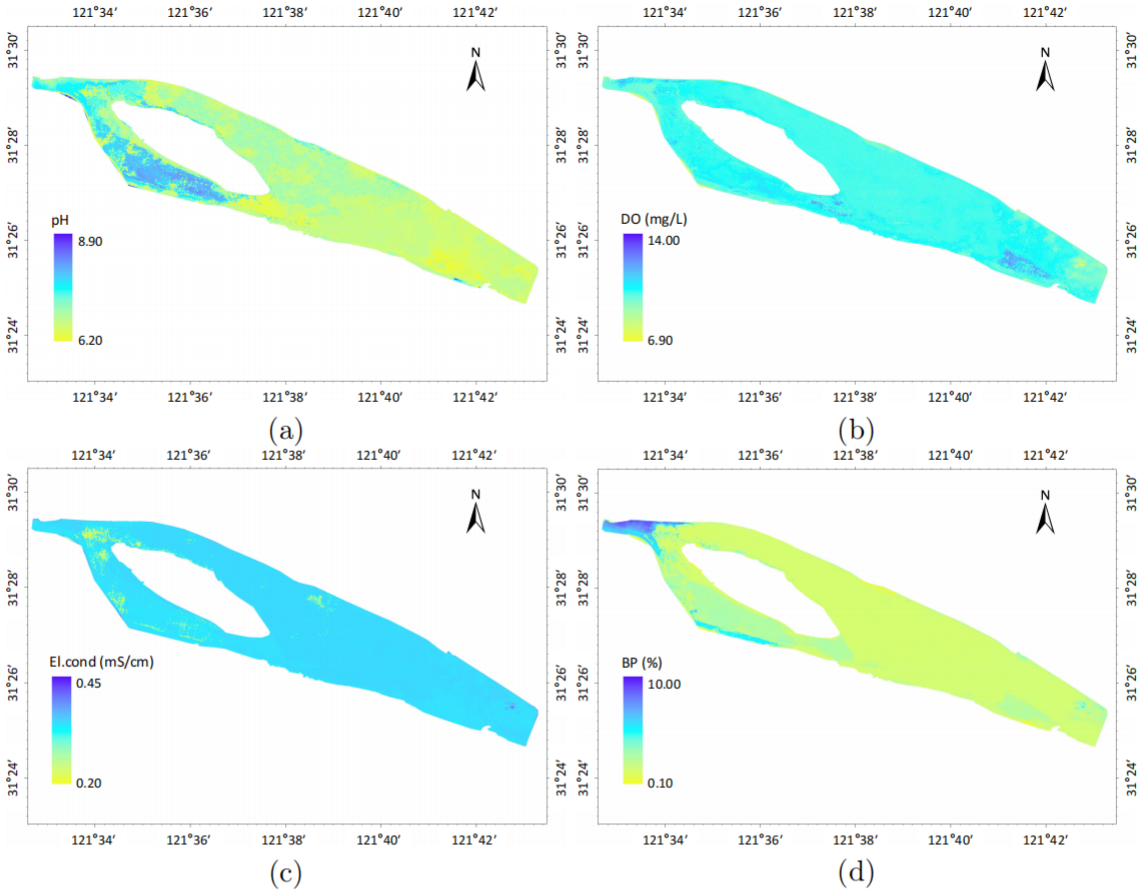


Figure 4.7: Distribution of (a)pH, (b)DO, (c)El.cond, (d)BP in Qingcaosha reservoir based on the framework

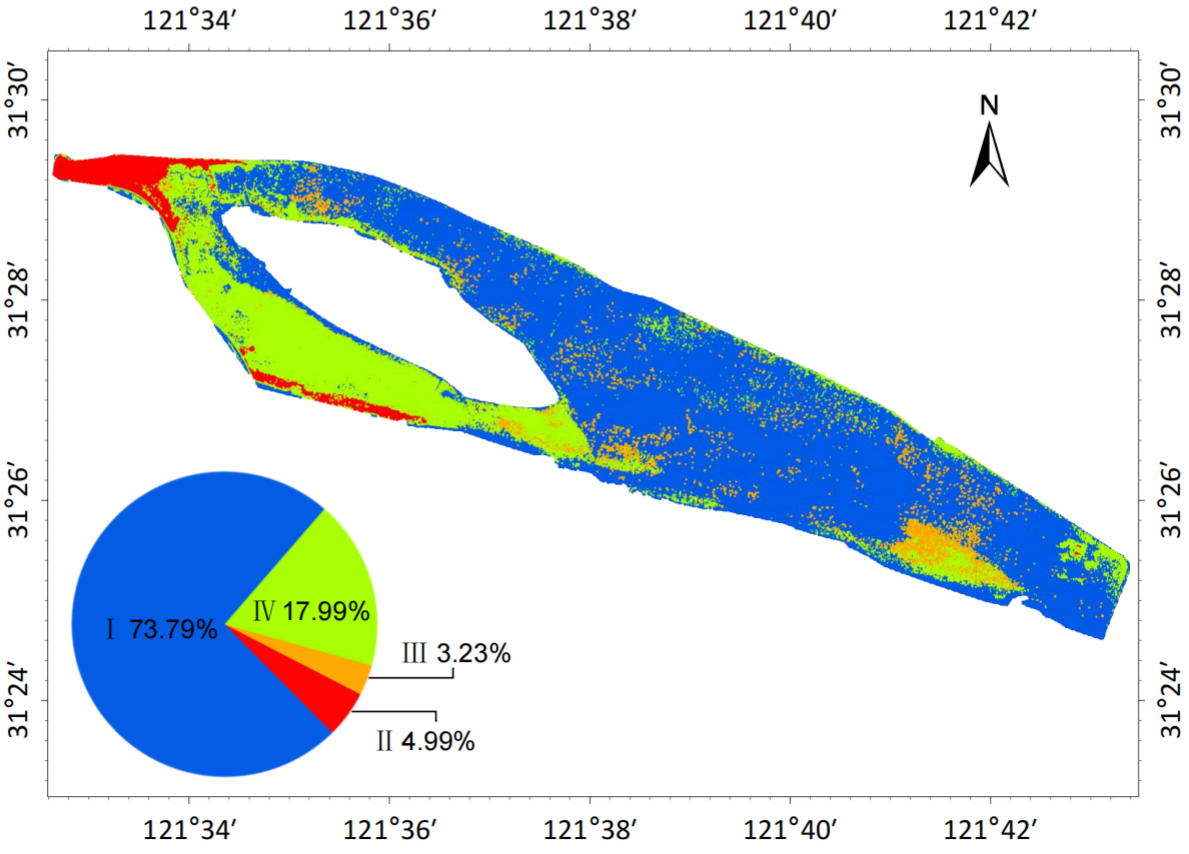


Figure 4.8: Clustering result of Qingcaosha reservoir

Table 4.4: Summary statistics of each group

Group	Parameter	Max	Min	STD	Median
Group I	pH	9.00	6.20	0.25	7.04
	DO	14.00	6.90	0.32	10.13
	El.cond	0.44	0.20	0.01	0.34
	BP	3.02	0.01	0.23	1.57
Group II	pH	8.25	6.50	0.27	7.64
	DO	13.94	6.90	0.64	9.94
	El.cond	0.42	0.20	0.02	0.34
	BP	10.08	4.58	1.26	6.11
Group III	pH	8.30	6.20	0.223	6.95
	DO	14.00	9.15	0.72	12.11
	El.cond	0.35	0.20	0.02	0.34
	BP	7.47	1.23	0.54	2.60
Group IV	pH	9.00	6.20	0.51	7.86
	DO	14.00	6.90	0.45	10.26
	El.cond	0.44	0.20	0.02	0.34
	BP	4.83	1.94	0.52	2.93

4.10 Conclusion of water quality monitoring and assessment task nodes

A groundbreaking framework has been erected, comprising three modules: RST, CMT, and DL. The latter capitalizes on the extensive data platform constructed by RST and CMT, to procure a robust efficacy in the monitoring and assessment of water quality. Examinations divulged that the DNN embedded within the framework manifests a superior proficiency in the monitoring of four water quality parameters (pH, DO, El.cond, and BP) when juxtaposed against MLR, SVR, and RFR. The deployment of IDEC for the appraisal of water quality demonstrated that the entirety of the QCSR was aptly segregated into four clusters, denominated as joint control areas, namely, Group I, Group II, Group III, and Group IV. The distinctive factors corresponding to each area were discerned, offering substantial contributions to the delineation of a unified regional control strategy for the QCSR.

5

Identification of Algal Growth Driving Factor Task Node

5.1 Published paper

Title: Identification of algal growth driving factors in the South-to-North Water Diversion Project by Transformer-based deep learning

Authors: Jing Qian, Nan Pu, Li Qian, Xiaobai Xue, Yonghong Bi and Stefan Norra

Journal: Water Biology and Security, https://doi.org/10.1016/j.watbs.2023.100184

Authorship statement: This peer-reviewed scientific journal article presents findings based on data collected during a multi-site, high-frequency field study conducted between August 1, 2018, and August 1, 2022. I was responsible for designing and overseeing the entire research process. My involvement included participating in eight field trips spanning the summer, fall, and winter of 2021, as well as the spring, summer, fall, and winter of 2022, and the spring of 2023. Owing to the confidential nature of the water quality data associated with the middle route of the South-to-North Water Division Project, standard sampling and storage procedures were carried out by maintenance staff at each monitoring site. Subsequently, the samples were sent to the Wuhan Institute of Aquatic Biology for analysis, for which I was responsible in the laboratory setting.

I collaborated with Nan Pu (Leiden University) and Li Qian (Ludwig Maximilian University of Munich) to develop the deep learning model, and together, I discussed the potential for industrial transformation of our research findings with Xiaobai Xue (MioTech Research). The project was supervised by Prof. Dr. Stefan Norra (KIT and Potsdam University) and Prof. Dr. Yonghong Bi (IHB), with funding provided by Prof. Dr. Yonghong Bi (IHB). All co-authors critically reviewed the manuscript and approved its publication.

Abstract: Accurate and credible identification of the algal growth drivers is essential for freshwater's sustainable utilization and scientific management. In this study, I developed a deep learning-based Transformer model, named Bloomformer-1, for end-to-end identification of algal growth drivers without extensive a priori knowledge and prior experiments. The Middle Route of the South-to-North Water Diversion Project (MRP) was used as the delegate to demonstrate that Bloomformer-1 exhibited more robust performance (with the highest R^2, 0.80 to 0.94, and the lowest RMSE, 0.22 to 0.43$\mu g/L$) compared to four widely used traditional machine learning models, including extra trees regression (ETR), gradient boosting regression tree (GBRT), support vector regression (SVR), and multiple linear regression (MLR). In addition, Bloomformer-1 had higher interpretability (including higher transferability and understandability) than the four traditional machine learning models, which meant that it was trustworthy and the results could be directly applied to real scenarios. Finally, it was determined that total phosphorus (TP) was the most important driver for the MRP, especially in Henan section of the canal. Total nitrogen (TN) had the highest effect on algal growth in the Hebei section. Based on the results, phosphorus loading controlling in the whole MRP was proposed as an algal control strategy.

5.2 Study area

The South-to-North Water Diversion Project (SNWDP) is an extensive infrastructure initiative in China designed to transfer water from the water-abundant southern regions to the arid and water-stressed northern areas (Figure 5.1). The project comprises three primary routes: the Eastern Route, the Middle Route, and the Western Route. Of these, the Middle Route is the largest and most crucial [51].

Originating at the Danjiangkou Reservoir in Hubei Province, created by damming the Han River, a significant tributary of the Yangtze River, the Middle Route channels water northward through an intricate network of canals, tunnels, and aqueducts. It traverses Henan and Hebei provinces before ultimately reaching its primary recipients, Beijing and Tianjin. Spanning approximately 1,200 kilometers (746 miles) in length, the Middle Route boasts a designed annual water transfer capacity of around 9.5 billion cubic meters [52]. The recipient areas mainly utilize the water for domestic, industrial, and agricultural purposes. Construction on the Middle Route began in 2003 and was officially inaugurated in late 2014. This route has been instrumental in mitigating water scarcity in northern China, enhancing water quality, and fostering sustainable socio-economic development in the region [53].

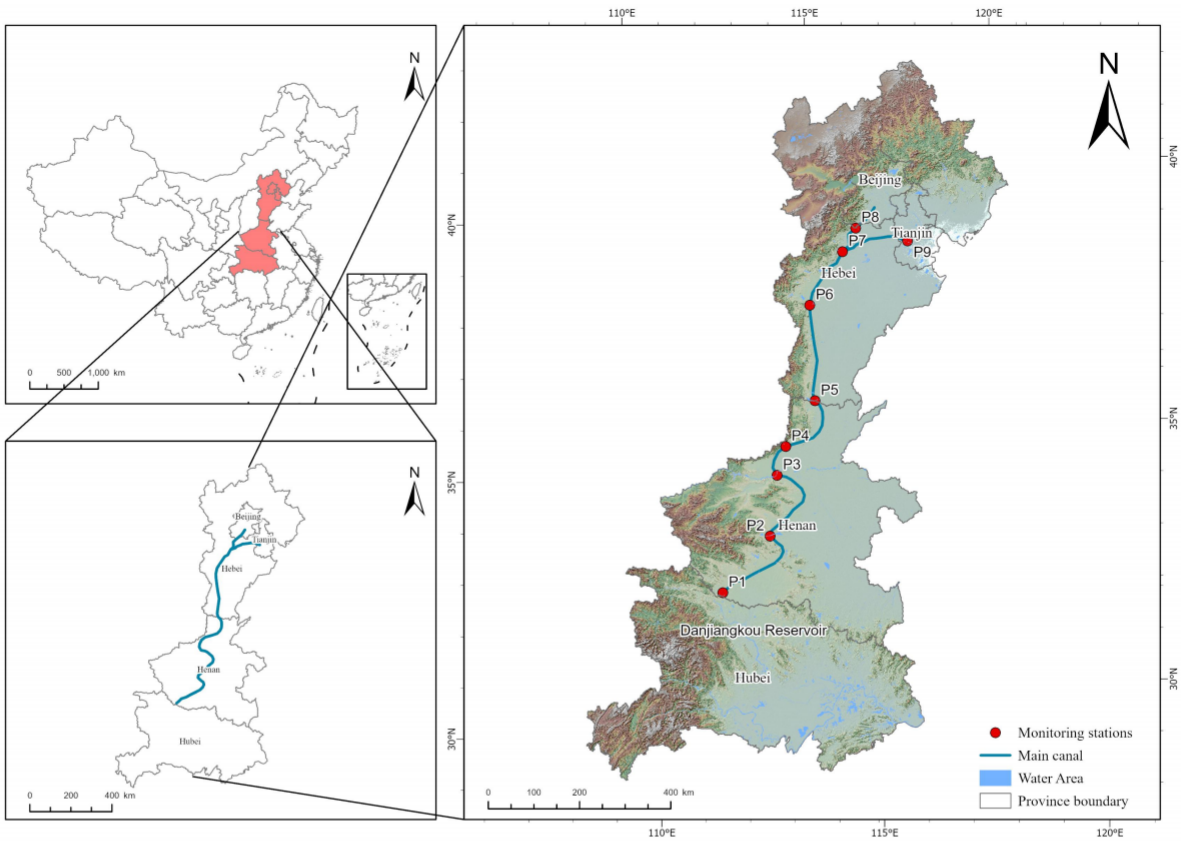


Figure 5.1: Sketch map of sampling stations distribution in the middle route of South-North Water Diversion Project

5.3 Sample collection and chemical analytics

Nine water quality monitoring stations, designated P1 to P9, have been established along the Middle Route of the SNWDP (MRP), stretching from south to north. Stations P1, P2, P3, and P4 are situated in the Henan section; P5, P6, and P7 in the Hebei section; P8 in the Tianjin section; and P9 in the Beijing section (Figure 5.1). The database used in this study includes water quality monitoring data from monitoring stations P1 to P9 for a total of 49 months (August 1, 2018 to August 30, 2022). Water samples were collected at a water depth of 0.5 m, stored at 4°C, and transported to the laboratory for determination of water quality parameters. This task node mainly focuses on the chemical water quality parameters, including total phosphorus (TP), phosphorous-phosphate (PO_4-P), total nitrogen (TN), nitrogen-nitrate (NO_3-N), nitrogen-ammonia (NH_3-N), potassium permanganate index (COD_{Mn}), and total organic carbon (TOC). These parameters were measured according to APHA [54]. The concentration of Chl-a was determined according to ASTM D3731-87 [55].

5.4 Multi-Head-Self-Attention

To analyze information more accurately and efficiently, humans have the ability to adjust their focus on the data they receive when faced with an abundance of information [56]. This suggests that the primary function of the focus mechanism is to assign weights to various

pieces of information. An attention function could be interpreted as mapping a Q(query) and a string of K(key)-V(value) to an output, where Q, K, V, and output were vectors [57]. The attention could be represented as:

$$Output_{Attention} = Attention(Q, K, V) \tag{5.1}$$

Multi-Head Attention was the projection of Q, K, and V by h different linear transformations, and finally, the different attention results were contacted together, which could be represented as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right) W^O \tag{5.2}$$

Where

$$\text{head}_\text{i} = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{5.3}$$

$$W_i^Q \in \mathbb{R}^{d_\text{model} \times d_k}, W_i^K \in \mathbb{R}^{d_\text{model} \times d_k}, W_i^V \in \mathbb{R}^{d_\text{model} \times d_v}$$

and

$$W^O \in \mathbb{R}^{hd_v \times d_\text{model}}$$

There is a direct correlation between the input and output weights in the appealed Attention mechanism, suggesting that the output vectors must participate in the weight calculation. As opposed to this, the weight of Self-Attention was an internal weight relationship between input vectors, which did not require participation from output vectors. Therefore, the multi-head-self-attention (MHSA) meant Q, K, and V were the same. In this study, we used the scaled dot-product to calculate Attention:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{5.4}$$

Where d_k was the vector dimension in both Q and K.

The encoder consisted of N same units(Figure 5.2). Each unit consists of two sub-layers, the multi-head-self-attention layer, and the fully connected feed-forward network, where each sub-layer was processed with the residual connection "Add" and normalization "Norm". The output of the sub-layer could be represented as:

$$Output_{Sublayer} = Norm(x + F(x)) \tag{5.5}$$

Where $F(x)$ was a function of the sublayer itself, multi-head-self-attention, or fully connected feed-forward network.

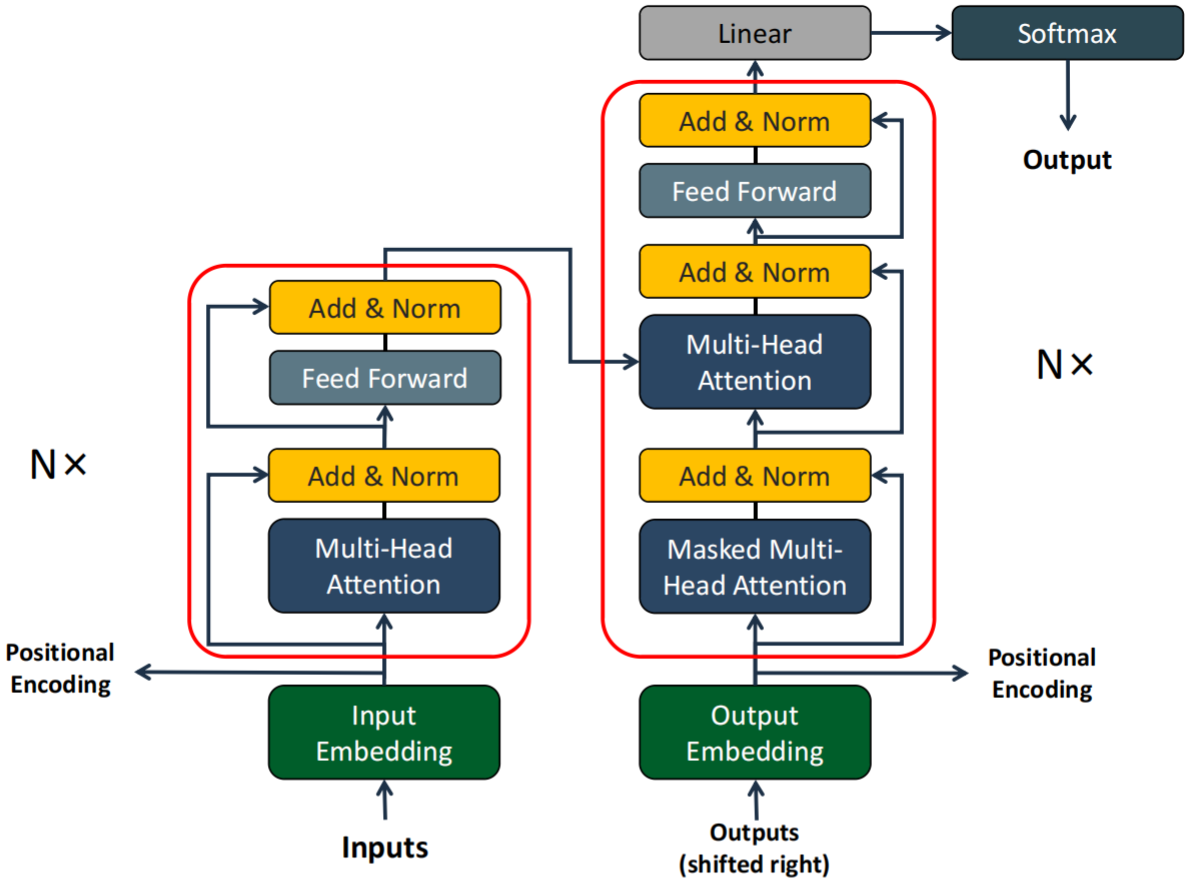


Figure 5.2: The architecture of standard Transformer [57]

Fully connected feed-forward networks provide a non-linear transformation that consists of two linear transformations with the active function ReLu [43]. Compared with the encoder, the decoder added another MHSA layer (Figure 5.2). A mask operation [58] was applied to this multi-head self-attention layer in order to prevent the model from being exposed to future information during training.

5.5 Bloomformer-1

Algal growth is a multifactorial process, and identifying its driving factors is a common application of multivariate regression. The key to addressing this issue lies in understanding the spatial relationships between the variables. However, the standard Transformer based on Multi-Head-Self-Attention (MHSA) is not specifically designed for this purpose, as it treats the value of each variable at a given time period as a single point in its graph, preventing each variable from having its own context prioritization [59].

To investigate spatial relationships, I have developed Bloomformer-1, based on the Transformer architecture. This enhanced method starts by converting the database's context sequence into an extensive spatial sequence. This sequence is then transposed to obtain its corresponding lengthy spatial sequence. Utilizing an encoder-decoder architecture based on Transformers, this sequence is processed to derive the expected values for each variable.

The predicted values are subsequently restructured into their original format and trained to minimize prediction error metrics.

Bloomformer-1's training structure comprises a reconstruction stage and a regression stage. The reconstruction task involves unsupervised pre-training and reconstruction of explanatory variables through the connected encoder stack and decoder stack to extract robust and compact features. The reconstruction task shares the encoder stack settings and position encoding with the relevant portion of the regression task. In this study, the number of units in the encoder and decoder layers is set at eight, representing the 7-dimensional water quality metrics and 1-dimensional station location data. The station location information for the substation task corresponds to the station number matched to each water quality measurement, ranging from 1 to 9. For the entire MRP procedure, the station location information is set to 1. Mean Square Error (MSE) is selected as the loss function for both the reconstruction and regression stages. Figure 5.3 illustrates the architecture of Bloomformer-1. The MHSA mechanism of Bloomformer-1 facilitates the positive and simultaneous derivation of driving factor identification results achieved during model training.

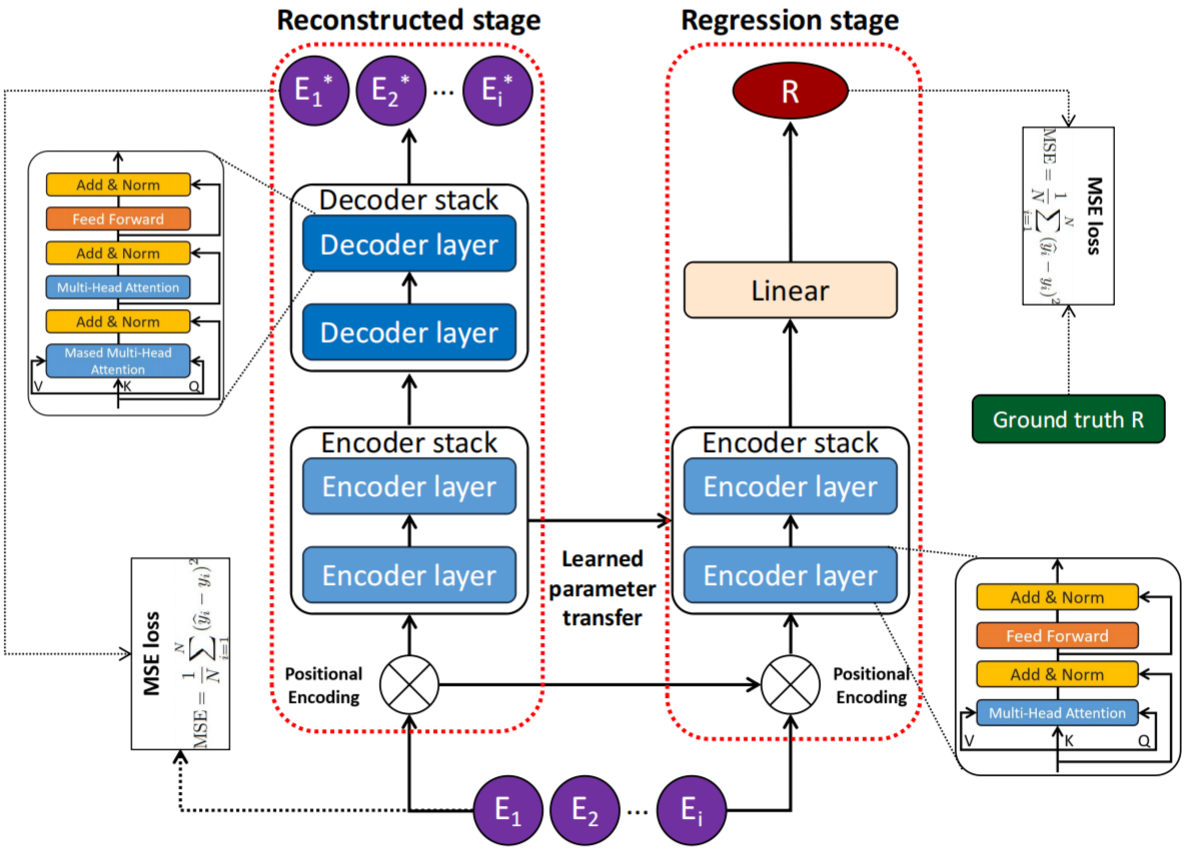


Figure 5.3: The architecture of Bloomformer-1

5.6 Training and performance evaluation of model

In this study, Chl-a and other previously described water quality parameters for the nine monitoring stations (P1 to P9) were incorporated into the model as response and explanatory variables. This was done for each station separately and for the entire project to identify

the drivers of algal growth. Prior to inputting the data into the model, normalization was performed using the Z-score (Section 4.3.1). Following the rule of randomly selecting one out of every five steps, the dataset was divided into training and testing sets. Consequently, 80% of the dataset was employed for model training, while the remaining 20% was used to evaluate the model's performance. Additionally, tenfold cross-validation was introduced during the training phase to prevent overfitting.

To assess the accuracy and stability of each regression model, two indicators were applied to the test set: R^2 and RMSE (Section 4.3.3).

5.7 Computational environment

The experiment was carried out on a PC with the following features: Hardware: CPU i7-6950X, RAM 64GB, dual GeForce RTX 3090, VRAM 24GB; Software: Ubuntu 20.04, Python3.6, Pytorch 1.10.0, Numpy 19.2.

5.8 Results

5.8.1 Model performance evaluation

The model performance results are summarized in Table 5.1 (The bold-italic values represent the best regression results and unit of RMSE is $\mu g/L$). Comparisons between model simulations and ground truth can be found in Figure 5.4 and Figure 5.5.

For stations P1, P2, and P3, Bloomformer-1 demonstrates a significantly better performance compared to the other four conventional machine learning models. Regarding stations P4, P5, P6, P8, and P9, Bloomformer-1 exhibits relatively high performance. Although the difference in R^2 values compared to Extra Trees Regression (ETR) is small (ranging from 0.03 to 0.06), there is still a considerable advantage in RMSE values (e.g., Bloomformer-1 has an RMSE value of 0.33 for P4, while ETR has the lowest RMSE value of 0.52 among the other four traditional machine learning models). In the case of P7, although ETR shares the same R^2 value as Bloomformer-1 (both at 0.94), Bloomformer-1 maintains a clear advantage in RMSE values (Bloomformer-1 with 0.23, ETR at 0.43,Gradient boosting regression tree (GBRT) with 0.47, SVR at 0.45, and MLR with 0.66). Consistent with the results for the individual stations, Bloomformer-1 continues to exhibit robustness across the entire MRP. In conclusion, Bloomformer-1 is the most effective model for describing the relationship between Chl-a concentrations and driving factors when compared to traditional machine learning models.

Table 5.1: Results of model performance evaluation

Stations	Indicator	Bloomformer-1	ETR	GBRT	SVR	MLR
P1	R^2	**0.85**	0.75	0.72	0.63	0.42
	RMSE	**0.32**	0.56	0.57	0.60	0.73
P2	R^2	**0.80**	0.66	0.51	0.63	0.25
	RMSE	**0.43**	0.62	0.68	0.63	0.82
P3	R^2	**0.83**	0.70	0.39	0.58	0.39
	RMSE	**0.40**	0.59	0.69	0.64	0.79
P4	R^2	**0.89**	0.84	0.68	0.46	0.35
	RMSE	**0.33**	0.52	0.62	0.61	0.76
P5	R^2	**0.90**	0.89	0.78	0.88	0.49
	RMSE	**0.30**	0.50	0.58	0.51	0.71
P6	R^2	**0.89**	0.85	0.74	0.88	0.46
	RMSE	**0.26**	0.45	0.49	0.43	0.68
P7	R^2	**0.94**	**0.94**	0.85	0.92	0.68
	RMSE	**0.23**	0.43	0.47	0.45	0.66
P8	R^2	**0.94**	0.91	0.84	0.89	0.71
	RMSE	**0.22**	0.43	0.48	0.44	0.62
P9	R^2	**0.93**	0.91	0.89	0.86	0.62
	RMSE	**0.28**	0.46	0.48	0.49	0.68
Whole MRP	R^2	**0.85**	0.79	0.73	0.80	0.39
	RMSE	**0.35**	0.54	0.55	0.51	0.70

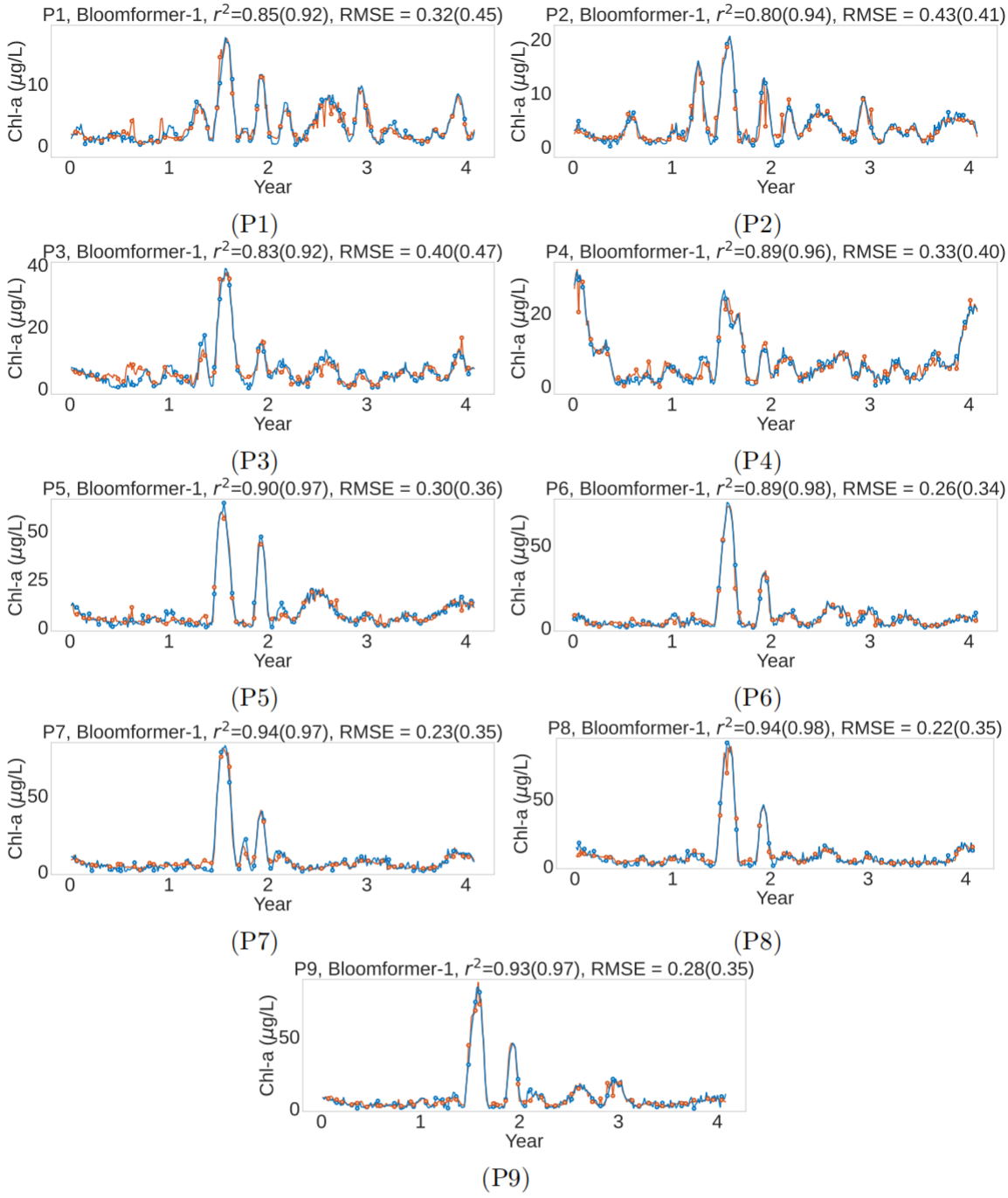


Figure 5.4: Performance of Bloomformer-1 in P1-P9 (Blue lines are observations, and red lines are model simulations. The circles are the test set, where the blue circles are the true values, and the red circles are the predicted values. The part of the blue line, except for the blue circles, is the training set. Numbers show RMSE and r^2 for model prediction and training data inside brackets.

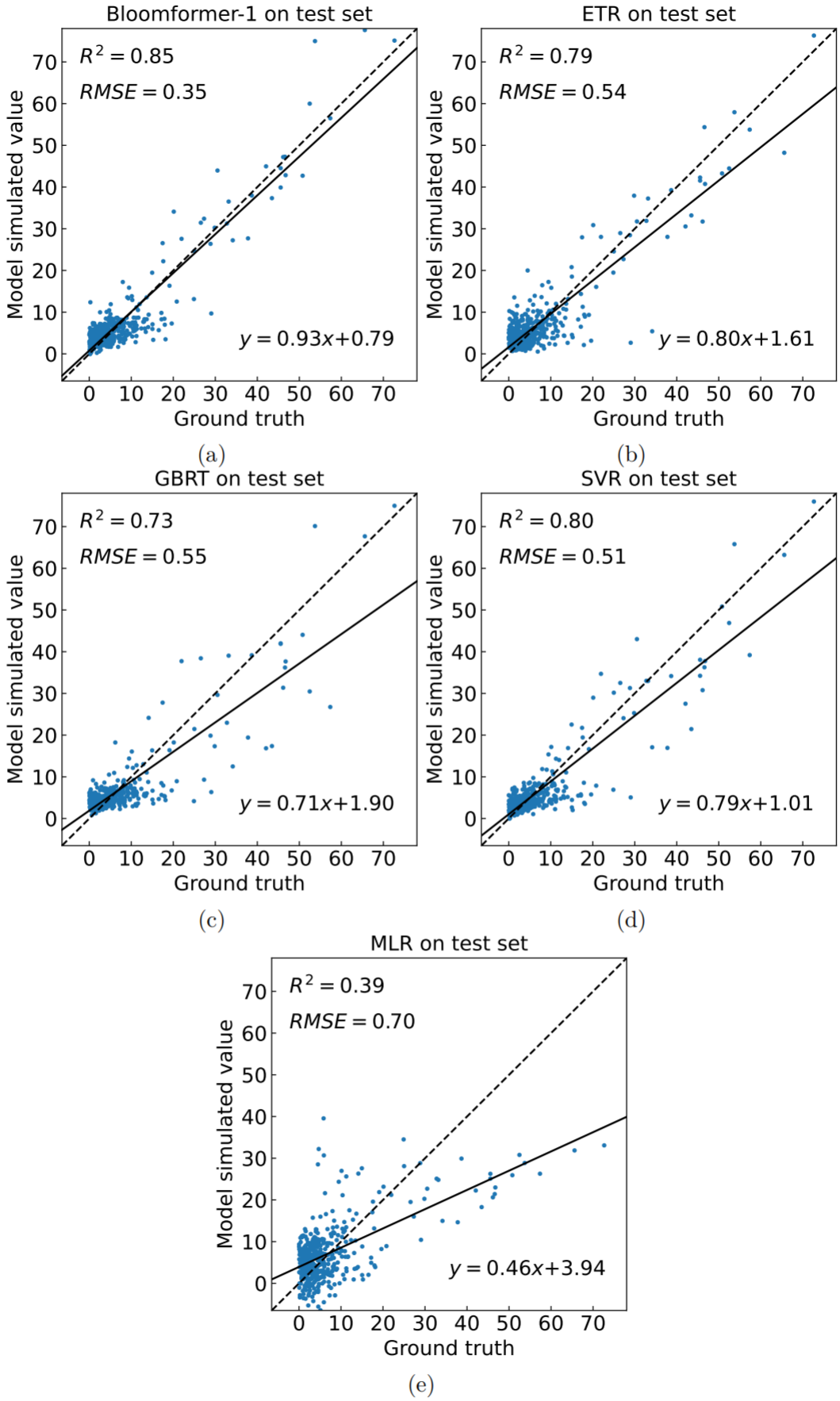


Figure 5.5: Model performance evaluation in the whole MRP, where (a), (b), (c), (d) and (e) represent the test results of the Bloomformer-1, ETR, GBRT, SVR and MLR, respectively.

5.8.2 Result of algal growth driving factors identification in MRP

Based on Bloomformer-1's MHSA, the drivers of algal growth in MRP are shown in Figure 5.6. The most dominant algal growth driver in P1, P2 and the whole MRP was TP with 18.73%, 19.20% and 22.28%, respectively. Notably, PO_4-P also showed a very close occupancy rate of 16.09% in the whole MRP. Results for P5, P6, P8 and P9 showed that the most dominant algal growth driver at these four sites was NO_3-N, with 20.24%, 28.27%, 20.16% and 17.16%, respectively. In P4 and P7, TN was the main algal growth driver with 22.16% and 17.96%, respectively. The results of P3 differed from the others with 23.84% NH_3-N as the most dominant algal growth driver.

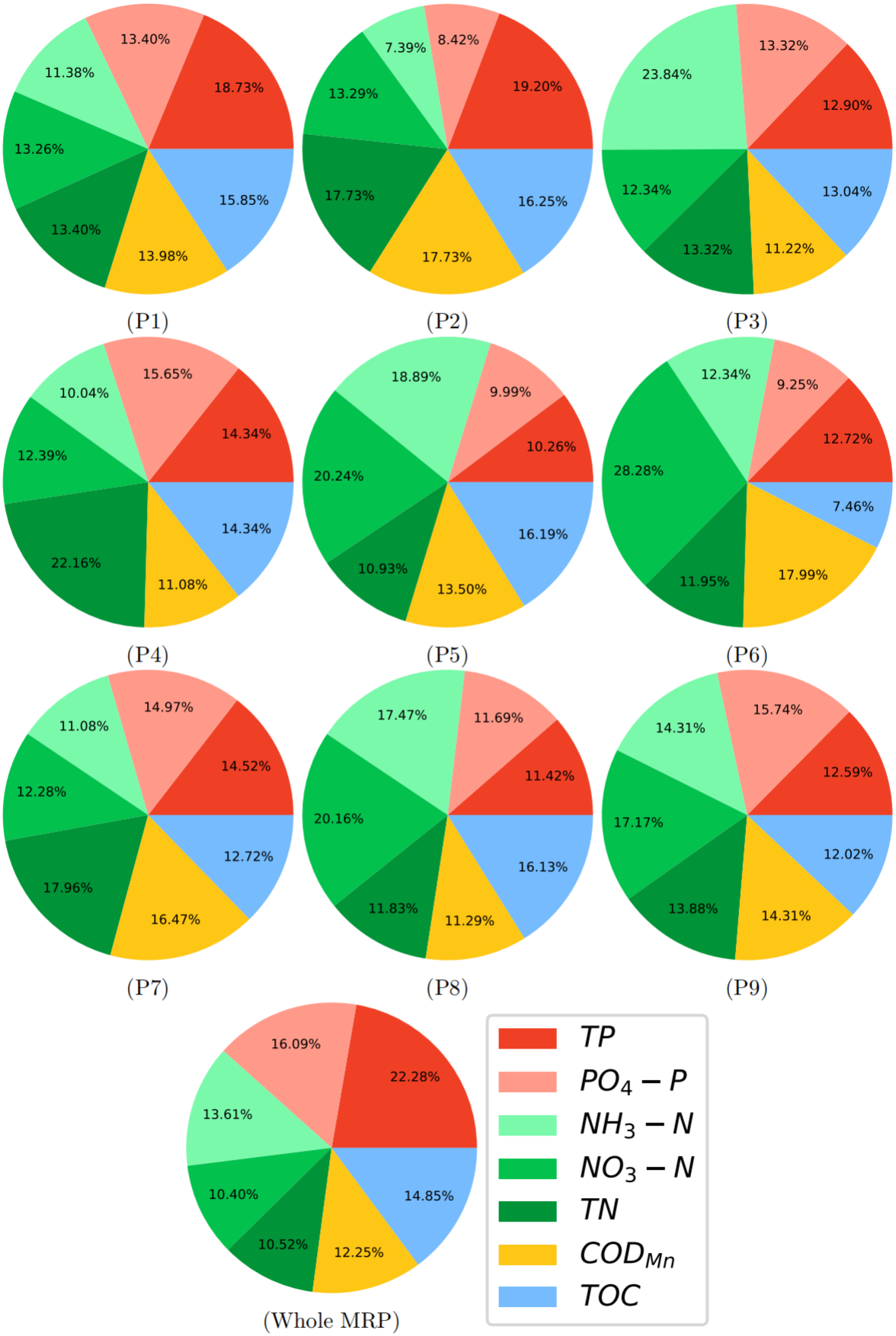


Figure 5.6: Results of algal growth driving factor in MRP

5.9 Conclusion of identifying algal growth driving factors task node

Bloomformer-1, a sophisticated deep learning-based Transformer model, conceived for the seamless identification of algae proliferation catalysts without the necessity of comprehensive antecedent knowledge or preliminary experimentation, garnered the apex R^2 (0.80 to 0.94) and the nadir RMSE (0.22 to 0.43$\mu g/L$) on both solitary subsites and all-encompassing simulations within the MRP. When contrasted against four prevalent conventional machine learning models (ETR, GBRT, SVR, and MLR) employed on monitoring data, Bloomformer-1 demonstrated superior interpretability, insinuating its reliability and the potential for the direct implementation of its resultant data in tangible world scenarios. TP emerged as the predominant instigator within the MRP. Consequently, the governance of phosphorus and its reduction would constitute a critical stratagem to curtail algal proliferation and sustain the stability of water quality within the MRP.

6

Algal Growth Early Warning Task Node

6.1 Submitted paper

Title: An Intelligent Early Warning System for Harmful Algal Blooms: Harnessing the Power of Big Data and Deep Learning

Authors: Jing Qian, Nan Pu, Li Qian, Yonghong Bi and Stefan Norra

Journal: This paper has been submitted to Environmental Science & Technology. The state is under review.

Authorship statement: The research of algal growth early warning and prevention is based on data obtained from the November 2018 and May 2019 field trips. I designed, implemented, and was responsible for the entire research process. During the field trip, Andre Wilhelms and I worked together to set up the BIOLIFT instrument and complete the field trip activities. The data pre-processing was done entirely by me alone. I proposed a new deep learning model, named Bloomformer-2, and completed the model building. In addition, I discussed the results with Nan Pu (Leiden University) and Li Qian (Ludwig Maximilian University of Munich) for productization and industrialization. The project was supervised by Prof. Dr. Stefan Norra (KIT and Potsdam University) and Prof. Dr. Yonghong Bi (IHB), with funding provided by Prof. Dr. Yonghong Bi (IHB).

Abstract: Harmful algal blooms (HABs) pose a significant ecological threat and economic detriment to freshwater environments. In an endeavor to manage these occurrences, we have harnessed the potential of big data and deep learning models to engineer an intelligent early warning system for HABs. Data acquisition is accomplished through a Vertical Aquatic Monitoring System (VAMS), which, in conjunction with the "DeepDPM-Spectral Clustering" methodology, facilitates an intricate analysis of the vertical algal distribution. This approach curtails the number of predictive models and enhances the adaptability of the system. Employing the Bloomformer-2 model, developed by our team, the system carries out both single-step and multi-step prognostications of HABs. Our case study corroborates the superior

performance of Bloomformer-2, exhibiting high congruity with actual value curves and a lower margin of predictive error. This system boasts the unique ability to identify the driving factors of HABs, thereby aiding in the formulation of targeted preventive measures. Additionally, the model's remarkable intelligence - the capacity to autonomously learn from preprocessed data - and its inherent adaptability pave the way for future system upgrades and broader applications. As part of future work, it is proposed to augment the big data platform and establish a VAMS monitoring network to bolster the system's geographical coverage and predictive capability. This research underscores the transformative potential of integrating big data and artificial intelligence in environmental management, and emphasizes the importance of model interpretability in machine learning applications.

6.2 Study area

Taihu Lake, situated in the rapidly developing Yangtze River Delta region of China, is the country's third-largest freshwater lake, with a surface area of 2,338 square kilometers [60] (Figure 6.1. This shallow lake has an average depth of 1.9 meters [61]. Meiliang Bay, one of the most eutrophic bays in the northern part of Taihu Lake, has an area of 124 square kilometers and an average depth of 1.5 meters. The Liangxi River and Zhenwugang River, two major rivers, transport urban pollutants from Wuxi and Changzhou City into Meiliang Bay. Since 1998, Meiliang Bay has experienced severe algal blooms in both summer and fall, owing to its role as a major recipient of human activities and an important source of drinking water.

To evaluate the performance of the algal early warning system I developed, I conducted experiments at the end of a 250-meter-long jetty at the Taihu Laboratory for Lake Ecosystem Research (TLLER) (31.418903 N, 120.213293 E), located on the southern side of Meiliang Bay. The panoramic photograph of the TLLER and the BIOLIFT installation position are shown in Figure 6.2.

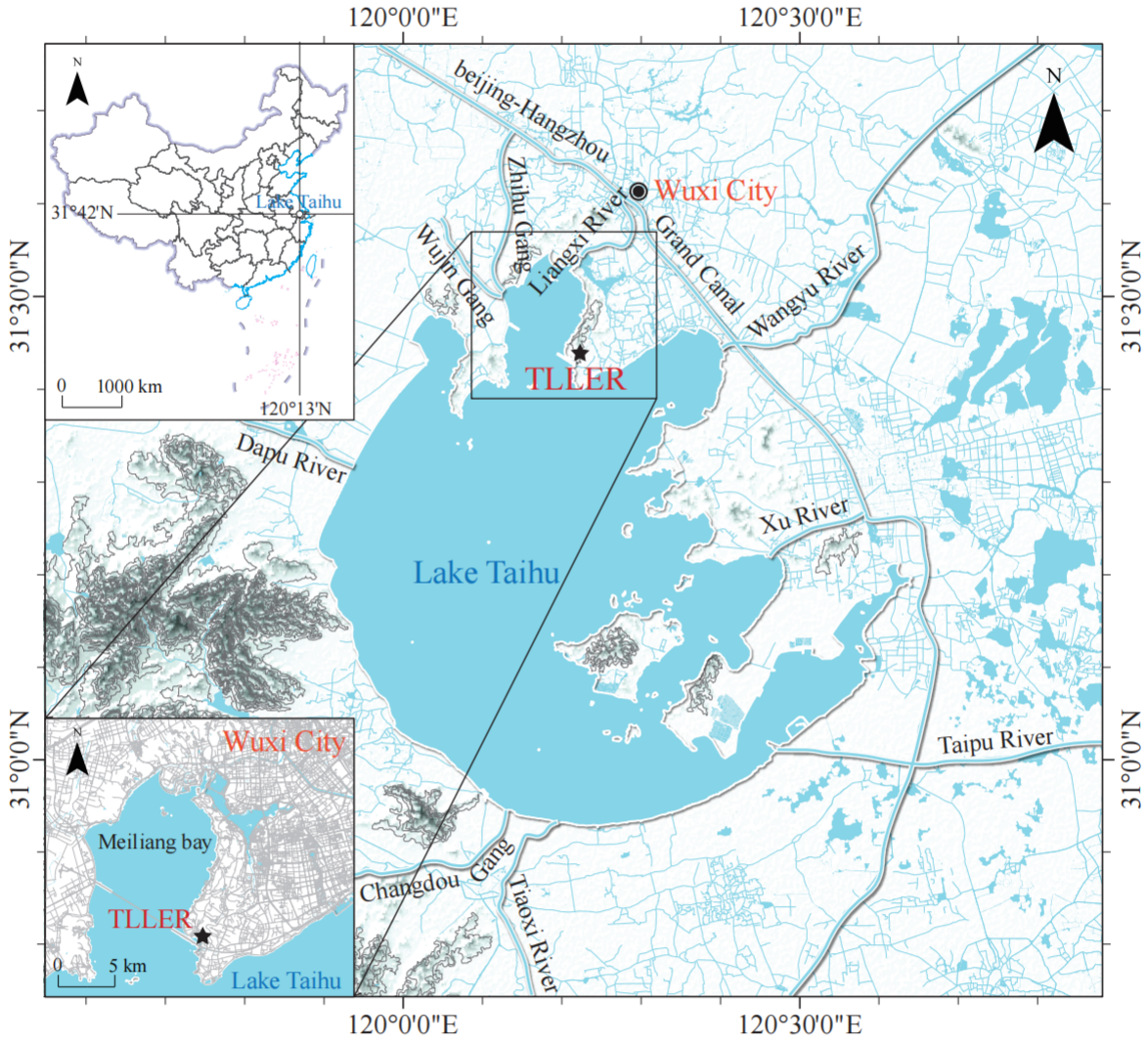


Figure 6.1: Location of Lake Taihu, Meiliang Bay and TLLER (pentagon)

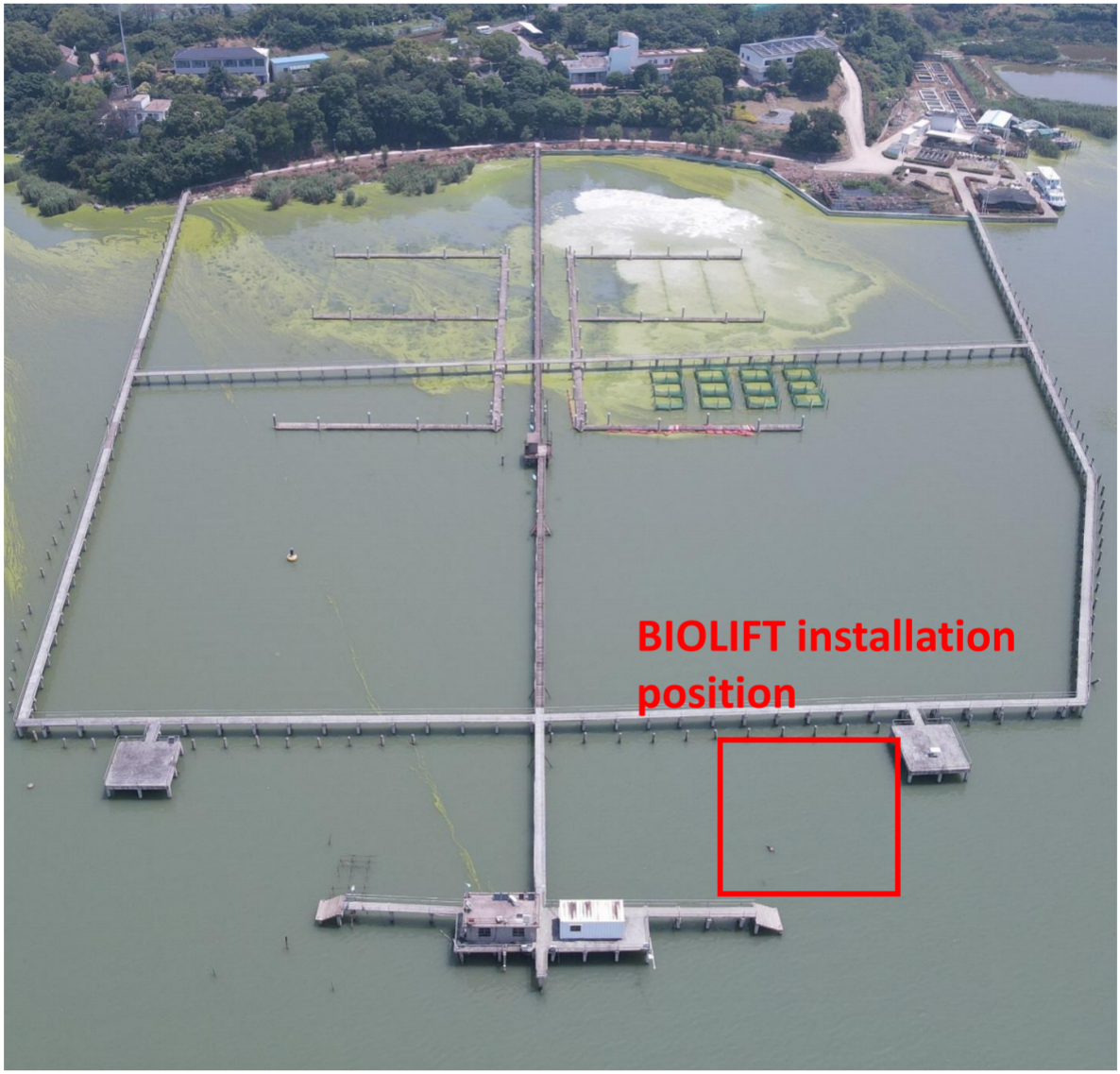


Figure 6.2: Panoramic photograph of the TLLER and the BIOLIFT installation position

6.3 BIOLIFT and EBDP construction

In this task node, multi-sensor vertical monitoring is conducted using BIOLIFT, an advanced system equipped with multiple sensors connected to a control box via a data transmission line (Figure 6.3). BIOLIFT performs real-time recording of water quality parameters and depth indicators while moving up and down at periodic intervals. Researchers can specify work cycles, which are a series of up and down movements, by setting the time interval between each cycle. The recorded parameters include conductivity (EC_{25}), temperature (Temp.), pH, turbidity (Turb.), colored dissolved organic matter (CDOM), and Chl-a at 25°C. A meteorological station (Vaisala meteorological Transmitter WXT520) is integrated with BIOLIFT to record wind speed (WS) and wind direction (WD) during its operation. The particulars pertaining to the BIOLIFT sensors, including their specifications, are elegantly delineated in Figure 6.1, whilst the meteorological station is vividly depicted in Figure 6.2.

Table 6.1: Sensors of BIOLIFT and their specifications

Parameter	Principle	Range	Resolution	Accuracy	Response time
Pressure	piezo-resistive	0-200dBar	0.005dBar	±0.1dBar	0.04s
Temp.	Pt 100	-2-38°C	0.001°C	±0.01°C	0.12s
pH	Potentiometric (Ag/AgCl)	0-14pH	0.02pH	0.02pH	1s(63%)
CDOM	Fluorescence exc.325nm fl. 470nm	0.15-1250 ppbQS	0.01ppbQS	±5%	1s
Chl-a	Fluorescence exc.465nm fl. 696nm	0.03-500 $\mu g/L$	0.01 $\mu g/L$	N.A.	1s
EC_{25}	7-pole-cell	0-6mS/cm	0.1uS/cm	±2uS/cm	0.05s
Turb.	Mie backscattering	0-750FTU	<0.001%	±2%	0.1s

Table 6.2: Sensors of meteorological station and their specifications

Parameter	Principle	Range	Resolution	Accuracy	Response time
WD	N.A.	0-360°	1°	±3°	0.25s
WS	N.A.	0-60m/s	0.1m/s	±3% at 10m/s	0.25s

BIOLIFT data from the "2018-Winter" and "2019-Summer" datasets collected by TLLER were used to construct the EBDP for this task node. Work cycle intervals were set at 10 minutes, and individual depth segments were set at 0.05 m. The "2018-Winter" dataset included 15 days of BIOLIFT data, with each day comprising 132 work cycles and water quality data for 23 depth segments (0.05m to 1.2m), as well as wind speed and direction. The "2019-Summer" dataset captured 13 days of BIOLIFT data, with each day consisting of 134 work cycles and water quality data for 37 depth segments (0.1m to 1.95m), along with wind speed and direction.

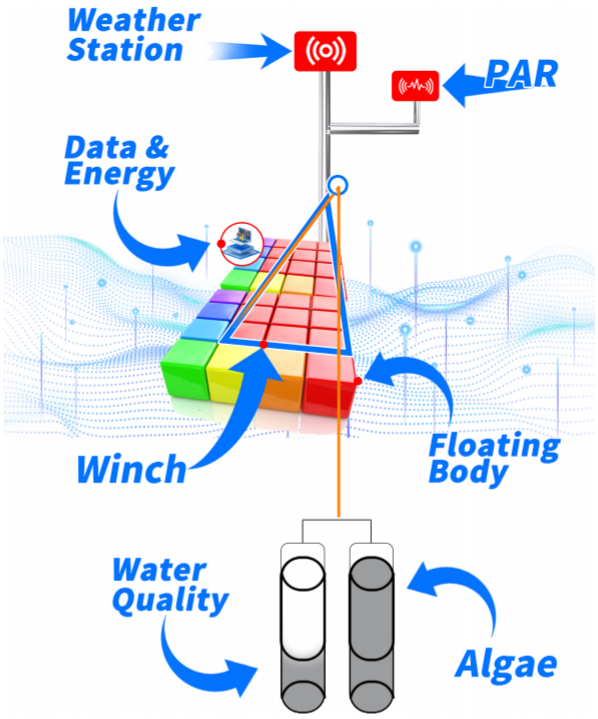


Figure 6.3: Diagram of BIOLIFT

6.4 Optimization of modeling strategy

The "DeepDPM-Spectral Clustering" method was developed to cluster depth segments into several reasonable groups. The efficiency of the system is improved by an optimization strategy that models each group rather than each depth segment.

DeepDPM is an inference algorithm capable of inferring and changing the number of clusters during training [62]. DeepDPM contains two main parts, the first part is the clustering network, and the second is K subclustering networks (one for each cluster k, $k \in \{1, \ldots, K\}$). The workflow of DeepDPM (Figure 6.4) is shown below [62]:

First given an arbitrary initial cluster number K, the data is fed to the clustering network f_{cl}, which generates K soft cluster assignments for each data point $\boldsymbol{x}_i$:

$$f_{\text{cl}}(\mathcal{X}) = \boldsymbol{R} = (\boldsymbol{r}_i)_{i=1}^{N} \tag{6.1}$$

$$\boldsymbol{r}_i = (r_{i,k})_{k=1}^{K} \tag{6.2}$$

Where $r_{i,k} \in [0, 1]$ is the soft cluster assignment $\boldsymbol{R}$ of $\boldsymbol{x}_i$ to cluster k and $\sum_{k=1}^{K} r_{i,k} = 1$.

Secondly, the hard assignments $\boldsymbol{z} = (z_i)_{i=1}^{N}$ are calculated according to the equation:

$$z_i = \arg\max_k r_{i,k} \tag{6.3}$$

Next, each subclustering network f_{sub}^{k} is fed the hard assignments data for its respective cluster and generates a soft subcluster assignment, as the following equations show:

$$f_{\text{sub}}^{k}\left(\mathcal{X}_k\right)=\tilde{\boldsymbol{R}}_k=\left(\tilde{\boldsymbol{r}}_i\right)_{i:z_i=k} \tag{6.4}$$

$$\tilde{\boldsymbol{r}}_i=\left(\tilde{\boldsymbol{r}}_{i,j}\right)_{j=1}^{2} \tag{6.5}$$

Where $\tilde{r}_{i,j}\in[0,1]$ is the soft assignment of $\boldsymbol{x}_i$ to subcluster $j(j\in\{1,2\})$, and $\tilde{r}_{i,1}+\tilde{r}_{i,2}=1\forall k\in\{1,\ldots,K\}$.

The clustering network f_{cl} and each subclustering network f_{sub}^{k} is a simple multilayer perceptron with a single hidden layer. The last layer of the clustering network has K neurons, while the last layer of each subclustering network has two.

Finally, the split or merge decisions are made for changing K according to the Metropolis-Hastings framework [63].

The split proposals are accepted stochastically with probability $\min\left(1,H_{\text{s}}\right)$, where H_{s} is Hastings ratio. In the split step, each cluster is split into its two subclusters. The merge proposals are accepted/rejected using the reciprocal number of the Hastings ratio H_{s}.

$$H_{\text{s}}=\frac{\alpha\Gamma\left(N_{k,1}\right)f_{\boldsymbol{x}}\left(\mathcal{X}_{k,1};\lambda\right)\Gamma\left(N_{k,2}\right)f_{\boldsymbol{x}}\left(\mathcal{X}_{k,2};\lambda\right)}{\Gamma\left(N_k\right)f_{\boldsymbol{x}}\left(\mathcal{X}_k;\lambda\right)} \tag{6.6}$$

Where H_{s} is the Hastings ratio, Γ is the Gamma function, $\mathcal{X}_k=\left(\boldsymbol{x}_i\right)_{i:z_i=k}$ stands for the points in the cluster k, $N_k=|\mathcal{X}_k|$, $\mathcal{X}_{k,j}=\left(\boldsymbol{x}_i\right)_{i:(z_i,\tilde{z}_i)=(k,j)}$ denotes the points in the subcluster, $j(j\in\{1,2\})$, $N_{k,j}=|\mathcal{X}_{k,j}|$, and $f_{\boldsymbol{x}}(\cdot;\lambda)$ is the marginal likelihood where λ represents the Normal-Inverse Wishart hyperparameters [64].

After the split and merge steps, the initial cluster number K, clustering network, and K subclustering networks are updated, and iterative operations are performed until the optimal cluster number K is found.

The DeepDPM algorithm's implementation is available on a public GitHub repository, https://github.com/BGU-CS-VIL/DeepDPM. In order to accommodate my initiative, I've undertaken a recompilation of the standard code's terminal segments. This entailed the integration of optimal cluster number statistics and the direct formulation of adjacency matrices. The code is written in Python and utilizes the PyTorch library.

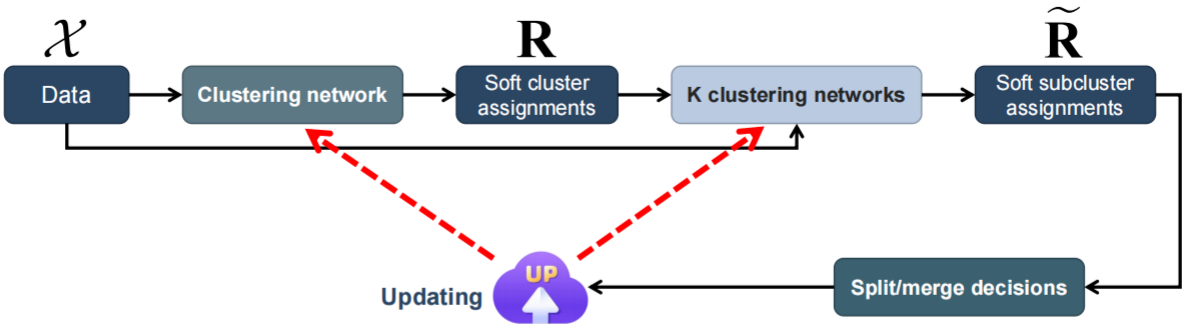


Figure 6.4: Workflow of DeepDPM [62]

Spectral clustering [65] is an unsupervised ML method used for partitioning data into groups or clusters based on the similarity between data points. It is particularly effective for data that does not fit the assumptions of traditional clustering algorithms like K-means [66] or hierarchical clustering [67]. The main idea behind spectral clustering is to analyze the eigenvectors and eigenvalues of the Laplacian matrix derived from the data's adjacency matrix. The steps of spectral clustering are as shown below [65].

Given a set of points $S = \{s_1, \ldots, s_n\}$ in $\mathbb{R}^l$ that we want to cluster into k subsets:

1. Form the affinity matrix $A \in \mathbb{R}^{n \times n}$ defined by $A_{ij} = \exp\left(-\|s_i - s_j\|^2 / 2\sigma^2\right)$ if $i \neq j$, and $A_{ii} = 0$.

2. Define D to be the diagonal matrix whose (i, i)-element is the sum of A 's i-th row, and construct the matrix $L = D^{-1/2} A D^{-1/2}$.

3. Find $x_1, x_2, \ldots, x_k$, the k largest eigenvectors of L, and form the matrix $X = [x_1 x_2 \ldots x_k] \in \mathbb{R}^{n \times k}$ by stacking the eigenvectors in columns.

4. Form the matrix Y from X by renormalizing each of X 's rows to have unit length (i.e. $Y_{ij} = X_{ij} / \left(\sum_j X_{ij}^2\right)^{1/2}$).

5. Treating each row of Y as a point in $\mathbb{R}^k$, cluster them into k clusters via K-means or any other algorithm (that attempts to minimize distortion).

6. Assign the original point s_i to cluster j if and only if row i of the matrix Y was assigned to cluster j.

This code was developed by me in Python to adapt the Spectral clustering algorithm to my project, utilizing the fundamental concepts proposed by the author. The adopted strategy enables spectral clustering to commence directly from a known cluster count and adjacency matrix, circumventing the initial adjacency matrix construction. This primarily aims to bolster the connectivity with DeepDPM.

In this task node, water quality parameter data, wind speed and direction from each work cycle in the "2018-Winter" and "2019-Summer" datasets were separately input into DeepDPM for deep clustering. The optimal number of clusters for each cycle was self-taught by DeepDPM, and the distribution of the optimal number of clusters was counted. In addition, the adjacency matrix can be obtained by calculating the probability of each depth segment being classified into the same cluster throughout the experiment. After that, the adjacency matrix was

clustered using spectral clustering. As the magnitude of the data scale swells (as in the scenario of an infinite data scale), the optimum cluster number across all temporal instances inclines towards congruity. Within the confines of the existing data scale, the predominant optimum clustering number, on account of its emblematic nature, is chosen to represent the totality of cluster number.

6.5 Long Short Term Memory

In this task node, I utilized a well-established DL model for time series prediction, known as LSTM [68], to serve as a comparison with Bloomformer-2. LSTM is a type of recurrent neural network (RNN) [69] architecture designed to address the vanishing gradient [70] problem commonly encountered in traditional RNN. LSTM have a more complex structure than standard RNN, incorporating memory cells and various gates to control the flow of information through the network.

The architecture of the LSTM is shown in Figure 6.5. For moment t, the LSTM has three inputs: the cell state C_{t-1}, the hidden layer state h_{t-1}, and the input vector at moment t, X_t. In addition, there are two outputs: the cell state C_t and the hidden layer state h_t, where h_t is also used as the output at moment t.

The gate layers of the LSTM is designed with some computational steps to adjust the input with the values of the two hidden layers. The gate layers in LSTM contain forget gate layer, input gate layer and output gate layer. The square components in Figure 6.5 represent neurons, and the difference between them is the difference in activation functions. σ denotes the Sigmoid function, whose output is between 0 and 1, and tanh is the hyperbolic tangent function, whose output is between -1 and 1 [71].

The forget gate layer plays a crucial role in determining which information is retained or discarded from the cell state. When the forget gate has a value of 1, it allows all the information from the previous cell state to persist. On the other hand, when the forget gate has a value of 0, it discards all the information, effectively resetting the cell state. The function is [72]:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{6.7}$$

The input gate layer is used to selectively record new information into the cell state, and the functions of input gate layer are [72]:

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{6.8}$$

$$\tilde{c}_t = \tanh\left(W_{\tilde{c}} \cdot [h_{t-1}, x_t] + b_{\tilde{c}}\right) \tag{6.9}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{6.10}$$

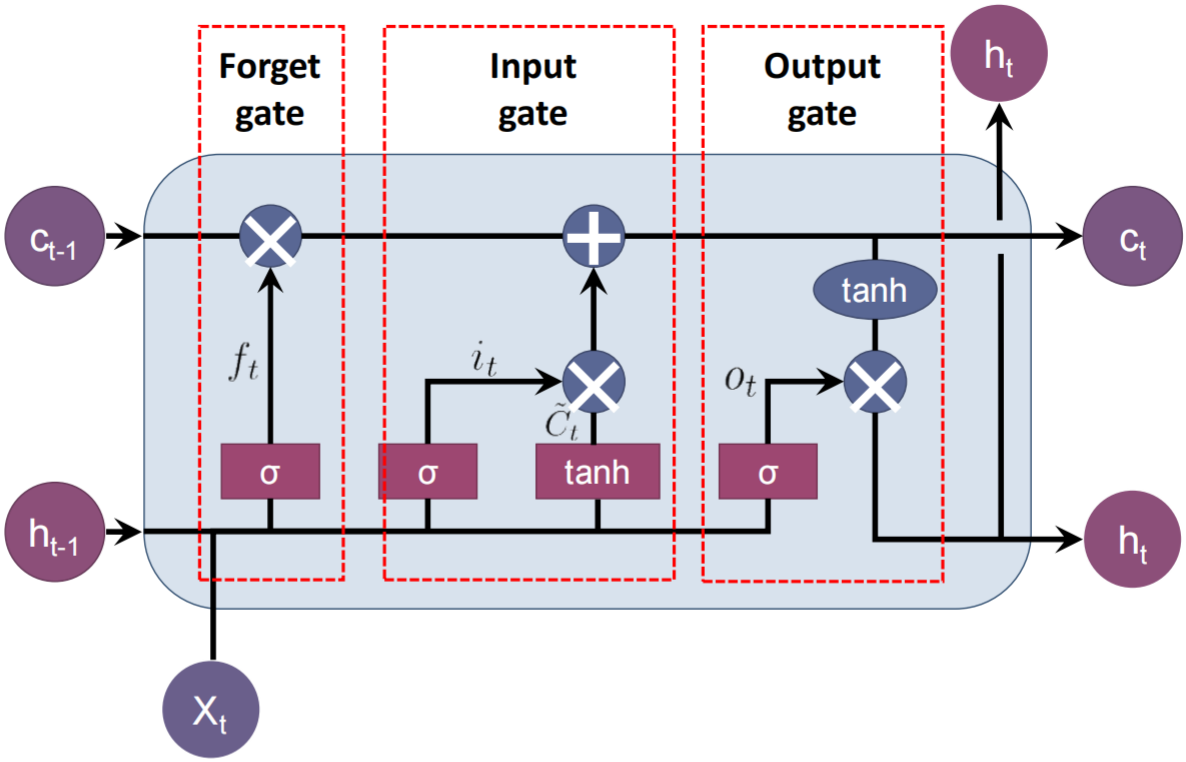


Figure 6.5: Architecture and workflow of LSTM [72]

The output gate layer is used to save the previous information into the hidden layer and output a time step value and the functions of output gate layer are [72]:

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{6.11}$$

$$h_t = o_t * \tanh\left(c_t\right) \tag{6.12}$$

where W_i, $W_{\tilde{c}}$, W_o are the weights of forget gate, input gate and output gate respectively, b_i, $b_{\tilde{c}}$, b_o are corresponding bias, and the operator ' $\cdot$ ' denotes the pointwise multiplication of two vectors. In the process of updating the cell state, the input gate is responsible for determining the new information that can be stored, while the output gate is tasked with deciding what information should be produced, all based on the cell state.

The LSTM architecture allows the network to learn and retain long-range dependencies in the input data by controlling the flow of information through the memory cell and gates.

6.6 Bloomformer-2

Bloomformer-2 is an enhanced version of the standard Transformer [57]. Its MHSA mechanism (Section 5.3.2) evaluates each token in the input sequence in relation to other tokens, gathering and learning dynamic contextual information. Simultaneously, it addresses the issue of the standard Transformer's inability to effectively manage complex multivariate Time Series prediction (TSP) [59], as it treats the value of each variable at a specific time period as a token in its graph, preventing each variable from having its unique contextual priority. The improved method initially transforms the contextual sequence of historical data and the target timestamp for prediction into an extensive spatio-temporal sequence. This sequence is then transposed to generate the corresponding long spatio-temporal sequence. Both sequences are processed using a Transformer-based encoder-decoder architecture to obtain the predicted values for each variable. Subsequently, the predicted values are restructured into their original format and trained to minimize prediction error metrics. The entire process consists of three stages: parallel dual-sequence input, spatiotemporal embedding, spatiotemporal attention and efficiency optimization. Figure 6.6 illustrates the architecture of Bloomformer-2.

In a bid to construct a genuine spatiotemporal prediction model, initially, I devised a parallel dual-sequence input at the ingress point, comprising a temporal sequence and a spatial sequence, which are transposes of each other. The design intention is to individually extract temporal and spatial feature information via subsequent structures. Now, an imperative query demands resolution: How do we embed this sequence so that the attention network parameters can accurately interpret the information within each sequence?

To embed the temporal sequence, I employed Time2vec [73]. Time2vec is the creation of a fully connected (dense) layer, accepting temporal input and outputting a vector of fixed size (a tunable parameter), thereby transforming temporal features into a learnable vector representation. This layer's activation function could be a periodic function; I opted for a sine function, capable of capturing the cyclical characteristics of the input data. Additionally, these activation functions can handle continuously increasing inputs without inducing gradient explosion or saturation, unlike common activation functions (such as ReLU, Softmax, Sigmoid, etc.). In Transformer-related models, spatial embedding is often achieved through position embedding. In conventional Transformer models, absolute position encoding is usually generated by predefined fixed functions (like sine and cosine), which do not involve learning. To render position encoding with enhanced flexibility and generalizability, I implemented learnable absolute position encoding. Specifically, it involves: 1) Initialization: creating a position vector for every possible position, which is randomly initialized. The dimension of the initialized vector is congruent with the dimension of the embedding vector. 2) Vector Addition: Prior to inputting the embedding vector into the model, the embedding vector is added to its corresponding position vector, thereby encoding the position information into the embedding vector. 3) Training: During the model's training process, the position vector is updated through the backpropagation of the fully connected layer.

Following the self-attention of the embedded time and space sequences, temporal and spatial feature information is extracted. The ensuing step is the amalgamation of temporal and spatial

feature information, thereby translating it into spatiotemporal feature information. Here, I have designed a spatiotemporal attention mechanism. Specifically, the "query" outputs of the temporal sequence self-attention layer (Q_t) and the "key" and "value" outputs of the spatial sequence self-attention layer (K_s and V_s) are used for attention computation. Simultaneously, the "query" outputs of the spatial sequence self-attention layer (Q_s) and the "key" and "value" outputs of the temporal sequence self-attention layer (K_t and V_t) are also employed for attention computation. Thus, time and space are integrated via the attention mechanism, which I dub spatiotemporal attention.

Considering the prediction task of algal growth is a prolonged forecasting assignment, utilizing the attention mechanism in this task would complexify learning, thereby augmenting the model's training time. To enhance the model's training efficiency and impart it with a degree of timeliness, I utilized Batch Normalization [74] and Pre-Norm structures [75]. Batch Normalization primarily resolves the Internal Covariate Shift problem, i.e., the alteration of input distribution between network layers, thereby increasing network training instability and reducing the time required to train a deep network. The Pre-Norm structure, while increasing model training stability, can also increase the model's depth, thereby endowing Bloomformer-2 with considerable potential.

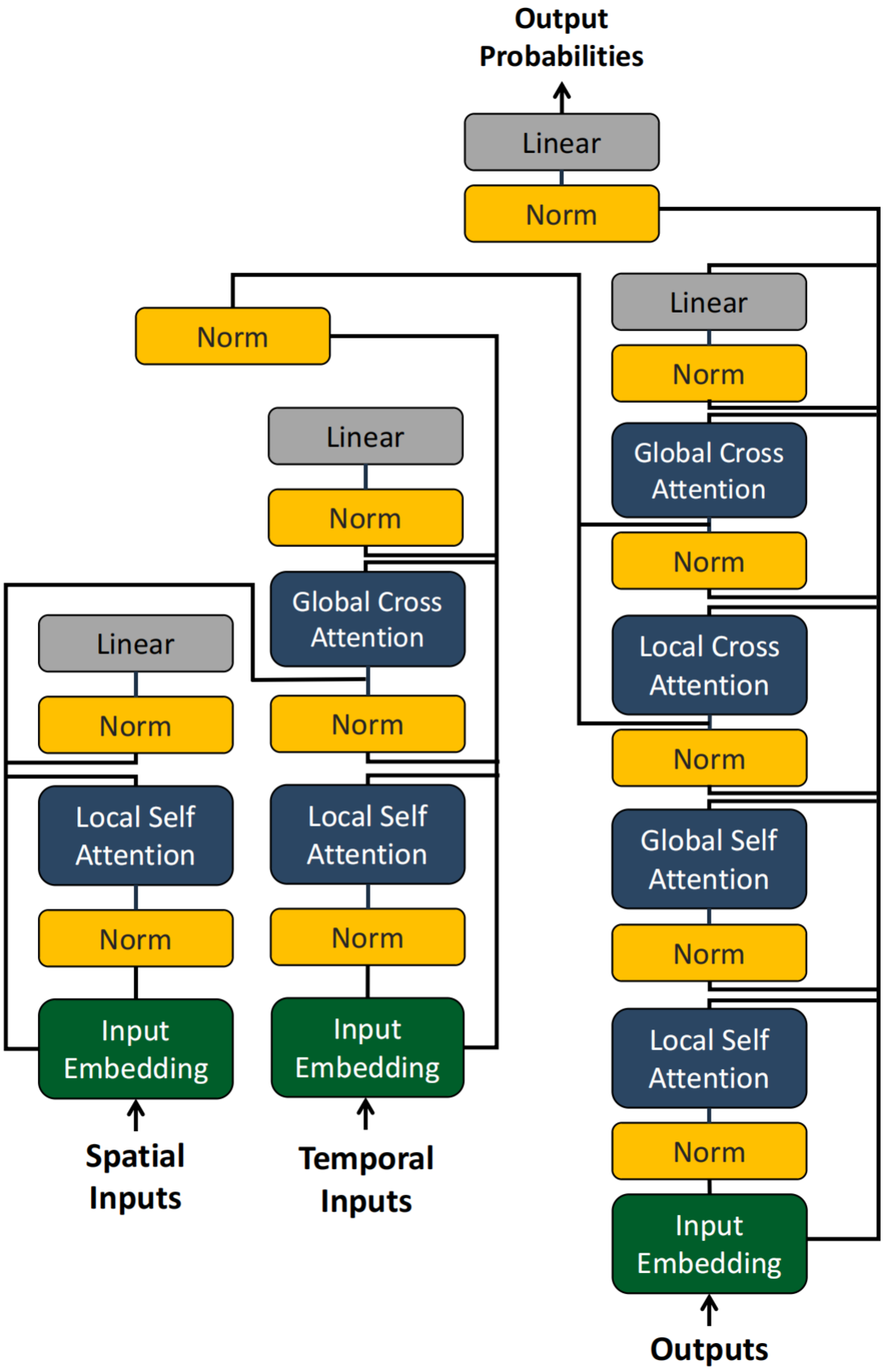


Figure 6.6: Architecture of Bloomformer-2

6.7 Prediction strategy

The prediction strategy can be divided into single-step prediction and multi-step prediction [76]. In single-step prediction, a 2-day time step (268 in summer and 264 in winter) is employed to predict the subsequent Chl-a data point. The prediction advances by sliding one step forward at a time. On the other hand, multi-step prediction utilizes the Seq2Seq prediction strategy [77], specifically designed for DL models. In this task node, 2-day time steps (268 in summer and 264 in winter) serve as input sequences to predict the following 3-day time steps (402 in summer and 396 in winter), which function as output sequences. Prior to incorporating all data into the model, data normalization is performed using the Z-score (Section 4.3.1).

6.8 Performance evaluation of model

Mean Absolute Error (MAE), Mean Squared Error (MSE), and MAPE (Section 4.3.3) are used to evaluate the performance of LSTM and Bloomformer-2. The Units of MAE and MSE are the same as the respective water quality parameter units, and the unit of MAPE is %. In addition, the 95% confidence interval of the predicted value for both DL modes was calculated, respectively.

MAE [45] was calculated as:

$$MAE(y, \hat{y}) = \frac{\sum_{i=1}^{n} |y_i - \widehat{y_i}|}{n}$$

MSE [78] was calculated as:

$$MSE(y, \hat{y}) = \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{n}$$

where $\widehat{y_i}$ is the predicted value of the ith sample, y_i is the corresponding true value of the total n samples, and $\overline{y_i}$ is the mean of true value.

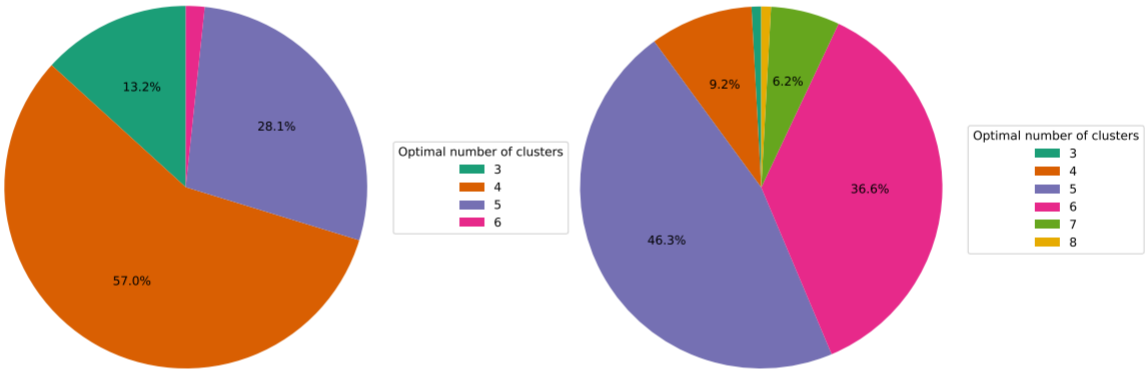
6.9 Computational Environment

The experiment was carried out on a PC with the following features: Hardware: CPU i7-6950X, RAM 64GB, dual GeForce RTX 3090, VRAM 24GB Software: Ubuntu 20.04, Python3.6, Pytorch 1.10.0, Numpy 19.2.

6.10 Results

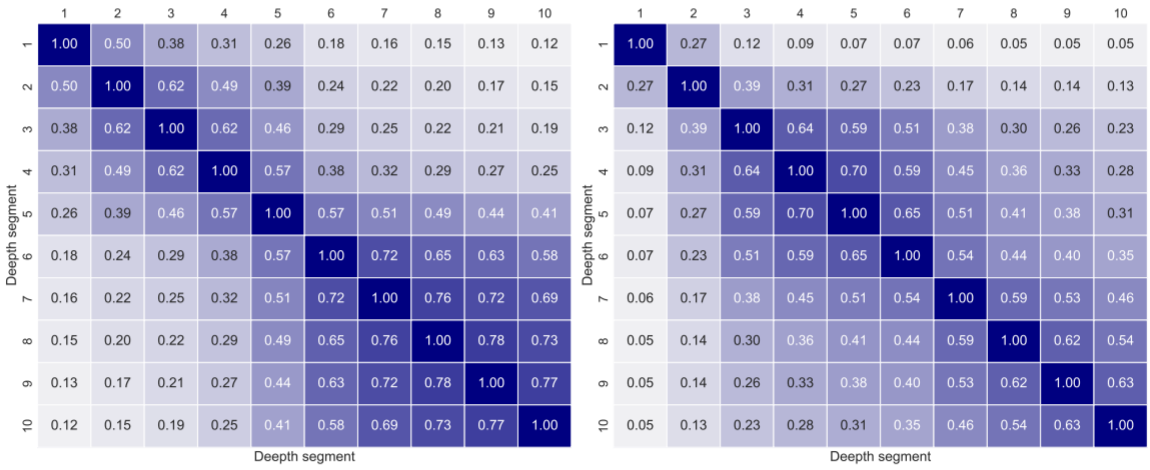
6.10.1 Result of water depth clustering

Each work cycle within the datasets pertaining to "Winter-2018" and "Summer-2019" was independently introduced into DeepDPM for deep clustering. The distribution of the optimal cluster number, as exhibited in Figure 6.7, reveals that the optimal cluster numbers are five during the four in 2018-winter (57.0%) and 2019-summer (46.2%). Partial adjacent matrices for winter and summer are delineated in Figure 6.8. These adjacent matrices underwent spectral clustering to refine the modeling strategy, the outcomes of which are cataloged in Table 6.3. In the summer of 2019, the aquatic depth within the target region was bifurcated into five factions, designated Group S1 through Group S5. Conversely, in the winter of 2018, the aquatic depth was segregated into four factions, marked as Group W1 through Group W4.



(2018-Winter) (2019-Summer)

Figure 6.7: Distribution of optimal cluster number



(2018-Winter) (2019-Summer)

Figure 6.8: Adjacency matrix of 2018-Winter and 2019-Summer (an example of 10 depth segment, a depth segment is 0.05m)

Table 6.3: Result of water depth clustering

Season	Water depth group				
Winter 2018	Group W1 0.05 - 0.1m	Group W2 0.1 - 0.3m	Group W3 0.3 - 0.95m	Group W4 0.95 - 1.2m	
Summer 2019	Group S1 0.1 - 0.15m	Group S2 0.15 - 0.4m	Group S3 0.4 - 0.95m	Group S4 0.95 - 1.55m	Group S5 1.55 - 1.95m

6.10.2 Model performance evaluation

The single-step prediction results for group S1 and W1, as examples, are depicted in Figure 6.9. The prediction errors of the two DL models in single-step prediction were calculated and presented in Table 6.4 (The bold-italic values represent the best performance). In comparison to the LSTM model, the prediction value curves of Bloomformer-2 for groups S1, S2, S4, W2, W3, and W4 exhibit a closer fit to the true value curves, accompanied by narrower confidence intervals. Moreover, the prediction error of Bloomformer-2 is smaller than that of LSTM. In the S3, S5, and W1 groups, Bloomformer-2's single-step prediction accuracy is comparable to that of LSTM.

Table 6.4: Errors of single-step prediction of Bloomformer-2 and LSTM

Water depth group	Model	MAE	MSE	MAPE
Group S1	Bloomformer	**0.254**	**0.305**	**2.279**
	LSTM	0.916	1.887	8.427
Group S2	Bloomformer	**0.394**	**0.246**	**2.108**
	LSTM	0.541	0.573	2.969
Group S3	Bloomformer	0.357	0.205	**0.733**
	LSTM	**0.309**	**0.154**	0.998
Group S4	Bloomformer	**0.288**	**0.142**	**0.848**
	LSTM	0.301	0.143	0.855
Group S5	Bloomformer	0.417	**0.249**	1.955
	LSTM	**0.373**	0.271	**1.162**
Group W1	Bloomformer	0.244	**0.072**	0.266
	LSTM	**0.159**	0.076	**0.191**
Group W2	Bloomformer	**0.213**	**0.056**	**0.269**
	LSTM	0.329	0.129	0.421
Group W3	Bloomformer	**0.201**	**0.052**	**0.247**
	LSTM	0.688	0.509	0.801
Group W4	Bloomformer	**0.175**	**0.042**	**0.228**
	LSTM	0.184	0.044	0.237

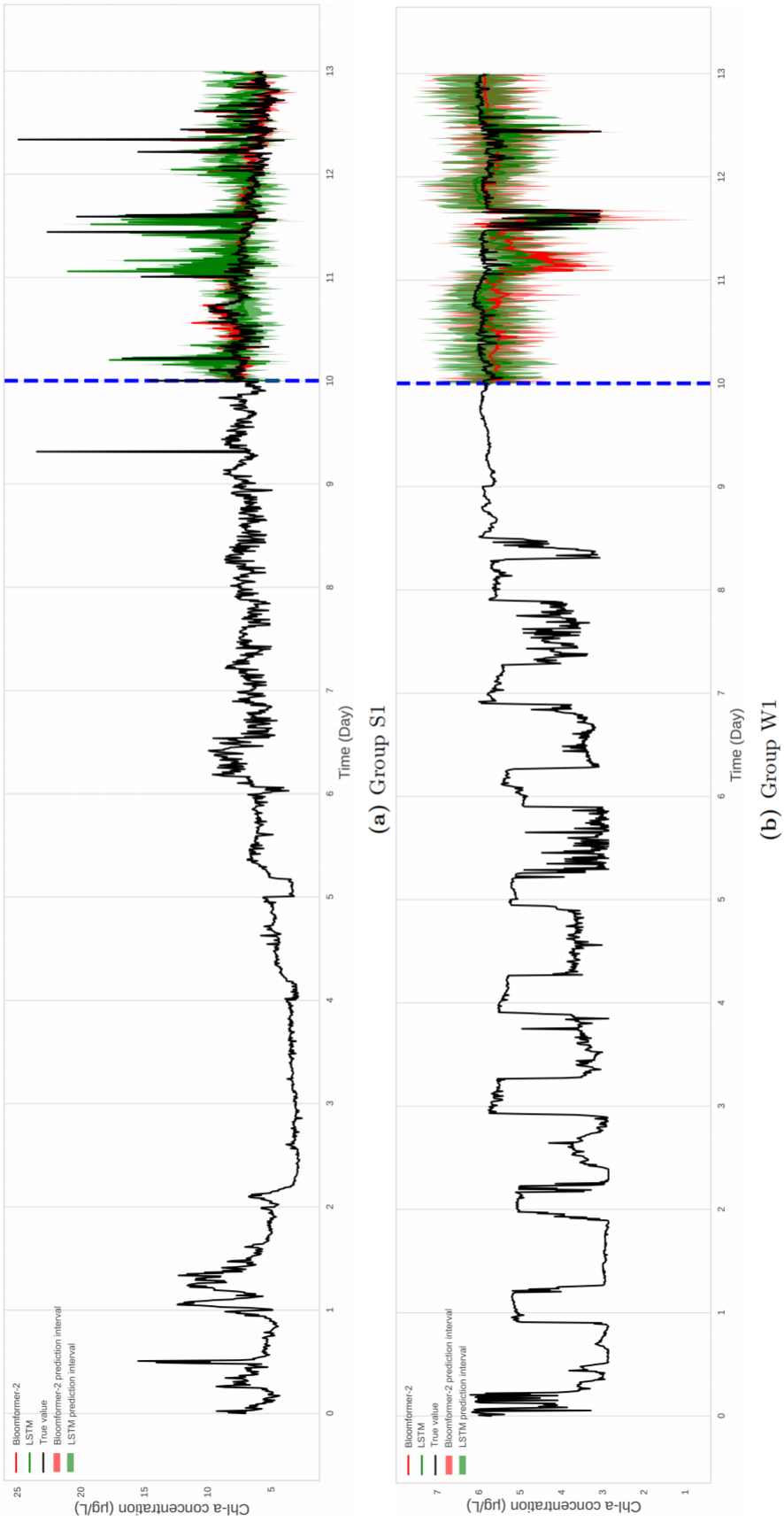


Figure 6.9: Comparison of model prediction in single-step prediction, (a) represents the result of Group S1 and (b) represents the result of Group W1

Multi-step predictions for day-11, 12, and 13 for each water depth group for both "2018-Winter" and "2019-Summer" were conducted using both models. The single-step prediction results for group S1 and W1, as examples, are displayed in Figure 6.10. The prediction errors of the two DL models in multi-step prediction are provided in Table 6.5 (The bold-italic values represent the best performance). Bloomformer-2 outperforms LSTM in multi-step prediction for all groups, as evidenced by the predicted value curves being closer to the true value curves and the smaller prediction errors.

Table 6.5: Errors of multi-step prediction of Bloomformer-2 and LSTM

Water depth group	Model	MAE	MSE	MAPE
Group S1	Bloomformer	**0.207**	**0.161**	**1.091**
	LSTM	0.613	1.086	5.264
Group S2	Bloomformer	**0.421**	**0.269**	**4.011**
	LSTM	0.474	0.361	4.034
Group S3	Bloomformer	**0.238**	**0.101**	**0.349**
	LSTM	0.526	0.473	2.629
Group S4	Bloomformer	**0.341**	**0.184**	**1.39**
	LSTM	0.549	0.508	2.418
Group S5	Bloomformer	**0.505**	**0.378**	**1.679**
	LSTM	0.512	0.402	3.748
Group W1	Bloomformer	**0.249**	**0.121**	**0.372**
	LSTM	0.339	0.337	0.621
Group W2	Bloomformer	**0.184**	**0.105**	**0.492**
	LSTM	0.353	0.283	0.799
Group W3	Bloomformer	**0.188**	**0.068**	**0.243**
	LSTM	0.291	0.301	0.352
Group W4	Bloomformer	**0.361**	**0.167**	**0.558**
	LSTM	0.397	0.307	0.603

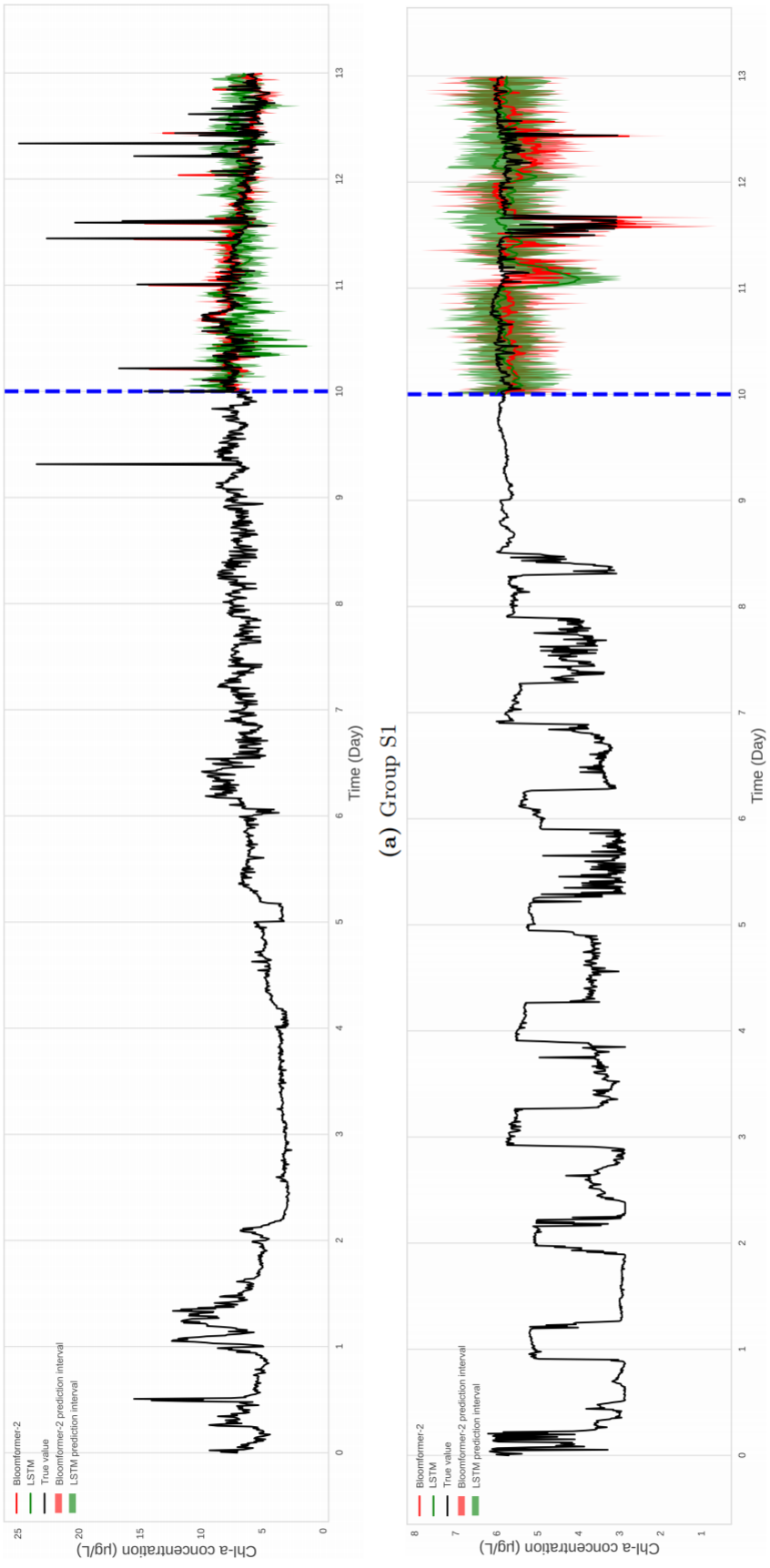


Figure 6.10: Comparison of model prediction in multi-step prediction, (a) represents the result of Group S1 and (b) represents the result of Group W1

6.10.3 Driving factors for the predicted value

The comprehensive driving factors of 11^{th} to 13^{th} prediction for Group W1 and S1 are shown in Figure 6.11 and Figure 6.13, respectively. And the driving factor of the predicted value for all work cycles for Group W1 and S1 on the 11th day is shown in Figure 6.12 and Figure 6.14, respectively.

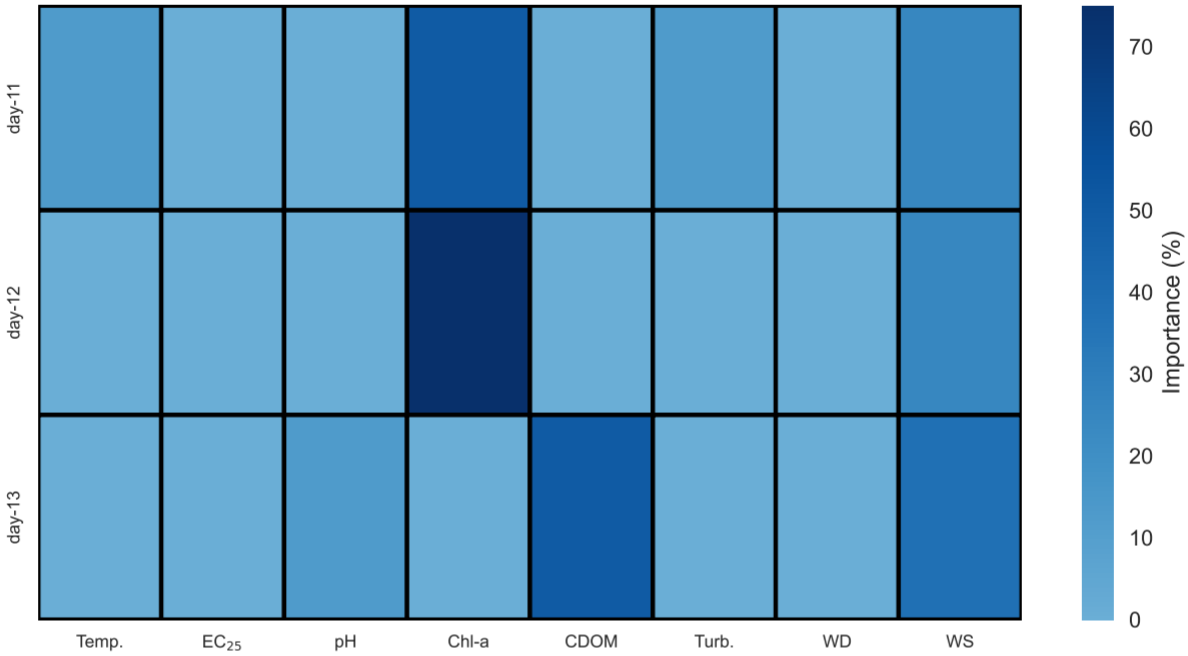
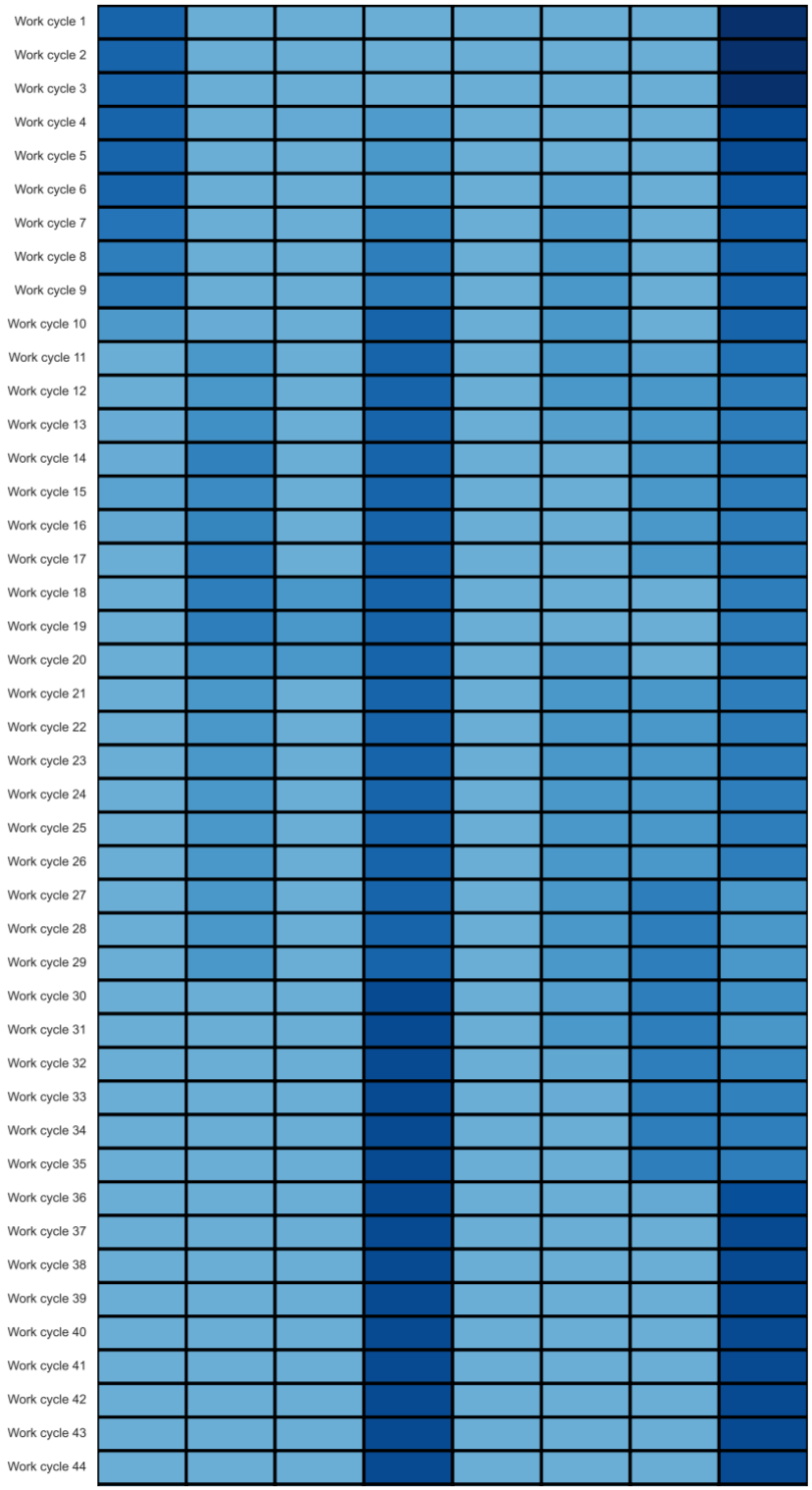
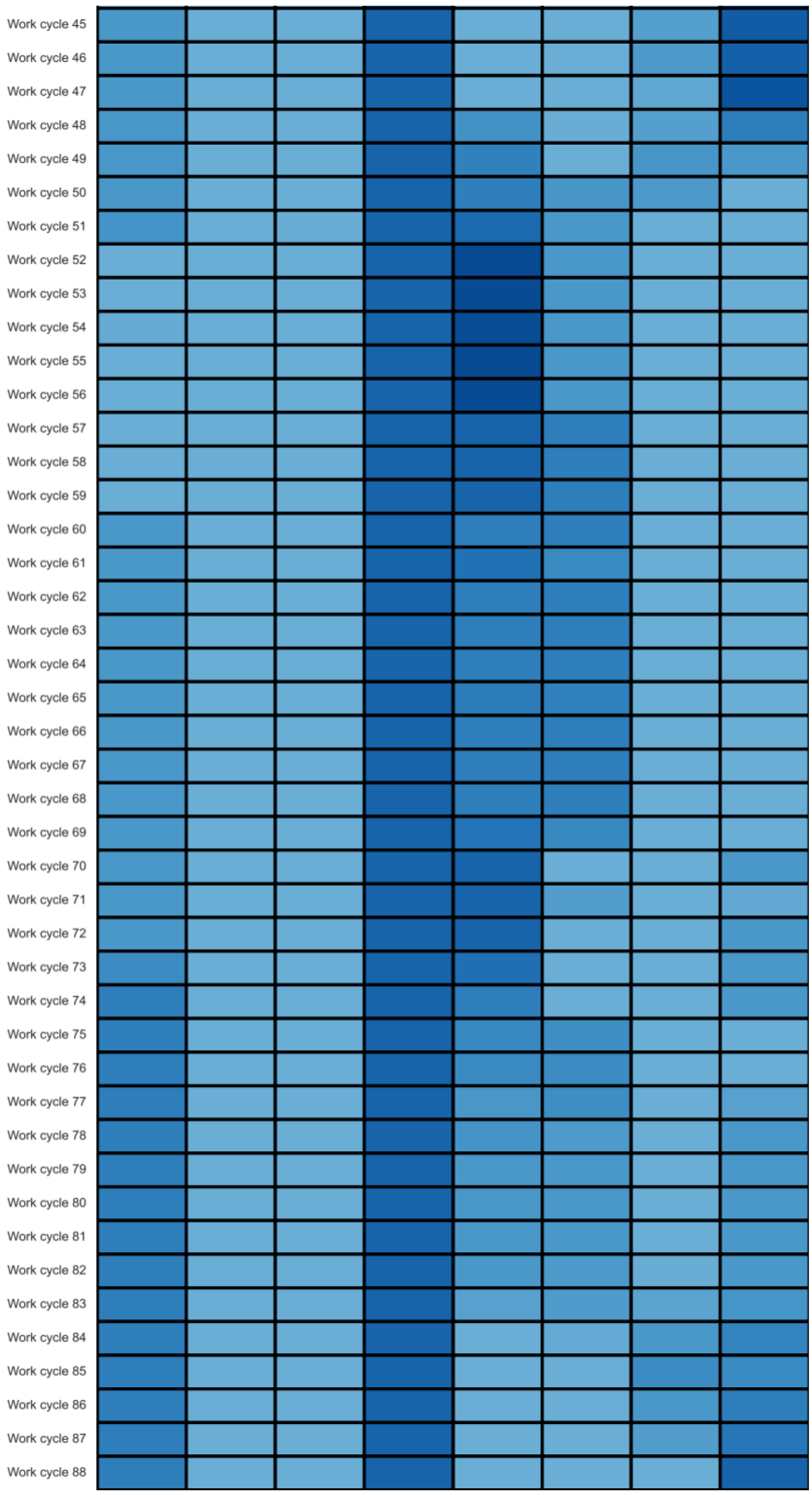


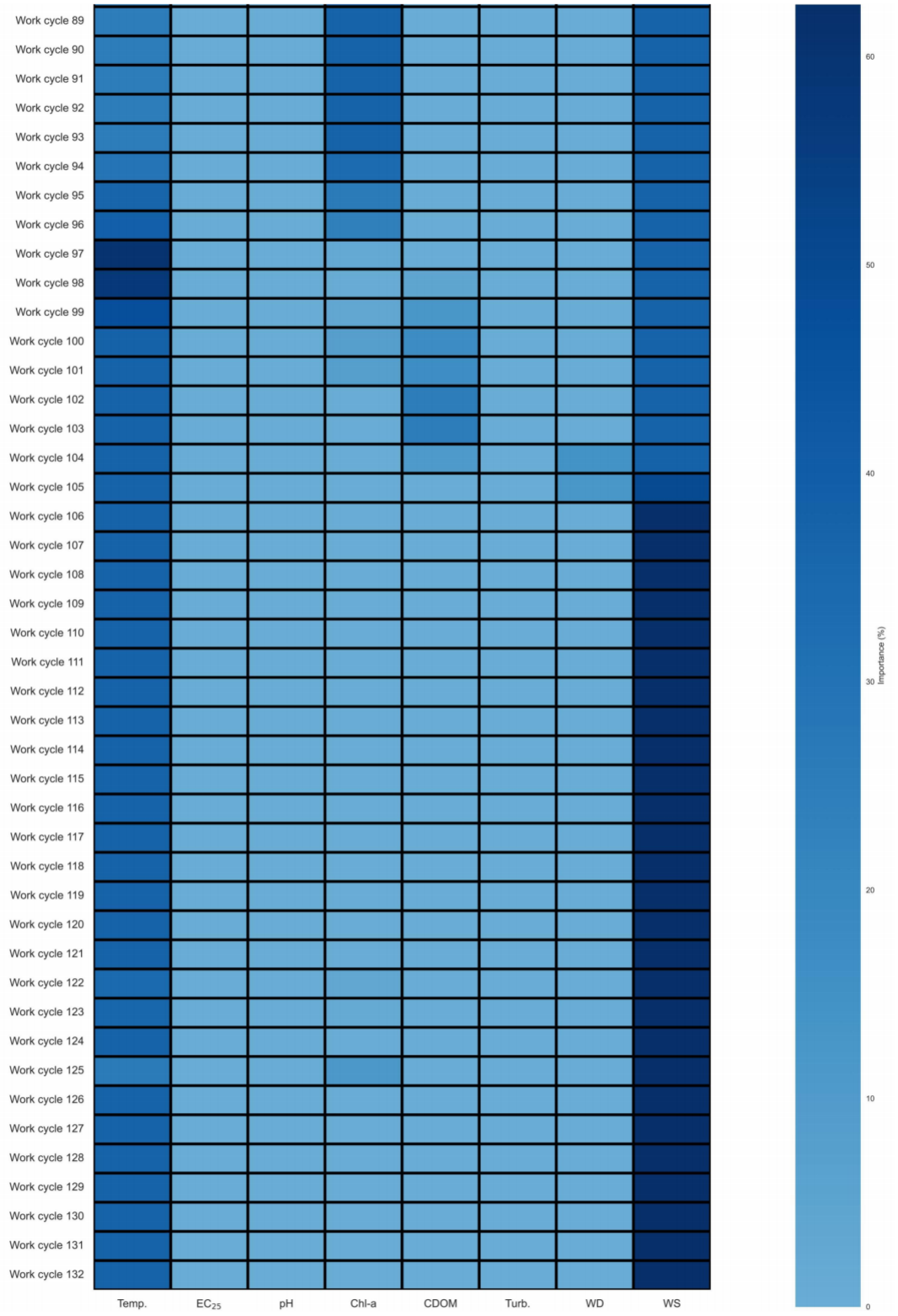
Figure 6.11: Driving factor of 11^{th} to 13^{th} day prediction for Group W1



(Work cycle 1 - Work cycle 44)



(Work cycle 45 - Work cycle 88)



(Work cycle 89 - Work cycle 132)

Figure 6.12: Driving factor of predicted value for all work cycles for Group W1 on 11^{th} day

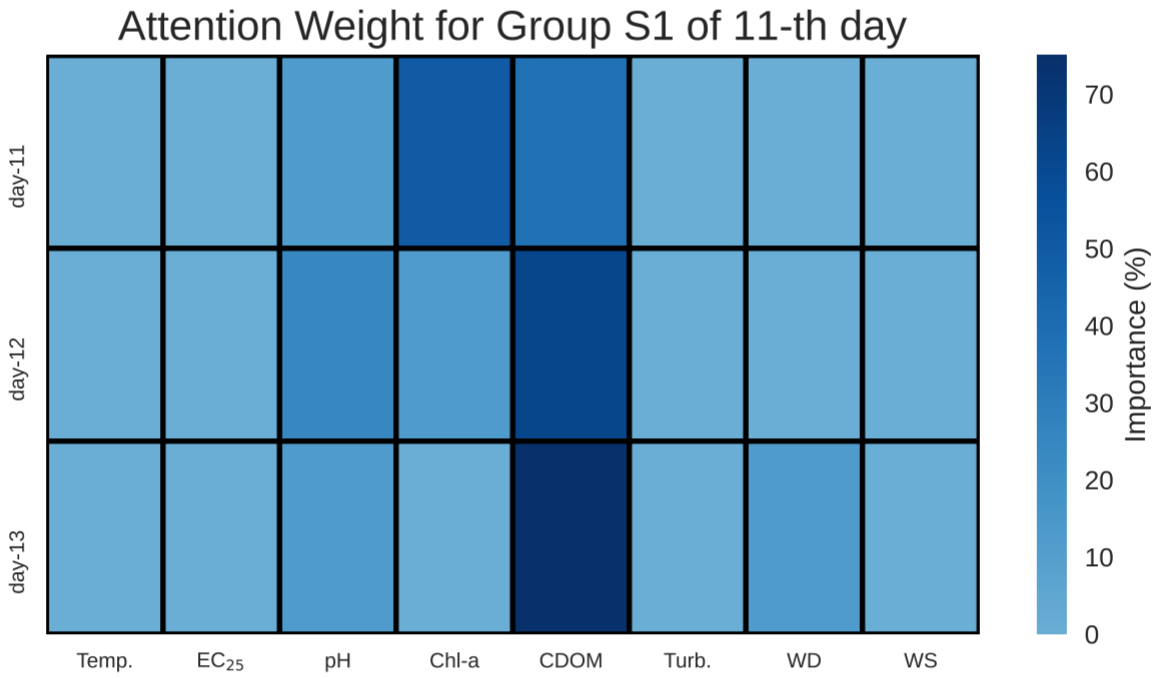
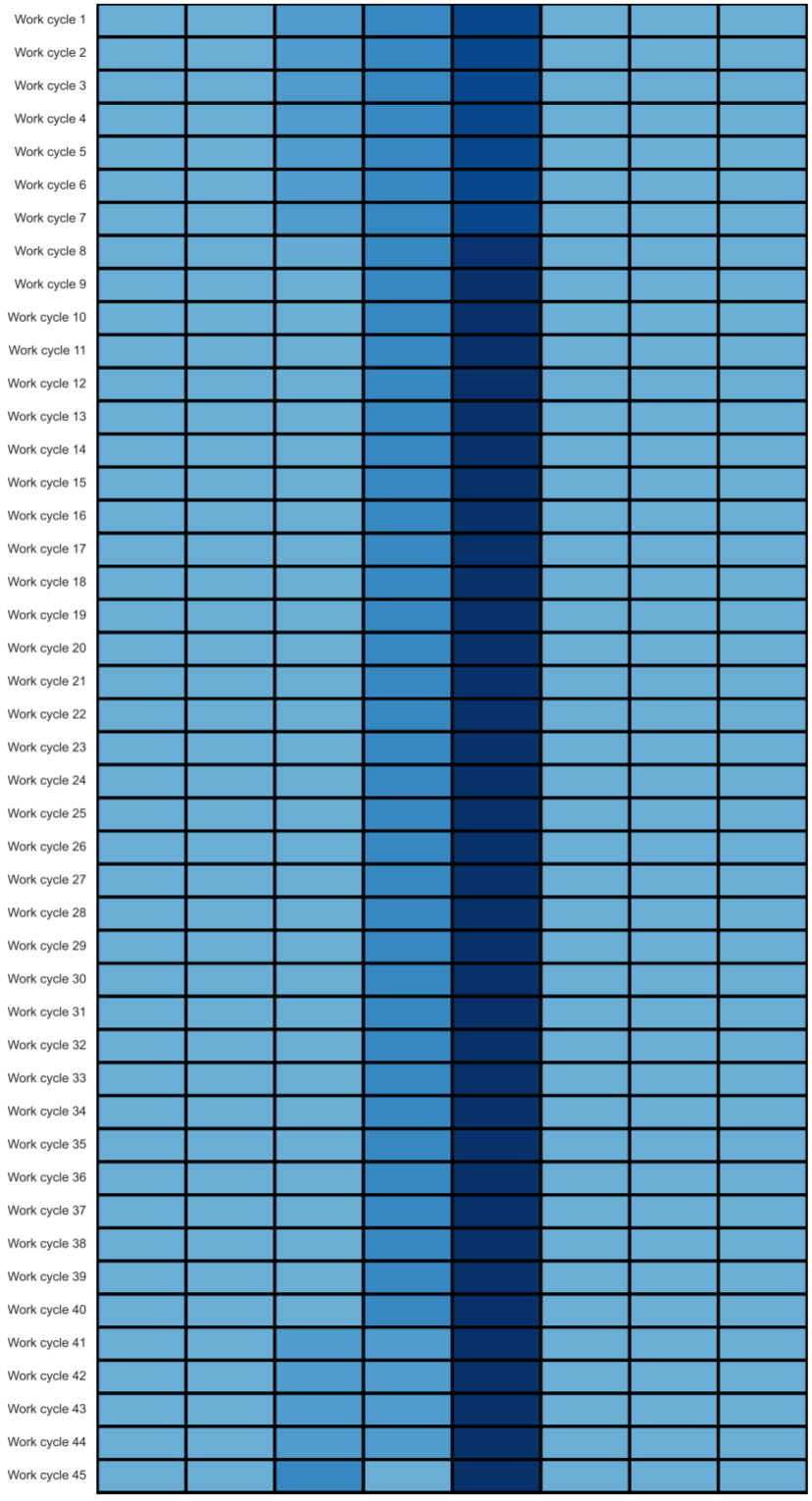
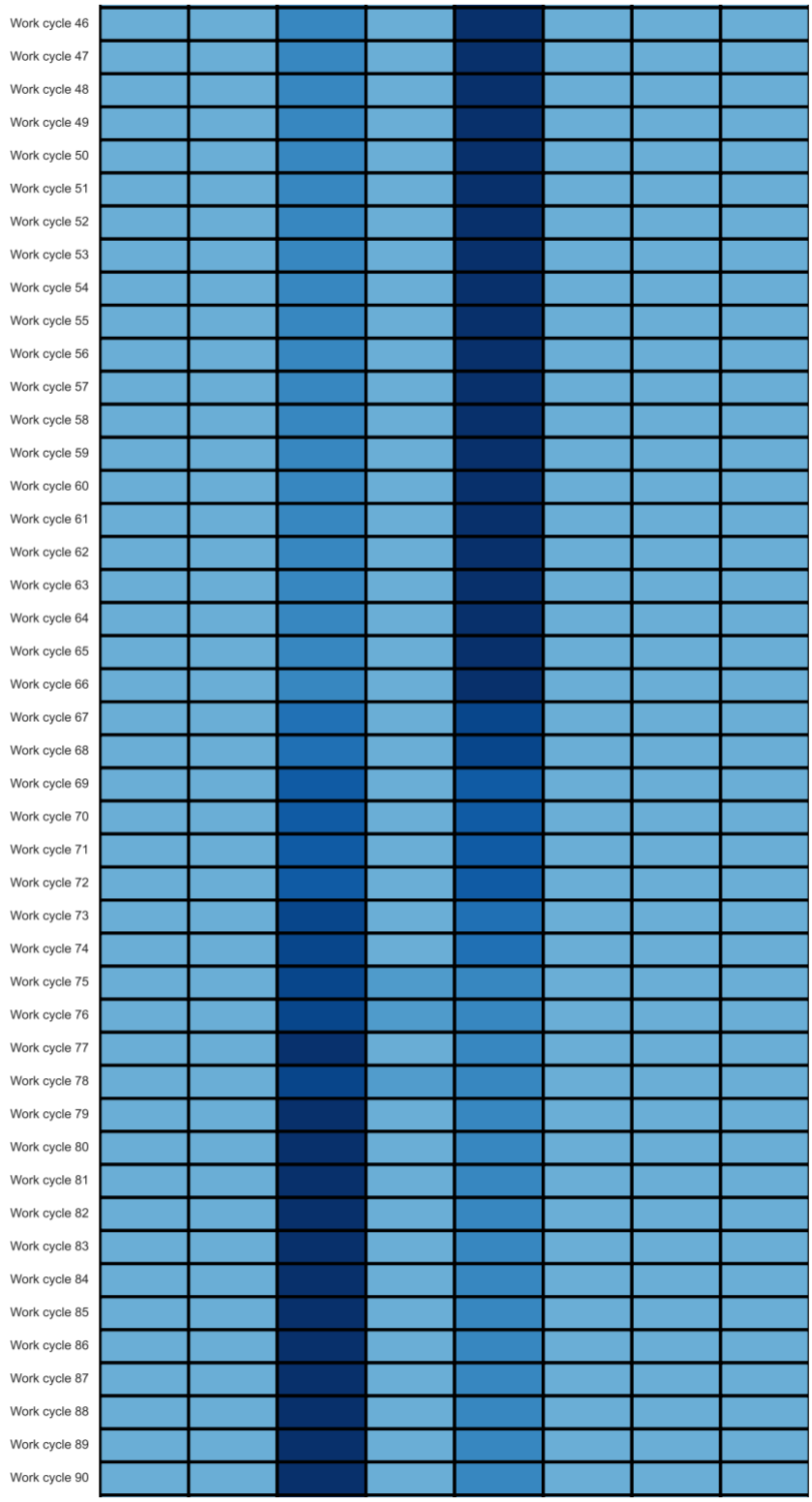


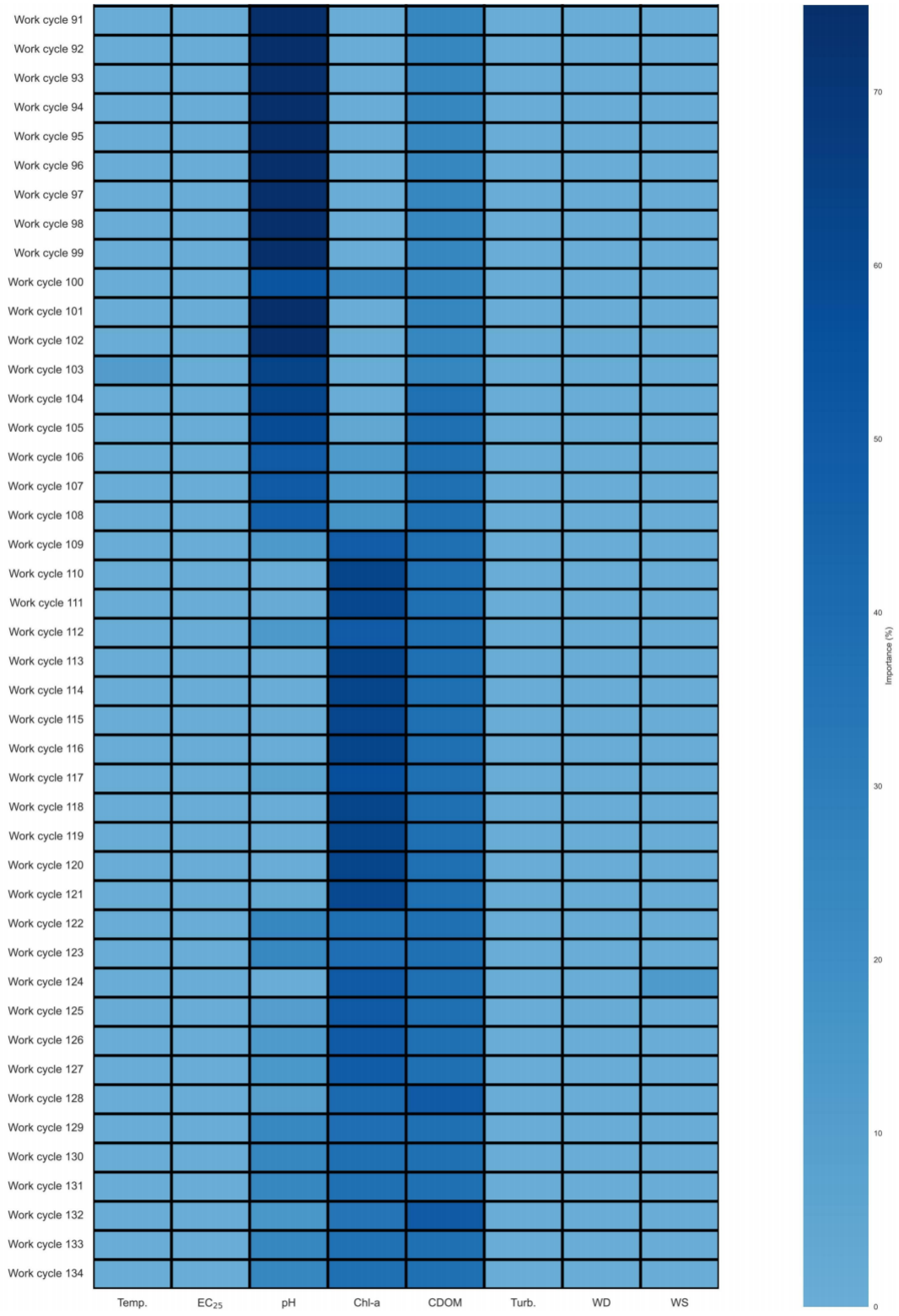
Figure 6.13: Driving factor of 11^{th} to 13^{th} day prediction for Group S1



(Work cycle 1 - Work cycle 45)



(Work cycle 46 - Work cycle 90)



(Work cycle 90 - Work cycle 134)

Figure 6.14: Driving factor of predicted value for all work cycles for Group S1 on 11^{th} day

In alignment with the World Health Organization's "Alert Level Framework," two Chl-a benchmarks (1 µg/L and 12 µg/L) are deployed to determine potential conditions fostering algal bloom eruptions [79]. The former threshold resonates with "Alert Level I," symbolizing the embryonic phase of HABs, while the latter corresponds to "Alert Level II," indicative of acute HABs.

Upon examination of predicted metrics spanning three days for Group W1, Chl-a concentrations consistently inhabited the interval between 1 µg/L and 12 µg/L, insinuating a persistent "Alert Level I" state across the experimental zones. Comprehensive mitigation strategies can thus be proposed for each day. Delving into the multi-step prognostication results on the 11^{th} day, Chl-a and wind velocity surfaced as paramount driving factors. The 12^{th} day elevated the importance of wind speed and Chl-a, while the 13^{th} day highlighted CDOM and wind velocity. Antecedent research posits that wind speed can trigger bottom sediment upheaval in shallow reservoirs such as Taihu Lake, releasing nutrients that stimulate algal proliferation [60]. Simultaneously, this resuspension process can significantly escalate turbidity. However, model outcomes illustrate that turbidity's influence is marginal. Hence, the accentuated weight of wind speed is likely attributed to its role in coalescing algae from surrounding regions at the target locale. The subsequent algal demise amplifies the CDOM content within the water strata. Subject to photochemical decay, CDOM can transition from macro-molecular organic matter to micro-molecular organic matter and inorganic nutrients, thereby nurturing conditions favorable for algal proliferation [80]. In summation, for Group W1, the prevention of algal confluence within the area forms a crucial cornerstone in devising preventative measures, such as the placement of algae interception barriers around the area's outskirts.

With respect to the predicted values for the triad of days concerning Group S1, nine temporal instances succumbed to Alert Level II, while the remainder maintained Alert Level I. Initially, exhaustive preventative measures for each day can be proposed, drawing from multi-step prediction results. Predominant driving factors on the 11^{th} day were identified as Chl-a and CDOM. The 12^{th} day underscored Chl-a and CDOM, while the 13^{th} day spotlighted CDOM and pH. Contrasting the winter season, neither wind speed nor direction emerged as primary influencers in summer, suggesting algal genesis primarily from localized growth. Additionally, factors such as algal mortality and surface runoff escalated CDOM and pH within the water strata at the experimental sites, fostering algal proliferation. As a result, the extraction of existing algae at the experimental sites, coupled with pH adjustment, may serve as an efficacious countermeasure against HABs outbreaks of Alert Level I, such as the employment of acidic algaecides and manual salvage.

6.11 Conclusion of algal growth early warning task node

In the face of escalating ecological threats and economic damages posed by harmful algal blooms (HABs) in freshwater environments, a novel approach has been developed that leverages the power of big data and deep learning models. This intelligent early warning system for HABs represents a significant stride forward in managing these phenomena. Data is gathered

through a Vertical Aquatic Monitoring System (VAMS), which, when combined with the "DeepDPM-Spectral Clustering" methodology, provides a detailed analysis of vertical algal distribution. This approach streamlines the number of predictive models and enhances the system's adaptability. The system utilizes the Bloomformer-2 model developed by me to perform both single-step and multi-step predictions of HABs. Our case study validates the superior performance of Bloomformer-2, demonstrating high alignment with actual value curves and a reduced margin of predictive error. One of the system's unique features is its ability to identify the driving factors of HABs, which aids in the development of targeted preventive measures. The model's impressive intelligence - its ability to autonomously learn from preprocessed data - and inherent adaptability set the stage for future system enhancements and wider applications.

7 Synoptic Discussion

The preceding chapter delineates the formulation of a comprehensive research framework anchored on big data, focused on water environment management, and marks the accomplishment of four exemplary task nodes at this juncture. These encompass:

1) Water quality monitoring and assessment task node: The EBDP for this task node is constructed from measurements procured by RST and CMT. The DNN is employed to apprehend the intricate relationships between remote sensing information and water quality parameters, thereby fulfilling the task node of water quality monitoring. Subsequent to this, IDEC is utilized to conduct a profound clustering of the water quality monitoring results, culminating in the accomplishment of the water quality assessment task node.

2) Algal growth driving factor identification task node: The EBDP for this task node is meticulously constructed from high-frequency, prolonged manual field sampling. Bloomformer-1 has been developed to precisely apprehend the spatial association information between Chl-a and other water quality parameters, thereby accomplishing the task node of identifying the driving factors of algal growth.

3) Algal growth early warning task node: The EBDP for this task node is constructed from the measurements gathered by VAMS (BIOLIFT). Bloomformer-2 has been developed to capture, with precision and efficiency, the spatio-temporal information in lengthy time series encompassing Chl-a, water quality parameters, and meteorological parameters, thereby facilitating the development of an algal growth early warning system.

This part proceeds to examine the merits of big data in the realm of WEM, the amalgamation of industry clusters for WEM big data, and the interpretability of the models.

7.1 Potentials of big data in WEM

As demonstrated by the fulfillment of task nodes, big data assumes a pivotal function in tackling the complexities of WEM, providing an array of benefits that contribute to optimizing

resource utilization, enhancing efficiency, and facilitating sustainable decision-making. The subsequent discourse delineates the merits of big data within the context of WEM.

First and foremost, big data exhibits considerable adaptability in addressing nodal tasks. In the water quality monitoring task, comprehensive water quality monitoring of a large reservoir (approximately 70 km^2) must be accomplished within a short time frame (5 hours) by a limited number of individuals (4 people). In the water quality assessment task, a multi-factor comprehensive evaluation of the water quality of large reservoirs is required, along with the accurate delineation of the joint control area. For the algae growth driver identification task, it is imperative to precisely identify the algae growth drivers without conducting preliminary experiments. In the algal growth warning task, accurate predictions of algal growth with high temporal resolution and depth-based information are necessary. The outcomes of each task node reveal that the big data research approach is capable of accomplishing these tasks, whereas the small data research approach is insufficient. For instance, at the current stage, the hypothesis-driven small data approach can only resolve the remote sensing inversion of water quality parameters with optical properties.

Another conspicuous benefit is the remarkable accuracy demonstrated by the models. Performance evaluations of various nodal tasks reveal that the results derived from the big data approach exhibit high precision. For instance, the remote sensing inversion accuracy of BP attains a value of 0.95.

Furthermore, big data offers the advantage of comprehensiveness. On one hand, owing to the attributes of EBDP and the high performance of value mining tools, big data can amalgamate and analyze information from multiple sources, such as satellite imagery, meteorological data, and water quality parameters, as evidenced in the task nodes. On the other hand, the data within EBDP represents the entire system rather than an isolated process, implying that the value extracted from EBDP also constitutes an integrated value interpretation of the whole system. This advantage can assist water managers in better comprehending the factors influencing water availability, quality, and distribution, and in devising more efficacious strategies to manage these resources.

Lastly, big data presents the advantage of fine granularity. Fine granularity entails collecting, storing, and analyzing data at a more detailed and intricate level, capturing nuanced fluctuations and patterns that may elude detection in coarser granular datasets. The automated equipment employed in the task node undoubtedly enhances the detail of data collection and storage, such as BIOFISH enabling the high-density collection of water quality data with GIS (Geographic Information Systems) labels, and BIOLIFT enabling high temporal resolution collection of water quality parameters with depth labels. This collection methodology can be characterized as a comprehensive process, mitigating the loss of crucial information to a certain extent. Concurrently, deep learning can perform deep data mining from these fine-granular datasets. This fine granularity advantage empowers decision-makers to base their choices on more detailed information and a deeper comprehension of various phenomena, thereby formulating superior and more targeted strategies, policies, or interventions.

7.2 Industry cluster for WEM big data

Industrial clusters pertain to an economic notion characterized by geographically concentrated enterprises, manufacturers in interconnected industries, and affiliated institutions within a specific region, which maintain competitive and cooperative relationships while exhibiting interdependence [81]. The genesis of such clusters is frequently ascribed to a multitude of factors, encompassing access to skilled labor, raw materials, specialized infrastructure, or proximity to markets. Industrial clusters give rise to cluster effects, encompassing resource agglomeration, synergy and spillover, as well as the division of labor effects, ultimately conferring benefits upon associated businesses and industries [82]. Notable exemplars of industrial clusters encompass Silicon Valley in the United States (technology and innovation), the automotive cluster in Stuttgart, Germany, and the fashion and design cluster in Milan, Italy [83].

From an economic standpoint, my doctoral research endeavors to devise a comprehensive framework for WEM Big Data, which can be conceptualized as a WEM Big Data industrial cluster. Within this framework, each task node assumes a distinct function in the WEM industry. For example, the water quality monitoring task node operates as an upstream raw material supplier, whereas the algal driving factor identification task node functions as a service provider, accountable for the facilitation of pertinent services. The inherent attributes of big data provide the basis for the establishment of a competitive and cooperative network amongst the nodes. Analogous to the cluster effect observed in the emergence of industrial clusters in the realm of economics [84], the WEM Big Data industrial cluster likewise engenders a cluster effect. The key term for the cluster effect, engendered by the successful completion of the four task nodes within my doctoral research, is precision.

Precision encompasses two dimensions: spatial and methodological, which pertain to the loci of intervention and the specific actions to be undertaken. The water quality monitoring task node carries out comprehensive evaluations of extensive water bodies. Initial analyses of the monitoring outcomes pinpoint relatively expansive risk areas and corresponding remedial strategies, such as the abrupt decline in dissolved oxygen (DO) at the tail end of the Qingcaosha Reservoir. Successive water quality assessment task nodes, employing big data techniques, partition the entire water body into multiple joint management and control zones. The lucid demarcation of each zone's boundaries augments the precision of risk area identification. Concurrently, the application of the driving factor identification task node to each joint management and control zone accurately ascertains the characteristic factors of every region, thus establishing a foundation for subsequent management interventions. The spatiotemporal dynamics of water quality monitoring and assessment outcomes can be efficiently acquired by performing water quality monitoring and assessment within a singular water body over a protracted duration. An examination of these spatiotemporal dynamics unveils specific regions within the reservoir that fall within different joint management and control zones at different times at varying time intervals, thereby designating them as fluctuating areas (typically situated at the peripheries of the joint management and control zones). These areas necessitate heightened vigilance. The placement of the BIOLIFT from the algal growth

warning task node within these areas furnishes a more precise foundation for management and control measures.

This transition from encompassing the entire water body to focusing on joint management and control zones, and ultimately pinpointing the fluctuating areas, epitomizes precision in the spatial dimension. Simultaneously, the progression from indistinct surface conditions to well-defined surface conditions, and subsequently to exact underwater conditions, embodies precision in the methodological dimension (Figure 7.1).

7.3 Model interpretability

As delineated in Section 1.1, it is posited that in the realm of WEM research, the pursuit of correlation and causality within the subject matter should be amalgamated. The sequential approach ought to initially focus on identifying correlation, followed by the progression from correlation to causality, facilitated by relevant prior knowledge. Model interpretability serves as a critical juncture in the process of deducing causality from correlation and can also be regarded as the embodiment of causality in the search for correlation.

Model interpretability refers to the degree to which a model's internal workings, predictions, and decision-making processes can be understood and explained by humans [85]. It directly determines whether the model is trustworthy. The interpretability of a model can be represented by transferability and understandability [86]. Humans possess a crucial ability to induce and transfer skills across various fields, and models should also be able to operate in such environments, such as when conditions are less stable. This is the transferability or generalizability of the model. Understandability represents the extent to which we comprehend the workings of the model, also known as transparency. A model lacking transparency in the decision-making stage is a black box model, whereas a transparent model is considered a white box model. Figure 7.2 shows the levels of transparency of some models [87]. It is evident that using highly interpretable models to seek correlation will subsequently allow us to obtain some causal clues by analyzing the model's internal workings, thus facilitating the inference from correlation to causality.

A pivotal aspect of big data methodologies pertains to the selection of value mining methods or models [88]. To procure more precise correlations, it is imperative to opt for models with superior performance. Concurrently, to facilitate the inference of causality from correlation, it is essential to select models with high interpretability. Presently, there exists a tension between the performance and interpretability of models. As illustrated in Figure 7.3, enhanced model performance typically implies more intricate algorithms, which may consequently result in diminished interpretability. Therefore, model selection necessitates striking an equilibrium between performance and interpretability. The task nodes encompassed within the WEM big data framework are complex and intimately connected to practical applications, rendering this balance even more crucial. In contrast to DNN employed in the water quality monitoring task node, the development and implementation of Bloomformer-1 and Bloomformer-2 in the subsequent two task nodes have thoroughly considered both dimensions of model performance and interpretability, enabling superior performance while concurrently

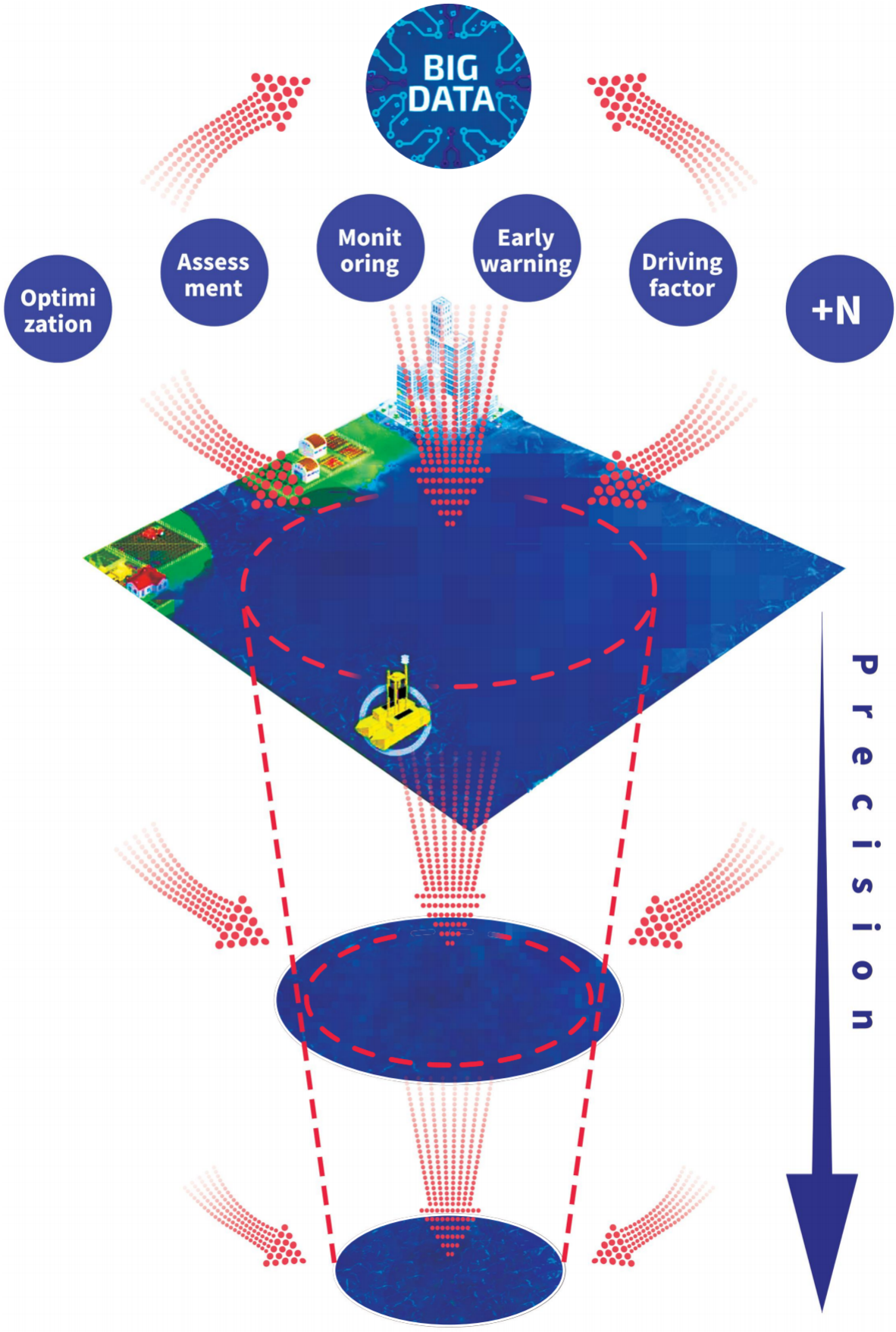


Figure 7.1: Precise management and control in spatial (Targeting of specific risk areas) and methodological (Coarse to fine granularity) dimensions

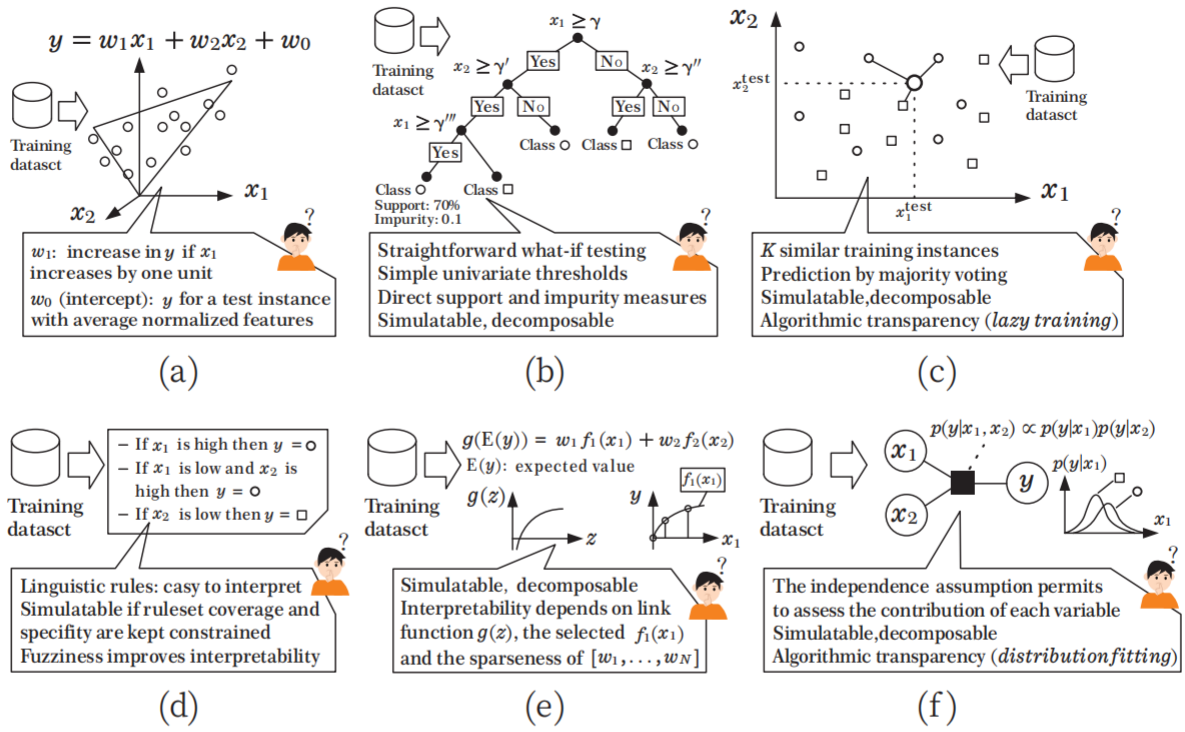


Figure 7.2: Levels of transparency of some models : (a) Linear regression; (b) Decision trees; (c) K-Nearest Neighbors; (d) Rule-based Learners; (e) Generalized Additive Models; (f) Bayesian Models. This figure comes from [87].

accounting for interpretability to a certain degree. Given the burgeoning popularity of model interpretability research and several technological advancements in recent years, the task nodes incorporated into the framework at a later stage will more readily grasp or even transcend this balance, thereby designing models with both high interpretability and exceptional performance.

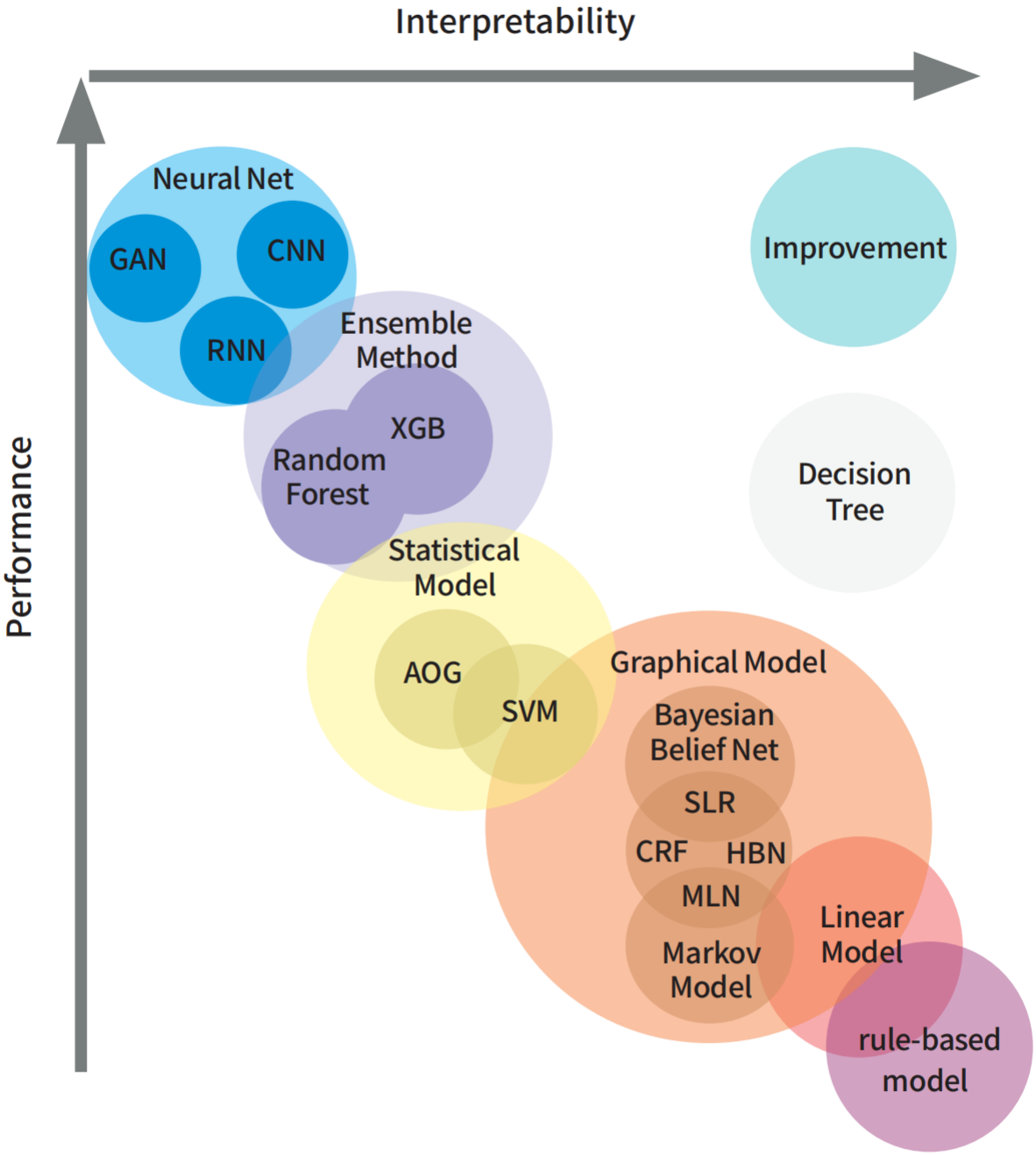


Figure 7.3: Model interpretability vs. model performance for some widely used models: HBN: Hierarchical Bayesian Networks; SLR: Simple Linear Regression; CRF: Conditional Random Fields; MLN: Markov Logic Network; SVM: Support Vector Machine; AOG: Stochastic And-Or-Graphs; XGB: XGBoost; CNN: Convolutional Neural Network; and GAN: Generative Adversarial Network. This figure comes from [89].

8 Conclusion and Outlook

8.1 Conclusion

WEM is a critical aspect of sustainable development. With the advancement of data science, big data has emerged as a transformative force in WEM, offering significant potential for addressing various challenges related to water resources. In this doctoral thesis, a big data framework in WEM was constructed, and four task nodes were finished: water quality monitoring, water quality assessment, identification of algal growth driving factors, and algal growth early warning.

The research area for the water quality monitoring task node is the Qingcaosha Reservoir. Its Environmental Big Data Platform is built using data collected by satellite remote sensing (Sentinel-2) and cruise monitoring devices (BIOFISH). A Deep Neural Network (DNN) is selected as the value mining tool. Results demonstrate that DNN performs exceptionally well in monitoring four water quality parameters (pH, DO, El.cond, and BP), accurately estimating the reservoir's overall water quality situation (R^2 values ranging from 0.77 to 0.95 for the different parameters).

The research area for the water quality assessment task node is also the Qingcaosha Reservoir. The Environmental Big Data Platform employed here consists of the results from the water quality monitoring task node. The Improved Deep Embedding Clustering (IDEC) is chosen as the value mining tool. Results indicate that the entire Qingcaosha Reservoir is distinctly divided into four joint management zones, with each zone's characteristic factors determined through statistical methods, providing the basis for formulating regional joint management strategies.

The research area for the identification of algal growth driving factors task node is the middle route of South-to-North Water Diversion Project. The EBDP used encompasses high-frequency, four-year manual sampling data. The value mining tool is Bloomformer-1, which I developed based on the Transformer core structure. Without extensive prior knowledge

and preliminary experiments, Bloomformer-1 achieves the highest R^2 (0.80 to 0.94) and the lowest RMSE (0.22 to 0.43 μ g/L) in both single sub-site and full-line simulations of MRP. Furthermore, Bloomformer-1 exhibits improved interpretability, ensuring reliability and direct applicability to real-world scenarios. The application of Bloomformer-1 in MRP indicates that total phosphorus (TP) is the most critical driving factor for MRP. Consequently, controlling and reducing phosphorus levels are essential strategies for managing algal growth and maintaining MRP water quality stability.

The research area for the algal growth early warning task node is Lake Taihu. The EBDP employed here is constructed from data collected by a vertical water quality monitoring system (BIOLIFT). Value mining tools include DeepDPM, Spectral clustering, and Bloomformer-2, which I developed based on the Transformer core structure. The combined use of DeepDPM and spectral clustering groups depth segments into several reasonable clusters, optimizing system efficiency through modeling strategies for each group rather than individual depth segments. Results show four depth groups in winter 2018 and five depth groups in summer 2019. Bloomformer-2 demonstrates outstanding performance in both single-step and multi-step predictions for all depth combinations. Moreover, like Bloomformer-1, Bloomformer-2 also exhibits enhanced interpretability, ensuring reliability and direct applicability to real-world scenarios.

In summary, the successful completion of the four example task nodes highlights the numerous advantages that big data offers in WEM, including but not limited to high adaptability, accuracy, comprehensiveness, and fine granularity within the task nodes. Moreover, the big data WEM framework established in this doctoral research forms an industry cluster exhibiting the characteristics typically associated with such clusters. The key theme emerging from the cluster effect generated by the four example task nodes is precision, which enables the development of accurate strategies and responsive measures for watershed water quality management. The implementation of the big data WEM framework exemplifies the potential of employing data-driven approaches to tackle complex water environment management challenges. By incorporating precision in both regional and methodological aspects, this research advances our understanding of water resources management and offers practical solutions for sustainable water environment management. As future research builds upon this framework and integrates additional task nodes, the benefits and effectiveness of the big data WEM approach are expected to further increase, ultimately contributing to the overarching goal of sustainable development and responsible water resource management.

8.2 Informatization of the WEM industry cluster – WEM foundation model

From the perspectives of environmental science, economics, and computer science, this doctoral thesis essentially digitalizes the WEM industry cluster through big data. This digital transformation retains the advantages of traditional industry clusters while also mitigating their geographical limitations to some extent. To achieve further clustering effects, the framework's

scale must be expanded with the addition of more task nodes, such as flood and drought prediction and mitigation, as well as the optimization of water infrastructure.

The rapid growth in scale introduces more diverse data formats and larger data capacities, resulting in fragmented and diverse value mining scenarios. The development cost of value mining tools, i.e., AI models, is extremely high, necessitating a process of development, parameter tuning, optimization, iteration, and application. Therefore, under these circumstances, the approach of customizing models for corresponding task nodes, as demonstrated in this doctoral thesis, also known as the workshop mode, proves challenging. To facilitate the transition from the workshop mode to the factory mode, foundation models offer a feasible solution, specifically the "pre-trained large model + downstream task fine-tuning" approach. Large-scale pre-training can effectively capture knowledge from vast amounts of labeled and unlabeled data, storing the knowledge in numerous parameters and fine-tuning the model for specific tasks to enable it to cope with multiple scenarios adequately. Moreover, designing foundation models for multi-modal, multi-task purposes is crucial to accommodate diverse data sources.

In summary, expanding the scale of the WEM big data framework and establishing a cross-scenario, multi-task, and multi-modal WEM foundation model based on this framework is the ultimate solution for addressing future WEM-related challenges.

reference

[1] William J. Cosgrove and Daniel P. Loucks. *Water management: Current and future challenges and research directions*. June 2015. DOI: 10.1002/2014WR016869.

[2] Judea Pearl et al. "Models, reasoning and inference". In: *Cambridge, UK: CambridgeUniversityPress* 19.2 (2000).

[3] Arika Virapongse et al. *A social-ecological systems approach for environmental management*. Aug. 2016. DOI: 10.1016/j.jenvman.2016.02.028.

[4] Alexander Y. Sun and Bridget R. Scanlon. *How can Big Data and machine learning benefit environment and water management: A survey of methods, applications, and future directions*. July 2019. DOI: 10.1088/1748-9326/ab1b7d.

[5] David Lazer et al. "The parable of Google Flu: traps in big data analysis". In: *science* 343.6176 (2014), pp. 1203–1205.

[6] Sirisha Adamala. "An Overview of Big Data Applications in Water Resources Engineering". In: *Machine Learning Research* 2 (1 2017), pp. 10–18. DOI: 10.11648/j.mlr.20170201.12. URL: http://www.sciencepublishinggroup.com/j/mlr.

[7] Rob Kitchin. "Big Data, new epistemologies and paradigm shifts". In: *Big Data and Society* 1 (1 July 2014). ISSN: 20539517. DOI: 10.1177/2053951714528481.

[8] Jeanne Behnke, Andrew Mitchell, and Hampapuram Ramapriyan. "NASA's earth observing data and information system - Near-term challenges". In: *Data Science Journal* 18 (1 2019). ISSN: 16831470. DOI: 10.5334/dsj-2019-040.

[9] Feng Zhang, Hui feng Xue, and Jing Cheng Zhang. "Multi-source big data dynamic compressive sensing and optimization method for water resources based on IoT". In: *Sustainable Computing: Informatics and Systems* 20 (Dec. 2018), pp. 210–219. ISSN: 22105379. DOI: 10.1016/j.suscom.2017.08.003.

[10] Ce Zhang et al. "Extracting databases from dark data with deepdive". In: *Proceedings of the 2016 International Conference on Management of Data*. 2016, pp. 847–859.

[11] Jamal Elhassan, Moumen Aniss, and Chao Jamal. "Big Data Analytic Architecture for Water Resources Management: A Systematic Review". In: Association for Computing Machinery, Mar. 2020. ISBN: 9781450375788. DOI: 10.1145/3399205.3399225.

[12] R. Rawat and R. Yadav. "Big Data: Big data analysis, issues and challenges and technologies". In: vol. 1022. IOP Publishing Ltd, Jan. 2021. DOI: 10.1088/1757-899X/1022/1/012014.

[13] Gaganjot Kang, Jerry Zeyu Gao, and Gang Xie. "Data-driven water quality analysis and prediction: A survey". In: Institute of Electrical and Electronics Engineers Inc., June 2017, pp. 224–232. ISBN: 9781509063185. DOI: 10.1109/BigDataService.2017.40.

[14] Wei Chen et al. "Research and Design of Distributed IoT Water Environment Monitoring System Based on LoRa". In: *Wireless Communications and Mobile Computing* 2021 (2021). ISSN: 15308677. DOI: 10.1155/2021/9403963.

[15] Michael V. Storey, Bram van der Gaag, and Brendan P. Burns. "Advances in on-line drinking water quality monitoring and early warning systems". In: *Water Research* 45 (2 2011), pp. 741–747. ISSN: 00431354. DOI: 10.1016/j.watres.2010.08.049.

[16] Marina Drosou et al. *Diversity in Big Data: A Review*. June 2017. DOI: 10.1089/big.2016.0054.

[17] Paolo Lo Giudice et al. "An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake". In: *Information Sciences* 478 (Apr. 2019), pp. 606–626. ISSN: 00200255. DOI: 10.1016/j.ins.2018.11.052.

[18] Muhammed Sit et al. "A comprehensive review of deep learning applications in hydrology and water resources". In: *Water Science and Technology* 82 (12 Dec. 2020), pp. 2635–2670. ISSN: 19969732. DOI: 10.2166/wst.2020.369.

[19] Mohammadhossein Amini and Shing Chang. *Assessing Data Veracity for Data-Rich Manufacturing Multiple Criteria Decision Making Facing Uncertain Attributes View project Enterprise Level Process Quality Control Using ML Techniques View project Assessing Data Veracity for Data-Rich Manufacturing*. 2017. URL: https://www.researchgate.net/publication/317604831.

[20] Fakhitah Ridzuan and Wan Mohd Nazmee Wan Zainon. "A review on data cleansing methods for big data". In: vol. 161. Elsevier B.V., 2019, pp. 731–738. DOI: 10.1016/j.procs.2019.11.177.

[21] Jinsong Wu et al. "Big Data Meet Green Challenges: Big Data Toward Green Applications". In: *IEEE Systems Journal* 10 (3 Sept. 2016), pp. 888–900. ISSN: 19379234. DOI: 10.1109/JSYST.2016.2550530.

[22] Yinghui Zhao and Ru An. "Big data analytics for water resources sustainability evaluation". In: *High-Performance Computing Applications in Numerical Simulation and Edge Computing: ACM ICS 2018 International Workshops, HPCMS and HiDEC, Beijing, China, June 12, 2018, Revised Selected Papers 2*. Springer. 2019, pp. 29–38.

[23] Christa D. Peters-Lidard et al. "Scaling, similarity, and the fourth paradigm for hydrology". In: *Hydrology and Earth System Sciences* 21.7 (2017), pp. 3701–3713. ISSN: 16077938. DOI: 10.5194/hess-21-3701-2017.

[24] Burhanullah Khattak et al. "Empirical Analysis of Recent Advances, Characteristics and Challenges of Big Data". In: *EAI Endorsed Transactions on Scalable Information Systems* 6 (23 2019), pp. 1–18. ISSN: 20329407. DOI: 10.4108/eai.13-7-2018.159621.

[25] Radoslaw M Cichy and Daniel Kaiser. "Deep neural networks as scientific models". In: *Trends in cognitive sciences* 23.4 (2019), pp. 305–317.

[26] Vasit Sagan et al. “Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing”. In: *Earth-Science Reviews* 205 (2020). ISSN: 00128252. DOI: 10.1016/j.earscirev.2020.103187.

[27] Andreas Holbach et al. “Three Gorges Reservoir: Density pump amplification of pollutant transport into tributaries”. In: *Environmental Science and Technology* 48.14 (2014), pp. 7798–7806. ISSN: 15205851. DOI: 10.1021/es501132k.

[28] V. Simeonov et al. “Assessment of the surface water quality in Northern Greece”. In: *Water Research* 37.17 (2003), pp. 4119–4124. ISSN: 00431354. DOI: 10.1016/S0043-1354(03)00398-1.

[29] Kunwar P. Singh, Amrita Malik, and Sarita Sinha. “Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques - A case study”. In: *Analytica Chimica Acta* 538.1-2 (2005), pp. 355–374. ISSN: 00032670. DOI: 10.1016/j.aca.2005.02.006.

[30] Yuming Su et al. “Identifying key drivers of harmful algal blooms in a tributary of the Three Gorges Reservoir between different seasons: Causality based on data-driven methods”. In: *Environmental Pollution* 297.August 2021 (2022), p. 118759. ISSN: 18736424. DOI: 10.1016/j.envpol.2021.118759. URL: https://doi.org/10.1016/j.envpol.2021.118759.

[31] David K. Ralston and Stephanie K. Moore. “Modeling harmful algal blooms in a changing climate”. In: *Harmful Algae* 91.November (2020), p. 101729. ISSN: 18781470. DOI: 10.1016/j.hal.2019.101729. URL: https://doi.org/10.1016/j.hal.2019.101729.

[32] Tianan Deng, Kwok Wing Chau, and Huan Feng Duan. “Machine learning based marine water quality prediction for coastal hydro-environment management”. In: *Journal of Environmental Management* 284.December 2020 (2021), p. 112051. ISSN: 10958630. DOI: 10.1016/j.jenvman.2021.112051. URL: https://doi.org/10.1016/j.jenvman.2021.112051.

[33] Peixuan Yu et al. “Predicting coastal algal blooms with environmental factors by machine learning methods”. In: *Ecological Indicators* 123 (2021), p. 107334. ISSN: 1470160X. DOI: 10.1016/j.ecolind.2020.107334. URL: https://doi.org/10.1016/j.ecolind.2020.107334.

[34] Quang Viet Ly et al. “Application of Machine Learning for eutrophication analysis and algal bloom prediction in an urban river: A 10-year study of the Han River, South Korea”. In: *Science of the Total Environment* 797 (2021), p. 149040. ISSN: 18791026. DOI: 10.1016/j.scitotenv.2021.149040. URL: https://doi.org/10.1016/j.scitotenv.2021.149040.

[35] Hongbo Liu et al. “Occurrence and Emergency Response of 2-Methylisoborneol and Geosmin in a Large Shallow Drinking Water Reservoir”. In: *Clean - Soil, Air, Water* 44.1 (2016), pp. 63–71. ISSN: 18630669. DOI: 10.1002/clen.201500077.

[36] Ting Xu et al. "Characteristics of antibiotics and antibiotic resistance genes in Qingcaosha Reservoir in Yangtze River Delta, China". In: *Environmental Sciences Europe* 32.1 (2020). ISSN: 21904715. DOI: 10.1186/s12302-020-00357-y. URL: https://doi.org/10.1186/s12302-020-00357-y.

[37] M. Drusch et al. "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services". In: *Remote Sensing of Environment* 120 (2012), pp. 25–36. ISSN: 00344257. DOI: 10.1016/j.rse.2011.11.026.

[38] Willibroad Gabila Buma and Sang Il Lee. "Evaluation of Sentinel-2 and Landsat 8 images for estimating Chlorophyll-a concentrations in Lake Chad, Africa". In: *Remote Sensing* 12.15 (2020). ISSN: 20724292. DOI: 10.3390/RS12152437.

[39] Dong Dong Zhang, Feng Xie, and Lei Zhang. "Preprocessing and fusion analysis of GF-2 satellite Remote-sensed spatial data". In: *Proceedings of 2018 International Conference on Information Systems and Computer Aided Education, ICISCAE 2018* (2019), pp. 24–29. DOI: 10.1109/ICISCAE.2018.8666873.

[40] Ian W. Housman, Robert A. Chastain, and Mark V. Finco. "An evaluation of forest health insect and disease survey data and satellite-based remote sensing forest change detection methods: Case studies in the United States". In: *Remote Sensing* 10.8 (2018). ISSN: 20724292. DOI: 10.3390/rs10081184.

[41] Gil Shamai and Ron Kimmel. "Geodesic distance descriptors". In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. Vol. 2017-Janua. 2017, pp. 3624–3632. ISBN: 9781538604571. DOI: 10.1109/CVPR.2017.386. arXiv: 1611.07360.

[42] David Rolnick and Max Tegmark. "The power of deeper networks for expressing natural functions". In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. 2018, pp. 1–14. arXiv: 1705.05502.

[43] Abien Fred Agarap. "Deep Learning using Rectified Linear Units (ReLU)". In: *arXiv e-prints*, arXiv:1803.08375 (Mar. 2018), arXiv:1803.08375. arXiv: 1803.08375 [cs.NE].

[44] Olivier Renaud and Maria Pia Victoria-Feser. "A robust coefficient of determination for regression". In: *Journal of Statistical Planning and Inference* 140.7 (2010), pp. 1852–1862. ISSN: 03783758. DOI: 10.1016/j.jspi.2010.01.008.

[45] T. Chai and R. R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature". In: *Geoscientific Model Development* 7.3 (2014), pp. 1247–1250. ISSN: 19919603. DOI: 10.5194/gmd-7-1247-2014.

[46] Paul Goodwin and Richard Lawton. "On the asymmetry of the symmetric MAPE". In: *International Journal of Forecasting* 15.4 (1999), pp. 405–408. ISSN: 01692070. DOI: 10.1016/S0169-2070(99)00007-2.

[47] Peter J. Rousseeuw and Christophe Croux. "Alternatives to the median absolute deviation". In: *Journal of the American Statistical Association* 88.424 (1993), pp. 1273–1283. ISSN: 1537274X. DOI: 10.1080/01621459.1993.10476408.

[48] Xifeng Guo et al. “Improved deep embedded clustering with local structure preservation”. In: *IJCAI International Joint Conference on Artificial Intelligence*. Vol. 0. August. 2017, pp. 1753–1759. ISBN: 9780999241103. DOI: 10.24963/ijcai.2017/243.

[49] Dmitry I. Belov and Ronald D. Armstrong. “Distributions of the Kullback-Leibler divergence with applications”. In: *British Journal of Mathematical and Statistical Psychology* 64.2 (2011), pp. 291–309. ISSN: 00071102. DOI: 10.1348/000711010X522227.

[50] Qinpei Zhao and Pasi Fränti. “WB-index: A sum-of-squares based index for cluster validity”. In: *Data and Knowledge Engineering* 92 (2014), pp. 77–89. ISSN: 0169023X. DOI: 10.1016/j.datak.2014.07.008. URL: http://dx.doi.org/10.1016/j.datak.2014.07.008.

[51] Huw Pohlner. “Institutional change and the political economy of water megaprojects: China’s south-north water transfer”. In: *Global Environmental Change* 38 (2016), pp. 205–216. ISSN: 09593780. DOI: 10.1016/j.gloenvcha.2016.03.015.

[52] Yuanzhu Wang et al. “Climatic changes and anthropogenic activities driving the increase in nitrogen: Evidence from the south-to-north water diversion project”. In: *Water (Switzerland)* 13.18 (2021). ISSN: 20734441. DOI: 10.3390/w13182517.

[53] Yan Long et al. “Comprehensive risk assessment of algae and shellfish in the middle route of South-to-North Water Diversion Project”. In: *Environmental Science and Pollution Research* 29.52 (2022), pp. 79320–79330. ISSN: 16147499. DOI: 10.1007/s11356-022-21210-0.

[54] Yuxuan Zhu et al. “Environmental Factors Drive Periphytic Algal Community Assembly in the Largest Long-Distance Water Diversion Channel”. In: *Water* 14.6 (2022). ISSN: 2073-4441. DOI: 10.3390/w14060914. URL: https://www.mdpi.com/2073-4441/14/6/914.

[55] ASTM. *Standard practices for measurement of chlorophyll content of algae in surface waters, D 3731-87*. 1993.

[56] Grace W. Lindsay. “Attention in Psychology, Neuroscience, and Machine Learning”. In: *Frontiers in Computational Neuroscience* 14.April (2020), pp. 1–21. ISSN: 16625188. DOI: 10.3389/fncom.2020.00029.

[57] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[58] Zhihao Fan et al. “Mask Attention Networks: Rethinking and Strengthen Transformer”. In: *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference* (2021), pp. 1692–1701. DOI: 10.18653/v1/2021.naacl-main.135. arXiv: 2103.13597.

[59] Jake Grigsby, Zhe Wang, and Yanjun Qi. “Long-range transformers for dynamic spatiotemporal forecasting”. In: *arXiv preprint arXiv:2109.12218* (2021).

[60] Boqiang Qin et al. “Dynamics of sediment resuspension and the conceptual schema of nutrient release in the large shallow Lake Taihu, China”. In: *Chinese Science Bulletin* 49.1 (2004), pp. 54–64. ISSN: 10016538. DOI: 10.1360/03wd0174.

[61] Yuchao Zhang et al. “Wind effects for floating algae dynamics in eutrophic lakes”. In: *Remote Sensing* 13.4 (2021), pp. 1–11. ISSN: 20724292. DOI: 10.3390/rs13040800.

[62] Meitar Ronen, Shahaf E. Finder, and Oren Freifeld. “DeepDPM: Deep Clustering With an Unknown Number of Clusters”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2022-June.Dl (2022), pp. 9851–9860. ISSN: 10636919. DOI: 10.1109/CVPR52688.2022.00963. arXiv: 2203.14309.

[63] W Keith Hastings. “Monte Carlo sampling methods using Markov chains and their applications”. In: (1970).

[64] Meitar Ronen, Shahaf E Finder, and Oren Freifeld. “DeepDPM: Deep Clustering With an Unknown Number of Clusters”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2022-June. Dl. 2022, pp. 9851–9860. ISBN: 9781665469463. DOI: 10.1109/CVPR52688.2022.00963. arXiv: 2203.14309.

[65] Ulrike Von Luxburg. “A tutorial on spectral clustering”. In: *Statistics and Computing* 17.4 (2007), pp. 395–416. ISSN: 09603174. DOI: 10.1007/s11222-007-9033-z. arXiv: 0711.0189.

[66] Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. “The k-means algorithm: A comprehensive survey and performance evaluation”. In: *Electronics (Switzerland)* 9.8 (2020), pp. 1–12. ISSN: 20799292. DOI: 10.3390/electronics9081295.

[67] Fionn Murtagh and Pedro Contreras. “Algorithms for hierarchical clustering: An overview”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.1 (2012), pp. 86–97. ISSN: 19424795. DOI: 10.1002/widm.53.

[68] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780. ISSN: 08997667. DOI: 10.1162/neco.1997.9.8.1735.

[69] Alex Sherstinsky. “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network”. In: *Physica D: Nonlinear Phenomena* 404 (2020), p. 132306. ISSN: 01672789. DOI: 10.1016/j.physd.2019.132306. arXiv: 1808.03314. URL: https://doi.org/10.1016/j.physd.2019.132306.

[70] Anthony Gillioz et al. “Overview of the Transformer-based Models for NLP Tasks”. In: *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020* 21 (2020), pp. 179–183. DOI: 10.15439/2020F20.

[71] Kamilya Smagulova and Alex Pappachen James. “A survey on LSTM memristive neural network architectures and applications”. In: *The European Physical Journal Special Topics* 228.10 (2019), pp. 2313–2324.

[72] Yong Yu et al. “A review of recurrent neural networks: LSTM cells and network architectures”. In: *Neural computation* 31.7 (2019), pp. 1235–1270.

[73] Seyed Mehran Kazemi et al. “Time2vec: Learning a vector representation of time”. In: *arXiv preprint arXiv:1907.05321* (2019).

[74] Shibani Santurkar et al. “How does batch normalization help optimization?” In: *Advances in neural information processing systems* 31 (2018).

[75] Ruibin Xiong et al. "On layer normalization in the transformer architecture". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 10524–10533.

[76] Bashar Alhnaity et al. "An autoencoder wavelet based deep neural network with attention mechanism for multi-step prediction of plant growth". In: *Information Sciences* 560 (2021), pp. 35–50. ISSN: 00200255. DOI: 10.1016/j.ins.2021.01.037. arXiv: 2012.04041.

[77] Zuchao Li et al. "Seq2seq dependency parsing". In: *Proceedings of the 27th International Conference on Computational Linguistics*. 2018, pp. 3203–3214.

[78] Kalyan Das, Jiming Jiang, and J. N.K. Rao. "Mean squared error of empirical predictor". In: *Annals of Statistics* 32.2 (2004), pp. 818–840. ISSN: 00905364. DOI: 10.1214/009053604000000201.

[79] Ingrid Chorus and Martin Welker. *Exposure to cyanotoxins*. Geneva: Taylor & Francis, 2021. Chap. 5, pp. 295–400. DOI: 10.1201/9781003081449-5.

[80] Shuhang Wang et al. "Characteristics of dissolved organic matter and its role in lake eutrophication at the early stage of algal blooms-A case study of Lake Taihu, China". In: *Water (Switzerland)* 12.8 (2020), pp. 1–17. ISSN: 20734441. DOI: 10.3390/w12082278.

[81] Piero Morosini. "Industrial clusters, knowledge integration and performance". In: *World Development* 32 (2 2004), pp. 305–326. ISSN: 0305750X. DOI: 10.1016/j.worlddev.2002.12.001.

[82] Richard Harris. "Models of regional growth: past, present and future". In: *Journal of economic surveys* 25.5 (2011), pp. 913–951.

[83] Ian R Gordon and Philip McCann. "Clusters, innovation and regional development: an analysis of current theories and evidence". In: *Industrial clusters and inter-firm networks* (2005), pp. 29–57.

[84] Chin-Huang Lin, Chiu-Mei Tung, and Chih-Tai Huang. "Elucidating the industrial cluster effect from a system dynamics perspective". In: *Technovation* 26.4 (2006), pp. 473–482.

[85] Zachary C Lipton. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3 (2018), pp. 31–57.

[86] Roberta Calegari, Giovanni Ciatto, and Andrea Omicini. "On the integration of symbolic and sub-symbolic techniques for XAI: A survey". In: *Intelligenza Artificiale* 14.1 (2020), pp. 7–32.

[87] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58 (2020), pp. 82–115.

[88] Xian Liu et al. "Data-driven machine learning in environmental pollution: Gains and problems". In: *Environmental science and technology* 56.4 (2022), pp. 2124–2133.

[89] Guang Yang, Qinghao Ye, and Jun Xia. "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond". In: *Information Fusion* 77 (2022), pp. 29–52.

A

Full Articles of Scientific Publications as the First Author

1. **Qian, Jing**, et al. "Water quality monitoring and assessment based on cruise monitoring, remote sensing, and deep learning: A case study of Qingcaosha Reservoir." Frontiers in Environmental Science 10 (2022).

2. **Qian, Jing**, et al. "Identification of driving factors of algal growth in the South-to-North Water Diversion Project by Transformer-based deep learning." Water Biology and Security (2023): 100184.

3. **Qian, Jing**, et al. "An Intelligent Early Warning System for Harmful Algal Blooms: Harnessing the Power of Big Data and Deep Learning." (Under review)

frontiers | Frontiers in Environmental Science

TYPE Original Research
PUBLISHED 11 October 2022
DOI 10.3389/fenvs.2022.979133

OPEN ACCESS

EDITED BY
Shuisen Chen,
Guangzhou Institute of Geography, China

REVIEWED BY
Heng Lyu,
Nanjing Normal University, China
Yulong Guo,
Henan Agricultural University, China

*CORRESPONDENCE
Jing Qian,
jing.qian@partner.kit.edu

SPECIALTY SECTION
This article was submitted to Environmental Informatics and Remote Sensing,
a section of the journal
Frontiers in Environmental Science

RECEIVED 28 June 2022
ACCEPTED 13 September 2022
PUBLISHED 11 October 2022

CITATION
Qian J, Liu H, Qian L, Bauer J, Xue X, Yu G, He Q, Zhou Q, Bi Y and Norra S (2022), Water quality monitoring and assessment based on cruise monitoring, remote sensing, and deep learning: A case study of Qingcaosha Reservoir. *Front. Environ. Sci.* 10:979133. doi: 10.3389/fenvs.2022.979133

Water quality monitoring and assessment based on cruise monitoring, remote sensing, and deep learning: A case study of Qingcaosha Reservoir

Jing Qian[1]*, Hongbo Liu[2], Li Qian[3], Jonas Bauer[1], Xiaobai Xue[4], Gongliang Yu[5], Qiang He[6], Qi Zhou[7], Yonghong Bi[5] and Stefan Norra[1]

[1]Institute of Applied Geosciences, Karlsruhe Institute of Technology, Karlsruhe, Germany, [2]School of Environment and Architecture, University of Shanghai for Science and Technology, Shanghai, China, [3]Institute of Informatics, Ludwig Maximilian University of Munich, Munich, Germany, [4]MioTech Research, Yingtou Information Technology (Shanghai) Limited, Shanghai, China, [5]State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, China, [6]Key Laboratory of Eco-environments in the Three Gorges Reservoir Region, Ministry of Education, College of Environmental and Ecology, Chongqing University, Chongqing, China, [7]College of Environmental Science and Engineering, Tongji University, Shanghai, China

Accurate monitoring and assessment of the environmental state, as a prerequisite for improved action, is valuable and necessary because of the growing number of environmental problems that have harmful effects on natural systems and human society. This study developed an integrated novel framework containing three modules remote sensing technology (RST), cruise monitoring technology (CMT), and deep learning to achieve a robust performance for environmental monitoring and the subsequent assessment. The deep neural network (DNN), a type of deep learning, can adapt and take advantage of the big data platform effectively provided by RST and CMT to obtain more accurate and improved monitoring results. It was proved by our case study in the Qingcaosha Reservoir (QCSR) that DNN showed a more robust performance (R^2 = 0.89 for pH, R^2 = 0.77 for DO, R^2 = 0.86 for conductivity, and R^2 = 0.95 for backscattered particles) compared to the traditional machine learning, including multiple linear regression, support vector regression, and random forest regression. Based on the monitoring results, the water quality assessment of QCSR was achieved by applying a deep learning algorithm called improved deep embedding clustering. Deep clustering analysis enables the scientific delineation of joint control regions and determines the characteristic factors of each area. This study presents the high value of the framework with a core of big data mining for environmental monitoring and follow-up assessment in a manner of high frequency, multidimensionality, and deep hierarchy.

KEYWORDS

deep learning, environmental big data mining, cruise monitoring, remote sensing, water quality, monitoring, assessment

1 Introduction

A growing population and climate change along with land use changes are increasing pollutant loads into freshwater ecosystems, making clean water an increasingly critical issue worldwide Sagan et al. (2020). As one of the indispensable foundations of clean water management, developing an economical, accurate, and practical water quality monitoring and assessing system has become unavoidable to scientists, policymakers, and environmental resource managers.

The traditional and widely applied water quality monitoring is point-based, placing a fixed site of varying density and dispersion in the area to measure the water quality within a given time series. However, limited research resources such as staff, time, equipment, money, and accessibility become a challenge. Thus, the spatial interpolation method was conducted to estimate water quality by limited monitoring points Li and Heap (2014). This method required a massive decentralized monitoring point across the study area, which is also subjected to limited research resources Lee et al. (2012).

With the significant development of sensors, cruise monitoring technology (CMT) has proven to be more effective for extracting environment-related parameters compared to point-based monitoring Holbach et al. (2014). It relies on a multisensor probe to record the water quality data as well as consecutive geographic information along the cruise route. Although CMT makes progress in monitoring compared to the point-based method because it can collect a large amount of *in situ* measurement data in a certain period of time, a route design is still necessary since the geographic information is a key parameter to spatial interpolation modeling.

In recent years, remote sensing technology (RST) has developed rapidly and played a significant role in the data collection and analysis of different Earth resources Feyisa et al. (2014). The data collected by RST are area-based since RST can scan the objective area directly. The status of water quality in a broader space is obtained according to an inversion model established using the *in situ* monitoring data (i.e., water quality parameters) and corresponding RST image data Yuan et al. (2020). According to the interaction with light, water-quality parameters can be categorized into optical parameters (i.e., chlorophyll-a and turbidity) and nonoptical parameters (i.e., dissolved oxygen); it should be noted that most of the studies have focused on optical parameters, and the detection accuracy for nonoptical parameters is not high Hassan et al. (2021). Specific internal correlations between spectral information and nonoptical parameters are very complex and challenging to find due to the absence of direct optical properties Niu et al. (2021). Therefore, data-driven machine learning has become an indispensable tool for finding this complex correlation Zhong et al. (2021) Sagan et al. (2020). In earlier studies, linear approaches such as multiple linear regression (MLR), partial least squares (PLSs), and genetic algorithms (GAs) were popular Ortiz-Casas and Peña-Martinez (1989); Stork and Autrey (2005); Zhan et al. (2003). Although linear models showed some degree of accuracy and feasibility, the nonlinear relationship between the *in situ* measured data and RST data makes the linear models less reliable in interpreting information from RST Chang et al. (2015). With the development of machine learning, several nonlinear approaches such as support vector regression (SVR), random forest regression (RFR), and gradient boosting decision tree (GBDT) have been developed and applied by many scientists to capture complex statistical relationships between RST and measured water quality parameters in recent years Kim et al. (2014); Forkuor et al. (2017); Abdel-Rahman et al. (2013). With the advances in algorithm development and computing power, the drawbacks of traditional machine learning become apparent, while deep learning, with its powerful big data processing capabilities, is receiving more attention. In our framework, deep neural networks (DNNs), one type of deep learning, were selected as a tool to approximate the complex nonlinear relationship between measured water quality parameters and RST observations through multilayer perception Marçais and de Dreuzy (2017).

It is important to note that the performance of deep learning methods is particularly dependent on a large number of training samples, which is difficult to obtain in real-world scenarios Sagan et al. (2020). The CMT mentioned earlier can significantly increase the speed of acquiring training samples, thus providing a sufficient database for deep learning RST inversion model building. On the other hand, RST, to a certain degree, liberates CMT from dependence on route design since geographic information is not involved in the inversion modeling.

As an important part of the water monitoring project, a representative and reliable assessment of water quality is necessary because of the spatial and temporal variability of water parameters Simeonov et al. (2003). The conventional methods for assessing the quality of water bodies are the single-factor assessment method, water quality grading method, and comprehensive pollution index method. These methods play an active role in the assessment process of water quality. However, the single-factor assessment method does not fully describe the overall water quality when there are multiple impairments. The water quality grading method ignores the influence of extreme contributing factors (maximum and minimum pollutant parameter values), making it difficult to assess the overall water quality conditions between sites when extreme conditions occur. The calculation result of the comprehensive pollution index method is a relative value and cannot indicate the specific water quality classification Ji et al. (2016). In particular, when faced with the huge and complex matrix of water quality attributes formed by the establishment of a big data platform like this study, making a meaningful water quality assessment is often difficult Singh et al. (2005). A cluster

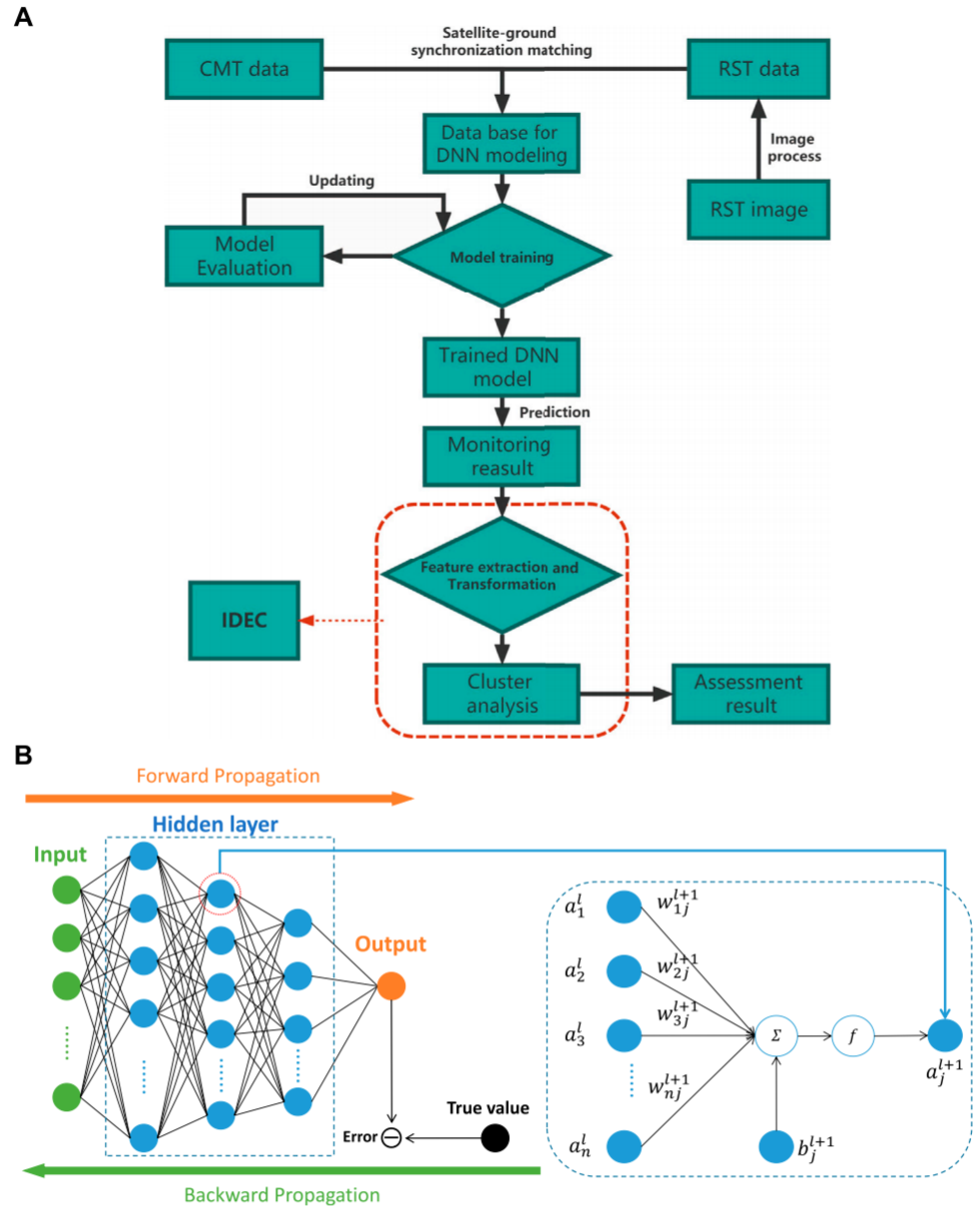


FIGURE 1
Schematic illustration of **(A)** the novel framework and **(B)** the architecture of DNN.

analysis can be applied to interpret these complex data matrices to help understand the water quality and ecological status of the studied systems, identifying the possible resources and finding rapid solutions to pollution problems by grouping the data so that similar elements are assigned to the same group and different elements are assigned to different ones Vo-Van et al. (2020); Simeonov et al. (2003). Additionally, considering that deep clustering is more effective at analyzing big data than traditional clustering methods Guo et al. (2017), such as K-means and C-means, an advanced deep learning clustering algorithm, improved deep embedded clustering (IDEC), was used for the water quality assessment in this study.

The aim of this study is to develop a novel framework with a core of big data mining, integrating (1) CMT data from multisensor monitoring systems, 2) RST information from the satellites, and 3) deep learning for rapid and effective overall water quality evaluation and the follow-up assessment of the environmental situation. This novel framework was applied in the Qingcaosha Reservoir (QCSR), located in Shanghai, to prove its reasonability and reliability.

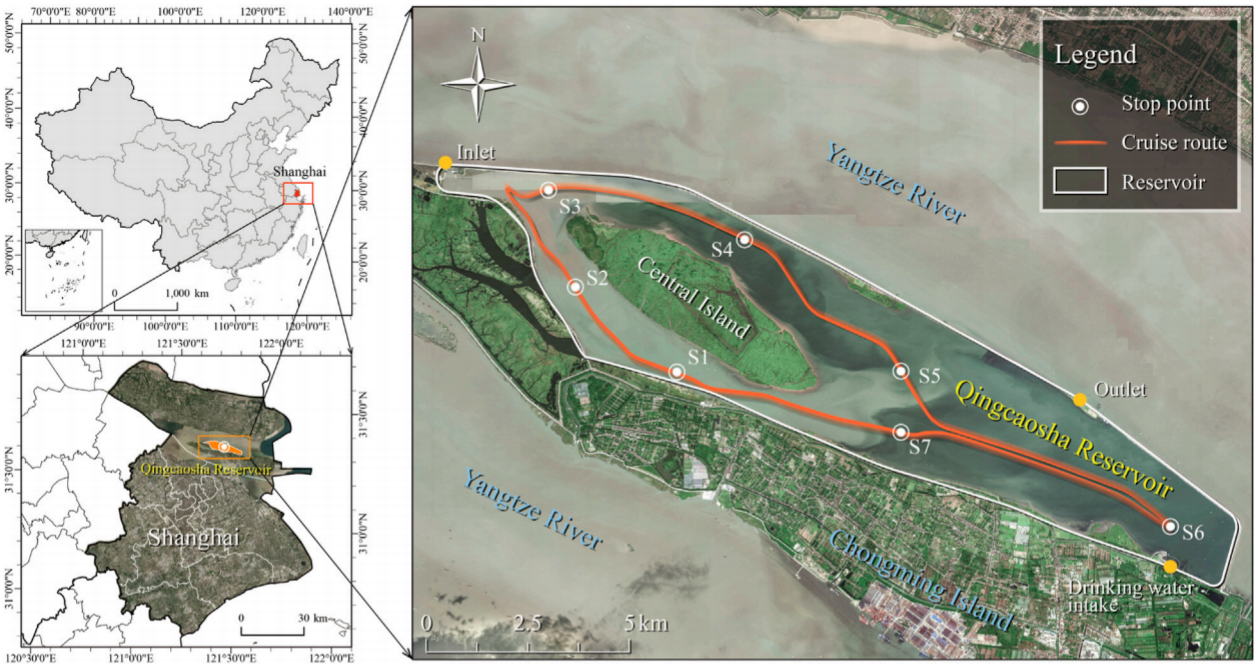


FIGURE 2
Study area and cruise route.

2 Materials and methods

To achieve a robust performance in water quality monitoring and assessment, the framework integrates the following three modularized parts: RST, CMT, and deep learning. The DNN model is responsible for efficient water quality monitoring on a big data platform created by CMT and RST, while the IDEC model is used for further assessment based on the previous monitoring results (Figure 1A). A sampling activity in QCSR (Supplementary Figure S1) was implemented to validate the performance of the framework. QCSR is one of the largest tidal reservoirs around the world. It is located in the middle of the Yangtze River Estuary (31.42–31.49N, 121.55–121.71E), and is the new largest drinking water supply for about 12 million Shanghai residents Liu et al. (2016a) since 2010 (Figure 2). The reservoir is long and narrow with a surface area of approximately 70 km^2 and an average depth of 2.7 m Liu et al. (2016b).

2.1 Remote sensing module

Sentinel-2 is an Earth observation satellite designed to systematically deliver optical imagery at high spatial resolution (10, 20, and 60m) over land and waters Drusch et al. (2012) (Supplementary Table S2). Due to its relatively high resolution and free accessibility, Sentinel-2 is widely used in environmental research. Its multi-spectral instrument (MSI) acquires 13 spectral bands from 440 nm to 2,200 nm. The image of Sentinel-2 on 19 January 2020 (the same day as the CMT *in situ* measurement) was downloaded from the official website of the U.S. Geological Survey (https://earthexplorer.usgs.gov/). The level-1C data product was selected in this study and this series of data has been radiometrically and geometrically corrected (including orthorectification).

RST image is processed in order of radiometric calibration, atmospheric correction, RST image fusion, and research area clipping to finish the conversion from images to spectral values (Figure 1A). One conventional atmospheric correction algorithm, Fast Line-of-Sight Atmospheric Analysis of Hypercubes (FLAASH) was set as an atmospheric correction algorithm in this study Buma and Lee (2020). The specific RST parameters set, including ground elevation, atmospheric model, aerosol retrieval, and water retrieval, were found in the files alongside their respective multispectral images. The RST images (not including bands 1, 9, and 10) were resampled to 10 m by the Gram-Schmidt pan sharpening method (Supplementary Material), one of the most widely used high-quality methods for RST image fusion Zhang et al. (2019). All of the RST data processing could be conducted using the packaged functions in ENVI®. The RST data were processed by Z-score normalization (Supplementary Material) before being input to the models. It should be noted that when new data are collected, the normalization part performs a new normalization of the

TABLE 1 Summary statistics of the BioFish data.

Parameter	*In situ* measuring(n)	Unit	Max	Min	Mean	STD*
pH	50081	N/A	8.31	6.20	7.34	0.43
DO	50005	*mg/L*	13.99	9.39	10.21	0.48
El.cond	50264	*mS/cm*	0.37	0.20	0.34	0.02
BP	50179	%	8.92	1.27	2.36	1.58

*Standard deviation.

overall data set (containing the previous data set and the new data set) for the training model.

2.2 Cruise monitoring module

Cruise monitoring with multiple sensors is conducted by BioFish in this study. It is an aquatic cruise monitoring system that is equipped with multisensors (Supplementary Table S1) and connected to a ship by a data transmission cable Udy et al. (2005). The data of water quality parameters were recorded in real-time with GPS longitudinal and latitudinal positions. In this study, the BioFish swam 10 cm below the water surface. One optical parameter, backscattered particles (BPs, similar to turbidity, measured by a beam attenuation probe to estimate water clarity) (Supplementary Material), and three nonoptical parameters, including electrical conductivity (El.cond), pH value, and dissolved oxygen (DO), are selected to validate the performance of the framework.

Due to the limitations of power supply, equipment, time, and accessibility, the *in situ* measuring in QCSR was finished within 1 day and the running time was 5 hours. The cruise route is shown in Figure 2, aiming to cover as much of the study area as possible. S1 is the start and end point of the cruise route. Seven stopping points were designed for (S1–S7, see Figure 2) the BioFish calibration with the YSL ProDSS to ensure the accuracy of the data. An overview of the data collected by BioFish in QCSR is displayed in Table 1. The BioFish data were processed by Z-score normalization (Supplementary Material) and satellite–ground synchronization matching (Supplementary Figure S2) before being input into the models (Figure 1A). It should be noted that the normalization section renormalizes the new overall dataset when new data are collected.

Since the high sampling density of BioFish means that multiple BioFish sampling points can be found randomly in a pixel block of size 10 m × 10 m, determining the BioFish sampling points within the same pixel block and deriving their representative values are required. The first step is to specify the spatial information of all BioFish sampling points and pixel grid centroids. The geodesic distance Shamai and Kimmel (2017) between the pixel grid centroid and the BioFish sampling point can be calculated by the *Python* package geopy, with an ellipsoidal model, WGS-84. Then, the pixel grid corresponding to the BioFish sampling point can be extracted by finding the shortest geodesic distance between them. The next step is calculating the representative values of BioFish measurements within each pixel block by the arithmetic mean (AM).

2.3 Water quality monitoring model

In this section, deep neural networks and three traditional machine learning models are used to find the relationship between RST and CMT and compare their performance, respectively.

2.3.1 Deep neural network

The deep neural network (DNN) is the basic form of deep learning and one of the most efficient and powerful tools to model complex nonlinear relationships Rolnick and Tegmark (2018). As the left side of Figure 1B shows, DNN is a connectionist system with multiple hidden layers between the input and output layers. Each hidden layer contains multiple neurons, called nodes. Any node in the lth layer must be connected to any node in the l + 1st layer, and the following equation indicates the nonlinear relationship between the DNN layers shown on the right side of Figure 1B:

$$a_j^{l+1} = f\left(\sum_{i=1}^{n} a_i^l w_{ij}^l + b_j^l\right),$$

where a_i^l is the activation value of the ith node in the lth layer, a_j^{l+1} is the activation value of the jth node in the l + 1st layer,$w_{ij}^{l+1}$is the weight between a_i^l and a_j^{l+1}, b_j^{l+1} is the bias value of the jth node in the l + 1st layer, and $f(\cdot)$ is the active function.

The training process is shown on the left side of Figure 1B. Forward propagation refers to the calculation and storage of intermediate variables (including outputs) from the input to the output layer. Back propagation refers to the method of calculating the gradient of neural network parameters and updating the parameters depending on the error between the output and true value. For tuning hyperparameters in this study, *relu* Agarap (2018) was set as the active function and *adam* as the

optimizer of all models. The layer and neural units of models were (256, 256, 256, 256, 256) except El.cond.-spectral value was (256, 256, 256). Additionally, batch size and learning rate were also tuned in a reasonable range.

2.3.2 Multiple linear regression

Linear regression, a typical traditional machine learning model, is a linear approach for estimating the relationship between a dependent variable and one or more independent variables. The case of one independent variable is called simple linear regression; for two or more, the process is called multiple linear regression (MLR) Berger et al. (2017). In this study, the MLR model was built by calling the function in the *Python* package scikit-learn. The parameter to be tuned in this study was the degree of the polynomial features.

2.3.3 Support vector regression

Support vector regression (SVR) is a traditional supervised machine learning that is applied widely in RST inversion Wagle et al. (2020). The SVR model was also conducted by calling the function in the *Python* package scikit-learn. The radial basis function was chosen as the kernel of SVR. The parameters that need to be tuned in this study are the regularization parameter and the kernel coefficient.

2.3.4 Random forest regression

Random forest regression (RFR) is a traditional machine learning algorithm for nonlinear regression. It uses an ensemble learning method that combines a large set of regression trees to make a more accurate regression than a single regression tree Kim et al. (2014). The RFR model was implemented by calling the function in the *Python* package scikit-learn. The n-estimators and random-state need to be tuned.

2.3.5 Evaluation metrics

Evaluating the performance of a model is an essential step before practical application. We split each dataset into a training set and a test set with a ratio of 4:1 and take one at every four intervals as the test data. Several indicators, including the coefficient of determination (R^2), root mean square error (RMSE), mean absolute percentage error (MAPE), and median absolute deviation (MAD), were used to evaluate each regression model's accuracy, stability, and inversion ability (Supplementary Material).

TABLE 2 Results of regression model evaluation.

Parameter	Model	R^2	RMSE*	MAPE**	MAD*
pH	MLR	0.55	0.64	0.86	0.41
	SVR	0.74	0.55	0.69	0.21
	RFR	0.73	0.50	0.57	0.17
	DNN	***0.89***	***0.33***	***0.52***	***0.10***
DO	MLR	0.22	0.85	2.83	0.30
	SVR	0.24	0.83	***1.30***	0.12
	RFR	0.57	0.65	1.59	0.14
	DNN	***0.77***	***0.49***	1.61	***0.06***
El.cond	MLR	0.23	0.88	9.62	0.20
	SVR	0.33	0.81	***1.54***	0.08
	RFR	0.52	0.67	1.78	0.10
	DNN	***0.86***	***0.38***	1.74	***0.06***
BP	MLR	0.78	0.44	3.07	0.14
	SVR	0.87	0.38	3.34	0.07
	RFR	0.87	0.38	***2.72***	0.06
	DNN	***0.95***	***0.26***	3.10	***0.03***

*Units are the same as the respective water quality parameter units.
**Unit is percentage.
The bold-italic values represent the best regression results, respectively.

2.4 Water quality assessment model

Improved deep embedded clustering is an unsupervised deep learning algorithm for clustering. The monitoring results obtained from the framework were clustered using IDEC, and points with similar environmental states were grouped based on the combined effect of all measured water quality parameters (pH, DO, BP, and El.cond in this study), thus dividing the entire reservoir into different areas possessing different environmental states. According to the clustering results, each group's specific water quality characteristics can be understood by analyzing the distribution of each group's characteristic water quality parameters. This characteristic of each group is the main reason why these measurement points are clustered into the same group, and it can also be described as the characteristic factor of this group.

The structure of IDEC includes an encoder and a decoder network Guo et al. (2017) (Supplementary Figure S3). The encoder network is set as a fully connected multilayer perceptron (MLP) with dimensions 4-125-125-500-10. The decoder network is a mirror of the encoder with dimensions 10-500-125-125-4. *relu* was set as the active function and *adam* as the optimizer of all models. The coefficient of cluster loss γ is set to 0.1 and batch size to 256. The convergence threshold δ is set to 0.1%. Also, the update interval T is one iteration. IDCE and CH method was conducted by PyTorch. The number of clusters was determined by the Calinski–harabasz (CH) method Zhao and Fränti (2014).

3 Results

Considering the conceptual merits of the developed framework, we applied the framework to a database of QCSR sampling activity to evaluate its performance on inversion and make the assessment of water quality through clustering results.

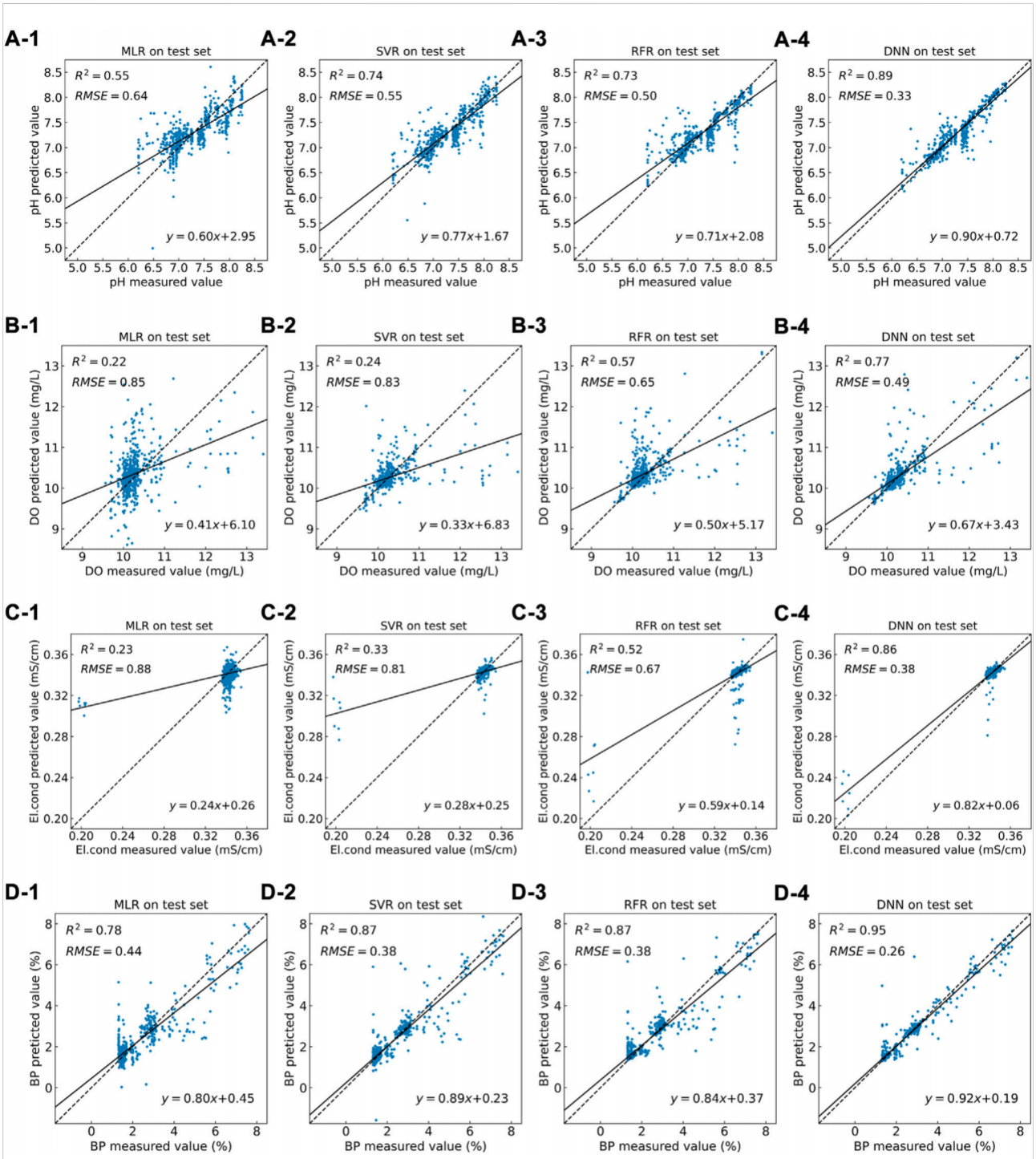


FIGURE 3
Regression model performance evaluation by comparison of the predicted data and measured data on a test set, where **(A)**, **(B)**, **(C)**, and **(D)** represent the test results of the pH, DO, El.cond, and BP, respectively and (1), (2), (3), and (4) represent the test results of the MLR, SVR, RFR, and DNN, respectively.

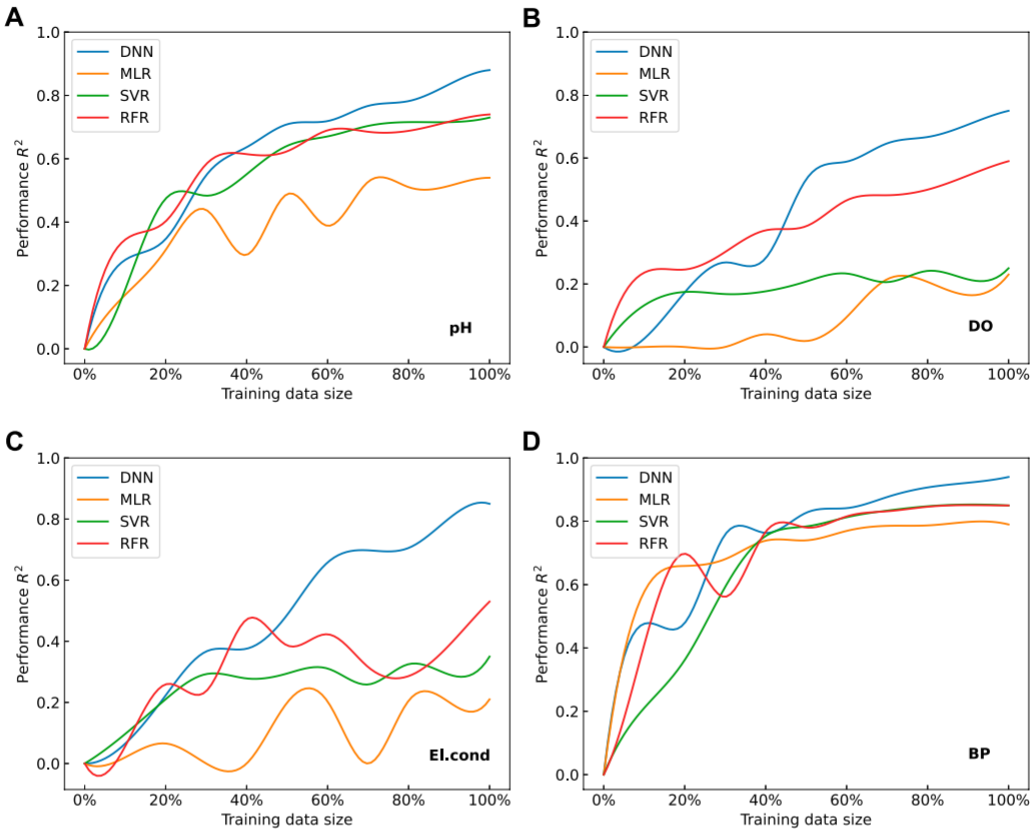


FIGURE 4
Performance of each regression model with increasing training data size, where **(A–D)** represent the results of the pH, DO, El.cond and BP, respectively.

3.1 Model performance evaluation

Based on the performance of the regression model for four water quality parameters, several results are achieved.

1) DNN represented the best performance in accuracy and stability compared to the other three algorithms. The results for the model performance are summarized in Table 2. Concerning the inversion of pH, SVR, RFR, and DNN delivered satisfactory results. DNN achieved the highest R^2 = 0.889 and the lowest RMSE = 0.33, MAPE = 0.52, and MAD = 0.10 that stand for the stability of DNN. As for the DO and El.cond inversion, DNN achieved the highest accuracy with R^2 = 0.77, 0.86 compared with MLR, SVR, and RFR. In addition, the lowest RMSE (0.49 for DO and 0.38 for El.cond) and MAD (0.06 for DO and 0.06 for El. Cond) demonstrated that DNN has high stability even though the MAPE of DNN is slightly less than that of SVR.

With respect to the inversion of BP, all models express relatively satisfactory results. In particular, DNN reached a very high accuracy with R^2 = 0.95 and relatively low RMSE, MAPE, and MAD.

The comparison of predicted values and the measured values on the test set are shown in Figure 3. It is found that the slope of DNN test results (0.90 for pH, 0.67 for DO, 0.82 for El.cond, and 0.92 for BP) is much larger than those of MRL, SVR, and RFR. Therefore, DNN significantly improved the inversion accuracy compared with MLR, SVR, and RFR.

2) The performance of each model increases with increasing training data size. We randomly select 0–100% of the data in the original dataset at 10% intervals for training and testing. This process is performed 50 times for each data size, and then the average performance of the model (denoted by R^2) and its standard deviation (Supplementary Material) are calculated at each data size. As the training data size increases, the results of each water quality parameter consistently showed an increasing trend of R^2 (Figure 4). It can be also found that DNN is highly sensitive to training data size. The performance of DNN was not the best among the four models with a small training data size, especially when less than 30% of the training data size was fed. (Figure 4D). When 40% or more of the training data are fed, a critical point is noted, where DNN performance surpasses the other models. In particular, a significant advantage of DNN can be observed as the training data size increases from 50% to 100%. Meanwhile, a more advanced performance of DNN could be expected.

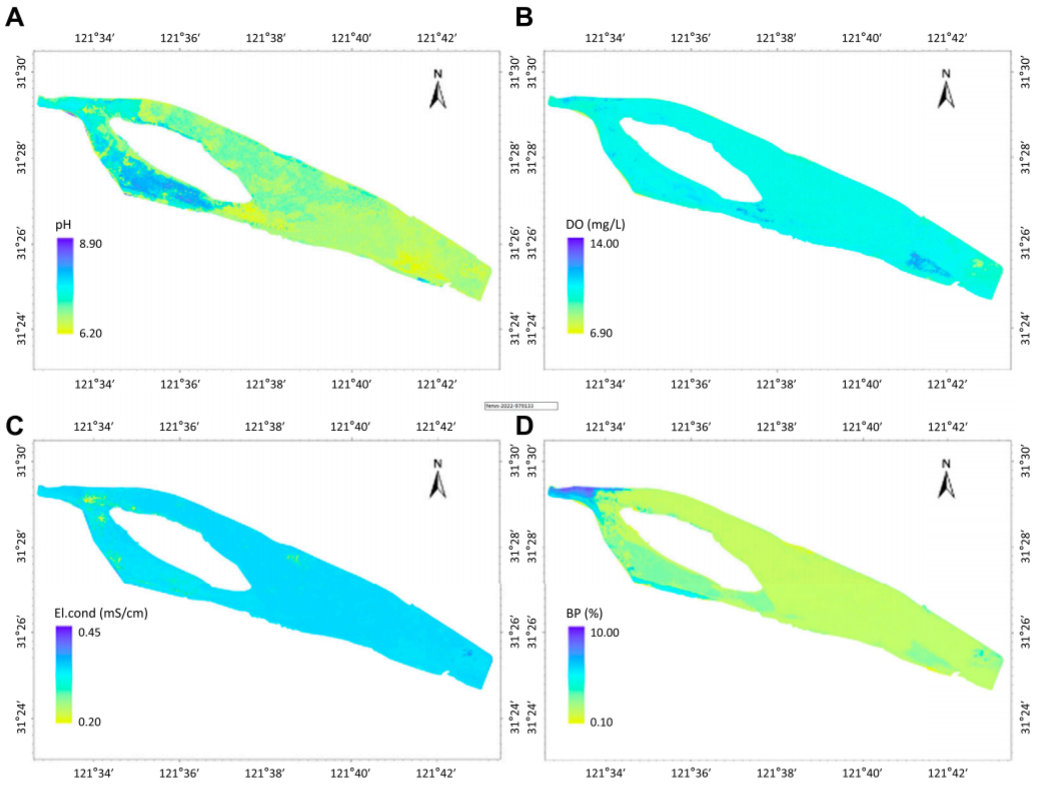


FIGURE 5
Distribution of **(A)** pH, **(B)** DO, **(C)** El.cond, and **(D)** BP in QCSR based on the framework.

3.2 Application of the developed framework in QCSR

The concentration heat-maps of each parameter (pH, DO, El.cond, and BP) are shown in Figures 5A–D, respectively. The pH value obtained by the developed framework ranges from 6.2 to 8.9, with a mean value of 7.2. The results reveal spatial difference in that pH value decreases from the head region of QCSR to the tail region (see Figure 5A). The inverted DO ranges from 6.90 mg/L to 14.00 mg/L, with a mean value of 10.30 mg/L. The results show a similar spatial difference as that of pH since the concentration of DO decreases from Figure 5Bthe head region to the tail region, as shown in Figure 5B. Differing from the pH value, a relatively low concentration of DO occurs in the eastern portion of the reservoir, the tail region. The result of El.cond ranges from 0.20 mS/cm to 0.44 mS/cm, with a mean value of 0.34 mS/cm. El.cond observed in the head region is lower than that of the rest area (see Figure 5C). The inverted BP ranges from 0.10 to 10.00%, with a mean value of 2.13%. The results show a similar spatial difference as that of pH since BP decreases from the head to the tail region, as shown in Figure 5D.

TABLE 3 Summary statistics of each group.

Group	Parameter*	Max	Min	STD**	Median
Group I	pH	9.00	6.20	0.25	7.04
	DO	14.00	6.90	0.32	10.13
	El.cond	0.44	0.20	0.01	0.34
	BP	3.02	0.01	0.23	1.57
Group II	pH	8.25	6.50	0.27	7.64
	DO	13.94	6.90	0.64	9.94
	El.cond	0.42	0.20	0.02	0.34
	BP	10.08	4.58	1.26	6.11
Group III	pH	8.30	6.20	0.223	6.95
	DO	14.00	9.15	0.72	12.11
	El.cond	0.35	0.20	0.02	0.34
	BP	7.47	1.23	0.54	2.60
Group IV	pH	9.00	6.20	0.51	7.86
	DO	14.00	6.90	0.45	10.26
	El.cond	0.44	0.20	0.02	0.34
	BP	4.83	1.94	0.52	2.93

*Unit of DO is *mg/L*; unit of El.cond is *mS/cm*; unit of BP is percentage.
**Standard deviation.

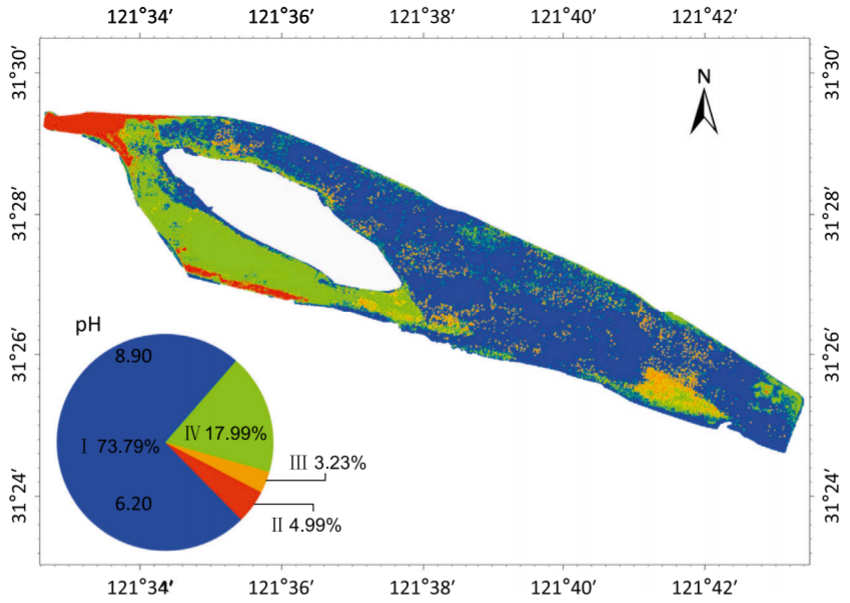


FIGURE 6
Clustering result of QCSR.

The entire QCSR can be divided into four groups of water bodies by the CH method (Supplementary Material, Figure 4), which are groups I, II, III, and IV. The summary statistics and distributions of each group obtained by IDEC are shown in Table3 and Figure 6. Group I occupied 73.79% of the entire QCSR area. It exhibits characteristics that the values for each parameter are in the middle position compared to others. Group I is dominated from the northeast of the central island to the tail region. The proportions of group II and III water were 4.99 and 3.23%, respectively. Group II is characterized by significantly higher BP values than the other groups and is distributed at the head of the reservoir and on the southern shore of the reservoir. Group III shows a higher DO value compared to other groups and is distributed close to the drinking water intake. Group IV has a higher pH value compared to others, indicating a mildly alkaline water body. Also, it is mainly found to the southwest of the central island to a lesser extent in the tail region and near the drinking water intake.

4 Discussion

4.1 Advantages of the novel framework

The design and application of the framework in our case study demonstrated its high performance in the monitoring and assessment of water quality. Compared to the previous studies, the advantages of this framework are summarized as follows.

CMT and RST are mutually integrated into the framework. CMT provides RST with sufficient *in situ* measurements, the prerequisite of the data set. In this study, the training data size provided by BioFish was nearly 500 times as many as the manual method within the same time interval Sagan et al. (2020); in addition, the geographic information of the data is not involved directly in the training and test process as input, which solves the problem of space-time limitation of the spatial interpolation methods to a certain extent. In the case of QCSR, the water quality parameters far from the cruise route, where no cruise route can be used nearby, can still be effectively inverted.

The environmental big data platform established by CMT and RST provides the basis for accurate environmental information interpretation. RST and CMT have the attributes of big data and good complementary so that the environmental big data platform can be built with the cooperation of the two parts. As shown in Figure 4, the results show an increasing trend of R^2 modelwise as the data size enlarges, indicating a significant advantage of environmental data analysis in contrast with the small or medium data platform.

On the big data platform, the adaptability and performance of deep learning ensure accuracy in monitoring and assessment. In Figure 4, break-even points can be observed at which the performance of DNN exceeds those of other traditional machine learning, especially when 40% or more data are fed. Through the encoder network with dimensions 4-125-125-500-10 and decoder network dimensions 10-500-125-125-4, original four-dimensional features (pH, DO, El.cond, and BP) are transferred into the new four-dimensional features, which contain much

more information. Based on the updated four-dimensional features, clustering results are better than those based on the original four-dimensional features, meaning they are very close to the real world.

The novel framework formed a closed loop of water quality research, into which data collection, processing, monitoring, and assessment are packaged. In the framework, monitoring results can be mined for further assessment. Joint regional control strategies are more efficient and effective than single-point control strategies in environmental management and pollution control Zhang and Yang (2022). Deep clustering analysis enables the scientific delineation of joint control regions. Through the character analysis of the divided joint control area, characteristic factors of each area can be identified, which can contribute to defining a joint regional control strategy for the objective area. In this study, each group is managed as a joint control area, in a way that depends on the characteristic factors. Elevated BP (low water clarity) noted in the group II area may cause poor underwater light climate and loss of submerged macrophytes to switch the water body from a macrophyte-dominated state to an algae-dominated one Huang et al. (2021). In addition, the alkaline water body is one of the stimulatives for algae growth Lin et al. (2021). This means that the two water quality parameters, BP and pH, will be the focus of subsequent management and control of the distribution areas of group II and group IV, respectively.

To further ensure the reliability and accuracy of data collection, we have several particular strategies. 1) Seven calibrating points keep BioFish in a well-calibrated condition during the *in situ* measuring in order to assure the measuring accuracy. 2) The day of the satellite transit with cloud cover of less than 10 % was selected as the sampling day.

4.2 Potentiality of the developed framework

The developed framework as well as its three modularized parts show high potential in extensibility.

1) The environmental quality parameters were inverted by the developed framework by a data-driven approach instead of a physics- or chemistry-based one. Being data-driven makes results from the developed framework easily and rapidly transform into inversion of other environmental parameters collected by different sensors or CMT systems. The implementation was in the water scenario in this study. Alternately, this framework can be applicable to the air scenario when using an air quality CMT system.

2) The developed framework can realize the water quality monitoring in a timely manner by shortening the revisit time. The revisit time is defined as the time interval between two successive a satellite or a system's observations on the same ground point on the surface of the Earth Luo et al. (2017). In this study, we chose the Sentinel-2 satellite system with a 5-day revisit time as the source of RST images. Accordingly, the need for 5-day monitoring of the whole target water bodies can be met with good weather conditions for RST observation and the availability of all parties (e.g., financing and labor). Selecting satellites with a shorter revisit time can increase the monitoring frequency, enabling the whole QCSR monitoring to keep pace with the environmental change of frequency, for example, replacing Sentinel-2 in this article with WorldView-3 (97 min revisit time) Ye et al. (2017).

A satellite with a spatial and spectral resolution provides a more precise inversion result and sharper clustering spatial boundaries by reducing the size of the raster within the objective area. As such, replacing Sentinel-2 in this article with WorldView-3 Ye et al. (2017) would obtain an up-to-date and more accurate result of inversion and clustering.

3) Our experiment in Section 3.1, 3.2 revealed that the performance of DNN is susceptible to the data size and gains a significant improvement as the data size increases. The reason can be seen in Figure 1B that each training iteration results in a model that is pretrained for the next training iteration after forward and backward calculations, and this process continues iteratively. Thus, we can expect a more robust and accurate model when more data are fed, such as more applications of the framework. More importantly, as the model was fed and trained by massive data, the *in situ* measuring might not be necessary.

4) The deep clustering method dealing with water quality assessment has advantages for big data sets with higher dimensional water quality parameters and multiple time periods. For processing high-dimensional water quality parameters, IDEC can have more objectives to extract and transform water quality features, which can make the clustering results closer to the real situation. In addition, deep clustering of the data for each time period separately allows for delineating newly integrated control areas and their characteristics. In this way, the overall state changes in the target water bodies can be seen at a glance, such as changes in the boundaries of each control area and changes in the characteristics of each control area.

4.3 Future work

Notwithstanding the developed work had several advantages, it is essential to note that improvements can be a part of future work.

RST images are significantly affected by weather conditions, especially cloud cover. An image with less than 35% cloud cover was regarded as a good practice to satisfy environmental monitoring requirements Marshall et al. (1994). To ensure accuracy, the "clear sky" images with less than 10% cloud cover were applied to the framework. Thus, there was a strict weather restriction during the *in situ* measuring. Sometimes the uncertainty of the weather can make *in situ* measurements

fruitless, even though the plan is made according to the weather forecast.

Furthermore, the cruise monitor's running does not synchronize with the satellite's visit to the objective area. For instance, it takes Sentinel-2 less than 2 s to cross the QCSR. This lag prevents the satellite from being in real-time synchronization with the CMT measurements. In order to hurdle the weather limitations and eliminate the lag between RST and CMT measurements, we plan to introduce unmanned aerial vehicles (UAVs) associated with multispectral sensors into the framework. Its lower-than-cloud flight altitude reduces the interference of the cloud. In addition, the synchronized working pace of UAVs allows for simultaneous data collection along with the cruise monitor. As a supplementary element of the framework, UAVs are particularly applicable to small surface water areas like river bays and estuaries.

Last but not least, the performance of deep learning was essentially dependent on the data size. Hence, collecting more data from diverse types of water bodies should be a critical and indispensable work.

5 Conclusion

An innovative framework was developed with three modules: RST, CMT, and deep learning. Deep learning uses the big data platform created by RST and CMT to achieve a robust performance in water quality monitoring and assessment. Our testing revealed that the DNN (a type of deep learning) in the framework has a higher performance in monitoring four water quality parameters (pH, DO, El.cond, and BP) than MLR, SVR, and RFR. DNN is highly sensitive to training data size compared to other models, and the performance increases significantly with the elevated training data size. The application of IDEC on the water quality assessment showed that the entire QCSR was well-defined and divided into four groups as joint control areas, which are group I, group II, group III, and group IV. The characteristic factors of each area were identified, which can contribute to defining a joint regional control strategy for the QCSR. Considering the big data platform is the foundation of this framework, our future work in priority would be collecting more measured data (RST and CMT) from different water bodies to increase the capacity of the big data platform and update the deep learning model in our framework.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

Conceptualization: JQ; data curation: JQ and JB; formal analysis: JQ and LQ; funding acquisition: HL and SN; investigation: JQ, JB, and XX; methodology: JQ; project administration: SN and HL; resources: SN and HL; software: JQ and LQ; supervision: SN, HL, and YB; visualization: JQ; writing—original draft: JQ; writing—review and editing: SN, YB, QH, GY, QZ, and JB.

Funding

This work was supported by the National Natural Science Foundation of China (U2040210 and 31971477), and SIGN II—Amoris, BMBF (02WCL1471J).

Acknowledgments

The authors are grateful for the help from Andre Wilhelms for providing the basic information of BioFish and Xiaojie Zhang during *in situ* measuring. They also acknowledge support from the KIT-Publication Fund of the Karlsruhe Institute of Technology.

Conflict of interest

XX was employed by the company MioTech Research.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fenvs.2022.979133/full#supplementary-material

References

Abdel-Rahman, E. M., Ahmed, F. B., and Ismail, R. (2013). Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. *Int. J. Remote Sens.* 34, 712–728. doi:10.1080/01431161.2012.713142

Agarap, A. F. (2018). *Deep learning using rectified linear units (ReLU). arXiv e-prints*arXiv:1803.08375.

Berger, P. D., Maurer, R. E., and Celli, G. B. (2017). "Experimental design: With applications in management, engineering, and the sciences," in *Experimental design: With applications in management, engineering and the Sciences*. Second edition. Cham, Switzerland: Springer Cham. chap. Multiple L. 505–507. doi:10.1007/978-3-319-64583-4

Buma, W. G., and Lee, S. I. (2020). Evaluation of sentinel-2 and landsat 8 images for estimating chlorophyll-a concentrations in lake Chad, africa. *Remote Sens.* 12, 2437. doi:10.3390/RS12152437

Chang, N. B., Imen, S., and Vannah, B. (2015). Remote sensing for monitoring surface water quality status and ecosystem state in relation to the nutrient cycle: A 40-year perspective. *Crit. Rev. Environ. Sci. Technol.* 45, 101–166. doi:10.1080/10643389.2013.829981

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., et al. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sens. Environ.* 120, 25–36. doi:10.1016/j.rse.2011.11.026

Feyisa, G. L., Meilby, H., Fensholt, R., and Proud, S. R. (2014). Automated water extraction index: A new technique for surface water mapping using landsat imagery. *Remote Sens. Environ.* 140, 23–35. doi:10.1016/j.rse.2013.08.029

Forkuor, G., Hounkpatin, O. K., Welp, G., and Thiel, M. (2017). High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: A comparison of machine learning and multiple linear regression models. *PLoS ONE* 12, 01704788–e170521. doi:10.1371/journal.pone.0170478

Guo, X., Gao, L., Liu, X., and Yin, J. (2017). Improved deep embedded clustering with local structure preservation. *IJCAI Int. Jt. Conf. Artif. Intell.* 0, 1753–1759. doi:10.24963/ijcai.2017/243

Hassan, G., Goher, M. E., Shaheen, M. E., and Taie, S. A. (2021). Hybrid predictive model for water quality monitoring based on sentinel-2A L1C data. *IEEE Access* 9, 65730–65749. doi:10.1109/ACCESS.2021.3075849

Holbach, A., Norra, S., Wang, L., Yijun, Y., Hu, W., Zheng, B., et al. (2014). Three Gorges Reservoir: Density pump amplification of pollutant transport into tributaries. *Environ. Sci. Technol.* 48, 7798–7806. doi:10.1021/es501132k

Huang, J., Qian, R., Gao, J., Bing, H., Huang, Q., Qi, L., et al. (2021). A novel framework to predict water turbidity using Bayesian modeling. *Water Res.* 202, 117406. doi:10.1016/j.watres.2021.117406

Ji, X., Dahlgren, R. A., and Zhang, M. (2016). Comparison of seven water quality assessment methods for the characterization and management of highly impaired river systems. *Environ. Monit. Assess.* 188, 15–16. doi:10.1007/s10661-015-5016-2

Kim, Y. H., Im, J., Ha, H. K., Choi, J. K., and Ha, S. (2014). Machine learning approaches to coastal water quality monitoring using GOCI satellite data. *GIsci. Remote Sens.* 51, 158–174. doi:10.1080/15481603.2014.900983

Lee, S. J., Serre, M. L., van Donkelaar, A., Martin, R. V., Burnett, R. T., and Jerrett, M. (2012). Comparison of geostatistical interpolation and remote sensing techniques for estimating long-term exposure to ambient PM2.5 concentrations across the continental United States. *Environ. Health Perspect.* 120, 1727–1732. doi:10.1289/ehp.1205006

Li, J., and Heap, A. D. (2014). Spatial interpolation methods applied in the environmental sciences: A review. *Environ. Model. Softw.* 53, 173–189. doi:10.1016/j.envsoft.2013.12.008

Lin, S. S., Shen, S. L., Zhou, A., and Lyu, H. M. (2021). Assessment and management of lake eutrophication: A case study in lake erhai, China. *Sci. Total Environ.* 751, 141618. doi:10.1016/j.scitotenv.2020.141618

Liu, H., Pan, D., and Chen, P. (2016a). A two-year field study and evaluation of water quality and trophic state of a large shallow drinking water reservoir in Shanghai, China. *Desalination Water Treat.* 57, 13829–13838. doi:10.1080/19443994.2015.1059370

Liu, H., Pan, D., Zhu, M., and Zhang, D. (2016b). Occurrence and emergency response of 2-methylisoborneol and geosmin in a large shallow drinking water reservoir. *Clean. Soil Air Water* 44, 63–71. doi:10.1002/clen.201500077

Luo, X., Wang, M., Dai, G., and Chen, X. (2017). A novel technique to compute the revisit time of satellites and its application in remote sensing satellite optimization design. *Int. J. Aerosp. Eng.*, 1–9. doi:10.1155/2017/6469439

Marçais, J., and de Dreuzy, J.-R. (2017). Prospective interest of deep learning for hydrological inference. *Groundwater* 55, 688–692. doi:10.1111/gwat.12557

Marshall, G. J., Dowdeswell, J. A., and Rees, W. G. (1994). The spatial and temporal effect of cloud cover on the acquisition of high quality landsat imagery in the European Arctic sector. *Remote Sens. Environ.* 50, 149–160. doi:10.1016/0034-4257(94)90041-8

Niu, C., Tan, K., Jia, X., and Wang, X. (2021). Deep learning based regression for optically inactive inland water quality parameter estimation using airborne hyperspectral imagery. *Environ. Pollut.* 286, 117534. doi:10.1016/j.envpol.2021.117534

Ortiz-Casas, J. L., and Peña-Martinez, R. (1989). Water quality monitoring in Spanish reservoirs by satellite remote sensing. *Lake Reserv. Manag.* 5, 23–29. doi:10.1080/07438148909354395

Rolnick, D., and Tegmark, M. (2018). The power of deeper networks for expressing natural functions." in 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, Vancouver, BC, Canada, April 30–May 3, 2018

Sagan, V., Peterson, K. T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B. A., et al. (2020). Monitoring inland water quality using remote sensing: Potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. *Earth-Science Rev.* 205, 103187. doi:10.1016/j.earscirev.2020.103187

Shamai, G., and Kimmel, R. (2017). Geodesic distance descriptors." in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition CVPR, Honolulu, HI, USA, July 21–26, 2017, 3624. doi:10.1109/CVPR.2017.386

Simeonov, V., Stratis, J. A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., et al. (2003). Assessment of the surface water quality in Northern Greece. *Water Res.* 37, 4119–4124. doi:10.1016/S0043-1354(03)00398-1

Singh, K. P., Malik, A., and Sinha, S. (2005). Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques - a case study. *Anal. Chim. Acta* 538, 355–374. doi:10.1016/j.aca.2005.02.006

Stork, C. L., and Autrey, B. C. (2005). Remotely mapping river water quality using multivariate regression with prediction validation. *Remote Sens. Model. Ecosyst. Sustain. II* 5884, 588408. doi:10.1117/12.616852

Udy, J., Gall, M., Longstaff, B., Moore, K., Roelfsema, C., Spooner, D. R., et al. (2005). Water quality monitoring: A combined approach to investigate gradients of change in the great barrier reef, Australia. *Mar. Pollut. Bull.* 51, 224–238. doi:10.1016/j.marpolbul.2004.10.048

Vo-Van, T., Nguyen-Hai, A., Tat-Hong, M. V., Nguyen-Trang, T., and Gomariz, F. (2020). *A new clustering algorithm and its application in assessing the quality of underground water*. Scientific Programming. doi:10.1155/2020/6458576

Wagle, N., Acharya, T. D., and Lee, D. H. (2020). Comprehensive review on application of machine learning algorithms for water quality parameter estimation using remote sensing data. *Sensors Mater.* 32, 3879–3892. doi:10.18494/SAM.2020.2953

Ye, B., Tian, S., Ge, J., and Sun, Y. (2017). Assessment of WorldView-3 data for lithological mapping. *Remote Sens.* 9, 1132–1219. doi:10.3390/rs9111132

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., et al. (2020). Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sens. Environ.* 241, 111716. doi:10.1016/j.rse.2020.111716

Zhan, H., Lee, Z., Shi, P., Chen, C., and Carder, K. L. (2003). Retrieval of water optical properties for optically deep waters using genetic algorithms. *IEEE Trans. Geosci. Remote Sens.* 41, 1123–1128. doi:10.1109/TGRS.2003.813554

Zhang, D. D., Xie, F., and Zhang, L. (2019). Preprocessing and fusion analysis of GF-2 satellite Remote-sensed spatial data." in Proceedings of 2018 International Conference on Information Systems and Computer Aided Education, Changchun, China, July 6–8, 2018. ICISCAE, 24. doi:10.1109/ICISCAE.2018.8666873

Zhang, L., and Yang, G. (2022). Cluster analysis of PM2.5 pollution in China using the frequent itemset clustering approach. *Environ. Res.* 204, 112009. doi:10.1016/j.envres.2021.112009

Zhao, Q., and Fränti, P. (2014). WB-Index: A sum-of-squares based index for cluster validity. *Data Knowl. Eng.* 92, 77–89. doi:10.1016/j.datak.2014.07.008

Zhong, S., Zhang, K., Bagheri, M., Burken, J. G., Gu, A., Li, B., et al. (2021). Machine learning: New ideas and tools in environmental science and engineering. *Environ. Sci. Technol.* 55, 12741–12754. doi:10.1021/acs.est.1c01339

Supplementary Material

1 CHALLENGES OF WATER QUALITY MONITORING IN REMOTE AND LARGE AREAS

In theoretical water monitoring, we should know the water quality parameters at every point and its corresponding time across the study area and period. However, the research sources, including time, equipment, accessibility, and money are usually limited. Thus, we have to rely on these limited research resources to get the conclusions we want. Especially in remote and large areas, where there are no sophisticated networks to monitor water quality, a framework to capture water quality at each point with limited research resources is needed. In our case, the Qingcaosha Reservoir is a very important source of drinking water, supplying over 50% of Shanghai's drinking water. We were facing the challenge of monitoring water quality in the $70km^2$ water body within one day (limited accessibility and time) by using only one BIOFISH (limited equipment). Therefore, this framework integrating remote sensing, cruise monitory, and deep neural network was developed to overcome these challenges.

Figure S1. Photo of BIOFISH running in Qingcaosha reservoir

2 BIOFISH MODULE AND DATA PROCESS

BIOFISH is a aquatic multi-sensors cruise monitor. The sensors and their specifications are shown in Table S1. BP (Backscatter particle) in "%" reflects the percentage of light scattered back to the sensor after emission and depends on the turbidity or water transparency. The more particles in the water, the more light is scattered back.

Z-score Normalization is a tool to standardize features by removing the mean and scaling to unit variance. This scaler is widely used for normalization in many machine learning algorithms (e.g., support vector

Table S1. Sensors of BIOFISH and their specifications

Parameter	Manufacturer	Principle	Range	Resolution	Accuracy
Pressure	ADM Elektronik	piezo-resistive	0-100dBar	0.01dBar	±0.1dBar
Temperature	ADM Elektronik	Pt 100	0-36$^{\circ}C$	0.001$^{\circ}C$	±0.01$^{\circ}C$
pH	ADM Elektronik	Potentiometric (Ag/AgCl)	0-12pH	0.01pH	±0.02pH
DO	ADM Elektronik	Potentiometric (Clark electrode)	0-100%	0.01%	±0.01%
El.cond	ADM Elektronik	7-pole-cell	0-60mS/cm	1uS/cm	±10uS/cm
Backscattered particles	ADM Elektronik	Mie backscattering	0-100%	0.01%	±0.01%

machines, logistic regression, and artificial neural networks). And it was utilized for each water quality parameter in this study i recorded by BIOFISH according to the following equationHousman et al. (2018):

$$Z_i = \frac{x_i - \overline{x_i}}{\sigma_i}$$

where Z_i is the standard score of i-th water quality parameter, x_i is the -th original water quality parameter, $\overline{x_i}$ is the mean of i-th water quality parameter, and σ_{i} is the standard deviation of i-th water quality parameter.

3 REMOTE SENSING

3.1 Sentinel-2

The information of Sentinel-2 is shown in Table S2.

Table S2. Bands of Sentinel-2 and their specifications

Bands	Specification	Central Wavelength (μm)	Resolution (m)
Band 1*	Coastal aerosol	0.443	60
Band 2	Blue	0.490	10
Band 3	Green	0.560	10
Band 4	Red	0.665	10
Band 5	Vegetation Red Edge	0.705	20
Band 6	Vegetation Red Edge	0.740	20
Band 7	Vegetation Red Edge	0.783	20
Band 8	NIR	0.842	10
Band 8A	Vegetation Red Edge	0.865	20
Band 9*	Water vapor	0.945	60
Band 10*	SWIR - Cirrus	1.375	60
Band 11	SWIRk	1.610	20
Band 12	SWIR	2.190	20

* Band 1, 9 and 10 were not used in this study

3.2 Pan-sharpening

Due to the technology at the time, Sentinel-2 could not also provide images with both high spectral and spatial resolutions. What it can offer is multispectral images have high spectral resolution but relatively low spatial resolution; panchromatic(PAN) images have high spatial resolution but low spectral resolution. In order to make the multispectral images have high spatial detail performance while preserving the spectral characteristics of multispectral images, merging the low-spatial-resolution multispectral images with high-spatial-resolution panchromatic optical images is a solution. This process is Pan-sharpening.

4 EVALUATION METRICS

To quantify the performance of the models, the coefficient of determinationRenaud and Victoria-Feser (2010)(R^2) was calculated as:

$$R^2\left(y,\hat{y}\right)=1-\frac{\sum_{i=1}^{n}\left(y_i-\widehat{y_i}\right)^2}{\sum_{i=1}^{n}\left(y_i-\overline{y_i}\right)^2}$$

Root-mean-square errorChai and Draxler (2014) (RMSE) was calculated as:

$$RMSE(y,\widehat{y})=\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\widehat{y_i}\right)^2}{n}}$$

Mean absolute percentage errorGoodwin and Lawton (1999) (MAPE) was calculated as:

$$MAPE\left(y,\hat{y}\right)=\frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i-\widehat{y_i}}{y_i}\right|$$

Median absolute deviationRousseeuw and Croux (1993) (MAD) was calculated as:

$$MAD\left(y,\hat{y}\right)=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\widehat{y_i}\right|$$

where $\widehat{y_i}$ is the predicted value of the ith sample, y_i is the corresponding true value of the total n samples, and $\overline{y_i}$ is the mean of true value.

5 RESULT OF DIFFERENT TRAINING DATA SIZE

The standard deviation of the model performance for each parameter at each data size are shown in S3, S4, S5, and S6.

Table S3. The standard deviation of the pH model performance at each data size

pH	DNN	MLR	SVR	RFR
10%	0.15	11.23	0.27	0.16
20%	0.09	0.61	0.17	0.09
30%	0.05	0.42	0.07	0.08
40%	0.04	0.12	0.07	0.05
50%	0.04	0.11	0.05	0.04
60%	0.03	0.12	0.04	0.03
70%	0.02	0.05	0.03	0.03
80%	0.03	0.03	0.03	0.02
90%	0.04	0.02	0.02	0.01
100%	0.02	0.02	0.01	0.01

Table S4. The standard deviation of the DO model performance at each data size

DO	DNN	MLR	SVR	RFR
10%	1.91	2081.34	3.94	1.21
20%	0.30	25.03	0.30	0.24
30%	0.27	6.51	0.21	0.23
40%	0.17	1.58	0.09	0.12
50%	0.14	0.90	0.08	0.10
60%	0.12	0.27	0.06	0.12
70%	0.10	0.09	0.05	0.07
80%	0.09	0.07	0.04	0.07
90%	0.06	0.11	0.03	0.05
100%	0.04	0.02	0.01	0.03

Table S5. The standard deviation of the El.cond model performance at each data size

El.cond	DNN	MLR	SVR	RFR
10%	10.10	1693.67	10.88	10.59
20%	8.74	21.66	3.20	4.88
30%	1.97	5.39	1.14	2.07
40%	0.82	1.29	0.26	1.10
50%	0.27	0.51	0.16	0.37
60%	0.17	0.14	0.10	0.24
70%	0.18	0.16	0.08	0.17
80%	0.12	0.04	0.07	0.13
90%	0.10	0.03	0.05	0.10
100%	0.07	0.01	0.01	0.06

Table S6. The standard deviation of the BP model performance at each data size

BP	DNN	MLR	SVR	RFR
10%	0.17	17.15	0.26	0.16
20%	0.09	2.28	0.12	0.09
30%	0.06	0.16	0.06	0.06
40%	0.04	0.09	0.06	0.06
50%	0.03	0.13	0.04	0.04
60%	0.03	0.05	0.04	0.03
70%	0.03	0.05	0.03	0.03
80%	0.03	0.02	0.02	0.02
90%	0.02	0.01	0.01	0.01
100%	0.02	0.01	0.01	0.01

6 IMPROVED DEEP EMBEDDED CLUSTERING

Improved deep embedded clustering (IDEC) is a deep clustering algorithms, concluding encoder and decoder. The network construction of IDEC is shown in Figure S2.

Calinski-harabasz (CH) method is used to determine of the number of clusters k. Several CH score of each cluster number were calculated based on sum-of-squares within cluster (SSW) and/or sum-of-squares between clusters (SSB) values, the equationsZhao and Fränti (2014) are:

$$SSW = \sum_{i=1}^{N} \|x_i - C_{pi}\|^2$$

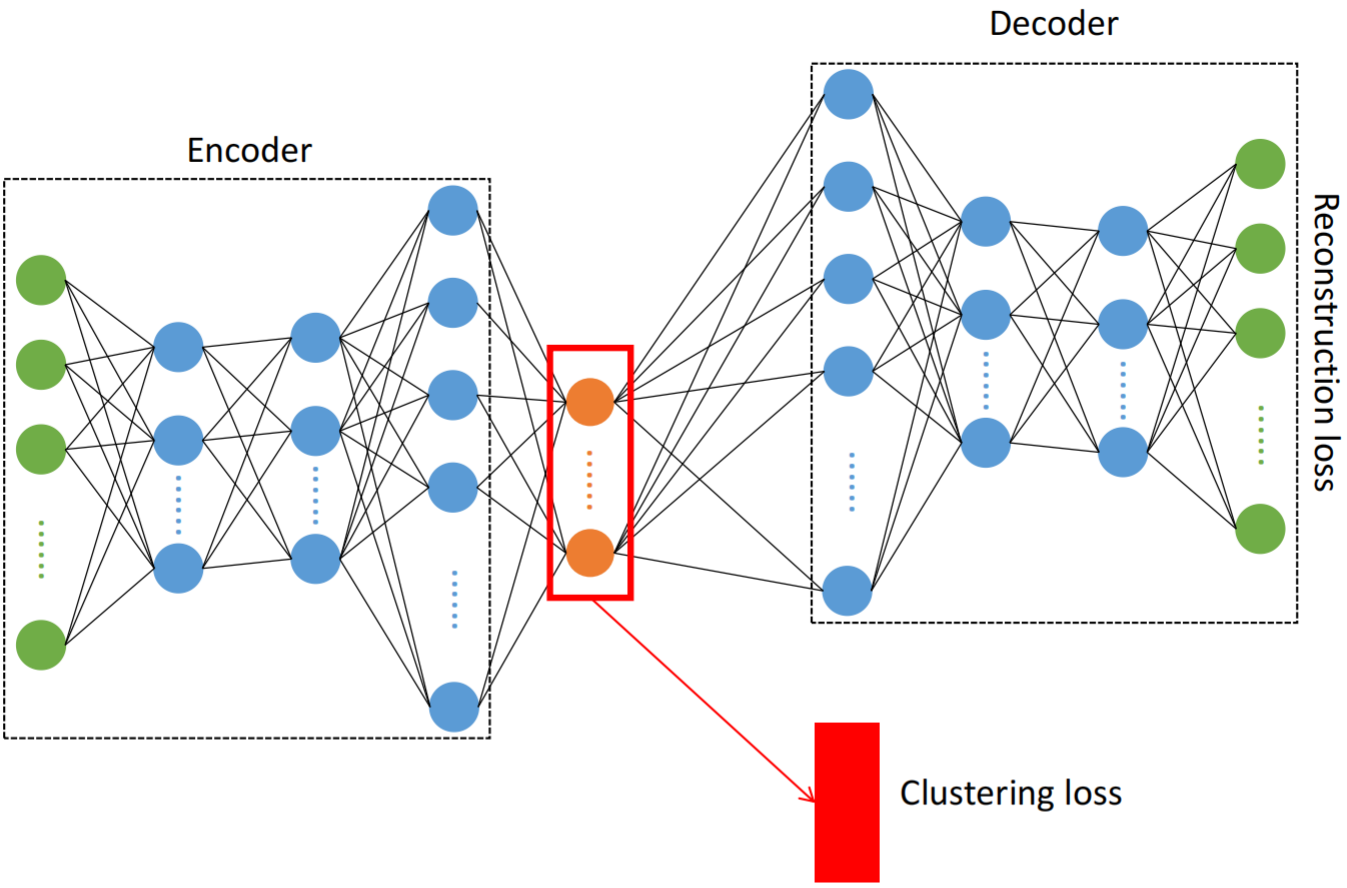


Figure S2. The network construction of IDEC

$$SSB = \sum_{i=1}^{k} n_i \left\| c_i - \bar{X} \right\|^2$$

$$CH = \frac{SSB/(k-1)}{SSW/(N-k)}$$

Where k is the number of clusters, P is the partitions, $X = \{x_1, x_2 ... x_N\}$ represents the data set with N-dimensional points, $\overline{X} = \sum_{i=1}^{N} x_i/N$ is the center of the entire data set, $C = \{c_1, c_2 ... c_k\}$ represents the centroids of cluster, c_i is the ith cluster.

The knee point detection algorithm finds the point of maximum curvature, which corresponds to most optimal clustering number. As Figure S3 shows, k=4 is the most optimal cluster number in the assessment of QCSR.

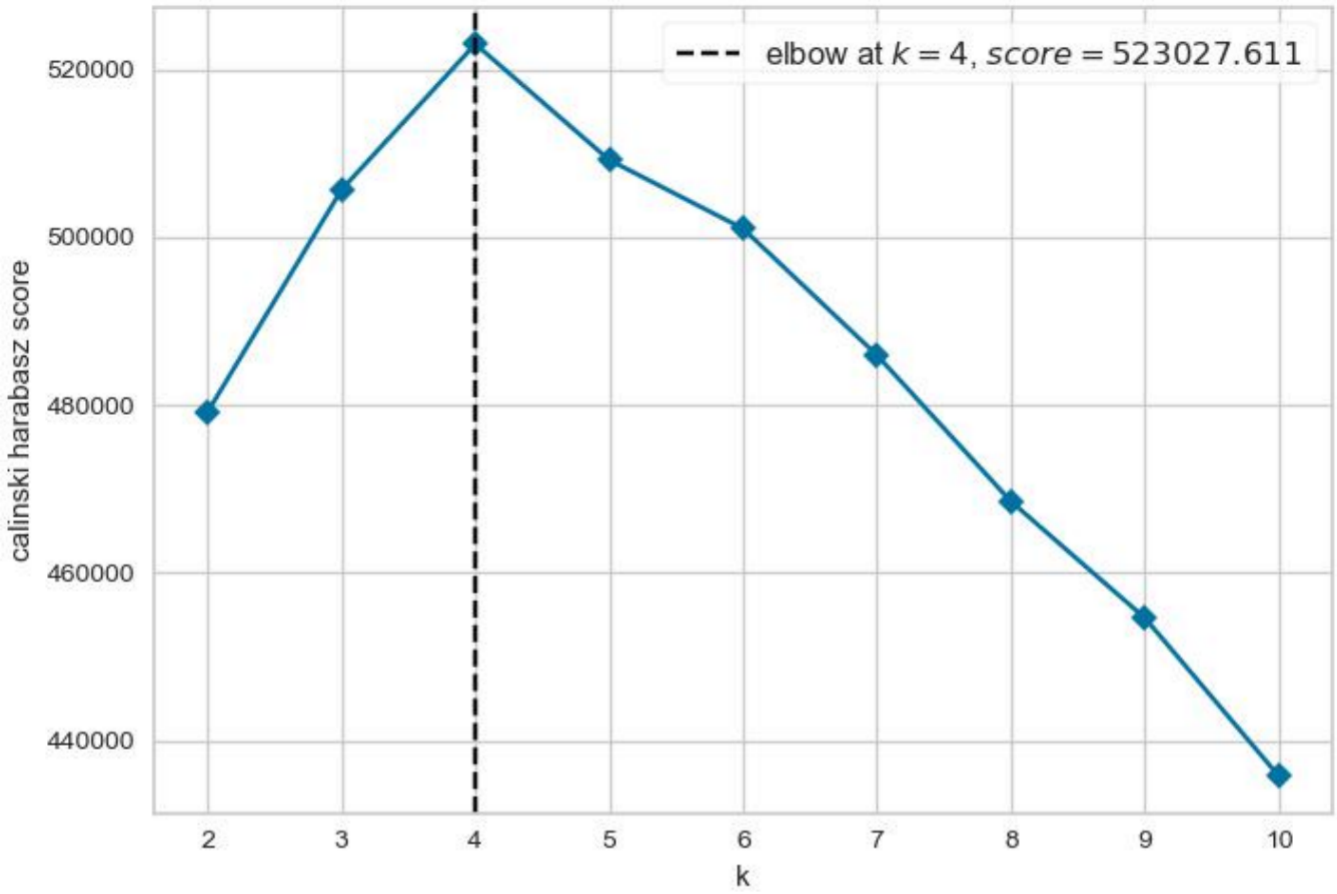


Figure S3. The result of Calinski-harabasz (CH) method

REFERENCES

Chai, T. and Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development* 7, 1247–1250. doi:10.5194/gmd-7-1247-2014

Goodwin, P. and Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting* 15, 405–408. doi:10.1016/S0169-2070(99)00007-2

Housman, I. W., Chastain, R. A., and Finco, M. V. (2018). An evaluation of forest health insect and disease survey data and satellite-based remote sensing forest change detection methods: Case studies in the United States. *Remote Sensing* 10. doi:10.3390/rs10081184

Renaud, O. and Victoria-Feser, M. P. (2010). A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference* 140, 1852–1862. doi:10.1016/j.jspi.2010.01.008

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association* 88, 1273–1283. doi:10.1080/01621459.1993.10476408

Zhao, Q. and Fränti, P. (2014). WB-index: A sum-of-squares based index for cluster validity. *Data and Knowledge Engineering* 92, 77–89. doi:10.1016/j.datak.2014.07.008

Water Biology and Security xxx (xxxx) xxx

Contents lists available at ScienceDirect

Water Biology and Security

journal homepage: www.keaipublishing.com/en/journals/water-biology-and-security

Identification of driving factors of algal growth in the South-to-North Water Diversion Project by Transformer-based deep learning

Jing Qian [a], Nan Pu [b], Li Qian [c], Xiaobai Xue [d], Yonghong Bi [e,*], Stefan Norra [f]

[a] Institute of Applied Geosciences, Karlsruhe Institute of Technology, Karlsruhe, 76131, Germany
[b] Institute of Advanced Computer Science, Leiden University, Leiden, 2333, CA, Netherlands
[c] Institute of Informatics, Ludwig Maximilian University of Munich, Munich, 80538, Germany
[d] MioTech Research, Yingtou Information Technology (Shanghai) Limited, Shanghai, 200120, China
[e] State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, 430072, China
[f] Institute of Environmental Sciences and Geography, Soil Sciences and Geoecology, Potsdam University Potsdam-Golm, 14476, Germany

ARTICLE INFO

Keywords:
Algal growth
Deep learning
Driving factor determination
Model interpretability
Transformer

ABSTRACT

Accurate and credible identification of the drivers of algal growth is essential for sustainable utilization and scientific management of freshwater. In this study, we developed a deep learning-based Transformer model, named Bloomformer-1, for end-to-end identification of the drivers of algal growth without the needing extensive a priori knowledge or prior experiments. The Middle Route of the South-to-North Water Diversion Project (MRP) was used as the study site to demonstrate that Bloomformer-1 exhibited more robust performance (with the highest R^2, 0.80 to 0.94, and the lowest RMSE, 0.22–0.43 μg/L) compared to four widely used traditional machine learning models, namely extra trees regression (ETR), gradient boosting regression tree (GBRT), support vector regression (SVR), and multiple linear regression (MLR). In addition, Bloomformer-1 had higher interpretability (including higher transferability and understandability) than the four traditional machine learning models, which meant that it was trustworthy and the results could be directly applied to real scenarios. Finally, it was determined that total phosphorus (TP) was the most important driver for the MRP, especially in Henan section of the canal, although total nitrogen (TN) had the highest effect on algal growth in the Hebei section. Based on these results, phosphorus loading controlling in the whole MRP was proposed as an algal control strategy.

1. Introduction

Algae, as a major footstone in the aquatic food chain, have a two-way and complex relationship with water quality. On the one hand, algae can affect water quality, since overgrowth and eventual death of algae cells can adversely influence water quality by producing toxic secondary metabolites and stench thereby affecting the survival of other aquatic organisms (Xia et al., 2019). On the other hand, algae can respond immediately to changes in physico-chemical properties of water, such as variations of temperature and nutrients, which can lead to changes in the species' qualitative and quantitative composition. Consequently, algae can often be used as reliable indicators for water quality assessment (Gökçe et al., 2016). However, increased knowledge and understanding of this relationship is necessary.

Modeling the interactions of algal biomass, expressed as chlorophyll-a(Chl-a) content, with multiple environmental factors based on a mathematical representation of the ecosystem is an effective approach to analyzing the relationship between water quality and algal growth, including process-based models and data-driven models (Su et al., 2022). Process-based models, such as the Lotka-Volterra model in ecology, are mathematical models that explicitly represent the processes occurring in the target system with equations. In the identification of the driving factors of algal growth, the process-based model is represented as an ecodynamic model that attempts to simulate process-based relationships by combining hydrodynamic processes with ecological processes and takes into account the interactions between multiple subsystems. Although ecodynamic models are capable of systematically representing relationships between a single output and multiple inputs, they usually require significant computational resource (Ralston and Moore, 2020). In addition, equations for process-based models are often derived from theory, but they are not necessarily credible (Knüsel and Baumberger, 2020), which leads to questionable correlations being obtained from the

* Corresponding author.
E-mail address: biyh@ihb.ac.cn (Y. Bi).

https://doi.org/10.1016/j.watbs.2023.100184
Received 12 December 2022; Received in revised form 21 February 2023; Accepted 19 April 2023
Available online xxxx

Please cite this article as: Qian, J. et al., Identification of driving factors of algal growth in the South-to-North Water Diversion Project by Transformer-based deep learning, Water Biology and Security, https://doi.org/10.1016/j.watbs.2023.100184

resulting models. In contrast, because this information is hidden in previous data, data-driven models can escape the limitations of theory and reveal patterns of interaction between algal growth and environmental factors from limited data and explain these patterns by correlation theory. Earlier data-driven approaches were empirical regression algorithms that used simple correlation and regression analyses to empirically model the relationship between a single water quality parameter (e.g., TP) and Chl-a (Xia et al., 2019). Since these models were generally unable to represent multi-factor interactions, multivariate analysis methods, such as cluster analysis (CA) and principal component analysis (PCA), were applied to explore algal growth (Bierman et al., 2011; Du et al., 2017; Qian et al., 2021). However, the relationship between environmental factors and algal biomass is, in many cases, non-linear (Nelson et al., 2018). As linear functions are the basis for most correlation coefficients and multivariate methods, they cannot be applied to nonlinear relationships (Su et al., 2022). In this context, machine learning has recently been widely used to understand aquatic ecological processes and to determine the strength of the association between environmental variables and algal growth (Yu et al., 2021; Ly et al., 2021; Deng et al., 2021).

Many studies have confirmed that traditional machine learning tools, such as support vector machine (SVM), logistic regression, extra trees regression (ETR), and multi-linear regression, are effective for the simulation of algal growth, (Su et al., 2022; Park et al., 2015; Liping and Binghui, 2013).As environmental research begins to migrate from small data to big data, the shortcomings of traditional machine learning is becoming more apparent, and deep learning, with its powerful big data processing capabilities, is receiving increased attention (Qian et al., 2022). Deep learning has been employed in previous studies to make predictions regarding Chl-a time series, but has rarely been applied to identify the critical factors associated with algal growth. This is because deep learning operates with less transparency than traditional machine learning and is implicitly expressive about the contributions of each factor. To solve this problem, deep learning models of algal growth are needed.

The Middle Route of the South-to-North Water Diversion Project (MRP) is a national large-scale project in China, which aims to transfer abundant water resources from the south to the north through artificial channels in order to balance the overall water distribution of the country (Zhu et al., 2022). The total length of the channel is 1432 km, including 155 km in Tianjin, serving a population of about 69 million people (Wang et al., 2021). As a long-distance and long-term drinking water supply corridor, water quality safety of the MRP is particularly important. Previous studies have shown that algal growth accelerated in parts of the MRP after 2016, with large clusters of filamentous algae causing problems such as blockage of the basin grate and rapid siltation in front of the outlet sluice (Zhu et al., 2019). Moreover, foul-smelling compounds and algal toxins produced by the siltation of decomposing algal debris also affected water quality levels and threatened water supply safety (Zhu et al., 2022). Consequently, during the 5–10 years since MRP operation, algal biomass has been a major factor affecting water quality. Furthermore, the instability of the overall system has made it difficult to identify the mechanisms and factors that determine algal growth in the MRP. It is noteworthy that most of the world's large water diversion projects are built for irrigation and power generation and that only a small percentage have provision of a drinking water supply as the main purpose (Long et al., 2022). The low attention to water quality changes in these large water diversion projects has resulted in a lack of case studies that can be applied to the management of water quality safety in MRP. Therefore, at this stage, the accurate identification of mechanisms that influence water quality and algae in MRP is lacking. Nevertheless, the effective prediction and management of algal growth are important for success of long distance and long-term drinking water delivery projects such as MRP.

This study aims to accurately and quantitatively identify the driving factors of algal biomass in the MRP with the core of big data mining. Our method involves developing a Transformer-based deep learning model, named Bloomformer-1, which runs on a big data platform derived from long-term manual monitoring data, in order to reveal the driving mechanisms of algal growth in the MRP accurately, transparently, and directly. The findings will be useful for the efficient management and sustainable utilization of the MRP.

2. Materials and methods

2.1. Study area and data collection

A total of nine water quality monitoring stations were evenly spaced along the MRP, labeled P1 to P9, extending from south to north, with P1, P2, P3 and P4 located in the Henan section, P5, P6 and P7 located in the Hebei section, P8 located in the Tianjing section, P9 located in the Beijing Section (Fig. 1). The database used in this study consists of 49 months (August 1, 2018, to August 30, 2022) of water quality monitoring data from each station. Water samples were collected at a depth of 0.5 m, stored at 4 °C, and transported to the laboratory to determine water quality parameters.

The chemical water quality parameters, which comprised total phosphorus (TP), phosphorous-phosphate (PO_4-P), total nitrogen (TN), nitrogen-nitrate (NO_3-N), nitrogen-ammonia (NH_3-N), potassium permanganate index (COD_{Mn}), and total organic carbon (TOC), were determined according to APHA (Zhu et al., 2022). The concentration of Chl-a was used as a response variable in the data-driven methods since it is considered to be an indicator of phytoplankton biomass and was determined according to ASTM D3731-87 (ASTM, 1993).

2.2. Bloomformer-1 model

Transformer is the state-of-the-art solution for natural language processing (NLP) tasks (Wolf et al., 2020). This method takes advantage of the Multi-Head Attention mechanism, which compares each token along the input sequence to other tokens in order to collect and learn dynamic contextual information. Attention is an important part of human cognitive function (Lindsay, 2020), and when faced with large amounts of information, humans can readily adjust the level of focus on the information they received to analyze it more accurately and efficiently. The essence of the attention mechanism was to provide weights. An attention function could be interpreted as mapping a Q (query) and a string of K(key)-V(value) to an output, where Q, K, V, and output were vectors (Vaswani et al., 2017). The attention could be represented as:

$$Output_{Attention} = Attention(Q, K, V)$$

Multi-Head Attention was the projection of Q, K, and V by h different linear transformations. The different attention results were then stitched together, which could be represented as:

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)\, W^O$$

where

$$\mathrm{head_i} = \mathrm{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

$$W_i^Q \in \mathbb{R}^{d_{\mathrm{model}} \times d_k},\ W_i^K \in \mathbb{R}^{d_{\mathrm{model}} \times d_k},\ W_i^V \in \mathbb{R}^{d_{\mathrm{model}} \times d_v} \text{ and } W^O \in \mathbb{R}^{hd_v \times d_{\mathrm{model}}}$$

In the appealed Attention mechanism, the weights were the direct weight correspondence between the input and output vectors, implying that the weight calculation required the participation of the output vectors. In contrast, the weight of Self-Attention was a weight relationship between the input vectors internally, which did not require the

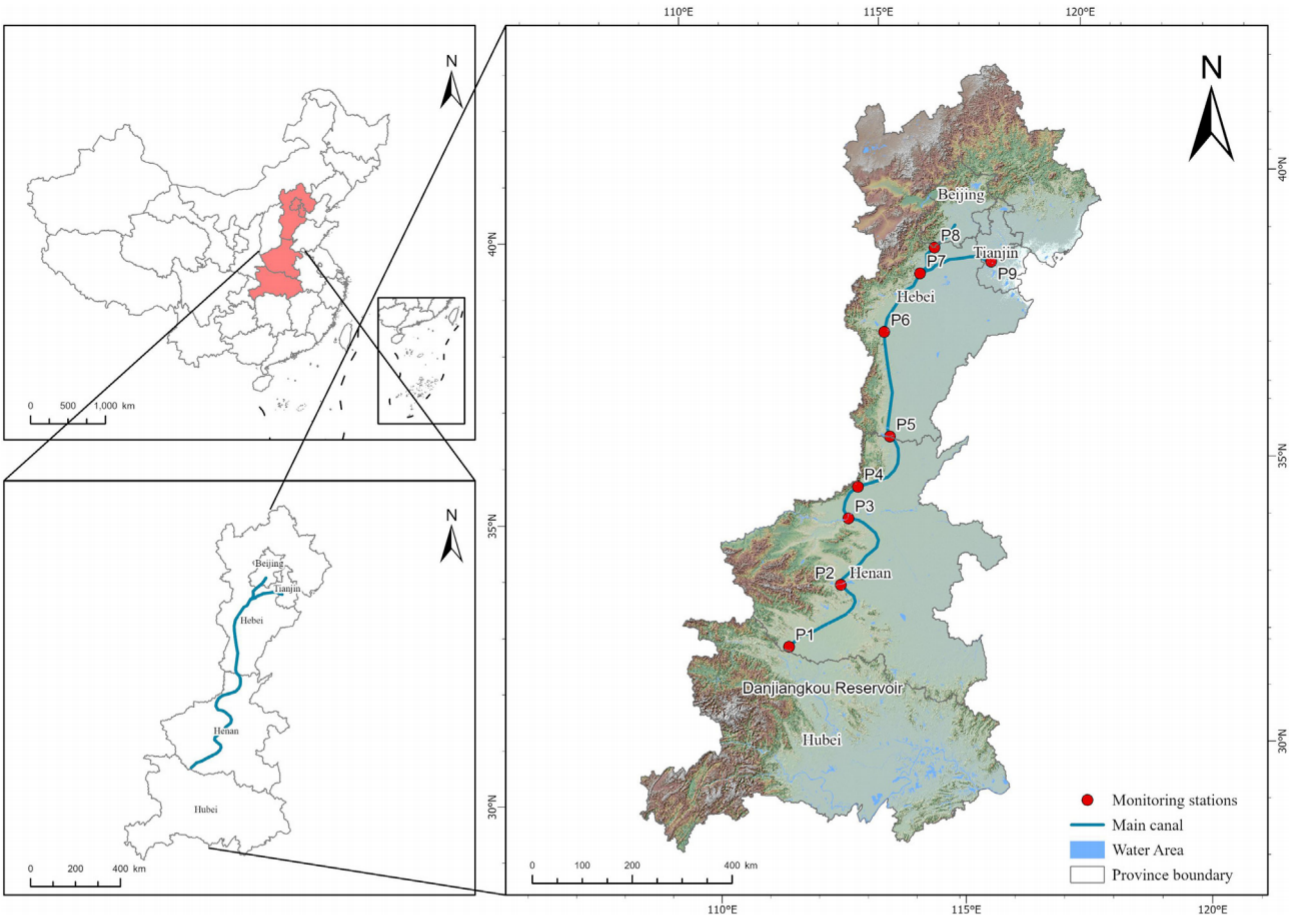


Fig. 1. Sketch map of sampling stations distribution in the middle section of the South-North Water Diversion Project.

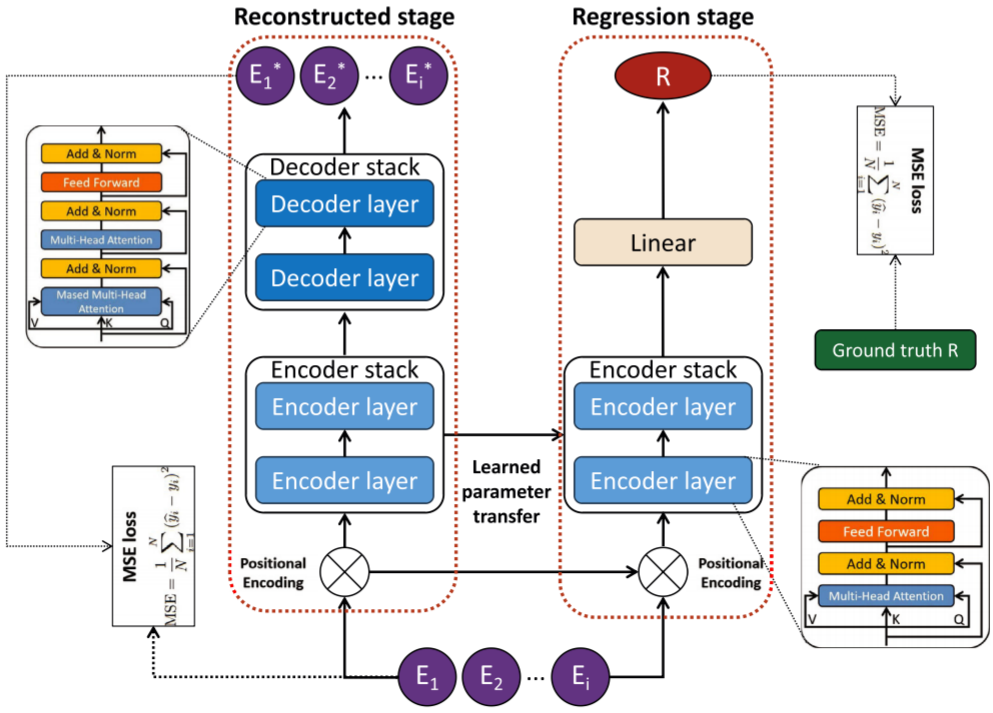


Fig. 2. The framework and architecture of Bloomformer-1.

participation of the output vectors. Therefore, the multi-head-self-attention meant Q, K, and V were the same.

In this study, we used the scaled dot-product to calculate Attention:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where d_k was the vector dimension in both Q and K.

The encoder consisted of N same units (Fig. 2). Each unit consists of two sub-layers, the multi-head-self-attention layer, and the fully connected feed-forward network, where each sub-layer was processed with the residual connection "Add" and normalization "Norm". The output of the sub-layer could be represented as:

$$Output_{Sublayer} = Norm(x + F(x))$$

Where $F(x)$ was a function of the sublayer itself, multi-head-self-attention, or fully connected feed-forward network.

The fully connected feed-forward network provided a non-linear transformation consisting of two linear transformations with the active function ReLu (Agarap, 2018). Compared with the encoder, the decoder added another MHSA layer (Fig. 2). A mask operation (Fan et al., 2021) was applied to this multi-head self-attention layer in order to prevent the model from being exposed to future information during training.

Because algal growth is a multi-factorial process, the determination of the driving factors of algal growth is a typical multivariate regression task. The key to solving this problem is to learn the spatial relationships to understand how the variables are related to each other. However, the standard Transformer is not designed for this because it treats the value of each variable at a given time period as a single marker on its graph: each variable cannot have its own view of the context it should prioritize (Grigsby et al., 2021). Therefore, we developed Bloomformer-1 for studying spatial relationships based on Transformer. The improved method first converted the context sequence in the database into a long spatial sequence. This sequence was also transposed to obtain the corresponding long spatial sequence. The sequence was then processed with a Transformer-based encoder-decoder architecture to obtain the predicted values for each variable. Finally, the predicted values were repackaged into their original format and trained to minimize prediction error metrics. The training framework of Bloomformer-1 consists of a reconstruction stage and a regression stage. The reconstruction task is an unsupervised pre-training and a reconstruction of the explanatory variables through the connected encoder and decoder stacks to extract their robust and compact features. The parameters of the encoder stack and position encoding obtained by the reconstruction task are shared with the corresponding part of the regression task. In this study, the number of units in encoder and decoder layer is 8, which represented the 7-dimensional water quality parameters and the 1-dimensional station location information. When performing the substation task, the station location information was the station number corresponding to each water quality parameter, from 1 to 9. When performing the whole MRP task, the station location information was set to 1. Mean square error (MSE, Supplementary material) was selected as the loss function both in the reconstructed stage and the regression stage. The framework and architecture of Bloomformer-1 is shown in Fig. 2. The MHSA mechanism of Bloomformer-1 allows the results of driving factor identification to be obtained during model training forward propagation direction and simultaneously derived.

2.3. Multiple linear regression

Multiple linear regression (MLR) is one of the typical traditional machine learning models that can be used to predict the result of an answer variable using a number of explanatory variables (Maulud and Abdulazeez, 2020). For the purpose of verifying performance, an MLR model was used in this study to compare with Bloomformer-1. The MLR model was built by using the Scikit-learn function from the Python package. The parameter to be tuned was the degree of the polynomial features. The driving factor analysis methods for MLR was sensitivity analysis (SA) (Saltelli, 2002).

2.4. Support vector regression

Support vector regression (SVR) is a powerful traditional learning machine for searching the relationship between the answer variable and several explanatory variables, including linear and non-linear correlations. The SVM approach is to map the training data non-linearly into a high-dimensional feature space and then construct a separated hyperplane there with maximum margin (Awad and Khanna, 2015). This study employed the SVR as a comparative model to assess the performance of Bloomformer-1. The SVR model was derived by calling the function in the Scikit-learn package in Python. Radial basis functions were selected as kernels because they provided better performance through the kernel test. The parameters that needed to be tuned in this study were the regularization parameter and the Kernel coefficient. The driving factor analysis methods for SVR was sensitivity analysis (SA) (Saltelli, 2002).

2.5. Gradient boosting regression tree

The gradient boosting regression tree (GBRT) algorithm is a combination of the classification and regression (CART) algorithm and the gradient boosting (GB) algorithm (He et al., 2013). CART allows for the modeling of non-linear relationships without requiring a priori information about the probability distribution of the variables (Nie et al., 2021). The gradient boosting algorithm combines weak learners by iteratively focusing on the error generated at each step until a suitable strong learner is obtained as a sum of successive weak learners (Friedman, 2001). The regression tree generated by the CART algorithm was used as the weak learner and was added to the model to correct errors in the previous model, thereby improving the accuracy of the model. This study employed GBRT as a comparative model to assess the performance of Bloomformer-1. The GBRT model was derived by calling the function in the Scikit-learn package in Python. The driving factor analysis methods for GBRT is to calculate the relative importance to the input variables, the idea being to score each input variable by estimating the reduction in relative variance (Su et al., 2022).

2.6. Extra trees regression

Extra trees regression (ETR) builds a collection of the unpruned decision or regression trees based on a classical top-down procedure that does not require a known underlying distribution of parameters or associated assumptions (Geurts et al., 2006). The main difference between this method and traditional tree ensemble methods is that it splits the nodes randomly and grows the tree based on the original training data set rather than using a bootstrap method. With these two features, ETR is able to produce outputs with lower variance and higher generalization than traditional tree-based models. In this study, the ETR was used to evaluate the performance of Bloomformer-1 as a comparative model. The ETR model was derived by calling the function in the Scikit-learn package in Python. As for GBRT, the driving factor analysis methods for ETR is to calculate the relative importance to the input variables (Su et al., 2022).

2.7. Training and performance evaluation of model

Data from each of the nine water quality monitoring stations (P1 to P9) were fed into the appealing model for training to identify the drivers of algal growth at each water quality monitoring station. Chl-a and the other water quality parameters described previously were placed in the models as responses and explanatory variables, respectively. Before entering all data into the model, data normalization was performed to

ensure equality in model comparisons. Data normalization followed the Z-equation (See Supplementary material).

Evaluation of model performance is a critical step prior to practical application. The data set was divided into a training set and a test set according to the rule of randomly taking one step out of every five, which means 80% of the whole data set was used to train the model and 20% was used to test the model performance. A tenfold cross-validation was introduced to avoid over-fitting in the training phase. For the purpose of evaluating the accuracy and stability of each regression model, two indicators were used on the test set: coefficient of determination (R^2) and root mean square error (RMSE), following the equations:

$$R^2 = 1 - \frac{\sum_{i=0}^{n} (y_i - \tilde{y}_i)^2}{\sum_{i=0}^{n} (y_i - \bar{y}_i)^2}$$

$$RMSE = \sqrt{\sum_{0}^{n-1} \frac{(y_i - \tilde{y}_i)}{n}}$$

2.8. Operation environment

The experiment was carried out on a PC with the following features: Hard-ware: CPU i7-6950X, RAM 64GB, dual GeForce RTX 3090, VRAM 24GB; Software: Ubuntu 20.04, Python3.6, Pytorch 1.10.0, Numpy 19.2.

3. Results

3.1. Model performance evaluation

The performance of regression models directly determined the accuracy and plausibility of the driver identification. After optimizing the proposed models, we compared the performance of five machine learning models across all monitoring stations using R^2 and RMSE in a tenfold cross-validation. The results for model performance are summarized in Table 1. The comparison between model simulation and the ground truth is shown in Figs. 3 and 4. In order to describe the training process of Bloomformer-1 more intuitively, the loss values during the training process are shown in Fig. S1.

The results of P1, P2, and P3 showed that Bloomformer-1 performed much better than the four traditional machine learning models because the difference in R^2 values between them was greater than 0.1. There was also a significant difference in RMSE values (e.g., in P1, Bloomformer-1 had an R^2 value of 0.85, while the four traditional machine learning models had R^2 values less than or equal to 0.75; Bloomformer-1 had an RMSE value of 0.32, while the other models had RMSE values greater than or equal to 0.56. The RMSE value of Bloomformer-1 was 0.32, while the RMSE values of the other models were all greater than or equal to 0.56).

Table 1
Results of model performance evaluation.

Stations	Indicator[a]	Bloomformer-1	ETR	GBRT	SVR	MLR
P1	R^2	**0.85**	0.75	0.72	0.63	0.42
	RMSE	**0.32**	0.56	0.57	0.60	0.73
P2	R^2	**0.80**	0.66	0.51	0.63	0.25
	RMSE	**0.43**	0.62	0.68	0.63	0.82
P3	R^2	**0.83**	0.70	0.39	0.58	0.39
	RMSE	**0.40**	0.59	0.69	0.64	0.79
P4	R^2	**0.89**	0.84	0.68	0.46	0.35
	RMSE	**0.33**	0.52	0.62	0.61	0.76
P5	R^2	**0.90**	0.89	0.78	0.88	0.49
	RMSE	**0.30**	0.50	0.58	0.51	0.71
P6	R^2	**0.89**	0.85	0.74	0.88	0.46
	RMSE	**0.26**	0.45	0.49	0.43	0.68
P7	R^2	**0.94**	**0.94**	0.85	0.92	0.68
	RMSE	**0.23**	0.43	0.47	0.45	0.66
P8	R^2	**0.94**	0.91	0.84	0.89	0.71
	RMSE	**0.22**	0.43	0.48	0.44	0.62
P9	R^2	**0.93**	0.91	0.89	0.86	0.62
	RMSE	**0.28**	0.46	0.48	0.49	0.68
Whole MRP	R^2	**0.85**	0.79	0.73	0.80	0.39
	RMSE	**0.35**	0.54	0.55	0.51	0.70

The bold values represent the best regression results.
[a] Unit of RMSE is $\mu g/L$.

According to the results of P4, P5, P6, P8, and P9, Bloomformer-1 showed relatively high performance. Although the difference with ETR in R^2 values was small (0.03–0.06), it still had a significant advantage in RMSE values (e.g., Bloomformer-1 had an RMSE value of 0.33 in P4, while ETR had the lowest RMSE of 0.52 among the four traditional machine learning models). In P7, except for MLR, the other three traditional machine learning models showed better performance, especially the R^2 value of ETR which was the same as Bloomformer-1 at 0.94. However, Bloomformer-1 still had a significant advantage in RMSE values (Bloomformer-1 0.23, ETR 0.43, GBRT 0.47, SVR 0.45, MLR 0.66). Consistent with the results from the individual stations, Bloomformer-1 showed superior performance on the whole MRP, as evidenced by the higher R^2 values (0.85) and lower RMSE values (0.35). In summary, Bloomformer-1 showed the highest R^2 with the lowest RMSE across all stations compared to traditional machine learning models and was, therefore, the best model in terms of performance to describe the relationship between Chl-a concentration and the water quality parameters.

3.2. Driving factors of algal growth

The driving factors of algal growth in the MRP based on the attention mechanism of Bloomformer-1 are shown in Fig. 5. In P1, P2 and the whole MRP, the most dominant driving factor of algal growth was TP, with 18.73%, 19.20% and 22.28%, respectively. It is noteworthy that $PO_4 - P$ also exhibited a very close occupancy rate in the whole MRP, at 16.09%. The results for P5, P6, P8, and P9 showed that the major driving factor of algal growth at these four stations was $NO_3 - N$ with 20.24%, 28.27%, 20.16%, and 17.16%, respectively. In P4 and P7, TN was the main driving factor of algal growth, with 22.16% and 17.96%, respectively. The results of P3 differed from the others, with 23.84% of $NH_3 - N$ as the most dominant driving factor of algal growth.

4. Discussion

4.1. Model performance

Inferring causation from correlation and determining the explanatory variables associated with the response variables is the basis for traditional model building, which requires a great deal of a priori knowledge and background information about the domain (Xia et al., 2019; Su et al., 2022). In traditional machine learning, feature extraction is manual-based and has limited learning capability, thus requiring the input terms (explanatory variables) have a clear one-way correlation with the response variables, which implies a high reliance on a priori knowledge. However, some explanatory variables are difficult to determine in practical applications, such as COD_{Mn} in this study. The relationship between COD_{Mn} and algal growth is bidirectional and complex (Li et al., 2020; Yan et al., 2016). The foundation of COD_{Mn} as an explanatory variable depends on which direction of the relationship is dominant, which requires a priori knowledge as well as prior experiments. Bloomformer-1 employs a combination of encoder and decoder structures as well as the MHSA mechanism to automatically extract features from raw data and to fully understand the raw data at the same time. This full understanding means that the complex relationship between COD_{Mn} and algal growth in the raw data is mined and quantified. In this way, a rigorous correlation analysis is not required before using Bloomformer-1. Moreover, building a model with excellent fitting performance is the first and most critical step to identify the driving factors

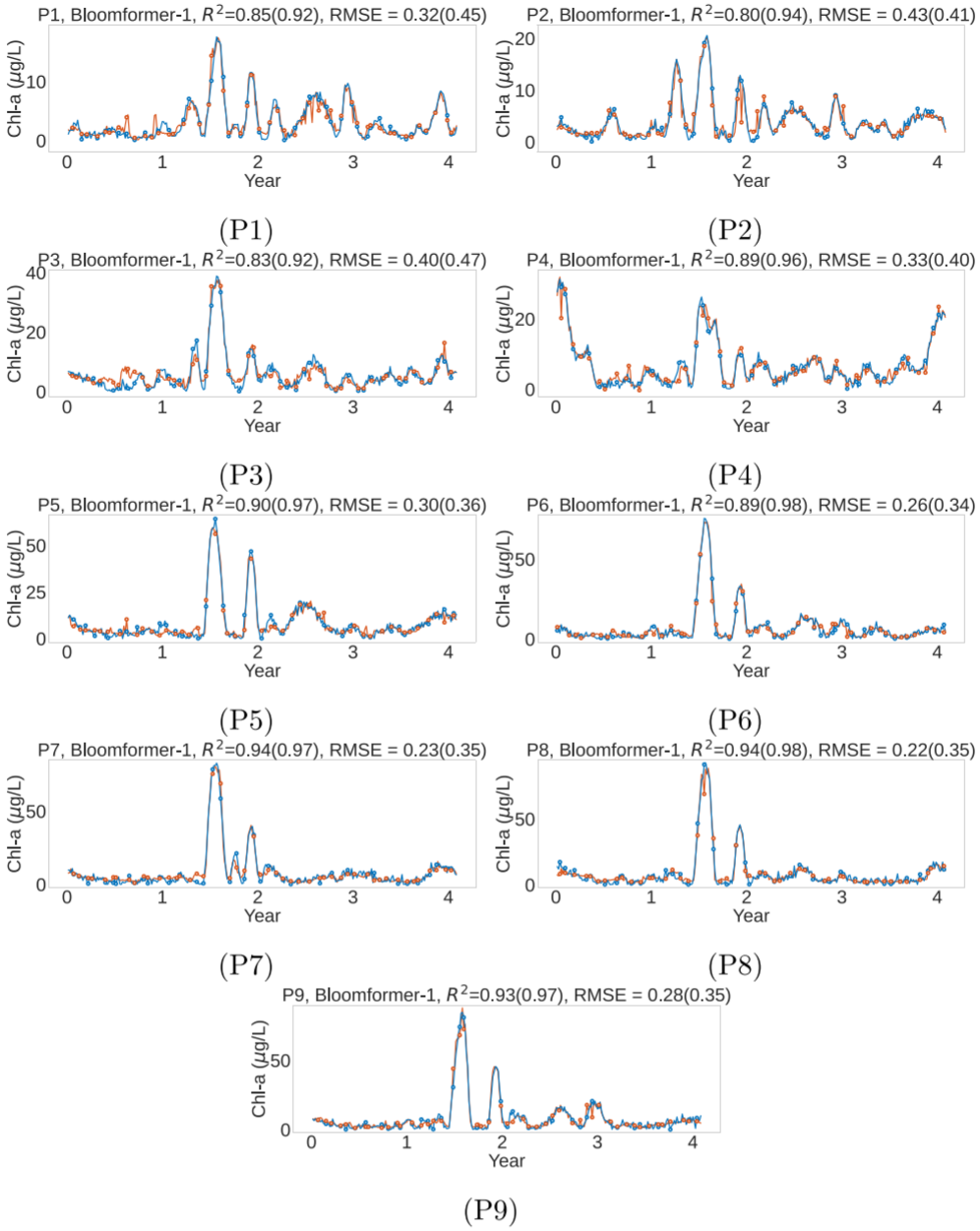


Fig. 3. Performance of Bloomformer-1 in P1–P9 (blue lines are observations, red lines are model simulations). The circles are the test set, where the blue circles are the true values and the red circles are the predicted values. The blue line, except for the blue circles, is the training set. Numbers show RMSE and R^2 for model prediction and training data (inside brackets).

of response variables. As a state-of-the-art deep learning model, Bloomformer-1 has an advantage in the accuracy of model fitting with R^2 (0.80–0.94). Compared with traditional machine learning, deep learning is more advanced and has a stronger learning ability to automatically extract, analyze and understand useful information from raw data to obtain better results (Chauhan and Singh, 2019; Janiesch et al., 2021).

In the present study, when training traditional machine learning models, each explanatory variable was completely independent, for example, each decision tree that made up the ETR was unrelated to each other. This meant that the traditional machine learning models only focused on the logical relationship between each explanatory variable and the corresponding variable, ignored the additional effects of the interactions between explanatory variables on the corresponding variable. Consequently, the traditional machine learning models could only partially identify the drivers of algal growth, because algal growth is not only related to a single water quality parameter, but also to the interactions between multiple water quality parameters in different spatial-temporal dimensions. The Transformer structure in Bloomformer-1 had the MHSA mechanism that could simultaneously focus on all relationship changes (Vaswani et al., 2017). Therefore, Bloomformer-1 can identify the drivers more reliably than traditional machine learning models.

4.2. Model interpretability

Model interpretability represents trustworthiness (Ridgeway et al., 1998), which can be expressed in terms of transferability and understandability (Lipton, 2016).

Transferability represents the ability to transfer learned skills to unfamiliar environments, especially in non-stationary environments (Lipton, 2016). In this study, Bloomformer-1 outperformed four traditional machine learning models on the test data set and was able to easily cope

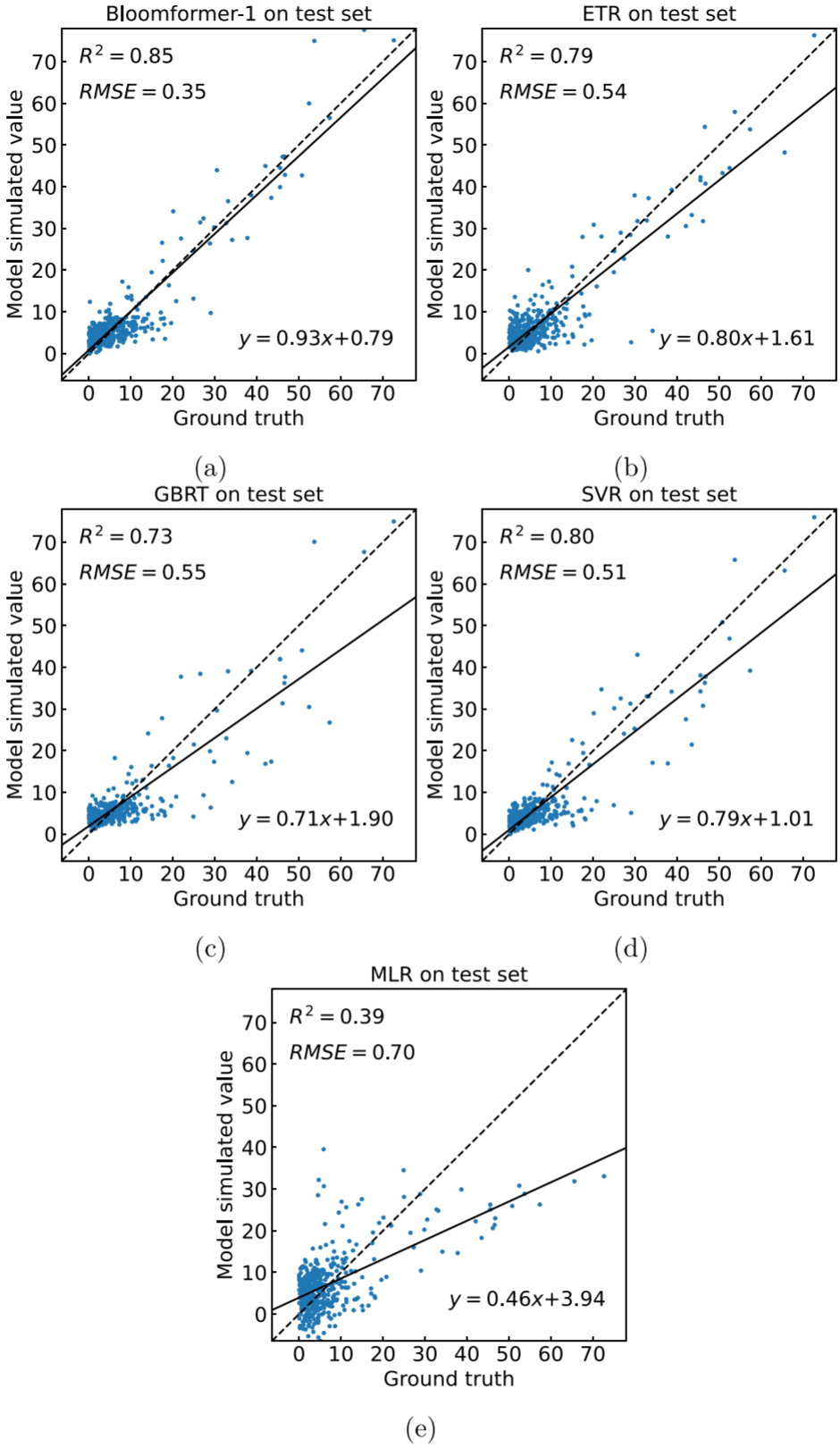


Fig. 4. Model performance evaluation in the whole MRP, where (**a**), (**b**), (**c**), (**d**) and (**e**) represent the test results of the Bloomformer-1, ETR, GBRT, SVR and MRL, respectively.

with abrupt changes in Chl-a concentration whereas traditional machine learning models were unable to do so (e.g., P3 in February 2020). These findings demonstrate that Bloomformer-1 has superior transferability.

Understandability represents our ability to understand how a model works (Lipton, 2016). When dealing with multidimensional variables, SVR is difficult to understand because the human brain is unable to visualize the hyperplane when the number of variables have more than three dimensions. Both GBRT and ETR also showed low

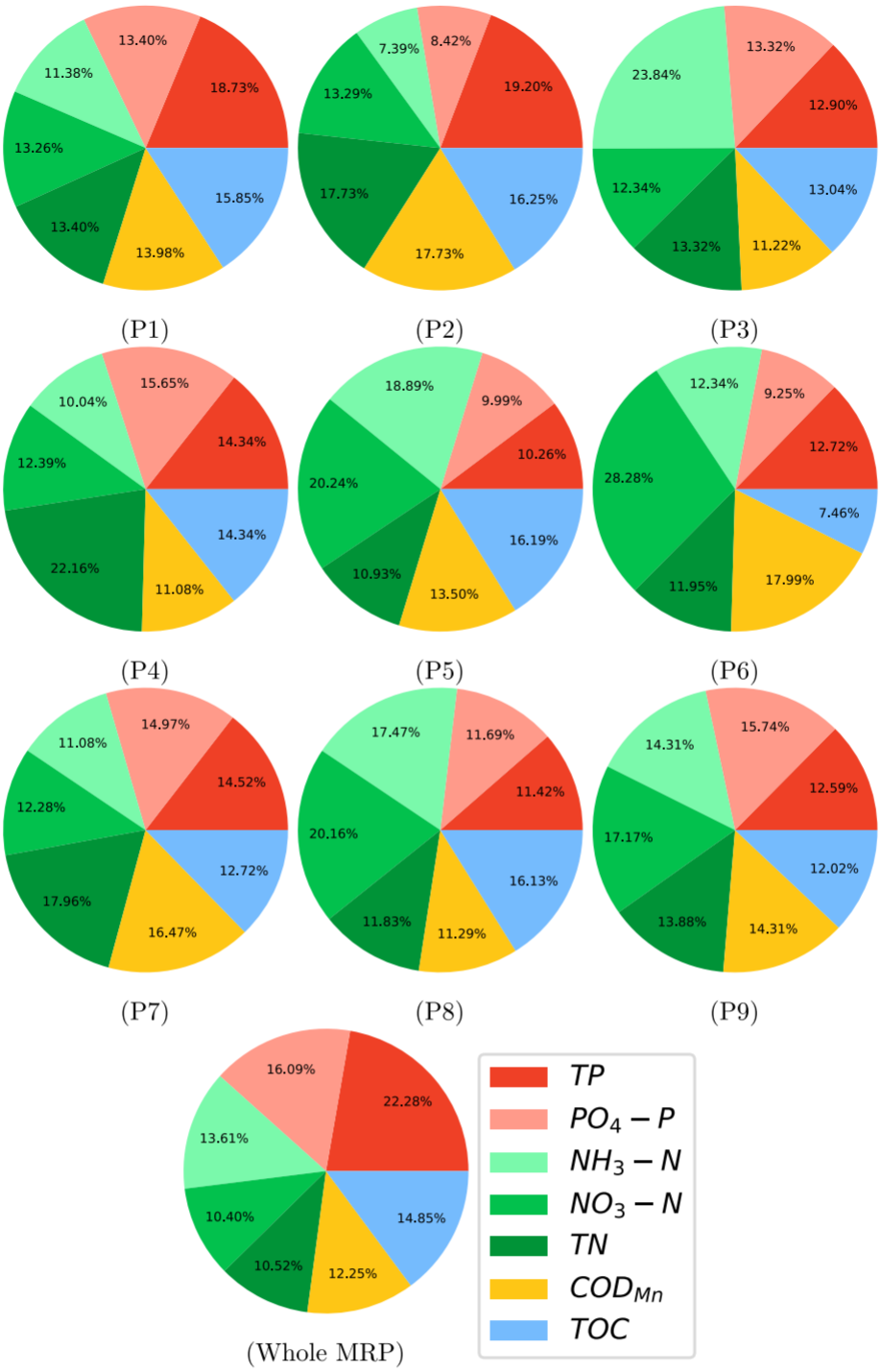


Fig. 5. Driving factors of algal growth at each of the sampling stations (P1–P9), and in the whole MRP, based on Bloomformer-1 modelling.

understandability. The direction and/or shape of covariate effects usually cannot be obtained by the simple interpretation of GBRT (Welchowski et al., 2022). ETR uses the same principles as Random Forest, except that the selection of attributes and cut points is strongly randomized when splitting the tree nodes (Geurts et al., 2006). Random forest is considered as a black box model in many studies (Wright, 2018), so ETR based on the same principle can also be considered as a black box model. On the contrary, Bloomformer-1 possessed a relatively high level of understandability. First, Bloomformer-1 worked by synthesizing the computational resources derived from the analysis and continuously adjusting the weights of each computational resource to obtain the desired results. This work pattern imitates that of humans and is therefore easy to understand. Secondly, the attribution algorithm (Hao et al., 2021) of the self-attentive mechanism could provide an interpretable description of the information interactions within Bloomformer-1 and construct attribution trees to visualize the direct information interactions in different layers. As a result, Bloomformer-1 has a high degree of interpretability, and the obtained results are highly applicable to real-world situations.

4.3. *Driving factors of algal growth*

Nutrients play a vital role in algal growth, in particular their supply and its variability affect algal biomass and net productivity (Yang et al., 2016; Koeller et al., 2009). Among them, nitrogen (N) and phosphorus

(P) are essential elements for algae (Hecky and Kilham, 1988). Nitrogen to phosphorus ratios (N:P) are often used to determine the nutrient limitation status of water bodies (Redfield, 1963), but difficulties remain because the optimal N:P ratio varies considerably, i.e., from 4 to 133, for different water bodies (Klausmeler et al., 2004). Previous studies on MRP have recognized phosphorus as the limiting factor for algal growth, but it was not definitive that it was the most critical nutrient limitation (Nong et al., 2020). The results of this study indicated that TP was the most critical factor in the whole MRP. These results agreed with other studies on algal growth and further confirmed the driving role of nutrients on algal growth.

Although the water quality of the MRP has been good and stable since 2014, the nutrient load has been increasing. Besides the increasing nutrient load of Danjiangkou reservoir, the rain runoff, dry and wet deposition along the channel were the important mechanisms of nutrient input (Wang et al., 2021; Nong et al., 2020). Inundation of farmland and mountainous areas led to the release of nitrogen, phosphorus and other nutrients from the soil into the water, resulting in increased nutrient concentrations in the Danjiangkou Reservoir. In recent years, rainfall along the MRP has increased and this, coupled with dry and wet deposition, has resulted in more nutrients, both from the land and the air, being deposited into the MRP, which made the rich material basis for algae rapid growth. It could be deduced that nutrient control, especially phosphorus, should be important strategy for controlling algal growth and maintaining water quality stability.

5. Future work

Bloomformer-1, as an advanced deep learning model, has obvious performance advantages over traditional machine learning models in processing high volume as well as high dimensional data (Fig. S2). As the database used in this study has medium capacity and dimensionality, the potential of Bloomformer-1 was not fully realized, which was also why traditional machine learning models were able to perform well in some scenarios. In addition, due to the complexity and size of the MRP, a deeper understanding of the relationship between algal growth and water quality is necessary. Therefore, future work should focus on building databases with higher data capacity and dimensionality (including collecting physical and hydrological data), increasing the density of monitoring stations, and using automated monitoring equipment. Using such databases, Bloomformer-1, with its excellent self-learning capability, could make more relevant and timely conclusions regarding the management of algal growth in the MRP.

6. Conclusion

Bloomformer-1, a deep learning-based Transformer model for end-to-end identification of the drivers of algal growth without the need for extensive prior knowledge and prior experiments, achieved the highest R^2 (0.80–0.94) and lowest RMSE (0.22–0.43 μg/L) on both individual subsites and full-line simulations in the MRP compared with traditional machine learning models, namely ETR, GBRT, SVR and MLR. Bloomformer-1 also had higher interpretability, implying that Bit was trustworthy and that the results obtained from this model could be directly applied to real-world scenarios. TP was the most important driver for the MRP. Phosphorus control and reduction would be an important strategy for controlling algal growth and maintaining water quality stability in the MRP.

Funding

This research was Jointly funded by National Key R&D plan (No.2021YFC3200900) and National Natural Science Foundation of China (No.31971477).

Credit author statement

Conceptualization: Jing Qian.
Data curation: Jing Qian, Li Qian.
Formal analysis: Jing Qian and Nan Pu.
Funding acquisition: Yonghong Bi and Stefan Norra.
Investigation: Jing Qian, Nan Pu, Li Qian and Xiaobai Xue.
Methodology: Jing Qian.
Project administration: Yonghong Bi and Stefan Norra.
Resources: Yonghong Bi.
Software: Jing Qian, Nan Pu and Li Qian.
Supervision: Stefan Norra and Yonghong Bi.
Visualization: Jing Qian, Nan Pu and Li Qian.
Writing – original draft: Jing Qian.
Writing –review & editing: Stefan Norra and Yonghong Bi.

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

We appreciate the help from Yuxuan Zhu and Gang Ruan with the experiments. We would also like to thank Di Wang for proofreading.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.watbs.2023.100184.

References

Agarap, A.F., 2018. Deep learning using rectified linear units (ReLU). arXiv e-prints arXiv:1803.08375arXiv:1803.08375.
ASTM, 1993. Standard Practices for Measurement of Chlorophyll Content of Algae in Surface Waters, pp. 3731–3787.
Awad, M., Khanna, R., 2015. Support vector regression. In: Efficient Learning Machines. Springer, pp. 67–80.
Bierman, P., Lewis, M., Ostendorf, B., Tanner, J., 2011. A review of methods for analysing spatial and temporal patterns in coastal water quality. Ecol. Indicat. 11 (1), 103–114. https://doi.org/10.1016/j.ecolind.2009.11.001.
Chauhan, N.K., Singh, K., 2019. A review on conventional machine learning vs deep learning. In: 2018 International Conference on Computing, Power and Communication Technologies, GUCON, vol. 2018, pp. 347–352. https://doi.org/10.1109/GUCON.2018.8675097.
Deng, T., Chau, K.W., Duan, H.F., 2021. Machine learning based marine water quality prediction for coastal hydro-environment management. J. Environ. Manag. 284 (December 2020), 112051. https://doi.org/10.1016/j.jenvman.2021.112051.
Du, X., Cai, Y., Wang, S., Zhang, L., 2017. Overview of deep learning. In: Proceedings - 2016 31st Youth Academic Annual Conference of Chinese Association of Automation, vol. 2016. YAC, pp. 159–164. https://doi.org/10.1109/YAC.2016.7804882.
Fan, Z., Gong, Y., Liu, D., Wei, Z., Wang, S., Jiao, J., Duan, N., Zhang, R., Huang, X., 2021. Mask attention networks: rethinking and strengthen trans-former. In: NAACL-HLT 2021-2021 Conference of the North American Chap-Ter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pp. 1692–1701. https://doi.org/10.18653/v1/2021.naacl-main.135 arXiv:2103. 13597.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting ma- chine. Ann. Stat. 29 (5), 1189–1232. https://doi.org/10.1214/aos/1013203451.
Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63 (1), 3–42. https://doi.org/10.1007/s10994-006-6226-1.
Gökçe, D., 2016. Algae as an indicator of water quality. In: Thajuddin, N., Dhanasekaran, D. (Eds.), Algae, IntechOpen, Rijeka, Ch. 4, pp. 81–101. https://doi.org/10.5772/62916.
Grigsby, J., Wang, Z., Qi, Y., 2021. Long-range Transformers for Dynamic Spatiotemporal Forecasting arXiv preprint arXiv:2109.12218.
Hao, Y., Dong, L., Wei, F., Xu, K., 2021. Self-attention attribution: interpreting information interactions inside transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 12963–12971.
He, Q., Kamarianakis, Y., Jintanakul, K., Wynter, L., 2013. Incident duration pre-diction with hybrid tree-based quantile regression. In: Advances in Dynamic Network Modeling in Complex Transportation Systems. Springer, pp. 287–305.
Hecky, R.E., Kilham, P., 1988. Nutrient limitation of phytoplankton in freshwater and marine environments: a review of recent evidence on the effects of enrichment.

Limnol. Oceanogr. 33 (4part2), 796–822. https://doi.org/10.4319/lo.1988.33.4part2.0796.
Janiesch, C., Zschech, P., Heinrich, K., 2021. Machine learning and deep learning. Electron. Mark. 31 (3), 685–695. https://doi.org/10.1007/s12525-021-00475-2.
Klausmeler, C.A., Litchman, E., Daufreshna, T., Levin, S.A., 2004. Optimal nitrogen-to-phosphorus stoichiometry of phytoplankton. Nature 429 (6988), 171–174. https://doi.org/10.1038/nature02454.
Knüsel, B., Baumberger, C., 2020. Understanding climate phenomena with data-driven models. Stud. Hist. Philos. Sci. 84, 46–56. https://doi.org/10.1016/j.shpsa.2020.08.003.
Koeller, P., Fuentes-Yaco, C., Platt, T., Sathyendranath, S., Richards, A., Ouellet, P., Orr, D., Skúladóttir, U., Wieland, K., Savard, L., Aschan, M., 2009. Basin-scale coherence in phenology of shrimps and phytoplankton in the North Atlantic Ocean. Science 324 (5928), 791–793. https://doi.org/10.1126/science.1170987.
Li, Y., Nwankwegu, A.S., Huang, Y., Norgbey, E., Paerl, H.W., Acharya, K., 2020. Evaluating the phytoplankton, nitrate, and ammonium interactions during summer bloom in tributary of a subtropical reservoir. J. Environ. Manag. 271 (May), 110971. https://doi.org/10.1016/j.jenvman.2020.110971.
Lindsay, G.W., 2020. Attention in psychology, neuroscience, and machine learn-ing. Front. Comput. Neurosci. 14 (April), 1–21. https://doi.org/10.3389/fncom.2020.00029.
Liping, W., Binghui, Z., 2013. Prediction of chlorophyll-a in the Daning River of Three Gorges Reservoir by principal component scores in multiple linear regression models. Water Sci. Technol. 67 (5), 1150–1158. https://doi.org/10.2166/wst.2013.679.
Lipton, Z.C., 2016. The Mythos of Model Interpretability, arXiv E-Prints arXiv:1606.03490arXiv:1606.03490.
Long, Y., Feng, M., Li, Y., Qu, J., Gao, W., 2022. Comprehensive risk assessment of algae and shellfish in the middle route of South-to-North Water Diver-sion Project. Environ. Sci. Pollut. Control Ser. 29 (52), 79320–79330. https://doi.org/10.1007/s11356-022-21210-0.
Ly, Q.V., Nguyen, X.C., Lê, N.C., Truong, T.D., Hoang, T.H.T., Park, T.J., Maqbool, T., Pyo, J.C., Cho, K.H., Lee, K.S., Hur, J., 2021. Application of Machine Learning for eutrophication analysis and algal bloom prediction in an urban river: a 10-year study of the Han River, South Korea. Sci. Total Environ. 797, 149040. https://doi.org/10.1016/j.scitotenv.2021.149040.
Maulud, D., Abdulazeez, A.M., 2020. A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends 1 (4), 140–147. https://doi.org/10.38094/jastt1457.
Nelson, N.G., Munoz-Carpena, R., Phlips, E.J., Kaplan, D., Sucsy, P., Hendrickson, J., 2018. Revealing biotic and abiotic controls of harmful algal blooms in a shallow subtropical lake through statistical machine learning. Environ. Sci. Technol. 52 (6), 3527–3535. https://doi.org/10.1021/acs.est.7b05884.
Nie, P., Roccotelli, M., Fanti, M.P., Ming, Z., Li, Z., 2021. Prediction of home energy consumption based on gradient boosting regression tree. Energy Rep. 7, 1246–1255. https://doi.org/10.1016/j.egyr.2021.02.006.
Nong, X., Shao, D., Zhong, H., Liang, J., 2020. Evaluation of water quality in the South-to-North Water Diversion Project of China using the water quality index (WQI) method. Water Res. 178, 115781. https://doi.org/10.1016/j.watres.2020.115781.
Park, Y., Cho, K.H., Park, J., Cha, S.M., Kim, J.H., 2015. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Sci. Total Environ. 502, 31–41. https://doi.org/10.1016/j.scitotenv.2014.09.005.
Qian, L., Plant, C., Böhm, C., 2021. Density-based clustering for adaptive density variation. In: 2021 IEEE International Conference on Data Mining (ICDM), IEEE, pp. 1282–1287.
Qian, J., Liu, H., Qian, L., Bauer, J., Xue, X., Yu, G., He, Q., Zhou, Q., Bi, Y., Norra, S., 2022. Water quality monitoring and assessment based on cruise monitoring, remote sensing, and deep learning: a case study of Qingcaosha Reservoir. Front. Environ. Sci. 10 (October), 1–13. https://doi.org/10.3389/fenvs.2022.979133.
Ralston, D.K., Moore, S.K., 2020. Modeling harmful algal blooms in a changing climate. Harmful Algae 91 (November), 101729. https://doi.org/10.1016/j.hal.2019.101729.
Redfield, A.C., 1963. The influence of organisms on the composition of seawater. Sea 2, 26–77.
Ridgeway, G., Madigan, D., Richardson, T., O'Kane, J., 1998. Interpretable boosted naïve bayes classification. In: The 4th International Conference on Knowledge Discovery and Data Mining (KDD-1998), pp. 101–104. URL citeseer.ist.psu.edu/ridgeway98interpretable.html.
Saltelli, A., 2002. Sensitivity analysis for importance assessment. Risk Anal. 22, 579–590. https://doi.org/10.1111/0272-4332.00040.
Su, Y., Hu, M., Wang, Y., Zhang, H., He, C., Wang, Y., Wang, D., Wu, X., Zhuang, Y., Hong, S., Trolle, D., 2022. Identifying key drivers of harmful algal blooms in a tributary of the Three Gorges Reservoir between different sea-sons: causality based on data-driven methods. Environ. Pollut. 297 (August 2021), 118759. https://doi.org/10.1016/j.envpol.2021.118759.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008.
Wang, Y., Li, Y., Liang, J., Bi, Y., Wang, S., Shang, Y., 2021. Climatic changes and anthropogenic activities driving the increase in nitrogen: evidence from the south-to-north water diversion project. Water (Switzerland) 13 (18). https://doi.org/10.3390/w13182517.
Welchowski, T., Maloney, K.O., Mitchell, R., Schmid, M., 2022. Techniques to improve ecological interpretability of black-box machine learning models: case study on biological health of streams in the United States with gra dient boosted trees. J. Agric. Biol. Environ. Stat. 27 (1), 175–197. https://doi.org/10.1007/s13253-021-00479-7.
Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al., 2020. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45.
Wright, R., 2018. Interpreting Black-Box Machine Learning Models Using Partial 594 Dependence and Individual Conditional Expectation Plots, Exploring SAS ® Enterprise Miner Special Collection, pp. 1950–2018. URL. https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/1950-2018.pdf.
Xia, R., Zhang, Y., Wang, G., Zhang, Y., Dou, M., Hou, X., Qiao, Y., Wang, Q., Yang, Z., 2019. Multi-factor identification and modelling analyses for managing large river algal blooms. Environ. Pollut. 254, 113056. https://doi.org/10.1016/j.envpol.2019.113056.
Yan, H., Huang, Y., Wang, G., Zhang, X., Shang, M., Feng, L., Dong, J., Shan, K., Wu, D., Zhou, B., Yuan, Y., 2016. Water eutrophication evaluation based on rough set and petri nets: a case study in Xiangxi-River, Three Gorges Reservoir. Ecol. Indicat. 69, 463–472. https://doi.org/10.1016/j.ecolind.2016.05.010.
Yang, Z., Zhang, M., Shi, X., Kong, F., Ma, R., Yu, Y., 2016. Nutrient reduction magnifies the impact of extreme weather on cyanobacterial bloom formation in large shallow Lake Taihu (China). Water Res. 103, 302–310. https://doi.org/10.1016/j.watres.2016.07.047.
Yu, P., Gao, R., Zhang, D., Liu, Z.P., 2021. Predicting coastal algal blooms with environmental factors by machine learning methods. Ecol. Indicat. 123, 107334. https://doi.org/10.1016/j.ecolind.2020.107334.
Zhu, Y., Mi, W., Tu, X., Song, G., Bi, Y., 2022. Environmental factors drive periphytic algal community assembly in the largest long-distance water diversion channel. Water 14 (6). https://www.mdpi.com/2073-4441/14/6/914.
Zhu, J., Lei, X., Quan, J., Yue, X., 2019. Algae growth distribution and key prevention and control positions for the middle route of the south-to-northwater diversion project. Water (Switzerland) 11 (9), 1–18. https://doi.org/10.3390/w11091851.

Supplementary Material to Deep learning approach towards accurate identification of spatial driving factors for algal growth in the South-to-North Water Diversion Project using Transformer-based model

Jing Qian[a], Nan Pu[b], Li Qian[c], Xiaobai Xue[d], Yonghong Bi[e,*], Stefan Norra[f]

[a]*Institute of Applied Geosciences, Karlsruhe Institute of Technology, Karlsruhe 76131, Germany*
[b]*Institute of Advanced Computer Science, Leiden University, Leiden, 2333 CA, Netherlands*
[c]*Institute of Informatics, Ludwig Maximilian University of Munich, Munich 80538, Germany*
[d]*MioTech Research, Yingtou Information Technology (Shanghai) Limited, Shanghai 200120, China*
[e]*State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China*
[f]*Institute of Environmental Sciences and Geography, Soil Sciences and Geoecology, Potsdam University Potsdam-Golm 14476, Germany*

1. Z-score Normalization

Z-score Normalization is a tool to standardize features by removing the mean and scaling to unit variance. This scaler is widely used for normalization in many machine learning algorithms (e.g., support vector machines, logistic regression, and artificial neural networks). And it was utilized for each water quality parameter in this study i according to the following equation[1]:

$$Z_i = \frac{x_i - \overline{x_i}}{\sigma_i}$$

where Z_i is the standard score of i-th water quality parameter, x_i is the -th original water quality parameter, $\overline{x_i}$ is the mean of i-th water quality parameter, and σ_i is the standard deviation of i-th water quality parameter.

*Corresponding author
Email address: biyh@ihb.ac.cn (Yonghong Bi)

2. MSE

MSE is a convenient way to measure the average error, defined as the expectation of the square of the difference between the estimated value and the true value, is a commonly used regression evaluation method[2].

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2$$

Where $\widehat{y}_i$ is the predicted value, y_i is the true value.

3. Training loss of Bloomformer-1

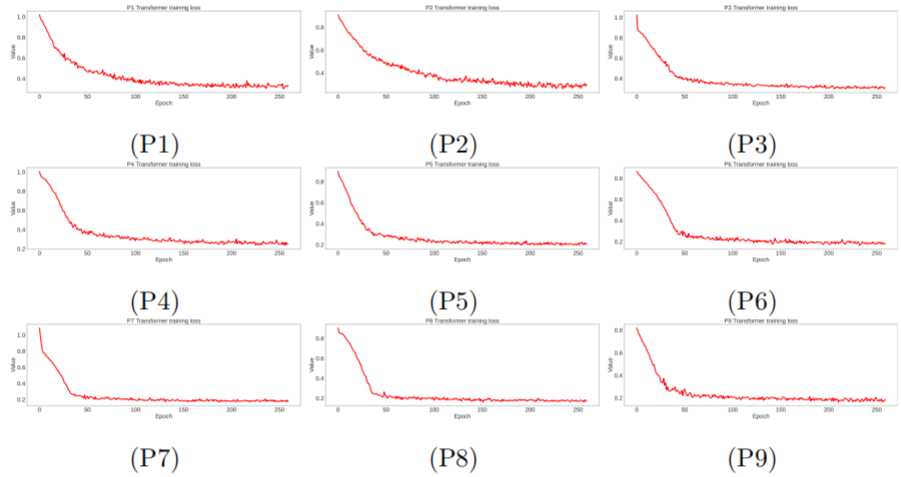


Figure S1: Training loss of Bloomformer-1

4. Deep learning

Deep learning is a subclass of machine learning that comprises neural networks with three or more layers. These neural networks seek to imitate the activity of the human brain, although inadequately, allowing them to "learn" from vast quantities of data. Despite the fact that a neural network with a single hidden layer can still produce approximations, multiple hidden layers can help to improve and enhance the network for precision. Figure S2 shows the

performance of various AI models as the data volume increases. Compared to others, deep learning is more sensitive to the data size and has a clear advantage when dealing with big data.

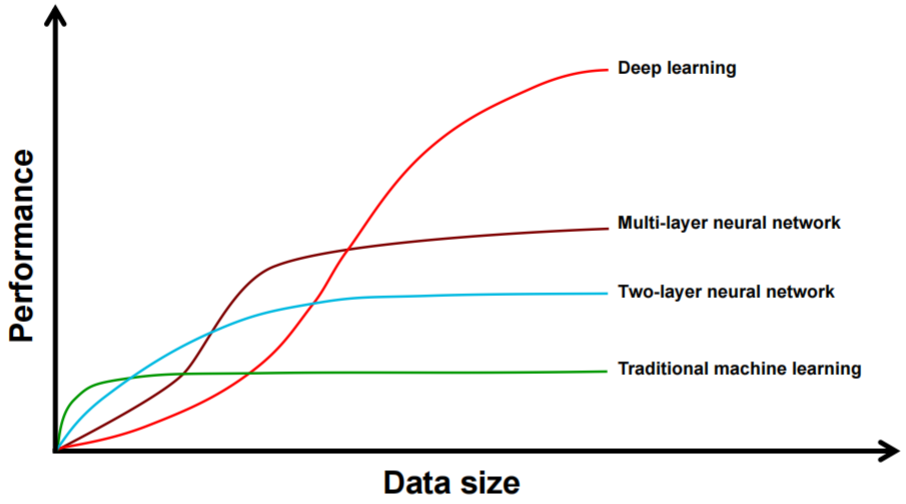


Figure S2: Performance vs. Data size

References

[1] X. Du, Y. Cai, S. Wang, L. Zhang, Overview of deep learning, Proceedings - 2016 31st Youth Academic Annual Conference of Chinese Association of Automation, YAC 2016 (2017) 159–164doi:10.1109/YAC.2016.7804882.

[2] P. Yu, R. Gao, D. Zhang, Z. P. Liu, Predicting coastal algal blooms with environmental factors by machine learning methods, Ecological Indicators 123 (2021) 107334. doi:10.1016/j.ecolind.2020.107334.
URL https://doi.org/10.1016/j.ecolind.2020.107334

An Intelligent Early Warning System for Harmful Algal Blooms: Harnessing the Power of Big Data and Deep Learning

Jing Qian,*,† Nan Pu,‡ Li Qian,¶ Yonghong Bi,§ and Stefan Norra‖

†*Institute of Applied Geosciences, Karlsruhe Institute of Technology, Karlsruhe 76131, Germany*

‡*Institute of Advanced Computer Science, Leiden University, Leiden, 2333 CA, Netherlands*

¶*Institute of Informatics, Ludwig Maximilian University of Munich, Munich 80538, Germany*

§*State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China*

‖*Institute of Environmental Sciences and Geography, Soil Sciences and Geoecology, Potsdam University, Potsdam-Golm 14476, Germany*

E-mail: jing.qian@partner.kit.edu

Abstract

Harmful algal blooms (HABs) pose a significant ecological threat and economic detriment to freshwater environments. In an endeavor to manage these occurrences, we have harnessed the potential of big data and deep learning models to engineer an intelligent early warning system for HABs. Data acquisition is accomplished through a Vertical Aquatic Monitoring System (VAMS), which, in conjunction with the "DeepDPM-Spectral Clustering" methodology, facilitates an intricate analysis of the vertical algal

distribution. This approach curtails the number of predictive models and enhances the adaptability of the system. Employing the Bloomformer-2 model, developed by our team, the system carries out both single-step and multi-step prognostications of HABs. Our case study corroborates the superior performance of Bloomformer-2, exhibiting high congruity with actual value curves and a lower margin of predictive error. This system boasts the unique ability to identify the driving factors of HABs, thereby aiding in the formulation of targeted preventive measures. Additionally, the model's remarkable intelligence - the capacity to autonomously learn from preprocessed data - and its inherent adaptability pave the way for future system upgrades and broader applications. As part of future work, it is proposed to augment the big data platform and establish a VAMS monitoring network to bolster the system's geographical coverage and predictive capability. This research underscores the transformative potential of integrating big data and artificial intelligence in environmental management, and emphasizes the importance of model interpretability in machine learning applications.

Keywords

HABs, Chl-a, Bloomformer-2, Early warning, Time series analysis

Synopsis

This study leverages big data and deep learning to create an early warning system for harmful algal blooms, improving predictive capabilities and adaptability, and enabling targeted preventive measures.

Introduction

Perilous algal efflorescences, or Harmful Algal Blooms (HABs), have burgeoned into a pervasive global environmental quandary.[1] As per the records maintained by the Harmful Algae

Event Database (HAEDT, http://haedat.iode.org), the chronology from 1990 to 2021 has witnessed a staggering 12,195 HABs events across the globe, exhibiting a disconcerting escalation in their frequency. These HABs, notably the cyanobacterial proliferations, pose formidable repercussions for various facets of our ecosystem, encompassing the water supply, fisheries, recreational utilities, tourism sector, and real estate.[2] Consequently, the exigency for devising a cost-effective, precise, and pragmatic early warning system to combat HABs has become an inescapable imperative for scientists, policy architects, and custodians of environmental resources.

The concentration of Chlorophyll-a (Chl-a), serving as a pivotal gauge of water quality, is typically employed to oversee HABs in both coastal and offshore aquatic environments.[3] Foretelling the spatial-temporal oscillations of Chl-a concentration proffers an early alert for the emergence of HABs, while concurrently providing a time-sensitive comprehension of ecological conditions. Numerous scholarly pursuits have endeavored to leverage numerical models to prognosticate Chl-a across diverse temporal resolutions, responding to the incidence of HABs. At this juncture, the prediction models for Chl-a are predominantly based on a monthly framework.[4,5] However, a subset of researchers have constructed prediction models for Chl-a with heightened temporal resolution, including semi-monthly, weekly, and daily intervals.[3,6,7] Albeit, considering the intricate nature of algae blooms, which correlate with physical, chemical, biological, and hydrological factors, the extant time-resolved studies fall short of efficiently tracking and predicting this multifaceted process. Thus, the procurement of high-resolution temporal data and the formulation of an apt prediction model become imperative. Presently, most HABs early warning system studies predominantly rely on Chl-a data derived from the water body's surface.[8] Nevertheless, water depth is recognized as a crucial determinant controlling algal biomass and growth,[9,10] suggesting that the circumstances of HABs at distinct depths within a precise location are subject to fluctuation, corroborated by our collected data. Hence, the development of Chl-a prediction models across various depths can enhance our understanding of the harmful algal circumstances

within the targeted area.

In light of the considerable advancements in sensor technology, a vertical aquatic monitoring system (VAMS) has been unveiled to scrutinize water. This system, armed with an array of sensors, autonomously gathers data pertaining to water quality parameters tagged with depth labels (from surface to bottom) at predetermined intervals. On the one hand, the data harvested by VAMS renders it feasible to construct a high temporal resolution and depth-dimensional early warning system. On the other hand, it lays down the foundation for transitioning from small data analysis to big data analysis of vertical water bodies, given its inherent characteristics of high-volume, real-time, and continuous data.

Given the unbroken nature of depth data, the number of models would burgeon if prediction frameworks were fashioned for each depth segment – for example, presuming a 0.05m depth segment, we would confront 40 models for a 2m water depth. Such a modeling approach is deemed superfluous, as it unsettles the balance between temporal investment and predictive prowess. As a result, a technique dubbed "DeepDPM-Spectral Clustering" was devised in this study to refine the modeling strategy. The most formidable challenge in this endeavor is clustering depth segments with an indeterminate cluster quantity. DeepDPM, a deep learning clustering algorithm, can address this quandary. Unlike conventional clustering algorithms, DeepDPM possesses the aptitude to autonomously discern the optimal cluster number to tackle expansive data with an undefined cluster figure through iterative cycles of amalgamation and division.[11] Subsequent to the deep clustering executed by DeepDPM, an adjacency matrix is obtained by computing the likelihood of each depth segment being classified within the same cluster. Thereafter, the spectral clustering algorithm can integrate pertinent depth segments based on this adjacency matrix, thus subdividing the water column at the focal site into several assemblages, each distinguished by water quality and demarcated by depth.

Serving as indicators of environmental alterations, algae swiftly react to a broad spectrum of pollutants, thus establishing a correlation between their proliferation and the quality

of the adjacent water bodies.[12] The forecast of Chl-a's future values hinges not solely on its antecedent values but also on the preceding or current values of other water quality parameters. Conventional time series forecasting (TSF) statistical techniques may grapple with interpreting lengthy contextual sequences and extending to intricate variable relationships.[13] Deep Learning models surmount these hurdles by capitalizing on extensive datasets. The Long-Short-Term-Memory (LSTM) model has recently gained traction for predicting Chl-a concentration.[2,14–16] Although LSTM can attain relatively superior prediction accuracy, its structure precludes direct visualization of the relationships and weights between the predicted value and other variables. As a result, preventative measures cannot be precisely delineated or timely instituted. Furthermore, LSTM can only mitigate vanishing gradient and exploding gradient issues to a certain degree when handling extensive time-series data, but it does not provide a comprehensive solution.[17] Therefore, LSTM is only suited to relatively long time-series data, whereas its predictive efficacy for more protracted time-series data is subpar. In addition, the HABs early warning system requires the capability to accurately predict Chl-a while concurrently offering a clear and tangible presentation of the driving factors for predicted values to aid in the creation of preventative measures. The mechanism and structure of LSTM dictate that it lacks this functionality. In contrast, the Transformer, another deep learning model, showcases a distinctive edge in processing TSF. Its Multihead-Self-Attention (MSA) mechanism allows the Transformer to achieve a prediction accuracy that is on par with or surpasses LSTM when dealing with long time-series data, while directly elucidating the temporal and spatial relationships between predicted values and other parameters. However, the standard Transformer struggles to effectively handle complex multivariate TSF problems.[13] To address these needs, we devised a transform-based prediction model for the HABs early warning system, dubbed Bloomformer-2. Owing to its structure, it retains all the advantages of standard Transformers while adeptly handling complex multivariate TSF.

This study endeavors to create a high temporal resolution and depth-dimensional HABs early warning system with a cornerstone of big data mining, grounded on (1) the Vertical

aquatic monitoring system for the construction of a big data platform; (2) "DeepDPM-Spectral cluster" for the optimization of modeling strategy; and (3) Bloomformer-2 for precise predictive outcomes and preventative measures. The early warning system was implemented in the Taihu Laboratory for Lake Ecosystem Research (TLLER) to validate its soundness and dependability.

Materials and methods

Study area

Reposing in the dynamically evolving Yangtze River Delta region of China, Taihu Lake, the nation's third-largest freshwater body, sprawls across an expansive 2,338 square kilometers,[18] as illustrated in Figure 1. This relatively shallow lake possesses an average depth of 1.9 meters.[19] Meiliang Bay, a notably eutrophic haven in the northern facet of Taihu Lake, encompasses a vast area of 124 square kilometers, with an average depth of 1.5 meters.[20] Two principal waterways, the Liangxi River and the Zhenwugang River, ferry urban pollutants from the metropolises of Wuxi and Changzhou into the heart of Meiliang Bay. Since 1998, Meiliang Bay has been grappling with severe algal blooms in both the summer and autumn seasons, a consequence of its significant function as a primary conduit of human activities and a crucial source of potable water.

To gauge the performance of our meticulously constructed algal early warning system, we orchestrated a series of experiments at the terminus of a 250-meter-long jetty at the Taihu Laboratory for Lake Ecosystem Research (TLLER) (31.418903 N, 120.213293 E), nestled on the southern fringe of Meiliang Bay.

Vertical aquatic monitoring system

In this study, we developed a vertical aquatic monitoring device, called BIOLIFT, to collect the data of the investigated water body (Seeing supporting information). Equipped with a

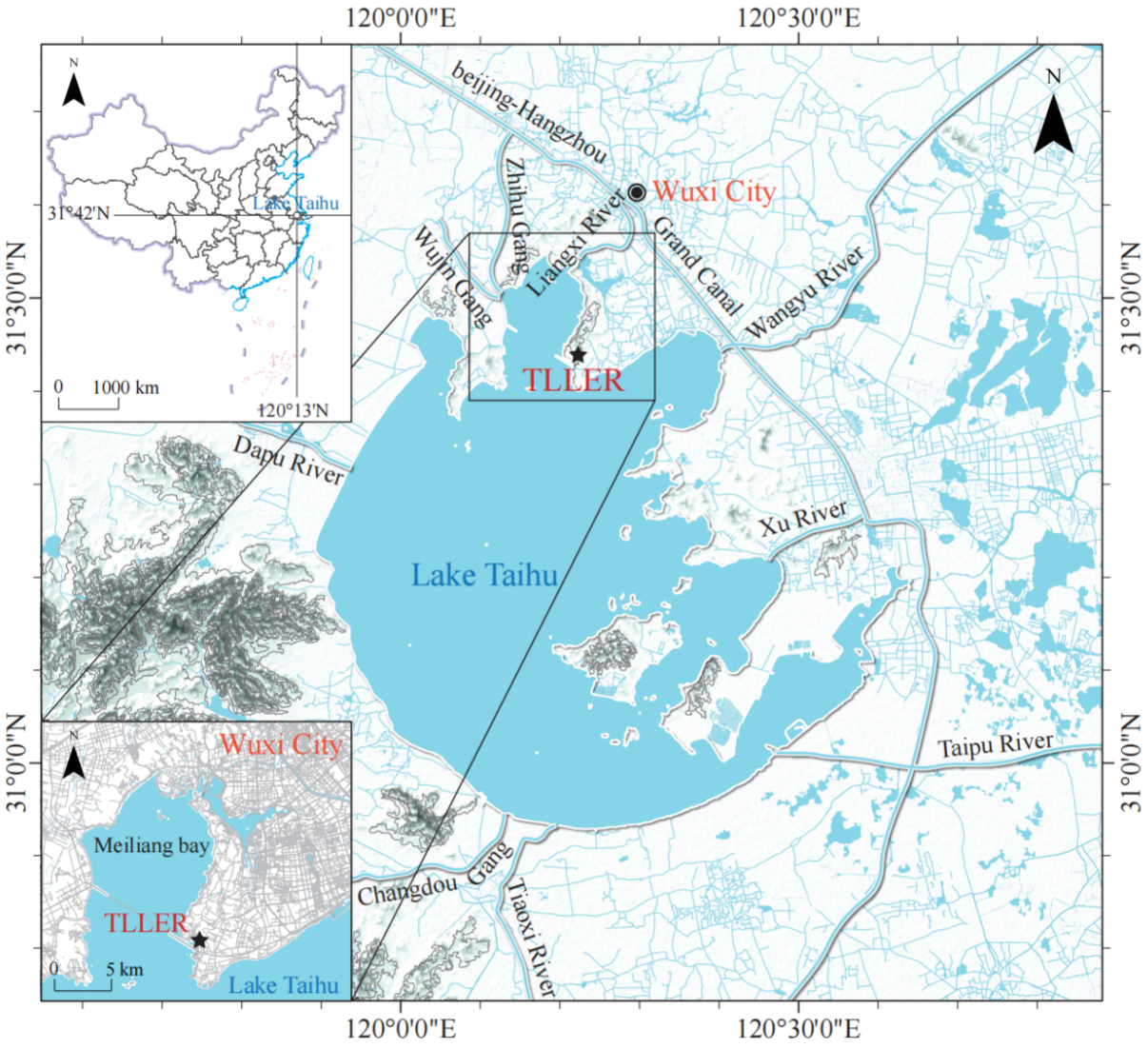


Figure 1: Location of Lake Taihu, Meiliang Bay and TLLER (pentagon)

suite of multi-sensors and tethered to a control box via a data transmission cable, it diligently records real-time water quality parameters, each data point meticulously labeled with its corresponding depth as the device ascends and descends at pre-set time intervals. Each complete cycle of ascension and descension is denoted as a work cycle, the temporal interval between which can be adjusted by the investigators. The recorded parameters encompass Electrical Conductivity at 25 °C (EC_{25}), Temperature (Temp.), pH, Turbidity (Turb.), Colored Dissolved Organic Matter (CDOM), and Chl-a. Moreover, given the potent correlation between wind patterns and the ecological evolution of large shallow lakes, epitomized by Lake Taihu, a weather station (Vaisala Weather Transmitter WXT520) is seamlessly integrated with BIOLIFT to capture wind speed (WS) and direction (WD) throughout the operational duration of BIOLIFT. The detail of the sensors is shown in supporting information.

Data harvested by BIOLIFT at TLLER during the winter of 2018 and the summer of 2019 have been utilized to construct a comprehensive data platform and validate the efficacy of this HABs early warning system. The temporal interval between work cycles was established at 10 minutes, and the depth of a single segment was determined to be 0.05m. Acknowledging the requisite stabilization period for sensors of BIOLIFT upon immersion, outlier analysis was employed to refine data from depth segments proximal to the surface. Moreover, the BIOLIFT system was routinely checked and calibrated each morning for a period of 1.5 to 2 hours, thereby ensuring the accuracy and reliability of its data. The "2018-Winter" data set encapsulates 13 days of BIOLIFT data, with each day comprising 132 work cycles and each cycle containing water quality data from 23 depth segments (ranging from 0.05m to 1.2m), in addition to wind speed and direction. The "2019-Summer" data set chronicles 13 days of BIOLIFT data, each day encompassing 134 work cycles and each cycle incorporating water quality data from 37 depth segments (ranging from 0.1m to 1.95m), along with wind speed and direction.

Modeling strategy optimization

The method christened "DeepDPM-Spectral clustering" was meticulously crafted and deployed to assemble depth segments into several cogent clusters, predicated on the cumulative impact of the measured water quality parameters. Based on this method, an optimization stratagem was devised to augment the efficiency of the system by modeling each cluster as an entity, rather than focusing on individual depth segments.

DeepDPM

DeepDPM is bifurcated into two primary components:[11] firstly, the clustering network and secondly, the K subclustering networks, each corresponding to a distinct cluster k, where $k \in 1, \ldots, K$. The role of the clustering network is to yield soft cluster assignments from the original data, while a subclustering network's task is to produce soft subcluster assignments from the same raw data. Each of these, the clustering network and the subclustering networks, constitutes a straightforward multilayer perceptron with a lone hidden layer. The terminal layer of the clustering network is equipped with K neurons, in contrast to the final layer of each subclustering network which houses a duo.

Guided by the Metropolis-Hastings framework,[21] decisions to split or merge are made, effecting changes in K, and the split/merge procedures are executed accordingly. Throughout these split/merge stages, the preliminary cluster number K, the clustering network, and the K subclustering networks undergo modifications, and an iterative process is implemented until the identification of the optimal cluster number K. The detailed technicalities are exhibited in the appended Supporting Information.

Within the parameters of this research, data pertaining to water quality parameters (excluding wind speed and direction) for each work cycle in the "2018-Winter" and "2019-Summer" datasets were independently processed through DeepDPM for intensive clustering. DeepDPM autonomously discerned the optimal cluster number for each cycle, resulting in a statistical distribution of the optimal cluster numbers. Furthermore, the adjacency matrix

can be derived by calculating the probability that each depth segment is classified into the same cluster during the course of the experiment.

Spectral clustering

The spectral clustering algorithm perceives[22] all data as spatial points that can be interconnected with edges. The weight of the edge connecting two distant points is diminished, while the weight between two proximate points is amplified. Through cleaving the graph, composed of all data points, the summation of edge weights between distinct subgraphs post-separation is minimized, and conversely, the aggregate weight within the subgraph is maximized, thereby facilitating the objective of clustering. In essence, the spectral clustering algorithm's process involves the construction of an adjacency matrix and its partition via normalized cut, necessitating a given value of cluster number K during the partitioning phase. The specifics of the spectral clustering algorithm are delineated in the appended supporting information.

The adjacency matrix was derived throughout this research via the analysis of the Deep-DPM results. The optimal cluster number K, boasting the highest percentage in the Deep-DPM results, was adopted as the value of the cluster number K in the spectral clustering.

Multivariate Time Series Forecasting

In this section, a juxtaposition of LSTM and Bloomformer-2, two intricate deep learning models tailored for multivariate time series prognostication, is presented. The amalgamated datasets stemming from the "2018-Winter" and "2019-Summer" intervals, encompassing wind speed and direction, are respectively furnished to these models for forecasting purposes.

LSTM

Long Short-Term Memory is distinctive Recurrent Neural Networks (RNNs) adept at discerning long-term dependencies. Conventional RNNs employ a loop to amalgamate the

input with the preceding output, thereby fostering information persistence. Despite recurrent networks theoretically embodying a straightforward yet potent model to grapple with the challenge of "long-term dependencies", their practical application often falls short.[23] The complications of vanishing and exploding gradients impede the consistent recollection of useful information.[24] To mitigate this issue, LSTMs were introduced. Equipped with a memory cell, they have the capability to store, discard, and append data contingent on the requirement of the information over an extended duration. As a result, LSTMs are proficient at predicting longer sequences.[25] The cornerstone of LSTM is the cell state.[26] The cell state traverses through the entire chain akin to a conveyor belt, undergoing only minor linear interactions. This allows information to flow along it unaltered. The constructs that facilitate the removal or addition of information to the cell state are termed gate layers. LSTM encompasses three such gate layers: the forget gate layer, input gate layer, and output gate layer. The architecture and intricate technicalities of LSTM are elucidated in the supporting information.

Bloomformer-2

The Transformer model represents the state-of-the-art solutions for natural language processing (NLP) tasks.[27] Employing the Multi-Head Attention mechanism (as delineated in the supporting information), the Transformer scrutinizes each token in the input sequence in relation to other tokens to gather and assimilate dynamic contextual information.[28] This underpins the model's proficiency in orchestrating the transfer of information among inputs. Unlike the LSTM, the Transformer, given its non-sequential analysis of inputs, remains impervious to the gradient vanishing issue that impedes the long-term predictive capabilities of RNNs.[29] Consequently, Transformers have found utility in datasets, including those pertinent to TSF, which harbor long-term historical information. The architecture and specifics of the standard Transformer are illustrated in the supporting information.

Nonetheless, the standard Transformer lacks optimization for complex multivariate TSF

tasks, as it perceives the value of each variable at a specified time period as a discrete marker in its graph, with each variable being incapable of prioritizing its own contextual view.[13] As an enhancement of the standard Transformer, the Bloomformer-2 is adept at tackling complex multivariate TSF tasks efficiently. This refined approach initially transmutes the contextual sequence of historical data and the anticipated target timestamps into an extensive spatio-temporal sequence. This sequence is then transposed to yield a corresponding extensive temporal-spatial sequence. Both sequences are subsequently processed with a Transformer-based encoder-decoder architecture to derive the predicted values for each variable. Ultimately, the predicted values are reassembled into their original format and trained to minimize the prediction error metrics. This entire process, constituting the flattening of input variables into spatio-temporal sequences, spatio-temporal embeddings, and long-term prediction, is vividly depicted in Figure 2, showcasing the architecture of the Bloomformer-2.

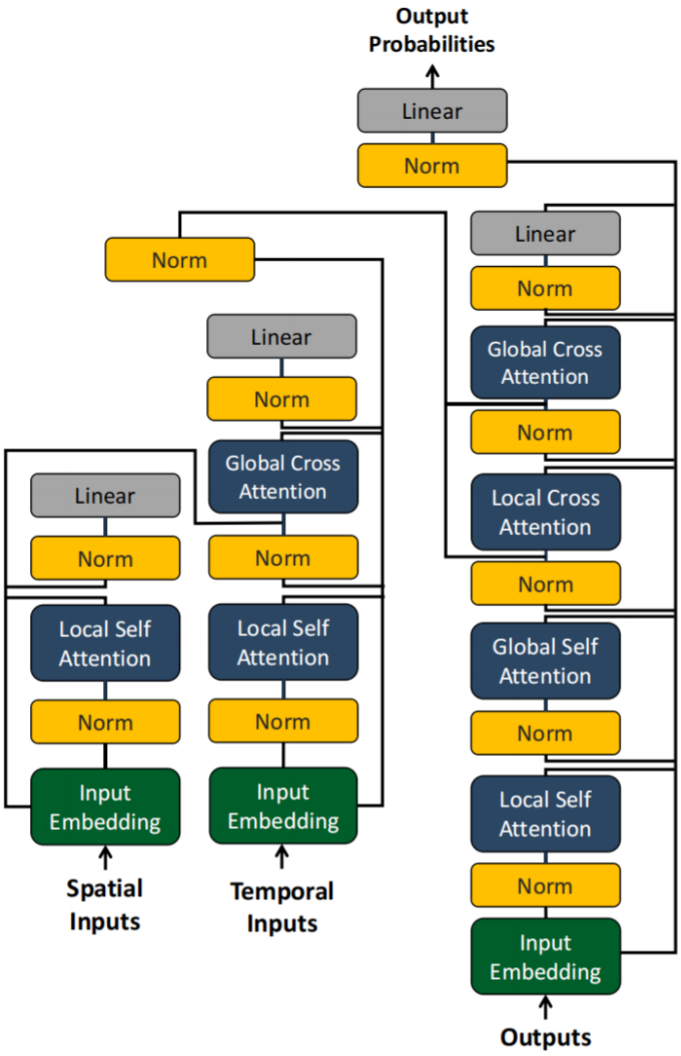


Figure 2: Architecture of the Bloomformer-2

Training process

In the realm of TSF, a prognostication for a solitary future timestep is termed as a single-step prediction, whereas a prediction encompassing multiple future steps is referred to as a multi-step prediction.[30] This procedure progressively slides forward one step at a time to envisage the ensuing data point. Traditional strategies for multi-step prediction encapsulate direct multi-step prediction, recursive multi-step prediction, and a fusion of both. Past research has substantiated that the amalgamation of direct and recursive strategies outperforms ei-

ther strategy in isolation. Nonetheless, due to the recursive operation, error accumulation transpires during multi-step prediction, diminishing the predictive efficacy.[31] To a certain degree, the Sequence-to-Sequence (Seq2Seq) architecture, tailored for deep learning models and delineated in the supporting information, mitigates this issue by directly outputting sequences of adjustable length.[32]

In this study, the datasets of 2018-Winter and 2019-Summer were divided into training sets and test sets. The training set encapsulates data from the 1st through the 10th day, whilst the test set comprises data spanning from the 11th to the 13th day. within the single-step prediction process, two-day timesteps—268 work cycles for 2019-Summer and 264 work cycles for 2018-winter—are employed to foresee the subsequent Chl-a data. In the multi-step prediction, two-day timesteps—268 work cycles for 2019-summer and 264 work cycles for 2018-winter—are utilized as the input sequence to predict the forthcoming three-day timesteps—402 work cycles for 2019-summer and 396 work cycles for 2018-winter—which constitute the output sequence. Prior to incorporating all data into the model, data normalization is performed using the Z-score(seeing supporting information).

Computational Environment

The experiment was carried out on a PC with the following features: Hardware: CPU i7-6950X, RAM 64GB, dual GeForce RTX 3090, VRAM 24GB Software: Ubuntu 20.04, Python3.6, Pytorch 1.10.0, Numpy 19.2

Evaluation metrics

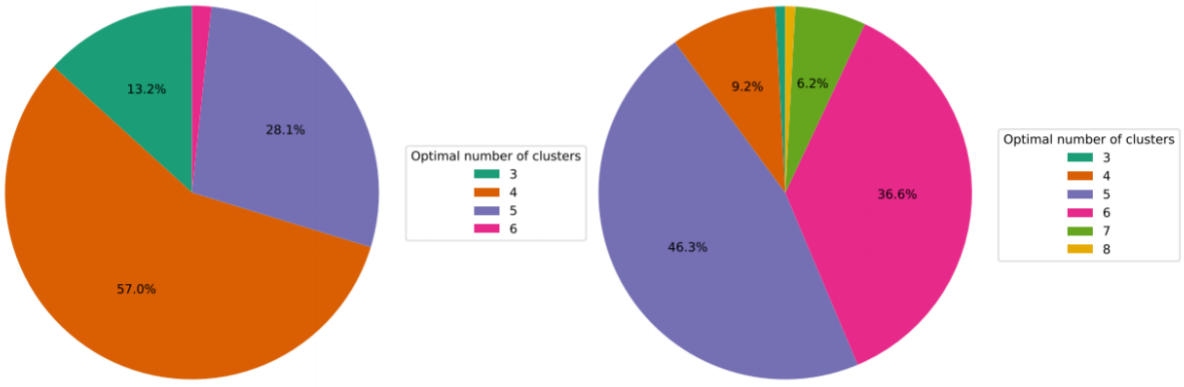
In the assessment of LSTM and Bloomformer-2 performance, we employ discerning metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).[33] Negligible quantities of MSE, MAE, and MAPE herald a harmonious congruence between the anticipated and the authentic values. A zero value epitomizes the zenith of accuracy. A comprehensive exposition of MSE, MAE, and MAPE is furnished

in the Supporting Information. Moreover, to depict their performance, the 95% confidence interval of the predicted value was meticulously calculated.

Results

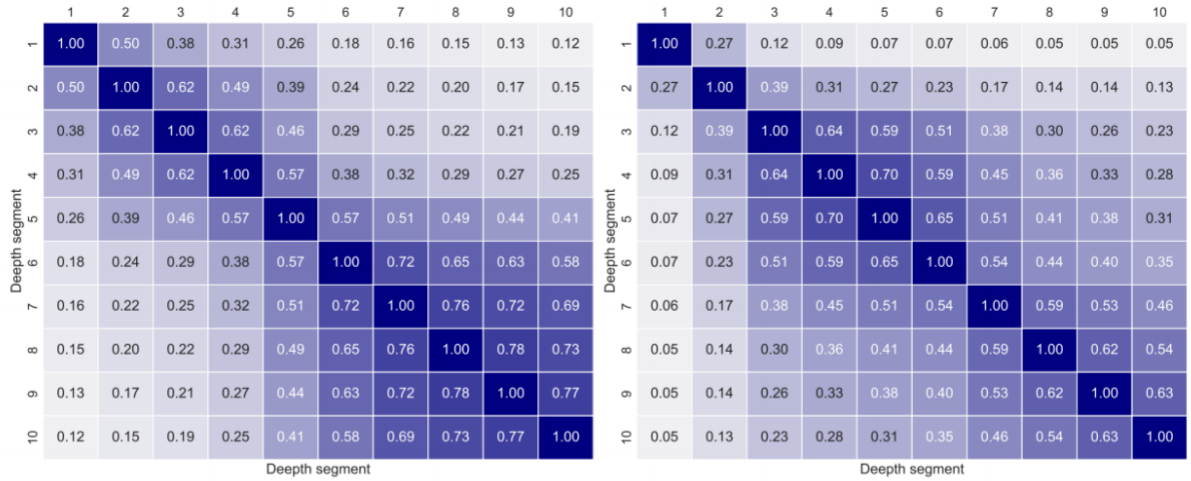
Results of water depth clustering

Each work cycle within the datasets pertaining to "Winter-2018" and "Summer-2019" was independently introduced into DeepDPM for deep clustering. The distribution of the optimal cluster number, as exhibited in Figure 3, reveals that the optimal cluster numbers are five during the four in 2018-winter (57.0%) and 2019-summer (46.2%). Partial adjacent matrices for winter and summer are delineated in Figure 4. These adjacent matrices underwent spectral clustering to refine the modeling strategy, the outcomes of which are cataloged in the table. In the summer of 2019, the aquatic depth within the target region was bifurcated into five factions, designated Group S1 through Group S5. Conversely, in the winter of 2018, the aquatic depth was segregated into four factions, marked as Group W1 through Group W4.



(2018-Winter) (2019-Summer)

Figure 3: Distribution of optimal cluster number



(2018-Winter) (2019-Summer)

Figure 4: Adjacency matrix of 2018-Winter and 2019-Summer (an example of 10 depth segment)

Table 1: Result of water depth clustering

Season	Water depth group				
Winter 2018	Group W1 0.05 - 0.1m	Group W2 0.1 - 0.3m	Group W3 0.3 - 0.95m	Group W4 0.95 - 1.2m	
Summer 2019	Group S1 0.1 - 0.15m	Group S2 0.15 - 0.4m	Group S3 0.4 - 0.95m	Group S4 0.95 - 1.55m	Group S5 1.55 - 1.95m

Results of prediction

In this study, a pair of models, LSTM and Bloomformer-2, were employed to prognosticate the datasets from Winter-2018 and Summer-2019. We executed single-step and multi-step forecasts for each aquatic depth category, subsequently juxtaposing the predictive performance of the two deep learning models.

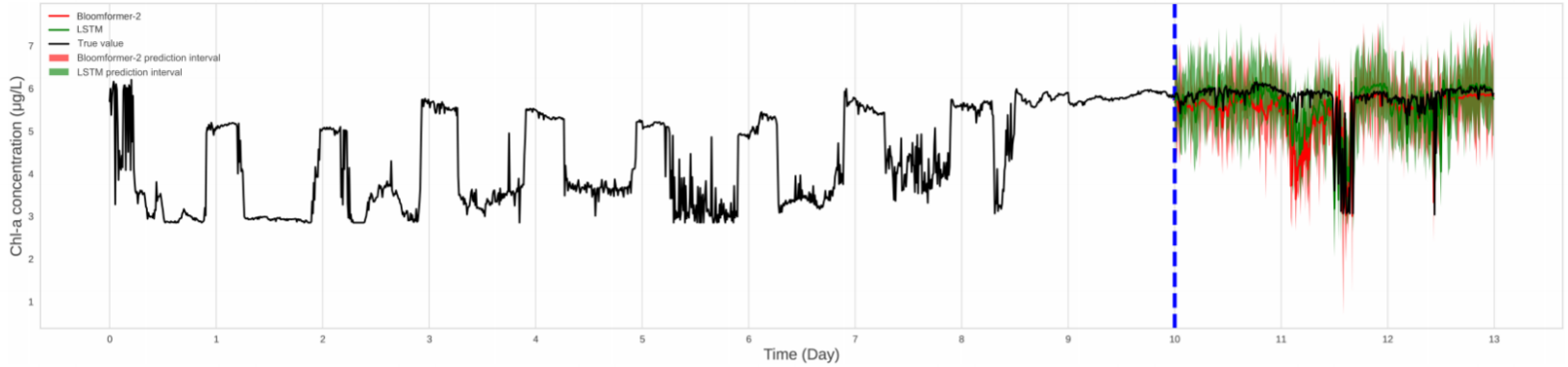
Single-step prediction

Computation of the predictive errors for the two deep learning models in single-step forecasting was undertaken (Table 2, the bold-italic values represent the best performance). The outcomes of the single-step predictions for Group W1 and Group S1 are portrayed in Fig 2

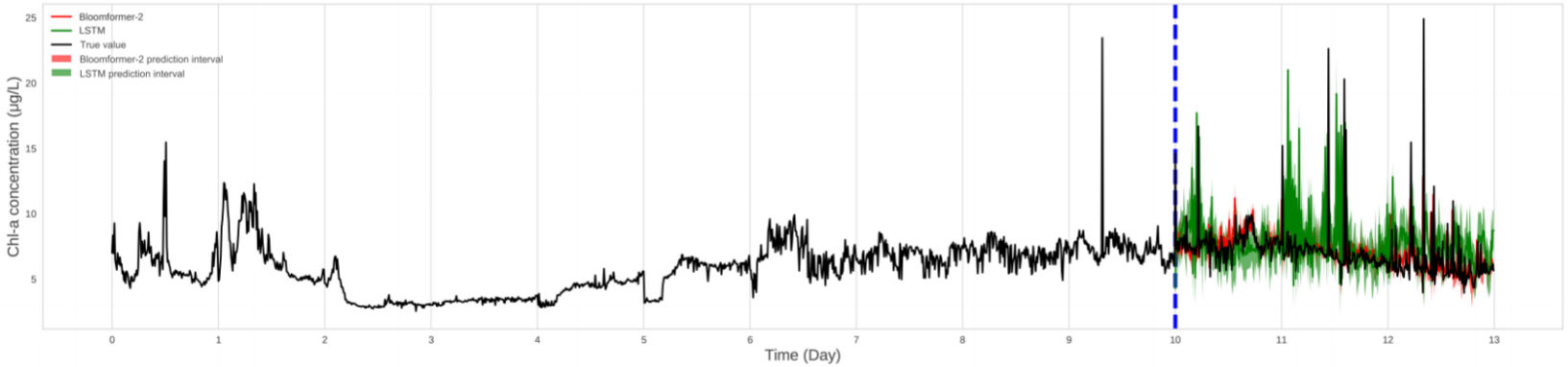
(The results of other water depth groups are shown in Figure S6 and S7). The predictive value trajectories of Bloomformer-2 for Groups S1, S2, S4, W2, W3, and W4 exhibit a superior alignment with the veritable value curves and possess constricted confidence intervals in comparison to LSTM. Furthermore, Bloomformer-2 yielded smaller predictive errors than LSTM. Using Group S1's single-step prediction as an illustration, the performance evaluative metrics of Bloomformer-2 (MAE=0.254, MSE=0.305, and MAPE=2.279) are universally inferior to those of LSTM (MAE=0.916, MSE=1.887, and MAPE=8.427). In Groups S3, S5, and W1, Bloomformer-2's single-step predictive precision parallels LSTM, as evidenced by their analogous predictive value curves, akin confidence intervals, and comparable performance evaluative metrics.

Table 2: Errors of Bloomformer-2 and LSTM in single-step prediction

Water depth group	**Model**	**MAE**	**MSE**	**MAPE**
Group S1	Bloomformer-2	**0.254**	**0.305**	**2.279**
	LSTM	0.916	1.887	8.427
Group S2	Bloomformer-2	**0.394**	**0.246**	**2.108**
	LSTM	0.541	0.573	2.969
Group S3	Bloomformer-2	0.357	0.205	**0.733**
	LSTM	**0.309**	**0.154**	0.998
Group S4	Bloomformer-2	**0.288**	**0.142**	**0.848**
	LSTM	0.301	0.143	0.855
Group S5	Bloomformer-2	0.417	**0.249**	1.955
	LSTM	**0.373**	0.271	**1.162**
Group W1	Bloomformer-2	0.244	**0.072**	0.266
	LSTM	**0.159**	0.076	**0.191**
Group W2	Bloomformer-2	**0.213**	**0.056**	**0.269**
	LSTM	0.329	0.129	0.421
Group W3	Bloomformer-2	**0.201**	**0.052**	**0.247**
	LSTM	0.688	0.509	0.801
Group W4	Bloomformer-2	**0.175**	**0.042**	**0.228**
	LSTM	0.184	0.044	0.237



(a) Group W1



(b) Group S1

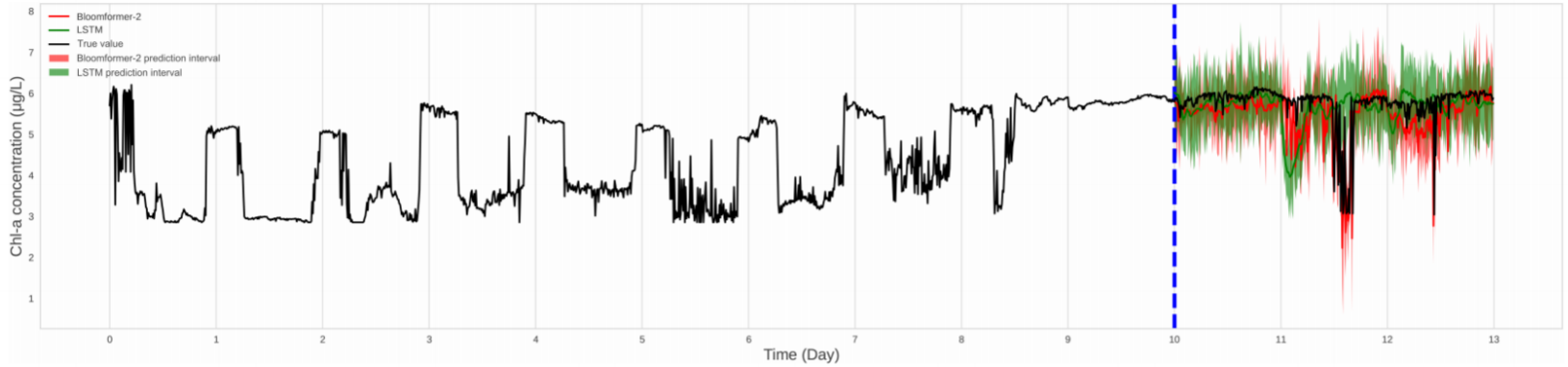
Figure 5: Comparison of model prediction in single-step prediction

Multi-step prediction

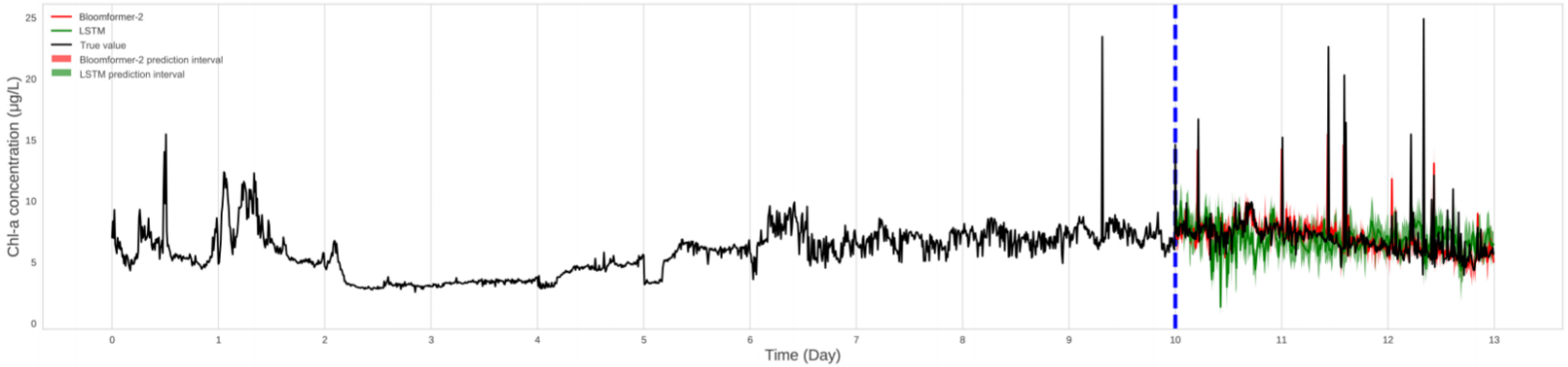
The prediction errors for both deep learning models in multi-step prediction were computed (Table 3the bold-italic values represent the best performance). The multi-step predictive outcomes for Group W1 and Group S1 are illustrated in Figure 6 (The results of other water depth groups are shown in Figure S8 and S9). Bloomformer-2 outperforms LSTM across all groupings in the context of multi-step forecasting, as the projected value trajectories more closely mirror the authentic value curves, and the forecast inaccuracies are notably reduced.

Table 3: Errors of Bloomformer-2 and LSTM in multi-step prediction

Water depth group	**Model**	**MAE**	**MSE**	**MAPE**
Group S1	Bloomformer-2	**0.207**	**0.161**	**1.091**
	LSTM	0.613	1.086	5.264
Group S2	Bloomformer-2	**0.421**	**0.269**	**4.011**
	LSTM	0.474	0.361	4.034
Group S3	Bloomformer-2	**0.238**	**0.101**	**0.349**
	LSTM	0.526	0.473	2.629
Group S4	Bloomformer-2	**0.341**	**0.184**	**1.39**
	LSTM	0.549	0.508	2.418
Group S5	Bloomformer-2	**0.505**	**0.378**	**1.679**
	LSTM	0.512	0.402	3.748
Group W1	Bloomformer-2	**0.249**	**0.121**	**0.372**
	LSTM	0.339	0.337	0.621
Group W2	Bloomformer-2	**0.184**	**0.105**	**0.492**
	LSTM	0.353	0.283	0.799
Group W3	Bloomformer-2	**0.188**	**0.068**	**0.243**
	LSTM	0.291	0.301	0.352
Group W4	Bloomformer-2	**0.361**	**0.167**	**0.558**
	LSTM	0.397	0.307	0.603



(a) Group W1



(b) Group S1

Figure 6: Comparison of model prediction in multi-step prediction

Identification of Driving factors for the predicted value

The outcome of the 11th-day prediction for Group S1 serves as an exemplar, showcasing the amalgamated driving factors for a one-day forecast and a solitary driving factor for a ten-work cycle prediction (as an example) in Winter-2018 and Summer-2019, respectively (Figure 7).

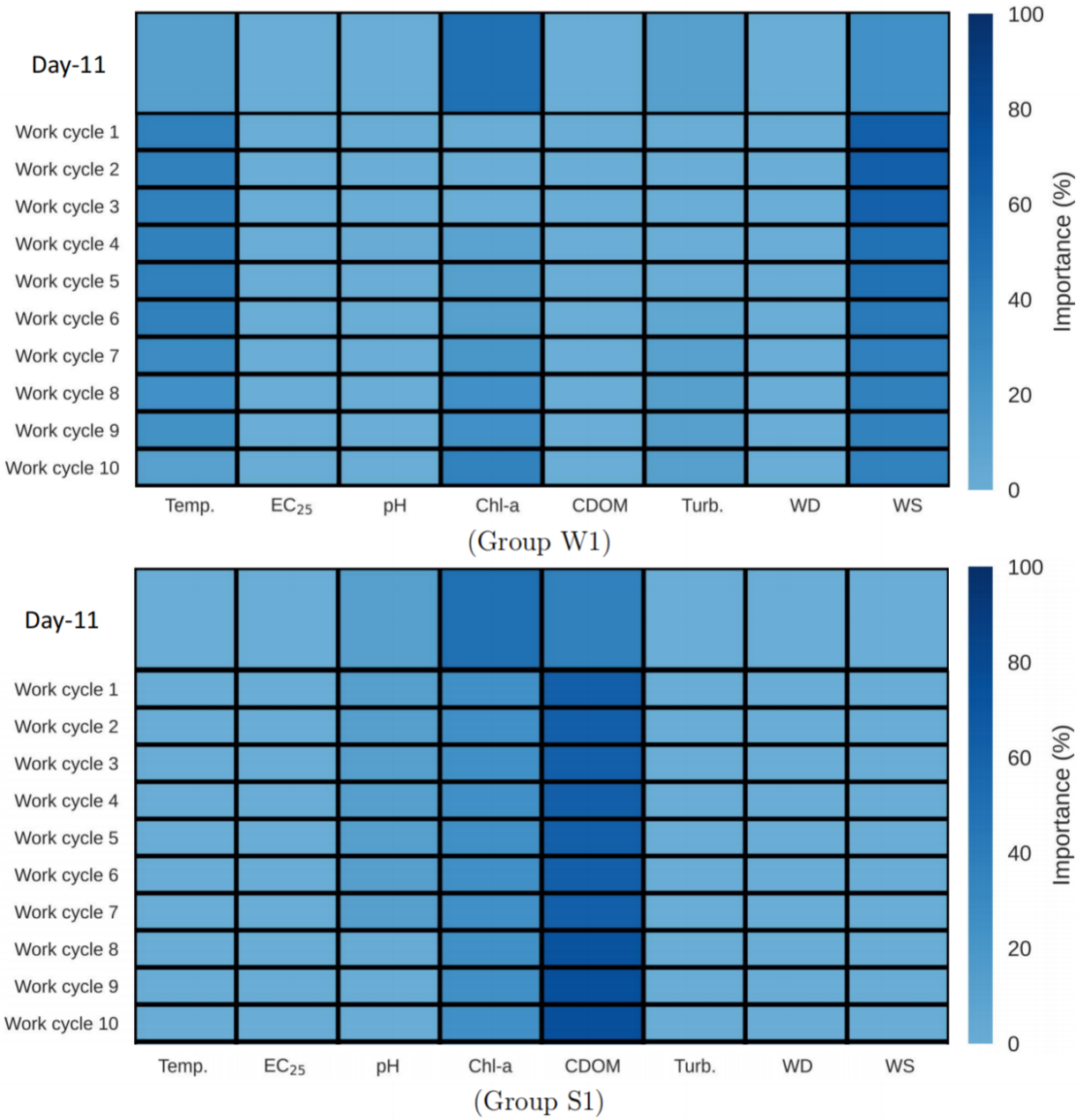


Figure 7: Results of the driving factors identification for day-11 and each 10 work cycle prediction

Discussion

Advantages of the developed system

The architectural design and application of our system, examined through a case study, demonstrate its substantial role in enabling early warnings for HABs.

The data infrastructure of the system is built upon VAMS, which offers a sturdy base for the HABs early warning system. The utility of big data methodologies over traditional methods is evident in scientific exploration.[34,35] These methodologies pave the way for creating a data platform with big data characteristics, overcoming the hurdles posed by conventional data collection methods, particularly in terms of human, material, and temporal constraints. Utilizing VAMS, we could successfully construct a data platform adhering to the 5Vs properties[36] of big data within the prescribed time limit.

A key aspect of our approach is the integration of water depth data, providing a comprehensive view of algal distributions. Notably, algal distributions and productivity have a significant correlation with water depth, indicating possible HABs occurrences at varying water depths.[9] Accordingly, we've developed prediction models focused on water depth, aiding in the analysis of algal vertical distribution at the study site. This provides predictive data with depth labels, leading to the development of targeted measures for specific water depths. The use of 'DeepDPM-Spectral Clustering' to optimize our modeling strategy resulted in a major reduction in model quantity, while also establishing a scientific suggestion for future sampling.

Predicting algal blooms, an intricate multivariate TSF procedure necessitates deciphering temporal and spatial correlations. Traditional modeling methodologies engage in sequential assimilation of these relationships – initially attaining knowledge of a singular relationship, then employing this amalgamated outcome as input to comprehend an additional relationship. Nonetheless, the loss function's direct affiliation with the apex of the model means complete optimization is attainable solely during the latter phase of the learning process. Thus,

models engineered in this manner possess an inherent predilection towards either the temporal or spatial domains. For instance, LSTM prioritizes spatial-temporal sequences, thereby emphasizing comprehension of temporal relationships. In essence, conventional models focus more on localized learning, frequently disregarding a comprehensive global understanding. Bloomformer-2, however, circumvents this predicament through its innovative design featuring parallel, interconnected temporal and spatial modules. This allows for real-time interaction and integrated learning of both dimensions. Consequently, Bloomformer-2 possesses unique capabilities for TSF tasks, such as early warning of HABs, exhibiting marked superiority over traditional models.

Preventive measures are formulated by identifying the factors influencing predicted values. The system, besides accurately predicting HABs occurrences, can determine the driving factors and their respective weights with precision based on its multi-head-self-attention mechanism.[37] Adjustments to relevant water quality parameters of medium and high weights can control or suppress HABs manifestation. A thorough prevention program can be developed for medium and long-term periods after identifying the combined driving factors in multi-step prediction outcomes. Specific high-risk temporal junctures can be addressed with tailored preventive measures.

A noteworthy feature of the HABs early warning system is its superior intelligence. The prediction module has achieved a high level of intelligence, suggesting that the deep-learning-based prediction model can learn autonomously from pre-processed data, eliminating the need for human intervention. During the data pre-processing phase, the 'DeepDPM-Spectral clustering' method independently learns the optimal cluster number, enabling deep clustering of raw data and enhancing the intelligence of the pre-processing stage. This 'end-to-end' operation improves adaptability to incoming data and holds promising implications for future system upgrades.

Model interpretability

Machine learning has been extensively employed across diverse fields, including environmental applications.[38] Yet, can we truly rely on these machine learning models? To address this concern, model interpretability is proposed as a solution, encompassing the concepts of transferability and comprehensibility.[39]

The innate human capacity to generalize and transfer acquired skills across various domains is crucial, and machine learning models must also function effectively in such contexts, especially under less predictable conditions.[40] Evaluation metrics demonstrate that Bloomformer-2 outperforms LSTM in both single- and multi-step predictions. Furthermore, in unstable situations, Bloomformer-2 can accurately discern fluctuations in trends, whereas LSTM falls short. For instance, in the multi-step prediction of Group W1 on the 12th day, LSTM failed to predict a sudden and substantial decrease in values, while Bloomformer-2 successfully accomplished the task. This underscores Bloomformer-2's superior transferability compared to LSTM.

Comprehensibility refers to our ability to understand a model's functioning. Transparent models are those that can be understood, while inscrutable models are deemed "black-box.[41]" The modeling logic of LSTM presupposes that data adheres to the Markov decision process,[42] considering only the relationship between two consecutive time steps. Specifically, it employs the Sigmoid function in the forget gate layer to selectively inherit information from the previous time step for predicting the next one. This selection mechanism lacks causality, potentially overlooking critical causal cues. Consequently, the model is classified as a black-box model, subject to stochastic inference. In contrast, Transformer-based models contemplate the relationship between any two-time steps directly. They utilize the Multi-Head-Attention (MHA) mechanism to consider all relationships between time steps (global, local, self, and cross) for predicting subsequent information. The Query, Key, and Value information interactions embedded in the attention mechanism exhibit causal tendencies, taking into account all causal cues for inferring the next time step data and revealing the

weight values of each attention head explicitly and in real-time. In addition, the attribution algorithm pertaining to the self-attention mechanism furnishes comprehensible delineations for information exchanges within the model, erecting attribution trees to manifest direct information interplays amidst disparate layers,[43] consequently amplifying the interpretability of such models grounded in the self-attention mechanism to a notable degree. Thus, Bloomformer-2 is more transparent and interpretable than the aforementioned black-box models.

Practical application

This section explores the pragmatic application of this system, which serves to offer a 3-day early warning (on the 11th, 12th, and 13th days) of HABs at experimental sites (using Group W1 and Group S1 as examples), suggesting suitable preventative measures. Adhering to the World Health Organization's "Alert Level Framework," two Chl-a thresholds (1 μg/L and 12 μg/L) are utilized to ascertain the conditions for lake algal bloom outbreaks.[44] The initial threshold corresponds to "Alert Level I," signifying the inception of HABs, whilst the latter corresponds to "Alert Level II," denoting severe HABs.

According to the predicted values for the three days to Group W1, Chl-a concentrations were consistently within the 1 μg/L to 12 μg/L range, suggesting the experimental sites were at "Alert Level I" throughout. Comprehensive preventative measures can be proposed for each day. Analysis of multi-step prediction results on the 11th day revealed Chl-a and wind speed as high-weight driving factors. On the 12th day, wind speed and Chl-a were prominent, while the 13th day spotlighted CDOM and wind speed. Previous research indicates that wind speed can cause bottom sediment resuspension in shallow lakes such as Taihu Lake, thereby releasing nutrients to promote algal growth.[45] Concurrently, this resuspension process can significantly augment turbidity. However, the model results demonstrate that turbidity's weight is nominal. Therefore, wind speed's high weight likely stems from its role in clustering algae from surrounding areas to the target site. The ensuing algae death elevates the CDOM

concentrations within the water column. CDOM is subject to photochemical degradation and can transition from large-molecule organic matter to small-molecule organic matter and inorganic nutrients, fostering conditions conducive to algal growth.[46] In conclusion, for Group W1, preventing algae aggregation around the area is instrumental in devising preventative measures, such as the installation of algae interception screens in the area's periphery.

Given the predicted values for the three days of Group S1, nine time points fell into Alert Level II, while the remaining were in Alert Level I. Initially, comprehensive preventative measures for each day can be proposed based on multi-step prediction results. High-weight driving factors on the 11th day were Chl-a and CDOM. The 12th day featured Chl-a and CDOM, while the 13th day highlighted CDOM and pH. Contrary to winter, wind speed and direction were not primary driving factors in summer, suggesting the algae predominantly originated from local growth. Additionally, factors such as algal mortality and surface runoff led to increased CDOM and pH in the water column at the experimental sites, fostering algal growth. Consequently, the removal of existing algae at the experimental sites complemented by pH adjustment may serve as an effective deterrent to HABs outbreaks of Alert Level I, such as the application of acidic algaecides and manual salvage.

Future work

Owing to Bloomformer-2's robust learning prowess, prioritizing the enhancement of the big data platform is paramount. We envision a twofold augmentation of the current big data platform. Initially, the installation of additional sensors on the VAMS is requisite, markedly amplifying data diversity. Subsequently, deploying the system across assorted water bodies can appreciably escalate the system's data capacity.

Moreover, by instituting a VAMS monitoring network, the system's application extends from a singular point (or modest area) to a more expansive territory, thereby capacitating it to execute tasks over a broader geographical expanse.

Credit Author Statement

Conceptualization: Jing Qian

Data curation: Jing Qian, Li Qian

Formal analysis: Jing Qian and Nan Pu

Funding acquisition: Stefan Norra and Yonghong Bi

Investigation: Jing Qian, Nan Pu, and Li Qian

Methodology: Jing Qian, Nan Pu, and Li Qian

Project administration: Stefan Norra

Resources: Stefan Norra and Yonghong Bi

Software: Jing Qian, Nan Pu and Li Qian

Supervision: Stefan Norra and Yonghong Bi

Visualization: Jing Qian, Nan Pu and Li Qian

Writing – original draft: Jing Qian

Writing –review & editing: Stefan Norra and Yonghong Bi

Funding

This work was supported by the Federal Ministry of Education and Research of Germany (BMBF, grant.-no.: 02WCL1336B).

Acknowledgement

The authors thank Jing-Chen Xue for her help and support in the fieldwork.

Conflicts of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

(1) Hallegraeff, G.; Enevoldsen, H.; Zingone, A. Global harmful algal bloom status reporting. *Harmful Algae* **2021**, *102*.

(2) Zheng, L.; Wang, H.; Liu, C.; Zhang, S.; Ding, A.; Xie, E.; Li, J.; Wang, S. Prediction of harmful algal blooms in large water bodies using the combined EFDC and LSTM models. *Journal of Environmental Management* **2021**, *295*, 113060.

(3) Park, Y.; Cho, K. H.; Park, J.; Cha, S. M.; Kim, J. H. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. *Science of the Total Environment* **2015**, *502*, 31–41.

(4) Lee, G.; Bae, J.; Lee, S.; Jang, M.; Park, H. Monthly chlorophyll-a prediction using neuro-genetic algorithm for water quality management in Lakes. *Desalination and Water Treatment* **2016**, *57*, 26783–26791.

(5) Su, J.; Wang, X.; Zhao, S.; Chen, B.; Li, C.; Yang, Z. A Structurally Simplified Hybrid Model of Genetic Algorithm and Support Vector Machine for Prediction of Chlorophyll a in Reservoirs. *Water (Switzerland)* **2015**, *7*, 1610–1627.

(6) Liu, X.; Feng, J.; Wang, Y. Chlorophyll a predictability and relative importance of factors governing lake phytoplankton at different timescales. *Science of the Total Environment* **2019**, *648*, 472–480.

(7) Cho, H.; Choi, U. J.; Park, H. Deep learning application to time-series prediction of daily chlorophyll-a concentration. *WIT Transactions on Ecology and the Environment* **2018**, *215*, 157–163.

(8) Na, L.; Shaoyang, C.; Zhenyan, C.; Xing, W.; Yun, X.; Li, X.; Yanwei, G.; Tingting, W.; Xuefeng, Z.; Siqi, L. Long-term prediction of sea surface chlorophyll- A concentration based on the combination of spatio-temporal features. *Water Research* **2022**, *211*, 118040.

(9) Henderson, K. A.; Murdock, J. N.; Lizotte, R. E. Water depth influences algal distribution and productivity in shallow agricultural lakes. *Ecohydrology* **2021**, *14*.

(10) Qin, B.; Zhou, J.; Elser, J. J.; Gardner, W. S.; Deng, J.; Brookes, J. D. Water Depth Underpins the Relative Roles and Fates of Nitrogen and Phosphorus in Lakes. *Environmental Science and Technology* **2020**, *54*, 3191–3198.

(11) Ronen, M.; Finder, S. E.; Freifeld, O. DeepDPM: Deep Clustering With an Unknown Number of Clusters. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2022; pp 9851–9860.

(12) McCormick, P. V.; Cairns, J. Algae as indicators of environmental change. *Journal of Applied Phycology* **1994**, *6*, 509–526.

(13) Grigsby, J.; Wang, Z.; Qi, Y. Long-Range Transformers for Dynamic Spatiotemporal Forecasting. Proceedings of ACM Conference (Conference'17). 2021.

(14) Yussof, F. N.; Maan, N.; Reba, M. N. M. LSTM networks to improve the prediction of harmful algal blooms in the west coast of Sabah. *International Journal of Environmental Research and Public Health* **2021**, *18*.

(15) Liu, M.; He, J.; Huang, Y.; Tang, T.; Hu, J.; Xiao, X. Algal bloom forecasting with

time-frequency analysis: A hybrid deep learning approach. *Water Research* **2022**, *219*, 118591.

(16) Cao, H.; Han, L.; Li, L. A deep learning method for cyanobacterial harmful algae blooms prediction in Taihu Lake, China. *Harmful Algae* **2022**, *113*, 102189.

(17) Hou, L.; Zhu, J.; Kwok, J. T.; Gao, F.; Qin, T.; Liu, T. Y. Normalization helps training of quantized LSTM. Advances in Neural Information Processing Systems. 2019; pp 1–11.

(18) Zhou, Y.; Wu, J.; Wang, B.; Duan, L.; Zhang, Y.; Zhao, W.; Wang, F.; Sui, Q.; Chen, Z.; Xu, D., et al. Occurrence, source and ecotoxicological risk assessment of pesticides in surface water of Wujin District (northwest of Taihu Lake), China. *Environmental Pollution* **2020**, *265*, 114953.

(19) Zhang, Y.; Loiselle, S.; Shi, K.; Han, T.; Zhang, M.; Hu, M.; Jing, Y.; Lai, L.; Zhan, P. Wind effects for floating algae dynamics in eutrophic lakes. *Remote Sensing* **2021**, *13*, 1–11.

(20) Chen, M.; Wang, D.; Ding, S.; Fan, X.; Jin, Z.; Wu, Y.; Wang, Y.; Zhang, C. Zinc pollution in zones dominated by algae and submerged macrophytes in Lake Taihu. *Science of the Total Environment* **2019**, *670*, 361–368.

(21) Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. **1970**,

(22) Ng, A. Y.; Jordan, M. I.; Weiss, Y. On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems. 2002.

(23) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **1997**, *9*, 1735–1780.

(24) Pascanu, R.; Mikolov, T.; Bengio, Y. On the difficulty of training recurrent neural networks. International conference on machine learning. 2013; pp 1310–1318.

(25) Siami-Namini, S.; Tavakoli, N.; Namin, A. S. The performance of LSTM and BiLSTM in forecasting time series. 2019 IEEE International Conference on Big Data (Big Data). 2019; pp 3285–3292.

(26) Sherstinsky, A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena* **2020**, *404*, 132306.

(27) Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M., et al. Transformers: State-of-the-art natural language processing. Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. 2020; pp 38–45.

(28) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.

(29) Gillioz, A.; Casas, J.; Mugellini, E.; Abou Khaled, O. Overview of the Transformer-based Models for NLP Tasks. 2020 15th Conference on Computer Science and Information Systems (FedCSIS). 2020; pp 179–183.

(30) Mohajerin, N.; Waslander, S. L. Multistep prediction of dynamic systems with recurrent neural networks. *IEEE transactions on neural networks and learning systems* **2019**, *30*, 3370–3383.

(31) Ben Taieb, S.; Hyndman, R. J. Recursive and direct multi-step forecasting: the best of both worlds. *International Journal of Forecasting* **2014**,

(32) Cai, J.; Zhang, Y.; Yang, L.; Cai, H.; Li, S. A context-augmented deep learning ap-

proach for worker trajectory prediction on unstructured and dynamic construction sites. *Advanced Engineering Informatics* **2020**, *46*, 101173.

(33) Tseng, F.-M.; Yu, H.-C.; Tzeng, G.-H. Applied hybrid grey model to forecast seasonal time series. *Technological Forecasting and Social Change* **2001**, *67*, 291–302.

(34) Chen, C. P.; Zhang, C.-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information sciences* **2014**, *275*, 314–347.

(35) Qian, J.; Liu, H.; Qian, L.; Bauer, J.; Xue, X.; Yu, G.; He, Q.; Zhou, Q.; Bi, Y.; Norra, S. Water quality monitoring and assessment based on cruise monitoring, remote sensing, and deep learning: A case study of Qingcaosha Reservoir. *Frontiers in Environmental Science* **2022**, *10*.

(36) Demchenko, Y.; De Laat, C.; Membrey, P. Defining architecture components of the Big Data Ecosystem. 2014 International conference on collaboration technologies and systems (CTS). 2014; pp 104–112.

(37) Qian, J.; Pu, N.; Qian, L.; Xue, X.; Bi, Y.; Norra, S. Identification of driving factors of algal growth in the South-to-North Water Diversion Project by Transformer-based deep learning. *Water Biology and Security* **2023**, 100184.

(38) Alzubaidi, L.; Zhang, J.; Humaidi, A. J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M. A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data* **2021**, *8*, 1–74.

(39) Calegari, R.; Ciatto, G.; Omicini, A. On the integration of symbolic and sub-symbolic techniques for XAI: A survey. *Intelligenza Artificiale* **2020**, *14*, 7–32.

(40) Lipton, Z. C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31–57.

(41) Loyola-Gonzalez, O. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE access* **2019**, *7*, 154096–154113.

(42) Puterman, M. L. Markov decision processes. *Handbooks in operations research and management science* **1990**, *2*, 331–434.

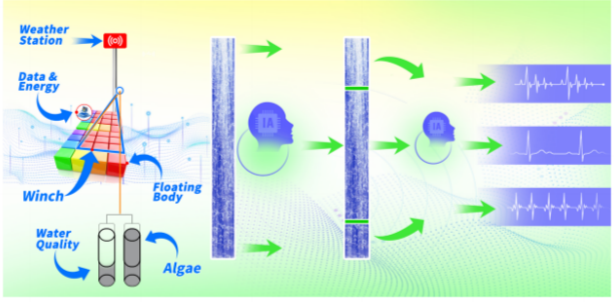
(43) Hao, Y.; Dong, L.; Wei, F.; Xu, K. Self-attention attribution: Interpreting information interactions inside transformer. Proceedings of the AAAI Conference on Artificial Intelligence. 2021; pp 12963–12971.

(44) Chorus, I.; Welker, M. *Toxic Cyanobacteria in Water*; Taylor & Francis: Geneva, 2021; Chapter 5, pp 295–400.

(45) Qin, B.; Hu, W.; Gao, G.; Luo, L.; Zhang, J. Dynamics of sediment resuspension and the conceptual schema of nutrient release in the large shallow Lake Taihu, China. *Chinese Science Bulletin* **2004**, *49*, 54–64.

(46) Wang, S.; Wang, W.; Chen, J.; Zhang, B.; Zhao, L.; Jiang, X. Characteristics of dissolved organic matter and its role in lake eutrophication at the early stage of algal blooms-A case study of Lake Taihu, China. *Water (Switzerland)* **2020**, *12*, 1–17.

TOC Graphic



For Table of Contents Only

Supporting information

Jing Qian,*,† Nan Pu,‡ Li Qian,¶ Yonghong Bi,§ and Stefan Norra‖

†*Institute of Applied Geosciences, Karlsruhe Institute of Technology, Karlsruhe 76131, Germany*

‡*Institute of Advanced Computer Science, Leiden University, Leiden, 2333 CA, Netherlands*

¶*Institute of Informatics, Ludwig Maximilian University of Munich, Munich 80538, Germany*

§*State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China*

‖*Institute of Environmental Sciences and Geography, Soil Sciences and Geoecology, Potsdam University, Potsdam-Golm 14476, Germany*

E-mail: jing.qian@partner.kit.edu

Vertical aquatic monitoring system - BIOLIFT

The photo of BIOLIFT at work is shown in Figure S1. And the panoramic photograph of the TLLER and the BIOLIFT installation position are shown in Figure S2. The sensors of BIOLIFT and their specifications are shown in Table S1. The producer of sensors of pressure, temperature and EC_{25} are ADM Elektronik GmbH. The sensors of Chl-a and CODM are produced by Turner Designs, Inc. The sensor of pH is produced by AMT GmbH. The sensor of Turbidity is produced by Seapoint Sensors, Inc.

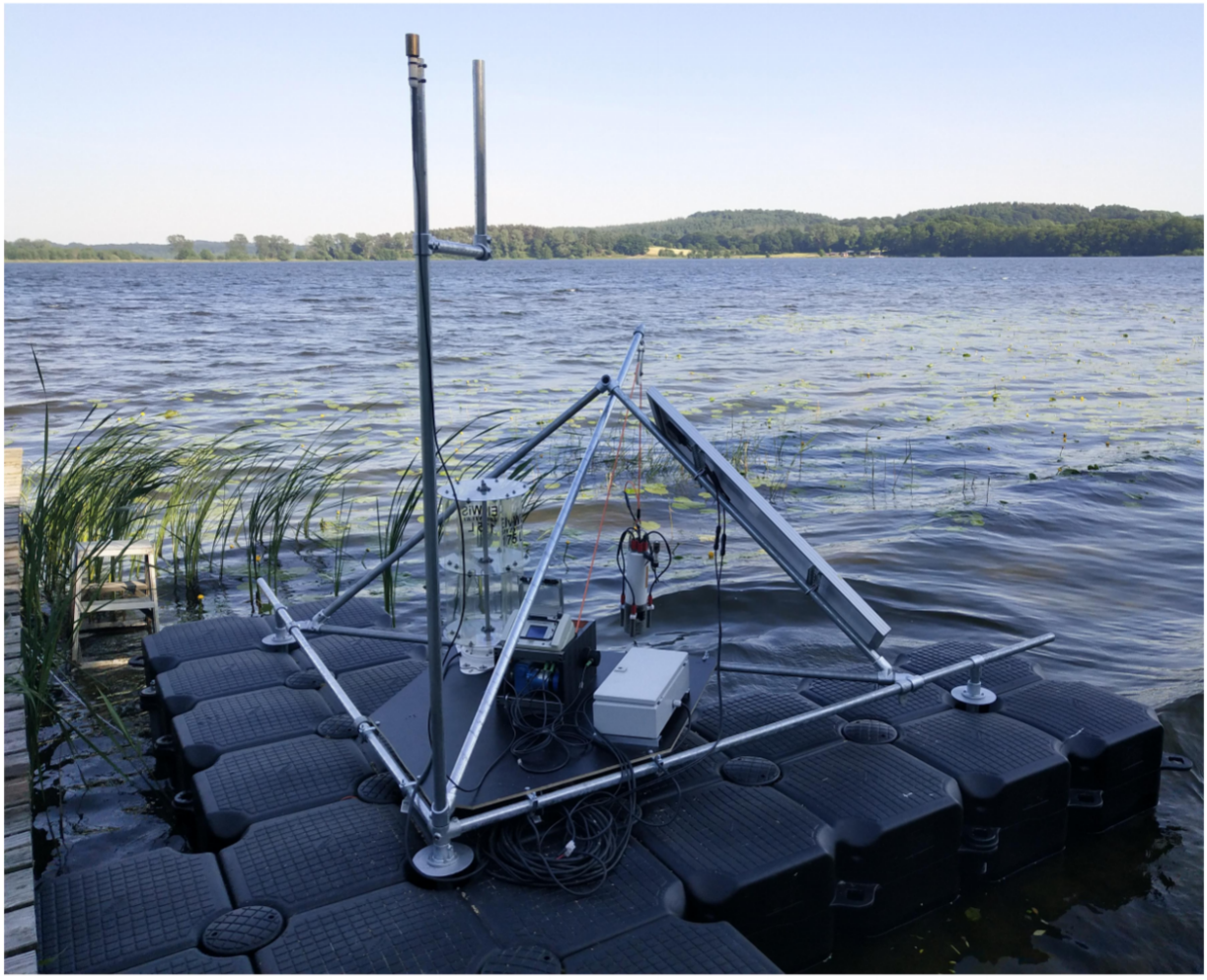


Figure S1: BIOLIFT at work ©Andre Wilhelms

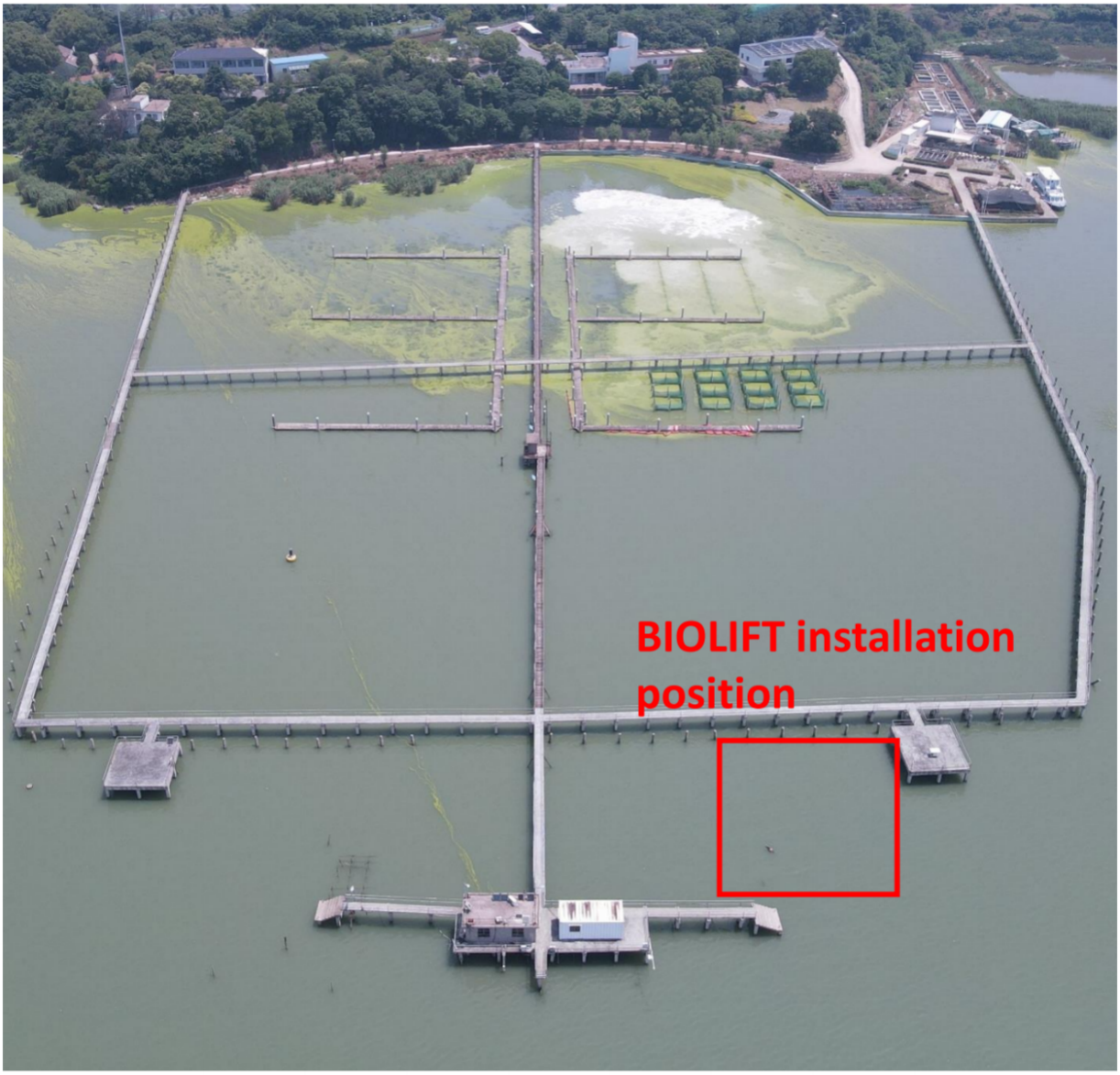


Figure S2: Panoramic photograph of the TLLER and the BIOLIFT installation position ©Jing Qian

Table S1: Sensors of BIOLIFT and their specifications

Parameter	**Principle**	**Range**	**Resolution**	**Accuracy**	**Response time**
Pressure	piezo-resistive	0-200dBar	0.005dBar	±0.1dBar	0.04s
Temp.	Pt 100	-2-38°C	0.001°C	±0.01°C	0.12s
pH	Potentiometric (Ag/AgCl)	0-14pH	0.02pH	0.02pH	1s(63%)
CDOM	Fluorescence exc.325nm fl. 470nm	0.15-1250 ppbQS	0.01ppbQS	±5%	1s
Chl-a	Fluorescence exc.465nm fl. 696nm	0.03-500 $\mu g/L$	0.01 $\mu g/L$	N.A.	1s
EC_{25}	7-pole-cell	0-6mS/cm	0.1uS/cm	±2uS/cm	0.05s
Turb.	Mie backscattering	0-750FTU	<0.001%	±2%	0.1s
WD	N.A.	0-360°	1°	±3°	0.25s
WS	N.A.	0-60m/s	0.1m/s	±3% at 10m/s	0.25s

DeepDPM

DeepDPM contains two main parts, the first part is the clustering network, and the second is K subclustering networks (one for each cluster k, $k \in \{1, \ldots, K\}$). The training process is shown in Figure S3.

In the training process, first given an arbitrary initial cluster number K, the data is fed to the clustering network f_{cl}, which generates K soft cluster assignments for each data point $\boldsymbol{x}_i$:

$$f_{\text{cl}}(\mathcal{X}) = \boldsymbol{R} = (\boldsymbol{r}_i)_{i=1}^{N}$$

$$\boldsymbol{r}_i = (r_{i,k})_{k=1}^{K}$$

Where $r_{i,k} \in [0, 1]$ is the soft cluster assignment $\boldsymbol{R}$ of $\boldsymbol{x}_i$ to cluster k and $\sum_{k=1}^{K} r_{i,k} = 1$.

Secondly, the hard assignments $\boldsymbol{z} = (z_i)_{i=1}^{N}$ are calculated according to the equation:

$$z_i = \arg\max_k r_{i,k}$$

Next, each subclustering network f_{sub}^{k} is fed the hard assignments data for its respective cluster and generates a soft subcluster assignment, as the following equations show:

$$f_{\text{sub}}^{k}\ (\mathcal{X}_k) = \tilde{\boldsymbol{R}}_k = (\tilde{\boldsymbol{r}}_i)_{i:z_i=k}$$

$$\tilde{\boldsymbol{r}}_i = (\tilde{\boldsymbol{r}}_{i,j})_{j=1}^{2}$$

Where $\tilde{r}_{i,j} \in [0, 1]$ is the soft assignment of $\boldsymbol{x}_i$ to subcluster $j (j \in \{1, 2\})$, and $\tilde{r}_{i,1} + \tilde{r}_{i,2} = 1 \forall k \in \{1, \ldots, K\}$.

The clustering network f_{cl} and each subclustering network f_{sub}^{k} is a simple multilayer perceptron with a single hidden layer. The last layer of the clustering network has K neurons, while the last layer of each subclustering network has two.

Finally, the split or merge decisions are made for changing K according to the Metropolis-Hastings framework.[1]

The split proposals are accepted stochastically with probability $\min(1, H_s)$, where H_s is Hastings ratio. In the split step, each cluster is split into its two subclusters. The merge proposals are accepted/rejected using the reciprocal number of the Hastings ratio H_s.

$$H_s = \frac{\alpha\Gamma(N_{k,1}) f_{\boldsymbol{x}}(\mathcal{X}_{k,1};\lambda)\Gamma(N_{k,2}) f_{\boldsymbol{x}}(\mathcal{X}_{k,2};\lambda)}{\Gamma(N_k) f_{\boldsymbol{x}}(\mathcal{X}_k;\lambda)}$$

Where H_s is the Hastings ratio, Γ is the Gamma function, $\mathcal{X}_k = (\boldsymbol{x}_i)_{i:z_i=k}$ stands for the points in the cluster k, $N_k = |\mathcal{X}_k|$, $\mathcal{X}_{k,j} = (\boldsymbol{x}_i)_{i:(z_i,\tilde{z}_i)=(k,j)}$ denotes the points in the subcluster, $j(j \in \{1,2\})$, $N_{k,j} = |\mathcal{X}_{k,j}|$, and $f_{\boldsymbol{x}}(\cdot;\lambda)$ is the marginal likelihood where λ represents the Normal-Inverse Wishart hyperparameters.[2]

After the split and merge steps, the initial cluster number K, clustering network, and K subclustering networks are updated, and iterative operations are performed until the optimal cluster number K is found.

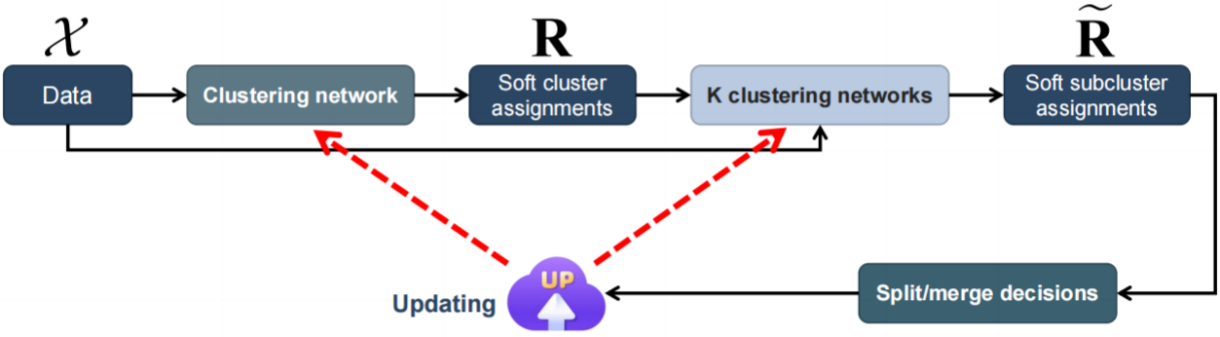


Figure S3: Training process of DeepDPM

Spectral cluster

Spectral clustering[3] is an unsupervised machine learning technique used for partitioning data into groups or clusters based on the similarity between data points. The main idea behind spectral clustering is to analyze the eigenvectors and eigenvalues of the Laplacian matrix

derived from the data's adjacency matrix. The steps of spectral clustering are as shown below.[3]

Given a set of points $S = \{s_1, \ldots, s_n\}$ in $\mathbb{R}^l$ that we want to cluster into k subsets:

1. Form the affinity matrix $A \in \mathbb{R}^{n \times n}$ defined by $A_{ij} = \exp\left(-\|s_i - s_j\|^2 / 2\sigma^2\right)$ if $i \neq j$, and $A_{ii} = 0$.

2. Define D to be the diagonal matrix whose (i,i)-element is the sum of A 's i-th row, and construct the matrix $L = D^{-1/2} A D^{-1/2}$.1

3. Find $x_1, x_2, \ldots, x_k$, the k largest eigenvectors of L, and form the matrix $X = [x_1 x_2 \ldots x_k] \in \mathbb{R}^{n \times k}$ by stacking the eigenvectors in columns.

4. Form the matrix Y from X by renormalizing each of X 's rows to have unit length (i.e. $Y_{ij} = X_{ij} / \left(\sum_j X_{ij}^2\right)^{1/2}$).

5. Treating each row of Y as a point in $\mathbb{R}^k$, cluster them into k clusters via K-means or any other algorithm (that attempts to minimize distortion).

6. Assign the original point s_i to cluster j if and only if row i of the matrix Y was assigned to cluster j.

LSTM

We utilized a well-established DL model for time series prediction, known as LSTM,[4] to serve as a comparison with Bloomformer-2. LSTM is a type of recurrent neural network (RNN)[5] architecture designed to address the vanishing gradient[6] problem commonly encountered in traditional RNN. LSTM have a more complex structure than standard RNN, incorporating memory cells and various gates to control the flow of information through the network.

The architecture of the LSTM is shown in Figure S4. For moment t, the LSTM has three inputs: the cell state C_{t-1}, the hidden layer state h_{t-1}, and the input vector at moment t, X_t. In addition there are two outputs: the cell state C_t and the hidden layer state h_t, where h_t is also used as the output at moment t.

The gate layers of the LSTM is designed with some computational steps to adjust the input with the values of the two hidden layers. The gate layers in LSTM contains forget gate layer, input gate layer and output gate layer. The square components in Figure S4 represent neurons, and the difference between them is the difference in activation functions. σ denotes the Sigmoid function, whose output is between 0 and 1, and tanh is the hyperbolic tangent function, whose output is between -1 and 1. The role of forget gate layer is to selectively forget the information in the cellular state, and the function is:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

The input gate layer is used to selectively record new information into the cell state, and the functions of input gate layer are:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

$$\tilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

The output gate layer is used to save the previous information into the hidden layer and output a time step value and the functions of output gate layer are:

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

The LSTM architecture allows the network to learn and retain long-range dependencies

in the input data by controlling the flow of information through the memory cell and gates.

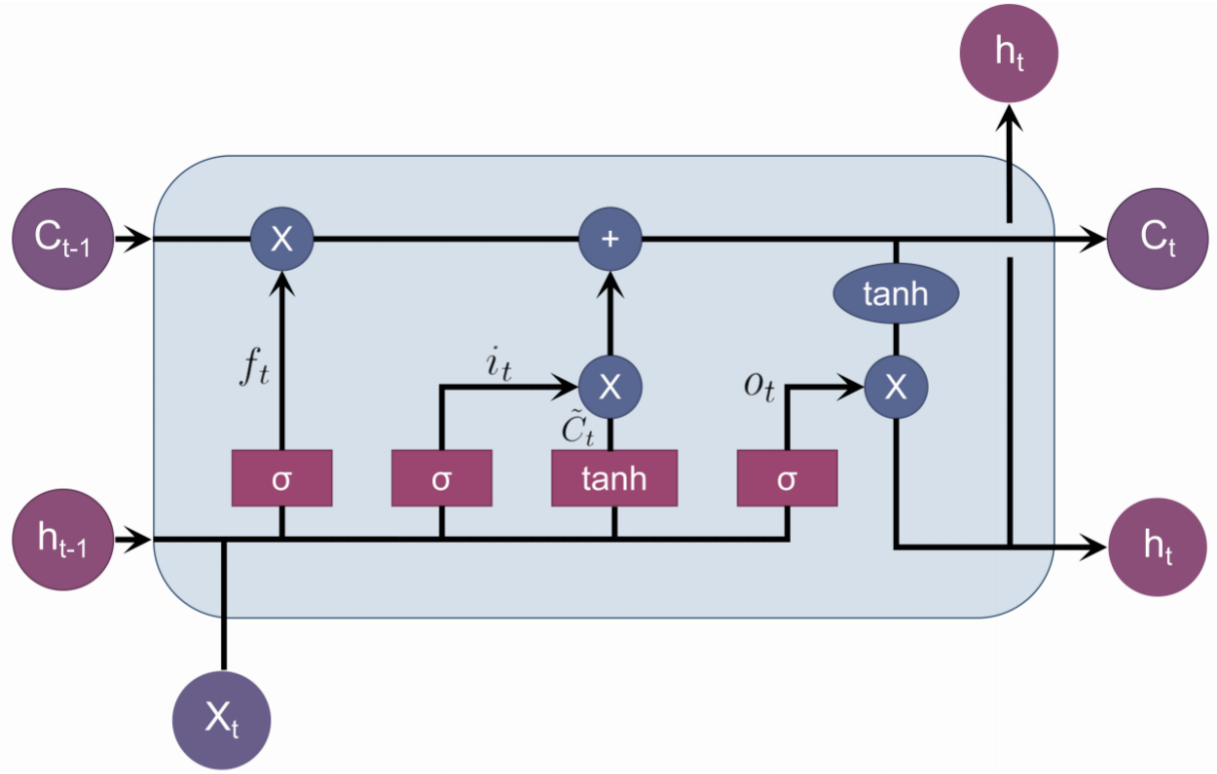


Figure S4: Architecture of LSTM

Transformer

Architecture

According to the article "Attention is all you need", the architecture of Transformer, depicted in Figure S5, operates on the fundamental principle of an encoder-decoder construct.[7] The leftmost section represents the encoder, while the decoder finds its placement on the right.

As for the encoder, it is constituted of six identically designed layers, epitomizing 'N' in the corresponding architectural schematic. Each encoder comprises dual sub-components: the inaugural one is the Multi-Head Attention, succeeded by a position-wise, fully connected feed-forward network forming the second. The two sub-layers are intricately linked via residuals, culminating in Layer Normalization, a process indicated as 'Add&Norm' in the associated architectural illustration.

Switching focus to the decoder, it too is fashioned out of six congruent layers. Each decoder encompasses three sub-layers, two of which are designed analogously to those in the encoder, with the supplementary inclusion of multi-head attention, tasked with processing the output produced by the encoder layer. The decoder, much like its encoder counterpart, utilizes residual connections and Layer Normalization.

Attention

Within the realm of the Attention function, the triad of Q(Query), K(Key), and V(Value) are transformed into an output, whereby the aforementioned triad and the output manifest as vectors. The output unfurls as a weighted summation of V(Value), the weightage of which is ascertained by scrutinizing the fusion of Q(Query) and K(Key).

Scaled dot-product attention

The equation of the scaled dot-product attention is as follows.

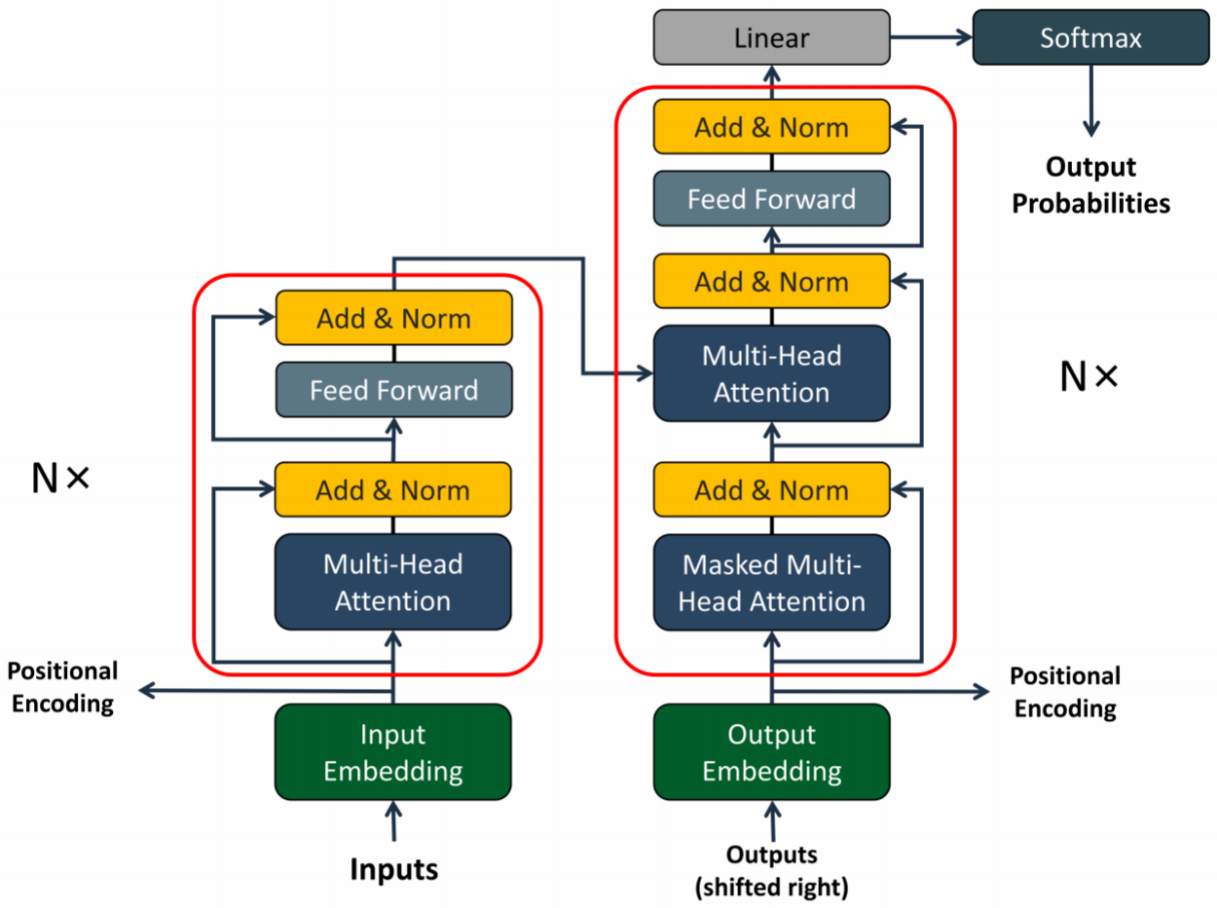


Figure S5: Architecture of standard Transformer

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The vector dimension in both Q and K in the above equation is d_k, and the vector dimension of V is d_v. In self-attention, $d_k = d_v = d_{wordEmbedding/numHeads}$. The commonly used attention function is dot-product (multiplicative) attention.

Multi-Head Attention

The function of Multi-Head Attention is:

$$\text{MultiHead}(Q, K, V) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right)W^O$$

Where $\text{head}_\text{i} = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$ Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$

The calculation process is shown below:

a. Assume that the number of heads is now h. First, the vectors are divided into h equal parts according to the length of the vectors on each time sequence.

b. Then the new values of Q, K, and W are obtained by mapping the above equal parts of h data with different weights.

c. The h copies of the above mapped data are used to calculate the value of the corresponding Attention.

d. It is reassembled according to the form of the previous segmentation and then mapped to the original vector dimension. Then we get the value of Multi-Head Attention.

Sequence to Sequence

The output of Sequence to Sequence (Seq2Seq) tactic is germane to the network endowed with an Encoder-Decoder framework, wherein both the input and output take the form of

sequences. In the Encoder, the sequence undergoes metamorphosis into a fixed-length vector, which is then transformed by the Decoder into the desired output sequence.

Evaluation metrics

MAPE was calculated as:

$$MAPE\left(y,\hat{y}\right)=\frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i-\widehat{y_i}}{y_i}\right|$$

MAE was calculated as:

$$MAE(y,\hat{y})=\frac{\sum_{i=1}^{n}\left|y_i-\widehat{y_i}\right|}{n}$$

MSE was calculated as:

$$MSE(y,\hat{y})=\frac{\sum_{i=1}^{n}\left(y_i-\widehat{y_i}\right)^2}{n}$$

where $\widehat{y_i}$ is the predicted value of the ith sample, y_i is the corresponding true value of the total n samples, and $\overline{y_i}$ is the mean of true value.

Z-score

All data were Z-score normalized before being input to the model according to the following equation:

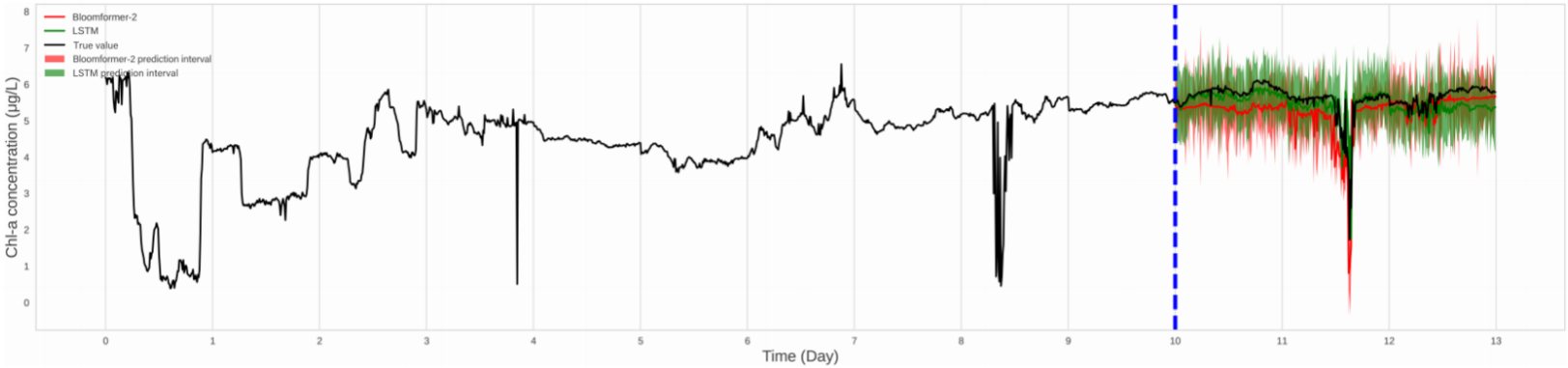
$$Z_i=\frac{x_i-\overline{x_i}}{\sigma_i}$$

where Z_i is the standard score of i-th data, x_i is the -th original data, $\overline{x_i}$ is the mean of i-th data, and σ_{i} is the standard deviation of i-th data.

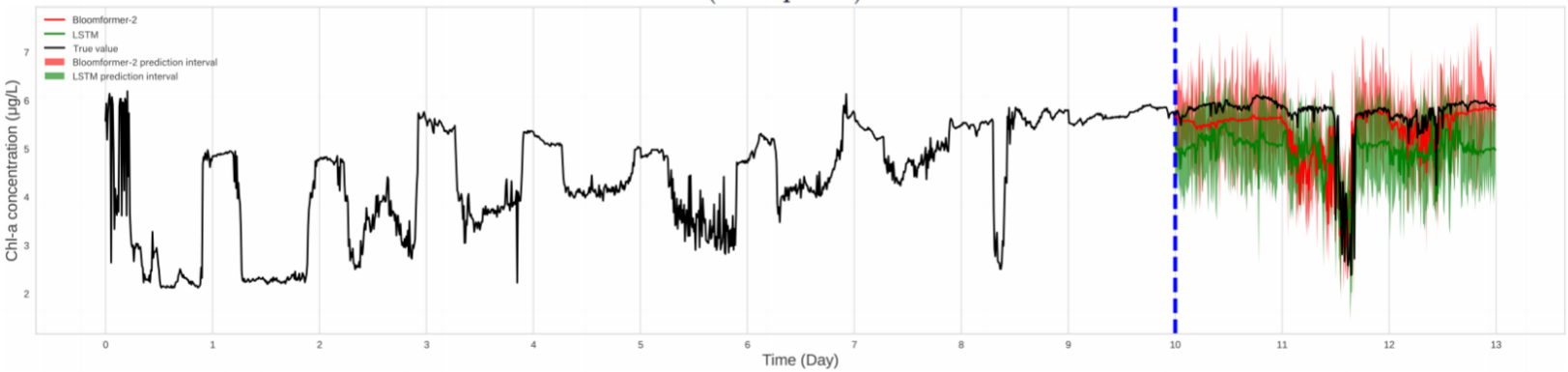
Result of prediction

Single prediction

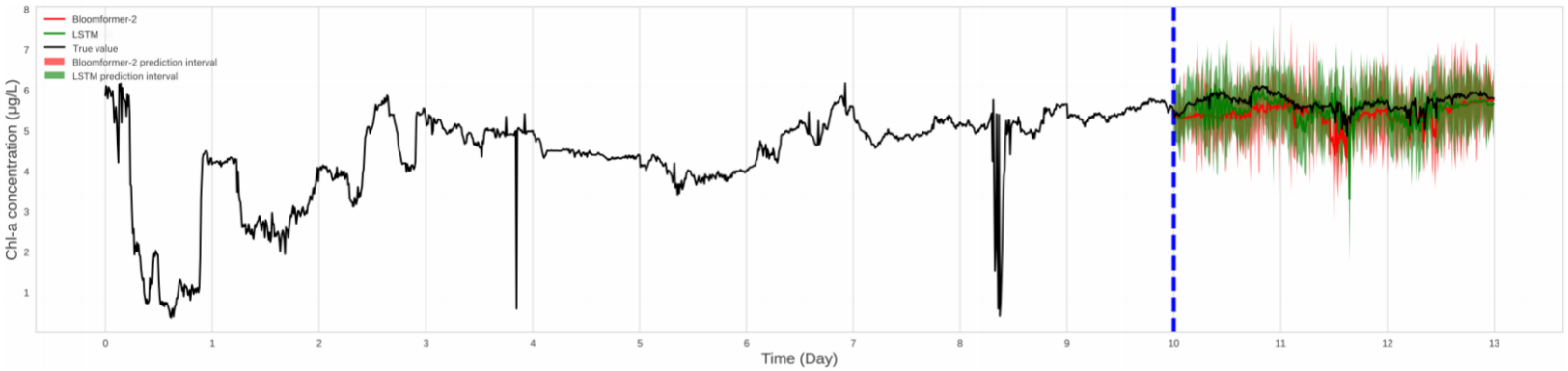
The single-step predictive outcomes for Group W2 to Group W4 are illustrated in Figure S6, while the results for Group S2 to S5 are shown in Figure S7.



(Group W2)

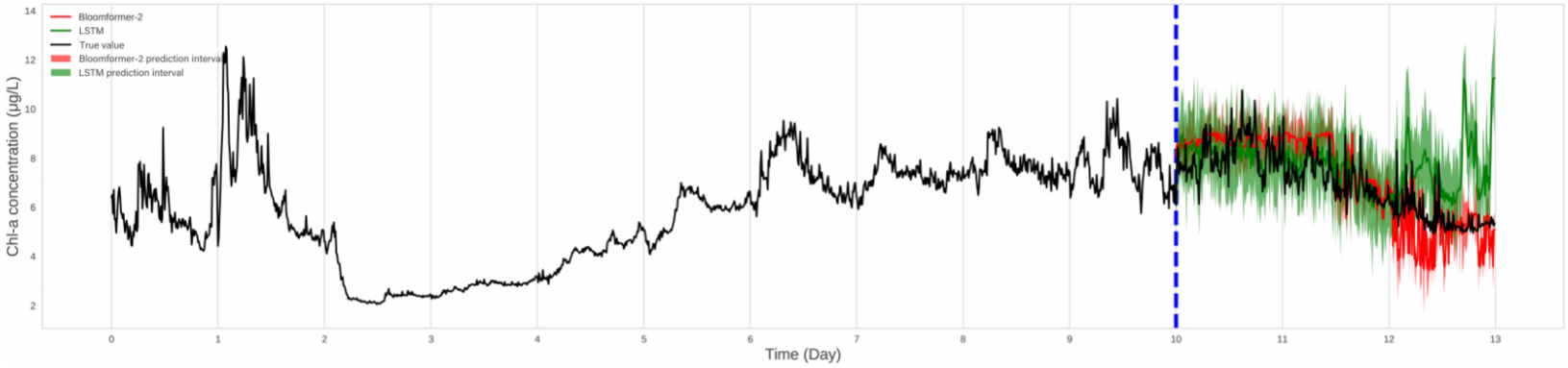


(Group W3)

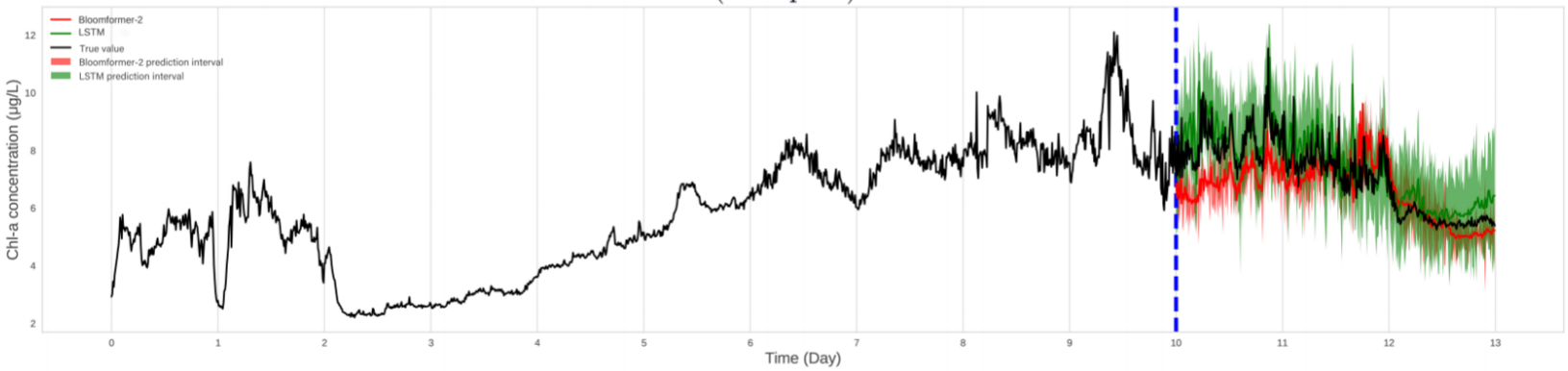


(Group W4)

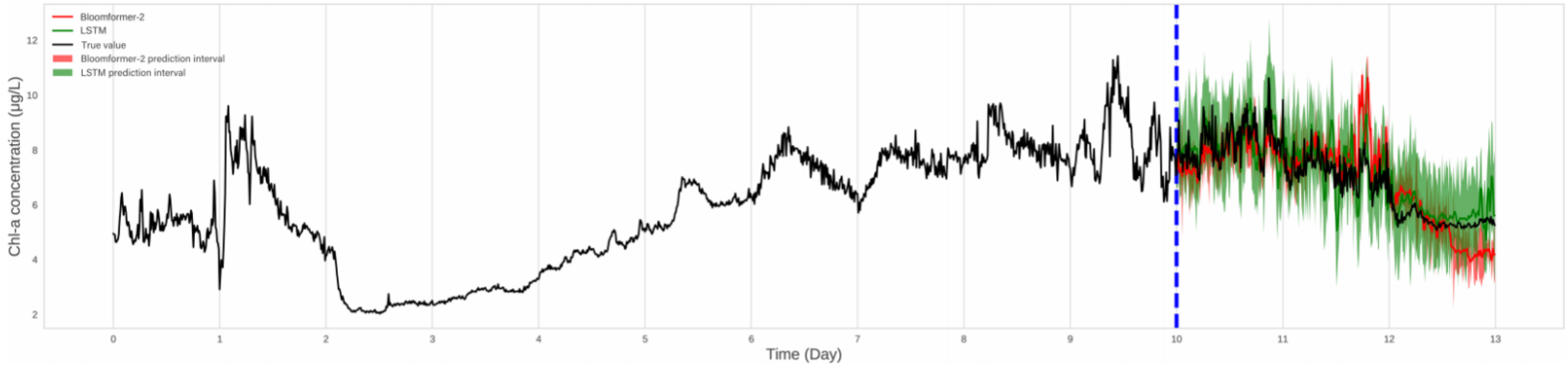
Figure S6: Comparison of model prediction for Group W2 to W4 in single-step prediction



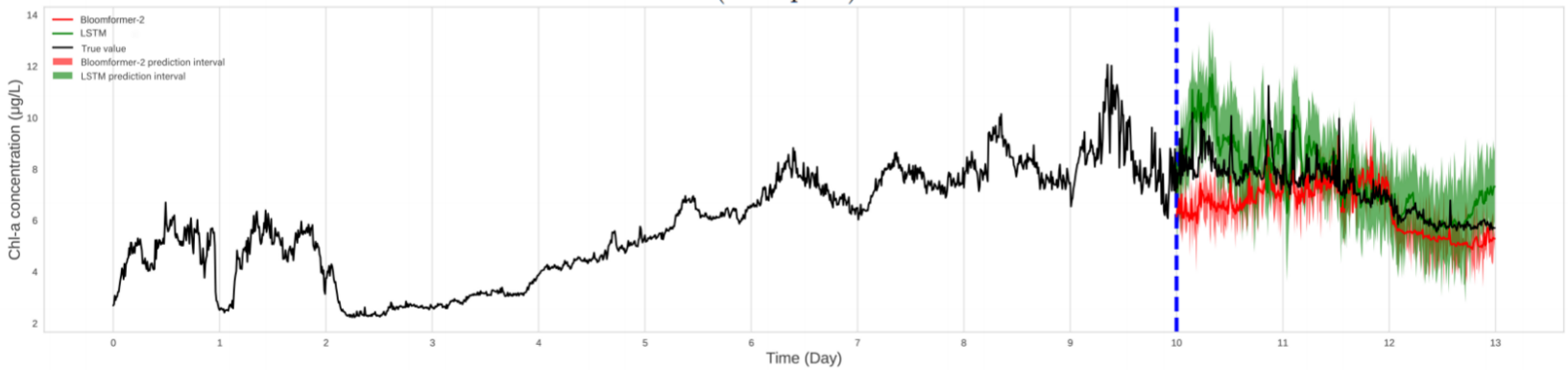
(Group S2)



(Group S3)



(Group S4)

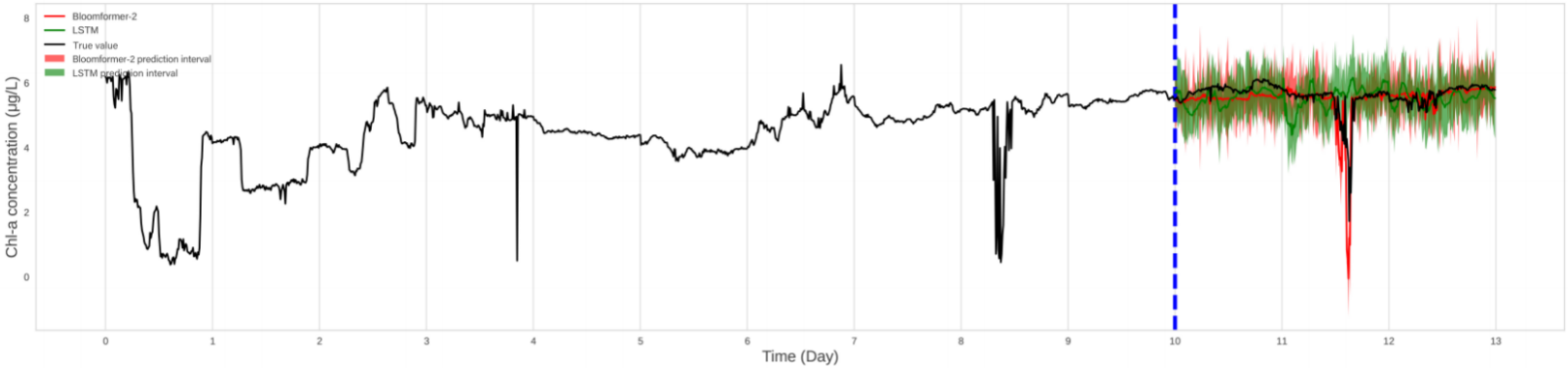


(Group S5)

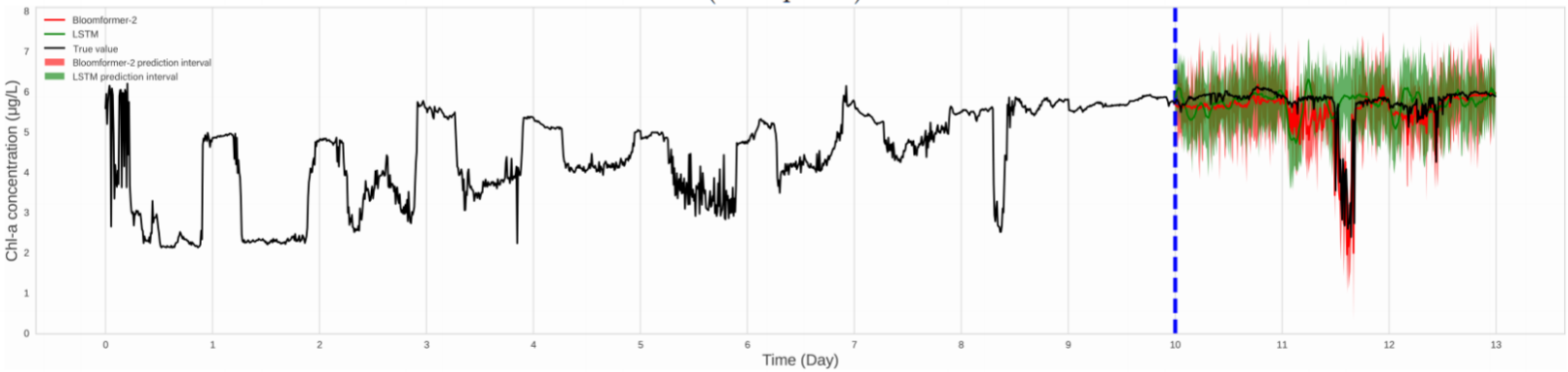
Figure S7: Comparison of model prediction for Group S2 to S5 in single-step prediction

Multi-step prediction

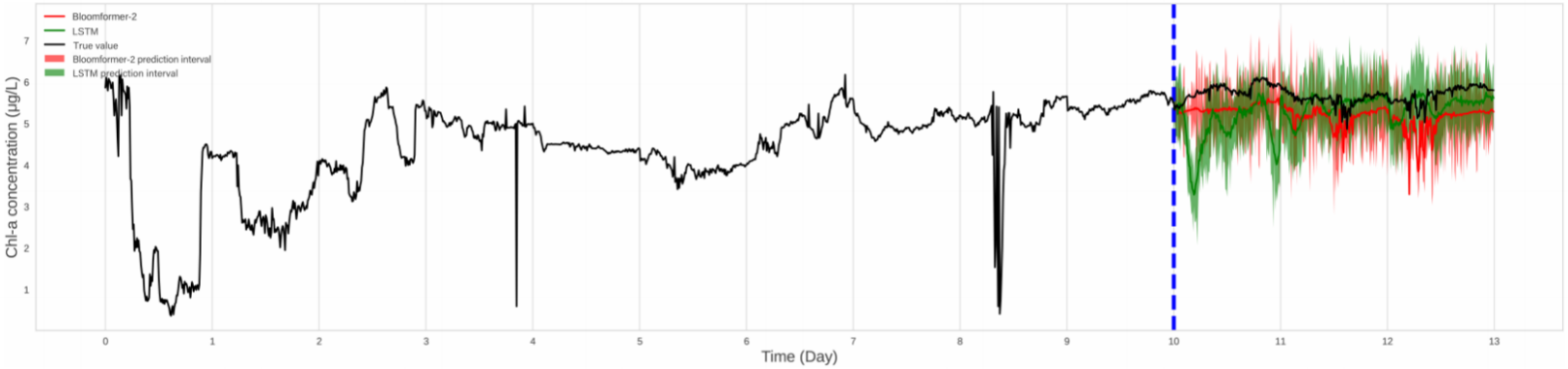
The multi-step predictive outcomes for Group W2 to Group W4 are illustrated in Figure S8, while the results for Group S2 to S5 are shown in Figure S9.



(Group W2)

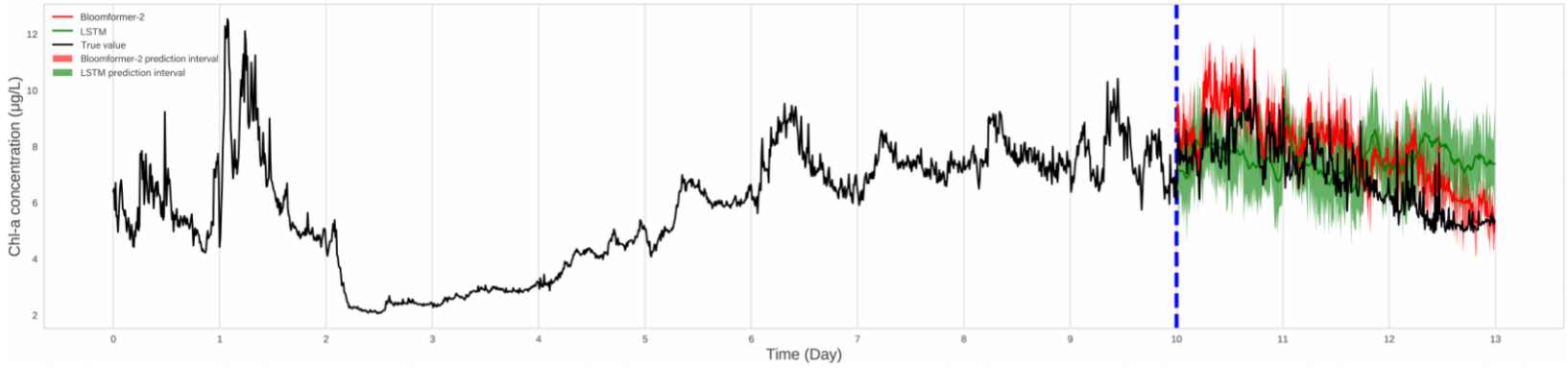


(Group W3)

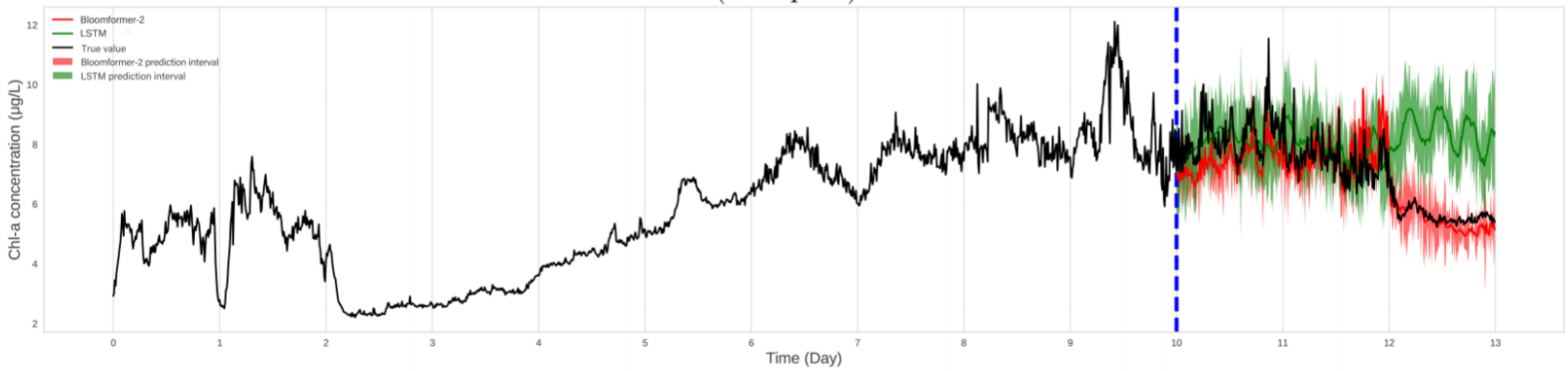


(Group W4)

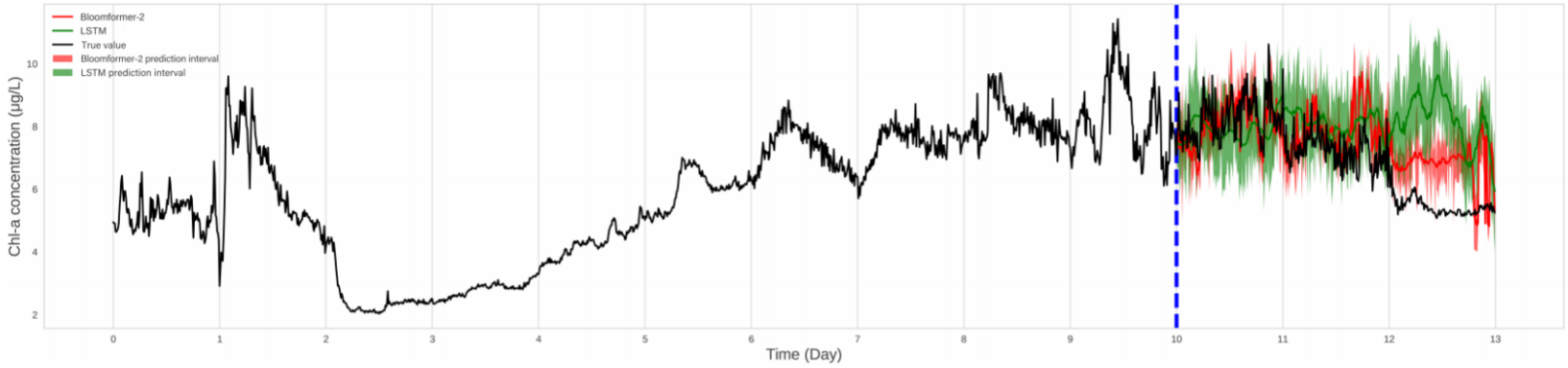
Figure S8: Comparison of model prediction for Group W2 to W4 in multi-step prediction



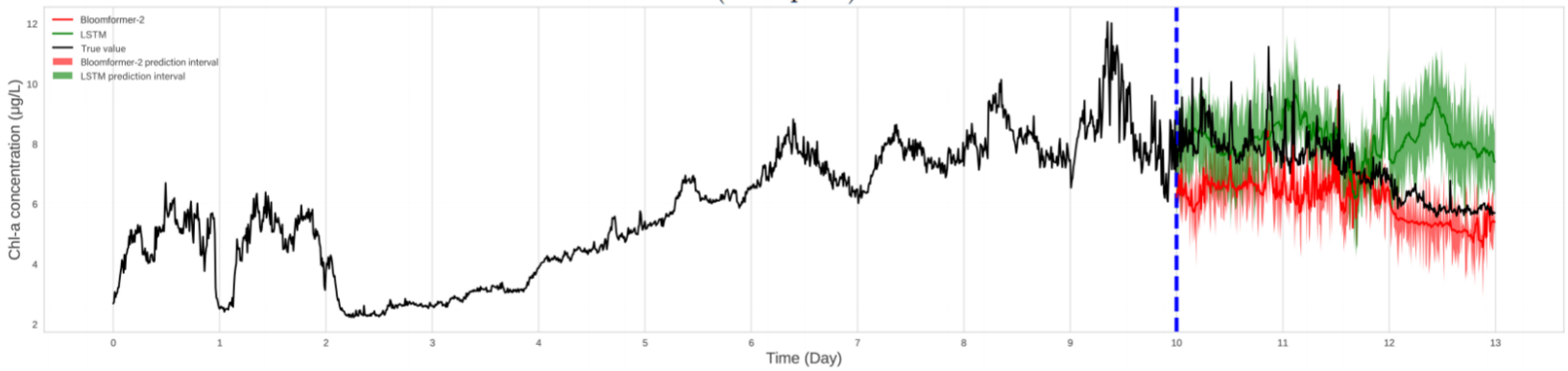
(Group S2)



(Group S3)



(Group S4)



(Group S5)

Figure S9: Comparison of model prediction for Group S2 to S5 in multi-step prediction

Identification of Driving factors for the predicted value

The comprehensive driving factors of 11th to 13th prediction for Group W1 and S1 are shown in Figure S10 and Figure S12, respectively. And the driving factor of the predicted value for all work cycles for Group W1 and S1 on the 11th day is shown in Figure S11 and Figure S13, respectively.

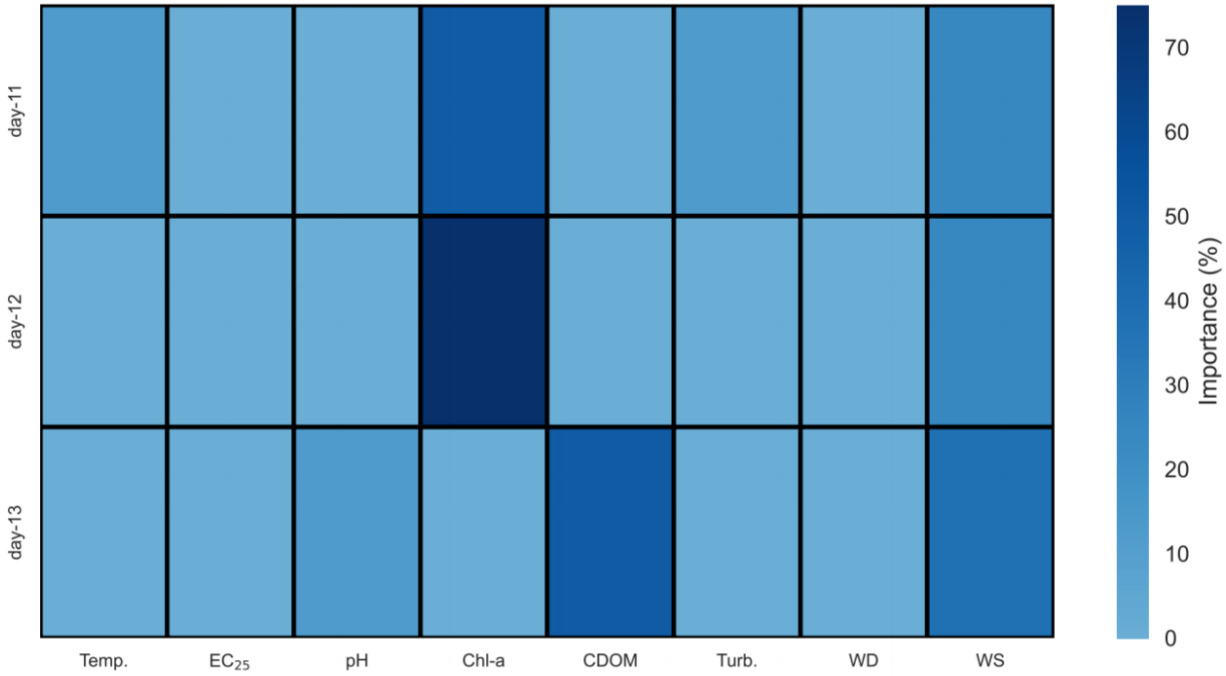
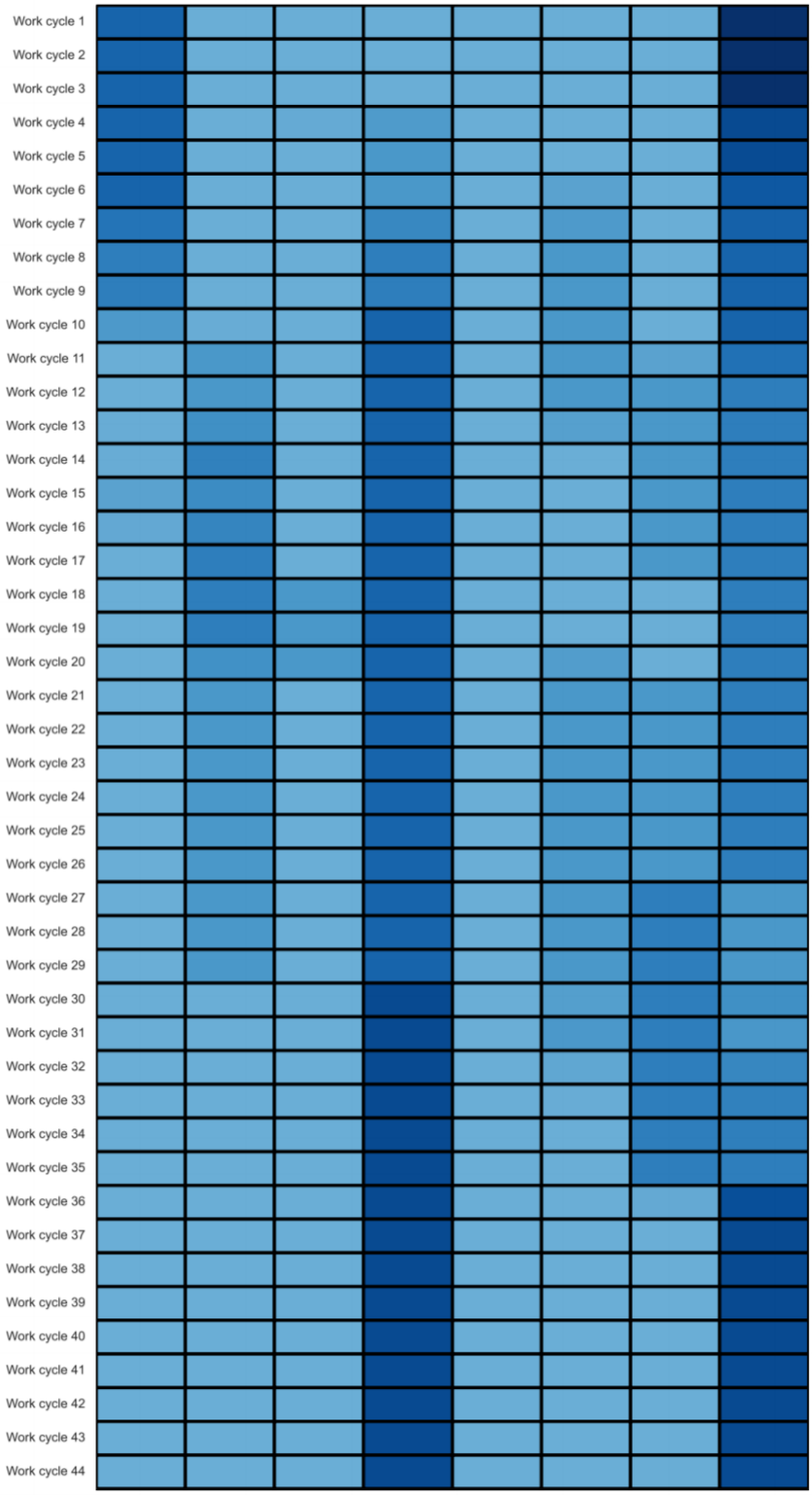
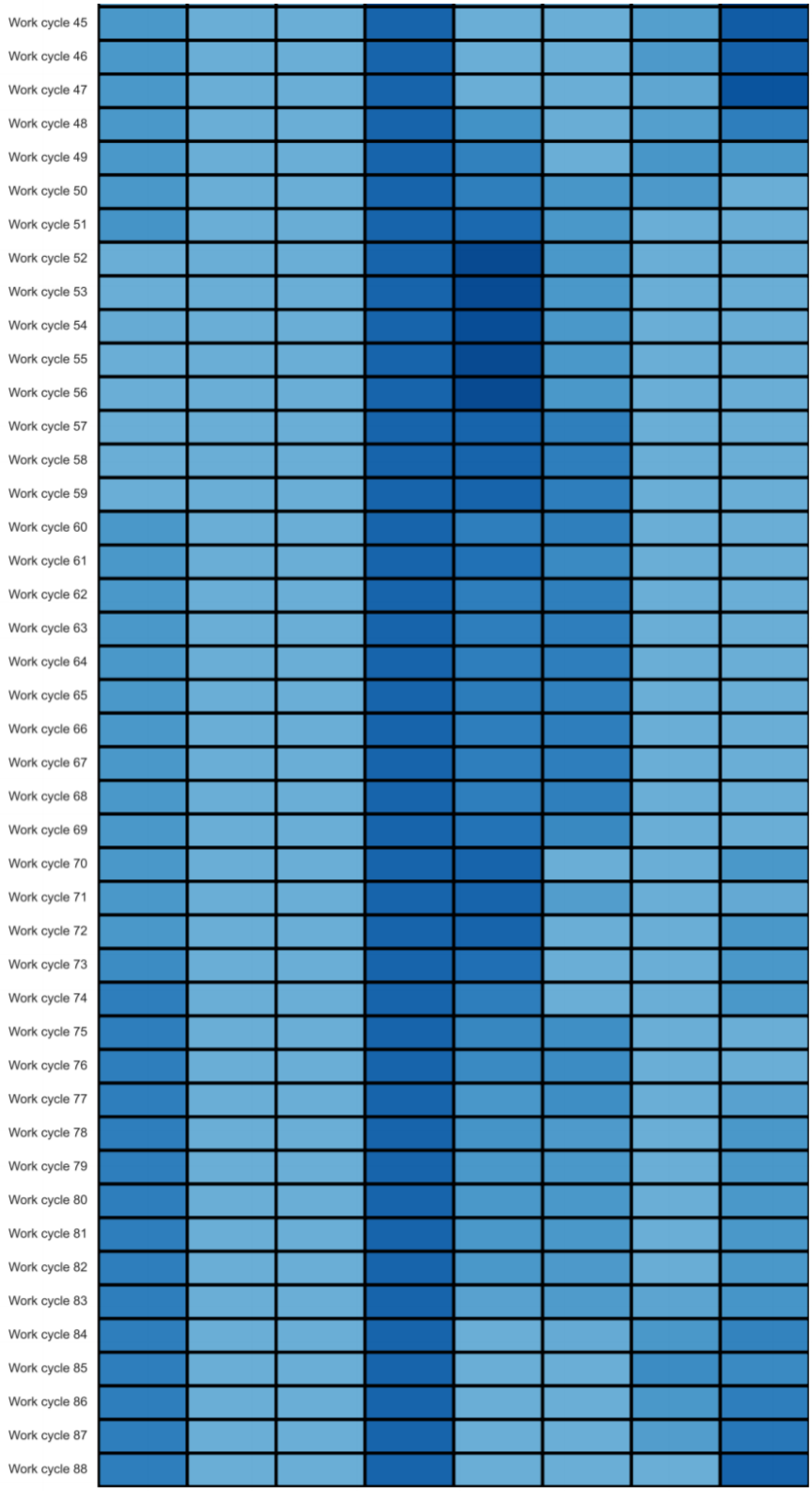


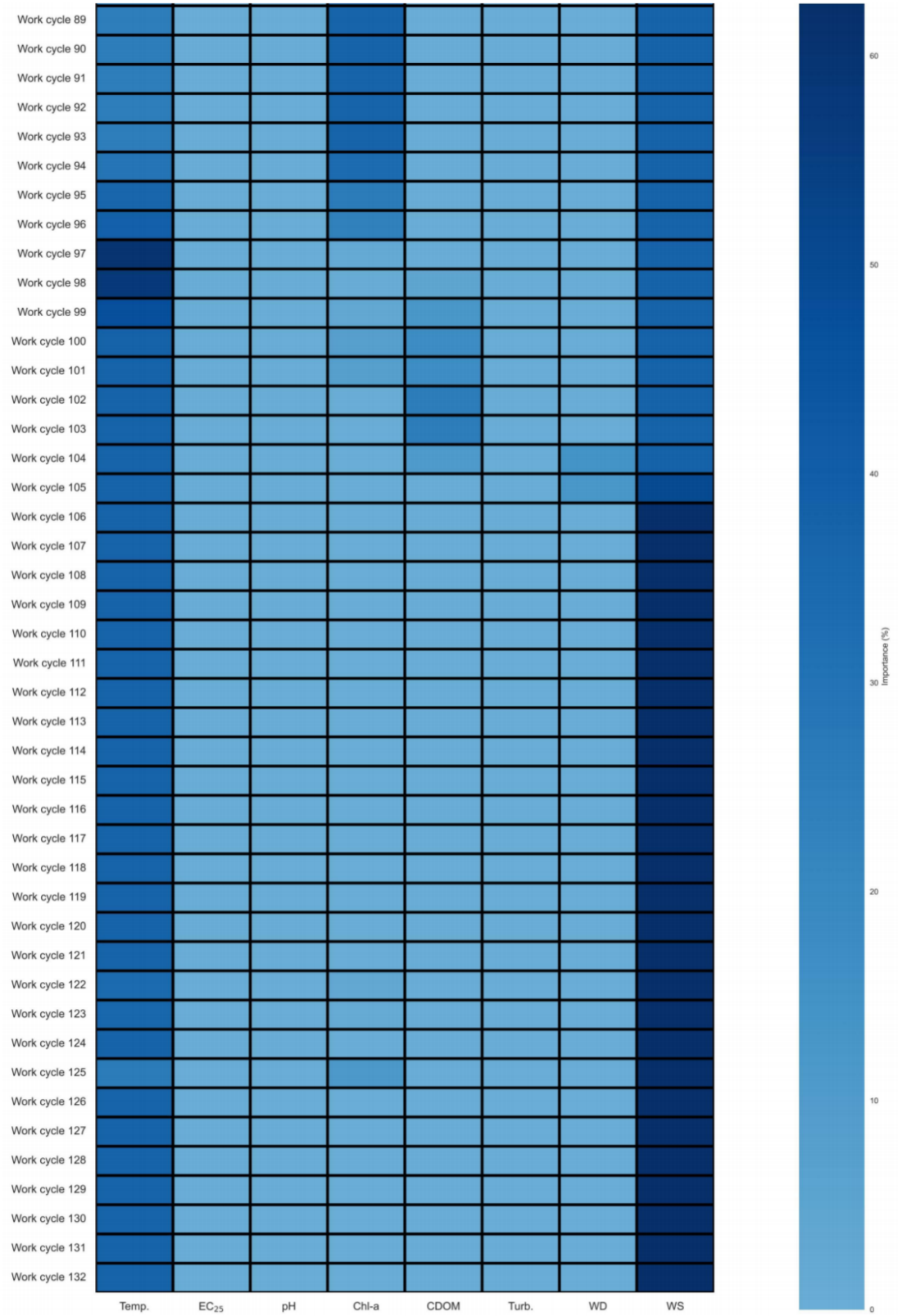
Figure S10: Driving factor of 11th to 13th day prediction for Group W1



(Work cycle 1 - Work cycle 44)



(Work cycle 45 - Work cycle 88)



(Work cycle 89 - Work cycle 132)

Figure S11: Driving factor of predicted value for all work cycles for Group W1 on 11th day

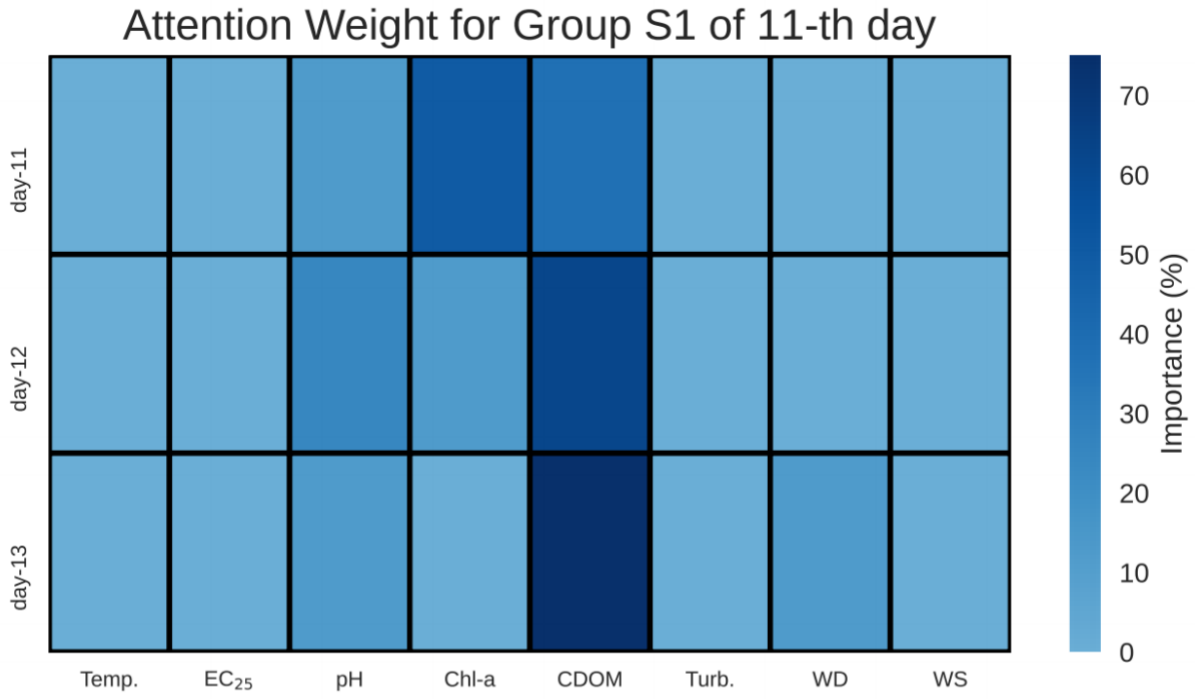
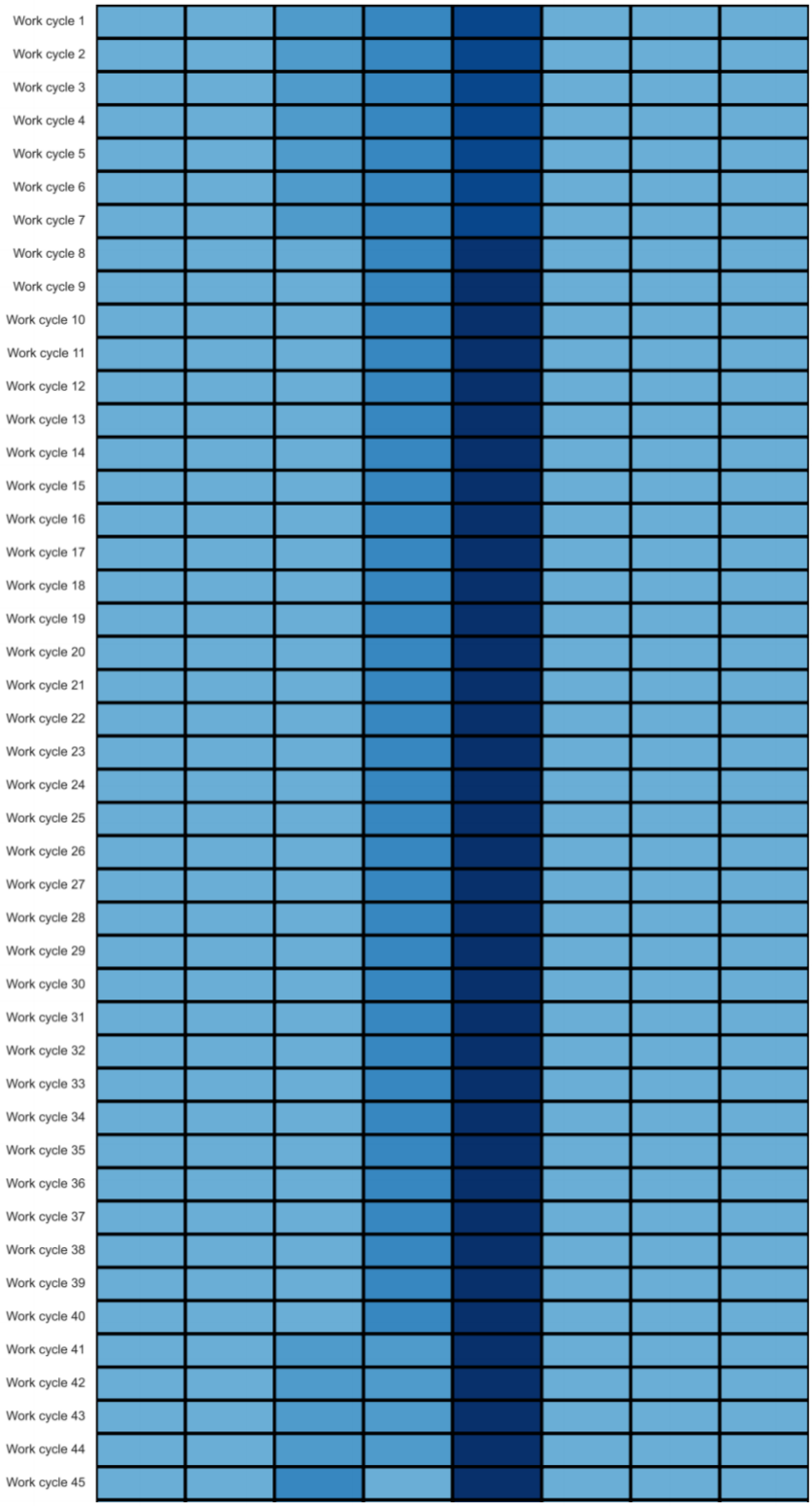
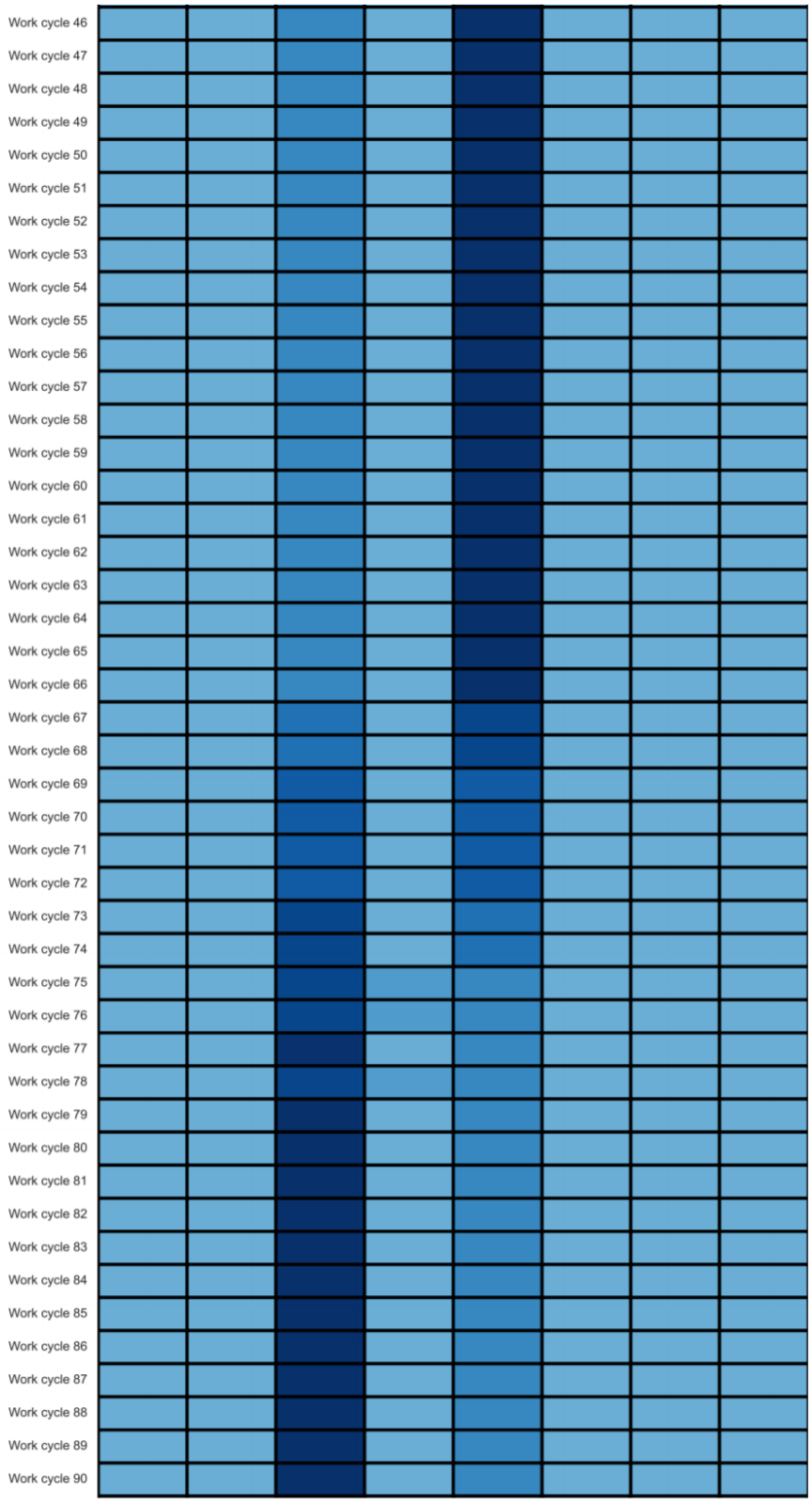


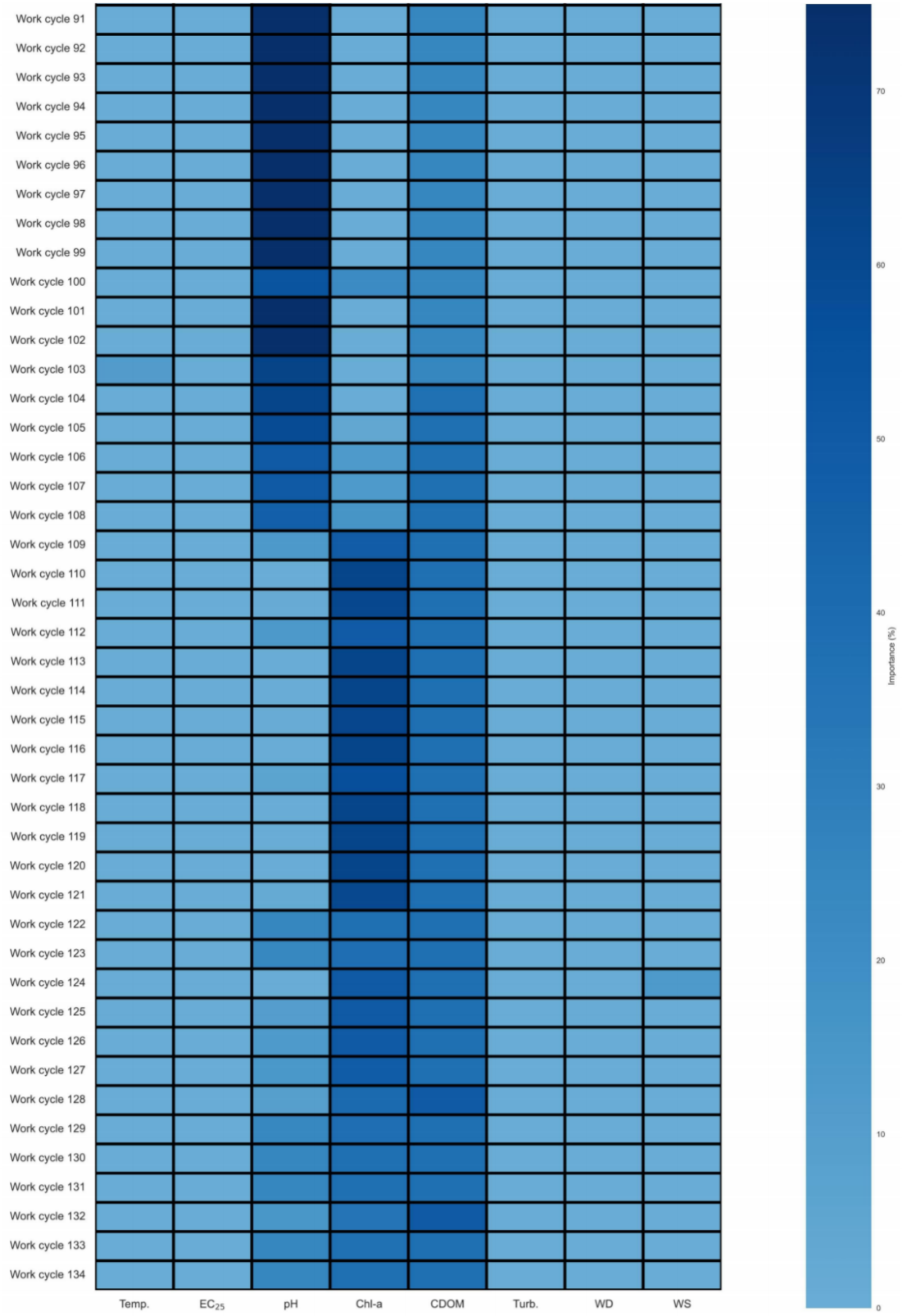
Figure S12: Driving factor of 11th to 13th day prediction for Group S1



(Work cycle 1 - Work cycle 45)



(Work cycle 46 - Work cycle 90)



(Work cycle 90 - Work cycle 134)

Figure S13: Driving factor of predicted value for all work cycles for Group S1 on 11th day

References

(1) Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. **1970**,

(2) Ronen, M.; Finder, S. E.; Freifeld, O. DeepDPM: Deep Clustering With an Unknown Number of Clusters. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2022; pp 9851–9860.

(3) Von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing* **2007**, *17*, 395–416.

(4) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **1997**, *9*, 1735–1780.

(5) Sherstinsky, A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena* **2020**, *404*, 132306.

(6) Gillioz, A.; Casas, J.; Mugellini, E.; Khaled, O. A. Overview of the Transformer-based Models for NLP Tasks. Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020. 2020; pp 179–183.

(7) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.

B

Media Reports

1. **EurekAlert**

   https://www.eurekalert.org/news-releases/989950

2. **PHYS**

   https://phys.org/news/2023-05-uncovering-factors-algal-growth-south-to-north.html

3. **KeAi**

   https://www.keaipublishing.com/en/news/uncovering-the-secret-masks-behind-algae-growth-in-the-south-to-north-water-diversion-project-using-advanced-ai/

NEWS RELEASE 21-MAY-2023

Uncovering the secret masks behind algae growth in the south-to-north water diversion project using advanced AI

Peer-Reviewed Publication

KEAI COMMUNICATIONS CO., LTD.

Identifying the factors contributing to algal growth accurately and reliably is vital for sustainable use and scientific management of freshwater resources. As scientific research evolves from using small data sets to larger ones, the shortcomings of traditional machine learning become clearer, and deep learning which is adept at processing large amounts of data, is getting more attention. Although it has been used occasionally for forecasting chlorophyll-a (Chl-a) time series, deep learning has hardly been employed to identify important factors concerning algal growth.

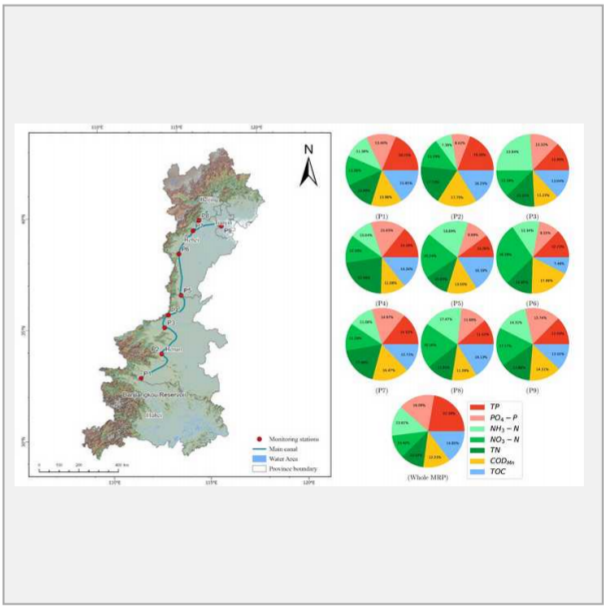


IMAGE: FIGURE 1 SKETCH MAP OF SAMPLING STATIONS DISTRIBUTION IN THE MIDDLE SECTION OF THE SOUTH-NORTH WATER DIVERSION PROJECT AND DRIVING FACTORS OF ALGAL GROWTH BASED ON BLOOMFORMER-1 MODELLING view more ›

CREDIT: JING QIAN, KARLSRUHE INSTITUTE OF TECHNOLOGY (KIT), GERMANY, AND INSTITUTE OF HYDROBIOLOGY (IHB), CHINA

To address this gap, a cross-national team of researchers from China, Germany and The Netherland developed a deep learning-based Transformer model, Bloomformer-1, designed for end-to-end identification of algal growth driving factors.

“Deep learning models have lower operational transparency compared to traditional machine learning, but they exhibit significant advantages in performance,” said Jing Qian, the first author of the paper. “The development of Bloomformer-1 aims to create a win-win situation in terms of

interpretability and performance, enabling the driving factors of algal growth to be identified transparently and accurately."

Qian, a doctoral student from the Karlsruhe Institute of Technology in Germany, conducted this research as a jointly-cultivated doctoral student at the Institute of Hydrobiology in China.

The Middle Route of the South-to-North Water Diversion Project (MRP), a national large-scale project in China, was selected as the study site to demonstrate the superior performance of Bloomformer-1. It was compared to four widely used traditional machine learning models—extra trees regression (ETR), gradient boosting regression tree (GBRT), support vector regression (SVR), and multiple linear regression (MLR)—with the highest R2 (0.80 to 0.94) and lowest RMSE (0.22 to 0.43 μg/L).

"Bloomformer-1 employs the multi-head-self-attention mechanism, which compares each token in the input sequence with other tokens to collect and learn dynamic contextual information, thus enabling a thorough understanding of all the field sampling data. This is one of the reasons for its superior performance," said co-author Stefan Norra from the University of Potsdam.

The results of study, published in the KeAi journal *Water Biology & Security*, revealed that total phosphorus (TP) was the most significant factor affecting the MRP, especially in the Henan section, while total nitrogen (TN) had the most substantial impact on algal growth in the Hebei section.

"Controlling and reducing phosphorus is an important strategy for controlling algal growth and maintaining stable MRP water quality, while nitrogen control in the Hebei region is also worth paying attention to," said Yonghong Bi from the Institute of Hydrobiology, Chinese Academy of Sciences, who is the corresponding author of the study. "Furthermore, the promotion and application of Bloomformer-1 in other water bodies will be an important task going forward."

###

Contact the author: Jing Qian, Karlsruhe Institute of Technology (KIT), Germany, and Institute of Hydrobiology (IHB), China, jing.qian@partner.kit.edu

The publisher KeAi was established by Elsevier and China Science Publishing & Media Ltd to unfold quality research globally. In 2013, our focus shifted to open access publishing. We now proudly publish more than 100 world-class, open access, English language journals, spanning all scientific disciplines. Many of these are titles we publish in partnership with prestigious societies and academic institutions, such as the National Natural Science Foundation of China (NSFC).

JOURNAL

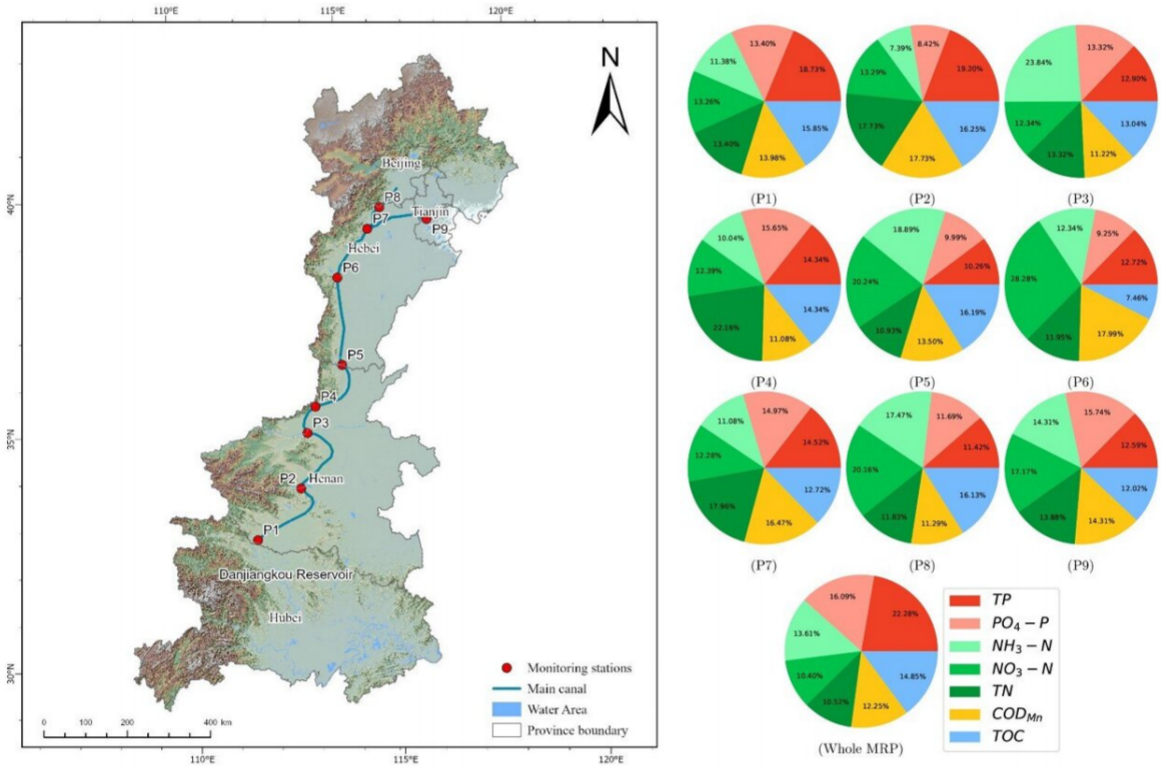
Water Biology and Security

DOI

10.1016/j.watbs.2023.100184

Uncovering the driving factors behind algal growth in the South-to-North Water Diversion Project using advanced AI

May 22 2023



Sketch map of sampling stations distribution in the middle section of the South-North Water Diversion Project and driving factors of algal growth based on Bloomformer-1 modeling. Credit: Jing Qian, Karlsruhe Institute of Technology (KIT), Germany, and Institute of Hydrobiology (IHB), China

Identifying the factors contributing to algal growth accurately and reliably is vital for sustainable use and scientific management of freshwater resources. As scientific research evolves from using small data sets to larger ones, the shortcomings of traditional machine learning become clearer, and deep learning which is adept at processing large amounts of data, is getting more attention.

Although it has been used occasionally for forecasting chlorophyll-a (Chl-a) time series, deep learning has rarely been employed to identify important factors concerning algal growth.

To address this gap, a cross-national team of researchers from China, Germany and The Netherlands developed a deep learning-based Transformer model, Bloomformer-1, designed for end-to-end identification of algal growth driving factors.

"Deep learning models have lower operational transparency compared to traditional machine learning, but they exhibit significant advantages in performance," said Jing Qian, the first author of the paper. "The development of Bloomformer-1 aims to create a win-win situation in terms of interpretability and performance, enabling the driving factors of algal growth to be identified transparently and accurately."

Qian, a doctoral student from the Karlsruhe Institute of Technology in Germany, conducted this research as a jointly-cultivated doctoral student at the Institute of Hydrobiology in China.

The Middle Route of the South-to-North Water Diversion Project (MRP), a national large-scale project in China, was selected as the study site to demonstrate the superior performance of Bloomformer-1. It was compared to four widely used traditional machine learning models—extra trees regression (ETR), gradient boosting regression tree (GBRT), support vector regression (SVR), and multiple linear regression

(MLR)—with the highest R2 (0.80 to 0.94) and lowest RMSE (0.22 to 0.43 μg/L).

"Bloomformer-1 employs the multi-head-self-attention mechanism, which compares each token in the input sequence with other tokens to collect and learn dynamic contextual information, thus enabling a thorough understanding of all the field sampling data. This is one of the reasons for its superior performance," said co-author Stefan Norra from the University of Potsdam.

The results of study, published in *Water Biology & Security*, revealed that total phosphorus (TP) was the most significant factor affecting the MRP, especially in the Henan section, while total nitrogen (TN) had the most substantial impact on algal growth in the Hebei section.

"Controlling and reducing phosphorus is an important strategy for controlling algal growth and maintaining stable MRP water quality, while nitrogen control in the Hebei region is also worth paying attention to," said Yonghong Bi from the Institute of Hydrobiology, Chinese Academy of Sciences, who is the corresponding author of the study. "Furthermore, the promotion and application of Bloomformer-1 in other water bodies will be an important task going forward."

More information: Jing Qian et al, Identification of driving factors of algal growth in the South-to-North Water Diversion Project by Transformer-based deep learning, *Water Biology and Security* (2023). DOI: 10.1016/j.watbs.2023.100184

Provided by KeAi Communications Co.

Uncovering the Secret Masks Behind Algae Growth in the South-to-North Water Diversion Project Using Advanced AI

Published 24 May, 2023

Identifying the factors contributing to algal growth accurately and reliably is vital for sustainable use and scientific management of freshwater resources. As scientific research evolves from using small data sets to larger ones, the shortcomings of traditional machine learning become clearer, and deep learning which is adept at processing large amounts of data, is getting more attention. Although it has been used occasionally for forecasting chlorophyll-a (Chl-a) time series, deep learning has hardly been employed to identify important factors concerning algal growth.

To address this gap, a cross-national team of researchers from China, Germany and The Netherland developed a deep learning-based Transformer model, Bloomformer-1, designed for end-to-end identification of algal growth driving factors.

“Deep learning models have lower operational transparency compared to traditional machine learning, but they exhibit significant advantages in performance,” said Jing Qian, the first author of the paper. “The development of Bloomformer-1 aims to create a win-win situation in terms of interpretability and performance, enabling the driving factors of algal growth to be identified transparently and accurately.”

Qian, a doctoral student from the Karlsruhe Institute of Technology in Germany, conducted this research as a jointly-cultivated doctoral student at the Institute of Hydrobiology in China.

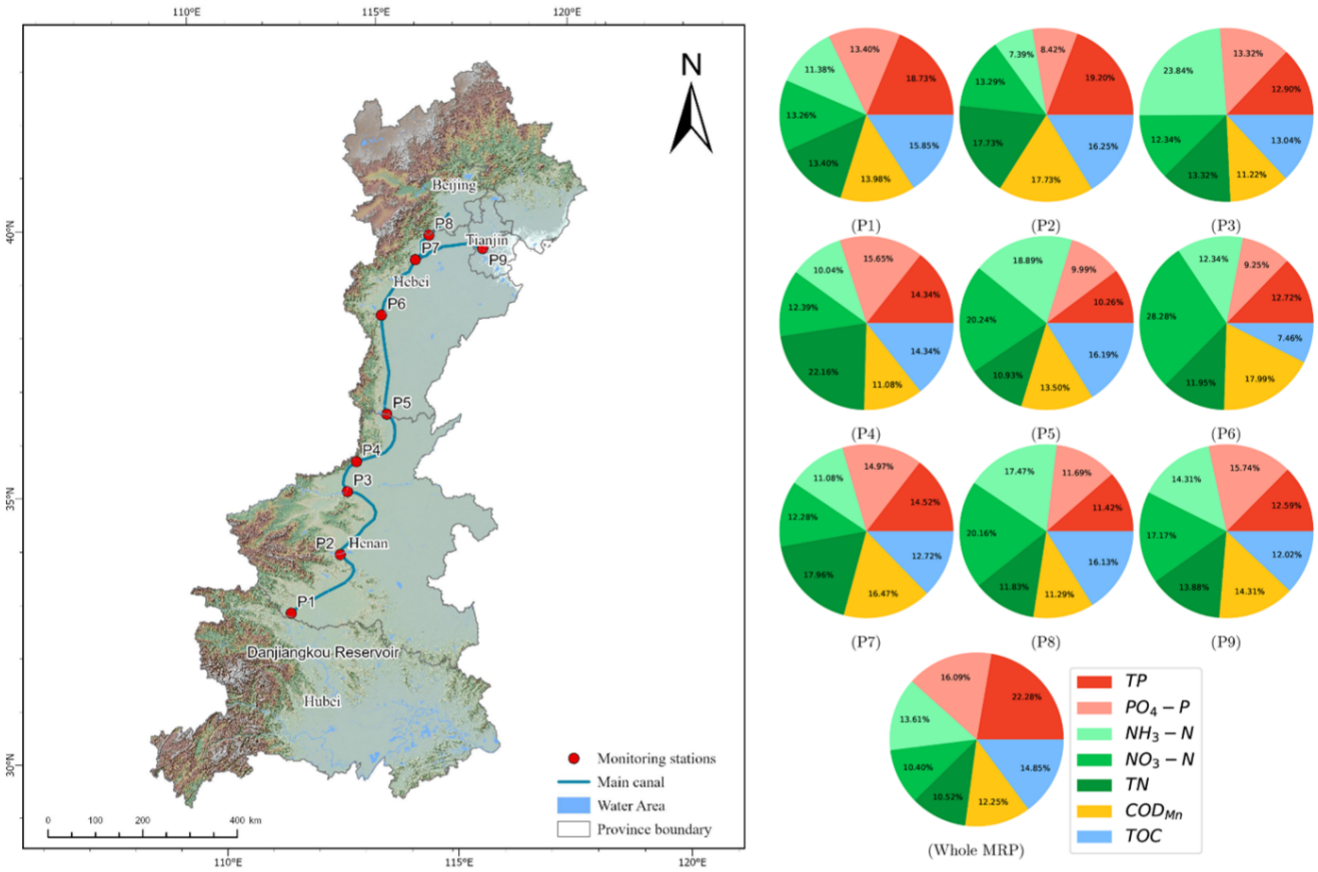
The Middle Route of the South-to-North Water Diversion Project (MRP), a national large-scale project in China, was selected as the study site to demonstrate the superior performance of Bloomformer-1. It was compared to four widely used traditional machine learning models—extra trees regression (ETR), gradient boosting regression tree (GBRT), support vector regression (SVR), and multiple linear regression (MLR)—with the highest R2 (0.80 to 0.94) and lowest RMSE (0.22 to 0.43 μg/L).

"Bloomformer-1 employs the multi-head-self-attention mechanism, which compares each token in the input sequence with other tokens to collect and learn dynamic contextual information, thus enabling a thorough understanding of all the field sampling data. This is one of the reasons for its superior performance," said co-author Stefan Norra from the University of Potsdam.

The results of study, published in the KeAi journal *Water Biology & Security*, revealed that total phosphorus (TP) was the most significant factor affecting the MRP, especially in the Henan section, while total nitrogen (TN) had the most substantial impact on algal growth in the Hebei section.

“Controlling and reducing phosphorus is an important strategy for controlling algal growth and maintaining stable MRP water quality, while nitrogen control in the Hebei region is also worth paying attention to," said Yonghong Bi from the Institute of Hydrobiology, Chinese Academy of Sciences, who is the corresponding author of the study. "Furthermore, the promotion and application of Bloomformer-1 in other water bodies will be an important task going forward.”

driving factors of algal growth based on Bloomformer-1 modelling



CREDIT: Jing Qian, Karlsruhe Institute of Technology (KIT), Germany, and Institute of Hydrobiology (IHB), China

Contact author name, affiliation, email address: Jing Qian, Karlsruhe Institute of Technology (KIT), Germany, and Institute of Hydrobiology (IHB), China, jing.qian@partner.kit.edu

Social media handles:

- Prof. Dr. Yonghong Bi: https://people.ucas.edu.cn/~biyonghong?language=en
- Prof. Dr. Stefan Norra: https://www.uni-potsdam.de/de/umwelt/forschung/ag-bodenkunde-und-geooekologie
- Jing Qian: https://egg.agw.kit.edu/66.php

See the article: Qian, Jing, et al. "Identification of driving factors of algal growth in the South-to-North Water Diversion Project by Transformer-based deep learning." Water Biology and Security (2023): 100184.

C Code

C.1 DNN

```
import torch
import torch.nn as nn
import torch.optim as optim
from torch.utils.data import DataLoader, TensorDataset
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error, r2_score,
    median_absolute_error

# Set seed for reproducibility
torch.manual_seed(42)
if torch.cuda.is_available():
    torch.cuda.manual_seed_all(42)

# Define model structure
class MLP(nn.Module):
    def __init__(self, input_size, hidden_layer_sizes):
        super(MLP, self).__init__()
        layers = []
        layer_sizes = [input_size] + list(hidden_layer_sizes)
        for i in range(len(layer_sizes) - 1):
            layers.append(nn.Linear(layer_sizes[i], layer_sizes[i+1]))
            layers.append(nn.ReLU())
        layers.append(nn.Linear(layer_sizes[-1], 1))
        self.model = nn.Sequential(*layers)

    def forward(self, x):
        return self.model(x)

# Load data
X_train, y_train, X_test, y_test = load_data()
```

```
31
32 # Initialize model
33 hidden_layer_sizes = (256, 256, 256, 256, 256)
34 model = MLP(X_train.shape[1], hidden_layer_sizes)
35
36 # Define a loss function and optimizer
37 criterion = nn.MSELoss()
38 optimizer = optim.Adam(model.parameters(), lr=0.001)
39
40 # Create dataloader for batch processing
41 dataset = TensorDataset(X_train, y_train)
42 dataloader = DataLoader(dataset, batch_size=200, shuffle=True)
43
44 # Train the model
45 for epoch in range(1000):
46     for X_batch, y_batch in dataloader:
47         model.train()
48         optimizer.zero_grad()
49         y_pred = model(X_batch)
50         loss = criterion(y_pred, y_batch)
51         loss.backward()
52         optimizer.step()
53
54 # After training, we switch to evaluation mode for testing
55 model.eval()
56 with torch.no_grad():
57     y_test_pred = model(X_test)
58
59 # Convert prediction tensor back to numpy array for metrics computation
60 y_test_pred = y_test_pred.numpy()
61 y_test = y_test.numpy()
62
63 test_mse = mean_squared_error(y_test, y_test_pred)
64 test_rmse = np.sqrt(test_mse)
65 test_r2 = r2_score(y_test, y_test_pred)
66 test_mape = np.mean(np.abs((y_test - y_test_pred) / y_test))
67 test_mad = median_absolute_error(y_test, y_test_pred)
```

C.2 Spectral clustering

```python
import numpy as np
from scipy.cluster.vq import kmeans, vq
from scipy.linalg import svd

def spectral_clustering(adj_matrix, num_clusters):
    """
    Perform spectral clustering from adjacency matrix.

    Parameters:
    adj_matrix: numpy.ndarray
        adjacency matrix
    num_clusters: int
        number of clusters

    Returns:
    cluster_labels: numpy.ndarray
        an array of cluster assignments for each data point
    """

    # Compute row sums (degrees for each node)
    rowsum = np.sum(abs(adj_matrix), axis=0)

    # Compute the degree matrix (diagonal matrix with degrees)
    degree_matrix = np.diag(1 / np.sqrt(rowsum + 1e-6))

    # Compute the Laplacian matrix using symmetric normalization
    laplacian_matrix = degree_matrix.dot(adj_matrix).dot(degree_matrix)

    # Compute the eigenvectors of the Laplacian matrix using Singular Value
    Decomposition (SVD)
    _, _, eig_vectors = svd(laplacian_matrix, full_matrices=False)

    # Stack the first num_clusters eigenvectors to create feature vectors
    features = np.array(eig_vectors[:num_clusters]).T

    # Perform k-means on these features
    centroids, _ = kmeans(features, num_clusters)
    cluster_labels, _ = vq(features, centroids)

    return cluster_labels
```