A framework for identifying factors controlling cyanobacterium Microcystis flos-aquae blooms by coupled CCM-ECCM Bayesian networks

Ofir Tal[1], Ilia Ostrovsky[1], and Gideon Gal[1]

[1]Israel Oceanographic and Limnological Research

June 21, 2023

Abstract

1. Cyanobacterial blooms in freshwater sources are a global concern, and gaining insight into their causes is crucial for effective resource management and control. 2. In this study, we present a computational framework for the causal analysis of cyanobacterial harmful algal blooms (cyanoHABs) in Lake Kinneret. Our framework integrates Convergence Cross Mapping (CCM) and Extended CCM (ECCM) causal networks with Bayesian Network (BN) models. 3. The constructed CCM - ECCM causal networks and BN models unveil significant interactions among factors influencing cyanoHAB formation. These interactions have been validated by domain experts and supported by evidence from peer-reviewed publications. Our findings suggest that M. flos-aquae levels are influenced not only by community structure but also by nitrate, nitrite, ammonium, phosphate, oxygen, and temperature levels in the weeks preceding bloom occurrences. 4. We have demonstrated a non-parametric computational framework for the causal analysis of a multivariate ecosystem. Our framework offers a more comprehensive understanding of the underlying mechanisms driving M. flos-aquae in Lake Kinneret. It captures complex interactions and provides an explainable prediction model. By considering causal relationships, temporal dynamics, and joint probabilities of environmental factors, the proposed framework enhances our understanding of cyanoHABs in Lake Kinneret.

A framework for identifying factors controlling cyanobacterium *Microcystis flos-aquae* blooms by coupled CCM-ECCM Bayesian networks

Affiliation

Tal, O., Ostrovsky, I. and Gal, G.

Kinneret Limnological Laboratory, Israel Oceanographic and Limnological Research

PO Box 447, Migdal 14950, Israel

Abstract

1. Cyanobacterial blooms in freshwater sources are a global concern, and gaining insight into their causes is crucial for effective resource management and control.
2. In this study, we present a computational framework for the causal analysis of cyanobacterial harmful algal blooms (cyanoHABs) in Lake Kinneret. Our framework integrates Convergence Cross Mapping (CCM) and Extended CCM (ECCM) causal networks with Bayesian Network (BN) models.
3. The constructed CCM - ECCM causal networks and BN models unveil significant interactions among factors influencing cyanoHAB formation. These interactions have been validated by domain experts and supported by evidence from peer-reviewed publications. Our findings suggest that *M. flos-aquae* levels are influenced not only by community structure but also by nitrate, nitrite, ammonium, phosphate, oxygen, and temperature levels in the weeks preceding bloom occurrences.
4. We have demonstrated a non-parametric computational framework for causal analysis of a multivariate ecosystem. Our framework offers a more comprehensive understanding of the underlying mechanisms driving *M. flos-aquae* in Lake Kinneret. It captures complex interactions and provides explainable prediction

model. By considering causal relationships, temporal dynamics, and joint probabilities of environmental factors, the proposed framework enhances our understanding of cyanoHABs in Lake Kinneret.

Keywords

Causality, CCM, ECCM, Ecosystem, Bayesian network, cyanoHAB, *Microcystis*, Freshwater

Introduction

Toxic cyanobacterial blooms (cyanoHABs) have a global impact, altering communities and producing toxins in lakes and water bodies. These events are influenced by climate change [1–3] and various environmental factors [4–8]. CyanoHABs negatively impact their environment by altering its chemical and physical properties [9], and releasing toxins and allelopathic compounds [10]. They also impact community structure and composition [11,12]. Cyanobacterium *M. flos-aquae* is the dominant toxin-producing species in Lake Kinneret and many other freshwater lakes and reservoirs [13]. *Microcystis* blooms often occur in warm, nutrient-rich waters with high levels of nitrogen and phosphorus. Additionally, *Microcystis* is known to thrive in alkaline conditions. The ability to cope with alkaline pH and to use different N species, grant *Microcystis* the advantage over other phytoplankton species. Factors such as changes in water temperature, light, and water circulation can also play a role in the development of *Microcystis* blooms [14]. Since 1995, Lake Kinneret has experienced significant and rapid ecological change, leading to increased frequency and magnitude of toxic blooms [15,16]. The development of cyanoHABs prediction models was studied before and reviewed by [17,18] and others. To date, cyanoHABs predictions have been carried out by both process-based, and data-driven approaches [17]. However, it’s important to understand the complex relations of cyanoHABs, community structure, and environmental factors.

The definition of causal relations between components of an ecosystem provides a valuable approach for understanding the key drivers and mechanisms behind specific events. By examining the relationships and interactions between the different components of an ecosystem, such as nutrients levels, phytoplankton communities, and environmental factors, we can identify the factors contributing to the development and persistence of cyanoHABs in Lake Kinneret. In recent years, causal relations between ecosystem components have been increasingly used to examine the drivers and impacts of the different components of ecological systems[19,20,21,22,23]. Traditionally, causal relations between variables of the same system, assuming X and Y, are measured by the amount of information of past X that is encoded into future Y [24–26]. Granger Causality (GC) is used to identify and measure causality in time series [27]. According to GC, X causes Y if the predictability of Y decreases when X is removed from the system. However, GC fails in dynamic systems consisting of variables that are not completely stochastic, with weak to moderate interactions. An alternative method, Convergent Cross Mapping (CCM), was recently presented by Sugihara et al. [19]. CCM assumes [28] that if two variables X and Y are of the same dynamic system, assuming X causes Y, then information about the state of X is embedded in Y and can be recovered. Interaction strength and directionality, between the two variables, can be quantified by measuring the prediction skill of the two variables using an increasing number of system states until convergence. CCM also captures causal interactions which are not necessarily linear [19]. This approach has been successfully implemented to reveal the causal effects in complex ecosystems [29–32]. Although CCM presents impressive performance in the identification of causal interactions in ecosystems, it carries essential drawbacks: (a) CCM does not supply information on synchrony between X and Y occurring by a strong driving force, (b) it does not specify whether the interaction is direct or indirect, and (c) CCM identifies causal interactions but does not supply information about their occurrence probabilities. The last point is extremely important for the understanding and possible management of complex and dynamic ecosystems. More recently, the Extended CCM (ECCM) was presented by Ye et al. [33], address the first two drawbacks. ECCM performs multiple CCM calculations at a range of time shifts of Y relative to X, to identify the lag of optimal prediction skill, which allows the identification of information flow direction. Bayesian networks (BN) are probabilistic graphical models that use conditional probability distributions to specify the influence of the system's variables on a target variable [34]. Yet, when the structure is learned from the data, it lacks of directionality and is strongly affected by correlations. Therefore, the reliable construction of BNs requires the knowledge of domain experts. BNs have been used before for the study of causality in ecosystems [35–37] due to their probabilistic nature.

Here, we suggest a novel causality analysis framework based on the use of CCM and ECCM for the construction of a target-focused interaction network, on which the BN is calculated. Using the complex Lake Kinneret ecosystem as a case study, we constructed a computational framework to investigate the causes of toxic cyanobacterium *M. flos-aquae* blooms.

Methods

Study site and data

Lake Kinneret (the Sea of Galilee) is a 170 km^2 warm meso-eutrophic lake located in northern Israel (Figure 1). The lake has a maximum depth of about 43 meters. CyanoHABs are especially critical in Lake Kinneret, the only freshwater lake in Israel and an essential source of drinking water, irrigation, fishing, and recreational activity. Understanding the dynamics leading to toxic blooms and producing accurate predictions of cyanoHABs would provide a powerful tool for proactive resource management and control of such events.

The Lake Kinneret Monitoring Program, which has been active since 1969, is conducted by the Kinneret Limnological Laboratory, IOLR. Routine measurements of physical, biological, and chemical variables are performed [38]. The current study utilizes a 21 year data set (2000-2020) consisting of measured phytoplankton biomass (Prasinophyte, Chlorophyta, Diatomaceae, Dinoflagellate, Cyanobacteria, Haptophytes, Cryptophytes) in the water column and measurements of the main environmental components (nitrite, nitrate, ammonium, oxygen, particulate organic nitrogen, organic nitrogen, chloride, total dissolved phosphorus,organic nitrogen dissolved, phosphate, turbidity, pH) in the upper 10-m stratum on a weekly-biweekly basis), as well as surface water temperature and inflow volume. The environmental variables were calculated as the sum per m^2 of the upper 10 meters, and the mean of biomass per m^2 of the upper 10 meters was used for the phytoplankton variables. All of the measurements used in this study are from the deepest station A, located at the center of the lake.

Data processing

The dataset was processed as follows. Z-scores of 3 (three standard deviations from the mean) were considered as outliers, therefor discarded and interpolated. The data set was re-sampled to a 7-day resolution, and was normalized to a range confined between 0 and 1 for the CCM calculations. For the BN approach, the data was re-sampled to a 5-day resolution and was categorized into three categories: ‘0’, ‘1’ and ‘2’ (Table S1).

Causal interactions

Convergent cross mapping

CCM was utilized to elucidate the presence and direction of weak to moderate non-linear causal interactions. The principle of CCM is based on the ability to predict the system state of a variable (X) by the system state of another variable (Y) [28]. If X is causal to Y, then information of X should be presented in Y, therefore the state space manifold reconstructed for Y should be able to predict the system's states of X [19]. Here, the 21-year time series was divided into multiple subsets by a 100-week sliding window and a 10-week gap between the windows. Lagged coordinate vectors of the different variables were calculated, where E is an embedding dimension, and l is the lag step. The optimal E was selected based on the simplex projection [39], and the optimal l was selected from the first minimum in the mutual information between the time series and a shifted version of itself, using the Python package skccm [40]. The S-map method [41] was used to test the nonlinearity of the system with the PyEDM [42] python package, since in the nonlinear system, the prediction skill improves as lag increases. In case the optimal E or the optimal l were larger than 20, default E = 5 and l = 2 were set. The data was split into train (0.75) and test (0.25) subsets. Prediction skill (*p*) scores were calculated on an increasing number of system states (library size). Here, *p* was calculated as the mean *p* that was calculated from the sliding windows of converged cross mapping, e.g. sliding windows of the same time series which are not of converged cross mapping were excluded.

A target-focused network of interactions was constructed as follows: (a) *M. flos-aquae* was set as the target; (b) CCM was calculated for all possible interactions of the target with all of the other variables; (c) the variables of causal interactions with the target were extracted; (d) CCM was calculated for all possible interactions within the causal variables that were extracted above; (e) mean prediction skill (*p*) < 0.01 was used to filter out very weak interactions; (f) we considered interaction if at least 10 of the sliding windows within a certain time series were of converged cross mapping and $p >= 0.01$.

We used surrogate time series to test the significance of CCM – ECCM results. Surrogate time series are created by modifying the data while preserving certain statistical properties of the original time series, such as its mean, variance, trend and autocorrelation. If the result of the CCM calculation for the original time series is higher than the same calculation done on multiple surrogate time series, then it is considered significant. This means that the nonlinear correlation between the two time series is likely to be real and not due to chance. Here, Ebisuzaki's (PyEDM) method [42] was used to generate surrogate data sets. If the observed prediction skill was greater than the 0.95 quantile of CCM prediction skill scores generated from 10,000 surrogate time series, it was considered significant.

Extended CCM

Extended CCM (ECCM) allows the detection of the optimal delay-lag and discriminates the real unidirectional causal relationship from bidirectional causation through adjusting the cross-map lag time (l) [33]. This method is capable of identifying synchronization effects and false interactions that decrease CCM performance. In this method, CCM is calculated from a series of shifted data sets [33]. In real causal interactions, the driving variable X can affect only the present or future Y. Therefore, Y can only predict the present or past values of X, but not its future values. Hence, the time lag between effect and cause must be non-positive. If the optimal prediction skill lag of both 'X causes Y' and 'Y causes X' is equal to 0 and of similar magnitude, it means that both respond instantaneously to a strong driving force. If the optimal prediction skill of 'X causes Y' is of a negative lag, and the optimal prediction skill 'Y causes X' is of a positive lag, then the influence of X is strong enough to 'enslave' Y due to synchrony [33]. When both 'X causes Y' and ‘Y causes X’ present optimal prediction skill of negative lag, the causal interaction is bi-directional. Given these guidelines, it is possible to validate the direction of causal interactions, determine the delay between the cause and the effect, and identify synchrony. Here, ECCM was tested for all causal interactions detected by the CCM analysis. For this analysis, a 400 data-points frame was considered using $E = 5$, $I = 2$, maximum library size of 200 and shift range of -20 to 20 weeks. This information was used to refine and filter the interactions network.

Simulations

To validate the framework presented here, and to understand the limitations of CCM and ECCM, a series of experiments was conducted based on a well-studied simulation of four species interactions adopted from equation (3) in [33]:

The simulated (eco)system consists of four species / factors whose direct time-dependent relationships can be expressed by the following synthetic relationships that occurred at time steps (lag) of (t+1):

$Y1(t+1) = Y1(t)[3.9 – 3.9Y1(t)]$,

$Y2(t+1) = Y2(t)[3.6 – 0.4Y1(t) – 3.6Y2(t)]$,

$Y3(t+1) = Y3(t)[3.6 – 0.4Y2(t) – 3.6Y3(t)]$,

$Y4(t+1) = Y4(t)[3.8 – 0.35Y3(t) – 3.8Y4(t)]$,

These equations simulate self-dynamics together with direct dependence of Y2 on Y1, Y3 on Y2, and Y4 on Y3, and consists of complex direct and indirect interactions. Although the network is small, it implies similar challenges associated with the analysis of real-world ecological interactions. We are aware of the fact that the current simulation is homogeneous in manners of system stability and influence of external drivers, while their influence in real-world systems fluctuates over time. Therefore, heterogeneous data were simulated by generating a weak causally connected data set by multiplying random components by Y2 and Y3, weakening the interactions of Y1, Y2, and Y3. The strong and the weak causally connected data were concatenated (1/3 strong causal interactions, and 2/3 weakened causal interactions) into a single heterogeneous data set. CCM was analyzed for the two synthetic types of data analyses - with and without implementing sliding windows.

Bayesian network

Python package bnlearn [43] was used to construct target-oriented BN model based on categorized (Table S3) 7-year historical data (2014-2020) (Figure S3). BNs are the non-parametric statistical method that describes the Bayesian probabilities of the system's components by directed acyclic graphs (DAG). Typically, the construction of BN involves multiple steps detailed in [44,45]. In our study, BN inference was used to elucidate the conditions that may promote the maximization or minimization of *M. flos-aquae* blooms. The BN was constructed based on causal interactions identified by CCM and ECCM. Since BN cannot represent feedback loops, it has to be calculated on a DAG structure. The CCM interactions network was processed as follows: (a) feedback loops were identified using the Python package Networkx [46]; (b) feedback loops were removed by identifying a feedback loop, and truncate the loop after the target node, or before a confounder node; and (c) sink nodes (nodes which don't consist of out-edges and are not *M. flos-aquae*) were removed. The conditional probability tables were calculated from the categorized data set. BN was also constructed by a structure learning approach directly from the categorized data sets using the HillClimbing algorithm [45].

Sensitivity analysis

Sensitivity analysis was used to assess the impact of changes in the input parameters of the BN model on the output of the model. It allows us to understand the robustness of the model and for identifying the input parameters that have the greatest impact on the output of the model. SHAP values (SHapley Additive exPlanations) [47] is a method used to increase transparency and interpretability of machine learning models by showing how each feature contributed to the prediction. We constructed 10,000 random permutations of different environmental scenarios, which were used as input vectors to the BN model. Then, the BN model's inputs and outputs were introduced to the shap.Explainer function, and SHAP values were used for the estimation of model sensitivity.

Computation

All calculations, analysis, and visualization were carried out under the Python environment and the relevant packages as described above.

Schematic illustration of the process is presented in (Figure 2).

Results

Blooming patterns of *M. flos-aquae*

Cyanobacterium *M. flos-aquae* has been observed in Lake Kinneret from the beginning of lake monitoring (1969), but only since 1996 it has frequently formed distinct winter-spring blooms (Figure 3a). During these blooms, the peak biomasses were moderate before 2009; then, between 2010 and 2016, higher peak biomasses were detected; since then, only irregular blooms have taken place. Annual dynamics observed during the last 21-year period show that *M. flos-aquae* abundance starts to increase in January, reaching the peak values usually during the second half of February – beginning of March. The minimal biomass is detected in August, with a following small increase in September (Figure 3b).

Causal interactions

Method validation using synthetic time-series

The CCM – ECCM approach implementation was validated using a well-studied synthetic data set of four components consisting of direct and indirect interactions (Figure 4a). CCM calculations of the simulated homogeneous time series, without sliding windows, successfully reconstructed the results presented by Ye et al. (2015). The influence of long-term indirect interactions was reduced when CCM was calculated using sliding windows (Figure 4a). In addition to the homogeneous data set, another version was created, in which the strong causal relations were weakened during the last two-thirds of the simulation, better representing the dynamics in a real ecosystem. The sliding window approach identified more causal interactions compared to the single-frame approach. The utilization of sliding windows reduced the identification of false interactions, which occurred when a single frame was used. The CCM calculation of the homogeneous time series using a sliding window technique identified the three direct interactions (y1 → y2, y2 → y3, and y3→ y4), however, missed a single indirect interaction of y1→ y4. CCM of the heterogeneous time series, without sliding windows, identified two direct interactions (y1 → y2 and y3 → y4), a single indirect interaction, and a single false interaction. Using sliding windows, CCM of the heterogeneous time series identified the three direct interactions but missed two indirect interactions (y1→, y4, and y2→ y4).

We compared the proposed CCM - ECCM approach to structure learning and Pearson correlations (Figure 4b). Structure learning failed to identify the simulated causal interactions, two false interactions in the homogeneous time series, and a single true direct interaction (y2 → y3) in the heterogeneous data set (Figure 4b). Pearson correlation results were too noisy and lacked information regarding directionality (Figure 4b).

Identification of causal interactions in Lake Kinneret historical records using coupled CCM – ECCM

CCM results present a complex array of interactions between the environmental (physical and chemical) variables and phytoplankton components (Table S2, Figure S1). Although more complex models can be more accurate, they may be more challenging to understand and interpret. In addition, too many nodes in a BN model can have a number of negative effects, including: reduced accuracy, increased computational complexity and overfitting [45]. Therefore, due to the complex network involving phytoplankton, individual species were aggregated by their taxonomic groups. Some of the causal interactions that were revealed by CCM (Figure 5a) are correlated (Figure 5b), while the other interactions correlate weakly or do not correlate at all. ECCM was used to calculate time-delayed interactions, and to identify synchrony and false discoveries in CCM results.

A total of 23 pairs, which represent all the possible interactions between all the environmental and biological variables and *M. floss-aquae,* were examined by CCM in the first iteration. Of those, 11 variables were of CCM prediction skill above 0.01 and used in the second iteration. All possible interactions (110, excluding self-interactions) between these variables were calculated in the second iteration. 39 interactions presented converged prediction skill above 0.01 and were validated using 10,000 surrogates for each interaction (Figure S2). We examined multiple surrogate cutoffs (0.9, 0.95, and 0.975, Table S2). We found that the BN of the interactions above the 0.9 quantiles are of higher accuracy compared to the models of the higher cutoffs (Table S2), and it consists of interactions supported by domain experts and previous research. Interestingly, the interactions of the lower cutoffs consist of more interactions involving biological variables (Figure S2). Only one biological variable is above the 0.975 cutoffs, while all the environmental variables are above the highest cutoff. The biological variables are more sensitive to long-term environmental trends. Following their weaker interactions, they are slightly masked by the surrogate results, which remain this characteristic of the original time series. Despite the slight decrease in accuracy, in order to gain information about the causes of *Microcystis* blooms and its interactions with the environment on the one hand and to avoid false positive results, we used the 0.95 cutoffs.

Of the 39 interactions, using the interactions above the 0.95 quantiles, 28 (71.8%) interactions were above this threshold and considered significant. The causal interactions were also examined by ECCM, which identified and discarded six false interactions. A total of 9 variables and 26 interactions constructed the final network.

The CCM scores of the environmental parameters were stronger (median=0.103, mean=0.099, std=0.021) in comparison to the scores of the phytoplankton community (median=0.068, mean=0.074, std=0.020) on *M. flos-aquae* (Figure 6a). The median lag between the cause to the effect of the environmental parameters (median=2, mean=3.4 mean=6, std=3.97) was of longer lag in comparison to the interactions of the biological parameters (median=1, mean=1.25, std=1.25) (Figure 6b).

Although we did not identify direct interaction between ammonium and *M. flos-aquae*, indirect effects were identified. We did identify this interaction with *M. aeruginosa* (ammonium, six weeks lag), another less abundant *Microcystis* species in the lake. Considering this interaction, domain experts' opinions, and evidence in the literature, we added 'ammonium causes *M. flos-aquae'* to the BN model (Figure 7a). Comparing structure learning and parameters learning approaches using the historical data set, four interactions were found to overlap (Figure 7b). The structure learning approach identified fewer (11) interactions than the CCM - ECCM structure learning (19). Like the simulation results, Pearson correlations are too noisy and do not supply information regarding interaction directionality (Figure 5b).

Directed acyclic graph (DAG) and Bayesian network model

The causal interactions network was converted to a DAG in order to construct a BN. The resulting DAG consisted of 9 nodes and 19 interactions, to which one direct interaction was added, 'ammonium causes *M. flos-aquae'*. To avoid over-fitting, we confirmed that the number of cases of each state, for each variable, was greater than 20 [44]. The dataset was split into training (0.75) and test (0.25) subsets. The BN model was calculated based on a shifted time series, where the cause and the effect were aligned according to the lags identified by ECCM analysis (Table S2). The model was evaluated based on confusion matrix (Figure S3), accuracy and AUC scores. *M. flos-aquae* BN model (Figure 8a) achieved an accuracy of 0.812 and AUC score of 0.817 (Table S3) considering a probability cutoff larger than, or equal to 0.5 (Figure 8b).

Sensitivity analysis and cyanobacteria blooming / non-blooming scenarios

Sensitivity analysis (Figure 9) shows the importance of the individual environmental parameters (oxygen, phosphate, nitrite, temperature, and ammonium) and the phytoplankton taxonomic groups on the model's output. The influence of nitrate is diluted due to its indirect interaction with *Microcystis*. The individual environmental variables present a lower effect on the model's output than the influence of the taxonomic groups. Mean scenarios were calculated based on permutations that produced high or low probabilities of *M. flos-aquae* bloom formation (Figure 10). Lower probabilities of *M. flos-aquae* blooms are associated with higher values of oxygen, nitrite, and Haptophytes but lower values of temperature, phosphate, nitrate, and ammonium. In contrast, higher blooming probabilities were associated with higher temperature, phosphate, nitrate, and ammonium values but lower nitrite, oxygen, and Haptophytes (mainly *Erkenia subaequiciliata*). These differences may pinpoint the potential factors favorable for developing higher *M. flos-aquae* biomass.

A notable finding from our study was the measured 6-week lag between changes in temperature and the subsequent effects on *M. flos-aquae*. This temporal relationship provides compelling evidence that winter temperatures in Lake Kinneret are causal to the occurrence of spring blooms. Furthermore, we observed temperature differences when comparing the mean maximization scenario (Figure 10a) to the mean minimization scenario (Figure 10b). Specifically, the mean maximization scenario depicted higher temperatures than the mean minimization scenario. This scenario-based analysis provides valuable insights into the potential consequences of elevated temperatures on the ecological dynamics of Lake Kinneret. It suggests that under warmer conditions, there may be an increase in the intensity or frequency of algal blooms.

In addition, the destratification of the lake, following lake overturn, which typically occurs in the second half of December or in January, elevates high concentrations of nutrients from the lower anoxic nutrient-rich layers of the lake to the nutrient-poor upper layer. Mixing occurs one to two months before the *M. flos-aquae* biomass peak. Mean scenarios also show the importance of higher ammonium and phosphate for *M. flos-aquae* growth, suggesting that the overturn-caused nutrient (ammonium and phosphate) supply to the upper productive layer is an important precondition for *M. flos-aquae* bloom development.

Discussion

Our framework for causal analysis represents an improvement over previous computational approaches in understanding the causes of cyanoHABs in Lake Kinneret. Traditional methods often relied on correlation-based analyses, which only provided limited insights into the complex interactions and causal relationships among different factors [17]. These approaches do not consider all three essential aspects - delayed effects, causality, and event probabilities. In contrast, our framework utilizes a targeted and focused approach by constructing CCM - ECCM causal networks and developing BN models based on these networks. This allows a more comprehensive understanding of the underlying mechanisms driving bloom formation. Our framework considers the temporal aspect by considering the weeks leading up to the blooming events. This temporal perspective provides a better understanding of various factors' delayed effects and cumulative influences, leading to an explainable predictive capability.

We validated the CCM - ECCM approach using a synthetic time series studied previously by Ye et al. (2015). Moreover, we reconstructed a more challenging system in which the causal relations of the above system were weakened during the simulation. Although the sliding window approach performed better than the single frame CCM calculations, it missed two indirect interactions. This approach may only partially identify indirect interactions of longer delayed effects. The targeted CCM - ECCM approach performs better than the structure learning approach, which failed to identify the interactions. The model has revealed several key relationships among the factors influencing cyanoHABs formation in Lake Kinneret. These interactions shed some light on the complex interplay between variables and provide a deeper understanding of the underlying mechanisms. Some of the noteworthy interactions include:

Effect of temperature on Microcystis. - Higher temperatures may affect increased *Microcystis* growth (Figure 10), although the optimal temperature is varied between the different *Microcystis* species. Higher temperatures were found to promote the development of toxic sub-populations [48,49]. Another study in which the interactions of BN were constructed by domain experts [50] shows that phosphorus and temperature are important for the development of cyanobacterial blooms. The monthly mean in Lake Kinneret (Figure 3b) shows the highest values of *M. flos-aquae* biomass in March, and relatively low values when the water gets warmest around June-August.

Effect of nitrate, nitrite and ammonium on Microcystis**.** - Ammonium, a reduced species of nitrogen, was found to be preferred by *Microcystis* over nitrate [13]. High nitrate levels were also found to promote development of toxic *Microcystis* populations [51]. As part of the nitrification process, nitrite is oxidized into nitrate in the presence of dissolved oxygen, which occurs in Lake Kinneret between January and April following the annual overturn. Here, nitrite was found to be related to lower values of *Microcystis,* suggesting that faster or earlier nitrification might promote *Microcystis* growth.

Effect of p*hosphate *on Microcystis. - Higher phosphorus values cause increased *Microcystis* growth [48], although biomass increase of various *Microcystis* species may be favored by different phosphorus levels [52]. According to our results, an increase of phosphate may affect *M. flos-aquae* biomass within 9 weeks. [53] showed that although the early growth stage of *Microcystis* population is affected by nitrate-to-ammonium ratio and phosphate concentration, its maximum growth rate is determined by a minimal phosphate concentration.

Effect of Oxygen on Microcystis. -*Microcystis* is highly tolerant to anoxic conditions [54,55]. The results show relatively immediate influence of oxygen on *Microcystis*. Our approach successfully captured these relations (Figure 5a, 10), and suggests that lower oxygen levels, in the right conditions, promote *M. flos-aquae* growth. This might be following the lack of other species blooms, which would increase oxygen levels due to photosynthesis. It also should be taken into account that a decrease in oxygen is associated with higher temperatures due to lower oxygen dissolution.

Inter-species interactions. - The formation of *Microcystis* colonies was found to be related with its bacterial microbiome [56,57] and the presence of multiple phytoplankton taxonomic groups [58]. The interactions between *Microcystis* and other community members is dynamic and bi-directional [59,60]. According to our results, *M. flos-aquae* is affected by other community members too. Inter-species interactions in the lake might be due to mutualism, amensalism, or competition. Both amensalism and competition may reduce *Microcystis* growth rate. Competition delayed effect lag might be longer, while amensalism through allelopathy is assumed to follow a shorter lag time. Indeed, the analysis of CCM scores, delayed effect lag time and model sensitivity results shed some light on the complex interactions of *M. flos-aquae* with the phytoplankton community in its environment. The interactions of *M. flos-aquae* with other phytoplankton species in the lake are weaker then its interactions with the environmental parameters. Even-though, *Microcystis* reacts faster to changes in community structure. These results are aligned with [61] which show that in diverse ecosystems biodiversity effects are more important than environmental effects as drivers of biomass.

Effect of temperature on nitrate and ammonium. - Higher temperatures can increase the rate of decomposition of organic matter and as a result, increase nitrate and ammonium levels in the water. Higher temperatures increase ammonia, nitrate, total nitrogen and phosphate concentrations in freshwater ecosystem [62]. On the other hand, increased winter inflows from watersheds and turnover-driven destratification are annual processes in the monomictic Lake Kinneret, occurring during the coldest season and affect nutrients level in the water body. In addition, following the turnover, large amounts of nitrite are oxidized to nitrate when arrive to the hypolimnion.

Effect of temperature on oxygen. - As lake temperatures increase, the amount of dissolved oxygen it can hold, decreases [63]. Higher temperatures also increase phytoplankton photosynthesis rate, which in turn increase dissolved oxygen levels in the upper layers [64].

Effect of nitrate on phosphate. - There are two counteracting effects of nitrate on phosphorus release from the sediments. Reduced nitrate increases the release of phosphate from some sediments, while nitrate also inhibits the release of iron-bound phosphorus from the sediment [65,66]. Higher phosphate levels promote phytoplankton growth [67,68], which leads to increased decomposition of organic matter due to the increase of total biomass [69]. This might also affect ammonium because both nitrate and ammonium are portions of total nitrogen.

***Effect of oxygen on phosphorus and* phosphate.** - Wu [70] showed that anaerobic conditions are more conducive to the release of phosphorus from the sediment than aerobic conditions. The effect of oxygen on the release and availability of phosphorus is complex and depends on the presence of other inorganic moieties in the lake [71]. Phosphate and oxygen concentrations are also related to phytoplankton biomass, which consumes oxygen at low light in the lower layers of the lake but increases oxygen levels in the upper layers where photosynthesis occurs.

It is important to note that these identified interactions are based on the specific context of Lake Kinneret.

The lags between cause and effect, calculated by ECCM (Figure 10, Table 2), support the above evidence. The effect of temperature on oxygen, nitrite, and nitrate is immediate (1-2 weeks) in ecological time scales. The interactions between phosphate and nitrogen species are more prolonged (8-11 weeks). These interactions are mediated by slow reduction processes in the sediment and biological processes in the lake. Long delayed effects are either slow processes, as mentioned above, or indirect effects through the sequence of events.

Intuitively, the importance of higher ammonium and phosphate for *M. flos-aquae* growth is in contrast to the temperature because with overturn, the temperature of the upper layer decreases. This may lead to shallow warm temperatures while most of the water column is cooler. However, *Microcystis* blooms occur only following sufficient warm, calm winter days [72]. Here we use temporal information of the lag between the cause and the effect (Figure 10), which shows that multiple processes of different time ranges affect *Microcystis* blooms. First, during the winter, around December – February, loads of nitrate, ammonium, and phosphate are washed from the drainage area into the lake. Later, warmer temperatures during the winter affect oxygen and nutrient levels. Global warming was found to delay the overturn in lakes [73], which reduces nitrite levels in the upper layer. In the absence of other prosper species in the lake, this sequence of events increases the probability of *Microcystis* blooms.

Synergistic effects refer to the phenomenon where the joint influence of multiple variables on an outcome is greater than the sum of their individual effects. This effect is particularly relevant when studying complex systems, where the interactions between variables can lead to nonlinear dynamics. In BNs, the network structure, represented as a DAG, allows for identifying synergistic and cumulative interactions among variables [74]. By considering the joint probability distribution of variables and their conditional dependencies, BNs can reveal synergistic effects within the system. Indeed, sensitivity analysis of the BN model presents a relatively weak effect of each parameter. However, their synergistic effect on the output values (*Microcystis* blooming probability) is of a broader range (Figure 8).

The results show that biological parameters are of higher importance in the sensitivity analysis of the BN model, while environmental parameters have higher CCM scores compared to the biological parameters. This can be explained by considering a synergistic effect between environmental and biological parameters. The apparent contradiction between the sensitivity analysis and the CCM scores could be attributed to the interplay and combined influence of these two sets of parameters.

The results could indicate that while biological parameters are individually important and exhibit strong statistical associations, their effects might be enhanced or modulated by the presence of certain environmental conditions. The combination of both sets of parameters working together may result in a more comprehensive understanding of the system's behavior, with the environmental parameters playing a crucial role in shaping and driving the overall dynamics.

The proposed framework presents a novel approach to understanding complex processes in ecological systems. However, there are still specific weaknesses that should be acknowledged: a. uncertainty in BN modeling: the accuracy of the BN models heavily relies on the availability and quality of data for training and validation. In addition, thresholds used for categorization and insufficient or noisy data may affect the reliability and generalizability of the models. b. data limitations: The framework's effectiveness is contingent on comprehensive and high-quality data availability. Incomplete or sparse data may limit the ability to accurately identify and capture all relevant causal relationships. c: Simplified representation: While the framework provides a more comprehensive understanding of the underlying mechanisms, it still relies on simplifications and assumptions to model the complex interactions among variables. This simplification may overlook specific interactions within the system, potentially leading to incomplete conclusions. d: hidden variables: not directly observed but impact the observed data. Incorporating hidden variables would allow for a more comprehensive representation of the causal structure. e: still depends on human interpretation of the ECCM results. And f: Interactions involving biological components tend to be masked by long-term trends, which should be considered.

Despite these weaknesses, the framework can offer valuable insights and contribute to understanding ecological systems. It provides a targeted and focused approach that considers the causal interactions and the temporal aspect, allowing for a better understanding of lagged effects and cumulative influences. Incorporating causality and dependent probabilities give the framework a more explainable predictive capability than traditional analyses.

Further improvements can be made in data collection and model refinement to address the weaknesses. Gathering more comprehensive and high-quality data, including long-term and continuous monitoring, can enhance the accuracy and robustness of the CCM causal networks and BN models. Additionally, incorporating unobserved variables can capture a more comprehensive representation of complex ecological systems.

We presented a computational framework of coupled CCM - ECCM and BN for causal analysis of complex ecosystems. As a case study, we focused on the bloom-forming *M. flos-aquae* species in deep subtropical Lake Kinneret. Given the causal interactions identified by CCM - ECCM and reviewed by domain experts, the structure of the causal network can be used as a basis for a BN model. Although it has been used in the ecological context, this is the first attempt to use BN models in conjunction with CCM and ECCM to understand the freshwater ecosystem. *M. flos-aquae* in Lake Kinneret is associated, by complex interactions, with the phytoplankton community but also driven by environmental variables such as temperature, nitrate, ammonium, nitrite and phosphate.

Declaration of competing interest

The authors have no conflicts of interest to declare.

Author Contributions

O.T. *conceived the ideas and designed methodology; O.T. analyzed the data; O.T. led the writing of the manuscript;* G.G. and I.O. *contributed critically to the drafts; All authors gave final approval for publicatio*n.

Data and supplementary information

Data availability

The code of this study is available on GitHub at https://github.com/ot483/ecol_evol_2023. The data used in this study have been deposited in the Zenodo repository and are publicly available and can be accessed and downloaded from the Zenodo [Zenodo URL] under the corresponding DOI [DOI].

References

1. Feng, L. *et al.* Dominant genera of cyanobacteria in Lake Taihu and their relationships with environmental factors. *J. Microbiol.* **54**, 468–476 (2016).

2. Paerl, H. W. & Huisman, J. Minireview Climate change : a catalyst for global expansion of harmful cyanobacterial blooms. **1**, 27–37 (2009).

3. Padisák, J. Cylindrospermopsis raciborskii (Woloszynska) Seenayya et Subba Raju, an expanding, highly adaptive cyanobacterium: worldwide distribution and review of its ecology. *Archiv Für Hydrobiologie Supplementband Monographische Beitrage* vol. 107 563–593 (1997).

4. Robarts, R. D. & Zohary, T. Temperature effects on photosynthetic capacity, respiration, and growth rates of bloom-forming cyanobacteria. *New Zeal. J. Mar. Freshw. Res.* **21**, 391–399 (1987).

5. Ninio, S., Lupu, A., Viner-Mozzini, Y., Zohary, T. & Sukenik, A. Multiannual variations in Microcystis bloom episodes – Temperature drives shift in species composition. *Harmful Algae* **92**, 101710 (2020).

6. Benincá, E. *et al.* Chaos in a long-term experiment with a plankton community. *Nature* **451**, 822–825 (2008).

7. Cao, H. S. *et al.* Effects of wind and wind-induced waves on vertical phytoplankton distribution and surface blooms of microcystis aeruginosa in lake taihu. *J. Freshw. Ecol.* **21**, 231–238 (2006).

8. Reichwaldt, E. S. & Ghadouani, A. Effects of rainfall patterns on toxic cyanobacterial blooms in a changing climate: Between simplistic scenarios and complex dynamics. *Water Res.* **46**, 1372–1393 (2012).

9. Ibelings, B. W., Vonk, M., Los, H. F. J., van der Molen, D. T. & Mooij, W. M. FUZZY MODELING OF CYANOBACTERIAL SURFACE WATERBLOOMS: VALIDATION WITH NOAA-AVHRR SATELLITE IMAGES. *Ecol. Appl.* **13**, 1456–1472 (2003).

10. Sukenik, A. *et al.* Inhibition of growth and photosynthesis of the dinoflagellate Peridinium gatunense by Microcystis sp. (cyanobacteria): A novel allelopathic mechanism. *Limnol. Oceanogr.* **47**, 1656–1663 (2002).

11. Svirčev, Z. B., Tokodi, N., Drobac, D. & Codd, G. A. Cyanobacteria in aquatic ecosystems in Serbia: effects on water quality, human health and biodiversity. *Syst. Biodivers.* **12**, 261–270 (2014).

12. Wang, K., Razzano, M. & Mou, X. Cyanobacterial blooms alter the relative importance of neutral and selective processes in assembling freshwater bacterioplankton community. *Sci. Total Environ.* **706**, 135724 (2020).

13. Harke, M. J. *et al.* A review of the global ecology, genomics, and biogeography of the toxic cyanobacterium, Microcystis spp. *Harmful Algae* **54**, 4–20 (2016).

14. Wilhelm, S. W., Bullerjahn, G. S. & McKay, R. M. L. The Complicated and Confusing Ecology of Microcystis Blooms. *MBio* **11**, (2020).

15. Hadas, O., Kaplan, A. & Sukenik, A. Long-term changes in cyanobacteria populations in lake kinneret (Sea of galilee), israel: An eco-physiological outlook. *Life* **5**, 418–431 (2015).

16. Sukenik, A., Zohary, T. & Markel, D. Lake Kinneret. *Lake Kinneret* 561–575 (2014) doi:10.1007/978-94-017-8944-8.

17. Rousso, B. Z., Bertone, E., Stewart, R. & Hamilton, D. P. A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. *Water Res.* **182**, 115959 (2020).

18. Cruz, R. C., Costa, P. R., Vinga, S., Krippahl, L. & Lopes, M. B. A review of recent machine learning advances for forecasting harmful algal blooms and shellfish contamination. *J. Mar. Sci. Eng.* **9**, (2021).

19. Sugihara, G. *et al.* Detecting causality in complex ecosystems. *Science (80-.).* **338**, 496–500 (2012).

20. Plowright, R. K., Sokolow, S. H., Gorman, M. E., Daszak, P. & Foley, J. E. Causal inference in disease ecology: investigating ecological drivers of disease emergence. *Front. Ecol. Environ.* **6**, 420–429 (2008).

21. Adams, S. M. Establishing Causality between Environmental Stressors and Effects on Aquatic Ecosystems. *Hum. Ecol. Risk Assess. An Int. J.* **9**, 17–35 (2003).

22. Doi, H., Yasuhara, M. & Ushio, M. Causal analysis of the temperature impact on deep-sea biodiversity. *Biol. Lett.* **17**, 2–7 (2021).

23. Bonotto, G., Peterson, T. J., Fowler, K. & Western, A. W. Identifying Causal Interactions Between Groundwater and Streamflow Using Convergent Cross-Mapping. *Water Resour. Res.* **58**, (2022).

24. Moraffah, R., Karami, M., Guo, R., Raglin, A. & Liu, H. Causal Interpretability for Machine Learning - Problems, Methods and Evaluation. *ACM SIGKDD Explor. Newsl.* **22**, 18–33 (2020).

25. Lucas, T. C. D. A translucent box: interpretable machine learning in ecology. *Ecol. Monogr.* **90**, 1–17 (2020).

26. Zhao, Q. & Hastie, T. Causal Interpretations of Black-Box Models. *J. Bus. Econ. Stat.* **39**, 272–281 (2021).

27. Granger, C. J. W. Investigating Causal Relations by Econometric Models and Cross-spectral Methods Authors (s): C . W . J . Granger Published by : The Econometric Society Stable URL : http://www.jstor.org/stable/1912791 Accessed : 25-03-2016 19 : 26 UTC Your use of the JS. *Econometrica* **37**, 424–438 (1969).

28. Deyle, E. R. & Sugihara, G. Generalized Theorems for Nonlinear State Space Reconstruction. *PLoS One* **6**, e18295 (2011).

29. Barraquand, F., Picoche, C., Detto, M. & Hartig, F. Inferring species interactions using Granger causality and convergent cross mapping. *Theor. Ecol.* **14**, 87–105 (2021).

30. Zhang, Z. *et al.* Identifying changes in China's Bohai and Yellow Sea fisheries resources using a causality-based indicator framework, convergent cross-mapping, and structural equation modeling. *Environ. Sustain. Indic.* **14**, 100171 (2022).

31. Nakayama, S., Takasuka, A., Ichinokawa, M. & Okamura, H. Climate change and interspecific interactions drive species alternations between anchovy and sardine in the western North Pacific: Detection of causality by convergent cross mapping. *Fish. Oceanogr.* **27**, 312–322 (2018).

32. Chang, C.-W. *et al.* Causal networks of phytoplankton diversity and biomass are modulated by environmental context. *Nat. Commun.* **13**, 1140 (2022).

33. Ye, H., Deyle, E. R., Gilarranz, L. J. & Sugihara, G. Distinguishing time-delayed causal interactions using convergent cross mapping. *Sci. Rep.* **5**, 1–9 (2015).

34. Milns, I., Beale, C. M. & Smith, V. A. Revealing ecological networks using Bayesian network inference algorithms. *Ecology* **91**, 1892–1899 (2010).

35. Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R. & Salmerón, A. Bayesian networks in environmental modelling. *Environ. Model. Softw.* **26**, 1376–1388 (2011).

36. McCann, R. K., Marcot, B. G. & Ellis, R. Bayesian belief networks: applications in ecology and natural resource management. *Can. J. For. Res.* **36**, 3053–3062 (2006).

37. Barton, D. N. *et al.* Bayesian networks in environmental and resource management. *Integr. Environ. Assess. Manag.* **8**, 418–429 (2012).

38. Sukenik, A., Zohary, T. & Markel, D. The monitoring program. *Lake Kinneret Ecol. Manag.* 561–575 (2014).

39. Sugihara, G. & May, R. M. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature* **344**, 734–741 (1990).

40. Nick Cortale. skcmm [Computer software]. (2016).

41. Chang, C. *et al.* Reconstructing large interaction networks from empirical time series data. *Ecol. Lett.* **24**, 2763–2774 (2021).

42. Sugihara, G. PyEDM [computer software].

43. Taskesen, E. bnlearn [Computer software]. (2020).

44. Chen, S. H. & Pollino, C. A. Good practice in Bayesian network modelling. *Environ. Model. Softw.* **37**, 134–145 (2012).

45. Marcot, B. G. Common quandaries and their practical solutions in Bayesian network modeling. *Ecol. Modell.* **358**, 1–9 (2017).

46. Aric A. Hagberg, D. A. S. and P. J. S. Exploring network structure, dynamics, and function using NetworkX. *Proc. 7th Python Sci. Conf.* 11–15 (2008).

47. Shapley, L. S. A value for n-person games. *Class. game theory* **69**, (1997).

48. Davis, T. W., Berry, D. L., Boyer, G. L. & Gobler, C. J. 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**, 715–725 (2009).

49. Imai, H., Chang, K.-H., Kusaba, M. & Nakano, S. -i. Temperature-dependent dominance of Microcystis (Cyanophyceae) species: M. aeruginosa and M. wesenbergii. *J. Plankton Res.* **31**, 171–178 (2008).

50. Moe, S. J., Haande, S. & Couture, R.-M. Climate change, cyanobacteria blooms and ecological status of lakes: A Bayesian network approach. *Ecol. Modell.* **337**, 330–347 (2016).

51. Yoshida, M., Yoshida, T., Takashima, Y., Hosoda, N. & Hiroishi, S. Dynamics of microcystin-producing and non-microcystin-producing Microcystis populations is correlated with nitrate concentration in a Japanese lake. *FEMS Microbiol. Lett.* **266**, 49–53 (2007).

52. Yue, T., Zhang, D. & Hu, C. Comparative studies on phosphate utilization of two bloom-forming Microcystis spp. (cyanobacteria) isolated from Lake Taihu (China). *J. Appl. Phycol.* **26**, 333–339 (2014).

53. Kim, H., Jo, B. Y. & Kim, H. S. Effect of different concentrations and ratios of ammonium, nitrate, and phosphate on growth of the blue-green alga (cyanobacterium) Microcystis aeruginosa isolated from the Nakdong River, Korea. *ALGAE* **32**, 275–284 (2017).

54. Chen, X. *et al.* The secretion of organics by living Microcystis under the dark/anoxic condition and its enhancing effect on nitrate removal. *Chemosphere* **196**, 280–287 (2018).

55. Brunberg, A.-K. Benthic overwintering of Microcystis colonies under different environmental conditions. *J. Plankton Res.* **24**, 1247–1252 (2002).

56. Wang, W. *et al.* Experimental evidence for the role of heterotrophic bacteria in the formation of Microcystis colonies. *J. Appl. Phycol.* **28**, 1111–1123 (2016).

57. Hoke, A. K. *et al.* Genomic signatures of Lake Erie bacteria suggest interaction in the Microcystis phycosphere. *PLoS One* **16**, e0257017 (2021).

58. Zhang, M. *et al.* Feedback Regulation between Aquatic Microorganisms and the Bloom-Forming Cyanobacterium Microcystis aeruginosa. *Appl. Environ. Microbiol.* **85**, (2019).

59. Omidi, A., Pflugmacher, S., Kaplan, A., Kim, Y. J. & Esterhuizen, M. Reviewing Interspecies Interactions as a Driving Force Affecting the Community Structure in Lakes via Cyanotoxins. *Microorganisms* **9**, 1583 (2021).

60. Schweitzer-Natan, O., Ofek-Lalzar, M., Sher, D. & Sukenik, A. The microbial community spatially varies during a Microcystis bloom event in Lake Kinneret. *Freshw. Biol.* **68**, 349–363 (2023).

61. Chang, C. W. *et al.* Causal networks of phytoplankton diversity and biomass are modulated by environmental context. *Nat. Commun.* **13**, 1–11 (2022).

62. Li, H. Y., Xu, J. & Xu, R. Q. The Effect of Temperature on the Water Quality of Lake. *Adv. Mater. Res.* **821**–**822**, 1001–1004 (2013).

63. Jankowski, T., Livingstone, D. M., Bührer, H., Forster, R. & Niederhauser, P. Consequences of the 2003 European heat wave for lake temperature profiles, thermal stability, and hypolimnetic oxygen depletion: Implications for a warmer world. *Limnol. Oceanogr.* **51**, 815–819 (2006).

64. Antonopoulos, V. Z. & Gianniou, S. K. Simulation of water temperature and dissolved oxygen distribution in Lake Vegoritis, Greece. *Ecol. Modell.* **160**, 39–53 (2003).

65. Søndergaard, M., Jeppesen, E. & Jensen, J. P. Hypolimnetic Nitrate Treatment to Reduce Internal Phosphorus Loading in a Stratified Lake. *Lake Reserv. Manag.* **16**, 195–204 (2000).

66. Tiren, T. & Pettersson, K. The influence of nitrate on the phosphorus flux to and from oxygen depleted lake sediments. *Hydrobiologia* **120**, 207–223 (1985).

67. Davison, W., George, D. G. & Edwards, N. J. A. Controlled reversal of lake acidification by treatment with phosphate fertilizer. *Nature* **377**, 504–507 (1995).

68. Gardner, E. M., McKnight, D. M., Lewis, W. M. & Miller, M. P. Effects of nutrient enrichment on phytoplankton in an alpine lake, Colorado, U.S.A. *Arctic, Antarct. Alp. Res.* **40**, 55–64 (2008).

69. Ma, S.-N. *et al.* Effects of nitrate on phosphorus release from lake sediments. *Water Res.* **194**, 116894 (2021).

70. Wu, Y., Wen, Y., Zhou, J. & Wu, Y. Phosphorus release from lake sediments: Effects of pH, temperature and dissolved oxygen. *KSCE J. Civ. Eng.* **18**, 323–329 (2014).

71. Hupfer, M. & Lewandowski, J. Oxygen Controls the Phosphorus Release from Lake Sediments - a Long-Lasting Paradigm in Limnology. *Int. Rev. Hydrobiol.* **93**, 415–432 (2008).

72. Hozumi, A., Ostrovsky, I., Sukenik, A. & Gildor, H. Turbulence regulation of Microcystis surface scum formation and dispersion during a cyanobacteria bloom event. *Inl. Waters* **10**, 51–70 (2020).

73. Anderson, E. J. *et al.* Seasonal overturn and stratification changes drive deep-water warming in one of Earth's largest lakes. *Nat. Commun.* **12**, 1688 (2021).

74. Simeoni, C. *et al.* Evaluating the combined effect of climate and anthropogenic stressors on marine coastal ecosystems: Insights from a systematic review of cumulative impact assessment approaches. *Sci. Total Environ.* **861**, 160687 (2023).

Figures and tables

Figure 1 – Study site. Lake Kinneret, northern Israel.

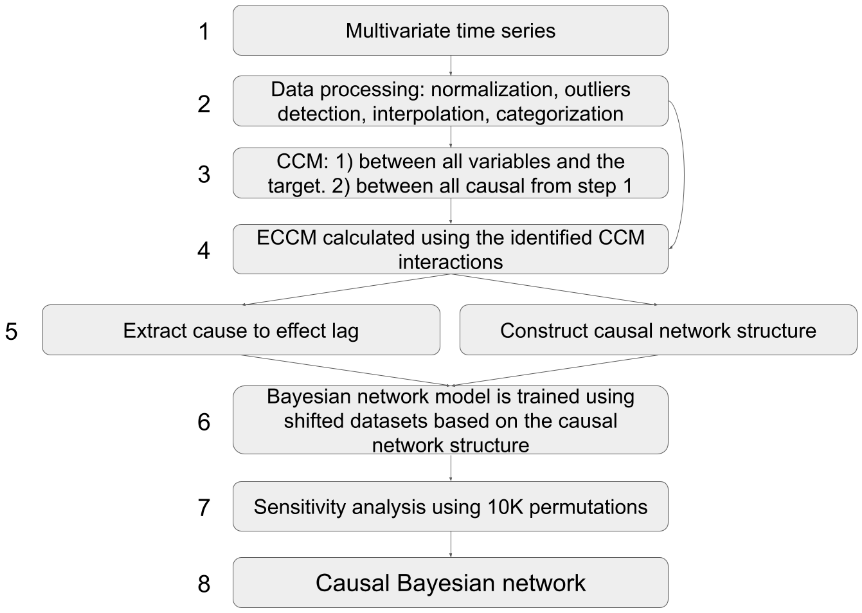


Figure 2 – Schematic description of the proposed framework. 1) Multivariate time series consists of the target (*M. flos-aquae*) and the other variables; 2) Data processing, including normalization, outliers detection, interpolation and categorization; 3) CCM is calculated between the target(s) and all of the other variables, followed by a second CCM step between all the causal variables; 4) Causal interactions from step 3 are validated and further filtered by ECCM analysis; 5) Cause-to-effect lags extractions, and construction of the causal network structure; 6) Parameters are shifted according to cause-to-effect lags, and used for BN model training, based on network structure from step 5; 7) Model sensitivity evaluation based on 10,000 permutations and random forest feature importance; and 8) Causal Bayesian network model.

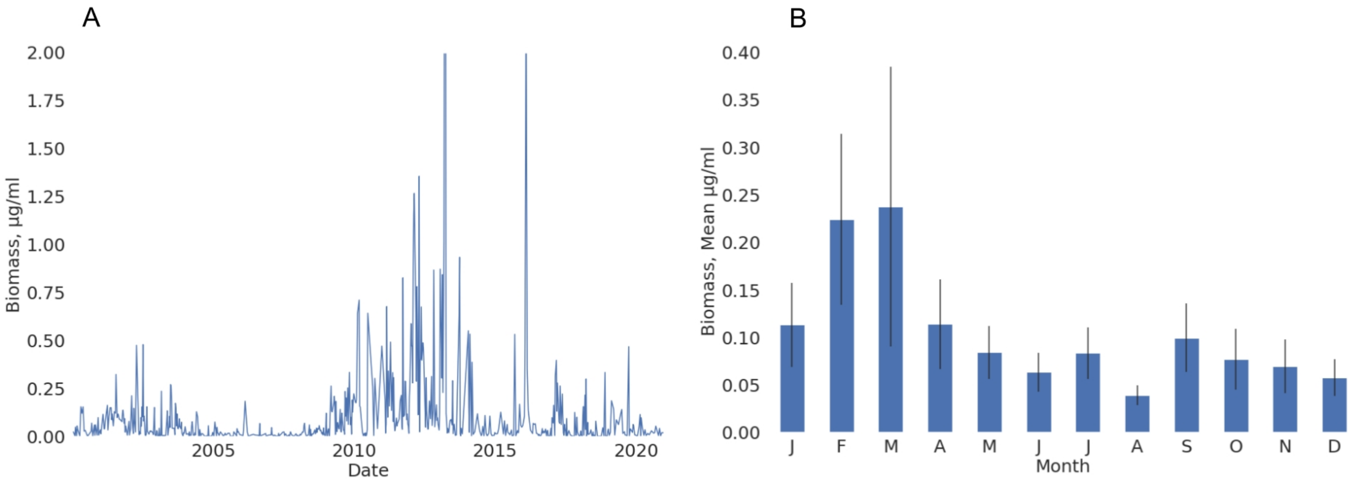


Figure 3 – a. *M. flos-aquae* values between years 2000 – 2020; and b. monthly means.

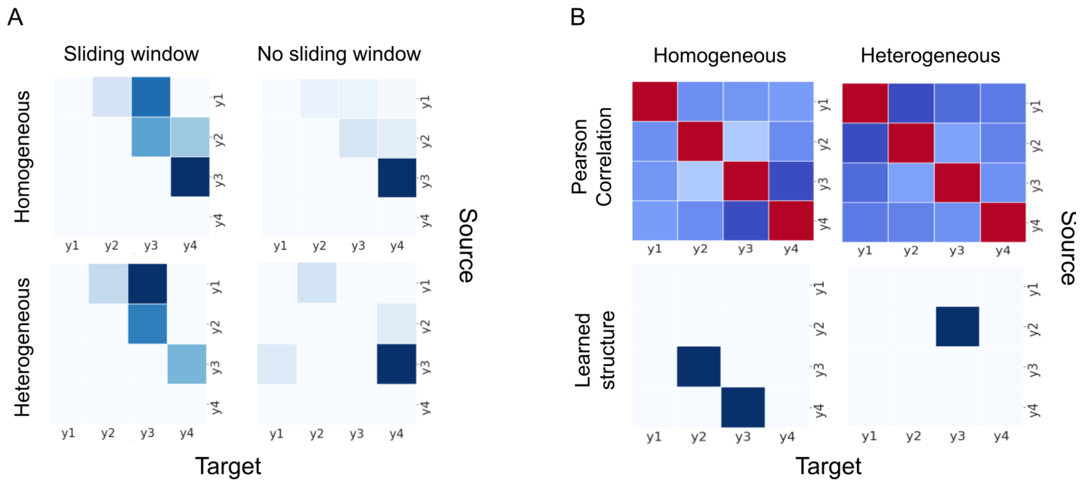


Figure 4 – Simulation results calculated by the proposed framework: (a) homogeneous and heterogeneous data sets CCM were calculated using sliding-window and without sliding-window (color scale shows CCM prediction skill. Light blue-weaker, Dark blue-stronger); and (b) Pearson correlations (color scale shows correlation score. Blues-negative. White-zero. Reds-positive) and Bayesian structure learning results suggest different interactions.

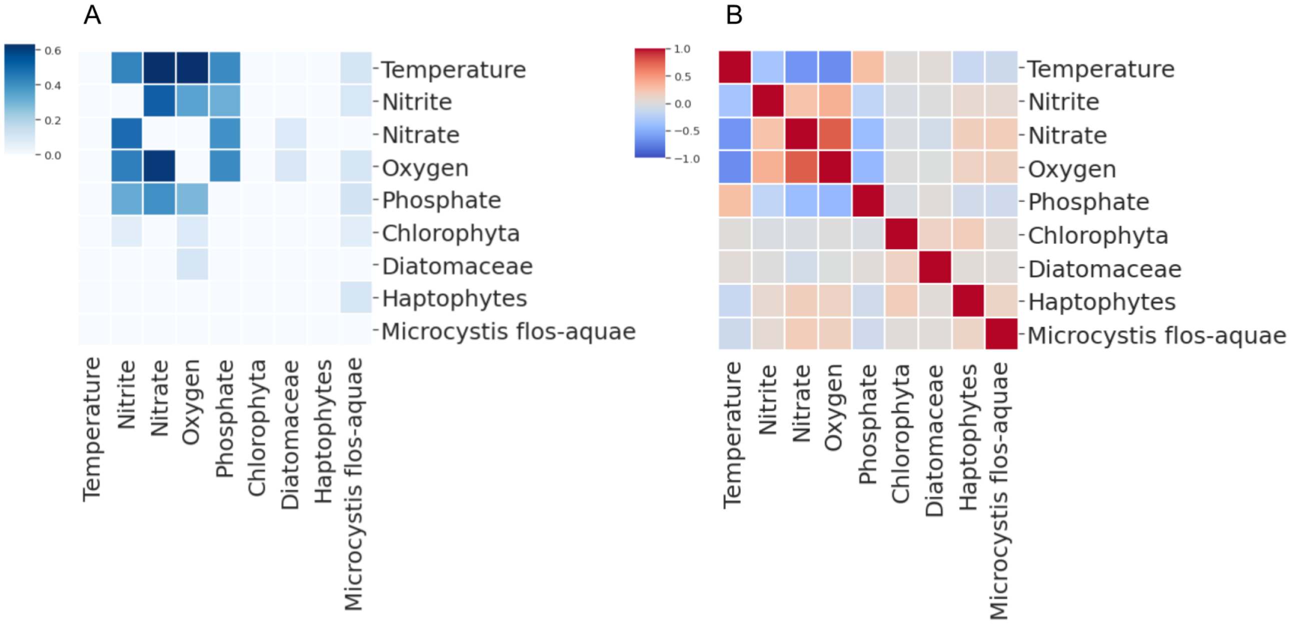


Figure 5 – (a) CCM and ECCM interactions were calculated from historical records of the years 2000-2020, as described in Methods (Figure 2, steps 1-4) (Color scale shows CCM score in the range of 0 (light blue) to 1 (dark blue)) (b) Correlations between the same causal variables (Color scale shows Pearson correlations between -1 (blue) to 1 (red)).

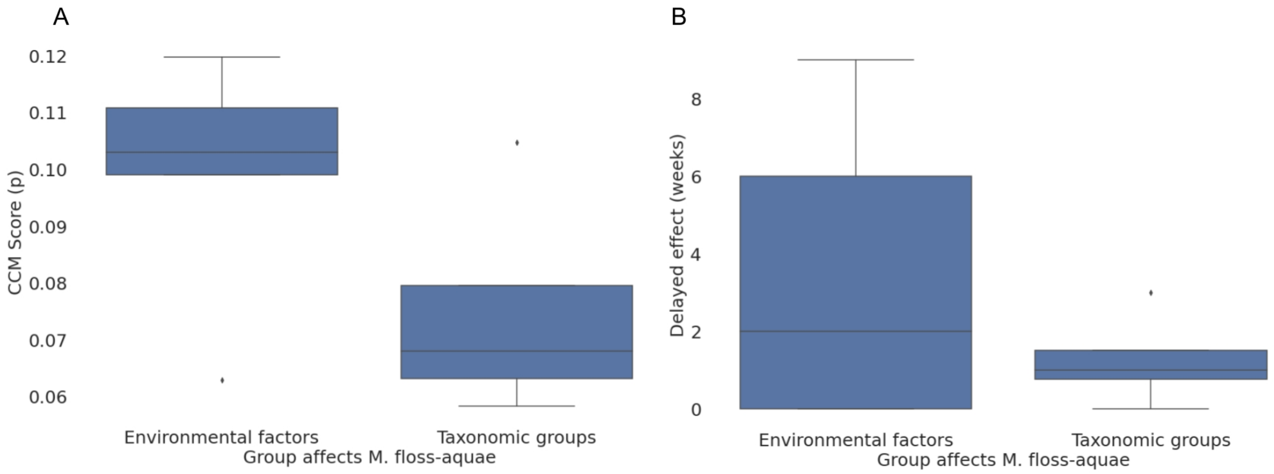


Figure 6 – Boxplots of the CCM prediction skill (a) results and delayed effect values (weeks) calculated from ECCM results. The boxplots are categorized to biological (n=4) and environmental (n=5) components.

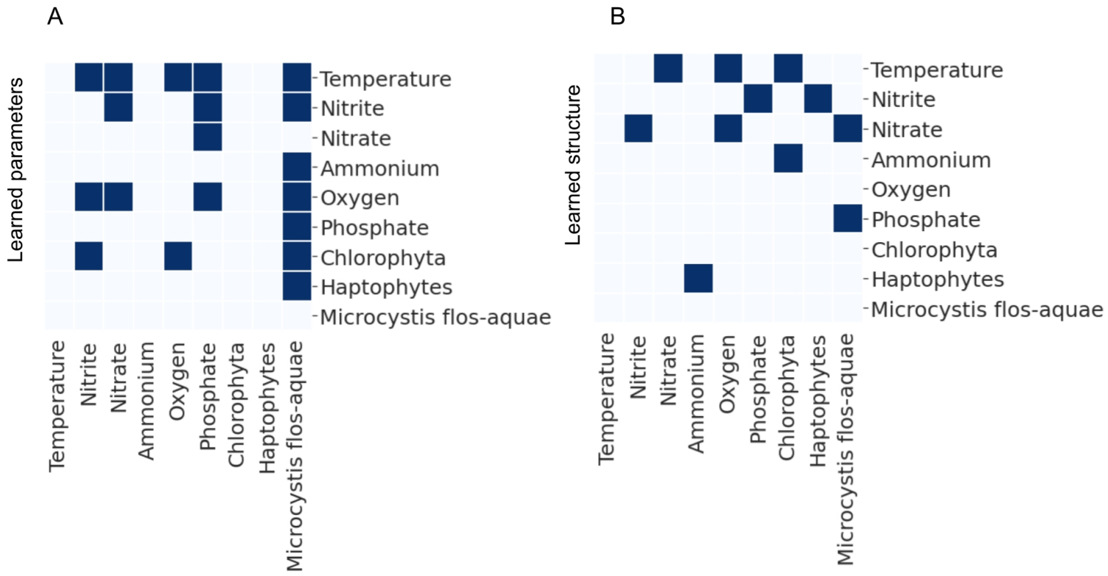


Figure 7 – Causal interactions identified by (a) CCM - ECCM approach, feedback interactions were removed and (b) structure learning. Y-axis is the causal parameter, and x-axis is the affected parameter.

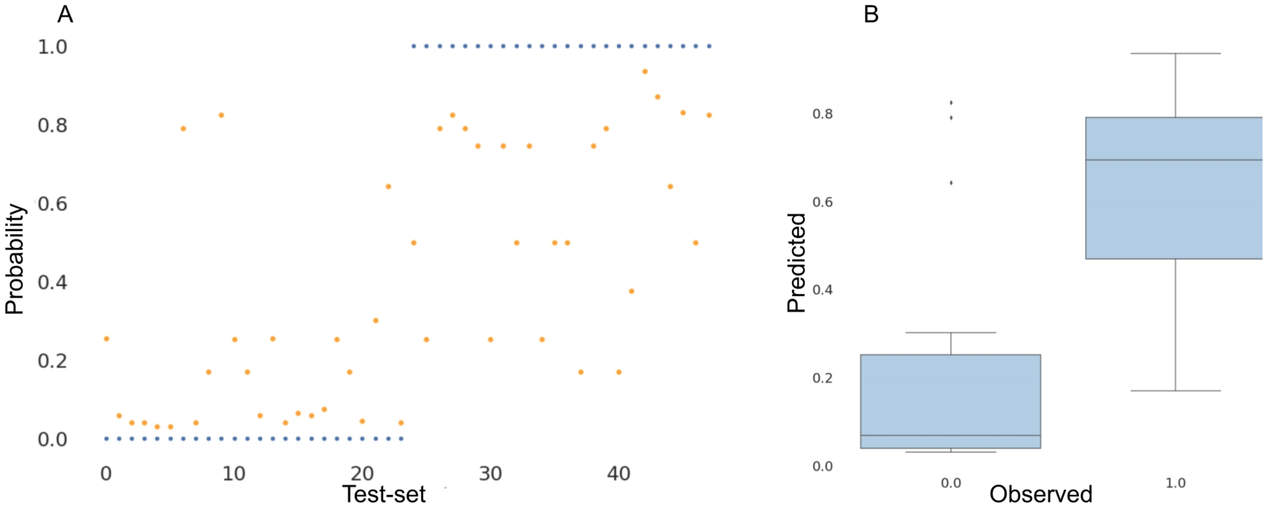


Figure 8 – Model validation. Bayesian network model was trained using the training dataset, and validated using the testing dataset. Probability < 0.5 considered ‘0’, and probability >= 0.5 considered ‘1’. (a) Observed (blue) and predicted probabilities (orange) of *M. flos-aquae* bloom formation. (b) Box plot of observed versus predicted probabilities of *M. flos-aquae* bloom formation.

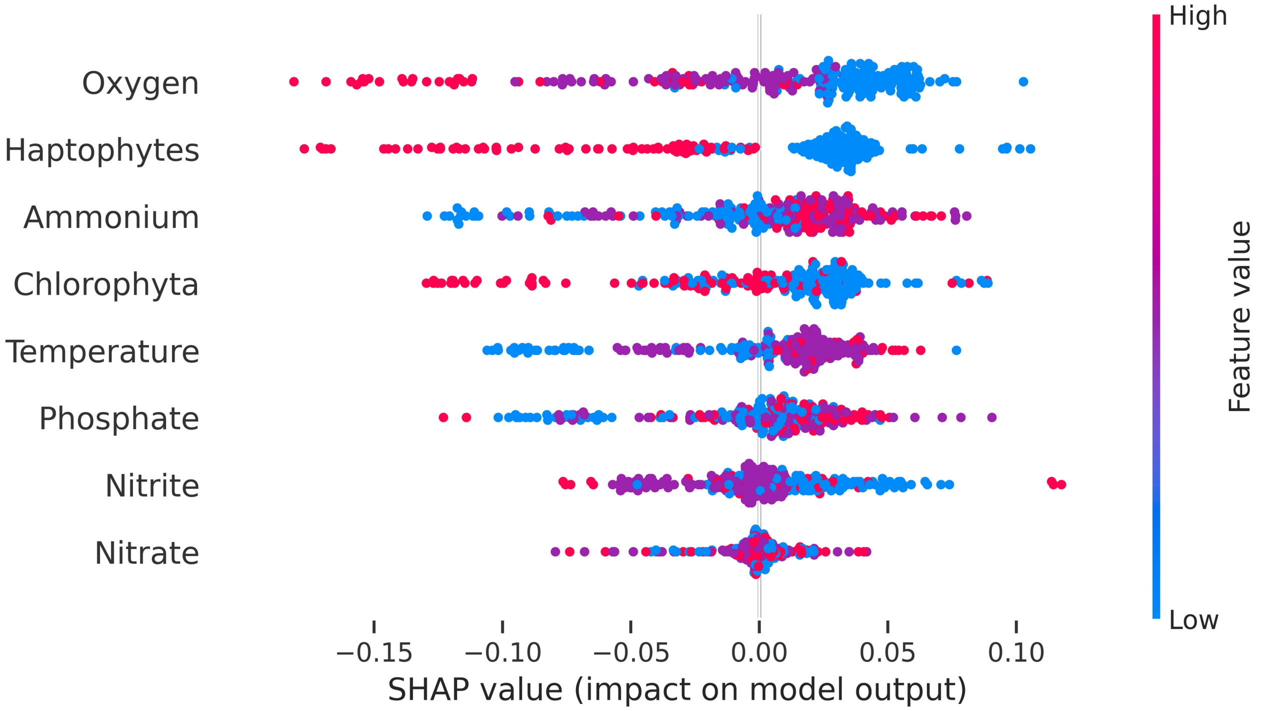
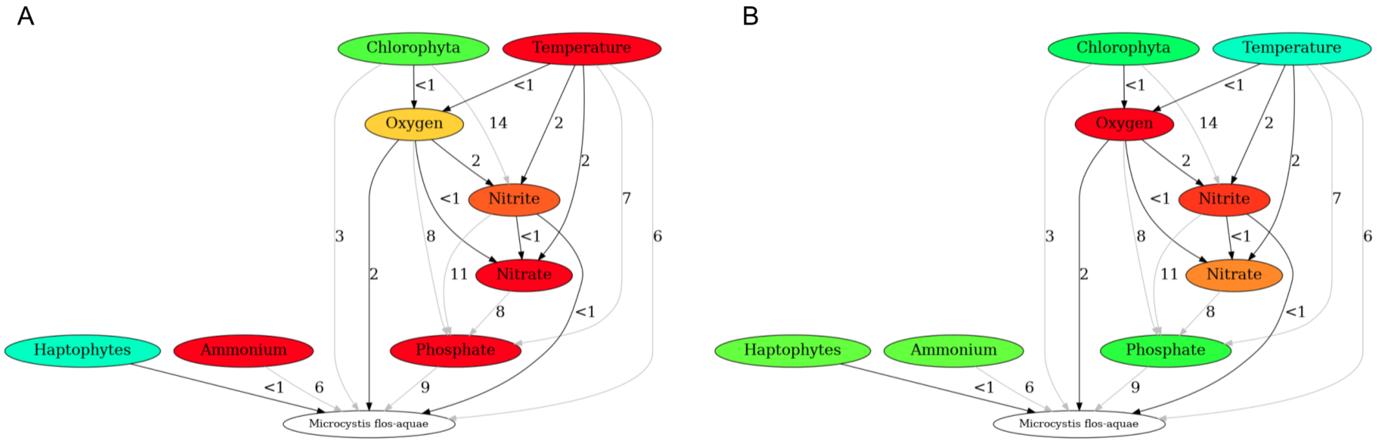


Figure 9 – Sensitivity analysis using SHAP. The beeswarm plot display the summary of how the different features impact the model's output. Each values of each permutation is represented by a single dot on each feature row. The X position of the dot is determined by the SHAP value of that feature. Color is used to display the original value of a feature. Features are sorted based on their impact on the BN model, with the most influential features at the top.



Mean maximization scenario

Mean minimization scenario

Figure 10 – Permutations mean scenarios that a. maximize blooming probabilities (probability > 0.5, n=272); and b. minimize blooming probabilities (probability < 0.5, n=104). (Cyan color indicates low meanvalue, red color indicates high mean value. e.g. red>orange>yellow>green>cyan). Numbers adjacent to the edges are the delayed effect in weeks. Black edges are of immediate effects (<= 2 weeks), suggested as direct interactions; Gray edges are of long-term effects (> 2 weeks), suggested as indirect interactions.

Supplementary – Figures and Tables

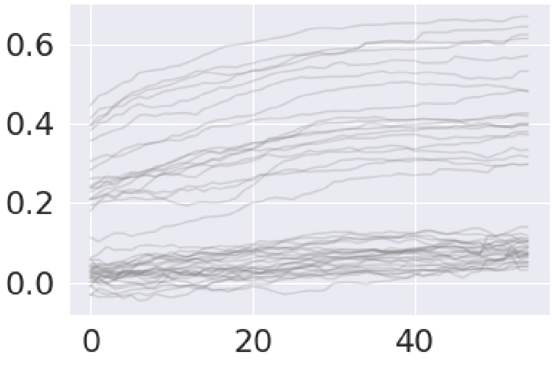


Figure S1 – Results of CCM converged prediction skill cases.

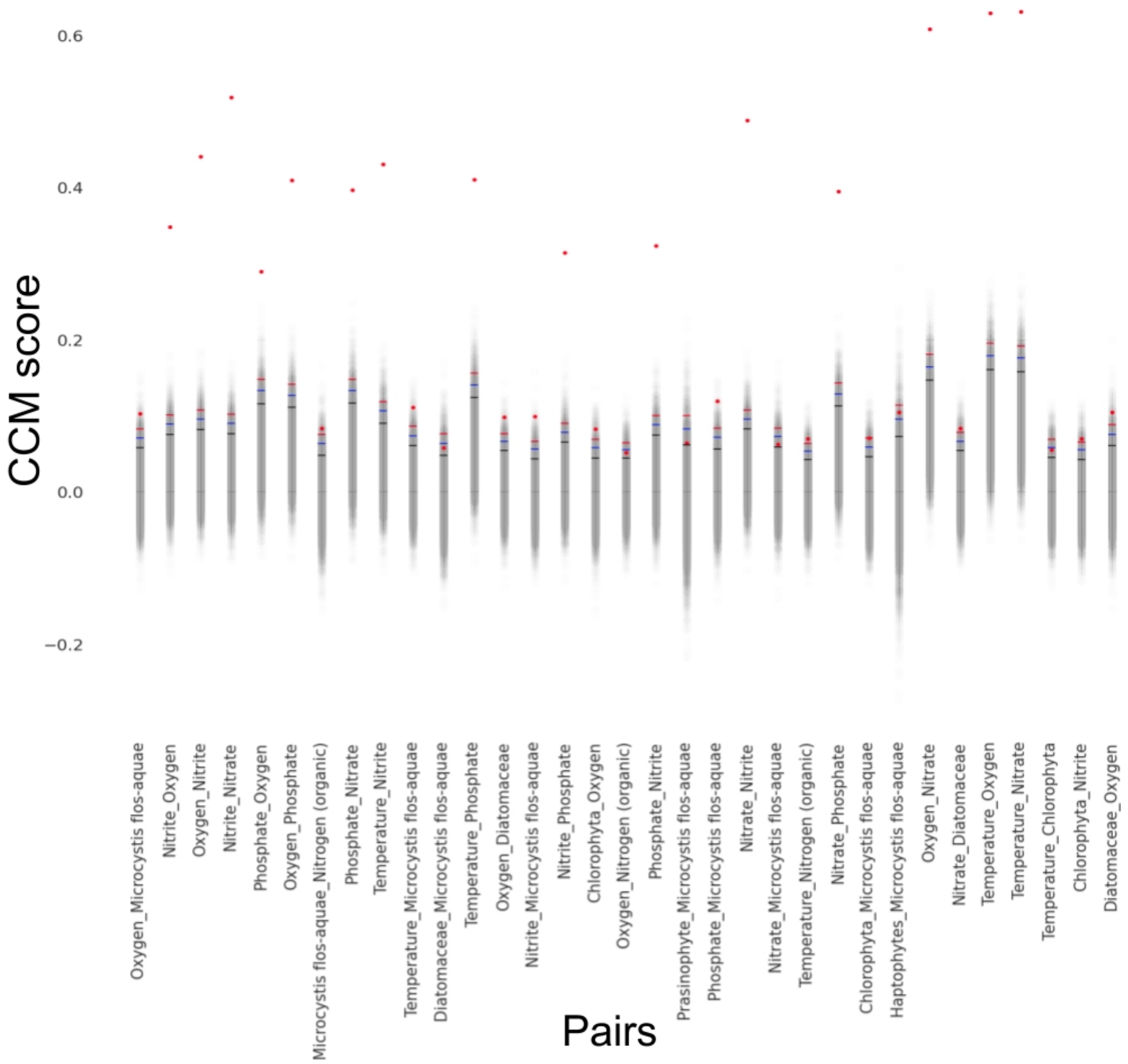


Figure S2 – Strip plot of the prediction-skill results that were calculated from 10,000 surrogate time-series between each of the pairs (gray). Hyphen sings represent the 0.9 (lower, black), 0.95 (middle, blue) and 0.975 (upper, red) quantile of the 10,000 surrogate CCM results. Red dots are the ‘true’ CCM values which were calculated from the time-series.

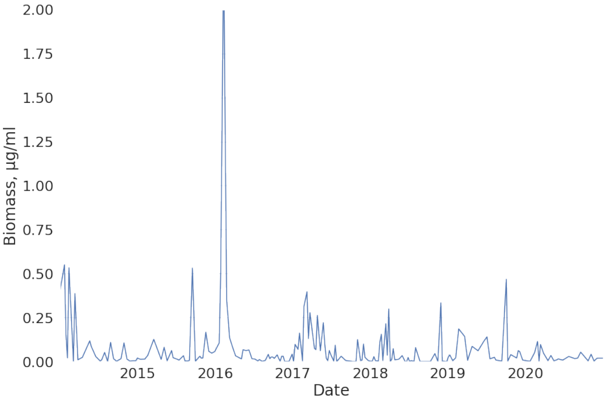


Figure S3 – *Microcystis* values between years 2014-2020, weekly, interpolated.

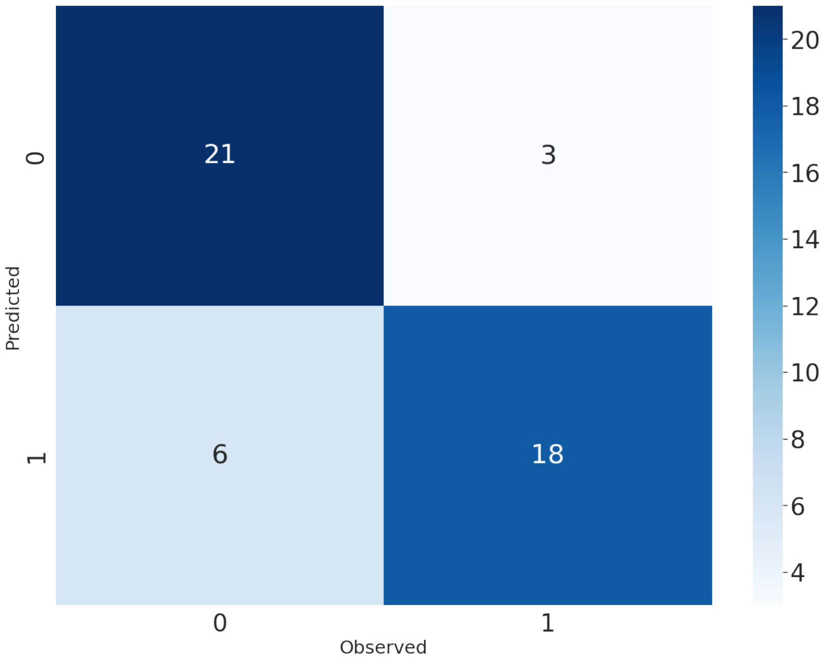


Figure S4 – Confusion matrix. X axis, observed; Y axis, predicted.

Table S1 – Categorization cutoffs.

Variable	**Unit**	**Categories and cutoffs**
Nitrite	mg/l	quantile 0.4/quantile 0.9, '0'/'1'/'2'
Nitrate	mg/l	quantile 0.4/quantile 0.9, '0'/'1'/'2'
Ammonium	mg/l	quantile 0.3/quantile 0.85, '0'/'1'/'2'
Oxygen	mg/l	quantile 0.55/quantile 0.75, '0'/'1'/'2'
Organic nitrogen (particulate)	mg/l	quantile 0.3/quantile 0.75, '0'/'1'/'2'
Organic nitrogen	mg/l	quantile 0.3/quantile 0.75, '0'/'1'/'2'
Chloride (Cl)	mg/l	quantile 0.3/quantile 0.75, '0'/'1'/'2'
Organic nitrogen (disolved)	mg/l	quantile 0.3/quantile 0.75, '0'/'1'/'2'
Phosphate	mg/l	quantile 0.5/quantile 0.75, '0'/'1'/'2'
Turbidity	NTU	quantile 0.3/quantile 0.75, '0'/'1'/'2'
pH	logaritmic units	8.25/8.45, '0'/'1'/'2'
Prasinophyte	Biomass, µg/ml	quantile 0.75, '0'/'1'
Chlorophyta	Biomass, µg/ml	quantile 0.75, '0'/'1'
Diatomaceae	Biomass, µg/ml	quantile 0.75, '0'/'1'
Dinoflagellate	Biomass, µg/ml	quantile 0.75, '0'/'1'
Cyanobacteria	Biomass, µg/ml	quantile 0.75, '0'/'1'
Haptophytes	Biomass, µg/ml	quantile 0.75, '0'/'1'
Cryptophytes	Biomass, µg/ml	quantile 0.75, '0'/'1'
Temperature	Celsius	18.5/21.5, '0'/'1'/'2'
Inflow	Weekly mean of L^3/sec	quantile 0.3/quantile 0.75, '0'/'1'/'2'
Microcystis flos-aquae	mg/l	quantile 0.75, '0'/'1'

Table S2 – CCM - ECCM results.

x1	**x2**	**Prediction skill**	**is_Valid**	**Delayed effect (weeks)**
Temperature	Oxygen	0.6290153436	1	0
Temperature	*Microcystis flos-aquae*	0.1107374773	1	6
Temperature	Nitrite	0.4304891828	1	2
Temperature	Chlorophyta	0.05535191585	1	19
Temperature	Nitrate	0.6315040961	1	2
Temperature	Phosphate	0.4107513852	1	7
Temperature	Organic nitrogen (particulate)	0.07022176498	1	6
Oxygen	*Microcystis flos-aquae*	0.102973285	1	2
Oxygen	Nitrite	0.4404129888	1	2
Oxygen	Nitrate	0.6085768545	1	0
Oxygen	Phosphate	0.4095347803	1	8
Oxygen	Organic nitrogen (particulate)	0.05177181098	1	2
Oxygen	Diatomaceae	0.098132401	1	10
Microcystis flos-aquae	Organic nitrogen (particulate)	0.08366328284	1	1
Nitrite	Oxygen	0.3481733729	1	0
Nitrite	*Microcystis flos-aquae*	0.09900214412	1	0
Nitrite	Nitrate	0.5189754535	1	0
Nitrite	Phosphate	0.3148284607	1	11
Chlorophyta	Oxygen	0.08254788438	1	0
Chlorophyta	Nitrite	0.07004027086	1	14
Haptophytes	Phosphate	0.07476616031	0	
Nitrate	Oxygen	0.6719748037	0	
Nitrate	*Microcystis flos-aquae*	0.0629061588	1	0
Nitrate	Nitrite	0.488140103	1	0
Nitrate	Phosphate	0.3952869711	1	8
Nitrate	Diatomaceae	0.08335491777	1	12
Phosphate	Oxygen	0.2897313105	1	17
Phosphate	*Microcystis flos-aquae*	0.1196515027	1	9
Phosphate	Nitrite	0.3234634705	1	20
Phosphate	Haptophytes	0.05701362075	0	
Phosphate	Nitrate	0.3972019158	1	0
Organic nitrogen (particulate)	Phosphate	0.05714645571	0	
Diatomaceae	Oxygen	0.1044963754	1	1
Diatomaceae	Nitrite	0.05068454162	0	
Diatomaceae	Phosphate	0.09336061053	0	
Chlorophyta	*Microcystis flos-aquae*	0.07108051147	1	3
Prasinophyte	*Microcystis flos-aquae*	0.06482822008	1	1
Haptophytes	*Microcystis flos-aquae*	0.1047188031	1	0
Diatomaceae	*Microcystis flos-aquae*	0.05837271527	1	1

Table S3 – BN model results using interactions above different surrogate quantile cutoffs (0.9, 0.95 and 0.975).

Quantile	**Number of nodes in DAG**	**Number of interactions in DAG**	**Accuracy**	**AUC**
0.9	11	22	0.812	0.826
0.95	9	19	0.812	0.817
0.975	8	18	0.791	0.8