The development and application of a universal trait-based model for rapid bioassessment of freshwater systems using diatoms

by

Katherine McKercher, BSc

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Science

in

Geography

Carleton University
Ottawa, Ontario

Abstract

Diatoms are commonly used as indicators of aquatic ecosystem health and can effectively predict values of specific environmental variables. Their use in environmental monitoring programs can be constrained by the expert knowledge that is required for identification to a species level as well as the time and costs associated with identification. Attempts have been made to simplify this identification process by lowering the taxonomic resolution of identification to family or genus level or through automation of identification. These techniques, though functional, do not result in an overall simplification of the identification process that is justifiable against the decrease in certainty of results. A trait-based identification technique using easily identifiable and influential traits for predicting environmental variables could justify the decrease in certainty when considering the decrease in time and cost associated with identification. This could make the use of diatoms as an indicator of ecosystem health more accessible to non-experts.

Acknowledgements

First and foremost I would like to acknowledge and thank my supervisors, Dr. Jesse Vermaire and Dr. Joe Bennett for their continued support and feedback as I wrote my thesis. It has been a wonderful experience learning from you academically and professionally as I start to navigated through the next chapter of my life. Second I would like to acknowledge Dr. Derek Mueller and Richard Schuster who both guided me through R and contributed to the writing of my code. Finally I would like to thank my friends, family and colleagues who have listened, edited, commented and reviewed not just my thesis but much of the other coursework throughout my degree. Thank you all for your support and encouragement.

Table of Contents

List of Tables

List of Illustrations

List of Appendices

Chapter 1: Introduction

Aquatic resources are under pressure from human activities and understanding these pressures and their short and long-term impacts will be important to manage these resources to ensure their future viability (Reid et al., 2019). In order to manage aquatic resources, researchers have relied upon physical, chemical and biological indicators of ecological health to better understand and predict how ecosystems may change in response to environmental stressors. Diatoms (Class Bacillariophyceae) are an extremely diverse group of microscopic algae and are abundant in freshwater and marine habitats. Diatoms are widely used as bio-indicators of aquatic and marine ecological status and are often relied upon as one of the biological indicators of a changing ecological system (Smol and Stoermer, 2010). Diatoms have a cosmopolitan distribution and their ecological tolerances are well researched, contributing to their use as bio-indicators of aquatic ecosystems (Smol and Stoermer, 2010). A large proportion of studies that incorporate diatoms are related to biomonitoring of aquatic habitats. Using diatom assemblages found within the water column and the sediment records of rivers and lakes, researchers are able to better understand the health of an aquatic system.

Methods of using diatoms for biomonitoring are well established across ecological disciplines and can be used to better understand the past and present ecological health of an aquatic system (Davis, 1987; Charles et al., 1990; Hall & Smol, 1992; Dixit & Smol, 1994; Bennion & Simpson, 2011; Pedziszewska et al., 2015; Boeff, Strock & Saros, 2016). Diatoms are particularly useful in biomonitoring as they are sensitive to changes in their environment (Smol & Stroemer, 2010). Their ability to preserve well

within the sediment record allows researchers to infer the state of the historical environment and the magnitude and timing of any changes that may have occurred in the environment over time (Smol & Stoermer 2010; Stevensen & Smol, 2015).

In-depth studies that rely on diatoms as indicators of water quality typically identify diatoms to the species level. It has been recognized that routine diatom identification can be very time consuming and is carried out by expert taxonomists whose services are expensive (Bayer et al., 2001; Bennett et al 2014, 2016). For all of the above reasons, efforts have been made to automate the identification of diatoms. In 1998 a pilot project called ADIAC (automated diatom identification and classification project) was initiated to speed up diatom identification through automation (DuBuf et al., 1999). Researchers involved with ADIAC recognized that many other fields had been successful in using automation to speed up the identification process and thought that diatoms, with their ornate shells and established morpho-taxonomy, could be the perfect candidate for automation (Hicket et al., 2006; Bayer et al., 2001; du Buf et al., 1999). The ADIAC project resulted in the technology being able to identify 37 diatom species with 97% accuracy, under ideal conditions with "perfect" specimens. The ADIAC pilot project ended and there was no further development. The time and effort associated with the project may have been too costly to continue given the tradeoffs of quicker automated diatom identification (Hicks, 2006).

Since the ADIAC project, there has been a trend towards exploring the use of lower taxonomic resolutions in ecological studies involving species richness (Heino & Soininen, 2007), assemblage richness (Bennett et al., 2014), and environmental changes (Heiri & Lotter, 2010). Decreasing taxonomic resolution may decrease certainty and

predictive ability for many of these tests, but it also adds benefits such as decreasing mis-identifications, increasing overall number of identifications for individuals that could not be identified to a species level, and most importantly, decreasing time spent on identification (Heiri & Lotter, 2010). Depending on the goal of the study, investing extensive resources into species identification may not be necessary to obtain desired results and information of comparable quality (Heiri & Lotter, 2010).

Diatoms, although having been previously identified as ideal candidates for automation, remain a group that has not advanced towards simpler or more cost-effective forms of identification. There have been attempts to simplify diatom identification through the use of lower taxonomic resolution but they do not always provide the necessary decrease in cost to justify the reduction in statistical certainty. Bennett et al. (2014) noted that species-level identification is still the most cost-effective method for obtaining the information often required for environmental assessment. The cost associated with field sampling and laboratory work in large studies (>50 lakes sampled) is great enough that the time saved by decreasing taxonomic resolution from species level to genus level will not provide a substantial impact to overall project costs. This being noted, smaller studies and recurring studies, such as monitoring programs, could benefit from cost savings associated with the amount of time saved through faster identification.

This study attempts to provide an alternative method of diatom identification to reduce time, costs, and error associated with traditional species level identification, using a list of previously compiled diatom traits from Cormier et al. (2020) that broke down hundreds of diatom species into a list of 20 physical traits. Using these data, in combination with lake water quality data associated with the environment in which the

diatoms were found, a list of potential traits that are correlated with specific or various water quality parameters were identified. This method can simplify the identification process by allowing non-experts or researchers who are new to the field to be able to extract meaningful information about the state of the environment using diatoms. Cormier et al. (2020) discovered that trait-based diatom identification methods were able to successfully predict the value of specific environmental variables with varying degrees of accuracy. With that success, the objective of this thesis was to identify which of these 20 physical traits, alone or in combination, had the strongest relationship to environmental variables and could be used to provide reliable estimates of indicators of ecological health.

Chapter 2: Literature Review

2.1 An overview of bio-indicators

A bio-indicator can be broadly defined as any organism that provides information about the quality of the environment in which it exists (Li, Zheng & Liu, 2010; Market, 1999). Bio-indicators are used to help create a better understanding of past and present environments and how human activities may be altering ecosystems. Bio-indicators that have been used as a proxy to understand the environment include but are not limited to bryophytes (Ford & Hasselbach, 2001; Markert et al., 1999), trees (Markert et al., 1999), macrophytes (Markert et al., 1999), phytoplankton (such as diatoms)(Dixit, Dixit, Smol, 1992), insects (Cain et al., 1992), parasites and bivalves (Sures, 2000) (Parmar, Rawtani & Agrawal, 2016; Li, Zheng & Liu, 2010).Unlike their chemical or physical counterparts, bio-indicators provide information about a system in a broader context as they are able to incorporate a temporal component corresponding to the lifespan of an individual (Holt & Miller, 2011).

Bio-indicators have been used broadly since the late 1800s when researchers discovered that species living in a "healthy" system were completely different from the species living in a contaminated, algae dominant system (Holt & Miller, 2010). The effectiveness and use of an organism as a bio-indicator is generally based on seven properties: easy identification, cosmopolitan distribution, low mobility, well studied, abundant, cost-effective, and sensitive to stressors (Li, Zheng & Liu, 2010).

While chemical and physical measurements of water quality provide a brief snapshot of the state of a system, bio-indicators can provide information on the state of the

environment and ecological stressors acting on an environment over short periods of time (i.e. oil spill), up to several years, and long periods of time (i.e. climate change; land use changes) (Bellinger & Sigee, 2010). Cumulative changes in the environment are considered long-term changes and can be seen through changes in species composition in the environment (Holt & Miller, 2011). Bio-indicators can also provide information about an individual water quality metric (i.e. pH, dissolved oxygen, total phosphorous), combinations of many metrics, and interactions among parts of their environment and stressors acting upon it (Iliopoulou-Georgudaki et al., 2003). Species assemblages are very commonly used as indicators of environmental change within paleolimnology and limnology. Changes in species assemblages are indicative of environmental change as individual species have known environmental preferences and tolerances to environmental stressors (Bellinger & Sigee 2010; Oertel & Salanki, 2003). In addition to what an individual bio-indicator may be able to tell us about the environment, multiproxy approaches involving the use of multiple indicator species that often coexist within a system (i.e. Chrysophytes and diatoms used concurrently in paleolimnology) can provide very meaningful information about the state of the ecosystem (Bellinger & Sigee 2010, 109; Dixit & Smol, 1994).

Functionally, the type of prediction a bio-indicator is making can be broken down into four categories: signal function, prediction function, control function and scientific research to inform policy (Oertel & Salanki, 2003). A signal function provides early warnings for any possible adverse changes in the environment. Plankton and other microorganisms are proxies that are often used as early detectors of a changing systems as some are capable of releasing stress proteins which can be interpreted as an early

warning sign of change (Parmar, Rawrani & Agrawal, 2016). The prediction function deals with the prediction of changes that will occur and if they can be improved. Marine plants are often used to predict changes that may be occurring within a system as they are immobile, their presence or absence in a system provides information about the health of its environment (Parmar, Rawtani, Agrawal, 2016). The control function is concerned with observing if restoration or conservation efforts are improving the quality of the environment. This can be observed using a number of indicators through their presence, absence or return to a system (Oertel & Salanki, 2003). Finally, many indicators will be used for the purpose of scientific research and in environmental monitoring for the purpose of informing policy and advancing our understanding of ecosystems and how we interact with them (Oertel & Salanki, 2003).

2.2 Advantages and disadvantages of bio-indicators

In addition to being excellent proxies for studying ecosystem dynamics, an advantage to using bio-indicators is that they are readily available and cost effective in comparison to technological alternatives (Bellinger & Sigeem, 2010; Zheng & Liu, 2010; Holt & Miller, 2011). The global community of researchers who rely upon bio-indicators have created an accessible and reliable method for using many different indicator species, with many guides available in multiple languages. Many indices have been developed to help researchers to measure pressures on the environment through the presence or absence of various species. Indices have been created to infer information such as trophic status (trophic diatom index; Kelly & Whitton, 1995), organic pollution (Saprobity index, % pollution tolerant taxa, Diatom Assemblage Index for Organic Pollution; Sládeček, 1986; Kelly and Whitton, 1995; Van Dam et al., 1994; Kalyoncu et al., 2009;

Karthick et al., 2010). These example indices, though specific to diatoms, represent a small fraction of the indices available for researchers in all fields.

Though the techniques associated with using bio-indicators are easily accessible to most researchers, there remain some drawbacks to their use. The greatest drawback may be the inability of bio-indicators to provide precise quantitative measurements of water quality variables that are most often used to monitor and manage ecosystems (Holt & Miller, 2010). Another drawback that is not necessarily unique to bi-indicators, but any singular method of monitoring, is that they cannot provide insight into changes within the environment that are due to natural variability and those that are due to anthropogenic influence. This distinction between natural variability and anthropogenic changes can be inferred through the use of multiple methods, but the use of a single type (i.e. specific species) or class (i.e. insects) cannot alone make the distinction (Holt & Miller, 2010). Another possible drawback of bio-indicators is the possibility that the distribution or presence of an indicator species may not be solely influenced by the pertinent stressor (Holt & Miller, 2010). Finally, for studies that intend to use the inferred information from the bio-indicators to influence water resource management, the indicator species may not provide a holistic view for all the potential species that exist within that ecosystem (Holt & Miller, 2010).

2.3 Environmental monitoring programs

Water quality monitoring programs throughout North America and Europe have implemented diatom indices in their monitoring efforts. The European Diatom Database Initiative (EDDI) was funded by the European Union (EU) as it showed potential in

biological monitoring (Battarbee et al., 2000; Bellinger & Sigee, 2010). The project took place between 1998 and 2000 and combined various diatom community datasets from across Europe, and parts of Africa and Asia to help enhance the use of diatoms in surface water analysis (Battarbee et al., 2001; Bellinger & Sigee, 2010). Many countries throughout Europe have implemented their own protocols and indices for water quality monitoring using diatoms. The EU adopted, through policy, the use of the *Water Framework Directive* in the year 2000 and it places considerable emphasis on the use of phytobenthos as bio-indicators, especially for monitoring and describing the effects of eutrophication (European Union, 2000). As a more locally relevant example, the *Algal Bioassessment Protocol* was initiated in Ontario to complement other water monitoring protocols in Ontario rivers (Ontario Ministry of the Environment, 2011). This protocol provides emphasis to diatoms as a new method to assess water quality (MOECC, 2011).

2.4 Diatom life history and ecology

Diatoms cells are composed of a siliceous outer shell surrounding and protecting the inner plasma and organelles (Julius & Theriot, 2010). More specifically, the outer shell is composed of two valves of slightly different sizes connected by a series of bands, all referred to as a frustule (Julius & Theriot, 2010). The frustule is extremely porous, allowing the internal organelles to interact with the surrounding environment (Julius & Theriot, 2010). During asexual mitotic reproduction, the two valves separate, and each become the larger valve of a new cell. This results in one daughter cell the same size as the parent cell and another smaller cell that is slightly smaller than the parent cell (Julius & Theriot, 2010). The continuous asexual reproduction results in a mean decrease in cell

size. The original cell size can be restored through sexual reproduction, which is thought to be triggered through a size or environmental cue (Julius & Theriot, 2010). Diatoms are one of the few classes among algae that have a diplontic life cycle, meaning they have both meiotic and mitotic reproductive stages using single cells and multicellular gametes (Mann, 1993). To add to the complexity, different families of diatoms are subject to variations in mitotic divisions, making it difficult to generalize the specifications of sexual reproduction (Mann, 1993).

2.5 Ecological controls on diatom assemblage structure and growth

Diatom growth rates are limited by various nutrients in both freshwater and marine systems, each with different physiological and metabolic implications. In marine systems, a major limiting nutrient for growth is silica; the silica deposition vesicle essentially takes up silicon dioxide from the environment and concentrates and deposits it in the cell wall (Jezequel, Hildebrand & Brzezinski, 2000; Julius & Theriot, 2010). In freshwater systems diatom growth is limited more often by other nutrients such as phosphorous which plays an important role in photosynthetic activities (Martin-Jezequel, et. al., 2000). Although phosphorous is notable and important in diatom life history, there remain trace elements such as lead, copper and aluminum that may also influence cell size and morphometry, though their influence is likely indistinguishable when considering the influence of nutrients through the lens of bio-indicator work (Gensemer, 1990).

2.6 Diatoms as bio-indicators in monitoring and paleolimnology

Many species of diatoms are known to be highly sensitive to environmental stressors, contributing to their frequent use in biomonitoring and other ecological studies (Ambasht & Ambasht, 2003; Smol & Stoermer, 2010). Diatoms are particularly effective as indicators of lake water pH, salinity and total phosphorous, which play an important role in shaping biotic components of aquatic ecosystems (Smol, 1992; Dixit & Smol, 1994; Wilson et al. 1996; Bellinger & Sigeem, 2010).

Diatoms possess many basic requirements of bio-indicators as stated by Dixit et al. (1992) including being easily identifiable, sensitive to environmental conditions in a quantifiable manner, are widespread and abundant, and provide background or reference information on the past state of the environment in which they existed. In addition, more general requirements have been outlined by various authors that include being cost-effective, well-studied, suitable for laboratory experiments, and highly sensitive to environmental stress (Bellinger & Sigee, 2010; Li, Zheng & Liu, 2010). Diatoms not only meet all of these requirements but also meet them more fully than other aquatic organism (Dixit et al., 1992). More specifically, diatoms are well preserved within the sediment record: their ornate silica shells preserve well within the sediment record of lakes and oceans (Julius & Theriot, 2010). Physiological properties such as fast reproduction time allow for rapid population turnover when conditions become more favorable for one species, and less favorable to another. Reproduction rates can reach up to one division per day for certain species under ideal conditions, contributing to their use in paleolimnology studies looking at changing lake conditions over time (Julius & Theriot, 2010). Their rapid reproduction and division times also allow diatoms to be easily studied in lab-based

experiments, providing researchers with opportunities to better understand how certain environmental factors may influence their life habits (Julius & Theriot, 2010).

The use of diatoms as a bio-indicator is based on their species-specific tolerances to various water quality parameters (Battarbee, 1984; Stoermer, 2001; Bellinger & Sigee, 2010; Holt & Miller, 2010). Simply put, species living in their optimum environment will be more abundant within a system, and as species move away from their environmental optima they become less abundant (Figure 1). This theory is widely applied to research involving bio-indicators (Holt & Miller, 2010).

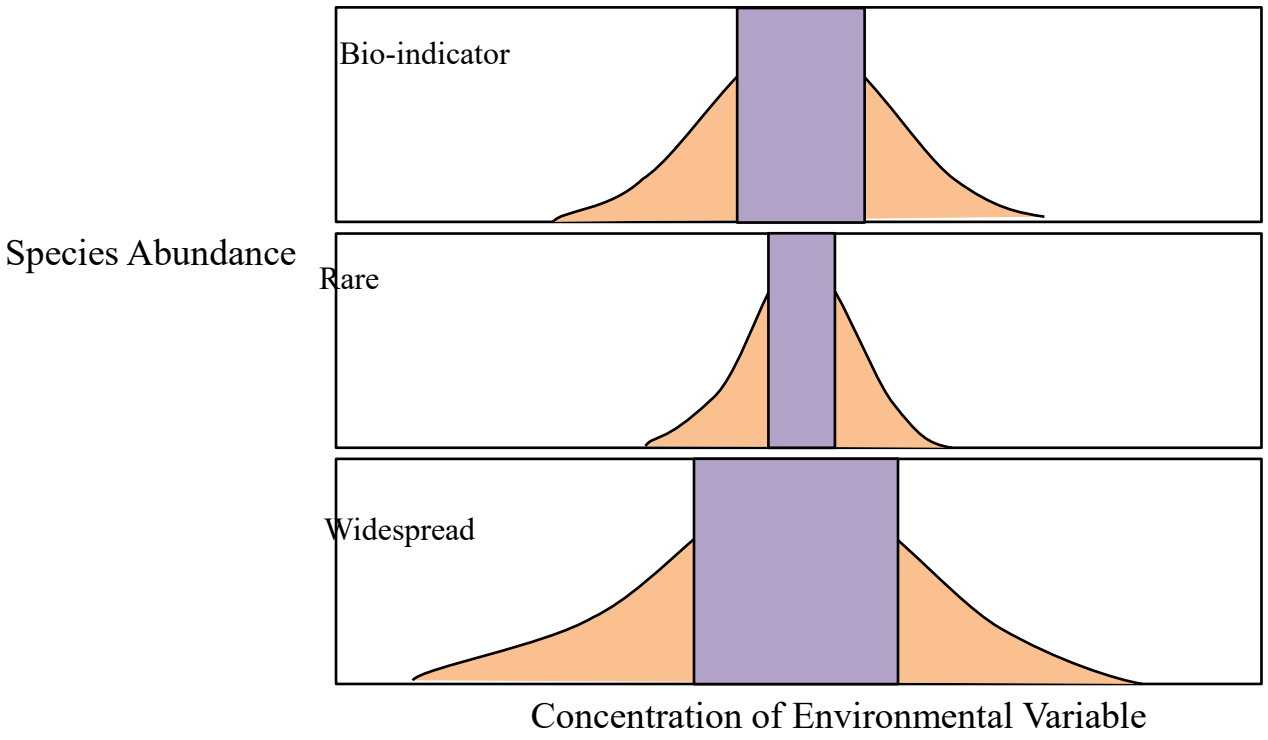


Figure 1. Relative environmental tolerances and optima of bio-indicators, rare species and ubiquitous species. Orange area represents a species distribution while the purple box shows an individuals' optimal tolerance (modified from Holt & Miller, 2010).

Diatoms are heavily relied upon in biomonitoring of rivers and lakes as they have predictable tolerances within both lentic and lotic systems (Oertel & Salanki, 2003; Bellinger & Sigee, 2010; Chang et al., 2014). Within rivers, live diatom samples are used as part of a multiproxy approach that relies on multiple indicators to provide insight into the physical, chemical and biological components that support a diverse community (Hering et al., 2006; Bellinger & Sigee, 2010; Li Zheng & Liu, 2010). Diatoms in lake systems are more often used as a tool for the reconstruction of past environments due to their abundance within the sediment record and the water column (Smol, 1992; Dixit & Smol, 1994).

Changes in diatom assemblages in river systems are associated with high and low flow, as well as nutrient inputs and seasonal changes (Garnier, et al., 1995). In addition, changes in diatom assemblages are influenced by various other physiographical, chemical and human drivers, requiring the use of multivariate analysis (Patapova & Charles, 2002; Leira & Sabater, 2005). According to Oertel and Salanki (2003) pollution in rivers effects four dimensions that must all be considered when assessing water quality. The longitudinal dimension relates directly to the river continuum concept and the common and predictable changes that occur along the length of a river system (Vanote et al., 1980; Oertel & Salanki, 2003). The latitudinal dimension refers to a cross section of the river including the riparian zone, floodplain, and watershed (Oertel & Salanki, 2003). The vertical dimension including water depth and sediment-water interactions shows seasonal variability and contributes to community assemblage changes (Oertel & Salanki, 2003). Finally, the temporal dimension can include anything from natural to human induced changes within the system that occur over time (Oertel & Salanki, 2003).

Similar to river biomonitoring, lake biomonitoring also relies on a multiproxy approach to water quality assessment. Common methods of examining changes in lake water chemistry using diatoms begins with taking sediment cores, followed by isolating diatom assemblages by cleaning and identification of species (Stockner & Benson, 1967; Dixit & Smol 1992; Wilson, Cumming & Smol, 1996; Smol et al., 2005; Ruhland, Paterson & Smol, 2008). Based on the diatoms' species-specific tolerances to various chemical properties of water, the past conditions of a lake can be inferred based on the changing species composition through time (Dixit & Smol, 1994).

2.7 Cost associated with Diatom based monitoring

The time and costs associated with diatom identification can be attributed to more than just the need for expert knowledge for accurate identification, but also the methods and costs associated with sample collection and preparation. Current methods of diatom frustule counting begin with extracting a subsample and mounting it on a microscope slide. A predetermined number of diatom frustules (usually between 300 and 600) are then counted and classified from the slide (Battarbee et al., 2002b). They are then subject to one or several methods of evaluating species abundance and inferring various water quality parameters (Kelly & Lewis, 1996; Werner, Adler & Drebler, 2016). It was estimated by Bennett et al (2014) that average costs of diatom identification to species level is ~$465 USD per sample (including sample prep). In multi lake studies, paleolimnological, and long-term studies it is easy to see how identification costs could quickly accumulate (Battarbee et al., 2002b; Bennett et al., 2016).

2.8 Diatom identification

Methods of diatom identification within the field of biomonitoring have remained relatively unchanged since the 1980s, with most studies relying upon traditional species identification using light microscopes. The general sequence of events carried out during the identification process remains very similar across most paleolimnological and ecological studies (Davis, 1987; Charles et al., 1990; Hall & Smol, 1992; Dixit & Smol, 1994; Bennion & Simpson, 2011; Pedziszewska et al., 2015; Boeff et al., 2016). Diatom identification can be done using either electron or light microscopes but many of the features that are being used for identification remain the same. Some key cell features including the shape, length, width, and the presence/absence of certain features as well as their size, shape, and number or density are used to identify cells to a genus or species level (Hasle & Fryxell, 1970).

2.8.1 Limitations of traditional diatom identification – Classification

Traditional diatom identification does not consider many aspects of diatom life history that may influence morphological features. The field of diatom taxonomy is constantly classifying and reclassifying both newly discovered and well-known species, in part due to the advancement of molecular techniques that have allowed researchers to better understand the linkages between species (Julius & Theriot, 2010). There have been many attempts to create a phylogenetic tree based on a monophyletic classification system, where species are grouped evolutionarily to a common ancestor. This has yielded successful results up to a certain point (Julius & Theriot, 2010; Tapolczai et al., 2017). Unfortunately this method is not possible with some ancient species from which genetic

material could not be obtained and it is therefore not possible to create a definitive evolutionary tree (Julius & Theriot, 2010). This can be particularly problematic for long-term studies, since changing species or genus classification can result in inconsistencies in species identification and in turn estimates of environmental change that rely upon species as a proxy for various environmental factors (Kahlert et al., 2009; Straile, Jochimsen & Kummerlin, 2013; Tapolczai et al., 2017).

2.8.2 Limitations of traditional diatom identification - Life history complications

Diatoms are a complex class of eukaryotes with many known and unknown cryptic species. To further this complexity, it has been observed that diatoms are subject to a certain level of phenotypic plasticity associated with the changing state of their environment (Falasco & Badino, 2011). This phenotypic plasticity can result in polymorphisms of a single species bearing different appearances and behaviours (Falasco & Badino, 2011). The most common polymorphism in diatoms relates to size or length to width ratio, which are commonly influenced by pH, nutrient limiting conditions and heavy metal concentrations (Falasco & Badino, 2011). Falasco and Badino (2011) summarized various environmental factors that influence cell morphology that can contribute to morphological plasticity, for example salinity can influence cell size and ornamentation while silicon dioxide limiting conditions can influence cell size, valve thickness and areolae density (Table 1) (Falasco & Badino, 2011). These naturally occurring forms of variability mean that diatom frustule characteristics may not be constant for species between or within aquatic systems. For identification purposes it is

assumed that all species within a system, subject to the same conditions, are likely to demonstrate the same form, whether it is a polymorphism or not (Falasco & Badino, 2011).

Table 1. Various environmental factors and their effects on diatom cell form. Modified from Falasco & Badino, 2011.

Environmental Factors	Morphological Plasticity (Valve Character Affected)
Dissolved organic carbon	cell size
Light intensity	cell size
Nutrient (limitation or eutrophication)	cell size; valve thickness; striae density (in 10um); mean number of areolae and fascicles
pH	cell size
Salinity	Valve thickness; ornamentation development (spine and costae); striae length; number and placement of fultoportulae and rimomportulae; fibule density; changes in striae punctiation (biseriate-uniseriate)
Silica limiting conditions	cell size; valve thickness; spine density or presence; areolae density; loss of central areolae or striae; loss of tangential areolae wall; velum position; pore membrane thickness; loculus depth
Temperature	cell size; valve thickness
Artificial conditions	cell size; valve thickness; morphotype production
Microtubule polymerization	number and placement of fultoportulae and rimoportulae
Mixed metal solution	number of cells forming colony

In addition, diatoms may also be subject to teratological forms which are changes in diatom form that are non-adaptive and made in response to a stressor, such as a toxic environment, the resulting cells are mis-shapen and can lead to changes in physiological characteristics (Falasco et al., 2009; Falasco & Badino, 2011). Unlike the polymorphisms associated with phenotypic plasticity the teratological forms often demonstrate more extreme deformations in the frustule and can dramatically alter identification. These forms do not have a genetic basis and are therefore considered non-adaptive (Falasco &

Badino, 2011). Within a taxon where reproduction can be rapid, these changes in cell form can quickly manifest themselves within the assemblage, as cells are known to respond in a similar way to the stressor (Falasco & Badino, 2011). In a field where diatoms are relied upon to help identify eutrophication and changes in water quality one should consider these changes as potentially influential to ecological studies and could go as far as to provide a useful insight into changes in the environment over time.

Although polymorphic and teratological diatom forms are known to occur under environmental stress it would be extremely difficult for a taxonomist or an ecological researcher relying upon published general guides to incorporate these factors into an already technically challenging identification scheme. There already exist discrepancies among researchers and cohorts when it comes to identifying some more difficult species, the inclusion of some potential alternate forms of species would add to these discrepancies and further complicate an already complex and difficult identification process (Stoermer, 200; Werner et al.,, 2016; Tapolczai, 2017). The use of published guides help make the use of diatoms in ecological studies possible, but it does not take away from the fact that this process remains time consuming and cost intensive.

2.9 Quantitative methods to evaluate environmental variables using diatoms

Paleolimnological and other ecological studies that rely upon diatoms (or other similar bio-indicators) will use various statistical methods to help evaluate and predict the value of the environmental variable of interest. Many paleolimnological studies rely on ordination techniques to help evaluate and display the relationships between species and components of their environments (Juggins & Birks, 2010). In this study, the goal was to

determine the relationship between trait groupings or individual traits in the prediction of environmental variables. To do so, I used weighted average regressions (or weighted averaging), a standard technique often used for relating species assemblages to their environment (Juggins & Birks, 2010). Other similar studies that aim to better understand the relationship between species, their traits and the environments also rely on this method (Bennett et al., 2014, 2016, Cormier et al., 2020).

Weighted averaging assumes that a species has a unimodal distribution along an environmental gradient and will be present in greater abundances at its environmental optima, and lower abundances as the environment moves farther away from its optima (Figure 2). Using large datasets with many overlapping species allows for better estimates of environmental variables. Species optima can be established based on their presence and abundance within differing environments, with the assumption that they will be more abundant in their optimum environment. This method is often used to infer what the environment may have been in the past when other physical or chemical measurements are not possible.

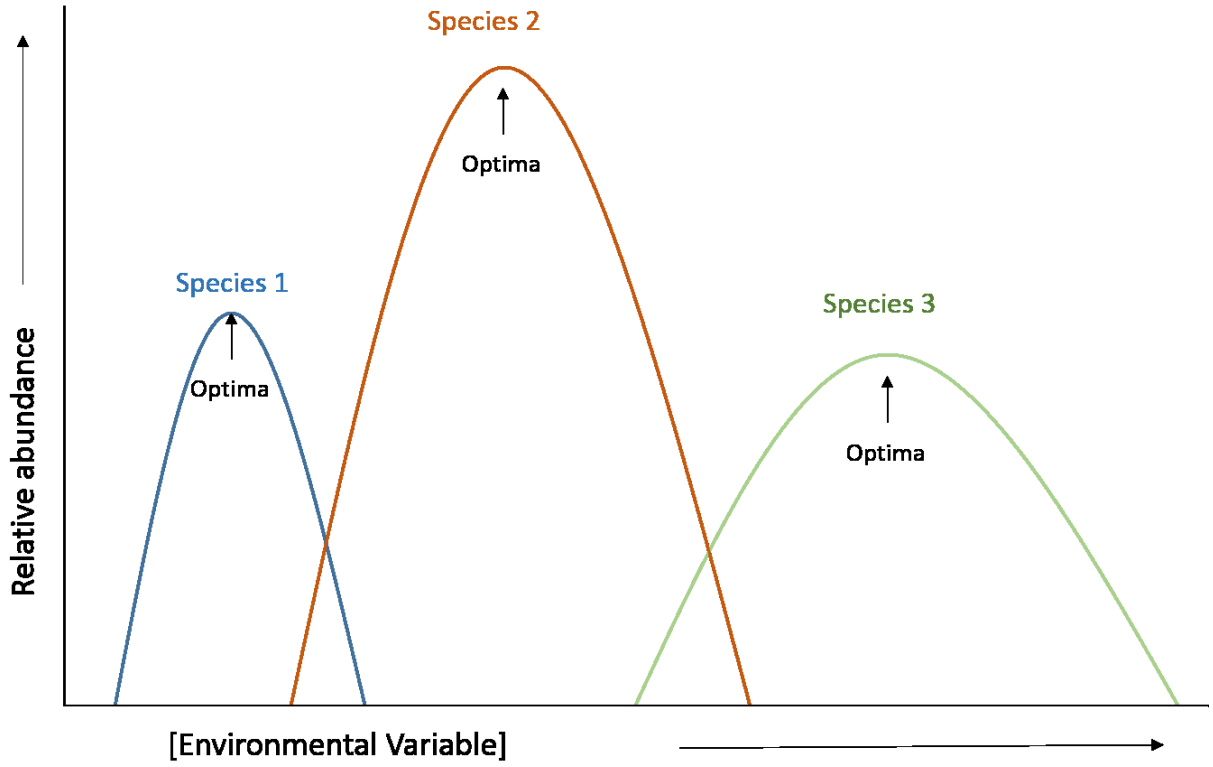


Figure 2. Relative abundance of species versus a gradient in an environmental variable. Species exist at a higher relative abundance at an optimal level of an environmental variable, and lower abundances as the environment moves away from optimal conditions.

Species optima can be determined using weighted averaging by averaging the values of the environmental variable against the relative abundance of the species present at various concentrations of that variable (Hall & Smol, 1992). In other words, the relative abundance of a species is multiplied by the value of the environmental variable and divided by the total sum of relative abundances of the species within a dataset or across multiple sites. In this study the same species exist within multiple sites at different relative abundances which are assumed to be influenced by the chemical, physical and biological properties of their environment. To determine the average value of a specific

variable, the optimum of each species is then multiplied against its abundance and divided by the sum of the relative abundances once more (Juggins et al., 2012). Weighted averaging can use relative species abundances within various sites to predict the average value/concentration of an environmental variable and therefore theoretical species optima for an individual species.

Tolerance downweighting is used for species that have the ability to survive in a broad range of conditions and therefore have broad tolerances to a specific environmental variable as they will not provide as much insight into their environment as a species with a narrow tolerance existing in the same environment. Downweighting will weigh taxa by the inverse of their squared tolerance to the environmental variable, giving them less weight in estimating environmental values compared to species with a narrower environmental tolerances.

Performing a weighted average regression is a two-step process, resulting in the abundance being averaged out twice. Each time the data is averaged the data will move closer towards to mean, possibly losing information for species that appear at the ends of the gradient (further from the mean). Deshrinking is used to take values further away from the mean as a result of the weighted averaging resulting in an outcome that may be greater or lesser than the actual test result (Birks et al., 2012; Horton, Edwards & Lloyd, 1999; Birks et al., 1990). In the case of the data used in this study, classical deshrinking may reduce the predictive power of the model by excluding data located near the mean environmental variable values of the lakes, while inverse deshrinking may result in an overestimation of data at the lower and higher ends (Birks et al., 2012).

Chapter 3: Methods

Using diatom and environmental data compiled from 1214 lakes across North America and Western Europe that had been previously compiled for Bennett et al. (2014), Cormier et al., (2020) created a trait based dataset classifying all diatom taxa present in the lakes by 20 morphological traits (hereafter referred to as the trait datasets) (Table 2**).** These traits were selected to be easily recognizable under light microscopy with minimal training in diatom morphology. The traits were used to identify species present in each lake and organized based on the environmental variable being observed.

Table 2. List of 20 traits selected by Cormier et al. (2019) to form trait-based

Trait Number	Description of trait
Trait 1	Centric, Disc-shaped: valve margins without spines and valve with outer zone of striae
Trait 2	Centric, Disc-shaped: valve margins with spines and valve with radial rows of punctae
Trait 3	Centric- Cylindrical: tubular
Trait 4	Elliptical to Lanceolate (oval shaped and could shape to a point at either end- could have polar inflations)
Trait 5	Linear with Central and Polar Inflations: valve length varies
Trait 6	Cruciform: Elliptical with central inflation
Trait 7	Very Long and Narrow/Skinny: needle to spindle-shaped, with or without polar and/or central inflations
Trait 8	Linear to Linear-Lanceolate: sides of valve are quite parallel (i.e. linear)
Trait 9	Rhombic and Rhombic-Lanceolate
Trait 10	Sigmoid
Trait 11	Crescent to semi-circular: Cymbelloid
Trait 12	Clavate:-Club/wedge-shaped, gomphonemoid
Trait 13	Elongate and Asymmetrical: heteropolar with inflated ends
Trait 14	Arcuate: curved like a bow
Trait 15	Bi-raphed: centrally located
Trait 16	Bi-raphed: shifted to one side
Trait 17	Bi-raphed: with siliceous struts/ribs (fibulae/transapical costae)- Keeled or Canal-bearing diatoms
Trait 18	Pro-raphed: small raphe at the pole ends
Trait 19	Mono-raphed: diatoms with a raphe on only one valve of the frustule
Trait 20	A-raphed: diatoms without a raphe on either valve

3.1 Dataset development

Environmental data for 1214 lakes was also collected at the same time as the diatom data (hereafter referred to as the environmental variable datasets) used in Bennett et al., (2014). These data have also been used in multiple studies relating to trait based analysis, optimization, and cost-effective sampling (Battarbee et al., 2001; Ginn et al., 2007; Bennett et al., 2010; 2014; 2016; Cormier et al., 2020). The environmental variable datasets were created by subsetting large publicly available datasets that document water quality throughout North America and Europe. The two environmental variable datasets containing surface water pH data were compiled from 493 North-Eastern North American lakes over 671 000 km^2, and 488 lakes from North-Western Europe over a combined 1.3 million km^2, and were originally compiled to examine the effects of acid precipitation on diatom assemblages (Battarbee et al., 2001; Ginn et al., 2007; Bennett et al., 2014). The pH for all lakes was measured using a handheld pH meter with the same protocol to ensure consistency across all lakes (Bennett et al., 2014 ; Ginn et al., 2007; Battarbee et al., 2001).

The environmental variable dataset containing surface water total phosphorous concentration data was created by taking a subset of 233 lakes from the USEPA Environmental Monitoring and Assessment Protocol (EMAP) (Bennett et al., 2014). These lakes were all located within a similar 410 000 km^2 area as the lakes sampled for pH in the North-Eastern North American dataset (Bennett et al., 2014; Bennett et al., 2010). Finally, the surface water salinity and lake depth environmental variable datasets were created from data for 207 lakes over 77 000 km^2 in Southern British Columbia

which were originally used to assess the reliability of salinity inference models using diatoms (Bennett et al., 2014; Wilson, Cumming & Smol, 1996).

3.2 Description of code

Statistical analyses were performed in R version 3.6.3 and Excel. Before analyses were performed, preliminary steps for each dataset were completed, including data formatting and organization, taking place both in R and in Excel. This step involved transforming the species trait dataset to allow for R to process the traits a species possess as a string of numbers. The data were presented in a way that allowed for a species to be identified using a combination of 20 different traits. This data represented a species through presence or absence of the 20 traits, similar to a binary string of ones and zeros (1 for presence of a trait, 0 for absence). This string can be interpreted as a trait-based barcode with each species now being recognized by the barcode rather than a species name. The barcode method also allowed for simple manipulation such as trait removal or replacement without having to reassign the newly grouped species to another name.

We created various functions to simplify the data analysis process. The "barcoder" function extracted the columns from the trait dataset containing the presence and absence data and created strings of ones and zeros that were then used to identify a species. Each species was now associated with a barcode and multiple species could be then identified with the same trait-based barcode. Following the transformation of the data into trait-based barcodes individual traits could be sequentially removed using a "trait remover" function. Upon the conversion into barcodes and combining all species with unique trait-based barcodes, all the species abundances for the species that had been

combined under one barcode would have to be summed. A function was then created that matched species barcodes to their respective species in the environmental variable dataset and summed their relative abundances. The resulting summed abundances were stored in a new dataset, which was then used in the final weighted average regression. A final function was made to run a weighted average regression and store the results in another dataset to easily compare the R^2 and root mean squared error (RMSE) values of the predicted environmental variables (Figure 3).

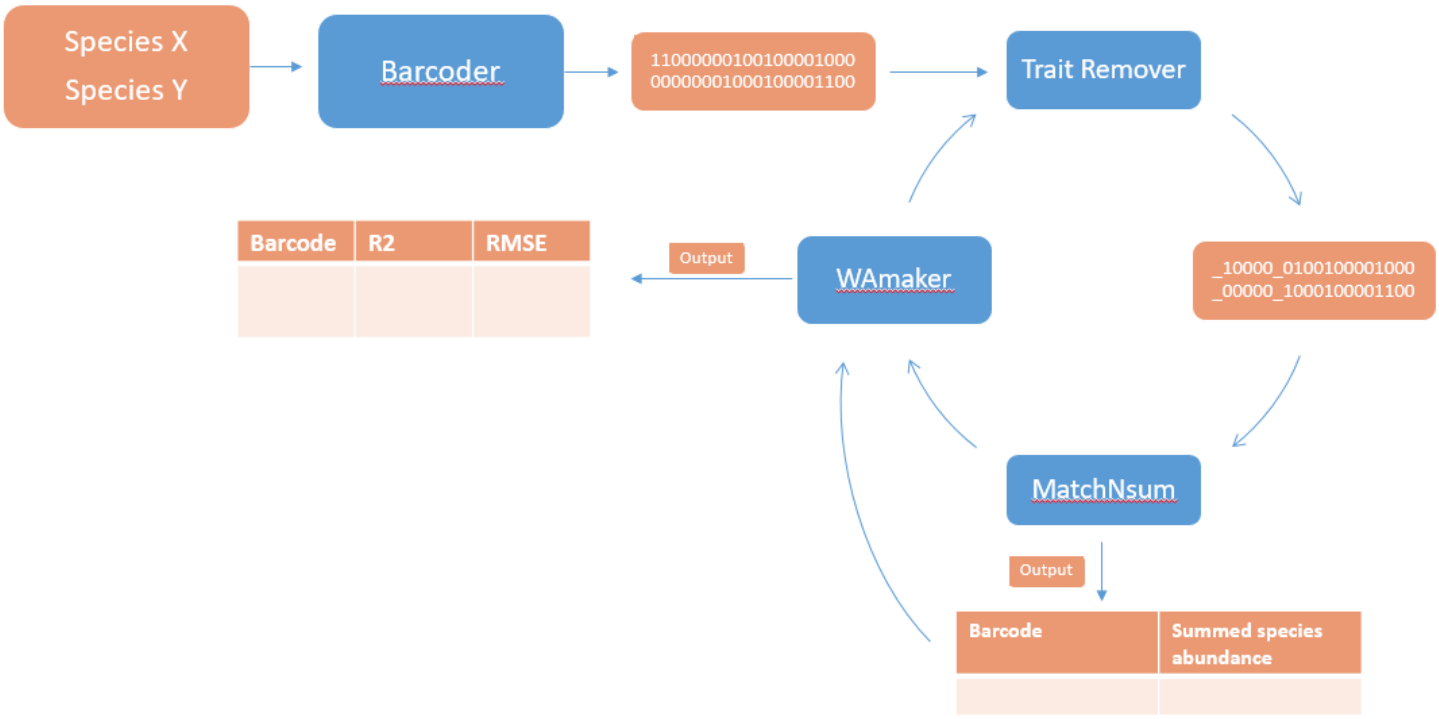


Figure 3. Flowchart showing functions (blue), and the inputs and outputs (orange) used to evaluate the effectiveness of functional traits at evaluating environmental variables

Upon completing all the initial setup, a loop was written that runs many weighted average regressions to help determine which traits showed the greatest predictive ability for each environmental variable based on the resulting R^2 and RMSE values. The loop sequentially removed traits each round until one million iterations were reached, ultimately permanently removing up to a maximum of 14 traits, under the assumption that the removal of more traits would result in R^2 values so low they would not provide meaningful insight into the quality of the environment. The loop was designed to remove a trait, regroup the barcodes by unique barcode, associate the new barcodes with any species matching the new code, sum their abundances, and run a weighted average regression. The results of each regression were added to a dataset which could later be graphed and used in other tests to determine which traits were the most or least effective at predicting the environmental variable as well as the optimum number of traits to use in tests.

In order to evaluate the predictive power of each trait, vectors were created evaluating how many times a trait appeared within the top 0.05% (500 of 1000 000) of values returned for each of the environmental datasets. The vectors were created to separate the R^2 values into groups defined by how many traits were removed. This was done to account for the fact that the R^2 values were generally higher when more traits were present, but the slightly lower R^2 values from a group with more traits removed could still offer valuable information about the relative importance of each trait. The top 0.05% of R^2 values and the trait-based barcodes which they were associated with were then recombined and plotted for comparison.

In addition to evaluating the effectiveness of each trait, the number of traits was also assessed to determine if there was an optimal number of traits that could be used to predict the environmental variable of interest. To do this, the vectors of R^2 and RMSE values grouped by number of traits removed were compared to the original values using all 20 traits and compared with the values resulting from the unaltered species abundance data.

3.3 Statistical tests

The goals of this study were to determine if a trait based method could effectively predict environmental variables and which traits are the most effective at predicting the environmental variables. Based on previous work on these datasets (Bennett et al., it was determined the best test for reconstructing environmental measurements using trait data would be a weighted average regression (Juggins & Birk, 2012). A weighted average regression incorporates relative abundance of organisms in various environments and the corresponding characteristics of the environment (Hall & Smol, 1992; Juggins & Birk, 2012).

I used weighted averaging on one million different trait combinations, created through sequential removal and replacement of 20 traits, to determine which traits and trait combinations have the greatest influence on predicting the value/concentration of an environmental variable. Models were developed based on all possible combinations of traits, using different total amounts of traits in the analysis, so that all possible combinations of traits were evaluated for decreasing number of total traits used, down to a minimum of 6 traits (14 traits removed) which was reached upon 1 million iterations. Tolerance downweighting and deshrinking were used in all model development to

improve the performance of the model in predicting the environmental variable of interest.

Prior to evaluating whether individual traits performed well in predicting environmental conditions, baseline tests were performed through weighted average regression on both the species-based data, and the trait-based data prior to any trait removal. This baseline data helps to determine how well the trait-based model is able to perform in comparison to the standard species-based model. In addition to determining if the trait-based models performed well compared to more traditional species based models, we were also interested in examining if trait-based models were equally effective for all of the environmental variables of interest (pH, TP, salinity, lake depth).

Correlation analysis was performed on all 20 traits to determine if any of the traits were correlated, as some traits were mutually exclusive. For example, there were multiple categories of centric-shaped diatoms that were mutually exclusive. A diatom could not be *centric-disk shaped* (traits 1 and 2)and be classified as having both *valve margins with* (trait 2) or *without spines* (trait 1). In addition to mutually exclusive traits, it is possible that other traits could have been associated with one another. Correlation analysis was performed in R using Pearson correlation tests. Pearson correlation is the most commonly used correlation analysis method which assumes normal distribution. The correlations were done on each unique dataset (North American and European pH datasets, and the BC dataset which provided information on phosphorous, salinity and depth).

Chapter 4: Results

4.1 Species-based versus trait-based transfer function models

The North American pH, European pH and Salinity models showed excellent predictive ability when using the species datasets with R^2 values of 0.86, 0.84, and 0.90, respectively and RMSE values of 0.44, 0.39 and 0.31, respectively (Table 3 and Figure 4 (a)-(f)). The water depth transfer function model using the species data also had strong predictive ability with an R^2 value of 0.75 and RMSE value of 0.33, while the phosphorous dataset had a lower R^2 value of 0.48 and RMSE value of 0.27 (Table 3 and Figure 4 (g)-(j)). Compared to the species-based models the trait- based models resulted in lower predictive ability with R^2 values consistently declining by 0.23 units, and the largest decrease in predictive ability being the salinity dataset at 0.28 (Table 3 and Figure 4).

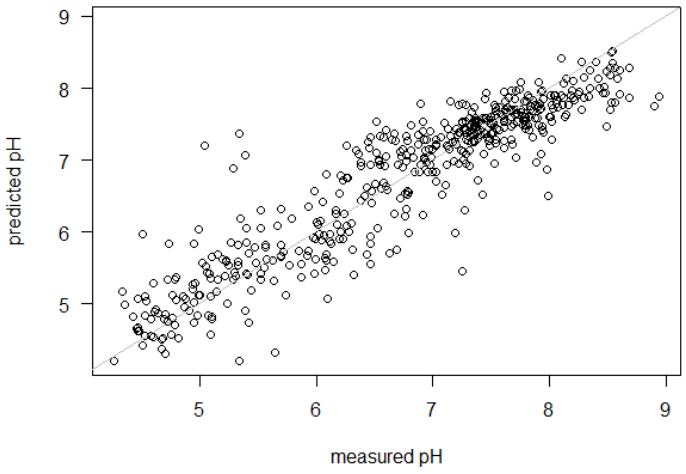
Table 3. Summary of preliminary results showing the difference between species-based results and trait-based results.

	Species Results		Trait-based Results			
Environmental variable	R^2	RMSE	R^2	RMSE	Difference in R^2	Difference in RMSE
pH North America	0.86	0.44	0.64	0.69	0.22	0.25
pH Europe	0.84	0.39	0.64	0.57	0.20	0.18
Phosphorous	0.48	0.27	0.26	0.32	0.22	0.05
Salinity	0.90	0.31	0.62	0.61	0.28	0.30
Depth	0.75	0.33	0.50	0.46	0.25	0.13

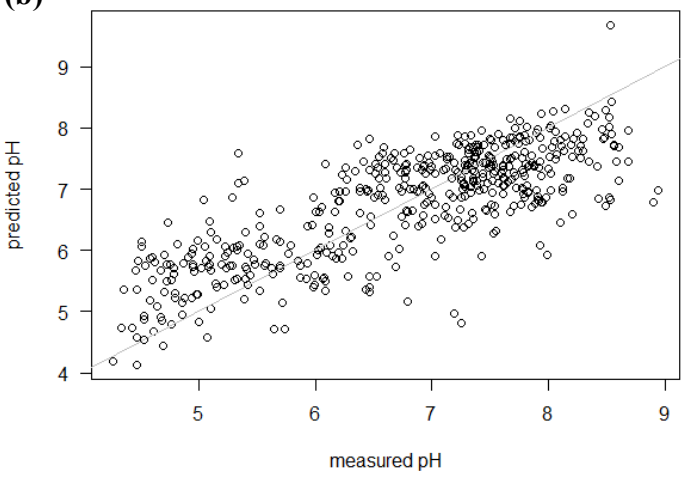
Species-based results

Trait-based results

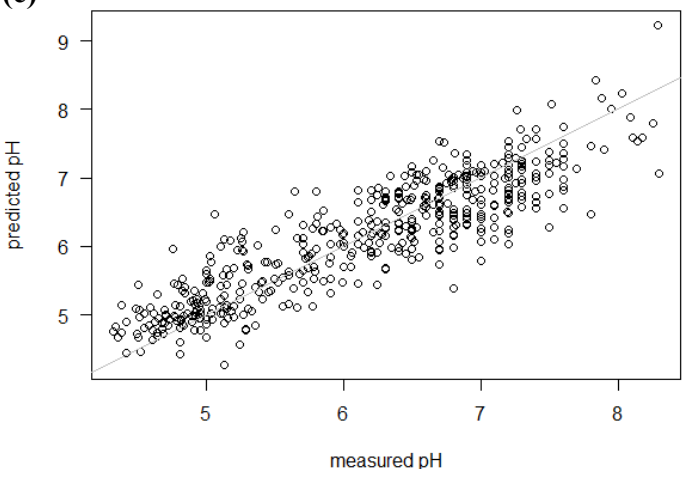
(a)



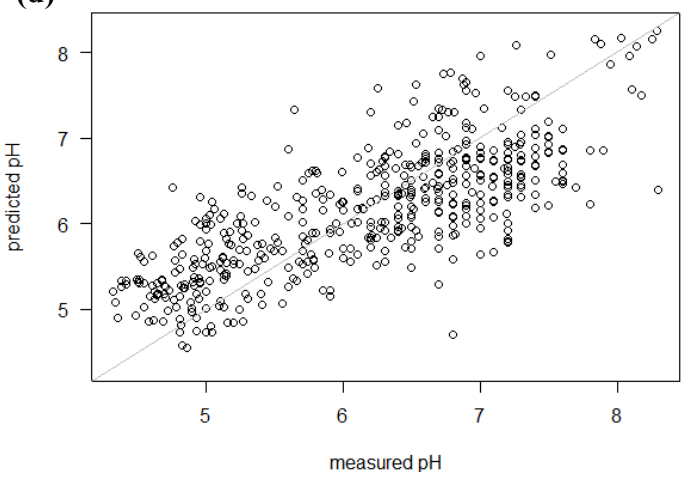
(b)



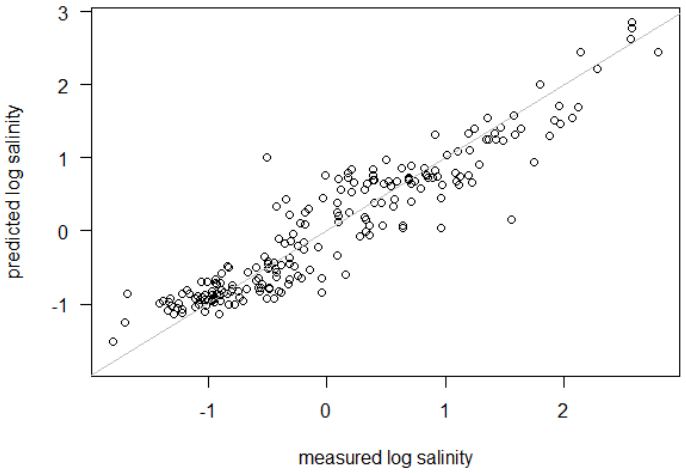
(c)



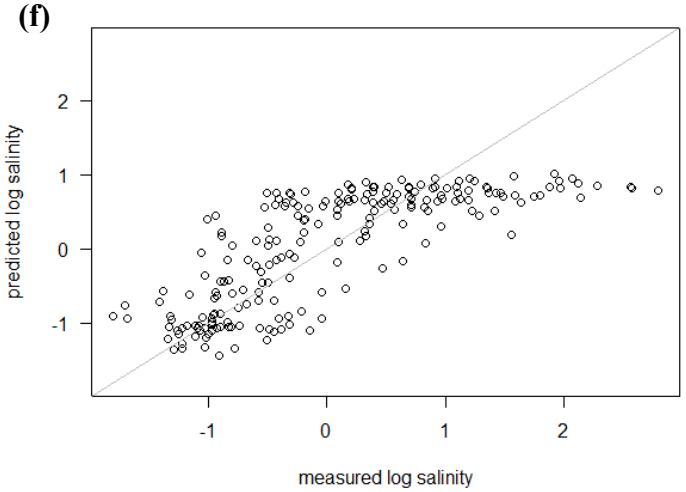
(d)



(e)



(f)



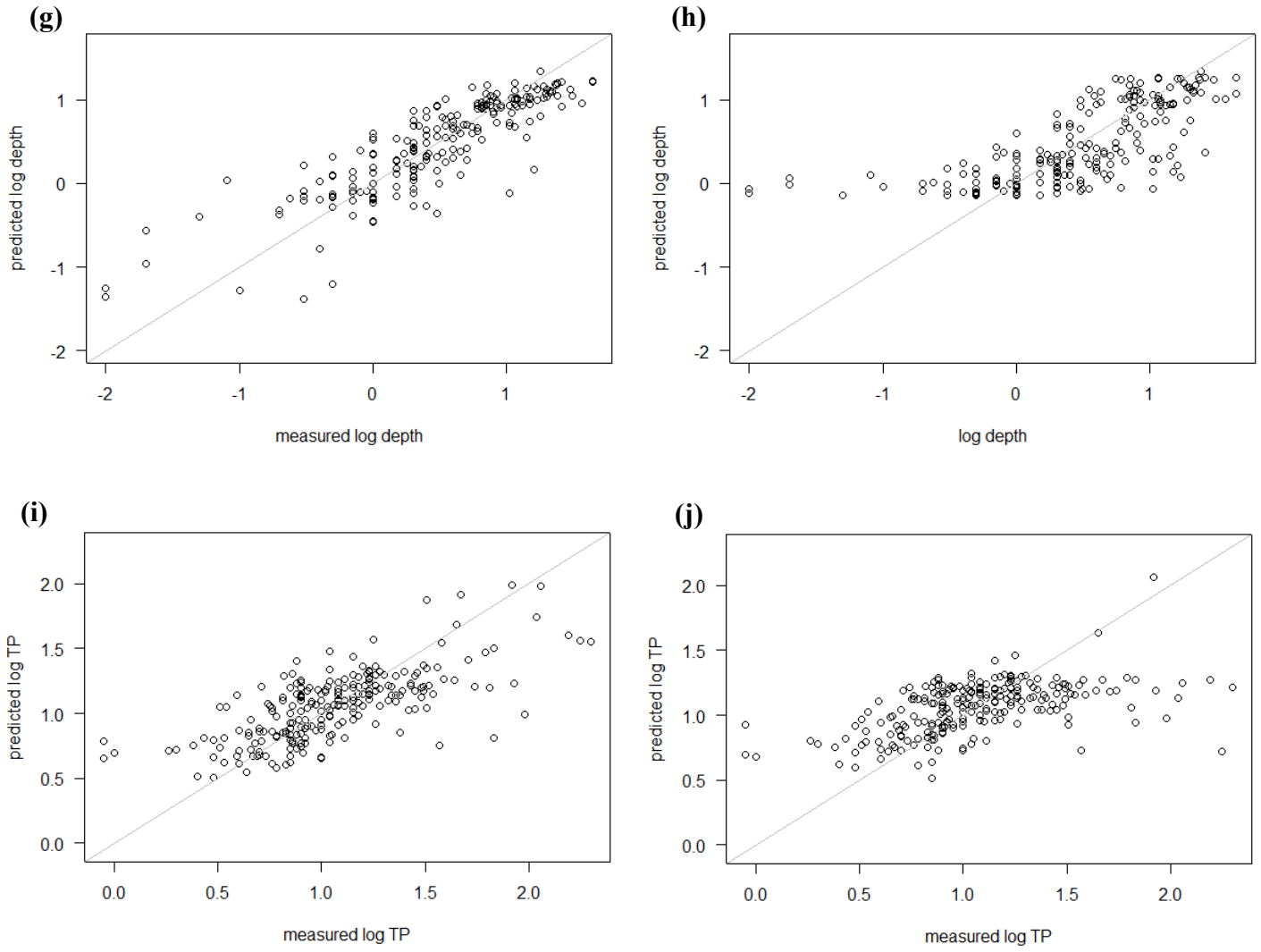


Figure 4. Predicted vs. measured environmental variables showing comparison between species-based and trait-based models for (a) North American pH species data and (b) trait data, (c) European pH species data and (d) trait data, (e) salinity species data and (f) trait data, (g) depth species data and (h) trait data, and (i) total phosphorous species data and (j) trait data. The line across each plot is the 1:1 line.

4.2 Trait removal test results

The North American pH dataset model showed 9 traits consistently appeared in the top 0.05% of tests (Figure 5). These highest performing trait included *arcuate* (trait 14), which appeared in 94% of the top 0.05% of tests, *elliptical to lanceolate* (trait 4) also performed well occurring in 93% of the top 0.05% of tests. In contrast, *centric, disk shaped: valve margins with spines* (trait 2) performed very poorly only appearing in 9% of the top tests to predict pH. The European pH dataset showed eight traits consistently appearing in the top 0.05% of tests (Figure 6). The highest performing trait for this dataset was the *very long and narrow* attribute (trait 7), which appeared in 92% of top tests. The trait *centric, disk shaped: valve margins with spines* (trait 2) once again had the poorest performance in this dataset appearing in only 1% of top tests. Other traits that performed poorly in both datasets (being defined as appearing less than 50% of the time) include the traits: *Centric, Disc-shaped- valve margins without spines; Cruciform; Clavate; Elongate and Asymmetrical; Bi-raphed- centrally located* (trait numbers 1, 6, 12, 13 and 15). Traits that performed well in both datasets (appearing > 50% of the time) include *linear with central and polar inflation, Rhombic, Bi-raphed: shifted to one side, and mono-raphed (*trait numbers 5, 9, 16, and 19).

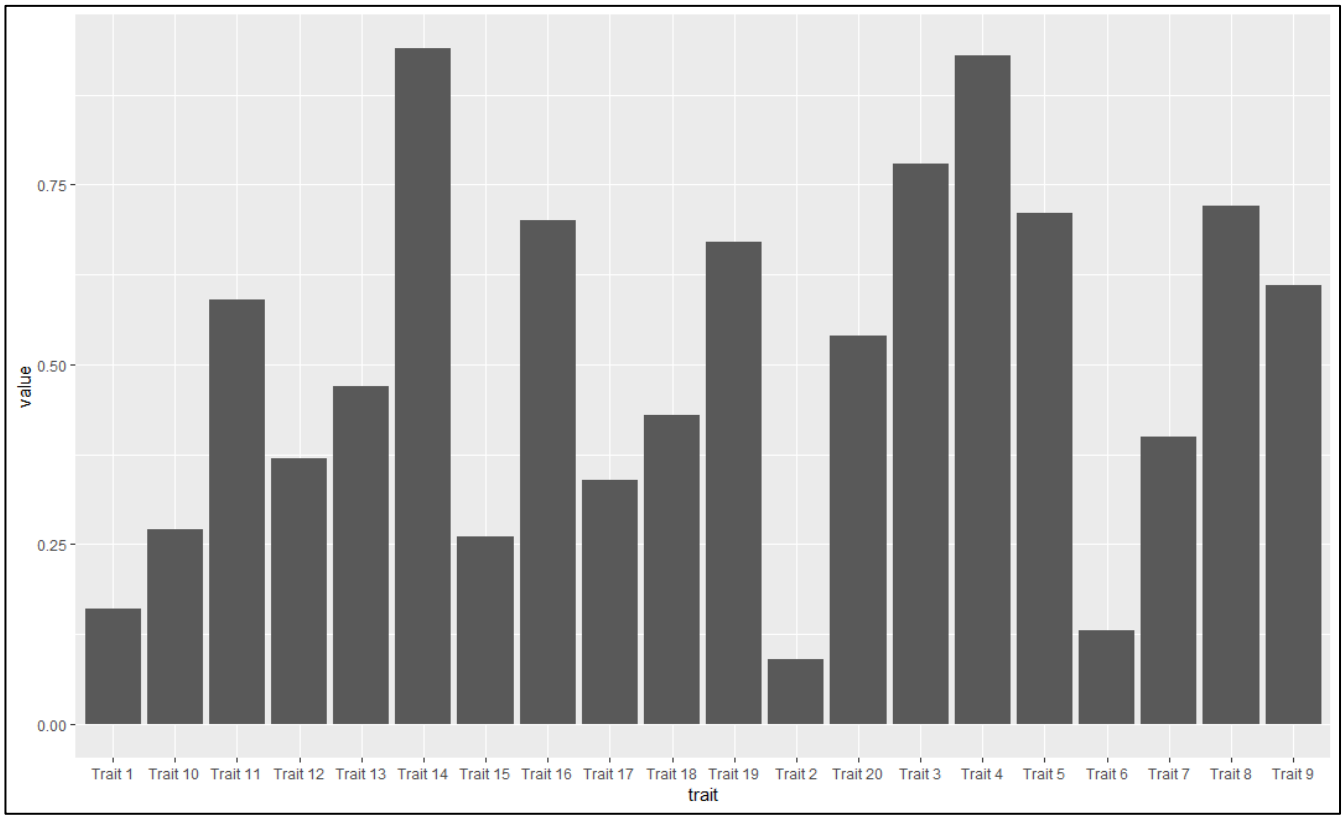


Figure 5. Frequency of traits appearing in the top 0.05% of tests for the North American pH dataset, descriptions of traits can be found in table 2.

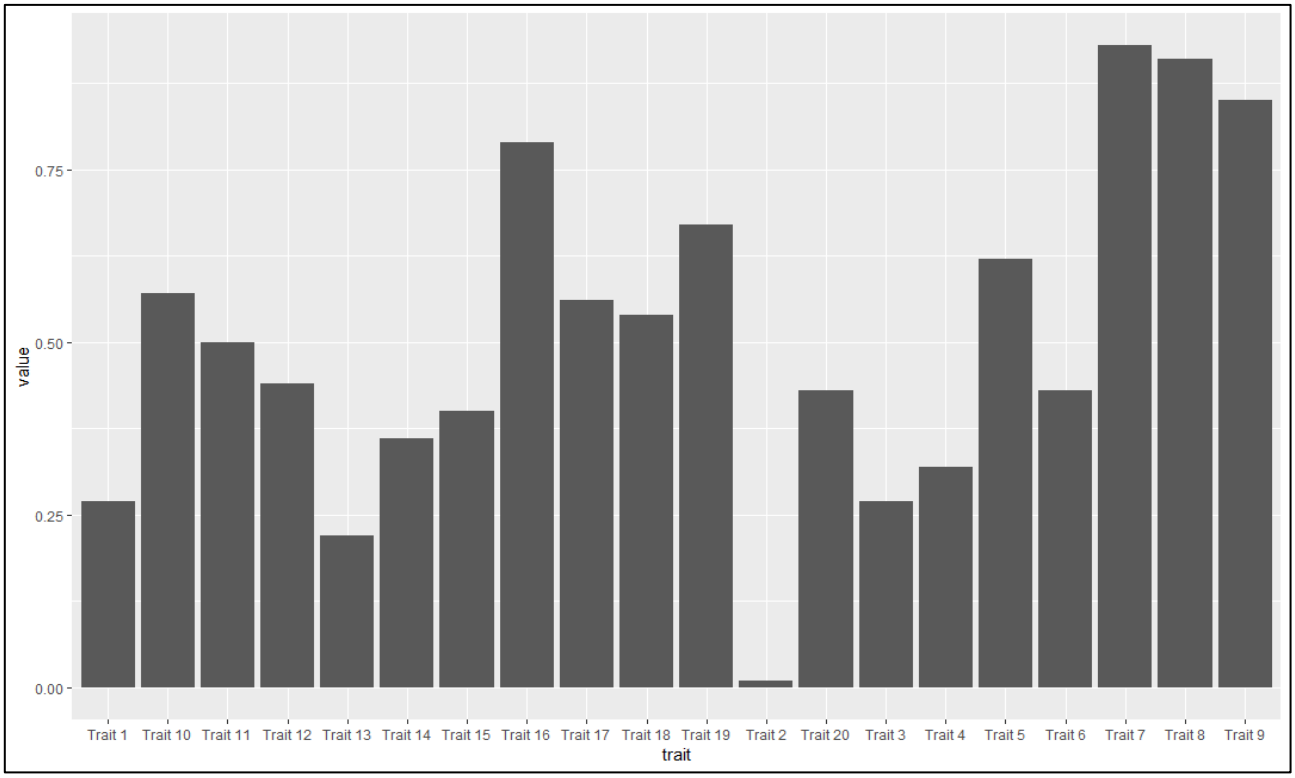


Figure 6. Frequency of traits appearing in the top 0.05% of tests for the European pH dataset, descriptions of traits can be found in table 2.

The phosphorous model indicated that 9 traits appeared in the top 0.05% of tests > 50% of the time. The highest performing traits were *disc-shaped with spines* (trait 2) and *centric-cylindrical* (trait 3), both of which appeared in 99% of the top tests. In addition to these traits many others also performed well, with *linear to linear-lanceolate* (trait 8), and *sigmoid* (trait 10) both appeared in 95% of the top tests. The poorest performing trait was *linear with central and polar inflations* (trait 5) which only appeared in 1% of top tests. The phosphorous results showed a large discrepancy among the highest performing traits and the poorest performing traits. The traits that performed well (appearing >50%

of the time) appeared in the top 0.05% at an average of 80%, while those that performed poorly appeared in an average of 30% of tests (Figure 7).

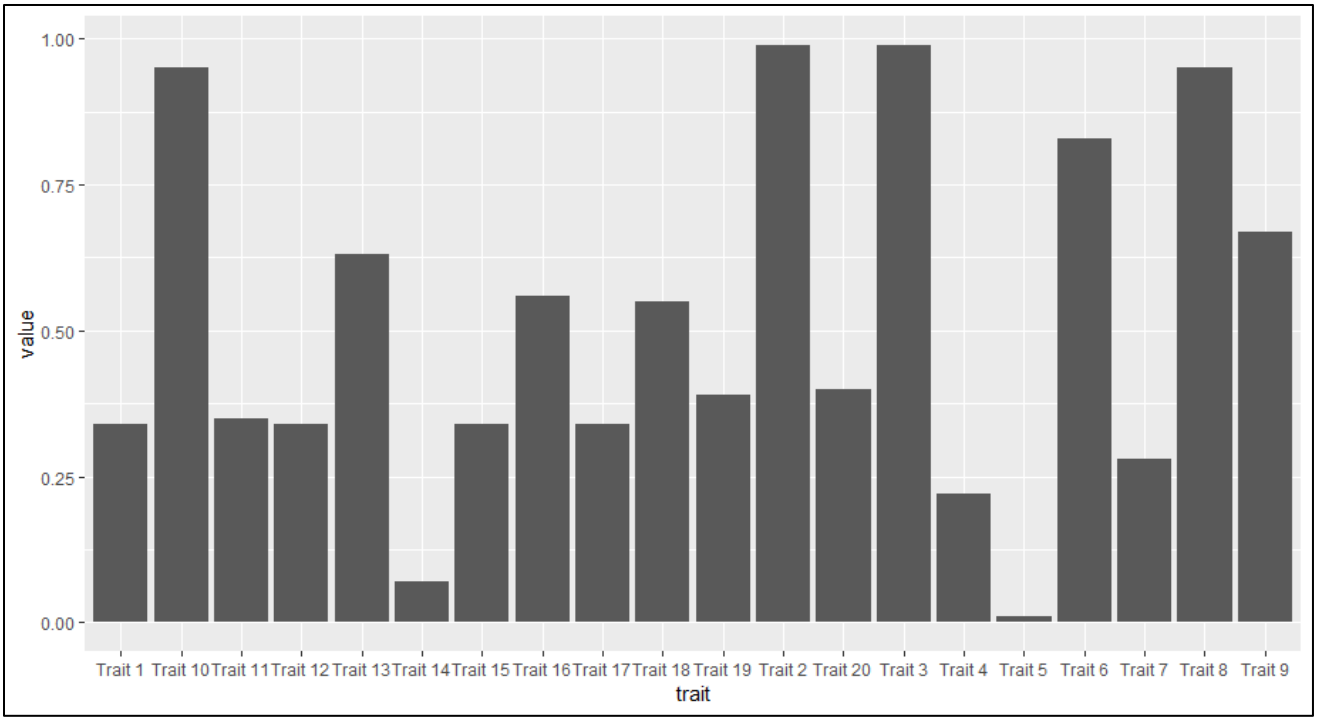


Figure 7. Frequency of traits appearing in the top 0.05% of tests for the phosphorous dataset, descriptions of traits can be found in table 2.

The salinity trait-based model showed that half the traits appeared more than 50% of the time in the top 0.05% of tests (Figure 8). The traits with the highest performance were *linear with central and polar inflations* (trait 5) and *rhombic* (trait 9), which appeared in 100% of the top tests. The traits *disc-shaped with spines* (trait 2) and *centric-cylindrical* (trait 3) did not appear in any of the top models.

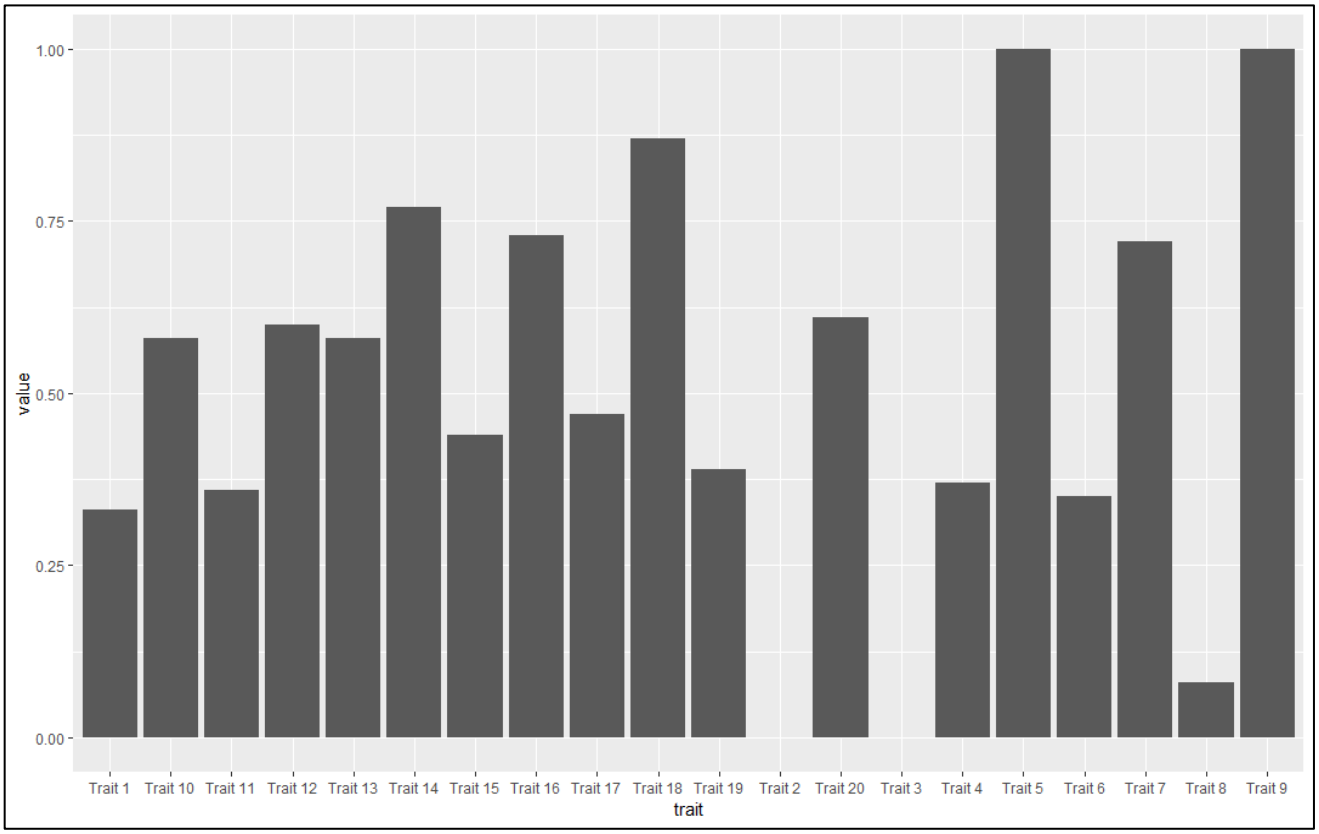


Figure 8. Frequency of traits appearing in the top 0.05% of tests for the salinity dataset, descriptions of traits can be found in table 2.

The lake depth trait based model showed that the traits *curved like a bow* (trait 14) and *disc-shaped with spines* appeared in 96% and 95% of the top performing models, respectively, while *centric-cylindrical* (trait 3) and *linear to linear-lanceolate* (trait 8) only appeared in 2% and 4% of top performing models, respectively (Figure 9).

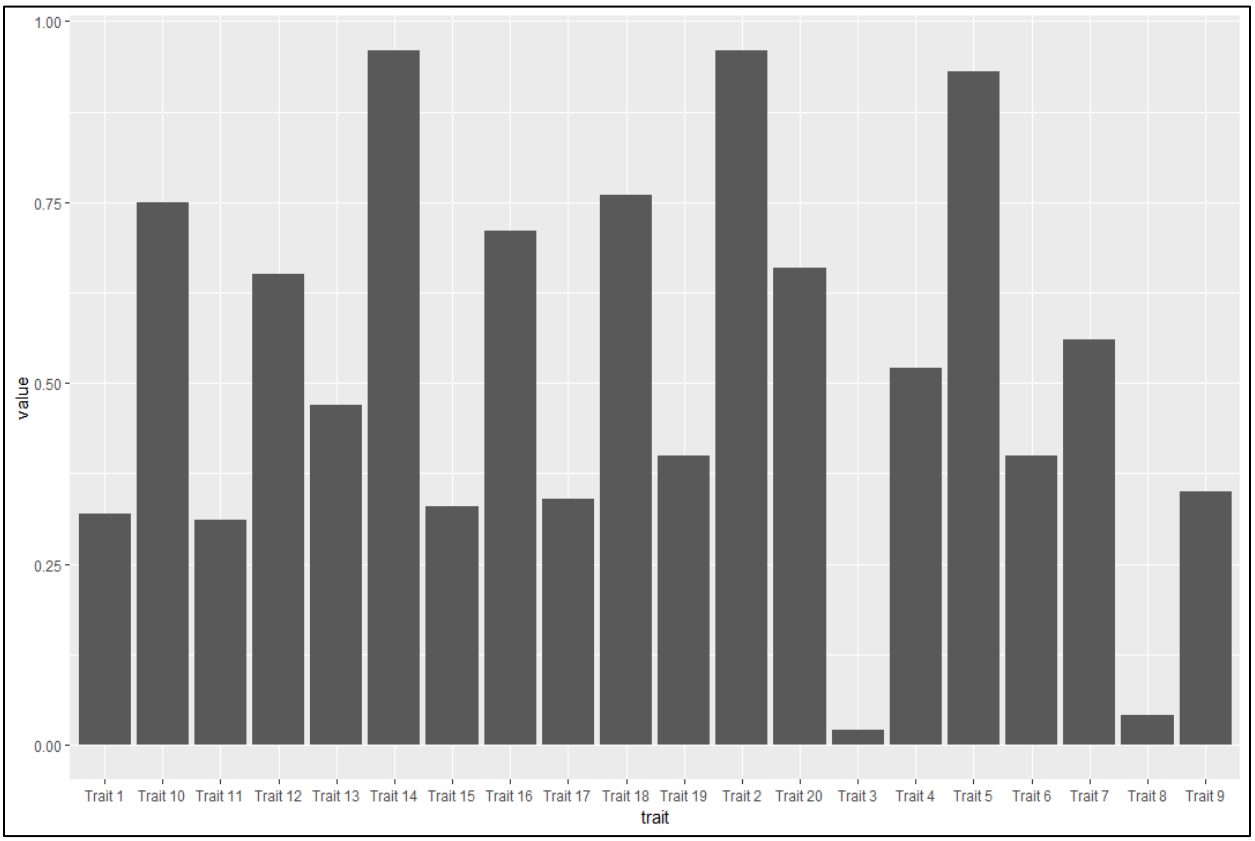


Figure 9. Frequency of traits appearing in the top 0.05% of tests for the depth dataset, descriptions of traits can be found in table 2.

Many of the same traits appeared as either among the highest or the lowest models in multiple datasets, indicating that they are likely more important for the prediction of environmental variables overall (Table 4). The traits*: centric disk-shaped: valve margins with spines; centric-cylindrical; linear with central and polar inflations; linear to linear-lanceolate and arcuate* (trait numbers 2, 3, 5, 8, and 14 in table 5) appeared more than once, suggesting they are more influential than some of traits that did not appear in any of the highest test results. Specifically, the traits *centric disk-shaped:*

valve margins without spines; *cruciform; crescent to semi-circular; clavate; elongate and asymmetrica*l; and all traits describing raphe presence and position (trait numbers 1, 6, 11, 12, 13, 15, 16, 17,18, 19 and 20 in table 2) did not appear in any of the top 0.05% of tests.

Table 4. Traits performing in the top and bottom 0.05% of tests among all environmental variables

Dataset	Top 0.05%			Bottom 0.05%		
pH NA	elliptical to lanceolate (trait 4)	arcuate (trait 14)		centric-disk shaped, valve margins with spines (trait 2)	centric, disk-shaped, valve margins without spines (trait 1)	
pH EUR	very long, narrow and skinny (trait 7)	Rhombic to rhombic-lanceolate (trait 9)				
Phosphorous	centric disk shaped – valve margins with spines (trait 2)	centric-cylindrical (trait 3)	linear to linear lanceolate (trait 8)	linear with central and polar inflations (trait 5)	arcuate (trait 14)	elliptical to lanceolate (trait 4)
Salinity	linear with central and polar inflations (trait 5)	Rhombic to rhombic-lanceolate (trait 9)		centric disk shaped – valve margins with spines (trait 2)	Centric-cylindrical (trait 3)	linear to linear lanceolate (trait 8)
Depth	centric disk shaped – valve margins with spines (trait 2)	arcuate (trait 14)		Centric-cylindrical (trait 3)	linear to linear lanceolate (trait 8)	

In order to test if the top performing traits across all environmental variables were in fact better at predicting the value of the environmental variables, a specific combination of traits was examined. The top overall performing traits and the top

performing trait for each variable created a unique barcode that was identified in tests for all environmental variables . The combination that was selected included the following ten traits: *centric disk shaped, valve margins with spines; centric-cylindrical; linear with central and polar inflations; linear to linear-lanceolate and arcuate; elliptical to lanceolate; rhombic to rhombic lanceolate; sigmoid; pro-raphed; and very long and narrow* (trait numbers 2, 3, 4, 5, 7, 8, 9, 10, 14 and 18 from Table 2). This specific combination had a higher R^2 value for four of the five environmental variables when compared to the average R^2 value that was achieved using all possible combinations of a random 10 traits (Table 5).

Table 5. Comparison of R^2 values between a specific trait combination using only influential traits, the average R^2 value using the same number of traits and the R^2 value

Environmental Variable	R^2 value of 10 most influential traits	Average R^2 value using 10 random traits	R^2 value with all 20 traits
North American pH	0.56	0.53	0.64
European pH	0.53	0.53	0.64
Phosphorous	0.27	0.20	0.26
Log Salinity	0.56	0.54	0.62
Log Depth	0.48	0.46	0.50

4.3 Trait removal trends

Determining if there was an optimum number of traits for predicting environmental variables was another important aspect of this study. Figure 11 shows the decreasing trends for all groups using the mean R^2 value for each grouping of traits

grouped by number of traits removed. It was predicted that there may be a point where a steep decrease in R^2 value was achieved, this was not the case.

Examining the differences between R^2 values revealed that the largest decreases in R^2 values happened upon the removal of >10 traits. For the North American pH dataset (Figure 10 (a)), the European dataset (Figure 10 (b)), and the salinity dataset (Figure 10 (c)) the largest decrease in R^2 value appeared between the removal of 12 and 13 traits with a difference in R^2 values of 0.028, 0.022, and 0.011, respectively. The phosphorous dataset showed the largest decrease between the removal of trait 13 and 14 at a value of 0.064, and the depth dataset between 10 and 11 at a value of 0.087.

Although it appears that there is some consistency with the largest decreases in average R^2 values occurring between the removal of 12 and 13 traits, the amount of error associated with these measures is substantial and when taking into account this variability there is no points of maximum decrease. It is also important to note that there appears to be a relatively consistent and steady decline in R^2 value (Figure 10). The phosphorous dataset showed a sharper decline in R^2 values between the removal of 13 and 14 traits, but the R^2 values were lower to begin with. It would not appear that there is any ideal number of traits overall, but certain combinations are more important and there is no huge advantage to using more than 10 traits.

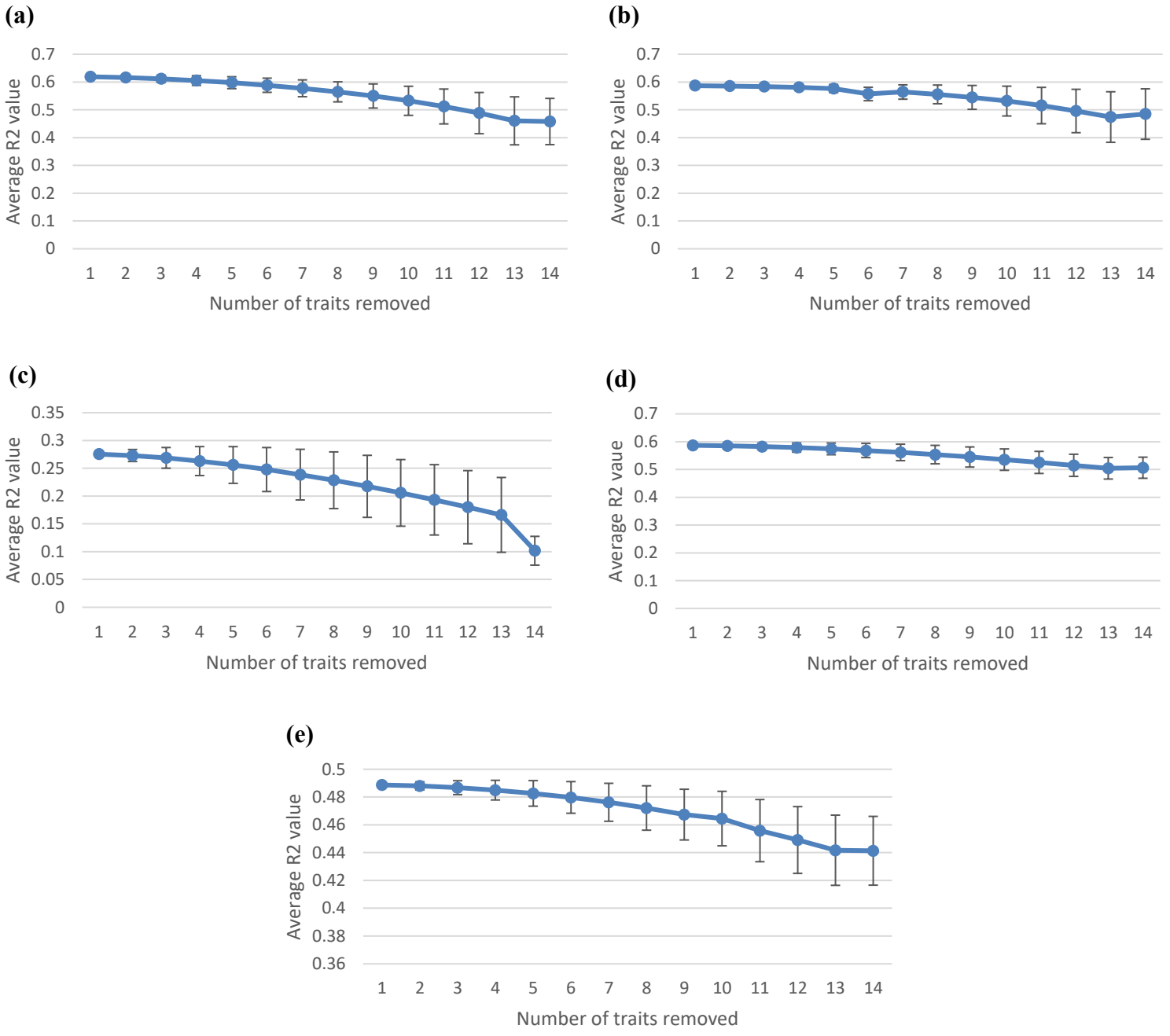


Figure 10. Change in average R^2 value as increasing numbers of traits are removed, up to a maximum of 14 for (a) North American pH dataset, (b) European pH dataset, (c) total phosphorous dataset, (d) salinity dataset and (e) depth dataset.

4.4 Correlation

Another important aspect of trait-based analysis is determining if any of the traits were highly correlated through analysis of the trait-based dataset. Since the five environmental variable datasets were extracted from the three datasets used in previous studies, there are only three individual correlation matrices. The Eastern North American and European data (Figure 11 (b) and (c)) showed high (r >+/-0.7) correlation between the traits *semi-circular* and *bi-raphed: shifted to one side* (traits 11 and 16), and between the traits *arcuate* and *pro-raphed* (traits 14 and 18), while the data from British Columbia (Figure 11 (d)) only showed strong positive correlation between the traits *crescent to semi-circular* and *bi-raphed: shifted to one side* (traits 11 and 16). A moderate negative correlation (+/- 0.3-0.6) could be observed in all three datasets between the traits *Bi-raphed: centrally located* and *A-raphed* (trait 15 and 20). Overall there was a strong negative correlation between the *semi-circular* and *bi-raphed: shifted to one side* (traits 11 and 16), and between the traits *arcuate* and *pro-raphed* (traits 14 and 18), and moderately positive correlation between the traits *bi-raphed: centrally located* and *A-raphed* (traits 15 and 20) (Figure 11 (a)). The trait *A-raphed,* was also negatively correlated with any of the traits that described raphe presence or location as well as mildly to moderately positively correlated with the traits *centric, disk shaped: valve margins without spines*, *centric, disk shaped: valve margins with spines* and *centric-cylindrical* (traits 1, 2 and 3, respectively). There was also a moderately positive correlation between the trait *elliptical to lanceolate* (trait 4) and many others. Some traits showed hardly any correlation with any of the other traits, these include the traits *centric,*

disk shaped: valve margins with spines, *linear with central and polar inflations*, *Rhombic*, *Sigmoid* and *elongate and asymmetrical* (traits 5, 6, 9, 10 and 13, respectively).

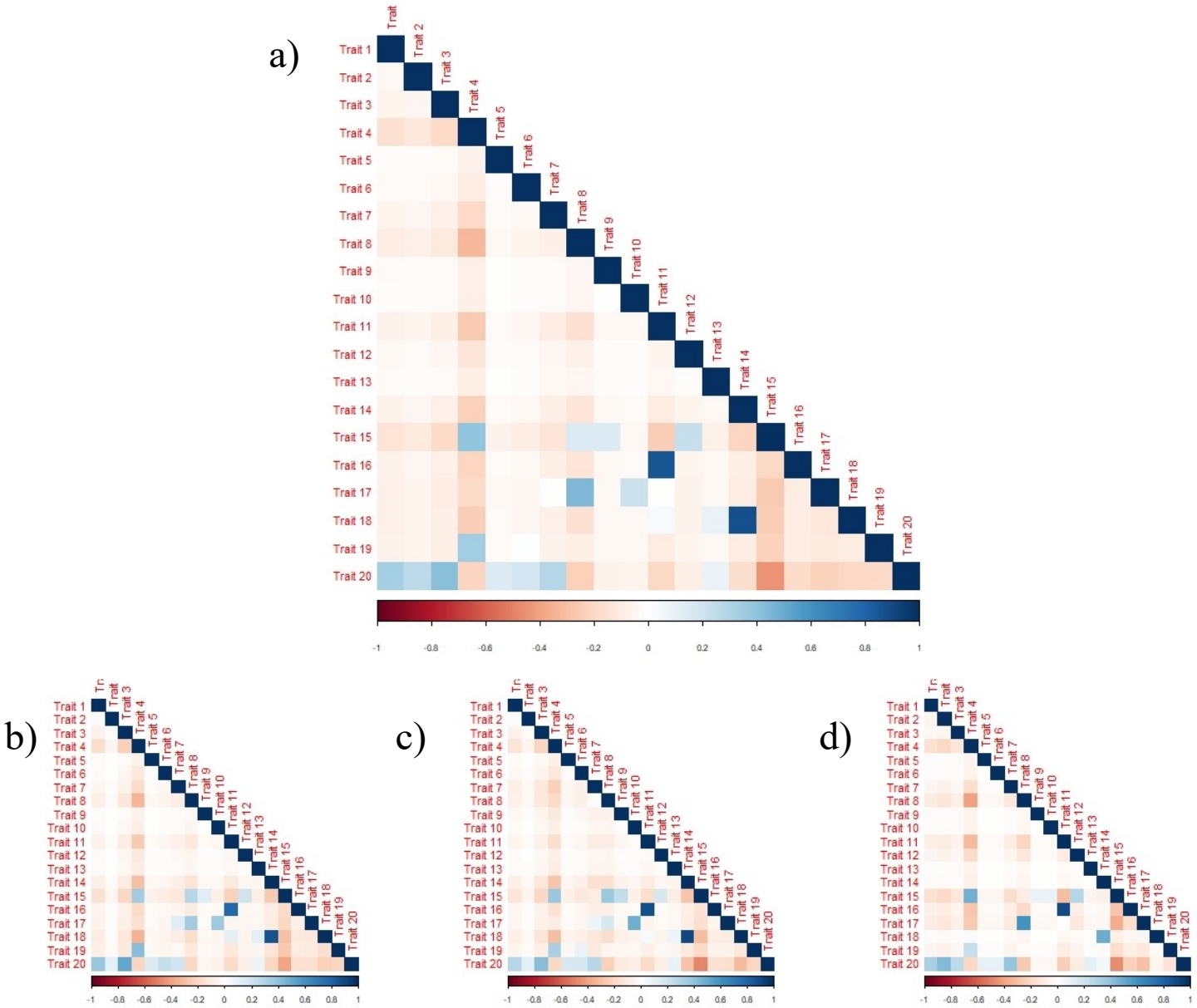


Figure 11. Trait correlation results using each location-based dataset, a) combined datasets trait correlation results, b) Correlation between traits from the eastern North American data, c) correlation between traits from the European data, d) correlation between traits of the data collected in British Columbia

Chapter 5: Discussion

This study examined the use of trait-based methods for the prediction of water quality variables using diatoms. The results of using species-based and trait-based inference models for the prediction of individual environmental variables provided valuable information into the importance individual traits play in this process. The ability to use a trait-based model allowed for faster evaluation of the state of a system, and could help to identify possible changes in a system without having to do an in depth species-based analysis, providing the researcher with the benefits of an overall decrease in the cost of the evaluation. However, the results provided less certainty surrounding the exact value of individual environmental variables as the trait-based method resulted in more error and lower R^2 values. The results indicated that an in depth look into the difference between the species-based and trait-based models, the performance of individual traits, and the optimum number of traits used for modeling could provide valuable information about the applicability of a trait-based identification technique for biomonitoring and bioassessment.

5.1 Species Based vs Trait based models

Baseline tests using the species data established the predictive ability of species within their environment using each of the four environmental variables (North American pH, European pH, total phosphorous, salinity, and lake depth). The species-based models did not perform equally well for all environmental variables. These discrepancies resulted from the variability of the selected environmental variables within a natural environment. Phosphorous was the environmental variable with the lowest performing predictive

ability based on the resulting R^2 value. Phosphorous is known to fluctuate seasonally within aquatic ecosystems and exhibits variable vertical distribution within the water column (Wetzel, 2001). This variability could contribute to the lower predictive ability of the species-based model to predict total phosphorus concentrations and ultimately resulted in lower overall R^2 and higher RMSE values for the trait based models as well.

Unlike phosphorous, the pH, salinity, and depth data do not typically show significant seasonal variability within the aquatic environment and are easier to model (Dixit & Smol, 1994 Hall & Smol, 2011; Bennett et al., 2014). The results for these environmental variables showed that salinity was the most accurately predicted by the species-based diatom models, pH (North American and European data) models were very effective and depth was reasonably accurately predicted. Diatom transfer function models are known to be excellent predictors of lake salinity and are frequently used in studies due to their species-specific tolerances to various concentrations of salts and consistent replacement along a salinity gradient. (Fritz et al., 1991; Stoermer & Smol, 1999). Diatoms are also known to be excellent predictors of pH and are frequently used in studies related to lake acidification (Stoermer & Soml, 1999). Based on the results from this study, depth was also predicted well by diatom assemblages but fewer studies rely upon diatoms to predict lake depth, as depth is not solely responsible for diatom species composition within a system. Rather, depth and lake morphometry in combination with other factors can influence species assemblage composition within a lake (Earle, 1988, Stroemer & Smol, 1999). This could possibly explain why depth is not as well-predicted by a trait-level model.

5.2 Trait-based and Trait Combination Results

The average drop in R^2 between species-based and trait-based models ranged from 0.19 (for European pH) to 0.28 (for Salinity) with an average reduction of 0.23 points across all environmental variables, which indicates that trait based model showed consistent decreases in predictive power across all environmental variables, rather than extremely large drops for some variables and not others.

The salinity dataset did not respond well to the change from species-level to trait-based groupings for analysis (also determined using different methods by Cormier et al., 2020). Salinity shows the greatest decrease in R^2 and increase RMSE values when switching from the species-based to the trait-based model. Diatom assemblages are frequently able to predict salinity to a high degree of confidence, but through the change in resolution and the aggregation of many diatom species some ecologically significant species for predicting salinity could have been masked, and may have disrupted the relationship that could have been seen in the species-based model (Cormier et al., 2020). Since the majority of the lakes sampled fell within a narrow range of salinity concentrations, with a small grouping of extreme measurements at either end, the statistical methods used, specifically deshrinking, would have moved these extreme measurements closer to the mean. In addition, some species show preferences for certain ionic salt compositions and by grouping these ecologically significant species (or species with narrow salinity tolerances) with more generalist species into the same trait combination could have decreased the overall predictive ability of the model (Cormier et al. 2020).

Both the North American and European pH datasets had similar species-level results and trait-based R^2-values, with the decrease in R^2 being lower than the average decrease of other environmental groups. In contrast, phosphorous models using trait combinations performed very poorly. Although the decrease in R^2 value was below the average decrease found among other variables, its initial R^2 value was the lowest by a large margin and the additional decrease associated with converting analyses to a trait-based method makes its predictive ability poor, especially at high values of total phosphorous (Table 3).

Depth also showed a significant decrease when switching from species based to trait-based analysis. This could potentially be attributed to the fact that prediction of depth by diatoms is not as accurate as other variables, and depth is often only one factor of many that contributes to species distribution. In addition, many species may show a wide range of depth optima, and through the weighted averaging process their optima would have been averaged out to a mean optima that may not have been representative of an individual (Cormier et al., 2020). Since this study only focused on depth, other factors contributing to variability were not measured and when switching to a broad scale analysis using trait combinations this variability only became more prevalent.

5.3 Top performing traits among all datasets

The trait removal process helped to identify the traits that had the greatest success rate in predicting the local value of the associated environmental variable. These traits were present in the top 0.05% R^2 values of trait combination tests. The traits that appeared in the top results of each of the datasets were: *centric disk shaped- valve*

margins with spines, centric-cylindrical, linear to linear-lanceolate, linear with central and polar inflations and *arcuate*. These traits are often associated with diatoms in the genera *Stephanodiscus*, *Aulacoseira*, *Nitzschia*, *Tabellaria* and *Eunotia,* respectively. Through the examination of each of these groups, some generalizations about the environment in which they are often found can be made. The family *Stephanodiscus*, represented in the *centric disc-shaped shaped* trait group is often associated with eutrophic lakes which have higher concentrations of phosphorous (Burge & Edlund, 2016; 2017; Edlund & Burge, 2016). A major genus in this family, *Cyclostephanos*, has been found in lakes with eutrophic or hypereutrophic statuses (Anderson, 2004). The genus *Aulacoseira*, represented in the *centric-cylindrical* shaped trait group, are often associated with oligotrophic-mesotrophic conditions, and can be found in lakes ranging from slightly acidic to slightly basic (Bicudo, 2016). The genus *Nitzschia*, as well as many other genera categorized as *Nitzschioid* are among the most common categories of diatoms, with autecological properties in a wide tolerance range, making this classification extremely broad and encompassing a large portion of the diatoms in each dataset (Battarbee et al., 2001; Ginn et al., 2007; Bennett et al., 2010; 2014; 2016; Cormier et al., 2020). The genus *Tabellaria*, represented in the *polar inflation* trait group, are often associated with slightly acidic to neutral lakes (Spaulding & Edlund, 2008). Some species of *Tabellaria* are sometimes associated with deeper lakes due to their ability to outcompete other species in lower light intensities (Kilinc & Sivaci, 2001). Finally, the genus *Eunotia*, represented by the trait group that are *curved or semicircular* are often epiphytic and associated with slightly higher levels of dissolved organic carbon,

and a slightly acidic to neutral pH levels and a wide range of trophic conditions (Whitmore, 1989).

Two of the traits found in the top tests of all the environmental variables, *centric disk-shaped- valve margins with spines* and *linear-linear lanceolate*, are associated with two of the most common genera of diatoms: *Stephanodiscus* (centric) and *Nitzschia* (nitzschioids). These types of diatoms can be found across a broad range of environmental conditions, making it likely that diatoms with these traits would appear in many of the top tests of each different environmental variable datasets. Since these traits are associated with a category of diatom that is extremely generalist they cannot provide specific information about their environments based solely upon their traits, but in combination with other species could increase the overall certainty when predicting an environmental variable. When evaluating the effectiveness of traits and the types of diatoms and environmental conditions that they may be associated with, it is important to consider all the dominant traits and the diatoms they may be associated with that are present. Stand-alone evaluation of only the top performing traits will not provide sufficient information to understand the state of a system.

It was clear through the individual evaluation of traits for each environmental variable group that some traits consistently appeared among the highest performing and lowest performing traits. Eight traits appeared among the highest and performing traits for each grouping of tests based on environmental variable, many of these same traits also appeared as the least influential for certain environmental variables. These traits cover a broad range of diatom species and their consistent appearance within these results indicates that they may in fact be better suited for prediction purposes than other traits.

Although some of the diatoms that are associated with these traits can show strong preference for specific values of an environmental variables, it is not necessarily these diatoms specifically that influence the outcome of the trait-based analyses. The combinations of the various diatoms with these traits may have overlapping environmental preferences and tolerances and together help to provide information about the environment in which they were sampled.

Unlike the highest performing traits, the lowest performing traits for each group did not show consistency across all environmental variables. Only three traits appeared in the bottom 25% of models across two of the five environmental variable datasets; these included *centric disk shaped- valve margins without spines, centric cylindrical,* and *linear to linear lanceolate*. Although there was less consistency among the bottom performing traits overall, these same three traits also appeared in the top models for two of the five environmental variable groups. Since some of the traits appear as both a top and bottom performing trait for different environmental variables, they would not be good candidates for permanent removal in future testing if these traits were selected for prediction of various environmental variables. Traits that did not appear to greatly influence the prediction of environmental variables include: *centric disk-shaped: valve margins without spines*; *cruciform; crescent to semi-circular; clavate; elongate and asymmetrica*l; and most traits describing raphe presence and position (trait numbers 1, 6, 11, 12, 13, 15, 16, 17, 19 and 20 in Table 2). These traits could be removed without having a large impact on predictive ability as many of them are redundant, and when removed resulted in an overall increase in predictive ability.

5.4 Trait Performance for Environemental Variables

5.4.1 pH

The highest performing traits for the North American pH dataset were *elliptical to lanceolate* and *arcuate*. Of these traits, *arcuate* has been noted as a trait that has appeared in the top 0.05% of tests of other environmental variables and can be associated with the genus *Eunotia*. The trait *elliptical to lanceolate* can be associated with the genus *Saurosirella*, and other diatom genera classified as *Fragilarioid* (long symmetrical diatoms with a pseudoraphe) (Finkelstein & Gajewski, 2008). These diatoms are frequently grouped and evaluated together, as many of them have broad tolerances and distributions. Weilhoefer and Pan (2007) found that the *Fragilarioid* type species including those in the genera *Staurosirella* and *Staurosira* could be grouped together based on their similar tolerances to TP, depth and turbidity.

The highest performing traits of the European pH dataset were traits *very long, narrow and skinny* and *rhombic*. The *very long, narrow and skinny* trait attribute can be associated with diatoms species such as *Fragilaria tenera* and genera such as *Synedra* and *Ulnaria*. *Fragilaria tenera* has commonly been found in forested lakes and are potentially associated with deeper lakes (Ruhland, Smol & Pienitz, 2003). Very broadly, some species belonging to the genera *Synedra* or *Ulnaria* can be associated with high nutrient levels and water with higher salinity concentrations (Spaulding & Edlund 2015).

Both the North American and European pH datasets had the traits *centric-disk shaped: valve margins with spines* and *centric, disk-shaped: valve margins without spines* as the lowest performing traits. In this case, centric diatoms do not show enough consistent preferences for pH levels in water to provide valuable information for

prediction purposes, or they are present so frequently it is difficult to evaluate their preferences. The centric diatoms with valve margins without spines can be associated with genera such as *Cyclotella*. The genus *Cyclotella* is one of the more widespread genera of diatoms and is often found in lakes with lower nutrient levels and overall productivity. It is thought that species within this genus are sensitive to environmental change and are often relied upon as indicators of environmental change (Saros and Anderson, 2015). The trait *centric shaped diatoms: valve margins with spines*" has already been examined with relation to common genera and is in fact one of the highest performing traits for the phosphorous, salinity and depth datasets. This trait is described as an overall well performing trait and is associated with eutrophic lakes and therefore higher phosphorous concentrations, though based on the models it along with other centric based traits do not provide valuable information about pH.

Diatoms are often used as indicators of lake pH and its potential change over time and the traits that had the most success in predicting lake pH in this study did not directly correlate with any diatom taxa that showed very specific pH tolerances (Dixit, Dixit & Smol, 1992). This indicates that it may not be any singular genus that is best at predicting pH, but rather the combination of very many different species with similar traits that are together able to give a very good idea of the pH of the environment in which they exist.

5.4.2 Total Phosphorous

The traits with the highest predictive ability for total phosphorous (TP) levels were *centric disk shaped: valve margins with spines*, *centric-cylindrical* and *linear to*

linear lanceolate, all of which appeared as overall well performing traits. These traits, as described previously, are all important indicators of lake trophic status. Trophic status is highly influenced by the nitrogen to phosphorous ratios in lakes and diatom traits that are linked with trophic status would show much greater success in the prediction of phosphorus than other traits. Phosphorous was not as clearly associated with these traits in literature and overall is not very well predicted by diatoms, and species with specific tolerances could be more commonly associated with these traits within the dataset.

The three lowest performing traits for the phosphorous datasets were *linear with central and polar inflations*, *arcuate* and *elliptical to lanceolate*. The first two traits also appeared in the overall high performing traits while *elliptical to lanceolate* also appeared as one of the highest performing traits for the North American pH.

5.4.3 Salinity

The traits with the best ability to predict salinity were *linear with central and polar inflations*, another trait found as a trait with overall high predictive ability, and *rhombic to rhombic-lanceolate* which also appeared as one of the highest predicting traits for the European pH dataset. The salinity traits with the lowest performance were *centric disk shaped: valve margins with spines*, *centric-cylindrical*, and *linear to linear lanceolate*. Important to note is that these three traits were the top three performing traits for the phosphorous dataset as well and are potentially influenced by trophic status. These traits did not relate to particular saline or non-saline environments. The relationship between diatoms and salinity relates back to the osmotic gradient between the cell and its environment and could influence nutrient update by the diatom (Fritz, 2013). Whether

this would influence the traits of diatoms in freshwater systems is unknown. Salinity was also the dataset with the highest decrease in predicted outcomes when switching from species-based to trait-based analyses, due to the loss of data at the gradient ends as a result of the statistical tests.

5.4.4 Depth

The trait-based analysis of the depth dataset resulted in traits that have previously been identified as either the highest or lowest performing traits for other environmental variables. The highest performing traits for predicting depth were *centric disk shaped: valve margins with spines* and *arcuate*, while the lowest performing traits were *centric-cylindrical* and *linear to linear lanceolate*. The highest performing traits were associated with genera of very common diatoms and were not found to have any specific relationship with depth, while the traits with the lowest predictive ability were found to be associated with trophic status. Literature on the predicting depth using diatoms is not as prominent as eutrophication, trophic status or other major influencers of ecosystem status. Depth alone is not a sole factor in predicting or understanding an aquatic ecosystem and is therefore often coupled with morphometry. Based on this, it would be difficult to predict the traits that might be more commonly found and associated with species living at specific depths.

5.5 Trait removal numbers – optimal number of traits

The results indicated that it is not necessary for the prediction of an environmental variable to use all 20 traits. There was no optimum number of traits consistent across all environmental variable groups, and for some groups there was not a clear minimum number of traits that should be used. Each environmental variable showed a steady decrease as increasing number of traits were removed.

Rather than relying on both an optimum number and type of trait, this method should rely upon the traits that consistently performed well across all environmental variables. Based on the individual trait based testing, eight traits consistently appeared in the top 0.05% of tests across all environmental variables (Table 5). As there is no optimum number of traits, these eight traits, in combination with two additional traits that were found to be influential to other variables could suffice in predicting the four environmental variables examined here.

5.6 Correlation

The trait correlation results showed that, overall, traits that were represented multiple times were negatively correlated with one another. Specifically, there were six separate traits that defined the presence or location of the raphe. This number of traits could be seen as redundant and mutually exclusive. Similarly, there were three traits describing different types of centric-shaped diatoms, though they did not show negative correlations to one another as strongly as the “raphe” traits did.

Through the removal of mutually exclusive traits, many of the intricacies of identification could be eliminated. None of the “raphe” traits appeared in the top

performing traits and could be consolidated into fewer traits. Conversely, all three traits related to centric shaped diatoms appeared as some of the most influential traits overall, indicating that this differentiation is valuable when making identification for these types of diatoms, though the presence or absence of valve margins may not be a necessary attribute. Another trait that was negatively correlated with many of the other traits was *elliptical to lanceolate*, which is described as being "oval shaped and could shape to a point at either end -could have polar inflations" (Table 2). This trait description is fairly vague and encompasses many of the other traits that describe a similar shape. For example, this trait could also encompass diatoms that are long and thin or those whose sides are quite linear or parallel. It appeared in the list of most influential traits, as did the trait *linear to linear-lanceolate*. Both of these traits likely encompass a broad range of common diatom species, making their presence within the datasets more common.

Chapter 6: Conclusions

Diatoms are known to be excellent proxies for predicting various water quality metrics and their use as indicators of water quality would be significantly more accessible if it did not require expertise for their identification. This study explored if diatom identification for the purpose of predicting environmental variables could be simplified using a trait based identification method. The largest decrease in predictive ability of the diatoms came when the data being used was transformed from species-level identification to trait-based identification with the use of 20 traits. Although there was a clear decline in predictive ability by shifting to trait-based models, the information could still be used to provide a broad understanding of a system and its changes. In order to push the trait-based method further, traits were removed sequentially and in combination and there was a slow decline in predictive ability as more traits were removed. The difference in predictive ability between the use of all 20 traits and a subset of 10 traits was minimal, with only an average decrease of 0.05 units, compared to the initial average decrease of 0.23 units that was seen when switching from a species-based to a trait-based identification. There were many traits that consistently appeared in the top results for all environmental variables indicating that a subset of traits containing only the most significant traits could be used and provide similar predictive ability to a trait-based method that relied upon more traits.

A subset of eight traits were found to be among the highest performing tests for all environmental variables: *centric disk-shaped: valve margins with spines*; *centric-cylindrical*; *linear with central and polar inflations*; *linear to linear-lanceolate* and *arcuate; elliptical to lanceolate; rhombic to rhombic lanceolate;* and *very long and*

narrow. These traits appear to be more influential when predicting all of the environmental variables. These traits alone are unlikely to provide enough certainty when evaluating the value of the environmental variables as there remain some outstanding influential traits for individual variables. A subset of the most influential traits from each of the environmental groups could also be included when using a trait-based identification method. Traits that could be added to the eight most influential traits in order to provide more certainty include: *sigmoid* (influential for phosphorous and depth); *pro-raphed* (influential for depth and salinity). These 10 traits were examined alone and proven to be more effective in predicting four of the five environmental variables than the average predictive value taken across all combinations using only 10 traits. In addition, the decrease in predictive ability between these 10 influential traits from the full 20 traits was minimal in comparison to the decrease that occurred between the species-level and trait-based analyses.

The ability to use 10 influential traits as opposed to complete species-level identification would allow for the use of diatoms as a bio-indicator to be accessible and cost effective. This aspect of the study proved to be very informative and shows that not only can trait-based identification be useful for predicting environmental variables, but there are traits that are more influential in their prediction. This brings another important question that remains unanswered: what is the interaction between specific traits and their environment? This question was broadly considered while examining each of the most influential traits for predicting each environmental variable, but there is not enough research to support this possible connection. In order to truly determine which traits may

be the most important to predicting environmental variables there must be more research into individual traits that influence diatom distribution.

This study highlighted some very interesting points surrounding the use of trait-based diatom identification for analyses for the purpose of predicting environmental variables. Though their predictive ability is not as consistent as a species-level identification, this technique in combination with other commonly used bio-indicators could provide valuable information about a system over time at a fraction of the cost of purely species-level based identifications.

References

Ambasht, N.K., and Ambasht, R, S. 2003. Modern trends in applied aquatic ecology. New York: Kluwer Academic/Plenu, Publishers

Battarbee, R., Grytnes, J. A., Thompson, R., Appleby, P., Catalán, J., Korhola, A., ... & Lami, A. (2002a). Climate variability and ecosystem dynamics at remote alpine and arctic lakes: the last 200 years.

Smol, J. P., Birks, H. J. B., & Last, W. M. (2002b). *Tracking environmental change using lake sediments, volume 3 : Terrestrial, algal, and siliceous indicators*. Retrieved from https://ebookcentral-proquest-com.proxy.library.carleton.ca

Battarbee, R. W., Juggins, S., Gasse, F., Anderson, N. J., Bennion, H., Cameron, N. G., ... & Telford, R. (2001). *European Diatom Database (EDDI): an information system for palaeoenvironmental reconstruction* (p. 94). Environmental Change Research Centre.

Battarbee, R. W. (1984). Diatom analysis an d the acidification of lakes. *Phil. Trans. R. Soc. Lond*, *305*, 451–477.

Bayer, M. M., Pullan, M. R., Mann, D. G., Juggins, S., Ciobanu, A., Santos, L., … Ludes, B. (2001). ADIAC: Using computer vision technology for automatic diatom identification. *Proceedings of the 16th International Diatom Symposium*, (January 2012), 537–562.

Bellinger, E.G. and Sigee, D.C. 2015. *Freshwater algae: identification and use as bio-indicators*. Chinchester, West Sussex, UK, NJ; Wiley-Blackwell

Bellinger, E. G., & Sigee, D. C. (2010). Introduction to freshwater algae. *Freshwater algae: Identification and use as bio-indicators*, 1-40.

Bennett, J. R., Rühland, K. M., & Smol, J. P. (2017). No magic number: determining cost-effective sample size and enumeration effort for diatom-based environmental assessment analyses. *Canadian journal of fisheries and aquatic sciences*, *74*(2), 208-215.

Bennett, J. R., Sisson, D. R., Smol, J. P., Cumming, B. F., Possingham, H. P., & Buckley, Y. M. (2014). Optimizing taxonomic resolution and sampling effort to design cost-effective ecological models for environmental assessment. *Journal of applied ecology*, *51*(6), 1722-1732.

Bennett, J. R., Cumming, B. F., Ginn, B. K., & Smol, J. P. (2010). Broad-scale environmental response and niche conservatism in lacustrine diatom communities. *Global Ecology and Biogeography*, *19*(5), 724-732.

Bennion, H., & Simpson, G. L. (2011). The use of diatom records to establish reference conditions for UK lakes subject to eutrophication. *Journal of Paleolimnology*, *45*(4), 469–488. https://doi.org/10.1007/s10933-010-9422-8

Birks, H. J. B., Lotter, A. F., Juggins, S., & Smol, J. P. (Eds.). (2012). *Tracking environmental change using lake sediments: data handling and numerical techniques* (Vol. 5). Springer Science & Business Media.

Birks, H. J. B., Juggins, S., & Line, J. M. (1990). Lake surface-water chemistry reconstructions from palaeolimnological data. *The Surface Waters Acidification Programme. Cambridge University Press, Cambridge*, 301-313.

Boeff, K. A., Strock, K. E., & Saros, J. E. (2016). Evaluating planktonic diatom response to climate change across three lakes with differing morphometry. *Journal of Paleolimnology*, *56*(1), 33–47. https://doi.org/10.1007/s10933-016-9889-z

Cain, D. J., Luoma, S. N., Carter, J. L., & Fend, S. V. (1992). Aquatic insects as bio-indicators of trace element contamination in cobble-bottom rivers and streams. *Canadian Journal of Fisheries and Aquatic Sciences*, *49*(10), 2141-2154.

Cairns, J. J., Almeida, S. P., & Fujii, J. (1982). Automated identification of diatoms. *BioScience*, *32*(2), 98–102. Retrieved from pdf

Charles, D. F., Binford, M. W., Furlong, E. T., Hites, R. A., Mitchell, M., Norton, S. A., … I, R. J. W. (1990). Paleoecological investigation of recent lake acidification in the Adirondack Mountains, N.Y. *Journal of Paleolimnology*, *3*(3), 195–241. https://doi.org/10.1007/BF00219459

Compton, J. C., 2011. Diatoms: Ecology and Life Cycle. New York: Nova Science Publishers

Cormier, E. C., Sisson, D. R., Rühland, K. M., Smol, J. P., & Bennett, J. R. (2020). A morphological trait-based approach to environmental assessment models using diatoms. *Canadian Journal of Fisheries and Aquatic Sciences*, *77*(1), 108-112.

Culverhouse, P. F., Williams, R., Benfield, M., Flood, P. R., Sell, A. F., Mazzocchi, M. G., … Sieracki, M. (2006). Automatic image analysis of plankton : future perspectives. *Marine Ecology Progress Series*, *312*, 297–309.

Davis, R. B. (1987). Paleolimnological diatom studies of acidification of lakes by acid rain: An application of quaternary science. *Quaternary Science Reviews*, *6*(2), 147–163. https://doi.org/10.1016/0277-3791(87)90031-X

Dixit, S. S., & Smol, J. P., 1994. Diatoms as indicators in the Environmental Monitoring and Assessment Program-Surface Waters (EMAP-SW). *Environmental Monitoring and Assessment*, ***31***(3), 275–307. https://doi.org/10.1007/BF00577258

Dixit, A. S., Dixit, S. S., & Smol, J. P.,1992. Long-term trends in lake water pH and metal concentrations inferred from diatoms and chrysophytes in three lakes near Sudbury, Ontario. Canadian Journal of Fisheries and Aquatic Sciences, **49**(S1),17-24.

Dixit, S. S., Smol, J. P., Kingston, J. C., & Charles, D. F. (1992). Diatoms: powerful indicators of environmental change. *Environmental science & technology*, *26*(1), 22-33.

Du Buf, H., Bayer, M., Droop, S., Head, R., Juggins, S., Fischer, S., ... & Cristóbal, G. (1999, September). Diatom identification: a double challenge called ADIAC. In *Proceedings 10th International Conference on Image Analysis and Processing* (pp. 734-739). IEEE.

Ellis, R., Simpson, R., Culverhouse, P. F., & Parisini, T. (1997). Committees , Collectives and Individuals : Expert Visual Classification by Neural Network. *Neural Comput. & Applic*, *5*, 99–105.

Embelton, K. V., Gibson, C. ., & Heaney, S. . (2003). Automated counting of phytoplankton by pattern recognition : a comparison with a manual counting method, *25*(6), 669–681.

Estep, K. W., & Macintyre, F. (1989). Counting , sizing , and identification of algae using image analysis. *Sarsia*, *4827*. https://doi.org/10.1080/00364827.1989.10413433

EC–European Commission. (2003). Common implementation strategy for the water framework directive (2000/60/EC). *Guidance Document N*, *8*.

Falasco, I. Bandino, G. 2011. The Role of environmental factors in shaping diatom frustule in Compton, J. C., 2011. Diatoms: Ecology and Life Cycle. New York: Nova Science Publishers

Falasco, E., Bona, F., Badino, G., Hoffmann, L., & Ector, L. (2009). Diatom teratological forms and environmental alterations: a review. *Hydrobiologia*, *623*(1), 1-35.

Fischer, S., Binkert, M., & Bunke, H. (2000). Feature Based Retrieval of Diatoms in an Image Database Using Decision Trees.

Ford, J., & Hasselbach, L. (2001). Heavy metals in mosses and soils on six transects along the Red Dog Mine Haul Road, Alaska. *Western Arctic National Parklands National Park Service*.

Forero, M. G., Šroubek, F., Flusser, J., Redondo, R., & Cristóbal, G. (2003). Automatic screening and multifocus fusion methods for diatom identification. *Proceedings of SPIE - The International Society for Optical Engineering*, *5200*(December 2003), 197–206. https://doi.org/10.1117/12.506841

Garnier, J., Nemery, J., Billen, G., & Théry, S. (2005). Nutrient dynamics and control of eutrophication in the Marne River system: modelling the role of exchangeable phosphorus. *Journal of Hydrology*, *304*(1-4), 397-412.

Gaston, K. J., & Neill, M. A. O. (2004). Automated species identification : why not ? *Phil. Trans. R. Soc. Lond*, *359*, 655–667. https://doi.org/10.1098/rstb.2003.1442

Gensemer, R. W. (1990). Role of aluminum and growth rate on changes in cell size and silica content of silica-limited populations of *Asterionella ralfsii* var Americana (Bacillariophyceae) 1. *Journal of phycology*, *26*(2), 250-258.

Greibach, S. (2003). *Lecture Notes in Computer Science*. https://doi.org/10.1007/3-540-45028-9

HALL, R. I., & SMOL, J. P. (1992). A weighted—averaging regression and calibration model for inferring total phosphorus concentration from diatoms in British Columbia (Canada) lakes. *Freshwater Biology*, *27*(3), 417-434.

Hasle, Grethe, R., & Fryxell, Greta, A. (1970). Diatoms: Cleaning and Mounting for Light and Electron Microscopy. *Transactions of the American Microscopical Society*, *89*(4), 469–474.

Haworth, E. Y. (1975). A scanning electron microscope study of some different frustule forms of the genus fragilaria found in scottish late-glacial sediments. *British Phycological Journal*, *10*(1), 73–80. https://doi.org/10.1080/00071617500650071

Hearn, D. J. (2009). Shape Analysis for the Automated Identification of Plants from Images of Leaves Linked references are available on JSTOR for this article : of leaves. *International Association for Plant Taxonomy*, *58*(3), 934–954.

Heino, J., & Soininen, J. (2007). Are higher taxa adequate surrogates for species-level assemblage patterns and species richness in stream organisms?. *Biological Conservation*, *137*(1), 78-89.

Heiri, O., & Lotter, A. F. (2010). How does taxonomic resolution affect chironomid-based temperature reconstruction?. *Journal of Paleolimnology*, *44*(2), 589-601.

Hering, D., Johnson, R. K., Kramm, S., Schmutz, S., Szoszkiewicz, K., & Verdonschot, P. F. (2006). Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organism response to stress. *Freshwater Biology*, *51*(9), 1757-1785.

Hicks, Y. A., Marshall, D., Rosin, P. L., Martin, R. R., Mann, D. G., & Droop, S. J. M. (2006). A model of diatom shape and texture for analysis, synthesis and identification. *Machine Vision and Applications*, *17*(5), 297–307. https://doi.org/10.1007/s00138-006-0035-1

Holt, E. A. & Miller, S. W. (2011) Bio-indicators: Using Organisms to Measure Environmental Impacts. *Nature Education Knowledge* 3(10):8

Horton, B. P., Edwards, R. J., & Lloyd, J. M. (1999). A foraminiferal-based transfer function: implications for sea-level studies. *The Journal of Foraminiferal Research*, *29*(2), 117-129.

Ishii, T., Adachi, R., Omori, M., Shimizu, U., & Trie, H. (1987). The identification , counting , and measurement of phytoplankton by an image-processing system. *J. Con. Int. Explor.*, (43), 253–260.

Martin-Jézéquel, V., Hildebrand, M., & Brzezinski, M. A. (2000). Silicon metabolism in diatoms: implications for growth. *Journal of phycology*, *36*(5), 821-840.

Juggins, S. (2013). Quantitative reconstructions in palaeolimnology: New paradigm or sick science? *Quaternary Science Reviews*, *64*, 20–32. https://doi.org/10.1016/j.quascirev.2012.12.014

Juggins, S., Lotter, A., Juggins, S., & Smol, J. (2012). Quantitative Environmental Reconstructions from Biological Data. In Tracking Environmental Change Using Lake Sediments: Data Handling and Numerical Techniques (2012th ed., Vol. 5, pp. 431–494). https://doi.org/10.1007/978-94-007-2745-8_14

Julius, Theriot, 2010. *The diatoms: a primer.* in Stoermer, E.F., and Smole, J.P., 2010.The diatoms: Applications for the environmental and earth sciences (2nd ed.). New York: Cambridge University Press.

John, J. (2003). Bioassessment of health of aquatic systems by the use of diatoms. In *Modern Trends in Applied Aquatic Ecology* (pp. 1-20). Springer, Boston, MA.

Kahlert, M., Albert, R. L., Anttila, E. L., Bengtsson, R., Bigler, C., Eskola, T., … Weckström, J. (2009). Harmonization is more important than experience-results of the first Nordic-Baltic diatom intercalibration exercise 2007 (stream monitoring). *Journal of Applied Phycology*, *21*(4), 471–482. https://doi.org/10.1007/s10811-008-9394-5

Kalff, J., & Knoechel, R. (1978). Phytoplankton and their Dynamics in Oligotrophic and Eutrophic Lakes. *Ann. Rev. Ecol. Syst.*, *9*(475–495).

Kalyoncu, H., Çiçek, N. L., Akköz, C., & Yorulmaz, B. (2009). Comparative performance of diatom indices in aquatic pollution assessment. *African Journal of Agricultural Research*, *4*(10), 1032-1040.

Karthick, B., Taylor, J. C., Mahesh, M. K., & Ramachandra, T. V. (2010). Protocols for Collection, Preservation and Enumeration of Diatoms from Aquatic Habitats for Water Quality Monitoring in India. *IUP Journal of Soil & Water Sciences*, *3*(1).

Kelly, M., & Lewis, A. (1995). Assessing the Quality of Water Quality Assessments: An Analytical Quality Control Protocol for Benthic Diatoms. *Freshwater FOrum*, *7*(1), 23–32.

Kelly, M. G., & Whitton, B. A. (1995). The trophic diatom index: a new index for monitoring eutrophication in rivers. *Journal of applied phycology*, *7*(4), 433-444.

Kirkpatrick, G., Millie, D., Moline, M., & Schofield, O. (2000). Optical discrimination of a phytoplankton species in natural mixed populations. *Limnol. Oceanogr.*, *45*(2), 476–471.

Leira, M., & Sabater, S. (2005). Diatom assemblages distribution in catalan rivers, NE Spain, in relation to chemical and physiographical factors. *Water Research*, *39*(1), 73-82.

Li, L., Zheng, B., & Liu, L. (2010). Biomonitoring and bio-indicators used for river ecosystems: Definitions, approaches and trends. *Procedia Environmental Sciences*, *2*, 1510–1524. https://doi.org/10.1016/j.proenv.2010.10.164

Mann, D. G. (1993). Patterns of sexual reproduction in diatoms. *Hydrobiologia*, *269*(1), 11-20.

Mann, D. G., & Droop, S. J. M. (1996). Biodiversity, biogeography and conservation of diatoms. In *Biogeography of freshwater algae* (pp. 19-32). Springer, Dordrecht.

Mann, D. G. (1989). The species concept in diatoms : evidence for morphologically distinct , sympatric gamodemes in four epipelic species *. *Plant Systematics and Evolution*, *164*, 215–237.

Markert, B., Wappelhorst, O., Weckert, V., Herpin, U., Siewers, U., Friese, K., & Breulmann, G. (1999). The use of bio-indicators for monitoring the heavy-metal status of the environment. *Journal of Radioanalytical and Nuclear Chemistry*, *240*(2), 425-429.

McCormick, P. V., & Cairns, J. (1994). Algae as indicators of environmental change. *Journal of Applied Phycology*, *6*(5–6), 509–526. https://doi.org/10.1007/BF02182405

Morales, E. a., Siver, P. a., & Trainor, F. R. (2001). Identification of diatoms (Bacillariophyceae) during ecological assessments: Comparison between Light Microscopy and Scanning Electron Microscopy techniques. *Proceedings of the Academy of Natural Sciences of Philadelphia*, *151*(1), 95–103. https://doi.org/10.1635/0097-3157(2001)151[0095:IODBDE]2.0.CO;2

Mosleh, M. A. A., Manssor, H., Malek, S., Milow, P., & Salleh, A. (2012). A preliminary study on automated freshwater algae recognition and classification system. *Bioinformatics*, *13*(17).

Mou, D., & Stoermer, E. F. (1992). Separating Tabellaria (Bacillariophyceae) shape groups based on fourier descriptors.

Nawrocka, A., Kandemir, İ., Fuchs, S., & Tofilski, A. (2018). Computer software for identification of honey bee subspecies and evolutionary lineages. *Apidologie*, *49*, 172–184. https://doi.org/10.1007/s13592-017-0538-y

Oertel, N., & Salánki, J. (2003). Biomonitoring and Bio-indicators in Aquatic Ecosystems, RS Ambasht, NK Ambasht (Eds.), Modern trends in applied aquatic ecology.

Parmar, T. K., Rawtani, D., & Agrawal, Y. K. (2016). Bio-indicators: the natural indicator of environmental pollution. *Frontiers in life science*, *9*(2), 110-118.

Passow, U., Alldredge, A, L., & Logan, B, E. (1993). The role of particulate carbohydrate exudates in the flocculation of diatom blooms. *Deep Sea Research*, *41*(2), 335–357.

Pech-Pachecho, J. L., Cristobal, G., Chamorror-Martinez, J., & Fernandez-Valdivia, J. (2000). Diatom autofocusing in brightfield microscopy : A comparative study. https://doi.org/10.1109/ICPR.2000.903548

Pedziszewska, A., Tylmann, W., Witak, M., Piotrowska, N., Maciejewska, E., & Latałowa, M. (2015). Holocene environmental changes reflected by pollen, diatoms, and

geochemistry of annually laminated sediments of Lake Suminko in the Kashubian Lake District (N Poland). *Review of Palaeobotany and Palynology*, *216*, 55–75. https://doi.org/10.1016/j.revpalbo.2015.01.008

Potapova, M. G., & Charles, D. F. (2002). Benthic diatoms in USA rivers: distributions along spatial and environmental gradients. *Journal of biogeography*, *29*(2), 167-187.

Reavie, E. D., Smol, J. P., Carignan, R., & Lorrain, S. (1998). Diatom paleolimnology of two fluvial lakes in the St. Lawrence River: A reconstruction of environmental changes during the last century. *Journal of Phycology*, *34*(3), 446–456. https://doi.org/10.1046/j.1529-8817.1998.340446.x

Reid, M. A., Tibby, J. C., Penny, D., & Gell, P. A. (1995). The use of diatoms to assess past and present water quality. *Australian Journal of Ecology*, *20*(1), 57-64.

Rhode, Kristina, M., Pappas, Janice, L., & Stoermer, Eugene, F. (2001). Quantitative analysis of shape variation in type and modern populations of. *Journal of Phycology*, *37*, 175–183. https://doi.org/10.1046/j.1529-8817.2001.037001175.x

Riesen, K., & Bunke, H. (2009). Approximate graph edit distance computation by means of bipartite graph matching. *Image and Vision Computing*, *27*(7), 950–959. https://doi.org/10.1016/j.imavis.2008.04.004

Roerdink, J. ADIAC: Automatic Diatom Identification and Classification www.cs.rug.nl/~roe/courses/ip/adiac-handout.pdf

Rühland, K. M., Paterson, A. M., & Smol, J. P. (2015). Lake diatom responses to warming: reviewing the evidence. *Journal of Paleolimnology*, *54*(1), 1–35. https://doi.org/10.1007/s10933-015-9837-3

Seckbach, J. and Kociolek. (2011). *The diatom world.* Vol. 19. Springer Science & Business Media.

Sims, P. A., Mann, D. G., & Medlin, L. K. (2006). Evolution of the diatoms: insights from fossil, biological and molecular data. *Phycologia*, *45*(4), 361-402.

Sládeček, V. (1986). Diatoms as indicators of organic pollution. *Acta hydrochimica et hydrobiologica*, *14*(5), 555-566.

Smol, J. P., & Stoermer, E. F. (Eds.). (2010). *The diatoms: applications for the environmental and earth sciences*. Cambridge University Press.

Smol, J. P., Wolfe, A. P., Birks, H. J. B., Douglas, M. S. V., Jones, V. J., Korhola, A., … Weckstrom, J. (2005). Climate-driven regime shifts in the biological communities of arctic lakes. *Proceedings of the National Academy of Sciences*, *102*(12), 4397–4402. https://doi.org/10.1073/pnas.0500245102

Smol, J. P. (1992). Paleolimnology: an important tool for effective ecosystem management. *Journal of Aquatic Ecosystem Health*, *1*(1), 49–58. https://doi.org/10.1007/BF00044408

Smol, J. P. (1985). The ratio of diatom frustules to chrysophycean statospores: A useful paleolimnological index. *Hydrobiologia*, *123*(3), 199–208. https://doi.org/10.1007/BF00034378

Sosik, H. M., & Olson, R. J. (2007). Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnol. Oceanogr. Methods*, 204–216.

Steinman, Alan, D., & Ladewsk, T. B. (1987). Quantitative shape analysis of Eunotia pectinalis (Bacillariophyceae) and its application to seasonal distribution patterns, *26*, 467–477.

Stevenson, R. J., & Smol, J. P. (2015). Use of algae in ecological assessments. In *Freshwater algae of North America* (pp. 921-962). Academic Press.

Stockner, J. G., & Benson, W. W. (1967). THE SUCCESSION OF DIATOM ASSEMBLAGES IN THE RECENT SEDIMENTS OF LAKE WASHINGTON 1. *Limnology and Oceanography*, *12*(3), 513-532.

Stoermer, E.F., and Smol, J.P., 2010. The diatoms: Applications for the environmental and earth sciences (2nd ed.). New York: Cambridge University Press.

Stoermer, E. F. (2001). Diatom taxonomy for paleolimnologists. *Journal of Paleolimnology*, *25*(3), 393-398.

Straile, D., Jochimsen, M. C., & Kummerlin, R. (2013). The use of long-term monitoring data for studies of planktonic diversity: A cautionary tale from two Swiss lakes. *Freshwater Biology*, *58*(6), 1292–1301. https://doi.org/10.1111/fwb.12118

Sures, B. (2001). The use of fish parasites as bio-indicators of heavy metals in aquatic ecosystems: a review. *Aquatic Ecology*, *35*(2), 245-255.

Tapolczai, K., Bouchez, A., Stenger-kovács, C., Padisák, J., & Rimet, F. (2017). Taxonomy- or trait-based ecological assessment for tropical rivers ? Case study on benthic diatoms in Mayotte island (France , Indian Ocean). *Science of the Total Environment*, *607*, 1293–1303. https://doi.org/10.1016/j.scitotenv.2017.07.093

Theriot, E., & Ladewski, Theodore, B. (1986). Morphometric Analysis of Shape of Specimens from the Neotype of Tabellaria flocculosa (Bacillariophyceae). *American Journal of Botany*, *73*(2), 224–229.

Van Dam, H., Mertens, A., & Sinkeldam, J. (1994). A coded checklist and ecological indicator values of freshwater diatoms from the Netherlands. *Netherland Journal of Aquatic Ecology*, *28*(1), 117-133.

Vannote, R. L., Minshall, G. W., Cummins, K. W., Sedell, J. R., & Cushing, C. E. (1980). The river continuum concept. *Canadian journal of fisheries and aquatic sciences*, *37*(1), 130-137.

Walker, R. F., Ishikawa, K., & Kumagai, M. (2002). Fluorescence-assisted image analysis of freshwater microalgae. *Journal of Microbiological Methods*, *51*, 149–162.

Weeks, P. J. D., & Gaston, K. J. (1997). Image analysis , neural networks , and the taxonomic impediment to biodiversity studies, *274*, 263–274.

Werner, P., Adler, S., & Dreßler, M. (2016). Effects of counting variances on water quality assessments: Implications from four benthic diatom samples, each 26 counted by 40 diatomists. *Journal of Applied Phycology*, *28*(4), 2287–2297. https://doi.org/10.1007/s10811-015-0760-9

Wetzel, C. E., Ector, L., Van De Vijver, B. V.D., Compère, P., & Mann, D. G. (2015). Morphology, typification and critical analysis of some ecologically important small naviculoid species (Bacillariophyta), *15*(2), 203–234. https://doi.org/10.5507/fot.2015.020

Weyenmeyer, G., Bleckner, T., & Pettersson, K. (1999). Changes of the plankton spring outburst related to the North Atlantic Oscillation. *Limnol. Oceanogr.*, *44*(7), 1788–1792.

Wilkinson, M. H. F., Roerdink, J., Droop, S., & Bayer, M. (2000). Diatom contour analysis using morphological curvature scale spaces Diatom Contour Analysis using Morphological Curvature Scale Spaces Stephen Droop. In *15th international confrence on pattern recognition, Vol 3, Proceedings* (Vol. 3, p. 652)).

Wilson, S. E., Cumming, B. F., & Smol, J. P. (1996). Assessing the reliability of salinity inference models from diatom assemblages: an examination of a 219-lake data set from western North America. *Canadian Journal of Fisheries and Aquatic Sciences*, *53*(7), 1580-1594.

Zapotoczny, P. (2011). Discrimination of wheat grain varieties using image analysis : morphological features. *Eur Food Res Technol*, 769–779. https://doi.org/10.1007/s00217-011-1573-y

Appendices

Appendix A

A.1 Frequency of traits appearing in the top 0.05% of tests for each environmental variable

Trait Number	Description of trait	North America pH	Europe pH	Phosphorus	Log Salinity	Log Depth
Trait 1	Centric, Disc-shaped: valve margins without spines and valve with outer zone of striae	0.16	0.27	0.34	0.33	0.31
Trait 2	Centric, Disc-shaped: valve margins with spines and valve with radial rows of punctae	0.09	0.01	0.99	0	0.95
Trait 3	Centric- Cylindrical: tubular	0.78	0.26	0.99	0	0.02
Trait 4	Elliptical to Lanceolate (oval shaped and could shape to a point at either end- could have polar inflations)	0.93	0.32	0.22	0.37	0.52
Trait 5	Linear with Central and Polar Inflations: valve length varies	0.71	0.62	0.01	1	0.93
Trait 6	Cruciform: Elliptical with central inflation	0.13	0.42	0.83	0.35	0.4
Trait 7	Very Long and Narrow/Skinny: needle to spindle-shaped, with or without polar and/or central inflations	0.4	0.92	0.28	0.71	0.56
Trait 8	Linear to Linear-Lanceolate: sides of valve are quite parallel (i.e. linear)	0.72	0.9	0.95	0.08	0.04
Trait 9	Rhombic and Rhombic-Lanceolate	0.61	0.85	0.67	1	0.35
Trait 10	Sigmoid	0.27	0.56	0.95	0.57	0.75
Trait 11	Crescent to semi-circular: Cymbelloid	0.59	0.5	0.35	0.36	0.31
Trait 12	Clavate:-Club/wedge-shaped, gomphonemoid	0.37	0.44	0.34	0.59	0.64
Trait 13	Elongate and Asymmetrical: heteropolar with inflated ends	0.47	0.22	0.63	0.58	0.47
Trait 14	Arcuate: curved like a bow	0.94	0.36	0.07	0.76	0.96
Trait 15	Bi-raphed: centrally located	0.26	0.4	0.34	0.44	0.33
Trait 16	Bi-raphed: shifted to one side	0.7	0.79	0.56	0.73	0.7
Trait 17	Bi-raphed: with siliceous struts/ribs (fibulae/transapical costae)- Keeled or Canal-bearing diatoms	0.34	0.56	0.34	0.46	0.33

Trait 18	Pro-raphed: small raphe at the pole ends	0.43	0.53	0.55	0.87	0.76
Trait 19	Mono-raphed: diatoms with a raphe on only one valve of the frustule	0.67	0.67	0.39	0.38	0.4
Trait 20	A-raphed: diatoms without a raphe on either valve	0.54	0.43	0.4	0.6	0.65

A.2 Correlation Results for Combined Datasets

	Trait 1	Trait 2	Trait 3	Trait 4	Trait 5	Trait 6	Trait 7	Trait 8	Trait 9	Trait 10	Trait 11	Trait 12	Trait 13	Trait 14	Trait 15	Trait 16	Trait 17	Trait 18	Trait 19	Trait 20
Trait 1	1.00	-0.03	-0.05	-0.16	-0.02	-0.03	-0.06	-0.09	-0.02	-0.02	-0.07	-0.04	-0.02	-0.06	-0.15	-0.06	-0.07	-0.07	-0.06	0.33
Trait 2	-0.03	1.00	-0.04	-0.12	-0.01	-0.02	-0.04	-0.07	-0.02	-0.02	-0.05	-0.03	-0.02	-0.05	-0.12	-0.05	-0.06	-0.06	-0.05	0.26
Trait 3	-0.05	-0.04	1.00	-0.20	-0.02	-0.03	-0.07	-0.12	-0.03	-0.02	-0.09	-0.05	-0.03	-0.08	-0.18	-0.08	-0.09	-0.09	-0.08	0.42
Trait 4	-0.16	-0.12	-0.20	1.00	-0.07	-0.10	-0.20	-0.33	-0.08	-0.07	-0.26	-0.14	-0.08	-0.23	0.40	-0.21	-0.18	-0.25	0.34	-0.21
Trait 5	-0.02	-0.01	-0.02	-0.07	1.00	-0.01	-0.02	-0.04	-0.01	-0.01	-0.03	-0.02	-0.01	-0.03	-0.06	-0.03	-0.03	-0.03	-0.03	0.14
Trait 6	-0.03	-0.02	-0.03	-0.10	-0.01	1.00	-0.04	-0.06	-0.01	-0.01	-0.04	-0.02	-0.01	-0.04	-0.09	-0.04	-0.05	-0.04	0.00	0.18
Trait 7	-0.06	-0.04	-0.07	-0.20	-0.02	-0.04	1.00	-0.07	-0.03	-0.03	-0.09	-0.05	-0.03	-0.08	-0.13	-0.08	0.00	-0.07	-0.07	0.28
Trait 8	-0.09	-0.07	-0.12	-0.33	-0.04	-0.06	-0.07	1.00	-0.05	-0.04	-0.15	-0.08	-0.05	-0.14	0.14	-0.13	0.45	-0.15	-0.08	-0.23
Trait 9	-0.02	-0.02	-0.03	-0.08	-0.01	-0.01	-0.03	-0.05	1.00	-0.01	-0.03	-0.02	-0.01	-0.03	0.14	-0.03	-0.04	-0.04	-0.03	-0.06
Trait 10	-0.02	-0.02	-0.02	-0.07	-0.01	-0.01	-0.03	-0.04	-0.01	1.00	-0.03	-0.02	-0.01	-0.03	-0.03	-0.03	0.22	-0.03	-0.03	-0.06
Trait 11	-0.07	-0.05	-0.09	-0.26	-0.03	-0.04	-0.09	-0.15	-0.03	-0.03	1.00	-0.06	-0.04	-0.10	-0.24	0.85	-0.01	0.03	-0.11	-0.21
Trait 12	-0.04	-0.03	-0.05	-0.14	-0.02	-0.02	-0.05	-0.08	-0.02	-0.02	-0.06	1.00	-0.02	-0.05	0.23	-0.05	-0.06	-0.06	-0.06	-0.09
Trait 13	-0.02	-0.02	-0.03	-0.08	-0.01	-0.01	-0.03	-0.05	-0.01	-0.01	-0.04	-0.02	1.00	-0.03	-0.08	-0.03	-0.04	0.10	-0.03	0.09
Trait 14	-0.06	-0.05	-0.08	-0.23	-0.03	-0.04	-0.08	-0.14	-0.03	-0.03	-0.10	-0.05	-0.03	1.00	-0.22	-0.09	-0.11	0.87	-0.09	-0.17
Trait 15	-0.15	-0.12	-0.18	0.40	-0.06	-0.09	-0.13	0.14	0.14	-0.03	-0.24	0.23	-0.08	-0.22	1.00	-0.21	-0.25	-0.24	-0.22	-0.44
Trait 16	-0.06	-0.05	-0.08	-0.21	-0.03	-0.04	-0.08	-0.13	-0.03	-0.03	0.85	-0.05	-0.03	-0.09	-0.21	1.00	-0.10	-0.10	-0.09	-0.18
Trait 17	-0.07	-0.06	-0.09	-0.18	-0.03	-0.05	0.00	0.45	-0.04	0.22	-0.01	-0.06	-0.04	-0.11	-0.25	-0.10	1.00	-0.12	-0.11	-0.22
Trait 18	-0.07	-0.06	-0.09	-0.25	-0.03	-0.04	-0.07	-0.15	-0.04	-0.03	0.03	-0.06	0.10	0.87	-0.24	-0.10	-0.12	1.00	-0.11	-0.21
Trait 19	-0.06	-0.05	-0.08	0.34	-0.03	0.00	-0.07	-0.08	-0.03	-0.03	-0.11	-0.06	-0.03	-0.09	-0.22	-0.09	-0.11	-0.11	1.00	-0.19
Trait 20	0.33	0.26	0.42	-0.21	0.14	0.18	0.28	-0.23	-0.06	-0.06	-0.21	-0.09	0.09	-0.17	-0.44	-0.18	-0.22	-0.21	-0.19	1.00

A.3 Correlation Results for North American Dataset

	Trait 1	Trait 2	Trait 3	Trait 4	Trait 5	Trait 6	Trait 7	Trait 8	Trait 9	Trait 10	Trait 11	Trait 12	Trait 13	Trait 14	Trait 15	Trait 16	Trait 17	Trait 18	Trait 19	Trait 20
Trait 1	1.00	-0.02	-0.06	-0.14	-0.02	-0.04	-0.08	-0.08	-0.03	-0.02	-0.05	-0.02	-0.03	-0.08	-0.14	-0.04	-0.05	-0.09	-0.06	0.28
Trait 2	-0.02	1.00	-0.02	-0.06	-0.01	-0.02	-0.03	-0.03	-0.01	-0.01	-0.02	-0.01	-0.01	-0.03	-0.06	-0.02	-0.02	-0.04	-0.02	0.11
Trait 3	-0.06	-0.02	1.00	-0.21	-0.04	-0.06	-0.12	-0.12	-0.05	-0.04	-0.08	-0.04	-0.04	-0.12	-0.21	-0.07	-0.07	-0.14	-0.09	0.42
Trait 4	-0.14	-0.06	-0.21	1.00	-0.08	-0.13	-0.24	-0.19	-0.11	-0.08	-0.17	-0.08	-0.10	-0.28	0.38	-0.15	-0.10	-0.31	0.32	-0.20
Trait 5	-0.02	-0.01	-0.04	-0.08	1.00	-0.02	-0.05	-0.04	-0.02	-0.01	-0.03	-0.01	-0.02	-0.05	-0.08	-0.03	-0.03	-0.05	-0.03	0.16
Trait 6	-0.04	-0.02	-0.06	-0.13	-0.02	1.00	-0.07	-0.07	-0.03	-0.02	-0.05	-0.02	-0.03	-0.07	-0.13	-0.04	-0.04	-0.08	0.08	0.18
Trait 7	-0.08	-0.03	-0.12	-0.24	-0.05	-0.07	1.00	0.01	-0.07	-0.05	-0.10	-0.05	-0.06	-0.16	-0.11	-0.09	0.08	-0.13	-0.12	0.27
Trait 8	-0.08	-0.03	-0.12	-0.19	-0.04	-0.07	0.01	1.00	-0.06	-0.04	-0.10	-0.04	-0.06	-0.16	0.36	-0.09	0.17	-0.17	-0.04	-0.24
Trait 9	-0.03	-0.01	-0.05	-0.11	-0.02	-0.03	-0.07	-0.06	1.00	-0.02	-0.04	-0.02	-0.02	-0.07	0.24	-0.04	-0.04	-0.07	-0.05	-0.12
Trait 10	-0.02	-0.01	-0.04	-0.08	-0.01	-0.02	-0.05	-0.04	-0.02	1.00	-0.03	-0.01	-0.02	-0.05	-0.08	-0.03	0.49	-0.05	-0.03	-0.08
Trait 11	-0.05	-0.02	-0.08	-0.17	-0.03	-0.05	-0.10	-0.10	-0.04	-0.03	1.00	-0.03	-0.04	-0.10	-0.17	0.88	-0.06	0.04	-0.07	-0.18
Trait 12	-0.02	-0.01	-0.04	-0.08	-0.01	-0.02	-0.05	-0.04	-0.02	-0.01	-0.03	1.00	-0.02	-0.05	0.17	-0.03	-0.03	-0.05	-0.03	-0.08
Trait 13	-0.03	-0.01	-0.04	-0.10	-0.02	-0.03	-0.06	-0.06	-0.02	-0.02	-0.04	-0.02	1.00	-0.06	-0.10	-0.03	-0.03	0.06	-0.04	0.10
Trait 14	-0.08	-0.03	-0.12	-0.28	-0.05	-0.07	-0.16	-0.16	-0.07	-0.05	-0.10	-0.05	-0.06	1.00	-0.27	-0.09	-0.09	0.90	-0.12	-0.29
Trait 15	-0.14	-0.06	-0.21	0.38	-0.08	-0.13	-0.11	0.36	0.24	-0.08	-0.17	0.17	-0.10	-0.27	1.00	-0.15	-0.16	-0.30	-0.20	-0.49
Trait 16	-0.04	-0.02	-0.07	-0.15	-0.03	-0.04	-0.09	-0.09	-0.04	-0.03	0.88	-0.03	-0.03	-0.09	-0.15	1.00	-0.05	-0.10	-0.06	-0.16
Trait 17	-0.05	-0.02	-0.07	-0.10	-0.03	-0.04	0.08	0.17	-0.04	0.49	-0.06	-0.03	-0.03	-0.09	-0.16	-0.05	1.00	-0.10	-0.07	-0.17

Trait 18	-0.09	-0.04	-0.14	-0.31	-0.05	-0.08	-0.13	-0.17	-0.07	-0.05	0.04	-0.05	0.06	0.90	-0.30	-0.10	-0.10	1.00	-0.13	-0.32
Trait 19	-0.06	-0.02	-0.09	0.32	-0.03	0.08	-0.12	-0.04	-0.05	-0.03	-0.07	-0.03	-0.04	-0.12	-0.20	-0.06	-0.07	-0.13	1.00	-0.21
Trait 20	0.28	0.11	0.42	-0.20	0.16	0.18	0.27	-0.24	-0.12	-0.08	-0.18	-0.08	0.10	-0.29	-0.49	-0.16	-0.17	-0.32	-0.21	1.00

A.4 Correlation Results for European Datasets

	Trait 1	Trait 2	Trait 3	Trait 4	Trait 5	Trait 6	Trait 7	Trait 8	Trait 9	Trait 10	Trait 11	Trait 12	Trait 13	Trait 14	Trait 15	Trait 16	Trait 17	Trait 18	Trait 19	Trait 20
Trait 1	1.00	-0.01	-0.06	-0.17	-0.02	-0.03	0.05	-0.09	-0.02	-0.02	-0.07	-0.03	-0.03	-0.08	-0.15	-0.06	-0.06	-0.09	-0.08	0.38
Trait 2	-0.01	1.00	-0.02	-0.05	-0.01	-0.01	-0.01	-0.02	-0.01	-0.01	-0.02	-0.01	-0.01	-0.02	-0.04	-0.02	-0.02	-0.03	-0.02	0.11
Trait 3	-0.06	-0.02	1.00	-0.24	-0.03	-0.04	-0.07	-0.12	-0.03	-0.03	-0.09	-0.04	-0.04	-0.11	-0.21	-0.08	-0.09	-0.12	-0.11	0.52
Trait 4	-0.17	-0.05	-0.24	1.00	-0.07	-0.11	-0.18	-0.33	-0.07	-0.09	-0.26	-0.11	-0.10	-0.29	0.34	-0.18	-0.16	-0.34	0.43	-0.22
Trait 5	-0.02	-0.01	-0.03	-0.07	1.00	-0.01	-0.02	-0.04	-0.01	-0.01	-0.03	-0.01	-0.01	-0.03	-0.06	-0.02	-0.03	-0.04	-0.03	0.15
Trait 6	-0.03	-0.01	-0.04	-0.11	-0.01	1.00	-0.03	-0.06	-0.01	-0.01	-0.04	-0.02	-0.02	-0.05	-0.10	-0.04	-0.04	-0.06	-0.05	0.24
Trait 7	-0.05	-0.01	0.07	0.18	-0.02	-0.03	1.00	-0.09	-0.02	-0.02	-0.07	-0.03	-0.03	-0.08	-0.12	-0.06	0.12	-0.09	-0.03	0.19
Trait 8	-0.09	-0.02	-0.12	-0.33	-0.04	-0.06	-0.09	1.00	-0.04	-0.04	-0.13	-0.06	-0.05	-0.15	0.27	-0.11	0.34	-0.17	-0.13	-0.21
Trait 9	-0.02	-0.01	-0.03	-0.07	-0.01	-0.01	-0.02	-0.04	1.00	-0.01	-0.03	-0.01	-0.01	-0.03	0.12	-0.02	-0.03	-0.04	-0.03	-0.05
Trait 10	-0.02	-0.01	-0.03	-0.09	-0.01	-0.01	-0.02	-0.04	-0.01	1.00	-0.03	-0.01	-0.01	-0.04	-0.07	-0.03	0.36	-0.04	-0.04	-0.06
Trait 11	-0.07	-0.02	-0.09	-0.26	-0.03	-0.04	-0.07	-0.13	-0.03	-0.03	1.00	-0.04	-0.04	-0.12	-0.22	0.79	-0.05	0.12	-0.12	-0.18

Trait 12	-0.03	-0.01	-0.04	-0.11	-0.01	-0.02	-0.03	-0.06	-0.01	-0.01	-0.04	1.00	-0.02	-0.05	0.19	-0.04	-0.04	-0.06	-0.05	-0.08
Trait 13	-0.03	-0.01	-0.04	-0.10	-0.01	-0.02	-0.03	-0.05	-0.01	-0.01	-0.04	-0.02	1.00	-0.04	-0.09	-0.03	-0.04	0.12	-0.05	0.07
Trait 14	-0.08	-0.02	-0.11	-0.29	-0.03	-0.05	-0.08	-0.15	-0.03	-0.04	-0.12	-0.05	-0.04	1.00	-0.26	-0.10	-0.11	0.87	-0.14	-0.21
Trait 15	-0.15	-0.04	-0.21	0.34	-0.06	-0.10	-0.12	0.27	0.12	-0.07	-0.22	0.19	-0.09	-0.26	1.00	-0.19	-0.21	-0.29	-0.27	-0.40
Trait 16	-0.06	-0.02	-0.08	-0.18	-0.02	-0.04	-0.06	-0.11	-0.02	-0.03	0.79	-0.04	-0.03	-0.10	-0.19	1.00	-0.08	-0.11	-0.10	-0.15
Trait 17	-0.06	-0.02	-0.09	-0.16	-0.03	-0.04	0.12	0.34	-0.03	0.36	-0.05	-0.04	-0.04	-0.11	-0.21	-0.08	1.00	-0.12	-0.11	-0.17
Trait 18	-0.09	-0.03	-0.12	-0.34	-0.04	-0.06	-0.09	-0.17	-0.04	-0.04	0.12	-0.06	0.12	0.87	-0.29	-0.11	-0.12	1.00	-0.16	-0.24
Trait 19	-0.08	-0.02	-0.11	0.43	-0.03	-0.05	-0.03	-0.13	-0.03	-0.04	-0.12	-0.05	-0.05	-0.14	-0.27	-0.10	-0.11	-0.16	1.00	-0.22
Trait 20	0.38	0.11	0.52	-0.22	0.15	0.24	0.19	-0.21	-0.05	-0.06	-0.18	-0.08	0.07	-0.21	-0.40	-0.15	-0.17	-0.24	-0.22	1.00

A.5 Correlation Results for British Columbian Datasets

	Trait 1	Trait 2	Trait 3	Trait 4	Trait 5	Trait 6	Trait 7	Trait 8	Trait 9	Trait 10	Trait 11	Trait 12	Trait 13	Trait 14	Trait 15	Trait 16	Trait 17	Trait 18	Trait 19	Trait 20
Trait 1	1.00	-0.05	-0.03	-0.15	-0.01	-0.01	-0.05	-0.11	-0.01	-0.01	-0.08	-0.05	-0.01	-0.02	-0.15	-0.07	-0.10	-0.02	-0.05	0.33
Trait 2	-0.05	1.00	-0.04	-0.20	-0.02	-0.02	-0.06	-0.14	-0.02	-0.02	-0.11	-0.06	-0.02	-0.02	-0.20	-0.09	-0.13	-0.02	-0.06	0.43
Trait 3	-0.03	-0.04	1.00	-0.12	-0.01	-0.01	-0.04	-0.09	-0.01	-0.01	-0.07	-0.04	-0.01	-0.02	-0.12	-0.06	-0.08	-0.02	-0.04	0.27
Trait 4	-0.15	-0.20	-0.12	1.00	-0.05	-0.05	-0.18	-0.40	-0.05	-0.05	-0.31	-0.19	-0.05	-0.07	0.47	-0.28	-0.26	0.03	0.22	-0.21
Trait 5	-0.01	-0.02	-0.01	-0.05	1.00	0.00	-0.02	-0.03	0.00	0.00	-0.03	-0.02	0.00	-0.01	-0.05	-0.02	-0.03	-0.01	-0.02	0.11

Trait 6	-0.01	-0.02	-0.01	-0.05	0.00	1.00	-0.02	-0.03	0.00	0.00	-0.03	-0.02	0.00	-0.01	-0.05	-0.02	-0.03	-0.01	-0.02	0.11
Trait 7	-0.05	-0.06	-0.04	-0.18	-0.02	-0.02	1.00	-0.12	-0.02	-0.02	-0.09	-0.06	-0.02	-0.02	-0.18	-0.08	-0.11	-0.02	-0.05	0.38
Trait 8	-0.11	-0.14	-0.09	-0.40	-0.03	-0.03	-0.12	1.00	-0.03	-0.03	-0.22	-0.13	-0.03	-0.05	-0.10	-0.19	0.59	-0.05	-0.03	-0.25
Trait 9	-0.01	-0.02	-0.01	-0.05	0.00	0.00	-0.02	-0.03	1.00	0.00	-0.03	-0.02	0.00	-0.01	0.09	-0.02	-0.03	-0.01	-0.02	-0.04
Trait 10	-0.01	-0.02	-0.01	-0.05	0.00	0.00	-0.02	-0.03	0.00	1.00	-0.03	-0.02	0.00	-0.01	0.09	-0.02	-0.03	-0.01	-0.02	-0.04
Trait 11	-0.08	-0.11	-0.07	-0.31	-0.03	-0.03	-0.09	-0.22	-0.03	-0.03	1.00	-0.10	-0.03	-0.04	-0.31	0.89	-0.01	-0.04	-0.09	-0.25
Trait 12	-0.05	-0.06	-0.04	-0.19	-0.02	-0.02	-0.06	-0.13	-0.02	-0.02	-0.10	1.00	-0.02	-0.02	0.28	-0.09	-0.12	-0.02	-0.06	-0.11
Trait 13	-0.01	-0.02	-0.01	-0.05	0.00	0.00	-0.02	-0.03	0.00	0.00	-0.03	-0.02	1.00	-0.01	-0.05	-0.02	-0.03	-0.01	-0.02	0.11
Trait 14	-0.02	-0.02	-0.02	-0.07	-0.01	-0.01	-0.02	-0.05	-0.01	-0.01	-0.04	-0.02	-0.01	1.00	-0.07	-0.03	-0.04	0.50	-0.02	0.05
Trait 15	-0.15	-0.20	-0.12	0.47	-0.05	-0.05	-0.18	-0.10	0.09	0.09	-0.31	0.28	-0.05	-0.07	1.00	-0.27	-0.37	-0.07	-0.18	-0.46
Trait 16	-0.07	-0.09	-0.06	-0.28	-0.02	-0.02	-0.08	-0.19	-0.02	-0.02	0.89	-0.09	-0.02	-0.03	-0.27	1.00	-0.17	-0.03	-0.08	-0.22
Trait 17	-0.10	-0.13	-0.08	-0.26	-0.03	-0.03	-0.11	0.59	-0.03	-0.03	-0.01	-0.12	-0.03	-0.04	-0.37	-0.17	1.00	-0.04	-0.11	-0.29
Trait 18	-0.02	-0.02	-0.02	0.03	-0.01	-0.01	-0.02	-0.05	-0.01	-0.01	-0.04	-0.02	-0.01	0.50	-0.07	-0.03	-0.04	1.00	-0.02	-0.06
Trait 19	-0.05	-0.06	-0.04	0.22	-0.02	-0.02	-0.05	-0.03	-0.02	-0.02	-0.09	-0.06	-0.02	-0.02	-0.18	-0.08	-0.11	-0.02	1.00	-0.14
Trait 20	0.33	0.43	0.27	-0.21	0.11	0.11	0.38	-0.25	-0.04	-0.04	-0.25	-0.11	0.11	0.05	-0.46	-0.22	-0.29	-0.06	-0.14	1.00

Appendix B

B.1 R Code used to analyze the data

```
#####Functions #######
barcoder = function(traitdf) {
  #function takes a copy of the trait df with:
  # row 1 being the species name, row 2:23 being SpeciesByTrait
  # that are either 1, 0 or _ for NA
  # outputs a barcode which represents a group
  n_sp = nrow(traitdf)
  code = vector(n_sp,mode='character')
  for (i in 1:n_sp) {
    code[i] = paste0(traitdf[i,2:ncol(SpeciesByTrait)], collapse = "")
  }
  return(code)
}

traitremover = function(traitdf, SpeciesByTrait2delete) {
  #function removes SpeciesByTrait from a copy of a trait dataframe based on the
  #column numbers in the vector SpeciesByTrait2delete
  # remember col 1 is the species name, so don't delete that
  # it passes back a trait df with deleted columns as "_"

  traitdf[,SpeciesByTrait2delete] = '_'
  return(traitdf)
}

matchNsum = function(species_df, barcode2spp) {
  # this takes 2 dataframes, matches the species names to barcodes and returns summed %
abundances, names the columns after the barcode

  #spp_df - a dataframe with all the original species and their abundance aka species_df
  #barcode2spp - a dataframe with col 2 as barcodes, col 1 as spp names aka species_code

  uniquecodes=unique(barcode2spp[,2])
  n_code = length(uniquecodes)
  sum_ab = as.data.frame(matrix(data=NA,nrow=nrow(spp_df),
ncol=ncol(barcode2spp)))  #the values stored in here arent changing each time around -
they should be  #dataframe

  for (i in 1:n_code) { #something in here isnt working
```

```
   tmpsp = as.character(barcode2spp[which( barcode2spp[,2] == uniquecodes[i]),1]) #this
doesnt always work
   barcodetemp=spp_df[,na.omit(match(tmpsp, names(species_df)))]
   if(is.null(ncol(barcodetemp))) {
     sum_ab[,i] = barcodetemp} else
     { sum_ab[,i] = rowSums(barcodetemp)  }
  }
  names(sum_ab) = uniquecodes
  return(sum_ab)
}

WAmaker=function(barcodesum_df,env_var) {
  # function reports R2 and RMSE for variable and rel abundance
  # barco...
  env_var= env_data$pH
  #barcodesum_df = barcodesum_df[,-which(colSums(barcodesum_df) == 0)] # there are
species missing data, must remove them before running a WA
  WA_out = WA(barcodesum_df, env_var, tolDW=T)
  #WApH = WA(species_df, pH, tolDW=T)
  #return(WApH)
  WAvals=performance(WA_out)
  R2= WAvals$object[2,2]
  RMSE= WAvals$object[1,1]
  return(c(R2,RMSE))
}
```

pH Europe as example of data management and dataset set up

```
##### TRAIT DATA ######
trait.Eur = read.csv('eur_ph.csv') #traits from Eur pH dataset
colnames(trait.Eur)[1]= "SppCode"

#### SET UP FOR BARCODING AND TESTS######
SpeciesByTrait =trait.Eur # Bind the european and NA datasets to include all the species
that might be present in ph lakes
SpeciesByTrait %>% mutate_all(as.character)  # convert SpeciesByTrait to character
#make sure none of the lakes in dataset repeat
SpeciesByTrait=unique(SpeciesByTrait) #accounts for species that are the same in both
datasets, results in 341 unique species
barcode = barcoder(SpeciesByTrait) #run barcoder function
spp_name=SpeciesByTrait$SppCode
species_code=data.frame(spp_name,barcode)
sort(barcode)
n_code = length(unique(barcode)) #shows how many unique barcodes there are
```

```
unique = (unique(barcode))

### Import Env Data used for tests####

env_data = read.csv("ph_data_eur.csv")

env_data[is.na(env_data)]<- 0 #switch all remaining NAs to 0s
is.na(env_data) # check to see if there are NAs left


###troubleshooting data, some rows not numeric##
#play=is.numeric(env_data)
#play=unlist(lapply(env_data, is.numeric))
#write.table(play)
#write.csv(play, file="pHasNumeric.csv")
#env_data[,25]

#looking for na values
which.nonnum <- function(x) {
  badNum <- is.na(suppressWarnings(as.numeric(as.character(x))))
  which(badNum & !is.na(x))
}
#which.nonnum(env_data)


species_df = subset(env_data, select = -c(LakeCode, pH))
#remove unecessary columns (not currently interested in looking at the coordinates or
speacies codes), the env_data df already contains this info
species_df[is.na(species_df)]<- 0


#### RUN PERMUTAIONS #####
results=as.data.frame(matrix(data=NA,nrow=nrow(SpeciesByTrait), ncol=3))
#hold the R2 and RMSE and trait removed here
#maybe change this to ncol for species traits so theres only 22 rows

#WAresult = data
spp_df = species_df #is this overwrite necessary?


#### RUN PERMUTAIONS #####
cols <- names(SpeciesByTrait)[-1]
n <- length(cols)
trt <- data.frame(t(rep(1,n)))
names(trt) <- cols
```

```
id <- unlist(
  lapply(1:n,
         function(i)combn(1:n,i,simplify=FALSE)
  )
  ,recursive=FALSE)


results <- data.frame(Trait = character(),
                      R2 = numeric(),
                      RMSE = numeric(),
                      stringsAsFactors = FALSE
)
#hold the R2 and RMSE and trait removed here
#maybe change this to ncol for species traits so theres only 22 rows

#WAresult = data
# spp_df = species_df #is this overwrite necessary?
# barcode2spp = species_code_new #can't do this until after for loop runs

for (tt in 1:100) {
  # for (tt in 1:length(id)) {
  #t=5
  code_new = barcoder(traitremover(SpeciesByTrait, id[[tt]]))
  species_code_new=data.frame(spp_name,code_new)
  #sort(code_new)
  length(unique(code_new))
  #uniquecodes = unique(code_new)
  #n_code=length(uniquecodes)
  codesums_df = matchNsum(species_df,species_code_new)
  WAresult = WAmaker(codesums_df, env_data$pH)
  #results[tt, "Trait"] <- paste0(code_new[tt]) # this does not work after 200 iterations
  results[tt, "Trait"] <- paste0(traitremover(trt,id[[tt]]), collapse = "")
  results$R2[tt] <- WAresult[1]
  results$RMSE[tt] <- WAresult[2]
  #maxtrait=as.numeric(which.max(results[,2]))
  #when this trait was removed there was little or no change in R2 value, meaning it is
least influencial
  #mintrait=as.numeric(which.min(results[,2]))
  #when this trait is removed there is the largest dip in R2, meaning it was the most
influencial
  ## Could we use the maxtrait to automate the removal of multiple traits? ##
  #Something like#
  #for (s in 1:ncol(SpeciesByTrait))
  #code_new = barcoder(traitremover(SpeciesByTrait, t, maxtrait) ##this line won't work
as t+maxtrait isn't valid, and the loop hit an error when it tries to remove the same trait
twice##
```

```
  if(! tt %% 1000){
    print(tt)
    flush.console()
  }

}



### Interpreation ###

setwd('/Users/mckercherk/Desktop/Thesis/pH Eur Results')

out1 <-read_csv("eur1.csv")
out2 <-read_csv("eur2.csv")
out3 <-read_csv("eur3.csv")
out4 <-read_csv("eur4.csv")
out5 <-read_csv("eur5.csv")
out6 <-read_csv("eur6.csv")
out7 <-read_csv("eur7.csv")
out8 <-read_csv("eur8.csv")
out9 <-read_csv("eur9.csv")
out10 <-read_csv("eur10.csv")
out11 <-read_csv("eur11.csv")
out12 <-read_csv("eur12.csv")
out13 <-read_csv("eur13.csv")
out14 <-read_csv("eur14.csv")



#probs value indicates what fraction of values to use. 0.995 results in only using the highest 0.5% of values
out_red1 <- out1[out1$R2 > quantile(out1$R2, probs = 0.995), ]
out_red2 <- out2[out2$R2 > quantile(out2$R2, probs = 0.995), ]
out_red3 <- out3[out3$R2 > quantile(out3$R2, probs = 0.995), ]
out_red4 <- out4[out4$R2 > quantile(out4$R2, probs = 0.995), ]
out_red5 <- out5[out5$R2 > quantile(out5$R2, probs = 0.995), ]
out_red6 <- out6[out6$R2 > quantile(out6$R2, probs = 0.995), ]
out_red7 <- out7[out7$R2 > quantile(out7$R2, probs = 0.995), ]
out_red8 <- out8[out8$R2 > quantile(out8$R2, probs = 0.995), ]
out_red9 <- out9[out9$R2 > quantile(out9$R2, probs = 0.995), ]
out_red10 <- out10[out10$R2 > quantile(out10$R2, probs = 0.995), ]
out_red11 <- out11[out11$R2 > quantile(out11$R2, probs = 0.995), ]
out_red12 <- out12[out12$R2 > quantile(out12$R2, probs = 0.995), ]
out_red13 <- out13[out13$R2 > quantile(out13$R2, probs = 0.995), ]
out_red14 <- out14[out14$R2 > quantile(out14$R2, probs = 0.995), ]
```

```
out_red <- rbind(out_red1,out_red2,out_red3, out_red4, out_red5,
                 out_red6, out_red7, out_red8, out_red9, out_red10,
                 out_red11, out_red12, out_red13, out_red14)

trt <- rbind(unlist(strsplit(out_red$Trait[1], split = "")))

for(ii in 2:nrow(out_red)){
  trt <- rbind(trt,
               rbind(unlist(strsplit(out_red$Trait[ii], split = ""))))
}

trt[ trt == "_" ] <- NA

trt <- data.frame(trt, stringsAsFactors=FALSE)

trt <- as.data.frame(sapply(trt, as.numeric))

cs <- colSums(trt, na.rm = TRUE)
fr <- round(cs/nrow(trt), 2) #fraction of times a trait is present in filtered results

df <- data.frame(trait = sprintf("Trait %s", 1:20), value = fr)

p<-ggplot(data=df, aes(x=trait, y=value), title= "") +
  geom_bar(stat="identity")


boxplot <- boxplot(out_red1$R2, out_red2$R2, out_red3$R2, out_red4$R2,
out_red5$R2, out_red6$R2,
                   out_red7$R2, out_red8$R2, out_red9$R2, out_red10$R2, out_red11$R2,
out_red12$R2,
                   out_red13$R2, out_red14$R2, ylab="Top 0.05% of R2 values", ann = TRUE,
xlab = "Number of traits removed",
                   main = "Comparison of top 0.05% of R2 values achieved upon the removal of
a different number of traits",
                   sub = "pH Europe", col.sub = "grey")
              axis(1, labels = TRUE)

#Could these plots be added to automated for loop?
plot(WA(species_df, env_data$pH, tolDW=T), xlab="Predicted pH", ylab='Actual
pH',main='Original data WA plot', sub="R2=something")

codesums_df = codesums_df[,-which(colSums(codesums_df) == 0)]
quartz()
plot(WA(codesums_df, env_data$pH, tolDW=T), xlab="Predicted pH", ylab='Actual
pH',main='WA with 7 traits total', sub="R2=0.6984")
```

```
### Average R2 value ###

out1AVG = mean(out1$R2)
out2AVG = mean(out2$R2)
out3AVG = mean(out3$R2)
out4AVG = mean(out4$R2)
out5AVG = mean(out5$R2)
out6AVG = mean(out6$R2)
out7AVG = mean(out7$R2)
out8AVG = mean(out8$R2)
out9AVG = mean(out9$R2)
out10AVG = mean(out10$R2)
out11AVG = mean(out11$R2)
out12AVG = mean(out12$R2)
out13AVG = mean(out13$R2)
out14AVG = mean(out14$R2)

outAVG = c(out1AVG, out2AVG, out3AVG, out4AVG, out5AVG, out6AVG,
out7AVG, out8AVG, out9AVG, out10AVG,
        out11AVG, out12AVG, out13AVG, out14AVG)


outAVG = as.table(c(out1AVG, out2AVG, out3AVG, out4AVG, out5AVG, out6AVG,
out7AVG, out8AVG, out9AVG, out10AVG,
                out11AVG, out12AVG, out13AVG, out14AVG))

write.csv(outAVG, "pH EUR avg r2.csv")

out1SD = sd(out1$R2)
out2SD = sd(out2$R2)
out3SD = sd(out3$R2)
out4SD = sd(out4$R2)
out5SD = sd(out5$R2)
out6SD = sd(out6$R2)
out7SD = sd(out7$R2)
out8SD = sd(out8$R2)
out9SD = sd(out9$R2)
out10SD = sd(out10$R2)
out11SD = sd(out11$R2)
out12SD = sd(out12$R2)
out13SD = sd(out13$R2)
out14SD = sd(out14$R2)
```

```
outSD = as.table(c(out1SD, out2SD, out3SD, out4SD, out5SD, out6SD, out7SD, out8SD,
out9SD, out10SD, out11SD,
                   out12SD, out13SD, out14SD))

tograph = rbind(outAVG, outSD)
write.csv(tograph, "ph EUR line graph.csv")

write.csv(outAVG, "pH EUR avg r2.csv")

avgR2 = read_csv("avgR2.csv")

avg_plot = plot(avgR2$`pH NA Average R2`, ylim = c(0.1, 0.65), type = "b",main =
"Average R2 value achieved when removing differing numbers of traits",
                xlab= "Number of traits removed", ylab= "Average R2 value")


#add 2nd line
points(avgR2$`pH EUR Average R2`, col="red")
lines(avgR2$`pH EUR Average R2`, col="red")


#add 3rd line
points(avgR2$`Phosphorous Average R2`, col="blue")
lines(avgR2$`Phosphorous Average R2`, col="blue")

#add 4th line
points(avgR2$`Salinity Average R2`, col="purple")
lines(avgR2$`Salinity Average R2`, col="purple")

#add 5th line
points(avgR2$`Depth Average R2`, col="green")
lines(avgR2$`Depth Average R2`, col="green")



#### Cross Validation ####

mod=WA(species_df, env_data$pH, tolDW=T)
cv.boot= crossval(mod, method = "bootstrap", nboot = 1000)

mod1=WA(codesums_df, env_data$pH, tolDW=T)
crossval(mod1, method = "bootstrap", nboot = 1000)
quartz()
plot(mod1)

Traits_R2=read.csv("Traits_R2.csv")
```

```
quartz()
plot(Traits_R2$Number.of.traits,Traits_R2$R2, type="o")



uniquecodes=unique(species_code_new[,2])
n_code = length(uniquecodes)
sum_ab = as.data.frame(matrix(data=NA,nrow=nrow(spp_df),
ncol=ncol(species_code_new)))  #the values stored in here arent changing each time
around - they should be  #dataframe

for (i in 1:n_code) { #something in here isnt working
  tmpsp = as.character(barcode2spp[which( barcode2spp[,2] == uniquecodes[i]),1]) #this
doesnt always work
  barcodetemp=spp_df[,na.omit(match(tmpsp, names(species_df)))]

  if(is.null(ncol(barcodetemp))) {
    sum_ab[,i] = barcodetemp} else
    { sum_ab[,i] = rowSums(barcodetemp)  }
}
names(sum_ab) = uniquecodes
return(sum_ab)


for (t in 1:ncol(SpeciesByTrait)) {
  #t=5
  code_new = barcoder(traitremover(SpeciesByTrait, t))
  species_code_new=data.frame(spp_name,code_new)
  #sort(code_new)
  length(unique(code_new))
  #uniquecodes = unique(code_new)
  #n_code=length(uniquecodes)
  codesums_df = matchNsum(species_df,species_code_new)
  WAresult = WAmaker(codesums_df, env_data$pH)
  results[t,1]= names(SpeciesByTrait[t])
  results[t,2]=WAresult[1]
  results[t,3]= WAresult[2]
  maxtrait=which.max(results[,2])
  #when this trait was removed there was little or no change in R2 value, meaning it is
least influencial
  mintrait=which.min(results[,2])
  #when this trait is removed there is the largest dip in R2, meaning it was the most
influencial
  ## Could we use the maxtrait to automate the removal of multiple traits? ##
  #Something like#
  #for (s in 1:ncol(SpeciesByTrait))
```

```
  #code_new = barcoder(traitremover(SpeciesByTrait, t + maxtrait) ##this line won't work as t+maxtrait isn't valid, and the loop hit an error when it tries to remove the same trait twice##
}



##### Original data results ####

codesums_df = matchNsum(species_df, species_code)

species_result = WA(species_df, env_data$pH, tolDW = TRUE)
Original_plot = plot(species_result,
                     xlab= 'measured pH', ylab= 'predicted pH')


barcode_result=WA(codesums_df, env_data$pH, tolDW=T)
barcode_plot = plot(barcode_result,
                    xlab= 'measured pH', ylab= 'predicted pH')

#### Cross Validation ####
cv.boot= crossval(species_result, method = "bootstrap", nboot = 1000)
crossval(barcode_result, method = "bootstrap", nboot = 1000)


### Correlation ####
rm(list=ls())
install.packages("corrplot")

 library(corrplot)

##### TRAIT DATA ######
trait.Eur = read.csv('eur_ph.csv') #traits from Eur pH dataset
colnames(trait.Eur)[1]= "SppCode"

trait.NA =read.csv('na_ph.csv') #traits from NA pH dataset same for Phos dataset
SpeciesByTrait =trait.NA[-c(116),] #species in row 116 does not match

trait.BC =read.csv('bc_sal.csv') #traits for Salinity and depth the same
colnames(trait.BC)[1]= "SppCode"

trait.data = rbind(trait.Eur, trait.NA, trait.BC)

trait.data = trait.data[,-1]

dev.new(width=20, height=18, unit="in")
```

```
par(mfrow = c(2,2))
layout(matrix(c(1,1,1,1,1,1,2,3,4), nrow = 3, ncol = 3, byrow = TRUE))

#layout(matrix(c(1,1,1,1), nrow = 3, ncol = 3, byrow = TRUE))

#### All datasets ######
colnames(trait.data) <- c("Trait 1", "Trait 2", "Trait 3", "Trait 4", "Trait 5", "Trait 6",
"Trait 7",
              "Trait 8","Trait 9", "Trait 10", "Trait 11", "Trait 12", "Trait 13", "Trait 14",
              "Trait 15", "Trait 16","Trait 17", "Trait 18", "Trait 19", "Trait 20")

cor1 = cor(trait.data, method = "pearson")
corrplot = corrplot(cor1, method = "color",type = "lower")

#### pH Eur #####
trait.Eur= trait.Eur[,-1]
colnames(trait.Eur) <- c("Trait 1", "Trait 2", "Trait 3", "Trait 4", "Trait 5", "Trait 6",
"Trait 7",
                    "Trait 8","Trait 9", "Trait 10", "Trait 11", "Trait 12", "Trait 13", "Trait
14",
                    "Trait 15", "Trait 16","Trait 17", "Trait 18", "Trait 19", "Trait 20")

cor2 = cor(trait.Eur, method = ("pearson")
corrplot = corrplot(cor2,method = "color", type = "lower",)

##### pH NA ######
trait.NA= trait.NA[,-1]
colnames(trait.NA) <- c("Trait 1", "Trait 2", "Trait 3", "Trait 4", "Trait 5", "Trait 6",
"Trait 7",
                    "Trait 8","Trait 9", "Trait 10", "Trait 11", "Trait 12", "Trait 13", "Trait
14",
                    "Trait 15", "Trait 16","Trait 17", "Trait 18", "Trait 19", "Trait 20")

cor3 = cor(trait.NA, method = ("pearson")
corrplot = corrplot(cor3,method = "color", type = "lower")


## Salinity and depth traits ##
trait.BC= trait.BC[,-1]
colnames(trait.BC) <- c("Trait 1", "Trait 2", "Trait 3", "Trait 4", "Trait 5", "Trait 6",
"Trait 7",
                    "Trait 8","Trait 9", "Trait 10", "Trait 11", "Trait 12", "Trait 13", "Trait
14",
                    "Trait 15", "Trait 16","Trait 17", "Trait 18", "Trait 19", "Trait 20")

cor4 = cor(trait.BC, method = ("pearson")
```

```
corrplot = corrplot(cor4,method = "color", type = "lower")

# tables ##
write.csv(cor1, "All traits Correlations.csv")
write.csv(cor2, "pH Eur traits Correlations.csv")
write.csv(cor3, "pH NA and phosphorous traits Correlations.csv")
write.csv(cor4, "Salinity and Depth traits Correlations.csv
```