SENSOR DATA ANALYTICS AND WEB APPLICATIONS TO IMPROVE MONITORING AND UNDERSTANDING OF LAKE PROCESSES

BY

WENZHAO XU

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Environmental Engineering in Civil Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Barbara Minsker, Chair
Professor Albert Valocchi
Dr. Paris Collingsworth, Purdue University
Associate Professor Feng Liang

ABSTRACT

Lakes are complex systems that involve numerous physical, chemical and biological processes. With modern sensor technology, large amounts of sensor data on lake water chemistry are being generated to help researchers understand the spatial and temporal patterns of these lake processes. Each sensor generates different datasets and effectively utilizing the resulting large and diverse datasets to improve understanding of lake processes and optimize sampling strategies is essential to protect and improve lake resources. For example, in the Great Lakes, the case study in this thesis, the US Environmental Protection Agency (USEPA) conducts several monitoring programs with various sensors, including the TRIAXUS undulating vehicle, the Sea-Bird CTD (Conductivity, Temperature, Depth) depth profiler, and a dissolved oxygen (DO) logger network that are the focus of this study. In this work, we develop three data analysis frameworks to support limnologists in more effectively collecting and analyzing these types of datasets, providing a lake system perspective. The frameworks have been made available to the research community as open-source code, including three prototype interactive Web applications.

For towed undulating vehicles such as TRIAXUS, we propose a geospatial analysis framework and software to interpret water-quality sampling data in near-real time. The framework includes data quality assurance and quality control processes, automated kriging interpolation along undulating paths, and local hotspot and cluster analyses. The approach is demonstrated using historical sampling data from an undulating vehicle deployed at three rivermouth sites in Lake Michigan during 2011. The normalized root-mean-square error (NRMSE) of the interpolation averages approximately 10% in 3-fold cross validation. The results show that the framework can be used to track river plume dynamics and provide insights on mixing, which could be related to wind and seiche events.

Next, we develop and test algorithms for rapid and consistent analysis of depth profiling data sampled from CTD profilers to identify lake stratifications and deep chlorophyll layers (DCL). We develop a segmentation method to approximate vertical temperature profiles with linear segments using Piecewise Linear Representation (PLR) algorithm, from which stratification patterns can be extracted. We also propose an automated peak detection algorithm to identify the fluorescence peak where the DCL lies. Testing the algorithms with data from the

Great Lakes, we obtained similar results to human judgments from historical surveys. The algorithms are able to reveal spatial and temporal trends of the thermocline and DCL, as well as analyzing the shape of temperature and fluorescence profiles to detect unusual patterns such as a double thermocline.

Finally, we develop a spatio-temporal interpolation framework that identifies the spatially varying temporal trend and estimates hourly hypoxia extent (dissolved oxygen [DO] concentration lower than 2mg/L) with estimation uncertainty. The framework is used to analyze spatio-temporal datasets of dissolved oxygen in Lake Erie, which were sampled from a logger network placed at the lake bottom in 2014, 2015, and 2016. The results show that hypoxia developed differently in these years. The locations with longest total hypoxic duration and longest continuous hypoxic duration are also different. Based on cross-validation results and DO time series patterns, some implications for optimizing logger locations are discussed.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support of my collaborators, my friends, and my family.

My most sincere thanks go to my advisor, Professor Barbara Minsker. I greatly appreciate her guidance, encouragement, and trust during my Ph.D. years. Thanks for letting me develop my interests in data science, advising me on critical thinking, teaching me mindfulness to manage stress, and helping me to improve my writing greatly. Those skills will continue to benefit me into the future.

I would also like to express my special thanks to Professor Paris Collingsworth, who guides me in the limnology field and provides great support, including academic guidance, funding information, and logistical arrangements that made this research possible.

To Professor Barbara Bailey, thank you for helping me explore the geostatistical world and continuously providing support throughout my study.

To my other thesis committee members, Professor Feng Liang and Professor Albert Valocchi, thank you for providing suggestions that improved my thesis work.

To Yuan Lei, my mentor in a Dow AgroSciences internship, thanks for the helping me to improve my data science skills.

To all of the members of my research group and my friends, thank you for your friendship and encouragement.

Finally, deepest gratitude goes to my parents.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

In this chapter, the research content and significance of the work is summarized, followed by background information on the case study, which focuses on the Great Lakes monitoring programs.

1.1 Thesis Content and Research Significance

With various monitoring programs launched to collect and analyze lake water quality data, increasing amounts of data are being generated, including large volumes of sensor data. Effectively and efficiently analyzing and utilizing these sensor data is becoming more and more important to researchers as well as decision makers.

The goal of this research is to use geospatial analytics to detect patterns in sensor data and support limnologists in better understanding lake processes. In this research, three analytical frameworks are developed to improve understanding of river plumes, lake stratification, deep chlorophyll layers, and hypoxia. For each type of data, we (1) propose and test the new analysis methods/frameworks, (2) explore the lake processes that can be explained by the results, and (3) provide suggestions to improve sampling activities. We also develop prototypes of related Web applications and provide open-source code for the research community to further test and deploy.

In Chapter 2, we provide a new geospatial near-real-time analysis framework to interpret river mouth water-quality sampling data from towed undulating vehicles. This study is novel in developing methods for plume pattern detection during sampling activities rather than in post-sampling analysis as in previous work. In Chapter 3, we propose algorithms to automatically detect lake stratification patterns and deep chlorophyll layers (DCL) from depth profiling data sampled by CTD (Conductivity, Temperature and Depth) profilers. The algorithms extend previous research to provide consistent and objective references for full lake feature detection, and algorithm parameters suitable for the Great Lakes are suggested. In Chapter 4, to study hypoxia extent, we use spatial-temporal interpolation to analyze dissolved oxygen data sampled by a newly-implemented sensor network. The seasonal patterns of hypoxia extent in Lake Erie are characterized and the results provide insights for optimizing future logger deployment locations. In each chapter, related lake processes are introduced first, followed by the data source, methodology, and results section. The last chapter discusses conclusions and future

research recommendations.

1.2 Case Study: Great Lakes Monitoring and Protection Programs

The analytical frameworks are tested with datasets collected by the U.S. EPA Great Lakes National Program Office (GLNPO) in the Great Lakes. The Great Lakes are a series of interconnected freshwater lakes located on the Canada-United States border, including Lakes Superior, Michigan, Huron, Erie, and Ontario. They form the largest group of freshwater lakes on Earth, containing 21% of the world fresh surface water by volume and 84% of North America's fresh surface water. More than 30 million people live in the Great Lakes basin and rely on the lakes for water supply, commercial fishing, and recreation.

In the 1950s, Lake Erie began to have massive algae blooms and hypoxia events due to synthetic fertilizers, nutrient-rich organic pollutants, and phosphate detergents being released to the lakes. Due to the importance of the Great Lakes for their economic and ecological value, and raising concerns about the deterioration of water quality, the United States and Canada first signed the Great Lakes Water Quality Agreement (GLWQA) in 1972 to address a wide range of water quality issues. It was amended in 1983, 1987 and 2012. The Agreement now aims to ensure the chemical, physical, and biological integrity of the Great Lakes (GLWQA, 2012).

To protect the Great Lakes, various monitoring programs were launched to collect and analyze water quality data from the Great Lakes. The United States Environmental Protection Agency (USEPA) has various Great Lakes monitoring programs, including the Great Lakes Fish Monitoring and Surveillance Program, the Great Lakes Biology Monitoring Program, and the Great Lakes Integrated Atmospheric Deposition Network that monitors concentrations of persistent toxic chemicals in Great Lakes air and precipitation. Cooperating with Canada, USEPA has conducted a binational effort called Coordinated Science and Monitoring Initiative (CSMI) since 2002. CSMI aims to assess conditions in one of the five lakes each year. Most of the sampling is conducted on USEPA's research vessel named Lake Guardian. Data from the monitoring programs are made available through a database called Great Lakes Environmental Database (GLENDA). Besides EPA programs, other major monitoring programs or research projects include the Great Lakes Restoration Initiative (launched in 2010) and Ecological Fore-casting: Hypoxia Assessment in Lake Erie (EcoFore-Lake Erie, launched in 2005).

CHAPTER 2: DETECTING SPATIAL PATTERNS OF RIVERMOUTH PROCESSES USING A GEOSTATISTICAL FRAMEWORK FOR NEAR-REAL-TIME ANALYSIS

In this chapter, we will first look at rivermouth systems and describe a near-real-time framework to maximize the yield of sampling processes. Section 2.1 gives an introduction on river plume dynamics. Section 2.2 describes the study area and data source. Section 2.3 presents the methodology. Section 2.4 and 2.5 are the results and discussion sections that show how the new approach is able to detect river plume patterns. Conclusions are given in Section 2.6.

2.1 Introduction

Rivermouth ecosystems are dynamic transitional river and lake mixing zones that can extend many kilometers upstream of the river/lake confluence and a similar distance into the lake. Rivermouths contain three parts: the lower river valley, receiving basin, and nearshore area (Larson et al., 2013). In this study, we mainly focus on the nearshore area that is influenced by the river plume.

2.1.1 Rivermouth and River Plume

Rivermouths provide a diversity of services such as fish production, water supply, erosion and sedimentation regulation, harbors, and recreation. They are also important biologically productive areas that support diverse habitats (Larson et al., 2013). River plumes affect nearshore water chemistry (Kaur et al., 2007; Makarewicz and Howell, 2012), bacteria transportation (Nekouee, 2012), and fish community composition (Janetski et al., 2013). Knowledge about rivermouth mixing patterns and especially plumes has become vital in understanding their role in maintaining nearshore and deepwater food webs (Hoffman et al., 2010; Larson et al., 2013).

For example, the recent (1990-present) invasion and proliferation of Dreissenid mussels have been implicated in the collapse of deepwater fish communities in Lakes Michigan and Huron (Riley et al., 2008; Madenjian et al., 2012). Mussels are thought to be sequestering energy and nutrients in nearshore areas that formerly supported fish in offshore and deepwater habitats (Hecky et al., 2004). Rivermouth ecosystems and their associated plumes may be one of the few areas where historical food webs are still intact, but food web assessments in such habitats have been limited due to the dynamic nature of plumes.

The complexity of the rivermouth system impedes understanding of the river plume

dynamics and their effects, which are controlled by many factors such as vertical/horizontal mixing, dispersion, density and seiche effects (Rao and Schwab, 2007, Jackson et al. 2008). Seiche events, for example, are wind-induced water-level fluctuations that bring large volumes of lake water into rivermouths and can create backflow, which may affect the location of mixing zones (Pebbles et al., 2013). The local topology and shoreline angle determine rivermouth exposure to wind and waves and also influence plume dynamics. Moreover, phytoplankton distributions not only depend on temperature, bathymetry and hydrologic features such as watershed type and riverine input (Pavlac et al., 2012; Snow et al., 2000), but also are influenced by wind and the presence of older plumes (Hickey et al., 2005; Frame and Lessard, 2009). Therefore, it is important to understand plume dynamics to fully comprehend rivermouth systems.

2.1.2 Existing Monitoring and Sampling Approaches

Understanding rivermouth dynamics requires comprehensive water quality data (Howell et al., 2012). Traditionally, rivermouth data are collected via fixed stations or buoys that continuously or periodically measure water chemistry. For example, the National Oceanographic and Atmospheric Administration (NOAA) have significant amounts of buoy data sampled at the coastline (https://coastwatch.glerl.noaa.gov). However, this approach provides data that are limited spatially by the existing buoy network of NOAA. Another approach is using a mobile sampling platform with flow-through systems that continuously pump water from a fixed depth through a series of sensors to obtain water chemistry data (Pavlac et al., 2012; Twiss and Marshall, 2012). This extends the spatial range of data collection, but fails to sample data throughout the water column.

A promising approach to sample data at extensive three-dimensional (3-D) spatial scales is to use towed or autonomous undulating vehicles that carry multiple sensors. Such a vehicle may be autonomous or towed behind a ship that moves along different survey paths, undulating throughout the survey between the water surface and the near bed region of the water column. Such vehicles currently in operation include ScanFish (Ludsin et al., 2009), SARAGO (Marcelli et al., 2005), TRIAXUS (Jones et al., 2011), V-Fin (Yurista et al., 2012), EcoMapper AUV (Jackson and Reneau, 2014) and various Gliders such as ROUGHIE (Page et al., 2017).

Monitoring with towed undulating vehicles requires expensive ship time so vehicles need to be deployed efficiently. Ships usually move along pre-defined transects or grid patterns and

the towed vehicles collect data along each transect while undulating to sample at multiple depths. However, grids that are too small may fail to capture the river plume, while those large enough to capture the river plume fully also may expend excess time and effort to capture data outside of the plume that are not needed. In addition, analyzing data from gridded sampling assumes stationarity of the river plume, and the river plume state may change markedly during the time spent sampling a large grid, thus introducing temporal change into the spatial variability of the data. The adaptive sampling strategy, which involves adjusting collection strategies based on previously collected data to minimize effort while maximizing river plume coverage, is one possible solution to this problem.

However, adaptive sampling raises a second serious problem: the large amount of high-frequency data that are collected by towed undulating vehicles are difficult to analyze quickly enough to adjust sampling. This is especially true for tow-yo sampling, where kriging interpolation is used to provide direct visualization of sampling results. Existing commercial software (such as Surfer, Golden Software) requires researchers to manually fit a variogram (Ludsin et al., 2009; Yurista et al., 2009), which is time-consuming and such data are usually analyzed after collection, making adaptive sampling impossible. New and efficient methods are needed to analyze data onboard the vessel as it is being collected.

In this study, we propose an automated kriging method that interpolates raw data onto grid maps that allow users to visualize patterns and adjust sampling in near-real time. To highlight the spatial distribution of variables in a distinct and informative way, we use hotspot analysis with local G statistics (Ord and Getis, 1995). We then further cluster the water chemistry data to explore the mixing structure of the river and lake water. The analysis framework has been implemented in an interactive Web application developed with the Shiny package in the R programming development environment. This will allow researchers on research vessels to easily perform analysis in near-real time.

2.2 Study Area and Data Description

We illustrate the utility of the methods developed in this work for illuminating details of the river plume dynamics using data collected by the TRIAXUS undulating vehicle at the Manitowoc, Muskegon and Pere Marquette rivermouth areas in Lake Michigan during the summer of 2011. TRIAXUS, developed by MacArtney Underwater Technology, was towed

behind the research vessel, Lake Guardian (operated by the EPA-Great Lakes National Programs Office), along pre-defined transects parallel or perpendicular to the shoreline. Figure 2.1 shows the transects located in nearshore areas outside of the Manitowoc River, Muskegon River, and Pere Marquette River in Lake Michigan that were sampled during summer 2011. At these three sites, the TRIAXUS vehicle was deployed in undulating trajectories to measure water chemistry at different depths as the ship moved along each transect. The sampling depth of all paths ranged from 3 to 34 meters. Average wavelengths of the undulating cycles (i.e., the distance between two peak points or two valley points) ranged from 0.126 kilometers to 0.6 kilometers.

The TRIAXUS carried multiple sensors that measured specific conductance, temperature, turbidity (measured as beam attenuation coefficients (BAT)), dissolved oxygen (DO), indices of chlorophyll concentration and algal accessory pigments, and zooplankton biomass and density. Chlorophyll concentrations were measured by a FluoroProbe sensor, which uses excitation light with varying wavelengths to distinguish algae fluorescence among different algal groups. The validation and potential cautions of using FluoroProbe to estimate phytoplankton community are given by Catherine et al. (2012). Zooplankton biomass and density were derived from a laser optical plankton counter (LOPC), which counts the number of particles in different size bins (from 105um to 1920um with step size 15um). The methods for comparing LOPC output to zooplankton biomass and density derived from traditional sampling methods is described in Watkins et al. (2016). Other variables were measured by a SeaBird CTD (conductivity, temperature, and depth) sensor attached to the vehicle. As a result, multi-dimensional spatial data with longitude, latitude, and depth as coordinates were generated.

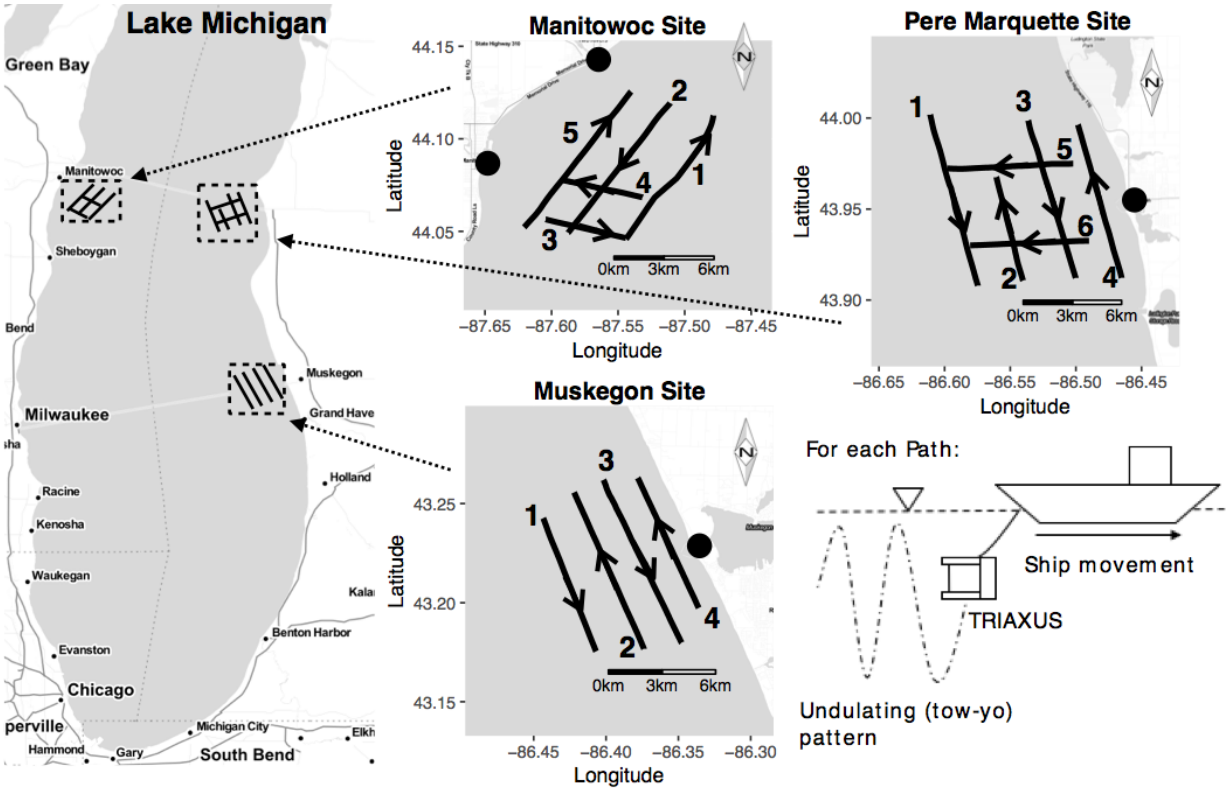


Figure 2.1. Sampling paths at three rivermouth zones in Lake Michigan. The black dots are the locations of rivermouths. Sampling at Manitowoc, Muskegon, and Pere Marquette began at 14:00 June 29, 21:30 June 27 and 13:30 June 28 in 2011 (UTC) and lasted for 5, 5.5, and 8.5 hours, respectively. The numbers besides each path are the path index, which are ordered according to the sampling order. The arrow indicates the ship direction. Maps are from Stamen (http://maps.stamen.com).

2.3 Methodology

2.3.1 General Description

Figure 2.2 shows the data analysis framework proposed and applied in this work. First, a data quality assurance/quality control (QA/QC) step removes outliers and anomalies in the data. Next, we use automated kriging interpolation to visualize water chemistry properties on grid maps from the sampling data. Based on the interpolations, two spatial statistical methods, local G statistics and k-means clustering algorithm, are implemented to identify patterns in the data. The proposed methods aim to extract the information from the raw data paths and require minimal human interaction. Such automated processes can extract information during the sampling activities, rather than as post-sampling analysis, enabling near-real-time adaptive observation.

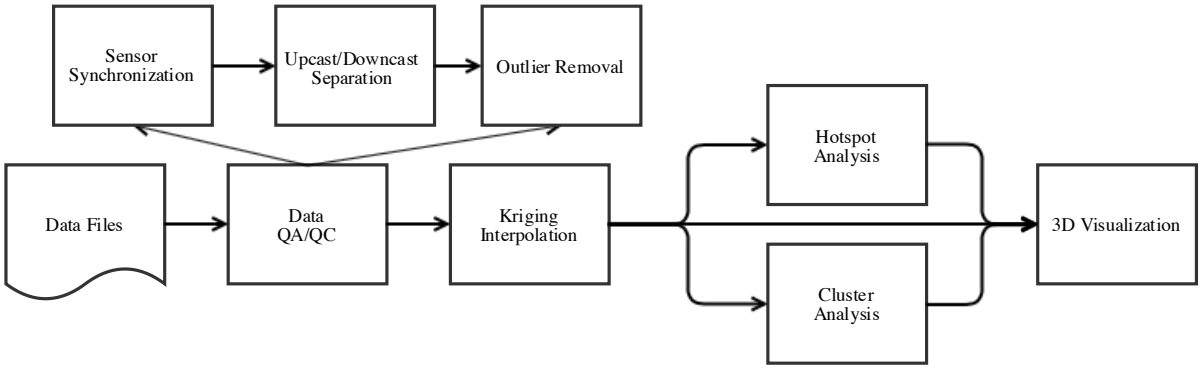


Figure 2.2. Flow chart of data analysis process

2.3.2 Data QA/QC

Direct visualization of the raw data in 2-D maps revealed several issues with the raw data streams. First, the sensors have different sampling frequencies (e.g., SeaBird CTD sensor sampled every 0.5 seconds while the FluoroProbe sensor sampled every 2 to 4 seconds) and need to be synchronized to correctly represent water-quality features at the same sampling location. Second, some variables, such as dissolved oxygen (DO) concentrations, have slightly different values for the same depth on upcasts and downcasts (e.g., see Figure 2.3a). As neither cast was more reliable than the other, we separated them, interpolated based on each cast, and averaged the interpolated estimates. Third, outliers often exist in these data sets, particularly in the zooplankton biomass and density (e.g., a spike at x = 11.4 km and y = -10 m in Figure 2.3b), which are biologically implausible and likely caused by bubbles or suspended sediment. These spikes and other anomalies were removed so that the variogram estimations used in kriging interpolation are more stable and accurate. Details on these steps are given below.

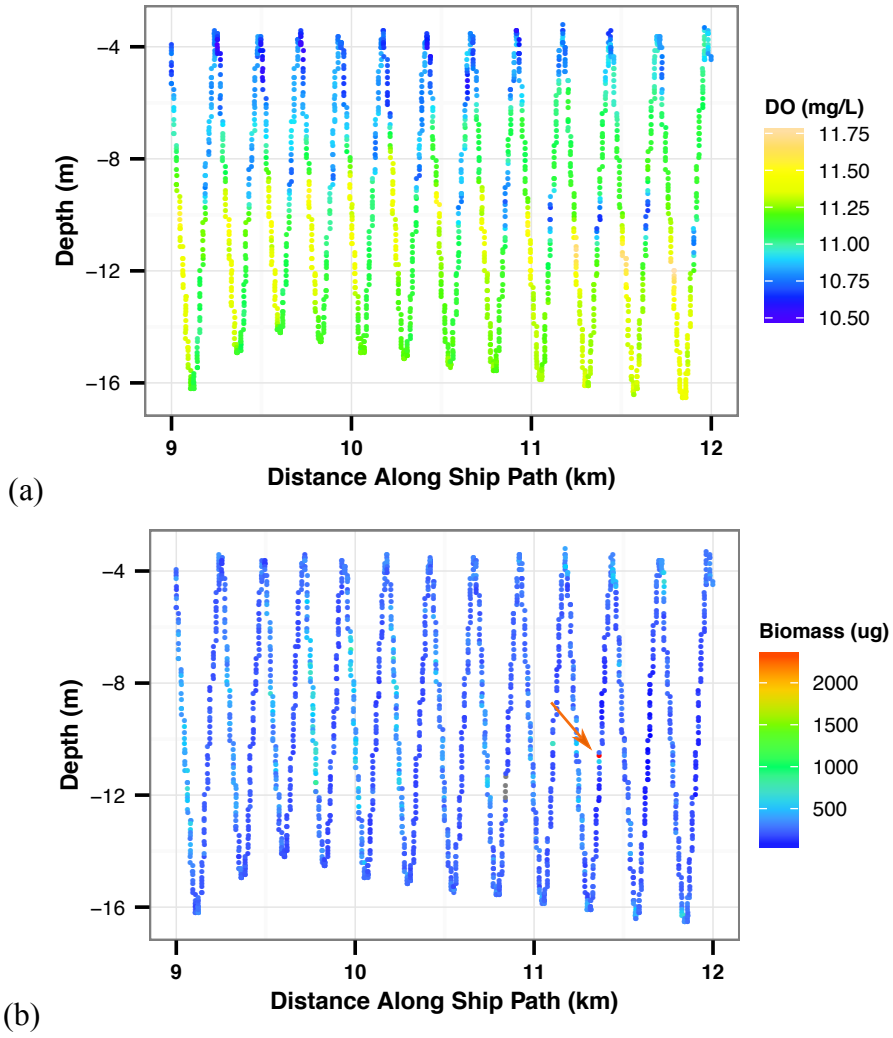


Figure 2.3. Raw data visualization and data-quality issues. (a) Inconsistency between upcast and downcast values in dissolved oxygen (DO); (b) Outliers in zooplankton biomass

To standardize measurements, we aligned the timestamps of each sensor to synchronize the data generated. For example, in our dataset, the starting times of SeaBird CTD sensor and FluoroProbe sensor are different so we first found timestamps ($t_{seabird}$, $t_{fluoroprobe}$) from both sensors that represent the same valley point or peak point of the sampling depths. Based on the differences of ($t_{seabird}$, $t_{fluoroprobe}$), we adjusted the FluoroProbe time to align with the SeaBird time and conducted a linear interpolation to map the FluoroProbe data onto SeaBird data points. The SeaBird sampling points are associated with ship GPS information, therefore sampling geo-location (longitude and latitude), sampling depth, and all water chemistry data can be correctly aligned. We then applied cubic smoothing spline ("smooth.spline" function in R "stat" package)

to reduce noise in the depth measurements and to find points more easily where the undulating vehicle changed direction. This is necessary to separate upcasts and downcasts for interpolation.

To filter zooplankton outliers, we first removed sampling points exceeding biomass or density values above the 99.5 percentile to eliminate implausible spikes. We then applied a spatial median algorithm (*Chen et al.*, 2008) to detect and remove significant spatial outliers. The median algorithm computes h_i as the difference between the value at point i and the median value in point i's neighborhood. The standardized h_i (zero mean and unit variance), denoted as y_i, satisfies a normal distribution. Therefore if $|y_i| \geq z_{\alpha/2}$ (α is the significance level), where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution, point i is a spatial outlier because it is significantly different from neighbor points. In our case study, neighbor points are defined as those within a circle of 0.75 kilometers in the x-axis and 0.75 meters in the y-axis around point i in the distance-depth map (similar to Figure 2.3). Using a predefined confidence interval, such as 99.9%, the sampling point i will be a spatial outlier when $|y_i| \geq 3.3$. For our case study, we used $|y_i| \geq 4$ to filter out only extreme outliers in the detrending residuals (discussed in the next subsection), which were assumed to be spatially stationary.

2.3.3 Kriging Interpolation

Kriging interpolation is a spatial interpolation method which assumes that the data points follow a Gaussian process and the covariance matrix is only related to the distance between points. Two nearby points will thus have highly correlated values. The covariance function is therefore only related to distance and it decreases as distance increases. Kriging interpolation has been used for similar undulating sampling data sets (Marcelli et al., 2005; Ludsin et al., 2009; Yurista et al., 2012). We did not choose the simplest inverse distance weighting interpolation (IDW) because: (1) the distances in IDW are hard to define as our datasets have different scales in x and y axis and (2) computational experiments have found that kriging methods are generally better than inverse distance weighting with different sampling patterns and surface types (Zimmerman et.al,1999), although not specifically for undulating patterns.

To apply the kriging method, a variogram is defined as $2\gamma(x, y) = var(Z(x) - Z(y))$, where $Z(s)$ is value at point *s* and γ is called the semivariogram. For a stationary process, the variogram can be related with the covariance function as $\gamma(x, y) = C(0) - C(h)$ where $C(0)$ is the variance of the spatial process, $C(h)$ is the covariance at h, and h the distance between point x and y. The variogram model (i.e. covariance function) must be estimated from the

observations (i.e. samples) and the variable values at unsampled locations are estimated based on sampled values and the estimated covariance matrix. Kriging is the best linear unbiased prediction (BLUP) (Oliver and Webster, 2015; Christensen, 2001). The Kriging interpolation equations are provided in the Appendix A.

We conducted two-dimensional (2-D) kriging interpolation for each variable at each path. We did not use some extended kriging methods such as 3-D kriging, spatio-temporal kriging or cokriging for the following reasons:

- In 3-D kriging, the covariance matrix is related with distance in 3-D space (i.e. longitude, latitude and depth). This approach was not used for the undulating datasets because: (1) water property patterns change significantly in different paths so we don't have enough data to capture the 3-D trend; and (2) the anisotropy ratio in 3-D space is not easy to estimate automatically and some datasets (e.g. Muskegon Site) only have paths parallel to the shoreline, so the correlation along the direction perpendicular to the shoreline may not have sufficient data to support 3-D kriging.
- In spatio-temporal kriging, the covariance matrix represents the covariance between points at different space-time coordinates. However, we do not have data sampled at the same location and different times, nor at the same time and different locations. Therefore, we have no information about the pure spatial covariance nor pure temporal covariance so that a spatio-temporal covariance matrix cannot be accurately estimated. In addition, our main interest is in modeling and characterizing the spatial distribution of the data rather than forecasting. The data were collected over a relatively short time period (5 to 8 hours at each location), during which the river plume dynamics can be assumed relatively steady. Therefore, we ignored the temporal component of the data to estimate a more complete spatial distribution.
- In co-kriging, the covariance matrix contains the covariance between two or more variables. Water properties may change even in one path, so that the correlations between variables are not spatially stationary. Thus, this type of dataset cannot adequately support co-kriging.

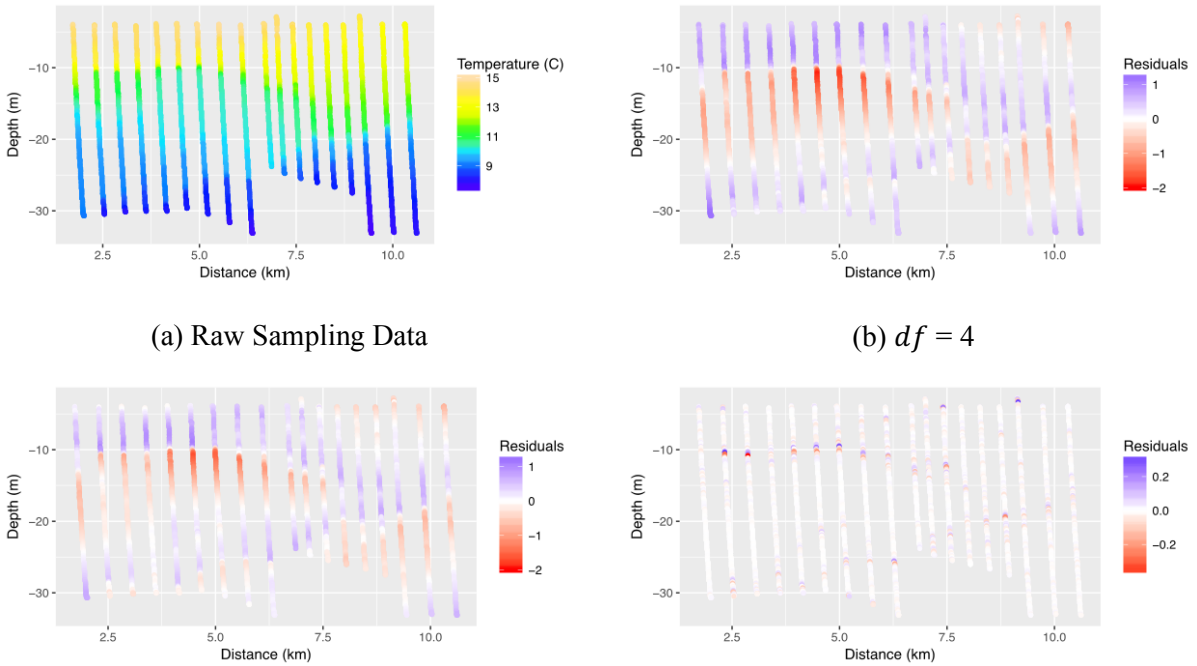
The kriging interpolation was implemented after data QA/QC to rapidly plot variation in water chemistry on a 2-D grid map. Raw upcast and downcast trajectories were kriged separately and then averaged on the same grid. We kriged the raw data separately for each sampling path

transect (Figure 2.1). Each kriging interpolation involves four steps: remove the trend to satisfy second-order spatial stationarity, fit a variogram model to the residuals, krig the residuals, and add back the trend values.

For detrending, we compared results from the linear regression model and thin plate spline (TPS) regression (Green and Silverman, 1993), which is a generalization of cubic smoothing spline. Both used x (distance) and y (depth) as independent variables. We chose the TPS regression, which involves fitting a flexible surface with a penalty that adjusts the smoothness of the fitted surface, because this method removed more of the trend pattern from our dataset. All TPS regressions were done with the "Tps" function in "fields" package in R.

TPS regression requires specifying a smoothing parameter (λ) to control the degree of data smoothing. Alternatively, this parameter can be indirectly specified via the effective number of parameters for the fitted surface (df). Figure 2.4 shows the effects on detrending residuals of different df values. As df increases, the large-scale patterns are gradually removed (Figure 2.4b to 4c). A df value that is too high generates a surface that is very flexible to fit the sampling data, even potential noise, and leads to little information in the residuals. Thus, interpolation on the residuals becomes meaningless (Figure 2.4d). On the other hand, a df that is too low (e.g., Figure 2.4b) still leaves some large-scale trend patterns in the residuals. Typically, generalized cross-validation (GCV) is used to choose the smoothing parameter. However, in our dataset, the GCV method gave a very small λ (large df on the order of hundreds) that captured too much detail or noise in the data (Figure 2.4d).

Through trial and error experimentations, we selected $df = 10$ as striking the right balance between these two extremes for our datasets, assuming that the residuals still have spatial autocorrelation left (Figure 2.4c). Our final interpolations are not overly sensitive to the value of df because changing df to 20 has very limited influence on the results. This is because the residual kriging stage can compensate for detrending differences caused by different df.



(a) Raw Sampling Data

(b) $df = 4$

(c) $df = 10$

(d) $df = 769$ (picked by GCV)

Figure 2.4. Detrending results using a different degree of freedom (df) in TPS. Data are temperature from downcast in Path1 at Manitowoc site as an example (Figure 2.1).

We used the widely-used R package "gstat" (Pebesma, 2004) to fit the variogram and perform kriging predictions. The cutoff to estimate the empirical variogram was chosen as one-third of the distance axis range. We chose this cutoff value based on the default value of "gstat" which uses one-third of the diagonal of the bounding box of the data. The y-axis (depth axis) of the map is also compressed or stretched by a factor K to eliminate the anisotropy revealed by the variogram. In this coordinate transformation, the x-axis is unchanged to include those horizontal points in the variogram estimation. The estimation of K is coupled with the variogram fitting processes. Assuming K is known, we estimated the empirical variogram by Cressie's robust variogram estimator (Cressie and Hawkins, 1980) in the vertical and horizontal directions and fit both spherical variogram models by the weighted least squares method, with the weights equal to N_j/h_j^2, where N_j are the number of points within binned distance h_j. Function "fit.variogram" in "gstat", which uses Levenberg-Marquard method, was used to fit the variogram model (nugget, range and sill) with default initial values. To increase the robustness of the fitting process, when unrealistic results were returned by "fit.variogram" due to singular model fits, we fixed nugget = 0 (or even fixed sill to the default values if necessary) and tried again. We fit different variogram

models for different variables in different paths.

The spherical model was chosen because the covariance matrix had a smaller conditional number so that the interpolation was more numerically stable, especially with a high-density sampling grid (Posa, 1989; Ababou et al., 1994). Our experiments also showed that Gaussian variogram models generated unstable results (i.e. very high or very low interpolated values) in some cases. We then optimized K to find K_{optm} that made the spherical variogram models in both directions as similar as possible so that the anisotropy was eliminated as much as possible. To optimize K, we first scaled both x and y-axis to [0,1] range and then used R's "optimize" function in "stats" package, which uses a combination of golden section search and successive parabolic interpolation, to optimize the value of K in the range [0.3, 5] (chosen by trial and error) with an accuracy of 0.05 to prevent non-convergence. We compared this approach with the genetic algorithm (GA) and found GA generated similar K_{optm} , but runs slower.

The vertical coordinates were then adjusted by K_{optm} and the final variogram model was also fitted with weighted least squares method and is used to perform the kriging prediction. We reduced the computational burden by conducting a local kriging that uses the 100 nearest data points. Considering further points is not necessary because of the screen effect (a distant sampling point will have very small influence if there are other sampling points between it and the interpolation point) in the kriging process (Armstrong, 1998). The kriging grid cell for our case study was set at 0.25 meters in depth (y-axis) and 0.2 kilometers in distance (x-axis) and was filtered by a convex hull to avoid extrapolation as much as possible. The number of grid points in each path depends on the depth and distance range of that path (Appendix B). Overall, the kriging processes used require no manual fitting of variogram models.

To evaluate the proposed kriging interpolation, we conducted ten 3-fold cross validations for each variable in both directions (moving up or down) on every path. Each 3-fold is randomly generated. The normalized root-mean-square error (NRMSE) is calculated for each cross-validation as:

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} = \frac{1}{y_{max} - y_{min}} \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \tag{2.1}$$

where y_{max}, y_{min} are the maximum and minimum values in the upcast or downcast data. $\hat{y}_i$, y_i are the interpolated values and true values at point i, and n is the total number of points in the

validation fold. For each variable, both directions at every path, a total of 30 (3×10) NRMSE values are computed, which are then averaged as the final NRMSE to serve as the model validation metrics.

2.3.4 Spatial Statistical Analysis

Hotspot and cluster analyses were applied to the interpolated dataset to automatically highlight interesting features. This provided near-real-time identification of areas of interest for adaptive sampling.

Hotspot Analysis

Hotspot analysis was performed to identify statistically high or low value zones in the 2-D grid maps. High/low value zones (hot and cold spots, respectively) can indicate areas of high or low-level biological activity (e.g., revealed by chlorophyll concentration) or unexpected spatial patterns. To identify hot and cold spots, local G statistics were calculated for each variable of interest (e.g. temperature, chlorophyll concentration, etc.) with distance and depth (i.e., x-axis and y-axis in Figure 3) as the coordinates of each path (e.g., Path 1, 2, in Figure 2.1). The local G statistics are defined as (Ord and Getis, 1995):

$$G_i = \frac{\sum_j w_{ij}x_j - W_i\overline{x}(i)}{s(i)\sqrt{\frac{(n-1)S_{1i} - W_i^2}{n-2}}}, (j \neq i) \tag{2.2}$$

where G_i is the local G statistics for point *i* and *n* is the number of sample points. $\overline{x}(i) = \frac{\sum_j x_j}{n-1}$(j ≠ i), $W_i = \sum_{j\neq i}\omega_{ij}$, $s(i) = \sqrt{\frac{\sum_j x_j^2}{n-1} - \overline{x}(i)^2}$ are the sample mean, sum of weights (ω_{ij} is the weight between point *j* to point *i*), and sample standard deviation, respectively. $S_{1i} = \sum_j \omega_{ij}^2\ (j \neq i)$ is the sum of squared weights, excluding point i. Weights w_{ij} follow a binary coding representation with a distance bound (weights in the neighborhood as 1, outside as 0). The statistical significance criteria of local G statistics is determined by the G statistics calculated from a null hypothesis that variables are randomly distributed rather than clustered to form hotspots or coldspots. Ord and Getis (1995) provided a reference for the largest G statistics under n independent and identical distributed normal random variables. For example, with n = 1000, the 0.95 percentile of the largest G values is 3.89, meaning that if G is larger than 3.89, it is generally safe to reject the null hypothesis (independent normal distribution) with an overall probability of type 1 error as 0.05, although not guaranteed (Ord and Getis, 1995). As a result,

we present results with cutoff values as the 95th (and 5th) percentile of the G statistics. That is, we assume if the G_i statistics are higher (or lower) than 3.89 (or -3.89), then point i is identified as a hot spot (or cold spot) in this path. We used the "localG" in "spdep" package in R (Bivand et al., 2008, 2013) to perform hotspot analysis.

In the hotspot analysis, we chose a neighborhood distance bound of 0.75, which means that neighbors are defined to be within a circle of radius 0.75, with kilometer and meter as horizontal and vertical units, respectively. Approximately 30 data points are located within the neighborhood based on our grid size (0.25 meters in depth (y-axis) and 0.2 kilometers in distance (x-axis)), which is well above the minimum number of eight suggested by Griffin et al. (1996). The size of the neighborhood can be adjusted to see the high/low value patterns in a multi-scale space (Deng, 2014; Wulder and Boots, 1998). With 30 points in the neighborhood, one can focus on small-scale hotspots and coldspots to identify more localized patterns (e.g., the center of high/low values rather than patches).

Cluster Analysis

One of the research objectives is to study mixing of lake and river water, which requires first separating these two water masses that differ in water chemistry characteristics. The k-means clustering algorithm (MacQueen, 1967) was used to group similar sampling points that are indicators of each water mass based on temperature, which reflects thermal stratification, and specific conductance, which is often used as a tracer for river water (Pavlac et al., 2012).

K-means clustering requires a user-specified cluster number, which can be determined by the average silhouette width (Rousseeuw, 1987). Silhouette width indicates the extent of data compactness in the cluster and an average silhouette width larger than 0.7 typically indicates a strong clustering structure (Kononenko and Kukar, 2007). The sampling activities only took about 5 to 8 hours at each location, so the river plume dynamics can be assumed relatively steady, and the dataset can be assumed to represent a snapshot in time of the river plume. Therefore, we clustered kriging interpolation estimates from all paths at each sampling site. Clustering results were then further assessed using boxplots to visualize overall trends in the characteristics of different river and lake water mixing areas.

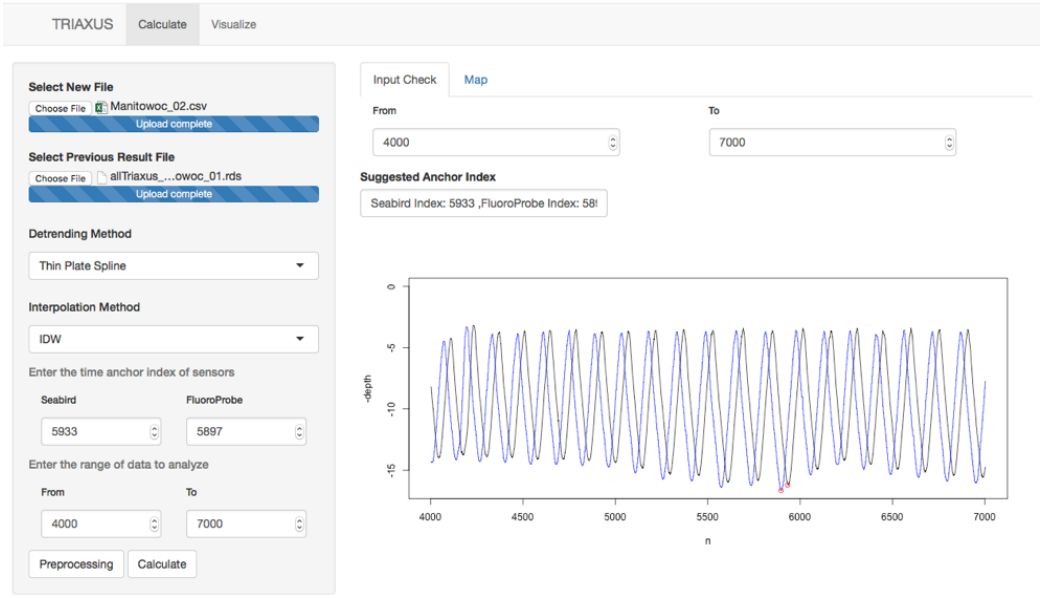
2.3.5 Web Application Development

We implemented the methods described in the previous sections in the open source R programming development environment. The "ggplot2" (Wickham, 2009) and "rgl" (Adler et al.,

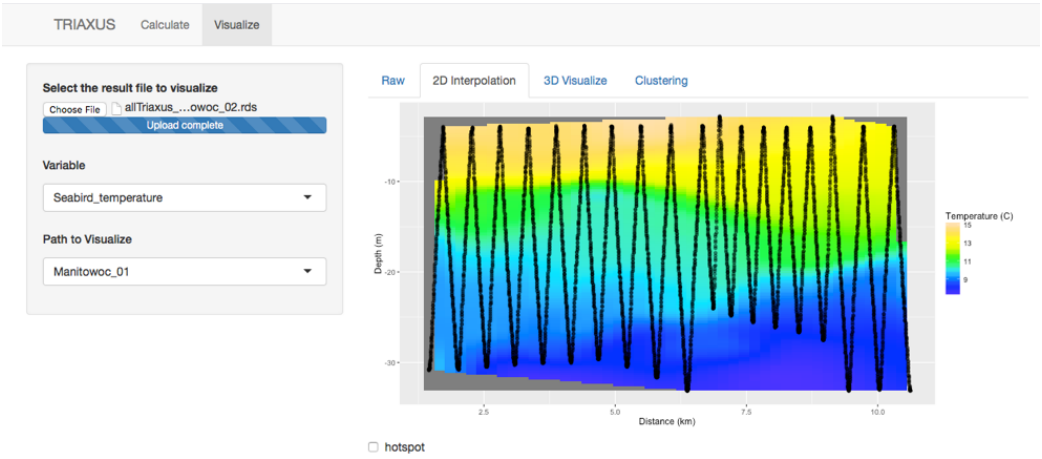
2016) packages were used to visualize results in 2-D and 3-D. We built an interactive user interface based on the "shiny" package (Chang et al., 2017), which allows deployment of R code as a Web application. The code can be found at http://stormxuwz.github.io/TUVTool/.

The Web user interface consists of two panels: a calculation panel and a visualization panel (Figure 2.5). Currently, the ship's workflow involves receiving the data stream from the undulating vehicle and manually saving the data files periodically. The user interface in the calculation panel is designed to allow users to upload the latest saved file. The app will then analyze the file and save the results in a local file, which also contains results from previous paths. It is also possible to save the results into a database system if needed. In this way, users are able to analyze previously saved data while the ship is continuing to sample water-quality data. In the visualization panel, users are able to visualize the raw sampled data, kriged data (2-D or 3-D), hotspot data (2-D or 3-D) and cluster data (3-D). Users can choose which paths and variables to visualize, as well as the number of clusters and variables to cluster for the cluster analysis.

To help compute correct timestamps to synchronize data, as discussed in Section 2.3.2, we also provide a simple algorithm in the calculation panel that finds the data index of maximum depths for both sensors (Figure 2.5a). The two indices usually represent the same valley (i.e., deepest) location in the undulating path so that the timestamps of multiple sensors (e.g., $t_{seabird}$, $t_{fluoroprobe}$ in Section 2.3.2) can be automatically determined. In some rare cases where the two indices represent different initial valley locations, users need to try another data range so that the program will recalculate the maximum depth points and find the correct indices.



(a) Calculation Panel



(b) Visualization Panel

Figure 2.5. Application user interface of the Shiny Web Application. Calculation Panel (a) is to input the raw data and execute the analysis. Visualization Panel (b) is to visualize the results.

2.4 Results

2.4.1 Kriging Interpolation

The boxplot of the NRMSEs for each variable at each rivermouth shows that most of the NRMSEs are within 10% and the highest error is under 15% (Figure 2.6). The errors of zooplankton density and biomass were higher than other variables because the zooplankton density and biomass are noisier in the sampled data (e.g. Figure 2.3b). Most of the outliers are from paths that were close to the rivermouth, where the water was not stable due to interactions with river plume water and wind disturbances.

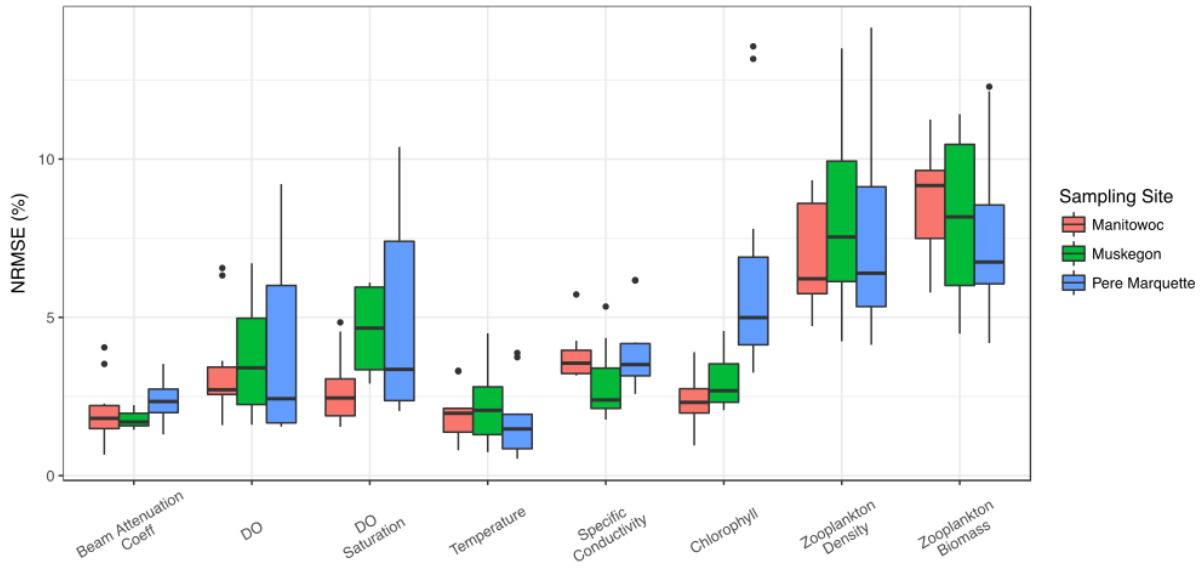
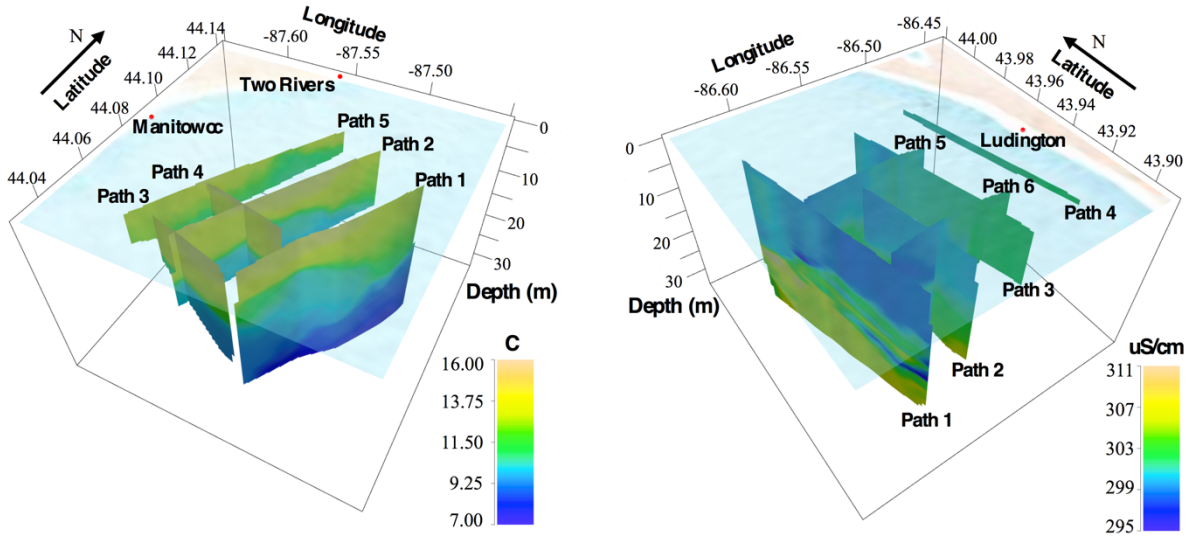


Figure 2.6. Boxplots of NRMSE of 3-fold cross validation for each variable at each site. Each box contains NRMSE of upcasts and downcasts from all paths in the site.

For each variable, visualization of kriging interpolations in 3-D geographical maps provides direct and intuitive understanding of feature distributions during the adaptive monitoring. Using temperature at the Manitowoc site as an example (Figure 2.7a), we can detect distinct thermoclines where temperature changed rapidly with depth. For specific conductance at the Pere Marquette site (Figure 2.7b), higher values were found in deeper areas, which may indicate an accumulation of solutes near the lake substrate.

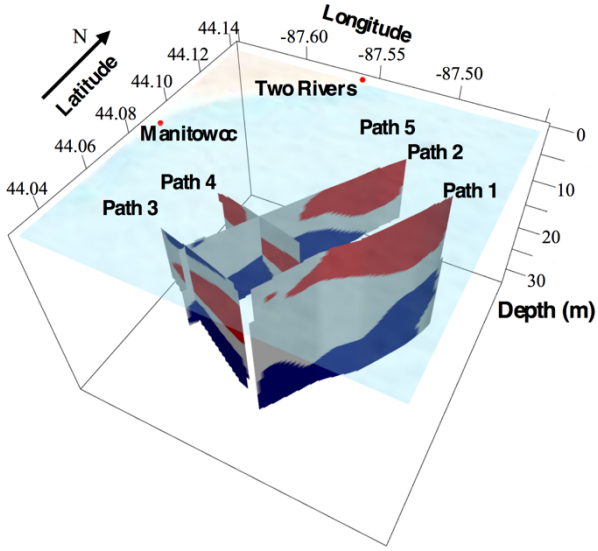


(a) Temperature at the Manitowoc site (b) Specific conductance at the Pere Marquette site

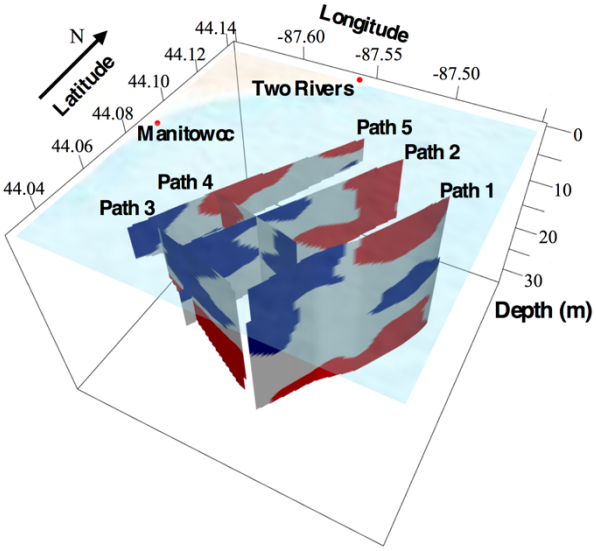
Figure 2.7. Examples of 3-D direct visualization of kriging results for two parameters at two of the sites.

2.4.2 Hotspot Analysis

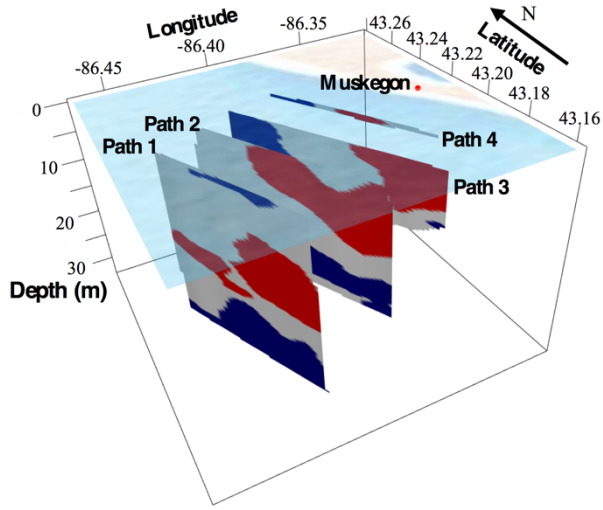
Local G values highlight the high (hot spot) and low (cold spot) values of a particular parameter in one path. Aggregating the results of hotspot analysis can reveal location changes of local high/low value clusters at different distances from the river mouth, from which the plume dynamics can be inferred. For example, the spatial distributions of chlorophyll concentration (Figure 2.8a, c, e) and specific conductance (Figure 2.8b, d, f) at each site are characterized below. Chlorophyll reflects river plume phytoplankton density and specific conductance is regarded as a common tracer of river water.



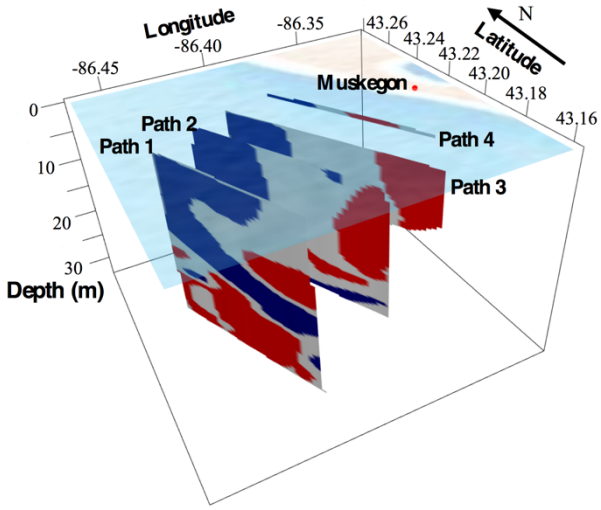
(a) Manitowoc total chlorophyll concentration



(b) Manitowoc specific conductance



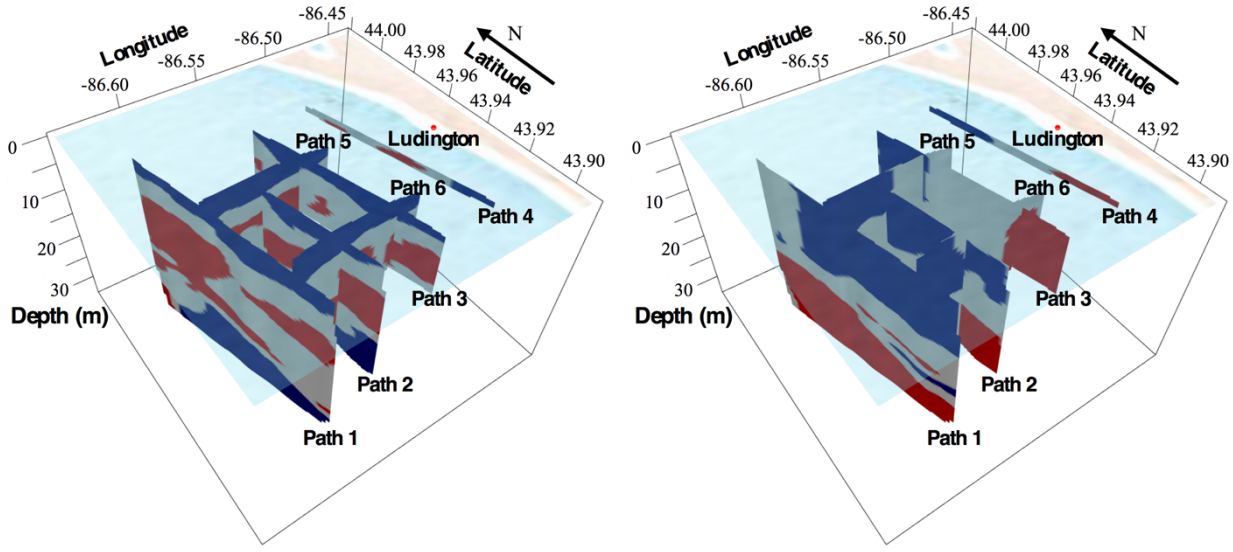
(c) Muskegon total chlorophyll concentration



(d) Muskegon specific conductance

Figure 2.8. Results of hotspot analysis, showing low-value areas in blue and high-value areas in red. White indicates areas with values near the mean.

Figure 2.8 (cont.)



(e) Pere Marquette total chlorophyll concentration

(f) Pere Marquette specific conductance

At the Manitowoc site (Figure 2.8a, b), high concentrations of phytoplankton and solutes were measured in the northern nearshore surface area (right side of Paths 1 and 2). Two hot spots of specific conductance in Path 3 (Figure 2.8b) could be the plumes from the two corresponding river mouths. High specific conductance expanded to the northern area of Paths 1 and 2, which indicated a river plume flowing northward within the sampling area (toward the right in Figure 2.8b). In Paths 1 and 2, high chlorophyll concentrations were detected on the surface in association with the observed river plume, while in Paths 3 and 4, higher chlorophyll concentrations were observed in the middle layer, different from Path 1 and 2. (Note that the Manitowoc site did not have chlorophyll data in Path 5).

At the Muskegon site (Figure 2.8c, d), high chlorophyll concentrations (Figure 2.8c) descended gradually from the surface to middle layers as the distance from the river mouth increase. The surface hotspot of specific conductance in Path 4 (Figure 2.8d) represented a river plume that continued toward the south and sank away from shore. Farther from shore, solutes accumulated in a large zone extended along most of the path, but was separated from the more recent, shallower plume by a cold spot.

A different spatial pattern was observed at the Pere Marquette site (Figure 2.8e, f).

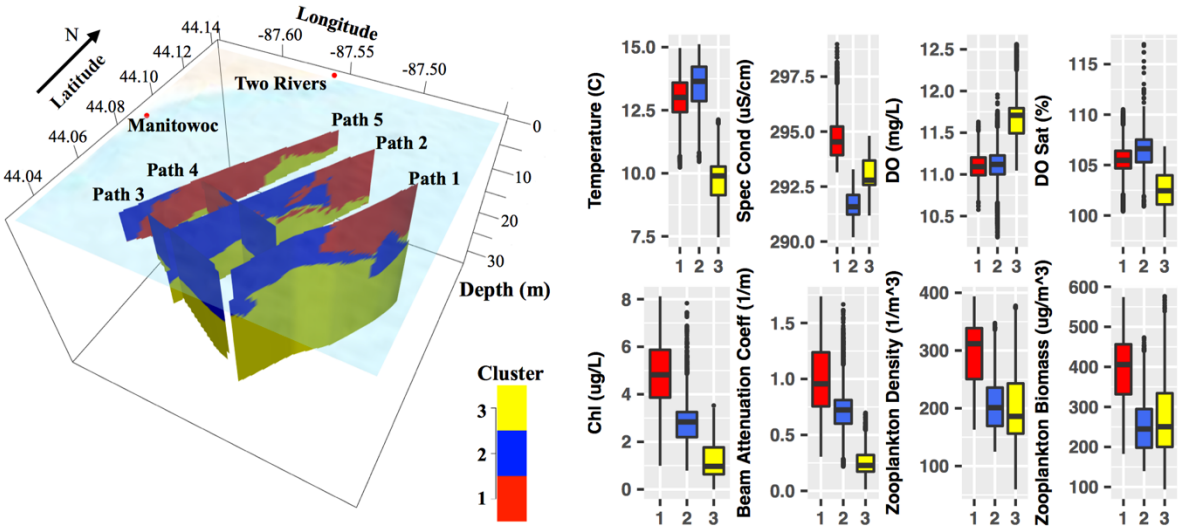
Chlorophyll concentrations were always higher in the middle of the water column for all paths. Specific conductance was higher at the bottom portions of Paths 1 and 2, which may indicate former river water sinking to the bottom. Higher specific conductance values also were observed at all depths along the southern extent of Paths 3 and 4 (toward the right in Figure 2.8f).

2.4.3 K-Means Clustering

K-means clustering was applied to the specific conductivity and temperature features, and clusters were used as indicators of river water. We calculated average silhouette widths (Table 2.1) and chose the number of clusters that had the highest average silhouette width for each site. As a result, the Manitowoc, Muskegon and Pere Marquette site were divided into three, three, and two clusters, respectively (Figures 2.9 to 2.11). This type of clustering can provide a quick analysis of mixing structure in the data, which can assist users on the ship in identifying where river plumes are and how they differ.

Table 2.1. Average silhouette widths of different cluster numbers. The highest silhouette widths for each site (bolded) were used to define the cluster number.

Clustering Number	2	3	4	5	6
Manitowoc	0.553	**0.705**	0.680	0.691	0.695
Muskegon	0.726	**0.810**	0.617	0.665	0.723
Pere Marquette	**0.862**	0.609	0.626	0.648	0.604

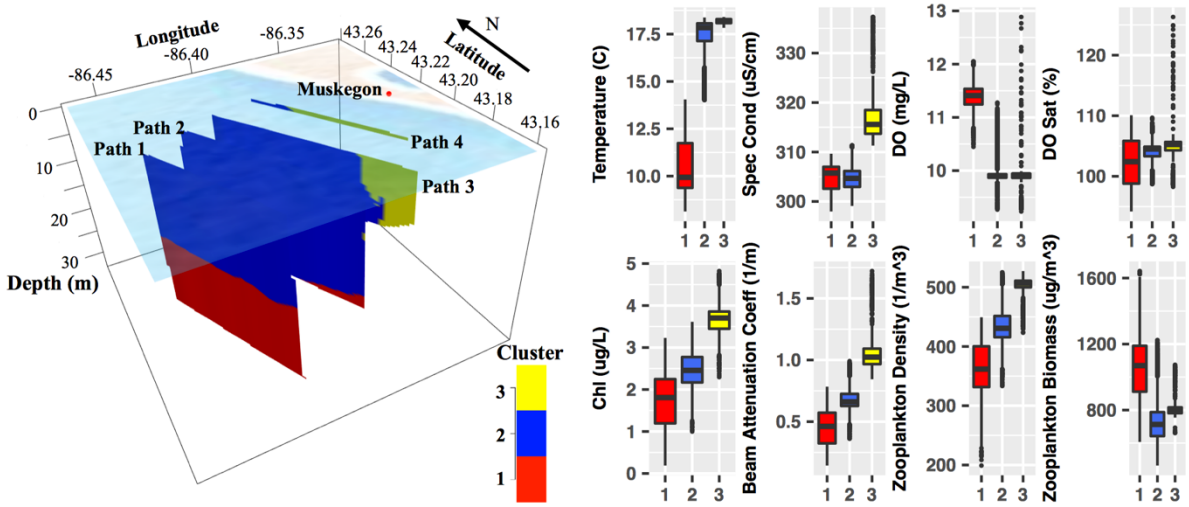


(a) Cluster distribution (b) Boxplot of water quality data

Figure 2.9. Cluster analysis results for the Manitowoc River site. The box in the boxplot shows the 1st (Q1) and 3rd quantile (Q3) and the black line shows the median value. The outliers are the values beyond Q3+1.5IQR or Q1-1.5IQR, where the interquartile range IQR = Q3-Q1.

The Manitowoc site (Figure 2.9) was divided into the surface zone on the north (Cluster 1, in red), a surface zone on the south (Cluster 2, in blue) and a bottom zone (Cluster 3, in yellow). From the boxplot results and geographic locations, Cluster 1, characterized by high temperature, specific conductance, chlorophyll, zooplankton biomass and density, and thus high biological activity, likely represented a distance river plume. Cluster 2 likely represented lake surface water or epilimnion (high temperature and low specific conductivity with low chlorophyll, zooplankton density, and zooplankton biomass). Cluster 3 mostly contained lake bottom water or hypolimnion (low temperature, medium specific conductance, high dissolved oxygen, low chlorophyll, and low density and biomass of zooplankton). The clustering results are consistent with the specific conductance hotspot analysis (Figure 2.8b), indicating that river water flowed northward from both river mouths. To check whether the differences between each group were significant, we used (a) ANOVA followed by Tukey's honest significant difference (HSD) test (Haynes, 2013) and (2) a nonparametric test, Kruskal-Wallis test (Kruskal and Wallis, 1952) followed by pairwise Wilcoxon rank sum test (Hollander et al., 2015). The second approach was selected because our data may not follow the normal distribution with homogeneity of variances, which is required for ANOVA. Both ANOVA and Kruskal-Wallis

test showed that significant differences existed in the groups ($p < 2.2e{-}16$). Tukey's HSD test and pairwise Wilcoxon rank sum test agreed that each pairwise group had significant differences ($p < 1e - 7$) for each variable except Cluster 1 and 2 for DO ($p = 0.93$ in Tukey's HSD but $p = 0.0003942$ in Wilcoxon rank sum).

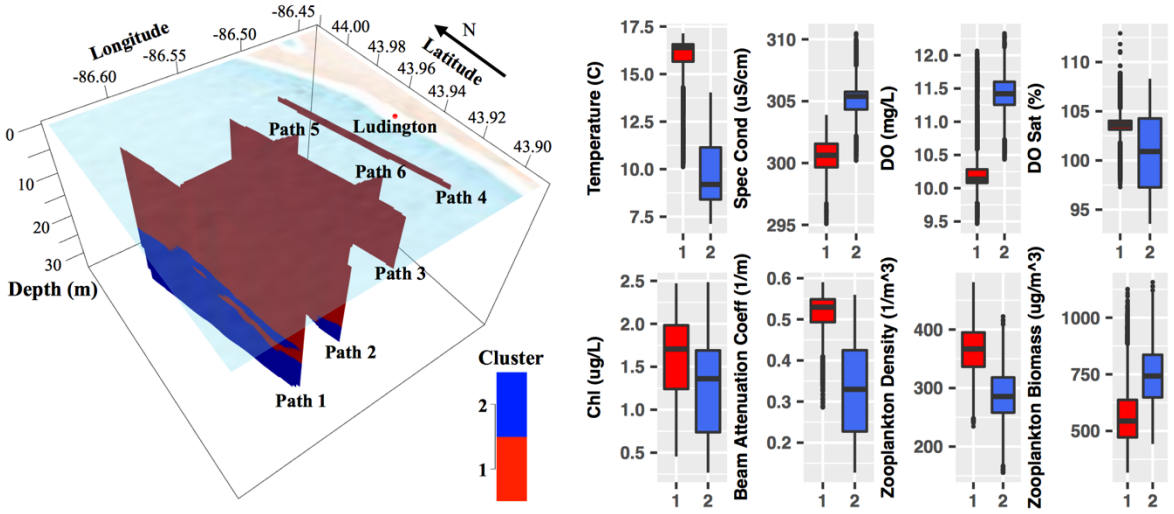


(a) Cluster distribution (b) Boxplot of water quality data

Figure 2.10. Muskegon cluster analysis results (The box in the boxplot shows the 1st (Q1) and 3rd quantile (Q3) and the black line in the middle shows the median value. The outliers are the values beyond Q3+1.5IQR or Q1-1.5IQR, where the interquartile range IQR = Q3-Q1.

The Muskegon site was divided into three clusters (Figure 2.10), similar to the Manitowoc site. River water flowing toward the south from the river mouth was evident in Cluster 3, which had high temperature and high specific conductance, as well as high chlorophyll concentrations. The blue zone (Cluster 2) was lake surface water, which had high temperature but low specific conductance. The red zone (Cluster 1), with low temperature and low specific conductance, represented lake bottom water. However, unlike the Manitowoc site, the Muskegon site had low zooplankton biomass but high zooplankton density in the river plume (Cluster 3, in yellow), indicating a zooplankton community characterized by a smaller size structure. The bottom water (Cluster 1, in red), conversely, tended to have a lower density but higher biomass in the zooplankton community, indicating larger zooplankton size structure. Again, both ANOVA and Kruskal-Wallis test showed significant differences existed in the groups ($p < 2.2e-16$). Tukey's HSD test and pairwise Wilcoxon rank sum test agreed on each pairwise group has

significant differences (p<2.2e-16) for each variable except for Cluster 3 and 2 for DO (p = 0.99 in Tukey's HSD test but p = 0.00073 in Wilcoxon rank sum) and Cluster 1 and 2 for specific conductance (p = 0.076 in Tukey's HSD but p = 1.043e-15 in Wilcoxon rank sum test).



(a) Cluster distribution

(b) Boxplot of water quality variables

Figure 2.11. Pere Marquette cluster analysis results (The box in the boxplot shows the 1st (Q1) and 3rd quantile (Q3) and the black line in the middle shows the median value. The outliers are the values beyond Q3+1.5IQR or Q1-1.5IQR, where the interquartile range IQR = Q3-Q1.

The Pere Marquette site was divided into two clusters (Figure 2.11), the surface zone (Cluster 1, in red) and the bottom zone (Cluster 2, in blue). The bottom zone had low temperature but high specific conductivity. This structure was quite different from the other two sites. Zooplankton biomass was higher in the red cluster, but density was higher in the blue cluster. Although specific conductance was relatively low in the surface zone of Paths 1 and 2 as shown in the direct visualization (blue, Figure 2.7b) and hotspot analysis (Figure 2.8f), temperature had more substantial stratification patterns. Therefore, the cluster was influenced strongly by temperature distributions. With only two clusters chosen by the highest average silhouette width, the analysis could not discern river versus lake water. Thus, the number of clusters evident in the sampling data can be used to determine whether evident river plumes exist at a given site. Both ANOVA and Kruskal-Wallis test showed significant differences exist in the two groups ($p < 2.2e-16$)

2.4.4 Limnological Inference - River Plume Dynamics

Combining hotspot analysis and cluster analysis with 3-D parameter visualizations using kriging helps identify river plumes and provides a more comprehensive understanding of river plume dynamics. At the Manitowoc site, the river plumes extended lakeward, with relatively high specific conductance serving as the primary tracer in the plume as supported by hotspot analysis (Figure 2.8b) and cluster analysis (Figure 2.9). At the Muskegon site, a river plume was evident until Path 3 in the cluster analysis (Figure 2.10). Further from shore at the Muskegon site, cluster analysis failed to identify the river plume. However, hotspot analysis (Figure 2.8d) shows the river plume might not have completely mixed as apparent river water is identified further into the lake as a mid-depth intrusion, as indicated in the relatively low/high distribution pattern of specific conductance and chlorophyll concentrations from the hotspot analysis. At the Pere Marquette site (Figure 2.8f), no river plume was detected during the sampling time; high specific conductance existed in the bottom area, which may indicate former river water sinking to the bottom.

To further assess these findings, we compared the plume dynamics with wind data from NOAA stations (Station MKGM4 at Muskegon site and Station LDTM4 at Pere Marquette site; no wind data were found for Manitowoc site in 2011). During the sampling period, the wind direction (i.e. where the wind was originating) at the Muskegon site changed from south (170 degrees, clock- wise from true N) to southwest (220 degrees) with a speed around 6.8 m/s. The direction seems inconsistent with the data showing that the river plume extended a short distance toward the south. The nearshore surface current toward the south might be related to internal longshore currents, which may be raised by previous wind gust (Ahmed et al., 2014). Additional data are needed for further analysis. The Pere Marquette site had wind directions changing from west (270 degrees) to northwest (330 degrees) then back to west (270 degrees), with a speed around 3.5 m/s, which may have pushed the water towards the shoreline. The lack of a typical river plume cluster may be due to the wind reversing the flow or simply low flows coming from the river during the sampling time.

The chlorophyll distribution was site-specific in this study and was largely influenced by river plumes. At the Manitowoc site (Figure 2.8a), for the area where river plumes did not intersect with the sampling path (Path 3), relatively high chlorophyll layers existed at mid-depth just above the thermocline rather than on the plume surface. The Muskegon site (Figure 2.8c)

had high chlorophyll concentrations near the shore, then decreased in surface waters away from the shore, and, finally, a higher value zone in the middle depth layers further away from the plume. Previous research (Lunven et al., 2005) showed similar patterns, which could be due to surface nutrient limitation, further indicating a nutrient subsidy from the river plume into the lake. The Pere Marquette (Figure 2.8e) site had low chlorophyll concentrations on the surface for all paths, which served as further evidence that the river plume was not observed in the lake at the Pere Marquette site during the sampling period.

2.5 Discussion

The framework developed in this work supports near-real-time data analysis and decision support via Web applications. The R Shiny package builds a Web application so that the computational power is reduced by performing all analysis on a remote server on the ship or, if Internet connections allow, on shore. On a laptop with Intel i7 2.3GHz, 8G memory, the analysis of each path was finished within minutes. In addition, with parallel computing of each water variable, these steps could be completed even faster. Therefore, the framework can rapidly analyze a saved data file while the ship is continuing to sample water chemistry.

The methods developed in this research contribute to adaptive decision support and sampling, enabling data collection personnel to respond to current conditions and collect more informative data while the ship is still on site. Adapting sampling paths will maximize the yield of monitoring activities for three reasons. First, using automated data QA/QC procedures, kriging interpolation and visualization, especially with highly interactive 3-D displays, means that sensor or data transmission problems can be detected rapidly and fixed prior to further data collection. Second, researchers can interpret and analyze parameter distributions and identify key patterns and processes while monitoring activities are underway, hence guiding the sampling activities to focus on the most important areas for further monitoring activities. Third, hotspot and cluster analysis highlight the difficult to discern river plume dynamics and interesting water-chemistry gradients where additional samples may be collected. Cluster analysis identifies the boundaries between river and lake water and provides useful statistics describing the attributes of each zone. Lastly, the hotspot analysis also reveals some patterns that are not easy to explain. For example, an isolated cold spot in specific conductance is noticeable in Path 1 at the Manitowoc site (Figure 2.8b). In addition, water with low specific conductance intruded into the high

specific conductance areas at the Muskegon site (Figure 2.8d). Lake seiche oscillations could potentially explain these phenomena, but further physical samples and continuous monitoring activities are needed to reach firm conclusions on the causes of this pattern. These methods can guide additional data collection to improve plume characterization and understanding of water mixing processes in rivermouth areas.

2.6 Conclusions

This chapter proposed and implemented techniques that can be applied concurrently with undulating sensor data collection to facilitate an adaptive monitoring program to rapidly determine where further data should be collected while surveys crews are onsite. Such a program should improve the benefit-to-cost ratio of monitoring programs and lead to better understanding of lake dynamics. Designed specifically for undulating sampling strategies implemented with gridded survey patterns, the framework reveals useful insights on the plume dynamics of the three rivermouths in Lake Michigan. The cross-validation of automated kriging routines show NRMSE are around 10% across our data sets. Direct 3-D visualization provides a comprehensive view of each water-quality parameter and hotspot analysis reveals more details on the spatial trends and dynamics of water-quality parameters such as chlorophyll concentration and specific conductance. Cluster analysis can delineate the boundaries of river water and lake water masses and provide descriptive statistics of other variables collected simultaneously, highlighting water-quality variability in different water bodies. The analysis framework is incorporated into an interactive Web application that allows researchers to obtain a comprehensive understanding of river plume dynamics during sampling activities, enabling a more adaptive approach.

Future work is given in Chapter 5.

CHAPTER 3: ALGORITHMIC CHARACTERIZATION OF THERMOCLINE AND DEEP CHLOROPHYLL LAYERS FROM DEPTH PROFILING WATER QUALITY DATA

In this chapter, the research focus moves from nearshore to offshore areas and another pattern detection approach is implemented to detect unusual patterns in the depth profiling data, particularly lake summer stratification and deep chlorophyll layers (DCL). Using piecewise linear segmentation and peak detection algorithms, we can automatically identify lake stratification and DCL patterns from the sampling data and detect anomalous profile shapes.

Section 3.1 describes the related lake process and current sampling processes for this data type. Section 3.2 explains the data sources used in this work. The methodology section (Section 3.3) describes the algorithms in detail. In the results section (Section 3.4), we apply our algorithms on data sampled from the Great Lakes. We first compare the two thermocline detection algorithms, Piecewise Linear Representations (PLR) and Maximum Gradient Hidden Markov Model model (MG-HMM). We then validate the better performing algorithm, PLR, against operators' historical notes and analyze profile shapes. In Section 3.5, we present a case study that uses the algorithms to reveal the spatiotemporal trends of lake stratifications and DCL in Lake Superior. This is followed by the discussions and conclusions in Section 3.6.

3.1 Introduction

CTD (Conductivity, Temperature, Depth) profilers are widely used to monitor the vertical distribution of water quality in lakes and oceans. Depth profiling data provide insights on key lake features such as lake thermo-stratification and deep chlorophyll layers (DCL). Background on these processes and approaches to detecting DCLs are given below.

3.1.1 Lake Stratification Processes

Thermal stratification occurs in summer lakes as the surface water are heated up and exceed 4 °C (the temperature for water to be at maximum density), so that it is no longer mixed with deep water. The lighter and warmer surface water stays on top, forming an "epilimnion" layer, and heavier water stays in the deep lakes and forms a "hypolimnion" layer. The transition zone between the epilimnion and hypolimnion is the metalimnion and the horizontal plane within the metalimnion that has the sharpest temperature changes is called the thermocline. In the

epilimnion, the water body is fully mixed mainly due to the breaking of wind-induced waves, while in the hypolimnion, the cold and heavy water is usually assumed static. In most of the lakes in the northern hemisphere, the thermocline is shallower in offshore areas than nearshore areas (i.e. dome shape). However, Lake Erie has a bow-shaped thermocline while offshore areas usually have a deeper thermocline, which is due to anticyclonic vorticity in the surface winds (Beletsky et al., 2012).

The wind is the most important factor that influences the depth of the thermocline (Boehrer and Schultze, 2008). The larger wind stress generated by strong wind will lead to a deeper thermocline, with more mechanical energy used to mix the lake. On the other hand, the lake will be stable with increased buoyancy strength due to a higher surface temperature that needs more mechanical energy to modify or destroy (Austin and Colman, 2007; Gorham and Boyce, 1989). In addition, the thermocline depth can be tilted by internal seiche events (warmer surface water is pushed downwind while the cooler water below the thermocline flows upwind), a process that may be affected by the Earth's rotation (Gorham and Boyce, 1989). For example, baroclinic motion generated a thermocline depth oscillation with a period of 17h in Lake Erie (Bouffard et al., 2012).

The surface temperature also plays a role in stratification as it mostly influences the water density that ultimately causes the lake to stratify. The surface temperature is controlled by solar radiation, cloud cover, wind-driven mixing, water clarity, ice cover, and lake bathymetry (Moukomla and Blanken, 2016). The energy or radiation that a water body absorbs can be reduced by more ice cover, leading to a later onset of stratification (Austin and Colman, 2007).

3.1.2 Deep Chlorophyll Layers

Deep chlorophyll layers (DCL), deep chlorophyll maximum (DCM) (White and Matsumoto, 2012; Camacho, 2006), or subsurface chlorophyll maximum (SCM) (Gong et al., 2015) are water layers with high chlorophyll concentrations that lie below the thermocline in stratified lakes or oceans. For example, in the Great Lakes, DCL is observed during the summer-stratification period and dissipates in late August due to deeper thermocline and unstable metalimnion (Watkins et al., 2015). In addition to laboratory analysis on bottle samples, measurement of chlorophyll concentration is substantially supplemented by the measurement of in-vivo fluorescence since Lorenzen (1966). Yet using fluorescence to deduce chlorophyll concentration may give large deviations as the ratio between fluorescence and chlorophyll

concentration varies (Cullen, 1982, 2015).

A significant portion of the lake primary production is determined by the DCL, where the highest chlorophyll concentrations exist. For example, in Great Lakes systems, studies have shown that the chlorophyll concentration in the DCL is 1.5 to 2.5 times the concentration in the epilimnion in Lake Superior (Barbiero and Tuchman, 2004) from 1996 to 2001 and 1.8 to 5.7 times in Lake Michigan from 1982 to 1984 (Fahnenstiel and Scavia, 1987). In Lake Michigan, DCL is responsible for 30 to 60 percentiles of areal primary production in Lake Michigan (Watkins et al., 2015). Furthermore, the phytoplankton community structures at the DCL are different from those in the epilimnion in Lake Superior (Barbiero and Tuchman, 2004). The DCL also affects energy and material transfer as it shifts the location and extent of food sources for grazers (Gong et al., 2015).

Many physical and biological factors influence the DCL formation and maintenance. Phytoplankton needs nutrients and light to grow. With deeper depth, light resources are reduced and DCL must exist above the euphotic depth with about 1% of surface illumination, so that photosynthesis just balances cellular respiration. On the other hand, more nutrients exist due to upward diffusion in deeper water. As a result, algae need to find the best location to compete for these two resources (Klausmeier and Litchman, 2001). Other influencing factors include: (a) lake stratification and surface water (Barbiero and Tuchman (2004), (b) the upper mixed water layers (Ryabov et al. ,2010), (c) phytoplankton mobility (Cullen, 2015) and phytoplankton photoacclimation or photoadaptation (phytoplankton increases the chlorophyll level when in a deficit of light) (Barbiero and Tuchman, 2004; White and Matsumoto, 2012; Watkins et al., 2015), and (d) zooplankton grazing (Watkins et al., 2015) and zooplankton excretion (Oliver et al., 2014).

Besides numerical models that have been developed to simulate DCL formations (Klausmeier and Litchman, 2001; White and Matsumoto, 2012; Ryabov et al., 2010; Mellard et al., 2011), several statistical or machine learning models have been used to study the DCL patterns. Richardson et al. (2002) used self-organizing maps (SOM) to cluster coastal chlorophyll-a profiles. Longhi and Beisner (2009) performed linear analysis between DCM depth and several environmental factors such as surface temperature and total phosphorus. Sauzède et al. (2015b) built the first database of in-situ fluorescence profiles for global oceans and also predicted chlorophyll concentrations from fluorescence profiles using artificial neural networks

(Sauzède et al., 2015a).

3.1.3 Current and Proposed DCL Detection Approaches

During sampling activities in the Great Lakes, USEPA operators manually view downcast profiling (i.e., CTD data sampled as the profiler descends to the lake bottom) to identify the depths of the thermocline, the boundaries among epilimnion, metalimnion, and hypolimnion, and deep chlorophyll layers. Next, bottle samples are taken at these depths during upcast profiling (i.e., as the CTD profiler ascends to the lake surface). However, manual identification of these features has the following drawbacks: (1) Each operator applies subjective criteria so that the depth recorded is not standardized and may not be comparable across operators; (2) As more and more data are recorded, it becomes more and more difficult to check the historical records to find mislabeled or missing judgements. Although researchers can manually relabel the depths of all profiles after collection, this is a time-consuming process and the subjective criteria are still problematic. Thus, historically sampled data may remain buried in databases and not adequately utilized to guide future sampling activities.

Automated feature identification provides a potential solution to these challenges. To automatically identify lake stratification patterns, the simplest approach is called the isotherm approach. In this approach, the thermocline depth is defined as the depth with a pre-defined temperature, which can be fixed dependent on locations (Wang et al. 2000) or profiles (Fiedler, 2010). Previous research also identified thermoclines using temperature gradients such as thermocline strength index (TSI) (Yu Hui, 2010) or indices based on density gradients such as Relative Thermal Resistance to Mixing (RTRM) (Hampton et al., 2014). The depth with maximum temperature gradient or maximum RTRM indicates the thermocline location. These methods are adequate for bottle-sampled data that are accurate and without much noise. However, with high-frequency sampled sensor data, the depth interval can be quite small and sampling noise or local fluctuations can lead to significant gradients that can bias the thermocline location. Smoothing and aggregating the data can reduce the effects of noise but may introduce errors from the choice of smoothing extent. Furthermore, these methods can only capture the location of the thermocline and are not able to characterize full stratification patterns such as the locations of the epilimnion and hypolimnion.

Another approach is to approximate the temperature profiles using piecewise linear segments. This algorithm, called piecewise linear representation (PLR), has three approaches

(Keogh et al., 2004): a) bottom-up (merging small segments to generate a larger segment), b) top-down (partitioning large segments into smaller segments), and c) sliding windows (building segments by adding points until approximation errors are too large). Thomson and Fine (2003) used split and merge algorithms developed by Pavlidis and Horowitz (1974), which is a variation of the top-down approach, to approximate density CTD profiles and identify mixed layers in ocean CTD profiles. Fiedler (2010) compared this algorithm with other thermocline detection algorithms mentioned above in oceans. Yet, the PLR method hasn't been extended to detect all features of lake stratifications (i.e. epilimnion, metalimnion, and hypolimnion).

For DCL identification, merely setting the maximum values as the location of the DCL is not reasonable when the fluorescence profile doesn't have a peak, or multiple peaks exist such as a bimodal distribution (Lips and Lips, 2014; Mellard et al., 2011). Researchers have previously approximated chlorophyll depth profiles using a Gaussian distribution shape (Abbott et al., 1984; Richardson et al., 2002; Gong et al., 2015). The Gaussian model assumes DCL shapes are symmetrical and have a constant chlorophyll concentration above and below the DCL, referred to as the background concentration. Uitz et al. (2006) slightly modified the Gaussian bell shape model by incorporating a linear term so that the background concentration can change below the DCL. However, in some systems such as the Great Lakes, a fluorescence profile may have an asymmetrical shape (examples are given in Section 3.4.3). Therefore, the Gaussian shape assumption is not always applicable.

In this study, we extended the PLR characterization method to not only detect the location of the thermocline but also study the shapes of temperature profiles to identify epilimnion and hypolimnion and other patterns such as double thermoclines in a robust way. We also compare this approach with another common segmentation method, hidden Markov models (HMM), which assume that the hidden states of observed data follow a Markov process. To analyze the fluorescence profiles and identify the characteristics of the DCL, including location, thickness, and intensity of chlorophyll concentration (Gong et al., 2015; Beckmann and Hense, 2007), we propose a peak detection algorithm based on the gradients of fluorescence concentrations first. Then we fit the peak with two half-Gaussian shapes to overcome the asymmetry of the peak shape, which is more flexible than fitting a single Gaussian shape. More details about the methodology are given in Section 3.3.

3.2 Data Sources

The algorithms developed in this work are tested using data collected by the EPA Great Lakes National Program Office in the Great Lakes. A Sea-Bird® CTD profiler measures water quality variables as it travels through the water. EPA uses Sea-Bird to generate depth profiles at fixed geo-locations in each lake (Figure 3.1) once per year. The operators determine the thermocline, epilimnion, hypolimnion and deep chlorophyll layers by visually assessing the raw data. The data analyzed in this study are sampled from 1996 to 2013, with 1665 profiles in total. Each depth profile contains data on temperature, dissolved oxygen, beam attenuation coefficients, specific conductivity, pH, and fluorescence. Fluorescence is widely used as a proxy for chlorophyll concentration (Lorenzen, 1966). Therefore, the peak of fluorescence concentrations is assumed to be the location of the DCL. We use the downcast part of the depth profiles for detection, as the EPA operators do, because the data measured during the upcast process are already disturbed by the downcast movement of the CTD.

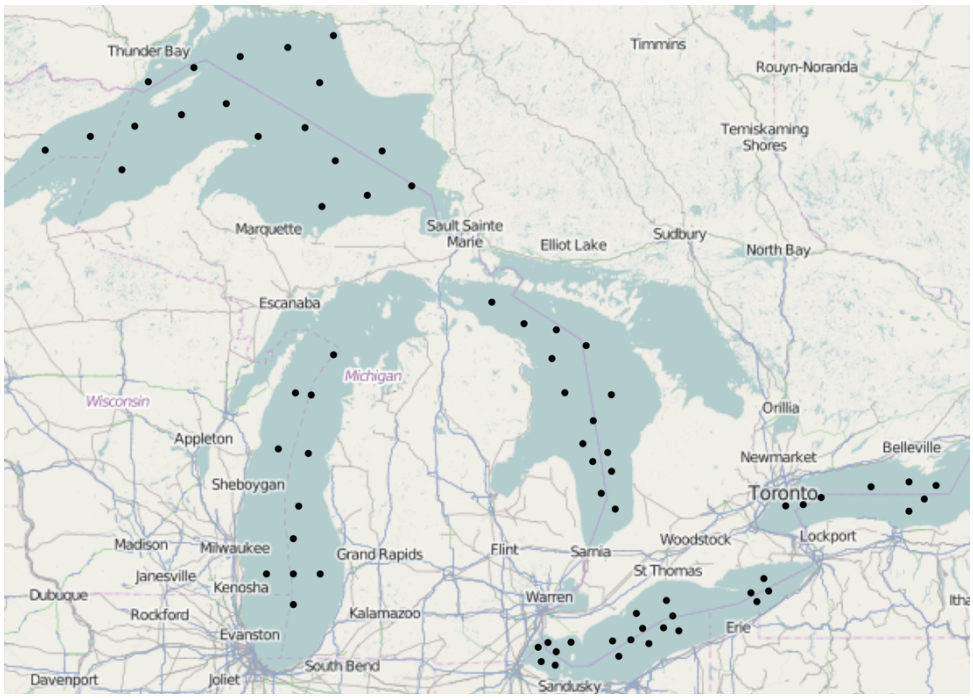


Figure 3.1. SeaBird CTD sampling locations in the Great Lakes

3.3 Methodology

The algorithm workflow is shown in Figure 3.2. Data preprocessing is first conducted to remove inaccurate data, aggregate data (i.e. standardize sampling depth intervals) and smooth

fluctuations by computing moving averages (Section 3.3.1). Then the algorithms detect lake stratification patterns using two time-series segmentation algorithms (Section 3.3.2), including piecewise linear segmentation (PLR) and hidden Markov model (HMM). Finally, the DCL is identified by detecting peaks of fluorescence concentrations (Section 3.3.3).

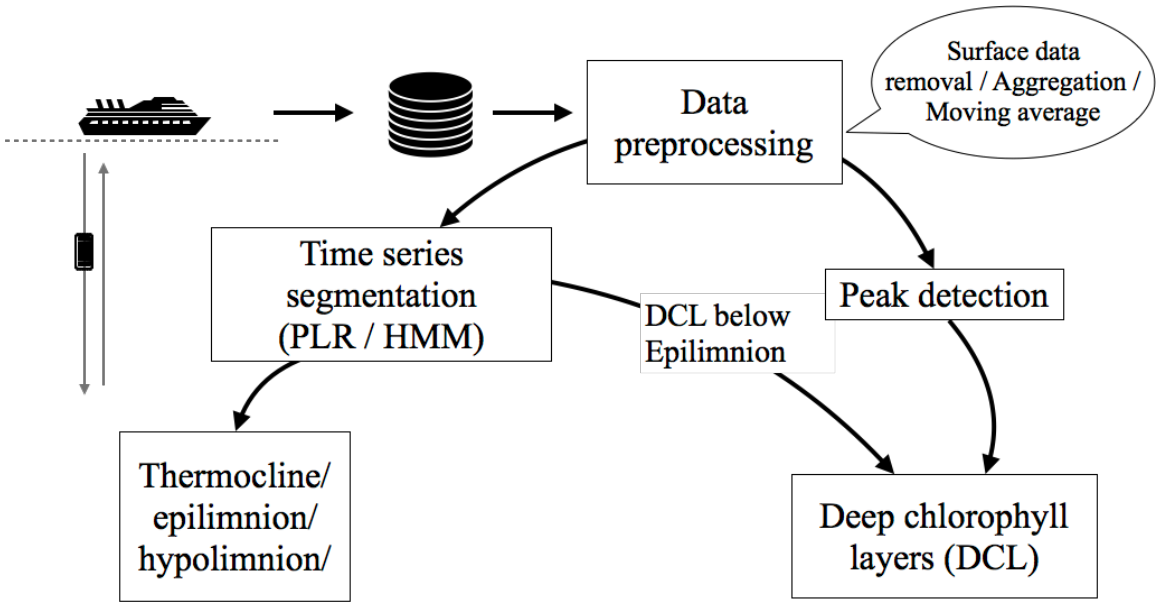


Figure 3.2. Algorithm workflow to detect lake stratification and deep chlorophyll layers

3.3.1 Data Preprocessing

Signal data contain considerable noise that can have significant effects on the results of the signal processing algorithm, including the Piecewise Linear Representation (PLR) algorithm used in this work (Keogh et al., 2004). To filter the noise, we first removed the lake surface data (depth less than 3 meters) because these data usually have many spikes and much noise. Since the raw data have inconsistent and small sampling depth intervals, we also used an averaging process to smooth the data at depths with a consistent interval h, which is defined as 0.25m, half of that in current EPA protocols, to characterize the profiles in more detail. That is, the data at depth D will be the mean values of data in $[D - h, D + h]$. For example, the water quality data at 1 meters deep will be the mean values of data from 0.75 m to 1.25 m depths. For profiles that contain a minimum depth interval larger than $2h$ (e.g., some profiles are already aggregated), we used linear interpolation to interpolate the data.

A moving average filter with Hann window (Oppenheim et.al., 1999) is then applied to the temperature and fluorescence profiles. The window size is 2 meters, meaning the data of 1

meter above and below the depth being estimated are used in the smoothing and the weights follow a Hann function. The size of this smoothing window was chosen by trial and error to avoid excessive distortion in the profile data. A moving average approach was selected because it performed better than other smoothing algorithms, such as Fourier transformation or wavelet transformation, on the profile data. Fourier transformation requires a threshold to remove low energy frequencies. Such a threshold is not easy to determine and has no clear physical meaning for limnological data. For the wavelet algorithm, the tuning threshold parameter is the decomposition level, which has only a few alternative values to choose among. However, according to our tests, the wavelet smoothing algorithm can significantly change the depth of the maximum peak so the peak detected after smoothing may not be the depth with the highest fluorescence in the raw data. Therefore, we chose the simplest approach of the moving average smoothing algorithm.

3.3.2 Lake Stratification Detection

We implemented two lake stratification detection algorithms based on (a) piecewise linear representation (PLR) and (b) maximum gradient with a hidden markov model (MG-HMM).

PLR algorithm

The PLR approach approximates the profiling data by several linear segments. We implemented the bottom-up approach, which has performed better than the top-down approach for various datasets (Keogh et al., 2004). The algorithm workflow is described in Figure 3.3 and the core steps are:

Step 1 Initial Segments: Create initial segments to approximate profiling data by connecting two consecutive points. Therefore, Points 1 and 2 (starting from the lake surface) form the first segment (Seg_1), Points 3 and 4 form the second segment Seg_2, etc.

Step 2 Compute Errors for Potential Merge Locations: Calculate the approximation errors (E) of one segment (Seg_i) combined with the next segment (Seg_{i+1}). The errors represent how closely the linear segment approximates the data when Seg_i and Seg_{i+1} are combined.

Step 3 Select Merge Locations: Find the combination of segment $i, i+1$ that generates the smallest error. If this smallest error is less than the error threshold, E_{max}, go to *Step 4*. Otherwise, stop and output the segments that remain.

Step 4 Merge: Combine the segment $i, i+1$, go to *Step 2*.

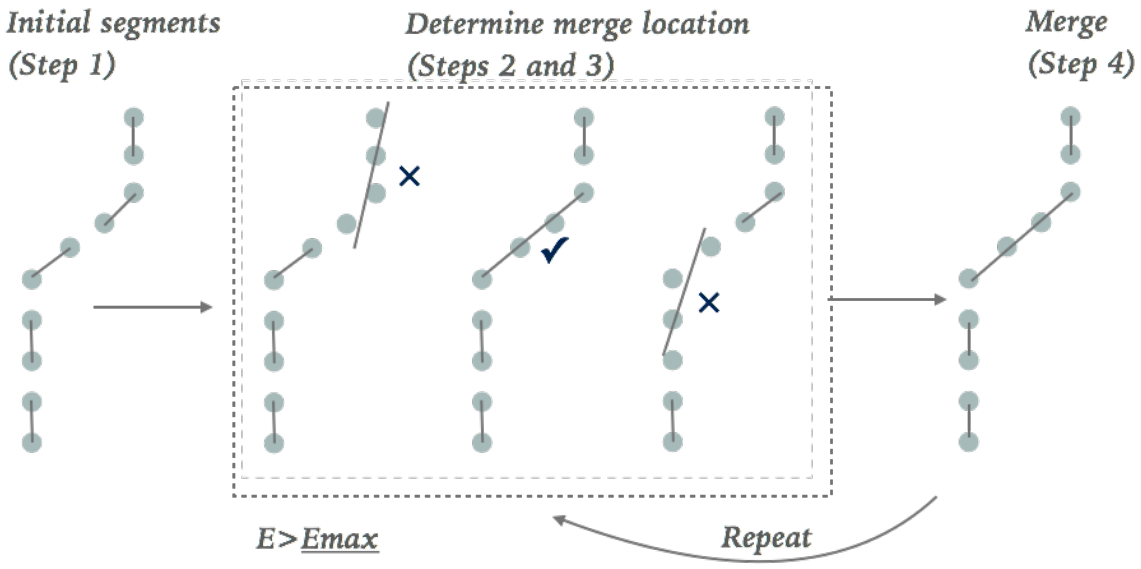
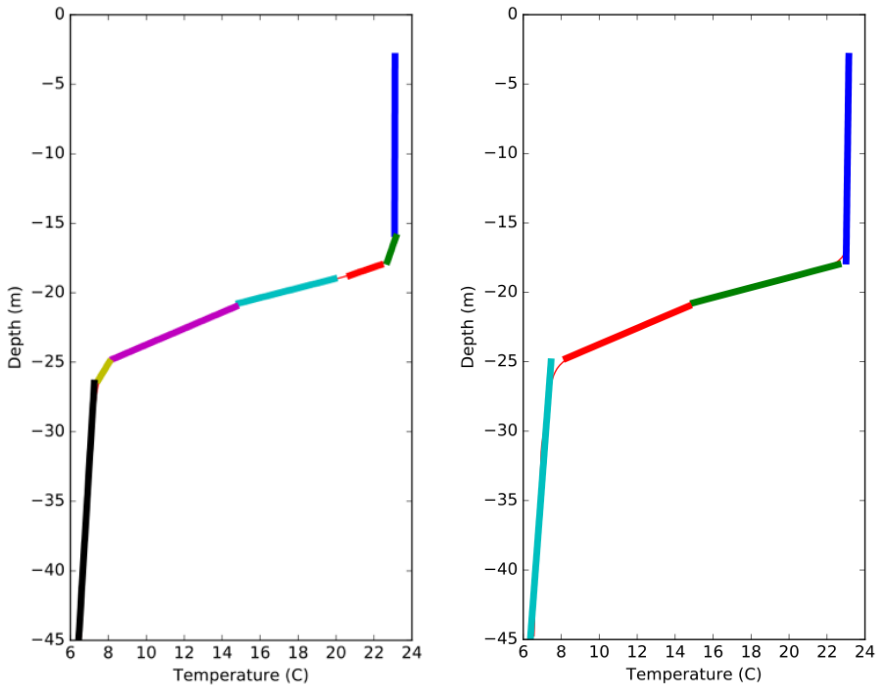


Figure 3.3. Algorithm workflow of piecewise linear representation

In our implementation, the linear segments that approximate the raw data are produced by linear regression. The approximation error E is determined by the maximum discrepancies of the linear segment approximation and the raw series. Through trial and error by manually comparing the detected layers with operators' notes, we selected $E_{max} = 0.3C$, indicating the linear segment approximation will have at most 0.3 Celsius degree differences with the temperature data. Reducing E_{max} will produce more segments, which decomposes the temperature profile into smaller pieces and is able to detect thin epilimnion or hypolimnon, while increasing E_{max} will create a rougher representation, ignoring more details in the profile (Figure 3.4). Fiedler (2010) used 3 percent of the temperature range from sea surface to sea bottom as E_{max}, which was too large and failed to detect thin epilimnion and hypolimnion in our datasets. An extremely large E_{max} value will produce only one segment, which is just the linear regression of temperature values with depth.



(a) $E_{max} = 0.3$ (b) $E_{max} = 0.7$

Figure 3.4. Segments generated by different E_{max}

The PLR algorithm generates L segments ($Seg_1, Seg_2, \ldots, Seg_L$ starting from lake surface to lake bottom), which are used to detect the epilimnion, thermocline, and hypolimnion via the following criteria:

(1) *Thermocline*:

The depth of the middle point of the segment with maximum gradient is set as the thermocline depth. To accommodate situations where no thermocline appears in the profiles, we apply a threshold g_{min}^{TRM} on temperature gradient, below which the thermocline does not exist in the profiles. Previous work gives different minimum gradients for the thermocline, with values ranging from $0.05°Cm^{-1}$ (McCullough et al., 2007) to $1°Cm^{-1}$ (Watkins et al. 2015). We used a smaller threshold, $0.15°Cm^{-1}$ rather than $1°Cm^{-1}$ because: (1) the segment approximation tends to smooth the raw gradient, so using $1°Cm^{-1}$ identify fewer thermoclines; and (2) in our data set, some profiles have mild temperature changes in the metalimnion due to already low surface temperatures.

(2) *Epilimnion and Hypolimnion*:

Epilimnion and hypolimnion are stable water columns where the temperature gradient is

small. For our datasets, only including the topmost segment as the epilimnion (Thomson and Fine, 2003) generated epilimnion layers that were sometimes too thin to match the data well. Therefore, in identifying the epilimnion, we examine the gradients of segments from the surface to bottom and find the index p so that Seg_i $(1 \leq i \leq p)$ has a temperature gradient less than a threshold $g(Seg_i) < g_{stable}$. Thus, the profile from Seg_1 to Seg_p is determined as the epilimnion.

Similarly, in identifying the hypolimnion, we check the gradients of segments from the bottom to the surface and find the index q so that for Seg_j $(q \leq j \leq L)$, $g(Seg_j) < g_{stable}$. Then Seg_q to Seg_L is the hypolimnion. It should be noted that in some cases the first (Seg_1) or last segments (Seg_L) sometimes have large gradients due to noise. To accommodate these edge cases, we relaxed the gradient threshold on the Seg_1 and Seg_L from g_{stable} to g_{stable}^{relax}, allowing the first and last segment to have larger gradients to be included in the epilimnion and hypolimnion, respectively.

To determine g_{stable}, we first calculated temperature gradients at the depths of upper hypolimnion and lower hypolimnion from operators' notes, which are summarized in Figure 3.5. Since the operators have different judging criteria and depths may be mislabeled (discussed in Section 3.4.2), the median values of the gradients (0.12 C/m for lower epilimnion and 0.097 C/m for upper hypolimnion) were selected as reasonable estimates of stable water columns. Finally, we choose $g_{stable} = 0.1°Cm^{-1}$ and $g_{stable}^{relax} = 0.25°Cm^{-1}$ by trial and error.

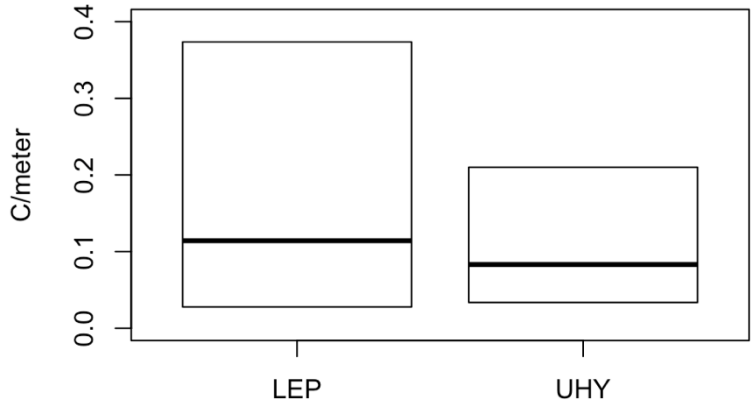


Figure 3.5. Boxplot of temperature gradients at depths of UHY (upper hypolimnion) and LEP (lower epilimnion) identified by operators in all lakes. The lower, middle, and upper lines in the box represent the 1st quantile, median, and 3rd quantile values.

The differences between the above algorithm and the previous approach (Fiedler, 2010) are: (1) We use a bottom-up approach rather than the split and merge approach; (2) we use a smaller error E_{max} to accommodate thin epilimion or hypolimnion in the profiles from the Great Lakes and (3) we extend the model by using stable gradient constraints g_{stable} to detect epilimnion and hypolimnion.

MG-HMM algorithm

For comparison, we also implemented another algorithm called MG-HMM (maximum gradient hidden Markov model) to detect stratification patterns. Instead of finding the segment with the maximum gradient, we set the thermocline as the depth with the highest temperature gradient calculated from the data after preprocessing (Section 3.3.1) rather than PLR segments. To separate the epilimnion, hypolimnion, and metalimnion, we apply hidden Markov models (HMM). In HMM model, the hidden states follow a Markov process with a transition matrix P where P_{ij} is the probability of transiting from state i to state j, where i and j can choose from "e"(epilimnion), "m"(metalimnion) and "h"(hypolimnion). The temperature and its gradient are the observations of each state and follow a bivariate Gaussian distribution.

To solve the HMM, expectation-maximization algorithm (EM) is used to determine the hidden states as well as the parameters of the model. EM algorithm is an iteration algorithm to maximize the likelihood of the observed data, starting from initial guesses of model parameters including the mean of each state (i.e. the expected values of the temperature and gradient in epilimnion, metalimnion, and hypolimnion) and the state transitional probability matrix P. Different initial values may lead to different results, causing unstable performance. We set the initial values based on prior knowledge as follows:

(1) Transitional probability matrix:

Denoting the total number of data points as n in the depth profile, we set the initial transition probabilities from epilimnion to metalimnion (P_{em}) and metalimnion to hypolimnion (P_{mh}) very small (=$3/n$), since the epilimnion will transition to the metalimnion at only one point, and similarly for metalimnion to hypolimnion. Hypolimnion is only allowed to transition to the hypolimnion itself as it is the deepest part of the depth profile, thus we set $P_{hh} = 1$ and $P_{he} = P_{hm} = 0$.

(2) The initial state mean:

In a typical three-layer stratification structure, the surface water temperature and gradient

define the epilimnion, the bottom water define the hypolimnion, and the middle values between the surface and bottom water determine the metalimnion. Therefore, we set the initial mean of epilimnion state and hypolimnion state as the median values of temperature and gradient in the lake surface and lake bottom, respectively. The temperature and gradient of metalimnion state are the averages of lake surface and bottom temperature and half of the 90th gradients, respectively. These initial conditions are likely near the desired solution (i.e. three layers representing epilimnion, metalimnion, and hypolimnion) and the expectation-maximization algorithm (EM) will iteratively update the parameters based on the detailed profile values.

3.3.3 DCL Detection Algorithm

Lastly, we identify the DCL based on a peak detection algorithm that identifies the number of peak points and their locations before analyzing the shapes of each peak. Denote f_i and g_i as the fluorescence concentration and gradient at the i^{th} point, starting from lake surface to the bottom, respectively. The algorithm first finds all zero crossing points of g_i, which occur where the fluorescence gradient changes from positive to negative, meaning fluorescence concentration increases and then decreases. Then the algorithm filters out peaks that are not significant. The detailed steps are (Figure 3.6):

Step 1: Find the points where the gradient changes from positive to negative, i.e. fluorescence concentration increases and then decreases. These zero crossing points are the potential peak points but may contain local minima (for example, see Figure 3.6, step 1). Denote $P^{(0)}$ as this set of peak candidate points.

Step 2: For each point in $P^{(0)}$, we apply a global threshold (f_{min}) so that point $P_i, P_i \in P^{(0)}$ will be filtered out if $f_i < f_{min}$. We also combine peaks that are close by only considering the peak with the highest magnitude if the peaks are within 2.5 meters. Denote the remaining point set as $P^{(1)}$.

Step 3: For each point in $P^{(1)}$, denote D_U^i and D_L^i as the set of data within the upper and lower sides of the peak point i (e.g., see Figure 3.6). If i is the first peak from the lake surface, then D_U^i are the set of data from the first data point to the peak point. If point i is not the first peak point, then D_U^i is defined as the set of data from P_i to P_{i-1}. D_L^i follows a similar definition: D_L^i is the set of data from peak i to peak $i+1$ or to the last data point. Calculate $h_U^i = f_i - min(D_U^i)$ and $h_L^i = f_i - min(D_L^i)$. Then the height of peak i is $h_i = min(h_U^i, h_L^i)$.

Step 4: Remove the peak point i from $P^{(1)}$ if $h_i < h_{min}$. h_{min} is a threshold parameter set by trial and error that is related to how significant the peak should be.

Step 5: Repeat step (3) and (4) until no points in $P^{(1)}$ can be removed. That is, all of the peaks left in $P^{(1)}$ are significant. If $P^{(1)}$ is empty, then there is no peak in the profile.

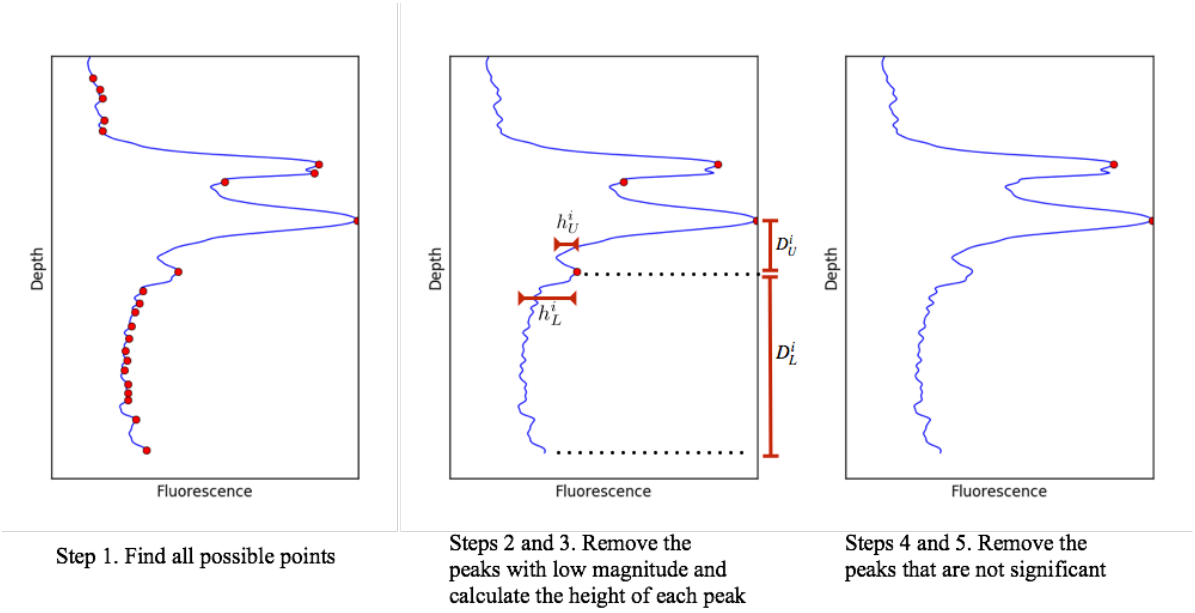


Figure 3.6. Peak detection algorithm for identifying possible peaks and then filtering out insignificant peaks

The peak boundaries (i.e. where the peak shape ends) are ill defined, especially for non-bell shape peaks or multiple peaks exists in the profiles. For profiles with only one peak, we fit two half Gaussian shapes with a linear trend of background concentration for the data above and below the peak, as follows:

$$\hat{y} = y_0 + k(x - x_0) + a\frac{e^{-(x-x_0)^2}}{2\sigma^2} \tag{3.1}$$

where x is the depth, x_0 is the location of the peak point, and $y_0 + k(x - x_0)$ is the background concentration with a trend k. We set $y_0 = max(y) - a$ to align the magnitude of the Gaussian peak with the magnitude of the peak in the data, leaving only parameters k, a and σ to be fit. To ensure that the increasing/decreasing stage is mostly captured by the Gaussian shape, we restricted k within a small range determined by trial and error as $[-0.15 \times K_{thres}, 0.15 \times K_{thres}]$, where K_{thres} is the gradient of the line connecting the peak point to the last point of the data to be fit.

The two half Gaussian shapes can have different parameter values so that non-symmetric shapes can be fit. We define the DCL upper boundary as $max(LEP,\ Depth_p - 2.5\sigma_1)$ and the lower boundary as $min(Max(depth),\ Depth_p + 2.5\sigma_2)$, where LEP is the boundary between epilimnion and metalimnon, $Depth_p$ is the depth of the peak point, and σ_1, σ_2 are the standard deviations of the upper and lower Gaussian shapes, respectively. The value of 2.5 is similar to the value recommended by Siswanto (2005), who defined the half peak size as 2σ when fitting with only one ordinary Gaussian shape without the trend of the background concentration. We calculated the squared correlation coefficient (r^2) to measure the fitness of Gaussian shapes to the data.

The final DCL depth is determined as the peak location with maximum fluorescence concentration and below the epilimnion, rather than the detected thermocline. This is because the depth of the thermocline may have different definitions among operators (discussed in the Section 3.4.2). Therefore, setting the depth of the lower epilimnion as the DCL depth upper bound reduces the effects of thermocline identification variations on DCL detection. As the moving average smoothing conducted in the data preprocessing stage may flatten the concentrations, the algorithm searches for the maximum fluorescence values C_m^{raw} in the raw data within 1 meter around the detected DCL depth. C_m^{raw} is then selected as the final DCL concentration.

The parameters used in the lake stratification and DCL detection algorithms are summarized in Table 3.1.

Table 3.1. Stratification and DCL detection algorithm parameters

Variable	Parameter	Definition	Value	Parameter Effects
Therm-ocline	E_{max}	Maximum error in PLR approximation	$0.30°C$	Smaller values will generate more and smaller segments, revealing more local structures
	g_{stable}	Maximum gradient of a stable water layer	$0.10°Cm^{-1}$	Smaller values will determine a deeper hypolimnion. Too small values will generate no epilimnion or hypolinion
	g_{stable}^{relax}	Maximum gradient of the first or last layer to be more robust in some profiles	$0.25°Cm^{-1}$	Smaller values may result in no epililimnion or hypolimnion detected
	g_{min}^{TRM}	Minimum gradient for thermocline	$0.15°Cm^{-1}$	Smaller values will tend to identify more profiles with mildly changing gradient

Table 3.1 (cont.)

Variable	Parameter	Definition	Value	Parameter Effects
Fluores-cence	f_{min}^{DCL}	Minimum magnitude of a peak	$min(f) + 0.3 \times (max(f) - min(f))$	Smaller values will identify more peaks with low absolute peak magnitude
	h_{min}^{DCL}	Minimum height of a peak	$0.2 \times (max(f) - min(f))$	Smaller values will identify more peaks with low relative peak magnitude (peak height)

3.3.4 Algorithm Implementation and Web Applications

The detection algorithms are implemented using Python and its libraries Numpy, Scipy, Pandas, hmmlearn, bekeh, and Flask. We also built an interactive Web application based on Flask web development framework (Figure 3.7 contains a snapshot of the Web app). The Web application allows users to upload raw SeaBird CTD data files and computes the depths of TRM, LEP, UHY and DCL, as well as DCL concentration with depth profiling plots. The code is open source (available at https://github.com/stormxuwz/SeabirdCode).

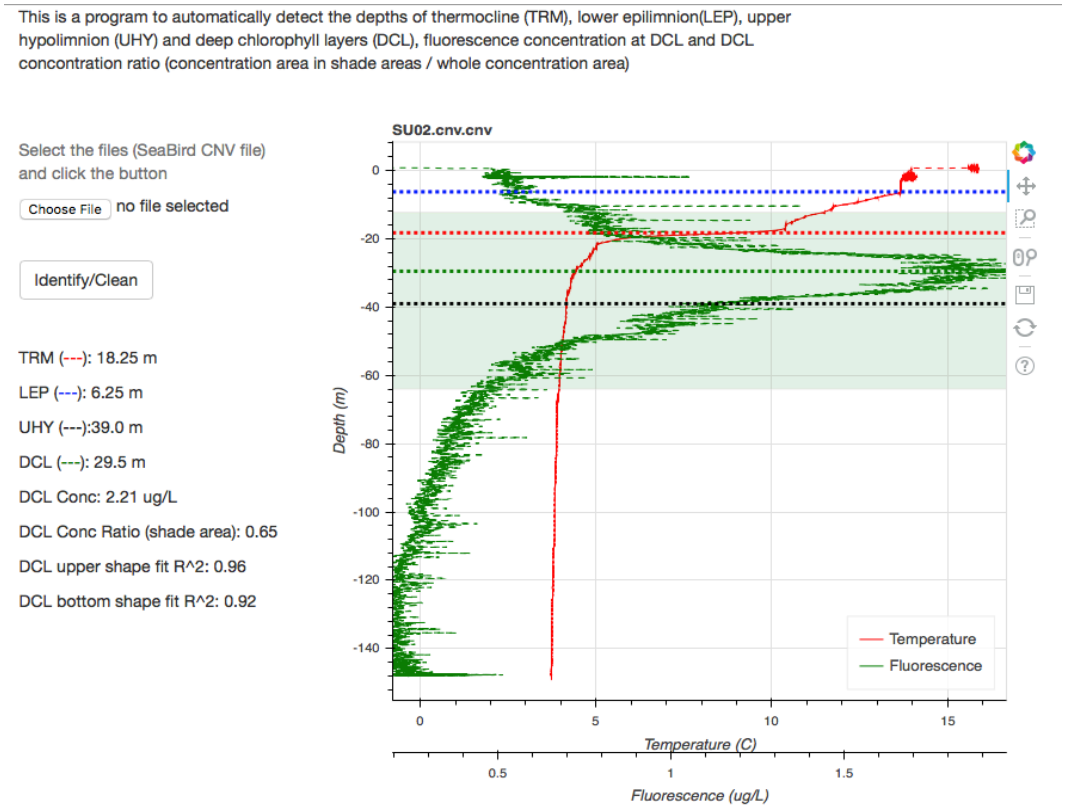


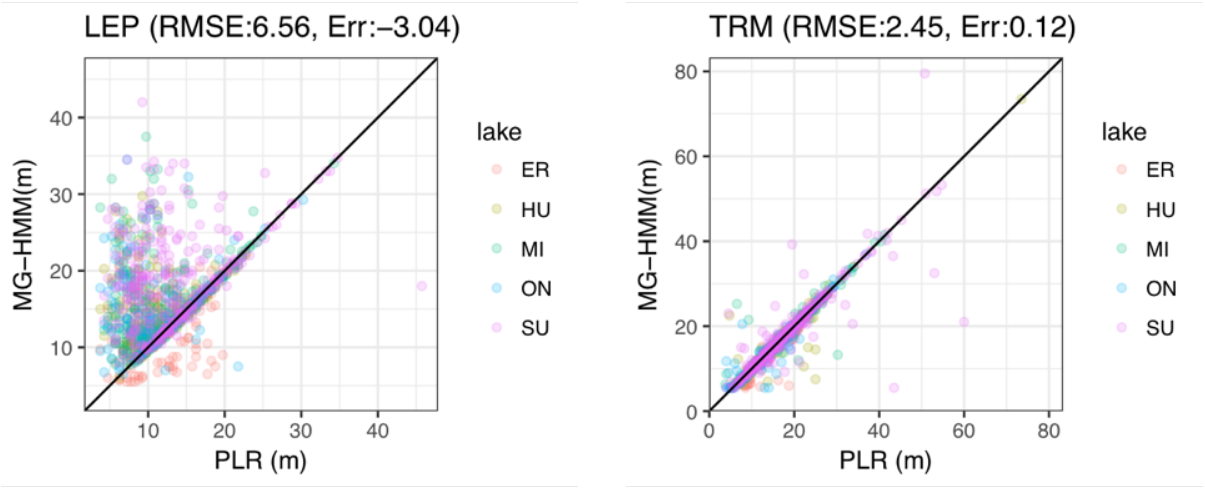
Figure 3.7. Web application interface. TRM, LEP, and UHY represents the depth of thermocline, lower epilimnion (the boundary between epilimnion and metalimnion), and upper hypolimnion (the boundary between metalimnion and hypolimnion)

3.4 Results

We applied the lake stratification and DCL detection algorithms on the Great Lakes profiles described previously. To validate the algorithms, we first compared the PLR and MG-HMM lake stratification detection algorithms. Then the results of the PLR lake stratification and peak detection algorithms are compared with historical operators' notes. Further discussion shows how the algorithms are able to characterize and detect unusual patterns in temperature and fluorescence profiles. Note that in this section, we use ER, HU, MI, ON, SU as the abbreviations for Lake Erie, Lake Huron, Lake Michigan, Lake Ontario and Lake Superior, respectively.

3.4.1 Lake Stratification Algorithm Comparison

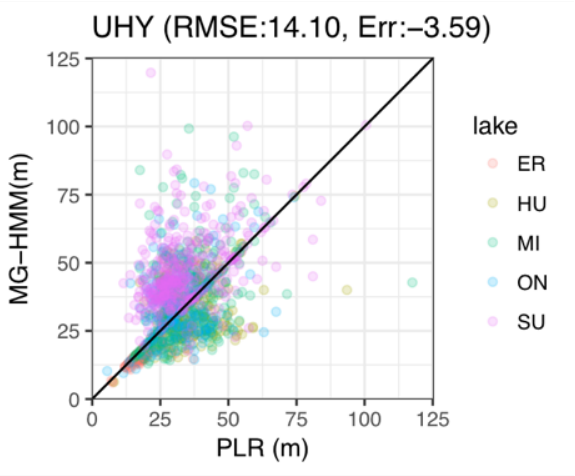
Figure 3.8 shows the differences between the PLR and MG-HMM algorithms in detecting the depth of thermocline (TRM), lower epilimnion (LEP, the boundary between epilimnion and metalimnion) and upper hypolimnion (UHY, the boundary between metalimnion and hypolimnion). The 1st /median/3rd quantile errors (mean depth difference between the algorithms) are -0.25/0.00/0.25 m for TRM, -4.88/-0.25/0.25m for LEP, and -11.25/0.00/2.50m for UHY, respectively.



(a) LEP (b) TRM

Figure 3.8. Comparisons between PLR algorithm and MG-HMM for all temperature profiles collected in all lakes from 1996 to 2013. "Err" represents the mean depth difference (m) between PLR and MG-HMM. Positive errors indicate depths generated from PLR are deeper.

Figure 3.8 (cont.)



(c) UHY

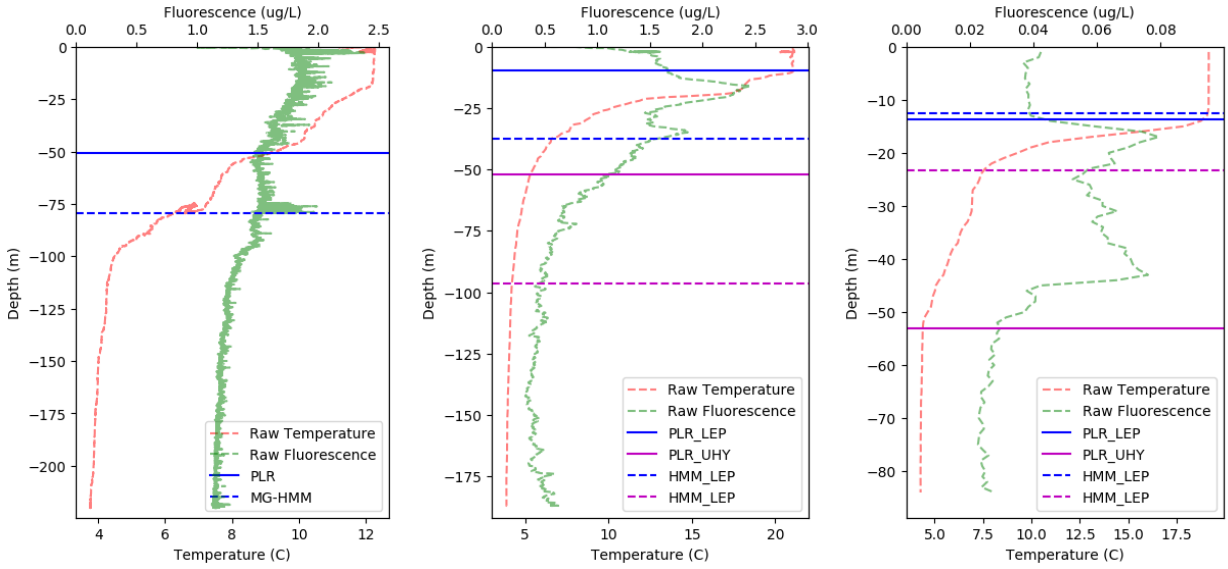
The results show that the two algorithms have often have similar performance, with median values around 0 and the 1st and 3rd quantile less than 5 meters for TRM and LEP. However significant discrepancies exist in some cases, for which we observe the following:

(1) For thermocline (TRM) detection (Figure 3.8b), PLR uses the gradients (or slope) of the linear segments, which are different from the gradients at each individual point used in MG-HMM (Section 3.3.2). For example, in Figure 3.9a, the PLR method (blue solid line) identifies a thermocline at an upper depth that has a maximum segment gradient (blue solid line) while the thermocline given by MG-HMM (blue dash line) is located at a deeper depth where the point gradient is maximum. The PLR algorithm's dynamic smoothing of the data provides more robustness to these types of small oscillations in the sensor data because it captures the gradient in a less localized manner.

(2) For lower epilimnion (LEP) detection (Figure 3.8a), the largest discrepancies are related to the HMM model fitting process. The HMM model requires setting the number of states in advance (i.e. three states for three lake layers) and the expected-maximization algorithm (EM) finds the parameters and state boundaries that fit the sampled data with maximum likelihood. In some profiles, where epilimnion or hypolimnion are quite large or small, the HMM model will produce unreasonable

results. For example, consider the 1999 Station MI47 in Figure 3.9b. The HMM model will detect a much deeper LEP in the profile, which has a very shallow epilimnion, because the EM algorithm will treat the data in the thin epilimnion as a few outliers rather than a separate state. Therefore, the epilimnion data will be grouped with the metalimnion data. The extreme case occurs with a profile that has no epilimnion or hypolimnion (e.g. due to sampling range), where the HMM model will divide the data from metalimnion and hypolimnion (or metalimnion and epilimnion) into three states regardless of the absence of one state. Most of the large discrepancies in LEP are because the HMM model generates deeper estimated LEPs due to the above limitations.

(3) For upper hypolimnion (UHY) detection, given that the HMM model detected a deep LEP for some profiles, the UHY will also be deeper (e.g., the purple dashed line in Figure 3.9b). However, there are some positive differences where HMM estimates a shallower depth. This is because the PLR algorithm uses absolute gradient threshold (g_{stable} in Table 3.1) to identify the hypolimnion while HMM models separate the data using the relative magnitude of gradients (recall that MG-HMM clusters water layers with similar gradients - see Section 3.3.2). HMM models produce a shallower UHY when the gradient around the thermocline is relatively sharper, so that the data points with smaller gradients are grouped with the hypolimnion data points (e.g., Figure 3.9c).



(a) SU15 2009 (b) MI47 1999 (c) HU53 2002

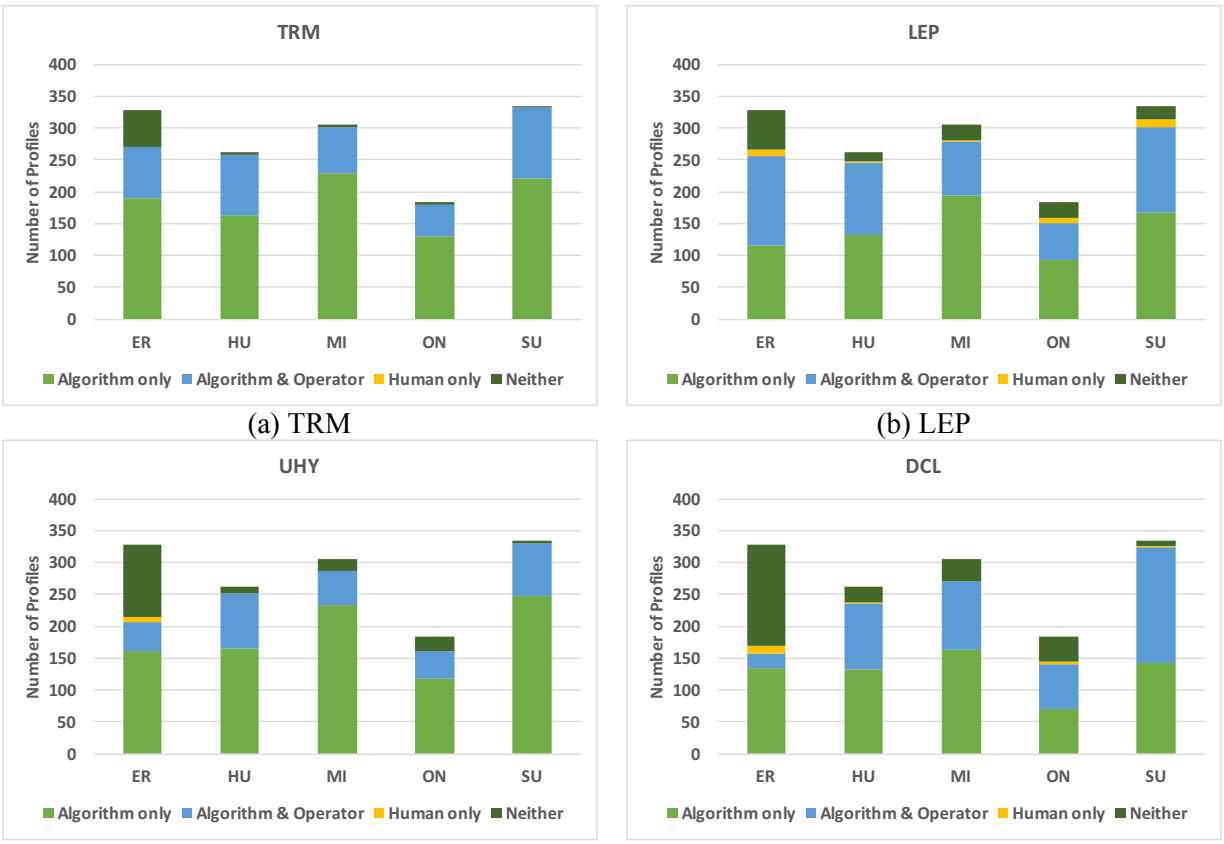
Figure 3.9. Comparison between PLR and MG-HMM methods (a) TRM comparison at SU15 in 2009 (b), and (c) LEP and UHY comparison at MI47 in 1999 and HU53 in 2002, respectively. The raw data in (b) and (c) are not noisy because these profiles are already aggregated.

Other algorithms could be used to detect the stratification structure, such as setting a gradient threshold such that above (or below) the thermocline, the first data point whose gradient is below the threshold is the LEP (or UHY). However, from our experiments, after data preprocessing, the profiles may still contain oscillations. Thus, the gradients at each point are not monotonically decreasing before and after the thermocline, and therefore a simple gradient threshold is not robust for this situation.

Because of these anomalies, we conclude that the PLR algorithm is more suitable than MG-HMM in detecting lake stratification patterns for heterogeneous profiles. Moreover, segmenting the profiles with PLR has other benefits such as detecting unusual shape patterns (e.g. double thermocline and increasing temperatures), which will be discussed in Section 3.4.3.

3.4.2 Algorithm Validation

The PLR and peak detection algorithms are validated with existing operators' historical notes from 1998 to 2012, providing a total of 1412 profiles. We first summarize in Figure 3.10 how many profiles in the dataset have features detected by the algorithms only (green), by the operators only (yellow), by both (blue), or by neither (dark green).

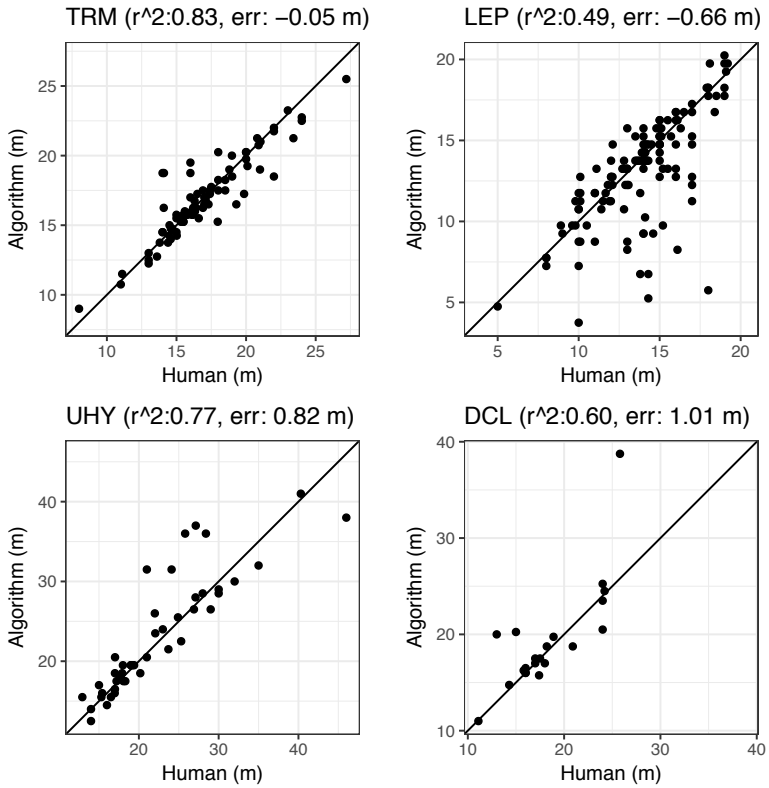


(a) TRM (b) LEP

(c) UHY (d) DCL

Figure 3.10. Number of profiles with features detected.

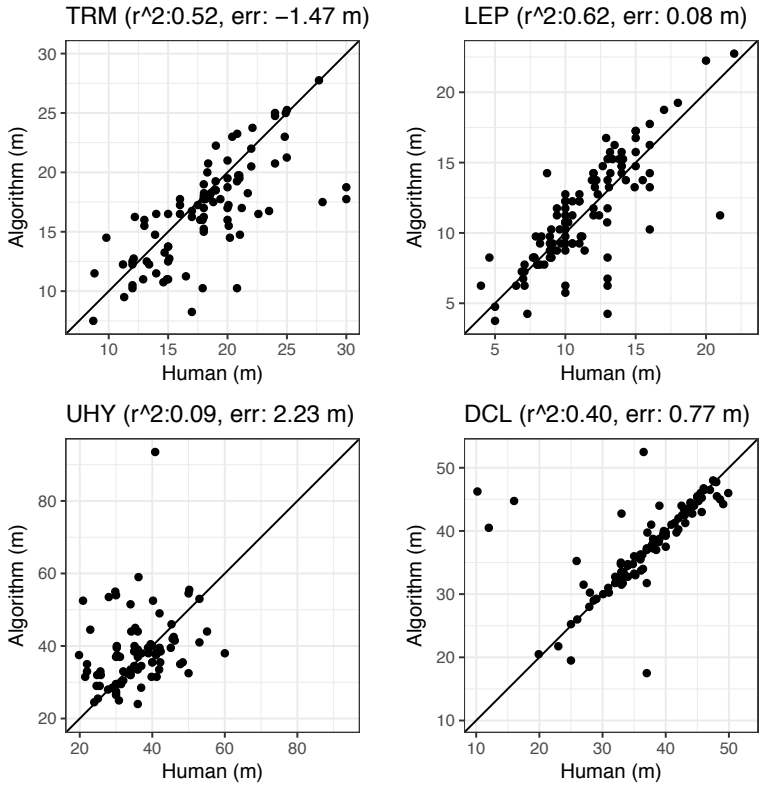
The algorithms identify most of the features in operators' notes, with only limited profiles containing features that are only detected by operators (yellow profiles in Figure 3.10). There are, however, many layers that are not captured in operators' records (green profiles in Figure 3.10). Comparing human and algorithm estimated depths (Figure 3.11), algorithm estimates are close to recorded human judgments, with $r^2 > 0.5$ in TRM, LEP, and DCL detection, except for DCL in Lake Huron ($r^2 = 0.40$). The discrepancies in UHY are generally large with small r^2 in Lake Huron ($r^2 = 0.09$), Michigan ($r^2 = 0.20$), Ontario ($r^2 = 0.35$) and Superior ($r^2 = 0.20$), but discrepancies are small in Lake Erie ($r^2 = 0.77$).



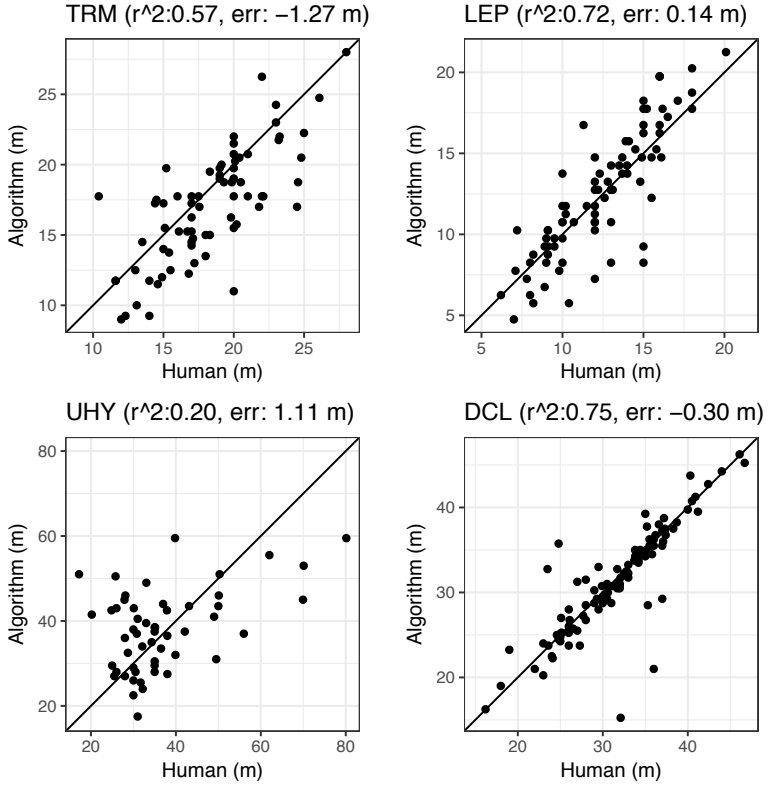
(a) Lake Erie (ER)

Figure 3.11. Comparisons between operators' and algorithm depth estimates for TRM, LEP, UHY and DCL; "r^2" represents the squared correlation coefficients (or Coefficient of Determination) and "err" represents the mean depth difference (m) between human and algorithms. Positive errors indicate that depths generated from algorithms are deeper.

Figure 3.11 (cont.)

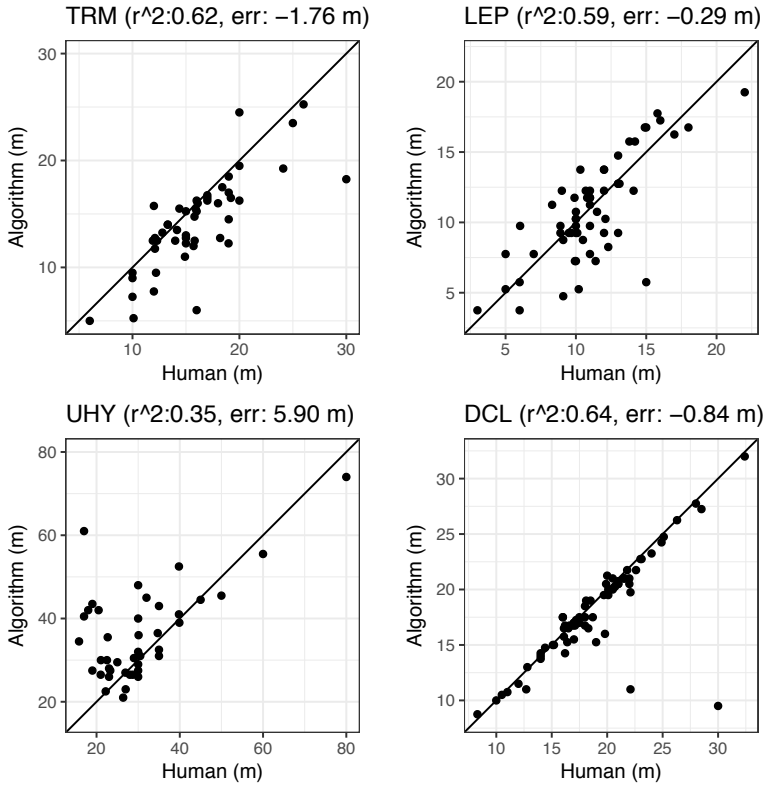


(b) Lake Huron (HU)

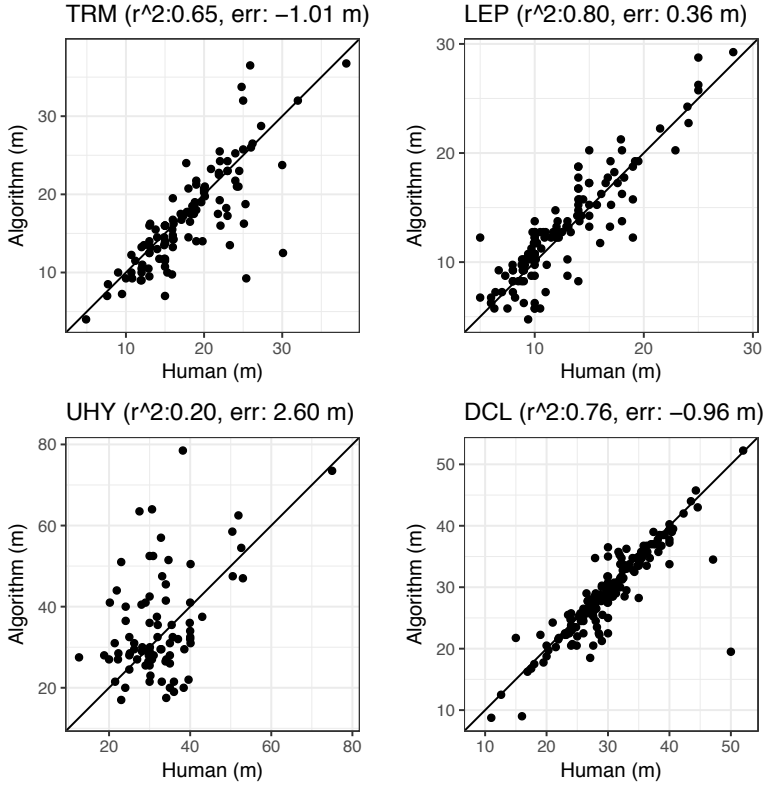


(c) Lake Michigan (MI)

Figure 3.11 (cont.)

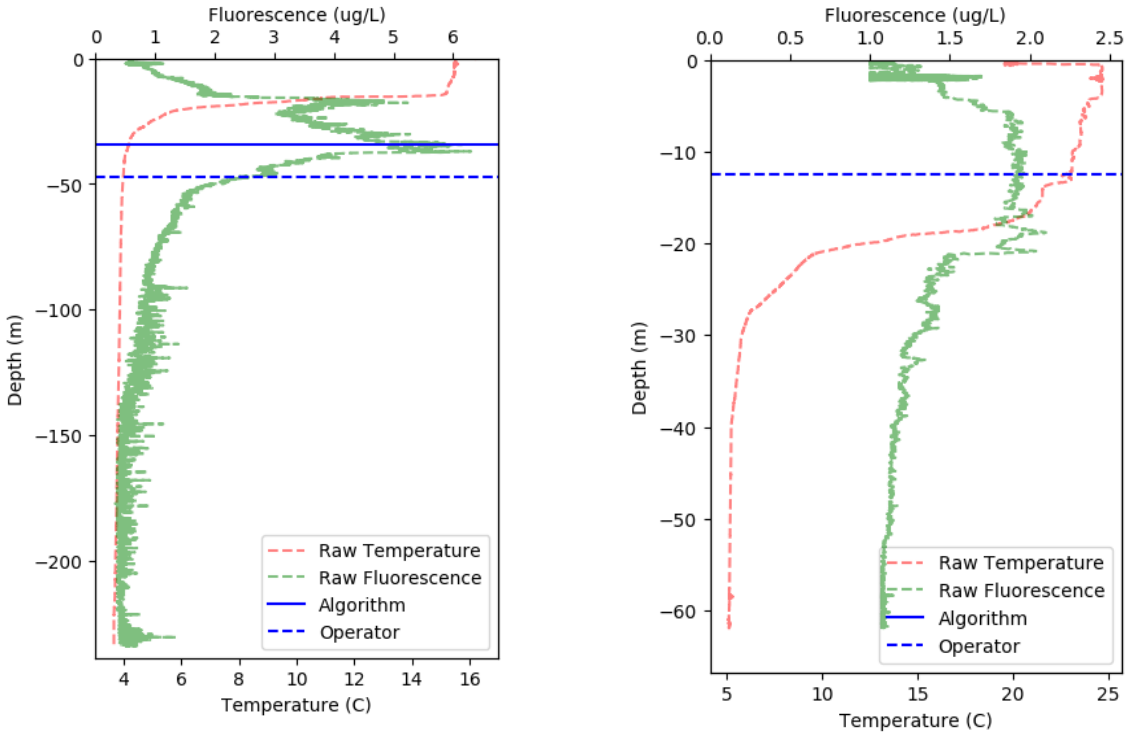


(d) Lake Ontario (ON)



(e) Lake Superior (SU)

By examining the profiles with larger discrepancies in more detail, we identify four categories that account for the differences, listed below. Figure 3.12 shows examples for these cases.

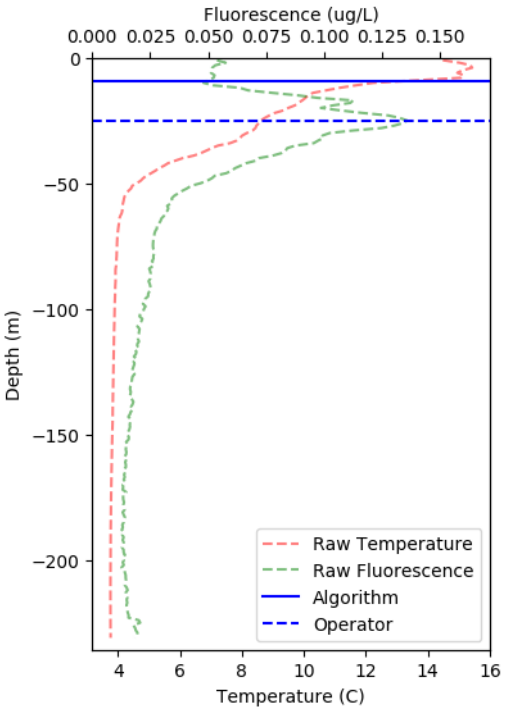


(a) SU09 in 2011, DCL mislabeled

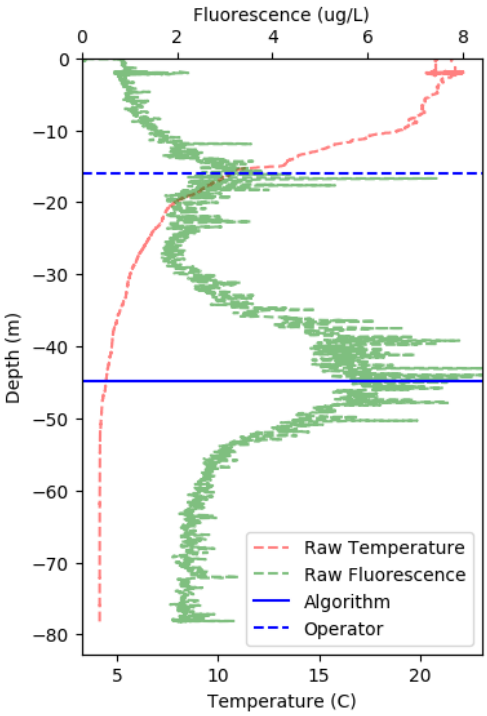
(b) ER15 in 2009, LEP failure

Figure 3.12. Outliers comparing algorithms and notes

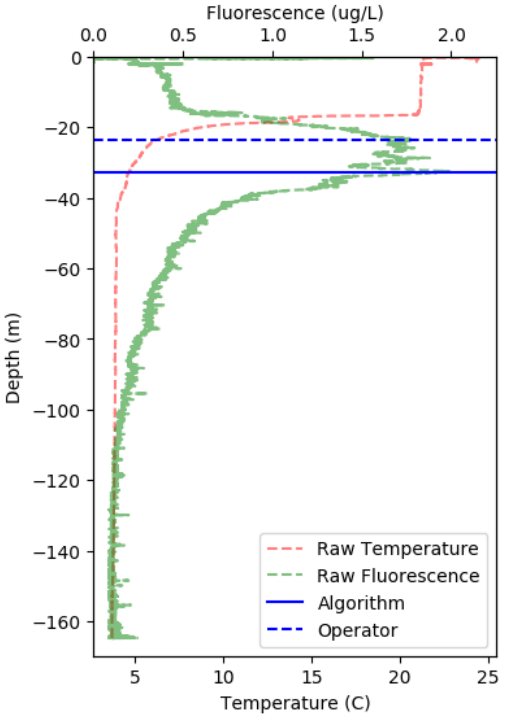
Figure 3.12 (cont.)



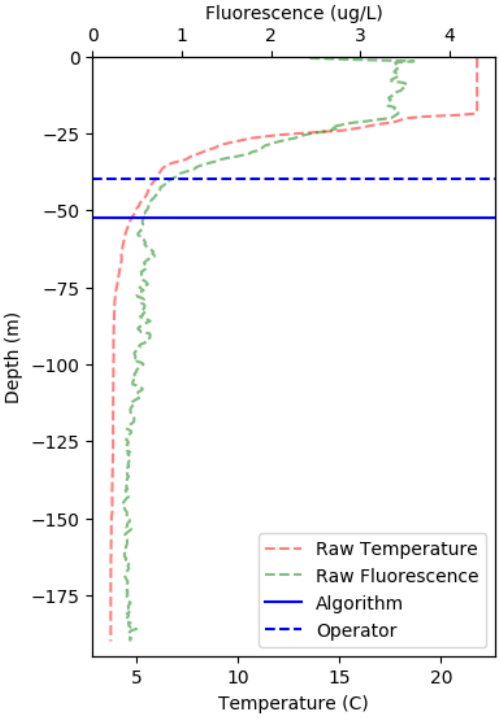
(c) SU09 in 2002, TRM definition



(d) HU32 in 2011, DCL definition



(e) MI40 in 2005, DCL subjectiveness



(f) ON55 in 1999, UHY subjectiveness

(1) *Data/expert mislabeled.* These outliers are due to operators' error or notes that are not corresponding to the correct profiles, which can occur when multiple casts are conducted by operators. For example, the DCL record of SU09 in 2011 is obviously incorrect in labeling the DCL as the peak point in the fluorescence profile (Figure 3.12a). This error was particularly prevalent in Lake Superior, where operators often labeled UHY using the depth of DCL, thus causing significant discrepancies (Figure 3.11e). Although some identifications may be reasonable (see *subjective ambiguity* discussion below), some operator labels are clearly incorrect.

(2) *Algorithm limitations*. There are several algorithm parameters (Table 3.1) that are related to the sharpness of the thermocline (g_{min}^{TRM}), the stability of the hypolimnion and epilimnion (g_{stable}), and the criteria to filter out shallow peaks (f_{min}^{DCL} and h_{min}^{DCL}). For these parameters, we used the same values in all of the Great Lakes. However, these parameters may not always reflect the variability in operators' subjective judgment nor differing conditions in each lake. This type of errors accounts for most of the discrepancies in LEP detections, especially in Lake Erie, as some profiles in Lake Erie have larger temperature gradients in the epilimnion due to its shallowness. Therefore, the algorithm tends to detect a shallower LEP (Figure 3.11a). If no segments have gradients less than g_{stable}, the algorithms will fail to detect LEP (Figure 3.12b)

For DCL detection, some peaks are not detected, which could be caused by the moving average smoothing that can reduce the peak value. Another situation is that when no LEP is detected, the algorithm will then fail to detect DCL since DCL is the peak below LEP (Section 3.3.3).

(3) *Definition ambiguity*. The operator notes indicate that the operators may not always label TRM as the point where temperature changes most rapidly, nor DCL as the peak with the largest magnitude below the TRM. Many TRM discrepancies in Lakes Michigan, Ontario, and Huron (Figure 3.11b, c, d where algorithms' depths are shallower) happen when the sharpest change was just below the LEP, but the operators define a deeper depth that is in the middle of the metalimnion as the thermocline (e.g. Figure 3.12c). Fiedler (2010) suggest merging adjacent segments to the segment with maximum slope using another error tolerance parameter (E'_{max}). However, this adjustment requires another parameter that must be tuned and still may not match the operator TRM, which

is not always adjacent to the maximum gradient segment (e.g. Figure 3.12c).

DCL definition ambiguity occurs when the first peak below the thermocline is the DCL, while the algorithms select the second peak with the largest magnitude, although the algorithm can detect both peaks (e.g. Figure 3.12d). Most of the large DCL differences in Lake Huron and Ontario are due to this double-peak structure in the profiles sampled during 2011. The double peak structures indicate two locations for phytoplankton growth, and phytoplankton photoacclimation probably existed in the deeper peak due to light limitations.

(4) *Subjective ambiguity*. In some cases, the fluorescence peak is very broad and the exact largest point identified by the algorithms may differ from the notes (e.g., Station MI40 in 2005 has a large DCL peak as shown in Figure 3.12e). In addition, when the transition zone from metalimnion to hypolimnion is relatively smooth, the exact depth to separate the two layers is highly subjective (e.g., see Station ON55 from 1998 in Figure 3.12f). Such ambiguity accounts for the large UHY errors in Lakes Michigan, Ontario, and Huron. Lake Erie, on the other hand, usually has a sharp transition between metalimnion and hypolimnion, thus the algorithm performs well.

Overall, the algorithm mostly reflects the operators' criteria since a majority of the detections are close to operators' notes (dots are near the 1:1 line in Figure 3.11), but cannot fully reflect the variability in highly subjective judgments made by individual operators. Moreover, it appears that some operators consider DCL depth as a close proxy to UHY since depths of operators' UHY and algorithms' DCL from some profiles are almost identical. Such UHY detection criteria are reasonable since DCL do often exist near the UHY.

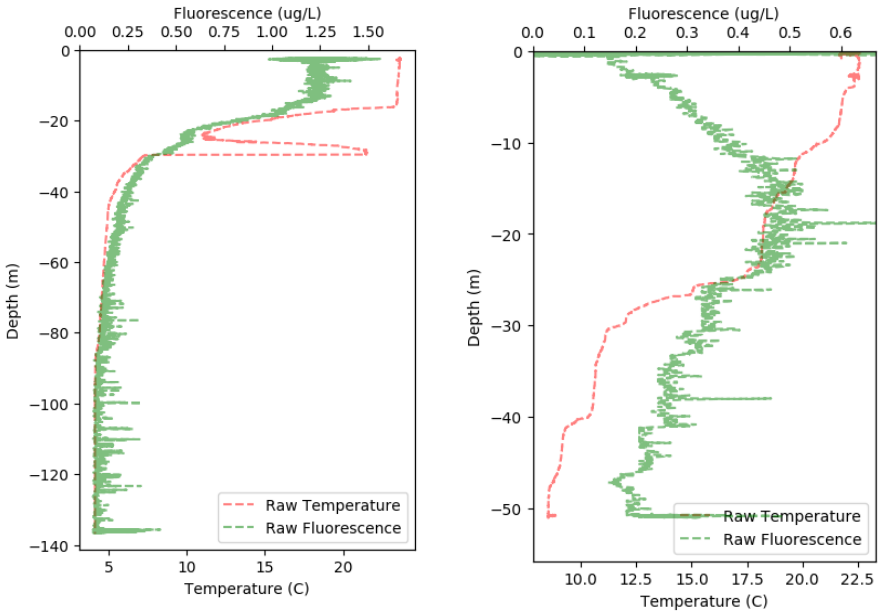
3.4.3 Shape Pattern Detection

By analyzing the gradient of each segment that approximates the profile data, we can automatically identify more features other than TRM, LEP, and UHY such as temperature anomalies. Similarly, the approximation of the profile peak with two half Gaussian shapes allows the size and symmetry of the DCL peak to be analyzed. These two capabilities are discussed in more detail below.

Temperature Anomalies

The gradients of segments generated by the PLR algorithm can be used to detect anomalies for data quality assurance and quality control as well as extracting unusual profiles for

further study. Profiles containing segments with significant positive gradients (temperature increases in deeper water) may have sensor errors. (e.g. ON33 in 2012 in Figure 3.13a). Another type of anomaly is a double thermocline, where the temperature drops rapidly, becomes stable, and then drops quickly again. Such phenomena may be caused by a remaining thermocline formed on a previous day (Ishikawa and Tanaka, 1993). We designed a simple double thermocline search algorithm to find the Seg_i where the temperature gradients (g) of segment Seg_i, Seg_{i+1} and Seg_{i-1} satisfy $g(Seg_{i+1}) < g_{stable}$, $g(Seg_{i-1}) < g_{stable}$ and $g(Seg_i) > g_{minTRM}$, meaning Seg_{i+1} and Seg_{i-1} are stable but Seg_i is not. Profile MI42 from 2006 (Figure 3.13b) is an example containing a double thermocline in the middle depth that was detected by the above search algorithm.



(a) ON33 2012, Positive gradient (b) MI42 2006 Double thermocline

Figure 3.13. Temperature anomalies

DCL Peak Shape

DCL shape features can also be analyzed automatically using the r^2 (squared correlation coefficients), which indicate how well Gaussian shapes fit the fluorescence data. For profiles with one and only one peak detected, we set a threshold T and calculate the proportion P of profiles that have both r^2 (i.e. r^2 for data above and below the DCL) greater than T. P will decrease with a larger T; i.e., fewer profiles will be selected that are similar to a Gaussian distribution shape, as shown in Figure 3.14.

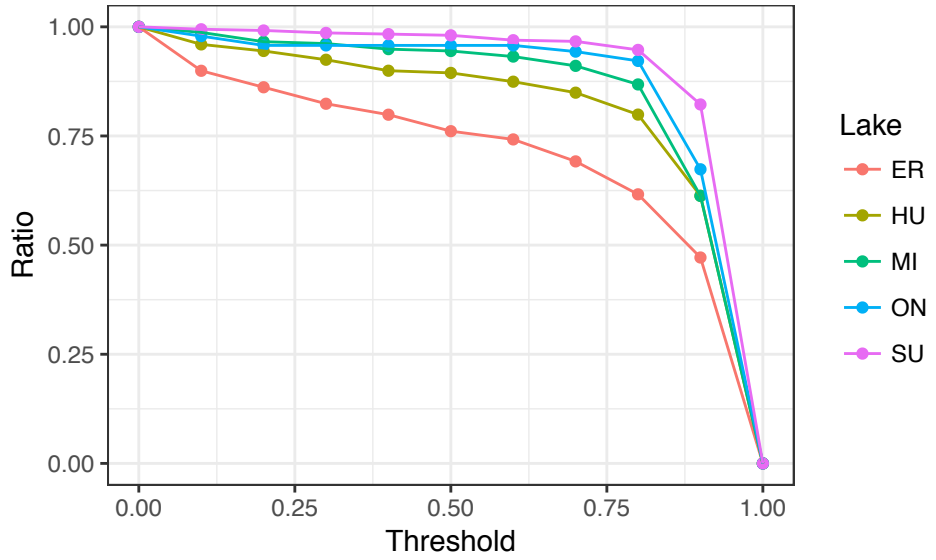


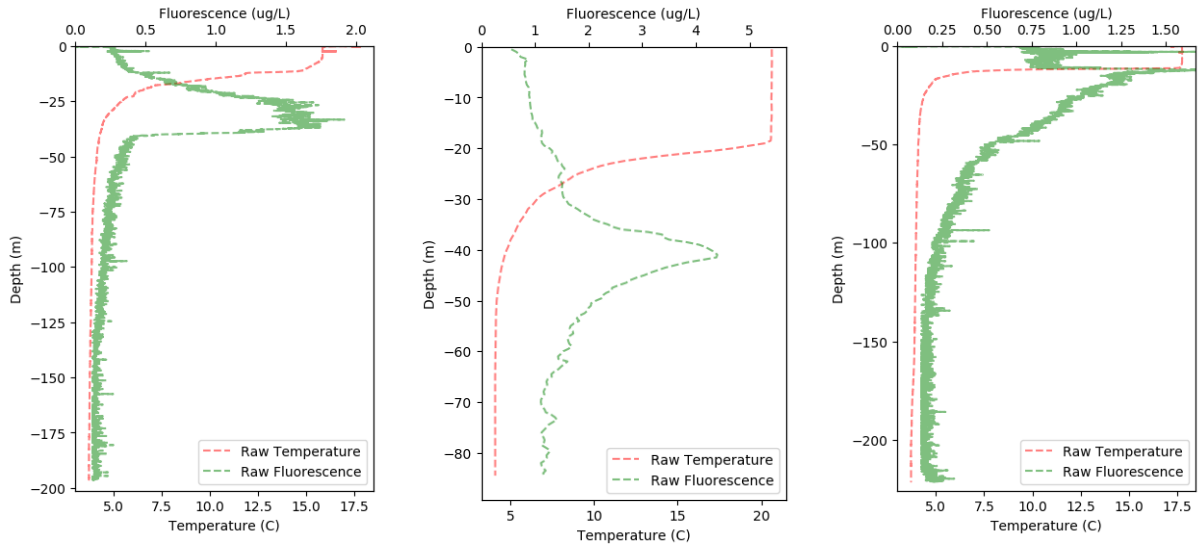
Figure 3.14. Profile proportions with different DCL fitness thresholds. The total number of profiles for each lake are: 159 (ER), 199 (HU), 235 (MI), 141 (ON) and 360 (SU).

Figure 3.14 shows that more than 75% of the fluorescence profiles with only one peak in Lake Superior can be characterized by Gaussian shapes with a strong fit ($r^2 > 0.9$). However only half of the detected DCL in Lake Erie can be well characterized, implying more complex biological and physical interactions that contribute to uncommon DCL shapes (i.e. non-Gaussian shapes) in Lake Erie.

Assuming that $r^2 > 0.9$ indicates a reasonable fit of two half-Gaussian shapes, we define the normalized peak symmetry metric γ as:

$$\gamma = \frac{\sigma_u - \sigma_l}{0.5(\sigma_u + \sigma_l)} \tag{3.2}$$

where σ_u, σ_l is the standard deviation of the half Gaussian shapes defined by Equation (3.1) using the data above and below the DCL, respectively. $\gamma > 0$ means that the peak has a milder gradient above the DCL than below the DCL. It should be noted that γ is dependent on the standard deviation of the Gaussian shapes. When Gaussian shapes characterize the data poorly, then γ is not a good representation of the DCL shape. Figure 3.15 shows some profile examples with different γ.



(a) SU17 2008 ($\gamma = 1.46$) (b) HU12 2000 ($\gamma = 0.003$) (c) SU11 2008 ($\gamma = -1.84$)

Figure 3.15. The shapes of DCLs with different symmetric patterns. (a) Peak has mild increasing gradient but sharp decreasing gradient; (b) Peak has relatively consistent increasing and decreasing gradient; (c) Peak has sharp increasing but mild decreasing gradient.

The DCL in Lake Ontario generally have sharper gradients in fluorescence concentrations above the DCL, with $\gamma < 0$ (Figure 3.16). One possible explanation for the asymmetric DCL shape could be due to predator grazing such as zooplankton and fish (Benoit-Bird et al., 2009). Fluorescence gradient above the DCL will be steeper when more zooplankton are presented, which can further be related to fish distribution (Benoit-Bird et al., 2009; Durham and Stocker, 2012). Symmetric shapes (e.g. Figure 3.15b) occur when zooplankton are in low abundance (Benoit-Bird et al., 2009). In addition, the steepness (σ_u, σ_l in Eq 3.2) could be related to turbulent mixing, direct swimming to the DCL, and vertical shear (Birch et al., 2009).

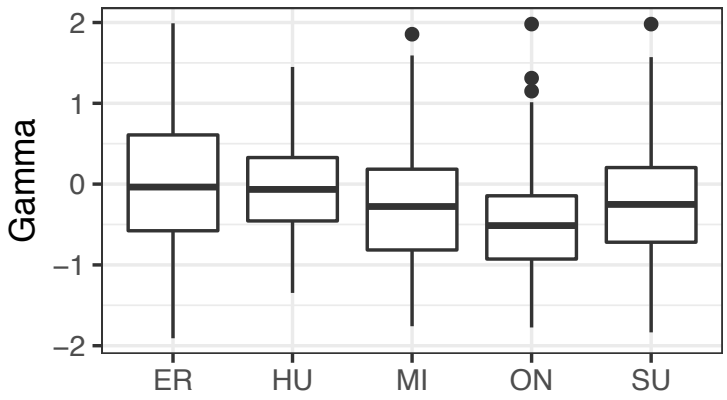


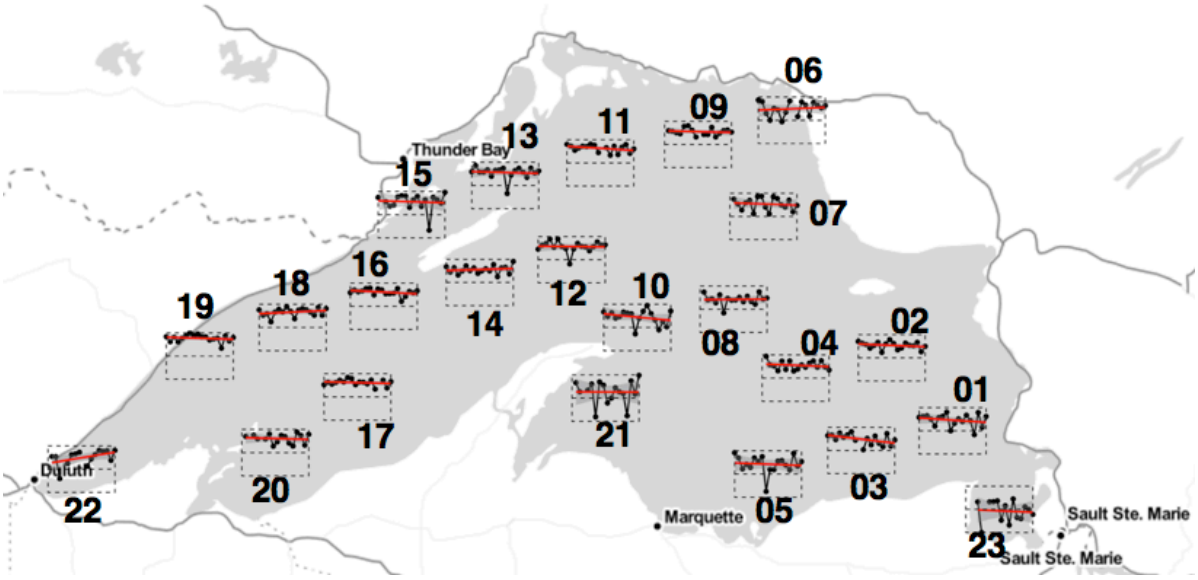
Figure 3.16. Peak symmetry (γ) of fluorescence profiles

3.5 Lake Superior Case Study

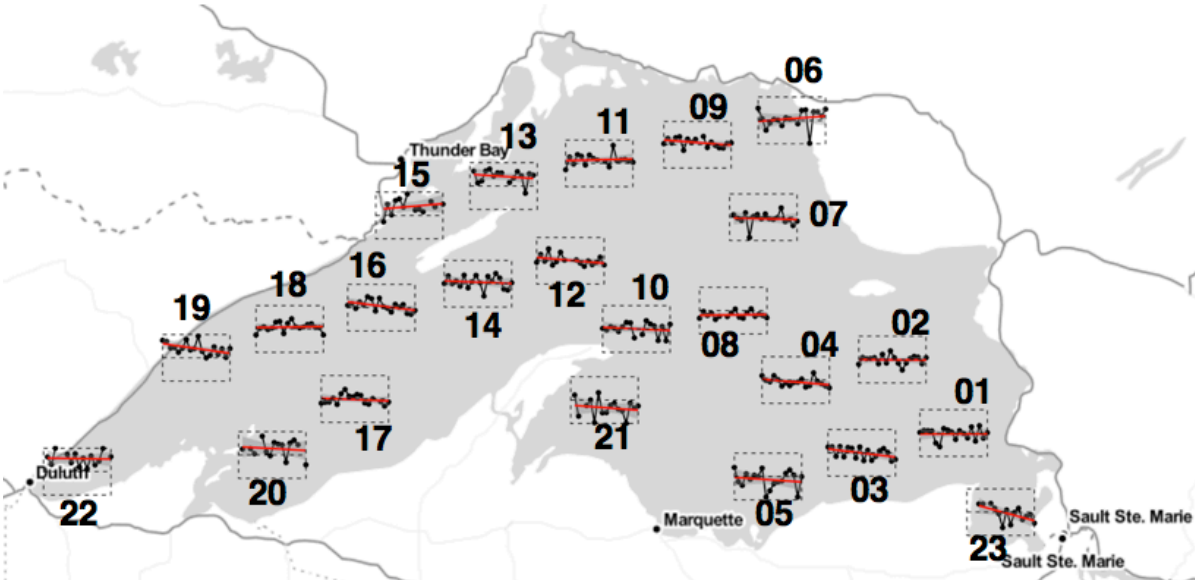
In this section, the detection results of profiles in Lake Superior are analyzed in more detail to explore the spatiotemporal changes of lake features (Section 3.5.1) and to illustrate how the algorithms can be used to perform rapid analysis of stratification and DCL patterns (Section 3.5.2).

3.5.1 Spatial and temporal trends of DCL and thermoclines in Lake Superior

Without manually analyzing and identifying features, we executed the proposed algorithms and used glyph-maps (Wickham et al., 2012) to visualize the spatiotemporal trend of the features detected. In glyph-maps (Figure 3.17), boxes with trend graphs are located at the geographic sampling locations on the map. The upper and lower line represent the maximum and minimum values of all years and all stations (i.e., a global range). The horizontal line in the mid-range is called the reference line, which is equal to the mean of the global maximum and minimum values. Glyph-maps help to understand the temporal trends as well as relative values across space. Note that for plots with depths (Figure 3.17a and b), higher points show data from shallower locations. A discontinuity (e.g. Station 23 in Figure 3.17a) indicates that no such features were detected in that year.



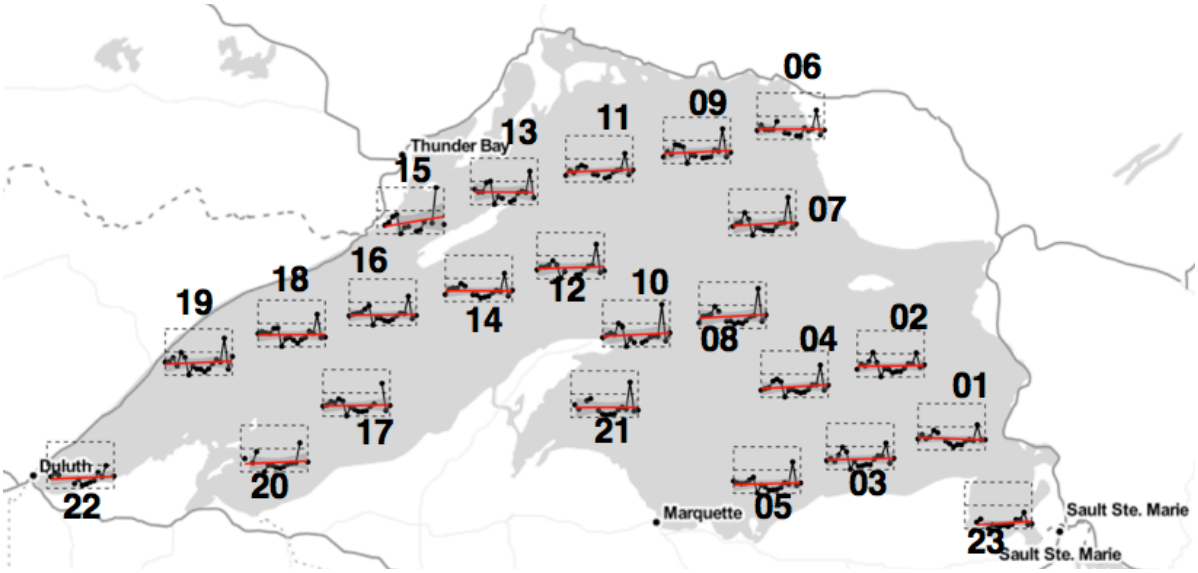
(a) Glyph-maps of thermocline depth (Range: 4.00 - 60.00 m)



(b) Glyph-maps of DCL depth (Range: 7.00 - 55.25 m)

Figure 3.17. Glyph-maps of thermocline depth, DCL depth, DCL concentrations from 1996 to 2013. The numbers are station indices. The red line is a linear trend. For plots with depths [(a) and (b)], a higher value indicates shallower depth. For plot (c), a higher value indicates high concentration.

Figure 3.17 (cont.)

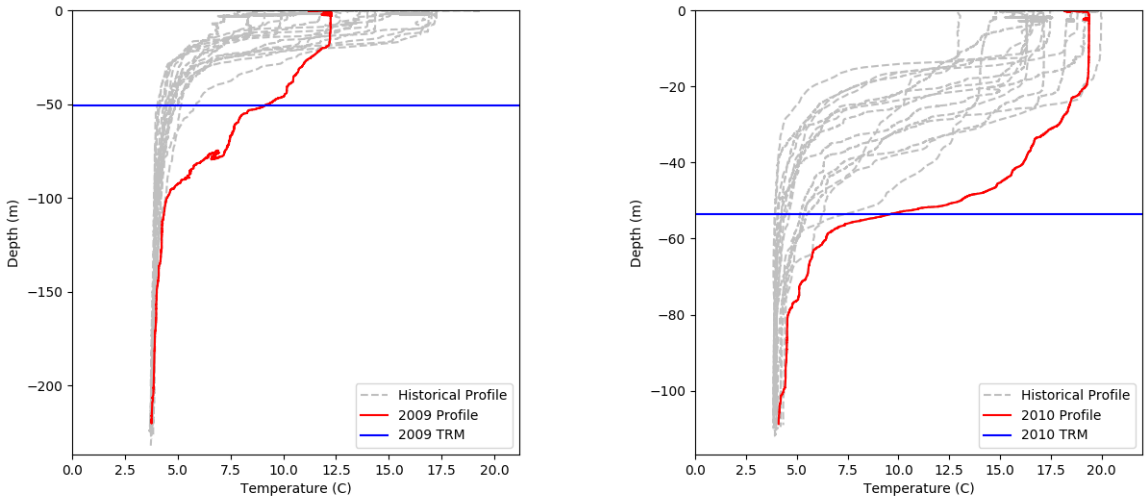


(c) Glyph-maps of DCL concentration (Range: 0.08 - 8.17 ug/L)

The thermocline in Lake Superior ranges from 4.0 meters to 60 meters. The thermo-stratifications in Lake Superior are heavily affected by wind, current, and corresponding upwelling (when warm surface water is pushed offshore by wind and cold bottom water rises toward the surface) and downwelling events (warm surface water is pushed toward the shore and sinks to the bottom). Specifically, the northwest shore (e.g. station SU16, 18, 19) had consistently shallow thermoclines, which could be related to low temperatures caused by the eastward or southward currents and resulting north shore upwelling events (Bennington et al., 2010, Bennett, 1978). In the eastern basin, thermoclines are generally deeper in east south shore areas (e.g. SU21, SU05 and SU23), which may be due to a strong anticlockwise current (Bennington et al., 2010) and downwelling events caused by Ekman drift, in which surface water moves toward the prevailing wind directions in the northern hemisphere (Emery and Csanady, 1973).

Some stations have fluctuating TRM. For example, Station SU15 in 2009 (Figure 3.18a) and SU21 in 2010 (Figure 3.18b) have different profile shapes compared to other years, with deeper thermoclines. Such different temperature profiles could be due to upwelling or down-welling events that happened at the sampling time. Station SU21 is heavily influenced by the southward current to the east of Keweenaw Peninsula (Bennington et al., 2010), which can lead to down-welling events and a deepened thermocline. The fluctuations could also be due to

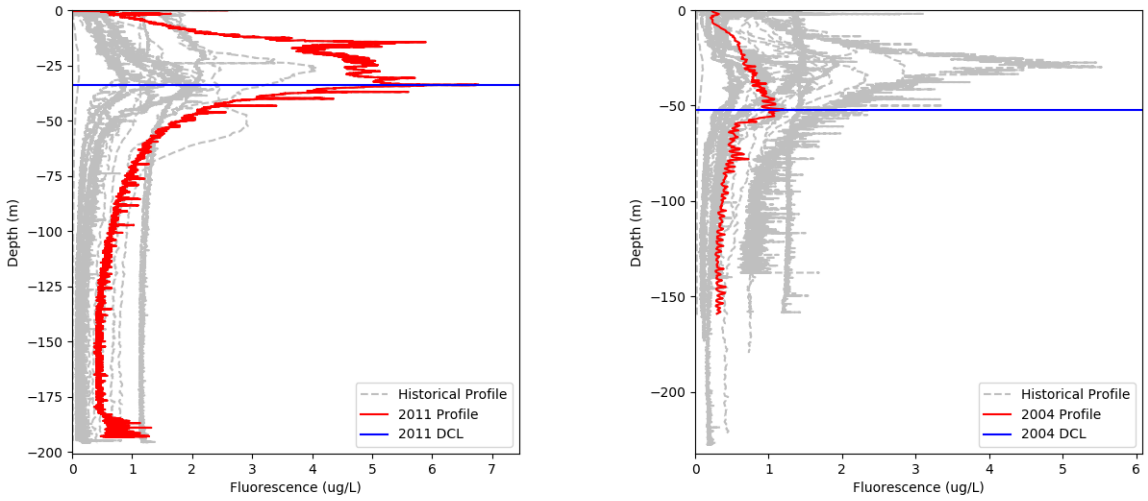
differences in total heat absorbed during that year, which is further discussed in Section 3.5.2.



(a) SU15 (b) SU21

Figure 3.18. Temperature profile of Stations SU15 and SU21. Red lines represent thermocline profiles that are different from other years.

Most of the DCL depths from 1996 to 2013 are relatively stable (Figure 3.17b). However, Station SU23 has an increasing DCL depth in this period. For DCL concentrations (Figure 3.17c), almost all of the stations experienced two spikes in fluorescence concentrations (i.e. algae bloom) during this period. One spike occurred in 2000 and the other in 2011 (e.g. Figure 3.19a at SU07). White and Matsumoto (2012) used a three-dimensional numerical model to study the DCL in Lake Superior. They showed that photoadaptation and the location of the nutricline (the depth where there is a rapid change in nutrients) are the primary factors determining DCL depths and concentrations, while zooplankton grazing and phytoplankton sinking secondarily affect DCL concentrations. Deeper nutriclines are also related to a longer stratification period (Barbiero and Tuchman, 2004). Therefore, the observed fluctuations in DCL depths (e.g., SU05 in Figure 3.19b) could be closely related to changes in light availability and lake stratifications. The spikes in DCL concentrations during 2000 and 2011 could be related to more nutrient inputs, most likely phosphorus as Lake Superior is usually phosphorus deficient.



(a) SU07 (b) SU05

Figure 3.19. Fluorescence profile of Stations SU07 and SU05. Red lines are profiles that are different in a specific year compared to other years.

3.5.2 Trend in Heat Storage from Spring to Summer in Lake Superior

The detected lake stratification can also be used to compute heat absorbed into the Great Lakes between spring and summer, which reflects changes in evaporation, lake freezing, and water temperature (Derecki, 1976). Previous research has examined trends in lake surface temperatures in Lake Superior, which have been increasing since 1979, most likely due to declining winter ice cover and increases in air temperature and wind speed (Austin and Colman, 2007).

Changes in heat storage Q_t from time t_1 to t_2 can be computed as $Q_t = (V_2T_2 - V_1T_1)$ where V_i, T_i are the volume of the lake and average temperature at time t_i (Derecki, 1976). In this study, we calculated the heat storage change of a water column with unit cross area at each sampling site from spring to summer to assess heat absorbance during this period. The equation to calculate heat storage Q_i at each station from spring to summer is:

$$Q_i = \int_{3m}^{h_i^{UHY}} (T_i^{summer}(h) - T_i^{spring}(h))dh \tag{3.3}$$

where h_i^{UHY} is the mean depth of the upper hypolimnion (UHY) at station i from 1996 to 2013 detected by the PLR algorithm (Section 3.3.2). $T_i^{summer}(h), T_i^{spring}(h)$ are the temperatures at depth h in summer and spring at station i, respectively. Thus Q_i computes the accumulated differences between the spring and summer temperature depth profiles above the mean UHY in

the same year, which represents the cumulative effects of heating in the upper lake. The h_i^{UHY} for different stations are summarized in Appendix B.

Figure 3.20 summarizes the local trends in Q_i at each sampling site in Lake Superior. A positive trend in Q_i means that the deep water is becoming increasingly heated from 1996 to 2013. Stations 03 (p = 0.04), 05 (p = 0.10), 07 (p = 0.10), 09 (p = 0.03), 10 (p = 0.02), 11 (p=0.04), 16 (p=0.04) all have positive and significant linearly increasing trends from 1996 to 2013. The stations in the east basin particularly show a positive trend for the entire area. Since EPA sampling dates are roughly the same every year and are after the ice covering period, such a positive trend could be related to increased air temperature and wind speed and thus more heat absorbed by the lake (Austin and Colman, 2007).

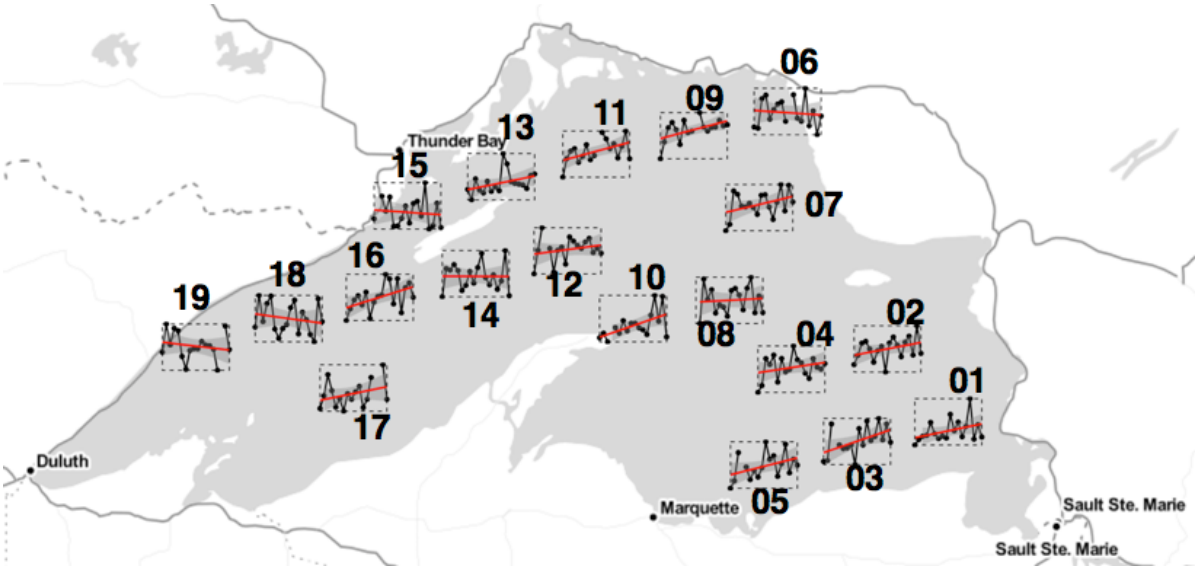


Figure 3.20. The local trend in heat storage changes from Spring to Summer at each site from 1996 to 2013. The values are scaled to [0,1] within each station. Red is the linear trend and the shaded area represents the 95% confidence interval. Discontinuities occur when no spring data were available that year.

Note that Station 08 does not show a positive trend. Such an exception is probably due to a low average current speed since Station 08 is located near the center of the counterclockwise gyre (Bennington et al. 2010). Bennington et al. showed that external environmental changes (e.g. increased wind speed and air-temperature) may have less effects on heat absorbance with low speed currents.

In the western basin, a positive trend is not obvious in some loggers, meaning that heat storage changes in the western basin from spring to summer are more consistent. Thus, the

heating of the western basin described by Austin and Colman (2007) may be caused by declining winter ice cover, which leads to longer periods of solar radiation absorbance before the spring sampling. In addition, upwelling events that occurred in the northwest shoreline (Bennington et al., 2010, Bennett, 1978) may complicate lake stratification patterns and water column heating processes, thus resulting in the varying trends seen in Figure 3.20.

3.6 Discussion and Conclusions

In this chapter, we extend PLR algorithms and propose a peak detection algorithm to automatically identify stratification and DCL patterns, providing a consistent reference to identify lake features. The algorithms generally produced similar results to human judgment, with squared correlation coefficients around 0.6 for LEP, TRM, and DCL detection. Differences between human judgment and the algorithm results revealed inconsistencies in the operators' judging criteria, such as using DCL as UHY and identifying TRM at a different depth from the sharpest gradient. The algorithms are also able to highlight anomalous patterns such as double thermoclines and unusual peak shapes. A visualization of the general patterns of lake stratification and DCL in Lake Superior revealed how the algorithms can be used to provide insights on spatiotemporal changes in lake processes.

Currently, the historical profiles are stored in databases, with some applications available to provide basic querying and visualization (e.g. https://greatlakesmonitoring.org). The characterization algorithms proposed here will provide more advanced functions to more effectively utilize the profile data. For example, the algorithms can aid users by allowing more queries such as identifying the DCL shape or multiple peaks. The approach can also help automatically detect unusual patterns in newly uploaded profiles by comparing the features detected with those from existing historical data at the same station.

This type of characterization of depth profiling data benefits future sampling activities by: (1) providing a consistent reference for detection of lake stratification patterns and deep chlorophyll layers, as well as comparisons with historical notes to highlight unusual patterns and correct notes if needed; (2) detecting additional features such as symmetry of DCL shapes and comparison of shapes from current and historical profiles, which could benefit adaptive sampling activities to immediately collect more data on unusual profiles; and (3) the Web application will facilitate operators and researchers without programming knowledge to easily use the algorithms.

It should be emphasized that the algorithms are not intended to replace human judgment, but rather to provide rapid information for applying judgment where needed, such as making assessments during sampling activities and extracting general patterns from a large database.

Recommendations on future research to extend this work are given in Chapter 5.

CHAPTER 4: SPATIO-TEMPORAL ANALYSIS OF HYPOXIA EXTENT IN LAKE ERIE

Looking deeper from the depth profiling data, in this chapter we analyze dissolved oxygen (DO) concentration data sampled from a lake bottom sensor network. To estimate hypoxia extent in the lakes from these data, we develop a spatio-temporal interpolation method with conditional simulations and a Bayesian framework to address uncertainty. The DO data sampled in Lake Erie in 2014, 2015, and 2016 are used as a case study. Section 4.1 gives a brief introduction to hypoxia events and previous research on modeling spatio-temporal data in lakes. Section 4.2 describes the data sources, while Section 4.3 gives the methodology. Section 4.4 presents the results, including DO data patterns, cross-validation of the interpolation methods, and estimations of hypoxia extent. Further discussion on the impacts of model parameters and sampling location optimization is provided in Section 4.5, followed by conclusions in Section 4.6.

4.1 Introduction

Hypoxia (dissolved oxygen [DO] concentrations lower than 2mg/L) is a major water quality issue in many lakes and estuaries. In this work, methods for analyzing hypoxia extent are developed and tested in central Lake Erie using DO data sampled via a sensor network. Lake hypoxia phenomena and previous spatio-temporal data analysis are introduced below in Sections 4.1.1 and 4.1.2, respectively.

4.1.1 Hypoxia in Lakes

Hypoxia often starts with eutrophication. Eutrophication caused by input of nitrogen and phosphorus leads to algae blooms. When the algae die, they sink towards the bottom of the lake and bacteria decompose the cells, which consumes oxygen. If oxygen concentrations become too low (hypoxic), many fish and some invertebrates will die.

Two principal factors for hypoxia development are water column stratification and decomposition of organic matter in the bottom sediment (Diaz, 2001). The oxygen in the hypolimnion is supplied by DO flux through the thermocline and is consumed by hypolimnetic oxygen demand (HOD) and sediment oxygen demand (SOD). HOD is related to photosynthesis, respiration, and decomposition (Edwards et al., 2005; Rucinski et al., 2010) and is correlated

with the thickness of the hypolimnion (Bouffard et al., 2013) and temperature (Rucinski et al., 2010). SOD can be estimated through experiments (Smith and Matisoff, 2008) and diffusion process modeling (Matisoff and Neeson, 2005). Models that link hydrodynamics and eutrophication have been developed to explore the influence of nutrient loads on hypoxia in Lake Erie. Examples include a one-dimensional (1-D) model developed by Rucinski et al. (2014), a 2-D model developed by Zhang et al. (2008), and a 3-D model developed by Di Toro and Connolly (1980) and Bocaniov et al. (2016).

Hypoxia events also directly affect fish communities. Rapid changes in DO may cause the death of fish that are trapped in hypoxia zones, especially when seiche events bring hypoxic waters to nearshore areas (Scavia et al., 2014; Rao et al., 2014). Fish may also avoid the hypoxia zone, which leads to a position shift of preferred diets and results in a change in the whole food web (Scavia et al., 2014).

In Lake Erie, the case study explored in this work, the western basin is eutrophic but seldom experiences hypoxia because it is shallow and has fully mixed water columns, although severe hypoxia may last for a few days annually. The oligotrophic and deep East basin has a thick hypolimnion column during a summer stratification period that leads to enough DO storage to prevent hypoxia. The Long Point – Erie Ridge also prevents hypoxic hypolimnion water in the central basin moving into the eastern basin. The mesotrophic central basin suffers the most from hypoxia due to its thin hypolimnion, so that oxygen demands often exceed replenishment rates.

Hypoxia in Lake Erie has been slow to improve in response to nutrient load reduction (Diaz, 2001; Bouffard et al., 2013). Since 1972, point-source phosphorus reduction has been implemented, but the extent of hypoxia in Lake Erie remained similar between 1970 and 1990 (Diaz, 2001). Further, since the mid-1990s, Lake Erie again suffered from cyanobacteria blooms and extensive hypoxia. Scavia et al. (2014) did a thorough review on such re-eutrophication of Lake Erie and proposed approaches to achieving new loading targets. They concluded that management actions are needed to reduce both total phosphorus and the more highly bioavailable dissolved reactive phosphorus (DRP). Efforts are needed to address non-point sources and consider the potential impacts of climate change.

4.1.2 Spatio-temporal Data Analysis and Interpolation

The hypoxia extent, including hypoxic area and hypoxic time, are related to nutrient loadings (Rucinski et al., 2014). To understand hypoxia development, a sensor network can be

deployed at the lake bottom to measure DO across time in different locations. Since only limited sensors are placed, interpolation is needed to estimate the values at unsampled locations.

Interpolations should utilize the spatial and temporal correlations. Previous research has modeled such spatio-temporal data using spatio-temporal kriging in other environmental systems, such as interpolating PM10 in atmosphere (Gräler et al., 2016) and soil water content (Snepvangers, et al., 2003). Another approach when time is discrete is the dynamic spatio-temporal models (DSTM) that model the process and relations at time t given time $t-1$, such as using a first-order vector autoregression model, leaving only the residuals that have spatial correlations (Bakar and Sahu, 2015). However, both models assume a determined spatio-temporal trend so that the spatio-temporal residuals satisfy temporal and/or spatial stationarity. Therefore, when there is no such spatio-temporal trend that can be easily extracted, the above methods are not applicable.

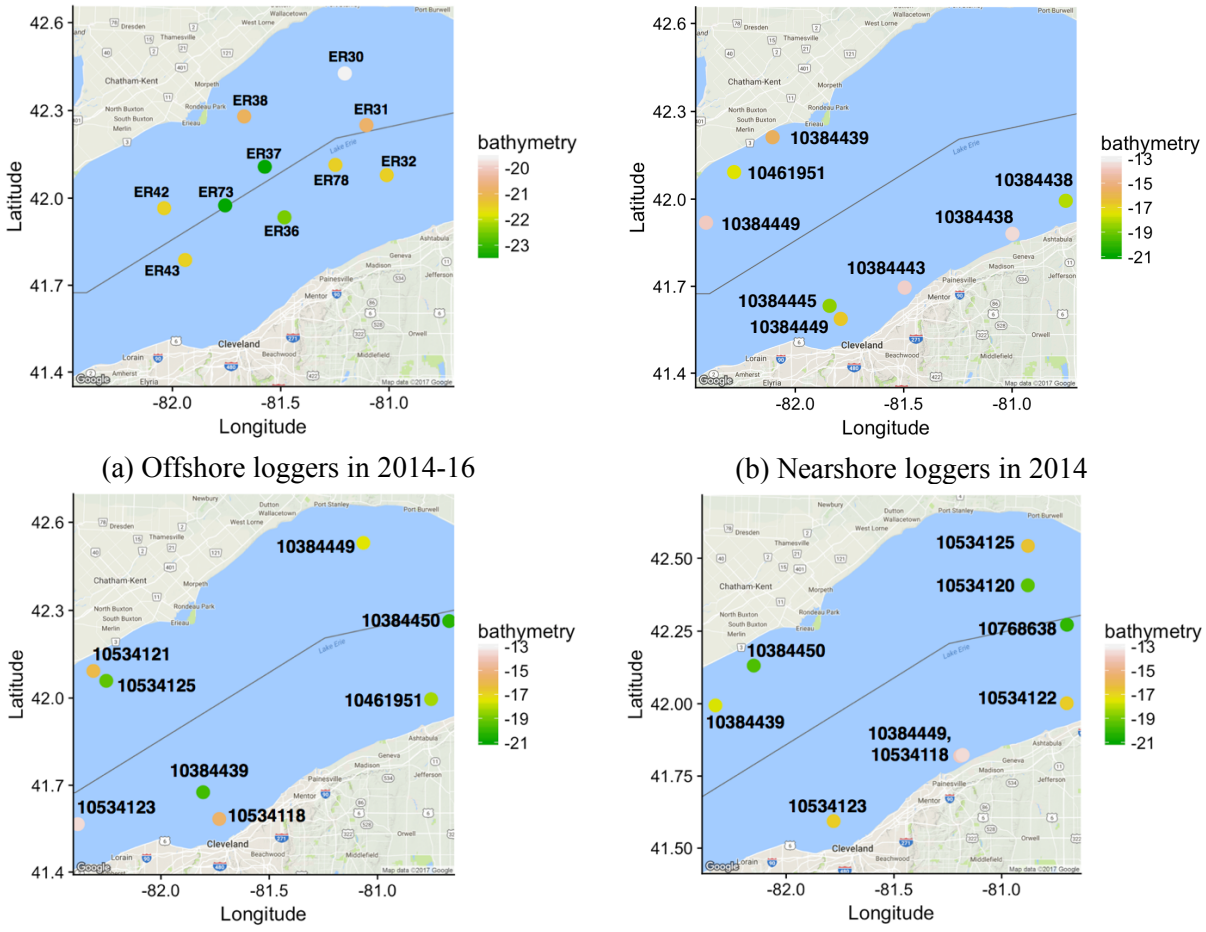
A more flexible approach from Lindström et al. (2014) is to model the data as a set of spatially varying temporal basis functions, or empirical orthogonal functions (EOF). Those temporal basis functions account for the temporal variability in data and are associated with spatially varying coefficients. Those coefficients are interpolated by universal kriging and are used to reconstruct the final interpolation at target locations. In this way, it is the temporal patterns that are to be interpolated in space.

In this chapter, we combine this decomposition-interpolation framework with previous research using conditional simulation (Zhou et al. 2013) to estimate hypoxia. We propose a spatio-temporal interpolation method to consider uncertainty in estimating hypoxia extent with conditional simulations. A Bayesian framework is incorporated to also consider interpolation model uncertainty. We apply and cross-validate the method with DO data in Lake Erie sampled in 2014, 2015, and 2016. Using these methods, seasonal changes to hypoxia extent in Lake Erie are characterized. An interactive Web application is also developed to allow other researchers to explore this and other DO datasets and estimate hypoxia extent.

4.2 Study Area and Data Description

In order to estimate hypoxia extent in Lake Erie, USEPA deployed a sensor network in 2014, 2015, and 2016 that continuously sampled bottom DO in the central basin of Lake Erie. Ten loggers in the offshore areas remained in the same sampling position in all three years

(Figure 4.1a). The locations of nearshore loggers changed every year (Figure 4.1b to d) and were usually deployed at intersections between the lake bed and thermocline. The loggers were deployed in early summer and retrieved in early fall and the deployment and retrieval dates are different for nearshore and offshore loggers. Offshore sampling locations usually have two loggers, one at the lake bottom and the other several meters above the lake bottom (3 meters for 2014 data, 1.5 meters for 2015 data, and 0.5 meters for 2016 data), while the nearshore loggers are only placed at the lake bottom. All loggers measured and recorded in-situ DO concentrations and temperatures every 10 minutes.



(a) Offshore loggers in 2014-16

(b) Nearshore loggers in 2014

(c) Nearshore loggers in 2015

(d) Nearshore loggers in 2016

Figure 4.1. Sampling logger deployment locations. Logger indices are labeled at each sampling location. Bathymetry data are from the National Oceanic and Atmospheric Administration (NOAA).

4.3 Methodology

The methodology is summarized in Figure 4.2 and described in more detail below. After data preprocessing (Section 4.3.1), an initial DO pattern analysis is performed to identify basic DO patterns. Then, spatio-temporal interpolations are implemented to estimate the hypoxia extent. The hourly averaged data are then interpolated at unsampled locations during the time period when all nearshore and offshore loggers have available data in each year. The simplest spatio-temporal interpolation is to use inverse distance weighting (IDW) spatial interpolation at every time step (every hour in our case, Section 4.3.2), which serves as the baseline method. As an alternative, we give a basis interpolation method (Section 4.3.3) that uses singular value decomposition to extract the temporal patterns, which are then spatially interpolated by kriging interpolation. Two methods are implemented to model the kriging variogram: (a) maximum likelihood estimation (MLE) and (b) Bayesian framework. Hypoxia estimation is described in Section 4.3.4.

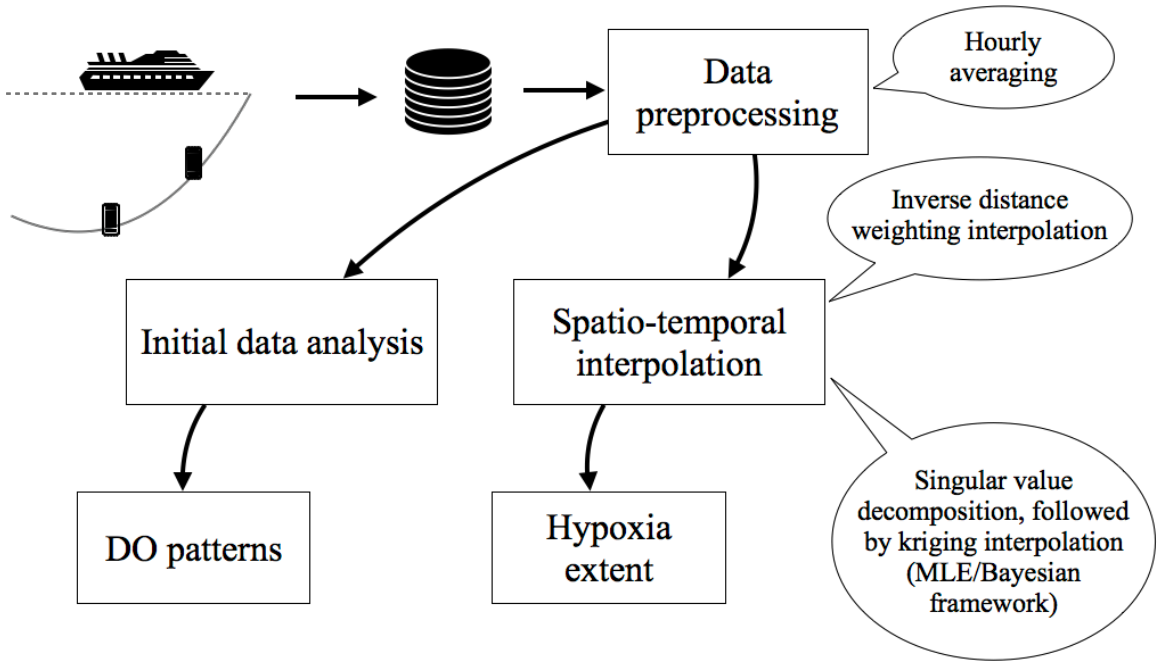


Figure 4.2. Hypoxia analysis framework and workflow

4.3.1 Data Preprocessing

Each logger starts measuring dissolved oxygen levels at different times. To standardize the data for analysis, we perform hourly averaging of the raw logger data. We also convert the longitude and latitude to Universal Transverse Mercator (UTM) using “WGS84” datum and zone

number “17T”. After this preprocessing, some spikes still exist in the data at the hourly scale. Nonetheless, no additional smoothing is undertaken because the spikes may indicate important lake processes such as the intrusion of surrounding hypoxic water or seiche events that mix the water column quickly.

4.3.2 Spatio-temporal IDW Interpolation

The inverse distance weighting (IDW) is a commonly used deterministic spatial interpolation technique. The target location is interpolated using a weighted average of values from other sampled locations and the weights are computed based on the inverse of the distance between the target locations and sampled points. The mathematical equations are:

$$y(x) = \begin{cases} \dfrac{\sum_{i=1}^{N} \omega_i(x) y_i}{\sum_{i=1}^{N} \omega_i(x)} & (d(x, x_i) \neq 0) \\ u_i & (d(x, x_i) = 0) \end{cases} \tag{4.1}$$

Where $y(x)$ is the interpolated value at target location x. N is the total number of sampled points surrounding x that are considered.

The weights ω are calculated as:

$$\omega_i(x) = 1/d(x, x_i)^2 \tag{4.2}$$

where $d(x, x_i)$ is the spatial distance between the target location x and sampled location x_i. We choose $N = 5$, meaning for a target location, the five nearest sampling loggers were used to interpolate. To extend IDW into temporal dimension, we performed IDW interpolation at every time step.

4.3.3 Spatio-temporal Basis Interpolation

As an alternative to IDW, we use spatio-temporal basis interpolation, following Lindström et al. (2014). DO data are then modeled using the form of:

$$y(s, t) = \mu(s, t) + \epsilon(s, t) \tag{4.3}$$

where $y(s, t)$ denotes the spatio-temporal DO values, $\mu(s, t)$ is the mean trend, and $\epsilon(s, t)$ is the space-time residuals. s, t are the space and time index, respectively.

The basis interpolation algorithm consists of the following steps (Figure 4.3):

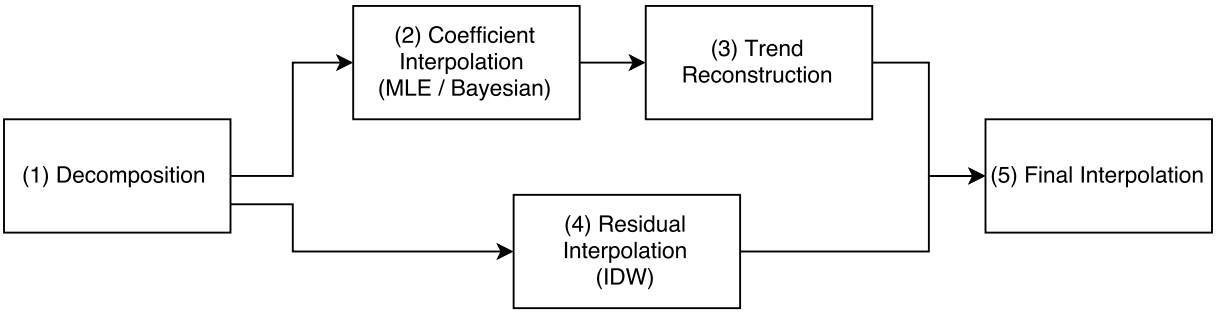


Figure 4.3. Workflow for spatio-temporal basis interpolation

(1) Decompose the mean trend term $\mu(s,t)$ using singular value decomposition (SVD) to obtain a temporal basis function and corresponding basis coefficients;

(2) For coefficients in each basis function, conduct universal kriging on the target grid using maximum likelihood estimation or Bayesian framework to fit a variogram model;

(3) Reconstruct the interpolated trend term $\mu(s,t)$ at the target grid from the interpolated coefficients, calculated from (2);

(4) Conduct spatio-temporal IDW interpolations (Section 4.3.2) on the residuals $\epsilon(s,t)$;

(5) Add the reconstructed trend and interpolated residuals as the final interpolated values.

More details on extracting the temporal basis function and conducting interpolations are given below.

Temporal Basis Function

As noted in Step 1 above, we use singular value decomposition (SVD) to obtain the basis functions (Lindström et al., 2014). Such techniques are also used in computer vision in facial recognition, where the basis is called eigenfaces (Sirovich and Kirby, 1987)

The SVD decomposition gives:

$$D' = U\Sigma V^T \tag{4.4}$$

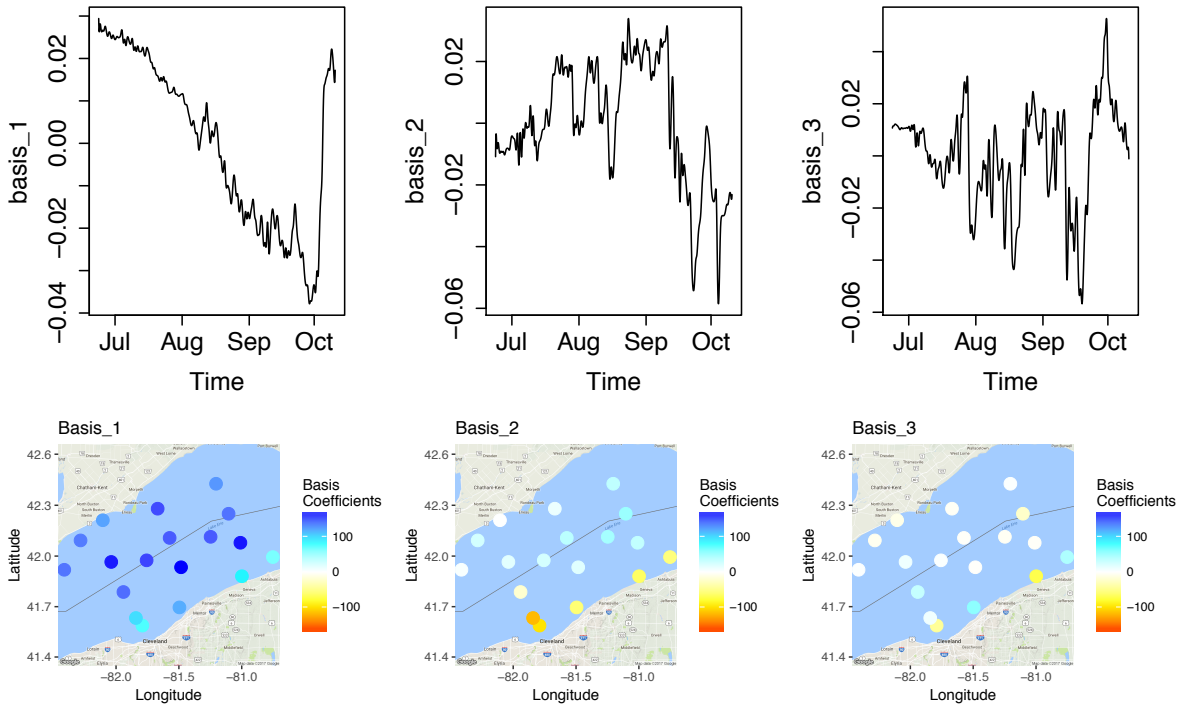
where D' are the raw data matrix that is column-wise centered and scaled to mean zero and unit variance, with element $d_{i,j}$ representing the centered and scaled DO values at time i from logger j. i and j are the row and column index. The first r columns of matrix U are the temporal basis functions which are kept and then smoothed using a 1-D spline. The basis functions are the fundamental "bricks" to build the logger data time series. They represent the mode or basic patterns. A simple example of the concept of basis is: in a two-dimensional space, vectors (0,1) and (1,0) are two basis vectors that can form any two-dimensional vectors.

We conduct a linear regression to obtain the coefficients of the basis function, so that:

$$y(s,t) = \sum_{k=1}^{r+1} \beta_k(s) f_k(t) \; + \; \epsilon(s,t) \tag{4.5}$$

where $y(s,t)$ are the DO values at sampling location s at time t, $f_k(t)$ is the k^{th} temporal basis function (i.e. the smoothed k^{th} column vector of matrix U), and $\beta_k(s)$ are the coefficients of $f_k(t)$ which vary across space. $f_k(t)$ and $\beta_k(s)$ can be positive and negative. The $(r+1)^{th}$ temporal basis function is the intercept term for which $f_{r+1}(t) \equiv 1$. $\epsilon(s,t)$ are the residuals. The function $f_k(t)$ is also called the empirical orthogonal function (EOF) and has been used for analyzing the modes in other spatio-temporal datasets such as sea level pressure (Hannachi et al., 2007) and DO in Chesapeake Bay (Scully, 2016). The number of temporal basis functions r in the SVD also needs to be pre-defined; cross-validation can be used to help choose the best r (following Lindström et al. [2014]).

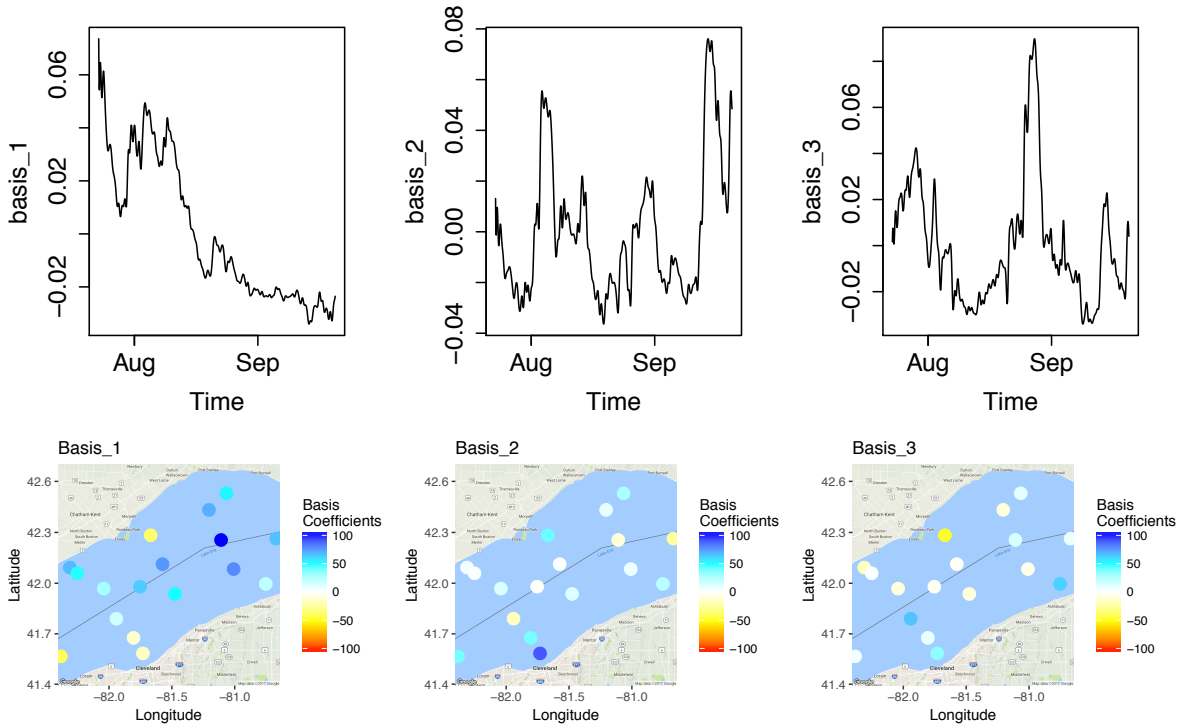
Figure 4.4 shows the first three temporal basis functions as examples from Lake Erie DO data in 2014, 2015 and 2016.



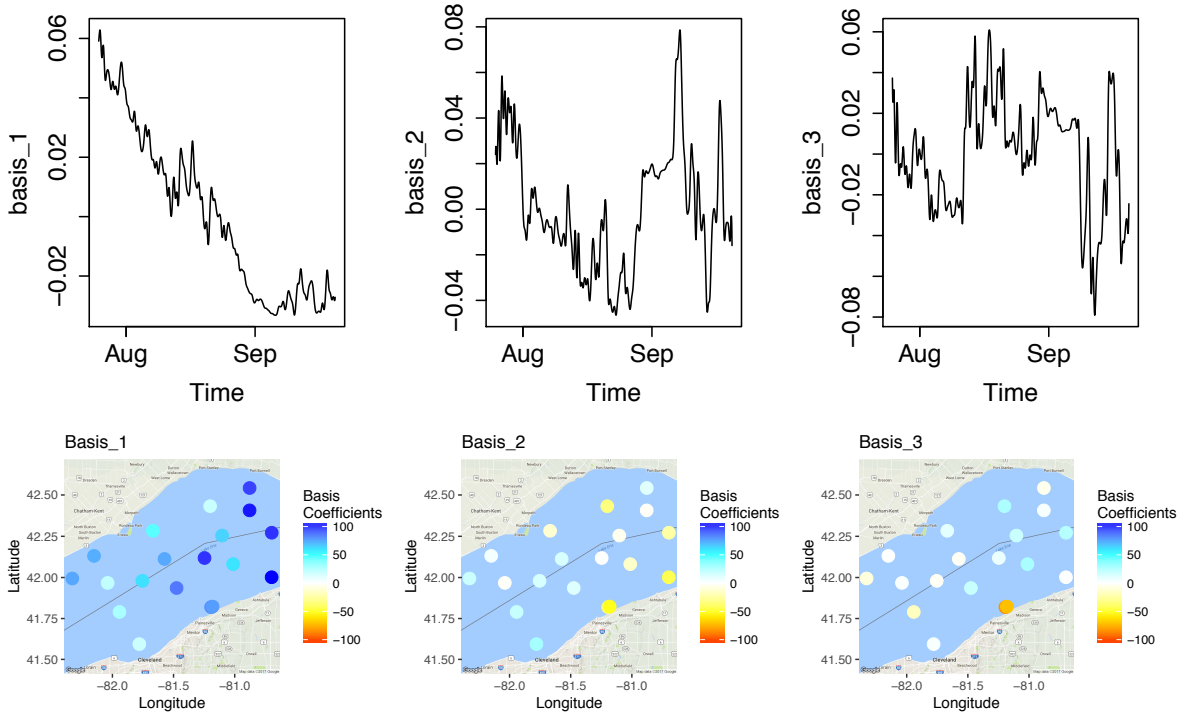
(a) Temporal basis functions $f_k(t)$ (above) and corresponding coefficients $\beta_k(s)$ (below) in 2014

Figure 4.4. Basis decompositions in 2014, 2015, and 2016.

Figure 4.4 (cont.)



(b) Temporal basis functions $f_k(t)$ (above) and corresponding coefficients $\beta_k(s)$ (below) in 2015



(c) Temporal basis functions $f_k(t)$ (above) and corresponding coefficients $\beta_k(s)$ (below) in 2016

The basis functions can provide insights on the data, particularly when they are ordered by the magnitude of eigenvalues. The basis with the largest magnitude of eigenvalues, i.e. Basis_1 (leftmost of the top row in Figure 4.4a, b and c), explains the most variance in the data. Different values of coefficients represent the extent to which such patterns are contained in the data series. Therefore, the magnitude of the coefficients of Basis_1 are the largest for data from most loggers (leftmost of the bottom row in Figure 4.4a, b, and c). With positive coefficients, Basis_1 reveals the most fundamental patterns: In this case, DO decreased during the sampling period in all three years, with 2014 having a sharp increase in October and 2015 having a large fluctuation at the beginning of August.

Basis_2 and remaining basis functions represent more detailed patterns, with coefficients generally becoming smaller and smaller, indicating that these patterns are less and less significant. Loggers with similar coefficients also exhibit particular DO patterns. For example, the south shore loggers have different patterns than the other loggers in 2014 (Figure 4.4a), as the coefficients of Basis_2 are negative while other loggers have positive coefficients. Some loggers have similar magnitude of coefficients for all basis functions (e.g., ER38 in 2015, see Figure 4.4b), which indicates that the patterns of these loggers are different and may not be accurately approximated with only a few basis functions.

Basis_1 mostly represents the offshore logger patterns and the values of the Basis_1 functions across different years are not comparable since they have been normalized in each year. We will discuss more details on various patterns of the offshore loggers and cross-year comparisons in the results section (Section 4.4.1).

Coefficient Kriging Interpolation

After obtaining the basis functions $f_k(t)$, the coefficient $\beta_k(s)$ in Eq. (4.5) are then interpolated using universal kriging (Step 2 in Figure 4.3). The detailed equations of universal kriging are illustrated in Appendix A. The covariates considered in the universal kriging interpolation include x and y in UTM system, bathymetry, and the square of the bathymetry. The square of the bathymetry is included because Zhou et al. (2013) found it important according to Bayesian information criterion (BIC) in their kriging interpolation on DO estimations.

We use two approaches to estimate the variogram model: maximum likelihood (MLE) and Bayesian framework. MLE is used to numerically maximize the likelihood of the Gaussian process (i.e., kriging). Bayesian kriging assumes a prior distribution of the variogram model

parameters and finds the parameter posterior distributions, based on which the expected prediction values are calculated. In this study, we use an "exponential" variogram model (i.e., covariance function).

To implement these approaches, we use "geoR" package in R (Ribeiro Jr and Diggle, 2016). The priors for the parameters of the exponential variogram model are summarized in Table 4.1. We assume the measurement errors are zero for simplicity, otherwise the nugget (τ) also needs a prior distribution that is unknown. The reciprocal distribution for sill (σ) and uniform distribution for covariates coefficient (β) are the default settings in the "krige.bayes" function in the "geoR" package. Choosing other priors did not have much influence on the interpolation results, as discussed in Section 4.5.1.

Table 4.1. Prior distributions for the variogram model

Parameter	Prior	Justification
Covariates coefficient (β)	Flat	We have no prior knowledge of this parameter
Sill (σ)	Reciprocal distribution (p(σ^2) is proportional to $1/\sigma^2$).	No parameters of prior distributions are needed.
Nugget (τ)	Fixed as 0	We assume the measurement errors are zero.
Range (ϕ)	Uniform distribution ranging from 20 to 70 km at 5 km intervals	Ranges are approximately 1/3 of the maximum pairwise distance (recommended default values in Golden Software and "gstat" R package)

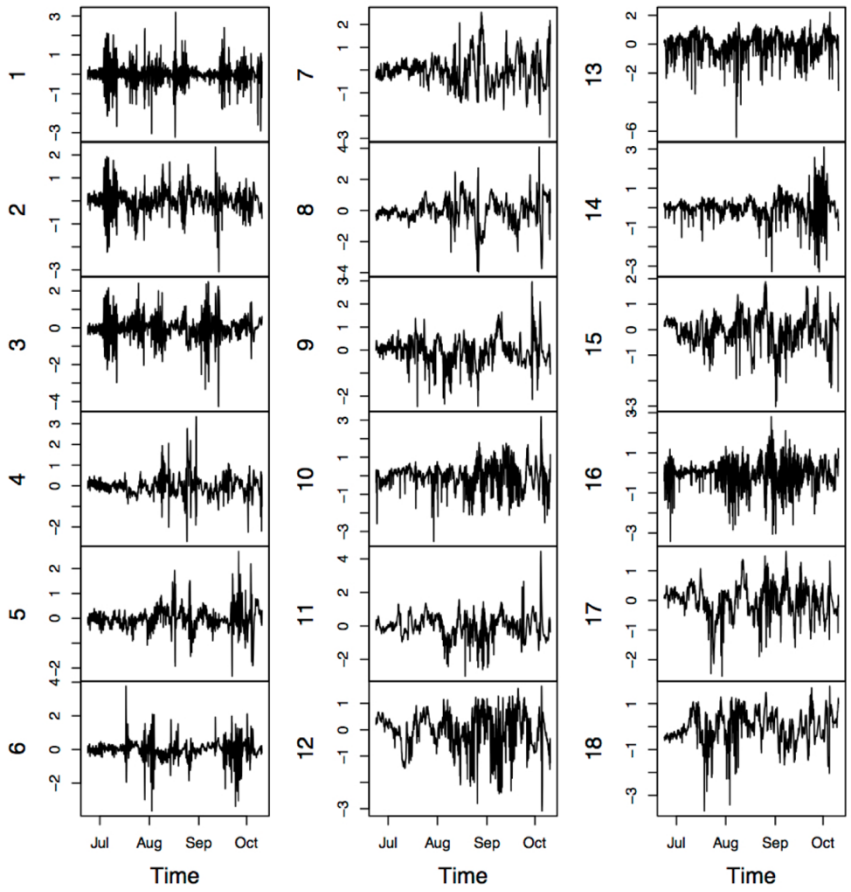
More details on the Bayesian kriging approach implemented in "geoR" are provided in the Appendix A. The number of basis functions determines how many spatial kriging interpolations are conducted. The trend predictions at a target cell from interpolated basis coefficients are reconstructed by applying Eq. (4.6) again, where $\beta_k(s)$ is now the interpolated coefficient at grid cell s and $\epsilon(s,t)$ is the interpolated residual, which is described in the next subsection. When the reconstructed DO is less than 0, we set it to zero.

Each kriging interpolation can generate the expected prediction values given the prediction distribution. Using conditional simulations, it can also generate multiple predictions on the target grid that satisfy the spatial correlations defined by the variogram. The equations to perform conditional simulations are provided in Appendix A. We denote $\hat{\beta}_k\ (s)$ as the expected prediction value and $\hat{\beta}_k^i(s)$ as the i^{th} prediction simulation.

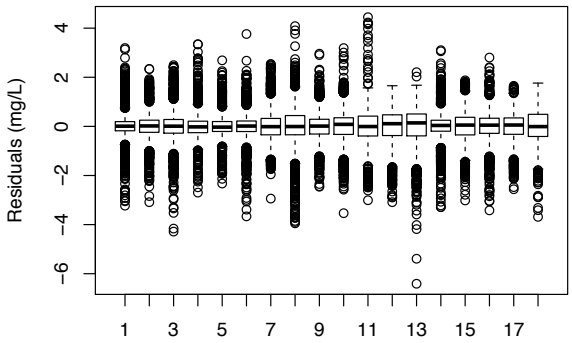
Residual Interpolation

After interpolating the coefficients, Step 4 in Figure 4.3 is next, interpolating residuals. The residuals $\epsilon(s,t)$ from Eq. (4.5) (Step 1 in Figure 4.3) need to be analyzed to find the appropriate method to interpolate. Figure 4.5 shows the residuals from all loggers when $r = 10$ (i.e., total of 11 basis functions, including the intercept basis) in 2014 as an example. Residuals are generally small with mean close to 0 and small 25% and 75% quantile values [Figures 4.5(a) and (b)], implying that they have insignificant effects on the interpolation results. The spatio-temporal variogram (Figure 4.5c) reveals that the residuals still have some temporal correlation as the semivariogram increases with time at spatial distance = 0. However, no global spatial correlation patterns are obvious since the semivariance quickly jumps to the plateau stage.

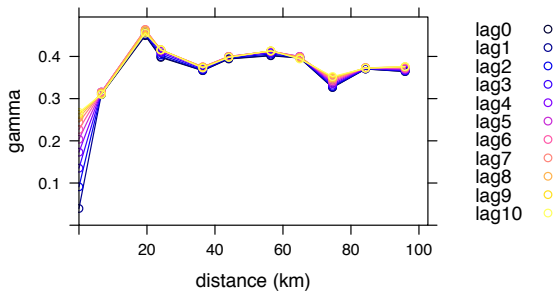
Yet completely ignoring the residuals may introduce errors since there are some spikes that may influence surrounding areas and some locations are still correlated (e.g., the first three loggers still have similar patterns; see Figure 4.5a). As no global spatial correlations are detected, we don't need to fit a complicated spatial variogram or spatial-temporal variogram as Lindström et al. (2014) did. Instead, we perform a simple spatio-temporal IDW interpolation (Section 4.3.2) on residuals. Since the IDW is a linear combination of values of surrounding points, the temporal correlations of logger data residuals are naturally preserved at the interpolated residuals.



(a) Residuals of the logger data



(b) Spatio-temporal variogram



(c) Boxplot of the residuals

Figure 4.5. Residuals statistics after detrending with 10 basis functions (r = 10) for hourly data in 2014. (A) residuals left for each logger; (b) Spatio-temporal semivariance (y axis) with difference spatial distance (x axis) and temporal lags (different colors: lag0 means no difference in time dimension; lag1 means 1 hour differences)

4.3.4 Hypoxia Extent Estimation

Lastly, in Step 5, hypoxia extent is estimated by interpolating DO values across the target area using both IDW and kriging interpolation methods. In the Lake Erie case study, we

interpolate DO on a grid with size 0.025 degree in longitude and latitude in the central basin of Lake Erie. Thus, one grid cell occupies approximately 5.6 km^2. The grid is also filtered by a convex hull defined by the sampling locations (Figure 4.1) so that only interpolations rather than extrapolations are conducted.

For IDW interpolation, we calculate $N_h(t)$, the number of grid cells $N_h(t)$ at time t where the interpolated DO was below the hypoxic threshold (e.g. 2mg/L). The number of cells are then converted to areas to obtain a time series of the hypoxic areas. The spatio-temporal IDW interpolation cannot give interpolation uncertainty so that the hypoxia extent estimation is also deterministic.

For kriging interpolation, the prediction at the target location follows a probability distribution (Gaussian distribution for traditional kriging and different distributions for Bayesian kriging, depending on the prior distributions). As a result, we are able to estimate the uncertainty of DO interpolations as well as the uncertainty of the estimated hypoxic areas by conditional simulations. The detailed steps are:

Step 1: In the kriging interpolation on the coefficients of each temporal basis function, generate 100 possible prediction simulations using conditional simulations.

Step 2: For each basis function, independently and randomly select a prediction realization among the 100 simulations. Reconstruct the spatial-temporal DO interpolations (Section 4.3.2, reconstruct the trend from the interpolated coefficients and add back the IDW residuals interpolation).

Step 3: Calculate the areas where interpolated DO values are smaller than the hypoxic threshold (e.g., 2mg/L) in this realization along time.

Step 4: Repeat Steps 2 and 3 1000 times. Then we have 1000 time series of the hypoxic area.

The 1000 time series are summarized to estimate the hypoxia extent with uncertainty. We choose 5% to 95% quantiles as the confidence interval for the kriging methods.

4.3.5. Program Design and Web Applications

To facilitate rapid computations, we implemented parallel programming in the spatio-temporal IDW interpolation (Section 4.3.1) so that the program computes interpolations at different time steps simultaneously. In addition, when performing conditional simulations to estimate hypoxia uncertainty, we calculated and saved only the hypoxia extent from each

realization and discarded detailed interpolation values. This saves significant computing memory and enables estimation of hypoxia extent and uncertainty on a modern laptop.

We also built an interactive web application using R "Shiny" packages (Chang et al., 2017) to explore the DO data and calculate hypoxia extent (Figure 4.6). We installed the RStudio Server on Amazon Elastic Compute Cloud (EC2) in one instance, and the DO data are stored in a MySQL database that is hosted using Amazon Web Services (AWS). The source code can be found at http://stormxuwz.github.io/Hypoxia_Lake_Erie/.

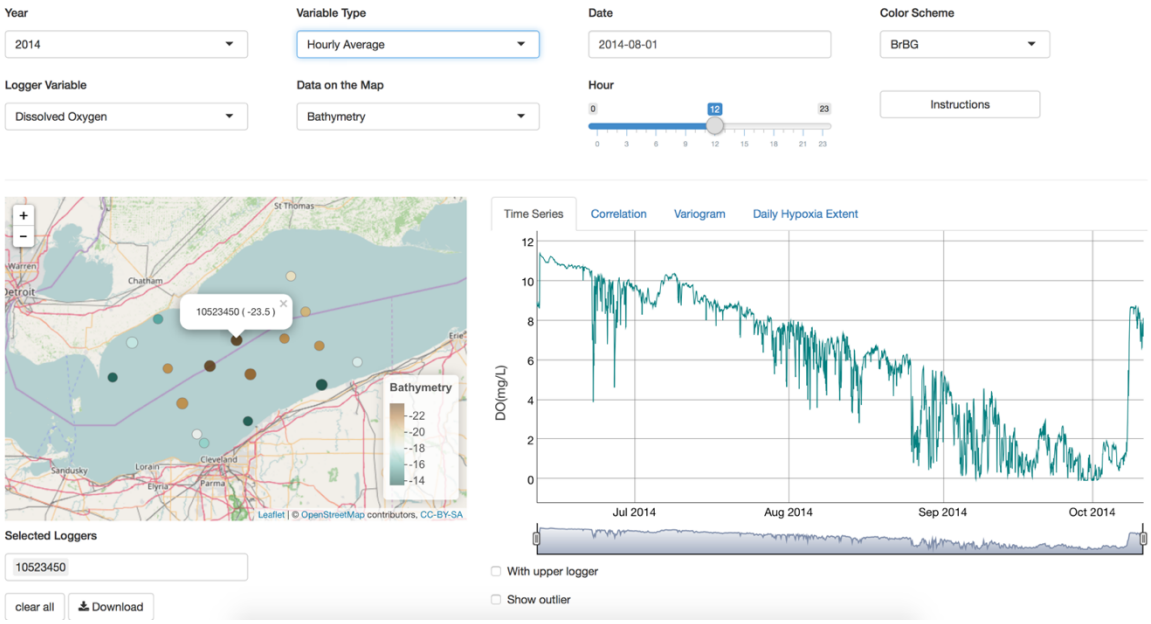


Figure 4.6. Interactive Web application for DO exploration in Lake Erie

4.4 Results

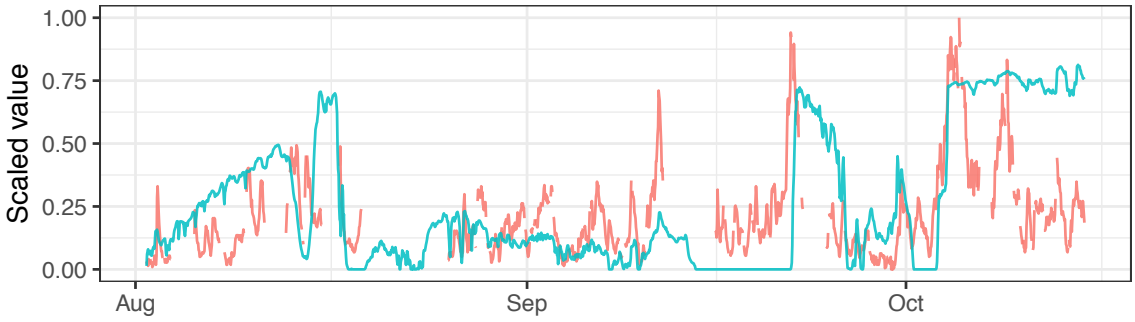
The DO patterns in all three years are first presented below in Section 4.4.1. Interpolation cross-validation results are then discussed in Section 4.4.2 to select the best method, which for this case study is the Bayesian method. This method is then used to explore spatio-temporal DO trends for the entire central basin in Section 4.4.3. Finally, the results from estimating hypoxia extent are presented in Section 4.4.4.

4.4.1 DO patterns

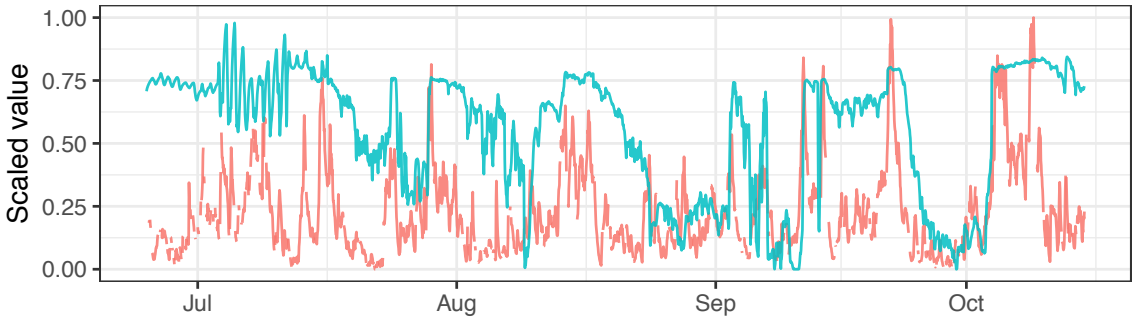
We first conduct an initial spatio-temporal data analysis to explore the DO patterns. Since the nearshore and offshore loggers have different sampling time ranges and the physical mechanisms involved in these locations are different, we analyze them separately.

Nearshore Patterns

Nearshore patterns are heavily influenced by internal seiche events. For example, when the winds push warm surface water to the shoreline and deepen the thermocline (i.e. downwelling events) below the sampling location, the whole water column mixes and DO is replenished. On the contrary, when winds push surface water away from the shoreline, upwelling events may bring low DO water to nearshore areas, causing a sharp drop in DO concentrations. This phenomenon can be observed in Figure 4.7, where the wave heights from two buoys on the south shore are compared with DO data from nearby loggers. Buoy 45169 (data available in 2015 and 2016) and Buoy 45164 (data available in 2014) are located near Cleveland City and Buoy 45167 is located near Erie City.



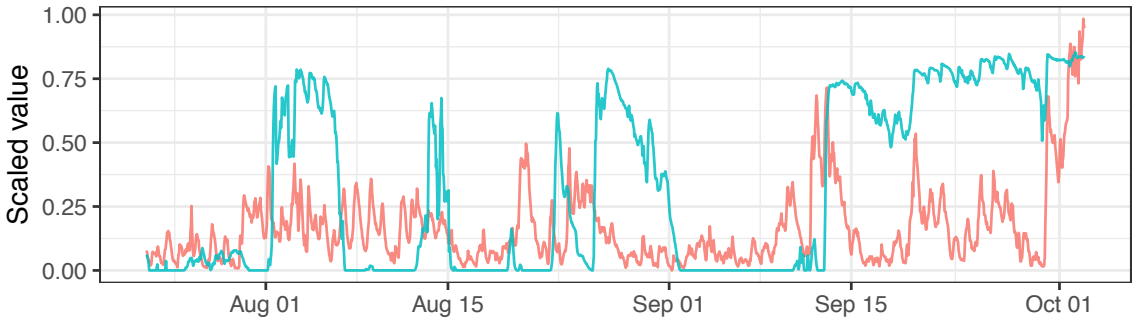
(a) Wave heights from buoy 45164 and DO logger 10384443 in 2014



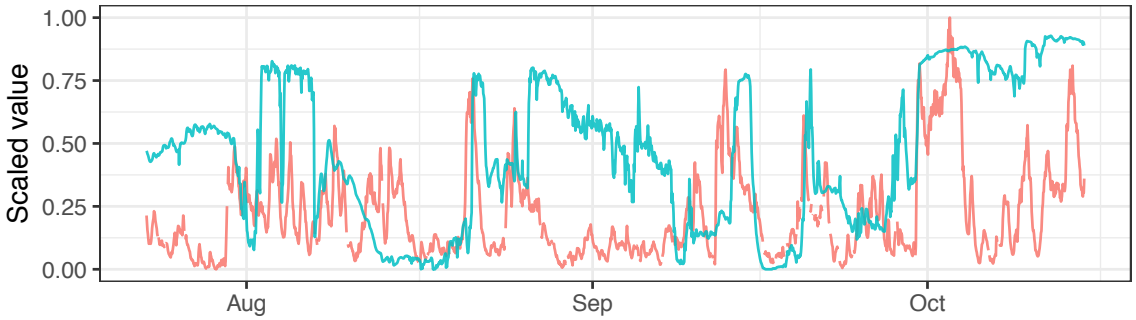
(b) Wave heights from buoy 45167 and DO logger 10384438 in 2014

Figure 4.7. Comparison between nearshore patterns and wave heights. Blue indicates DO concentration and red shows wave height. Both are scaled to [0, 1] values.

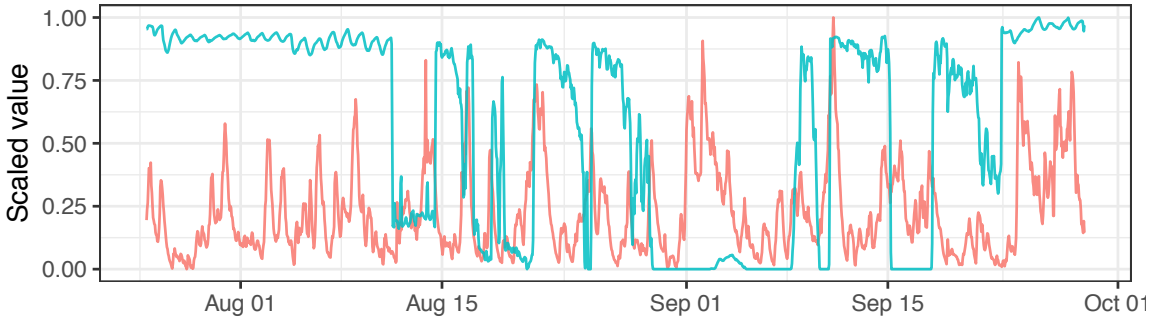
Figure 4.7 (cont.)



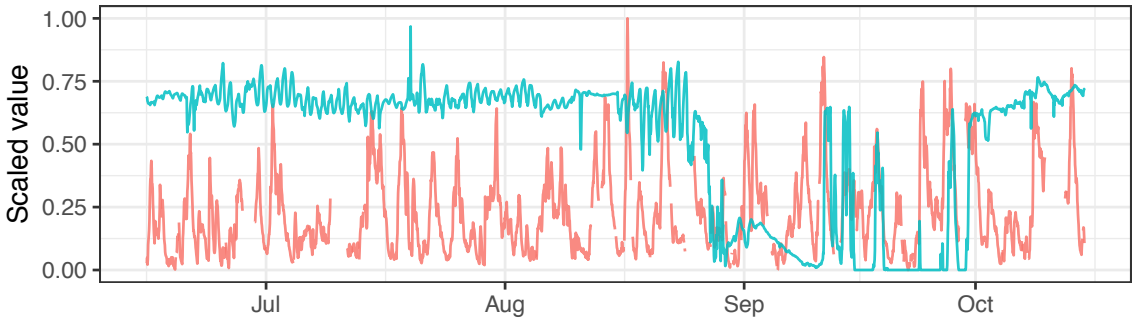
(c) Wave heights from buoy 45169 and DO logger 10534118 in 2015



(d) Wave heights from buoy 45167 and DO logger 10461951 in 2015



(e) Wave heights from buoy 45169 and DO logger 10384449 in 2016



(f) Wave heights from buoy 45167 and DO logger 10534122 in 2016

Figure 4.7 shows that DO concentrations in nearshore areas have large fluctuations. Sharp increases in DO concentration (blue line) are usually associated with a peak of wave heights. For example, in Figure 4.7a, the two sharp increases in DO around October are overlapped with sharp increases in wave heights, indicating that DO patterns are closely related to seiche events and their corresponding upwelling and downwelling phenomena. Sometimes increased wave heights did not lead to an increase in DO, most likely because the thermocline was not deepened sufficiently and the sampling logger remained in the hypolimnion with low DO concentrations.

In 2015, Figure 4.7c, note that the steep increases DO at the end of August and in the middle of September are after the wave height peaks. A consistent high wave may push the warm water downward slowly, thus causing a lag between the time of wave height peak and the time when the thermocline crosses the lake bottom, depending on the initial thermocline depth.

Offshore DO Patterns

Unlike nearshore loggers (Figure 4.7), DO concentrations drop gradually without significant fluctuations in offshore areas. Higher DO levels are reserved in the thicker hypolimnion after stratification and deep water is not as easily mixed with surface water as in the nearshore areas.

Spatially, offshore loggers in the eastern areas of Lake Erie have more DO than the western areas. For example, the logger at ER32, located in the eastern area, recorded higher DO concentrations than loggers at ER43, located in the western areas (Figure 4.8). This may be due to less hypolimnion oxygen demand in the eastern area. As a large portion of phosphorus loads usually come from tributaries in the western basin (Scavia, 2014), fewer nutrients may be available in the eastern zone.

Cross-year patterns in DO are illustrated using loggers at ER30 (Figure 4.9). It is clear that DO concentrations in 2014 were generally greater than 2015 and 2016, and 2014 also had lower lake bottom temperatures. These temperature differences may be the cause of the higher DO depletion rate in 2015 and 2016, since increased temperature would enhance biochemical reactions, thus increasing hypolimnetic oxygen demand as well as sediment oxygen demand (Matisoff and Neeson, 2005; Rucinski et al, 2010).

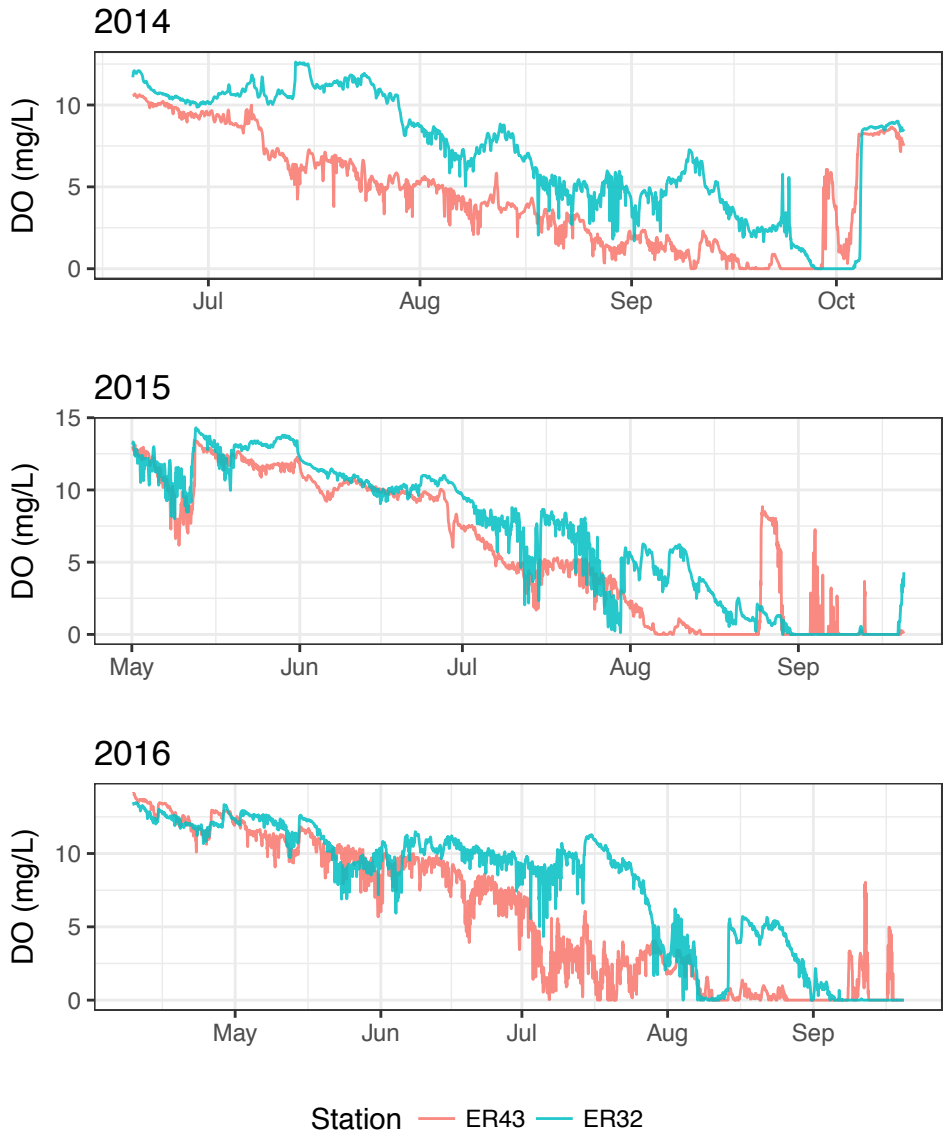


Figure 4.8. DO concentrations of loggers at stations ER43 and ER32.

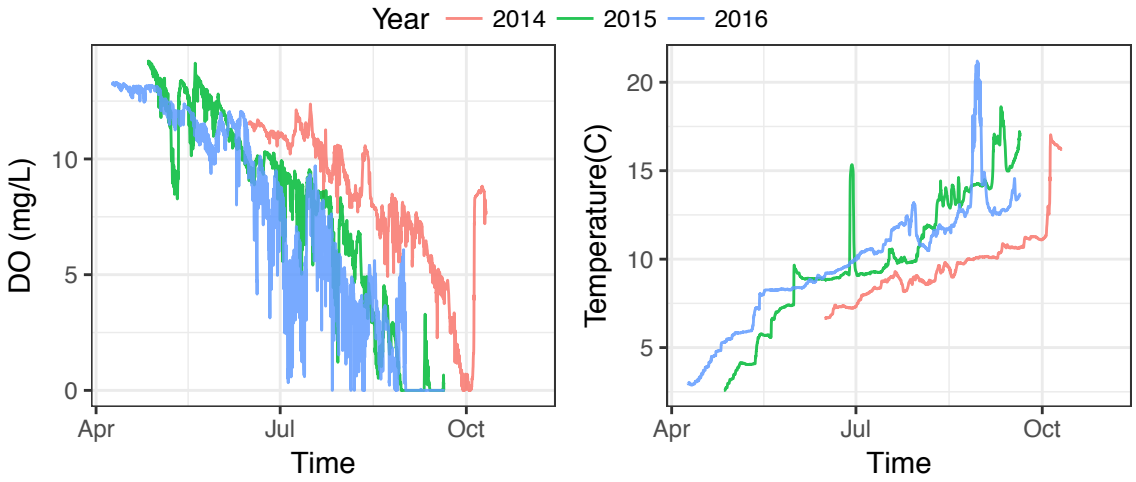


Figure 4.9. DO (left) and Temperature (right) in different years at Station ER31

The cause of lower lake bottom temperatures in 2014 needs further investigation. The Lake Erie surface temperature (Figure 4.10) shows that surface temperature patterns are similar between 2014 and 2015, which should lead to similar lake bottom temperatures during the spring turnover period. However, 2014 had significantly lower lake bottom temperatures in July. One possible factor is that the lake surface temperatures in 2015 are higher between the 120th and 140th days of the year (Figure 4.10), which was just before lake stratification occurred and the bottom water was separated from the surface water. The bottom water in 2015 thus became higher temperature than in 2014.

The lake bottom temperature in 2016 is higher than 2015 in April and May (Figure 4.10), which is likely due to a warm 2015-2016 winter (in fact, 2013-2014 and 2014-2015 winters were abnormally cold). This may have caused less ice coverage so that the whole lake started to warm earlier. More data, especially temperature depth profiling data, are needed to study these temperature changes in the lake bottom and provide more definitive explanations.

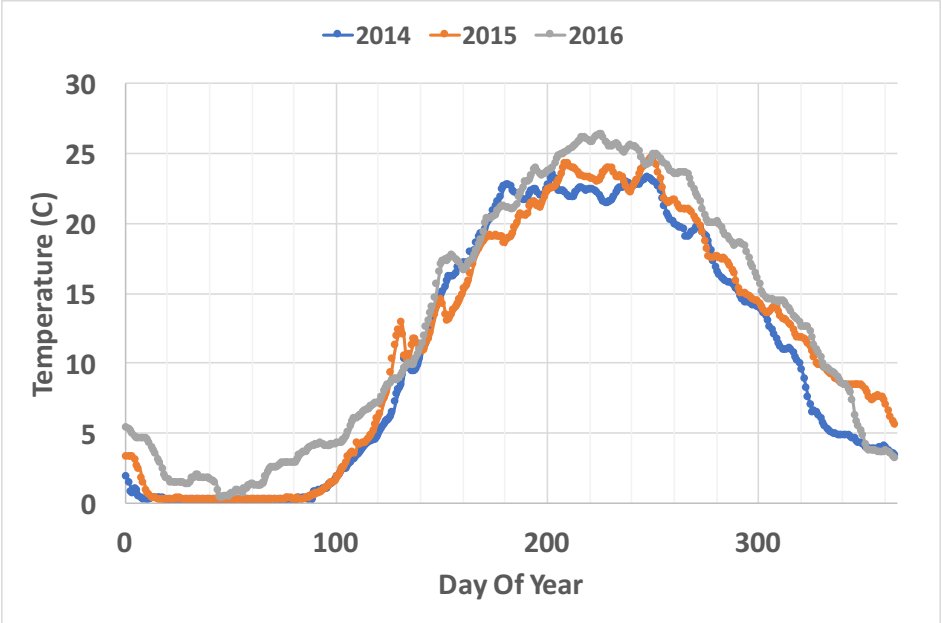
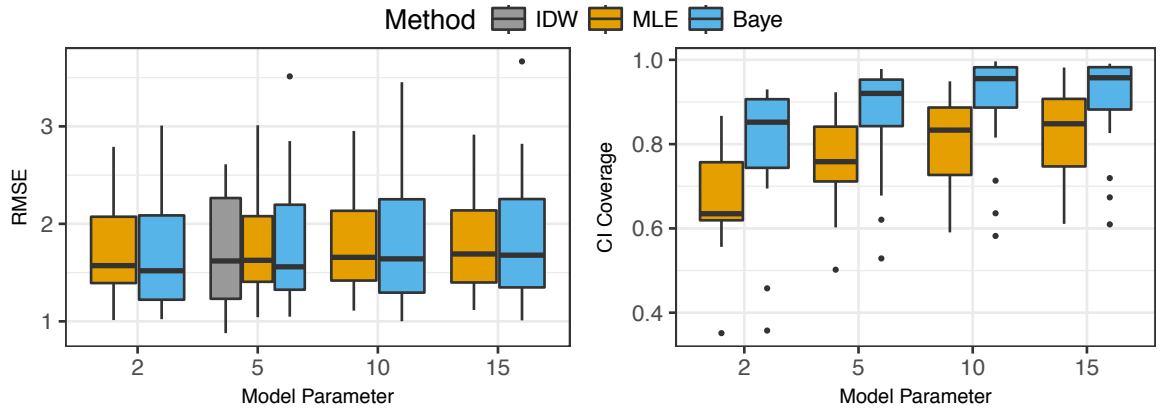


Figure 4.10. Lake Erie surface water temperature (data from NOAA)

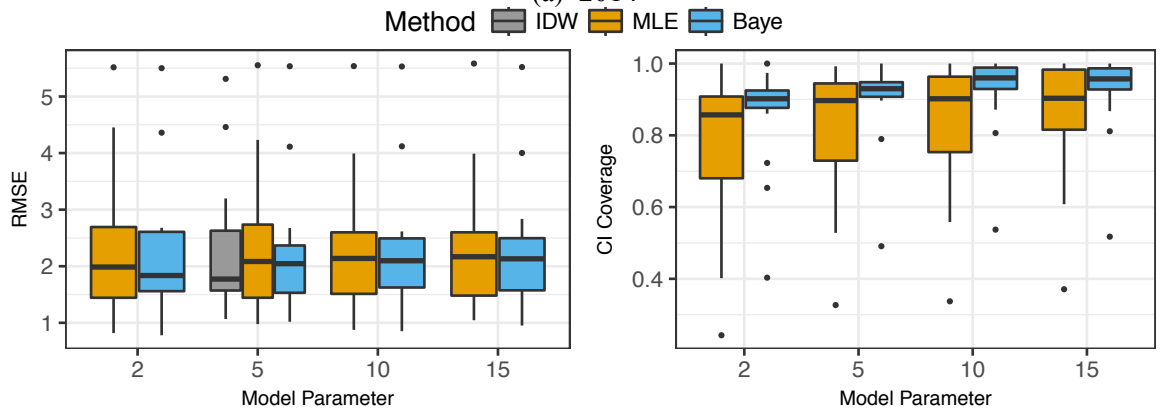
4.4.2 Spatio-temporal Interpolation Cross-Validation

For the spatio-temporal interpolations, we used cross-validation to validate both methods. In Figure 4.11 (left), cross-validation shows that spatio-temporal IDW (Section 4.3.2) and spatio-temporal basis interpolation with MLE and Bayesian framework (Section 4.3.3) have similar root-mean-square errors (RMSEs) for different numbers of basis functions (r in Eq. 4.5). The model also had generally larger RMSE in 2015 compared to 2014, indicating that in 2015, the patterns of each logger are harder to capture.

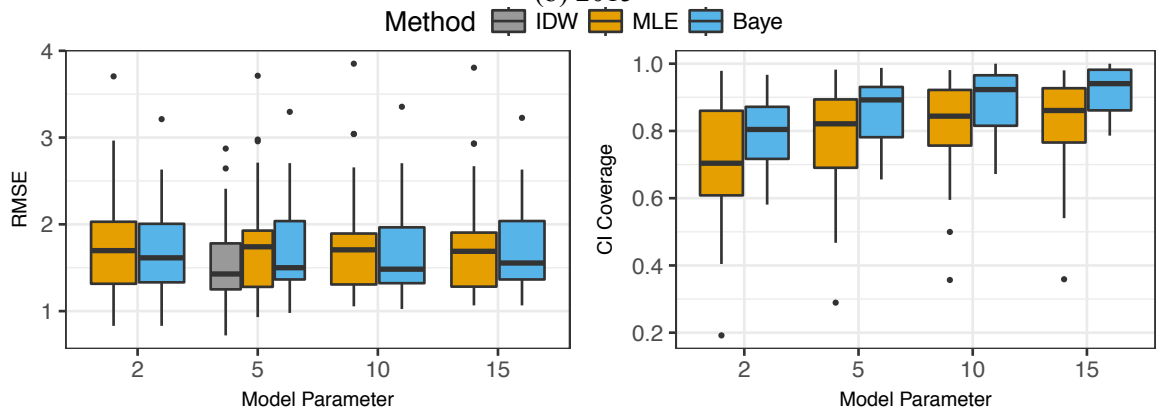
Unlike IDW, kriging models using MLE or Bayesian framework are able to provide estimation uncertainty. Figures 4.11 (right plot) shows the metric called CI coverage, which is the proportion of sampled data that fall into the confidence interval. 0 means no sampled data are within the confidence interval and 1 means all of the observed data are within the confidence interval. The Bayesian method has larger confidence intervals due to consideration of the variogram model uncertainty, so that more observed data fall into the confidence interval. MLE underestimates the prediction uncertainty by not incorporating the variogram model uncertainty.



(a) 2014



(b) 2015



(c) 2016

Figure 4.11. Box plots of cross-validation results for hourly aggregated data. IDW refers to the spatio-temporal IDW method, and MLE and Baye represent the variogram model fitted by MLE method or Bayesian framework in the spatio-temporal basis interpolation. Left: RMSE with different methods. For MLE and Bayesian, RMSEs are calculated using the median values in all interpolation simulations; Right: Coverage of 90% confidence intervals [CI] (5% to 95%) with MLE and Bayesian method. The box in the boxplot shows the 1st (Q1) and 3rd quantile (Q3) and the black line shows the median value. The outliers are the values beyond Q3+1.5IQR or Q1-1.5IQR, where the interquartile range IQR = Q3-Q1. The model parameter refers to the number of basis functions (r in Eq 4.5) for MLE and Bayesian method or the number of nearest sampling locations used in the IDW method.

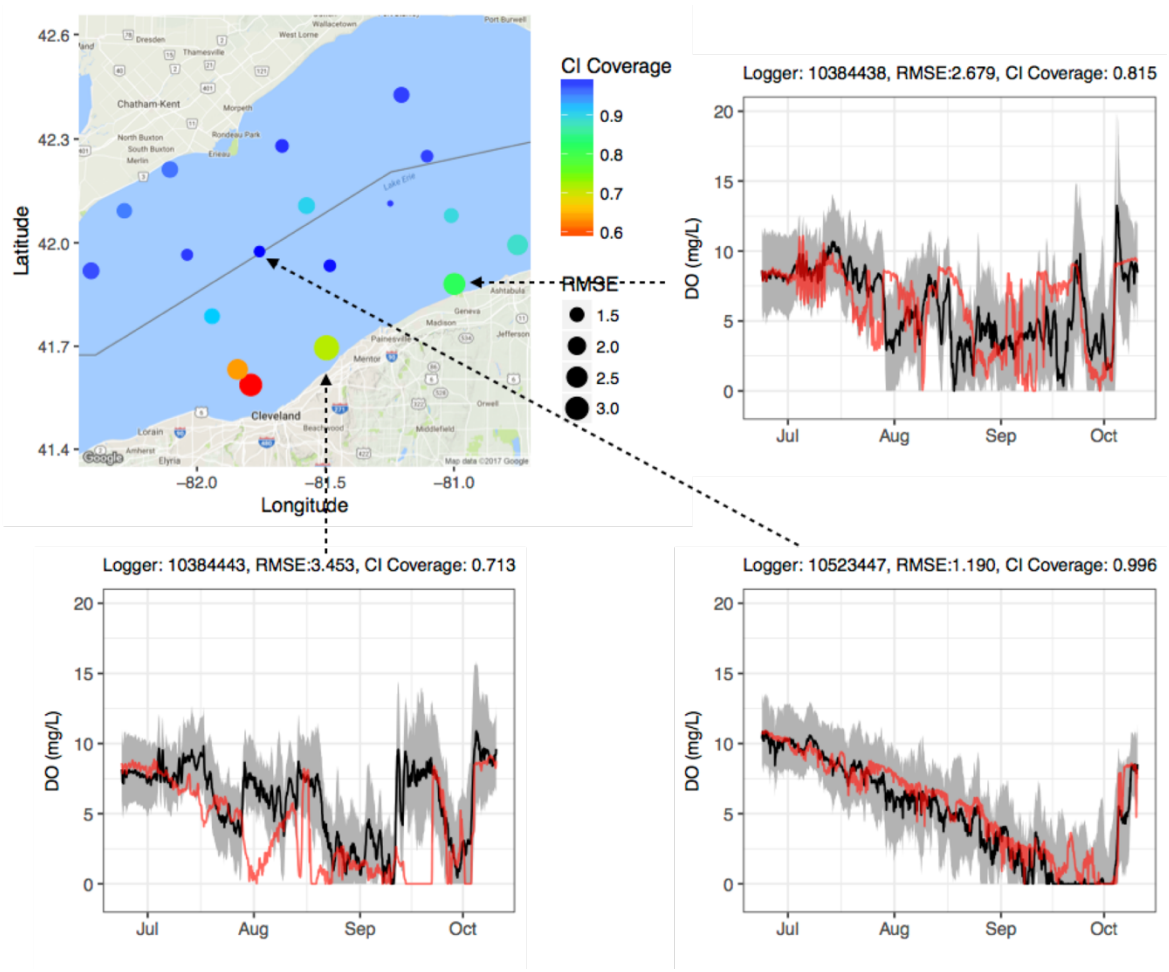
Overall, the Bayesian method with 10 basis functions performed best for this case study because: (1) compared to spatio-temporal IDW, although more complicated, basis interpolation provides uncertainty estimates for interpolation and hypoxia extent; (2) compared to MLE method, the confidence interval from the Bayesian method contains more observed data; and (3) r = 10 has better CI coverage performance compared to r=2 and r=5. The model with 15 basis functions has similar performance but is more complicated and the computational cost is heavier. In general, the final hypoxia extent estimation is not overly sensitive to the number of basis functions for this case study.

Figure 4.12 shows visualizations of RMSE and CI coverage spatially in the cross-validation, with spatio-temporal basis interpolation using Bayesian framework and r = 10. The results show that in 2014, loggers near the shoreline have high RMSE errors and low CI coverage (top left in Figure 4.12a), indicating the patterns are fundamentally different from other locations such that the uncertainty in the spatial random process and uncertainty in the variogram model cannot completely account for the differences. In Figure 4.12a, Logger 10384443 (bottom left) and Logger 10384438 (top right) have quite different patterns from surrounding loggers. They are hard to predict because: (1) the predictions on nearshore loggers are extrapolation, which has worse performance in general; and (2) the DO dynamics are complicated in nearshore areas and are influenced by seiche events as discussed in Section 4.4.1.

In 2015 (Figure 4.12b), the prediction is generally worse compared to 2014 (RMSE are higher, Figure 4.11b). Logger 10523449 (top right) has the highest RMSE and lowest CI Coverage. The bottom water there started to mix with surface water around September with a sudden increase of DO, yet the neighboring loggers were still in hypoxic states. Logger 10534123 (bottom right) also has low CI coverage and high RMSE since it is in the extrapolation range. Predicting the western zone DO given only the central zone patterns is intrinsically unreliable when the dynamics in the west were different.

In 2016, besides the south shore, the interpolation also performs worse in eastern areas, with lower CI coverage and high and larger RMSE due to varying DO patterns. The prediction for logger 10534125 is significantly higher than observed. One possible reason is that the sampling depth at that location is shallower than the surrounding loggers, so that during universal kriging with basis coefficients, the effects of the bathymetry are inaccurately extrapolated, which eventually leads to higher predicted DO.

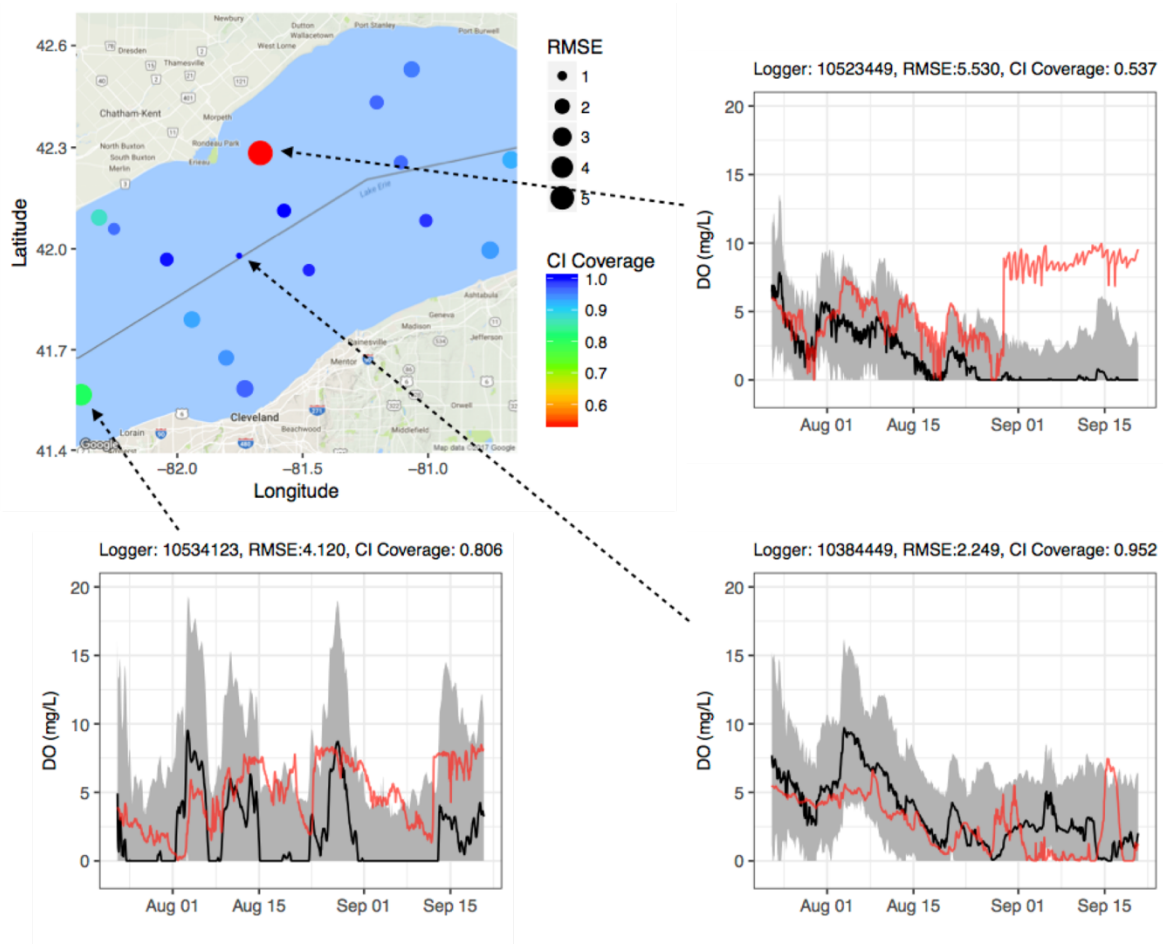
The loggers in central offshore locations are easier to predict, such as Logger 10523447 in 2014 and Logger 10384449 in 2015. They have similar patterns as the nearby loggers, thus having a low RMSE as well as high CI coverage.



(a) 2014

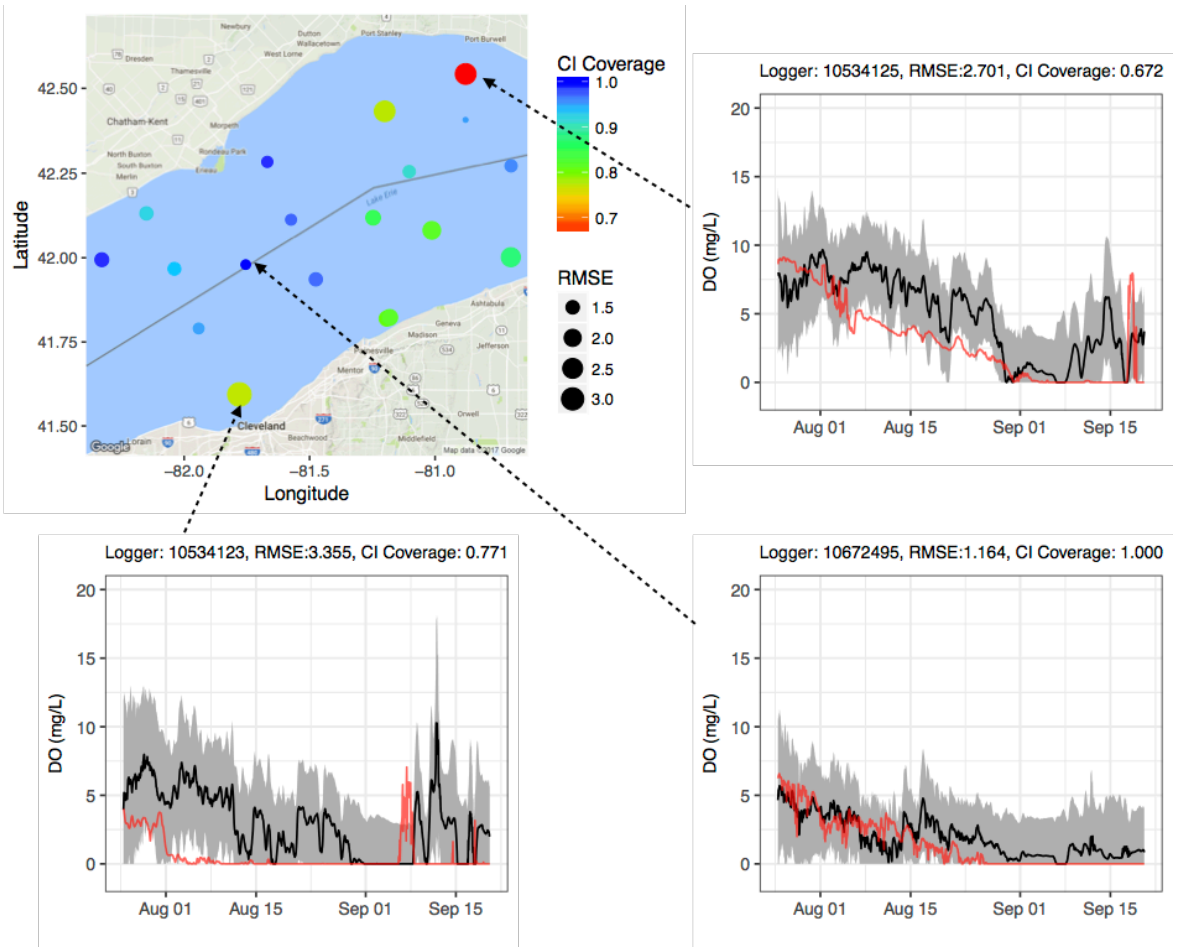
Figure 4.12. Cross-validation results in hourly aggregated data in (a) 2014, (b) 2015 and (c) 2016. In the top left, each dot indicates the sampling location, with size representing RMSE and color representing CI coverage ratio. Other plots represent cross-validation predictions on specific loggers, where red is the observed data, black is the median values of all simulations, and grey areas are the 90% confidence intervals.

Figure 4.12 (cont.)



(b) 2015

Figure 4.12 (cont.)



(c) 2016

The range of the confidence interval is related to the values of the fitted variogram model. If the variogram model has smaller sill (i.e. the spatial correlation is high) or the variogram model uncertainty is narrower, conditional realizations will generate less variable predictions so that the confidence interval will be narrower. Smaller sill in the variogram model can be achieved by detrending the data first using an accurate physical model.

To narrow the posterior distributions of the variogram model, one approach is to reduce the uncertainty of the prior distributions, but with a risk that if the prior distribution is not accurate or biased, significantly more data may be required to correct the posterior variogram. Another approach is to increase the number of loggers so that more data points will contribute to a more confident estimation of the variogram parameters. In addition, increasing the density of sampling points also makes the distance between sampling points and target points smaller, so

that the spatial correlation will be stronger, leading to smaller prediction variances in the kriging interpolation.

4.4.3 Spatio-temporal DO Interpolation

Figures 4.13 to 4.15 show the interpolated DO in 2014, 2015 and 2016, using spatio-temporal basis interpolation with $r = 10$ and expected prediction coefficients (i.e. $\hat{\beta}_k\ (s)$, Section 4.3.2) within the Bayesian framework.

In 2014 (Figure 4.13), low DO concentrations started from the south shore and western zone, and then gradually expanded to the whole interpolation area. The hypoxia areas were reduced around 09/22 as the south shore zone had regained some DO, but then the hypoxia areas increased again. By early October, as the lake fall turnover started, the bottom water quickly mixed with DO-saturated surface water and the hypoxic area rapidly decreased. The phenomenon of hypoxia starting from the nearshore or shallower zones is also consistent with previous numerical modeling (Bocaniov and Scavia, 2016).

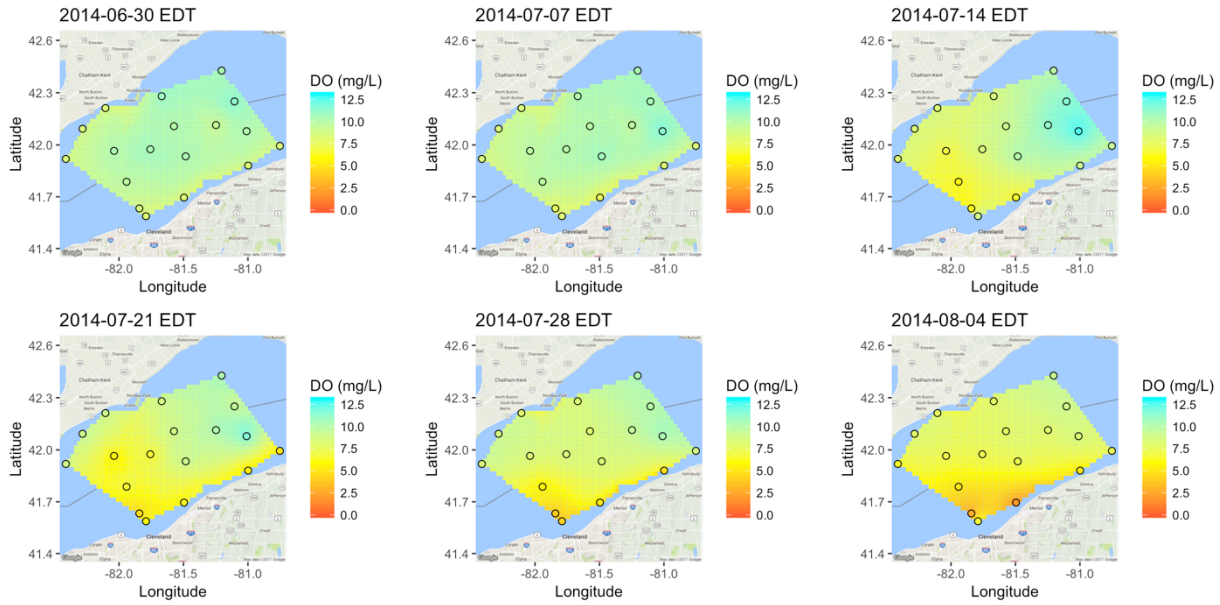
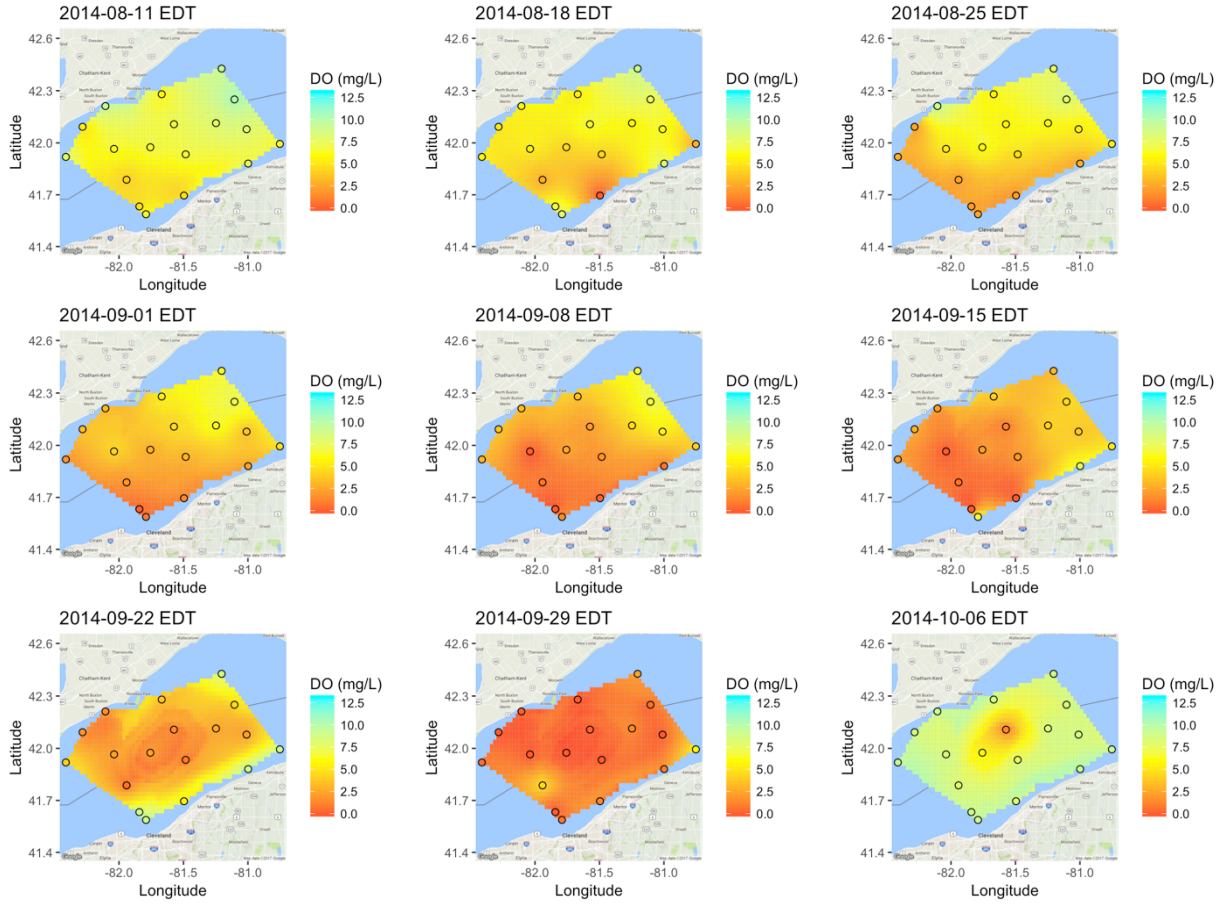


Figure 4.13. DO Interpolation results in 2014 for multiple dates at 00:00:00 EDT. The circles are the sampling locations.

Figure 4.13 (cont.)



Unfortunately, the sampling ranges of nearshore and offshore loggers overlap for less than two months in 2015 and 2016, starting from late July and end in late September, thus the hypoxia development and dissipation are not completely captured. In 2015 (Figure 4.14), hypoxia already existed by late July (07/23) and then expanded quickly (07/30). DO concentrations in the eastern zone recovered in early August (08/06) but then worsened again. The worst situation appeared around 08/19, when almost all areas were in a low DO state. After that, DO increased in the western zone (08/26) but was later reduced again. Then the north shore had a patch of relatively high DO concentrations. The whole lake seemed to start recovering around 09/16 as the DO in western zones increased to non-hypoxic DO level and maintained (Figure 4.7c), indicating the thermocline has dropped below the nearshore bathymetry and lake fall turnover would begin shortly.

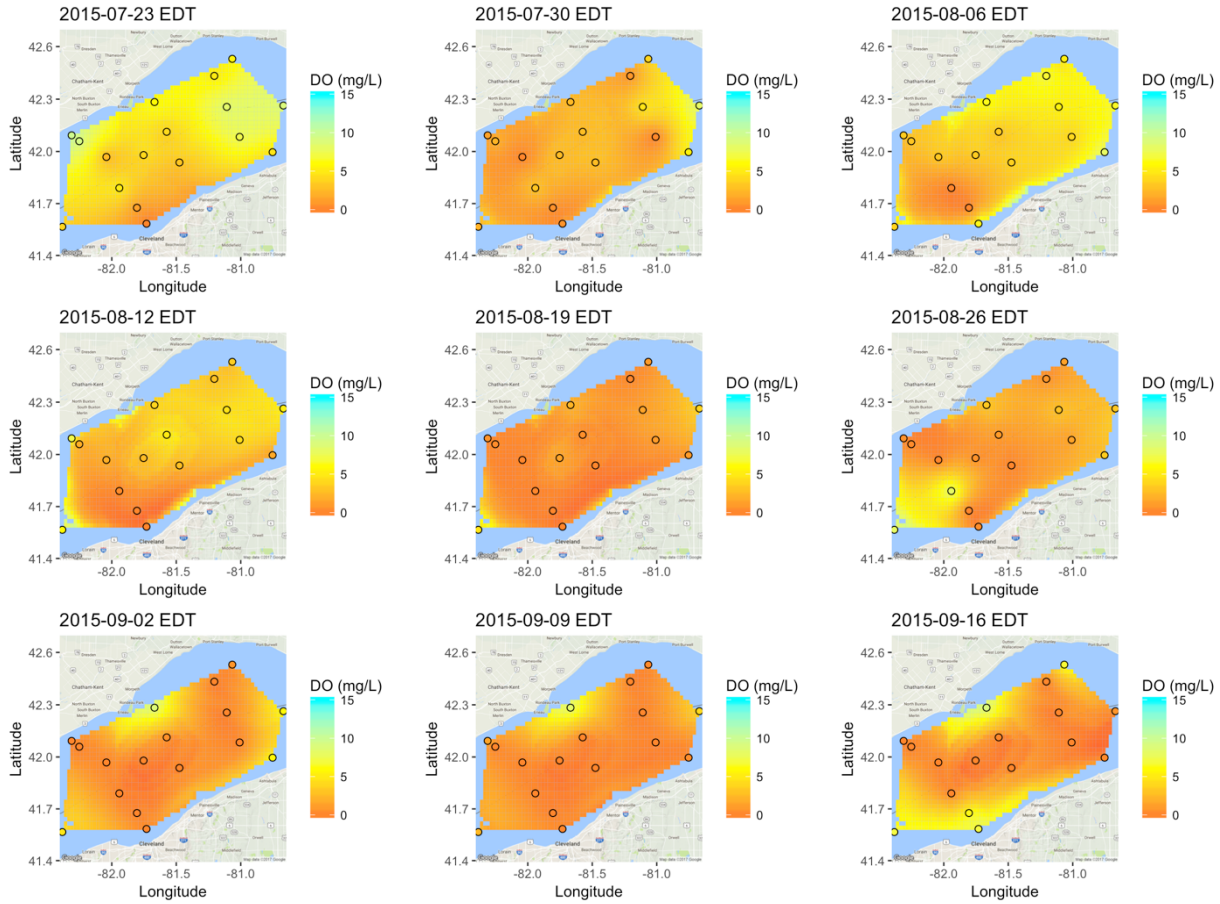


Figure 4.14. DO Interpolation results in 2015 for multiple dates at 00:00:00 EDT. The circles are the sampling locations.

In contrast, during 2016 (Figure 4.15) within the interpolation time range, hypoxia starts from the North shore and western areas (07/26) and quickly expands to the whole central basin. The whole central basin remains hypoxic during most of September.

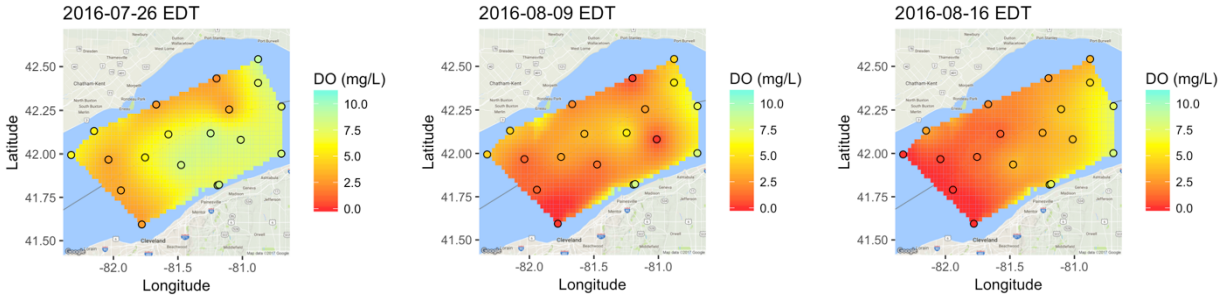
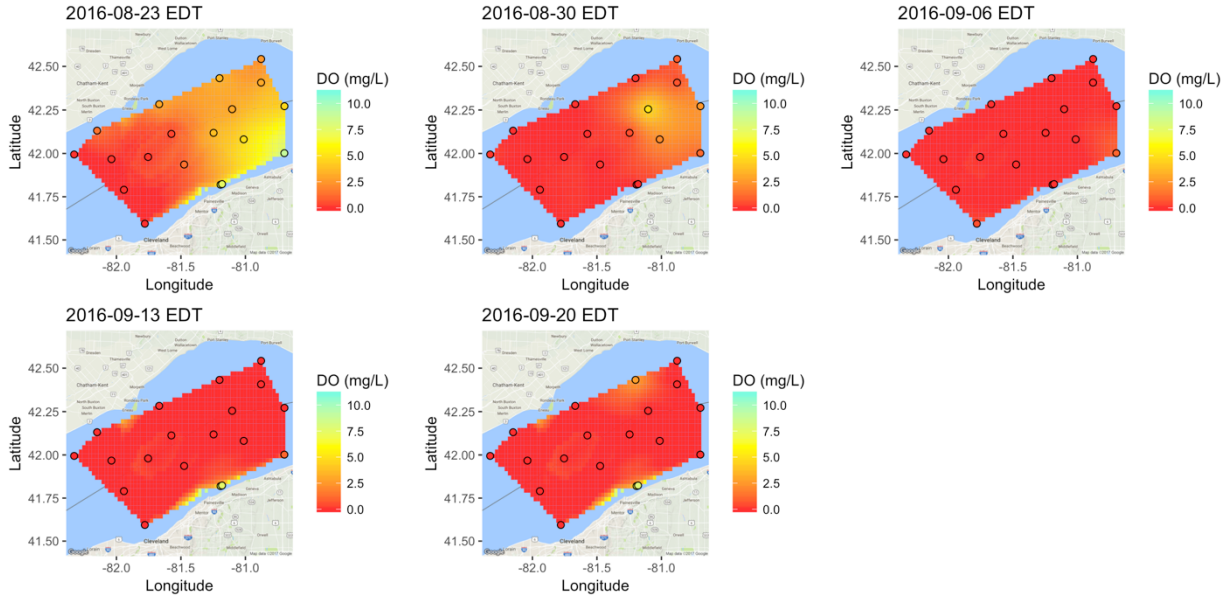


Figure 4.15. DO interpolation results in 2016 for multiple dates at 00:00:00 EDT. The circles are the sampling locations.

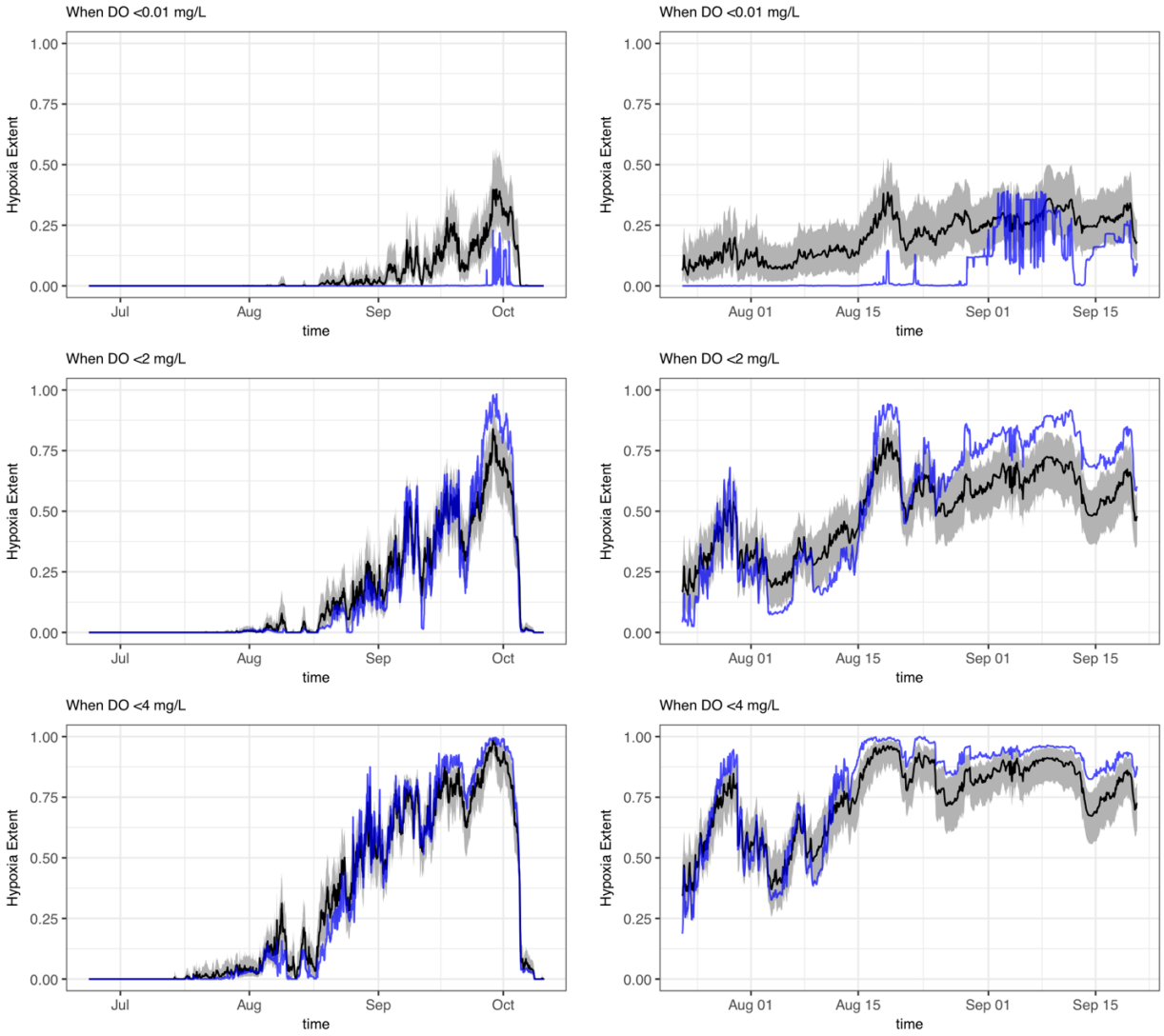
Figure 4.15 (cont.)



4.4.4 Hypoxia Extent

Hypoxia extent is summarized with uncertainty bounds estimated by conditional simulations (Section 4.3.4) in Figure 4.16. We also provide hypoxia extent calculated by spatio-temporal IDW interpolation for comparison (blue line). Hypoxia started to emerge in August 2014 (Figure 4.16a), while in 2015 (Figure 4.16b) and 2016 (Figure 4.16c), some areas were already hypoxic in late July. There were two DO recovery events as two points with low hypoxia appeared in the 2014 series around mid-September (in the middle row, DO < 2mg/L). In 2015, the DO increased in early August (Figure 4.16b), while in 2016, DO was replenished to some extent only in the middle of August (Figure 4.16c).

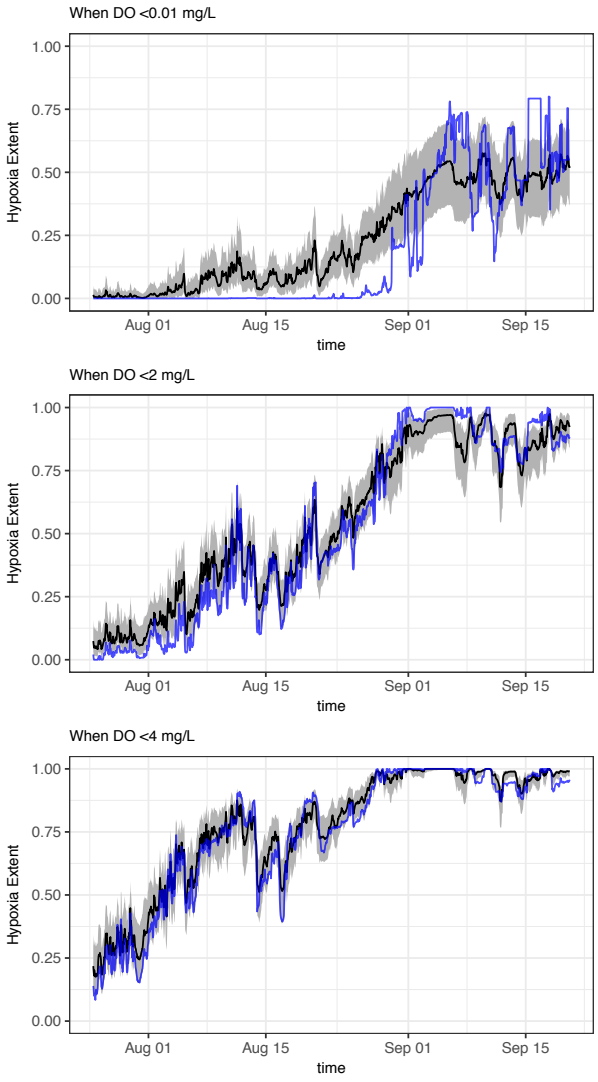
During the periods with the largest hypoxia extents (late September 2014, mid-August 2015, and early September 2016), more than 75% of the interpolated areas had less than 2 mg/L DO and almost 100% of the areas had less than 4mg/L DO. The uncertainty for hypoxia extent (the range of confidence interval) can be as large as 0.25 for both years (around 1700 km^2 [0.25 * 6985km^2] for 2014, around 2200 km^2 [0.25* 8917 km^2] for 2015 and 1900 km^2 [0.25 * 7543 km^2] for 2016. More loggers are needed to reduce the uncertainty and obtain more accurate estimations, as discussed in Section 4.3.2.



(a) 2014 (b) 2015

Figure 4.16. Hypoxia extent ratio (compared to the total interpolation area) with different hypoxic DO thresholds in 2014 (a), 2015 (b) and 2016 (c). Shaded area: 90% confidence interval generated by conditional simulations; Black: Median value from conditional simulations generated by Bayesian approach; Blue: Hypoxic area calculated by IDW. The total interpolation areas in 2014, 2015 and 2016 are around 6985, 8917 and 7543 km^2, respectively.

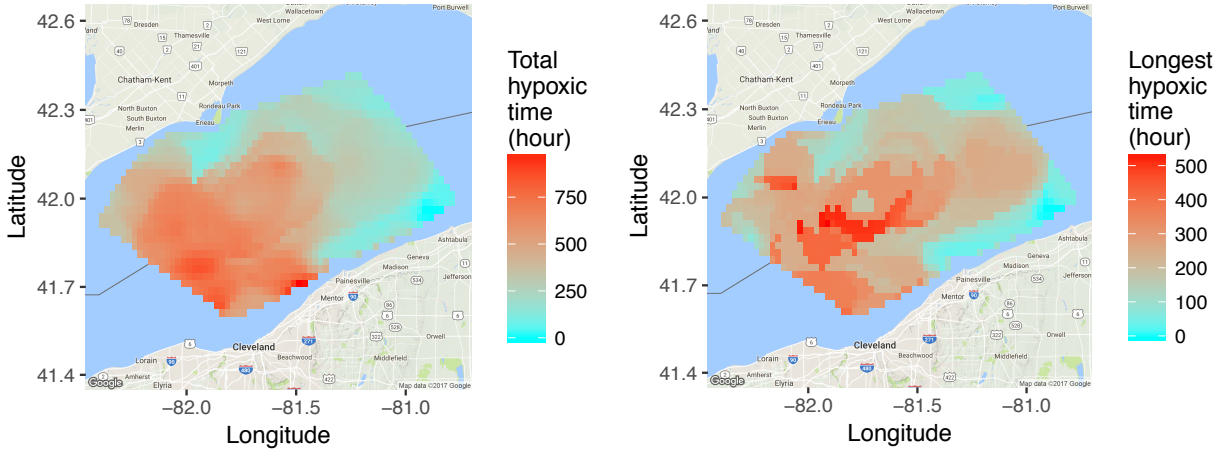
Figure 4.16 (cont.)



(c) 2016

The spatio-temporal IDW interpolation (Section 4.3.2; blue line in Figure 4.16) underestimates the hypoxia extent in the extreme cases (Figure 4.16, top row, DO < 0.01mg/L). This is because IDW uses the nearest 5 loggers so that the interpolated DO will be less than 0.01mg/L only when the nearest 5 loggers are all below 0.01 mg/L, which are rare cases. Also, IDW tends to have larger fluctuations that overestimate the extent during hypoxic situations and underestimate when DO concentrations are recovering, generating a less smooth time series. The differences between spatio-temporal IDW and basis interpolation are due to IDW only considering the spatial coordinates, while the basis interpolation framework also considers the bathymetry information in the universal kriging interpolations of the coefficients (Section 4.3.3).

The total hypoxia time and the continuous hypoxia time are also important to the lake ecosystem. Total hypoxia time is related to nutrient loading (Rucinski et al., 2014), and fish may die or shift diets when exposed to long-term hypoxic states (Scavia et al., 2014). To calculate the hypoxia time, we used the expected coefficient predictions $\hat{\beta}_k\ (s)$ to reconstruct the DO interpolation. In 2014 (Figure 4.17a), southwestern areas had the longest total hypoxic time, while the eastern areas had less. Although the areas near Mentor city had the longest total hypoxic time (Figure 4.17a, around 900 hours), the longest continuous hypoxic time (Figure 4.17b, around 500 hours) is instead located in the central part of the interpolation areas. 2015 has a much longer hypoxic time (Figure 4.17c, the maximum is near 1250 hours) in the western and south shore zones, among which a northwestern patch and a near south shore zone have the longest continuous hypoxic periods (Figure 4.17d). In 2016, the western areas had the longest total hypoxia time as well as the longest continuous hypoxic time.

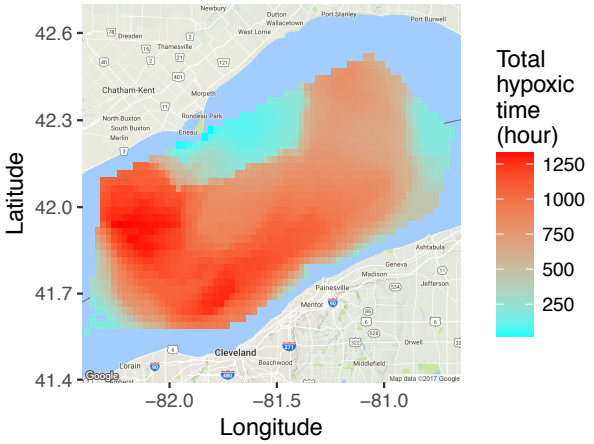


(a) 2014 Total Hypoxic Time

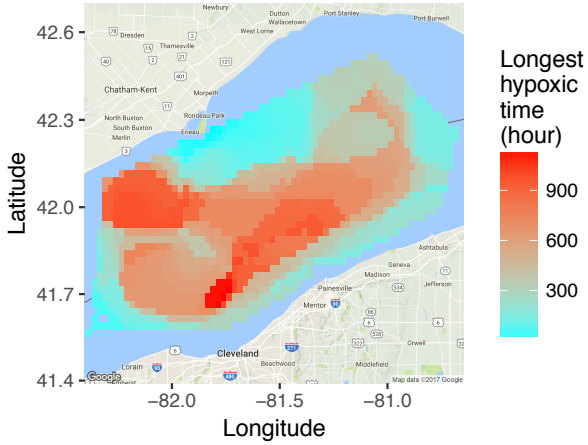
(b) 2014 Longest Hypoxic Time

Figure 4.17. Total hypoxic time and longest hypoxic time (DO < 2mg/L) in the central basin

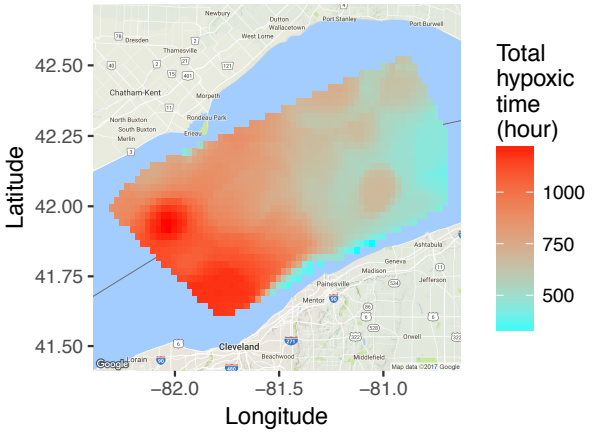
Figure 4.17 (cont.)



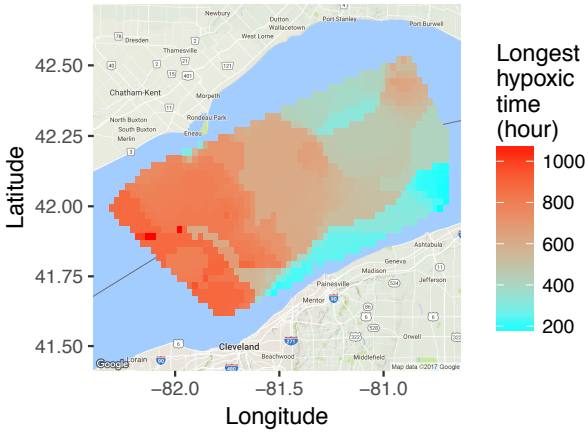
(c) 2015 Total hypoxic time



(d) 2015 Longest hypoxic time



(e) 2016 Total hypoxia time



(f) 2016 Longest hypoxic time

4.5. Discussion

In this section, we discuss a sensitivity analysis to explore the effects of different priors on the Bayesian interpolations (Section 4.5.1) and suggestions for optimizing logger deployment locations (Section 4.5.2)

4.5.1 Sensitivity Analysis on Bayesian Interpolation Parameters

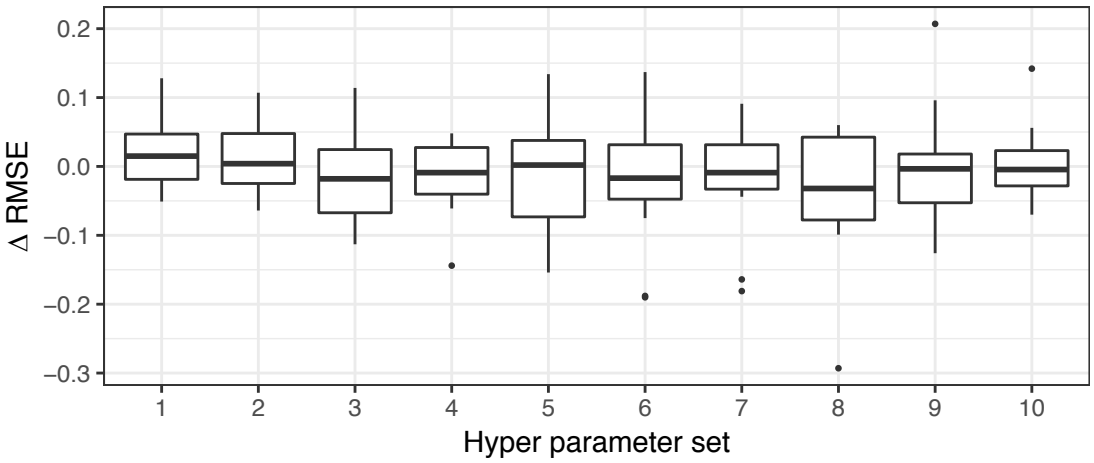
The parameters that define prior distributions in Bayesian interpolation were initially chosen based on the default parameters used in the "geoR" package. To explore the effects of this selection on the spatial and temporal interpolations, we also ran the models using other parameters. These alternative parameter sets are defined in Table 4.2. Parameter Sets 1 to 3 explore different priority distributions for the Range (ϕ). Parameter Set 4 to 7 explore nugget

effects. Parameter Set 8 changes the variogram model to another widely used "Spherical model". Parameter Sets 9 and 10 change the priority distribution of the Partial Sill (σ^2).

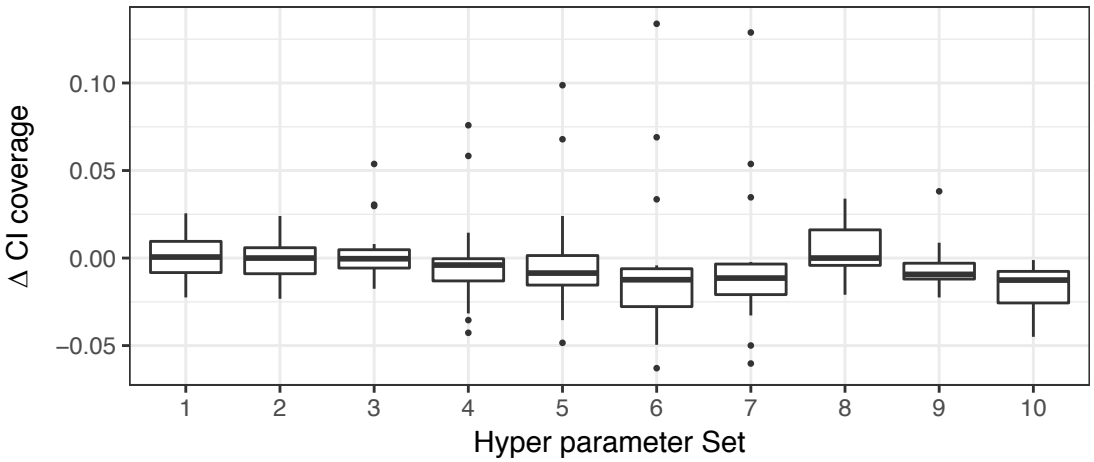
Table 4.2 Alternative prior distributions for Bayesian kriging.

Parameter Set #	Changed Parameter	Parameter Setting
1	Range (ϕ)	Reciprocal distribution from 20km to 70km with 5km intervals
2	Range (ϕ)	Uniform distribution with a sequence of 51 values from 0 to 2 times the maximum distance between the data locations
3	Range (ϕ)	Squared reciprocal distribution from 20km to 70km at 5km intervals
4	Nugget (τ^2) / Partial Sill (σ^2)	Fixed at 0.1
5	Nugget (τ^2) / Partial Sill (σ^2)	Fixed at 0.2
6	Nugget (τ^2) / Partial Sill (σ^2)	Uniform distribution from 0 to 1 with an interval of 0.1
7	Nugget (τ^2) / Partial Sill (σ^2)	Reciprocal distribution from 0 to 1 with an interval of 0.1
8	Variogram model	Spherical model
9	Partial Sill (σ^2)	Scaled inverse chi-squared distribution with 1 degree of freedom
10	Partial Sill (σ^2)	Scaled inverse chi-squared distribution with 3 degrees of freedom

We tested these new parameters with cross-validation on the 2014 data. The boxplots in Figure 4.18 summarize the differences in RMSE and CI coverage between the base and alternative parameters in Table 4.2 at the same logger.



(a) RMSE differences (positive values indicate the base parameter is better)



(b) CI Coverage differences (positive values indicate the alterative parameter is better)

Figure 4.18. RMSE and CI coverage differences with alternative prior distributions

The results show that the parameters for prior distributions do not have significant effects on the interpolations, with small differences in RMSE and CI coverage (Note that the original RMSE is approximately 2 mg/L and CI coverage is about 0.9. Furthermore, no alternative parameters are able to achieve better results on every logger in the cross-validation. Parameter sets 4 to 7 explore scenarios where measurement noise exists (nugget $[\tau^2] > 0$); these scenarios show inconsistent effects, with interpolation on some loggers performing better and on other loggers performing worse.

4.5.2 Logger Deployment Optimization

One way to measure hypoxia extent accurately is to deploy loggers densely, however this is quite costly. With limited numbers of loggers, how to optimize the sampling locations without

compromising the interpolation accuracy is very important to understanding the lake DO dynamics and maximizing monitoring yields.

Cross-validation results (Section 4.4.2) give information on logger importance within the current deployment topology. Locations that are difficult to predict (large RMSE and low CI coverage) by other loggers in cross-validation results (Figure 4.12) are the locations that should not be removed and would probably most benefit from more loggers. In all three years, the DO series from south shore loggers are consistently difficult to predict from other loggers, while the offshore areas, especially loggers around ER73, consistently have good prediction accuracy from other loggers (Figure 4.12). The northern and eastern areas are more complicated. One logger in the north in 2015 (Figure 4.12b) and multiple loggers in the west in 2016 (Figure 4.12c) have low cross-validation prediction accuracy, implying heterogeneous patterns could happen in these areas.

Overall, compared to the current logger deployment, better information may be obtained by moving some loggers from the central offshore area to the nearshore. It should be noted that the above discussion is a retrospective analysis based on three years in which the sampling data are already known. With more data in the future, better generalizations can be made as to which locations have consistently homogeneous or heterogeneous patterns that may need fewer or more loggers, respectively, for accurate interpolation.

4.6. Conclusions

In this chapter, we proposed a spatio-temporal interpolation that incorporates model uncertainty with a Bayesian framework, along with conditional simulations to estimate hypoxia extent with uncertainty. DO sensor data from 2014 to 2016 in the central basin of Lake Erie are analyzed as a case study and a Web application is developed using R Shiny.

We conduct initial analysis on the patterns of the nearshore and offshore loggers. The nearshore patterns are closely related with internal seiche events and related upwelling/downwelling events. The offshore patterns are likely related to nutrient distributions and lake thermo-stratification changes. Colder bottom water temperatures in 2014 and less nutrient transport to the eastern areas led to less DO depletion. We compared three different spatio-temporal DO interpolation models based on IDW, MLE kriging, and Bayesian kriging via cross-validation. The results show that these methods have similar RMSE but Bayesian kriging

estimates the prediction uncertainty more accurately than the other methods.

The seasonal changes in hypoxia within the central basin of Lake Erie show that hypoxia developed differently in each year. In 2014, hypoxia started to emerge in early August while in 2015 and 2016, hypoxia began in July. The peak hypoxia extent occurred in late September 2014, mid-August 2015, and early September 2016, and the uncertainty of the hypoxia extent can be as large as 25% of the interpolation areas.

Further discussion of cross-validation results suggests that some offshore loggers may be redundant and could be relocated to nearshore areas to improve estimations of hypoxia extent.

Future research recommendations are given in Chapter 5.

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

In this chapter, we present conclusions (Section 5.1) and discuss recommendations for future research to extend the methodology and findings of this dissertation (Section 5.2). The near-real-time geostatistical framework for undulating data in Chapter 2 can be extended to other sampling activities and improved by incorporating external data sources, as discussed in Section 5.2.1. Section 5.2.2 discusses how the lake stratification and DCL detection algorithms can be applied with other water chemistry depth profiling data and can guide future sampling activities. Lastly, we discuss some future directions for better understanding the Lake Erie dissolved oxygen processes in Section 5.2.3.

5.1. Conclusions

This thesis has developed three data analysis frameworks for lake water chemistry data generated by different sensors. Each framework detects patterns from raw sampling data to better understand lake processes and provide guidance on future sampling strategies. The algorithms are validated either by cross-validation or comparison with operator logs. We also provide open-source code in R or Python and prototype Web applications for each framework.

In Chapter 2, a proposed new geostatistical framework is able to detect river plume patterns from undulating sampling data. The framework includes automated interpolation and visualization that enables near-real time use to support adaptive sampling. Hotspot analysis and cluster analysis reveal river plume orientation and water mixing areas. Application of the framework reveals three different river plume dynamics in three rivermouth areas of Lake Michigan, 2011, which could be due to wind and seiche events.

In Chapter 3, we extend previous PLR characterization approaches and develop a peak detection algorithm that automates detection of lake stratification features and deep chlorophyll layers (DCL). This provides a consistent and reproducible reference for identifying lake features to guide future sampling activities. Application to available depth profiles from the Great Lakes identified depths of the thermocline (TRM), lower epilimnion (LEP), and upper hypolimnion (UHY) and deep chlorophyll layers (DCL), which were then validated by comparing with operator judgements. The r^2 is around 0.6 (from 0.40 to 0.83) for LEP, TRM and DCL detections in all of the Great Lakes. Moreover, the algorithms characterize profiles to detect unusual profile

shapes and serve as a tool to quickly visualize spatio-temporal patterns of lake stratifications and the DCL.

In Chapter 4, we propose a spatio-temporal interpolation framework that uses a Bayesian framework and conditional simulations to estimate hypoxia extent in the lakes. With dissolved oxygen data (DO) sampled from loggers in Lake Erie in 2014, 2015, and 2016, we compare and validate three interpolation methods by cross validation. The basis interpolation with Bayesian framework performs the best, with RMSE around 2mg/L and confidence interval coverage around 0.9. Seasonal changes in hypoxia in Lake Erie are then characterized. It is found that the hypoxia started from nearshore or shallower zones in 2014 and the peak hypoxia extent occurred in late September 2014, mid-August 2015, and early September 2016. The interpolation cross-validations also provide insights on logger deployment strategies, indicating that moving some offshore loggers to nearshore areas would reduce uncertainties in interpolated DO levels.

5.2. Future Work

While this thesis has demonstrated successful applications of the analysis frameworks on several datasets in the Great Lakes, many opportunities for extending the analysis remain.

5.2.1 Extending Applications to Other Types of Undulating Data

The methodology, data flow, and Web application proposed in Chapter 2 can be extended to analyze other types of undulating data collected to characterize other phenomena besides river plumes. For example, every year the EPA conducts undulating nearshore monitoring along the shoreline of the Great Lakes. The methods described in this study could allow limnologists to visualize how water-quality parameters change along the shoreline and identify hotspot locations with unusual nearshore features that merit further examination. In offshore areas, this methodology can help limnologists identify and focus on specific phenomena such as deep chlorophyll layers (DCL). Either by direct visualization or input of chlorophyll concentration data into the clustering algorithm, limnologists may quickly identify the locations of clusters of high chlorophyll concentration (i.e. DCL) and compare the differences in zooplankton density and biomass inside and outside of the DCL.

The Web applications can also be improved by (1) coupling with a database so that users don't need to manually save and upload files; and (2) providing real-time information from other external data sources such as USGS river discharges and wave/wind information from Great

Lakes Coastal Forecasting Systems (GLCFS) developed by NOAA-Great Lakes Environmental Research Laboratory. By providing these external data, limnologist onboard the Lake Guardian will be able to conduct a more thorough analysis such as comparing the wind and wave direction with the plume direction revealed by the data to identify possible mechanisms causing the plume dynamics.

5.2.2 Characterization of More Depth Profiling Data

In Chapter 3, we used piecewise linear representations and peak detection algorithms to detect lake stratification patterns and the DCL. These two algorithms describe typical ways to characterize the depth profiling data and can be extended to other water quality data in the lakes. Conductivity and other nutrient data can be similarly analyzed with temperature profiles to detect the halocline and chemocline where a sharp salinity or chemical gradient exists (Boehrer and Schultze, 2008). The peak of dissolved oxygen can be automatically detected by the peak detection algorithm and compared with the DCL data, and limnologists will be able to study patterns in the locations of DCL and DO peaks. In addition, these two algorithms can be easily extended to profiles in other lakes, serving as a tool to help limnologist identify lake stratification and DCL patterns.

However, a careful model parameter tuning process is needed (Fiedler, 2010). If depth labels (i.e. the locations of thermocline, lower epilimnion and upper hypolimnion) are available as in this study, users can calculate the gradients of the labeled data (i.e., temperature gradients of the thermocline as well as gradients of the epilimnion and hypolimnion). These gradients derived from labeled data can then be used to determine the parameters of the detection model. The temperature gradient of the thermocline can be used to determine the minimum TRM gradient threshold (g_{min}^{TRM} in Table 3.1) and the gradients of the epilimnion and hypolimnion can be used as the initial guess of the gradient of stable water layers (g_{stable} in Table 3.1)

Characterization of the depth profiling data reduced whole profiles into a few features, from which cluster analysis can be conducted to group similar profiles, similar to previous research (Richardson, 2002). By analyzing the clusters along space and time, the seasonal and location variability of the depth profiles can be revealed.

The upcast or downcast data sampled by the undulating vehicles in Chapter 2 are depth profiling data, thus the piecewise linear representations and peak detection algorithms could be applied to detect lake stratification and DCL in the raw undulating data as well as the

interpolated data. Then the changes in thermocline or DCL along the ships' path can be more clearly revealed.

5.2.3 Further Analyses with the DO Sensor Network

The hypoxia extent from 2014 to 2016 estimated in Chapter 4 provides a reference for assessing the nutrient load target in Lake Erie. By comparing nitrogen and phosphorus concentrations in the rivers, relationships between nutrient load and hypoxia extent can be derived. In addition, if we assume that the lake bottom DO represent the sediment-water interface, we could estimate sediment oxygen demand using Michaelis-Menton kinetics (Matisoff and Neeson, 2005) in the central basin.

Another possible application is to check the DO interpolations against the fish data. Fish are sensitive to hypoxic conditions and could be used to validate the interpolation model if sufficient data exist. Other external data, especially temperature depth profiling data, can be checked in order to better understand thermocline locations and fluctuations in DO. In addition, the interpolation model can be used in other lakes. We have provided open-source codes for readily conducting the interpolation analysis with the data stored in a MySQL database. (See http://stormxuwz.github.io/Hypoxia_Lake_Erie/ for the database schema).

Finally, the interpolation in Chapter 5 only utilized the location information (longitude, latitude and bathymetry) in the kriging processes. To reduce the uncertainty, more relevant hydrodynamics data (e.g. wave data or water temperature data) can be used as covariates to remove the driving trend in the raw data first and then perform spatial-temporal interpolations on the residuals so that the prediction uncertainty can be reduced (Section 4.4.2). One potential data source would be numerical models such as the Great Lakes Coastal Forecasting Systems (GLCFS).

REFERENCES

Ababou, R., Bagtzoglou, A. C., Wood, E. F., 1994. On the condition number of covariance matrices in kriging, estimation, and simulation of random fields. *Math. Geol. 26* (1), 99–133.

Abbott, M. R., Denman, K. L., Powell, T. M., Richerson, P. J., Richards, R. C., Goldman, C. R., 1984. Mixing and the dynamics of the deep chlorophyll maximum in Lake Tahoe. *Limnol. Oceanogr*. 29 (4), 862–878.

Adler, D., Murdoch, D., others, 2016. rgl: 3D Visualization Using OpenGL.

Ahmed, S., Troy, C. D., Hawley, N., 2014. Spatial structure of internal poincare waves in lake michigan. *Environ. Fluid Mech*. 14 (5), 1229–1249.

Armstrong, M., 1998. Practical aspects of kriging. In: *Basic Linear Geostatistics*. Springer Berlin Heidelberg, pp. 103–116.

Austin, J. A., Colman, S. M., 2007. Lake Superior summer water temperatures are increasing more rapidly than regional air temperatures: A positive ice-albedo feedback. *Geophys. Res. Lett*. 34 (6), L06604.

Bakar, K. S., Sahu, S. K., 2015. sptimer: Spatio-temporal bayesian modelling using r. J. *Stat. Softw*. 63 (15), 1–32.

Barbiero, R. P., Tuchman, M. L., 2004. The deep chlorophyll maximum in Lake Superior. *J. Great Lakes Res*. 30, 256–268.

Beckmann, A., Hense, I., 2007. Beneath the surface: Characteristics of oceanic ecosystems under weak mixing conditions – a theoretical investigation. *Prog. Oceanogr*. 75 (4), 771–796.

Beletsky, D., Hawley, N., Rao, Y. R., Vanderploeg, H. A., Beletsky, R., Schwab, D. J., Ruberg, S. A., 2012. Summer thermal structure and anticyclonic circulation of Lake Erie. *Geophys. Res. Lett*. 39 (6), L06605.

Bennett, E. B., 1978. Characteristics of the thermal regime of Lake Superior. *J. Great Lakes Res*. 4 (3), 310–319.

Bennington, V., McKinley, G. A., Kimura, N., Wu, C. H., 2010. General circulation of Lake Superior: Mean, variability, and trends from 1979 to 2006. *J. Geophys. Res*. 115 (C12), C12015.

Benoit-Bird, K. J., Cowles, T. J., Wingard, C. E., 2009. Edge gradients provide evidence of ecological interactions in planktonic thin layers. *Limnol. Oceanogr*. 54 (4), 1382–1392.

Bivand, R., Hauke, J., Kossowski, T., 2013. Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods. *Geogr. Anal*. 45 (2), 150–179.

Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V., 2008. Areal data and spatial autocorrelation. In: *Applied Spatial Data Analysis with R. Use R!* Springer New York, pp. 237–272.

Bocaniov, S. A., Leon, L. F., Rao, Y. R., Schwab, D. J., Scavia, D., 2016. Simulating the effect of nutrient reduction on hypoxia in a large lake (Lake Erie, USA-Canada) with a three-dimensional lake model. *J. Great Lakes Res*. 42 (6), 1228–1240.

Bocaniov, S. A., Scavia, D., 2016. Temporal and spatial dynamics of large lake hypoxia: Integrating statistical and three-dimensional dynamic models to enhance lake management criteria. *Water Resour. Res*. 52 (6), 4247–4263.

Boehrer, B., Schultze, M., 2008. Stratification of lakes. *Rev. Geophys*. 46 (2), RG2005.

Bouffard, D., Ackerman, J. D., Boegman, L., 2013. Factors affecting the development and dynamics of hypoxia in a large shallow stratified lake: Hourly to seasonal patterns. *Water Resour. Res*. 49 (5), 2380– 2394.

Camacho, A., 2006. On the occurrence and ecological features of deep chlorophyll maxima (DCM) in Spanish stratified lakes. *Limnetica* 25, 453–478.

Catherine, A., Escoffier, N., Belhocine, A., Nasri, A. B., Hamlaoui, S., Yéprémian, C., Bernard, C., Troussellier, M., 2012. On the use of the FluoroProbe, a phytoplankton quantification method based on fluorescence excitation spectra for large-scale surveys of lakes and reservoirs. *Water Res.* 46 (6), 1771– 1784.

Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J., 2017. shiny: Web Application Framework for R. R package version 1.0.3.

Chen, D., Lu, C.-T., Kou, Y., Chen, F., 2008. On detecting spatial outliers. *Geoinformatica* 12 (4), 455–475.

Christensen, R., 2001. Linear models for spatial data: Kriging. In: *Advanced Linear Modeling. Springer Texts in Statistics*. Springer, New York, NY, pp. 269–311.

Cressie, N., Hawkins, D. M., 1980. Robust estimation of the variogram: *I. Math. Geol*. 12 (2), 115–125.

Cullen, J. J., 1982. The deep chlorophyll maximum: Comparing vertical profiles of chlorophyll a. *Can. J. Fish. Aquat. Sci*. 39 (5), 791–803.

Cullen, J. J., 2015. Subsurface chlorophyll maximum layers: enduring enigma or mystery solved? *Ann. Rev. Mar. Sci*. 7 (1), 207–239.

Deng, S., 2014. Local-measure-based landslide morphological analysis using airborne LiDAR data. Ph.D. thesis, The Hong Kong Polytechnic University.

Derecki, J. A., 1976. Heat storage and advection in Lake Erie. *Water Resour. Res*. 12 (6), 1144–1150.

Di Toro, D. M., Connolly, J. P., 1980. Mathematical models of water quality in large lakes part 2: Lake Erie. Tech. rep.

Diaz, R. J., 2001. Overview of hypoxia around the world. *J. Environ. Qual*. 30 (2), 275–281.

Diggle, P. J., Ribeiro, P. J., 2002. Bayesian inference in Gaussian model-based geostatistics. *Geographical and Environmental Modelling* 6 (2), 129–146

Durham, W. M., Stocker, R., 2012. Thin phytoplankton layers: characteristics, mechanisms, and consequences. *Ann. Rev. Mar. Sci*. 4, 177–207.

Edwards, W. J., Conroy, J., Culver, D. A., 2005. Hypolimnetic oxygen depletion dynamics in the central basin of Lake Erie. *J. Great Lakes Res*. 31, 262–271.

Emery, K. O., Csanady, G. T., 1973. Surface circulation of lakes and nearly land-locked seas. *Proc. Natl. Acad. Sci. U. S. A*. 70 (1), 93–97.

Fahnenstiel, G. L., Scavia, D., 1987. Dynamics of Lake Michigan phytoplankton: the deep chlorophyll layer. *J. Great Lakes Res*. 13 (3), 285–295.

Fiedler, P. C., 2010. Comparison of objective descriptions of the thermocline. *Limnol. Oceanogr. Methods* 8 (6), 313–325.

Frame, E. R., Lessard, E. J., 2009. Does the Columbia river plume influence phytoplankton community structure along the Washington and Oregon coasts? J. *Geophys. Re*s. 114 (C2), C00B09.

Gong, X., Shi, J., Gao, H. W., Yao, X. H., 2015. Steady-state solutions for subsurface chlorophyll maximum in stratified water columns with a bell-shaped vertical profile of chlorophyll. *Biogeosciences* 12 (4), 905–919.

Gorham, E., Boyce, F. M., 1989. Influence of lake surface area and depth upon thermal stratification and the depth of the summer thermocline. *J. Great Lakes Res*. 15 (2), 233–245.

Gräler, B., Pebesma, E., Heuvelink, G., 2016. Spatio-Temporal interpolation using gstat.

Green, P. J., Silverman, B. W., 1993. Nonparametric Regression and Generalized Linear Models:

A roughness penalty approach. *Chapman & Hall/CRC Monographs on Statistics & Applied Probability*. CRC Press.

Griffin, P., Getis, A., Griffin, E., 1996. Regional patterns of affirmative action compliance costs. *Ann. Reg. Sci.* 30 (3), 321–340.

Hampton, S. E., Gray, D. K., Izmest'eva, L. R., Moore, M. V., Ozersky, T., 2014. The rise and fall of plankton: long-term changes in the vertical distribution of algae and grazers in Lake Baikal, Siberia. *PLoS One* 9 (2), e88920.

Hannachi, A., Jolliffe, I. T., Stephenson, D. B., 2007. Empirical orthogonal functions and related techniques in atmospheric science: A review. *Int. J. Climatol*. 27 (9), 1119–1152.

Haynes, W., 2013. Tukeys test. In: Dubitzky, W., Wolkenhauer, O., Cho, K.-H., Yokota, H. (Eds.), *Encyclopedia of Systems Biology*. Springer New York, New York, NY, pp. 2303–2304.

Hecky, R. E., Smith, R. E. H., Barton, D. R., Guildford, S. J., Taylor, W. D., Charlton, M. N., Howell, T., 2004. The nearshore phosphorus shunt: a consequence of ecosystem engineering by dreissenids in the laurentian great lakes. *Can. J. Fish. Aquat. Sci*. 61 (7), 1285–1293.

Hickey, B., Geier, S., Kachel, N., MacFadyen, A., 2005. A bi-directional river plume: The Columbia in summer. *Cont. Shelf Res.* 25 (14), 1631–1656.

Hoffman, J. C., Peterson, G. S., Cotter, A. M., Kelly, J. R., 2010. Using stable isotope mixing in a great lakes coastal tributary to determine food web linkages in young fishes. *Estuaries Coasts* 33 (6), 1391–1405.

Hollander, M., A. Wolfe, D., Chicken, E., Hollander, M., A. Wolfe, D., Chicken, E., 2015. The Two-Sample location problem. In: *Nonparametric Statistical Methods*. John Wiley & Sons, Inc., pp. 115–150.

Howell, E. T., Chomicki, K. M., Kaltenecker, G., 2012. Tributary discharge, lake circulation and lake biology as drivers of water quality in the Canadian nearshore of Lake Ontario. *J. Great Lakes Res*. 38, 47–61.

Ishikawa, T., Tanaka, M., 1993. Diurnal stratification and its effects on wind-induced currents and water qualities in Lake Kasumigaura, japan. *J. Hydraul. Res*. 31 (3), 307–322.

Jackson, P. R., García, C. M., Oberg, K. A., Johnson, K. K., García, M. H., 2008. Density currents in the Chicago river: characterization, effects on water quality, and potential sources. *Sci. Total Environ*. 401 (1-3), 130–143.

Jackson, P. R., Reneau, P. C., 2014. Scientific investigations report. Tech. rep.

Janetski, D. J., Iii, C. R. R., Bhagat, Y., Clapp, D. F., 2013. Recruitment dynamics of age-0 yellow perch in a drowned river mouth lake: Assessing synchrony with nearshore Lake Michigan. Trans. *Am. Fish. Soc*. 142 (2), 505–514.

Jones, B., Lee, C., Toro-Farmer, G., Boss, E., Gregg, M., Villanoy, C., 2011. Tidally driven exchange in an archipelago strait: Biological and optical responses. *Oceanography* 24 (01), 142–155.

Kaur, J., Jaligama, G., Atkinson, J. F., DePinto, J. V., Nemura, A. D., 2007. Modeling dissolved oxygen in a dredged Lake Erie tributary. *J. Great Lakes Res*. 33 (1), 62–82.

Keogh, E., Chu, S., Hart, D., Pazzani, M., 2004. SEGMENTING TIME SERIES: A SURVEY AND NOVEL APPROACH. In: *Data Mining in Time Series Databases*. Vol. 57 of Series in Machine Perception and Artificial Intelligence. WORLD SCIENTIFIC, pp. 1–21.

Klausmeier, C. A., Litchman, E., 2001. Algal games: The vertical distribution of phytoplankton in poorly mixed water columns. *Limnol. Oceanogr*. 46 (8), 1998–2007.

Kononenko, I., Kukar, M., 2007. *Machine Learning and Data Mining: Introduction to Principles and Algorithms*. Horwood Publishing.

Kruskal, W. H., Wallis, W. A., 1952. Use of ranks in One-Criterion variance analysis. *J. Am. Stat. Assoc*. 47 (260), 583–621.

Larson, J. H., Trebitz, A. S., Steinman, A. D., Wiley, M. J., Mazur, M. C., Pebbles, V., Braun, H. A., Seelbach, P. W., 2013. Great lakes rivermouth ecosystems: Scientific synthesis and management implications. *J. Great Lakes Res*. 39 (3), 513–524.

Lindström, J., Szpiro, A. A., Sampson, P. D., Oron, A. P., Richards, M., Larson, T. V., Sheppard, L., 2014. A flexible Spatio-Temporal model for air pollution with spatial and Spatio-Temporal covariates. *Environ. Ecol. Stat*. 21 (3), 411–433.

Lips, U., Lips, I., 2014. Bimodal distribution patterns of motile phytoplankton in relation to physical pro- cesses and stratification (Gulf of Finland, Baltic sea). Deep Sea Res. Part 2 *Top. Stud. Oceanogr*. 101, 107–119.

Longhi, M. L., Beisner, B. E., 2009. Environmental factors controlling the vertical distribution of phyto-plankton in lakes. *J. Plankton Res*. 31 (10), 1195–1207.

Lorenzen, C. J., 1966. A method for the continuous measurement of in vivo chlorophyll concentration. *Deep Sea Research and Oceanographic Abstracts* 13 (2), 223–227.

Ludsin, S. A., Zhang, X., Brandt, S. B., Roman, M. R., Boicourt, W. C., Mason, D. M., Costantini, M., 2009. Hypoxia-avoidance by planktivorous fish in Chesapeake Bay: Implications for food web interactions and fish recruitment. *J. Exp. Mar. Bio. Ecol*. 381, Supplement, S121–S131.

Lunven, Michel, Jean François Guillaud, Agnès Youénou, Marie Pierre Crassous, Roger Berric, Erwan Le Gall, Roger Kérouel, Claire Labry, and Alain Aminot. 2005. "Nutrient and Phytoplankton Distribution in the Loire River Plume (Bay of Biscay, France) Resolved by a New Fine Scale Sampler." *Estuarine, Coastal and Shelf Science* 65 (1–2): 94–108.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. *In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. The Regents of the University of California.

Madenjian, C. P., Bunnell, D. B., Desorcie, T. J., Chriscinske, M. A., Kostich, M. J., Adams, J. V., 2012. Status and trends of prey fish populations in Lake Michigan, 2011. Report to the Lake Michigan Committee. Windsor, ON.

Makarewicz, J. C., Howell, E. T., 2012. The Lake Ontario nearshore study: Introduction and summary. *J. Great Lakes Res*. 38, 2–9.

Marcelli, M., Caburazzi, M., Perilli, A., Piermattei, V., Fresi, E., 2005. Deep chlorophyll maximum distribution in the central Tyrrhenian sea described by a towed undulating vehicle. *Chemistry and Ecology* 21 (5), 351–367.

Matisoff, G., Neeson, T. M., 2005. Oxygen concentration and demand in Lake Erie sediments. *J. Great Lakes Res*. 31, 284–295.

McCullough, G. K., Barber, D., Cooley, P. M., 2007. The vertical distribution of runoff and its suspended load in Lake Malawi. *J. Great Lakes Res*. 33 (2), 449–465.

Mellard, J. P., Yoshiyama, K., Litchman, E., Klausmeier, C. A., 2011. The vertical distribution of phyto-plankton in stratified water columns. *J. Theor. Biol*. 269 (1), 16–30.

Moukomla, S., Blanken, P. D., 2016. Remote sensing of the north American laurentian Great Lakes surface temperature. *Remote Sensing* 8 (4), 286.

Nekouee, N., 2012. Dynamics and Numerical Modeling of River Plumes in Lakes. Proquest, UMI Dissertation Publishing.

Oliver, M. A., Webster, R., 2015. Geostatistical prediction: Kriging. In: Oliver, A. M., Webster, R. (Eds.), Basic Steps in *Geostatistics: The Variogram and Kriging. SpringerBriefs in*

Agriculture, Cham, pp. 43–69.

Oliver, S. K., Branstrator, D. K., Hrabik, T. R., Guildford, S. J., Hecky, R. E., 2014. Nutrient excretion by crustacean zooplankton in the deep chlorophyll layer of Lake Superior. *Can. J. Fish. Aquat. Sci*. 72 (3), 390–399.

Oppenheim, A. V., Schafer, R. W., Buck, J. R., 1999. Discrete-Time signal processing, 468.

Ord, J. K., Getis, A., 1995. Local spatial autocorrelation statistics: Distributional issues and an application. *Geogr. Anal*. 27 (4), 286–306.

Page, B. R., Ziaeefard, S., Pinar, A. J., Mahmoudian, N., 2017. Highly maneuverable Low-Cost underwater glider: Design and development. *IEEE Robotics and Automation Letters 2* (1), 344–349.

Pavlac, M. M., Smith, T. T., Thomas, S. P., Makarewicz, J. C., Edwards, W. J., Pennuto, C. M., Boyer, G. L., 2012. Assessment of phytoplankton distribution in the nearshore zone using continuous in situ fluorometry. *J. Great Lakes Res*. 38, 78–84.

Pavlidis, T., Horowitz, S. L., 1974. Segmentation of plane curves. *IEEE Trans. Comput. C*-23 (8), 860–870.

Pebbles, V., Larson, J., Seelbach, P., 2013. Great lakes rivermouths: a primer for managers. Tech. rep., Great Lakes Commission.

Pebesma, E. J., 2004. Multivariable geostatistics in s: the gstat package. *Comput. Geosci.* 30 (7), 683–691.

Posa, D., 1989. Conditioning of the stationary kriging matrices for some well-known covariance models. *Math. Geol*. 21 (7), 755–765.

Rao, Y. R., Howell, T., Watson, S. B., Abernethy, S., 2014. On hypoxia and fish kills along the north shore of Lake Erie. *J. Great Lakes Res*. 40 (1), 187–191.

Rao, Y. R., Schwab, D. J., 2007. Transport and mixing between the coastal and offshore waters in the great lakes: a review. *J. Great Lakes Res*. 33 (1), 202–218.

Ribeiro Jr, P. J., Diggle, P. J., 2016. geoR: Analysis of Geostatistical Data. R package version 1.7-5.2.

Richardson, A. J., Pfaff, M. C., Field, J. G., Silulwane, N. F., Shillington, F. A., 2002. Identifying characteristic chlorophyll a profiles in the coastal domain using an artificial neural network. *J. Plankton Res*. 24 (12), 1289–1303.

Riley, S. C., Roseman, E. F., Nichols, S. J., O'Brien, T. P., Kiley, C. S., Schaeffer, J. S., 2008.

Deepwater demersal fish community collapse in Lake Huron. *Trans. Am. Fish. Soc*. 137 (6), 1879–1890.

Rousseeuw, P. J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math*. 20, 53–65.

Rucinski, D. K., Beletsky, D., DePinto, J. V., Schwab, D. J., Scavia, D., 2010. A simple 1-dimensional, climate based dissolved oxygen model for the central basin of Lake Erie. *J. Great Lakes Res*. 36 (3), 465– 476.

Rucinski, D. K., DePinto, J. V., Scavia, D., Beletsky, D., 2014. Modeling Lake Erie's hypoxia response to nutrient loads and physical variability. *J. Great Lakes Res*. 40, Supplement 3, 151–161.

Ryabov, A. B., Rudolf, L., Blasius, B., 2010. Vertical distribution and composition of phytoplankton under the influence of an upper mixed layer. *J. Theor. Biol*. 263 (1), 120–133.

Sauzède, R., Claustre, H., Jamet, C., Uitz, J., Ras, J., Mignot, A., D'Ortenzio, F., 2015a. Retrieving the vertical distribution of chlorophyll a concentration and phytoplankton community composition from in situ fluorescence profiles: A method based on a neural network with potential for global-scale applications. *J. Geophys. Res. C: Oceans* 120 (1), 451–470.

Sauzède, R., Lavigne, H., Claustre, H., Uitz, J., Schmechtig, C., D'Ortenzio, F., Guinet, C., Pesant, S., 2015b. Vertical distribution of chlorophyll a concentration and phytoplankton community composition from in situ fluorescence profiles: a first database for the global ocean. *Earth Syst. Sci. Data Discuss*. 8 (1), 365–399.

Scavia, D., David Allan, J., Arend, K. K., Bartell, S., Beletsky, D., Bosch, N. S., Brandt, S. B., Briland, R. D., Daloglu, I., DePinto, J. V., Dolan, D. M., Evans, M. A., Farmer, T. M., Goto, D., Han, H., Hook, T. O., Knight, R., Ludsin, S. A., Mason, D., Michalak, A. M., Peter Richards, R., Roberts, J. J., Rucinski, D. K., Rutherford, E., Schwab, D. J., Sesterhenn, T. M., Zhang, H., Zhou, Y., 2014. Assessing and addressing the re-eutrophication of Lake Erie: Central basin hypoxia. *J. Great Lakes Res*. 40 (2), 226–246.

Scavia, Donald, J. David Allan, Kristin K. Arend, Steven Bartell, Dmitry Beletsky, Nate S. Bosch, Stephen B. Brandt, et al. 2014. "Assessing and Addressing the Re-Eutrophication of Lake Erie: Central Basin Hypoxia." *J. Great Lakes Res* 40 (2): 226–46.

Scully, M. E., 2016. The contribution of physical processes to inter-annual variations of hypoxia

in chesapeake bay: A 30-yr modeling study. *Limnol. Oceanogr*. 61 (6), 2243–2260.

Sirovich, L., Kirby, M., 1987. Low-dimensional procedure for the characterization of human faces. *J. Opt. Soc. Am*. A 4 (3), 519–524.

Siswanto, E., Ishizaka, J., Yokouchi, K., 2005. Estimating chlorophyll-a vertical profiles from satellite data and the implication for primary production in the kuroshio front of the east china sea. *J. Oceanogr*. 61, 575–589.

Smith, D. A., Matisoff, G., 2008. Sediment oxygen demand in the central basin of lake erie. *J. Great Lakes Res*. 34 (4), 731–744.

Snepvangers, J. J. J. C., Heuvelink, G. B. M., Huisman, J. A., 2003. Soil water content interpolation using spatio-temporal kriging with external drift. *Geoderma* 112 (3), 253–271.

Snow, G. C., Adams, J. B., Bate, G. C., 2000. Effect of river flow on estuarine microalgal biomass and distribution. *Estuar. Coast. Shelf Sci*. 51 (2), 255–266.

Thomson, R. E., Fine, I. V., 2003. Estimating mixed layer depth from oceanic profile data. *J. Atmos. Ocean. Technol*. 20 (2), 319–329.

Twiss, M. R., Marshall, N. F., 2012. Tributary impacts on nearshore surface water quality detected during a late summer circumnavigation along the 20 m isopleth of Lake Ontario. *J. Great Lakes Res*. 38, Supplement 4, 99–104.

Uitz, J., Claustre, H., Morel, A., Hooker, S. B., 2006. Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. *J. Geophys. Res*. 111 (C8), C08005.

Wang, B., Wu, R., Lukas, R., 2000. Annual adjustment of the thermocline in the tropical Pacific Ocean. *J. Clim.* 13 (3), 596–616.

Watkins, J. M., Collingsworth, P. D., Saavedra, N. E., OMalley, B. P., Rudstam, L. G., 2016. Fine-scale zooplankton diel vertical migration revealed by traditional net sampling and a laser optical plankton counter (LOPC). *J. Great Lakes Res.*

Watkins, J. M., Weidel, B. C., Rudstam, L. G., Holeck, K. T., 2015. Spatial extent and dissipation of the deep chlorophyll layer in lake ontario during the lake ontario lower foodweb assessment, 2003 and 2008. *Aquat. Ecosyst. Health Manag.* 18 (1), 18–27.

White, B., Matsumoto, K., 2012. Causal mechanisms of the deep chlorophyll maximum in Lake Superior: A numerical modeling investigation. *J. Great Lakes Res*. 38 (3), 504–513.

Wickham, H., 2010. *ggplot2: Elegant Graphics for Data Analysis*. Springer New York.

Wickham, H., Hofmann, H., Wickham, C., Cook, D., 2012. Glyph-maps for visually exploring temporal patterns in climate data and models. *Environmetrics* 23 (5), 382–393.

Wulder, M., Boots, B., 1998. Local spatial autocorrelation characteristics of remotely sensed imagery assessed with the getis statistic. *Int. J. Remote Sens*. 19 (11), 2223–2231.

Yu, H., Tsuno, H., Hidaka, T., Jiao, C., 2010. Chemical and thermal stratification in lakes. *Limnology* 11 (3), 251–257.

Yurista, P. M., Kelly, J. R., Miller, S., Van Alstine, J., 2012. Lake Ontario: Nearshore conditions and variability in water quality parameters. *J. Great Lakes Res*. 38, Supplement 4, 133–145.

Yurista, P. M., Kelly, J. R., Miller, S. E., 2009. Lake Superior zooplankton biomass: Alternate estimates from a probability-based net survey and spatially extensive LOPC surveys. *J. Great Lakes Res*. 35 (3), 337–346.

Zhang, H., Culver, D. A., Boegman, L., 2008. A two-dimensional ecological model of lake erie: Application to estimate dreissenid impacts on large lake plankton populations. *Ecol. Modell*. 214 (2–4), 219–241.

Zhou, Y., Obenour, D. R., Scavia, D., Johengen, T. H., Michalak, A. M., 2013. Spatial and temporal trends in Lake Erie hypoxia, 1987-2007. *Environ. Sci. Technol*. 47 (2), 899–905.

Zimmerman, D., Pavlik, C., Ruggles, A., Armstrong, M. P., 1999. An experimental comparison of ordinary and universal kriging and inverse distance weighting. *Math. Geol*. 31 (4), 375–390.

APPENDIX A: KRIGING INTERPOLATION

We provide the related kriging equations in this Appendix. Denote the estimation and true value at location s as $\hat{z}(s)$ and $z(s)$, respectively. $Z_s = [z(s_1), z(s_2), \ldots, z(s_n)]^T$ as the vector of sampled values. Denote the covariance matrix between the sampled values is K_{ss}, the covariance vector between sampled location and upsampled or target location is K_{st}. The element in the covariance matrix $k_{ij} = C(h_{ij})$ where $C(h_{ij})$ is the covariance function with h_{ij} is the distance between location i, j.

In traditional ordinary kriging (Section A.1) or universal kriging (Section A.2), the covariance matrix K or covariance function C is known and accurate, and can be estimated by fitting the empirical variogram or numerically maximizing the likelihood of the sampled data. The covariance function can be used in the conditional simulations (Section A.3). In Bayesian kriging (Section A.4), the uncertainty of the covariance matrix is considered.

A.1 Ordinary Kriging

Kriging is a best linear unbiased estimator (BLUE). The estimation at unsampled location s_0 can be estimated by the linear combination of sampled values, as:

$$\hat{z}(s_0) = \sum_{i=1}^{n} \lambda_i z(s_i) \tag{A.1}$$

where λ_i is the weight for sampled value $z(s_i)$ at location s_i. Ordinary kriging assumes a consistent yet unknown mean μ across the space.

To obey unbiasness, we have:

$$E\big(\hat{z}(s_0) - z(s_0)\big) =$$
$$E\left(\sum_{i=1}^{n} \lambda_i z(s_i) - z(s_0)\right) = \sum_{i=1}^{n} \lambda_i E\big(z(s_i)\big) - E\big(z(s_0)\big) = \mu \sum_{i=1}^{n} \lambda_i - \mu = 0 \tag{A.2}$$

Thus:

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{A.3}$$

To minimize the prediction variances, we have:

$$V(\hat{z}(s_0) - z(s_0)) = E((\hat{z}(s_0) - z(s_0))^2) = E((\sum_{i=1}^{n} z(s_i) - z(s_0))^2) = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j E(z(s_i)z(s_j)) - 2\sum_{i=1}^{n}\lambda_i E(z(s_i)z(s_0)) + E(z(s_0)^2) \quad (A.4)$$

Since

$$Cov(x,y) = E((x-\mu_x)(y-\mu_y)) = E(xy) - \mu_x E(y) - \mu_y E(x) + \mu_x\mu_y \quad (A.5)$$
$$= E(xy) - u^2$$

,we have:

$$E(z(s_i)z(s_j)) = C(h_{ij}) + \mu^2$$
$$E(z(s_i)z(s_0)) = C(h_{i0}) + \mu^2 \quad (A.6)$$
$$E(z(s_0)^2) = C(0) + \mu^2$$

And the variance will become to:

$$V = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C(h_{ij}) - 2\sum_{i=1}^{n}\lambda_i C(h_{i0}) + C(0) + \mu^2(\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j - 2\sum_{i=1}^{n}\lambda_i + 1) \quad (A.7)$$

Since $\sum_{i=1}^{n}\lambda i = 1$, the last term in is reduced to zero. Adding the Lagrange multiplier α, we have:

$$l = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C(h_{ij}) - 2\sum_{i=1}^{n}\lambda_i C(h_{i0}) + C(0) + 2\alpha(\sum_{i=1}^{n}\lambda_i - 1) \quad (A.8)$$

Take the derivative of each weight λ_i and Lagrange multiplier α, and set them to zero, we have:

$$\frac{\partial l}{\partial \lambda_i} = 0 \quad \Rightarrow \quad \sum_{j=1}^{n}\lambda_i C(h_{i,j}) + \alpha = C(h_{i,0}) \quad (A.9)$$

Using matrix representations, we have:

$$\begin{bmatrix} K_{ss} & 1 \\ 1^T & 0 \end{bmatrix}\begin{bmatrix} \Lambda \\ \alpha \end{bmatrix} = \begin{bmatrix} K_{st} \\ 1 \end{bmatrix} \quad (A.10)$$

Where Λ is the weight vector, K_{ss} is the covariance matrix between sampling locations. K_{st} is the covariance matrix between sampling locations and the target location. Then the prediction on the target location will be:

$$z^*(s_0) = \Lambda Z_s \quad (A.11)$$

A.2 Universal Kriging

Universal kriging assumes the mean is a function of external covariates (e.g. longitude, latitude), i.e.:

$$\mu = \mu(s) = \beta^T X(s) \tag{A.12}$$

Where $X(s) = [x_1(s), \ldots, x_p(s)]^T$ (a column with size of p, p is the number of covariates) are the covariates at location s. Then the predicted values are:

$$\hat{z}(s_0) = \sum_{i=1}^{n} \lambda_i z(s_i) = \sum_{i=1}^{n} \lambda_i \sum_{k=1}^{p} \beta_k x_k(s_i) + \sum_{i=1}^{n} \lambda_i \delta(s_i) \tag{A.13}$$

The true value is $z(s_0) = \beta^T X(s_0) + \delta(s_0) = \sum_{k=1}^{p} \beta_k x_k(s_0) + \delta(s_0)$ where $\delta(s)$ is the residual at location s. Comparing the estimated and true value, the un-biasness $E(\hat{z}(s_0) - z(s_0) = 0$ leads to:

$$\sum_{i=1}^{n} \lambda_i X_k(s_i) = X_k(s_0) \text{ for any k} \tag{A.14}$$

Similar to ordinary kriging, the minimization of the prediction variance $V(\hat{z}(s_0) - z(s_0))$ gives:

$$\begin{bmatrix} K_{ss} & X_s \\ X_s^T & 0 \end{bmatrix} \begin{bmatrix} \Lambda \\ \alpha \end{bmatrix} = \begin{bmatrix} K_{st} \\ X_s \end{bmatrix} \tag{A.15}$$

Then the prediction will be:

$$\begin{aligned} z^*(s_0) &= \Lambda Z_s \\ V(s_0) &= K_{tt} - K_{st}^T \Lambda^T + \alpha X_s \end{aligned} \tag{A.16}$$

Where $X_s = [X(s_1), X(s_2), \ldots X(s_n)]$ is the covariates matrix at sampled locations. The coefficients of the covariates can be estimated by the generalized linear square (GLS) method so that:

$$\beta = (X_s^T K_{ss}^{-1} X_s) X_s^T K_{ss}^{-1} z_s \text{ , with the variance of } (X_s^T K_{ss}^{-1} X_s)^{-1} \tag{A.17}$$

A.3 Conditional Simulations

Conditional simulations at the unsampled location u can be calculated by the following equations:

$$Z^{CS}(u) = \Lambda Z_s + R'_u - \Lambda R'_s \tag{A.18}$$

Where Λ are the weights for estimation locations u. R'_u and R'_s are the the non-conditional residual simulations for sampled location and target location, which can be generated by

$$\begin{bmatrix} R_s^{nCS} \\ R_u^{nCS} \end{bmatrix} = L R_N \tag{A.19}$$

where R_N is sampled values from a normal distribution with mean zero and unit variance. L is the matrix via Cholesky decomposition:

$$\begin{bmatrix} K_{ss} & K_{su} \\ K_{su}^T & K_{uu} \end{bmatrix} = LL^T \tag{A.20}$$

A.4 Bayesian Kriging

In Bayesian kriging, three parameters are critical, the β (covariate coefficients) in the trend term, the range ϕ and the scale factor σ in the covariance function.

The inference in bayesian framework is:

$$p(z(s_0)|z(s)) = \int p(z(s_0)|z(s),\theta)p(\theta|z(s))d\theta \tag{A.21}$$

Where $\theta = (\beta, \phi, \sigma)$ are the covariance model parameters. With flat prior of β, improper prior of σ that $p(\sigma^2) \propto 1/\sigma^2$, GeoR package uses the following algorithms (Diggle and Ribeiro, 2002) to generate the parameter posterior distribution, $p(\theta|z(s))$

Step 1: Create a uniform prior distribution for ϕ. In our application, we discretized the ϕ for 30km to 70km with interval 5km.

Step 2: Compute the posterior probabilities $p(\phi|y)$ as

$$p(\phi|y) \propto p(\phi)|V_{\hat{\beta}}|^{1/2}|R_y|^{-1/2}(S^2)^{-(n-p)/2}.$$

Where $V_{\hat{\beta}} = (X'R_y^{-1}X)^{-1}$, $S^2 = \frac{1}{n-1}(y - X\hat{\beta})^T R_y^{-1}(y - X\hat{\beta})$, p is the number of elements of β.

Step 3: Sample a value of ϕ from $p(\phi|y)$

Step 4: Sample from $p\ (\beta, \sigma^2|Y, \phi)$. $p(\beta, \sigma^2|Y, \phi)$ follows a Normal-Scaled-Inverse-Chi square distribution (a product of Normal and Scaled-Inverse-χ2 distribution) that

$p(\beta, \sigma^2|Y, \phi) \sim N - \chi^2_{ScI}(\hat{\beta}^2, V_{\hat{\beta}}, n-p, S^2)$ and $\hat{\beta} = (X'R_y^{-1}X)^{-1}X'R_y^{-1}y$

Step 5: Iterate steps (3) - (4) to get sampled values $\theta = (\beta, \phi, \sigma)$ from $p(\beta, \phi, \sigma|z(s))$.

APPENDIX B: SUPPLEMENTAL INFORMATION

Table B.1 provides some grid information in the kriging interpolation in Chapter 2.

Table B.1 Interpolation Grid Information

Site	Path	Distance (km)	Depth Range [minimum, maximum] (m)	Number of grid points after filtered by convex hull
Manitowoc	1	9.0	[3.00, 33.00]	5195
	2	9.8	[3.00, 18.50]	2670
	3	4.8	[3.00, 27.75]	1724
	4	5.0	[3.25, 25.25]	1496
	5	10.2	[3.50, 11.25]	1375
Muskegon	1	8.0	[3.00, 33.00]	4537
	2	9.4	[3.25, 24.75]	3912
	3	10.0	[3.75, 15.75]	2336
	4	8.0	[2.75, 4.75]	252
Pere Marquette	1	10.6	[1.75, 33.75]	6500
	2	6.4	[3.75, 29.00]	2720
	3	9.8	[3.75, 16.25]	2380
	4	9.6	[2.75, 4.50]	355
	5	7.6	[3.50, 30.25]	2669
	6	7.2	[1.25, 32.00]	2916

Table B.2 provides some mean depth of UHY of Lake Superior in Chapter 3.

Table B.2 Mean depth of UHY (h_i^{UHY}) at each station

Station	Mean UHY Depth (m)	Station	Mean UHY Depth (m)
SU01	36.86	SU11	27.24
SU02	28.19	SU12	30.58
SU03	33.94	SU13	30.61
SU04	28.18	SU14	29.69
SU05	39.25	SU15	41.65
SU06	46.29	SU16	30.28
SU07	31.79	SU17	30.21
SU08	35.56	SU18	25.03
SU09	33.09	SU19	23.64
SU10	46.67		