DEVELOPMENT OF A PREDICTIVE MODEL FOR TASTE AND ODOR EPISODES IN REGIONAL DRINKING WATER RESERVOIRS

by

Peyton E. Goodling

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science
in Biosystems Engineering

Auburn, Alabama
August 7, 2021

Keywords: Geosmin, MIB, cyanobacteria, actinobacteria, synthase genes, drinking water, taste-and-odor control, predictive models

Approved by

Dr. Brendan Higgins, Chair, Assistant Professor, Biosystems Engineering
Dr. David Blersch, Associate Professor, Biosystems Engineering
Dr. Natalie Capiro, Assistant Professor, Civil Environmental Engineering
Dr. Alan Wilson, Professor, School of Fisheries, Aquaculture, & Aquatic Sciences

ABSTRACT

Taste-and-odor episodes affect water quality in reservoirs throughout the world, and utilities including the City of Auburn Water Resources Department, Opelika Utilities, and Columbus Water Works have each identified having taste and odor issues in recent years and consider them high priority for resolution. These episodes are caused by high concentrations of odorous compounds, predominantly 2-methylisoborneal (MIB) and geosmin, in drinking water reservoirs. MIB and geosmin are volatile compounds that are produced by microorganisms, primarily cyanobacteria and actinobacteria, in natural water bodies. Though these compounds are not harmful, they produce musty odors in drinking water supplies that lead to distrust and complaints from consumers because humans are highly sensitive to these compounds. Both compounds are recalcitrant in traditional water treatment processes, thus activated carbon is typically used for advanced temporary treatment. The high cost of advanced treatment makes continuous treatment of raw water unreasonable for most facilities, leaving a short period between an episode and consumer complaints. To determine when these T&O episodes are most likely to occur, predictive models are needed for better water-quality management. We developed CART and multiple linear regression models for geosmin using R. One of the key advances of this work was the integration of geosmin synthase gene abundance which was determined by qPCR. Modeling of the data revealed the best model fits were built when the datasets had high (>30 ng/L) geosmin peaks shown with Auburn, whereas the current Opelika and Columbus datasets gave us limitations, display low-moderate peaks and variability, having models with lower predictive power. The inclusion of the qPCR data proves to be most effective at predicting the high geosmin levels. Sequencing of the qPCR products revealed *Anabaena* and *Planktothrix* as likely producers. Another component of this work was the evaluation of PCR primers for the MIB synthase gene in cyanobacteria and actinobacteria. Specific and efficient primers are needed for accurate quantification of the MIB synthase gene which can be incorporated into models, similar to what was done for geosmin. In evaluating MIB primers, we discovered good efficiency for 4 primer sets and good specificity for one of them. Results from select samples sent for sequencing helped in discovering the primary MIB producers for reservoir in our region.

ACKNOWLEDGMENTS

Throughout the writing of this dissertation, I have received a great deal of support and assistance.

I would first like to thank my supervisor, Professor Brendan Higgins, whose expertise was invaluable in formulating the research questions and methodology for this study. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would also like to thank the utility partners we paired with to complete water sampling and water quality analysis. This includes Dusty Kimbrow with the City of Auburn Water Resource Management Department, Dan Hilyer, PE with Opelika Utilities, and William Kent with Columbus Water Works.

In addition, I would like to thank my thesis committee not only for their time and effort, but for their intellectual contributions to my development as an engineer. This includes Professor David Blersch, Professor Natalie Capiro, and Professor Alan Wilson. A special thanks to Professor Alan Wilson for supplying additional fishpond samples that aided in our MIB primer set evaluations. I sincerely hope to contribute to the environmental field in the future as well as to inspire others to do so just as my professors here at Auburn have done for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

T&O	taste-and-odor
MIB	2-methylisoborneal
Auburn	Auburn Water Works
CWW	Columbus Water Works
Cgeo1	Primer set for cyanobacterial geosmin synthase gene abundance
d	Depth (continuous)
CHLOR_A	chlorophyll-a
TEMP	water temperature
Specific	specific conductance
Wind.spe	wind speed
TKN	total Kjeldahl Nitrogen
TURB	turbidity
Phosphat	phosphate anion abundance
BG.Count	blue-green algae count
MIB3324	MIB3324F/4050R primer set for cyanobacterial MIB synthase gene abundance
Gaget	Gaget-MS-F1/R1 primer set for cyanobacterial MIB synthase gene abundance
MIB-Rf/Rr	MIB-Rf/Rr primer set for cyanobacterial MIB synthase gene abundance
Str-Rf/Rr	Str-Rf/Rr primer set for actinobacterial MIB synthase gene abundance

INTRODUCTION

The purpose of this research is to develop models for the prediction of taste and odor outbreaks in drinking water reservoirs, caused by high concentrations of musty-odor compounds geosmin and 2-methylisoborneal (MIB). Episodes of high geosmin and MIB have been a perennial problem for water utilities in Alabama and Georgia and are also an issue in countries globally (Devi et al., 2021). These occurrences, though non-harmful, create excessive amounts of customer complaints and erode community trust in the water utilities, creating the need to address the issue. Geosmin and MIB are both difficult to remove during traditional water treatment processes and require more advanced treatment techniques such as powdered activated carbon to better resolve the problem compounds. These more advanced treatment practices come with heightened costs for the drinking water utilities and the low human detection limit of 10 parts per trillion (ppt) requires these costly treatments to be used in a short period to prevent customer concerns. To allow the water utilities time to treat the water only when necessary, to be economically efficient, predictive modeling tools are necessary for the predictions of these taste and odor events. A lack of a clear universal relationship between geosmin and/or MIB production with environmental factors (Devi et al., 2021) has made such predictive modeling efforts difficult. It was our goal to develop a predictive model for taste and odor episodes. This would allow the water utilities to respond with temporary treatment options or provide the community with an alternate water source, as they have done before. Our hypothesis was that by focusing on the organisms that synthesize the geosmin, rather than focusing on factors correlated with cyanobacterial blooms in general, better predictions would be made. We anticipated that this analysis and modeling framework could be more broadly applied for geosmin and MIB outbreaks in the Southeastern regional drinking water reservoirs.

Predictive models have previously been made (Jüttner & Watson, 2007; Dzialowski, 2009), though many models tend to focus on factors correlated with algal blooms although only a small subset of algae actually produce MIB or geosmin (Jüttner & Watson, 2007; Suurnäki et al., 2015). Algal blooms tend to occur when warm and sunny weather is met with elevated nutrient levels in a waterbody. Efforts to use proxies for algae blooms, like chlorophyll-a, weather, and other water quality parameters, to predict these T&O events are only successful if the reservoir is dominated by a T&O producing algae (Dzialowski, 2009). When empirical models are developed in these particular cases, the effectiveness could fade over time as the microbial population evolves or if the model is applied to a different water body. Rather than focusing on algae blooms, our focus was on the abundance of the compound synthase genes. Knowing that cyanobacteria and actinobacteria are the primary producers for these compounds, our focus was on their geosmin synthase and MIB synthase genes. The presence of these genes is a necessary state in order to experience a T&O episode. Thus, a high abundance of such genes, accompanied by other environmental factors could allow for better prediction of T&O episodes. Sequencing these synthase genes can also help utilities understand which specific organisms are contributing most to particular T&O episodes, providing new opportunities for reservoir management.

1.1. Overall Objective

The objective of this research was to develop Classification and Regression Tree (CART) and multiple linear regression models for predicting T&O episodes. These models include synthase gene abundance data for MIB and geosmin. Levels of MIB were low and sometimes undetectable in the reservoirs analyzed during the study period from March to October of 2020, causing us to focus our modeling efforts on geosmin model development. Our hypothesis was that by focusing on the organisms that synthesize the compounds, rather than predicting algal blooms in general, enhanced prediction tools would be developed. The expected outcome was that the analysis and modeling framework can be applied more broadly throughout the Southeastern regional drinking water reservoirs for other utilities' use. We quantified water quality parameters (including geosmin and MIB levels) across three seasons in three drinking water reservoirs, quantified the abundance of geosmin gene synthase in water samples and generated predictive modeling tools using CART and multiple regression to predict the geosmin. Water samples were collected from three drinking water reservoirs in Auburn, AL (Ogletree), Opelika, AL (Saugahatchee), and Columbus, GA (Oliver). Three separate reservoirs were measured to increase data variability and better understand the comprehensiveness of the developed models. A secondary objective of this work was to evaluate qPCR primers targeting the MIB synthase gene. Developing primers targeting this gene has proven challenging (Devi et al., 2021). Studies using qPCR for T&O detection have demonstrated usefulness, though oftentimes primers have a limited capacity for detecting all cyanobacterial or actinobacterial species involved in T&O episodes, and some primers simply have poor efficiency or specificity (Devi et al., 2021). Validation of previously developed primer sets in different geographical locations is helpful in their verification and development of standard protocols (Devi et al., 2021). Our aim was to identify primers with high specificity and efficiency that could be used to develop future models for MIB, similar to what we would develop for geosmin.

This thesis is composed of three chapters, the first being a literature review, the second being our efforts for predictive modeling of geosmin in drinking water reservoirs, and the third chapter being an evaluation of primers targeting the MIB synthase gene. This study was performed over a 14-month period, with the sampling taking place from March to October of 2020. There were 19 water quality parameters incorporated in the datasets in this study, including geosmin and MIB. Because MIB levels were very low in the three drinking water reservoirs throughout the study period, we obtained additional samples from fishponds and other sources to aid in the evaluation of the MIB synthase primers.

CHAPTER 1: LITERATURE REVIEW

1. Background

Taste-and-odor episodes effect water quality in drinking water reservoirs throughout the world, and regional utilities have reported having these issues arise in past years. These episodes are caused by high concentrations of odorous compounds, predominantly 2-methylisoborneal (MIB) and geosmin (Suurnäki et al., 2015), in drinking water reservoirs. These unpleasant tastes and odors in drinking water reservoirs, though not currently known to be harmful to human health, lead to consumer complaints (Giglio et al., 2010) and subsequent high cost of advanced treatment for the utilities due to the low human detection limit of around 1.3 ng/L for geosmin and 6.3 ng/L for MIB (Kehoe et al., 2015), though 10 ng/L is the accepted lower limit for utilities. In recent history, consumer complaints for regional drinking water utilities ranged from 20-60 complaints per day during taste and odor events. These utilities consider the resolve of these T&O episodes a very high priority for their customers, spending extra funds on management strategies to treat the reservoirs in response. These tastes and odors have caused users to instead purchase bottled water, mineral water, or otherwise processed water (Parinet et al., 2010), though these issues extend to affect other uses of tap water such as showering and cooking. Just in the US, an estimated $813 million is spent annually by the consumers on bottled water to avoid the tastes and odors associated with these compounds (Kehoe et al., 2015). The regional utilities that have identified having these issues arise near the border of Alabama and Georgia in the USA include the City of Auburn Water Resources Department, Opelika Utilities, and Columbus Water Works, with Columbus having to outsource their water from a different Georgia utility when a massive outbreak occurred in the past.

The two most common compounds associated with taste and odor episodes are geosmin and MIB, which are volatile musty-odor compounds produced by microorganisms in water bodies (Suurnäki et al., 2015). These compounds have been concerns globally, with research occurring in many countries including America, Canada, China, Korea, Europe, and Australia (Dzialowski et al., 2009; Kehoe et al., 2015; Chen & Zhu, 2018; Chung et al., 2016; Suurnäki et al., 2015; Asquith et al., 2018) attempting to find resolutions to this issue. Geosmin and MIB are not easily removed during the conventional water treatment methods, and advanced treatment methods therefore must be used, which again are costly. The low human detection and difficulty for removal during typical processes leaves little time for the drinking water utilities to remove geosmin and MIB below the level of 10 ppt. Being able to predict when these geosmin and MIB episodes might occur would provide a more viable option for time and cost efficiency for drinking water providers by allowing the utilities to initiate advanced water treatment practices before major customer complaints occur, rather than treating the water around-the-clock or having to outsource their water from surrounding cities.

2. *Most Common Taste and Odor Compounds: Geosmin and MIB*

Geosmin and MIB are both odorous metabolites that cause musty taste and odor problems in water bodies worldwide (Suurnäki et al., 2015). They are secondary metabolites (small organic molecules produced by organisms, but are not essential for their growth, development, or reproduction), and are the main causes for T&O episodes in drinking water reservoirs.

There has been considerable research on these compounds since their isolation and identification from actinomycetes in the 1960s (Gerber & Lechevalier, 1965; Medsker et al., 1969). Geosmin (-4,8a-dimethyloctahydronaphthalen-4a-ol) has the molecular formula of $C_{12}H_{22}O$ and MIB (-1,2,7,7-tetramethylbicyclo [2.2.1] heptan-2-ol) has the molecular formula $C_{11}H_{20}O$ (NCBI, 2021). Geosmin is an irregular sesquiterpene which has lost an isopropyl group and MIB is a methylated monoterpene (Fig. 1) (Jüttner & Watson, 2007), and both exist as (+) and (-) enantiomers, though the (-) enantiomers which are naturally produced are the cause for odor outbreaks because it is 10 times more potent than the (+) enantiomers (Jüttner & Watson, 2007). The genes identified for the biosynthesis of the off-flavor compounds have been found in bacterial genomes and the biosynthetic pathway reported on in detail for MIB (Komatsu et al., 2007) and geosmin (Giglio et al., 2011).

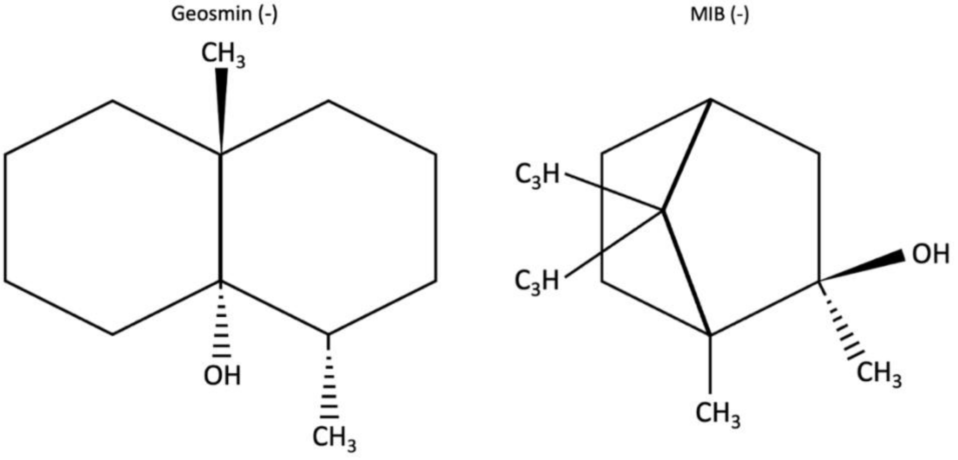


Figure 1. Structures of microbial volatile terpenoid metabolite: (a) geosmin and (b) MIB (image adapted from Watson, 2003).

Geosmin and MIB in freshwater systems are produced by a subset of benthic and pelagic microorganisms in water bodies and are also known to be produced by terrestrial microorganisms, like actinobacteria (Jüttner & Watson, 2007). Recognizing the sources of these compounds is essential to our ability to predict and treat these outbreaks.

2.1. Geosmin and MIB Producers

Geosmin and MIB are produced primarily by prokaryotes including cyanobacteria, actinobacteria, proteobacteria, myxobacteria, and are also produced by some eukaryotes including some fungi, liverworts, and amoeba (Jüttner & Watson, 2007). There are often misperceptions that these compounds are synthesized by "algae" which is a generic term that includes both eukaryotic taxa and cyanobacteria ("blue-green algae") together (Watson et al., 2008), though no green-algae have been cited as geosmin/MIB producers. Green-algae is not known to produce these T&O compounds. Pelagic and benthic cyanobacteria are considered the primary producers of these T&O compounds in freshwater systems (Asquith et al., 2018), with a total of 132 strains from 21 genera having been observed to produce geosmin, and 72 cyanobacterial strains from 13 genera known to produce MIB (Devi et al., 2021). However, only 58 sequences associated with geosmin synthase, and 28 sequences associated with MIB synthase have been assembled in the NCBI database. Known cyanobacterial taxa producing geosmin and MIB are listed in Tables 1 and 2. The primary actinobacteria producer of these compounds is thought to be *Streptomycetes* (Asquith et al., 2013, Asquith et al., 2018) due to it being the first producer that was isolated and identified. However, it is now known that *Nocardia* and *Micromonospora* are also producers (Otten et al., 2016; Lindholm-Lehto & Vielma, 2018).

Table 1. List of reported cyanobacterial producers (genus) of geosmin (Geo) and MIB compounds (adapted from supplemental data from Devi et al., 2021).

Cyanobacteria genera	Geo	MIB
Anabaena	+	+
Aphanizomenon	+	
Calothrix	+	
Coelosphaerium	+	
Cylindrospermum	+	
Fischerella	+	
Geitlerinema	+	
Gloeotrichia	+	
Hyella	+	+
Leptolyngbya	+	+
Lyngbya	+	+
Microcoleus	+	+
Neowollea	+	
Nodosilinea	+	
Nostoc	+	
Oscillatoria	+	+
Phormidium	+	+
Planktothricoides		+
Planktothrix	+	+
Pseudanabaena	+	+
Schizothrix	+	
Scytonema	+	
Spirulina		+
Symploca	+	
Synechococcus	+	+
Tychonema	+	

Table 2. List of reported actinobacterial and other non-cyanobacterial producers of geosmin (Geo) and MIB (with data from Juttner & Watson, 2007; Asquith et al., 2013).

Actinobacteria genera	Geo	MIB
Actinomadura		+
Aspergillus	+	+
Microbispora	+	
Micromonospora		+
Nocardia	+	+
Penicillium	+	+
Streptomyces	+	+
Symphyogyna	+	
Vannella	+	

The genes and biosynthetic pathway for geosmin and MIB synthase in some *Streptomycetes* and myxobacteria are found in a simplified biosynthetic scheme in Figure 2 where there are two main pathways for synthesis, with the 2-methylerythritol-4-phosphate (MEP) pathway (also called the non-mevalonate pathway) suggested for being the major route for many bacterial (cyanobacterial) groups (Jüttner & Watson, 2007; Giglio et al., 2011; Komatsu et al., 2007). Depending on growth stages, the Mevalonate (MVA) pathway is also suggested for being the chosen pathway by some organisms (typically *Streptomyces*) (Jüttner & Watson, 2007), though both pathways can lead to the production of geosmin and/or MIB. For Figure 2, after IPP and DMAPP have been produced by *Streptomyces*, they can generate GPP (geranyl diphosphate) which generates MIB using MIB synthase, or farnesyl diphosphate, then germacradienol (geosmin synthase) to produce geosmin (Jüttner & Watson, 2007; Komatsu et al., 2007) and the cyanobacteria *Nostoc punctiform* PCC 73102 (Giglio et al., 2008). The dynamics of production vary among species and strains and the product also varies by amount (Watson, 2003), which makes monitoring and prediction more difficult. The biological and ecological function of geosmin and MIB have not been explicitly identified through research (Tyc et al., 2017). Asquith et al (2013) addressed that the production in *Streptomyces* is generally the defense strategy for its survival of the next generation of germinating spores, suggesting the geosmin and MIB could be allelogens acting against one or many species, or that it may act as an antibiotic by reducing parasitic bacteria or fungi that harm the algae around it. However, the levels would need to be > 0.45 mg/L and >480 ng/L for geosmin and MIB, respectively, to be effective. Antifungal properties of geosmin and MIB have also been reported being produced by *Streptomyces alboflavus* (Wang et al., 2013). It is interesting then that many fungi also produce these compounds.

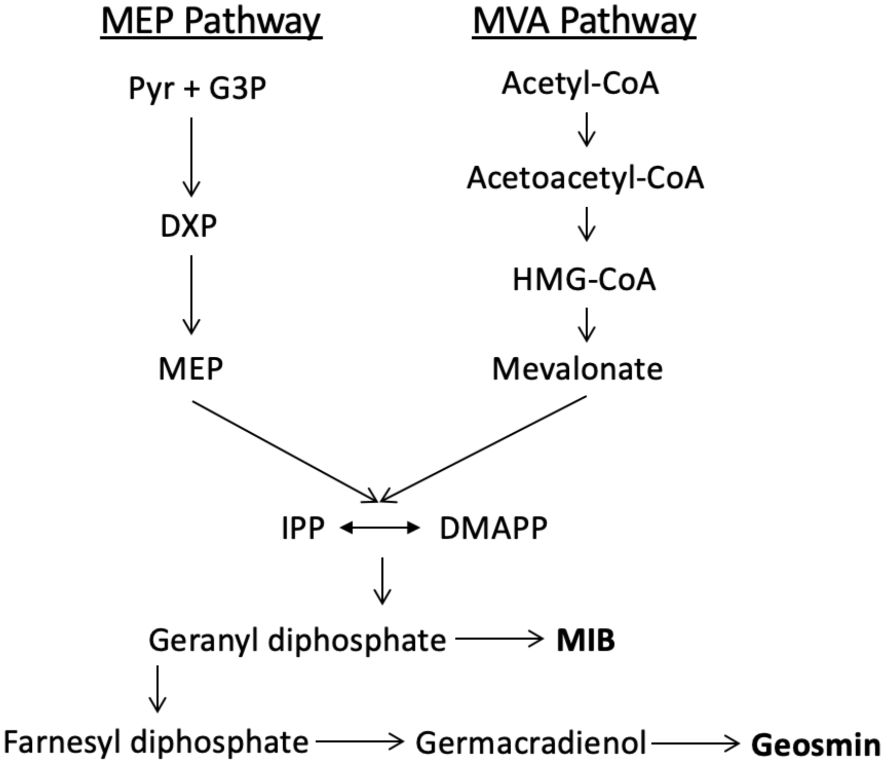


Figure 2. Simplified biosynthetic scheme for the formation of MIB and geosmin in some streptomycetes and myxobacteria (Jüttner & Watson, 2007; Giglio et al., 2011; Komatsu et al., 2007). (Pyr: Pyruvate; G3P: glyceraldehyde-3-phosphate; DXP: 1-Deoxy-D-xylulose-5-phosphate; MEP: 2-C-methyl-D-erythritol-4-phosphate; HMG-CoA: 3-Hydroxy-3-methylglutaryl-CoA; IPP: Isopentenyl diphosphate; DMAPP: Dimethylallyl diphosphate)

The genes associated with the synthesis of geosmin and MIB have also been elucidated. Devi et al (2021) reports that the main target for both cyanobacteria and actinobacteria (actinomycetes) for geosmin synthase is the *geo* gene and the MIB synthase (cyclase) gene is the *mic* gene, also called *mibC*.

The geosmin and MIB in water bodies are developed and released through metabolism along with the biodegradation of cyanobacteria and their metabolites (Kim et al., 2014). The blooming cyanobacteria can produce extracellular metabolites during their decay process (Xuwei et al., 2019), and most of the MIB and geosmin are released during their death and decay (Srinivasan & Sorial, 2011). Degradation of dissolved geosmin includes photolysis (photo-oxidation), biolysis (biodegradation), volatilization, and adsorption, whereas degradation of cell-bound (particulate) geosmin includes cell lysis or damage which then becomes dissolved geosmin (Chung et al., 2016). It has been observed that bacterial biodegradation was more important for MIB loss than volatilization, photolysis, or adsorption (biosorption) and some soil and aquatic bacteria are capable of that biodegradation though little is known about MIB and geosmin degradation in water supply reservoirs (Westerhoff et al., 2005). Through batch experimentation in a lab for both MIB and geosmin degradation, Westerhoff et al. (2005) found bacterial biodegradation rates ranging from 0.5-1.0 ng/L-day (~30 ng/L-month), which was comparable to the total loss rate of 0.23-1.7 ng/L-day. Using sand filters and bioreactors, rates of biodegradation have also been found to be 0.10-0.58 ng/L-day (Ho et al., 2007), comparable to the study before. Due to volatilization, the half-life of geosmin and MIB was found (using each compound's Henry's constant) to be around 1 year, which is too long to be important to total degradation, and biosorption removed both compounds at a negligible rate as well (Westerhoff et al., 2005). Natural lake bacteria (heterotrophic aerobes) could be stimulated by higher biomass densities and potentially increase the MIB and geosmin biodegradation rate, observed as up to 5-50 ng/L-day removal, though it was not specified how much of that removal was due directly to bacterial biodegradation (Means & McGuire, 1986). *Pseudomonas, Bacillus, and Enterobacter* sp. have been found to be able to deplete MIB (Izaguirre et al., 1988; Tanaka et al., 1996), and *Sphingopyxis alaskensis*, *Novosphingobium stygiae* and *Pseudomonas veronii* (only when all three are present) to deplete geosmin (Hoefel et al., 2006). *Alphaproteobacterium*, *Sphingomonas*, *Acidobacteriaceae, Methylobacterium,* and *Oxalobacteraceae* have also been identified as capable of geosmin biodegradation (Ho et al., 2007; Xue et al., 2011). There is no comprehensive list of bacteria responsible for either geosmin or MIB biodegradation, though research continues to discover more strains responsible. Photolysis is also a possibility for the degradation of geosmin and MIB, though the average 254 nm radiation is ineffective in removing these compounds from water, thus additional techniques such as oxidation must be used for effective degradation (Kutschera et al., 2009). Degradation can also occur through grazing of zooplankton, such as the dominant freshwater planktonic herbivore *Daphnia*. With some copepods exhibiting high food selectivity and sensory-based feeding, it has been suggested that odor moderates the food selection of some small herbivores, though this has not been characterized, and the robust *Daphnia* do not exhibit this sensitivity to volatiles (Watson et al., 2003). Grazing of the zooplankton almost completely transferred particulate geosmin into the dissolved form (Jüttner & Watson, 2007). This transition into a dissolved form would then, though, make it easier for bacterial biodegradation.

Because cyanobacteria can produce geosmin and MIB as well as cyanotoxins, there is a little overlap which is species specific though there is no robust relationship between toxins and T&O compounds in source waters (Watson et al., 2008). The utilities that we were paired with for this research have not reported cyanotoxin abundances, and otherwise low incidence of human poisoning by cyanotoxins might be related to the avoidance of unfiltered drinking water when significant odor is perceived. Major producers of cyanotoxins are *Microcystis*, which is not known to produce either geosmin or MIB (Watson, 2003), and *Clyindrospermopsis* (Chiu et al., 2017), which can produce geosmin (Devi et al., 2021). There are species of Anabaena which can produce geosmin along with saxitoxins, microcystins, and anatoxins, as well as *Phormidium* and *Planktothrix* which can produce geosmin and MIB, along with lipopolysaccharide and microcystins (Watson, 2003). *Aphanizomenon* and *Oscillatoria* have also been accounted for being important cyanobacterial species which can produce toxins and geosmin/MIB (Kim et al., 2020), though it is species specific. Watson (2003) provides a table which lists all cyanobacterial species which can produce volatile organic compounds, which toxins they can also produce, and their typical habitat.

Knowing this genetic information allows for the development of further molecular tools for the monitoring and prediction of T&O outbreaks. However, there is significant variability in the synthase gene sequences among taxa. While this can be helpful for organism identification, it makes it challenging to develop universal primers that target these genes across a wide range of taxa.

2.2. Conditions that Promote Producers

These off-flavor compounds are often found in raw water sources, though the conditions that promote their production by organisms is varied by location and the dominant organisms in each water body. Aside from the varied locations of water sources, the differences in taste and odor production are likely due to different microbial compositions of the organismal community, which is driven by the changing environmental conditions, where the T&O concentrations could rapidly increase if a particular or dominating species or genotype becomes favored by environmental conditions (Peter et al., 2009). Peter et al (2009) found that the main producers of geosmin in a lake they sampled had only very small numbers present after 10 years, and some species had disappeared completely. The quality of the water and the local environmental conditions play a large role in the proliferation of different taste and odor outbreaks, including those outside of geosmin and MIB. For a drinking water utility, the most widely available water quality parameters are the ones that are easily and routinely measured (e.g. temperature, pH, DO, conductivity, etc.) (Parinet et al., 2013). It is also important to keep in mind that it is difficult to detect relationship between the biomass of the T&O producers and the T&O compound if it is not measured as both the cell-bound, or particulate, and dissolved, or extracellular, fractions (Jüttner & Watson, 2007). This is because particulate geosmin can build up in the cyanobacterial cells and be released as dissolved geosmin depending on the environmental conditions as they biodegrade or are eaten by zooplankton, like *Daphnia* (Dzialowski et al., 2009; Jüttner & Watson, 2007).

Climate change is a factor that can affect environmental conditions in favor of cyanobacterial blooms, like eutrophication, increasing water temperatures, increasing severe storms, and shorter winter seasons (Huisman et al., 2018; Srinivasan & Sorial, 2011) that could lead to elevated T&O issues in drinking water resources. The effects of environmental factors on cyanobacterial production of odorous compounds have been explored and found that warm, nutrient-rich waters are prone to the cyanobacterial blooms that can be accompanied by T&O episodes, though the linkage between nutrients and the T&O compounds are still not understood clearly (Jüttner & Watson, 2007). Some studies have relied on the abundance of cyanobacteria as a predictor of T&O compound concentrations, yet Dzialowski et al. (2009) states that the measurements of cyanobacterial biovolume were not consistent predictors of the dissolved geosmin detected in their study, and that their efforts even found a negative relationship between the variables. This contrasts with the finding that geosmin has a positive relationship to the cyanobacterial biovolume (Jüttner & Watson, 2007). This underscores that the predominant T&O producers vary significantly across systems. It has been found that *Anabaena* sp. FACHB-1384, a major MIB producer, is sensitive to low temperature (<20° C), and that *Anabaena* sp. Chusori and *Anabaena* sp. NIER (both geosmin producers) are sensitive to high light intensity above 100 µmol/m^2/sec but are not sensitive to low temperatures (Oh et al., 2017). Another species, *Oscillatoria limosa* was found to have maximum growth at 25° C (Cai et al., 2017). Thus, developing a model with universal predictive power is very difficult.

Chlorophyll-a has been shown to have r = 0.47 correlation with particulate MIB levels detected in a Lake Taihu, China, possibly due to the high concentrations of chlorophyll-a providing biodegrading cellular material and carbon source that is useful for the growth of odor-producing organisms (Xuwei et al, 2019). Research on the correlation of dissolved oxygen and odorants is scarce (Xuwei et al, 2019), with Jüttner (1984) pointing to lower MIB concentrations with anaerobic environments for shallow eutrophic lakes.

Geosmin concentrations (linearly proportional to cell density) were also found to increase under high nitrate concentrations, and that phosphorus stress decreased the geosmin production in *Anabaena* sp. NIER (Oh et al., 2017). This differed from MIB productivity which only had reduced productivity in *Planktothrix* sp. FACHB-1374 under phosphorus stress (Oh et al., 2017). It was concluded that high light intensity and P-stress contribute to lower geosmin levels and lower temperature contribute to lower MIB levels, but they state that no universal optimal conditions for cyanobacterial odor compound production have been found likely due to different strains having their own specificities (Oh et al., 2017; Dzialowski et al., 2009). Researchers at Chaohu Lake in China found a -0.709 (P<0.05) Pearson correlation between TP and geosmin levels, but no significant correlation for MIB (Zhang et al., 2019). Treatments in another experiment revealed that the highest MIB concentrations were produced when low TN:TP ratios were supplied, though it was stated that it is still hard to determine whether the specific nutrient levels (TN or TP) or the individual TN:TP ratio effected the experiment greater (Olsen et al., 2016). There is little other evidence supporting that TN, TP, or TN/TP ratios are important in cyanobacterial dominance or whether they make a difference in concentrations of MIB, though their odor-producing potential and concentrations could be affected by cyanobacterial species and volume (Bai et al., 2017). An aerobic and organic-rich environment is known to stimulate the growth of actinomycetes and the subsequent production of geosmin and MIB in the system (Guttman & Rijn, 2008).

Another study in China found that geosmin was correlated to dissolved oxygen, ammonia, and nitrate with Pearson correlation coefficients of -0.798 ($P<0.05$), 0.935 ($P<0.01$), and 0.744 ($P<0.05$), respectively in their Songhua Lake (Zhang et al., 2019), and no significant variables correlated with MIB. Although there are studies finding correlations between water quality parameters and the T&O compounds, there is no single factor that causes these episodes due to the dynamics of the aquatic systems.

3. T&O Mitigation and Treatment Methods

Before treating these T&O outbreaks, it is essential to be able to detect them in water bodies and the main approach for direct detection is solid-phase micro-extraction coupled with gas chromatography/mass spectrometry (SPME GC/MS), which can sensitively and accurately measure the compounds at levels as low as parts per trillion (ppt) (Suurnäki et al., 2015; Wang et al., 2016; Devi et al., 2021). GC/MS can be coupled with other advanced methods that improve the sensitivity of detection, like closed-loop stripping analysis (CLSA), resin adsorption (RA), solid-phase extraction (SPE), stir-bar sorptive extraction (SBSE), liquid phase microextraction (LPME), purge and trap (P&T), static headspace (SH), and dynamic headspace (DH) (Devi et al., 2021). The SPME pre-conditioning method is simple, easy, and fast making it the most used standard method for geosmin and MIB in water samples. With the low detection thresholds of these compounds for humans being so low (<10 ng/L) it is then necessary for the drinking water utilities to control and treat.

Because the compounds are recalcitrant in traditional water treatment processes like coagulation, sedimentation, and filtration (Srinivasan & Sorial, 2011; Jüttner & Watson, 2007), further treatments must be employed to stop their persistence in open waters. Oxidation with the chemicals chlorine (Cl_2), chlorine dioxide (ClO_2), and potassium permanganate ($KMnO_4$) are not found to be effective for the compound removal (Srinivasan & Sorial, 2011), and though chlorine can mask the musty odor it may also produce altered unwanted odors and chlorinated biproducts. Ozonation is a process that has shown great efficiency at destroying T&O compounds, but it has downfalls in that it can form carcinogenic bromate (BrO_3) during the ozonation depending on the pH, temperature, ozone dosage, and bromide ion concentration (Liato & Aider, 2017). To improve the use of ozonation, other processes can be used in combination, such as ultraviolet (UV), hydrogen peroxide (H_2O_2), or (manganese) Mn^{2+} treatments. More turbid waters can use ultrasonic irradiation (sonication) for the proposed decomposition of odorous compounds, and UV treatment alone can also be used but is only effective at high doses (>254 nm) though it can form unwanted nitrite products (Liato & Aider, 2017; Srinivasan & Sorial, 2011). A relatively newly (2011) researched method is electrochemical treatment, which has been found to have high efficacy at destroying T&O molecules along with being simple and robust in structure and operation at a low cost (Liato & Aider, 2017). Biodegradation using biofilters has also been used to remove geosmin and MIB from drinking waters but have issues in drinking water reservoirs with degrading the micropollutants, though pairing it with ozonation can improve usefulness (Nerenberg et al., 2000). One of the most effective methods for odor compound removal is through adsorption by powdered/particulate activated carbon (PAC), which is successful at reducing concentrations, though it is less successful for geosmin and MIB than other T&O compounds and the presence of other chemical oxidants (chlorine or chloramines) reduces the effectiveness (Liato & Aider, 2017; Srinivasan & Sorial, 2011). To increase efficacy, membrane systems have been studied in combination of both coarse and fine PAC (Kim et al., 2014).

4. Predictive Models for T&O

The removal of geosmin and MIB can have high costs and constant PAC dosing becomes impractical for water utilities, therefore predictive tools are needed to help foresee when these T&O episodes are most likely to occur. Multiple types of modeling have been used in previous research in efforts to predict T&O outbreaks, such as empirical, hydrologic, numeric, machine learning, and classification and regression modeling.

Empirical modeling can be used to relate water quality parameters with measured geosmin concentrations. Dzialowski et al (2009) developed two regression models to predict dissolved geosmin levels in Kansas reservoirs. The first was a best-subsets regression, providing two "n"-variable models with the highest coefficient of determination (r^2) values, and the second was a stepwise regression method which added a candidate explanatory variable to the regression model one at a time ($P = 0.10$ to enter) if they increased the model's predictive power significantly (Dzialowski et al., 2009). This was done until they got the highest coefficient of multiple determination (R^2), providing a measure of how much of the geosmin level variation was explained by predictor variables in the models. Their one and two-variable cross-sectional models developed had a 24-35% explanation in geosmin level variation for all five of their reservoir locations combined. For their five reservoir sampling locations in the Kansas reservoir, their highest R^2 was 0.94, whereas one of their locations was not able to develop any significant regression model. They found significant relationships in separately developed equations for geosmin with Secchi Disk depth, dissolved PO_4-P, dissolved NO_3-N, *Anabaena* biovolume, chlorophyll-a, total algal biovolume, total phosphorus, total cyanobacterial biovolume, and dissolved oxygen (Dzialowski et al., 2009). Favorably, they only included variables that are easy and inexpensive, and likely routine, for the water utilities to collect data on. The models they developed only used samples that exhibited geosmin levels of 5 ng/L or higher due to detection limits. They concluded that they were not able to identify any consistent spatial patterns or any positive relation to cyanobacterial biovolume for their reservoirs (Dzialowski et al., 2009). They also did not use varying depths for their sampling depths, potentially hindering their ability to account for all geosmin or cyanobacteria. Previous research on similar reservoirs in Kansas found similar and contrasting variables to be significant, such as turbidity and specific conductance (Christensen et al., 2006), chlorophyll-a (Smith et al., 2002), Secchi Disk depth, specific conductance, and turbidity (Mau et al., 2004), and total phosphorus, chemical oxygen demand, and dissolved oxygen (Suguira et al., 2004). Temperature, velocity, and phosphorus have also been linked in a simple binary model to the occurrence and prediction of cyanobacteria (Kim et al., 2020), though not all cyanobacteria are linked to geosmin and MIB production, therefore not being useful. A negative of empirical modeling is that, like the above reviewed research displays, the models for the individual reservoirs have the best predictive power because the dissolved geosmin levels are largely based on local environmental factors (Dzialowski et al., 2009) and are useful only if the water body is dominated by one major producer of T&O compounds. Linear models rely on correlation analysis, which relies on independency of variables and therefore not robust because odor production relies on multiple variables.

A comparison of the use of multi-linear regression (MLR), principal component analysis regression (PCA), and multi-layer perceptron (MLP) found that among them the simple multi-linear regression had reasonable predictive capabilities ($R^2 = 0.657$, $P<0.001$), but the use of PCA did not lead to better model performance (Parinet et al., 2013). Multiple regression could be realistically useful for utilities to use, though the parameters described in the Canadian research are not routinely monitored in our regional drinking water reservoirs, like phaeophytin, sum of algae, and redox potential.

Aquatic systems are dynamic, so the models for predictions of odorous compounds have also been developed to be dynamic. Empirical Dynamic Modeling (EDM) is nonparametric and is a robust method that can be used in R. Through EDM on a reservoir in Japan (Wang et al., 2019), it was found that *Phormidium* spp. (a cyanobacteria) was the most important cause and predictor of MIB production. A better indication for MIB production in this type of modeling would be of MIB-producing genes as a better index than cyanobacterial abundance itself (Wang et al., 2019) to improve EDM. Another type of predictive modeling is hydrodynamic modeling, which uses inflow and outflow, production of geosmin by metabolism of cyanobacteria (*Anabaena*), and degradation and decay (Chung et al., 2016). A particular model developed in China, a 2D hydrodynamic and water quality model called CE-QUAL-W2, involved the physical and biological processes of "total" geosmin, which left out the long-term transformation processes of particulate geosmin. They state that their model reasonably predicted outbreaks for their reservoir, though their Root Mean Square Error (RMSE) was 69.14 and their Relative Error (RE) was 78.22% (Chung et al., 2016). To build an accurate hydrodynamic model better incorporation of the sources of the T&O compounds and long-term degradation need to be implemented, which are difficult to accurately measure, as well as changing from species and location for the waterbody.

Few numerical models have been developed to directly simulate and predict the generation of MIB in freshwater, though another hydrodynamic simulation was developed in China by Chen and Zhu (2018) using an adapted ECOMSED model for their Qingcaosha, Shanghai Reservoir. An ecological model was developed with RCA (Row Column Advanced), an ecological systems operating program frequently used in water quality modeling. The model integrated nutrients and cyanobacterial species, which successfully simulated MIB levels in their reservoir, though with an average RE% of 50.56, which are acceptable to describe the variations (Chen & Zhu, 2018). Their highest MIB level captured was 10,000 ng/L (10 mg/m^3), which the model simulated, though not at the correct concentration (Chen & Zhu, 2018).

With odor production being influenced by multiple variables, machine learning has become more widely used for prediction of odorous compounds. Using physiochemical water quality data from a reservoir in Kansas, twelve linear and non-linear regression modeling techniques were compared by Harris and Graham (2017). They concluded that random forest (RF), support vector machine (SVM), boosted tree (BT), and Cubist modeling were most predictive out of the 12 models for geosmin concentrations, along with cyanobacterial abundance, and microcystin levels, with the Cubist modeling being able to predict the geosmin maximum concentrations (Harris & Graham, 2017). This is a helpful technique because the reservoir's management pertains mostly to those large concentrations. Cubist modeling is a popular regression method in R, which has results of low RMSE (Yang et al., 2017; Harris & Graham, 2017) and are made by creating a set of rules that split data at terminal nodes, each using a linear equation to predict the response variable but is best used for predicting maxima metabolite levels. Harris and Graham's study used one-way ANOVA (Analysis of Variance) to compare variances within their groups of data, which was important for their large-scale environmental study. Their random forest model performed best with geosmin levels >20 ng/L, though the Cubist model displayed a more robust fit for the same geosmin threshold. The Cubist model has potential for improving regression modeling efforts for metabolite maxima, it has not been widely used or proven (Harris & Graham, 2017) to be useful for ordinary geosmin or MIB level predictions. Random forest and SVM models are machine learning algorithms that take multiple variables into consideration to create nonlinear models, but are complicated for interpretation (Wang et al., 2019) and would not be easy to integrate into a water utility's routine management.

A Canadian study (Kehoe et al., 2015) used linear regression and random forest modeling in comparison for forecast modeling of T&O episodes based on a 24-year time series of data. This team was able to develop forecasting random forest models of TON using chlorophyll-a, turbidity, TP, temperature, and odor producing algal taxa with 0–26-week lag-times. TON is the Threshold Odor Number which represents the persistence of geosmin, MIB, and other odors. Their highest performing random forest model ($R^2 = 0.71$) was able to forecast 12 weeks in advance with a 93% true positive rate and 0% false positive rate, with the taxonomic data being the most important in the model (Kehoe et al., 2015). The ability to use this type of modeling on just geosmin and/or MIB would be harder than simply TON due to a lack of long-term data for calibration and validation and would be difficult for utilities to implement.

Because aquatic ecosystems are nonlinear systems with chemical, biological, and physical variables in constant interactions with temporal and spatial variations, nonlinear tools are a necessity for odor predictions. Classification and Regression Tree (CART) modeling is another form of nonlinear modeling in R using the 'rpart' package and is one of the most well-known decision-tree learning algorithms in literature (Yang et al., 2017). CART uses a recursive binary splitting process on a dataset by identifying an input variable and a breakpoint and partitioning the samples into two child-nodes (Yang et al., 2017). Split-nodes in the trees are made with the expected sum variances for two resulting nodes to be minimized (squared residuals minimization algorithm or Gini index) (Choubin et al., 2017). The tree is then pruned to optimize the prediction accuracy and reducing overfitting by minimizing the number of branches and can then be analyzed through performance metrics. In past research, CART has been used in determining the quality of water in a reservoir (Chou et al., 2018) and has been shown to best predict suspended sediment loading (R^2 ranging from 0.53 to 0.74) with "very good" performance (Choubin et al., 2017). It has shown to be better at predictions than support vector machines (SVM) and artificial neural networks (ANN) (Choubin et al., 2017). CART is computationally fast and is robust to outliers and although no previous research has been done on predictions of taste and odor compounds in water bodies using CART, literature proves that where hydrological data are available CART modeling can be useful in developing easy to read visual decision trees for predicting a categorical variable. To date, CART modeling has not been used for the prediction of T&O levels using genetic data.

5. *qPCR and Primer Production*

Quantitative Polymerase Chain Reaction (qPCR) has proven to be a promising tool for the detection of geosmin and MIB events (Devi et al., 2021; Gaget et al., 2020; Suurnäki et al., 2015;

Wang et al., 2016; Tsao et al., 2014) by focusing on the gene encoding the geosmin or MIB synthesis. To better improve modeling efforts, rather than focusing on cyanobacterial abundance itself, enhanced indicators for episodes are the T&O producing genes (Wang et al., 2019). qPCR has the ability to quantify the number of copies of a particular strand of DNA (gene) in a sample using specific primer tools. SYBR green qPCR uses a fluorescent dye which binds to the amplicon (DNA fragment being amplified), which changes the fluorescent pattern, which allows the qPCR machine to detect how much DNA is in the sample after each cycle of amplification (MilliPore, n.d.). A primer is a tool that is used to target the specific gene sequence in the DNA of an organism, such as the conserved gene regions *geo* (geosmin synthase enzyme) or *mic* (MIB synthase enzyme). There are existing primer tools that have been developed (Suurnäki et al., 2015; Gaget et al., 2020; Wang et al., 2016) that have accurately developed PCR protocols for the detection of geosmin and MIB synthase in water samples.

A high-quality primer is of importance, where a low primer concentration decreases the accumulation of primer-dimer formation, which is critical for SYBR green use in qPCR (MilliPore, n.d.). Efficiency and specificity are important factors in evaluating primer sets for qPCR use. Efficiency is determined by preparing a dilution series for each primer set, and the efficiency will be 100% if the number of molecules of the target gene sequence double during each cycle of amplification. Specific amplification of the intended target gene requires that primers don’t match other target sequences that would allow for unwanted amplification, and software tools such as BLAST or Primer-BLAST can aid in the complex design of primers (Ye et al., 2012). Existing primer tools for geosmin and MIB have been developed but many have issues with non-specificity or exclusion of important organisms.

As part of the research reported in this thesis, evaluation of primers targeting the MIB synthase gene was carried out. Using Primer-BLAST, Suurnäki et al. (2015) developed a primer set, MIB3324F/4050R, that was designed to amplify the MIB synthase gene based on *Oscillatoria*, *Planktotricoides*, and *Pseudanabaena* species and had specific and efficient results. In developing a primer that targets the *mic* gene in *Pseudanabaena*, *Planktotrhricoides*, and *Leptolyngbya*, Wang et al. (2016) developed MIB-Rf/Rr and found no non-specific amplification and a wide coverage for MIB-producing cyanobacterial species for the first time, along with developing a *Streptomyces* spp. specific MIB primer, Str-Rf/Rr. Gaget et al. (2020) was able to develop a primer set, MIB-

MS-F1/R1, based on the alignment of Castaic *Pseudanabaena limnetica* (HQ630883; California, USA), NIVA-CYA111 *Pseudanabaena* sp. (HQ630887; Japan) and LBD305b *Oscillatoria limosa* (HQ630885; South Korea) and conclude that it worked effectively.

These primer developments are great advances on developing efficient and specific primers in their sampling systems and help in proving the usefulness and necessity of qPCR in the detection of synthase genes for T&O outbreak predictions rather than cyanobacterial and actinobacterial abundance alone. Variations in sampling, DNA extraction, and quantification errors, species variation in water bodies, and many primer sets having limited capacity to detect a range of species, are all difficulties behind the use of qPCR (Devi et al., 2021), and therefore the evaluation of these primer sets on varied water bodies is necessary for improvement.

CHAPTER 2: GEOSMIN MODELING

1. Introduction

Geosmin is a naturally occurring taste and odor compound that is produced by bacteria and released into drinking water reservoirs, causing musty flavors. In aquatic systems, geosmin production has been primarily attributed to freshwater cyanobacteria and actinobacteria. This compound is an issue because it can be detected at such low levels, as low as 10 ng/L (Devi et al., 2021). It is a stable compound that is not easily oxidized, and requires advanced treatment methods, such as sorption to activated carbon, ozone/GAC, ozone/UV, hydrogen peroxide/UV and membrane water treatment processes (Kim et al., 2014). The utilities involved in this research project use activated carbon in order to minimize consumer complaints, however, this approach is expensive. The other alternative would be to remove the issue at the source, by being able to detect which organisms are producing T&O compounds and developing a reservoir management strategy to minimize the growth of such organisms. Such an approach may not work because water utilities often have limited control over influent into their drinking water reservoirs. To treat the water quickly and efficiently, it is helpful to be able to predict the geosmin episodes and treat the T&O event with temporary advanced water treatment, such as with activated carbon. Such an effective predictive tool is not currently available.

The cause for geosmin production by organisms is largely unknown, though it is noted that geosmin production might act as a possible chemical defense for lichen (Suurnäkki et al., 2015). It has also been linked to functioning as a quorum sensing signal molecule, allelogen, and competitive-organism inhibitors (Zaitlin and Watson, 2006). The geosmin production cause being highly unclear slightly hinders our predictive capabilities. Also, with no clear/universal relationships between geosmin levels and easily measured water quality parameters (Watson and Ridal, 2004; Zaitlin and Watson, 2006), more sophisticated predictive tools and models are needed. Empirical regression models have been developed, and that of Dzialowski et al. (2009) was able to predict T&O episodes in multiple Kansas reservoirs, only some with reasonable predictive power. The downside to their approach of using cyanobacterial correlated factors is that their model is not able to be transferred geographically. Kehoe et al. (2015) was able to develop a "random forest" model, which is a machine learning algorithm, using a long-term dataset. They found that cyanobacterial counts of likely producers of T&O compounds was the most significant in their modeling, and they were able to forecast T&O events reasonably (Kehoe et all, 2015). Among varying models made, a hydrodynamic and mass transport model has also been developed by Chung et al. (2016) which relied on inflow and outflow, production of geosmin by cyanobacteria, and degradation and were able to capture outbreaks of geosmin in the North Han River in Korea. This model also relied on cyanobacteria as a proxy for T&O. A water body could exhibit high cyanobacterial counts, but because only a subset of cyanobacteria produces geosmin or MIB, the T&O levels could still be low. You could not have high geosmin or MIB counts at the same time as having low geosmin or MIB synthase genes present in the corresponding T&O producers. Although the studies above were able to develop models with practical predictive capabilities, we expected that the use of geosmin and MIB synthase gene abundance, rather than cyanobacterial abundance, will lead to higher predictive power.

The purpose of this study was to develop a predictive model for geosmin episodes in regional reservoirs using quantification of the geosmin synthase gene in water samples. We hypothesized that it is useful to measure the abundance of the organisms that actually have the gene to make geosmin, and we can quantify the abundance of the synthase gene using qPCR. By then overlaying this data with the routinely collected water quality parameters, we hypothesized that the model would have improved predictive capability compared to models that lack this information. This hypothesis is supported by the results of previous researchers who have shown that cell count data on known geosmin producers was the most significant predictive factor for their success in empirical models (Kehoe, 2015; Journey et al., 2013). Other researchers have also shown that synthase gene abundance alone can have good predictive power over geosmin concentration (e.g. Tsao et al. 2014; Wang et al., 2016). To be able to most accurately predict T&O outbreaks, we intended to utilize the ability to quantify the synthase gene abundances in the water samples, and then in combination with the water quality parameter data use Classification and Regression Tree (CART) modeling in R to predict the conditions and genetic states that contribute to the critical thresholds for the T&O compounds.

2. *Methods and Materials*

2.1. Reservoir Geography and Water Sample Collection

The three utilities we paired with for this project were the City of Auburn Water Resources Department located in Auburn, Alabama, Opelika Utilities located in Opelika, Alabama, and Columbus Water Works (CWW) located in Columbus, Georgia. These three utilities routinely collect water samples for their own characterization and compliance with regulations for water quality. In addition to these samples, each utility agreed to collect and deliver 100-200 ml samples for our lab to conduct further molecular biology work from March 1st through October 31st of 2020. The map in Figure 3 shows the watersheds and sub-basins that include the reservoirs in this study. The starred locations represent the intake locations where sampling occurred in Lake Ogletree (Auburn, AL) and Lake Saugahatchee (Opelika, AL). While these intake samples may not be fully reflective of all upstream sources of taste and odor compounds, they were collected on a frequent basis by the utilities at multiple depths, which reflects the extent of stratification in the reservoir. The goal was to develop a predictive tool that would not be burdensome for the utilities to be able to use, thus the single intake location minimizes the burden to the utilities. The Saugahatchee Creek basin and the Chewacla Creek basin both drain into the Tallapoosa River. Lake Oliver (Columbus, GA) lies on the Chattahoochee River and is a much larger and deeper reservoir than Saugahatchee and Ogletree.

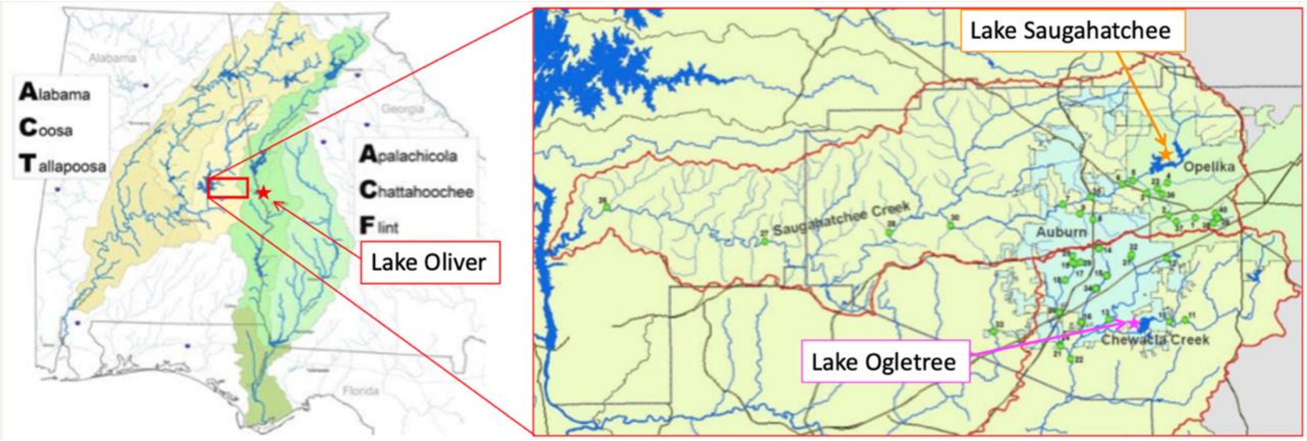


Figure 3. Lake Ogletree (Auburn, AL) and Lake Saugahatchee (Opelika, AL) shown within the red square, each with only 1 sampling location at the intake, represented by the pink and orange stars, respectively.

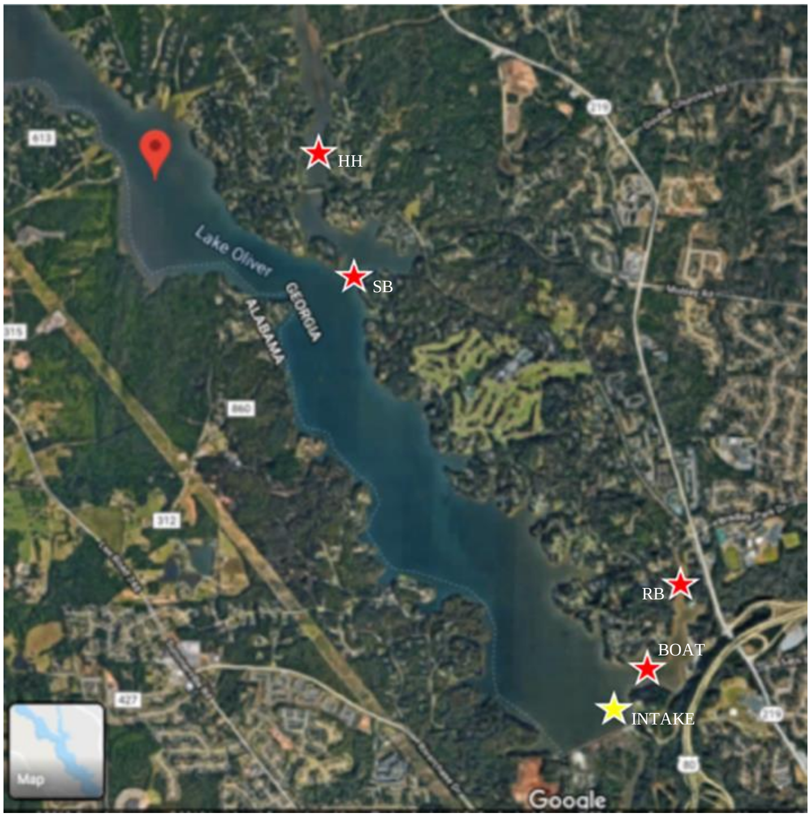


Figure 4. Lake Oliver (Columbus, GA) with intake sampling location denoted by the yellow star, and the 4 other sampling locations denoted by the red stars.

Figure 4 shows the sampling locations in the watershed for Lake Oliver that were carried out by Columbus Water Works (CWW). Columbus agreed to engage in more diverse sampling locations throughout their large reservoir. The red stars show upstream sampling locations, and the yellow star shows the treatment plant's water intake location, having five total sampling locations for CWW. The five different locations within Lake Oliver are the Intake, Boathouse (BOAT), Roaring Branch (RB), Standing Boy (SB), and Heiferhorn (HH) locations. Each of the utilities also agreed to conduct water quality analysis and provide the data to us, described in the following section.

2.2. Standard Methods of Water Quality Characterization by Utilities

The Auburn, Opelika, and Columbus water utilities individually performed the water quality analyses shown in Table 3 for each of their reservoirs. The sampling occurred during peak season for taste and odor episodes, allowing our lab to best capture the data needed for predictive tools. Auburn and Opelika samples were analyzed every two weeks for most variables, with air temperature, precipitation, and wind speed being collected daily by weather stations. Columbus sampled at their five locations twice per week while also collected daily probe data parallel to Auburn and Opelika. All three of the utilities agreed to engage in this more intensive sampling and analysis than is typical. The geosmin quantification was performed by gas chromatography/mass spectrometry (GC/MS) using external laboratories, and by in-house analysis by Columbus Water Works. The anion chromatography panel for soluble anions, described in section 4 below, was conducted in our lab.

Table 3. Sampling and analysis parameters and frequency from March 1, 2020, through October 31, 2020.

	Auburn	Opelika	Columbus	
Parameter	Ogletree Intake	Saugahatchee Intake	Oliver Intake	Oliver Watershed
Total phosphorus	2 months	2 weeks	Twice/week	Twice/week
Total Kjeldahl nitrogen	2 months	2 weeks	Twice/week	Twice/week
Nitrite + nitrate	2 months	2 weeks	Twice/week	Twice/week
Water temperature	2 weeks	2 weeks	Twice/week	Twice/week
pH	2 weeks	2 weeks	Twice/week	Twice/week
Dissolved oxygen	2 weeks	2 weeks	Twice/week	Twice/week
Specific conductance	2 weeks	2 weeks	Twice/week	Twice/week
Turbidity	2 weeks	2 weeks	Twice/week	Twice/week
Secchi depth	2 weeks	2 weeks	Twice/week	Twice/week
MIB	2 weeks	2 weeks	Twice/week	Twice/week
Geosmin	2 weeks	2 weeks	Twice/week	Twice/week
Anion panel*	2 weeks	2 weeks	Twice/week	Twice/week
Chlorophyll-a	Daily	2 weeks	Twice/week	Twice/week
Phycocyanin	Daily	2 weeks	Twice/week	Twice/week
Blue-green algae count	NA	NA	Twice/week	Twice/week
Air temperature	Daily	Daily	Daily	Daily
Precipitation	Daily	Daily	Daily	Daily
Wind speed	Daily	Daily	Daily	Daily
Sample depths	5 ft, 10 ft, 15 ft	5 ft, 10 ft, 15 ft	20 ft	0 ft (Surface)
Sample time points	16	16	69	69
No. samples/site	3	3	1	1
No. of sites	1	1	1	4
Total water samples	**48**	**48**	**69**	**276**

*Conducted in our lab

2.3. Water filtration and DNA Extraction

In the eight-month sampling season, the water samples were delivered to our lab for further research. Each sample was filtered using 0.2 mm nitrocellulose (VWR) filters to concentrate the solid material. Up to 200 ml of each sample was filtered using a vacuum flask, stopping when the filter had completely fouled, and the filtered amount was recorded for each sample. After each filtration, 2 ml of the filtered sample was collected for ion chromatography analysis. The vacuum flask was then cleansed with deionized water and nano-pure water between each filtration, and a new filter was applied to the apparatus.

For each sample, the respective filter was then cut into strips and extracted directly in the bead homogenization tubes using the PowerSoil DNA Isolation kit by MoBio Laboratories, Inc. The manufacturer's instructions were followed regarding the use of this kit, which has already been shown to be effective for extraction of DNA from both soil and water samples (Kaevska, 2015). This procedure involved vortexing the PowerBead tubes with the cut-up filters inside for rapid and thorough homogenization, then adding different solutions and transferring the liquid step-by-step to ensure cell lysis occurs by the mechanical and chemical methods. The total genomic DNA was captured on the silica membrane in a spin column and was then washed and eluted from that membrane. This kit allowed for further sample analysis using qPCR analysis as well as sequencing.

2.4. Anion Chromatography

The anion chromatography panel complemented the nutrient analyses carried out by the utilities and aided in providing a more complete depiction of the soluble anions including chloride, nitrite, nitrate, phosphate, and sulfate. Soluble ions were of particular interest for algae and cyanobacteria because these are the key nutrient sources for growth, as nitrate and phosphate are the typical limiting nutrients in freshwater systems (Zhang et al., 2019, Oh et al., 2017). Anion chromatography was carried out per the methods described in Chaump et al 2018. Mobile phase was prepared with 4.5 mM sodium carbonate plus 1.5 mM sodium bicarbonate using Milli-Q water. Anion standards were prepared and included chloride, nitrite, nitrate, phosphate, and sulfate at concentrations ranging from 0.195-800 mg/L to detect the low concentrations. A Shimadzu Prominance High Performance Liquid Chromatograph with a conductivity detector was used in conjunction with an AS22 column (Dionex Thermo). An AERS500 suppressor was used to reduce noise and improve sensitivity with detection limits on the order of 50 ng/L. The flow rate was set at 1 mL/min with a system pressure of around 10-12 MPa. With all the samples loaded in their individual HPLC vials, the anion chromatography batches were run.

2.5. Quantitative Polymerase Chain Reactions (qPCR)

2.5.1. Primer Sets

Primers are the tools used to target a specific gene, such as geosmin synthase. Existing primer tools for geosmin are not well developed and those that are documented have many issues with non-specific amplification or exclusion of important organisms. The primers in Table 4 have been referenced and used in publications by at least one other research group other than their own developers. The last two primer sets in the table were developed in our lab and were previously tested for efficiency and specificity via gel electrophoresis and sequencing.

Table 4. Primer sets from literature and those developed in our lab for geosmin.

Primer Set	Target Taxa	Reference
Giglio 250F/971R	Cyanobacteria	Giglio et al. 2008
288A F/R	Cyanobacteria	Giglio et al. 2008
SGF1/JDR1	Cyanobacteria	Tsao et al. 2014
AMgeo F/R	Actinobacteria	Auffret et al. 2011
Cgeo1	Cyanobacteria	Our lab
ActGeo2	Actinobacteria	Our lab

Prior work in our lab showed that at high and low tested annealing temperatures the Giglio 250 primer set led to significant non-specific amplification, and that the 288A primer set led to almost exclusive primer-dimer complexes, diminishing its usefulness in our future work. The SGF1/JDR1 primer set showed some promise and amplified both actinobacteria and cyanobacteria. The AMgeo primer set showed extreme non-specificity but this could be partly the result of low levels of actinobacteria in certain samples. Prior work in our lab also showed that the ActGeo primer set performed acceptably in areas of specificity and efficiency and was further sequenced to determine the producers of these geosmin gene sequences. ActGeo sequencing mapped to actinobacteria, though showing up to 30% of cyanobacteria, proteobacteria, and bacteroides as well.

The Cgeo1 primer set showed high specificity for geosmin synthase in cyanobacteria based on gel electrophoresis and sequencing of the products. The sequencing results showed that 100% of the amplicons mapped to the correct target of cyanobacteria geosmin synthase, confirming the gel results. The Cgeo1 sequencing showed high specificity for targeting only genes in the cyanobacterial phylum while still being broad enough to include multiple producers, including several common cyanobacteria known to produce geosmin, such as *Anabaena* and *Planktothri*x. The primers in this set are: 5’-GATCACTTCCTGGAAATCTAT-3’ and the reverse primer sequence 5’-GCCATTCTACAGACTTAGTAA-3’, with melting temperature, annealing temperature, and target gene referenced in Table 5. Given this high level of specificity, we chose to include the qPCR results using Cgeo1 primers in our modeling effort.

Table 5. Cgeo1 primer set melt temperature, annealing temperature, and gene that is targeted.

	Melt Temperature (Tm, 0.5 µM primer), °C	Annealing Temperature, °C	Target Gene
Cgeo1	50.1	50	*geoA*

2.5.2. qPCR and DNA Quantification

qPCR was used to measure the abundance of geosmin synthase gene abundance in the water samples for each utility. The goal of this project was to concentrate on cyanobacterial contributions to the geosmin synthesis given past findings that they are typically the largest contributor to taste and odor compounds in freshwater systems compared to other organisms (Watson et al., 2008; Jüttner & Watson, 2007). The qTower3 qPCR instrument (Analytic Jena) was used to quantify the abundance of geosmin synthase in the DNA extracts from the water samples using SYBR green detection. The primer set used to amplify the cyanobacterial geosmin synthase gene abundance was Cgeo1. Cycling parameters were 95° C for 5 min, 36 cycles of 95° C for 15 s, 50° C for 30 s, and 72° C for 45 s. It then ran at 72° C for 5 min before the machine shut down and the qPCR products were able to be removed and the data analyzed. The DNA standards required for qPCR were made using previously amplified products from DNA obtained from raw fishpond water samples. These samples were sequenced and found to contain predominantly *Oscillatoria* sp. PCC 9240 geosmin synthase (geoA) and *Planktothrix* sp. 328 geosmin synthase (geoA). This previously amplified product was cleaned using the QIAquick PCR Purification Kit (Qiagen) and then quantified using a double stranded DNA fluorescent quantification kit (Promega QuantiFluor dsDNA System). Dilutions of this reference material were used to generate standard curves in subsequent qPCR runs with the Cgeo1 primers. In this manner, I was able to quantify the abundance of the geosmin synthase genes in the collected water samples by using the constant reference point with the standards. This allowed us to track the change in abundance of genes over the sampling season.

2.5.3. Sequencing

After qPCR for the amplification of the synthase genes, it was important to check for the correct product being amplified using sequencing. Select qPCR products from samples taken across multiple water bodies using the Cgeo1 primer set were sent to an external laboratory for Sanger sequencing. This approach is inexpensive but generally not effective in cases where there are many forms of geosmin synthase amplicons present. Rather, it works best when only one dominant organism is present. This proved to be the case in several samples and allowed us to confirm the primary geosmin producer in certain samples with high geosmin levels.

2.6. CART Model Development and Multiple-Linear Regression

In previous literature, Classification and Regression Tree (CART) modeling has proven to be a powerful alternative to traditional multiple regression-based models for taste and odor episodes (Kehoe, 2015), especially when dealing with aquatic systems (Harris & Graham, 2017; Downing et al., 2001). They are non-linear decision tree models that divide data into more homogenous groups to explain how an outcome variable's values can be predicted based on other variables. A CART output is a tree where each fork is split in a predictor variable and each end node contains the prediction value for the outcome variable, which for this research is geosmin levels. This CART analysis was carried out using the free statistical software package R so that it could be freely and easily implemented by the regional utilities. This is a much simpler option compared to other machine learning algorithms, like random forests, to complete our goal of having developed a simple tool that could be implemented. For use of this software by the utilities, a file containing instructions was also developed.

Once all water quality parameters were obtained, the full datasets for each utility were uploaded into the R software. The CART fit was then made, taking all parameters into consideration. The program then reports back the most significant variables and produces an unpruned regression tree, typically with multiple branches. The program also reports the adjusted R^2 value of the CART model produced. Next, using the important variables listed a new fit was made and a pruned regression tree was output with a new adjusted R^2 value. Each tree developed gives an output of the predicted geosmin levels. This was done for each of the three water utilities. To test our hypothesis that the inclusion of the genetic data results found through qPCR would result in better geosmin predictions, I ran CART models for each utility that included and then excluded the genetic data to compare the adjusted R^2 values. Successful predictions of geosmin level allow the utilities more knowledge on how to best handle their reservoir treatment. For the most successful prediction capabilities, sample variation is key. If there is only a small level of variation in the datasets (e.g. geosmin levels are very low in all samples), the predictions capabilities are unlikely to be meaningful.

After CART modeling, multiple linear regressions were used for inspecting more closely the relationships between water quality variables to aid in answering the question of when geosmin levels might spike. Multiple regression, an extension of simple linear regression, was able to be used in R using the ‘lm’ function using all variables as input variables initially. In all cases, stepwise regression was manually used to reduce the number of variables used to attain the most impressive adjusted R^2 value using significant variables. To do so, all 19 original variables from the water quality parameters and genetic data were entered into the regression function in R. The least significant variable listed was then removed. This process was repeated multiple times until all variables were included and the adjusted R^2 peaked. The adjusted R^2 value is a modified R-squared value that has been adjusted for the number of predictors in the model. It increases only if the new term improves the model more than would be expected by chance and decreases when a predictor improves the model by less than expected by chance. In other words, it shows whether adding additional predictor variables improves the regression model or not. This is in comparison to the multiple R-squared value that is also given, which doesn't provide any incentive to stop adding more variables. Too many variables in a model can produce results that cannot be trusted (Gardener, 2012), so we use the adjusted R^2.

3. Results and Discussion

3.1. Standard Methods of Characterization by Utilities

3.1.1. Geosmin Levels

Through GC/MS, the geosmin levels (ng/L) were measured in all water samples. In our sampling time from March to October of 2020, the geosmin levels were rarely elevated above the 10 ng/L threshold level, with Auburn elevated 60% of the time (Fig. 5), Opelika 0% of the time (Fig. 6), and Columbus only 16.2% of the time (Fig. 7). Typically, the customer complaints begin when geosmin exceeds 20 ng/L. In 2020, water samples from Auburn were above that higher threshold 27% of the time, Opelika 0% of the time, and Columbus only 0.87% of the time. With these low levels of geosmin for the 2020 sampling months, the modeling efforts were not expected to show great adjusted coefficients of determination (adjusted R^2) or fit, due to that low variation of data. Because of its higher variation in geosmin level, the best model for this research was expected to be developed for the Auburn samples, but lower predictive power was expected for Opelika and Columbus.

3.1.1.1. Auburn

The spikes occurred primarily around April and then again in July to August of 2020, with the highest spike occurring at 67.3 ng/L in the lowest depth sample on August 18th. This compares similarly with known spikes in water odor persistence levels in previous studies where the most intense annual peak was in August, while April was when the first increase occurred after a winter minimum (Kehoe et al., 2015).

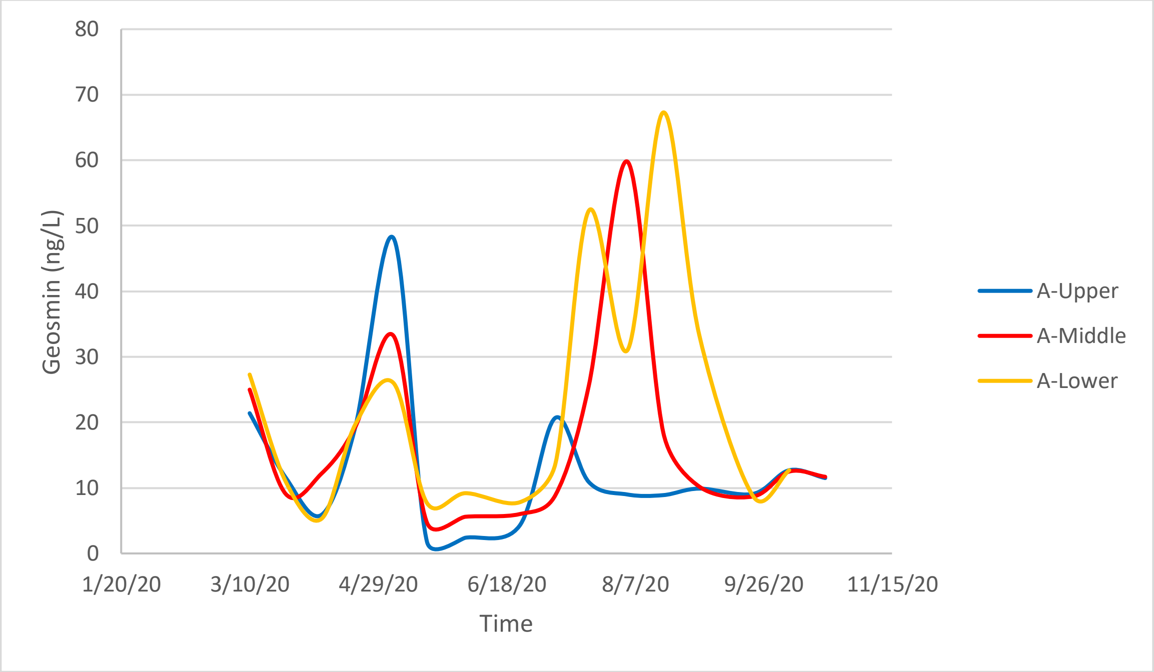


Figure 5. Geosmin over Time: Auburn. "A-Upper" represents samples taken from the upper layer (11.5-17.4 ft.), "A-Middle" represents the middle layer samples (18-23.9 ft.), and "A-Lower" represents the lower layer samples (25.5-30.9 ft.).

3.1.1.2. Opelika

Figure 6 shows similar trends to that of Auburn, with the initial spike occurring in late March/early April and then another spike occurring from late August into late September of 2020, though the most intense spike only reached a level of 8.67 ng/L on March 26th, and subsequently at 7.85 ng/L on September 22nd.

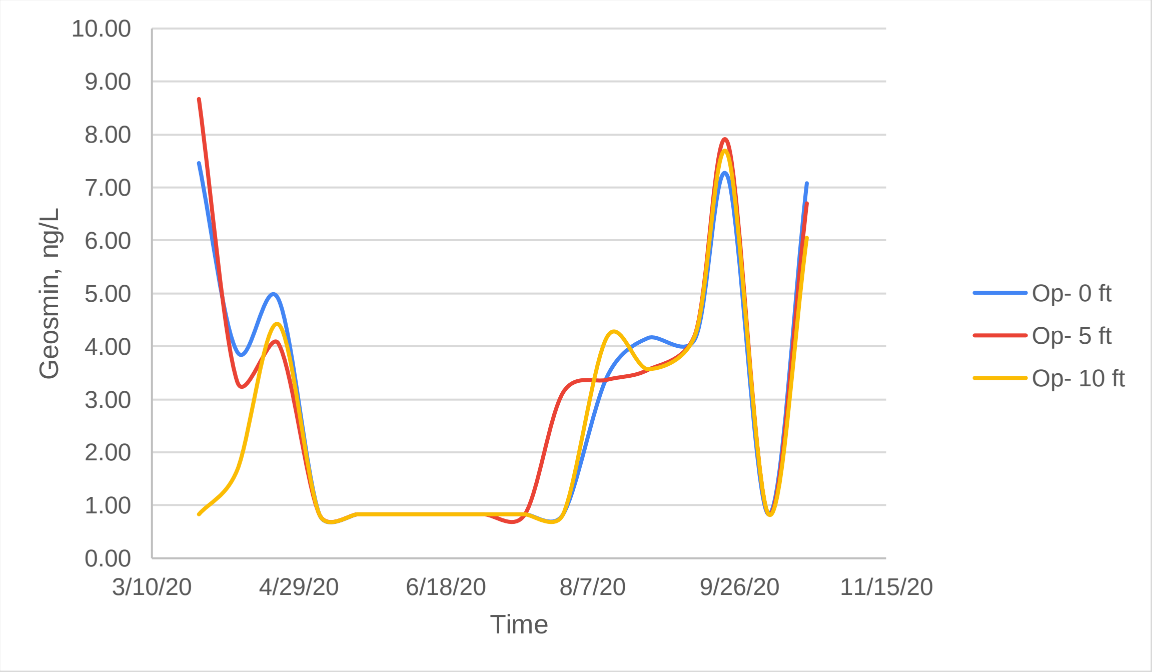


Figure 6. Geosmin over Time: Opelika Utilities.

3.1.1.3. Columbus

The highest geosmin concentrations were observed in March and June (Figure 7). The first spike occurred on March 2nd at a level of 25 ng/L in the Roaring Boat sampling location (the second upstream location from the intake), and subsequent smaller spikes throughout March to June ranging from 14.9-21 ng/L. The largest spike occurred on June 10th at a level of 27.7 ng/L at the Standing Boy sampling location (the third upstream location from the intake). Another cluster of smaller spikes (~15 ng/L) occurred in early-mid September.

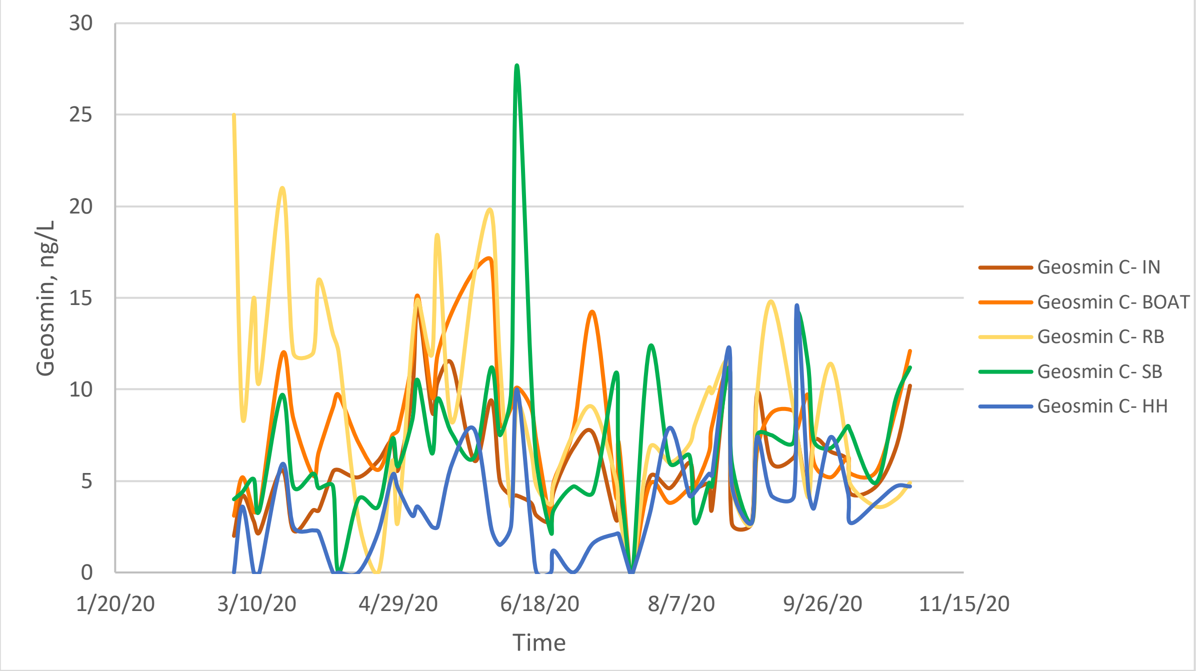


Figure 7. Geosmin over Time: Columbus Water Works.

3.2. Anion Chromatography

Below are graphs of soluble nitrate and phosphate measured in all water samples taken post-filtration through the 0.2 mm nitrocellulose filters. The levels of nitrogen and phosphorus were important to focus on in this study, as they typically promote cyanobacterial growth in lakes when high concentrations are found (Oh et al., 2017; Harris & Graham, 2017). The anion concentrations were integrated into the Pearson's correlations analysis to determine correlations between nitrate or phosphate levels and geosmin levels.

3.2.1. Auburn

The phosphate levels (Fig. 8) for Auburn only show small increases in early April and slight increases in late June. The nitrate levels show spikes in late March and again with smaller spikes from August to late September. The mean soluble nitrate ion (NO_3^-) level was 0.23 ± 0.21 mg/L and soluble phosphate had a mean of 0.14 ± 0.88 mg/L. Nitrate levels tend to be higher than the phosphate levels, but these levels were rather low in comparison to other lakes (Dzialowski et al., 2009).

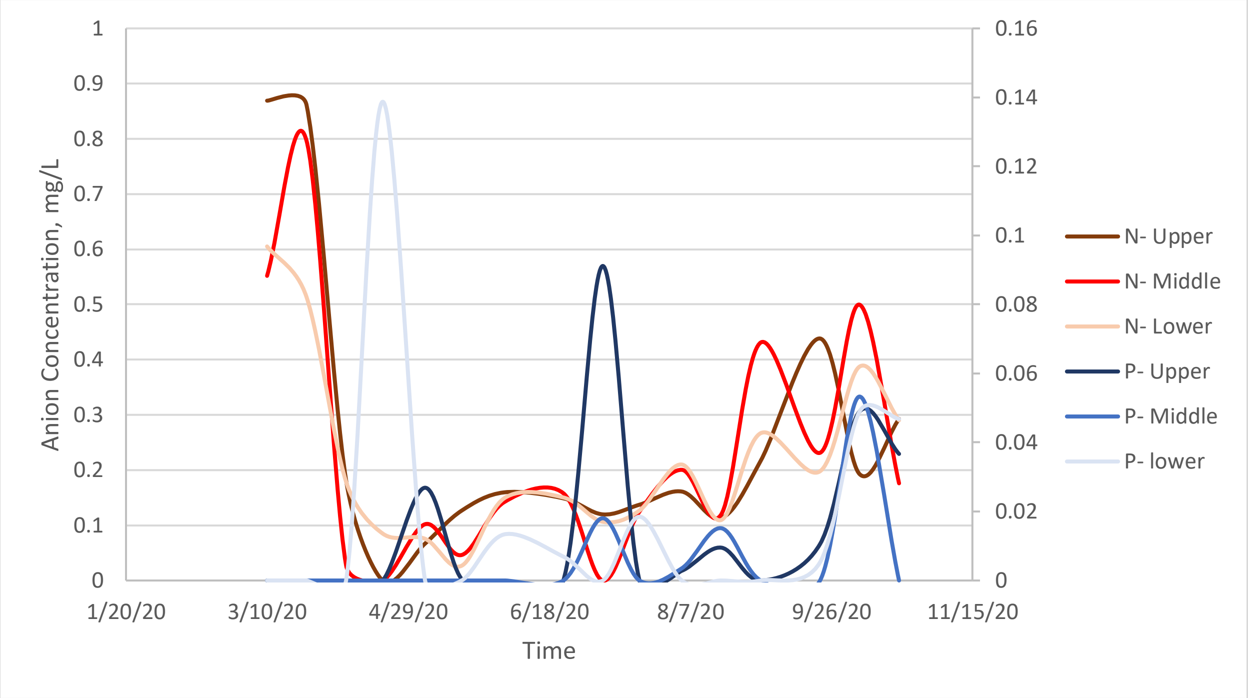


Figure 8. Auburn: Concentrations of nitrate ion (NO_3^-) on the primary scale, and phosphate ion (PO_4^{2-}) on the secondary scale from anion chromatography. "Upper" represents samples taken from the upper layer (11.5-17.4 ft.), "Middle" represents the middle layer samples (18-23.9 ft.), and "Lower" represents the lower layer samples (25.5-30.9 ft.).

3.2.2. Opelika

Opelika's soluble anion levels (Fig. 9) spiked for both nitrate and phosphate around April, with levels of around 0.5 mg/L. The mean phosphate level throughout the sampling period was 0.03 ± 0.07 mg/L, and nitrate had a mean of 0.12 ± 0.09 mg/L.

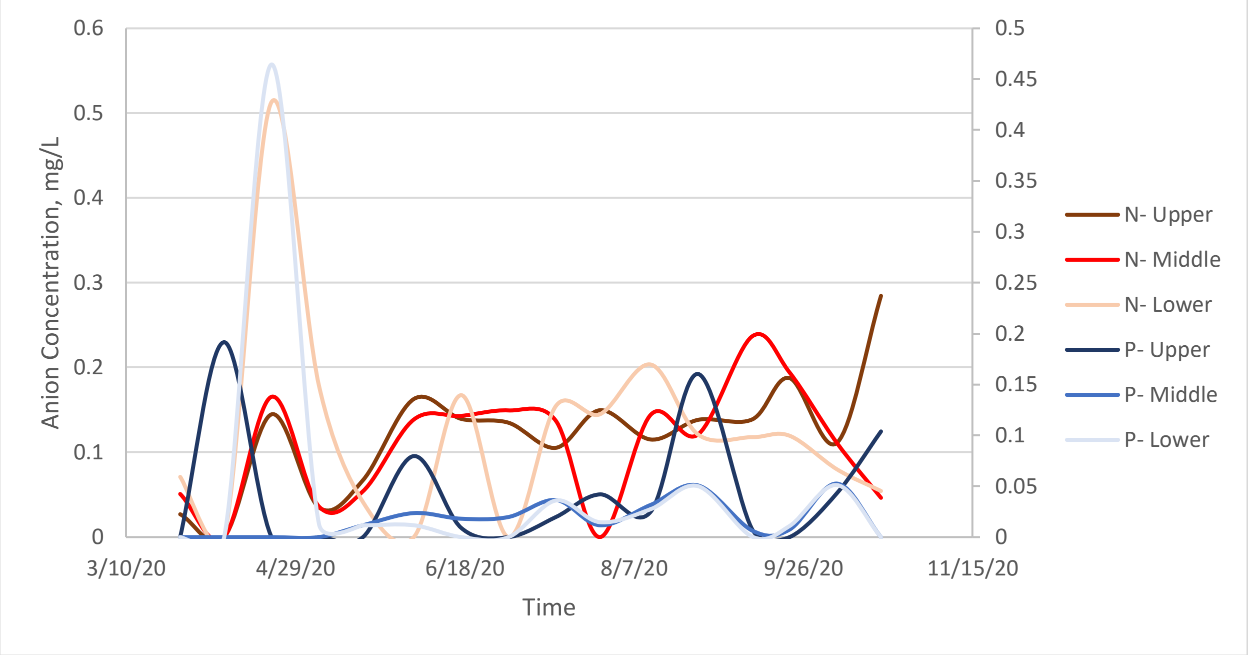


Figure 9. Opelika: Concentrations of nitrate ion (NO_3^-) on the primary scale, and phosphate ion (PO_4^{2-}) on the secondary scale from anion chromatography. "Upper" represents samples taken from the upper layer (0 ft.), "Middle" represents the middle layer samples (5 ft.), and "Lower" represents the lower layer samples (15 ft.).

3.2.3. Columbus

In CWW's drinking water reservoir, the phosphate levels peaked most intensely in October, primarily in the location for Standing Boy at a level of 1.26 mg/L (Fig. 10). All the location's phosphate levels spiked slightly from late March to mid-April, though they never surpassed 0.55 mg/L. In Figure 11, the nitrate levels increased from ~2.5-4.7 mg/L in late April to early May, and then again even higher throughout October, at 4.3-5.1 mg/L. The mean phosphate level was 0.03 ± 0.10 mg/L and the mean nitrate level was 1.29 ± 1.00 mg/L.

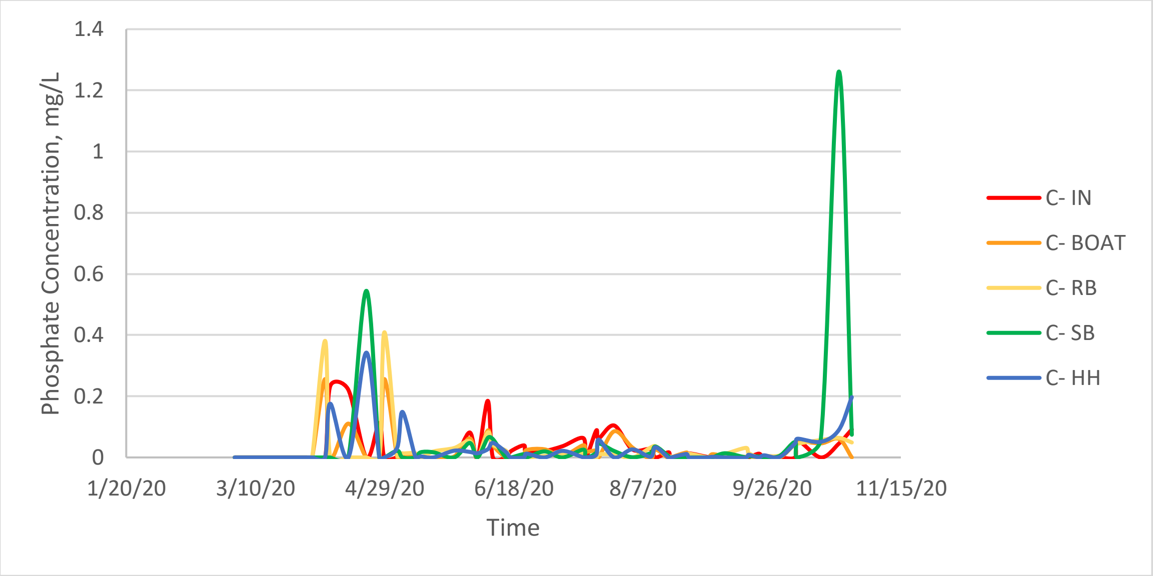


Figure 10. Columbus: Phosphate ion concentration from anion chromatography.

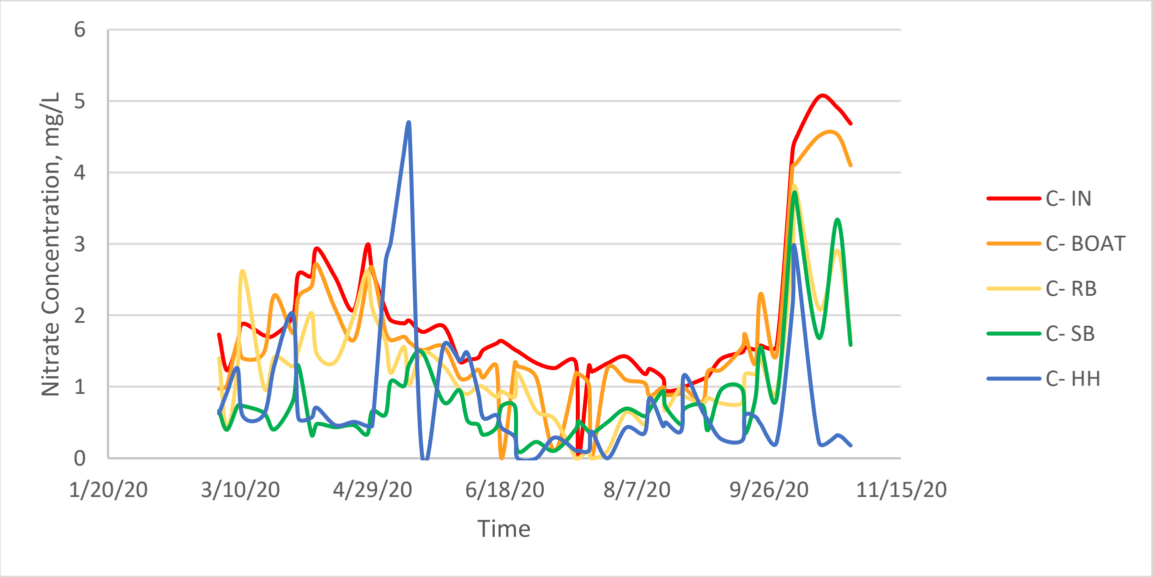


Figure 11. Columbus: Nitrate ion concentration from anion chromatography.

The key take-aways from these figures are that the phosphate levels were elevated primarily in spring at first and then again from late summer to fall. The nitrate levels were also elevated most obviously around the same times in spring and fall from this water sampling period. The nitrate and phosphate are likely being taken up by for productivity during the summer months by the cyanobacteria and other organisms (Gobler et al., 2007).

3.3. qPCR

The following figures (Figures 12-14) show results of cyanobacterial geosmin synthase gene abundance over time calculated after the use of our lab's previously developed Cgeo1 primer set in qPCR. These results were used in subsequent modeling and regression.

3.3.1. Auburn

In Figure 12, the Cgeo1 synthase gene abundance over time for Auburn shows spikes in May and then again in July, August, and October. These spikes in gene abundance are similar in trend to the spikes we saw in the geosmin level detection for Auburn, which also primarily occurred in mid-spring and then again in late summer to early fall months, though with decreases in-between.

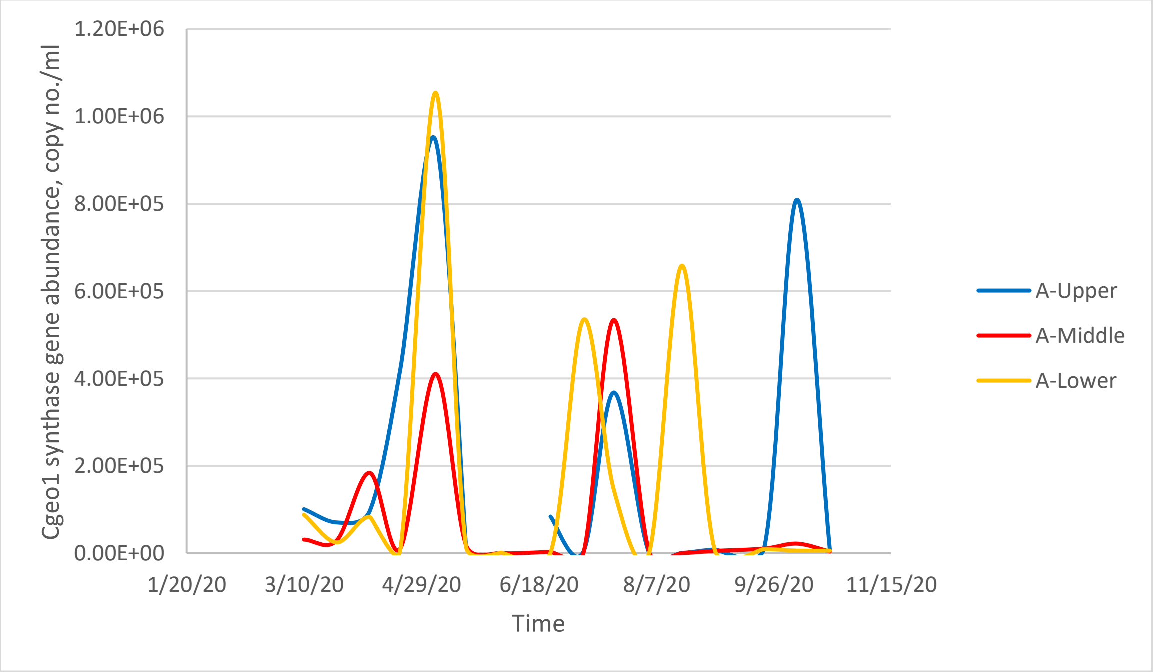


Figure 12. Cgeo1 synthase gene abundance over Time: City of Auburn Water Resources Department.

3.3.2. Opelika

The increases in Cgeo1 gene abundances over time occur at slightly different time periods for Opelika Utilities, with the first large spike on June 3rd at 1,180,000 (1.18E6) copy no./ml of water sample for Cgeo1 synthase gene abundance in the lower sampling depth. The upper and middle depths have increases in gene abundance in mid-June, and then they all spike in late August (Figure 13).

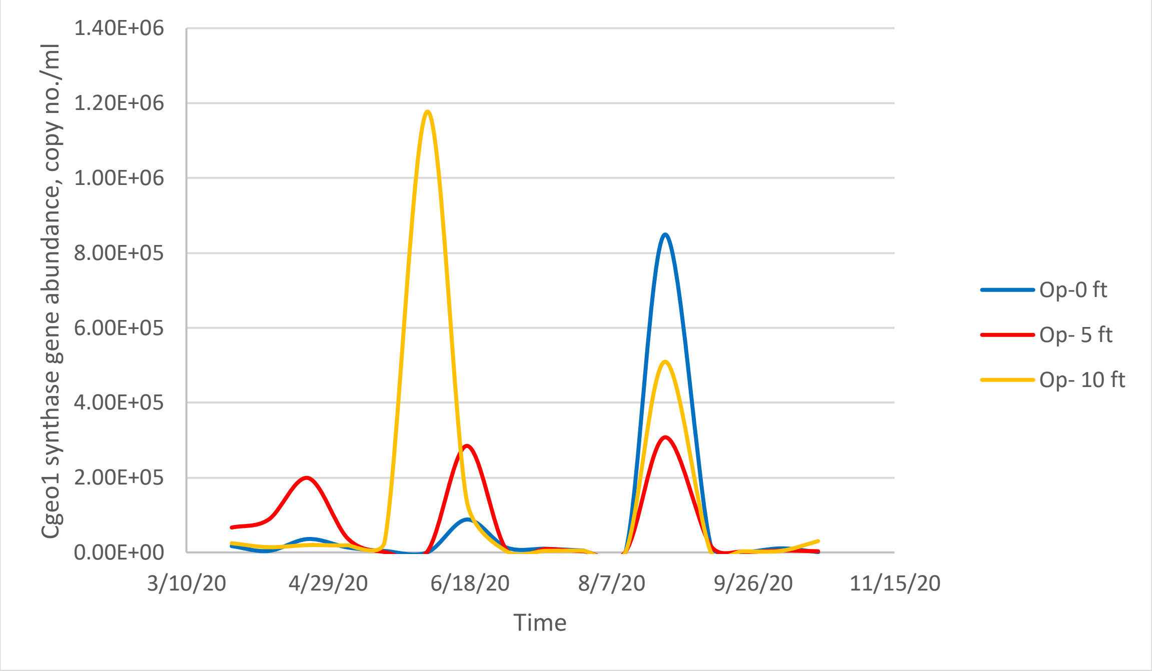


Figure 13. Cgeo1 synthase gene abundance over Time: Opelika Utilities.

3.3.3. Columbus

The largest spike for CWW's Cgeo1 gene abundance was on March 19th at 2,650,000 copy numbers of gene synthase per ml of water sample, though the rest of the smaller spikes happen in early June into mid-July, with the smallest abundances being found from August through September (Figure 14).

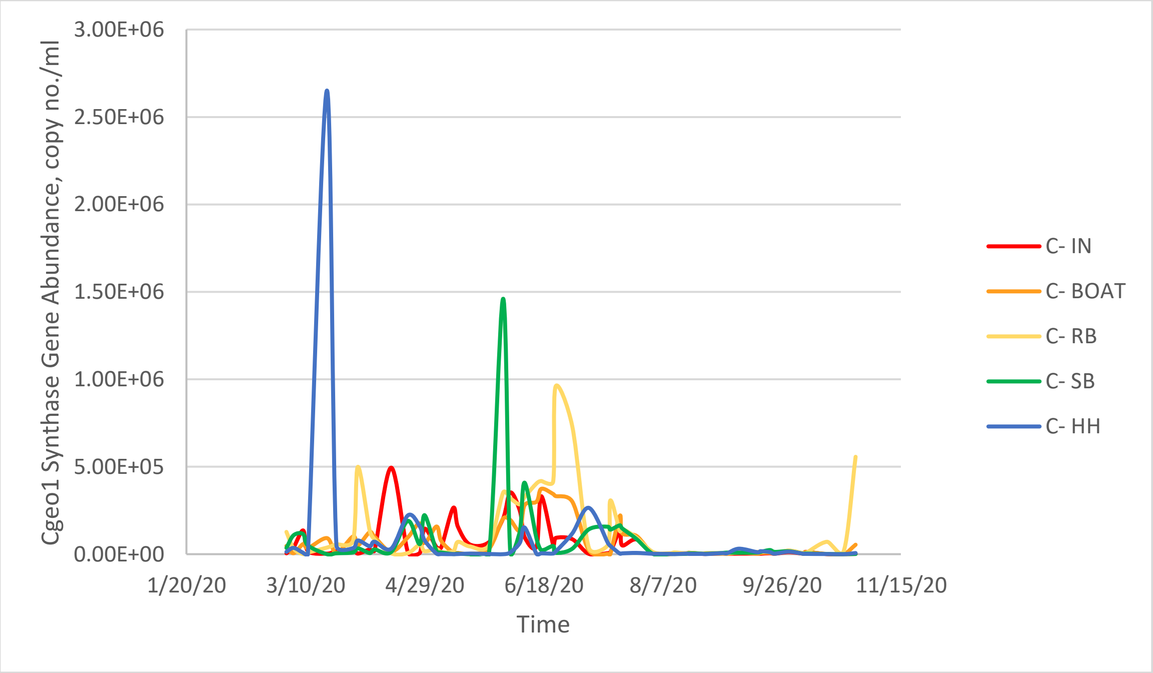


Figure 14. Cgeo1 synthase gene abundance over Time: CWW.

3.4. CART Model Development and Linear Regressions

3.4.1. Pearson's Correlations

Prior to modeling efforts, the individual Pearson's correlations between geosmin levels and water quality parameters were determined to check for existing significant correlations (Table 6). None of the correlations indicated strong predictive capability (r > ±0.5), and multi-parameter models were therefore explored. In addition, the geosmin synthase abundance (as measured using the CGeo1 primer set) was included in these multi-parameter models.

Table 6. Correlations (r) found between geosmin and each environmental variable.

	Auburn	Opelika	Columbus
TP	-0.0009	0.1455	0.1204
TKN	0.3740	0.0000	-0.0418
Nitrite + Nitrate	0.0133	0.0983	0.0993
OP	0.2564	-	-
Water temp	0.0723	-0.2298	0.0077
pH	-0.2695	-0.1081	-0.1879
DO	-0.0406	0.0526	-0.1023
Specific Conductance	0.3804	0.1979	-0.0774
Turbidity	0.1180	0.1150	0.0790
Secchi Depth	-0.3266	-0.3868	0.0253
Chlor-a	-0.2181	0.3388	0.2159
Phycocyanin	-0.2646	-0.2117	0.1675
Air temp	0.3821	-0.4337	-0.0417
Rainfall	0.0082	-0.2500	-0.0218
Wind Speed	-0.1671	0.2959	0.1417
BG Count	-	-	0.1124
Nitrate	-0.0659	0.1929	0.0153
Phosphate	-0.0541	0.0630	0.0117

r	
± 0-0.05	🟥
± 0.05-0.2	🟨
± 0.2-0.4	▨
> ± 0.4	🟩

3.4.2. CART Modeling

3.4.2.1. Auburn

A CART model (Figure 15) was generated using water quality data and geosmin synthase abundance data for Auburn water samples to predict geosmin levels at the Lake Ogletree water intake sampling location. To build the tree, 20 water quality variables were used as input, including each listed variable in Table 3. The variable depth was used twice: once as a measured continuous value and once as a categorical value of either upper, middle, or lower sampling depth. The unpruned tree algorithm uses 2 of those variables due to significance found using the 'rpart' package in R. Specific conductance and phosphate anion concentration were found to be significant in this initial (unpruned) tree. The initial tree has a coefficient of determination, R^2, of 0.4, explaining approximately 40% of variation in geosmin levels. Further, a pruned regression tree was attempted in order to reduce the chances of overfitting the tree to the data and to reduce any complexity of the tree (Yang et al., 2017), but the pruned regression tree was unable to be made, resulting in a root only. This is due to the minimum number of observations that must exist in a node in order for a split to be attempted not being met by this dataset which only had 48 datapoints. This parameter in the 'rpart' package of R can be adjusted manually using the control parameters, though it leads to overfitting of the tree, so it was not employed in this study. It was interesting that the genetic information was not an important factor in the CART model, in contrast to our hypothesis.

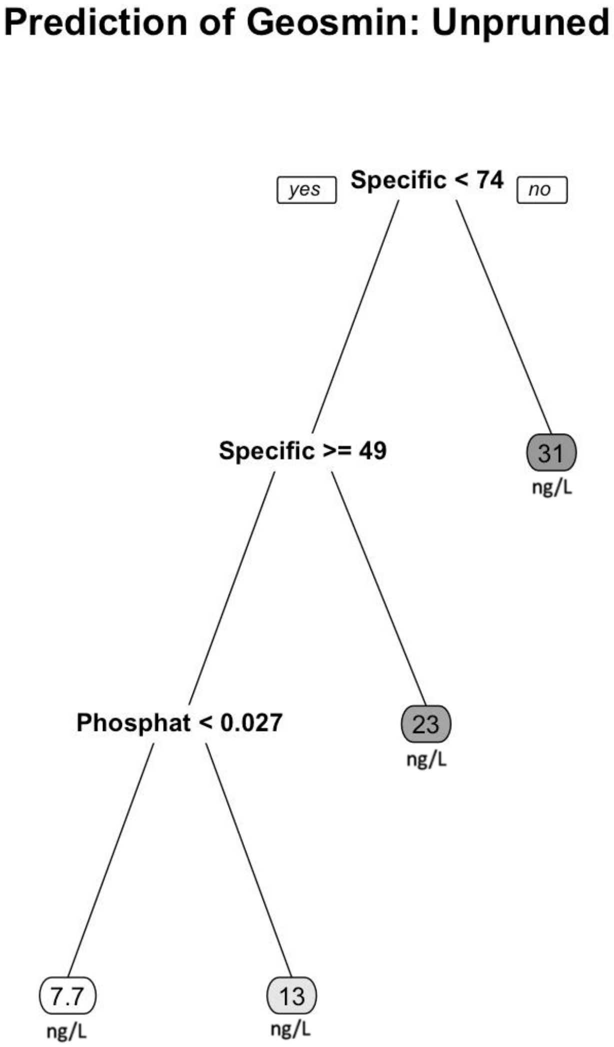
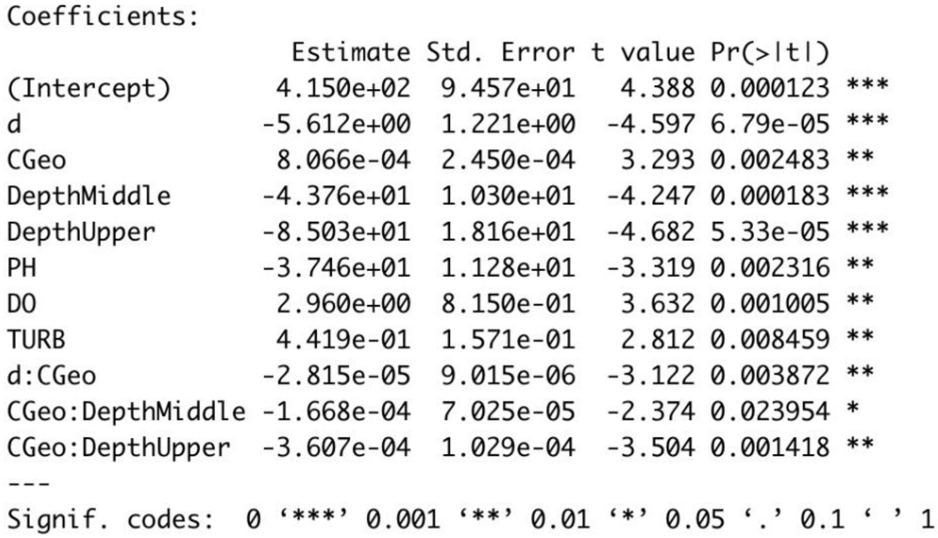


Figure 15. Unpruned CART model for City of Auburn Water Resources Department.

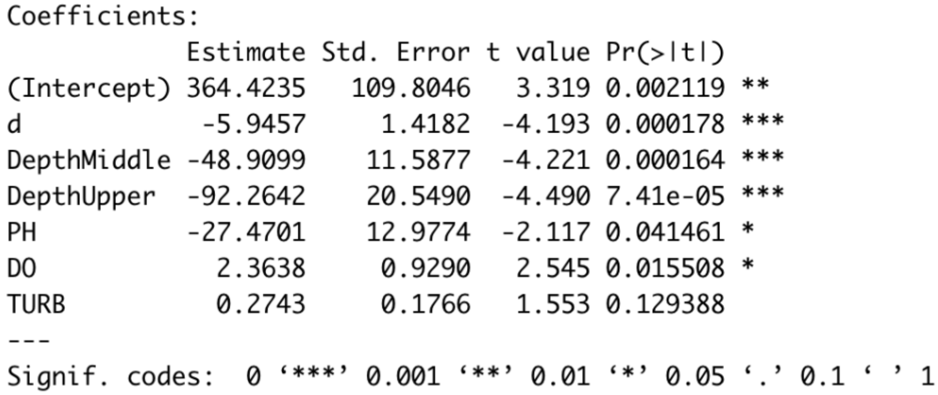
The CART model could potentially be used by the City of Auburn Water Resources Department in being able to determine up to 40% of variation in taste and odor concentrations, allowing them to treat the water in the reservoir. In using this CART model, the use of qPCR would not be necessary, as they could follow the branches with the data on specific conductance which they currently routinely monitor, and phosphate anion levels, which they could also measure. However, there is high probability this model is overfit and won't perform well in future years – additional data points are needed to improve the model. The performance of the model is determined by the variation of the input data from the reservoir samples collected through 2020.

Multiple regression was also conducted on the Auburn dataset using R. To fit the model, stepwise regression was used to choose the predictive variables to be used in the multiple regression. To do so, all 20 original variables from water quality parameters and genetic data were entered into the regression function in R (Figure 16 shows only parameters with p values less than 0.05). The least significant variable was then removed, and this process was repeated multiple times until all variables were included and the adjusted R^2 peaked (Figure 17).



```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.150e+02  9.457e+01   4.388 0.000123 ***
d                -5.612e+00  1.221e+00  -4.597 6.79e-05 ***
CGeo              8.066e-04  2.450e-04   3.293 0.002483 **
DepthMiddle      -4.376e+01  1.030e+01  -4.247 0.000183 ***
DepthUpper       -8.503e+01  1.816e+01  -4.682 5.33e-05 ***
PH               -3.746e+01  1.128e+01  -3.319 0.002316 **
DO                2.960e+00  8.150e-01   3.632 0.001005 **
TURB              4.419e-01  1.571e-01   2.812 0.008459 **
d:CGeo           -2.815e-05  9.015e-06  -3.122 0.003872 **
CGeo:DepthMiddle -1.668e-04  7.025e-05  -2.374 0.023954 *
CGeo:DepthUpper  -3.607e-04  1.029e-04  -3.504 0.001418 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 16. Multiple regression for Auburn with Cgeo1 gene abundance (adjusted R^2 = 0.5405).



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 364.4235   109.8046   3.319 0.002119 **
d            -5.9457     1.4182  -4.193 0.000178 ***
DepthMiddle -48.9099    11.5877  -4.221 0.000164 ***
DepthUpper  -92.2642    20.5490  -4.490 7.41e-05 ***
PH          -27.4701    12.9774  -2.117 0.041461 *
DO            2.3638     0.9290   2.545 0.015508 *
TURB          0.2743     0.1766   1.553 0.129388
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 17. Multiple regression for Auburn without Cgeo1 gene abundance (adjusted R^2 = 0.2879).

It was clear in testing our hypothesis again for whether Cgeo1 gene abundance is an important variable to include in predictions, regressions were made with and without the Cgeo1 variable being included. The variables found to be significant in multiple regression differ from those found to be significant in the CART models for Auburn. Specific conductance and phosphate anion concentration were the most important variables for CART, the depth (continuous), depth (categorical), Cgeo1 gene abundance, pH, dissolved oxygen, turbidity, the interaction between Cgeo1 gene abundance with categorical depth, and the interaction between continuous depth with the Cgeo1 gene abundance are the most significant for the multiple regression with Cgeo1 gene abundance. This regression gave an adjusted R^2 value of 0.5405. When Cgeo1 gene abundance was taken out of the analysis, depth (continuous), depth (categorical), dissolved oxygen, and pH were the only significant variables, with an adjusted R^2 of 0.2879. There was no similar significant value between both multiple regression analyses and the CART model previously made. The adjusted R^2 without using Cgeo1 was lower by 0.2526 from the multiple regression using Cgeo1 gene abundance as a variable, exhibiting the improvement of our predictive modeling with the use of the qPCR for cyanobacteria geosmin gene abundances.

Figure 18 shows single regression of geosmin versus geosmin synthase abundance for each depth. Auburn Upper depth (11.5-17.4 ft.) shows reasonable correlation between the Cgeo1 synthase gene abundance and the detected geosmin levels. Auburn Middle (18-23.9 ft.) and Lower (25.5-30.9 ft.) depths do not show good correlations. This agrees with the multiple regression in showing that the interaction between the Cgeo1 gene abundance and "Auburn upper" depth has a significance of $p < 0.01$, and that the lower depth was insignificant. Cyanobacteria tend to prefer higher temperatures and light availability (Guttman & Rijn, 2008; Oh et al., 2017) and these are associated with the upper level of water bodies. The Cgeo1 primers appear to exclusively amplify geosmin synthase genes in cyanobacteria. The high correlation between gene abundance and geosmin suggests that geosmin levels in the upper layers of Lake Ogletree were driven by cyanobacteria producers. the higher geosmin peaks occurring in the Auburn Upper depth range (Figure 18) help in articulating this correlation.

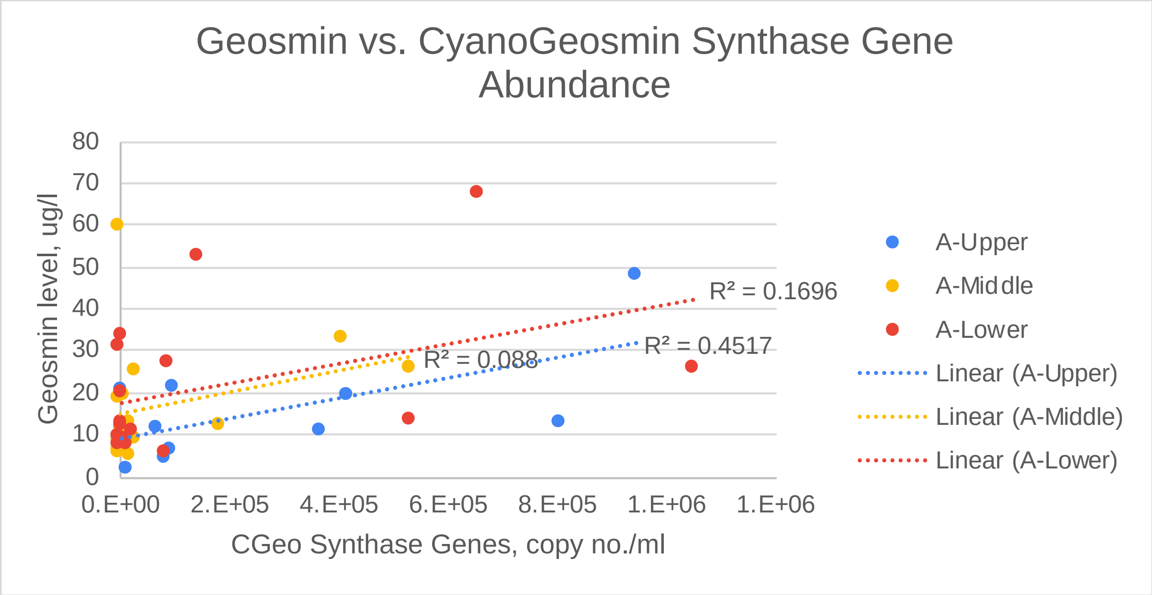


Figure 18. Correlation between cyanobacteria geosmin synthase genes and the detected geosmin level.

Overall, multiple regression with geosmin synthase gene abundance resulted in better predictive capability (adjusted R^2 = 0.5405) compared to CART modeling. It also outperformed the model without gene abundance data.

3.4.2.2. Opelika

The Opelika CART model was generated using the water quality data and genetic data for Opelika Utilities water samples to predict the geosmin levels at the Lake Saugahatchee water intake sampling location. All 19 water quality variables were used for input to build the original tree, with the variable, depth, still being used twice: once as a measured continuous value and once as a categorical value of either upper (0 ft), middle (5 ft), or lower (15 ft) sampling depth. The unpruned tree algorithm found that 2 of those 19 variables were significant (Figure 19).

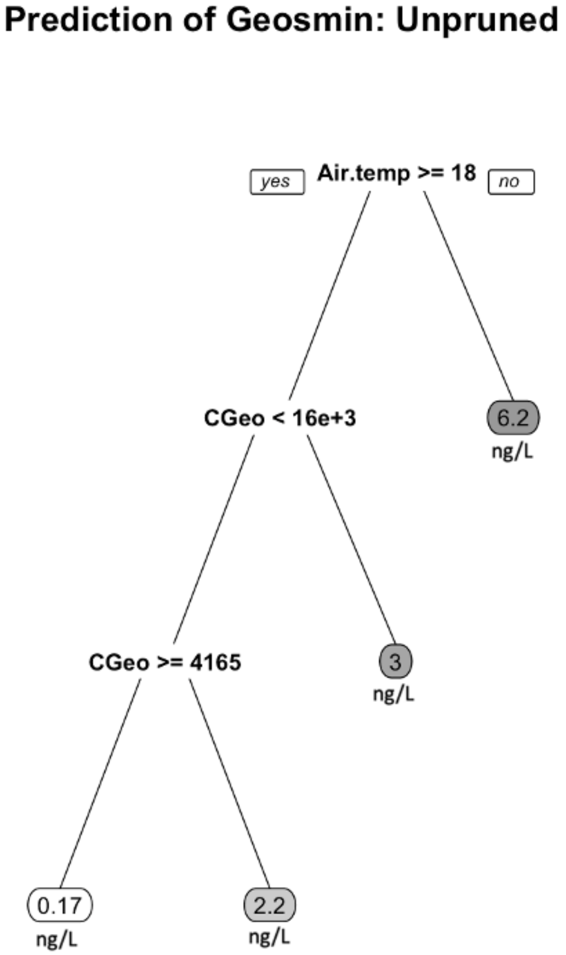


Figure 19. Unpruned CART model for Opelika Utilities.

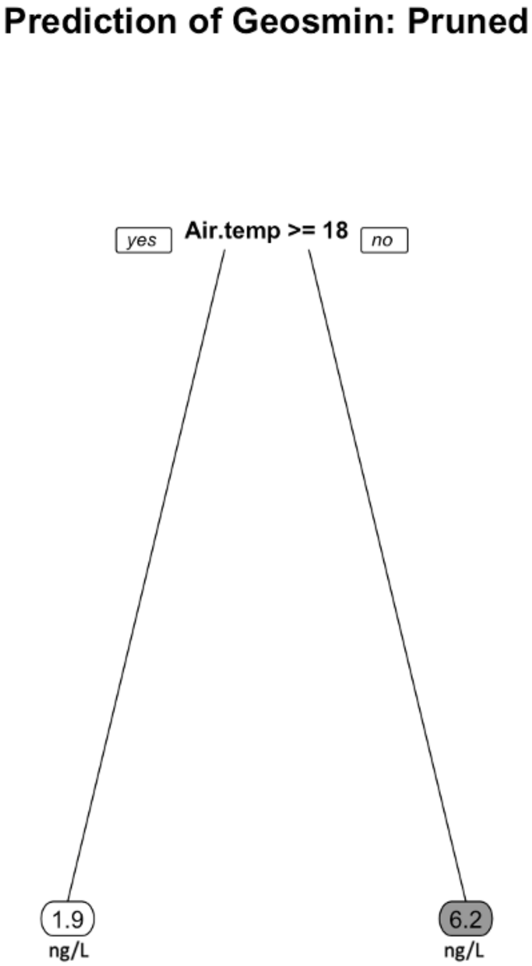
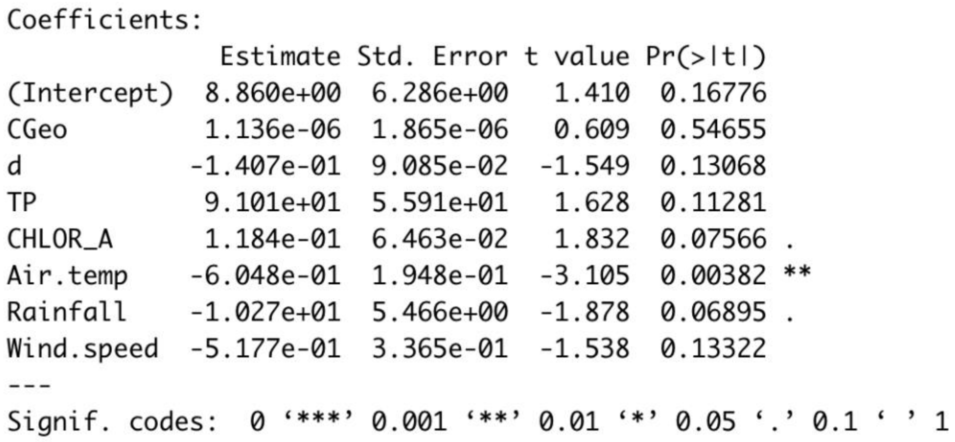


Figure 20. Pruned CART model for Opelika Utilities.

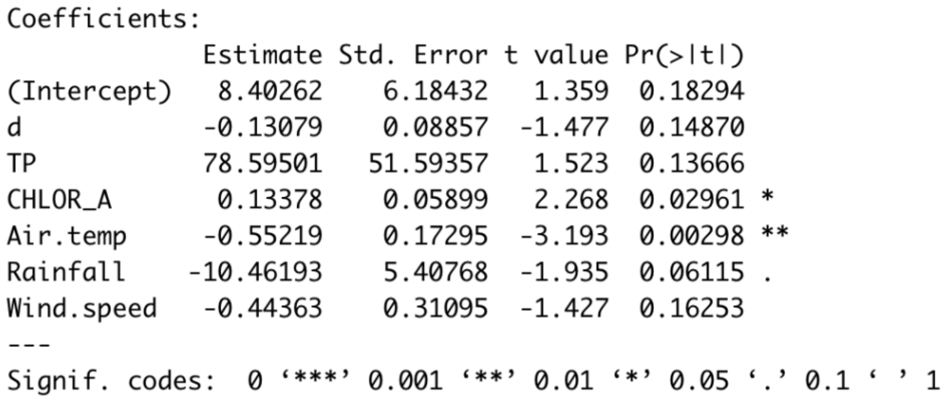
Air temperature and Cgeo1 gene abundance were initially found to be significant. This initial tree has a coefficient of determination, R^2, of 0.6, explaining 60% of the geosmin levels, though those geosmin levels were only predicted at a maximum of 6.2 ng/L, which is not of concern. The pruned regression tree (Figure 20) then shows that only air temperature was an important variable for predicting the geosmin concentrations for Opelika Utilities, then with an R^2 of 0.4. The CART model was then built again without the inclusion of the geosmin synthase gene abundance data to test our hypothesis that the inclusion of genetic data would result in better prediction capabilities. The R^2 remained the same for this new model without Cgeo1 gene abundance included, now with phosphate ion concentration cited as significant. Again though, air temperature was the most significant variable for the model, although it is not logical in the CART model that if the air temperature is below 18° C, the geosmin levels would be higher. It is typical that higher temperatures result in higher geosmin levels due to most cyanobacteria preferring higher temperatures (>20° C) (Oh et al., 2017). These CART models for Opelika were not expected to provide meaningful management information due to the geosmin level having low variation and not once reaching above the threshold value of 10 ng/L for the sampling season.

As with the Auburn data, multiple regression was then performed using data from Opelika Utilities. Stepwise regression was again implemented to choose the predictive variables to be used for the multiple regression, with the significance of each variable also listed in Figures 21 and 22. Out of the 19 input variables, 3 were found to be significant: air temperature ($P<0.01$), chlorophyll-a ($P<0.1$), and rainfall ($P<0.1$). Air temperature was the only variable that is comparable to the significant variable found in the CART model. With Cgeo1 gene abundance included in the multiple regression the adjusted R^2 was found to be 0.3537, and without Cgeo1 gene abundance the adjusted R^2 was 0.3653, not surprising given that it was not a significant parameter in the model. Further removal of insignificant variables in the multiple regressions begins to decrease the adjusted R^2, so the significance ($P<1$) is noted. Overall, the models for Lake Saugahatchee are not likely to be helpful management tools given that they predict geosmin levels below the threshold for human detection.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.860e+00  6.286e+00   1.410  0.16776
CGeo         1.136e-06  1.865e-06   0.609  0.54655
d           -1.407e-01  9.085e-02  -1.549  0.13068
TP           9.101e+01  5.591e+01   1.628  0.11281
CHLOR_A      1.184e-01  6.463e-02   1.832  0.07566 .
Air.temp    -6.048e-01  1.948e-01  -3.105  0.00382 **
Rainfall    -1.027e+01  5.466e+00  -1.878  0.06895 .
Wind.speed  -5.177e-01  3.365e-01  -1.538  0.13322
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 21. Multiple regression for Opelika Utilities with Cgeo1 (adjusted $R^2 = 0.3537$).



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.40262    6.18432   1.359  0.18294
d            -0.13079    0.08857  -1.477  0.14870
TP           78.59501   51.59357   1.523  0.13666
CHLOR_A       0.13378    0.05899   2.268  0.02961 *
Air.temp     -0.55219    0.17295  -3.193  0.00298 **
Rainfall    -10.46193    5.40768  -1.935  0.06115 .
Wind.speed   -0.44363    0.31095  -1.427  0.16253
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 22. Multiple regression for Opelika Utilities without Cgeo1 (adjusted $R^2 = 0.3653$).

3.4.2.3. Columbus

CART models were also created using the water quality data and genetic data for Columbus Water Works. Columbus samples come from five different locations within Lake Oliver: Intake, Boathouse, Roaring Branch, Standing Boy, and Heiferhorn. To build the trees in Figures 23 and 24, 18 variables were used as input. For this utility, the variable depth was only used as a continuous measurement, at either 0 feet (Boathouse, Roaring Branch, Standing Boy, and Heiferhorn sampling) or 20 feet deep (Intake sampling).

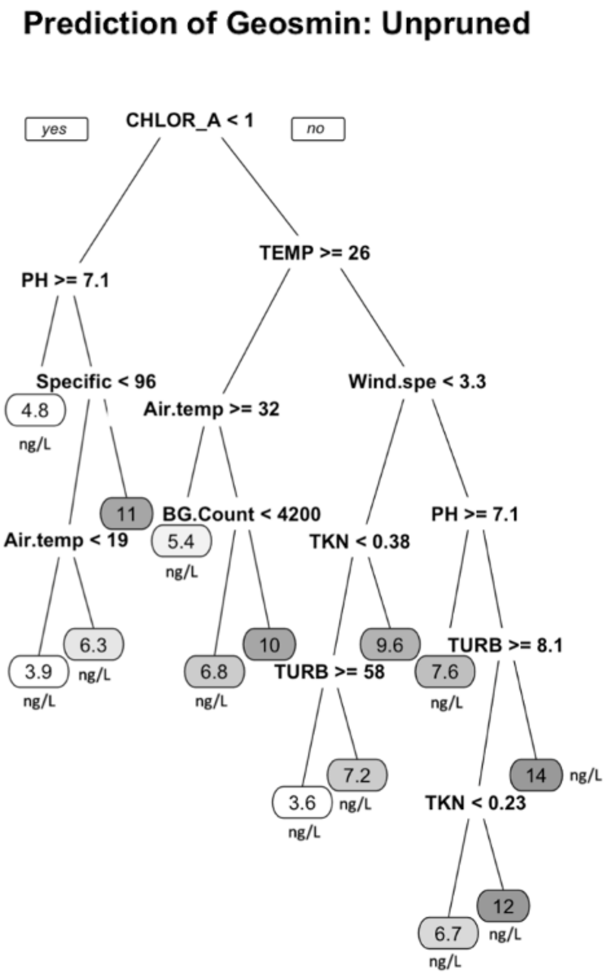


Figure 23. Unpruned CART model for CWW.

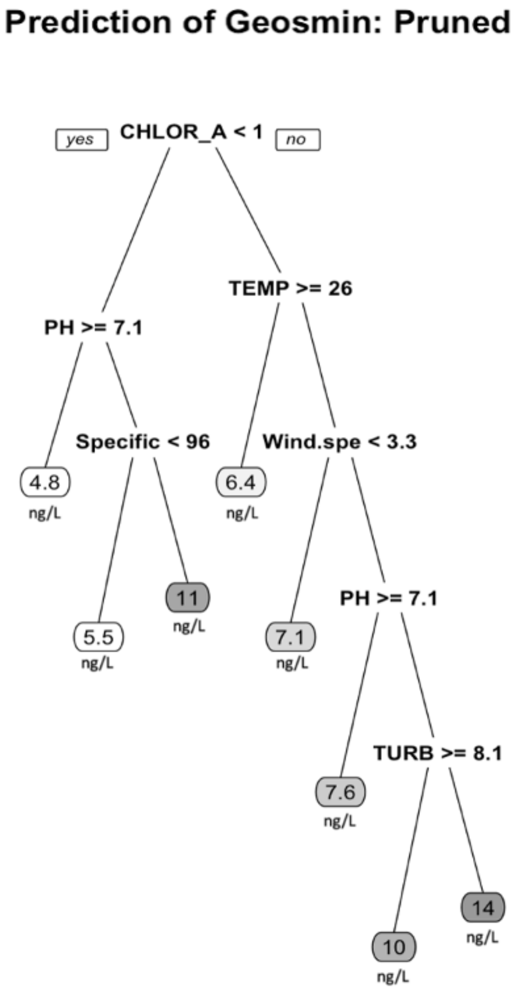


Figure 24. Pruned CART model for CWW.

The above figures (Figures 23 and 24) do not take individual sampling location into account. Out of those 18 input variables, 9 were found to be significant initially in the unpruned CART model. These include chlorophyll-a concentration, pH, water temperature, specific conductance, wind speed, air temperature, blue-green algae count (BG Count), Total Kjeldahl Nitrogen (TKN), and turbidity. The unpruned tree has an R^2, of 0.4, explaining 40% of the geosmin levels. These geosmin levels predicted by the model do not exceed 14 ng/L. Humans can begin to taste these compounds at 10 ng/L, but truly become an issue for complains to the utilities at around 20-30 ng/L. Thus, the usefulness of this model in predicting significant outbreaks is limited. Further, the pruned CART model narrows the significant variables to now exclude air temperature, BG Count, and TKN. The pruned tree also has an R^2 of 0.4. The inclusion of synthase gene abundance data did not improve the CART models.

To evaluate the significance of location being used as a variable, it was then included in the CART model, seen in Figures 25 and 26 below. Here, the only significant variables were initially location, dissolved oxygen, chlorophyll-a levels, nitrate anion concentration, pH, water temperature, turbidity, and wind speed. Both models have R^2 values of 0.4, though the pruned tree in this scenario was much less intricate.

Either pruned CART model could be used by Columbus Water Works in being able to predict up to 40% of observed geosmin levels. Model performance again relies on the variation in geosmin levels in the drinking water reservoirs, and Columbus Water Works only surpassed the 10 ng/L threshold 16.2% of the time, and only surpassed 20 ng/L 0.87%. This likely explains the low utility of the CART model in predicting outbreaks of significance.

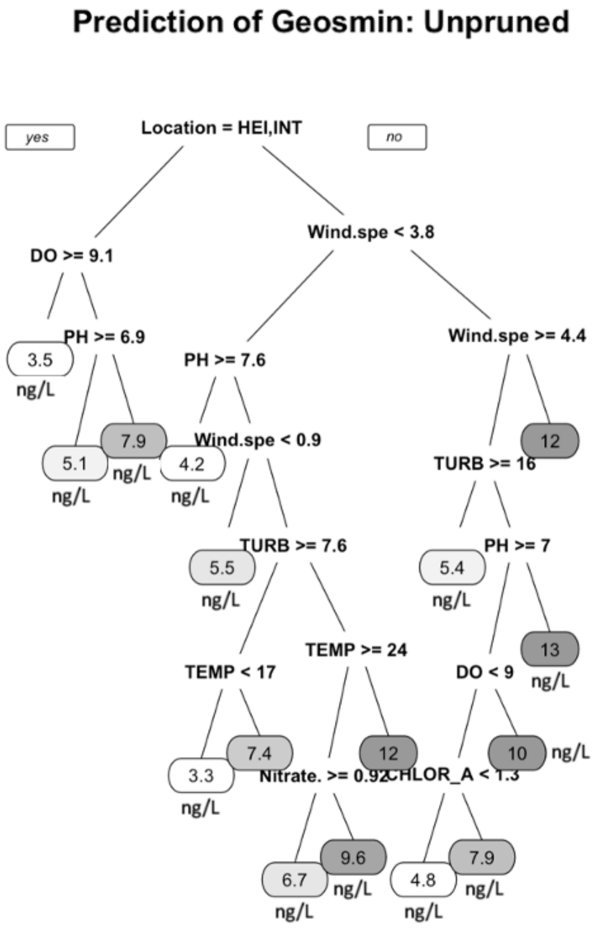


Figure 25. Unpruned CART model for CWW with location (HEI represents the Heiferhorn sampling location, INT represents the Intake sampling location).

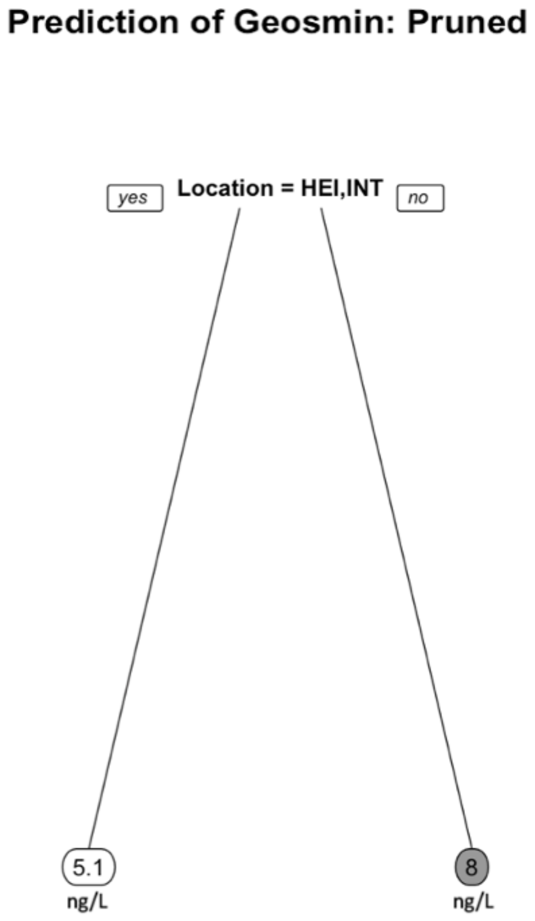
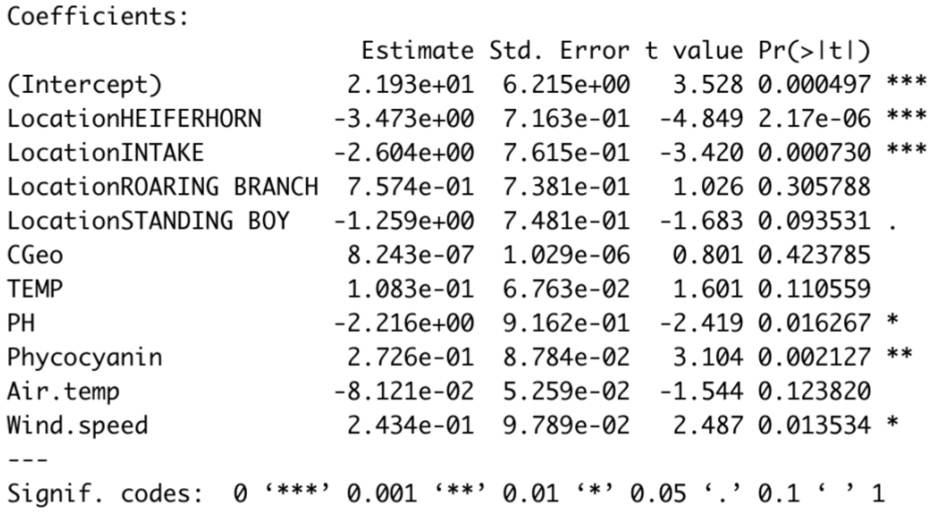


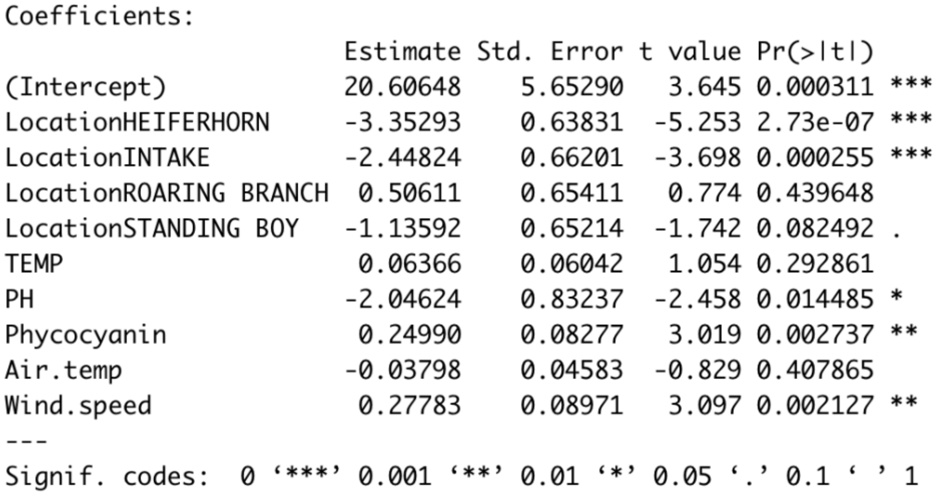
Figure 26. Pruned CART model for CWW with location (HEI represents the Heiferhorn sampling location, INT represents the Intake sampling location).

Multiple regression analysis was performed on CWW water data to determine the significant variables and their relationship to the prediction of geosmin levels. The same stepwise regression steps were implemented to finalize the correct input variables for the multiple regression. Among the 19 input variables, only location, pH, phycocyanin, and wind speed were considered statistically significant ($p < 0.05$). For Columbus, location and depth are seen as the same variable due to the only depths being 20 feet at the intake, and 0 feet at every other sampling location. With Cgeo1 gene abundance included in the multiple regression the adjusted R^2 was 0.2311 (Fig. 27), and without Cgeo1 gene abundance the adjusted R^2 was 0.2016 (Fig. 28). This small difference reflects the fact that synthase gene abundance was not a significant factor in the model. Water temperature and pH were the only variables that were both of significance in comparison between the multiple regression and the pruned CART model.



```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            2.193e+01  6.215e+00   3.528 0.000497 ***
LocationHEIFERHORN    -3.473e+00  7.163e-01  -4.849 2.17e-06 ***
LocationINTAKE        -2.604e+00  7.615e-01  -3.420 0.000730 ***
LocationROARING BRANCH  7.574e-01  7.381e-01   1.026 0.305788    
LocationSTANDING BOY  -1.259e+00  7.481e-01  -1.683 0.093531 .  
CGeo                   8.243e-07  1.029e-06   0.801 0.423785    
TEMP                   1.083e-01  6.763e-02   1.601 0.110559    
PH                    -2.216e+00  9.162e-01  -2.419 0.016267 *  
Phycocyanin            2.726e-01  8.784e-02   3.104 0.002127 ** 
Air.temp              -8.121e-02  5.259e-02  -1.544 0.123820    
Wind.speed             2.434e-01  9.789e-02   2.487 0.013534 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

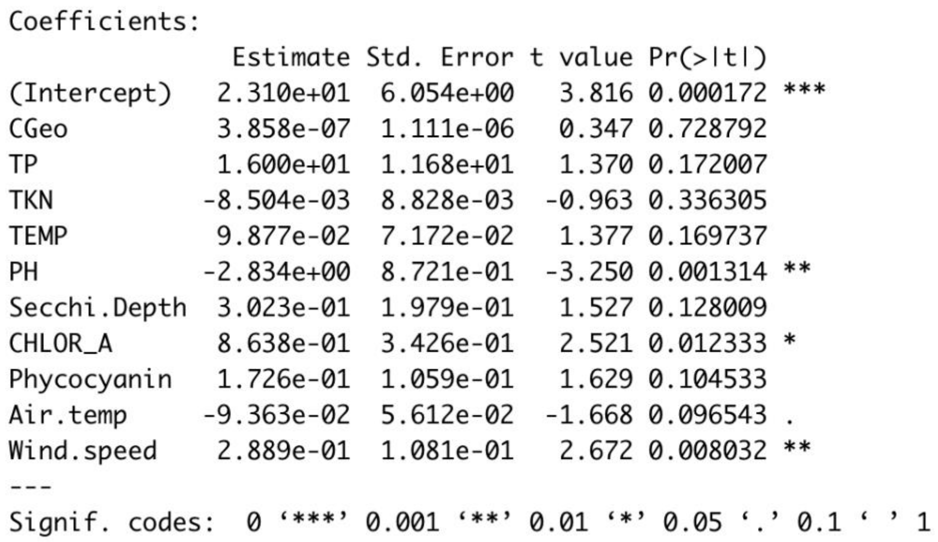
Figure 27. Multiple regression for CWW with Cgeo1, including location (adjusted $R^2 = 0.2311$).



```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          20.60648    5.65290   3.645 0.000311 ***
LocationHEIFERHORN   -3.35293    0.63831  -5.253 2.73e-07 ***
LocationINTAKE       -2.44824    0.66201  -3.698 0.000255 ***
LocationROARING BRANCH  0.50611    0.65411   0.774 0.439648    
LocationSTANDING BOY -1.13592    0.65214  -1.742 0.082492 .  
TEMP                  0.06366    0.06042   1.054 0.292861    
PH                   -2.04624    0.83237  -2.458 0.014485 *  
Phycocyanin           0.24990    0.08277   3.019 0.002737 ** 
Air.temp             -0.03798    0.04583  -0.829 0.407865    
Wind.speed            0.27783    0.08971   3.097 0.002127 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

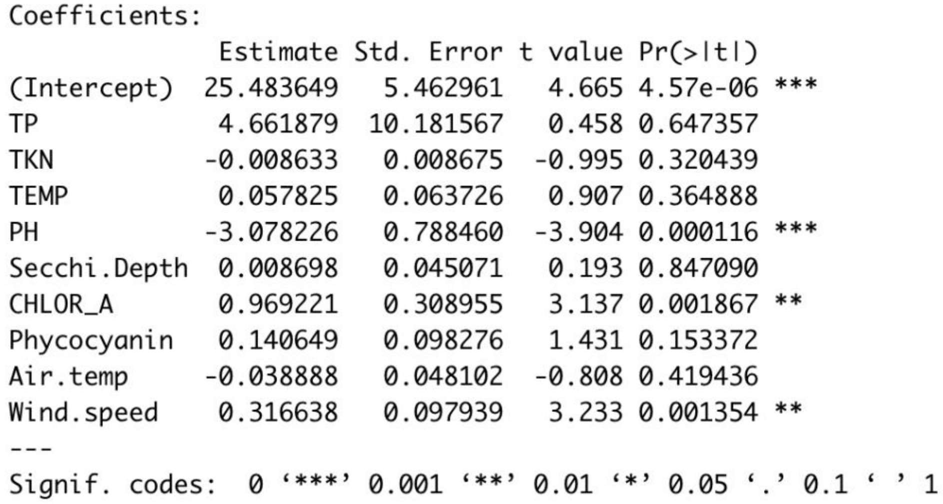
Figure 28. Multiple regression for CWW without Cgeo1, including location (adjusted R^2 = 0.2016).

I also performed multiple regression for the entire Lake Oliver dataset with the location and water depth variables excluded. With Cgeo1 gene abundance included, the adjusted R^2 was 0.1277 and it was not a significant model parameter. When Cgeo1 gene abundance was removed from the model, the adjusted R^2 goes down to 0.1131. The significant variables can be seen in Figures 29 and 30, with pH, wind speed, chlorophyll-a, and air temperature showing the most significance in both scenarios. We again had the situation where removing any further least significant values begins to reduce the adjusted R^2 value, though neither of these regressions show high adjusted R^2 values.



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.310e+01  6.054e+00   3.816 0.000172 ***
CGeo          3.858e-07  1.111e-06   0.347 0.728792    
TP            1.600e+01  1.168e+01   1.370 0.172007    
TKN          -8.504e-03  8.828e-03  -0.963 0.336305    
TEMP          9.877e-02  7.172e-02   1.377 0.169737    
PH           -2.834e+00  8.721e-01  -3.250 0.001314 ** 
Secchi.Depth  3.023e-01  1.979e-01   1.527 0.128009    
CHLOR_A       8.638e-01  3.426e-01   2.521 0.012333 *  
Phycocyanin   1.726e-01  1.059e-01   1.629 0.104533    
Air.temp     -9.363e-02  5.612e-02  -1.668 0.096543 .  
Wind.speed    2.889e-01  1.081e-01   2.672 0.008032 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 29. Multiple regression for CWW with Cgeo1, excluding location (adjusted $R^2 = 0.1277$).



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  25.483649   5.462961   4.665 4.57e-06 ***
TP            4.661879  10.181567   0.458 0.647357    
TKN          -0.008633   0.008675  -0.995 0.320439    
TEMP          0.057825   0.063726   0.907 0.364888    
PH           -3.078226   0.788460  -3.904 0.000116 ***
Secchi.Depth  0.008698   0.045071   0.193 0.847090    
CHLOR_A       0.969221   0.308955   3.137 0.001867 ** 
Phycocyanin   0.140649   0.098276   1.431 0.153372    
Air.temp     -0.038888   0.048102  -0.808 0.419436    
Wind.speed    0.316638   0.097939   3.233 0.001354 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 30. Multiple regression for CWW without Cgeo1, excluding location (adjusted $R^2 = 0.1131$).

3.4.3. CART Summary

To compare, the empirical models that Dzialowski et al. (2009) developed for each of their individual reservoirs in Kansas had a range of R^2 values, with the low being 0 (unable to be made) and the high being 0.94 using a two-variable equation with water quality variables. Christensen (2006) also developed a two-variable equation to predict geosmin levels, which gave an R^2 of 0.709, and another regression model for geosmin (Mau et al., 2004) gave an R^2 of 0.70 (p-value = 0.0016). A simple multiple linear regression had a reasonable predictive capability for geosmin (R^2 = 0.657, P<0.001) using water quality parameters (Parinet et al., 2013). More advanced modeling, like random forest models, have been found to have R^2 values of 0.71 to predict the threshold odor number for MIB (Wang et al., 2019), R^2 of 0.81 for geosmin (Harris & Graham, 2017), though Harris and Graham got lower R^2 values for their Support Vector Machine, Boosted Tree, and Cubist models (R^2 = 0.34, 0.34, 0.75, respectively). Kehoe et al (2015) developed linear regression and random forest models for TON (threshold odor number), with the linear regression having low and high R^2 values of 0.61 and 0.71, and the random forest model having low and high R^2 values of 0.48 and 0.52. When comparing these previously found predictive capabilities of models, our values of 0.4 demonstrate low predictive power using the CART modeling with the current low geosmin level datasets (Table 7). The Opelika and Columbus multiple regressions also have low predictive power, though the Auburn dataset has reasonable predictive power in the multiple regression.

Table 7. Summary of CART model R^2 and multiple regression adjusted R^2 values using R software.

	CART Unpruned	CART Pruned	Multiple Regression with Cgeo1	Multiple Regression without Cgeo1
City of Auburn Water Resources Department	0.4	-	0.5405	0.2879
Opelika Utilities	0.6	0.4	0.3537	0.3653
Columbus Water Works (Locations Combined)	0.4	0.4	0.1277	0.1131
Columbus Water Works (Separate Locations)	0.4	0.4	0.2311	0.2016

4. *Sequencing*

Select samples from Auburn, Opelika, and Columbus that had higher geosmin levels were sent for Sanger sequencing of the product amplified by the CGeo1 primer set. The top likely geosmin producers from the episodes in the upper layer of Lake Ogletree were *Anabaena* and *Planktothrix* during March and May 2020, whereas *Planktothrix* was the only dominant organism in April 2020 for the upper layer. In March in Lake Ogletree, there were small amounts (<5%) of *Aphanizomenon* and *Nostoc* found, and in May there was *Aphanizomenon* present. The likely producers found from Lake Oliver at the Boathouse sampling location during April 2020 includes *Planktothrix* (85%) and *Anabaena* (15%), and Lake Saugahatchee sequenced to almost 97% *Planktothrix* at 5-foot depth during April 2020. *Dolichospermum* (*Anabaena*) *planctonicum* and *Planktothrix* sp. 328 represent the majority. *Anabaena* and *Aphanizomenon* are a widely known cyanobacteria which produces musty-odor compounds, which exist as pelagic plankton and can produce toxins (Watson et al., 2008). *Planktothrix* is also a widely known T&O producing cyanobacteria which accumulated in large blooms as planktonic and/or benthic species and have the ability to produce toxins, dependent on the species. With this knowledge, the individual reservoirs have the ability to apply mechanical, biological, chemical or physical methods to control the growth of these specific T&O producers.

5. *Conclusions*

The inclusion of geosmin synthase gene abundance data was most effective at predicting geosmin in Auburn's Lake Ogletree. This was the only reservoir to experience high (>30 ng/L) geosmin levels. The datasets from Lake Saugahatchee (Opelika) and Lake Oliver (Columbus) had low geosmin variability and rarely experienced geosmin levels above the threshold value for human detection. The multiple regression model for Auburn had the best model fit with an adjusted R^2 of 0.5405 when geosmin synthase gene abundance was included. Models predicting geosmin levels for Opelika and Columbus had lower predictive power, due to the low geosmin concentrations detected through the sampling period. The results of this study show significant correlations between geosmin and a few water quality parameters.

- For Auburn CART modeling, air temperature, specific conductance, depth, and phosphate anion abundance were relevant. For the City of Auburn Water Resources Department multiple regression, dissolved oxygen, depth, pH, and Cgeo1 gene abundance were significant.

- For Opelika Utilities' CART modeling, air temperature and Cgeo1 gene abundance were included, and their multiple regression included chlorophyll-a, air temperature, and rainfall, though Opelika's predictive tools were the least useful.
- CART modeling for CWW found the most relationships for geosmin level determination, including the variables chlorophyll-a, pH, water temperature, specific conductance, wind speed, air temperature, blue-green algae count, turbidity, and TKN levels. Columbus was also found to have significance in the location of the water sampling. In CWW's multiple regression, pH, wind speed, air temperature, and chlorophyll-a levels showed some significance, with location also being an applicable factor.

The most significant finding in this section of study is the relationship between Cgeo1 gene abundance and geosmin levels for Auburn's multiple regressions. With this gene abundance included, the fit was 54.05%, whereas without the gene abundance the fit was only 28.79%. In order to create models with better fits for the other two reservoirs, it is necessary to collect water samples with higher geosmin levels. It is likely that our molecular approach is the most effective under these conditions.

CHAPTER 3: MIB PRIMER EVALUATION

1. Introduction

Taste and odor episodes are an important issue when it comes to drinking water reservoirs, and cyanobacterial 2-methyisoborneal (MIB) is one of the most commonly detected and problematic taste and odor compounds worldwide (Wang et al., 2019). This terpenoid compound is not harmful to the human body, occurrences of increased T&O compounds lead to surges in customer complaints and often decreases the public's confidence in the quality and safety of their water resource. The consequence, just as with geosmin, is that the regional utility must then undertake the cost of more advanced water treatment to remove the recalcitrant compound. To develop improved water quality treatment, an early detection and monitoring system for these MIB events is necessary. For the detection of source organisms at low concentrations, quantitative PCR has proven, in recent research, to be one of the most promising tools (Devi et al., 2020). Cyanobacteria are the most widely attributed organism in the production of MIB in freshwater system, though not all cyanobacteria are responsible for them (Asquith et al., 2018), with only a small percentage of cyanobacteria actually producing MIB (Jüttner & Watson, 2007). One of the major challenges with carrying out qPCR assays for MIB synthase is design of primers that are both specific to this gene yet offer coverage of multiple taxa. There is a gap in current knowledge on MIB synthase gene sequencing, with only 28 of the 72 cyanobacterial synthesis regions having been researched and input into the NCBI database. Moreover, MIB synthase has fewer conserved regions than geosmin synthase, further limiting opportunities for primer design (Devi et al., 2020). Consequently, there is difficulty in developing the design of a universal primer that amplifies MIB synthase. Primers targeting the MIB synthase gene are already published in the literature as reviewed by Devi et al. and their *in-silico* analysis suggests these primer sets may offer poor coverage across multiple taxa. Here, we expanded on their work by testing multiple published primers on environmental samples existing moderate to high levels of MIB and evaluating PCR efficiency and specificity to the target. Establishing these metrics is a necessary step for using these molecular tools for future sequencing and modeling efforts, similar to what we have done with geosmin synthase in Chapter 2.

MIB gene synthesis is most frequently found as a secondary metabolite of actinomycetes, filamentous cyanobacteria, myxobacteria, and some fungi in freshwater (Komatsu et al., 2008). Suurnäkki et al. (2015), developed a primer set for detecting MIB synthase in qPCR, called MIB3324, based on gene sequences from *Oscillatoria*, *Planktotricoides*, and *Pseudanabaena*. The primer was tested and analyzed using a database to confirm it was MIB synthase specific. MIB was detected in *Oscillatoria* and *Planktothrix* strains via SPME GC/MS, and from the qPCR the same strains were identified as producers. Consequently, the authors concluded that their developed MIB primer set accurately targeted only MIB-producing cyanobacteria (Suurnäkki et al, 2015). The shortcoming of this work is that they were only able to identify an MIB primer that targets a few genera of cyanobacteria, and none that amplify MIB synthase in actinobacteria. They also were only able to identify two genera, whereas they state that there are 8 known genera of producers. Consequently, other researchers have continued to develop new primer tools targeting MIB synthase. Gaget et al. (2020) was able to develop and validate a primer set to be used in qPCR for the detection of the MIB synthase gene in cyanobacteria, which was developed using a reservoir in Australia. With the high specificity for primers for planktonic species, the detection method produced should allow for detection of low cell numbers and be easily applicable on a range of environmental samples (Gaget et al., 2020). Another study by Wang et al. (2014) in China established SYBR Green qPCR assays for field monitoring of cyanobacterial MIB producers by targeting the *mic* gene that could be used for the early detection of T&O episodes, and Wang et al. (2016) developed primers for qPCR sequencing of the actinobacterial MIB producers by targeting the pentalenene synthase in *Streptomyces*. These developed protocols for qPCR in the detection of MIB producers in water bodies will be useful in the future for T&O prediction models.

A previous study focused on reviewing the current status of developed qPCR primers and probes in identifying the cyanobacterial blooms along with geosmin and MIB events. The review by Devi et al., explains how majority of the current research on MIB producers is on the cyanobacterial genera, with *Anabaena*, *Oscillatoria*, *Planktothrix*, *Pseudanabaena*, and *Phormidium* as the top producers researched, though none of the developed primers is universal for all MIB producing cyanobacteria based on their *in-silico* analysis mapping these primers onto MIB synthase sequences in the NCBI database. The review concludes that SYBR green qPCR detection methods have high specificity and can quantify low cell numbers while dealing with multiple samples in a single run (Devi et al., 2020). It has also previously been concluded that the gene-based approach for MIB event systems is more accurate and specific compared to the conventional laborious cell count method (Lu et al., 2019). This molecular method developed is therefore a useful tool in monitoring cyanobacterial and actinobacterial producers of the MIB gene synthase. Our process described below aimed to evaluate formerly developed primer sets on samples taken from our local drinking water reservoirs to evaluate their efficiencies and specificities. We also sequenced the amplified products to identify what taxa are represented among the amplified gene products. A group of samples with moderate to high MIB (ng/L) levels detected by the utilities and fishpond samples was run with SYBR green qPCR to detect the MIB gene copy numbers and this was correlated to the MIB concentration in the water column. Having precise primer sets assists in the universal goal of knowing the producers and factors influencing the proliferation of T&O synthesis so as to better prevent the off-flavor problems more accurately in the future.

2. Methods

2.1. Reservoir Geography and Water Sample Collection

Again, the three utilities we paired with for this project were the City of Auburn Water Resources Department (Auburn) located in Auburn, Alabama, Opelika Utilities located in Opelika, Alabama, and Columbus Water Works (CWW) located in Columbus, Georgia. The 100-200 ml additional samples collected were delivered to our lab to further molecular research, including the subsequent MIB primer evaluation. The goal was to evaluate MIB primers for these utilities to be used in successive research leading to potential MIB modeling, similar to that which we completed for geosmin. See Figures 3 and 4 for more detailed information and figures on the reservoir geography and water sample collection. One sample from an aquaculture pond with high MIB concentration was also included in this analysis. This sample was collected by Alan Wilson.

2.2. Standard Methods of Characterization by Utilities and Pearson's Correlations

All three of the utilities agreed to participate in more intensive sampling and analysis described in detail in Table 3, including the detection of MIB levels (ng/L) in each of their reservoirs. The MIB quantification was performed by gas chromatography/mass spectrometry (GCMS) using external laboratories, and by in-house analysis by Columbus Water Works, which are further described in Chapter 2, section 2.2 above. With each of the values for these water quality variables recorded, simple Pearson's correlations were carried out to examine correlations between water quality parameters and the MIB levels detected.

2.3. Water Filtration and DNA Extraction

The water samples were filtered using the 0.2 mm nitrocellulose (VWR) filters to concentrate the solid material. The filtered amount, up to 200 ml for each sample, was recorded. After each filtration, the vacuum flask was cleansed with deionized water and nano-pure water between each filtration, and a new filter was applied to the apparatus to continue the process. The filter for each sample was then cut into strips to be used for the extraction of DNA. See above section 3b for the protocol for DNA extraction using the PowerSoil Kit to capture total genomic DNA, which allowed for further sample analysis for MIB primer evaluation.

2.4. qPCR and DNA Quantification

2.4.1. Primer Sets

To target the specific gene for MIB synthesis in samples from reservoirs, primer sets are necessary. Existing primer sets for MIB have been developed and reported in the literature (Table 8), though there are concerns about specificity or the exclusion of important MIB-producing taxa.

Table 8. Primer sets from literature and those developed in our lab.

Primer Set	Target Taxa	Reference
MIB3324	Cyanobacteria	Suurnäkki et al., 2015
Gaget	Cyanobacteria	Gaget et al., 2020
MIB-Rf/Rr	Cyanobacteria	Wang et al., 2016
Str-Rf/Rr	Actinobacteria	Wang et al., 2016

The above listed primer sets were shown to be useful in targeting both cyanobacterial and actinobacterial produced MIB synthase genes. The MIB3324F/4050R (noted as MIB3324) primer set was designed to amplify the MIB synthase gene similarly based on the alignment of the sequences from *Oscillatoria limosa* LBD305 (HQ630885), *Planktotricoides raciborskii* CHAB3331 (HQ830029), *Pseudanabaena* sp. dqh152 (HQ830028), *Pseudanabaena* sp. NIVA-CYA111 (HQ630887) and *Pseudanabaena limnetica* str. Castaic Lake (HQ630883) (Suurnäkki et al., 2015). Suurnäkki et al performed *in-silico* analysis on the primers against putative cyanobacterial, proteobacterial, and actinobacterial producers available in the nr database using Primer-BLAST to ensure the designed primer was specific to cyanobacterial MIB synthase. The Gaget primer set was also made to target the MIB synthase gene, using the cyanobacterial positive control of *Pseudanabaena galeata*, and was one of the most efficient assays that was developed for qPCR (Gaget et al., 2020). The MIB-Rf/Rr primer set by Wang et al. was developed to target cyanobacterial MIB synthase during SYBR green detection in qPCR, and it specifically targeted fragments from *Pseudanabaena* sp., *Planktothricoides raciborskii*, *Planktothricoides* sp., and *Leptolyngbya* sp., verifying its specificity and wide coverage for MIB-producing cyanobacterial species. The Str-Rf/Rr primer was made specific to *Streptomyces* spp. for MIB-producer detection (Wang et al., 2015). Table 9 references their forward and reverse sequences used. These primer sets were purchased from Invitrogen (Thermo).

Table 9. Primer sets from literature and their forward and reverse sequences.

Primer Set	Forward Sequence	Reverse Sequence
MIB3324	CATTACCGAGCGATTCAACGAGC	CCGCAATCTGTAGCACCATGTTGA
Gaget	CAGCACGACAGCTTCTACACCTCCATGAC	GGTGGCTGCTCGTCTGCCAGATC
MIB-Rf/Rr	CGACAGCTTCTACAYCYCCATGAC	CGCCGCAATCTGTAGCACCAT
Str-Rf/Rr	GGTGGACGACYKCTACTGCGAG	CAGGGVCGGAAGTTGTTGAA

Geneious Prime software (www.geneious.com/features/) was used to determine potential amplicons of each primer set through alignment with known MIB synthase gene fragments. The molecular weight and length of each amplicon was then calculated (Table 10).

Table 10. Molecular weights and expected base-pair (bp) length of amplicons for each primer set from literature.

Primer Set	MW (µg/mol)	Expected Amplicon Length (bp)
MIB3324	223,175	726
Gaget	7,900	179
MIB-Rf/Rr	6,784	202
Str-Rf/Rr	6,520	339

Initial rounds of PCR were carried out on the environmental samples to generate an initial product that was pooled and cleaned using the QIAquick PCR Purification Kit (Qiagen). The resulting product was quantified using the Promega QuantiFluor dsDNA System and serially diluted to create DNA standards for subsequent qPCR. The Promega QuantiFluor dsDNA System enables sensitive quantification of small amounts of double-stranded DNA (dsDNA) in a purified sample using a dye-based system. A multi-well detection instrument for measuring fluorescence, nuclease-free water, a flat-bottom 96-well plate, and 1.5 ml tubes for standards preparation were each needed for this procedure. Briefly, a 1X TE buffer was made by diluting the given 20x TE buffer 20-fold with nuclease-free water, and a working solution was made by diluting the QuantiFluor dsDNA dye 1:400 in 1x TE buffer. A standard curve was made to result in 0.05-200 ng/well because the quantification process requires the comparison of the unknown samples (primers) to a dsDNA standard curve using the Lambda DNA Standard. 200 ul of the working solution was pipetted into each well intended for quantification of standard samples, unknown samples, and the blanks for comparison, with 10 ul of each standard, sample, and blank (1x TE buffer) into each well. After a 5-second plate shake and 5-minute incubation period, the fluorescence was read at 504 nm/531 nm with the plate reader. The dsDNA concentration was then calculated using the standard curve to be used in further efficiency calculations for each primer set. This allowed for calculation of the gene copy number per ml of DNA reference material for each primer set.

2.4.2. qPCR

qPCR was carried out using the qTower3 by Analytic Jena to quantify the abundance of the MIB synthase genes in the DNA extractions from the samples collected. Each primer melting temperature was found using the IDT OligoAnalyzer tool, and each primer set was then run for qPCR using an annealing temperature 2° C lower and 2° C higher than the melting temperature to test for potential tradeoffs in specificity and efficiency. The annealing temperature (and melting temperature) relies directly on the length and GC composition of the primer sets, where the melting temperature (Tm) of each primer set was found using the IDT OligoAnalyzer. If the annealing temperature is set too low, the primer sets are more likely to anneal to DNA sequences other than the intended target which then leads to non-specific PCR amplification. Also, if the annealing temperature is too high the efficiency may be reduced because of poor primer annealing to the template DNA. The optimal annealing temperature will give the best PCR product yield of the correct amplicon (IDT, n.d.). These values are found for each primer set in Table 11.

Table 11. Low and high annealing temperatures and melting temperatures for primer sets from literature.

Primer Set	Low Annealing Temperature, °C	Melt Temperature (T_m, 0.5 μM primer), °C	High Annealing Temperature, °C
MIB3324	57.1	59.1	61.1
Gaget	62.3	64.3	66.3
MIB-Rf/Rr	58.0	60.0	62.0
Str-Rf/Rr	56.1	58.1	60.1

For MIB3324, cycling parameters were 95° C for 5 min, 51 cycles of 95° C for 15 s, gradient 57-62° C for 30 s, and 72° C for 45 s. For Gaget, cycling parameters were 95° C for 5 min, 55 cycles of 95° C for 15 s, gradient 62-66.3° C for 30 s, and 72° C for 45 s. For MIB-Rf/Rr, cycling parameters were 95° C for 5 min, 45 cycles of 95° C for 15 s, gradient 56.1-62.9° C for 30 s, and 72° C for 45 s. For Str-Rf/Rr, cycling parameters were 95° C for 5 min, 45 cycles of 95° C for 15 s, gradient 56.1-62.9° C for 30 s, and 72° C for 45 s. Each primers set also ran at 72° C for 5 min before the machine shut down and the qPCR products were removed, and the data analyzed. Each primer was run at the low and high annealing temperature. Nuclease-free water was used as a negative control. The DNA standards required for qPCR were then made using known DNA concentrations to quantify MIB synthase gene abundance in each sample amplified in qPCR by using the constant reference point with the DNA standards. The efficiency of each primer set at its low and high annealing temperature were found to analyze which annealing temperature was the most efficient. This was found using the equation,
$E\ (\%) = 2^{-S} - 1$, where S is the slope of the standard curve. If during each cycle of qPCR, the targeted DNA template is doubled, the efficiency will be 100%. The qPCR product quality was then evaluated after gel electrophoresis on agarose gels.

2.5. Gel Electrophoresis

After qPCR was completed, gel electrophoresis was run to check the size of the qPCR products. It is a standard lab procedure for separating DNA by size (length of base pairs) for visualization and purification. Electrophoresis uses an electrical field to move the negatively charged DNA through an agarose gel matrix toward a positive electrode. Shorter DNA fragments migrate through the gel more quickly than longer ones. Consequently, determined the approximate length of a DNA fragment by running it on an agarose gel alongside a DNA ladder, which is a collection of DNA fragments of known lengths. The procedure has been adapted from Addgene (www.addgene.org/protocols/gel-electrophoresis/#faq) to produce a 25x TAE buffer solution, using 0.5 M EDTA combined with Tris solution and dH_2O. One gel requires 500 ml of the 1x TAE to prepare the gel (50 ml) and the running buffer solution (400-450 ml), so the 25x TAE was diluted to create 1x TAE buffer solution. A 1% agarose gel was then prepared in a refrigerated gel tray with well combs in place in the first and middle slots to allow for double gels.

To run the gel, the DNA samples were first prepared on parafilm with a dilution mixture of 10 ul of product, 5 ul of dH2O, and 5 ul of 4x loading buffer. The ladder was diluted with 15 ul of ladder, and 5 ul of dH2O since the ladder already contains the loading buffer. All 20 ul of each sample was loaded into their respective wells in the gel. The 100 base-pair DNA ladder was added in the fifth well on both rows for comparisons of lengths. The gel was run at 75 volts and then a 0.5 µg/ml EtBr solution was prepared. When the gel was done running, it was soaked in this diluted EtBr solution for 20-30 minutes on a shaker table at 60 rpm and 20°C. The gel was then transferred to dH2O in a separate container to de-stain for 5 minutes on the shaker. It was then viewed under UV light to visualize the bands.

2.6. Sequencing

Sequencing for these samples was completed through Mr. DNA Lab Molecular Research, LP. The MIB synthase gene sequences were amplified using the four provided primer sets for 30-35 cycles using the HotStarTaq Plus Master Mix Kit (Qiagen, USA) under the following conditions: 95°C for 5 minutes, followed by 30-35 cycles of 95°C for 30 seconds, the higher annealing temperatures shown in Table 11 for 40 seconds, and 72°C for 1 minute, after which a final elongation step at 72°C for 10 minutes was performed. After amplification, PCR products were checked in 2% agarose gel to determine the success of amplification and the relative intensity of bands. A bioanalyzer was also used to confirm product sizes. Only samples that amplified a product close to the target size (Table 10) underwent sequencing. Samples are multiplexed using unique dual indices and are pooled together (e.g., 100 samples) in equal proportions based on their molecular weight and DNA concentrations. Pooled samples were purified using calibrated Ampure XP beads. Then the pooled and purified PCR product was used to prepare an Illumina DNA library. Sequencing was performed at MR DNA (www.mrdnalab.com, Shallowater, TX, USA) on a MiSeq following the manufacturer's guidelines. Sequence data were processed using MR DNA analysis pipeline (MR DNA, Shallowater, TX, USA). In summary, sequences were joined, sequences <100bp removed, and sequences with ambiguous base calls removed. Sequences were quality filtered using a maximum expected error threshold of 1.0 and dereplicated. The dereplicated or unique sequences were denoised; unique sequences identified with sequencing and/or PCR point errors were removed, chimera were removed, thereby providing a denoised sequence or zOTU. Final zOTUs were taxonomically classified using BLASTn against a curated database derived from NCBI for the MIB synthase gene (www.ncbi.nlm.nih.gov).

3. Results and Discussion

3.1. Water Quality Variable Detection and Pearson's Correlations

3.1.1. MIB level detection

The overall detection for MIB levels across all three reservoirs was low for the 2020 sampling season. The average for all three reservoirs together was only 2.25 ng/L.

3.1.1.1. Auburn

The City of Auburn Water Resources Department saw MIB spikes from mid-May to early-August, though the average for the reservoir was a mere 3.23 ng/L with a range of 0-7.9 ng/L (Fig. 31). The detected MIB levels never reached above the 10 ng/L taste threshold.

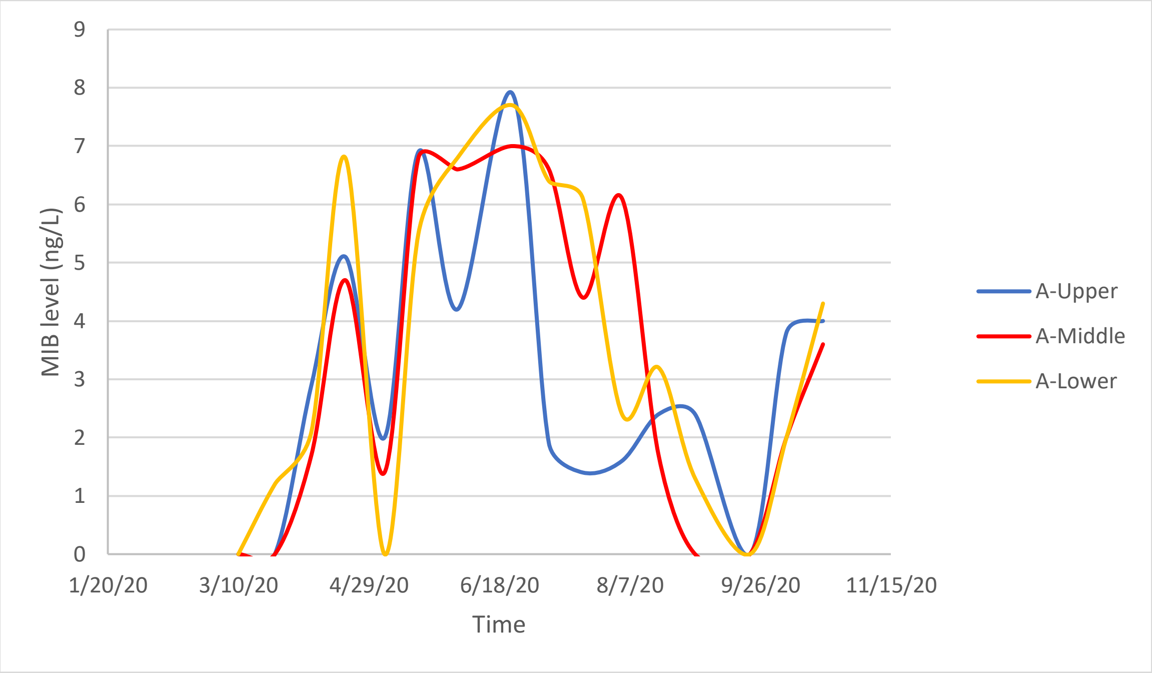


Figure 31. MIB over time: City of Auburn Water Resources Department.

3.1.1.2. Opelika Utilities

Small spikes in MIB occurred for Opelika Utilities first on March 26th, then again on June 3rd and July 28th. The highest MIB level detected was only 5.43 ng/L (Fig. 32), lower than the human detection level and therefore not a customer issue for the 2020 year.

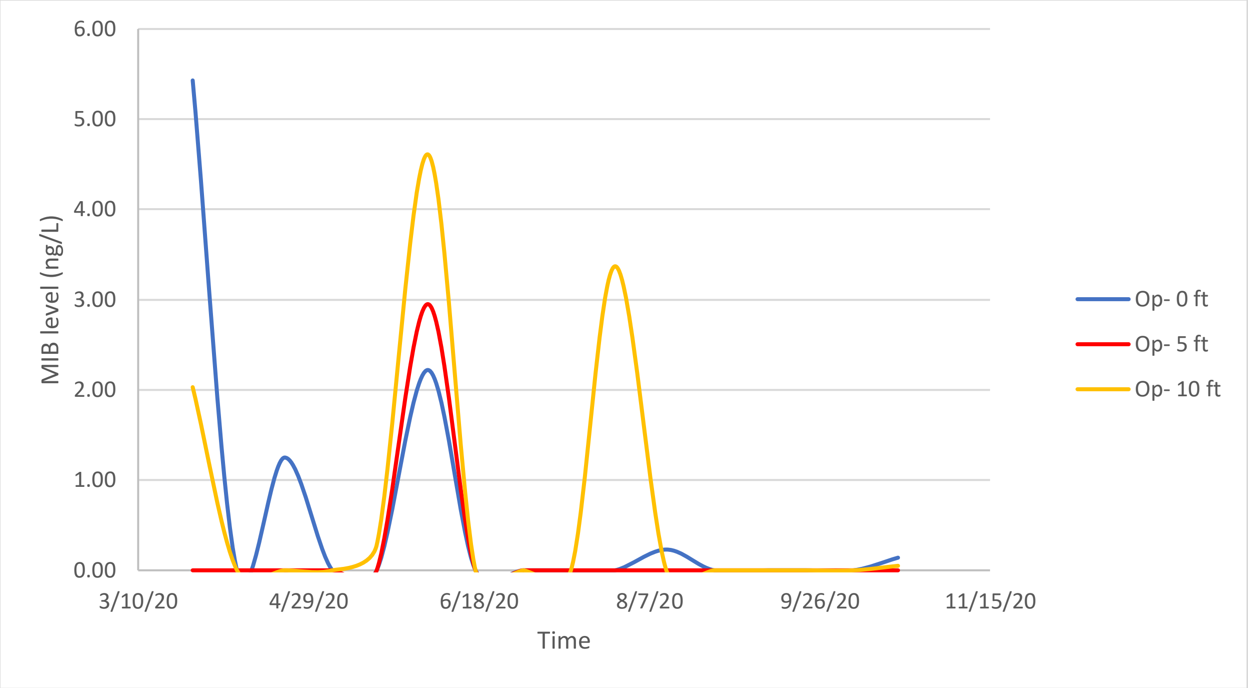


Figure 32. MIB over time: Opelika Utilities.

3.1.1.3. Columbus Water Works

The largest MIB spike for CWW occurred on September 17th for 3 locations in the reservoir: Roaring Boy, Standing Boy, and Heiferhorn, with the largest being in Heiferhorn at 18.5 ng/L. Other increases above the taste threshold ranged from 11.7-15.4 ng/L in mid-March, late April, and mid-June. There was a cluster of increases in MIB level detection from early August until the end of September, though none of these levels reached above 10 ng/L, aside from the 9/17 spike.

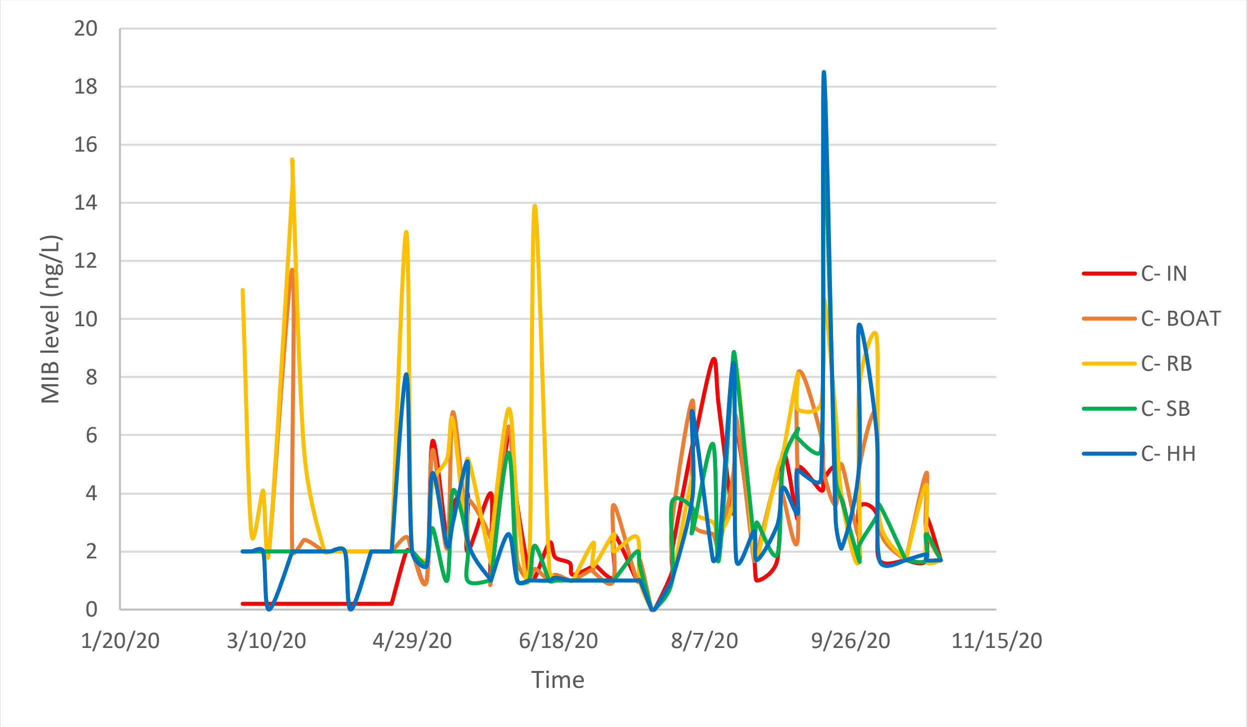


Figure 33. MIB over time: CWW.

The key take-aways from these figures (Figures 31-33) are that the MIB levels detected never reached above 10 ng/L for the City of Auburn Water Resources Department or Opelika Utilities, and only reached above that threshold 2.6% of the time for Columbus, which is not helpful for modeling, similar to what was done for geosmin. More samples with elevated MIB levels are necessary to develop useful predictive models.

3.2. Pearson's Correlations

A variety of common water quality metrics were measured and recorded by each utility (Table 12). Correlation analysis of MIB concentration (ng/L) versus each water quality variable was performed. Dissolved oxygen and nitrate showed an r value greater than ±0.5, which could be further investigated for importance in possible future multiple linear regression or CART modeling. Modeling with this 2020 dataset was limited by the fact that so little MIB was detected in the 2020 sampling season.

Table 12. Correlation (r) between MIB and each water quality parameter.

	Auburn	Opelika	Columbus
TP	-0.1772	0.0235	0.1455
TKN	0.0645	0.0000	0.0748
Nitrite + Nitrate	0.1034	-0.0720	0.0372
OP	0.0716	-	-
Water temp	0.0420	0.0132	0.0716
pH	-0.1504	-0.1155	-0.0124
DO	-0.5386	-0.0924	-0.0682
Specific Conductance	0.0555	-0.2295	0.0680
Turbidity	0.1639	-0.0658	0.2520
Secchi Depth	0.1742	-0.2520	-0.0471
Chlor-a	0.0432	-0.1682	0.2536
Phycocyanin	-0.1956	-0.1497	0.3589
Air temp	0.0447	0.1700	0.0269
Rainfall	0.2582	-0.1447	0.0660
Wind Speed	-0.2662	-0.0952	0.0166
BG Count	-	-	
Nitrate	-0.5641	-0.1664	-0.0420
Phosphate	0.0257	-0.1149	-0.0707

r	
± 0-0.05	🟥
± 0.05-0.2	🟨
± 0.2-0.4	⬜
> ± 0.4	🟩

3.3. Primer Set Efficiencies

For each primer set targeting the MIB synthase gene, the efficiencies were found for the low and high annealing temperatures used during qPCR. The low annealing temperatures show very similar efficiencies to the higher annealing temperatures, with the low annealing temperature having a slightly higher efficiency for 3 out of the 4 primer sets (Table 13). However, in all cases, efficiency fell between 90% and 105%, a range that is generally considered acceptable for qPCR. This suggests that there is only a small efficiency tradeoff at the higher annealing temperature.

Table 13. Primer set efficiencies found from efficiency tests.

Primer Set	Low Annealing Temperature Efficiency (%)	High Annealing Temperature Efficiency (%)
MIB3324	99	96
Gaget	97	92
MIB Rf/Rr	99	103
Str Rf/Rr	99	98

3.4. qPCR

Figures 34 and 35 below express the abundance of MIB synthase gene copies found per milliliter of water sample after qPCR using each of the four primer sets above at the higher and lower annealing temperatures. The 8 chosen samples have a range of MIB levels of 0.5-166.62 ng/L and MIB synthase levels of 1,431-346,962 copies/ml across the 4 primer sets at the low annealing temperatures and 8-2,250,331 copies/ml for the high annealing temperatures, spanning from sampling times of May 6th to September 22nd. The log of the MIB synthase gene abundance using the Gaget primer set has the highest R^2 value for both the low and high annealing temperature when correlated with the log of the MIB level, with an R^2 of 0.6739 and 0.5523 respectively. This suggests that gene abundance alone may have high predictive power for moderate to high MIB concentration across water samples from multiple sources.

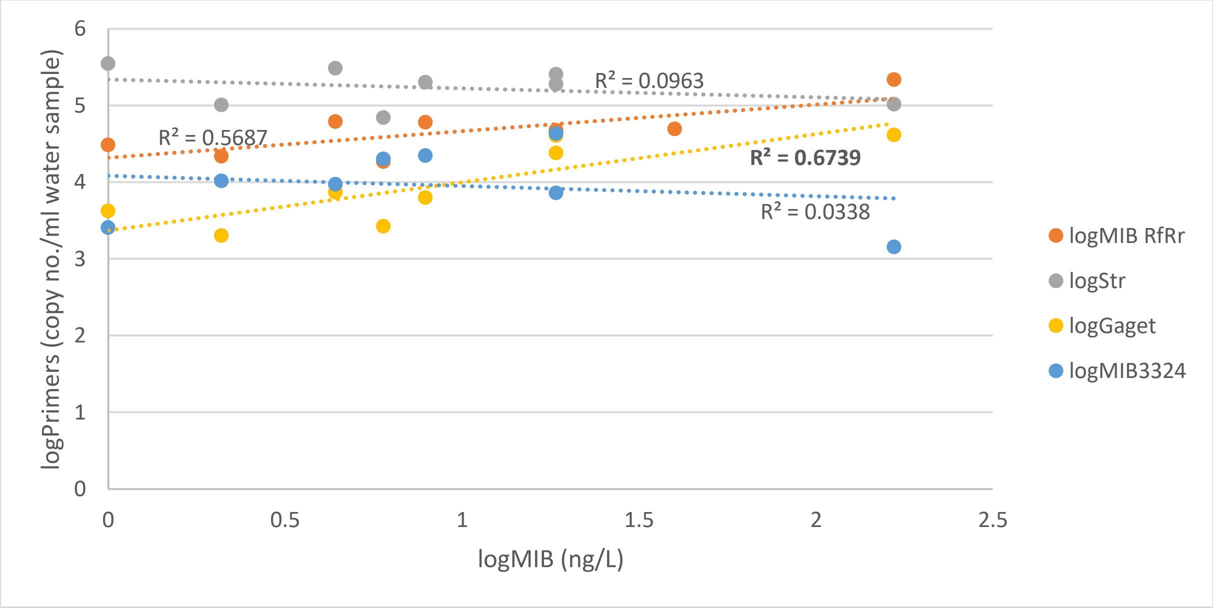


Figure 34. Linear regression of the log of detected MIB levels versus the log of the abundance of copy no. of the targeted gene sequence with each primer set, all run at the low annealing temperature during qPCR.

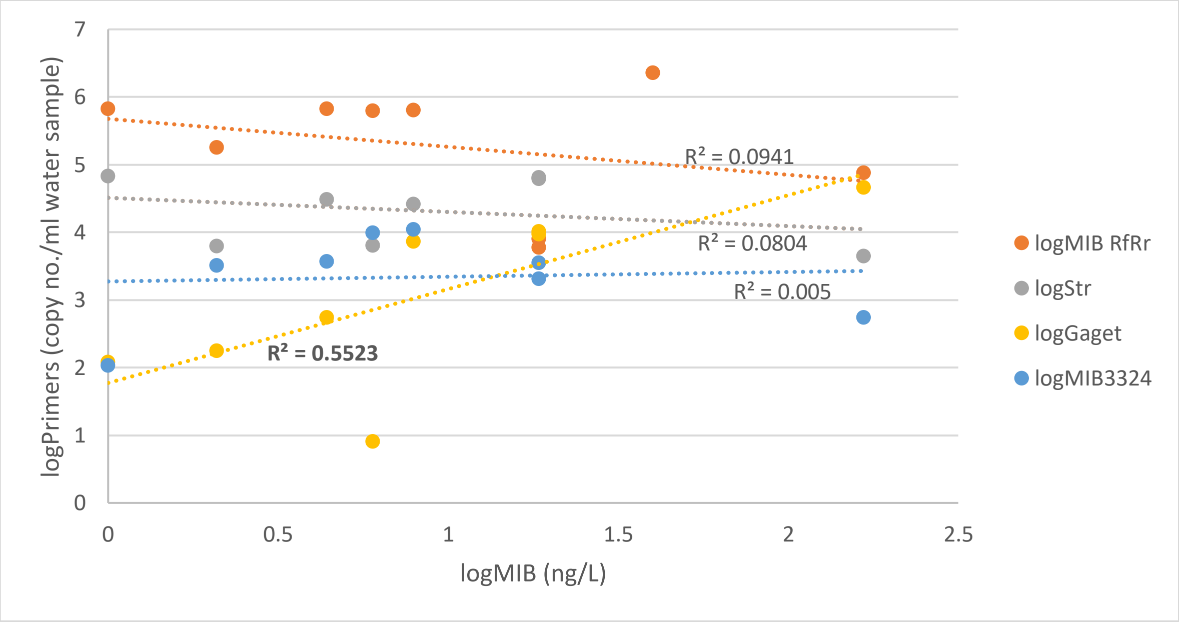


Figure 35. Linear regression of the log of detected MIB levels versus the log of the abundance of copy no. of the targeted gene sequence with each primer set, all run at the high annealing temperature during qPCR.

These R^2 values show us that the Gaget primer set was able to explain around 67% and 55% of the variation in MIB levels detected using its low and high annealing temperatures during qPCR and be useful in future predictive modeling using multiple linear regression. At the low annealing temperature, MIB synthase gene copy as measured using the MIB-Rf/Rr primers could explain 56% of the variation in MIB level in these samples. Because both of these primer sets target cyanobacteria, it is likely that MIB in these water bodies are largely controlled by cyanobacteria MIB producers. To better investigate the usefulness of each of these primer sets, I checked the amplified product size using gel electrophoresis, and then further by sequencing the qPCR products.

3.5. Gel Electrophoresis and Specificity

By using gel electrophoresis to check for DNA fragment lengths that have been amplified during qPCR, we can verify the primer specificity toward the single intended target. The specificity of primer is controlled by the length of the primer, the annealing temperature used during qPCR, and the frequency with which mis-priming occurs during PCR (Dieffenbach et al., 1993). Poor specificity of a primer set is clear when extra unrelated amplicons are present in the gel image (multiple bands, incorrect band location). In Figure 36, the higher annealing temperature looks to be more specific for Gaget, MIB-Rf/Rr, and Str-Rf/Rr. The specific samples that were of greatest interest are in the last three columns on the right-hand side of the gel: A-Upper 6/23, USDA 8 8/20, and C-HH 9/17 since they had higher detected MIB levels of 7.9, 166.6, and 18.5 ng/L, respectively, compared to the other samples. The four samples on the left have lower (but still detectable) MIB levels of < 6 ng/L and therefore would not be expected to have high MIB synthase gene abundance. For the three noteworthy samples, Gaget shows the highest specificity for the ~179 bp (appears as 200 bp on gel) amplicon length for the DNA sequence. Some variation in amplicon length is expected in environmental samples since not all taxa have the same gene sequence length. Hence, some bands may appear hazy. For the other three primers, there is evidence of poor specificity as seen through multiple bands on the gels or streaking.

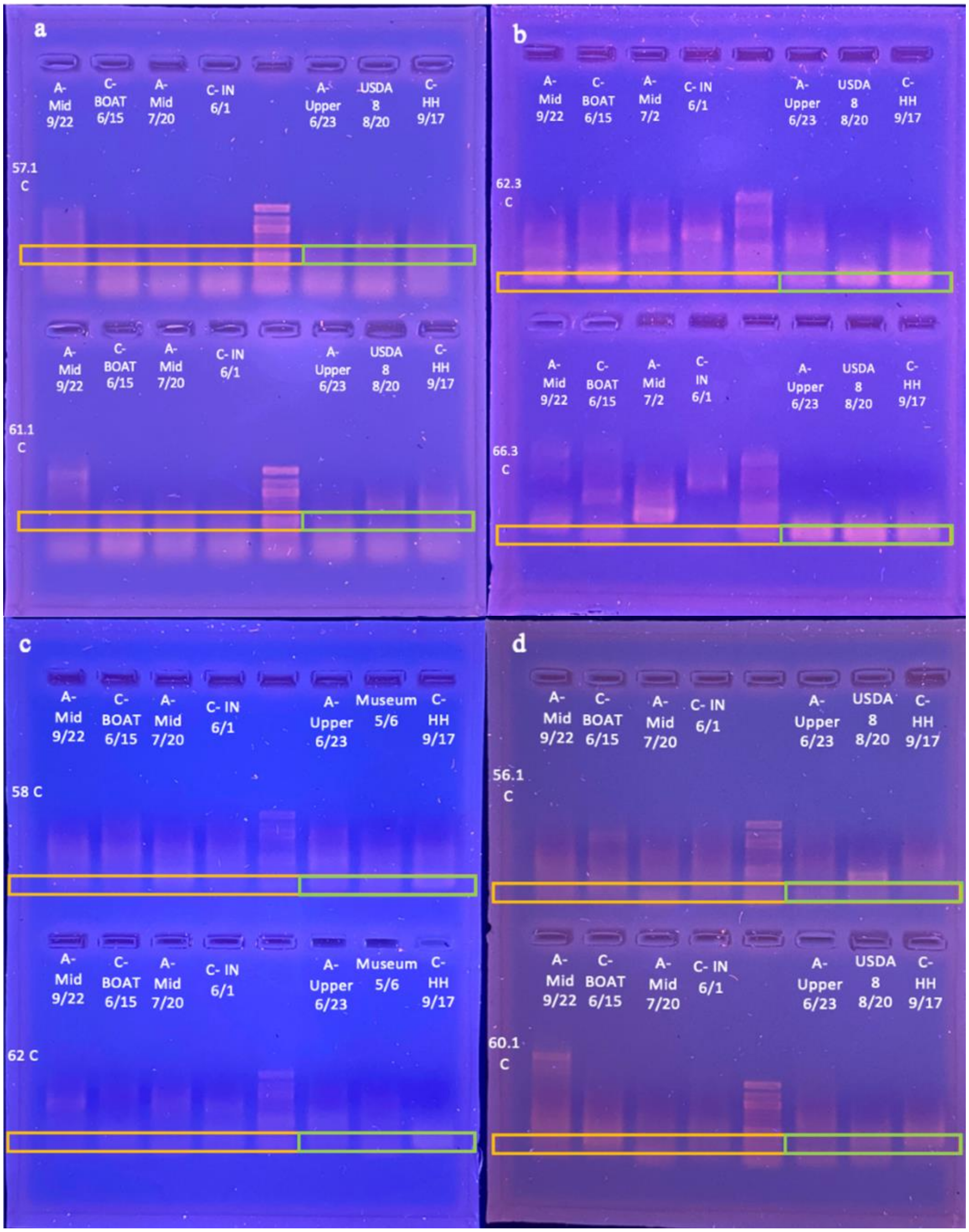


Figure 36: (a) MIB3324 (726 bp) gel electrophoresis results, (b) Gaget F/R (179 bp) gel electrophoresis results, (c) MIB Rf/Rr (202 bp) gel electrophoresis results, (d) Str Rf/Rr (339) gel electrophoresis results.

3.6. Sequencing

3.6.1. Sequencing Results

If MIB is present, there must be elevated levels of MIB synthase present in the DNA, and as mentioned above the samples with higher levels of MIB are easier to amplify and subsequently run sequencing on. Sometimes, the samples did not have a strong enough PCR product to be able to sequence. Hence, the following sequencing results may only be shown for a few samples that had sufficient product for sequencing. Again, this was not surprising given the overall low levels of MIB present in most samples. The samples with the highest MIB levels (Table 14) generally had enough amplified product for sequencing and were the focus of this analysis.

Table 14. MIB (ng/L) concentrations detected in chosen samples.

	MIB (ng/L)
A-Upper 6/23	7.9
USDA 8 8/20	166.6
C-HH 9/17	18.5

The Gaget primer set mapped 100% to its intended Cyanobacterial phylum (Fig. 37), with *Planktothricoides* and *Pseudanabaena* as the major genera identified, and low abundance of *Oscillatoria* and *Leptolyngbya* identified as well (Fig. 38). All of these genera have previously been identified as MIB producers (Devi et al., 2021). These results mean that the primer set was able to amplify the correct target DNA sequence specifically within its intended phylum, while also being broad enough to include four important genera. The A-Upper sample from 6/23 did not generate sufficient product with this primer set for sequencing and is therefore not shown in Figures 37 and 38.

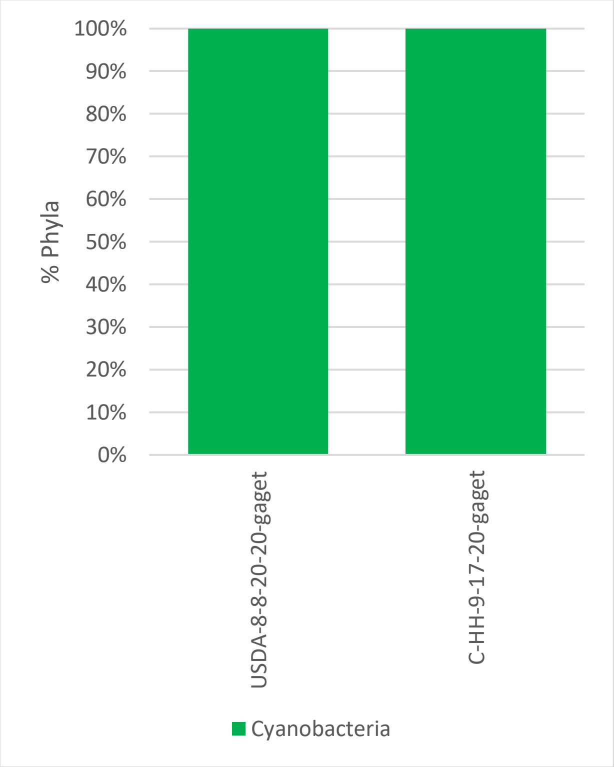


Figure 37. Phyla identified through sequencing of PCR products using the Gaget primer set.

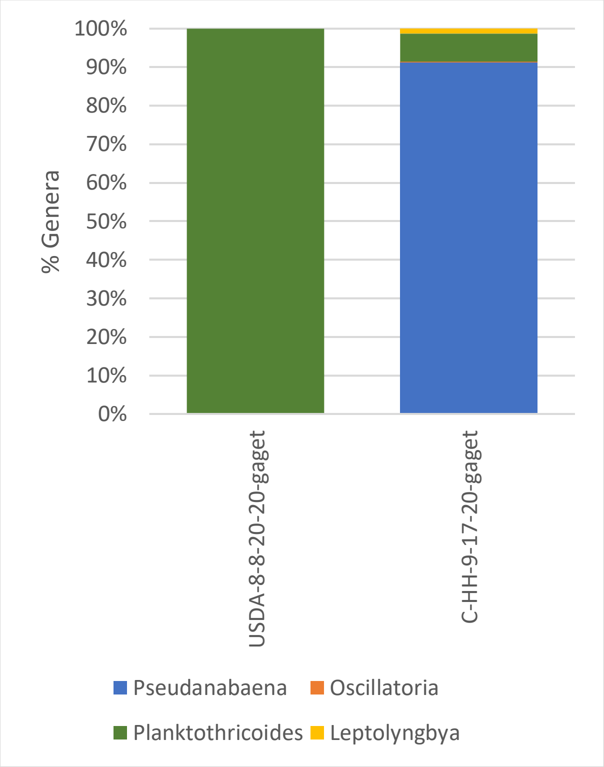


Figure 38. Genera identified through sequencing of PCR products using the Gaget primer set.

The MIB3324 primer set mapped almost 100% to actinobacterial phylum rather than its intended target of cyanobacteria (Fig. 39), with *Frankia* and *Streptomyces* as the major genera identified, with very low abundance of *Nostoc* (cyanobacteria) identified as well (Fig. 40). While *Streptomyces* is a well-known MIB producer, *Frankia* and *Nostoc* are not known for their MIB production, yet *Nostoc* can produce geosmin (Devi et al., 2021) and *Frankia* is presumed to have geosmin synthase (Giglio et al., 2008). It is telling that the two samples with the highest MIB abundance (and likely produced by cyanobacteria based on two of the other primer sets) did not even amplify with the MIB3324 set.

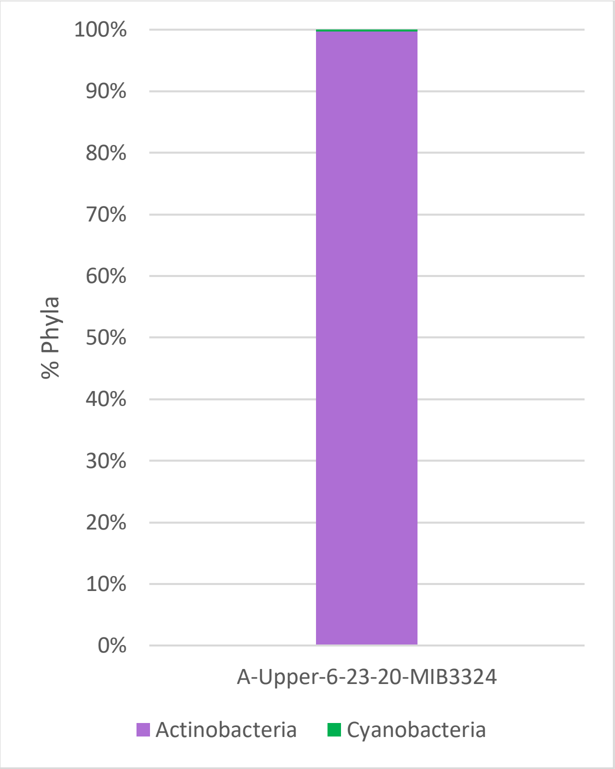


Figure 39. Phyla identified through sequencing of PCR products using the MIB3324 primer set.

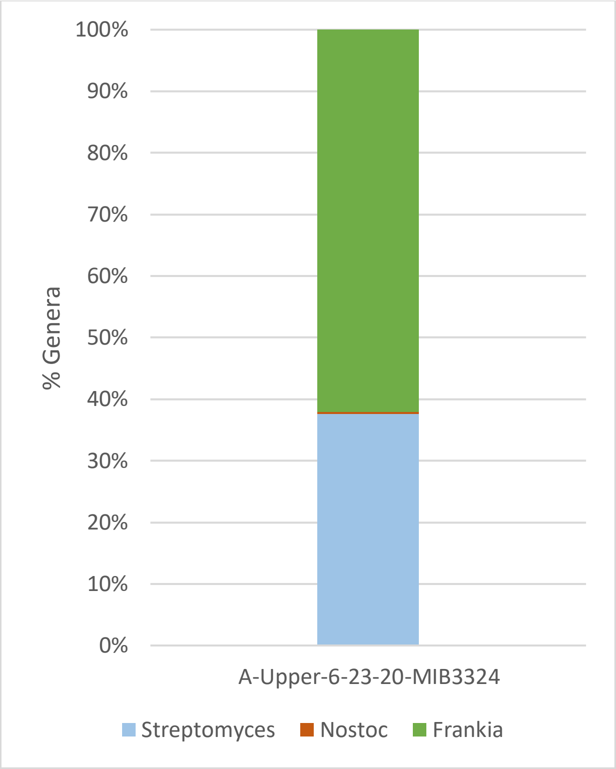


Figure 40. Genera identified through sequencing of PCR products using the MIB3324 primer set.

The MIB-Rf/Rr primer set also targets cyanobacteria, but sequencing of the PCR products reveals that it also amplified actinobacteria and proteobacteria in one of the samples (Fig. 41). Of the cyanobacterial phylum, *Planktothricoides*, *Pseudanabaena*, *Kamptonema*, *Novosphingobium*, and *Leptolyngbya* were identified (Fig. 42), though not all are known to produce T&O compounds. This primer set, despite its tendency toward non-specific amplification based on the gel results, was able to amplify known MIB producers in all three water samples. Moreover, there was general agreement with the Gaget primer set: the USDA aquaculture pond was dominated by *Planktothrix* and the Heiferhorn location in Lake Oliver was dominated by *Pseudanabaena*. These were very likely the major source organisms for these elevated MIB levels, particularly given the strong correlation between gene abundance and MIB for the Rf/Rr and Gaget primer sets.

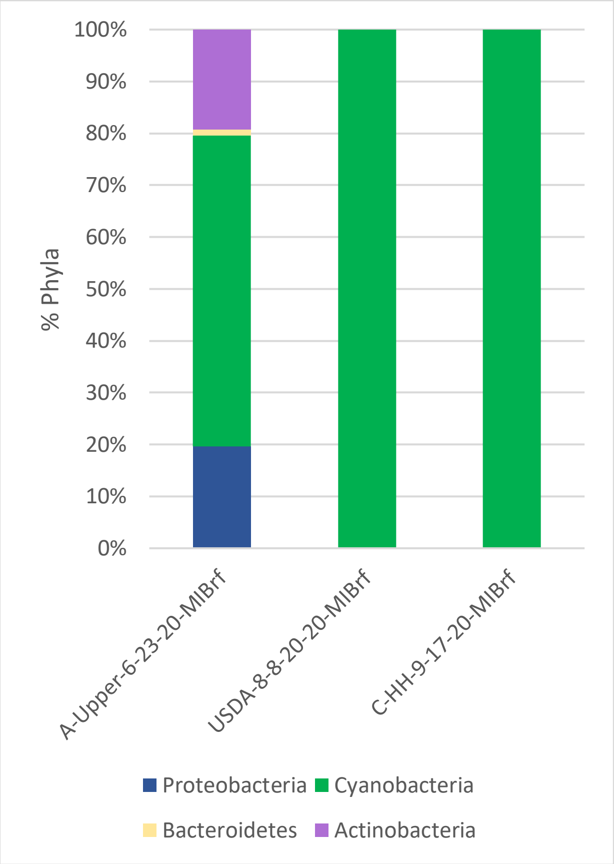


Figure 41. Phyla identified through sequencing of PCR products using the MIB-Rf/Rr primer set.

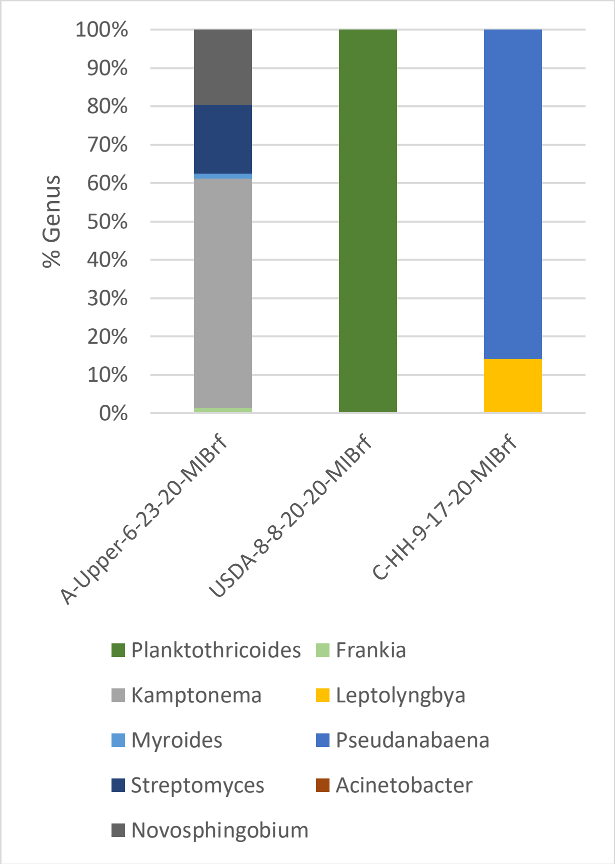


Figure 42. Genera identified through sequencing of PCR products using the MIB-Rf/Rr primer set.

The Str-Rf/Rr primer set was intended to amplify only actinobacteria, specifically within the *Streptomyces* genus. It amplified mostly *Streptomyces*, though it clearly amplified many other genera within the actinobacterial, cyanobacterial, and proteobacterial phyla (Figures 43 and 44). In two of the three samples, this set effectively amplified Streptomyces, a known MIB producer. In the third sample, it primarily amplified a gene mapping to *Sorangium*. *Sorangium* is likely a geosmin producer (Lukassen et al., 2019) but there are no known reports of it producing MIB.

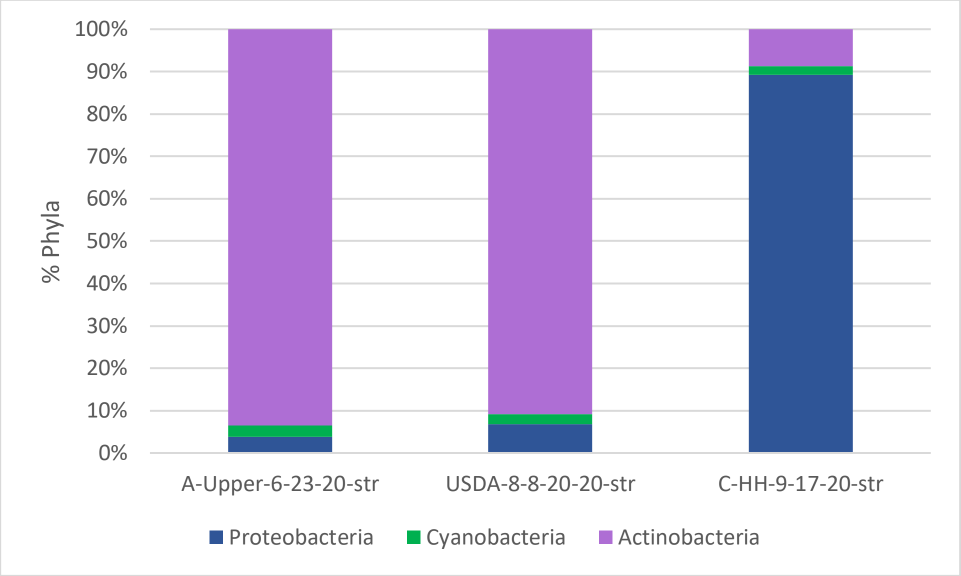


Figure 43. Phyla identified through sequencing of PCR products using the Str-Rf/Rr primer set.

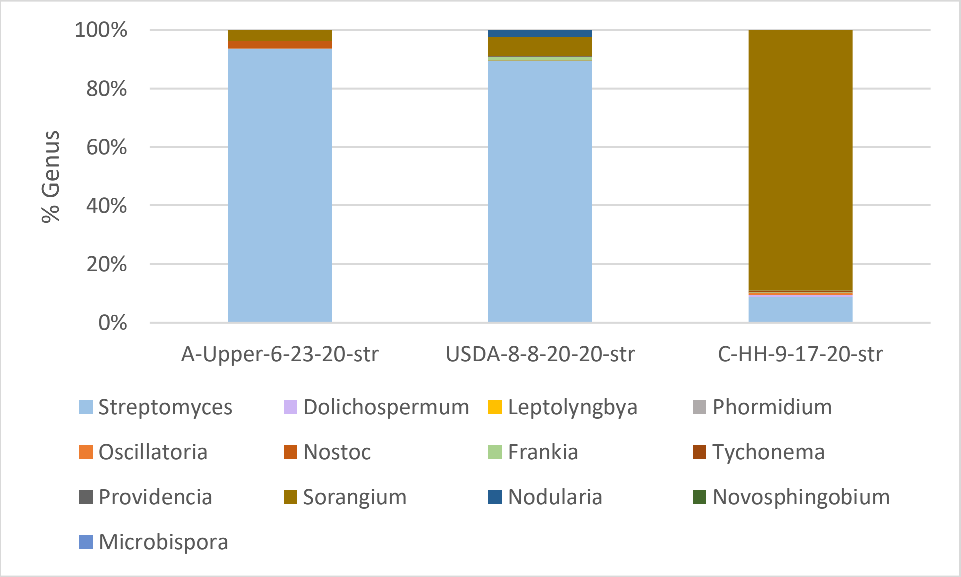


Figure 44. Phyla identified through sequencing of PCR products using the Str-Rf/Rr primer set.

4. MIB Conclusions

All primer sets displayed high efficiency (within the acceptable range of 90-105%) at their low and high annealing temperatures. We were able to amplify 8 water samples with varied MIB levels using each of the four primer sets during qPCR. After this, we were able to run the correlations of the calculated MIB synthase gene abundance with the detected MIB levels and found that the Gaget primer set had the highest correlation at its low and high annealing temperatures (R^2 = 0.5523 and 0.6739, respectively). The Gaget primer set also displayed the best specificity on the gel electrophoreses, especially at its higher annealing temperature of 66.3° C. At this annealing temperature, the efficiency was found to be adequate, at 92%. After qPCR, the sequencing revealed that the Gaget primer set was the only one to amplify 100% of its intended target, cyanobacteria. It also displayed the ability to include several cyanobacterial genera, exhibiting the right amount of specificity that is essential for further modeling processes. With this, Gaget proves to be the best choice of primer for qPCR in the reservoirs sampled in this study and could be used in the future for predictive modeling, just as has been done for geosmin. Limitations of this study include the extremely low levels of MIB (average of 2.25 ng/L) in the Auburn, Opelika, and Columbus reservoirs throughout 2020.

5. Overall Conclusions

- A series of predictive models were developed relating water quality variables combined with genetic data (synthase gene abundance) to concentrations of dissolved geosmin in three Southeastern regional drinking water reservoirs. All models have 40-60% predictive capabilities for geosmin levels, with Cgeo1 gene abundance being significant in only one reservoir. Models were limited by the low-moderate geosmin levels detected throughout the sampling season. We anticipate that with greater geosmin levels captured in sampling the models would be improved.
- Multiple regressions were also completed using the water quality and genetic data, with the best fit for the City of Auburn Water Resources Department, with an adjusted R^2 of 0.5405. Auburn had the highest geosmin peaks (>30 ng/L), whereas Opelika Utilities and Columbus Water Works had consistently low geosmin levels and lower predictive power in their subsequent multiple regressions.

  - Cgeo1 gene abundance improved the Auburn multiple regression ($P<0.01$). With higher geosmin levels and significance shown with the Auburn data, it can be concluded that the inclusion of the qPCR data was most effective at predicting higher geosmin levels.
- Sequencing of products using Cgeo1 primer set found *Anabaena* and *Planktothrix* as the key geosmin producers in the lakes sampled.
- MIB primer set evaluations found that the Gaget primer (Gaget et al., 2020) has the best specificity and adequate efficiency when set at an annealing temperature of 66.3° C. MIB gene synthase abundance found using the Gaget primer set during qPCR has the best correlation to the MIB levels detected ($R^2 = 0.5523$).
- Sequencing of MIB qPCR products using the Gaget primer set found *Planktothricoides* and *Pseudanabaena* as the major genera identified. This knowledge aids in the possibilities for the utilities to adapt reservoir management practices, whether it is chemical/physical, biological, or mechanical.

5.1. Future Work

- For the best use of the CART modeling framework, higher geosmin and MIB levels should be captured. Though the models produced have reasonably useful predictive power, it is senseless to be able to predict such low levels of T&O compounds. The future ability to incorporate T&O outbreak data will likely improve prediction power.
- After evaluation of the four primer sets using the same water sample data, the Gaget primer set could be used in future modeling efforts using CART modeling in R to predict MIB levels as well. Just as geosmin had low concentrations, MIB had even lower levels during this study. To have significant predictions made with this modeling, it is necessary to capture higher MIB concentrations.
- With more time, I would have liked to analyze the confidence with which the organisms were matched from the sequencing results for both geosmin and MIB.
- It would also be of interest to do analysis using both primer sets using sediment samples to better include the actinobacterial role for geosmin and MIB.
- The ability to use mRNA rather than DNA extraction and analysis could be beneficial in the future to be able to capture the genetic transcription and therefore the actual gene

expression in these samples. The drawback of the use of mRNA is that it is much more expensive and difficult to perform, so with the hopes of being able to easily implement this at the water utility department this might not be as applicable.

- With multiple-year datasets incorporating more of this information, drinking water utilities could incorporate these models into their routines for better water quality management from taste and odor outbreaks.

REFERENCES

Asquith, E., Evans, C., Dunstan, R. H., Geary, P., & Cole, B. (2018). Distribution, abundance and activity of geosmin and 2-methylisoborneol-producing Streptomyces in drinking water reservoirs. Water Research, 145, 30–38. https://doi.org/10.1016/j.watres.2018.08.014

Asquith, E. A., Evans, C. A., Geary, P. M., Dunstan, R. H., & Cole, B. (2013). The role of Actinobacteria in taste and odour episodes involving geosmin and 2-methylisoborneol in aquatic environments. Journal of Water Supply: Research and Technology-Aqua, 62(7), 452–467. https://doi.org/10.2166/aqua.2013.055

Bai, X., Zhang, T., Wang, C., Zong, D., Li, H., & Yang, Z. (2016). Occurrence and distribution of taste and odor compounds in subtropical water supply reservoirs and their fates in water treatment plants. Environmental Science and Pollution Research, 24(3), 2904–2913. https://doi.org/10.1007/s11356-016-7966-5

Cai, F., Yu, G., Zhang, K., Chen, Y., Li, Q., Yang, Y., Xie, J., Wang, Y., & Li, R. (2017). Geosmin production and polyphasic characterization of Oscillatoria limosa Agardh ex Gomont isolated from the open canal of a large drinking water system in Tianjin City, China. Harmful Algae, 69, 28–37. https://doi.org/10.1016/j.hal.2017.09.006

Chaump, K. et al. (2018). Leaching and anaerobic digestion of poultry litter for biogas production and nutrient transformation. Waste Management, doi:10.1016/j.wasman.2018.11.024.

Chen, Y., & Zhu, J. (2018). Observation and simulation of 2-methylisoborneol in the Qingcaosha Reservoir, Changjiang estuary. Journal of Oceanology and Limnology, 36(5), 1586–1596. https://doi.org/10.1007/s00343-018-7124-7

Chou, J.-S., Ho, C.-C., & Hoang, H.-S. (2018). Determining quality of water in reservoir using machine learning. Ecological Informatics, 44, 57–75. https://doi.org/10.1016/j.ecoinf.2018.01.005

Choubin, B., Darabi, H., Rahmati, O., Sajedi-Hosseini, F., & Kløve, B. (2018). River suspended sediment modelling using the CART model: A comparative study of machine learning techniques. Science of The Total Environment, 615, 272–281. https://doi.org/10.1016/j.scitotenv.2017.09.293

Christensen, V.G., Graham, J.L., Milligan, C.R., Pope, L.M., Zeigler, A.C., 2006. Water Quality and Relation to Taste-and-odor Compounds in the North Fork Ninnescah River and Cheney Reservoir, South- central Kansas, 1997–2003. United States Geological Survey Scientific Investigations Report 2006–5095, 43 pp.

Chung, S.-W., Chong, S.-A., & Park, H.-S. (2016). Development and Applications of a Predictive Model for Geosmin in North Han River, Korea. Procedia Engineering, 154, 521–528. https://doi.org/10.1016/j.proeng.2016.07.547

Devi, A., Chiu, Y.-T., Hsueh, H.-T., & Lin, T.-F. (2021). Quantitative PCR based detection system for cyanobacterial geosmin/2-methylisoborneol (2-MIB) events in drinking water sources: Current status and challenges. Water Research, 188, 116478. https://doi.org/10.1016/j.watres.2020.116478

Downing, J. A., Watson, S. B. & McCauley, E. (2001). Predicting Cyanobacteria dominance in lakes. Canadian Journal of Fisheries and Aquatic Sciences 58, 1905-1908, doi:10.1139/f01-143.

Dzialowski, A. R. et al. (2009). Development of predictive models for geosmin-related taste and odor in Kansas, USA, drinking water reservoirs. Water Research 43, 2829-2840, doi:https://doi.org/10.1016/j.watres.2009.04.001.

Gaget, V., Hobson, P., Keulen, A., Newton, K., Monis, P., Humpage, A. R., Weyrich, L. S., & Brookes, J. D. (2020). Toolbox for the sampling and monitoring of benthic cyanobacteria. Water Research, 169, 115222. https://doi.org/10.1016/j.watres.2019.115222

Gardener, M. (2012). Beginning R: the statistical programming language. John Wiley & Sons.

Gerber, N. N., & Lechevalier, H. A. (1965). Geosmin, an earthy-smelling substance isolated from actinomycetes. Appl. Environ. Microbiol., 13(6), 935-938.

Giglio, S., Chou, W. K. W., Ikeda, H., Cane, D. E., & Monis, P. T. (2010). Biosynthesis of 2-methylisoborneol in cyanobacteria. Environmental science & technology, 45(3), 992-998.

Giglio, S., Saint, C. P., & Monis, P. T. (2011). EXPRESSION OF THE GEOSMIN SYNTHASE GENE IN THE CYANOBACTERIUM ANABAENA CIRCINALIS AWQC3181. Journal of Phycology, 47(6), 1338–1343. https://doi.org/10.1111/j.1529-8817.2011.01061.x

Guttman, L., & van Rijn, J. (2008). Identification of conditions underlying production of geosmin and 2-methylisoborneol in a recirculating system. Aquaculture, 279(1-4), 85–91. https://doi.org/10.1016/j.aquaculture.2008.03.047

Harris, T. D. & Graham, J. L. (2017). Predicting cyanobacterial abundance, microcystin, and geosmin in a eutrophic drinking-water reservoir using a 14-year dataset. Lake and Reservoir Management, doi:10.1080/10402381.2016.1263694.

IDT: OligoAnalyzer Tool - primer analysis. Integrated DNA Technologies. (n.d.). https://www.idtdna.com/calc/analyzer.

Juttner, F., 1984. Dynamics of the volatile organic-substances associated with cya- nobacteria and algae in a eutrophic shallow lake. Appl. Environ. Microbiol. 47, 814e820.

Jüttner, F. & Watson, S. B. (2007). Biochemical and Ecological Control of Geosmin and 2-Methylisoborneol in Source Waters. Applied and Environmental Microbiology 73, 4395, doi:10.1128/AEM.02250-06.

Kaevska, M. & Slana, I. (2015). Comparison of filtering methods, filter processing and DNA extraction kits for detection of mycobacteria in water. Annals of Agricultural and Environmental Medicine 22, 429-432, doi:10.5604/12321966.1167707.

Kehoe, M. J., Chun, K. P. & Baulch, H. M. (2015). Who Smells? Forecasting Taste and Odor in a Drinking Water Reservoir. Environ Sci Technol 49, 10984-10992, doi:10.1021/acs.est.5b00979.

Kim, C., Lee, S. I., Hwang, S., Cho, M., Kim, H., & Noh, S. H. (2014). Removal of geosmin AND 2-methylisoboneol (2-MIB) by membrane system combined with powdered activated CARBON (PAC) for drinking water treatment. Journal of Water Process Engineering, 4, 91-98. doi:10.1016/j.jwpe.2014.09.006

Kim, S., Kim, S., Mehrotra, R., & Sharma, A. (2020). Predicting cyanobacteria occurrence using climatological and environmental controls. Water Research, 175, 115639. https://doi.org/10.1016/j.watres.2020.115639

Komatsu, M., Tsuda, M., Omura, S., Oikawa, H., & Ikeda, H. (2008). Identification and functional analysis of genes controlling biosynthesis of 2-methylisoborneol. Proceedings of the National Academy of Sciences, 105(21), 7422–7427. https://doi.org/10.1073/pnas.0802312105

Liato, V., & Aïder, M. (2017). Geosmin as a source of the earthy-musty smell in fruits, vegetables and water: Origins, impact on foods and water, and review of the removing techniques. Chemosphere, 181, 9–18. https://doi.org/10.1016/j.chemosphere.2017.04.039

Lindholm-Lehto, P. C., & Vielma, J. (2018). Controlling of geosmin and 2-methylisoborneol induced off-flavours in recirculating aquaculture system farmed fish—A review. Aquaculture Research, 50(1), 9–28. https://doi.org/10.1111/are.13881

Lu, K.Y., Chiu, Y.T., Burch, M., Senoro, D., Lin, T.F. (2019). A molecular-based method to estimate the risk associated with cyanotoxins and odor compounds in drinking water sources. Water Res. 164, 114938. doi:10.1016/j.watres.2019.114938.

Lukassen, M. B., Podduturi, R., Rohaan, B., Jørgensen, N. O., & Nielsen, J. L. (2019). Dynamics of geosmin-producing bacteria in a full-scale saltwater recirculated aquaculture system. Aquaculture, 500, 170-177.

Mau, D.P., Ziegler, A.C., Porter, S.D., Pope, L.M., 2004. Surface- water-quality Conditions and Relation to Taste-and-odor Occurrences in the Lake Olathe Watershed, Northeast Kansas, 2000–02. United States Geological Survey Scientific Investigations Report 2004–5047, 95 pp.

Medsker, L. L., Jenkins, D., Thomas, J. F., & Koch, C. (1969). Odorous compounds in natural waters. 2-Exo-hydroxy-2-methylbornane, the major odorous compound produced by several actinomycetes. Environmental Science & Technology, 3(5), 476-477.

MilliporeSigma: Merck KGaA. (n.d.). Universal SYBR Green qPCR Protocol. https://www.sigmaaldrich.com/US/en/technical-documents/protocol/genomics/qpcr/sybr-green-qpcr.

NCBI: National Center for Biotechnology Information (2021). PubChem Compound Summary for CID 15559490, (+)-Geosmin. https://pubchem.ncbi.nlm.nih.gov/compound/15559490.

NCBI: National Center for Biotechnology Information (2021). PubChem Compound Summary for CID 16913, 2-Methylisoborneol. https://pubchem.ncbi.nlm.nih.gov/compound/2-Methylisoborneol.

Nerenberg, R., Rittmann, B. E., & Soucie, W. J. (2000). ozone/biofiltration for removing MIB AND GEOSMIN. Journal - American Water Works Association, 92(12), 85–95. https://doi.org/10.1002/j.1551-8833.2000.tb09073.x

Oh, H.-S., Lee, C. S., Srivastava, A., Oh, H.-M., & Ahn, C.-Y. (2017). Effects of Environmental Factors on Cyanobacterial Production of Odorous Compounds: Geosmin and 2-Methylisoborneol. Journal of Microbiology and Biotechnology, 27(7), 1316–1323. https://doi.org/10.4014/jmb.1702.02069

Olsen, B. K., Chislock, M. F., & Wilson, A. E. (2016). Eutrophication mediates a common off-flavor compound, 2-methylisoborneol, in a drinking water reservoir. Water Research, 92, 228–234. https://doi.org/10.1016/j.watres.2016.01.058

Otten, T. G., Graham, J. L., Harris, T. D., & Dreher, T. W. (2016). Elucidation of Taste- and Odor-Producing Bacteria and Toxigenic Cyanobacteria in a Midwestern Drinking Water Supply Reservoir by Shotgun Metagenomic Analysis. Applied and Environmental Microbiology, 82(17), 5410–5420. https://doi.org/10.1128/aem.01334-16

Parinet, J., Rodriguez, M. J., & Sérodes, J. (2010). Influence of water quality on the presence of off-flavour compounds (geosmin and 2-methylisoborneol). Water Research, 44(20), 5847–5856. https://doi.org/10.1016/j.watres.2010.06.070

Parinet, J., Rodriguez, M. J., & Sérodes, J.-B. (2012). Modelling geosmin concentrations in three sources of raw water in Quebec, Canada. Environmental Monitoring and Assessment, 185(1), 95–111. https://doi.org/10.1007/s10661-012-2536-x

Peter, A., Köster, O., Schildknecht, A., & von Gunten, U. (2009). Occurrence of dissolved and particle-bound taste and odor compounds in Swiss lake waters. Water Research, 43(8), 2191–2200. https://doi.org/10.1016/j.watres.2009.02.016

Smith, V.H., Sieber-Denlinger, J., deNoyelles, F., Campell, S., Pan, S., Randtke, S.J., Blain, G., Strasser, V.A., 2002. Managing taste and odor problems in a eutrophic drinking water reservoir. Journal of Lake and Reservoir Management 18 (4), 319–323.

Srinivasan, R., & Sorial, G. A. (2011). Treatment of taste and odor causing compounds 2-methyl isoborneol and geosmin in drinking water: A critical review. Journal of Environmental Sciences, 23(1), 1–13. https://doi.org/10.1016/s1001-0742(10)60367-1

Sugiura, N., Utsumi, M., Wei, B., Iwami, N., Okano, K., Kawauchi, Y., Maekawa, T., 2004. Assessment for the complicated occurrence of nuisance odours from phytoplankton and environmental factors in a eutrophic lake. Lakes & Reservoirs: Research and Management 9 (3–4), 195–201.

Suurnäkki, S., Gomez-Saez, G. V., Rantala-Ylinen, A., Jokela, J., Fewer, D. P., & Sivonen, K. (2015). Identification of geosmin and 2-methylisoborneol in cyanobacteria and molecular detection methods for the producers of these compounds. Water Research, 68, 56–66. https://doi.org/10.1016/j.watres.2014.09.037

Tsao, H.-W., Michinaka, A., Yen, H.-K., Giglio, S., Hobson, P., Monis, P., & Lin, T.-F. (2014). Monitoring of geosmin producing Anabaena circinalis using quantitative PCR. Water Research, 49, 416–425. https://doi.org/10.1016/j.watres.2013.10.028

Tyc,, O., Song, C., Dickschat, J. S., Vos, M., & Garbeva, P. (2017). The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria. Trends in Microbiology, 25(4), 280–292. https://doi.org/10.1016/j.tim.2016.12.002

Wang, Z., Song, G., Shao, J., Tan, W., Li, Y., & Li, R. (2015). Establishment and field applications of real-time PCR methods for the quantification of potential MIB-producing cyanobacteria in aquatic systems. Journal of Applied Phycology, 28(1), 325–333. https://doi.org/10.1007/s10811-015-0529-1

Wang, C. L., Wang, Z. F., Qiao, X., Li, Z. J., Li, F. J., Chen, M. H., Wang, Y. R., Huang, Y. F. & Cui, H. Y. (2013). Antifungal activity of volatile organic compounds from Streptomyces alboflavus TD-1. FEMS Microbiol. Lett. 341 (1), 45–51.

Wang, M., Yoshimura, C., Allam, A., Kimura, F., & Honma, T. (2019). Causality analysis and prediction of 2-methylisoborneol production in a reservoir using empirical dynamic modeling. Water Research, 163, 114864. https://doi.org/10.1016/j.watres.2019.114864

Watson, S. B. (2003). Cyanobacterial and eukaryotic algal odour compounds: signals or by-products? A review of their biological activity. Phycologia, 42(4), 332-350.

Watson, S., Ridal, J., (2004). Periphyton: A primary source of widespread and severe taste and odour. Water Science Research.49.9, 33-39.

Watson, S. B., Ridal, J. & Boyer, G. L. (2008) Taste and odour and cyanobacterial toxins: impairment, prediction, and management in the Great Lakes. Canadian Journal of Fisheries and Aquatic Sciences 65, 1779-1796, doi:10.1139/F08-084.

Xuwei, D., Min, Q., ren, R., Jiarui, L., Xiaoxue, S., Ping, X., & Jun, C. (2019). The relationships between odors and environmental factors at bloom and non-bloom area in Lake Taihu, China. Chemosphere, 218, 569–576. https://doi.org/10.1016/j.chemosphere.2018.11.121

Yang, L., Liu, S., Tsoka, S., & Papageorgiou, L. G. (2017). A regression tree approach using mathematical programming. Expert Systems with Applications, 78, 347–357. https://doi.org/10.1016/j.eswa.2017.02.013

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., & Madden, T. L. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics, 13(1). https://doi.org/10.1186/1471-2105-13-134

Zaitlin, B., Watson, B., (2006). Actinomycetes in Relation to Taste and Odour in Drinking Water. Myths, Tenetsa nd Truths. Water Research.40.9, 1741-1753.

Zhang, R., Qi, F., Liu, C., Zhang, Y., Wang, Y., Song, Z., Kumirska, J., & Sun, D. (2019). Cyanobacteria derived taste and odor characteristics in various lakes in China: Songhua Lake, Chaohu Lake and Taihu Lake. Ecotoxicology and Environmental Safety, 181, 499–507. https://doi.org/10.1016/j.ecoenv.2019.06.046