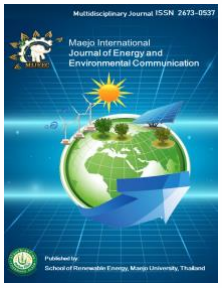
Maejo International Journal of Energy and Environmental Communication

Journal homepage: https://ph02.tci-thaijo.org/index.php/MIJEEC



ARTICLE

Early-warning modeling of water blooms using machine learning algorithm

Tianxiao Liu[1]*, Yuan Tian[2], Zhongfang Lei[2], Zhenya Zhang[2], Motoo Utsumi[2], Kunihiro Okano[3], Kazuya Shimizu[4]

[1] Graduate School of Science and Technology, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8572, Japan
[2] Institute of Life and Environmental Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8572, Japan
[3] Graduate School of Bioresource Science Akita Prefectural University, 241-438 Kaidobata-Nishi, Nakano Shimoshinjo, Akita 010-0195, Japan
[4] Faculty of Life Sciences, Toyo University, 1-1-1 Izumino Itakura Oura-gun, Gunma 374-0193, Japan

ARTICLEINFO

Article history:
Received 28 May 2024
Received in revised form
23 June 2024
Accepted 26 June 2024

Keywords:
Machine learning
Water bloom prediction
Random forest
Model generalizability

ABSTRACT

Many environmental studies have implemented machine learning methods, including water bloom prediction. However, the need for robustness and generalization of models across different environmental conditions and geographical locations is urgent. This study implements. The Random Forest algorithm to develops a predictive model for water bloom occurrences. Using a classification model, it is a novel approach for predicting water bloom occurrences across various freshwater bodies, from reservoirs to lakes. The model demonstrates an overall accuracy of 0.74 for predicting the water blooms from multiple freshwater sources, including Inba Swamp, Takizawa Dam, Shourenji Dam, and Kasumigaura, using surveillance data from Japan Water Agency websites. The promising accuracy and intentionally introduced one-month lag of water bloom occurrence in the constructed database shows its effectiveness in anticipating water blooms. The Out-Of-Bag error rate suggests the optimal number of trees to grow in the model is 500 for space and time efficiency. Cross-validation across multiple sampling sites reveals nuanced prediction accuracies, emphasizing the importance of considering spatial variability. Predictor importance analysis identifies chlorophyll a, total nitrogen, total phosphorus, and temperature as crucial factors, with the importance of 0.18, 0.04, 0.02, and 0.02, respectively. It does not only contribute to the accuracy of the prediction model but also provides insights into the underlying dynamics of water bloom events, enhancing the understanding of water bloom dynamics and informing proactive environmental management strategies.

1. Introduction

Water blooms, characterized by the rapid proliferation of cyanobacteria in freshwater bodies, have emerged as a pressing

* Corresponding author.
E-mail address: liu.tianxiao127@gmail.com (Liu T.)

environmental concern due to their negative impacts on water quality and ecosystems (Piontek et al., 2023). In addition to threatening aquatic life, water blooms have cascading effects on human health, recreational activities, and the usability of water resources (Vidal et al., 2017; Lad et al., 2022). Cyanotoxin-producing cyanobacteria can contaminate drinking water supplies and cause adverse health effects when consumed by humans or animals (Drobac et al., 2013; Du et al., 2024). Exposure to cyanotoxins has been linked to liver damage, gastrointestinal problems, and even neurological effects, highlighting the urgent need to protect water resources (Mittelman et al., 2016; Sandhu et al., 2024). As these events become more common, proactive and accurate predictive models become imperative for timely intervention and sustainable water management practices (Burford et al., 2020).

In recent years, the intersection of environmental studies and machine learning has gained significant traction in addressing complex ecological challenges. Machine learning techniques, driven by advances in computational power and the availability of large datasets, have demonstrated the potential to unravel intricate patterns and relationships within environmental data (Anandhi & Iyapparaja, 2024; Nath et al., 2024; Rather et al., 2024; Walsh et al., 2024). The application of machine learning to predict water quality phenomena, such as water blooms, holds promise for improving our understanding of the dynamics underlying these events, and numerous machine learning models have been used in previous studies (Marrone et al., 2023).

Given the diverse machine learning models available, it is crucial to consider the strengths and weaknesses of each approach when selecting the appropriate model for a specific application. For example, Bayesian frameworks offer interpretability and uncertainty estimation but can struggle with complex datasets and require significant computational resources (Marwala et al., 2023). Gradient boost regressors (GBR) and long short-term memory (LSTM) networks excel at capturing temporal patterns and handling sequential data but may suffer from overfitting and sensitivity to hyperparameters (Brophy & Lowd, 2022; Varsha et al., 2023). Convolutional neural networks (CNNs) are adept at extracting spatial features from images but may require large datasets and extensive preprocessing. Clustering algorithms such as K-means provide insight into data grouping but may struggle with nonlinear relationships. Neural Networks, XGBoost, and Logistic Regression, provide flexibility and scalability but may require extensive feature engineering and tuning (Bonaccorso, 2018). Random Forest is a widely used machine learning algorithm known for its versatility in handling classification and regression problems. This algorithm works by aggregating the outputs of multiple decision trees to derive a single result, with robustness to the issue of input variable noise, overfitting, and a significant improvement in prediction accuracy. The simplicity and flexibility of Random Forest have contributed to their widespread adoption. Random Forest provides robustness against overfitting and handles high dimensional data well but may lack interpretability (Genuer et al., 2010; Liaw & Wiener, 2018).

These algorithms and models have been implemented in previous studies of water bloom parameter predictions. Several challenges remain considering the implementation and characteristics of each algorithm and model. One notable challenge is the need for robustness and generalization of models across different environmental conditions and geographic locations. In addition, issues related to data quality, feature selection, model interpretability, and scalability remain. Addressing these challenges will improve the effectiveness and applicability of machine learning-based early warning systems for water quality phenomena.

Leveraging the capabilities of machine learning algorithms, notably the widely used Random Forest algorithm, this study seeks to provide insights into the factors that influence the occurrence of water blooms. The capability of Random Forest to manage diverse datasets and aggregate the outputs of multiple decision trees makes it particularly suitable for our objective of predicting water quality phenomena across varying environmental conditions. Moreover, the flexibility in parameter tuning of Random Forest allows for optimization in terms of time and space complexity, enhancing the model's efficiency and adaptability to different scenarios. This study aims to contribute to the growing body of research at the intersection of machine learning and environmental studies by presenting a predictive model for water blooms in freshwater bodies with optimization strategies.

2. Material and methods

2.1 Study Design

Using the Random Forest algorithm, this study aimed to predict water bloom occurrences in freshwater lakes and reservoirs. Monthly water quality data were collected from multiple freshwater reservoirs and lakes in Japan, covering 2011 to 2022. The data included various environmental parameters and were used to build a classification model to identify the presence of water blooms in subsequent months. The study integrated automated data imports and manual entries to ensure comprehensive data coverage. The model's performance was evaluated using metrics such as the Area Under the Receiver Operating Characteristic Curve, feature importance, and Mean Decrease in Accuracy. Additionally, the model was optimized for space and time efficiency, ensuring compatibility with future larger datasets and enhancing its scalability and robustness for broader applications.

2.2 Data Source and Collection

The data for this study were sourced from the Annual Water Quality Reports publicly published by the Japan Water Agency. These reports spanned around the years 2011 to 2022. The focus of the study involved monthly water quality sampling in freshwater reservoirs and lakes, specifically Inba Swamp, Takizawa Dam, Shourenji Dam, and Kasumigaura. Each water body comprised 3 to 4 sampling sites, with consistent monthly analysis of the same set of parameters. The study sites and period were picked based on water bloom occurrence detection, public surveillance data availability, water body type diversity, and geographical location diversity.

The constructed database, derived from these water surveillance reports, includes categorical variables such as sampling location and sampling date. While sampling location is a distinguishing identifier for each data entry, sampling date, though categorical, possesses a meaningful chronological order. Although the latter is essential for specific machine learning simulations, it remains an identifying feature rather than a training feature in this study's context of the Random Forest algorithm. The data structure of the database is organized to

facilitate the classification task of identifying water bloom occurrences. Specifically, for prediction, the outcome of the water bloom is encoded as whether it occurred in the next month of a particular data entry rather than the actual month of the sampling. The database was constructed, cleaned, adjusted, and integrated from the automatic import of datasets from the Excel file reports and the manual entering of statistics from other incompatible file formats.

2.3 Forecasting Model and Statistical Methods

This study's classification task involves determining whether water bloom is present. The algorithm examines the training features to understand the relationship between these features and the target outcome. Decision trees are constructed for subsets of data entries, evaluating the conditions of the features to predict the outcome. For each node in a decision tree, the split is determined by calculating the distribution of class labels (whether there is a water bloom) among the data points associated with that node that outputs the minimum Gini index. The formula is as follows:

$$Gini(t) = 1 - \sum_{i=0}^{1} (p_i)^2$$

Where t is the node, pi is the proportion of data points in class i at node t (specifically, p0 = no bloom, p1 = bloom)

After the number of trees constructed has reached a predetermined value, the result is determined through a majority vote, given its classification nature. A detailed algorithm flowchart can be referenced in Figure 1. The ensemble nature of this algorithm, combining predictions from multiple decision trees, is the reason for its name, Random Forests. Algorithm construction, data management, and analysis are performed in RStudio version 2023.06.2+561 (Posit Software, PBC, 2023) based on R 4.3.1 (R Core Team, 2023).

The Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve was employed as a critical evaluation metric to assess the model's predictive performance in this study. The ROC curve represents a model's ability to discriminate between classes, particularly in binary classification tasks. It plots the actual positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The AUC comprehensively measures the model's discriminative ability, representing the area under the ROC curve. A higher AUC value model indicates better performance in distinguishing between positive and negative instances. In the context of this study, where the classification task involves identifying water bloom occurrences, the AUC of the ROC curve serves as a robust indicator of the model's predictive accuracy and ability to balance sensitivity and specificity.

The predictive ability of predictors in Random Forests is measured by importance and mean decrease in accuracy (MDA). MDA was calculated as the mean accuracy decreases when the predictors are excluded across all trees in one Random Forest run. Should a value be harmful, the model's performance increases after excluding a particular predictor, indicating that this predictor does no better than a random guess in classifying water bloom occurrence. Thus, it is recommended to be eliminated from the set of predictors for the model.

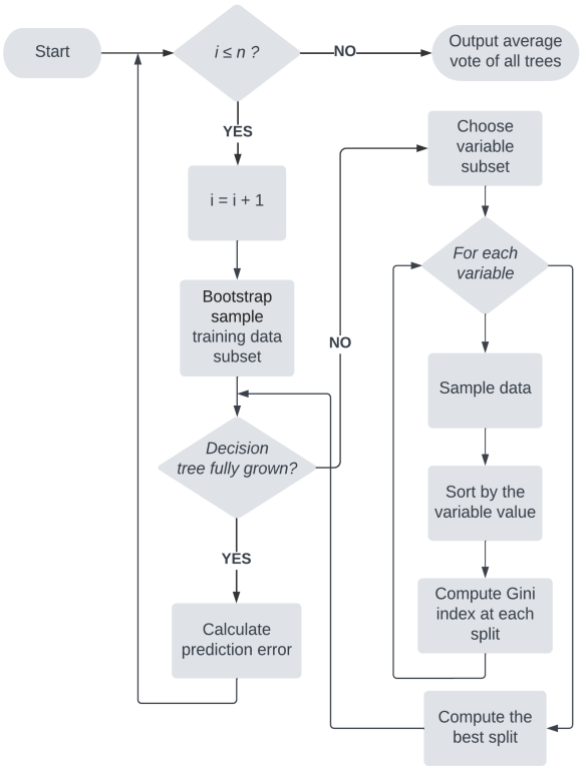


Figure 1. Random forest algorithm flowchart. n is the number of trees set in a particular run. The prediction error is calculated from the Out-Of-Bag samples.

The Random Forest method also implements an Out-Of-Bag (OOB) error rate that measures the prediction error. It is the average error for a training sample that does not contain the OOB sample in the bootstrap sample during model building. This measure helps determine the optimal number of trees to grow in a Random Forest run, where the error rate converges and stabilizes. The averaged MDA and error rate of 50 Random Forest runs were used to determine the significant predictors and the optimal number of trees to grow for each model to achieve prediction accuracy and optimization of the maximum efficiency of time and space complexity.

2.4 Ethical Considerations

This study utilized publicly available surveillance environmental data from the Annual Water Quality Reports published by the Japan Water Agency. The data did not involve any personal or sensitive information. All research activities adhered to ethical standards for environmental research, ensuring transparency and integrity in data handling and analysis. No specific ethical approval was required as the study relied solely on secondary data from public sources.

3. Results and Discussion

3.1 Model Performance and Optimization

The application of the Random Forest algorithm yielded an overall prediction accuracy with an Area Under Curve (AUC) of 0.74 (Figure 2), indicating the model's potential effectiveness in anticipating water bloom occurrences and associated water quality concerns. To delve deeper into the model's performance, we examined the accuracy of training and test data across four sampling sites and the combined sets of freshwater bodies with recorded water blooms (Table 1). The analysis revealed variations in prediction accuracy across different water bodies, stabilizing at approximately 0.69 to 0.89 after cross-validation, utilizing one sampling site as the training set and the remaining sites as the testing set allowed for a comprehensive assessment of the model's robustness and non-overfitting. The results indicate that while accuracy may differ across water bodies, the Random Forest model demonstrates stability in predicting water bloom occurrences after appropriate cross-validation. However, it is crucial to acknowledge the challenges associated with developing comprehensive machine-learning models for water bloom prediction. The complex nature of environmental, chemical, and biological factors, alongside geographical diversity, presents significant complications. Moreover, the heterogeneity of data across different sites and the relatively small datasets poses additional challenges (Marrone et al., 2023).

Table 1. Accuracy of training and test data for 4 sampling sites and combined sets of freshwater bodies with recorded water blooms applied to random forest prediction model.

Prediction set	Prediction Accuracy	
	Training	Testing
Overall	0.74	0.74
Inba Swamp	0.85	0.79
Takizawa Dam	0.75	0.73
Shourenji Dam	0.89	0.83
Kasumigaura	0.74	0.69

*Cross-validation is used and for each site.

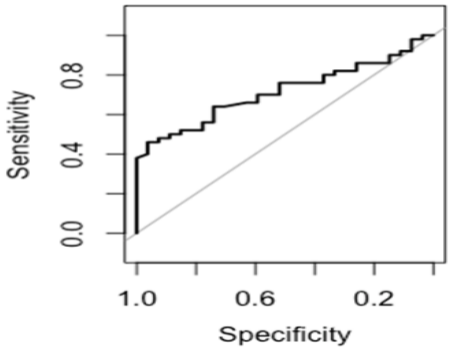


Figure 2. Overall result of prediction accuracy using Random Forest algorithm. Area Under Curve = 0.74

The OOB error rate was calculated and plotted against the number of trees grown in the Random Forest to optimize the model's time and space complexity performance. The error rate exhibited initial fluctuations but stabilized around 500 trees, remaining constant to 3000 trees in this study (Figure 3). Consequently, the presented results are based on a Random Forest run with 500 trees.

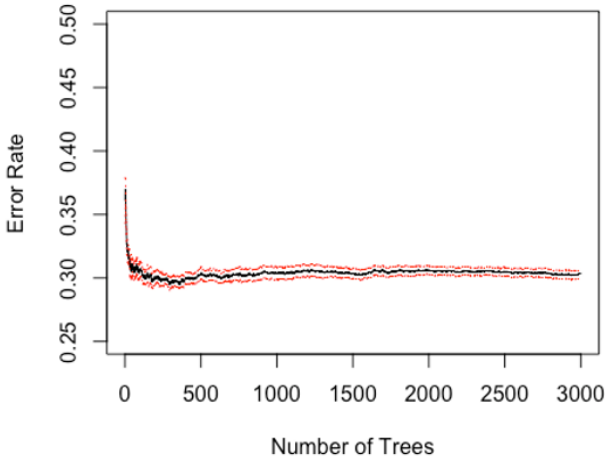


Figure 3. Mean (black line) and 95% confidence interval (red line) of Out-Of-Bag error rate in water bloom prediction using overall data from 50 runs of random forests

Further insights into the factors influencing water bloom events were obtained by examining predictor importance and MDA (Table 2). In pursuit of an accurate predictive model for water blooms, we acknowledge, and address concerns related to overfitting and collinearity within the complex freshwater environment. The interdependence of variables such as chlorophyll a, total nitrogen, total phosphorus, and temperature introduces the risk of collinearity, where high correlations among predictors can distort the model's estimates and reduce its interpretability. Additionally, the intricate dynamics of the water ecosystem pose challenges, making the model susceptible to overfitting. To counteract this, regularization techniques, such as cross-validation and shrinkage methods, are employed to enhance the model's generalizability. Furthermore, parameters with negative MDA values, which indicate they have minimal impact on the model's accuracy, could be omitted in the generated model to enhance the reliability and robustness. By carefully addressing these concerns and applying appropriate techniques, the predictive model can better capture the underlying relationships between predictor variables and water bloom occurrences, improving its utility for water resource management and mitigation efforts.

Table 2. Importance and mean decrease in accuracy (MDA) of predictors from the Random Forest model run in MDA descending order.

Variables	Importance	MDA (descending order)
Chlorophyll a (mg/m^3)	0.1763	0.2037
T-N (mg/l)	0.0371	0.0815
T-P (mg/l)	0.0224	0.0725
Temperature (°C)	0.0221	0.0331
BOD (mg/l)	0.0021	0.0038
COD (mg/l)	0.0015	0.0020

Coliform count (MPN/100ml)	0.0014	-0.0005
SS (mg/l)	0.0010	-0.0017
DO (mg/l)	0.0009	-0.0019

Permutation tests involve randomly shuffling the values of the response variable while keeping the predictor variables unchanged, thereby creating a distribution of scores under the null hypothesis. For each predictor variable, the observed importance and MDA scores were compared with the distribution of permuted scores obtained through 1000 permutations. Significance was determined based on the proportion of permuted scores greater than or equal to (for importance) or less than or equal to (for MDA) the observed score, yielding p-values.

This analysis revealed that chlorophyll a, total nitrogen, total phosphorus, and temperature exhibited highly significant importance scores ($p < 0.001$) and significant MDA scores ($p < 0.001$), providing robust evidence of their crucial roles in water bloom prediction. These results underscore the integral relationship between water quality parameters and bloom dynamics. Specifically, the significant scores highlight the importance of chlorophyll a, total nitrogen, total phosphorus, and temperature as indicators of nutrient enrichment and environmental conditions conducive to cyanobacterial growth and bloom formation. The observed high MDA scores further support this notion by indicating that changes in these variables substantially impact the model's predictive accuracy.

3.2 Model Interpretation and Supporting Results

The application of machine learning methods to predict water bloom and related parameters has been extensively studied in previous research. A recent study developed a Bayesian modeling framework with intracellular and extracellular microcystins that characterizes the relationships between various environmental and biological factors with monitoring data from three bloom-affected lakes (Shan et al., 2020). Another study applied two machine learning models, GBR and LSTM network, to predict algal blooms and seasonal changes in algal chlorophyll concentrations in a mesotrophic lake (Lin et al., 2023). Previous research by Pyo et al. (2020; 2021) specifically targeted cyanobacterial density using CNN, which predicts cyanobacteria such as Microcystis biomass. A separate investigation applied K-means and Random Forest, targeting the bloom-forming *Planktothrix rubescens* with a combined clustering and subsequent regression approach to predict harmful algal blooms (Derot et al., 2020). There is also a study that developed three types of models - Neural Network, XGBoost, and Logistic Regression – using data from water samples collected in a freshwater lake through corrected over several years to predict the occurrence of water blooms in advance (Villanueva et al., 2023). These models primarily focused on single water bodies or used regression models, limiting their generalizability across different geographic locations and interpretability regarding the outcome of interest. A summary of the studies as mentioned earlier compared to this study could be referenced in Table 3.

Table 3. Comparison of this study with previous studies that use machine learning methods in water quality prediction

Task	Target variable	Algorithms	Waterbody	References
Classification	Water bloom	RF	Multiple lakes and reservoirs	This study
Clustering & regression	*Planktothrix rubescens*	K-means and RF	Single, surface water	Derot et al., 2020
Clustering & regression	Intracellular and extracellular MCs	Bayesian modeling	Multiple bloom-plagued lakes	Shan et al., 2020
Regression	Chlorophyll concentrations	GBR and LSTM	Single, mesotrophic lake	Lin et al., 2023
Regression	Chlorophyll *a* concentration	RF	Single, reclaimed lake	Wang et al., 2023
Classification	Water bloom	Neural Network and XGBoost	Single, freshwater lake	Villanueva et al., 2023

In this study, we analyzed water bloom occurrences across multiple freshwater bodies, each with distinct environmental and geographical characteristics. The comparative analysis revealed that different water bodies exhibit varying influence factors contributing to water bloom events, resulting in different prediction accuracy (Table 1). Understanding these differences is essential for developing targeted preventive measures. For example, for nutrient-sensitive water bodies, management strategies should prioritize reducing nutrient inputs through better agricultural practices and wastewater treatment; for water bodies where temperature is the significant factor, monitoring and regulating thermal pollution could be more effective. Implementing customized management plans based on the dominant influence factors for each water body will enhance the effectiveness of water bloom prevention and control efforts, ultimately contributing to more sustainable water quality management.

To contextualize our findings further, we look specifically into parameters that have significant predictability for water bloom occurrences. As the results of our model show, chlorophyll emerged as the most crucial predictor, accompanied by total nitrogen, total phosphorus, and temperature, which also played substantial roles. The intricate relationship between these variables indicates the complex dynamics influencing water bloom events. In particular, the association between chlorophyll and nutrient levels may be influenced by the dynamics of reclaimed water quality reports. This could suggest a relative lag in phytoplankton growth, impacted by fluctuations in nutrient levels. The time lag of the predicted value and actual value for cyanobacterial density can

be observed, especially for Microcystis and Phormidium (Wei et al., 2001). Thus, the intentional introduction of a one-month time lag between water quality parameters and water bloom occurrences in the database of this study may adhere to the observed complexities. Previous studies, such as the work by Wang et al. (2023), have highlighted the intricate interplay between nutrient fluctuations and phytoplankton growth. The observed variability in water samples, including those with faster nitrate decrease, underscores the complexity of nutrient dynamics and their influence on algal growth. This underscores the need for a deeper understanding of temporal relationships and the potential lag effects in predictive modeling for water bloom events.

Some other previous research findings provide valuable context and enhance the understanding of the factors influencing water bloom events and the growth of cyanobacteria and cyanotoxins. Schaeffer et al. (2024) highlight the significance of water surface temperature, precipitation, and lake geomorphology in predicting chlorophyll with cyanobacteria dominance using satellite remote sensing technologies, among which water temperature is recognized as the most significant factor. The findings align closely with the results of this study, which acknowledges temperature as an essential variable across diverse freshwater lakes with ahead-of-time predictivity. Moreover, Kim et al. (2022) identified water temperature and total nitrogen as key hydro-environmental predictors for cyanobacterial blooms based on cyanobacteria cell count, illustrating the significance of water temperature and total nitrogen in driving water bloom occurrences. The quantitative analysis conducted by Kim et al. (2022) offers insights into the relationship between hydro-environmental factors and cyanobacteria biomass, complementing this study's focus on predictive modeling. Similarly, Wang et al. (2022) emphasize the importance of dissolved inorganic nitrogen concerning the toxicity of Microcystis, as well as the role of total phosphorus in predicting the biomass of total Microcystis and toxic Microcystis aeruginosa. The findings underscore the complex interplay between nutrient availability and water bloom dynamics, further supporting the relevance of the identified predictors of this study, namely chlorophyll a, total nitrogen, and total phosphorus, in natural environmental settings.

Vézie et al. (2002) find that high levels of nitrogen and phosphorus in freshwater significantly impact the growth of toxic Microcystis strains over nontoxic ones, potentially contributing to the onset of harmful water blooms during eutrophication. Intriguingly, the study also reveals a paradoxical observation: while intracellular MC concentrations peak under high nitrogen and phosphorus levels, they also exhibit elevated levels under minimal combined nutrient concentrations. This unexpected finding prompts speculation regarding the underlying mechanisms driving intracellular MC production and accumulation in Microcystis populations, hinting at complex regulatory pathways influenced by nutrient availability. Further research endeavors to unravel the intricate interplay between nutrient dynamics and toxin production in water blooms hold promise for advancing our understanding of bloom formation mechanisms.

Similarly, Sivonen (1990) observes that strains of Oscillatoria, a common hepatotoxin-producing cyanobacteria genus in freshwater, exhibited higher cyanotoxin production with increased nitrogen levels. Additionally, toxin production was influenced by phosphorus concentration at lower levels. Temperature also impacted both bacteria growth and cyanotoxin production. Rapala et al. (1997) discover a significant correlation between nitrogen levels and dissolved cyanotoxin concentrations, including MCs and cyanobacterial hepatotoxins, emphasizing nitrogen's role in cyanotoxin accumulation. It suggests that reducing phosphorus loads in water bodies could help mitigate toxic water blooms by limiting cyanobacterial growth and cyanotoxin production.

Other studies also demonstrated the significant roles of nitrogen, phosphorus, and temperature in recognizing cyanobacterial growth and cyanotoxin production (Downing et al., 2005; Yoshida et al., 2007; Davis et al., 2009) in laboratory settings. Together, these studies provide a comprehensive understanding of the multifaceted nature of the dynamics of cyanobacteria and cyanotoxin in water bloom and reinforce the importance of integrating multiple predictors to enhance predictive models for water resource management and bloom mitigation strategies.

3.3 Limitations and Future Steps

Acknowledging the inherent limitations in the research methodology is crucial for interpreting the findings accurately. First, it is important to recognize that in some predictive models, chlorophyll concentration is treated as an outcome rather than a parameter due to its intuitive correlation with water bloom occurrences as well (Lin et al., 2023). Thus, it might have more implications if future research focuses on other environmental and chemical parameters as predictors.

Constrained by the data collection methods, study sites are limited to specific regions and freshwater bodies in Japan. The model's generalizability would be further improved if study sites could be expanded to include diverse geographic locations, such as in other countries. Meanwhile, the collected annual water quality reports only have monthly data, introduced one-month lag may only capture some dynamics of water blooms. Depending on data availability, it would introduce more accurate and detailed implications if different temporal lags of water quality data and their impacts could be explored in future studies.

This study's predictive analysis relies on retrospective surveillance data from multiple water resources, a methodology prone to biases and constraints (Ramirez-Santana, 2018). Recently, remote sensing and satellite imagery have made real-time and large-scale surveillance data capture feasible (Schaeffer et al., 2024). However, real-time forecasting of water bloom still requires extensive computer resources to be developed and achieved, suggesting a potential venue for future research to explore. Compared to surveillance data analysis on chemical data, real-time monitoring through remote sensing and satellite imagery on physical data has also shown promising potential for water bloom observation and prediction. However, it should be noted that platforms such as Landsat and Sentinel satellites have significant costs and temporal constraints (Radeloff et al., 2024).

Additionally, the classification criteria must clarify the variability in defining water bloom occurrence thresholds across different geographic regions (Welker et al., 2021). An example of innovative monitoring approaches is the implementation of Alert

Levels by CARU (Comisión Administradora del Río Uruguay), recommended by WHO, for recreational waterbody use, based on observed color variations in the water obtained from Sentinel 2 satellite imagery. These variations correlate with cyanobacteria cell density and cyanotoxin concentrations, identifying average densities of 200 cells/mL and mapping of CARU's Alert Levels (CARU, 2016; 2017), and have been used in some prediction models of previous studies regarding water blooms. As implemented in Schaeffer et al. (2024), Alert Level 1 is used as the threshold for water bloom identification. However, it is essential to recognize that sometimes, the identification of water bloom occurrences could be addressed by artificial monitoring methods, as demonstrated by the data presented in this study. Thus, it is essential for stakeholders to set standardized criteria for water bloom evaluation if a more robust and generalizable model is to be simulated.

Addressing these challenges is imperative for developing more robust and accurate predictive models, ensuring that reliable and comprehensive data sources inform water resource management strategies. Continued research efforts focusing on refining methodologies, integrating advanced monitoring technologies, and enhancing collaboration among stakeholders will be vital to overcoming these challenges and improving the effectiveness of water bloom prediction and mitigation efforts.

This study's results highlight the Random Forest algorithm's predictive capabilities in anticipating water blooms and provide valuable insights into the influential factors and dynamics contributing to these events. Identifying key predictors, such as chlorophyll, total nitrogen, total phosphorus, and temperature, lays the foundation for targeted and informed strategies in mitigating the impact of water blooms on freshwater ecosystems.

The novelty of this research lies in predicting water bloom occurrences in multiple freshwater bodies, from reservoirs to lakes, using a classification model. This contrasts with earlier studies that primarily focused on a single water body and often relied on the measurement of a regression model instead of a classification model. This research has advanced the understanding of water quality dynamics and contributed to developing proactive strategies to mitigate the environmental consequences of water blooms.

4. Conclusion

In this study, a Random Forest model was developed to predict water bloom occurrences in various freshwater bodies, achieving an overall accuracy of 0.74. The model identifies chlorophyll a, total nitrogen, total phosphorus, and temperature as significant predictors of water blooms. These findings highlight the importance of nutrient levels and environmental conditions in understanding and forecasting water blooms. Our results demonstrate the effectiveness of machine learning in ecological monitoring and suggest practical applications for water resource management. By providing insights into the dynamics of water bloom events, our study contributes to developing proactive strategies for environmental management and mitigating water quality issues. Future research should focus on enhancing model robustness across diverse ecological conditions, enlarging data scales, and improving data quality. Additionally, comparing and exploring the integration of other machine-learning techniques could further refine predictive accuracy and generalizability.

Acknowledgment

This study was supported by a Grant-in-Aid for Scientific Research (B) of the Japan Society for Promotion of Science (Research No. 22H03769) to K.S. The authors thank the Japan Science and Technology Agency (JST) for their financial support to T.L. (Grant No. JPMJSP2124).

References

Anandhi, G., & Iyapparaja, M. (2024). Systematic approaches to machine learning models for predicting pesticide toxicity. Heliyon, 10(7), e28752. https://doi.org/10.1016/j.heliyon.2024.e28752

Bonaccorso, G. (2018). Machine Learning Algorithms - Second Edition: Popular algorithms for data science and machine learning (2nd ed.). Packt Publishing. ISBN: 978-1789347999.

Brophy, J., & Lowd, D. (2022). Instance-Based Uncertainty Estimation for Gradient-Boosted Regression Trees. arXiv preprint arXiv:2205.11412 [cs.LG]. https://doi.org/10.48550/ arXiv.2205.11412

Burford, M. A., Carey, C. C., Hamilton, D. P., Huisman, J., Paerl, H. W., Wood, S. A., & Wulff, A. (2020). Perspective: Advancing the research agenda for improving understanding of cyanobacteria in a future of global change. Harmful Algae, 91, 101601. https://doi.org/10.1016/j.hal.2019.04.004

CARU. (2016). Estudio de la calidad del agua del Río Uruguay. Bienio 2013–2014. Paysandú: Comisión Administradora del Río Uruguay.

CARU. (2017). Programa de vigilancia de playas del Río Uruguay. Paysandú: Comisión Administradora del Río Uruguay. Retrieved from http://www.caru.org.uy/web/2017/12/programa-de-vigilancia-de-playas-del-rio-uruguay/.

Davis, T. W., Berry, D. L., Boyer, G. L., & Gobler, C. J. (2009). The effects of temperature and nutrients on the growth and dynamics of toxic and non-toxic strains of Microcystis during cyanobacteria blooms. Harmful Algae, 8(5), 715-725. https://doi.org/10.1016/j.hal.2009. 02.004

Derot, J., Yajima, H., & Jacquet, S. (2020). Advances in forecasting harmful algal blooms using machine learning models: A case study with *Planktothrix rubescens* in Lake Geneva. Harmful Algae, 99, 101906. https://doi.org/10.1016/j.hal.2020.101906.

Downing, T. G., Sember, C. S., Gehringer, M. M., & Leukes, W. (2005). Medium N:P ratios and specific growth rate comodulate microcystin and protein content in Microcystis aeruginosa PCC7806 and M. aeruginosa UV027. Microbial ecology, 49(3), 468–473. https://doi.org/10.1007/s00248-004-0054-2

Drobac, D., Tokodi, N., Simeunović, J., Baltić, V., Stanić, D., &

Sviršev, Z. (2013). Human exposure to cyanotoxins and their effects on health. Arh Hig Rada Toksikol, 64(2), 119-130. https://doi.org/10.2478/10004-1254-64-2013-2320

Du, X., Liu, J., Wang, X., Chen, X., Mao, Z., Yu, F., Wang, P., Wu, C., Guo, H., & Zhang, H. (2024). Environmentally related microcystin-LR-induced ovarian dysfunction via the CCL2-CCR10 axis in mice ameliorated by dietary mulberry. Environmental pollution (Barking, Essex: 1987), 349, 123929. https://doi.org/10.1016/j.envpol.2024.123929

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using Random Forests. Pattern Recognition Letters, 31(14), 2225-2236. https://doi.org/10.1016/j.patrec.2010. 06.005

Kim, K. B., Uranchimeg, S., & Kwon, H. H. (2022). A multivariate Chain-Bernoulli-based prediction model for cyanobacteria algal blooms at multiple stations in South Korea. Environmental pollution (Barking, Essex: 1987), 313, 120078. https://doi.org/10.1016/j.envpol.2022.120078

Lad, A., Breidenbach, J. D., Su, R. C., Murray, J., Kuang, R., Mascarenhas, A., Najjar, J., Patel, S., Hegde, P., Youssef, M., Breuler, J., Kleinhenz, A. L., Ault, A. P., Westrick, J. A., Modyanov, N. N., Kennedy, D. J., & Haller, S. T. (2022). As We Drink and Breathe: Adverse Health Effects of Microcystins and Other Harmful Algal Bloom Toxins in the Liver, Gut, Lungs and Beyond. Life (Basel, Switzerland), 12(3), 418. https://doi.org/ 10.3390/life12030418

Liaw, A., & Wiener, M. (2018). Breiman and Cutler's Random Forests for Classification and Regression. CRAN.

Lin, S., Pierson, D. C., & Mesman, J. P. (2023). Prediction of algal blooms via data-driven machine learning models: An evaluation using data from a well-monitored mesotrophic lake. Geoscientific Model Development, 16, 35–46. https://doi.org/10.5194/gmd-16-35-2023.

Marrone, B. L., Banerjee, S., Talapatra, A., Gonzalez-Esquer, C. R., & Pilania, G. (2023). Toward a predictive understanding of cyanobacterial harmful algal blooms through AI integration of physical, chemical, and biological data. ACS ES&T Water, 4(3), 844-858. https://doi.org/10.1021/acsestwater.3c00369

Marwala, T., Mongwe, W. T., & Mbuvha, R. (2023). Introduction to Hamiltonian Monte Carlo. In T. Marwala, W. T. Mongwe, & R. Mbuvha (Eds.), Hamiltonian Monte Carlo Methods in Machine Learning (pp. 1-29). Academic Press. ISBN: 978-0443190353. https://doi.org/10.1016/B978-0-44-319035-3.00013-6

Mittelman, N. S., Engiles, J. B., Murphy, L., Vudathala, D., & Johnson, A. L. (2016). Presumptive iatrogenic microcystin-associated liver failure and encephalopathy in a Holsteiner Gelding. Journal of Veterinary Internal Medicine, 30(5), 1747-1751. https://doi.org/10.1111/ jvim.14571

Zarrouk, C. 1966. Contribution a l'etude d'une Cyanophycee. Influence de Divers Facteurs Physiques et Chimiques sur la croissance et la photosynthese de *Spirulina mixima*. Thesis. University of Paris, France

Piontek, M., Czyżewska, W., & Mazur-Marzec, H. (2023). Effects of harmful cyanobacteria on drinking water source quality and ecosystems. Toxins (Basel), 15(12), 703. https://doi.org/ 10.3390/toxins15120703

Posit Software, PBC. (2023). RStudio: Integrated Development Environment for R (Version 2023.06.2+561 "Mountain Hydrangea" Release). Retrieved from https://www.rstudio.com/

Pyo, J., Cho, K. H., Kim, K., Baek, S. S., Nam, G., & Park, S. (2021). Cyanobacteria cell prediction using interpretable deep learning model with observed, numerical, and sensing data assemblage. Water Research, 203, 117483. https://doi.org/10.1016/j.watres.2021.117483

Pyo, J., Park, L. J., Pachepsky, Y., Baek, S. S., Kim, K., & Cho, K. H. (2020). Using convolutional neural network for predicting cyanobacteria concentrations in river water. Water Research, 186, 116349. https://doi.org/10.1016/j.watres.2020.116349

Radeloff, V. C., Roy, D. P., Wulder, M. A., Anderson, M., Cook, B., Crawford, C. J., Friedl, M., Gao, F., Gorelick, N., Hansen, M., Healey, S., Hostert, P., Hulley, G., Huntington, J. L., Johnson, D. M., Neigh, C., Lyapustin, A., Lymburner, L., Pahlevan, N., Pekel, J. F., Scambos, T. A., Schaaf, C., Strobl, P., Woodcock, C. E., Zhang, H. K., & Zhu, Z. (2024). Need and vision for global medium-resolution Landsat and Sentinel-2 data products. Remote Sensing of Environment, 300, 113918. https://doi.org/10.1016/j.rse.2023.113918

Ramirez-Santana, M. (2018). Limitations and Biases in Cohort Studies. In R. Mauricio Barría (Ed.), Cohort Studies in Health Sciences. InTech. https://doi.org/10.5772/intechopen. 74324

Rapala, J., Sivonen, K., Lyra, C., & Niemelä, S. I. (1997). Variation of microcystins, cyanobacterial hepatotoxins, in Anabaena spp. as a function of growth stimuli. Applied and environmental microbiology, 63(6), 2206–2212. https://doi.org/10.1128/aem.63.6.2206-2212.1997

Rather, M. A., Ahmad, I., Shah, A., Ahmad Hajam, Y., Amin, A., Khursheed, S., Ahmad, I., & Rasool, S. (2024). Exploring opportunities of Artificial Intelligence in aquaculture to meet increasing food demand. Food Chemistry X, 22, 101309. https://doi.org/10.1016/j.fochx. 2024.101309

R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Sandhu, P. K., Solonenka, J. T., & Murch, S. J. (2024). Neurotoxic non-protein amino acids in commercially harvested Lobsters (*Homarus americanus* H. Milne-Edwards). Scientific Reports, 14(1), 8017. https://doi.org/10.1038/s41598-024-58778-1

Schaeffer, B. A., Reynolds, N., Ferriby, H., Salls, W., Smith, D., Johnston, J. M., & Myer, M. (2024). Forecasting freshwater cyanobacterial harmful algal blooms for Sentinel-3 satellite resolved U.S. lakes and reservoirs. Journal of environmental management, 349, 119518. https://doi.org/10.1016/j.jenvman.2023.119518

Shan, K., Wang, X., Yang, H., Zhou, B., Song, L., & Shang, M. (2020). Use statistical machine learning to detect nutrient thresholds in Microcystis blooms and microcystin

management. Harmful Algae, 94, 101807. https://doi.org/10.1016/j.hal.2020.101807

Sivonen K. (1990). Effects of light, temperature, nitrate, orthophosphate, and bacteria on growth of and hepatotoxin production by *Oscillatoria agardhii* strains. Applied and environmental microbiology, 56(9), 2658–2666. https://doi.org/10.1128/aem.56.9.2658-2666.1990

Varsha, S., Adalarasu, K., Jagannath, M., & Arunkumar, T. (2023). Chapter 7 - IoT in modern healthcare systems focused on neuroscience disorders and mental health. In B. Bhushan, S. K. Sharma, M. Saračević, & A. Boulmakoul (Eds.), Cognitive Data Science in Sustainable Computing, Blockchain Technology Solutions for the Security of IoT-Based Healthcare Systems (pp. 133-149). Academic Press. https://doi.org/10.1016/B978-0-323-99199-5.00006-9

Vézie, C., Rapala, J., Vaitomaa, J., Seitsonen, J., & Sivonen, K. (2002). Effect of nitrogen and phosphorus on growth of toxic and nontoxic Microcystis strains and on intracellular microcystin concentrations. Microbial ecology, 43(4), 443–454. https://doi.org/10.1007/ s00248-001-0041-9

Vidal, F., Sedan, D., D'Agostino, D., Cavalieri, M. L., Mullen, E., Parot Varela, M. M., Flores, C., Caixach, J., & Andrinolo, D. (2017). Recreational Exposure during Algal Bloom in Carrasco Beach, Uruguay: A Liver Failure Case Report. Toxins, 9(9), 267. https://doi.org/10.3390/toxins9090267

Villanueva, P., Yang, J., Radmer, L., Liang, X., Leung, T., Ikuma, K., Swanner, E. D., Howe, A., & Lee, J. (2023). One-Week-Ahead Prediction of Cyanobacterial Harmful Algal Blooms in Iowa Lakes. Environmental Science & Technology, 57(49), 20636-20646. https://doi.org/10.1021/acs.est.3c07764

Wang, C., Liu, J., Qiu, C., Su, X., Ma, N., Li, J., Wang, S., & Qu, S. (2023). Identifying the drivers of chlorophyll-a dynamics in a landscape lake recharged by reclaimed water using interpretable machine learning. Science of The Total Environment, 906, 167483. https://doi.org/10.1016/j.scitotenv.2023.167483

Wang, X., Wang, L., Shang, M., Song, L., & Shan, K. (2022). Revealing Physiochemical Factors and Zooplankton Influencing Microcystis Bloom Toxicity in a Large-Shallow Lake Using Bayesian Machine Learning. Toxins, 14(8), 530. https://doi.org/10.3390/toxins14080530

Wei, B., Sugiura, N., & Maekawa, T. (2001). Use of artificial neural network in the prediction of algal blooms. Water research, 35(8), 2022–2028. https://doi.org/10.1016/s0043-1354(00)00464-4

Welker, M., Chorus, I., Schaeffer, B. A., & Urquhart, E. (2021). Planning monitoring programmes for cyanobacteria and cyanotoxins. In I. Chorus & M. Welker (Eds.), Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring and Management (2nd ed., pp. 641-667). CRC Press. https://doi.org/10.1201/9781003081449

Yoshida, M., Yoshida, T., Takashima, Y., Hosoda, N., & Hiroishi, S. (2007). Dynamics of microcystin-producing and non-microcystin-producing Microcystis populations is correlated with nitrate concentration in a Japanese lake. FEMS microbiology letters, 266(1), 49–53. https://doi.org/10.1111/j.1574-6968.2006.00496.x