CLEARWATER
ANALYTICA

FORECASTING HARMFUL ALGAL BLOOMS FOR ADAPTIVE WATER RESOURCES MANAGEMENT

WHITE PAPER

Original version: January 2020

This version: October, 2021

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The ideas in this paper are a collaborative result of many discussions with many wonderful minds that have been working on water quality issues for many years. The ClearWater team would like to say "thank you" especially to the following people, who added considerable value: Devin Doring, Nitin Joshi, and Todd McDonnell and finally to all those collaborators that have given us data, helped us with ecological theory and the development of our machine learning algorithms, and provided us vital knowledge about freshwater systems.

AUTHORS

James R. Watson, CEO, Clearwater Analytica LLC
Mathew Titus, CTO, Clearwater Analytica LLC

KEYWORDS

Harmful algal bloom, Water quality, Prediction, Machine Learning, AI, Big Data, Bayesian Modeling, Neural Networks, Data Assimilation, Decision Support

ABOUT THIS DOCUMENT

In 2018 the City of Salem in Oregon, USA, suffered numerous toxic algal bloom events in their sole source of drinking water – Detroit Lake. Algal toxins were found in their water treatment facility and citizens were advised to not drink tap-water for a month. To make sure this would never happen again, the City of Salem partnered with The Prediction Lab LLC to develop a harmful algal bloom prediction system. The Prediction Lab is a data science company that was founded by James Watson and Mathew Titus, who met at Oregon State University where James is a Professor of Environmental Science. Through an iterative process, The Prediction Lab developed a HAB prediction system for the City of Salem. This was able to provide the City of Salem accurate forecasts of HABs in Detroit Lake, and it has been used by the City of Salem since 2018 to manage their source waters. This success led to others: James and Mat were awarded a grant from the Water Research Foundation to deploy their harmful algal bloom forecasting system in 30 other water systems spread across the USA. At this point, James and Mat started ClearWater Analytica LLC as the successor to The Prediction Lab.

Clearwater Analytica is a public benefit company whose mission is to is to help achieve universal and equitable access to safe drinking water for all people.

This document provides an overview of the water quality prediction system deployed at Detroit Lake in 2018. Since then, ClearWater Analytica has developed other machine learning approaches for predicting harmful algal blooms, and the overall prediction system (including data ingestion, QCQA and visualization) has also been updated. Details about these techniques can be found in other whitepapers.

REDUCED WATER QUALITY AS A THREAT

Poor water quality in lakes, reservoirs, and rivers is a significant and growing threat for water utilities, food and agricultural industries, and the public. In particular, Harmful Algal Blooms (HABs), which produce cyanotoxins, can cause direct harm to people and animals. Decades of research on the world's lakes, coasts and estuaries have revealed immense complexity in the conditions that promote harmful algal bloom development and the diversity of HAB species. This has created an extreme challenge to their prediction, mitigation, and removal. Discussion of HABs in the scientific literature has focused on the disruptive or even "catastrophic" nature of high toxicity and/or high-biomass blooms. However, the caveat is often made that such blooms are not new or unnatural phenomena, and they have long been part of a region's local ecology, primary productivity, and important biogeochemical cycling. That said, there is increasing recognition that the effects of HABs on public health, marine and freshwater ecosystems, and economies are worsening, and that new solutions are required.

Central to new opportunities for managing HABs are the Internet of Things revolution, Big and Small Data, and novel machine learning approaches for prediction.

The list of potential impacts from HABs include (1) the production of dangerous phycotoxins that enter food webs, the atmosphere (if aerosolized) and the potential contamination of water supplies in freshwater reservoirs; (2) the depletion of dissolved oxygen and/or the smothering of benthic biota as algal biomass decays; and (3) diminished fisheries through physical damage to fish gill tissue. Mirroring the diversity of HAB impacts, a major struggle in the study and management of HABs has been the sheer breadth of species, life histories and ecosystems involved. The algae that are categorized as potentially harmful do not belong to a single, evolutionarily distinct group. Rather, they span most algal taxonomic clades, including eukaryotic protists, for example dinoflagellates, raphidophytes and diatoms, euglenophytes, cryptophytes, haptophytes, and chlorophytes and microbial prokaryotes, and most commonly the ubiquitous, sometimes nitrogen-fixing cyanobacteria that occur in both marine and freshwater systems. Increasing prevalence of HABs, caused by this long list of microbial organisms, include eutrophication, climate change, ballast water dispersal.

RESILIENCE THROUGH SMART TECHNOLOGIES

Predicting HABs – State of the Industry

The number of approaches for monitoring, detecting, and forecasting the onset, fate, and demise of algal blooms is arguably comparable to the diversity of harmful algal species being studied. Over the past two decades, there has been an increasing desire to use big data methods from machine learning to develop early-warning systems that will alert managers and communities of impending danger. An ideal early-warning system provides quantitative predictions of HAB likelihood, intensity, and movement through a waterbody. These machine learning approaches rely on data from a range of platforms from space-based, airborne, and in-water sensors to traditional environmental sampling. In general, for a HAB to occur, there needs to be light and food (i.e., nutrients), and it needs to be warm.

However, the devil is in the details, as not all algal blooms occur when these conditions are met, and it is especially difficult to know and predict when these algae will release toxins into the water. To go beyond these simple rules of thumb and create quantitatively accurate predictions of HABs, additional data and computational methods are required. HAB risk quantification and prediction is a wicked problem for several reasons. There are regional differences in the (often hidden) drivers of HABs, the volume, velocity, and quality of water quality data (both historical and real-time) vary dramatically from place to place, and the tools that are used to make inferences about HAB risk are many.

Quantifying HAB Risk

Numerous modeling approaches have been developed to quantify the risk of HABs for a given waterbody. These studies have emphasized that great differences in the drivers of HABs exist among waterbodies. This means that a model (statistical, phenomenological, or data-driven in nature) trained in one place will not necessarily perform well at predicting HABs in another place. The uniqueness of different waterbodies presents a challenge to scaling-up the application of HAB risk models, and it is one of the main reasons for the absence of regional and national HAB prediction systems. Another important challenge is the disparate nature of data collected for waterbodies, in terms of both the variables collected and the length of historical records.

Early advances to using water quality data to quantify the risk of HABs, and to predict their occurrence, used primarily linear and non-linear statistical approaches. These approaches balanced predictive skill with learning–that is, these methods helped create a deeper understanding of the processes involved in the creation of a HAB for a given water body. This understanding then allows for the creation of heuristic rules by which to quantify and manage HAB risk. For example, in addition to statistical approaches, relatively complicated hydrodynamic models linked to water quality models can be used to simulate the occurrence of HABs in a given waterbody.

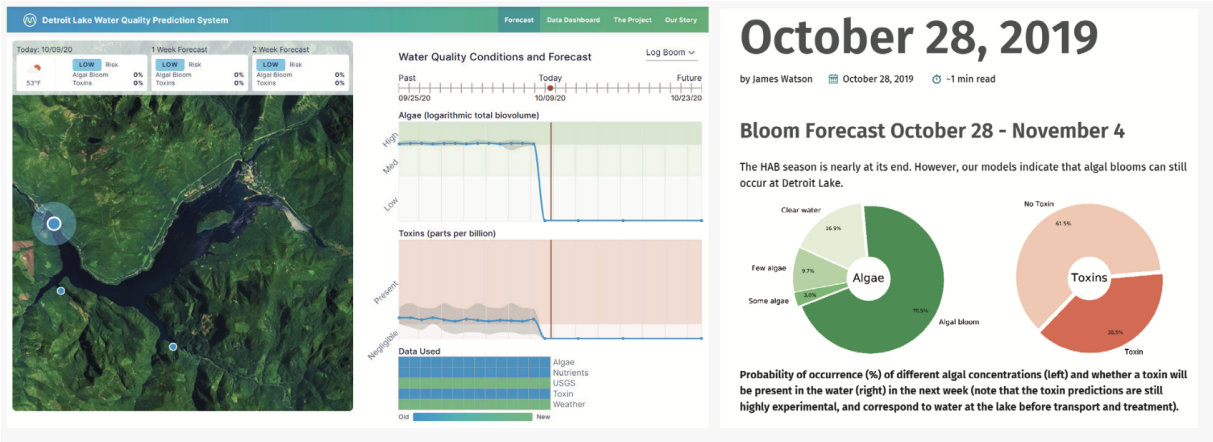

Figure 1. Left) *A screenshot of the landing page of the HAB prediction web-app developed for the City of Salem in 2018. This web-app is powered by machine learning predictions of HABs in Detroit Lake, based upon real-time data feeds, developed for the 2020 HAB season. Forecasts (and a data dashboard for visualizing historical data) are used by the City of Salem public works staff for making decisions about treatment plant staffing and the frequency of source water monitoring efforts. Right) A screenshot of the landing page of the prototype HAB prediction system developed for the City of Salem in 2018/2019. This first-of-its-kind HAB prediction system laid the groundwork for continued development of quantitative HAB risk assessment and prediction by ClearWater.*

Case-Study: The City of Salem HAB Prediction System

Our team previously developed a HAB risk assessment and prediction algorithms for use as a web-app by water quality stakeholders. Specifically, in 2018 the City of Salem, Oregon suffered numerous HABs in the sole source of their drinking water–Detroit Lake. This led to a communications catastrophe with their citizenry and a subsequent loss of trust in the City's ability to manage Detroit Lake. To overcome this, the City of Salem asked us to develop an HAB prediction system for Detroit Lake. First, a prototype web-app was developed, based on the use of Bayesian Model Averaging. Over the 2019 HAB season, the City of Salem used this model, with results posted to a weekly blog (link), to target source water monitoring and to communicate more transparently with the public. For the 2020 HAB season, this prototype was further developed into an operational data dashboard, for use by City of Salem staff (link). See Figure 1 for screenshots of these HAB prediction web-apps. Lessons learned from this collaboration with the City of Salem have laid the foundation for the next steps for advancing our HAB forecasting system, namely, (1) how to design a database and algorithm front-end architecture for effective use as an operational HAB risk assessment and prediction system, (2) the challenges of making accurate HAB predictions, and (3) the limitations of standard approaches securing resilient and accurate HAB forecasts, and for communicating understanding about the drivers of HABs.

The City of Salem collects a wide range of data. These data reveal the complexity and dynamics of freshwater ecosystems:

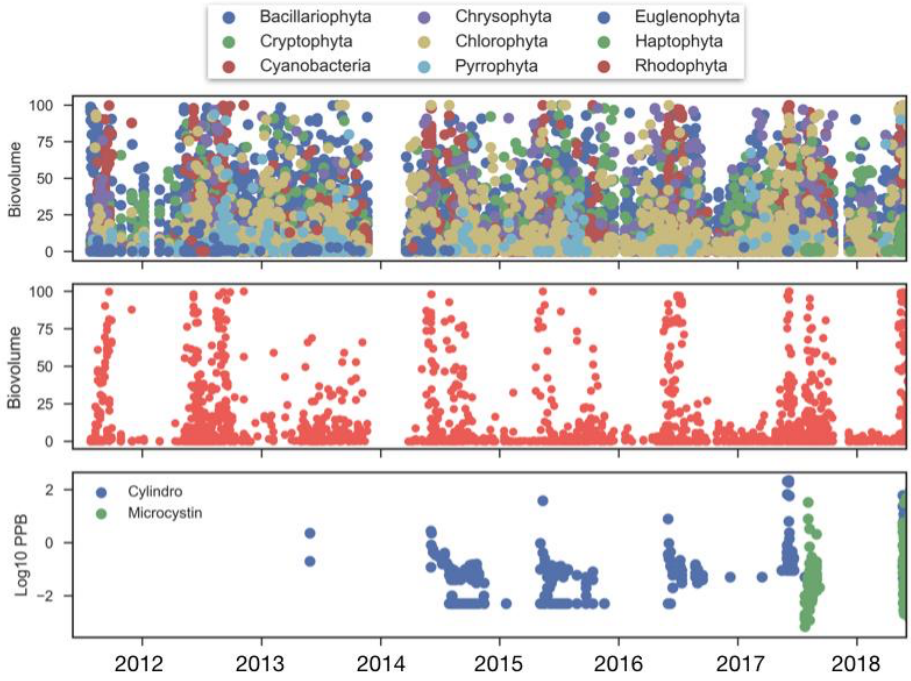

***Figure 2**. Time series of data collected for Detroit Lake including the biovolume (the volume of cells in a unit amount of water expressed as a %) of major microbial families (**top**; the complexity of the lake system is evidenced by the overlapping time-series), the biovolume of cyanobacteria, the main HAB causing microorganism in Detroit Lake (**middle**) and toxin concentrations in parts per billion (**bottom**: note that these toxin data were only collected systematically after 2014).*

ClearWater Analytica's goal was to leverage large volumes of data gathered from Detroit Lake and make weekly predictions of HABs. The data that we used include those listed in Table 1 below.

Time Series	Unit	Time Period	Sampling Frequency
Cylindrospermopsin	ppb	5/28/13 - 6/30/18	Sporadic
Microcystin	ppb	7/25/17 - 6/30/18	Sporadic
Algal Density	cells/mL	7/27/11 - 7/6/17	Weekly
Algal Biovolume	μmeters^3/mL	7/27/11 - 7/6/17	Weekly
Ammonia	mg/L	5/17/16 - 6/13/17	(Bi)weekly
NO3 + NO2	mg/L	1/20/16 - 4/25/18	(Bi)weekly
O-Phos	mg/L	1/20/16 - 4/25/18	(Bi)weekly
TN	mg/L	1/20/16 - 4/25/18	(Bi)weekly
T-Phos	mg/L	1/20/16 - 4/25/18	(Bi)weekly
Rain Accumulation	in	1/1/09 - present	Hourly
Solar Radiation	Langleys (cumulative)	1/1/09 - present	Hourly
Barometric Pressure	inHg	1/1/09 - present	Hourly
Wind Direction	degrees	1/1/09 - present	Hourly
Wind Speed / Gusts	mph	1/1/09 - present	Hourly
Temperature	°F	1/1/09 - present	Hourly
Humidity	%	1/1/09 - present	Hourly
Dew point	°F	6/21/18 - 7/16/18	Hourly
Sentinel 2 (11 bands)	512-by-332 pixel image	1/3/18 - 10/3/18	5 days (sporadic)
LandSat 8	512-by-332 pixel image	2/20/17 - 9/28/18	8 days (sporadic)

Table 1. *List of data currently collected for Detroit Lake, and in future data (i.e. satellite data collected from Sentinel 2 and Landsat 8).*

In recent HAB seasons, the City of Salem installed and continues to maintain a YSI vertical profiler, from which continuous measurements are obtained. ClearWater Analytica makes use of these data for making HAB predictions, and for presenting these continuous data in clear and compelling ways:

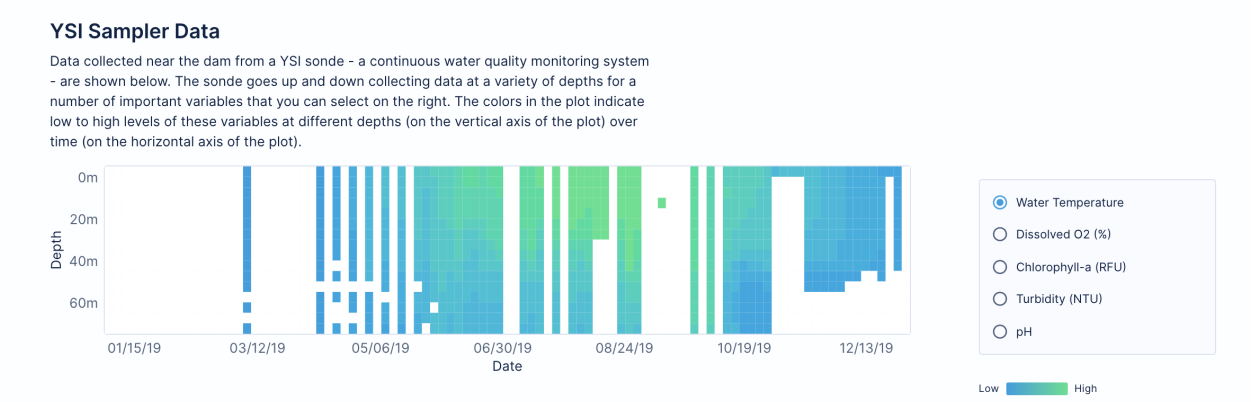

Figure 3. *The YSI vertical profiler at Detroit Lake provides continuous data for a number of variables, which ClearWater makes use of in its predictions, and help visualize with our web- and mobile-apps (the above figure is a screen shot from our app: https://detroitlake-staging.thepredictionlab.com/dashboard).*

The Value of a HAB Prediction System

The City of Salem has benefited from our HAB prediction system in the following ways:

1) HAB predictions have helped city officials avoid follow-on emergency events. This has reduced costs: because of the 2018 HABs, the City of Salem has spent approximately $80 million to improve its resiliency to HABs. If HAB predictions were available in 2018, these costs would likely have been reduced.
2) HAB predictions inform efficient source water monitoring. Employing staff to take water samples from multiple locations in a water body is extremely costly. Having predictive insights to augment decision making about when and where to test can reduce costs.
3) The HAB prediction system, including our front-end web- and mobile applications helps to build trust with the public by displaying metrics of water quality and other environmental factors to better connect them to their watershed.

The Limits to Blackbox Machine Learning

Recent advances in machine learning have improved our ability to predict the onset, duration and intensity of HABs in waterbodies across the world. This step-up in predictive skill has given water quality stakeholders an improved ability to manage their water, through avoiding extreme events and targeted source water monitoring. However, as is true of machine learning in general, there are two main challenges associated with water quality machine learning algorithms: 1) they are data-hungry and 2) they are black boxes, with a limited ability to improve understanding of the processes governing the phenomenon of interest, in this case HABs. In response to these challenges, our development of our HAB forecasting system has been cognizant of:

- Basic models are too simplistic; they don't factor regional differences in HAB drivers; therefore, they have too narrow an application
- There exists a disparate and variable nature of available data (i.e., this is a "lots of small data" problem, and not a "big data" problem).

- Existing machine learning algorithms are (1) data hungry, and (2) potentially black boxes, which limit the user's ability to understand the fundamental processes governing HAB phenomena.

SMALL OR BIG, DIRTY OR CLEAN: ALL DATA

ClearWater Analytica makes use of a wide variety of data sources (see Table 1 for a list of all the data presently used): from weekly water sample data (e.g., nutrient, algae and toxin concentrations), to continuous weather station data (air temperature, precipitation, etc.), to intermittent satellite imagery (specifically, Sentinel 2 and Landsat 8 reflectance data). ClearWater can leverage small data sets (e.g., excel spreadsheets), as well as Big Data, e.g., terabytes of images and time series stored on servers. ClearWater also overcomes a critical challenge in the water quality space in terms of the disparate formatting and standards used for environmental data. For each freshwater system that ClearWater Analytica applies its harmful algal bloom prediction system to, all relevant data streams are connected to a PostgreSQL database hosted in the cloud. Strict data QC/QA is performed and, where possible, APIs are leveraged to provide automated updates to the database. In general, the data the ClearWater makes use of includes:

Data Type	**Variable**	**Spatial coverage**	**Temporal frequency**	**Used in ClearWater?**
Water sample	Nutrient, algal, and toxin concentrations	Low	Weekly	Yes
Weather station	Air temperature, precipitation, barometric pressure, peak wind speed, wind direction, relative humidity, solar irradiance	Intermediate	Hourly	Yes
Stream gauge	Flow rate, water temperature, water level	Intermediate	Minutes	Yes
Vertical profiler	Water temperature, dissolved oxygen, chlorophyll-a, turbidity and pH.	Low	Minutes	Yes
Satellite imagery	Landsat 8, Sentinel 2b spectral band reflectance	High	Weekly	Yes
Contextual information	Catchment type, elevation,	NA	NA	Yes

Table 1. *The variety of data that ClearWater uses to make predictions of changes in the concentration of harmful algae. In essence, ClearWater makes use of all available data: from small to big datasets, and those that are cleaned and those that are dirty, in terms of data format and standards.*

ACTIONABLE INSIGHTS FROM MACHINE LEARNING

To overcome the challenges stated above ClearWater Analytica developed a water quality prediction system that is:

1) **Flexible**: able to be implemented quickly for different waterbodies
2) **Resilient**: able to continue to make predictions even as data dropouts occur
3) **Accurate**: making predictions that are useful for informing decisions
4) **Transferable**: able to inform HAB risk in other waterbodies that do not have large amounts of historical data.
5) **Understandable**: able to advance learning about a given system, in terms of the main drivers of HABs

The approach is based on foundational academic work on the application of Bayesian Model Averaging (BMA) for predicting changes in the concentration of cyanobacteria and cyanotoxins. We have developed a new implementation of the BMA approach, one that leverages machine learning. We have embedded this new HAB prediction algorithm into a holistic water quality prediction system which includes a streamlined data backend, where a wide range of data are automatically synthesized and cleaned, and a compelling front-end, where results of the analysis are presented in real-time to users managing the quality of their source waters.

Machine and Transfer Learning Algorithms for Harmful Algae Prediction

There exist many methods for predicting harmful algal blooms and changes in water quality in general. We make use of a broad range of techniques: Econometric approaches, Generalized Additive Modeling, Bayesian Model Averaging, Neural Networks, and Random Forests to name a few. Each approach has their pros and cons in terms or resilience, accuracy, flexibility, transferability and understandability. In addition to these core classes of algorithms, we develop new approaches to solve specific problems. For example, new approaches to Transfer Learning to scale predictions up. Below, we describe in detail the use of Bayesian Model Averaging for predicting HABs in source waters.

ClearWater's HAB prediction system leverages the Bayesian Model Averaging approach because it satisfies the main needs of water quality stakeholders: flexibility, resilience, accuracy, transferability, and understandability. Despite the considerable advantages that predictive algal bloom models may have for water quality management, it is important to recognize the need to acknowledge uncertainty in any modeling approaches. Models have a structure, including the parameters that are used in the model and estimates of the parameters that are particular to that structure. If model predictions are incorrect, for instance because parameter estimates are wrong, this may prove costly in water quality management programs. All water quality stakeholders have informed us that a clear quantification of uncertainty in results from predictive models must be considered.

Together with parameter uncertainty, however, there is often also uncertainty regarding the selection of the models, in terms of their structure, to best explain observed responses. Typically, there are at least several, and often many models from which to select. In ecological studies it is still routine to assume that a single best model choice exists, and to proceed as though this choice were known to be

correct in making predictions. If the predictions from alternative plausible models are different, there are hazards in relying on a single model. This may lead to overconfident predictions, making management decisions based on these predictions riskier than might be supposed. Given that the scale of HAB impacts as well as management programs may be large, this creates a substantial problem for modelers and managers to ensure that all sources of uncertainty are adequately accounted for.

The Bayesian paradigm has been recognized as a useful framework for the effective management of ecological problems. Bayesian analysis also allows practitioners to sift through a multitude of possible predictive factors and relationships to determine which models are the most plausible given the observed data. Rather than ignoring model uncertainty in the search for a ‘‘best’’ model, a more satisfactory solution is to use BMA techniques, where a summary model is constructed by the combination of individual model results weighted by their degree of plausibility. By averaging over many different competing models BMA incorporates model uncertainty into conclusions about parameters and prediction.

Advances to Bayesian Model Averaging: Machine and Transfer Learning

Bayesian Model Averaging has been successfully applied to solve the HAB prediction problem previously. However, in these previous studies only linear models have been used. ClearWater takes another step and uses the Bayesian Model Averaging approach to make HAB predictions presented by a suite of Neural Networks, varying in network topology, as well as a host of linear and non-linear models. This is a much more powerful implementation of Bayesian Model Averaging, as it makes use of the latest machine learning models that have the best predictive skill. In addition to including machine learning into the Bayesian Model Averaging framework, ClearWater can produce predictions at source waters with little to no historical data. To do so, we have advanced the use of Transfer Learning in combination with Bayesian Model Averaging. Transfer Learning, as the name suggests, allows one to transfer learning from one place (i.e., a data-rich system) to another (i.e., a data-poor system). In other words, if we know that two fresh-water systems are similar in some ways (e.g. catchment type, size, history) we can use models created in one place to inform models generated at the other.

Training and Cross Validation

By accounting for model uncertainty BMA minimizes prediction risk and has also been shown to improve model prediction accuracy on average. A practical consideration in the use of a BMA strategy is the potentially large number of competing models in the posterior distribution (also known as the combined or averaged model). ClearWater Analytica’s HAB prediction system employs posterior predictive and calibration checks to ascertain the utility of each of these approaches as predictors of harmful algal occurrences under the provided data. Cross validation is a method which allows for the estimation of approximately unbiased prediction error/misclassification rates. The procedure involves splitting the original data set into training and test sets. The model is then fitted to the training set and predictions of the data in the test set are formed using this model. The predictions are compared to the test set and a summary of the accuracy is made. The HAB prediction system uses a ‘‘leave one out’’ or a k-fold cross validation procedure, where observations are sequentially excluded from the original

data set and predicted using the remaining training set. Results of each of the cross-validation procedures were summarized in a receiver operator characteristic (ROC) curve and other metrics of accuracy (e.g., mean-square error). ROC curves assess the predictive power of a model.

APPLICATION CASE-STUDY: DETROIT LAKE, OREGON

ClearWater Analytica's Bayesian Model Averaging approach was applied to predict cyanobacterial blooms in Detroit Lake, Oregon (USA) as part of our project with the City of Salem. Data from 2011 to 2019 were used to train neural networks for use in the BMA framework. Retrospective predictions were then made for the 2018 HAB season. Similarly, the prediction system was employed during the 2019 (using only linear models) and 2020 (using linear, non-linear and neural network models) HAB seasons. All data and predictions were recorded, allowing for the direct comparison of predictions with observations. For example, for the City of Salem, ClearWater provided a report on the utility of this approach, paying particular attention to the quantification of uncertainty, the accuracy of the predictions (see Figure below; we achieved HAB prediction accuracy of ~80%) and the identification of the drivers of HABs in Detroit Lake.

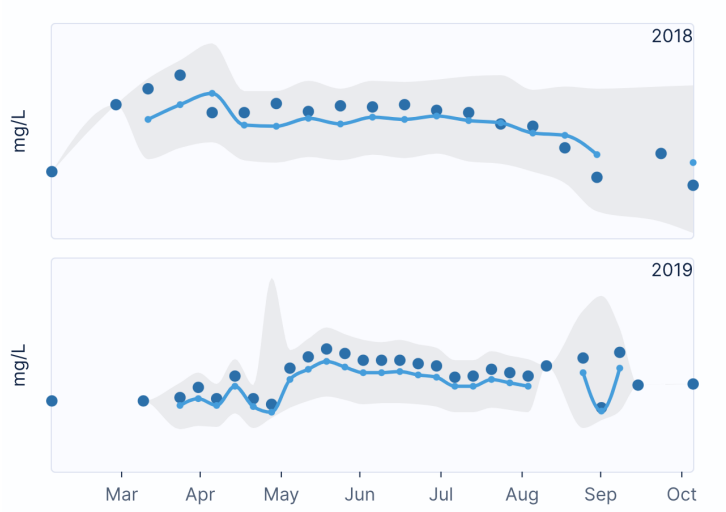

Figure 4. *Time series of predictions (blue line and greyed area) and observations (blue dots) for 2018 and 2019. In this early proof-of-concept our Bayesian Model Averaging predictions performed well at predicting HAB blooms, with an accuracy of ~80%.*

The Bayesian Model Averaging also allowed ClearWater Analytica to identify the most important variables contributing to predictive skill:

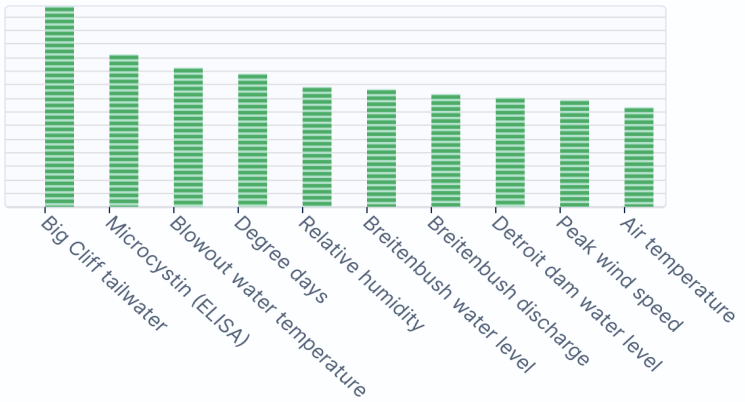

Figure 5. *Variables were ranked in terms of their importance to the accuracy of the HAB predictions produced by the Bayesian Model Averaging algorithm.*

FRONT-END DEVELOPMENT AND VISUALIZATION

The data that water stakeholders collect is varied in terms of its volume, veracity, and richness. As we have described above, it comes in a variety of formats and describes a wide range of water-body properties. However, on their own, these data are not valuable. To mine value from these data, it must be transformed into information. This is what the machine learning algorithms do by discovering relationships among the different data streams. To capitalize on the value of the information provided by the algorithms, one also needs an effective communication tool, able to present the (often complex) output of the machine learning algorithms to a broad audience: from utility staff working in water resources professionally to members of the public. To that end, the HAB results of the prediction algorithm are embedded in an interactive web and mobile application, where custom charts allow users to explore the data that has been collected, and the predictions that the machine learning algorithms make. Additional webpages describe the "story" of any given waterbody and summarize the product itself in terms of the overall accuracy of the algorithms and long-term changes in water quality dimensions. In addition, all the data collected and cleaned are made available to users through direct download.

THE FUTURE OF SMART WATER ANALYTICS

Advances to water quality prediction were made by incorporating machine learning models (i.e., neural nets and Random forests) into a Bayesian Modeling Averaging framework. This new prediction algorithm was embedded in a full-stack system, including both a data focused back-end for the synthesis and clearing of water quality data, and a web and mobile responsive front-end, where users can explore the data and predictions. This system was prototyped for Detroit Lake, the sole source of drinking water for the City of Salem in Oregon, USA, where data from a range of sources -- water samples, weather stations, vertical profilers -- are automatically synthesized and provided as inputs to the machine learning algorithms. Daily (for 2020) and weekly (for 2019) predictions were made at 3d, 7d and 14d forecast horizons. The prediction system performed well for the 2019 season (~80% accuracy), and the 2020 season. The prediction system was funded to continue operations into the 2021 HAB season. The application of Bayesian Modeling Averaging to make use of a suite of linear, non-linear regression-type models, as well as neural networks and random forests is an advance to the state

of the art in terms of harmful algal blooms prediction. This system is an example of "Robust AI", where predictions are guaranteed, regardless of data-dropouts.

It is important to recognize that there exist many other approaches to modeling changes in water quality, in terms of both phenomenological approaches like machine learning, where functions are fit to the data, and process-based models, where dynamical equations are used to model specific processes occurring in a given water body, from fluid dynamics to food-web interactions. For the latter, these mechanistic models have dominated harmful algal bloom prediction studies in coastal waters, and there are examples of "assimilation models" where data is used to drive model dynamics. However, there is a general sense that these process-based models are best used for exploring the sensitivity of system dynamics to perturbations (e.g., this is how we explore the impact of carbon emissions on climate change), and that for short-term prediction, data-driven methods like those created here perform better. Certainly, there is promise in a synthesis of both approaches: machine learning operating along-side process-based models of system dynamics.

It is important to note that there are various dimensions of success with regards to water quality prediction. In the case-study of Detroit Lake, the machine learning algorithms achieved a high level of accuracy when predicting changes in the concentration of algae. However, predicting changes in the occurrence of toxins proved to be more challenging. It is well known in the limnological literature, that there remain large knowledge gaps about when and why certain algal species will exude toxins. As more is learned about this process, from genetic studies, the machine learning algorithms will also have more and more data with which to learn from. One of the challenges of machine learning is that there needs to be a suitably large dataset to learn from, and toxin blooms (i.e., not just non-toxic blooms) are somewhat rare, creating a "low n" situation. But as time goes on and data collected by utilities expand (as the frequency of HABs increases), the machine learning algorithms will improve in their ability to predict changes in toxin concentrations.

Predictive skill is key, but so is the visual representation and communication of results. For many municipalities/utilities, HABs are not a data problem, nor even a prediction problem. They are a communications challenge. Big data is often already in hand, and the struggle lies in knowing how best to use this data/information to augment one's decision-making processes. Therefore, it is vital that ClearWater Analytica's HAB prediction system includes a visually compelling and easily accessed front-end. Looking to the future, there is an opportunity to expand the automation provided by the water quality prediction systems. Many utilities operate SCADA systems, to which the output of ClearWater Analytica's HAB prediction system can be connected. In doing so, automated actions on the part of water treatment can be seamlessly informed by water quality prediction systems. Of course, full automation is not desired, and human oversight is required at all stages. But there exist many opportunities for increasing the speed at which preventive measures can be taken, ultimately reducing the cost of HAB mediation for example.

In building the HAB prediction system, several paths forward for improving predictions became evident. First, the current emphasis is on algae and toxins. For many utilities this is a priority. However,

for all utilities other water quality dimensions are also of interest. For example, water temperature, turbidity, pH and dissolved oxygen are all factors that impact decision making at treatment plants. In addition, the timing of toxin transport (and other pollutants more generally) is also of central importance. Utilities often draw water some distance from a given reservoir, and changes in the concentration of toxins occur along the path from the source water to the treatment plant. Given a positive identification at the source water, knowing how long one has before it reaches the treatment plant factors greatly into the water quality manager's decision apparatus. Last, this first iteration of the HAB prediction system focused on short-term predictions, i.e., 3d, 7d and 14d forecast horizons. There is also a need for seasonal predictions: in the winter, water quality managers have a need for some quantification for how good or bad the coming summer will be. This can be measured relative to a benchmark, such as the running average or the preceding year's mean algae and toxin concentrations. This information can be used to help in the stocking of HAB remediation products and staffing.

A key next step is to expand the application of the HAB prediction system to multiple waterbodies. Not only will this bring value to the utilities that source drinking water from these waterbodies, but it provides an opportunity to expand the application of Transfer Learning to water quality prediction. With Transfer Learning, there is an opportunity to develop regional, national, and even global water quality prediction models. In doing so, actionable information will be generated not just for the lucky few municipalities that have the resources to support such an endeavor, but for all communities around the world, including those which have the least resources in terms of collecting data, and in terms of their vulnerability to diminished water quality and quantity. Central to the mission of Clearwater Analytica is to provide *actionable and equitable* insights, so that everyone has the best chance to deal with the grand environmental challenges facing us in the coming century.