PREDICTION OF WATER QUALITY PARAMETERS FROM VIS-NIR
RADIOMETRY: USING LAKE ERIE AS A NATURAL LABORATORY
FOR ANALYSIS OF CASE 2 WATERS

A dissertation submitted
to Kent State University in Partial
fulfillment of the requirements for the
degree of Doctor of Philosophy

by

Khalid Adem Ali

August, 2011
Dedicated with love and affection to my beautiful wife, Hana A. Hussein, and our daughters Parveen Adem and Sabreen Adem.
Dissertation written by

Khalid Adem Ali

B.Sc., Addis Ababa University, 2000
M.Sc., Addis Ababa University, 2003
Ph.D., Kent State University, 2011

Approved by

____________________________________ Chair, Doctoral Dissertation Committee
Dr. Joseph D. Ortiz

_____________________________________ Doctoral Dissertation Committee
Dr. Abdul Shakoor

_____________________________________ Doctoral Dissertation Committee
Dr. Elizabeth M. Griffith

_____________________________________ Doctoral Dissertation Committee
Dr. Darren Bade

Accepted by

____________________________________ Chair, Department of Geology
Dr. Daniel K. Holm

____________________________________ Dean, College of Arts and Sciences
Dr. Timothy Moerland
# TABLE OF CONTENTS

LIST OF FIGURES ....................................................................................................................... IX

LIST OF TABLES ............................................................................................................................ XV

CHAPTER 1 Water quality parameters in the optically complex water of the Western Basin of Lake Erie ........................................................................................................................................ 1

1.1 Introduction ............................................................................................................................. 2

1.2 Purpose of Study ......................................................................................................................... 4

1.3 The study area ........................................................................................................................... 7

1.4 Remote sensing of aquatic environment ................................................................................... 12

1.5 Atmospheric correction ............................................................................................................ 18

1.6 Optical properties of Case 2 waters ......................................................................................... 21

1.7 Research Objectives ................................................................................................................. 26

1.8 Dissertation layout .................................................................................................................... 27

1.9 Data used in this study .............................................................................................................. 29

1.9.1 Field data ............................................................................................................................ 29

1.9.2 Remote sensing data ........................................................................................................... 30

CHAPTER 2 Application of satellite algorithms to estimate CPAs .............................................. 34

ABSTRACT ...................................................................................................................................... 35

2.1 Introduction ............................................................................................................................. 36

2.2 Data acquisition and methodology ......................................................................................... 44
2.2.1 Field measurements ........................................................................... 44
2.2.2 Satellite imagery ................................................................................. 45
2.2.3 Spectral analysis ................................................................................ 47
2.3 Chlorophyll $a$ prediction ..................................................................... 49
  2.3.1 Empirical models .............................................................................. 49
  2.3.2 Semi-empirical models ..................................................................... 52
  2.3.3 Fluorescence Line height ................................................................. 53
2.4 CDOM prediction .................................................................................. 55
2.5 Calibration-validation .......................................................................... 57
2.6 Results and Discussion ......................................................................... 58
  2.6.1 CPAs in the WBLE ......................................................................... 61
2.7 Chlorophyll $a$ prediction ..................................................................... 65
  2.7.1 Empirical models .............................................................................. 65
  2.7.2 Semi-analytical models .................................................................... 72
  2.7.3 Fluorescence Line Height (FLH) ..................................................... 76
  2.7.4 Calibration-validation of chlorophyll $a$ algorithms ......................... 79
2.8 CDOM .................................................................................................. 80
2.9 Uncertainties ........................................................................................ 82
  2.9.1 Assumption of negligible effects due to other components ............. 82
  2.9.2 Temporal variability ........................................................................ 83
  2.9.3 Pixel heterogeneity .......................................................................... 84
2.9.4 Path radiance correction

2.10 Conclusion

CHAPTER 3 Application of principal component analysis to estimate color producing agents in turbid waters using spectrophotometer data

ABSTRACT

3.1 Introduction

3.2 Materials and methods

3.2.1 Study area

3.2.2 Data acquisition

3.3 Methodology

3.3.1 Multivariate Statistics

3.4 Results and discussion

3.5 PCA analysis of the lab based VIR derivative spectra

3.5.1 Optical properties representing the joint spatiotemporal signal variability

3.5.2 Optical properties representing the temporal signal variability

3.6 Prediction of CPAs

3.6.1 PC regression against chlorophyll $a$ levels

3.6.2 PC regression against concentrations of TSM

3.6.3 Calibration-validation

3.7 Conclusion
CHAPTER 4 Retrieval of chlorophyll a, phycocyanin and TSM from hyperspectral reflectance data using discrete wavelet transformation ........................................ 138

ABSTRACT ........................................................................................................................................ 139

4.1 Introduction .................................................................................................................................. 141

4.2 Wavelet background ..................................................................................................................... 143

4.2.1 Discrete Wavelet Transforms ................................................................................................. 148

4.3 Materials and Methods .............................................................................................................. 151

4.3.1 Lake Erie .................................................................................................................................. 151

4.3.2 Data acquisition ....................................................................................................................... 152

4.3.3 Laboratory analysis .................................................................................................................. 153

4.3.4 Multi-resolution hyperspectral data analysis ......................................................................... 153

4.3.5 Multiple linear regression analysis (MLA) ............................................................................ 158

4.3.6 Calibration-validation ............................................................................................................. 160

4.4 Results and discussion .............................................................................................................. 161

4.4.1 Spatial and temporal variability of CPAs .............................................................................. 161

4.4.2 Discrete wavelet transformation ............................................................................................. 165

4.5 Regression models ..................................................................................................................... 175

4.5.1 Performance of the selected wavelet functions ..................................................................... 175

4.5.2 Prediction models using various frequency components ....................................................... 176

4.5.3 Prediction models using the whole range of scales ............................................................... 185

4.5.4 Calibration-validation ............................................................................................................. 189
4.6 Conclusion .......................................................................................................................... 191

CHAPTER 5 Using FR-MERIS multispectral satellite data to estimate CPAs in the Western Basin of Lake Erie: a neural network approach ......................... 193

ABSTRACT .......................................................................................................................... 194

5.1 Introduction ..................................................................................................................... 196

5.2 Artificial Neural Networks .............................................................................................. 202

5.3 Materials and Methods .................................................................................................. 203

5.3.1 Western Basin of Lake Erie ....................................................................................... 203

5.3.2 Data acquisition ......................................................................................................... 203

5.3.3 Artificial Neural Network ......................................................................................... 210

5.3.4 Performance measurements ....................................................................................... 216

5.4 Results and discussion ................................................................................................... 217

5.4.1 The ANN structure .................................................................................................... 217

5.4.2 Cross-validation of the ANN ..................................................................................... 218

5.4.3 CPA dynamics using the ANN algorithm ................................................................. 225

5.5 Conclusion ...................................................................................................................... 228

CHAPTER 6 Summary and Future work ............................................................................... 230

6.1 Summary ........................................................................................................................ 231

6.1.1 Spectral indices .......................................................................................................... 233

6.1.2 Principal Component Analysis .................................................................................. 234

6.1.3 Wavelets .................................................................................................................... 236
6.1.4 Neural Networks ................................................................. 236
6.2 Model performance ............................................................... 237
6.3 Best model selection .............................................................. 239
6.4 Future directions ................................................................. 241
REFERENCES ........................................................................... 243
LIST OF FIGURES

Figure 1. 1. The Great Lakes watershed consists of 5 large lakes; Lake Superior, Lake Michigan, Lake Huron, Lake Ontario and Lake Erie in the southern most region..........................................................10

Figure 1. 2. Lake Erie bathymetry with longitudinal profile section.......................11

Figure 1. 3. Absorption (log (1/R)) spectra of pure water increases with wavelength showing an exponentially increase in the near-infrared region...............17

Figure 1. 4. Average reflectance spectra of the Western Basin of Lake Erie from the MERIS bands after path radiance correction. ........................................21

Figure 1. 5. In-situ spectra of the Western Basin of Lake Erie at 18 stations corresponding to the sampling points for the study. .........................23

Figure 1. 6. Example spectra of the inherent optical properties of photoactive components in a Case 2 waters.........................................................25

Figure 2. 1. A schematic of the FLH model on MERIS based spectra of a pixel representing the WBLE. .................................................................54

Figure 2. 2. Spectral plot of pixels representing 18 stations based on MERIS data......60

Figure 2. 3. Spectral plot of pixels representing 18 stations based on MODIS data.....60

Figure 2. 4. Spatial and temporal variability of CDOM in the WBLE during summer of 2010 based on data from in-situ measurements .................................64
Figure 2. 5. In-situ measurement of CPAs shows the variability of chlorophyll $a$ and CDOM in the WBLE during summer of 2010. ........................................64

Figure 2. 6. Regression plot of MERIS based blue-green chlorophyll $a$ model versus in-situ chlorophyll $a$ concentrations .................................................................66

Figure 2. 7. Regression plot of MERIS based two-band red/NIR chlorophyll $a$ model versus the chlorophyll $a$ concentrations measured in-situ.................................69

Figure 2. 8. Regression plot of MODIS based two-band red/NIR chlorophyll $a$ model versus the chlorophyll $a$ concentrations measured in-situ.................................69

Figure 2. 9. Regression plot of MERIS based three-band red/NIR chlorophyll $a$ model versus the chlorophyll $a$ concentrations measured in-situ...............................71

Figure 2. 10. Regression plot of MERIS based Simis chlorophyll $a$ model versus the chlorophyll $a$ concentrations measured in-situ........................................73

Figure 2. 11. Regression plot of MERIS based Gons chlorophyll $a$ model versus the chlorophyll $a$ concentrations measured in-situ........................................75

Figure 2. 12. Regression plot of MERIS based FLH model versus the chlorophyll $a$ concentrations measured in-situ ................................................78

Figure 2. 13. Regression plot of ALI based CDOM concentrations versus in-situ measurements. ........................................................................................................81

Figure 3. 1. Some examples of remote-sensing reflectance spectra from different types of waters ..............................................................93
Figure 3.2. Reflectance spectra (400-860nm) recorded by the Medium Resolution Imaging Spectrometer (MERIS) in the WBLE at pixels corresponding to 18 sampling stations.

Figure 3.3. The Western Basin of Lake Erie and the locations of 18 sampling stations visited during multiple cruises of Research Vessels Gibraltar and Erie.

Figure 3.4. Reflectance spectra (400-2500nm) generated using lab-based spectrophotometer for samples collected during the four research expeditions.

Figure 3.5. Center-weighted first derivative of the reflectance spectra in figure 2.2, the derivative removes and accentuates absorption features in the visible and NIR.

Figure 3.6. Plots of concentrations of water constituents along the cruise track (a) chlorophyll a concentrations in µg/l at each station and (b) gravimetrically estimated total suspended matter in mg/l at each sampling point.

Figure 3.7. The “scree plot” represents eigenvalues scaled as the percentage of variance explained by each factors.

Figure 3.8. Plots of principal factors as function of wavelength. The three eigenvectors account for approximately 87% of the optical variability observed in the WBLE during the summer season of 2009 and 2010.

Figure 3.9. Factor loadings for the PCA and selected reference derivative spectra.

Figure 3.10. Factors scores from the PCA model.
Figure 3. 11. Factor loadings and selected reference derivative spectra for 2009........124

Figure 3. 12. Factor loadings and selected reference derivative spectra for 2010........125

Figure 3. 13. Regression analysis between predicted chlorophyll $a$ levels based on model 3 and in-situ chlorophyll $a$ values. .................................................................131

Figure 3. 14. Regression analyses between predicted TSM levels and gravimetrically measured TSM in the WBLE: .................................................................133

Figure 4. 1. Examples of wavelet function filters used to decompose the hyperspectral signal .................................................................146

Figure 4. 2. Signal decomposition using wavelets that have been scaled (stretched) and shifted. .................................................................148

Figure 4. 3. High-pass (H) and low-pass (L) filters are applied to the input signal, $f(t)$, having 64 discrete bands. .................................................................157

Figure 4. 4. Spatial variability of (a) chlorophyll $a$ and (b) TSM, in the WBLE during the different cruise periods. .................................................................163

Figure 4. 5. Sandusky River discharge in 2009 and 2010. Note the vertical axis for 2009 and 2010 data...............................................................................164

Figure 4. 6. Temporal and spatial variability of phytoplankton pigments in WBLE ...164

Figure 4. 7. First-derivative transformed hyperspectral data of 18 stations used as an input for DWT.................................................................167

Figure 4. 8. A 4 level discrete transformation of the first-derivative hyperspectral signal of 18 stations (a), and 4 levels of decomposition (b). .........................170
Figure 4. 9. Accuracy of chlorophyll a regression models generated using wavelet coefficients at various scales. .................................................................174

Figure 4. 10. Plots of chlorophyll a regression model output (R^2 and RMSE) as function of wavelet frequency components. .........................................................178

Figure 4. 11. Performance of regression models for chlorophyll a prediction developed using wavelet coefficients generated based on MRA of the first-derivative spectrum. .........................................................................................180

Figure 4. 12. Results from MRA on wavelet coefficients at each scale (cD1, cD2, cD3, cD4, cD5 and cA5) using the Db4 wavelet function........................................182

Figure 4. 13. Plots of Phycocyanin and TSM regression model output (R^2) as function of multiple frequency components. .........................................................184

Figure 4. 14. Results of MLR on the wavelet coefficients derived from the various wavelet functions by applying DWT...............................................................186

Figure 4. 15. Frequency statistics of the R^2 values generated for each type of CPA proxies using the 17 wavelet functions. .........................................................188

Figure 5. 1. MERIS data acquired on 24 June 2009 with locations of in-situ stations.205

Figure 5. 2. Single hidden layer ANN configuration.................................................................211

Figure 5. 3. The SHLP structure constructed for the WBLE.....................................................217

Figure 5. 4. Effects of the network’s structural complexity on ANN performance for (5.4a) chlorophyll a and (5.4b) TSM, estimation.....................................................219
Figure 5. 5. Relationship between predicted and observed concentrations of (5.5a) chlorophyll $a$ and (5.5b) TSM. Algorithm is trained using the Random Holdback approach. 223

Figure 5. 6. Relationship between predicted and observed concentrations of (5.6a) chlorophyll $a$ and (5.6b) TSM. Algorithm is trained using the 8-fold cross-validation technique. 224

Figure 5. 7. Observed and predicted (Prd) values for chlorophyll $a$ (5.7a), and TSM (5.7b) at 18 stations across the WBLE. 227
LIST OF TABLES

Table 1.1. MERIS instrument specification onboard ENVISAT (Environmental Satellite) ........................................................................................................................................31

Table 1.2. MODIS instrument specification onboard Terra and Aqua satellites .... 33

Table 2.1. Summary of calibration-validation of chlorophyll a algorithms ............79

Table 3.1. Statistics of chlorophyll a and total suspended matters (TSM) .............111

Table 3.2. Descriptive statistics of selected principal components. .....................114

Table 3.3. Shows the PCAs with leading PC patterns, the corresponding factor loadings, ........................................................................................................................................115

Table 3.4. Correlation analysis between factor scores ........................................127

Table 3.5. Collinearity indices of three standardized principal component regression models for chlorophyll a prediction. .........................................................................................128

Table 3.6. Collinearity indices of three standardized principal component regression models for TSM prediction. .................................................................128

Table 3.7. Regression analysis of PC scores against chlorophyll a (n=89). .........130

Table 3.8. Regression analysis of PC scores against TSM (n=89). .......................132

Table 3.9. Summary of the validation parameters computed using the calibration and validation dataset for the PCs and two CPAs.........................................................135

Table 4.1. Descriptive statistic of the selected CPA proxies..................................162
Table 4. 2. Percent variance of the various frequencies in the original hyperspectral reflectance data……………………………………………………………………………169

Table 4. 3. Percent variance of the various frequency components that exist in the first-derivative transformed hyperspectral data. .................................................172

Table 4. 4. Performance of the three wavelets functions in explaining chlorophyll $a$ variability in the WBLE. .................................................................................179

Table 4. 5. Results of multiple linear regression models (R2 and RMSE) using various wavelet functions to estimate selected CPA proxies.........................186

Table 4. 6. Descriptive statistics of the R$^2$ values generated for each type of CPA proxies using various wavelet functions. ......................................................187

Table 4. 7. Summary of the calibration-validation parameters of wavelet based bio-optical models...............................................................................................190

Table 5. 1. MERIS Level-1b characteristics (after European Space Agency, ..........208

Table 5. 2. MERIS bands wavelength, bandwidth and applications ..........................209

Table 5. 3. Dates of each cruise and the date of the corresponding MERIS overpass of WBLE.................................................................................................209

Table 5. 4. Summary of the ANN validation parameters using the Random Holdback approach based on training and validation data set. .........................221

Table 5. 5. Summary of the ANN validation parameters using the K-fold cross-validation approach based on training and validation data set. .............222
Table 6.1. Performance of the various applied models in the retrieval of CPAs. The Bias% is the RMSE bias relative to a 1:1 regression line………………238

Table 6.2. Akaike’s best model selection criteria for chlorophyll a and TSM estimation …………………………………………………………………………………. 240
ACKNOWLEDGEMENT

I would like to express my gratitude to my advisor Dr. Donna L. Witter for her support and guidance throughout the course of this project. The research tips and the diligent guidance I received from her was hugely responsible for the successful completion of this research. I would like to thank the Chairman of my Supervisory committee and my current advisor Prof. Joseph D. Ortiz. I enjoyed discussing the various aspects of this research with him. I have learnt a lot of science and its practices during our communications over the years. It was a privilege for me to work with Dr. Witter and Dr Ortiz. I would like to thank members of my Supervisory Committee, namely, Prof. Abdul Shakoor, Dr Elizabeth Griffith, Dr Darren Bade and Dr Shanhu Lee.

Financial support came through two grants from the Ohio Sea Grant. The satellite images were provided free of charge by NASA and the European Space Agency.

Field data collection was a huge component of this research. I would like to thank all of the undergraduates and graduates who volunteered to go out on the Ship cruise. I would especially like to thank Renee Crane who was instrumental in ensuring that all equipments were tested and fully functional prior to setting out to the lake. She also helped processing the samples in the labs. Matt Thomas at the Stony Lab was very generous in sparing his time and energy and driving us on the Research Vessel when we went out to collect data.
I am also very thankful to the faculty and students at the Department of Geology, Kent State who helped keep an atmosphere that was conducive to my studies. I appreciate all the support from Karen smith, secretary of the department and all of my friends at Kent State including Yonathan Admassu, Tej Gautum and James Fisher.

Words cannot adequately express my thanks to my parents who have been the source of motivation for me to pursue my education. Finally, my heart-felt gratitude and thanks goes to my wonderful wife Hana Hussien, whose love and unmatched dedication encouraged me in ways that cannot be measured. My two daughters, Parveen and Sabreen have made sacrifices during my studies and they always helped relieve my stress during peak moments.
CHAPTER 1

Water quality parameters in the optically complex water of the Western Basin of Lake Erie
1.1 Introduction

The color of naturally occurring waters is often considered as an indicator of the physical, chemical and biological characteristics. These characteristics, in turn, define its quality. The interaction of incident light with the various in-water constituents results in optical features that can be detected by satellite sensors. Therefore, data from optical sensors can be used to interpret water quality using spectral properties of the organic and inorganic components of the water body. In recent years, satellite based optical remote sensing has gained considerable attention as a tool to monitor water quality of large water bodies (Dekker et al., 1997; Ritchie et al., 2003; Dall ‘Olmo et al., 2003; Gons, 1999; Gons et al., 2002; Gitelson et al., 1985, 2008). The major photoactive proxies that are used to characterize water quality include chlorophyll $a$, a dominant pigment which is present in all eukaryotic algae and the cyanobacteria, suspended material and the concentrations of dissolve organic matter. Morel and Prieur (1977) have classified marine waters as Case 1 or Case 2 based on optical properties. In Case 1 marine waters, optical properties are dominated by a single component, phytoplankton, a number of algorithms have been developed for estimating chlorophyll $a$ concentration and some consensus is being reached with regards to an appropriate algorithm (IOCCG, 2006; Gordon and Morel, 1983; O’Reilly et al., 1998). Coastal and most inland waters are classified as Case 2 waters, because in addition to phytoplankton, terrestrially derived constituents such as dissolved organic matter and suspended sediments occur in abundance and therefore the optical properties of such waters represents an aggregate spectrum of the various
components. Although coastal and inland waters are of greater importance to human interests such as drinking water, industry, fishing and recreation, algorithm development is less mature for these environments.

Optical properties of coastal and inland waters are influenced by multiple components which may include phytoplankton, suspended materials (SM), colored dissolved organic matter (CDOM) and bottom reflection. The diversity of optically active constituents make these water bodies optically complex, and the task of optical remote sensing for retrieving estimates of the CPAs more challenging. This optical complexity is the result of terrestrial interaction, anthropogenic inputs via river discharges and human activities. These interactions load various components into lacustrine waters, thus influencing the spectral signature of these waters. In these cases, the multiple in-water constituents may have complex associations, resulting in non-linear relationships.

The goal of this study is to assess the utility of satellite remote sensing to estimate water quality parameters in Case 2 waters. This problem can be approached by accurately determining the detectable signatures of CPAs within the reflectance spectrum. The spectral signatures, which are the result of absorption and scattering of light, can be used to estimate concentrations of the optically independent constituents within surface water layers (Schofield et al., 2004). Multiple techniques including multivariate statistics and mathematical signal transformations are applied to develop algorithms that relate remote sensing reflectance and in-water constituents. The derivation of water quality parameters from satellite visible/near-infrared (VIR) spectral radiometry can provide frequent spatial
information critical to the understanding of biogeochemical processes of interest to climate studies, such as carbon exchange and phytoplankton and sediment dynamics (IOCCG, 2008).

1.2 Purpose of Study

The Great Lakes represent the largest surface freshwater system on the Earth (Fig 1). These lakes contain about 84 percent of North America’s surface fresh water supply and about 21 percent of the world’s surface fresh water supply (USEPA, 2006). The residence time of water in the Great Lakes is highly variable, ranging from less than 3 years for Lake Erie to about 190 years for Lake Superior (USEPA, 2006). The short residence time for Lake Erie makes this lake particularly sensitive to environmental stresses.

Lake Erie is the most populated among the Great Lakes with over ten million residents depending on the lake for drinking water purposes (USEPA, 2006). As a result of population density and land use patterns, Lake Erie receives particularly high volumes of sediment loading from agricultural erosion and effluent from sewage treatment plants (USEPA, 2006). Such input has resulted in symptoms such as high concentrations of dissolved organic matter and suspended particulates which creates favorable condition for the formation of Harmful Algal Blooms (HABs) (Becker et al., 2009). The eutrophication process causes algal blooms that have detrimental effects on water bodies resulting in surface scums, unpleasant taste, odor problems and reduced economic vitality in the
recreation and fishing industry. A minority of algal blooms are composed of toxin producing taxa e.g. *Microcystis* and *Anabaena* (Boyer, 2008). Introduction of *Microcystis* into the human body may have acute hepatotoxic effect with symptoms such as diarrhea and vomiting. There are also potential risks to skin exposure.

When algae die, the decay process naturally consumes oxygen and can cause hypoxic or anoxic conditions that are harmful to other species. This is exacerbated during summer due to thermal stratification. In recent years, oxygen depletion and the extent of both harmful and nuisance algal blooms have increased in Lake Erie, resulting in the development of the “dead zone” (USEPA, 2006). Despite a general decline in eutrophication (Matisoff and Ciborowski, 2005), high concentrations of *Microcystis* have been recorded in the Western Basin of Lake Erie in late summer to fall of 2000 through 2009 as a result of influx of higher amounts of nutrients (Boyer, 2008; Millie et al., 2009).

Water quality assessments for Lake Erie are largely based on in-situ biogeochemical measurements as well as collection of water samples for further laboratory analysis. This approach has limited spatial and temporal resolution and makes it difficult to understand the spatial and temporal dynamics of color producing agents (CPAs) such as phytoplankton, suspended materials and dissolved organic matter. Satellite-based synoptic measurements using multi-spectral sensors (MODIS, MERIS) have better spatial and temporal resolution, but require the development of algorithms
that transform measured spectral reflectance to estimates of CPAs (e.g., Doerffer and Fischer, 1994; IOCCG, 2000; McClain, 2009).

Currently, isolating remote sensing signals of a single component (e.g. chlorophyll a) from Case 2 waters is a challenge. In this dissertation, I develop models that apply techniques such as multivariate statistics, wavelet analysis and Neural Network (NN) to remote sensing data to better differentiate among and quantitatively estimate the concentrations of CPAs in the optically complex environment of the Western Basin of Lake Erie. The results from this analysis, and the conceptual approach, will apply to a variety of optically complex environments such as coastal areas and other large inland water bodies.
1.3 The study area

Lake Erie is the twelfth largest lake in the world, and the fourth largest and southernmost of the Great Lakes (Fig 1.1). It gets most of its inflow from the Detroit River and feeds the Niagara Falls. It consists of three basins: the shallow Western Basin which has an average depth of 7m, the Central Basin with an average depth of 18m and the deeper Eastern Basin with an average depth of 30m (Fig 1.2). Lake Erie is about 367km long and has variable width ranging between 48 and 97km.

The presence of three separate basins is attributed primarily to the geological formations making up the bedrock underlying the basins. The Eastern Basin is composed of soft Devonian shale surrounded by hard limestone on the eastern side making the Niagara Falls. During the Pleistocene, glaciers scoured the shale to form the deep bottom of the Eastern Basin. Farther west, where the width of the shale belt is greater, glacial erosion resulted in a broader but shallower Central Basin. The Devonian and Silurian dolostone and limestone make up the Western Basin and its Islands (Bolsenga and Herdendorf, 1993). These harder formations resisted glacial and other erosional forces, resulting in shallower depths in the Western Basin.

The Western Basin of Lake Erie (WBLE) is shallow enough to have dramatic wind- and wave-driven turbidity. Its relatively warm temperature makes it conducive to high biologic productivity. In some regions of the basin, algal mats form annually during
the summer. A number of rivers serve as conduits for fluxes of nutrients and sediments into the WBLE, influencing water clarity, particularly near the mouths of the Sandusky and Maumee Rivers (Fig 1.1 inset). These rivers collect water mainly from the adjacent agricultural fields.

Being the smallest by volume and the shallowest of the Great Lakes, the water circulation through the WBLE is faster and it responds more rapidly to ecologic stressors such as pollutants, invasive species, and climate change (Quinn, 1992). In recent years, oxygen depletion and the extent of both harmful and nuisance algal blooms have increased in Lake Erie, despite a general decline in eutrophication (Becker et al., 2009; Rinta_Kanto et al., 2005). Recent reports by the National Center for Water Quality (NCWQR) indicate that, although total phosphorus levels remained within the recommended limits of the US Environmental Protection Agency, levels of Soluble Reactive Phosphorus (SRP) have increased. SRP is continuously being loaded from the Maumee and Sandusky Rivers to the Western Basin of Lake Erie. According to Kane et al. (2008), there is positive linear correlation between concentrations of cyanobacteria and SRP.

A potentially toxic filamentous cyanobacterium, *Lyngbya wolleti* has been documented in the WBLE, where it forms algal mat along with the summer bloom of the *Microcystis* (Bridgman and Penamon, 2010). As a result, the bloom formation is prolonged, extending into the winter period. This causes an expansion of the “dead zone” in the Central Basin of Lake Erie, impacting its aesthetic and human use.
Because Lake Erie has dynamic in-water constituents, satellites can enhance monitoring efforts by providing synoptic measurements. The current generation of imaging satellite sensors can synoptically monitor large areas at relatively high pixel resolution with repeat sampling intervals ranging from days to weeks. This study focuses on the Western Basin of Lake Erie (83 to 82.5 W and 41.2 to 41.7 N, Figure 1.1 inset), to assess the utility of remote sensing data to detect and map the various optically active constituents.
Figure 1. The Great Lakes watershed consists of 5 large lakes; Lake Superior, Lake Michigan, Lake Huron, Lake Ontario and Lake Erie (bathymetry) in the southern most region. Inset: Bathymetry map of the Western Basin of Lake Erie. The red-brown lines represent shallowest area region (0-4m), and greenish lines are for intermediate depths (5-7m) and the light and dark blue lines corresponds to deeper waters (8 - 12m).
Figure 1. 2. Lake Erie bathymetry map. The Western Basin is shallow with average depth of 7m consisting of dolomite and limestone, the central basin has average depth of 18m and the Eastern Basin is covered by shale formation that is deeply scarped by the glaciers, the scour has depth of up to 64m (NOAA).
1.4 Remote sensing of aquatic environment

Multispectral remote sensing of CPAs involves the acquisition of surface reflectance in discrete bands in the visible and near-infrared spectrum. Satellite-based remote sensing is increasingly being used to monitor water quality of large lakes, such as Lake Erie at higher temporal and spatial resolution than is available from in-situ sensors. Light reaching a water body is reflected from the surface, but most of it is transmitted into water, where it will be attenuated by the various in-water constituents that interfere with the radiation path. The optical property of the water is defined by spectral variations in reflectance, \( R(\lambda) \), which is defined as the ratio of up-welling radiance \( L_u(\lambda) \) to down-welling irradiance \( E_d(\lambda) \), mathematically expressed as:

\[
R(\lambda) = \frac{L_u(\lambda)}{E_d(\lambda)} \quad (1.1)
\]

Optical properties of aquatic bodies are classified as Inherent Optical Properties (IOPs) and Apparent Optical Properties (AOPs). The IOPs represent the scattering and absorption characteristics of the in-situ water and its constituents, whereas the AOPs are directly derived from the light field, radiance reflectance, and therefore the scattering and absorption values are a function of the light geometry. Changing the light field that enters the water body, for example, by a changing solar zenith or nadir changes the AOP, but not the IOP.
Primary objectives of aquatic visible and near-infrared remote sensing include: identifying the optical properties of in-water constituents and to quantify them. To address these questions, radiative transfers have been employed to express results of reflectance in-terms of backscattering and absorption coefficients of the various in-water constituents. One approach that was used for radiative transfer is the Monte Carlo simulation of photon propagation through aquatic media (Gordon et al., 1988; Plass and Kattawar, 1972). Radiative transfer models show that reflectance can be expressed as the backscattering coefficient ($b_b(\lambda)$) and the absorption coefficient ($a(\lambda)$) i.e.,

$$R(\lambda) = \frac{f(\lambda)}{Q(\lambda)} \frac{b_b(\lambda)}{a(\lambda)+b_b(\lambda)}$$  \hspace{1cm} 1.2

The $f$ and $Q$ factor represent the geometry of the light field that affects the $R(\lambda)$.

Most of the orbital satellites that are used for our study are synchronous, and therefore solar zenith angle and the geometry of the light field emerging from the water remain relatively constant for a particular time (Morel and Gentili, 1993). A modified version of equation 1.2 with a light field property reduced to single parameter is defined as:

$$R(\lambda) = C \frac{b_b(\lambda)}{a(\lambda)+b_b(\lambda)}$$ \hspace{1cm} 1.3

Monte-Carlo studies by Kirk (1984) have found the coefficient $C$ to be a function of the cosine of the solar zenith angle, $\mu_o$, during satellite overpass and it is expressed as:
\[ C(\mu_o) = -0.629\mu_o + 0.97 \]  \hspace{1cm} 1.4

Following the equation above, the reflectance just below water’s surface can be calculated by:

\[ R(\lambda) = (-0.629\mu_o + 0.975) \frac{b_b(\lambda)}{a(\lambda)+b_b(\lambda)} \]  \hspace{1cm} 1.5

For e.g., in the Western Basin of Lake Erie solar zenith angle is 67° and 80° during overpass of MERIS (dated 07/28/2010) and MODIS (dated 07/28/2010), respectively. For the MERIS, the relationship predicts \[ C(cos(67)) = 0.729 \] and for MODIS, \[ C(cos(80)) = 0.865 \].

As light passes through the water-air interface, it refracts to higher angle. This reduces the total backscattering by up to 50% (Morel and Prieur, 1977). This factor is accounted for in the equation, and therefore the reflectance just above water surface is described as:

\[ R(\lambda) = 0.5(-0.629\mu_o + 0.975) \frac{b_b(\lambda)}{a(\lambda)+b_b(\lambda)} \]  \hspace{1cm} 1.6

The total backscattering \( (b_b(\lambda)) \) and absorption, \( a(\lambda) \), coefficients recorded at sensor can be spectrally partitioned into the constituents of the medium (water, particles and dissolved materials). The absorption spectra in Lake Erie are assumed to be the sum contributions of pure water, phytoplankton, dissolved organic matter and suspended matter (Bukata et al., 1995).
The total spectral absorption coefficient \(a(\lambda)\) in m\(^{-1}\), within the water body can be described by:

\[ a(\lambda) = a_w(\lambda) + a_{ph}(\lambda) + a_{dom}(\lambda) + a_{SM}(\lambda) \] 1.7

Or

\[ a(\lambda) = a_w(\lambda) + a_{ph}(\lambda)c_{Chl} + a_{dom}(\lambda) + a_{SM}(\lambda)c_{SM} \] 1.8

Where \(a_w(\lambda), a_{ph}(\lambda), a_{dom}(\lambda)\) and \(a_{SM}(\lambda)\) are the absorption coefficients of pure water, phytoplankton, dissolved organic matter and suspended matter, respectively. \(a_{ph}^*(\lambda)\) is chlorophyll-specific spectral absorption coefficient of phytoplankton and \(a_{SM}^*(\lambda)\) is the specific spectral absorption coefficient of suspended matter. \(c_{Chl}\) and \(c_{SM}\) are concentrations of chlorophyll a and suspended matter.

The spectral backscattering coefficient is another inherent optical property that affects the magnitude of the reflectance from water bodies (Gons, 1999). The total spectral absorption coefficient, \(b_b(\lambda)\) with in the water body can be described by:

\[ b_b(\lambda) = 0.5b_{bw}(\lambda) + b_{bph}(\lambda) + b_{bSM}(\lambda), \quad b_{bdom}(\lambda) \approx 0 \] 1.9

or

\[ b_b(\lambda) = 0.5b_{bw}(\lambda) + b_{bph}(\lambda)c_{Chl} + a_{dom}(\lambda) + b_{bSM}(\lambda)c_{SM} \] 1.10
Where \( b_{bw}(\lambda) \) is the scattering coefficient of pure water which accounts for refraction due to path transmittance between medium transitions (air - water), it is assumed that backscattering is always 50% of total backscattering by water and therefore expressed as \( 0.5b_{bw} \) (Mobley, 1994). The \( b_{bph}(\lambda) \) represents the backscattering coefficient of phytoplankton and \( b_{bSM}(\lambda) \) is the backscattering coefficient of suspended matter. The degree of scattering of dissolved organic matter is negligible and assumed to be null, \( b_{bdom}(\lambda) \approx 0 \). In equation 1.10, \( b^{*}_{bph}(\lambda) \) is chlorophyll-specific spectral backscattering coefficient of phytoplankton and \( b^{*}_{bSM}(\lambda) \) is the specific spectral absorption coefficient of suspended matter.

Absorption due to pure water is small across the visible range but increases exponentially in the near-infrared (Bricaud et al., 1981; Kutser et al., 2004). In our study the absorption coefficients of pure water are taken from Buiteveld et al. (1994) (Fig 1.3). In Lake Erie, in-water constituents reflect radiance back to the sensor and the in-water constituents influence the spectral characteristics of reflected radiance from the water differently. In the VIR, various biogeochemical components theoretically cause wavelength-dependant variations upon interaction with the incoming irradiance. These discrepancies produce distinct spectra, allowing identification of specific in-water constituents from the position and character of absorption features. Understanding these optical characteristics is fundamental for assessing the capability of remote sensing devices to detect in-water constituents.
Independent studies have shown that the magnitude of reflectance from water bodies varies as a function of the IOPs of the media and associated constituents. Measurement of IOPs can theoretically separate the spectral fingerprints of water, particulates and dissolved materials. However, this approach is not feasible in practice because it requires in-situ measurements at high spatial and temporal resolution. Moreover, measurement of attenuation and absorption coefficients requires expensive instrumentation. However, reflectance measurements recorded at satellite sensors can give AOPs which may be used to infer IOPs using inversion techniques.

Figure 1.3. Absorption (log (1/R)) spectra of pure water increases with wavelength showing an exponentially increase in the near-infrared region (after Buiteveld, et al., 1994).
1.5 Atmospheric correction

In turbid environments such as the Western Basin of Lake Erie where multiple constituents are present, the recorded spectra represent a convolution of the spectral responses from multiple constituents. The challenge is discriminating and quantifying the concentrations of the components independently. Further complication arises from spectral contributions from atmospheric constituents along the path between the water surface and the satellite sensor. These atmospheric materials include aerosols, dusts, vapor and gaseous molecules which can also change the spectral characteristics of the signal, influencing the signal to noise ratio.

Satellite-based sensors record irradiance reflectance which is the sum of the AOP of surface and atmospheric constituents. Due to high absorption by the water, signals from the atmospheric component can be an order of magnitude larger than those from the water (Siegel et al, 2000). Thus it is crucial to remove or account for the atmospheric interference on the reflectance data before spectral analysis can be performed for the retrieval of the various in-water constituents. Effects due to gaseous absorption, molecular scattering and their temporal and spatial variations can be corrected using look-up tables with computed values for different locations and illumination condition and the Rayleigh theory of scattering (Wang and Gordon, 1994; Gordon et al., 1983). However, accounting for effects due to aerosol particles is challenging.
Bio-optical algorithms are developed by relating remote sensing reflectance with biophysical properties such as AOPs of the various in-water components. In order to apply bio-optical algorithms, records of up-welling radiance from aquatic bodies require absolute atmospheric correction (Gordon et al., 1998). Correction of path radiance (i.e., the radiance received at the satellite from sources other than the water body of interest) requires an atmosphere-surface water model that varies as a function of aerosol optical thickness (Takashima et al., 2003).

In ocean remote sensing, it is assumed that the water-leaving radiance is zero in the infrared region. Any signal recorded in the near-infrared (NIR) after being corrected for the gaseous absorption and molecular scattering is considered to be due to aerosol particles (Gordon and Wang, 1994; Shutler et al., 2007). Reflectance due to aerosols is extrapolated from the near-infrared bands to shorter wavelengths and used to subtract for the signals across the spectrum, this method is referred to as “dark object subtraction” or the “black pixel correction”. This is valid for open ocean environment, where optical properties are only governed by water itself and phytoplankton which do not cause scattering and therefore no water leaving radiance beyond the visible range. However, this approach is not valid in highly turbid waters, e.g. Lake Erie, where optical properties are governed by multiple components. The presence of in-water constituents such as suspended sediments results in significant reflectance in the infrared region (Ruddick et al., 2000; Stumpf et al., 2003; Han and Rundquist, 1996). Application of dark object
subtraction in turbid water causes negative reflectance values at the shorter wavelengths (Fig 1.4).

Several algorithms have been developed to perform atmospheric correction on spectral radiance from turbid waters, which have non-zero reflectance in the infrared region. Several bio-optical reflectance models that account for reflectance due to suspended materials in the NIR have been developed, including the 2-band NIR black pixel algorithm of Gordon and Wang, (1994), the iterative NIR correction model developed by Stumpf et al. (2003), and the Management Unit of the North Sea Mathematical Models (MUUM) developed by Ruddick et al. (2000). Other atmospheric correction models developed recently include the Case 2 Regional (C2R) processing model which is based on neural-network procedure specifically developed for turbid waters (Doerffer and Schiller, 2007). In the WBLE the 2-band NIR and its iterative version clearly enabled extraction of important optical features such as the peaks near 560 and 710nm and the troughs at 620nm. However, these atmospheric algorithms resulted in negative reflectance values at short wavelengths (Fig 1.4). For this study the MUUM and the C2R are used to correct for the path radiance effects on the MODIS and MERIS data, respectively.
Figure 1. 4. Average reflectance spectra of the Western Basin of Lake Erie from the MERIS bands after path radiance correction.

1.6  Optical properties of Case 2 waters

Some of the challenges encountered when dealing with Case 2 waters include: (a) phytoplankton communities are complex and their composition show high seasonal variations, as these communities are sensitive to the availability of nutrients and light, and to other factors, such as water temperature, and (b) large variations in the spatial, temporal and particle size distribution of suspended sediments that are added to the system by erosion and fluvial transport. These challenges preclude the direct application of ocean-derived bio-optical algorithms to satellite observations, with results suggesting
that a fundamentally different approach is needed in the Western Basin (Witter et al., 2009).

In productive and turbid waters, such as the Western Basin of Lake Erie, dissolved and suspended matter attenuates incoming irradiance by absorption and scattering thus changing the spectral radiance measured by the satellite sensor. When the backscattering increases, \( R(\lambda) \) values increase more or less uniformly throughout the spectrum (Schalles et al., 1998). Conversely, the rise in absorption diminishes \( R(\lambda) \), especially in the spectral bands corresponding to the specific absorption of the various substances present. Thus the shape of the spectral reflectance curve may be diversely modified (Morel and Prieur, 1977).

Three main components affect the optical properties of Case 2 waters: phytoplankton, suspended matters and dissolved organic matter (gelbstoff). In the presence of phytoplankton, the \( R(\lambda) \) values in the blue region decrease progressively and a minimum is formed around 440nm signified as the Soret band (Fig 1.5) (Gordon and Morel, 198; Schalles et al., 1998) which is mainly due to the maximum absorption of chlorophyll \( a \). This feature is often masked by dissolved organic matter. Maximum reflectance occurs near 560nm also known as the green peak (GP), which is the wavelength where simultaneously the absorption due to pigments is at its minimum and scattering by algal cells is maximum (Fig 1.5; Gitelson, 1992). Increased carotenoid concentration causes irregularities near 500nm and depresses the left limb of the peak curve (560nm) shifting the green peak towards longer wavelength (Fig 1.5).
Figure 1. Spectra of the Western Basin of Lake Erie at 18 stations corresponding to the sampling points for the study. The Reflectance data are taken on 7-08-2010 using the Lab based UV-VIS-NIR Spectrometer. Effect of water absorption is removed by drying the sample over 24 hour period. Phytoplankton (chlorophyll $a$ (Chl$_a$) is used as a proxy) has maximum absorption near the 440nm and 680nm. Minimum absorption of Chl$_a$ combined with backscattering from algal cells is signaled near 550nm (Green-peak, GP). Accessory pigment, Phycocyanin (PC), can be detected from it maximum absorption near 620nm. Carotenoids (Caro) depress the left limb of the green peak (GP). The signal after 720nm is mainly due to backscattering from the suspended materials.
The presence of phycocyanin and other accessory pigments, phycobilins, cause an absorption trough near 620nm (Rowan, 1989; Schalles et al., 1998; Gitelson et al., 1999). High concentrations of phycocyanin affect the right limb of the GP causing it to shift towards shorter wavelength (~540nm) (Rowan, 1989). The absorption by chlorophyll \textit{a} also creates a $R(\lambda)$ minimum between 660 and 675nm. This minimum is not particularly prominent because the increase in absorption due to the presence of chlorophyll \textit{a} remains weak compared with the absorption due to the water itself. The minor reflectance peak at approximately 650nm is attributed to backscattering from dissolved organic matter (Rowan, 1989). This feature is also known to be affected by phycocyanin concentration because it is the location of the phycocyanin fluorescence emission maxima (Ahn et al., 1992). The trough feature at approximately 675nm represents absorption maxima of chlorophyll \textit{a} in vivo (Rundquist et al., 1996; Han and Rundquist, 1997) (Fig 1.6). Reflectance further into the NIR depends mainly on the presence of non-organic suspended matter (Gitelson et al., 1993; Dekker, 1993; Yacobi et al., 1995).

The reflectance peak in the NIR portion of the spectrum is the result of scattering from algal cells, and particulate matter. The location and height of this peak is a function of chlorophyll \textit{a} concentration. The position of the peak shifts toward longer wavelengths and the height increases with increasing chlorophyll \textit{a} concentration. (Gitelson, 1992; Dekker et al., 1991; Han and Rundquist, 1997; Mittenzwey et al., 1992; Rundquist et al., 1996; Schalles et al., 1998). Reflectance features at wavelengths longer than 750nm are attributed to organic and non-organic suspended matter concentrations (Han et al., 1994).
Therefore this region may be considered optimal to estimate suspended matters in the water column (Fig 1.6).

**Figure 1.6.** Example spectra of the inherent optical properties of photoactive components in a Case 2 waters. (a) is spectra of absorption coefficient and (b) spectra of backscattering coefficients. The $a^{*}\text{chl}$ is the absorption coefficient of chlorophyll $a$; $a^{*}\text{tsm}$ is absorption coefficient of total suspended matter; $a^{*}\text{cdom}$ is absorption coefficient of colored dissolved organic matter. The $a_{w}$ is the absorption coefficient of pure water. $b^{*}$ represents the backscattering coefficients of the various constituents (chlorophyll $a$, TSM). Phytoplankton in the blue region is fundamentally masked by the CDOM absorption. Effects of suspended materials can be independently extracted using values of higher backscattering after 700nm, where backscattering effects due to other components becomes negligible (1.6b) (after Dekker et al., 1997)
1.7 Research Objectives

The aim of this work is to develop robust algorithms that can estimate CPAs from satellite data in an aqueous environment that possesses a wide variety of optical properties. The Western Basin of Lake Erie has high concentrations of suspended matter due to high loading of sediment from major rivers which respond to weather conditions. For example, particulate loadings can exceed 1600 tons/day during the rainy season at Sandusky Bay (USGS, 2007). The Western Basin also has several distinct phytoplankton communities including chlorophyta, cyanobacteria and diatoms (Makarewicz, 1993). These phytoplankton communities vary spatially and temporally within the Lake system. These multiple in-water constituents define the color of the WBLE and pose a great challenge to remotely retrieve the concentrations of the various CPAs. Lake Erie can thus serve as a physically compact, natural laboratory for improving satellite-based retrievals of CPAs in optically complex coastal environments. Results can then potentially be translated to other settings, improving the capability of resource managers to assess and predict impacts on the nation’s environmental and economic resources in coastal and lake environments.

The specific objectives of the proposed study are to:

- Evaluate the concentration and composition of CPAs in WBLE based on field cruise data analyzed using HACH Hydrolab, Turner Cyclops CDOM sensor, Lab based Pro FR UV/VIS/NIR Spectrophotometer, and the Malvern Mastersizer 2000.
Parameterize spectral windows that can be used to distinguish among selected CPAs by analyzing the optical properties of field-collected samples.

- Develop algorithms for retrieval of CPAs from remotely sensed radiance values using multivariate statistics, wavelet and neural network approaches.

- Assess performance of MODIS and MERIS derived data in predicting CPA proxies.

### 1.8 Dissertation layout

Several approaches will be used to improve satellite-based retrievals of CPAs in optically complex environments. The techniques used to develop algorithms include multivariate statistics, wavelet analysis, spectral indices, and neural networks. In chapter 2 traditional techniques are employed where selected band ratios are used. In this section, several existing empirical and semi-empirical algorithms are applied to current generation ocean satellites (MERIS, MODIS) to estimate the concentrations of phytoplankton in the WBLE, using chlorophyll $a$ as a proxy. A band ratio algorithm developed by Kutser et al., 2005 is used to retrieve the concentration of colored dissolved organic matter.

In chapter 3, Principal Component Analysis (PCA) is applied to hyperspectral data to estimate concentrations of the main biogeochemical constituents (chlorophyll $a$ and total suspended matter). PCA seeks to establish combinations of variables capable of
describing the principal tendencies and patterns observed within a data matrix (Shlens, 2005). Computationally, PCA relies upon the eigenvector and eigenvalue decomposition of the variance-covariance or correlation matrix. Results of work done by Ortiz et al. (2004, 2009), demonstrates the efficiency of PCA as a tool for isolating patterns of features, with distinct spectral signature in remote sensing data. This chapter discusses the results from WBLE remotely sensed data.

The 4th chapter discusses the application of discrete wavelet transformation techniques to hyperspectral data for capturing various spectral characteristics that are used to improve the signal to noise ratio and develop bio-optical algorithm based on detectable optical features of CPAs. This technique involves signal decomposition and frequency analysis for retrieving the concentrations of the diverse photoactive components in the WBLE.

The 5th chapter discusses the application of Neural Network (NN) to estimate the concentrations of water quality parameters. A single layered artificial network is designed and trained using in-situ observations and satellite data. A regional neural network algorithm is developed to retrieve concentrations of selected CPAs. In this case, the training data consists of collocated satellite-measured radiometric values and in-situ measurements. Based on the parametric learning the remote sensing data are fed into the network, and the model calculates the output CPA concentrations. Chapter 6 is a summary chapter that presents the primary findings, best model selection process, and future prospects.
1.9 Data used in this study

1.9.1 Field data

Field data is crucial to assess variability of in-water constituents and perform calibration and validation of the satellite algorithms. Physical measurements (Secchi depth, pH, dissolved oxygen, electrical conductivity, total dissolved solids, turbidity, temperature, chlorophyll a) are collected using a Hydrolab multiparameter sonde. The sonde has multiple sensor probes. The fluorescence probes include the chlorophyll a sensor and the phycocyanin sensor. The chlorophyll a sensor operates by exciting photons at 460nm and emitting at 680nm. The phycocyanin sensor detects the pigment by exciting photons at 570nm and recording the emission at 650nm. In-situ Colored Dissolved Organic Matter (CDOM) fluorescence measurement was taken during cruises in 2010 using the Turner Designs Cyclops-7 CDOM sensor. The CDOM sensor is equipped with excitation filter at 370nm and emission filter at 420nm. In the lab, total suspended particulate concentration was measured gravimetrically and using Malvern Mastersizer 2000. The reflectance properties of Lake Erie constituents were measured in the lab from dry filters using a Lab-based Pro FR Spectrophotometer.
1.9.2 Remote sensing data

a) Lab-based Spectrophotometer

The lab-based spectrophotometer is an active FR (Full Resolution, 2nm) hyperspectral device that uses the visible – near infrared (VIR: 350- 2500nm) range to measure optical energy that is reflected by, absorbed into, or transmitted through a sample. For the analysis I averaged 800 spectra to remove noise related frequency and increase the signal to noise ratio.

b) Medium Resolution Imaging Spectrometer

The MEdium Resolution Imaging Spectrometer (MERIS) is a passive sensor that operates in the solar spectral range. The primary mission of MERIS is to measure sea color in the oceans and in coastal areas. The color signals across the spectrum may be converted into a measurement of chlorophyll pigment concentration, suspended sediment concentration and the atmospheric aerosol loads over water. It transmits fifteen spectral bands to the ground receiving station. MERIS is designed so that it can acquire data over the Earth whenever illumination conditions are suitable. The instrument's 67° field of view around nadir covers a swath width of 1150km and MERIS gets global coverage in two to three days which is essential in water quality studies. MERIS is able to deliver reduced spatial resolution of 1.2km x 1.04km and full spatial resolution of 290m by 260m
at satellite points. The observation is performed simultaneously in 15 spectral bands, ranging from the visible to the near infrared (390nm to 1040nm) (Table 1.1).

<table>
<thead>
<tr>
<th>No.</th>
<th>Band centre (nm)</th>
<th>Band width (nm)</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>412.5</td>
<td>10</td>
<td>Yellow substance and detrital pigments</td>
</tr>
<tr>
<td>2</td>
<td>442.5</td>
<td>10</td>
<td>Chlorophyll absorption maximum</td>
</tr>
<tr>
<td>3</td>
<td>490</td>
<td>10</td>
<td>Chlorophyll and other pigments</td>
</tr>
<tr>
<td>4</td>
<td>510</td>
<td>10</td>
<td>Suspended sediment, red tides</td>
</tr>
<tr>
<td>5</td>
<td>560</td>
<td>10</td>
<td>Chlorophyll absorption minimum</td>
</tr>
<tr>
<td>6</td>
<td>620</td>
<td>10</td>
<td>Suspended sediment</td>
</tr>
<tr>
<td>7</td>
<td>665</td>
<td>10</td>
<td>Chlorophyll absorption &amp; fluorescence reference</td>
</tr>
<tr>
<td>8</td>
<td>681.25</td>
<td>7.5</td>
<td>Chlorophyll fluorescence peak</td>
</tr>
<tr>
<td>9</td>
<td>708.75</td>
<td>10</td>
<td>Fluorescence reference, atmosphere corrections</td>
</tr>
<tr>
<td>10</td>
<td>753.75</td>
<td>7.5</td>
<td>Vegetation, cloud, O₂ absorption band reference</td>
</tr>
<tr>
<td>11</td>
<td>760.625</td>
<td>3.75</td>
<td>O₂ R- branch absorption band</td>
</tr>
<tr>
<td>12</td>
<td>778.75</td>
<td>15</td>
<td>Atmosphere corrections</td>
</tr>
<tr>
<td>13</td>
<td>865</td>
<td>20</td>
<td>Atmosphere corrections</td>
</tr>
<tr>
<td>14</td>
<td>885</td>
<td>10</td>
<td>Vegetation, water vapour reference</td>
</tr>
<tr>
<td>15</td>
<td>900</td>
<td>10</td>
<td>Water vapour</td>
</tr>
</tbody>
</table>

Table 1.1. MERIS instrument specification on board ENVISAT (Environmental Satellite), launched March 1st 2002, spatial resolution (FR & RR)- with ~ 300 and 1000m, respectively. Collects reflectance at 15 spectral bands with variable band width, spectral coverage ranges between 412 – 1000nm, (ESA, 2006)
c) **Moderate Resolution Imaging Spectroradiometer**

Moderate Resolution Imaging Spectroradiometer (MODIS) is an instrument aboard the Terra and Aqua satellites. Both satellites are sun synchronized, Terra passes north to south while Aqua passes south to north. Terra acquires data from the Western Basin of Lake Erie usually around 16:00hr and Aqua footprints the Western Basin of Lake Erie later during the day (~18:00hr). Together Terra MODIS and Aqua MODIS view the entire surface of the Earth every 1 to 2 days. The instrument records data at radiometric sensitivity (12 bits) and splits the reflectance measurements into 36 spectral bands ranging in wavelength from 0.4µm to 14.4µm (Table 1.2). Two bands are imaged at a nominal resolution of 250m at nadir, with five bands at 500m, and the remaining 29 bands at 1km. It has 80º field of view at the Earth Observation Satellite orbit of 705km and covers a 2,330km swath (NASA).
<table>
<thead>
<tr>
<th>Band</th>
<th>Wavelength (nm)</th>
<th>Resolution (m)</th>
<th>Primary Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>620-670</td>
<td>250m</td>
<td>Land/Cloud/Aerosols</td>
</tr>
<tr>
<td>2</td>
<td>841-876</td>
<td>250m</td>
<td>Boundaries</td>
</tr>
<tr>
<td>3</td>
<td>459-479</td>
<td>500m</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>545-565</td>
<td>500m</td>
<td>Land/Cloud/Aerosols</td>
</tr>
<tr>
<td>5</td>
<td>1230-1250</td>
<td>500m</td>
<td>Properties</td>
</tr>
<tr>
<td>6</td>
<td>1628-1652</td>
<td>500m</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2105-2155</td>
<td>500m</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>405-420</td>
<td>1000m</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>438-448</td>
<td>1000m</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>483-493</td>
<td>1000m</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>526-536</td>
<td>1000m</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>546-556</td>
<td>1000m</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>662-672</td>
<td>1000m</td>
<td>Ocean Color/Phytoplankton/Biogeochemistry</td>
</tr>
<tr>
<td>14</td>
<td>673-683</td>
<td>1000m</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>743-753</td>
<td>1000m</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>862-877</td>
<td>1000m</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>3.929-3.989</td>
<td>1000m</td>
<td></td>
</tr>
</tbody>
</table>

Table 1.2. MODIS instrument specification on board Terra and Aqua satellites that fly over the entire Earth every one to two days. It has relatively high spectral windows, 36 spectral bands between 405 and 14,385 nm, and it acquires data at three spatial resolutions -- 250m, 500m, and 1,000m. The table shows only the required bands for this study.
CHAPTER 2

Application of satellite algorithms to estimate chlorophyll \( a \) and CDOM in the Western Basin of Lake Erie.
Abstract

The Western Basin of Lake Erie (WBLE) is heavily influenced by the surrounding terrestrial ecosystem via rivers such as the Sandusky River and the Maumee River. As a result, the optical properties of the WBLE are influenced by multiple color producing agents (phytoplankton, colored dissolved organic matter (CDOM), organic detritus, and terrigenous inorganic particles). Many of these vary independently and have non-linear interactions. Therefore, the applications of satellite remote sensing for deriving biophysical variables present a great challenge. Chlorophyll $a$, which is the primary light harvesting pigment in all phytoplankton, is used as an index for the estimation of phytoplankton density in open water bodies including the WBLE. Bio-optical algorithms developed by Dall’Olmo and Gitelson (2005), Moses et al. (2009), Gons (1999), Gower et al. (1999) and Simis et al. (2005) are applied to the reflectance data from multi spectral satellite sensors (MERIS and MODIS) to predict chlorophyll $a$ concentrations. Colored Dissolved Organic Matter (CDOM) in the water column is estimated using the algorithm developed by Kutser et al. (2005). Efficiency of the sensors and the algorithms performance were tested through a least squares regression and residual analysis. Results show that for prediction models of chlorophyll $a$ concentrations, the Dall’Olmo and Gitelson, (2005), three-band empirical algorithm yielded coefficients of determination as high as 0.57 with RMSE=1.19µg/l for an aggregated dataset (n=71, $P<0.05$). A two-band algorithm by Moses et al. (2009) gave a slightly better chlorophyll $a$ prediction model, with $R^2 =0.61$, RMSE=0.89µg/l. Estimation of chlorophyll $a$ concentrations using the
Simis et al. (2005) semi-empirical algorithm yielded a coefficient of determination of 0.65, RMSE 0.85µg/l, (n=71, P<0.05). A semi-empirical model developed by Gons (1999) explained about 61% of the chlorophyll a variability observed in the WBLE. The magnitude of chlorophyll a fluorescence near 680nm was used to determine concentrations of chlorophyll a. A regression analysis between Gower et al. (1999) based fluorescence line height algorithm and in-situ chlorophyll a level in the WBE yielded an R² value of 0.54. Results indicate that, among this suite of existing bio-optical, empirical and semi-empirical algorithms, the Simis et al. (2005) algorithm provided better measures of algal biomass in the turbid waters of the WBLE. Results from regression analysis between in-situ CDOM absorption measurements and the Kutser et al. (2005) empirical estimate produced an R² value of 0.67. The absorption of CDOM in the WBLE ranges between 0.115 to 1.38m⁻¹.

**Keywords:** chlorophyll a, CDOM, MERIS, Spectral indices.

### 2.1 Introduction

Monitoring large inland fresh water bodies such as the Great Lakes through conventional in-situ methods can be time consuming and expensive. Currently, Lake Erie
is monitored from observation made twice a year through the Environmental Protection Agency’s Great Lakes National Program Office (GLNPO) at twenty stations throughout the lake (EPA, 2006b). Additional data come from other research initiatives, such as the Great Lakes Environmental Research Laboratory’s (GLERL) 2005 International Field Year on Lake Erie and the 2004 Lake Erie Collaborative Comprehensive Survey (ECCS), as well as some additional government and academic researchers (USEPA, 2006). These observations are restricted in space and time and provide incomplete information regarding the spatial and temporal dynamics of the biological and geochemical process in Lake Erie. Observing the complex dynamics of the various in-water constituents at high spatial and temporal resolution via satellite remote sensing instrumentation allows effective water quality monitoring particularly in surveillance programs of harmful algal blooms. This approach also provides researchers high resolution data for ecological and climate studies.

Effective water quality monitoring is critical for addressing the question of how various natural and anthropogenic factors affect the health of these environments. A remote sensing approach overcomes some of the shortcoming of in-situ observations by providing a synoptic view of the aquatic bodies at high spatial and temporal resolution.

Morel and Prieur (1977) have classified marine waters as Case 1 and Case 2 based on optical properties. In Case 1 waters, such as the epi-pelagic ocean environments, phytoplankton pigments govern the optical properties of the water. Coastal and most inland waters are classified as Case 2 waters, because, in addition to phytoplankton,
constituents such as colored dissolved organic matter and suspended sediments occur in abundance, and therefore the optical properties of such waters is governed by multiple components. Case 2 waters are also generally turbid and highly productive. Remote sensing has been successfully applied in Case 1 waters, where optical properties are a function of single in-water component (phytoplankton) and its co-varying by products. Estimation of in-water constituents remotely has been less successful in Case 2 waters because of the complex interaction among diverse CPAs (e.g., phytoplankton, suspended sediments and colored dissolved organic matter).

Use of field spectroscopy has substantially increased over the past decade. Many researchers have attempted to develop empirical and semi-empirical algorithms to estimate concentrations of various in-water constituents based on field spectrophotometer measurements (Dekker et al., 1997; Schalles, 2006; Gitelson et al., 2009, 1992; Gons et al., 2002; Dall’Olmo and Gitelson, 2005; Moses et al., 2009; Mathews et al., 2010; Mayo et al., 1995). Empirical methods are developed by relating spectral reflectance values and in-situ measurements of selected OACs taken simultaneously. A particular challenge of this approach, which involves costly field observations, is differentiating the optical properties of the various in-water constituents. The semi-empirical method employs an inversion technique based on physical properties of the inherent optics. It eliminates the need for field observations and the optical effects of the various in-water properties can be accounted for independently.
Nutrient controls in Lake Erie led to significant reduction in the biomass density in the 1980’s, however large blooms were recorded in the Western Basin of Lake Erie in 1995 (Budd et al., 2002) and 2009 (USEPA, 2009). The distribution of phytoplankton in space and time has major implications for water quality and ecosystem function. Traditional methods of water quality assessment makes it difficult to understand the dynamics of color producing agents (CPAs) such as chlorophyll *a*, total suspended materials (TSM) and colored dissolved organic matter (CDOM).

Chlorophyll *a* is the major light harvesting pigment in phytoplankton. The presence of chlorophyll *a* changes the optical properties of water by absorbing and scattering incident light (Bukata et al., 1995; Rowan, 1989). As a universal phytoplankton pigment, chlorophyll *a* concentrations have been considered as important index in water quality monitoring and in surveillance programs for harmful algal blooms (Glasgow et al., 2004; Kahru and Mitchell, 1999) as well as for monitoring primary productivity.

In Case 1 waters, where optical properties are influenced by single component (phytoplankton), a distinctive broad green peak in the reflectance spectrum occurs near 560nm and absorption trough occurs near 490nm (blue region) (Schalles, 2006). A second peak occurs near 680nm, and is associated with the fluorescence. Studies have indicated that the fluorescence peak may shift to higher wavelength and/or its magnitude may increase proportionally with increasing chlorophyll *a* concentration (Gower et al., 1999). In these waters, spectral algorithms that use reflectance in the blue and green
regions of the spectrum (blue–green models) have been shown to yield accurate estimates of chlorophyll \( a \) concentration (Gordon and Morel, 1983).

In inland waters, such as the Western Basin of Lake Erie, the optical properties of the water depend on multiple components with overlapping and correlated inherent optical properties, complicating retrieval of chlorophyll \( a \) concentration via satellite sensors. Moreover, inland waters are characterized by biochemical complexity that results in temporally and spatially variable optical properties. These waters include multiple constituents including phytoplankton with chlorophyll \( a \) and other accessory pigments, TSM and CDOM (Gordon and Morel, 1983).

CDOM is a major structuring component of lake system constituting a major source of carbon for heterotrophic bacteria within lakes (Kutser et al., 2005). The WBLE usually has significant amounts of CDOM that are derived from the terrestrial environment (Binding et al., 2008). Light absorption by CDOM affects the availability of light for primary producers (Jones, 1998). Moreover, CDOM efficiently absorbs short-wavelength solar radiation, which results in photochemical reactions. CDOM absorption also protects aquatic biota from ultraviolet-blue (UV-B) radiation induced damage—an effect that may be especially important in high-latitude lakes (Pienitz and Vincent, 2000). Given the role of CDOM in the structure and function of lake ecosystems, a method to estimate the amounts in lake waters would be highly desirable. Absorption due to the CDOM results in low reflectance in the visible range and therefore its detection requires satellite with high radiometric sensitivity.
Accessory pigments and CDOM significantly affect the Soret bands accounting for most of the total absorption. TSM plays role in scattering light across the visible and NIR region (Schalles, 2006). These multiple constituents tend to mask fundamental absorption and scattering relationship that represent characteristic spectral features of chlorophyll \( a \).

Current generation satellite sensors such as the MEdium Resolution Imaging Spectrometer (MERIS) and the MOderate Resolution Imaging Spectrometer (MODIS) are equipped with radiometric sensitivities optimized for measurements in turbid waters. These waters normally have low reflectance due to the high absorption by water itself and its associated constituents such as phytoplankton and CDOM. These satellites have high spectral and temporal resolution (10nm and 1-3 days, respectively) and therefore, allow better characterization of subtle variations in spectral signals associated with the various water quality parameters.

Application of satellites to estimate concentrations of CPAs requires the development of empirical or semi-empirical algorithms which correlate reflectance measured remotely with in-situ data collected for ground truthing (e.g., Doerffer and Fischer, 1994, IOCCG, 2000, McClain, 2009). Spectral variations in the backscattered flux of Case 1 waters are primarily related to concentration of chlorophyll \( a \), which varies as a function of the phytoplankton population. In many marine environments satisfactory relationships have been described for chlorophyll \( a \) concentrations ranging between 0.02 to 78mg/m\(^3\) (O'Reilly et al., 2000; Gordon et al., 1988; Gons, 1999). Many of these
relationships make use of the blue-green spectral region because optical properties are determined mainly by a single component, phytoplankton pigment (Babin et al., 2003). In many coastal waters, the ratio of upward irradiance and downward irradiance is affected by multiple optically active components. As shown by Witter et al. (2009), optical teledetection of chlorophyll $a$ in the Case 2 waters of the WBLE using Case 1-derived algorithms is problematic. The challenge is attributed to optical heterogeneity of Case 2 waters as compared to the oceans due to interference from higher concentrations of total suspended matters and CDOM. Therefore the differentiation of the overlapping and correlated in-water constituents requires the use of both the visible and the near-infrared portions of the spectrum.

The objective of this study is to statistically assess the potential of existing visible/NIR based empirical and semi-empirical methods (Dall’Olmo and Gitelson, 2005; Moses et al., 2009; Gons, 1999; Gower et al., 1999; Simis et al., 2005; Kutser et al., 2005) for estimating chlorophyll $a$ in the WBLE from satellite data. Important spectral features of chlorophyll $a$ include: absorption near 440nm, minimum absorption coupled with cellular scattering near 560nm and the red absorption near 665nm, and the chlorophyll $a$ fluorescence peak near 680nm (Fig 1.5). Absorption due to phytoplankton beyond 720nm is minimal and is not detected (Schalles, 2006), therefore, remote sensing algorithms that develop estimates of phytoplankton are confined to the visible/NIR spectral range. The study also assesses the efficiency of an existing CDOM prediction algorithm, developed by Kutser et al. (2005), in estimating CDOM concentration in the
WBLE. The spectral signature of CDOM is described as an exponentially decaying curve with increasing wavelength (Dekker et al., 1993; Kutser et al., 2005; Binding et al., 2008).
2.2 Data acquisition and methodology

2.2.1 Field measurements

Assessment of models that predict concentrations of biogeochemical proxies requires in-situ data. Field expeditions were carried out during the early and late summers of 2009 and 2010. This involved measurements of water quality properties (e.g., Secchi depth, biogeochemical proxies) at 18 GPS-referenced locations that extended between Sandusky Bay and Middle Bass Island in the WBLE. The field campaign included collection of water samples using screw capped Nalgene bottles for laboratory analysis. A submersible HACH Hydrolab DS5X equipped with multi-parameter sonde was used to measure water temperature, pH, conductivity, salinity, dissolved oxygen, chlorophyll \( a \), and phycocyanin. Absorbance of CDOM, which is indicative of its concentration, was measured in-situ using UV CYCLOPS-7 submersible fluorometer.

All of the sensor probes were calibrated using standards. The chlorophyll \( a \) sensor was calibrated by the manufacturer, HACH. Quinine sulfate which has similar excitation and emission as CDOM compounds was diluted to low concentrations ranging from 0-30ppb in weak acid (0.05 Molar H2SO4) to calibrate the Cyclops-7 CDOM fluorometer.
2.2.2 Satellite imagery

The capability of MERIS and MODIS to retrieve concentrations of CPAs in the WBLE is assessed. The various color producing agents in this aquatic environment are mainly sensitive to the visible and NIR part of the electromagnetic spectrum. Therefore, in this study only visible/NIR bands are used from both satellite sensors. A total of five MERIS FR Level-1b images were acquired for summer of 2009 and 2010 from the European Space Agency (ESA). These dates were selected on the basis of the coincident field campaign for in-situ data and the cloud coverage conditions over the WBLE. The Basic Toolbox for ENVISAT (A) ATSR and MERIS (BEAM)-4.6 visualization, analyzing and processing software developed by Brockmann-consult, Germany, was used for converting the images from Level-1b to Level-2b and to perform radiometric corrections. The Full Resolution (FR)-MERIS sensor is equipped with 15 bands representing the visible/NIR spectral range with 300m and 10nm spatial and spectral resolutions, respectively.

MODIS images of similar dates were downloaded from the Ocean Color facility of the NASA/Goddard Space Flight Center. The SeaWiFS Data Analysis System (SeaDAS) was used to convert images from Level-1 to Level-2. The instrument records data using 36 spectral bands ranging in wavelength from 0.4 to 14.4µm. Two bands are imaged at a nominal resolution of 250m at nadir, five bands at 500m, and the remaining 29 bands at 1km. MODIS has 55° field of view at the Earth Observation Satellite orbit of
705km and covers a 2,330km swath (NASA). In this study, only bands with wavelengths between 0.4µm and 0.9µm are used.

Digital image processing including atmospheric correction was carried out on the images to account for multiple scattering and absorption due to atmospheric constituents. Removal of atmospheric effects is crucial to reduce path radiance effects. In turbid waters such as the WBLE, the assumption of negligible reflectance at NIR wavelength becomes invalid due to increased scattering from non-algal components. Therefore, the conventional method for image correction, the black pixel approach does not hold (Schalles, 2006; Morel and Prieur, 1977). A recently developed atmospheric correction model for MERIS data is the Case 2 regional (C2R) atmospheric correction. This model is based on a neural network procedure specifically developed for turbid waters (Doerffer and Schiller, 2007). It involves a two-step procedure—(a) a forward neural network for the retrieval of water-leaving radiances and, subsequently, remote sensing reflectance from the radiances, and (b) a backward neural network for the retrieval of the inherent optical properties of water and, subsequently, the concentrations of constituents by inverting the remote sensing reflectance. The neural networks were trained based on radiances simulated by radiative transfer solutions and built to parameterize the relationships between the top-of-atmosphere radiances and the water-leaving radiances (for the forward model) and between the remote sensing reflectance and the inherent optical properties (for the backward model). The recorded radiances at 12 wavebands (in visible and NIR wavelengths) are used in the neural network.
For the MODIS data, several atmospheric procedures, including the 2-band and NIR pixel approach, the fixed model and iterative NIR correction, and the Management Unit of the North Sea Mathematical Model (MUMM) were applied to check for performance. These procedures resulted in negative reflectance values in the shortest wavelengths, except for MUMM. In the MUUM model, the usual assumption of zero water-leaving radiance in the NIR bands is replaced by the assumption of spatial homogeneity of the 748/869 reflectance ratio for aerosol and water reflectance within an image (Ruddick et al., 2000).

Atmospheric correction procedures have significant effects in the Western Basin where atmospheric constituents may contain significant continental components (Shutler et al., 2007). In addition, absorption by gases and water vapor must also be accounted for in the correction procedure. Two of the atmospheric corrections that have been employed in this study are the MUUM for MODIS and the Case 2 Regional algorithm for MERIS. Both of these include corrections for continental aerosol, gas absorption and water vapor.

2.2.3 Spectral analysis

Remote sensing estimation of chlorophyll $a$ concentrations make use of particular spectral patterns that result from the interaction between the incident radiation and the attenuation (absorption and scattering) induced by the pigment. The interaction of chlorophyll $a$ and water absorption provides important insights for the interpretation of maxima and minima in the spectral reflectance curves. In Case 2 waters, absorption due
to CDOM is inversely proportional to wavelength (Kutser et al., 2005). Backscatter from TSM also controls the reflectance across the VIR spectrum. However, the reflectance in the NIR region is mainly controlled by scattering effects due to SM and absorption of water itself (Bukata et al., 1995). Isolating the spectral windows that correlate with the apparent optical property of chlorophyll $a$, independent of other water constituents requires a robust spectral analysis technique. Bale et al. (1994) showed that increasing the concentrations of phytoplankton biomass or CDOM have no effect on the spectral shape of the reflectance near the NIR region, thus the optical signatures of materials, such as suspended sediments may be traced in these longer wavelength regions.

In the WBLE, the optical properties of the multiple in-water constituents overlap and therefore the differentiation of the signatures requires the selection of optimal bands. Various empirical and semi-analytical algorithms have been developed for estimating chlorophyll $a$ in coastal waters (Gitelson et al., 2000; Schalles, 2006; Dall’Olmo and Gitelson, 2005; Simis et al., 2005; Moses et al., 2009; Gons, 1999; Gower et al, 1999). All of the algorithms are based on sensitivity analysis of the chlorophyll $a$ pigment at different spectral regions. The challenge in algorithm development in these waters comes with the overlapping magnitude of the spectral coefficients from the various CPAs and their interaction.
2.3 Chlorophyll a prediction

2.3.1 Empirical models

Many studies (Gons, 1999; Bukata et al., 1995; Gitelson, 1992, Witter et al., 2009) have attempted to identify spectral regions that are dominantly influenced by specific, individual in-water constituents. Several empirical algorithms have been developed based on important optical features of chlorophyll a in the blue-green and the NIR-red spectral region. Empirical algorithms developed by Witter et al. (2009) (eq. 2.1 below) and Moses et al. (2009) (eq. 2.2 and 2.3 below) are based on simple ratios of spectral reflectance, $R(\lambda_1)/R(\lambda_2)$. Other studies have used a normalized difference approach e.g., the normalized difference vegetation index (NDVI) model. All of these procedures attempt to isolate the effect of the target CPA, in this case chlorophyll a, against the others. Two-band models used for this study include:

\[
chl_{a(MERIS)} \propto \frac{R^{-1}_{490}}{R^{-1}_{560}} \tag{2.1}
\]

\[
chl_{a(MERIS)} \propto \frac{R^{-1}_{665}}{R^{-1}_{708}} \tag{2.2}
\]

\[
chl_{a(MODIS)} \propto \frac{R^{-1}_{66.7}}{R^{-1}_{74.8}} \tag{2.3}
\]
Algorithm developed by Witter et al. (2009) uses the blue-green spectral region and is based on a comparison of remote sensing reflectance for bands in the blue (490nm) and green (560nm) parts of the spectrum. The two-band models from Moses et al. (2009) shown above take advantage of the maximal chlorophyll $a$ absorption region in the Q bands (660-670nm) and a normalizing band that accounts for absorption due to other in-water constituents. The effect of backscattering is also removed since the variability of scattering for wavelengths between 660 to 750nm is assumed to be small. For the MODIS based model, the spectral band near 700nm does not exist therefore the 748nm band is employed.

Recently Dall’Olmo and Gitelson, (2005), described a three-band algorithm for estimating chlorophyll $a$ concentration (eq. 2.4 below). This approach aims at retrieving the concentration of chlorophyll $a$ independent of the other in-water constituents. This model requires the use of three spectral bands and is mathematically expressed in the form of:

$$chl_a \propto \left( \frac{R^{-1}(\lambda_1)-R^{-1}(\lambda_2)}{R(\lambda_1)} \right)$$ 2.4

The choice of spectral bands is based on sensitivity of the spectral region to chlorophyll $a$ (Zimba and Gitelson, 2006). Reflectance at $\lambda_1$ is highly sensitive to the absorption of chlorophyll $a$, while $\lambda_2$ represents a spectral region that is minimally sensitive to the chlorophyll $a$ absorption, but sensitive to the absorption of the other
constituents, and $\lambda_3$ accounts for the variability due to scattering, where reflectance in the NIR wavelength regions is due to TSM, independent of CDOM and chlorophyll $a$. The three-band model was tested using observations from lakes with chlorophyll $a$ concentration varying between 0.02 and 70mg/m³ (Gitelson et al., 2008). This includes the chlorophyll $a$ concentration range observed in our current study area. Moses et al. (2009) found that this model gives accurate estimates of chlorophyll $a$ concentration in productive waters when the waveband location and bandwidth are chosen to match bands of MERIS. For this work, $\lambda_1 = 665\text{nm}$, $\lambda_2 = 708\text{nm}$ and $\lambda_3 = 753\text{nm}$ are used as inputs to model for MERIS data. The mathematical expression using the MERIS bands is presented as follows,

$$
chl_a^{(MERIS)} \propto \frac{R_{665}^{-1} - R_{708}^{-1}}{R_{753}^{-1}}
$$
2.3.2 Semi-empirical models

Additional models used in this study include physically-based inversion techniques that estimate concentrations of color producing agents from reflectance data without the need of collocated synchronous field observations. Such techniques have been employed to retrieve concentrations of chlorophyll \( \alpha \) in freshwater systems using spectroradiometric data (Gons, 1999; Simis et al., 2005). The semi-empirical algorithms developed by Gons (1999) (eq. 2.6 below) and Simis et al. (2005) (eq. 2.7 below) account for the interference by other constituents including correction for \( b_b \), biogenic backscatter (Gons, 1999), and water absorption (Buiteveld et al. 1994).

\[
\text{chl}_\alpha(\text{MERIS}) \propto \left( \frac{R_{708} - b_b}{R_{665} + b_b} \right) \left( \frac{R_{665}^{-1} + b_b}{R_{708}^{-1} + b_b} \right) - b_b - R_{665}^{-1}
\]

\[
\text{chl}_\alpha(\text{MERIS}) \propto \left( \frac{R_{708}}{R_{665}} \right) \left( \frac{R_{665}^{-1} + b_b}{R_{708}^{-1} + b_b} \right) - b_b - R_{665}^{-1} \times \delta^{-1}
\]

The Simis algorithm (eq. 2.7) goes further to correct for the interference from the phycocyanin pigments (\( \delta \)), which is a major accessory pigment in the WBLE, where blue-green algae is dominant.
2.3.3 **Fluorescence Line height**

Light absorbed by pigments in phytoplankton is mostly used for the photosynthetic process, however, excess energy will be lost as heat or can be re-emitted at a longer wavelength centered near 680nm also known as the chlorophyll \( a \) fluorescence peak (Maxwell and Johnson, 2000). This phenomenon is referred to as inelastic scattering; it involves the scattering of an incident photon during which part of the photon’s energy is changed to the internal energy of the interacting matter and the scattered photon is emitted at longer wavelength. The magnitude of chlorophyll \( a \) fluorescence gives important information regarding the (a) physiological state of phytoplankton, (b) causes of phytoplankton bloom and their collapse, and (c) concentration of chlorophyll \( a \). Gower et al. (1999) studied the effect of chlorophyll \( a \) fluorescence variability on peak reflectance near the infra-red spectral region. The study which was conducted in Case 2 waters demonstrated that the fluorescence peak height increased with increasing chlorophyll \( a \) concentrations. Studies by Schalles, (2006) indicated that although fluorescence peak is primarily focused near 680nm, the fluorescence peak may shift to higher wavelengths (680 - 710nm). The study also ruled out the effect of scattering due to tripton on the NIR peak.

A modified version of the Fluorescence Line Height (FLH) algorithm developed by Gower et al. (1999) is used in this study to assess its potential in describing the chlorophyll \( a \) variability in the WBLE. The original algorithm is modified based on inspection of the MERIS-derived spectral pattern of the WBLE. Fluorescence line height
was quantified using data from MERIS spectra by comparing the reflectance at 680nm, where the peak NIR reflectance is observed, and to a baseline interpolated from measurements at 664nm (the absorption trough) and 778nm (Fig 2.1). Accordingly, the FLH algorithm is given by:

\[
chl_a(\text{MERIS}) \propto R_{680} - R_{664} - (R_{778} - R_{664})^{(680-664)/(778-664)}
\]

**Figure 2.1.** A schematic of the FLH model on MERIS based spectra of a pixel representing the WBLE. The dash/solid vertical red lines represent fluorescence magnitude above a baseline extending between 664 and 778nm.
2.4 CDOM prediction

CDOM is composed of substances that originate from the decomposition of organic matter. It is usually abundant in inland water bodies and coastal habitats, where there are highly productive plant communities that are influenced by human activities. CDOM concentration is often reported in units of optical absorption (m$^{-1}$) at a fixed wavelength (usually at 420nm) (Kutser et al., 2005). It is a major color producing component in the WBLE, absorbing significant amounts of the radiation in ultraviolet (UV) and blue spectral ranges. The CDOM absorption spectrum is characterized by an exponentially decaying curve with increasing wavelength (Bukata et al., 1995, Binding et al., 2008). CDOM has a fluorescence effect near 650nm which is usually masked by the absorption due to phytoplankton. Tank experiments were conducted by Schalles et al. (2002) to assess the spectral changes of Case 2 waters constituting cyanobacteria. Their results indicated that, with increasing CDOM, reflectance deceased mainly in the VIS range. CDOM showed higher influence on the blue-green (443/556) chlorophyll $a$ model than the red/NIR (670/700) models by causing the blue-green based chlorophyll $a$ estimates to be biased high relative to the results from red/NIR models. This indicates that the sensitivity of the blue-green model to multiple variables (chlorophyll $a$ and CDOM) causes bias as well as errors in the prediction of the CPAs.
A study conducted by Binding et al. (2008) showed that the CDOM in Lake Erie does not have a significant correlation with phytoplankton concentration suggesting that allochthonous CDOM acts as a CPA that is independent of chlorophyll $a$. However, it has an impact on photosynthetic process by absorbing some of the light at wavelengths that are used for photosynthesis which in turn affects the structure of the lake’s ecosystem. Detailed understanding of the spatial and temporal dynamics of CDOM in the WBLE requires monitoring this CPA using high resolution satellite data.

Ocean sensors (e.g. MERIS, MODIS) have high spatio-temporal resolution offering frequent global CDOM data, moreover, they are equipped with detectors that have radiometric sensitivities (16-32bits) optimized for measurements in aquatic bodies. The high radiometric resolution is crucial especially in the Case 2 waters where the net water leaving radiance is low. A drawback of these satellites is their low pixel resolution (250 to 1200m) which can potentially introduce errors in prediction models. Terrestrial satellite sensors (e.g., Landsat ETM+, SPOT) have a spatial resolution adequate for measurements in the WBLE (4–30m). However, radiometric sensitivity of these sensors (8bits) makes it difficult to detect optical changes due to CDOM activity. The Advanced Land Imager (ALI) sensor launched aboard the Earth Observation Satellite has 10-30m pixel resolution, 3 days revisit period, 12bit radiometric resolution and spectral bands sensitive to CDOM making it an ideal sensor to detect and map CDOM in large water bodies such as the Lake Erie.
In order to relate the radiances values recorded at sensor to CDOM concentrations, a satellite algorithm is required. I used the band-ratio algorithm developed by Kutser et al. (2005), which is defined using the following equation:

\[ a_{CDOM(ALI)}(420) = 5.13\left(\frac{B_2}{B_3}\right)^{-2.67} \]

Where, \( a_{CDOM} \) is the CDOM absorption, \( B_2 \) is band 2 (525-605nm) and \( B_3 \) is band 3 (630-690nm) from the Advanced Land Imager (ALI) instrument. The equation is calibrated against measurements of CDOM absorption from filtered water samples obtained from lakes in Sweden, which have optical properties similar to the WBLE. The performance of the algorithm in the WBLE is assessed using linear regression analysis against the in-situ CDOM measurements made using a Cyclop-7 CDOM sensor (Turner design, Inc.) equipped with an excitation filter at 370nm and emission filter at 420nm.

2.5 Calibration-validation

The strength of regression based algorithm depends on its prediction capability. Models can have large prediction errors despite good calibration results. The performance of three of the best empirical and semi-empirical chlorophyll \( a \) algorithms (i.e., with good calibration results) are evaluated by running model validation using data outside the calibration dataset. Statistical measures such as mean residuals (MRES), \( R^2 \), and RMSE values are used as criteria to assess model stability and efficiency. Additionally, the stability of beta coefficients between the calibration and validation models is evaluated.
2.6 Results and Discussion

Figure 2.2 is a spectral plot of the MERIS reflectance representing apparent optical properties of the water in the WBLE. The reflectance values are highly variable in the visible and NIR range. The blue-green spectral range shows reflectance patterns typically observed in turbid waters (Schalles et al., 2002), with significant absorption due to chlorophyll \textit{a}, CDOM and other associated constituents. A local reflectance maximum is observed near 560nm, also referred to as the green peak. This is the mainly due to minimum absorption by chlorophyll \textit{a} combined with scattering effects. The reflectance peak at the red/NIR edge, which is mainly due to fluorescence from phytoplankton, is not strongly apparent in the satellite data. The data indicates local minimum near 620nm due to phycocyanin, which is the major accessory pigment in blue-green algae. The local minimum observed near 664nm is due to chlorophyll \textit{a} absorption. The absorption troughs in the red region, near 620 and 664nm and the reflectance peaks at 560 and 680nm are more noticeable in spectra from stations closer to Sandusky Bay than in the open waters of WBLE (Fig 2.2 and 2.3).

The fact that (a) MERIS data has relatively course spatial resolution of 300 m, (b) phytoplankton density in the WBLE is relatively high and localized, and (c) absorption effects due to water and CDOM are high, the average water leaving radiances from the WBLE is low. This makes distinguishing optical features that result from the interaction of the various CPAs with the incoming irradiance more challenging. As will be indicated in Chapters 3 and 4, the spectral data generated using the lab-based spectrophotometer
clearly shows the various apparent optical characteristics including the troughs and peaks associated with the various CPAs (Fig 4.7, section 4.4.2). The spectrophotometer data is not affected by the absorption effects of water, and the effects from CDOM are significantly lower. The data is also free from the attenuation effects of path radiance.

The spectral resolution of MODIS in the visible/NIR is even lower than that of MERIS (Fig 2.3). Therefore, important signals that would otherwise characterize the various optically active in-water constituents are further masked and the optical features are less easily identified in the spectral signatures of the WBLE. MODIS-based spectra of the WBLE are mainly characterized by (a) low reflectance values in the blue region which is attributed to absorption effects of CDOM and chlorophyll $a$ and (b) local maxima in the green region due to low absorption capacity of chlorophyll $a$. The red absorption maxima are not well represented near the 667nm band. In the MODIS data, the fluorescence peak (680-715nm), is not detected due to the lack of a band near this spectral window. The NIR region beyond 750nm has low reflectance values due to high absorption by water. Stations having high reflectance in the green region also show relatively higher values in the NIR range. This is mainly due to the scattering effects by the associated detritus and suspended sediments.
**Figure 2. 2.** Spectral plot of pixels representing 18 stations based on MERIS data. Dashed red lines represent spectra of stations closer to Sandusky Bay.

**Figure 2. 3.** Spectral plot of pixels representing 18 stations based on MODIS data. Dashed red lines represent spectra of stations closer to Sandusky Bay.
2.6.1 CPAs in the WBLE

The average in-situ concentration of chlorophyll \( a \) was 5.68\( \mu g/l \) and 8.07\( \mu g/l \) in June and September of 2009, respectively. In September, the concentrations across the basin varied between 4.96 and 9.61\( \mu g/l \) with the higher concentration values recorded at stations located in the Sandusky Bay. In summer of 2010, chlorophyll \( a \) concentrations in the basin ranged from 2.14 to 10.67\( \mu g/l \). The standard deviation of the chlorophyll \( a \) concentrations decreased from 2.71\( \mu g/l \) to 0.79\( \mu g/l \) between July and September, in 2010. The high standard deviation of chlorophyll \( a \) concentrations in July is attributed to the relatively higher seasonal algal density in the Sandusky Bay. During the early summer, river discharges are high and significant amounts of terrestrial matter, including nutrients, are dumped into Sandusky Bay (See Fig 4.5, section 4.4.1). This increases the combined concentrations of allochthonous and autochthonous phytoplankton and consequently higher chlorophyll \( a \) concentrations are recorded at stations located in Sandusky Bay (Stations 17, 19 and 20). This condition results in large variability of chlorophyll \( a \) between the waters in Sandusky Bay and the central WBLE. By late summer (September), the river discharges decrease and the waters between the two sub basins undergo mixing and the variability of chlorophyll \( a \) in the WBLE lowers.

Descriptive statistics of CDOM in the WBLE shows that the absorbance ranged from 0.115 to 1.38\( m^{-1} \), with an average value of 0.375\( m^{-1} \). This result is consistent with the findings of Binding et al. (2008) from their filtrate samples collected in Lake Erie.
During July of 2010, significantly higher absorbance was recorded at stations 3, 4, 17, 19 and 20 (Fig 2.4). Stations 3 and 4 are located close to the discharge zones of the Toussaint and Portage Rivers and stations 17, 19, and 20 are located in the Sandusky Bay, which is a shallow basin heavily influenced by terrestrial matters. This suggests that the optical properties of the waters near discharge zones are heavily influenced by terrestrially derived CDOM. As shown in Figure 2.4, the spatial variability of CDOM concentrations across the WBLE decreases during the summer period. This may be attributed to the decreasing stream flow from late spring to summer and at the same time dispersion of in-water constituents into central WBLE.

The spatial structure of CDOM is similar to the chlorophyll \( a \) dynamics, providing strong evidence of the influence of rivers on the algal distributions observed in the WBLE (Fig 2.5). Riverine input makes nutrients readily available for primary producers, and this increases the productivity of the lake system. Constituents such as CDOM absorb significant amount radiation in the UV-B region, thus reducing the energy available for primary producers. This may influence the community structure of the lake system by favoring primary producers that can utilize energy in the longer wavelength region. In the WBLE, the cyanobacteria represent the dominant bio-community (Makarewicz, 1993). These organisms posses assessor pigments (e.g., phycocyanin and carotene) allowing them to harvest radiant energy beyond the blue spectral region and making their productivity less sensitive to the presence of CDOM.
Away from the littoral zone, in the central WBLE, CDOM concentration is relatively low and the influence of rivers is reduced. As described in the earlier part of the dissertation, a different bio-community structure makes up the biomass. Studies by Ortiz et al. (in press) have indicated that the blue-green algae are generally lower in central WBLE than in Sandusky Bay, suggesting that this assemblage is more important in the Sandusky Bay than in central WBLE. The study has further indicated that central WBLE is dominated by the diatom/green algae community, which is consistent with the idea of CDOM interference.
Figure 2.4. Spatial and temporal variability of CDOM in the WBLE during summer of 2010 based on data from in-situ measurements. Missing values due to instrument errors.

Figure 2.5. In-situ measurement of CPAs shows the variability of chlorophyll $a$ and CDOM in the WBLE during summer of 2010.
2.7 Chlorophyll $a$ prediction

2.7.1 Empirical models

2.7.1.1 Empirical two-band blue-green model

Band ratios such as $R_{443}/R_{565}$ and $R_{520}/R_{565}$ were commonly used to estimate chlorophyll $a$ in ocean waters (Gordon and Morel, 1983; Morel and Prieur, 1977). However, O’Reilly et al (1998) and other studies have clearly indicated that blue-green models did not perform well in Case 2 waters due to the presence of multiple constituents (CDOM, suspended matters) affecting the absorption characteristics of the Soret bands. Witter et al. (2009) developed a regional blue-green chlorophyll $a$ algorithm based on co-located satellite observations and field-collected samples from the WBLE. The algorithm was applied using MERIS data from 2009 and 2010 and the result is illustrated in Figure 2.6. The regional algorithm gave $R^2$ values of 0.46 which indicated that the blue-green algorithm overestimates the chlorophyll $a$ values in the WBLE. Careful examination of Figure 2.6 shows that for the low chlorophyll $a$ concentrations ($< 6 \, \mu g/l$), the predicted values plot closer to the 1:1 line, but with higher chlorophyll $a$ concentrations the model overestimates chlorophyll $a$ values showing exponential relationship with the field measurements.

Close observations of the data set indicated that, the bias towards high values correspond to samples collected from stations that are closer to river discharge zone or in
areas where Secchi depth was low. The overestimation is due to optical interference from other in-water constituents such as CDOM.

**Figure 2.6.** Regression plot of MERIS based blue-green chlorophyll $a$ model versus in-situ chlorophyll $a$ concentrations.
2.7.1.2 Empirical two-band red /NIR model

MERIS is equipped with chlorophyll $a$ red absorption band (664nm) and a narrow NIR band (708nm, 7.5nm band width) which is sensitive to non-algal absorption. Regression analysis between the two-band red/NIR MERIS model and chlorophyll $a$ concentration for the WBLE is shown in Figure 2.7. The MERIS data is able to explain approximately 60% of the chlorophyll $a$ variability in the WBLE. The two band ratio removed effects of other in-water constituents and therefore the magnitude of the model will increase proportionally with chlorophyll $a$ concentration. Application of band ratios removes systematic errors from instrumental disturbances and minimizes effects due to path radiance especially when working in smaller area where atmospheric effects can be assumed to be uniform. Another important strength of the two band ratio model is capability to remove back scattering effects. However, scattering effects can significantly deteriorate the performance of the model. In this case, some of the factors that may have limited the performance of the two band red/NIR model are absorption effects of accessory pigments e.g., phycocyanin, and absorption due to the dissolved organic matter which varies as a function of $(\lambda^{-1})$.

Application of the red/NIR two-band MODIS based model data gave a coefficient of determination of 0.51 for chlorophyll $a$ (Fig 2.8). The relatively low $R^2$ value is due to multiple factors. The first factor is attributed to the fact that the normalizing spectral band used in the model is located at 753nm where water leaving radiance is extremely low making it extremely sensitive to stochastic errors and path radiance effects. The second
factor is that related to MODIS based model using bands at 678 and 753nm, which are relatively further apart. This can easily lead to differences in the amount of backscattering recorded at the two wavelength regions and therefore the ratio procedure will not remove the backscattering effects. A third factor for the low $R^2$ value is the pixel resolution of the MODIS data. MODIS sensor has a coarser spatial resolution (>500m) relative to the MERIS data (300m), with the exception of MODIS band 1 which is designed to take image with 250m resolution. The coarse resolution of the MODIS data will merge optical data from larger areas, reducing the capability to identify local variability in chlorophyll $a$ concentrations. In cases where chlorophyll $a$ field is highly spatially inhomogeneous, the sensor will average out the phytoplankton density and therefore the modeled concentrations of chlorophyll $a$ will not reproduce the highest chlorophyll $a$ concentrations. As shown in Figure 2.8, the effect of pixel size will also result in overestimation of the lowest chlorophyll $a$ values and this will flatten out the best fit line. In Figure 2.3, the red absorption trough in the MODIS spectra is of low magnitude or almost absent. The above factors reduce the efficiency of the two band red/NIR model when being applied to MODIS data.
**Figure 2.7.** Regression plot of MERIS based two-band red/NIR chlorophyll $a$ model versus the chlorophyll $a$ concentrations measured in-situ

**Figure 2.8.** Regression plot of MODIS based two-band red/NIR chlorophyll $a$ model versus the chlorophyll $a$ concentrations measured in-situ
2.7.1.3 Empirical three-band red /NIR model

Regression analysis of three-band MERIS model based on $R_{664}$, $R_{708}$ and $R_{753}$ versus chlorophyll $a$ gave $R^2$ value of 0.57, indicating that this model has some potential for explaining chlorophyll $a$ variability in the turbid WBLE. Model performance is illustrated using regression analysis (Fig 2.9).

The $R^2$ values obtained from the MERIS based three-band model is relatively low and this can be generally attributed to the absorption effects of water on the longer wavelengths (> 720nm). In comparison to the two-band model, the performance of the three-band model is lower. This is because firstly, with increasing wavelength ranging from 664 to 708nm, the absorption of water exponentially increases due to the low energy incident photons being easily attenuated (Buitenveld et al., 1994). This in turn reduces reflectance from scattering effects of suspended particles and therefore subtracting absorption effects between the red and the NIR bands may not effectively remove the effects of backscattering causing the relationship between the model derived chlorophyll $a$ estimate and in-situ chlorophyll $a$ to be relatively low. A second potential contribution to errors in the three-band prediction model of chlorophyll $a$ is due to use of $R_{753}$ in the model. This band is a multiplication factor that is usually applied to account for the backscattering from inorganic matters. However, in aquatic bodies, the irradiance of water at this wavelength region produces very low reflectance values and this low values are easily influenced by effects of path radiance and are also more susceptible to
stochastic errors (IOCCG, 2000). This further contributed to errors in the prediction model of chlorophyll $a$.

**Figure 2.9.** Regression plot of MERIS based three-band red/NIR chlorophyll $a$ model versus the chlorophyll $a$ concentrations measured in-situ.
2.7.2 Semi-analytical models

2.7.2.1 Simis red/NIR model

The relationship between the Simis NIR/red model and in-situ chlorophyll $a$ concentration is stronger, with a coefficient of determination of 0.65. Figure 2.10 shows a plot of the Simis model versus observed chlorophyll $a$ concentrations. The best fit line is relatively close to the 1:1 line, indicating the potential of the Simis model in retrieving CPAs. The higher predictive power of the Simis model is attributed to the incorporation of correction factors such as absorption effects due to pure water ($a_w$), backscatter ($b_b$) from suspended matter and interference due to phycocyanin ($\delta^{-1}$). The relatively higher $R^2$ value clearly indicates the importance of accounting for optical interference of other in-water constituents such as phycocyanin and suspended matter on the chlorophyll $a$ reflectance.

Although it is demonstrated that the Simis model performs well in the retrieval of chlorophyll $a$ in a wide range of eutrophic lakes (Simis et al., 2005), the predictive power of the Simis model can further be improved by optimizing correction factors such as $\delta^{-1}$ based on local measurements of pigment absorption. This will provide an opportunity to correct for any offset values due to absorption by tripton and CDOM, which are significant water quality parameters in the WBLE (Binding et al., 2008). The Simis model only corrects for the absorption effects due to phycocyanin using correction factors derived from slope of a line constructed between the measured absorption values of...
phytoplankton and total absorption at 665nm. Generally, the model ignores the offsets values due to CDOM and tripton assuming it to be negligible near the red/NIR region and does not subtract it from the computed chlorophyll $a$ absorption.

**Figure 2.** Regression plot of MERIS based Simis chlorophyll $a$ model versus the chlorophyll $a$ concentrations measured in-situ

\[
y = 1.4291x - 2.42
\]
2.7.2.2 Gons red/NIR model

The Gons red/NIR model is very similar to the Simis model, except that the model does not account for absorption by accessory pigments. Plots of Gons red/NIR based chlorophyll $a$ estimate versus in-situ chlorophyll $a$ concentrations show coefficient of determination of 0.61 (Fig 2.11). Relative to the Simis algorithm, the Gons model gave lower prediction accuracy with very flat best fit line. This is mainly due to simplified assumptions taken in the development of the model. These include the assumption of negligible pigment reflectance near 708nm region. In addition, absorption by CDOM and tripton in the red/NIR is not accounted for in the model.

The Gons model overestimates low to moderate chlorophyll $a$ concentrations, and tends to underestimate the higher chlorophyll $a$ concentrations (Fig 2.11). The simplified assumption of negligible absorption of pigment in the NIR region leads to overestimation by the reflectance in the red region (665nm) and therefore lower ratio between 708 and 665nm, causing underestimation of chlorophyll $a$ concentrations. Similarly, the assumption of the zero-absorption effects of CDOM and tripton may cause an underestimation of reflectance in the red region (lower $R_{708}/R_{665}$ ratio), which may not necessarily be correlated with chlorophyll $a$ concentration. This leads to erroneous prediction of higher chlorophyll $a$. In the WBLE, the correlation between chlorophyll $a$ and CDOM is 0.37.
Regression of MERIS Gons Model against in-situ chlorophyll $a$ ($R^2=0.612$)

$y = 0.2804x + 6.1956$

Figure 2.11. Regression plot of MERIS based Gons chlorophyll $a$ model versus the chlorophyll $a$ concentrations measured in-situ
2.7.3 Fluorescence Line Height (FLH)

According to studies conducted by Abbott et al. (1982), the fluorescence intensity can vary by over a wide range due to changes in light intensity and nutrient stress. Other studies have indicated that FLH magnitude is dependent on the fluorescence quantum yield which is a function of the availability of electron acceptors in the photosynthetic pathways. Once a photosynthetic unit absorbs light and the electron acceptor accepts an electron, the unit will not be able to transfer the extra photon through the photosynthetic pathway and the excess energy is emitted as chlorophyll fluorescence (Maxwell and Johnson, 2000).

Regression analysis of the computed FLH values against in-situ chlorophyll $a$ gave moderately good relationship of $R^2 = 0.54$. Figure 2.12 shows the linear relationship between the MERIS based fluorescence peak height model and in-situ chlorophyll $a$. The standardized residuals did not show any pattern, indicating that the model has accounted for most of the systematic chlorophyll $a$ variability observed in the WBLE. This illustrates that an understanding of fluorescence (inelastic scattering effect) can result in the ability to use it to estimate concentrations of chlorophyll $a$.

The regression plot in Figure 2.12 shows that the MERIS based FLH model consistently underestimates the magnitude of in-situ chlorophyll $a$. This can be attributed to several factors, (a) the radiance leaving the water undergoes several modifications before it reaches the sensor. For example, location of the fluorescence band is in the
region of the spectrum where there are several narrow absorption features. In particular, there is an oxygen absorption band at 687nm and (b) in turbid waters such as the WBLE, the presences of suspended particulates scatter the red wavelength and attenuate the fluorescing photons. This partly explains the lack of distinct peaks at 680nm in the spectra of certain locations where in-situ chlorophyll $a$ has been recorded. Therefore, the absence of well-defined fluorescence peak is not necessarily an indication for the absence of chlorophyll $a$ in WBLE, a sub-basin composed of multiple optically active components.

A limiting factor of targeting fluorescence band is that, the peak may shift to higher wavelengths with increased chlorophyll $a$ concentration (Gower et al., 1999). Field analysis performed by Gitelson (1992) showed that the peak position varied from 680 to 715 nm. MERIS does not have the spectral resolution required to detect this subtle shift in the fluorescence peak and therefore the FLH model can fail if shifting occurs. In this work, fluorescence peaks in the MERIS spectra were recorded at 680nm.
Regression of MERIS Fluorescence Line Height (FLH) against in-situ chlorophyll $\alpha$ ($R^2=0.54$)

\[ y = 0.5419x + 2.1275 \]

Figure 2.12. Regression plot of MERIS based FLH model versus the chlorophyll $\alpha$ concentrations measured in-situ
2.7.4 Calibration-validation of chlorophyll \( a \) algorithms

The two-band Moses et al. (2009) and the semi-empirical algorithms by Gons (1999) and Simis et al. (2005) have produced the best calibrated results among the different empirical and semi-empirical chlorophyll \( a \) algorithms, with \( R^2 \) of 0.61, 0.61 and 0.65, respectively. The stability of the algorithms is shown in Table 2.1. All three models show significant changes in the validation-calibration parameters. The MRES errors have increased by an order of two to three magnitudes and the RMSE has also increased for all three models. In addition, contrary to the higher values of the RMSE for validation, the validation \( R^2 \) values increased. Slope coefficients (Beta values) of the validation models also decreased. This clearly demonstrates that these algorithms lack stability when applied to small dataset. Despite the increases in the errors and the weakness of the models, the Simis shows relatively better performance as compared to the Gons and the two-band Moses algorithm.

<table>
<thead>
<tr>
<th>Spectral Indices</th>
<th>Dataset-Chlorophyll ( a )</th>
<th>N</th>
<th>MRES</th>
<th>( R^2 )</th>
<th>RMSE</th>
<th>Beta values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simis Training</td>
<td>50</td>
<td>-0.00</td>
<td>0.60</td>
<td>0.65</td>
<td>1.54</td>
<td></td>
</tr>
<tr>
<td>validation</td>
<td>21</td>
<td>0.07</td>
<td>0.78</td>
<td>0.84</td>
<td>1.29</td>
<td></td>
</tr>
<tr>
<td>Gons Training</td>
<td>50</td>
<td>-0.00</td>
<td>0.57</td>
<td>0.68</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>validation</td>
<td>21</td>
<td>0.02</td>
<td>0.74</td>
<td>0.90</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>Moses Training</td>
<td>50</td>
<td>-0.00</td>
<td>0.57</td>
<td>0.68</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>validation</td>
<td>21</td>
<td>0.03</td>
<td>0.76</td>
<td>0.89</td>
<td>0.55</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1. Summary of calibration-validation of chlorophyll \( a \) algorithms based on Simis et al. (2005), Gons (1999) and two-band Moses et al. (2009).
2.8 CDOM

Linear regression between the ALI-based Kutser et al. (2005) model and the in-situ CDOM concentrations in the WBLE show relatively strong correlation, with $R^2 = 0.67$. The regression plot shows that the Kutser et al. (2005) algorithm tends to overestimate CDOM concentration by approximately a factor of 2 (Fig 2.13). Although band 1 (450-515nm) is more sensitive to CDOM, studies conducted by Kutser et al. (2005), justified the use of band 2 to minimize effects of absorption by chlorophyll $a$ which is greater in band 1 relative to band 2. Hence, the band2/band3 ratio is considered to be more suitable than the band1/band3 ratio for purpose of CDOM retrieval.

However, in Case 2 waters such as the WBLE, the concentration of the suspended particulates is significant and the minimum chlorophyll $a$ absorption effect coupled with particulate scattering in band 2 results in higher reflectance. This causes higher band2/band3 ratio which leads to overestimation of CDOM values when applying the Kutser et al. (2005) algorithm. This is effect is illustrated in Figure 2.11. A second source of bias arises from the use of band 3. Kutser et al. (2005) used band 3 for normalizing purposes assuming the influence of CDOM to be negligible. Studies conducted by Schalles (2006) clearly show that absorption effects of CDOM extend well into band 3 for concentration ranges observed in the WBLE. Furthermore, band 3 of ALI overlaps with the spectral region where chlorophyll $a$ has its maximal absorption. These combined absorption effects from two independent constituents (CDOM and chlorophyll $a$) causes the ALI based Kutser et al. (2005) algorithm to be biased towards higher values.
Regression of ALI based CDOM estimates (Kutser al., 2005 model) against in-situ CDOM measurements ($R^2=0.67$)

\[ y = 0.6701x + 0.1239 \]

**Figure 2.13.** Regression plot of ALI based CDOM concentrations versus in-situ measurements. CDOM algorithm (equation 2.9) from Kutser et al. (2005).
2.9 Uncertainties

The satellite algorithms presented in this work illustrate the strength of the red/NIR and blue-green bands in detecting the variability of chlorophyll $a$ and CDOM, respectively, in turbid waters such as the WBLE. However, studies conducted by Gitelson et al. (2008); Moses et al. (2009); Simis et al. (2007); Kutser et al. (2005) indicated that the performance of the satellite algorithm comes with certain uncertainties. The models are not stable and calibration work has always been difficult due to various assumptions and physical factors. Several factors limit the potential of satellite algorithms to estimate the concentrations of in-water constituents.

2.9.1 Assumption of negligible effects due to other components

In Case 2 waters, the red/NIR spectral range is often used to develop algorithms that estimate pigment concentration in the water column. The red-edge of the spectrum is usually invoked in order to minimize the complicating influences of other in-water constituents such as CDOM and inorganic particles. However, as shown in section 4.7.2.2, Gons (1999) assumption of negligible effects of CDOM and inorganic particles in the red/NIR models led to significant bias. In a similar case, in the prediction of CDOM, the contribution of chlorophyll $a$ caused the Kutser et al. (2005) algorithm to be biased towards higher values.
2.9.2 Temporal variability

Data collected using conventional in-situ methods cannot sample multiple water points at the same time. However, a satellite has the capability to capture optical properties of large inland water bodies instantly. The time difference between the sampling period and the satellite data may be enough to have significant changes in the optical properties of the water, especially in shallow-turbid waters such as the WBLE. Instant wind action over the waters of WBLE causes mixing, significantly changing the optical and hydrochemical properties of the water column. The small difference in the time of data acquisition can introduces random errors that significantly reduce the correlation factor between the retrieved constituents from the two platforms hence degrading model performance.
2.9.3 Pixel heterogeneity

In optically complex aquatic systems such as in the WBLE, the spatial variability of in-water constituents may be large. The spatial resolutions (0.09km$^2$ for MERIS and 1km$^2$ for MODIS) may be too coarse to correlate with in-situ point measurements. In such cases collocated in-situ data may not represent conditions averaged over the satellite pixel. This can reduce the correlation between the in-situ measurements and the satellite retrievals.

2.9.4 Path radiance correction

The influence of path radiance due to atmospheric scattering can be substantial and the success of satellite remote sensing in predicting water quality parameters depends on the accuracy of the applied atmospheric correction module. As indicated in chapter one, several atmospheric correction procedure have been tested to correct for path radiance. The MUUM and the Case 2 waters regional algorithms have been selected as the optimal models. However, without synchronized in-situ measurements of water leaving radiances, it is not possible to assess accuracy of the procedures from the reflectance curves. Implementing accurate corrections due to interference of atmospheric constituents remains a significant challenge within optical remote sensing.
2.10 Conclusion

The results have illustrated the potential of the remote sensing algorithms utilizing satellite data in red/NIR wavelength region to estimate chlorophyll \( a \) concentrations in optically complex productive waters. Among the various models applied, results indicate that the Simis et al. (2005) algorithm provides the best estimates of chlorophyll \( a \) in the optically complex waters of the WBLE. The Simis algorithm accounts for back scattering effects and the absorption effects due to major accessory pigment, phycocyanin. These corrections significantly increased its performance relative to others, such as the Gons (1999) model which does not account for effects of phycocyanin. The addition of correction factors for other optically active constituents including suspended sediments and CDOM will improve model performance and the task of formulating the correction factors is a promising direction for future research.

The results clearly indicate that the MERIS sensor with its higher spectral and spatial resolution yielded better results in predicting CPAs when compared with data generated from the MODIS sensor. The increased spatial and spectral resolution of the FR-MERIS sensor aboard Envisat along with its chlorophyll \( a \)-sensitive bands such as 664 nm and 680nm significantly increases the effectiveness of satellite-based remote sensing to estimate chlorophyll \( a \) distributions and hence effectively monitor of water quality in large water bodies such as the WBLE. The ALI sensor with its relatively high radiometric and spatial resolution in predicted CDOM with a factor of 2 bias. A satellite
with similar radiometric and spatial resolution having more focused and narrower spectral bands would provide better CDOM estimates.

To develop more robust bio-optical models, future efforts must focus on: (a) improving atmospheric correction methods through calibration-validation techniques, (b) designing better field methods to accommodate for the temporal difference and effects of ground resolution between in-situ and satellite data, and (c) applying mathematical transformation techniques on full spectral information instead of narrow individual bands signatures. This will enhance retrievals of CPAs from satellite based sensors and improve model stability.
CHAPTER 3

Application of principal component analysis to estimate color producing agents in turbid waters using spectrophotometer data.
Abstract

The complex composition and distribution of CPAs (color producing agents) in turbid waters such as the Western Basin of Lake Erie (WBLE) presents a challenge to the application of remote sensing data for differentiating among in-water constituents and estimating their concentrations independently. In this paper, multivariate procedures are applied to lab-based spectrophotometer data to estimate the concentration of chlorophyll \( a \) and suspended matters in the WBLE. Principal Component Analysis (PCA) of first-derivative transformed hyperspectral data from the spectrophotometer extracted three significant spectral components for each cruise, explaining up to 88% of the spectral variability (PCA returns more than three components, in this case only the 1\textsuperscript{st} three components were meaningful).

Spectral matching using reference spectra indicated that the extracted patterns represent signatures of in-water constituents that govern the optical properties of the WBLE namely, cyanobacteria and diatoms associated with green algae. The third factor represents a complex mixture of phytoplankton and inorganic minerals. The spectrophotometer data clearly revealed known spectral features associated with phytoplankton, such as the absorption minima near 550 and 700nm which can be attributed to the minimum of absorption and fluorescence of chlorophyll \( a \), respectively. The method also extracted the absorption peaks due to chlorophyll \( a \), near 670nm, and due to phycocyanin, near 620nm. Spectral responses from inorganic particles were primarily picked up in the 3\textsuperscript{rd} factor. Principal Component (PC) regression of chlorophyll
a on the PC scores indicated that 63.4% of variation of chlorophyll a in the WBLE can be explained by two components. Factors 2 and 3 explain 60% of the joint spatiotemporal variability of suspended matters in the WBLE. This indicates that the multivariate technique efficiently isolates patterns that represent single constituents including chlorophyll a and suspend material. The R$^2$ between the PC based regression and in-situ chlorophyll a measurements was 0.701 with RMSE = 0.81μg/l, and for total suspended matters the regression model has R$^2$ of 0.75, RMSE = 2.861mg/l. This illustrates the potential of remote sensing data in accounting for the bio-optical variability due to phytoplankton and suspended matters in turbid Case 2 waters.

**Keywords:** chlorophyll a, remote sensing, Lake Erie, PCA.
3.1 Introduction

One of the greatest environmental challenges anticipated for the 21st century is meeting society’s demand for fresh water. In the United States, the increasing utilization of open water bodies has led to the deterioration of water quality and aquatic ecology, placing the future of these resources at risk. Effective water quality monitoring is critical to address the question of how various natural and anthropogenic factors affect the health of these environments. Obtaining these observations through in-situ methods is challenging for large water systems, such as Lake Erie.

Nutrient controls in Lake Erie led to significant reduction in the biomass density in the 1980’s, however large blooms were recorded in the Western Basin of Lake Erie in 2009 (USEPA, 2009). Distributions of phytoplankton in space and time have major implications for water quality and ecosystem function. Water quality assessments for Lake Erie are largely based on conventional in-situ measurements with limited spatial and temporal resolution. This makes it difficult to understand the dynamics of Color Producing Agents (CPAs) such as phytoplankton, Total Suspended Matter (TSM) and Colored Dissolved Organic Matter (CDOM).

Satellite-based measurements using multi-spectral sensors (MERIS, MODIS, SeaWiFS) have better spatial coverage and temporal resolution to characterize dynamic water quality properties, but require the development of algorithms that relate the spectral reflectance measured by the sensor to the concentrations of CPAs (e.g., Doerffer and
Ocean satellites have radiometric sensitivities optimized for measurements in turbid waters which normally have low reflectance due to the high absorption by water and its constituents such as CDOM. A potential limiting factor for satellite-based estimates of the concentrations of in-water constituents is the depth of penetration at which the recorded reflectance signal originates. According to work done by Ortiz et al. (in press) in the Western Lake Basin of Lake Erie, surface water conditions are representative of the bio-optical properties. This makes satellite applications to estimate concentrations of CPAs feasible.

CPAs change the optical properties of water by absorbing and scattering incident light (Bukata et al., 1995; Gordon and Morel, 1983). Chlorophyll $a$ concentrations have been considered as important index in assessing water quality for coastal waters in surveillance programs for harmful algal blooms (Glasgow et al., 2004). Morel and Prieur (1977) have classified marine waters as Case 1 and Case 2 based on optical properties. In Case 1 waters, such as the epi-pelagic ocean environments, phytoplankton pigments govern the optical properties of the water. Many coastal and most inland waters are classified as Case 2 waters, because in addition to phytoplankton, constituents such as dissolved organic matter and suspended matter occur in abundance. The optical property of such waters is thus governed by multiple components.

Application of satellite remote sensing methodologies to estimate concentrations of individual CPAs requires the development of empirical or semi-empirical algorithms. These algorithms are developed by relating the reflectance measured remotely with the...
in-situ observations of individual CPA concentrations. Ground-truthing is achieved by comparing spatially and temporally co-located satellite and in-situ observations. Spectral variations in the backscattered flux of Case 1 waters are primarily related to concentration of chlorophyll \( a \), which varies as a function of the phytoplankton population. In marine environments satisfactory relationships have been described for chlorophyll \( a \) concentration ranging between 0.02 to 78mg/m\(^3\) (O'Reilly et al., 2000; Gordon et al., 1988; Gons, 1999). In many coastal waters, the water-leaving irradiance is affected by multiple optically active components. As shown by Witter et al. (2009), optical detection of chlorophyll \( a \) in the Case 2 waters of the Western Basin of Lake Erie (WBLE) using Case 1-derived algorithms is problematic. The challenge is attributed to optical heterogeneity of Case 2 waters due to interference from higher concentrations of suspended matter and dissolved organic matter.

One strategy for differentiating the overlapping and correlated in-water constituents is utilizing both the visible and the infrared portions of the spectrum (Gons, 1999; Gitelson, 1992). The combined effects of absorption due to phytoplankton, suspended matter, dissolved organic matter and pure water produces distinct spectral features in the visible and NIR (near-infrared). The main spectral features of Case 2 waters include absorption maxima at the blue and red wavelengths, and absorption minima at green wavelengths and in the NIR (Fig 3.1). Several models have been developed to estimate the in-water constituents such as colored dissolved organic matter.
independently (Becker et al., 2009; Binding et al., 2008). However, the models do not account for the variability in the characteristics of the CPA with time and space.

**Figure 3.1.** Some examples of remote-sensing reflectance spectra from different types of waters, including waters with (a) very high sediment and CDOM concentrations, (b) high sediment and CDOM concentrations, (c) moderate sediment and CDOM with some phytoplankton, (d) clear water, (e) waters with moderate chlorophyll and sediment concentrations, (f) waters with moderate chlorophyll concentration. Lines labeled as a, b, c and e represent Case 2 waters, while d and f are signatures of Case 1 waters (IOCCG 2003).

The color of Case 2 waters is influenced by multiple constituents with distinct spectral properties that combine non-linearly. Hence remote sensing of Case 2 waters is a non-linear multivariate problem, requiring a multivariate approach to discriminate among, and estimate concentrations of, the various in-water constituents. The spectral signature of the WBLE (Fig 3.2) has similar spectral pattern with the line labeled as “a” in Figure 3.1. The spectral discrepancies observed from the reflectance spectra of the
MEdiuim Resolution Imaging Spectrometer (MERIS) clearly indicate the presence of multiple biotic and inorganic components in the WBLE (Fig 3.2). Based on this comparison, reflectance pattern from the WBLE are consistent with those measured in Case 2 water.

A powerful multivariate statistical tool that can be applied in water quality studies that have hyper-spectral data is Principal Component Analysis (PCA). This method maximizes the variances of leading factors and reduces invalid factors in the matrix data, thus reducing the complexity of multidimensionality of the matrix systems (Kaiser, 1958). This approach helps to extract relevant information that explains the optical complexity of Case 2 waters. Decomposition of hyper-spectral remote sensing data taken from filter samples in Lake Erie can be used to detect important water quality indices such as chlorophyll $a$, and estimate the concentrations across the WBLE (Ortiz et al., in press). The multivariate approach can also used to estimate total suspended matters in the WBLE.
Reflectance spectra (400-860nm) recorded by the Medium Resolution Imaging Spectrometer (MERIS) in the WBLE at pixels corresponding to 18 sampling stations. Pixel resolution is 290 by 290m. The spectra shows typical signatures of Case 2 waters with the high absorption of CDOM in the blue region (440nm), chlorophyll $a$ in the red region (665nm, red absorption) and phycocyanin in the yellow region (620nm). The peak centered at 550 nm represents effect of backscattering from suspended matter.

In this study, spectral signatures of various in-water constituents that contribute to the optical complexity of the WBLE are extracted using multivariate statistics. The main objectives of the study are to: (a) analyze the efficiency of remote sensing data to detect the various water constituents in the WBLE and (b) examine the effectiveness of multivariate regression techniques for predicting concentration chlorophyll $a$ and total suspended matters using remote sensing reflectance data.
3.2 Materials and methods

3.2.1 Study area

The Western Basin of Lake Erie, with an average depth of 10 meters, is shallow enough to have dramatic wind- and wave-driven turbidity. Its relatively warm temperature makes it conducive to high biologic productivity. A number of rivers serve as conduits for fluxes of nutrients and sediments into the WBLE, influencing water clarity, particularly near the mouths of the Sandusky and Portage rivers (Fig 3.3). In recent years, oxygen depletion and the extent of both harmful and nuisance algal blooms have increased in Lake Erie, despite a general decline in eutrophication (Budd et al., 2002; Rinta_Kanto et al., 2005). Because Lake Erie has dynamic nutrient and trophic interactions, satellites can enhance monitoring efforts. The current generation of imaging satellite sensors can synoptically monitor large areas at relatively high spatial resolution with repeat sampling intervals ranging from days to weeks, depending on the cloud cover. This study assesses the utility of using remote sensing data to detect and map chlorophyll $a$ and suspended matters in the Western Basin of Lake Erie (83 to 82.5 W and 41.2 to 41.7 N, Figure 3.3),
Figure 3. The Western Basin of Lake Erie and the locations of 18 sampling stations visited during multiple cruises of Research Vessels Gibraltar and Erie. The lake interacts with terrestrial environment through fluxes of the Sandusky, Portage, Toussaint and other rivers such as the Maumee and Detroit, which drain into the far western side of the basin (not shown). The inset shows, the Lake Erie bathymetry.
3.2.2 Data acquisition

3.2.2.1 Field and Lab data

In summer of 2009 and 2010, in-situ data and water samples were collected from the Western Basin of Lake Erie aboard the Research Vessel *Gibraltar* and *Erie Monitor* during cruises in June, July, August and September. These cruises were concurrent with ENVISAT and/or MODIS satellite overpasses. A total of 108 stations were visited between the two summer periods. Stations were located along a transect that extends between Sandusky Bay and Middle Bass Islands (Fig 3.3). Station locations were pre-selected to be at least 2 km away from terrestrial bodies to avoid contamination of satellite pixels by land.

Water quality parameters such as chlorophyll *a*, oxygen, pH, total dissolved solids, temperature and electrical conductivity were measured using HACH Hydrolab along a vertical profile at 0.1m resolution between 0-1m; measurements were taken at 0.5m resolution below 1 m depth. Secchi depth was also recorded at each station using a standard Secchi disk with alternating black and white quadrants. One-liter water samples were collected at each station and filtered through pre-weighed 0.47μm Glass Fiber Filters (GF/F) to collect the lake water residue. The samples were wrapped with aluminum foil to inhibit radiation interaction and stored in a cooler with ice in the dark. The samples were taken to laboratory the same day and the filtered residues were allowed to dry for 24 hours at 60°C in an oven. Dry particulate mass was measured with accuracy
of 0.1mg with a scale. The dry filtered samples were placed below a high intensity 20mm diameter Spectralon integrating sphere and reflectance spectra were recorded between 250 and 2500nm using lab-based Pro Full-Resolution ultraviolet/visible/near-infrared (UV/VIS/NIR) spectrophotometer. Spectra were measured at 2nm intervals in the UV-VIS range and 4-10nm intervals in the NIR range. A Spectralon reference panel was used for white referencing. Percent reflectance was transformed to Log (1/Reflectance) units. To enhance signal to noise ratios, the records for each sample was obtained by averaging 800 individual measurements of the spectrum. The absorption values were corrected for the GF/F filter background using averaged measurements of the blank filters.
3.3 Methodology

3.3.1 Multivariate Statistics

3.3.1.1 Principal Component Analysis

Principal component analysis (PCA) is a multivariate procedure that involves the transformation of a number of possibly correlated variables into a smaller number of uncorrelated indices (eigenvectors). In this remote sensing application, PCA accounts for the correlation between the signals in different spectral channels, and enhances the potential discrimination and reconstruction accuracy of retrieved constituents. In mathematical terms, PCA relies upon an eigenvector decomposition of the covariance or correlation matrix (Kim and Mueller, 1978). It seeks to establish combinations of variables capable of explaining most of the variations in a dataset. For this study, the data matrix is composed of $m$ rows, where $m$ is number of stations sampled (18 stations per cruise, total of five cruises) and $n$ columns, where $n$ is the number of spectral bands (64 bands ranging from 360 nm to 1000 nm). The Principal Components (PCs) returned by PCA represent combinations of co-varying data. These are ordered based on the percent of total variance that they explain. In most applications, a small number of PCs explain the bulk of the total variance, with remaining PCs representing noise. In PCA, eigenvalues are used to determine the number of PCs that should be interpreted and retained. Projection of the original variables on the subspace of PCs are called factor loading and represent the correlation between the PCs and the original variables (e.g. percent reflectance). Principal components are projected onto orthogonal axis with the
application of varimax-rotation to address the problem of variables loading equally on one or more of the axes (Smith, 2002; Schlens, 2005). Each PC is a linear combination of the original variables, potentially describing a different source of variation. The largest or 1\textsuperscript{st} PC is oriented in the direction of the largest variation of the original variables and passes through the center of the data. The 2\textsuperscript{nd} largest PC lies in the direction of the next largest variation, passes through the center of the data and is orthogonal to the first PC, and so forth.

PCA is commonly used in environmental application, including surface and ground water quality studies (Tauler et al., 2000; Yu et al., 1998). Works by Balsam and Deaton (1991); Deaton and Balsam (1991); Mix et al. (1995, 1999; Harris and Mix (1999) have demonstrated the utility of this method for extracting information regarding variations in sediment composition both spatially and temporally. The method works equally well with CPAs, allowing me to isolate spectral patterns that are correlated in space and time across each cruise track. In many cases, the spectral patterns identified by PCA can be interpreted physically based on known spectral signatures (Ortiz et al., 1999; Ortiz et al., 2004; Ortiz et al., 2009).

The reflectance spectra are sensitive to background noise which may alter the shape of reflectance spectra. These effects can be minimized by calculating the first derivative of the reflectance spectra (Ortiz, in press). I conducted PCA on the correlation matrix obtained from a data matrix of first-derivative-transformed Visible-Infrared (VIR) spectra. The columns of each data matrix explicitly define the 1\textsuperscript{st} derivative transform
represent wavelength in 10nm increments and the rows represent a station in one of the cruise tracks. Data from cruise tracks were combined into a single matrix to allow evaluation of the joint variation in space and time within the dataset, the data matrix contained a total of 89 rows. Data from one station for a particular cruise was removed due to instrument errors. PCA was also conducted for each cruise date individually to evaluate the temporal dynamics of the leading optical components. A varimax-rotation was applied to the component matrix extracted from the correlation matrix (Kaiser, 1960). This allows sufficient orthogonality between the derived component axes to enable spectrally distinguishing among several independent in-water constituents. Use of the correlation matrix weights each band equally within the analysis since the correlation coefficient is the cross product of the z-scores of the two bands. In this study, only factors exhibiting an eigenvalue of over 1 were retained (Kaiser, 1960). Prior to conducting the PCA calculation, a standard normal variate (SNV) transformation was applied to the data to remove a bias due to concentration differences between stations. The trend due to concentration is removed using the following mathematical expression (Barnes et al., 1989):

\[ SNV(\lambda) = \frac{(y(\lambda)-\bar{y})}{\sqrt{\frac{\sum(\chi(\lambda)-\bar{x})^2}{n-1}}} \]  

A complete PCA provides statistics that can be used to assess the percentage of variance explained by the extracted factors, identify the components represented by these, and explore their spatial and temporal variation. The eigenvalues account for the amount
of variance in the data set attributed to each factor. A plot of eigenvalues scaled as the percentage of variance explained as a function of factor rank is often referred to as a “scree plot”. It can be used to determine the number of leading factors that can be distinguished from the remaining ones or “noise floor” within the dataset (Smith, 2002; Schlens, 2005). In practice, these significant factors will account for considerably greater variance than the trailing factors within the “noise floor”.

The ability of PCA to discriminate among the multiple in-water constituents is attributed to its capability to resolve for the satellite signal variability caused by the constituents including interference from atmospheric medium. By truncating the noise factors, the atmospheric interference and stochastic errors, PCA reduces the dimensionality of the dataset to that of the significant or leading factors that represent only signals caused by in-water constituents. Theoretically, some components of atmosphere interference may come out as a leading mode of PCA. However, in practice, this does not happen because the atmospheric correction procedures applied during pre-processing of the data remove much of the systematic atmospheric interference. The communality of the PCA identifies the variance in each input variable explained by the varimax-rotated factors extracted from the data set.

The varimax factor loadings or eigenvectors define the data-adaptive filters, and thus the spectral signature of each component. The factor loadings can be plotted, in this case, as a function of wavelength. The factor scores represent the projection of the factor loadings back onto the original data matrix, providing information about the relative
importance of each factor at each location and time. The method can be thought of as a data-adaptive filtering technique because the multivariate structure of the data set - the factor loadings - form the set of basis functions through which the data matrix is filtered.

To identify the origin of the factors, I compared the resulting factor loading patterns with the center-weighted derivatives of plant pigment reflectance spectra for algal groups using spectra from Moberg et al. (2002) and Toepel et al. (2005) and mineral diffuse spectral reflectance signatures from known standards measured in our laboratory or available from version 5 of the USGS Digital Spectral Library (Clarke et al., 2003). This is analogous to X-ray diffraction (XRD) methods that use peak or whole spectrum matching methods to compare samples against known standards (Will, 2006).

3.3.1.2 Principal component regression

To accurately estimate the color producing agents in the optically complex waters of the WBLE, robust remote sensing models are required. Linear regression models can be effective, but the estimation of the least-squares may be affected by any linear relationship that exists between independent variables. This effect is known as multicollinearity and can produce erratic results including affecting stability of a model. For this study, the independent variables are assumed to be the measured spectral reflectance in 10nm bands and the dependant variable is the required water quality component e.g. chlorophyll $a$ or TSM. The linear relationship among the hyperspectral bands is evaluated using collinearity statistics. This includes determination of simple
correlation coefficients. Variables with high correlation coefficients, high variance inflation factor (VIF) and low Tolerance (T) indicates issues of collinearity. VIF is reciprocal of T, where, T = 1-R².

One approach that avoids effects of collinearity is using the results from the PCA and performing multiple linear regression analysis on the principal factors. This procedure is known as Principal Component Regression (PCR) (Draper and Smith, 1998). This is a more efficient than regression of the CPAs directly on the hyperspectral bands. In this case, the factor scores can be used in multivariate regression analyses with water quality indices, chlorophyll a and TSM. Using the uncorrelated PC scores instead of the original hyperspectral data removes effects of multicollinearity between the variables. Moreover, the dimensionality of the regressors is reduced by taking only the first few principal factors that explain most of the variance. This makes the procedure computationally more efficient. I performed stepwise selection of the three leading PCs starting with the PC resulting in the best prediction of the response variable and increased the number of PCs in each step until the quality of prediction was not improved. The coefficient of determination was used as the standard metric of the success of the prediction.
3.3.1.3 Calibration-validation

Regression models are generally used for predictive purposes and a good fitting model does not necessarily guarantee that predictions will be good, especially, when applied to data outside the calibration dataset (Martens and Dardenne, 1998). One factor that can potentially affect results of a model is dataset size. An efficient and stable model can be developed if large datasets are available. Unfortunately, in practice, the number of field observations is limited because of the high costs of monitoring. To increase confidence in prediction ability of a particular model, validation is performed.

A split sample validation method is used to perform calibration and validation and this consists of splitting available data into two samples (Mourad et al., 2005). In this study, about 66% of dataset are drawn randomly without replacement to calibrate the model and the remaining 34% are used to test the stability of the model with respect of dataset size. Statistical measures such as mean residuals, coefficient of determination and root mean square values are used as criteria to assess model stability and efficiency. Standardized coefficients (beta coefficients) are also used to assess the stability of the constructed models.
3.4 Results and discussion

Figure 3.4 represents the reflectance spectra recorded between 400 and 2500nm using lab-based spectrophotometer. The reflectance values are highly variable in the visible and NIR range. The blue-green spectral range shows reflectance patterns typically observed in turbid waters (Schalles et al., 2002; Gitelson et al., 2000; Dall’Olmo and Gitelson, 2005). These patterns reflect the effects of multiple factors including absorption by phytoplankton, CDOM, and scattering due to SM. Reflectance peaks near the red/NIR edge and absorption peaks near 620 and 670nm are apparent; these peaks confirm the presence of phytoplankton biomass. The red/NIR peaks were much higher than the green peaks for chlorophyll \( a \) concentrations > 1.5\( \mu \)g/l. The absorption minima near 550nm and the maxima near 670nm were more distinct for samples representing stations closer to Sandusky Bay.

The center-weighted first-derivative of the reflectance spectra minimizes background noise and accentuates absorption features in the visible and NIR. This reveals troughs at 410nm, peak at 460nm, a broad peak centered at 550nm, a trough at 672nm and a peak at 710nm in the visible range (Fig 3.5, inset). Hydroxyl absorption is present at 1400 and 1910nm indicating the presence of inorganic suspended sediments (Fig 3.5) (Ortiz et al., in press).
Figure 3. Reflectance spectra (400-2500nm) generated using lab-based spectrophotometer for samples collected during the four research expeditions. The absorption effect of colored dissolved organic matter (CDOM), chlorophyll $a$ and phycocyanin (PC) are observed in the visible range (400-700nm). The tough near 440nm is due to high absorption due to CDOM and phytoplankton, toughs at 620 and 675nm represent the presence of PC and chlorophyll $a$ pigment. High reflectance values near 550nm and beyond 720nm occur due to the combined effects of backscattering from suspended material and the minimum absorption by plant pigments. The three line markers indicate the green peak near 550 nm, the 620nm absorption due to PC, and the red absorption near 675nm due to chlorophyll $a$. 
Figure 3.5. Center-weighted first derivative of the reflectance spectra in figure 2.2, the derivative removes and accentuates absorption features in the visible and NIR. The NIR absorption features between 1400 and 2500nm correlate well with suspended sediment. The most prominent of these is at 1910nm. The inset to Figure 3.5 illustrates prominent features in the visible range, including absorption troughs at 440, 620 and 675nm, and a fluorescence peak near 710nm.
In the WBLE, the average in-situ concentration of chlorophyll $a$ was 5.68μg/l in June 2009 and concentrations during September period ranged between 4.96-9.64μg/l, with an average concentration of 8.07μg/l. Stations 19 and 20 contributed to the high variability during the August 2009 cruise and appeared as outliers in the dataset. Average concentrations on July 8th and 27th of 2010 were 6.05 and 7.45μg/l, respectively. In September of 2010, average chlorophyll $a$ concentration was 7.23μg/l (Fig 3.6b).

The standard deviation of the chlorophyll $a$ concentration decreases from early summer period to late summer period for both 2009 and 2010 (Table 3.1). Relatively high chlorophyll $a$ concentrations were recorded at stations, 17, 19 and 20. These stations represent the waters of Sandusky Bay which is heavily influenced by terrestrial influx via the Sandusky River that loads significant nutrients (Fig 3.6). The high standard deviation in chlorophyll $a$ during the early summer period reflects the optical difference between the waters of Sandusky Bay and the central WBLE. During the summer, prolonged periods of strong sunlight and the continuous input of nutrients allows expansion of algal blooms into the central WBLE (Becker et al., 2009). Moreover, turbulent mixing between the terrestrially influenced Sandusky Bay and the central WBLE water, coupled with wind-induced mixing homogenizes the water resulting in lower standard variation in the concentrations of the in-water constituents across the basin. The spatial and temporal variability of the concentrations of total suspended matter in the WBLE is similar to that of chlorophyll $a$. Higher than average values of TSM are recorded in early summer at the stations corresponding to Sandusky Bay relative to the stations representing the central
WBLE, indicating difference in the optical properties between the WBLE and Sandusky Bay. The standard deviation of TSM among the sampling stations decreases over the summer period (Fig 3.6b). This signals lake circulation that involved material advection and mixing throughout the WBLE.

<table>
<thead>
<tr>
<th>Cruise dates</th>
<th>Average Chlorophyll a (µg/l)</th>
<th>Standard deviation</th>
<th>Average Total Suspended Matters (TSM) (mg/l)</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>6/24/2009</td>
<td>5.68</td>
<td>1.76</td>
<td>7.46</td>
<td>5.28</td>
</tr>
<tr>
<td>9/2/2009</td>
<td>8.07</td>
<td>0.98</td>
<td>10.75</td>
<td>8.02</td>
</tr>
<tr>
<td>7/8/2010</td>
<td>6.05</td>
<td>2.716</td>
<td>8.1</td>
<td>8.07</td>
</tr>
<tr>
<td>7/28/2010</td>
<td>7.45</td>
<td>1.08</td>
<td>8.12</td>
<td>6.45</td>
</tr>
<tr>
<td>9/13/2010</td>
<td>7.23</td>
<td>0.79</td>
<td>8.08</td>
<td>4.96</td>
</tr>
</tbody>
</table>

Table 3.1. Statistics of chlorophyll $a$ and total suspended matters (TSM) in the WBLE. A low concentration of chlorophyll $a$ is obtained during early summer and the concentration increases throughout summer period. Similar pattern is observed for TSM.
Figure 3.6. Plots of concentrations of water constituents along the cruise track (a) chlorophyll $a$ concentrations in µg/l at each station and (b) gravimetrically estimated total suspended matter in mg/l at each sampling point. Relatively higher concentrations of the CPAs are recorded in the Sandusky Bay. Red horizontal bars represent average concentrations and the red vertical lines are standard variation of the concentrations.
3.5 PCA analysis of the lab based VIR derivative spectra

3.5.1 Optical properties representing the joint spatiotemporal signal variability

The center-weighed first-derivative spectra highlight fine-scale differences in the reflectance spectra, particularly in the visible region between 600 and 700nm (Fig 3.5). Application of varimax-rotated PCA to the derivative-transformed reflectance spectra for the 2009 and 2010 cruises produced three leading principal components with eigenvalues greater than one. The leading factors are associated with their empirical orthogonal factors that explain most of the joint spatiotemporal variability observed in the WBLE (42.4, 31.1 and 14.3%, respectively) (Table 3.2). The number of significant factors can also be inferred from the “scree plot” (Fig 3.7). PCs with eigenvalues less than one did not give interpretable factors and therefore were not considered in this study. The factor loading values sharply decrease within the first three principal vectors and then slowly stabilize for the remaining ones which may contain a great deal of interference stochastic error, and therefore are discarded. This procedure contributes to dimensionality reduction in the data matrix while preserving the relationship that exists in the data. The presence of multiple factors with high percentage of signal variability suggests that the optical variability observed in the Western Basin of Lake Erie is attributed to multiple, independent constituents described by the various orthogonal vectors, further confirming the categorical classification the WBLE, as Case 2 water type.
Table 3.2. Descriptive statistics of selected principal components. This represents the joint spatial and temporal optically variability across the WBLE during summer of 2009 and 2010. The first four principal components represent approximately 95% of the optical variability observed during the two summer periods.

Table 3.3 shows the results of the general and specific PCs with the leading factor loadings, the corresponding eigenvalues, which are the amount of variance, extracted by each factor, and the variance percentages (accounted for) corresponding to the principal components. The factor loading indicates the significant bands contributing to each PC. The loadings express the correlation between the hyperspectral bands and the newly formed PCs. The communality, which is the proportion of variation in each of hyperspectral band explained by the PCs, is greater than 0.85 for all the bands, indicating that most of the optical variability detected by each band is explained by the PC model. The total communality is 27.2 out of 30 and the proportion of the total optical variation explained by the three factors is 87.8%.
Table 3.3. Shows the PCAs with leading PC patterns, the corresponding factor loadings, which represent the largest amount of variance extracted by each factor and the variance percentages for each principal component.

<table>
<thead>
<tr>
<th>PCs</th>
<th>Factor</th>
<th>Loadings</th>
<th>Total Variance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.26</td>
<td>400</td>
<td>0.759</td>
<td>42.4</td>
</tr>
<tr>
<td></td>
<td>410</td>
<td>0.830</td>
<td></td>
</tr>
<tr>
<td></td>
<td>420</td>
<td>0.862</td>
<td></td>
</tr>
<tr>
<td></td>
<td>430</td>
<td>0.893</td>
<td></td>
</tr>
<tr>
<td></td>
<td>440</td>
<td>0.614</td>
<td></td>
</tr>
<tr>
<td></td>
<td>570</td>
<td>0.797</td>
<td></td>
</tr>
<tr>
<td></td>
<td>580</td>
<td>0.814</td>
<td></td>
</tr>
<tr>
<td></td>
<td>590</td>
<td>0.842</td>
<td></td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>0.838</td>
<td></td>
</tr>
<tr>
<td></td>
<td>610</td>
<td>0.662</td>
<td></td>
</tr>
<tr>
<td></td>
<td>620</td>
<td>0.517</td>
<td></td>
</tr>
<tr>
<td></td>
<td>640</td>
<td>0.521</td>
<td></td>
</tr>
<tr>
<td></td>
<td>660</td>
<td>0.421</td>
<td></td>
</tr>
<tr>
<td></td>
<td>670</td>
<td>0.511</td>
<td></td>
</tr>
<tr>
<td></td>
<td>690</td>
<td>0.580</td>
<td></td>
</tr>
<tr>
<td></td>
<td>700</td>
<td>0.631</td>
<td></td>
</tr>
<tr>
<td>6.28</td>
<td>450</td>
<td>0.889</td>
<td>31.1</td>
</tr>
<tr>
<td></td>
<td>460</td>
<td>0.819</td>
<td></td>
</tr>
<tr>
<td></td>
<td>470</td>
<td>0.935</td>
<td></td>
</tr>
<tr>
<td></td>
<td>480</td>
<td>0.901</td>
<td></td>
</tr>
<tr>
<td></td>
<td>490</td>
<td>0.892</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.726</td>
<td></td>
</tr>
<tr>
<td></td>
<td>510</td>
<td>0.574</td>
<td></td>
</tr>
<tr>
<td></td>
<td>520</td>
<td>0.459</td>
<td></td>
</tr>
<tr>
<td></td>
<td>680</td>
<td>0.418</td>
<td></td>
</tr>
<tr>
<td>1.96</td>
<td>530</td>
<td>0.338</td>
<td>14.3</td>
</tr>
<tr>
<td></td>
<td>540</td>
<td>0.464</td>
<td></td>
</tr>
<tr>
<td></td>
<td>550</td>
<td>0.557</td>
<td></td>
</tr>
<tr>
<td></td>
<td>560</td>
<td>0.725</td>
<td></td>
</tr>
</tbody>
</table>
Figure 3.7. The “scree plot” represents eigenvalues scaled as the percentage of variance explained by each factor. The first three factors (F1, F2 and F3) explain about 88% of optical variability. Factors beyond the F3 represent background noise and instrumental error, and therefore can be discarded reducing the dimensionality of the dataset.
3.5.1.1 Identification of the eigenvectors

Plots of the first three factor loadings (PCs) as a function of wavelength indicates the spectral patterns of the major in-water constituents that governed the optical variability across the WBLE during the summer of 2009 and 2010. The plot clearly indicates the critical bands contributing for each factor (Fig 3.8). Factor one is heavily influenced by the blue (400-460nm), green (560-640nm) and red region (670-680nm). The second factor represents the variance contribution from the signals in the region between 450 and 560nm, high loading to this factor also comes from 680nm. Maximum loading for the third factor comes from the green region (520-560nm).

Figure 3.8. Plots of principal factors as function of wavelength. The three eigenvectors account for approximately 87% of the optical variability observed in the WBLE during the summer season of 2009 and 2010. These plots represent the first-derivative spectral signature of the various water constituents.
Spectral matching of the derived eigenspectra as a function of wavelength with derivative-transformed reflectance spectra for known classes of in-water constituents provides a means of identifying the eigenvectors. This approach indicates that the first principal vector represents the cyanobacteria population (Fig 3.9a). The second principal vector, which describes 31.1% of the optical variability observed in the WBLE, corresponds to diatoms associated with cyanobacteria signaling the absorption features due to phycobilins (Fig 3.9b). The third factor represents a more complex mixture of phytoplankton biomass and oxy hydroxides (Fig 3.9c). The importance between the first two principal vectors flipped when performing factor analysis excluding samples from Sandusky Bay. This indicates that the Sandusky Bay is optically different from the open water of the WBLE. These results are consistent with those obtained from previous studies in the WBLE (Ortiz et al., in press).
Figure 3. 9. Factor loadings for the PCA and selected reference derivative spectra. (a) The first factor relates to cyanobacteria, (b) the second factor relates to diatoms and cyanobacteria, (c) the third factor relates to a mixture of cyanobacteria and goethite, an iron oxyhydroxide component of the suspended sediment.
3.5.1.2 Spatial variations of the CPAs

The raw hyperspectral data can be projected onto the eigenvector space, defined by the PCs, to generate factor scores. These scores are the linear combination of all of the original bands that were relevant in making the new factor loading. A plot of the factor scores, representing each vector space, against station number indicates the station’s contribution towards each principal factor. This in turn indicates the dominate type of CPAs along the cruise track (Fig 3.10).

The first component, cyanobacteria, is dominant in the Sandusky Bay, and its significance decreases towards the stations in the central WBLE. This suggests heavy presence of riverine algae discharged into the Sandusky Bay. Additionally, nutrient loading from the rivers supply the required nutrients for the cyanobacteria increasing lake productivity and causing algal blooms. In the central basin, relatively higher factor 1 scores were computed for Stations 3 and 4, both are stations closer to the outlets of the Toussaint and Portage Rivers, further suggesting the influence of terrestrially derived algal components in the lake water quality. The concentration of the cyanobacteria increases between June and July. Factor score values for the cyanobacteria are low during late summer and or early fall period. The relatively lower September temperatures coupled with reduced amount of nutrients in the water inhibit algal growth leading to shrinkage of the cyanobacteria community. Water temperatures in July were approximately 75° and by later summer decreased to 65° (NOAA, http://www.erh.noaa.gov/buf/laketemps/laketemps.php). According to the NCWQR, in
2009, during summer, concentrations of phosphorus and nitrate in Lake Erie dropped from an average of 0.092 and 10.15 mg/l, respectively, to 0.01 and 0.01 mg/l in late summer, respectively. The second factor represents the variations of the diatoms associated with blue-green algae. Factor 2 scores are generally lower in Sandusky Bay than in the central WBLE, suggesting that these are more dominant members of the community in the central WBLE.

The third component represents a complex mixture of iron bearing oxide minerals and blue-green algae and shows a general increase in its importance from the central WBLE to the Sandusky Bay. Sandusky Bay and stations 3 and 4 represent aquatic environments that are heavily influenced by terrestrial influx as a result of their close proximity to river discharge zones (Fig 3.3). These environments showed higher values in the factor 3 scores which suggests that the third PCA component represents optical variability attributed to the terrestrially derived suspended matters. Factor 3 shows a general decreasing trend over the summer period which may be associated with the decrease in discharge rates of the Sandusky River during the same period. Sandusky River discharge decreases from approximately 1000 cubic feet per sec in early spring of 2010 to 40 cubic feet per sec in late summer of 2010 (USGS, http://waterdata.usgs.gov). The relatively higher factor scores in the Sandusky Bay for July 2010 cruise data is a result of high discharge in the days preceding the cruise date. Figure 4.4 in section 4.4.1 presents the Sandusky River discharge for 2009 and 2010.
Figure 3. 10. Factors scores from the PCA model. The importance of each component at each station is proportional to its magnitude; the sample locations are as defined in Figure 2.3. The data is presented in chronologic order with the earliest cruise on the left.
3.5.2 Optical properties representing the temporal signal variability

In order to assess the temporal dynamics of CPAs, the spectral reflectance matrix data was divided into specific cruise periods representing early and late summer. In 2009, cruise was made in late June and early September. In 2010, cruise was made in mid July and early September. Varimax-principal component analyses were then performed for each of the four cruise periods. Factor loadings are plotted as a function of wavelength, indicating the spectral pattern of the principal in-water constituents that governed the optical variability across the WBLE during these specific periods. As discussed earlier, spectral matching was determined based on the reference library spectra (Fig 3.11 and 3.12).

PCA of the 2009 first-derivative transformed reflectance data indicates that maximum variance occurs due to the presence of blue green algae (Figs 3.11a and 3.12a). Factor 2 represents the diatoms which form the phytoplankton community in the central WBLE (Fig 3.11b). The third factor indicates optically complex signal that relates mixture of terrigenous oxide minerals and riverine cyanobacteria (Fig 3.11c). Varimax-PC analysis of the 2010 optical data gave eigenvectors similar to the 2009 data. The third PC component for September of 2009 lack signatures of inorganic sediments (Figs 3.11f). This is possibly the result of the low discharge rate of the Sandusky River during September of 2009 (< 100 ft³/sec, http://waterdata.usgs.gov) leading to low volume of material influx, and the settlement of the suspended clays to the benthic environment.
Figure 3. 11. Factor loadings and selected reference derivative spectra for 2009. The first factor relates mainly to cyanobacteria (a, d), the second factor relates to diatoms (b, d) and the third factor relates to a mixture of cyanobacteria and goethite (c), an iron oxyhydroxide component of the suspended sediment. The third factor represents predominantly blue-green algae (f).
Figure 3.12. Factor loadings and selected reference derivative spectra for 2010. The first factor relates to cyanobacteria (a, d), the second factor relates to diatoms (b, d) and the third factor relates to a mixture of cyanobacteria and goethite (c), an iron oxyhydroxide component of the suspended sediment. The third factor for the September period represents predominantly algal components but signatures of iron bearing minerals are observed from the peak near the 550nm (f).
Close observation of Figs 3.11a and d shows that optical properties of the WBLE changed throughout summer of 2009. In early summer, the optical property of the WBLE was primarily controlled by cyanobacteria, however, by late summer (September) goethite became equally important and showed up in PC one. Spectral patterns for 2010 (Fig 3.12a and d) shows that cyanobacteria remained as the primary optical constituent throughout the summer period. Records of 2010 Sandusky River discharge shows that the discharge was significantly high compared to 2009 and fluxes of high discharge rates remained until August of 2010. Higher and persistent river flux allows for continuous loading of riverine algae community and nutrients, increasing the productivity of the lake and hence phytoplankton play major role in defining the optical properties of the WBLE.
3.6 Prediction of CPAs

One of the major strengths of performing PC regression as opposed to using least-squares methods is it removes the effects of collinearity that exist between the spectral bands. An important index to justify the removal of collinearity effects between the PCA-based independent variables, i.e., the factor scores, is their correlation. Table 3.4 shows that the correlation coefficients between all PCs are 0. PC regression models used for the prediction of chlorophyll $a$ and TSM have eigenvalues and condition indices close to 1 (Tables 3.5 and 3.6). All of these collinearity statistics suggest that all principal components are independent of each other. A general equation that shows the form of the regression model that is used is given by:

$$\hat{Y}_i = \sum \beta_i F_i$$ \hspace{1cm} 2.2

Where, $\hat{Y}_i$ is predicted value, $\beta_i$ is the regression coefficients and $F_i$ = factors

<table>
<thead>
<tr>
<th>Correlations coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor1</td>
</tr>
<tr>
<td>Factor1</td>
</tr>
<tr>
<td>Factor2</td>
</tr>
<tr>
<td>Factor3</td>
</tr>
</tbody>
</table>

Table 3. 4. Correlation analysis between factor scores
<table>
<thead>
<tr>
<th>Model</th>
<th>Dimension</th>
<th>Eigenvalue</th>
<th>Condition Index</th>
<th>Variance Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Constant)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1.05</td>
<td>1.00</td>
<td>.47</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.95</td>
<td>1.05</td>
<td>.53</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1.19</td>
<td>1.00</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.02</td>
<td>1.08</td>
<td>.89</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.78</td>
<td>1.24</td>
<td>.11</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1.57</td>
<td>1.00</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.19</td>
<td>1.15</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.02</td>
<td>1.24</td>
<td>.89</td>
</tr>
</tbody>
</table>

a) Dependent Variable: chlorophyll \(a\)

Table 3. 5. Collinearity indices of three standardized principal component regression models for chlorophyll \(a\) prediction.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dimension</th>
<th>Eigenvalue</th>
<th>Condition Index</th>
<th>Variance Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Constant)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1.04</td>
<td>1.02</td>
<td>.50</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.02</td>
<td>0.95</td>
<td>.50</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1.20</td>
<td>1.00</td>
<td>.14</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.01</td>
<td>1.10</td>
<td>.81</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.86</td>
<td>1.23</td>
<td>.06</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1.14</td>
<td>1.00</td>
<td>.14</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.24</td>
<td>1.12</td>
<td>.21</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.02</td>
<td>1.21</td>
<td>.59</td>
</tr>
</tbody>
</table>

a) Dependent Variable: TSM

Table 3. 6. Collinearity indices of three standardized principal component regression models for TSM prediction.
3.6.1 PC regression against chlorophyll $a$ levels

The score values obtained from PCA are used as independent variables in the stepwise multiple linear regression analysis to determine the most significant PCs for predicting the concentrations of chlorophyll $a$. Table 3.7 indicates the zero collinearity of the independent variables using statistical measures such as the tolerances and VIFs, both of which are equal to 1.

Factors that significantly increase the coefficient of determination are included in the model. Accordingly, scores 1, 2 and 3 were found to have significant linear relationship with chlorophyll $a$ (Table 3.7). The regression statistics shows that 70% of the chlorophyll $a$ variation is explained by the linear combination of factor scores 1, 2 and 3. Addition of 4<sup>th</sup> PC scores in the model raised the coefficient of determination to 72%. However, this difference is not statistically significant, $P > 0.05$. Figure 3.13 shows the regression between predicted chlorophyll $a$ values from PC regression model and the chlorophyll $a$ concentrations. The model explains approximately 70% of chlorophyll $a$ variation in the WBLE, RMSE =0.81μg/l.
## Table 3.7

Regression analysis of PC scores against chlorophyll *a* (n=89). The first three principal component account for 87.8% of the optical variability in the WBLE during the summer period of 2009 and 2010. The regression model explains about 70% of the chlorophyll *a* variation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Regression Coefficients</th>
<th>t</th>
<th>P</th>
<th>Collinearity Statistics</th>
<th>$R^2$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Tolerance</td>
<td>VIF</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>6.98</td>
<td>.14</td>
<td>50.35</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Factor3</td>
<td>1.41</td>
<td>.15</td>
<td>9.09</td>
<td>.00</td>
</tr>
<tr>
<td>2</td>
<td>(Constant)</td>
<td>7.01</td>
<td>.12</td>
<td>60.14</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Factor3</td>
<td>1.57</td>
<td>.13</td>
<td>11.82</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Factor1</td>
<td>-.79</td>
<td>.13</td>
<td>-6.09</td>
<td>.00</td>
</tr>
<tr>
<td>3</td>
<td>(Constant)</td>
<td>7.12</td>
<td>.11</td>
<td>64.72</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Factor3</td>
<td>2.03</td>
<td>.17</td>
<td>12.32</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Factor1</td>
<td>-1.16</td>
<td>.17</td>
<td>-7.80</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Factor2</td>
<td>.95</td>
<td>.23</td>
<td>4.17</td>
<td>.00</td>
</tr>
</tbody>
</table>
Figure 3. 13. Regression analysis between predicted chlorophyll $a$ levels based on model 3 and in-situ chlorophyll $a$ values.

A total increase in chlorophyll $a$ level would lead to a decrease in significant variables of score 1, namely, bands 400, 410, 420, 440, 580, 610, 660 and 670 nm. On the other hand, chlorophyll $a$ abundance would cause an increase in reflectance values at 460, 470, 490, 500, 510, 530, 540, 550, 560 and 680 nm. This is consistent with chlorophyll $a$'s spectral features, i.e. absorption in the Soret bands and the red absorption near 670 nm. The positive relation between chlorophyll $a$ abundance and the backscattering recorded in the range of 490 to 560 nm is due to minimum absorption capacity of the pigment and backscattering effects from cell walls.
3.6.2 PC regression against concentrations of TSM

Multiple linear regression analysis is performed between the independent variables, PC scores, and dependent variable, TSM. The first three factors are found to have significant correlation with the TSM, \( P < 0.05 \) (Table 3.8). Approximately, 75\% of the TSM variability is accounted by the linear multivariate regression analysis of the first three principal factors (Fig 3.14), \( R^2 = 0.754 \) and \( \text{RMSE} = 2.861 \text{mg/l} \).

<table>
<thead>
<tr>
<th>Model</th>
<th>Regression Coefficients</th>
<th>T</th>
<th>P</th>
<th>Collinearity Statistics</th>
<th>( R^2 (%) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Const) 8.50 .58</td>
<td>14.79 .00</td>
<td></td>
<td>1.00 1.00</td>
<td>33.4</td>
</tr>
<tr>
<td></td>
<td>Factor3 3.84 .58</td>
<td>6.64 .00</td>
<td></td>
<td>1.00 1.00</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(Const) 8.50 .45</td>
<td>19.04 .00</td>
<td></td>
<td>1.00 1.00</td>
<td>60.2</td>
</tr>
<tr>
<td></td>
<td>Factor3 3.84 .45</td>
<td>8.54 .00</td>
<td></td>
<td>1.00 1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Factor2 -3.44 .46</td>
<td>-7.66 .00</td>
<td></td>
<td>1.00 1.00</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>(Const) 8.50 .35</td>
<td>24.07 .00</td>
<td></td>
<td>1.00 1.00</td>
<td>75.4</td>
</tr>
<tr>
<td></td>
<td>Factor3 3.84 .35</td>
<td>10.80 .00</td>
<td></td>
<td>1.00 1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Factor2 -3.44 .36</td>
<td>-9.68 .00</td>
<td></td>
<td>1.00 1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Factor1 -2.58 .35</td>
<td>-7.28 .00</td>
<td></td>
<td>1.00 1.00</td>
<td></td>
</tr>
</tbody>
</table>

Dependent Variable: TSM

**Table 3.8.** Regression analysis of PC scores against TSM (n=89). The first three principal components explain about 75\% of the TSM variation.
The TSM variability in the WBLE is explained primarily by factor 3 and the coefficient of determination increased with the addition of factor 2 and 1. The PCs contribute significantly to the performance of the model with all three factors having correlation coefficients statistically different from zero. TSM has a positive impact on factor 3, while it possesses a negative correlation with PC Scores 2 and 1. In other words, TSM levels would be expected to increase the values of Score 3, which refers to the reflectance level recorded at 540, 550 and 560 nm. The increase in reflectance in the green region due to suspended matter is corroborated in many previous works (Ortiz et al., in press; Gordon and Morel, 1983). Regression analysis between PCs and TSM shows that factors 2 and 1 have negative coefficients (Table 3.8). Factor 1 represents the Soret bands.
and the red region. Factor 2 represents mainly the spectral regions 440-510nm. Figure 3.14 illustrates the regression is strongly controlled by the outliers. These high value data represent concentrations of the TSM in the turbid waters of Sandusky Bay, and therefore are considered to be valid measurements that should be incorporated in the model.

The presence of multiple constituents in the total suspended matter in the WBLE causes various levels of absorption and scattering across the spectrum. The inverse relationship between the reflectance near the Sort bands and TSM signals the presence of multiple constituents including CDOM which may be bound to filtered particles as suggested in the works of Binding et al. (2008). TSM also showed inverse relation with bands near 670 nm, signaling the presence of phytoplankton component.
3.6.3 Calibration-validation

Calibration-validation procedures were applied to the PCR models for chlorophyll \(a\) and TSM prediction. Table 3.9 shows that the validation parameters (mean residuals (MRES), \(R^2\) and RMSE) for both chlorophyll \(a\) and TSM are only slightly worse than the calibration statistical parameters. Parameters from validation datasets (28 dataset) were expected to be worse than those attained from the calibration dataset (68 dataset) because the validation data were not included in the development of the model. The beta coefficients in the calibration and validation models are within the error margin for the chlorophyll \(a\). The fact that all of the validation statistics have similar values to that of calibration dataset demonstrated that the model is stable and it is not affected by dataset size making the model more robust.

<table>
<thead>
<tr>
<th>PCA</th>
<th>Data set-Chlorophyll (a)</th>
<th>N</th>
<th>MRES</th>
<th>(R^2)</th>
<th>RMSE</th>
<th>Beta Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>62</td>
<td>0.02</td>
<td>0.70</td>
<td>0.98</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td>28</td>
<td>0.86</td>
<td>0.71</td>
<td>2.041</td>
<td>0.64</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data set-TSM</th>
<th>Training</th>
<th>N</th>
<th>MRES</th>
<th>(R^2)</th>
<th>RMSE</th>
<th>Beta Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>62</td>
<td>0.09</td>
<td>0.77</td>
<td>2.89</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td>28</td>
<td>-0.39</td>
<td>0.69</td>
<td>4.17</td>
<td>0.63</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.9. Summary of the validation parameters computed using the calibration and validation dataset for the PCs and two CPAs (chlorophyll \(a\) and TSM).
3.7 Conclusion

Multivariate decomposition of reflectance measured from filtered sample using a lab-based spectrophotometer clearly indicated the presence of multiple factors that affect the optical characteristics of the WBLE, namely, cyanobacteria, diatoms and inorganic minerals. This approach is efficient to: a) reduce number of the variables in multiple regression models, b) remove scattering effect of variables and c) eliminate multicollinearity problems. The consistency and reproducibility of the extracted Lake Erie water constituents from the PCA analysis, both in the case of the joint spatiotemporal data and temporally specific data, indicates that the multivariate approach is stable method that may be applied efficiently in Case 2 waters.

Various models have been applied to explain the biogeophysical characteristics of Case 2 waters from remote sensing data. (Witter et al., 2009; Moses et al., 2008; Simis et al., 2005; Dekker, 1993; Gitelson, 1992; Gitelson et al., 2009; Gitelson et al., 2008; Gons et al., 2002; Han et al., 1994; Gower et al., 1999, 2005; Moore, 1980; O’Reilly et al., 1998; Ruddick et al., 2003; Schalles, 2006). The successes of bio-optical models have been commonly described using simple statistical criteria such as the coefficient of determination and the root mean square error of the predictions. A number of studies have defined the relations of reflectance with biogeophysical factors as univariate and others have adapted a multivariate linear or nonlinear approach. Recent studies show that multivariate models are capable of assessing large number of variables and interrelations
and therefore, more successful in defining and predicting complex biogeophysical processes.

In the present study the relationships between the first derivative hyper-spectral bands with chlorophyll $a$ and total suspended matters have been investigated using a linear multivariate approach. This model has been able to predict concentration of chlorophyll $a$ and total suspended matters in the Western Basin of Lake Erie with a maximum predictive success of 70% and 75%, respectively, using approximately 88% of the variation in the optical data. Calibration-validation procedures clearly demonstrated that the PCA based models are stable and the results are not affected by dataset size.
CHAPTER 4

Retrieval of chlorophyll $a$, phycocyanin and total suspended matter from hyperspectral reflectance data using discrete wavelet transformation.
Abstract

Remote sensing has become very promising in providing temporal and spatial information regarding biogeodynamics in large and open freshwater bodies. However, in optically complex environments, such as in the Western Basin of Lake Erie (WBLE), the water contains multiple biogeochemical constituents or Color Producing Agents (CPAs) such as, phytoplankton, sediment, and dissolved organic carbon. Identifying and analyzing in-water constituents in these waters is crucial for understanding and assessing many biogeochemical processes. For example, concentrations of chlorophyll $a$ and suspended matter can be used as proxies to assess phytoplankton dynamics and particulate loading. In aquatic environments such as the WBLE, where *Microcystis* is abundant, phycocyanin constitutes a major photosynthetic accessory pigment that can serve as an important proxy to detect the dynamics of the cyanobacteria. However, quantitatively estimating concentrations of in-water constituents from satellite observations is complicated when working with mixed spectral signatures.

This study focuses on improving quantification of chlorophyll $a$, phycocyanin and Total Suspended Matter (TSM) by applying wavelet analysis to hyperspectral reflectance data derived from a lab-based spectrophotometer. Wavelet analysis of a reflectance spectrum is performed by scaling and shifting wavelet functions to produce wavelet coefficients that are assigned to different frequency components. By selecting appropriate wavelet coefficients, a spectral model can be established between wavelet coefficients and concentrations of biogeochemical components. Regression analysis between the
proxies and derived wavelet coefficients show coefficients of variation of up to 0.89. The high predictive capability of the regression models is due to the localization property of wavelets and to the extraction of critical frequency components that are highly sensitive to selected CPA proxies. In this work it is demonstrated that wavelet analysis has the potential to discriminate between various water constituents in complex signals, making this a robust tool for predicting multiple constituents in optically complex waters.

**Keywords:** chlorophyll a, hyperspectral, phycocyanin, wavelets.
4.1 Introduction

Case 2 waters such as the Western Basin of Lake Erie (WBLE) are often characterized by high concentrations of suspended and dissolved organic matter mostly derived from discharge of particle laden rivers. River influx directly affects many water column processes such as phytoplankton productivity and nutrient dynamics (May et al., 2003). The distribution and flux of in-water constituents are highly variable in lake environments and vary over broad scales in time and space. This variability renders many traditional field sampling methods inadequate in studies that require resolving water quality dynamics in complex waters such as the WBLE.

Satellite remote sensing provides a synoptic view of in-water constituents and an opportunity to conduct longer-term and more consistent monitoring of water quality parameters (Witter et al., 2009). Sensors with a wide variety of spectral, spatial and temporal resolution have been used to evaluate water quality parameters (e.g., Godin et al., 1993, Kutser et al., 2005). Major factors affecting water quality in Case 2 waters include phytoplankton, dissolved organic matter and suspended particulate concentrations. These factors change the visible and infrared reflectance spectra from surface waters, which can be measured using remote sensing techniques (Morel and Prieur, 1977; Doerffer and Fischer, 1994).

Case 2 waters have posed challenges to remote sensing application because of the presence of multiple optically active constituents (Morel and Gorden, 1983).
Discrimination between the various constituents requires separation of the mostly overlapping scattering and absorption properties of water and its constituents. Many empirical and semi-empirical techniques have been employed to retrieve CPAs independently using band ratios based on conveniently selected spectral bands that contain relevant information (e.g., Gons, 1999; Mayo et al., 1995, Kutser et al., 2004, 2005; Gitelson, et al., 2008, 2009). However, a recent review by Blackburn (2006) indicates the lack of consensus on optimal band selection when using conventional spectral indices. Moreover, spectral indices tend to focus on a small number of bands while discarding the majority of the spectrum, and, potentially, opportunities for distinguishing more effectively among CPAs. Other approaches such as Fourier transformation failed to represent local variability as well as trends, drift and abrupt changes in the constituents (Ovanesova et al., 2004). Fourier-based transformation only retains frequency information, and loses the spectral domain, and therefore it is not possible to determine the optimal spectral bands contributing to the various frequency components that may in-turn represent signatures of CPAs. Another even bigger issue with the traditional Fourier transformation is the lack of localization i.e., to zoom in on any spectral interval.

One particular technique that holds considerable promise that overcomes the deficiencies in the analysis of hyperspectral reflectance data is wavelet analysis. The wavelet transformation is capable of retaining both, the frequency and spectral domains, enhancing sensitivity of the retrievals to the various CPAs.
Using the Discrete Wavelet Transform (DWT) algorithm developed by Mallat (1999), wavelet analysis has the capability to project signals onto a basis of wavelet functions that can capture the high frequency and low frequency information of the multispectral signal independently. This can be used as a feature extraction method which would help identify emission or absorption features at various scales.

The objective of this work is to evaluate the robustness of wavelet techniques in detecting the various optical features that result due to interaction between incident radiance and in-water constituents. The specific objective is to determine the wavelet basis functions that provide wavelet coefficients that are the most sensitive to variation in the concentrations of selected CPAs.

4.2 Wavelet background

Wavelets are mathematical functions that express an irregular waveform with limited duration making them ideal to describe anomalies within finite events. Unlike Fourier analysis, in which I analyze signals using infinitely oscillating sine and cosine basis functions, I use locally oscillating finite “wavelet” functions to decompose data into different frequency components, where each component is characterized with a resolution appropriate to its scale (Mallat, 1999; Strang & Nguyen, 1996, Fig 4.1).

For hyperspectral signals, wavelet analysis represents the decomposed spectral signal as localized in the wavelength and frequency domain. However, Fourier analysis only represents frequency information (Strang & Nguyen, 1996). The basis functions of
wavelet analysis are called wavelets and the original operator is referred to as the mother wavelet. A necessary condition for a function to be a valid wavelet function is:

$$\int_{-\infty}^{\infty} \psi(\lambda) d\lambda = 0 \quad 4.1$$

Where, $\psi(\lambda)$ is the basis wavelet function which oscillates with a zero average value. There are numerous wavelet functions that can be used to perform signal processing and the selection of the optimal wavelet function can be a challenge (Fugal, 2008). Generally, there are two major categories of wavelets functions; continuous wavelets and discrete wavelets. These wave functions differ based on their regularity, orthogonality, effective support (range), and symmetry. The continuous wavelets are regular (smooth), symmetrical with wide range support (beginning and ending points are further apart), and are used for continuous signals. They have only one stretchable filter, therefore are not orthogonal. The discrete wavelets are more irregular; they have compact support (smaller range) and can be asymmetrically oscillating waves. In this study, since the input hyperspectral reflectance data is restricted to discrete values, in the selected wavelength bands, only DWTs are applied. This requires orthogonal wavelets functions that are applicable to perform a DWT. Unlike continuous wavelet functions, the discrete wavelet functions have two filters (the scaling and the wavelet function filters) that form an orthogonal basis to decompose the signal, and therefore the frequency information carried by the wavelet coefficients retrieved using the two filters are independent (Fig 4.1). Popular wavelet functions, such as the Mexican hat and the Morlet do not have the scaling functions, therefore cannot be used to perform DWT.
A few examples of orthogonal wavelet systems that can be used to decompose hyperspectral reflectance data from the WBLE are shown in Figure 4.1. The Daubechies (DbN) are the most widely used orthogonal wavelet functions in the field of signal processing. These waves are irregular, asymmetric and have compact support. Here, N represents the number of non-zero points used to represent the wavelet function, e.g., Db4 has 4 points and Db6 is estimated using 6 points. An increase in N smooths the wavelet function and makes it more continuous; this offers better frequency resolution and increases the wavelet's ability to explain polynomial behavior in a signal at the expense of decreased spectral resolution. I have to determine the optimal N values to be used to capture the required frequencies without the loss of spectral resolution. The DbN’s are well-suited for identifying irregular and asymmetrical features within signals (Fugal, 2008).

The Symlet (SymN) are the second type of wavelets used in this study. They have compact support and are more symmetrical than the Daubechies. The N factor also applies to the Symlets. The third type of wavelet function used belongs to the Coiflet’s family (CoifN). The Coiflets have even higher degree of symmetry and thus avoid dephasing in signal processing that may help to detect symmetrical signals. Furthermore, they have orthogonality property which makes them suitable for extracting independent optical features from discrete hyperspectral signals.
Figure 4.1. Examples of wavelet function filters used to decompose the hyperspectral signal in this study: Db = Daubechies, Sym = Symlet and Coif = Coiflet. Note that all have the scaling function.
Once a mother wavelet is selected, the wavelet analysis is performed using scaled (dilated) and shifted versions of the mother wavelets, $\psi (\lambda)$, to produce a set of basis wavelet functions $\psi_{s,k} (\lambda)$ defined as:

$$\psi_{s,k} (\lambda) = \frac{1}{\sqrt{s}} \psi\left(\frac{\lambda - k}{s}\right),$$

where $s$ and $k$ represent the scaling and the spectral shifting parameters, respectively.

The various wavelet basis functions are shifted smoothly to different positions over the full domain of the signal. With their ability to stretch and translate, wavelets are extremely adaptable, allowing a multiple resolution analysis to separate the fine-scale and large-scale properties of a hyperspectral signal. To show the robustness of the discrete wavelet transformation, it is necessary to illustrate the continuous wavelet transformation. The continuous wavelet transform of reflectance spectra $f(\lambda)$ is given by:

$$W_f (s, k) = \int_{-\infty}^{\infty} f(\lambda) \frac{1}{\sqrt{s}} \psi\left(\frac{\lambda - k}{s}\right) d(\lambda),$$

where $s$ and $k$ are real numbers, wavelet coefficients $W_f (s, k)$ are continuous and the transformation can operate at every scale, from that of the original signal up to some maximum scale which is theoretically determined by the record length. In practice this is often truncated at smaller scales as a trade off for the need for detailed analysis against available computational power. The wavelet coefficients that are produced represent the correlation between the wavelet and the original signal at different scales for different sections of the signal. Calculating wavelet coefficients at every possible...
scale is computationally intensive and can generate a significant amount of redundant data. By choosing a subset of scales and positions, the analysis becomes much more computationally efficient (Blackburn, 2006).

### 4.2.1 Discrete Wavelet Transforms

In a finite, irregular and discrete dataset such as the hyperspectral reflectance data generated for this study, an efficient and more accurate technique for extracting the various optical features within the integral spectra is the dyadic DWT. This technique provides sufficient information for both the analysis and the synthesis of the original hyperspectral signal, with a significant reduction in the computation time. In DWT, the mother wavelet is dilated by a factor of two (dyadically) and the fixed scaled wavelet function is shifted over the range of the signal. The scaling and shifting of the wavelets allow the identification of the high frequency and low frequency components of the hyperspectral signal (Fig 4.2).

![Wavelet Transform Diagram](image)

**Figure 4.2.** Signal decomposition using wavelets that have been scaled (stretched) and shifted. The procedure involved comparing multiple levels of dilated wavelets to the original hyperspectral signal across the wavelength domain.
For the hyperspectral reflectance data, the DWT is computed separately for the different segments of the reflectance spectrum at various frequencies. Wavelet coefficients are calculated as the inner products between a segment of the measured spectrum and the discretized wavelet functions. Each inner product results in one wavelet coefficient. A one-dimensional DWT is mathematically expressed as:

\[ W_{p,q} = f(\lambda), \phi_{p,q}(\lambda), \]  

4.4

Where, \( W_{p,q} \) is used to denote the coefficients of the DWT, \( f(\lambda) \) is the hyperspectral signal, and \( \phi_{p,q} \) is the DWT scaling function, i.e., the scaled and shifted version of the mother wavelet. The scaling function is defined by:

\[ \phi_{p,q}(\lambda) = \frac{1}{2^{p/2}} \phi(2^{-p}\lambda - q) \]  

4.5

Here, \( p \) represents the decomposition level, indicating the wavelet’s width and \( q \) is the wavelet coefficient at the \( p \)'s level. The \( 2^p \) is the dyadic scaling factor and the translation parameter in the DWT takes values \( k=q2^p \), where, \( k \) is the translation factor in CWT.

According to Fugal (2008) and Pu et al. (2004), instead of continuously stretching the wavelet function dyadically, the signal can, in practice, be decimated by a factor of 2 while keeping the wavelet function fixed. The decimated signal can then be convolved with the impulse response of a selected filter. This method is known as the Fast Wavelet Transform (FWT) algorithm. The algorithm uses a wavelet filter \( \psi_{p,q} \) and a scaling
function filter $\phi_{p,q}$ to perform a high-pass and low-pass filter on the hyperspectral reflectance data, respectively. The wavelet filter generates wavelet coefficients that represent the small-scale or high-frequency components of the signal. Filtering the signal through the scaling function generates large-scale or low-frequency coefficients. After decomposing the signal through a half-band filter, since the half of the frequency is filtered out, the scale of the signal doubles and according to the Nyquist rule this can be adjusted by discarding half of the samples without the loss of the signal information (Mallat, 1999). The maximum level of decomposition is $p=\log_2 (N)$, where $N$ is the length of the hyperspectral reflectance record ($N = 64$ bands for this study). In this study, our signal consists of reflectance in bands 360 to 1000nm averaged at 10nm, as this is the spectral range where electromagnetic radiation is most sensitive to properties of the CPAs of interest. The output of the low-pass filters are called approximation coefficients, $cA$, and the residuals of the high pass filter are called detail coefficients, $cD$.

\[
\begin{align*}
    cA(p) &= \sum_n f[\lambda] \cdot \phi_{p,q}[\lambda] \\
    cD(p) &= \sum_n f[\lambda] \cdot \psi_{p,q}[\lambda]
\end{align*}
\]

In the context of signal processing, several studies (Chen and Wang, 2001; Mittermayr et al., 2001; Varotsos et al., 2003) have demonstrated the robustness of wavelet analysis in distinguishing signals that have similar appearance but are emitted from different dynamic features. Wavelets are relatively insensitive to background spectral variations, and are capable of removing noise and reducing the dimensionality of
hyperspectral signal without losing the signal’s fine spectral features (Bruce et al., 2001; Cai et al., 2001). Hsu (2007) demonstrated the capability of wavelet analysis in extracting relevant optical features from NASA’s Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) for vegetation mapping and the superiority of its prediction models. The capability to discriminate among various features in complex signals coupled with efficiency available from dimensional reduction makes the wavelet technique an ideal candidate for retrieving CPAs in optically complex waters from hyperspectral remote sensing data.

4.3 Materials and Methods

4.3.1 Lake Erie

The Western Basin of Lake Erie is an optically complex environment characterized by diverse and inhomogeneous distributions of in-water constituents (Witter et al., 2009). The WBLE is the shallowest sub-basin of Lake Erie that is heavily influenced by the terrestrial environment, with a number of rivers discharging suspended and dissolved material. Re-suspension of bed loads (Marvin et al., 2007) coupled with loading from the terrestrial environment has resulted in biological productivity and turbidity that is the highest within the Great Lakes. In addition, in recent years, toxic algal blooms (Microcystis and Lyngbya Wollei) have been documented in the WBLE (Becker et al., 2009; Bridgmann and Penamon, 2010) providing additional motivation for developing new techniques for monitoring water quality in the WBLE.
4.3.2 Data acquisition

In summer of 2009 and 2010, in-situ data and samples were collected from the WBLE aboard the Research Vessel Gibraltar and Erie during 5 cruises. A total of 18 stations were visited across the WBLE on 6/24/09, 9/02/09, 7/08/2010, 7/27/2010 and 9/13/2010. Sampling locations were stored in a marine GPS to verify re-visit of the same site during the multiple cruises. In-situ measurements included concentrations of multiple water quality parameters as well as the time and date of collection. Water quality parameters, such as chlorophyll \(a\), phycocyanin, and dissolved oxygen concentrations, turbidity, pH, total dissolved solids, electrical conductivity and temperature were measured in-situ using the HACH Hydrolab along a vertical profile with 0.1m resolution. One liter samples were collected at each station from 0.2m depth and were filtered onto a 0.45\(\mu\)m pore size glass-fiber filter (GF/F) filter to collect the suspended matter. These samples were stored in a cooler with ice in the dark, to avoid material degradation, and taken to the lab for analysis the next day. Secchi depths were measured on the shaded side of a vessel using a standard 20 cm diameter Secchi disk with alternating black and white quadrants.
4.3.3 Laboratory analysis

Spectroscopy was performed on the filter residue within 24 hours using a lab-based spectrometer which has a spectral range of 250–2500nm. The spectrophotometer has 2nm VIS resolution; 4-10nm NIR resolution and is equipped with a high-intensity contact probe. The instrument was calibrated using a white Spectralone reference panel. Eight hundred reflectance measurements were made on each filter and averaged to increase the signal to noise ratio. The hyperspectral reflectance data was corrected for the filter background by dividing each of the sample spectra by the reflectance spectrum of a blank GF/F. In order to reduce data redundancy and increase computational efficiency, each spectrum was averaged to 10nm. The concentration of the total suspended matter was determined gravimetrically by placing the dried filter residues on a highly sensitive analytical balance with 0.1mg readability.

4.3.4 Multi-resolution hyperspectral data analysis

Hyperspectral reflectance data from aquatic samples represents a convolution of contributions to the spectrum from various optically active components. For this study a centered-weighed first-derivative (i.e., high-pass) filter is applied to the original reflectance spectrum (Fig 4.4, section 4.4). This reduces the sensitivity of the spectrum to variations in illumination intensity and possible stochastic errors. After this operation, any features observed in the spectrum are more likely to be due to the CPAs. Moreover, application of this high-pass filter can enhance smaller peaks that are obscured by larger
peaks due to noise. This procedure can be used to facilitate the preliminarily identification of key spectral features such as the chlorophyll $a$ absorption peak (Fig 4.5, section 4.4). However, the precise identification and extraction of the various optically active components from a reflectance signal requires further signal transformation. In this study, Multiple Resolution hyperspectral Analysis (MRA), which separates the fine-scale and large-scale properties of an input signal, is used as a feature extraction method. When applied to the wavelet transform of the reflectance signal, this method can be used to identify emission or absorption features that exist at various scales.

In this study, a dyadic filter tree algorithm that implements the FWT is applied to the hyperspectral reflectance data (Fig 4.3). The algorithm uses the first-derivative hyperspectral signal as input data, and the signal is then convolved through a series of orthogonal wavelet functions, including high-pass (H) and low-pass (L) filters. Results from wavelet analyses of the first-derivative filtered data are compared with the wavelet transformed reflectance spectra, which had not been treated with the first-derivative filter (see section 4.4.2 below). The input spectra are measured using an UV/VIS/NIR spectrophotometer. After each low-pass filter the signal is decimated by 2 (Morales and Shih, 2000). This filtering removes half of the frequencies, and therefore the decimation-process will reduce the signal to half the original number of points. Note that due to successive subsampling by 2, the record length must be a power of 2. The spectrum used for the wavelet analyses consists of 64 bands at 10nm intervals covering the range between 360 and 1000nm. This spectral range includes prominent spectral features from the various water quality parameters of interest to this study. Information contained
within the radiance beyond 1000nm is mostly confounded by noise factors. The input signal for this study has its derivate centered at 64 bands, and therefore 5 levels of decomposition are carried out. Due to low signal to noise ratio outside the visible and near-infrared range, all graphical representations of the wavelet analyses are between 400 and 1000nm range.

The DWT analysis was done using the wavelet tool in the IDL software package, within which a number of different families of wavelets are available. For this study, compactly supported orthogonal wavelets are implemented to apply the FWT, and visual assessment was used to determine the correlation between the shape of the hyperspectral signal and the various choices of orthogonal wavelet. Based on this criterion, three wavelet functions were selected. These include the Daubechies with order number ranging from 4 to 9, the Symlets (order number 4 to 9) and the Coiflets (order number 1 to 5). These wavelets were then assessed quantitatively using predictive models to identify the best performing wavelet function in this study. Seventeen wavelet functions were used to perform a 5-level one-dimensional DWT for each derivative spectrum. Approximation coefficients for the 5th level and detail coefficients for levels 1 to 5 were extracted for each of these fifteen functions.

The original signal can be synthesized from the wavelet coefficients (cD and cA) by applying the Inverse Discrete Wavelet Transform (IDWT). The IDWT reverses the order of the DWT; the wavelet coefficients at every level are up-sampled by two, passed through the synthesis filters H’, and L’ (high-pass and low-pass, respectively), and then
added. The H and H’ produce a high-pass half band filters while L and L’ produce low-pass orthonormal filters (Fig 4.3). All of the information in the original hyperspectral signal is contained in the approximation coefficients at final level (p) plus all the detail coefficients from the preceding levels of decimation. Mathematically, the signal, $f(\lambda)$, is reconstructed from the approximation coefficients, $cA_p$, and the sum of the detail coefficients,$cD_p$.

$$f(\lambda) = cA_p + \sum_p cD_p$$
Wavelet decomposition

Figure 4. 3. High-pass (H) and low-pass (L) filters are applied to the input signal, $f(\lambda)$, having 64 discrete bands. The wavelet coefficients of the high-pass filters can be stored in the last 32 places of the original length 64 vector. The low-pass wavelet coefficients are then passed through five levels of decomposition. The signal reconstruction involves dyadically up-sampling the wavelet coefficients and passing them through orthonormal equivalent wavelet filters.
4.3.5 Multiple linear regression analysis (MLA)

The performance of the wavelet transformed signals in capturing signals pertinent to the various optical constituents (chlorophyll \(a\), phycocyanin and TSM) is assessed using predictive models. A popular Multiple Linear Regression (MLR) analysis technique applied in the field of spectroscopy is stepwise regression (Kokaly et al., 1999; Osborne et al., 2002; Blackburn, 2006). Regression models are constructed between the wavelet transformed coefficients and the concentrations of CPAs, chlorophyll \(a\), phycocyanin and TSM. A 95% confidence interval on the regression coefficients was used to determine which variables were admitted to the model at each step. To be able to optimize the predictors (wavelet coefficients) from the entire signal, stepwise regressions are performed using all coefficients representing all scales (cA5, cD5, cD4, cD3, cD2 and cD1). Stepwise regression was also performed on coefficients from each scale, in turn, to assess the frequency components that characterize the selected CPA.

One potential problem of stepwise regression is potential for high coefficient of determination values due to over fitting. This is usually caused when using data such as the hyperspectral signal where independent variables, the wavelet coefficients in this case, exceed the number of observation representing the dependant variables (CPAs). This causes higher ratios between the regression sum of square and the total deviation, which leads to high \(R^2\) values, even if the contributing variable is inefficient in reducing model error.
One way to evaluate the accuracy of model predictions based on stepwise MLR is using the adjusted coefficient of determination (adjusted $R^2$). This accounts for the number of independent variables included in the regression model, thus lowering the degree of freedom (Davis, 2002). In stepwise regression, the adjusted $R^2$ only increases if the addition of an independent variable lowers the prediction error of the model. Additionally, Fisher’s F-test is used to test goodness of fit of the regression model i.e., a measure of the relative improvement in the explained variance that has taken place after the model was extended with new additive terms. Additionally, t-test is applied to assess the significance of the associations between the dependant (e.g., chlorophyll $a$) and the independent variables (wavelet coefficients). In this study, 95% confidence interval is taken as a criterion to perform the F and t statistical tests.

In this study, regression models are built in two ways: (a) the regression model is constructed using the wavelet coefficient at each scale independently, and (b) predictive models are developed using wavelet coefficients from the full range of level of detail. The first case provides information regarding the optimal frequency range that characterizes a particular CPA. The second case collects wavelet coefficients across the various frequencies and generates a model that utilizes all of the spectral information. The adjusted $R^2$, F-test, t-test, and the Root Mean Squared Error (RMSE) are then used to select the set of wavelet coefficients that provide the most predictive skill and to identify the most predictive models.
4.3.6 Calibration-validation

Wavelet based regression models can be sensitive to calibration data, a high $R^2$ value for models does not guarantee that prediction will be good when using data outside the calibration data. To increase confidence in a particular model, its stability and performance in predicting future event must be evaluated. Experience shows that, a potential factor affecting model stability is the amount of data used to develop it. In this study, the number of available data is limited; therefore, model validation is crucial to assess its stability.

A split sample calibration-validation method similar to that described in section 3.3.1.3, is used to calibrate and validate the wavelet regression models. Prediction errors, $R^2$ and RMSE values were used as criteria to assess model performances. The stability of the beta coefficients was also assessed by comparing the values for calibration and validation datasets.
4.4 Results and discussion

4.4.1 Spatial and temporal variability of CPAs

Descriptive statistics for the concentrations of chlorophyll $a$, phycocyanin and TSM in the WBLE are summarized in Table 4.1. The spatial variation of chlorophyll $a$ is depicted using data from the 18 stations, which extend between the Sandusky Bay and Middle Bass Island in central region of the WBLE (Fig 4.4a). Higher chlorophyll $a$ concentrations were observed at stations 19 and 20 in Sandusky Bay and at stations 3 and 4 in the central region of the WBLE, all of which correspond to river discharge zones. This clearly demonstrates the influence of riverine plumes in the WBLE. Variability across the basin is more apparent during early summer (June and July) than in September. As the summer progresses, the phytoplankton assemblage disperse into the open waters of WBLE and towards the Central Basin and therefore spatial variability among the station decreases (Fig 4.4a). As shown in Fig 4.5, the Sandusky River discharge data for 2009 and 2010 clearly indicates significant riverine flux into the WBLE during early summer period. After July, rivers flow at lower and more constant discharges rates. As evidenced from local rainfall data (NOAA, 2010), few of the peak discharges observed in mid to late July of 2010 are associated with local rainfall events.
Table 4. 1. Descriptive statistic of the selected CPA proxies

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chlorophyll a (µg/l)</td>
<td>89</td>
<td>2.14</td>
<td>10.67</td>
<td>6.91</td>
<td>1.81</td>
</tr>
<tr>
<td>Phycocyanin (µg/l)</td>
<td>89</td>
<td>.64</td>
<td>4.37</td>
<td>2.69</td>
<td>.89</td>
</tr>
<tr>
<td>TSM (mg/l)</td>
<td>89</td>
<td>2.42</td>
<td>37.00</td>
<td>8.50</td>
<td>6.58</td>
</tr>
</tbody>
</table>

High river discharge during the early summer loads significant amounts of terrestrial material in the WBLE leading to high spatial heterogeneity in the concentrations of CPAs between near-shore waters and the central WBLE. Low and more constant discharge rates after July allows homogenization of the various in-water constituents through processes of advection and settlement to the benthic environment.

Assessment of the phycocyanin distribution in the WBLE shows that, the spatial and the temporal variability of phycocyanin matches that of the chlorophyll a pigment (Fig 4.6). This is also an indication of the dominant role of the blue-green algae in these waters. The concentration values of the phycocyanin obtained in this study are within the range of the measured values by other researchers (Vincent et al., 2004).
Figure 4.4. Spatial variability of (a) chlorophyll a and (b) TSM, in the WBLE during the different cruise periods. High variability is recorded during early summer season (June and July). The variance in the concentrations among the stations becomes smaller by late summer (September).
Figure 4. 5. Sandusky River discharge in 2009 and 2010. Note the vertical axis for 2009 and 2010 data. Relatively higher discharge rates are record in 2010 (USGS).

Figure 4. 6. Temporal and spatial variability of phytoplankton pigments in the WBLE. Each date includes data from 18 stations (ST-2 to ST-20 left to right). The Sims based concentrations of phycocyanin co-vary with the primary chlorophyll a pigment, ($R^2=0.81$).
TSM concentration in the WBLE decreased throughout the summer period in 2010 (Fig 4.4). In July of 2010, average concentration was 8.1mg/l and by late summer of 2010 (September), the average concentrations dropped to 7.05mg/l. This is consistent with the trends observed in the magnitudes of main rivers emptying in the basin (Fig 4.5). Higher than average values of TSM are recorded in early summer at the stations corresponding to Sandusky Bay relative to the stations representing the central region of WBLE, indicating difference in the optical variability between the WBLE and Sandusky Bay.

4.4.2 Discrete wavelet transformation

A fundamental property of discrete wavelet analysis is its multiresolution approach to signal processing. This provides a detailed as well as a broad-scale of the hyperspectral reflectance observations. This capability allows: (a) the detection and removal of noise and stochastic errors, thereby reducing the dimensionality of the data set, (b) the identification of the various frequency elements and their localization within the wavelength range of interest (400 – 1000nm). This aids in identifying the optical features of the various in-water constituents.

Derivative spectroscopy has been employed to (a) remove high frequency noise components that may be present across the full spectral range, increasing the signal to noise ratio and therefore, enhancing the signals that characterize the various in-water components and, (b) emphasize peaks and troughs that occur due to the various CPAs.
The latter produces an effective input for multi-resolution analysis. Several remote
sensing studies have employed similar approaches to derivative spectroscopy. For
example, Gong et al. (1997); Han and Rundquist (1997) have used the first-derivatives of
hyperspectral signals to increase correlation between signal and vegetation species.
Demetriades-Shah et al. (1990); Pu et al. (2003) have used first-derivative techniques to
detect the “red-edge” in spectral signatures of vegetation.

In this case, first-derivative hyperspectral data is also used as an input to the 4
level DWT. Preliminary analysis of the first-derivative spectrum of the WBLE clearly
indicate prominent spectral features including absorption troughs at 440, 620 and 675nm,
and the reflectance/fluorescence peak near 710nm (Fig 4.7). This results from the
interaction of incident light with the multiple constituents that are present within the
water column. High inflections positioned near 620, 675 and 710nm indicate the presence
phycocyanin and chlorophyll a, which are used as proxies for detecting phytoplankton
(Simis et al., 2005). The features that are observed in the spectral range of 400 to 600nm
represent absorption and scattering from multiple constituents including detritus and
inorganic sediments. The first-derivative transformed spectrum consists of series of
troughs and peaks which oscillate around zero, resembling the wavelet functions (i.e.,
oscillations around zero that satisfy the zero integrated area admissibility condition). This
enables a more robust convolution of the first-derivative signal using the wavelet
functions. The multi-scale representation of the DWT can distinguish the various in-water
constituents.
Figure 4.7. First-derivative transformed hyperspectral data of 18 stations used as an input for DWT. The figure illustrates prominent optical features in the visible range, including absorption troughs at 440, 620, 675 nm, and fluorescence peak near 710 nm. Each line in the figure represents a spectrum from each field station. The full range spectrum (400 to 2500 nm) is shown in Fig 2.5, section 2.4.
The Daubechies 4 (Db4) rather than the other DWT cases employed (Fig 4.1) is arbitrarily selected to demonstrate the decomposition of the first-derivative hyperspectral data into multi-scale wavelet coefficients. The products of the wavelet transformation of 89 reflectance spectra are shown in Figure 4.8. As shown in Figure 4.8, the DWT produces decreasing number of wavelet coefficients at the larger scales. The wavelet coefficients at scale cD1 represent the highest frequency information picked from the hyperspectral measurements. With increasing scales, the wavelet coefficients (cD2, cD3 and cD4) characterize low-frequency information and the wavelet coefficients in cA4 give a global view of the signal. The original spectrum can be completely reconstructed by adding the detailed coefficients from scale 1 to 4 and the approximation coefficients at scale 4. This demonstrates that there is no signal loss in the discrete decomposition process.

A four level DWT (cA4, cD4, cD3, cD2 and cD1) of the original hyperspectral signal shows that 93.7% of the signal variability can be represented by the medium to low-frequency components (Table 4.2). In this case, the information carried by cD1 (a high frequency component) represents the contribution to the signal due to noise, instead of any characteristic of the in-water constituents. The second frequency component (cD2) represents approximately 3% of the signal variability. Therefore, cD1 and cD2 may be removed and the signal may be reconstructed using only the remaining 25% of the data volume in cD3, cD4 and cA4. This achieves a significant data reduction while maintaining approximately 94% of variance in the signal.
The reduction of the dimensionality of the data set is one key application of the wavelet transformation that is demonstrated here. The frequency components represented in cD1 are considered to be noise and artifacts, and the removal of cD1 increases the signal to noise ratio which in turn improves the detection of the in-water constituents. However, although cD2 carries very low signal variability, as is discussed later in the chapter, it characterizes some spectral features of the in-water constitutes. Removal of cD1 alone results in data reduction by half and retains 97% of the data variance which is still considered to be efficient. Such a reduction in dimensionality contributes to computational efficiency and reduces the complexity of the analysis and interpretation. In this case, intermediate (cD2, cD3) and large-scales (cD4, cA4) characterize most of the absorption and scattering features of the CPAs.

<table>
<thead>
<tr>
<th>Scale</th>
<th>(%) Variances</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>cD1</td>
<td>3.02</td>
<td>3.02</td>
</tr>
<tr>
<td>cD2</td>
<td>3.26</td>
<td>6.28</td>
</tr>
<tr>
<td>cD3</td>
<td>7.75</td>
<td>14.03</td>
</tr>
<tr>
<td>cD4</td>
<td>15.54</td>
<td>29.57</td>
</tr>
<tr>
<td>cA4</td>
<td>70.43</td>
<td>100.00</td>
</tr>
</tbody>
</table>

**Table 4.2.** Percent variance of the various frequencies in the original hyperspectral reflectance data. Note that most of the signal is contained in high-scale range (cD2 – cA4)
Figure 4.8. A 4 level discrete transformation of the first-derivative hyperspectral signal of 18 stations (a), and 4 levels of decomposition (b). The original signal $f(\lambda)$ can be reconstructed by $\sum cA4 + cD4 + cD3 + cD2 + cD1$. 
Visual inspection of Figure 4.8 clearly indicates that the absorption and scattering features characterizing the in-water constituents are represented at various scales. The highest frequency signals (cD1) isolate minor features that exist mainly in the red and near-infrared regions, specifically between 640 and 750nm. Longer wavelength regions are more sensitive to noise and therefore, the certain amount of the signal features observed in cD1 may be associated with noise. The amplitude of the signal increases at intermediate-scales (cD2 and cD3) and prominent absorption and fluorescence peaks that characterize the phytoplankton become more apparent. The second scale of the signal (cD2) isolates anomalies near 620, 665 and 680nm which represent optical features of phytoplankton pigments (Fig 3.5, section 3.4). At this scale, the effects of the constituents on shorter wavelengths are not represented. The effect of phytoplankton and co-varying products on the red/NIR is also retrieved at a lower frequency (cD3). This demonstrates that single frequency signal analysis techniques such as standard spectral analysis, which do not allow localization of frequency content, do not characterize in-water constituents effectively. This in turn demonstrates that retrieval of optically active constituents from remote sensing reflectance requires multi-scale signal analysis using the full spectral range.

The optical features represented in the lower wavelength region of the spectrum are mostly isolated at the lower frequencies (cD4 and cA4). Although these low frequency signals may result from the combined effects of multiple in-water constituents, the individual wavelet coefficients can be optimized using regression models to target the various spectral features.
In Table 4.3, the sums of the variances at the multiple scales represent the total variability within the first-derivative transformed signal. About 4% of the signal variability represents very high frequency signal (cD1), while 26, 23 and 33% of the signal variances are represented in the 2\textsuperscript{nd}, 3\textsuperscript{rd} and 4\textsuperscript{th} levels of frequency resolution, respectively. The approximation frequency (cA4) component represents a global view of the spectral signal.

<table>
<thead>
<tr>
<th>Scale</th>
<th>(% Variances)</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>cD1</td>
<td>3.90</td>
<td>3.90</td>
</tr>
<tr>
<td>cD2</td>
<td>26.60</td>
<td>30.50</td>
</tr>
<tr>
<td>cD3</td>
<td>22.50</td>
<td>53.00</td>
</tr>
<tr>
<td>cD4</td>
<td>33.10</td>
<td>86.10</td>
</tr>
<tr>
<td>cA4</td>
<td>13.80</td>
<td>99.90</td>
</tr>
</tbody>
</table>

Table 4.3. Percent variance of the various frequency components that exist in the first-derivative transformed hyperspectral data.
To develop a more robust wavelet based bio-optical model, a complete 5 level DWT was applied to the derived 64 center-weighted first-derivative bands. Prediction models presented in Figure 4.9 show the performance of multi-resolution analysis using two types of data inputs, original reflectance and first-derivative spectra. Here, the Db4 is used as the wavelet function to decompose the signal and regression analysis is performed between the wavelet coefficients at each scale and in-situ chlorophyll $a$ concentrations.

High coefficient of determination values are obtained between chlorophyll $a$ and wavelet coefficients in the medium to small-scale (high frequency) components (cD3 to cD1). This suggests that important optical features characterizing phytoplankton have medium to high frequency components. Results indicate that the first-derivative of the hyperspectral signal removes high frequency noise, improving the signal to noise ratio and reducing the RMSE of the chlorophyll $a$ prediction model (i.e., the obtained $R^2 = 0.57$ using original hyperspectral signal increased to 0.81 for the first-derivative transformed at cD1, Fig 4.9).
Figure 4.9. Accuracy of chlorophyll a regression models generated using wavelet coefficients at various scales. Fig 4.9 a) represents the $R^2$ values at multiple scales and b) shows corresponding model accuracies. Values calculated from the first–derivative spectrum (red) gives better prediction values than the original reflectance data (blue), high $R^2$ and low RMSE.
4.5 Regression models

A total of 306 regression analyses were performed between the wavelet coefficients derived using MRA and the selected bio-optical indices (chlorophyll \( a \), phycocyanin and TSM). Selected results also illustrated in Figure 4.12.

4.5.1 Performance of the selected wavelet functions

From the many wavelet functions that exist in the field of wavelet analysis, the DWT’s criteria of orthogonality and the requirement of a scaling function, restricts the available wavelet functions to Daubechies, Symlet, and Coiflets. These functions are used to generate wavelet coefficients that characterize the chlorophyll \( a \), phycocyanin and TSM in the WBLE.

All three water quality indices are used as dependant variables to assess the performance of the wavelet functions for predicting the multiple constituents. The analysis was done at individual frequencies as well as using data from the full range of scales.
4.5.2 Prediction models using various frequency components

For prediction of the CPAs, wavelet families of the Db, SYM and COIF were used to generate wavelet coefficients that exist at different frequencies. These wavelets coefficients are used as independent variables in regression analysis to develop models for predicting the concentrations of the selected CPAs. For chlorophyll \( a \) prediction, different DbNs were used to generate wavelet coefficients. Figure 4.10a and b shows the performance of the various DbN’s in estimating chlorophyll \( a \) concentrations at each scale. The variability of the \( R^2 \) values is small among the Daubechies family with an average variance of 0.004. The Db4 shows consistently higher \( R^2 \) and lower RMSE values at the various scales. Further analysis illustrates the highest \( R^2 \) and lowest RMSE value to be in cD3 and cD1 frequency range. Hence, these frequencies can be considered to carry important features that characterize phytoplankton.

Regression models for estimating chlorophyll \( a \) using the wavelet coefficients generated from the Symlet wavelet families (Sym4, Sym5, Sym6, Sym7, Sym8, and Sym9) indicate that, the various Symlets perform equally well at specific signal resolution with \( R^2 \) variance of 0.02, which is even lower than the among the DbNs. As in the case of the Daubechies, the Symlets show high coefficients of determination and low RMSE values when using wavelet coefficients representing the high (cD1) and medium (cD3) frequency components (Fig 4.10c and d). Among the Symlet families the Sym9 shows relatively higher \( R^2 \) and low RMSE, however, the improvement in \( R^2 \) relative to the other Symlets is not statistically significant at \( P > 0.05 \).
The third type of wavelet functions that were applied was the Coiflets family (Coif1, Coif2, Coif3, Coif4, and Coif5). Regression analyses calculated using the Coiflet-derived wavelet coefficients and chlorophyll \( a \) concentrations show correlation and RMSE pattern similar to Daubechies and Symlets (Fig 4.10e and f). In this case, the Coif1 functions gave better predictive accuracy relative to other Coiflet members, \( R^2 \) of 0.79 and RMSE of 0.82μg/l.
Figure 4.10. Plots of chlorophyll a regression model output ($R^2$ and RMSE) as function of wavelet frequency components. Panels a and b represent regression outputs based on Daubechies coefficients, panels c and d are outputs based on Symlet coefficients and panels e and f show the model outputs when using coefficients generated from Coiflet application.
To identify the wavelet function that best predicts chlorophyll $a$, a direct comparison of the $R^2$ and RMSE was done at each scale from the three wavelet families (Daubechies, Symlets and Coiflets) and the best wavelet function was identified from each family member that has relatively good predictive capability. Results indicate that Db4, Coif1 and Sym9 perform the best among their family members. Further comparison between the three selected wavelet functions shows that Db4 best describes chlorophyll $a$ variability with the least RMSE, followed by Sym9 and Coif1, respectively (Fig 4.11). However, it should be noted that the variance in $R^2$ values among the functions at each scale is very low (Table 4.4). Akaike’s best model selection criterion (AIC) was used in this case to scale the difference in the number of decomposition steps. Low AIC values correspond to efficient scale at which the model performs well and in this case it is at cD1. This serves as a method to determine the number of decomposition steps required to extract optical water quality features.

<table>
<thead>
<tr>
<th>Scales</th>
<th>$R^2$</th>
<th>AIC</th>
<th>$R^2$</th>
<th>AIC</th>
<th>$R^2$</th>
<th>AIC</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>cA5</td>
<td>0.61</td>
<td>12.28</td>
<td>0.55</td>
<td>12.44</td>
<td>0.69</td>
<td>12.05</td>
<td>0.005</td>
</tr>
<tr>
<td>cD5</td>
<td>0.40</td>
<td>12.84</td>
<td>0.41</td>
<td>12.80</td>
<td>0.45</td>
<td>12.71</td>
<td>0.003</td>
</tr>
<tr>
<td>cD4</td>
<td>0.47</td>
<td>10.66</td>
<td>0.33</td>
<td>10.98</td>
<td>0.40</td>
<td>10.84</td>
<td>0.005</td>
</tr>
<tr>
<td>cD3</td>
<td><strong>0.76</strong></td>
<td>7.77</td>
<td>0.69</td>
<td>8.05</td>
<td>0.73</td>
<td>7.87</td>
<td>0.001</td>
</tr>
<tr>
<td>cD2</td>
<td>0.74</td>
<td>5.87</td>
<td>0.76</td>
<td>5.84</td>
<td>0.69</td>
<td>6.10</td>
<td>0.001</td>
</tr>
<tr>
<td>cD1</td>
<td><strong>0.81</strong></td>
<td>3.59</td>
<td>0.79</td>
<td>3.71</td>
<td>0.80</td>
<td>3.65</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 4.4. Performance of the three wavelets functions in explaining chlorophyll $a$ variability in the WBLE. Bold labels show the highest $R^2$ values.
Figure 4.11. Performance of regression models for chlorophyll a prediction developed using wavelet coefficients generated based on MRA of the first-derivative spectrum. MRA is applied using the three wavelets functions (Db, Sym and Coif).
Figure 4.12 shows the regression plots of the stepwise regressions for chlorophyll a prediction using the wavelet coefficients generated by convolving the signal using the Db4 (Daubechies 4) wavelet function. High $R^2$ and low RMSE values are recorded for cD1 and cD3. Although most of the models are linear, a slightly curvilinear relationship is evidenced when using coefficients from low-frequency components (cD4, cD5 and cA5). This curvilinearity may be an indication of the complex low-frequency signal that characterizes aggregate of multiple in-water constituents.

Regression models for predicting phycocyanin indicate that the Db4 from the Daubechies family, the Coif1 from the Coiflets family and the Sym9 from the Symlets family are in general the wavelet functions with the highest predictive accuracy (Fig 4.13). The results are consistent with that of chlorophyll a prediction models. As shown in Figure 4.13, critical bio-optical information describing phycocyanin is captured in the cD3 and cD1 levels in all of the wavelet functions. In the WBLE, the blue-green algae are the dominant phytoplankton species with chlorophyll a as the main photosynthesizing pigment, along with the accessory pigment, phycocyanin. Although pure chlorophyll a and pure phycocyanin show distinct spectral patterns, in cyanobacteria denominated waters, these distinct spectral patterns will both occur together in the spectral signatures of the water body. Based on the DWT analysis, optical information characterizing both pigments is captured at cD3 and cD1, further confirming their association.
Figure 4.12. Results from MRA on wavelet coefficients at each scale (cD1, cD2, cD3, cD4, cD5 and cA5) using the Db4 wavelet function. Note the curvilinear relationship when using the low frequency components (red polynomial lines).
Predictive models for TSM using the wavelet coefficients generated from the DWT indicate that coefficients that best explain the variability of TSM are captured in the medium to high frequency components. Since most the optical features related to phytoplankton are captured in the high frequency components, the fact that TSM features are also captured in the high frequency domain suggests that most of the TSM is composed of biomass component. This is consistent with the PCA results in section 3.5.1, where approximately 74% of the variance in the recorded reflectance is attributed to phytoplankton component.

Figure 4.13 shows that the Coif1 provide the best predictions of TSM and phycocyanin. However, when applying the Daubechies and the Symlet family wavelet functions, the best function that describes TSM is different from that of phycocyanin. For TSM, Db5 and Sym7 provide the best models, whereas, for phycocyanin, Db4 and Sym9 generate the best models. This clearly indicates that, although TSM may mainly be composed of phytoplankton component, it also consists of other constituents such as detritus and inorganic matter that have non-linearly relationship with the phytoplankton. Note that in this case, two of the three applied wavelet functions were able to detect this variability, which clearly demonstrates the wavelet’s capabilities in detecting subtle changes in signals.
Figure 4.13. Plots of Phycocyanin and TSM regression model output ($R^2$) as function of multiple frequency components. Panels a and b represent regression outputs based on Db coefficients for phycocyanin and TSM, respectively, panels c and d are outputs based on Coiflet coefficients and panels e and f show the model outputs for Phycocyanin and TSM prediction when using coefficients generated from Symlet application.
4.5.3 Prediction models using the whole range of scales

Predictive models constructed using the wavelet coefficients from the full range of frequencies (cD1, cD2, cD3, cD4, cD5, and cA5) can give $R^2$ of up to 0.95 for each of the CPAs. However, when model efficiency is assessed using adjusted $R^2$, goodness of fit with F-test, and significance of the model coefficients from t-tests, these metrics show that all of these frequencies do not contribute significantly to the actual predictive skill. These statistical assessments overcome problems of overfitting in the prediction models when using a high numbers of independent variables compared to the number of observations.

The performance of the predictive models that are constructed using wavelet coefficients is represented in Table 4.5. The variability at various scales due to chlorophyll $a$ is captured by more than one wavelet function Db4, COIF2, COIF4, SYM5 and SYM8. Strong predictive models are also generated for phycocyanin and TSM using Db4, COIF1, and SYM9. Figure 4.14 shows coefficients from DB4, COIF2, and SYM8 consistently generate high $R^2$ models for the selected CPA proxies.

In the Daubechies family, Db4 appears to be dominant in generating good predictive models for all the proxies. The Coiflets did not have a single function within their families that consistently gave high predictive accuracy. Among the Symlets, Sym6 and Sym8 generated good predictive models for all the CPAs. Results in this study indicated that, all of the major wavelets (Daubechies, Symlets and Coiflets) performed relatively well in generating wavelet coefficients for predicting CPAs.
Table 4. Results of multiple linear regression models (R² and RMSE) using various wavelet functions to estimate selected CPA proxies in the WBLE. All of R² values ≥ 0.84 are highlighted in red.

<table>
<thead>
<tr>
<th></th>
<th>Chlorophyll a</th>
<th>Phycocyanin</th>
<th>Total suspended Matter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R²</td>
<td>RMSE</td>
<td>R²</td>
</tr>
<tr>
<td>DB4</td>
<td>0.85</td>
<td>0.72</td>
<td>0.83</td>
</tr>
<tr>
<td>Db5</td>
<td>0.77</td>
<td>0.87</td>
<td>0.65</td>
</tr>
<tr>
<td>Db6</td>
<td>0.79</td>
<td>0.82</td>
<td>0.83</td>
</tr>
<tr>
<td>Db7</td>
<td>0.81</td>
<td>0.80</td>
<td>0.81</td>
</tr>
<tr>
<td>Db8</td>
<td>0.77</td>
<td>0.88</td>
<td>0.52</td>
</tr>
<tr>
<td>Db9</td>
<td>0.60</td>
<td>1.14</td>
<td>0.71</td>
</tr>
<tr>
<td>COIF1</td>
<td>0.70</td>
<td>0.99</td>
<td>0.84</td>
</tr>
<tr>
<td>COIF2</td>
<td>0.86</td>
<td>0.67</td>
<td>0.83</td>
</tr>
<tr>
<td>COIF3</td>
<td>0.82</td>
<td>1.48</td>
<td>0.84</td>
</tr>
<tr>
<td>COIF4</td>
<td>0.85</td>
<td>0.70</td>
<td>0.76</td>
</tr>
<tr>
<td>COIF5</td>
<td>0.83</td>
<td>0.76</td>
<td>0.85</td>
</tr>
<tr>
<td>SYM4</td>
<td>0.82</td>
<td>0.76</td>
<td>0.83</td>
</tr>
<tr>
<td>SYM5</td>
<td>0.85</td>
<td>0.74</td>
<td>0.76</td>
</tr>
<tr>
<td>SYM6</td>
<td>0.83</td>
<td>0.75</td>
<td>0.82</td>
</tr>
<tr>
<td>SYM7</td>
<td>0.84</td>
<td>0.72</td>
<td>0.51</td>
</tr>
<tr>
<td>SYM8</td>
<td>0.85</td>
<td>0.71</td>
<td>0.82</td>
</tr>
<tr>
<td>SYM9</td>
<td>0.68</td>
<td>1.02</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Figure 4. Results of MLR on the wavelet coefficients derived from the various wavelet functions by applying DWT.
Predictive models constructed using wavelet coefficients derived from 17 wavelet functions have coefficient of determination ranging from 0.51 to 0.89 for phycocyanin. The descriptive statistics of the derived $R^2$ values for each type of proxy is shown in Table 4.6.

<table>
<thead>
<tr>
<th>R² statistics</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Min</th>
<th>Max</th>
<th>Skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chlorophyll a</td>
<td>0.80</td>
<td>0.07</td>
<td>0.60</td>
<td>0.86</td>
<td>-1.63</td>
</tr>
<tr>
<td>Phycocyanin</td>
<td>0.77</td>
<td>0.12</td>
<td>0.51</td>
<td>0.89</td>
<td>-1.29</td>
</tr>
<tr>
<td>TSM</td>
<td>0.79</td>
<td>0.05</td>
<td>0.70</td>
<td>0.88</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Table 4.6. Descriptive statistics of the $R^2$ values generated for each type of CPA proxies using various wavelet functions.

Among the 17 models tested, the mean $R^2$ value of chlorophyll $a$ is 0.8 with a standard deviation of 0.07 (Fig 4.15a). A higher range of $R^2$ values are obtained for the phycocyanin prediction models, with most of the models having $R^2$ value at 0.83. The negative skewness in $R^2$ values is shown in Figure 4.15b. For TSM, the $R^2$ range from 0.70 to 0.88, however, the frequency distribution of the $R^2$ value is broad. As shown in Figure 4.15c, the models for TSM vary in their predictive skill much more than the models for chlorophyll $a$ and phycocyanin. As pointed out in section 3.5.2, this is additional indication that the TSM is characterized by multiple constituents that include both organic and inorganic materials which have different spectral properties. No single wavelet function appeared to generate the best predictive model; however, Db4 model does relatively well for TSM, as well as for chlorophyll $a$ and phycocyanin.
Figure 4.15. Frequency statistics of the $R^2$ values generated for each type of CPA proxies using the 17 wavelet functions. a) Chlorophyll $a$, b) Phycocyanin and c) TSM.
4.5.4 Calibration-validation

Calibration-validation procedures were applied to the wavelet based models that are used to retrieve the concentrations of chlorophyll $a$ and TSM. Table 4.7a and 4.7b show the calibration-validation parameters (mean residuals (MRES), $R^2$ and RMSE) for chlorophyll $a$ and TSM, respectively. Among the individual wavelet frequency components, the cD1 model has stronger prediction capability, i.e., higher $R^2$, lower RMSE for both calibration and the validation datasets. As expected, the validation parameters for both CPA models show lower coefficient of determination and higher RMSE values. Assessment of the beta coefficients shows that, for chlorophyll $a$ the beta values for the calibration and validation datasets are not significantly different. In the case of TSM, the beta values of the cD3 and cD4 based models shows large variation. This is due to poor performance of the models which is also shown by the low $R^2$ values of the models (Fig 4.7b).

Application of calibration-validation to the models developed using full range spectral frequencies (scales) show higher $R^2$ values, and lower RMSE. This shows that the use of full range spectral frequencies to estimate CPAs allows me to capture multi-scale features that characterize a given CPA. Evaluation of the statistical parameters shows states that the $R^2$ and RMSE values for the calibration and validation dataset are similar. This indicates that the wavelet based models are stable, i.e., not affected by dataset size.
<table>
<thead>
<tr>
<th>Wavelets</th>
<th>Data set-Chlorophyll a</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>N</td>
<td>MRES</td>
<td>R²</td>
<td>RMSE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cD1</td>
<td>61</td>
<td>0.00</td>
<td>0.67</td>
<td>0.95</td>
<td>0.69</td>
</tr>
<tr>
<td>cD2</td>
<td>61</td>
<td>-0.00</td>
<td>0.56</td>
<td>1.14</td>
<td>0.43</td>
</tr>
<tr>
<td>cD3</td>
<td>61</td>
<td>-0.00</td>
<td>0.73</td>
<td>0.88</td>
<td>0.70</td>
</tr>
<tr>
<td>cD4</td>
<td>61</td>
<td>0.00</td>
<td>0.50</td>
<td>1.21</td>
<td>0.42</td>
</tr>
<tr>
<td>Full spectra</td>
<td>61</td>
<td>0.00</td>
<td>0.84</td>
<td>0.68</td>
<td>0.83</td>
</tr>
</tbody>
</table>

| Validation                       | cD1                    | 28    | -0.08  | 0.67   | 0.94   | 0.68   |
|                                  | cD2                    | 28    | -0.22  | 0.42   | 1.40   | 0.39   |
|                                  | cD3                    | 28    | -0.00  | 0.71   | 0.91   | 0.64   |
|                                  | cD4                    | 28    | -0.02  | 0.38   | 1.42   | 0.39   |
|                                  | Full spectra           | 28    | 0.01   | 0.81   | 0.80   | 0.79   |

<table>
<thead>
<tr>
<th>Wavelets</th>
<th>Data set-TSM</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>N</td>
<td>MRES</td>
<td>R²</td>
<td>RMSE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cD1</td>
<td>61</td>
<td>0.00</td>
<td>0.72</td>
<td>1.70</td>
<td>0.69</td>
</tr>
<tr>
<td>cD2</td>
<td>61</td>
<td>-0.00</td>
<td>0.30</td>
<td>2.71</td>
<td>0.32</td>
</tr>
<tr>
<td>cD3</td>
<td>61</td>
<td>-0.00</td>
<td>0.65</td>
<td>1.89</td>
<td>0.62</td>
</tr>
<tr>
<td>cD4</td>
<td>61</td>
<td>0.00</td>
<td>0.57</td>
<td>2.10</td>
<td>0.52</td>
</tr>
<tr>
<td>Full range spectra</td>
<td>61</td>
<td>0.00</td>
<td>0.81</td>
<td>1.41</td>
<td>0.79</td>
</tr>
</tbody>
</table>

| Validation                       | cD1                    | 28    | -0.04  | 0.61   | 2.66   | 0.58   |
|                                  | cD2                    | 28    | -0.08  | 0.30   | 3.16   | 0.28   |
|                                  | cD3                    | 28    | -0.20  | 0.64   | 1.99   | 0.70   |
|                                  | cD4                    | 28    | 0.03   | 0.34   | 3.29   | 0.16   |
|                                  | Full range spectra     | 28    | -0.40  | 0.80   | 1.74   | 0.84   |

Table 4. 7. Summary of the calibration-validation parameters of wavelet based bio-optical models.
4.6 Conclusion

This study demonstrated that hyperspectral reflectance data from Case 2 waters can be decomposed using wavelet transforms to extract optical features characterizing various color producing agents. Application of MRA using DWT showed that about 94% of the variability in the data can be retained within 25% of the original data volume and 97% of the information can be retained within 50% of data. This demonstrates the capability of wavelets in reducing the dimensionality of large datasets such as hyperspectral observations. This also contributes greatly to efficiency of data processing. Comparison of the results obtained from the wavelet analysis of the original observation versus first-derivative transformed hyperspectral reflectance data showed that derivative spectroscopy enhances the performance as well as the accuracy wavelet applications by removing high frequency noise components from the data.

Disparate from the traditional methods where spectral indices (band ratios) are used to predict concentrations of in-water constituents, the localization and scaling properties of wavelet functions captures spectral signatures of in-water constituents of interest at multiple scales across the spectrum. Moreover, this method accounts for many of the confounding factors including variations in the material characteristics of the selected CPAs, background noise, and interference by other optically active constituents. Results showed that most of the spectral properties of chlorophyll $a$, phycocyanin and TSM are captured within the medium to high frequency wavelet components. This clearly shows that sensors with high spectral resolution have better capability to detect
spectral properties of phytoplankton and the associated aquatic constituents in optically complex waters such as the WBLE.

Application of wavelet functions of various order number shows that for the Daubechies family the lower order wavelet (Db4) performed better relative to the higher order members. In the single frequency among the Symlets, Sym9 performed well, however, when using whole frequency ranges, Sym6 and Sym8 generated the best predictive models for all CPAs. Members of the Coiflet family performed more or less equally in extracting optimal wavelet coefficients to predict concentrations of CPAs. This variation can be attributed to the difference in symmetry and compactness among the Daubechies, Symlet and the Coiflets.

The reproducibility of high $R^2$ values from multiple iterations using different wavelet functions illustrates the robustness of wavelet functions in extracting meaningful spectral signatures from within complex reflectance spectra. Calibration-validation procedures using a total of 89 observations showed that the developed CPA algorithms are stable and are robust in predicting concentrations of CPAs in the WBLE. The consistency and stability in the performance of the wavelet functions such as the Db4, COIF1 and SYM9, makes these wavelets ideal candidates for developing robust global algorithms for the retrieval of color producing agents in optically complex waters.
CHAPTER 5

Using FR-MERIS multispectral satellite data to estimate color producing agents in the Western Basin of Lake Erie: a neural network approach.
Abstract

There is considerable interest in accurately estimating water quality parameters for the Western Basin of Lake Erie (WBLE) because of the recurring harmful algal blooms that affect the functioning of the ecosystem. Accurate retrieval of water quality parameters using satellite algorithms is still a challenging task in optically complex, turbid, waters such as the WBLE. Important parameters for analysis of inland water quality include phytoplankton and Total Suspended Matter (TSM). In Case 2 waters e.g. WBLE, optical properties are a function of multiple, independent and nonlinearly varying in-water constituents and, therefore, water quality parameters may not be quantified independently from remote sensing water leaving radiance based on simple empirical methods. Nonlinear transfer functions used in Artificial Neural Network (ANN) techniques are a promising approach for developing algorithms for retrieval of the concentrations of water quality proxies such as chlorophyll \( a \) and TSM. Full Resolution-Medium Resolution Imaging Spectrometer (FR-MERIS) images paired with in-situ water quality data were used for the development and validation of an ANN model. A single layered geophysical transfer function is used to relate in-water constituents (phytoplankton density and TSM) and the normalized water-leaving radiances acquired from the MERIS sensor. The network has been trained using a dataset obtained through in-situ sampling which contains measured concentrations of the in-water constituents. The results obtained from the validation data were fairly good and proved the capability of the tool to predict chlorophyll \( a \) (a phytoplankton proxy) and TSM concentrations. Model performance in estimating chlorophyll \( a \) shows \( R^2 = 0.76 \), and \( \text{RMSE} = 0.58 \mu g/l. \)
For the TSM, $R^2 = 0.82$ with RMSE = 2.59mg/l. The ANN developed in this study effectively detected the optical features of the selected color producing agents (CPAs), in both the training and validation sets. This resulting locally derived algorithm for retrieval of CPAs from the MERIS satellite sensor provides a new tool for future monitoring and prediction of water quality parameters within Lake Erie at a high spatial and temporal resolution.

*Keywords:* chlorophyll $a$, total suspended matter, Neural Networks, MERIS.
5.1 Introduction

The Western Basin of Lake Erie has had a history of eutrophication over the last four decades and, consequently, the ecosystem has continuously changed during that period (Conroy et al., 2005). Being the shallowest and southernmost Great Lake, Lake Erie has warm temperatures, which are conducive to biologic productivity. The tendency towards excessive biologic productivity, eutrophication, is further exacerbated by the influx of rivers that flush nutrients from agricultural runoff and Combined Sewer Overflows (CSOs). A minority of algal blooms are composed of species that produce toxins, e.g. Microcystis, or taste or odor problems with drinking water, e.g. Cladophora (Conroy et al., 2005).

Before phosphorus reduction programs were mandated by the Great Lakes Water Quality Agreement (GLWQA) in 1972, the central and western part of the lake was declared as “dead zone” due to development of hypoxia. This is mainly a result of the process of decay that is activated by bacterial action. This process naturally consumes oxygen and primarily takes place in the hypolimnion. The passage of the Great Lakes Water Quality Agreement significantly reduced the amount of phosphorus loading from the adjacent terrestrial environment, leading to improvement of water quality. As a result, there a reduction in algal biomass, as well as in the incidence and extent of harmful algal blooms (e.g., cyanobacterial blooms). Reduction in algal biomass reduced the formation hypoxia/anoxia in the relatively deeper central basin of Lake Erie (McGucken, 2000; Beeton, 2002). Recent introduction of invasive species such as filter-feeding zebra
mussels (dreissenids) into Lake Erie has further diminished plankton density, resulting in the positive shift in the trophic state of the Lake (Munawar and Munawar, 1999).

Although, some researchers have observed decreases in phytoplankton biomass (measured as chlorophyll \textit{a} concentration) in the central and eastern basins (Matisoff and Ciborowski, 2005), recent data and studies indicate that productivity is among the highest seen in fresh water ecosystems (Boyer, 2008; Millie et al., 2009). In addition, the recurrence of cyanobacterial blooms is increasing (Becker et al., 2009). Budd et al. (2002) and Vincent et al. (2004) suggested that Lake Erie is shifting to more eutrophic conditions. Ouellette et al. (2006) and Millie et al. (2009) have documented that in 2009, the Western Basin has experienced an increase in blooms of \textit{Microcystis aeruginosa}, a toxic blue-green alga.

The geology of the Western and Central Basin of Lake Erie is mostly characterized as dolomitic-limestone which resisted the glacier scouring effect during the most recent glaciation resulting in a small and shallow Lake Erie. Because of its smaller volume, water residence time in Lake Erie is short with an average of 2.6 years (Quinn, 1992). The relatively short residence time makes Lake Erie sensitive to stressors such as pollutants, invasive species, and climate change (Quinn, 1992). Moreover, Lake Erie is heavily influenced by the adjacent terrestrial environment through agricultural runoff and urbanization, further increasing the stress on the lake’s ecosystem. The maximum depth of the Western Basin is only 11 meters, and rivers draining eight major watersheds deliver a mix of sediment, Colored Dissolved Organic Matter (CDOM), and riverine
algae to the region. The rivers bring significant volume of sediment loads from agricultural erosion as well as partially treated sewage from treatment plants and untreated sewage from CSOs (USEPA, 2006). Efficient and accurate monitoring of water quality is essential as this region of the lake serves as an economic and social resource through the fishing and recreation industries, and is an important component of the regional drinking water supply.

Traditionally water quality monitoring is done through analysis of water samples collected from pre-selected sites. A method that does not provide the spatial and temporal coverage needed to continuously assess water quality on basin-wide scales. Use of satellite remote sensing provides complementary information which has the potential to greatly improve our understanding of variations in water quality. Satellite imaging systems that detect variations in spectral reflectance from the Earth’s surface in carefully selected visible bands are promising for water quality assessment in both freshwater and marine settings (Baban, 1995; Gons, 1999; Nellis et al., 1998; Islam et al., 2001; Schmugge et al., 2002; De Cauwer et al., 2004; Ouillon et al., 2004; Shuchman et al., 2006; Torbick et al., 2008; Tarrant and Neuer, 2009). These systems assess of water quality by applying algorithms that relate satellite-measured reflectance to the concentrations of specific CPAs (see Martin, 2004). Ground-truthing is often achieved by comparing co-located satellite and in-situ observations. Applying this technology to the WBLE has proven challenging due to the difficulty of separating the spectral signatures of the multiple CPAs in the environment.
Bio-optical algorithms such as empirical and semi-analytical have been derived for NASA’s MODIS, SeaWiFS, and CZCS sensors for determining CPAs (Pozydnyakov et al., 2005; Werdell and Bailey, 2002; Gordon, 1997; Martin, 2004). Based on optical properties, Morel and Prieur (1977) have classified marine waters as Case 1 and Case 2. In Case 1 waters, phytoplankton pigments govern the optical properties of the water. Spectral variations in the backscattered flux of Case 1 waters are primarily related to concentration of chlorophyll $a$, which varies as a function of the phytoplankton population. In open ocean environments, where chlorophyll $a$ dominates the spectral signal, the spectral signals are linear functions of chlorophyll $a$ concentration; therefore simple linear regression analysis can be used to model theoretical relationships between in-situ concentration of chlorophyll $a$ and radiance (O'Reilly et al., 1998). For Case 1 marine waters, a number of algorithms have been developed and tested for the retrieval of the concentrations of in-water constituents from water reflectance spectra. In Case 2 waters, optical properties are influenced by chlorophyll $a$ and by one or more other agents such as Total Suspended Matter (TSM), Colored Dissolved Organic Matter (CDOM) and, in some cases, bottom reflection. This makes the water bodies optically more complex, and the task of independently retrieving estimates of the CPAs is more challenging.

Many existing algorithms employ spectral indices, which usually involve band ration and/or band differencing between two or more spectral windows. These are then used to establish statistical relationships between satellite reflectance values and synchronously collected collocated in-situ water quality data (Gitelson et al., 2000;
Dall’Olmo and Gitelson, 2005; Gordon & Morel, 1983; O'Reilly et al., 1998; Kishino et al., 1998; IOCCG, 2006). Two band and three band models have been developed to estimate chlorophyll $a$ in turbid waters using the Medium Resolution Imaging Spectrometer (MERIS) sensor data (Dall'Olmo and Gitelson, 2005; Gitelson et al., 2007, 2008). Semi-analytical models constitute another approach for estimating chlorophyll $a$ from satellite sensors. These models are based on known spectral features that are discovered from the correlation between the apparent optical properties (AOPs) and the inherent optical properties (IOPs) of the water (Carder et al., 1999; Gordon and Wang, 1994). In Case 2 waters, most of the semi-analytical algorithms used to retrieve chlorophyll $a$ utilize the red and near infrared (NIR) spectral region (650-750 nm) (Simis et al., 2005; Gons et al., 2002, 1999; Gitelson et al., 1991)

In the Western Basin of Lake Erie, the spectral signature of the water is affected by concentrations of a number of independently and co-varying optically active constituents. The development of a single chlorophyll $a$ algorithm in this type of waters can be challenging, especially if environmental factors (e.g., wind, temperature, nutrient influxes) are significant. Some of the challenges encountered when dealing with Case 2 waters include: (a) phytoplankton communities are complex and their composition is highly variable through the seasons as these communities are sensitive to the availability of nutrients and light, and to other factors, such as water temperature, and (b) large variations in the spatial, temporal and particle size distribution of suspended matter that is added to the system by erosion and fluvial transport.
Optical complexity results in a situation that is best addressed as a nontrivial non-linear modeling problem (Keiner and Yan, 1998). In this case, standard regression analysis may not model these non-linear relationships accurately, except over small ranges. Moreover, non-linear regression analysis generally requires \textit{a priori} knowledge of the nature of the non-linear behavior, which may be unknown (Krasnopolsky et al., 1995). In Lake Erie, these challenges preclude the direct application of existing bio-optical algorithms to satellite observations, with results suggesting that a fundamentally different approach is needed in the Western Basin (Witter et al., 2009). Artificial Neural networks (ANN) may flexibly model a variety of nonlinear behavior among water quality parameters. Tanaka et al. (2004) and Schiller and Doerffer, (1999) demonstrated that NN modeling can be used operationally to derive the concentrations of in-water constituents of very turbid coastal waters from satellite data. They have also showed that NN based algorithm can estimate multiple water quality parameters simultaneously. The main goal of this study is to develop ANN-based algorithm for estimating phytoplankton density (chlorophyll \textit{a} being used as a proxy) and TSM using high resolution MERIS satellite data, which would be specific for the optically complex waters of the WBLE.
5.2 Artificial Neural Networks

Computing technology has enhanced our capability to describe complex and dynamic physical systems using mathematical modeling. One such tool is the ANN, which is inspired by the way the brain processes information. The main concept with ANN algorithms is to develop models capable of capturing and representing complex input/output relationships (Fausett, 1994). In recent years, the ANN, a data-adaptive modeling tool, has become increasingly popular among researchers who model dynamic physical processes. Many researchers (Gross et al., 2000; Keiner and Brown, 1999; Dzwonkowski and Yan, 2005) have implemented ANN to estimate phytoplankton from Case 1 waters, where influence of TSM and CDOM is negligible.

Modeling using ANN involves parametric learning by which the network adapts itself to a pattern, and eventually, it approximates the desired output. The adaptation process is achieved as the network adjusts its parameters (i.e., the weights of synapses) in response to inputs, so that it produces output within acceptable error margin. In this study, the training data consists of synchronously measured satellite radiometric data and in-situ measurements of CPAs. Based on the parametric learning, the inputs are fed into the network and the network calculates the outputs as concentrations of CPAs. One challenge that needs to be overcome when applying ANN techniques is choosing an efficient network size. This involves deciding on the number of layers in the network, the number of neurons with a given layer and the weight of the synapses.
5.3 Materials and Methods

5.3.1 Western Basin of Lake Erie

The Western Basin of Lake Erie is an optically complex environment due to the diversity and inhomogeneous distributions of in-water constituents (Witter et al., 2009). It represents the shallowest part of Lake Erie, and is heavily influenced by influx of rivers discharging suspended and dissolved terrestrial matter. Re-suspension of bed sediments (Marvin et al., 2007), coupled with loading from the terrestrial environment, has resulted in a biologically productive and turbid WBLE. At present, variations in water quality in the WBLE are suggested to be generally due to seasonal and inter-annual variability in river runoff to the Lake (Marvin et al., 2007). In recent years, toxic algal blooms (Microcystis and Lyngbya Wolleii) have been documented in the WBLE (Becker et al., 2009; Bridgman and Penamon, 2010).

5.3.2 Data acquisition

5.3.2.1 Field/Lab data set

Field cruises were conducted in the summers of 2009 and 2010, during days that have relatively low cloud covers and days that matched-up with satellite overpasses. The Ohio Stone Lab’s Research Vessel (RV) Gibraltar, and Erie monitor were used for field cruises. Each cruise sampled 18 pre-selected stations over the course of approximately 11 hours, usually from 8:00 in the morning to 7:00pm local time. MERIS overpasses were
available within few hours of the time of in-situ measurements. Water samples were also collected for laboratory analysis. All sampling stations were at least 2km further from land to avoid contamination of the collocated satellite pixel by land (Fig 5.1).

In-situ measurements were made using the HACH Hydrolab submersible sonde equipped with multiple sensors, including a chlorophyll a detector that has excitation and emission wavelengths at 460 and 680nm, respectively. Concentration of chlorophyll a in the water column was taken as an averaged value of measurements from the upper-most 0.5m of the depth profile. A liter of water was collected at each station and filtered onto a 0.45µm pore size Glass-Fiber Filter (GF/F) to collect Total Suspended Matter (TSM). These samples were stored in a cooler with ice in the dark, to avoid material degradation, and then taken to the laboratory for analysis. The concentration of the TSM was determined gravimetrically by placing the dried filter residues on a highly sensitive analytical balance. An additional liter of water was collected at each station from 0.2m depth and taken to the laboratory to determine TSM using a Malvern Mastersizer 2000 laser particle size analyzer.
Figure 5.1. MERIS data acquired on 24 June 2009 with locations of in-situ stations. The Western Basin of Lake Erie (WBLE) and the Sandusky Bay is shown. The pseudo true color MERIS image clearly shows variations of color producing agents in the WBLE.
5.3.2.2 Satellite data

The Full Resolution - Medium Resolution Imaging Spectrometer (FR-MERIS) is a current-generation multispectral sensor that is mainly designed for remote sensing of coastal waters. The sensor was launched by the European Space Agency (ESA) aboard the Environmental Satellite-1, in March 2002. It has several features that provide remote sensing capabilities for monitoring and assessing color producing agents that are typically found in Case 2 waters. Some of the main features of the design include, high spatial resolution, with individual pixels measuring 290 by 290 m, a latitude-dependant revisit period of one to three days, and fifteen narrow spectral bands in the range 400–900 nm including bands (680 or 708nm) for measuring the sunlight-stimulated fluorescence of phytoplankton. The sensor has a high signal to noise ratio and high radiometric sensitivity (16 bit). The complete characteristic of MERIS Level-1b and the MERIS spectral bands is provided in Table 5.1 and Table 5.2.

In this study, four FR-MERIS Level-1b images over the study area (during 2009–2010) were downloaded from the ESA data gateway. The dates were selected on the basis of the availability of ground data and the cloud coverage conditions on the WBLE (Table 5.3). Field measurements were taken throughout the day, whereas, the satellite images covering the entire WBLE were taken instantly in the afternoons. Field sampling campaign usually started early in the morning; therefore, data from stations that were visited in the morning will not have a perfect temporal correspondence with the data from MERIS. However, in this case hourly variability of the optical properties is assumed to be
insignificant. The study did not find trends in the residual errors to suggest effects of temporal drifts between satellite and in-situ match-ups. Therefore, the in-situ measurements can be used as match-up data with the satellite. The radiance-based values in the Level-1b data were converted to Level-2b reflectance, which are used for the development and validation of the ANN.

The Doerffer and Schiller (2007) Case 2 Regional processor algorithm, included in the Basic ENVISAT Toolbox for (A) ATSR and MERIS (Beam-4.8) was used to carry out atmospheric correction on the MERIS images of WBLE. This algorithm is adapted to a wide range of optical properties, and was optimized for eutrophic European lakes that have extreme concentrations of in-water constituents. The algorithm is based on radiative transfer simulations and is designed primarily for performing atmospheric correction over turbid Case 2 waters. Several researchers including Moses et al. (2009) and Luis Guanter et al. (2009) have successfully applied the Case 2 Regional processor for atmospheric correction on MERIS data in turbid waters. The performance of the atmospheric correction over the WBLE is not within the objectives of the present study, however, compared with other near-infrared based atmospheric correction procedures, the Case 2 Regional processor performed relatively well in the WBLE, producing positive values across the spectrum (Fig 1.4, section 1.5). During data processing, flags were raised for pixels representing coastline, land, clouds and invalid reflectance using the Beam software. Data points delivered from cloud-free scenes, and areas that were not flagged for coastline and invalid reflectance, were considered to be valid match-up data. In order
to retain pixel resolution, values from center pixels were used to match up with the
location of sampling stations.

<table>
<thead>
<tr>
<th>Name</th>
<th>Level-1b Full Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifier</td>
<td>MER FR 1P</td>
</tr>
<tr>
<td>Product Level- 1b</td>
<td>TOA radiance for the 15 MERIS bands at full resolution, calibrated,</td>
</tr>
<tr>
<td>Description</td>
<td>Geo-located, annotated with Product Confidence Data, Calibration data, classification flags, and environment parameters.</td>
</tr>
<tr>
<td></td>
<td>The user will be able to order adjacent scenes covering the full MERIS swath and any number of adjacent scenes pairs leading up to the complete MERIS orbit of 17500 km, depending on the availability of the Full-Resolution data. Radiometric and geometric continuity is guaranteed between adjacent scenes.</td>
</tr>
<tr>
<td>File Size</td>
<td>2241 lines x 2241 pixels (scene)</td>
</tr>
<tr>
<td>Pixel Spacing</td>
<td>approximately 300 m x 300 m (along-track x across track)</td>
</tr>
<tr>
<td>Coverage</td>
<td>approximately 575 km x 575 km (scene)</td>
</tr>
<tr>
<td>Bits/Pixel</td>
<td>16</td>
</tr>
<tr>
<td>Unit</td>
<td>$10^{-3}$Wm$^{-2}$sr$^{-1}$nm$^{-1}$</td>
</tr>
</tbody>
</table>

Table 5.1. MERIS Level-1b characteristics (after European Space Agency, (ESA, 2006))
<table>
<thead>
<tr>
<th>Centre (nm)</th>
<th>Bandwidth(nm)</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>412.5</td>
<td>10</td>
<td>Yellow substance and detrital pigments.</td>
</tr>
<tr>
<td>442.5</td>
<td>10</td>
<td>Chlorophyll absorption maximum.</td>
</tr>
<tr>
<td>489.9</td>
<td>10</td>
<td>Chlorophyll and other pigments.</td>
</tr>
<tr>
<td>509.8</td>
<td>10</td>
<td>Suspended sediment, red tides.</td>
</tr>
<tr>
<td>559.6</td>
<td>10</td>
<td>Chlorophyll absorption minimum.</td>
</tr>
<tr>
<td>619.2</td>
<td>10</td>
<td>Suspended sediment.</td>
</tr>
<tr>
<td>664.8</td>
<td>10</td>
<td>Chlorophyll absorption and fluorescence reference.</td>
</tr>
<tr>
<td>681.25</td>
<td>7.5</td>
<td>Chlorophyll fluorescence peak.</td>
</tr>
<tr>
<td>705</td>
<td>10</td>
<td>Fluorescence reference, atmospheric corrections.</td>
</tr>
<tr>
<td>753.75</td>
<td>7.5</td>
<td>Vegetation, cloud.</td>
</tr>
<tr>
<td>760.62</td>
<td>3.75</td>
<td>O2R-branch absorption band.</td>
</tr>
<tr>
<td>865</td>
<td>20</td>
<td>Vegetation, water vapour reference.</td>
</tr>
<tr>
<td>885</td>
<td>10</td>
<td>Atmosphere corrections.</td>
</tr>
<tr>
<td>900</td>
<td>10</td>
<td>Water vapour, land.</td>
</tr>
</tbody>
</table>

Table 5.2. MERIS bands wavelength, bandwidth and applications (ESA, 2006)

<table>
<thead>
<tr>
<th>Dates</th>
<th>Field Cruise</th>
<th>Time</th>
<th>MERIS overpass</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25-Jun-09</td>
<td>all day</td>
<td>25-Jun-09</td>
<td>15:33</td>
</tr>
<tr>
<td></td>
<td>2-Sep-09</td>
<td>all day</td>
<td>2-Sep-09</td>
<td>16:05</td>
</tr>
<tr>
<td></td>
<td>27-Jul-10</td>
<td>all day</td>
<td>27-Jul-10</td>
<td>15:53</td>
</tr>
<tr>
<td></td>
<td>13-Sep-10</td>
<td>all day</td>
<td>13-Sep-10</td>
<td>16:05</td>
</tr>
</tbody>
</table>

Table 5.3. Dates of each cruise and the date of the corresponding MERIS overpass of WBLE. Field campaign took all day to cover 18 stations. The time of MERIS image acquisition is given using the local time.
5.3.3 Artificial Neural Network

The application of ANN involved five steps, (a) compiling the input and output dataset, (b) designing the structure of the ANN (i.e., determining the coefficients of the ANN), (c) testing the performance of the ANN, (d) applying the ANN to in-situ data and satellite data, and (e) validating the results. The Statistical Analysis Software (SAS) package was used to implement the steps for developing the analytical algorithms. Compiled data included spatially and temporally collocated satellite reflectance data and field-based observations of CPAs.

A Single Hidden Layer-Perceptron (SHLP) is implemented to retrieve the concentrations of chlorophyll $a$ and TSM. This network architecture comprises three layers: an input layer, one hidden layer and an output layer (Fig 5.2). A geophysical transfer function is used to approximate the relationship between the reflectance measured within MERIS bands and the concentrations of the in-water constituents in the WBLE. Following the fundamentals of the McCulloch and Pitts model (MCP) approach, the outputs (chlorophyll $a$ and total suspended matter) are a function of the weighed sum of the spectral bands plus a bias (Fig 5.2).
In the first layer, MERIS data is distributed into the network. The first operation is a summation function which sums up the products of the radiometric values ($\rho_i$) along with geometric information ($g_i$) and the corresponding weights of the links ($\omega_{i,j}$). The sum is then added to a ‘bias’ ($\alpha_j$) and undergoes sigmoidal transformation. The transformed values are kept in the hidden neurons. These are then multiplied with the weight links, $\mu_{j,k}$, that connects the hidden neurons with the outputs. The results are added to a bias correction factor, $\beta_k$, and a second sigmoid transformation is carried out to estimate the concentrations of chlorophyll $a$ and TSM. Subscripts $i$, $j$ and $k$ refer to neurons in the input, hidden and output layers, respectively.
In the WBLE, the biogeochemical process and interactions are nonlinear and complex, therefore the most widely used logistic function (a sigmoid function) was applied as a transfer function to predict the selected CPAs. The function has the capability of describing nonlinear processes that lack detailed description and that show progression from small beginnings that accelerates and approaches a climax over time (Tanaka et al., 2004). The logistic function is mathematically defined by,

$$f(x) = \frac{1}{1 + e^{-x}}$$ \hspace{1cm} 5.1

MERIS reflectance data are used as input variables for the ANN. The reflectance data included MERIS bands 2 to 12, which range from 442.5nm to 778nm. According to González Vilas et al. (2011), band 1 can be a source of noise, and in this case since the reflectance values in band 1 are extremely low the signals are susceptible to noise factors. Therefore, band 1 was not used in this analysis. In addition to reflectance values, satellite geometrical data (i.e., sun zenith angle) is used in the input layer. Zenith angle affects the amount of down-welling irradiance, and consequently the normalized water leaving radiance. MERIS images taken at angles beyond a solar zenith angle of 60° were not considered because of possible sun-glint effects (ESA, 2006). In-situ measurements of the optical properties of Lake Erie are used in combination with parametric learning processes to train the ANN algorithm. These measurements include observations from representative sites within the WBLE.
Based on the parametric learning, the inputs are fed into the network, and the network calculates the CPA concentrations. Variations on this methodology include studies such as Dzwonkowski and Yan (2005) that used the MCP where the inputs are 'weighted' and the effect that each input has on decision making is dependent on the weight of that particular input.

A back-propagation learning procedure was used during the training phase. This involves estimating an error function for each layer by computing the Mean Square Difference (MSD) between the model estimates and the actual measurements (Hetcht-Nielsen, 1988). In this study, a modified error surface is used to avoid overfitting. This can occur due to the effects of local minima in a small network, affecting the predictive skill of the ANN algorithm (Wynne-Jones, 1991). The optimization criterion is formed by the MSD plus overfitting penalty times the sum of squares of weights (eq. 5.2). This method modifies the error function in a way such that the unnecessary connections with near zero weights can be eliminated or reduced without degrading the performance of the network. When the neural network is overfit, the number of weights is too large, thus this criterion penalizes the use of large number of weights and keeps the weights small. The penalty coefficient, referred to as the weigh decay, encourages the algorithm to find solutions which use as few weights as possible. Setting this value to zero will result in an overfit and the numerical iteration procedure may have difficulty converging. In this study, the penalty coefficient is set to 0.01, which is a default used in the SAS and recommended when dealing with relatively small dataset (JMP and Proust, 2009). If the dataset is large the over fitting penalty will have less effect.
The functional form of the modified error surface is defined by:

\[ MSD = MSD^0 + \gamma \sum_i \sum_j w_{ij}^2 \]  \hspace{1cm} 5.2

Where, \( \gamma \) is the overfitting penalty, a positive constant used to control the contribution of the error due to the weights, and \( w_{ij} \) is the weight of the synapses between two adjacent layers. This procedure is carried out on each layer backwards within the network structure, thus giving back-propagation its name. The adjustment continues until no more significant variations in the error function are observed. To find the best performing model, the algorithm was run multiple times using different starting values. The number of runs was set to 16 as recommended by JMP and Proust (2009), and each run was iterated 75 times before non-convergence was reported.

Cross-validation is a standard method for assessing the predictive accuracy of a model in a test sample and to avoid overfitting. This is done by reserving a portion of the existing dataset to verify the algorithm generated by the training data. SAS has several options for cross-validation, the Random Holdback and K-fold. In the Random Holdback, the data are randomly partitioned into groups: a training group and a validation (Holdback) group. The training group is used to construct the model and the Holdback group is used to validate the model. In this case, approximately two-third of the data was used to train the model and the remaining one-third was used to validate the model. In K-fold cross-validation, the algorithm is initially built using the complete dataset. Then the data are partitioned randomly into a number of groups (K). Using the overall fit for starting values, for \( i=1, 2, 3 \ldots, K \), the \( i^{th} \) group is treated as a holdback set and the
algorithm is built using the remaining N-N/K observations and the $R^2$ value is calculated using the $i^{th}$ group as a holdback. After iterating through the K starting values, the predictive accuracy of the model is expressed as the average of the $R^2$ for the K iterations. In this study, the best value for K was determined after multiple runs using K values ranging from 2 to 10. JMP and Proust (2009) indicated that K-fold cross-validation works well for K between 2 and 10.

One of the challenges in the application of neural networks is determining an appropriate number of nodes in the hidden layer. The use of a large number of nodes in the hidden layer may give good approximation of the training dataset, however, it affects the generalization capabilities of the network, i.e., its ability to produce accurate results when using a validation dataset containing withheld information. Studies by Chen et al. (2008) have shown that using a smaller network (i.e., fewer nodes) has several advantages over using a larger network. These include a more efficient learning process, which is computationally less complex, requires less memory, and, most importantly, has linear decision boundaries that allow for better generalization, even for data outside the training dataset. In this study, the number of hidden nodes to be used was determined by running the algorithm using a variable number of nodes and searching for divergence in the statistics of the training and validation datasets.
5.3.4 Performance measurements

Model performance was evaluated from the validation dataset using three parameters which compare the in-situ CPA (chlorophyll $a$ and TSM) concentrations and the predicted values from the ANN. The computed parameters are the sum of square errors ($SS_E$), the coefficient of determination ($R^2$) and the Root mean square error (RMSE):

$$SS_E = \sum_{i=1}^{n} (C_i^m - \frac{1}{n} \sum_{i=1}^{n} C_i^m)^2$$  \hspace{1cm} (5.3)

$$R^2 = \frac{\sum_{i=1}^{n} (C_i^m - \frac{1}{n} \sum_{i=1}^{n} C_i^m)(C_i^n - \frac{1}{n} \sum_{i=1}^{n} C_i^n)}{\sqrt{\sum_{i=1}^{n} (C_i^m - \frac{1}{n} \sum_{i=1}^{n} C_i^m)^2 \sum_{i=1}^{n} (C_i^n - \frac{1}{n} \sum_{i=1}^{n} C_i^n)^2}}$$  \hspace{1cm} (5.4)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (C_i^m - C_i^n)^2}{n}}$$ \hspace{1cm} (5.5)

In equations 5.3 – 5.6, $C_i^m$ is the ANN predicted concentration value. $C_i^n$ is the measured value, and $n$ represents the number of observations. $SS_E$ measures how much variation there is in the modeled values and the coefficient of determination indicates ratio of $SS_E$ to the total variation observed in the in-situ data. In other words, it is a measure of the correlation between the observed and the predicted data sets. The RMSE is a measure of the differences between values predicted by the model and the measured values. It is a good measure of precision. These individual differences are also called
residuals, and thus the RMSE serves to aggregate them into a single measure of absolute error, with units equivalent to the units of the measurements.

5.4 Results and discussion

5.4.1 The ANN structure

The ANN architecture is a function of the dimensionality of the input and output data. A SHLP is used to describe the relationship between MERIS reflectance bands and two independently varying CPAs (chlorophyll $a$ and TSM) (Fig 5.3).

Figure 5.3. The SHLP structure constructed for the WBLE. The network shows an input layer with MERIS data, a hidden layer and an output layer representing two CPAs.
5.4.2 Cross-validation of the ANN

Eleven MERIS bands and the geometric (i.e., the sun zenith angles) provide the number of nodes in the input layer. The output data layer has two nodes, one for chlorophyll $a$ and another for TSM. The number of nodes in the hidden layer is determined by analyzing the performance of the algorithm using a variable number of nodes (1 to 5). Figures 5.4a and 5.4b illustrate that the predictive accuracy of the ANN algorithm increases on the training dataset when using a higher number of nodes and the RMSE of the algorithm decreases linearly, with a minimum RMSE when using five hidden nodes. However, assessment of algorithm performance using the validation dataset shows that the predictive accuracy of the model is highest when using three nodes. This demonstrates that a network structure that has a higher number of neurons (greater than 3 for this case) performs well on patterns included in the training dataset, but leads to problems of overfitting and therefore exhibits poor performance when applied outside the training dataset. On the other hand, a small ANN with few nodes in the hidden layer (less than 3 in this case), will poorly approximate the training as well as the validation dataset. As shown in Figures 5.4a and 5.4b, the RMSE of a network is high when using fewer nodes. In this work three nodes are selected for the hidden layer as an optimal number of processing units to relate the MERIS reflectance and the CPAs.
Figure 5.4. Effects of the network’s structural complexity on ANN performance for (5.4a) chlorophyll $a$ and (5.4b) TSM, estimation.
ANN models can often be overparameterized and may provide high but false predictive accuracy due to overfitting. These models have poor generalization when applied to data outside the training set. Therefore, it is crucial to identify models that have a high $R^2$ for training data as well as the validation dataset. Tables 5.4 and 5.5 summarize the results of the cross-validation for the ANN algorithm using the Random Holdback and the K-fold approach, respectively.

The Random Holdback approach trained the ANN using 66% of the dataset. A well-fitting model was developed with $R^2$ of 0.76 and 0.81 for estimating chlorophyll $a$ and TSM, respectively. The performance of the algorithm on the validation dataset gave $R^2$ values of 0.66 and 0.68, as the highest predictive accuracy for the estimation of chlorophyll $a$ and TSM, respectively. As compared to the $R^2$ obtained from the training data, the relatively lower $R^2$ values for the validation dataset were expected because the data were not used during the training phase.

The K-fold cross-validation is the second technique that is used to train the ANN. Optimal K value was obtained after training data using range of K values (between 2 and 10). Table 4 shows that performance parameters (SSE, RMSE and $R^2$-training) do not vary significantly when using different K values. However, slight variation were observed in the performance of the algorithm when using the validation dataset. In this case, the 8-fold cross-validation gave the best $R^2$-validation values of 0.652 and 0.697 for chlorophyll $a$ and TSM, respectively.
<table>
<thead>
<tr>
<th></th>
<th>Hidden nodes</th>
<th>SSE</th>
<th>RMSE</th>
<th>$R^2$-Training</th>
<th>$R^2$-Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chlorophyll $a$</td>
<td>1</td>
<td>30.08</td>
<td>0.81</td>
<td>0.51</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>22.24</td>
<td>0.70</td>
<td>0.64</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>15.06</td>
<td>0.59</td>
<td>0.76</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>12.52</td>
<td>0.53</td>
<td>0.79</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8.71</td>
<td>0.45</td>
<td>0.86</td>
<td>0.29</td>
</tr>
<tr>
<td>TSM</td>
<td>1</td>
<td>642.00</td>
<td>3.73</td>
<td>0.64</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>451.54</td>
<td>3.17</td>
<td>0.75</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>347.72</td>
<td>2.81</td>
<td>0.81</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>165.74</td>
<td>1.96</td>
<td>0.91</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>214.44</td>
<td>2.26</td>
<td>0.88</td>
<td>0.31</td>
</tr>
</tbody>
</table>

**N=54 Training set**
**N=17 Validation set**
**Random Holdback**

*Table 5. 4. Summary of the ANN validation parameters using the Random Holdback approach based on training and validation data set.*
<table>
<thead>
<tr>
<th>K-fold</th>
<th>Chlorophyll a</th>
<th>S SE</th>
<th>RMSE</th>
<th>R² Training</th>
<th>R² Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-fold</td>
<td>24.94</td>
<td>0.61</td>
<td>0.75</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>3-fold</td>
<td>24.96</td>
<td>0.61</td>
<td>0.75</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>4-fold</td>
<td>24.96</td>
<td>0.61</td>
<td>0.75</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>5-fold</td>
<td>24.94</td>
<td>0.60</td>
<td>0.75</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>6-fold</td>
<td>24.95</td>
<td>0.61</td>
<td>0.75</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>7-fold</td>
<td>24.95</td>
<td>0.61</td>
<td>0.75</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>8-fold</td>
<td>24.98</td>
<td>0.60</td>
<td>0.75</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>9-fold</td>
<td>24.96</td>
<td>0.61</td>
<td>0.75</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>10-fold</td>
<td>24.96</td>
<td>0.61</td>
<td>0.75</td>
<td>0.55</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>K-fold</th>
<th>TSM</th>
<th>SSE</th>
<th>RMSE</th>
<th>R² Training</th>
<th>R² Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-fold</td>
<td>451.20</td>
<td>2.59</td>
<td>0.82</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>3-fold</td>
<td>451.23</td>
<td>2.59</td>
<td>0.82</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>4-fold</td>
<td>451.23</td>
<td>2.59</td>
<td>0.82</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>5-fold</td>
<td>451.20</td>
<td>2.59</td>
<td>0.82</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>6-fold</td>
<td>451.22</td>
<td>2.59</td>
<td>0.82</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>7-fold</td>
<td>451.27</td>
<td>2.59</td>
<td>0.82</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>8-fold</td>
<td>451.08</td>
<td>2.58</td>
<td>0.82</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>9-fold</td>
<td>451.24</td>
<td>2.59</td>
<td>0.82</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>10-fold</td>
<td>451.26</td>
<td>2.59</td>
<td>0.82</td>
<td>0.57</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.5. Summary of the ANN validation parameters using the K-fold cross-validation approach based on training and validation data set.

Figures 5.5 and 5.6 show the relationship between the CPAs predicted using ANN and the actual values obtained using the training dataset. Application of a t-test indicated that the coefficients of determination obtained from the regression analyses are statistically significant ($P < 0.05$). Assessment of the residual plots also confirms the goodness of fit of the ANN models.
**Figure 5.5.** Relationship between predicted and observed concentrations of (5.5a) chlorophyll \( a \) and (5.5b) TSM. Algorithm is trained using the Random Holdback approach. The red dashed lines represent the 1:1 Line.
ANN based regression ($R^2 = 0.82$)
$y = 0.8155x + 1.5467$

Predicted TSM (mg/L) vs. TSM (mg/l) - Training dataset

5.6a

ANN based regression ($R^2 = 0.75$)
$y = 0.7384x + 1.7809$

Predicted chlorophyll a (ug/l) vs. chlorophyll a (ug/l) - Training dataset

5.6b

Figure 5. 6. Relationship between predicted and observed concentrations of (5.6a) chlorophyll $a$ and (5.6b) TSM. Algorithm is trained using the 8-fold cross-validation technique. The red dashed lines represent the 1:1 Line.
5.4.3 **CPA dynamics using the ANN algorithm**

The ANN algorithm developed in this study shows relatively high $R^2$ values of 0.76 and 0.82 for chlorophyll $a$ and TSM, respectively. The root mean square error (RMSE) values for the training and validation datasets are low (0.58µg/l for chlorophyll $a$ and 2.6mg/l for TSM). Regression plots from the CPAs do not show significant deviation from the 1:1 line (red line, Fig. 5.6), indicating that the ANN algorithm did not underestimate or overestimate the predictive values. Scattering effects due to inorganic matter is not observed on the chlorophyll $a$ and TSM plots. This indicates that most of the residue on the filter is composed of biogenic material that potentially co-varies with phytoplankton (chlorophyll $a$). CDOM in the WBLE is another CPA that could potentially mask effects of phytoplankton and other suspended matter and hence degrade the ANN performance. The effects of CDOM would show up especially on samples collected near the shore and Sandusky Bay. However, all the regression plots show linear trends without curvature.

The effect of CDOM on chlorophyll $a$ may have been negligible because in-situ chlorophyll $a$ measurements were made using the HACH submersible flourometer, which is an active instrument that detects chlorophyll $a$-induced emission at 685nm. This emission is attributed to the process of inelastic scattering. The 685nm wavelength region is minimally affected by CDOM absorption and hence an ANN-trained based on HACH chlorophyll $a$ measurements makes it less susceptible to effects of CDOM.
In the case of TSM, effects of CDOM may be resolved because TSM is estimated from the gravimetric analysis of a filter residue and the dissolved organic matter may potentially pass through the filter (0.47µm) as a filtrate, without significant retention. Therefore, the ANN is trained using TSM values that have negligible amount of CDOM.

The plots for TSM show that most of the data is confined to low-level concentrations, with only few data points plotting at higher concentrations. This is actually influenced by the distribution of the pre-selected sampling stations. As discussed earlier in the chapter, most of the sampling stations are in the open waters of the WBLE where terrestrial influence is relatively low. The high TSM values represent samples taken from within Sandusky Bay and from stations 3 and 4. These sampling points are much closer to mouths of major rivers such as the Sandusky, Toussaint and Portage, where significant loads of TSM are continuously discharged. Moreover, the sampling campaign started in late spring, a period when river discharge increases due to snow melting and precipitation is relatively high.

Figures 5.7a and 5.7b show the variability of the predicted CPA concentrations across the WBLE. The figures below show the close relationship between the observed data and the predictive values for both biophysical parameters. These plots illustrate that the ANN algorithm did well in picking up the optical variability that is observed between the shallow waters of Sandusky Bay (stations 17, 19 and 20) and the relatively deeper waters in the central region of the WBLE. This is especially true of the concentrations of TSM.
Figure 5.7. Observed and predicted (Prd) values for chlorophyll $a$ (5.7a), and TSM (5.7b) at 18 stations across the WBLE.
The efficiency of the ANN algorithm in capturing local optical variability may also be influenced due to a number of other factors including: (a) mismatch in spatial resolution between MERIS data (pixel resolution is 290 by 290m) and the ground data which is a point data, (b) mismatch in the time of data acquisition between the two platforms, during which optical variability may change, especially in fast circulating bodies such as the WBLE, and (c) effects of path radiance on the MERIS data, which may mask detection of optically active constituents that cause variability in the water color.

5.5 Conclusion

The WBLE of Lake Erie represents optically complex aquatic environment where multiple color producing agents dictate the optics of the water. In this study, a neural network approach is applied to develop a regionally specific algorithm for retrieval of chlorophyll a and TSM from normalized water leaving reflectance that is recorded on MERIS sensor. Regional algorithms such as this use satellite sensors that could provide a promising tool for quantitative monitoring and studying harmful algal blooms in the WBLE at high spatial and radiometric resolution.

The best-performing regional ANN is a single hidden layer network with 3 hidden nodes optimized for high predictive accuracy of water quality parameters. The validation of the ANN after training was performed by two techniques: using the Random Holdback and the 8-fold cross-validation, which consists of 71 observation values. This study
showed that this ANN is capable of estimating multiple components of optically active constituents simultaneously. Prediction performances showed correlation values of 0.76 and 0.82 for chlorophyll $a$ and TSM, respectively. The residual mean errors of the predictive models were less than a percentage of the average raw concentration values, for both parameters.
CHAPTER 6

Summary and Future work
6.1 Summary

Many inland and coastal aquatic systems are quickly responding to multiple environmental stressors such as urbanization, agricultural practices, and climate changes. These stressors will continue to increase with the increasing population density, placing the health of ecosystems at risk. Lake Erie is one of Ohio’s most valuable natural resources that provide water for drinking, industry, recreation and fishing. The lake generates about a 10.5 billion dollars annually from tourism which accounts for about a third of Ohio’s tourism dollars (Great Lakes Commission, 2011). The lake is surrounded by terrestrial stressors and its water quality has deteriorated affecting local ecosystems (Boyer, 2008; Becker et al., 2009). This is continuing to have negative impacts on human health and the local economy. Water quality monitoring is essential for guiding appropriate measures to improve and sustain the health of the ecosystem.

Traditional methods of field-based water quality monitoring generated only sparse point data, hampering understanding of the dynamics of water quality parameters. The color of naturally occurring waters is often considered as an indicator of its quality. Therefore, the optical sensors may be used to determine many important physical, chemical and biological characteristic of water. In recent years satellites have been used to carry out optical monitoring of aquatic systems using bio-optical models developed empirically and semi-empirically. An extensive literature shows that most of the spectral approaches use spectral indices which employ ratios of reflectance measured in narrow bands of the visible and near-infrared portion of the spectrum. There is little agreement
among researchers on the optimal wavelength band and modeling approaches. The limitation of previous methods calls for the evaluation of novel full waveform spectral analytical approaches.

The aim of this research was to develop robust satellite algorithms to estimate color producing agents in optically complex waters by taking a different approach from the traditional spectral indices. The specific objectives of the research were to: (a) evaluate the concentrations and compositions of Color Producing Agents (CPAs) in the Western Basin of Lake Erie (WBLE), (b) parameterize absorption features that can be used to distinguish among selected CPAs, (c) develop algorithms for retrieval of water quality parameters from VIS/NIR spectroscopy, and (d) assess the performance of MODIS and MERIS satellite sensors in estimating CPA proxies.

In chapter 1, the optical signature of the WBLE is presented, indicating the presence of multiple optically active constituents including phytoplankton, suspended material and CDOM making the WBLE an optically complex aquatic environment. The optical complexity of the WBLE categorizes it as Case 2 type water according to Morel and Prieur (1977) optical classification system. The composition and concentrations of CPAs in the WBLE are discussed in Chapter 2. Chapters 2 to 5 of the dissertation are written in manuscript format, each chapter, with the exception of chapter 2, applied a new technique to develop the remote sensing algorithms for estimating the various in-water constituents. Chapter 2 evaluates the potential of existing band ratio methods of retrieving CPAs in the WBLE using satellite data. In Chapters 3 and 4, a hyperspectral
data generated using a lab-based spectrophotometer was used to assess the efficiency of the spectral analytical techniques in the retrievals of CPAs. In chapter 5, satellite based multi-spectral data are employed in the retrievals of selected CPAs using non-linear functions. In this study, chlorophyll $a$, phycocyanin and total suspended matter (organic and inorganic) are selected as important CPAs that need to be analyzed to understand the water quality issues of the WBLE.

6.1.1 Spectral indices

In chapter 2, several existing empirical and semi-empirical chlorophyll $a$ algorithm were assessed using two ocean satellites (MODIS and MERIS). Results showed that the blue-green algorithm did not perform well in the WBLE due to interference from CDOM. The effects of CDOM exponentially decreases towards the longer wavelengths (NIR region) and application of the NIR-red spectral indices showed relatively high $R^2$ values for chlorophyll $a$ estimation. Among the existing semi-empirical and empirical algorithms (Gons, 1999; Gower et al., 1999; Moses et al., 2009), the NIR-red based Simis et al. (2005) algorithm performed better in estimating the chlorophyll $a$ in the WBLE. This is attributed to addition of multiple correction factors to account for effects from other optically active constituents (i.e., suspended sediments and phycocyanin).

The Performance of the MODIS and MERIS satellite sensors in retrieving in-water constituents in optically complex waters was assessed. Regression analysis indicated that MERIS performs better than the MODIS in Case 2 waters. MERIS has a significant
advantage over MODIS with respect to the estimation of chlorophyll $a$ concentration due to its possession of a narrow spectral band at 708nm and its higher spatial resolution (260 m × 290m compared to 1km× 1km for MODIS). However, this study has clearly outlined some of the limitations associated with using space-based platform sensors. Some of the challenges include: (a) modification of water leaving radiance due to path radiance effects, which in turn affects the slope and intercept factors of model regressions, (b) optical variability within individual satellite pixels, especially in dynamic conditions such as those found in the WBLE, and (c) optical variability due to the time gap between satellite overpass and the in-situ sampling.

6.1.2 Principal Component Analysis

Chapter 3 discussed the application of Principal Component Analysis (PCA) on lab-based hyperspectral VIS/NIR spectrometer data. This technique separated the reflectance data into three Principal components accounting for about 88% of the optical variability observed in samples collected from the WBLE. The first two factors, which account for 42 and 31% of the variability, represented prokaryotic blue-green algae and eukaryotic diatoms, respectively. The third factor arose from a complex association of cyanobacteria and inorganic suspended sediments. Plots of factor scores at each station indicated that factors 1 and 3 were dominant in areas which were heavily influenced by river influx (i.e., in Sandusky Bay and at stations 3 and 4). This suggests that some of the cyanobacteria and the inorganic sediments observed in the Sandusky Bay and at stations 3 and 4 are terrestrially derived. The central region of the WBLE was dominated by the
diatoms. Despite the coarse temporal resolution of our sampling campaign, optical variations were detected between late spring (June-July) and late summer (September). This is mainly attributed to the high river discharge which is the result of snow melt and increased precipitation. Analysis showed that, by September, inorganic sediments originating from the terrestrial environment disperse further into the lake and become equally important as the phytoplankton in characterizing the lake’s optical properties. PCA of hyperspectral data from September, 2009 showed that the suspended sediments associated with cyanobacteria appeared as the first principal factor, accounting for about 36% of the optical variability. In contrast, in earlier in spring (June, 2009) the PCA showed Cyanobacteria was the only constituent that appeared as the first principal factor, accounting for 54% of the variations.

Application of PCA technique has several advantages: (a) reduced the dimension of the hyperspectral data, this significantly improved computation efficiency and provided insight to the most important features within the dataset, b) eliminated multicollinearity problems that existed among the independent variables, and c) improved signal to noise ratio by sequestering noise in low order principal components. The Principal component regression was able to explain about 70 and 75 % of chlorophyll \( a \) and TSM variability, respectively, in the WBLE.
6.1.3 Wavelets

Another powerful mathematical tool that is applied in this work (chapter 4) is wavelet analysis. This technique has the ability to perform Multi-Resolution Analysis (MRA) based on discrete wavelet decomposition of the full spectral range. The technique allows local frequency analysis of the spectrum. This effectively zooms in on each part of the spectrum to analyze the local frequency variability. This method precisely targeted the scale and position of optical features that characterize chlorophyll $a$, phycocyanin and TSM. Multiple iterations using different wavelet functions produced consistently high $R^2$ values, illustrating the robustness of the wavelet functions in estimating CPAs in optically complex waters. Application of wavelets clearly demonstrated the fact that single frequency analysis does not fully characterize in-water constituents, a deviation from current thinking regarding the nature of optical features of CPAs. This tool has several advantages including: (a) data-noise reduction, (b) matrix dimensionality reduction, c) signal scaling and localization properties, and (d) fast signal transformation.

6.1.4 Neural Networks

One of the most promising achievements in this study was the development of an Artificial Neural Network (ANN) algorithm for retrieval of CPA concentrations from MERIS observations of the WBLE. A statistical analysis was performed to determine the optimal design of the network. The best performing network has three layers: an input layer consisting of eleven MERIS bands and geometric information to account for
radiance variability due to solar zenith angle, a hidden layer with three processing neurons, and an output layer with two nodes (chlorophyll $a$ and TSM). The ANN-based model was able to explain 75 and 82% of the chlorophyll $a$ and TSM variability. A calibration-validation procedure that was performed on the ANN models showed that the models are stable. Statistical measures such as the Mean Residuals (MRES) and Root Mean Square Errors (RMSE) indicated that calibration and validation parameters are similar. The ANN can provide a tool for future monitoring and prediction of Lake Erie at a high spatial and temporal resolution. The technique used is novel in applying a single ANN that use a non-linear transfer function to predict two CPAs (chlorophyll $a$ and TSM) from the full resolution MERIS data.

6.2 Model performance

The strengths of the various models developed in this study were assessed by performing calibration-validation. These analyses demonstrate the stability of the models and their capability in predicting future events using new datasets. Assessments showed that algorithms that are built using the full range of the spectrum produced validation parameters consistent with calibration results, indicating that these algorithms are relatively stable.

However, validation parameters from the empirical and semi-empirical band ratioing algorithms were relatively poor. These algorithms had large RMSE and MRES values, with significant deviations between the calibration and validation parameters.
Moreover, the biases of the models with respect to the 1:1 line were relatively high for models that were based on the spectral indices (Table 6.1). This suggests that the use of full spectral information yields more stable algorithms than the conventional methods that usually make use of only a few narrow spectral bands. The PCA, wavelet and ANN techniques use the full range of spectral information in retrieving the various in-water constituents.

<table>
<thead>
<tr>
<th>Chlorophyll a</th>
<th>Slope</th>
<th>Intercept</th>
<th>R²</th>
<th>RMSE%</th>
<th>Bias%</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue-green</td>
<td>0.43</td>
<td>3.91</td>
<td>0.46</td>
<td>20.30</td>
<td>95.25</td>
<td>71</td>
</tr>
<tr>
<td>NIR-2 band</td>
<td>0.62</td>
<td>2.56</td>
<td>0.61</td>
<td>15.50</td>
<td>23.46</td>
<td>71</td>
</tr>
<tr>
<td>NIR-3 band</td>
<td>0.57</td>
<td>2.90</td>
<td>0.57</td>
<td>14.58</td>
<td>23.96</td>
<td>71</td>
</tr>
<tr>
<td>Gons</td>
<td>0.28</td>
<td>6.19</td>
<td>0.61</td>
<td>14.70</td>
<td>23.41</td>
<td>71</td>
</tr>
<tr>
<td>Sims</td>
<td>1.43</td>
<td>-2.42</td>
<td>0.65</td>
<td>14.10</td>
<td>21.05</td>
<td>71</td>
</tr>
<tr>
<td>PCA</td>
<td>0.72</td>
<td>2.05</td>
<td>0.70</td>
<td>11.50</td>
<td>14.64</td>
<td>89</td>
</tr>
<tr>
<td>Wavelets</td>
<td>0.86</td>
<td>0.98</td>
<td>0.86</td>
<td>9.70</td>
<td>9.98</td>
<td>89</td>
</tr>
<tr>
<td>ANN</td>
<td>0.74</td>
<td>1.78</td>
<td>0.76</td>
<td>8.40</td>
<td>10.70</td>
<td>71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TSM</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>0.77</td>
<td>2.14</td>
<td>0.75</td>
<td>32.00</td>
<td>34.65</td>
<td>89</td>
</tr>
<tr>
<td>Wavelets</td>
<td>0.87</td>
<td>0.74</td>
<td>0.88</td>
<td>15.00</td>
<td>22.26</td>
<td>89</td>
</tr>
<tr>
<td>ANN</td>
<td>0.83</td>
<td>1.54</td>
<td>0.82</td>
<td>30.00</td>
<td>30.02</td>
<td>71</td>
</tr>
</tbody>
</table>

Table 6.1. Performance of the various applied models in the retrieval of CPAs. The Bias% is the RMSE bias relative to a 1:1 regression line. All of the values are statistically significant at 0.05 significant levels.
6.3 Best model selection

The dissertation study has demonstrated that algorithms developed using the PCA, and wavelet techniques produced stable models that gave high $R^2$ and low RMSE values for estimating the various in-water constituents in WBLE. In this study, the Akaike’s Information Criteria (AIC) is used to assess performance within this group of methods (Burnham and Anderson, 2002). AIC seeks a model that has highest predictive accuracy with the fewest parameters. In this study, the RMSE was used to compare between models independent of the size of the dataset.

For linear standard models, the AIC is mathematically defined as:

$$AIC = n\log(RMSE^2) + 2M$$

Where $n$ represents the number of data, RMSE is the root mean square error and, $M$ is the number parameters used. The AIC includes a bias correction ($M$) to impose a penalty for increasing the number of parameters.

The first term on the right-hand side tends to decrease as more parameters are added to the approximating model, while the second term ($2M$) gets larger as more parameters are added to the approximating model. This is a tradeoff between bias and variance explained. This is, essentially, a tradeoff between underfitting and overfitting the data during model development (Burnham and Anderson, 2002). Lower values of the AIC indicate the preferred model, that is, the one with the fewest parameters that still provides an adequate fit to the data.
Table 6.2 shows the comparison of the AIC values among the techniques that were used to develop bio-optical algorithms in the WBLE. The wavelet technique produced relatively low AIC values for both CPAs. This indicates that wavelet functions are more ideal candidates for developing robust algorithms for the retrieval of color producing agents in optically complex waters.

<table>
<thead>
<tr>
<th>Chlorophyll a</th>
<th>N</th>
<th>RMSE</th>
<th>M</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wavelets</td>
<td>89</td>
<td>0.67</td>
<td>3</td>
<td>-24.9</td>
</tr>
<tr>
<td>PCA</td>
<td>89</td>
<td>0.86</td>
<td>3</td>
<td>-5.7</td>
</tr>
<tr>
<td>TSM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wavelets</td>
<td>89</td>
<td>1.32</td>
<td>3</td>
<td>27.5</td>
</tr>
<tr>
<td>PCA</td>
<td>89</td>
<td>2.86</td>
<td>3</td>
<td>87.2</td>
</tr>
</tbody>
</table>

Table 6.2. Akaiik’s best model selection criteria for chlorophyll a and TSM estimation
6.4 Future directions

Advances in aquatic remote sensing are expected to be more heavily focused on improving satellite retrieval products of optical properties of coastal waters. These waters are optically complex offering more challenges than open marine waters where satellite observations and retrieval algorithms such as the blue-green algorithms are already reasonably effective. Therefore, new measurement and retrieval techniques and algorithms for coastal waters is important for characterizing the optical environment connected with coastal waters, which are of importance because of their ecological diversity and their susceptibility to anthropogenic impacts. As evidenced from this work, hyperspectral remote sensing data provides better spectral resolution to isolate and characterize single in-water constituents in Case 2 waters.

Recently, a new generation of satellites such as the Hyperspectral Imager for Coastal Ocean (HICO) has been launched. These sensors will provide hyperspectral data across the full range of the ultraviolet to infrared spectrum (350 to 2500nm) with 100m spatial resolution. With hundreds of narrow spectral bands, this data will contain enough detail to allow spectroscopy to be applied for understanding of optical properties of turbid waters. The extension of aquatic remote sensing to hyperspectral sensors is a major improvement for CPA retrieval in Case 2 water as well as for developing more robust atmospheric correction algorithms. One problem that will accompany hyperspectral data is the complexity of working with such high-dimensional data.
This analysis of mathematical transformation techniques (PCA, wavelet and NN) on multi- and hyperspectral reflectance data has revealed that these techniques can produce results that are superior to existing ratio-based spectral analysis approaches. Moreover, the PCA and wavelet techniques have proven to be efficient data reduction methods that can overcome the complexity associated with hyperspectral data. These aspects provide justification for further work with these techniques, particularly in the context of testing the robustness and developing globally applicable CPA retrieval algorithms from hyperspectral remote sensing data. A refinement that can be made to the wavelet techniques is the development of automated approaches for the selection of appropriate wavelet basis functions.

It will also be valuable to test effects of the mixture of biogeochemical constituents, differing sensor characteristics, within pixel heterogeneity, and atmospheric effects on the robustness of the transformation techniques. Finally, to develop accurate optical retrieval models, it is crucial to perform calibration-validation on the present and future hyperspectral sensors. Some of the in-situ instruments that can be employed for sensor calibration purpose include: absorption and attenuation meters (e.g., acs), Laser In-Situ Scattering and Transmissometry-100X, and submersible water quality sensor (e.g., Hydrolab).
REFERENCES


Frouin, R. Atmospheric correction of satellite ocean-color imagery in the presence of sunglint, sea ice, and clouds.


IOCCG. (2008). *Why ocean color? the societal benefits of ocean-color technology* No. 7

IOCCG. (2000). *Remote sensing of ocean colour in coastal, and other optically-complex, waters* No. 3


Smith, L. I. (2002). A tutorial on principal components analysis. *Cornell University, USA, 52*


USEPA. (2009). *Status of nutrients in the Lake Erie basin*


