2013

Integrated Data Fusion And Mining (idfm) Technique For Monitoring Water Quality In Large And Small Lakes

Benjamin Vannah
University of Central Florida

INTEGRATED DATA FUSION AND MINING (IDFM) TECHNIQUE FOR MONITORING WATER QUALITY IN LARGE AND SMALL LAKES

by

BENJAMIN VANNAH
B.S. Georgia Institute of Technology, 2009

A thesis submitted in the partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Civil, Environmental and Construction Engineering
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2013

Major Professor: Ni-Bin Chang

ABSTRACT

Monitoring water quality on a near-real-time basis to address water resources management and public health concerns in coupled natural systems and the built environment is by no means an easy task. Furthermore, this emerging societal challenge will continue to grow, due to the ever-increasing anthropogenic impacts upon surface waters. For example, urban growth and agricultural operations have led to an influx of nutrients into surface waters stimulating harmful algal bloom formation, and stormwater runoff from urban areas contributes to the accumulation of total organic carbon (TOC) in surface waters. TOC in surface waters is a known precursor of disinfection byproducts in drinking water treatment, and microcystin is a potent hepatotoxin produced by the bacteria *Microcystis*, which can form expansive algal blooms in eutrophied lakes. Due to the ecological impacts and human health hazards posed by TOC and microcystin, it is imperative that municipal decision makers and water treatment plant operators are equipped with a rapid and economical means to track and measure these substances.

Remote sensing is an emergent solution for monitoring and measuring changes to the earth's environment. This technology allows for large regions anywhere on the globe to be observed on a frequent basis. This study demonstrates the prototype of a near-real-time early warning system using Integrated Data Fusion and Mining (IDFM) techniques with the aid of both multispectral (Landsat and MODIS) and hyperspectral (MERIS) satellite sensors to determine spatiotemporal distributions of TOC and microcystin. Landsat satellite imageries have high spatial resolution, but such application suffers from a long overpass interval of 16 days. On the other hand, free coarse resolution sensors with daily revisit times, such as MODIS, are incapable of providing detailed water quality information because of low spatial resolution. This issue can be resolved by using data or sensor fusion techniques, an instrumental part of IDFM, in which the high spatial resolution of Landsat and the high temporal resolution of MODIS imageries are fused and analyzed by a suite of regression models to optimally produce synthetic images with both high spatial and temporal resolutions. The same techniques are applied to the hyperspectral sensor MERIS with the aid of the MODIS ocean color bands to generate fused images with enhanced spatial, temporal, and spectral properties. The performance of the data mining models derived using fused hyperspectral and fused multispectral data are quantified using four statistical indices. The second task compared traditional two-band models against more powerful data mining models for TOC and microcystin prediction. The use of IDFM is illustrated for monitoring microcystin concentrations in Lake Erie (large lake), and it is applied for TOC monitoring in Harsha Lake (small lake). Analysis confirmed that data mining methods excelled beyond two-band models at accurately estimating TOC and microcystin concentrations in lakes, and the more detailed spectral reflectance data offered by hyperspectral sensors produced a noticeable increase in accuracy for the retrieval of water quality parameters.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Remote sensing is the collection of information from a target object or event from a distance. The sensors for observation can be used from ground, air, and space-borne installations to monitor the environment and observe ecological changes occurring over time. Space-borne sensors possess the ability to monitor large spatial extents at a high spatial or temporal frequency. This presents a unique advantage for monitoring the environment, which is in a constant state of flux due to anthropogenic influences. Surface waters in the 20th were subject to rapid changes and contaminates due to population rise, economic development, and climate change. This brings forth the emerging challenge of monitoring the health of surface waters in urban areas, which can be degraded by excess nutrients in wastewater, residential runoff, and agricultural runoff; suspended solids in runoff, and additional pollutants that alter water quality. Manual sampling of the waters has been the traditional method to garner water quality information, yet it is a costly, time-consuming, and tedious approach. The use of remote sensing to assess water quality is quick and economical approach, and its accuracy at predicting the spatiotemporal distributions of water quality parameters in optically complex waters is constantly improving.

Remote sensing of water quality parameters is fundamentally rooted in the detection of electromagnetic radiation from the target. This is made possible by the fact that all substances absorb, transmit, and reflect electromagnetic waves. However, the amount of light that is reflected or absorbed is a function of wavelength. For example, chlorophyll-a is known for its predominantly green hue, which indicates that it strongly reflects light from the “green portion” of the electromagnetic spectrum. On the other hand, a blue object would reflect electromagnetic waves from the blue portion of the spectrum. This simplified example illustrates how objects with differing spectral properties can be identified and differentiated based upon their spectral reflectance curves, aka. their propensity to reflect electromagnetic waves at specific frequencies. Monitoring case 1 waters is relatively straightforward, since their optical properties are primarily determined by phytoplankton and other biological organics. Case 2 waters are known for their optical complexity resulting from the presence of suspended solids, colored dissolved organic matter, phytoplankton, and additional substances. The real challenge is deciphering the spatiotemporal distribution of individual water quality parameters when the observed spectral curve is a multifaceted function of the various constituents contained in the water.

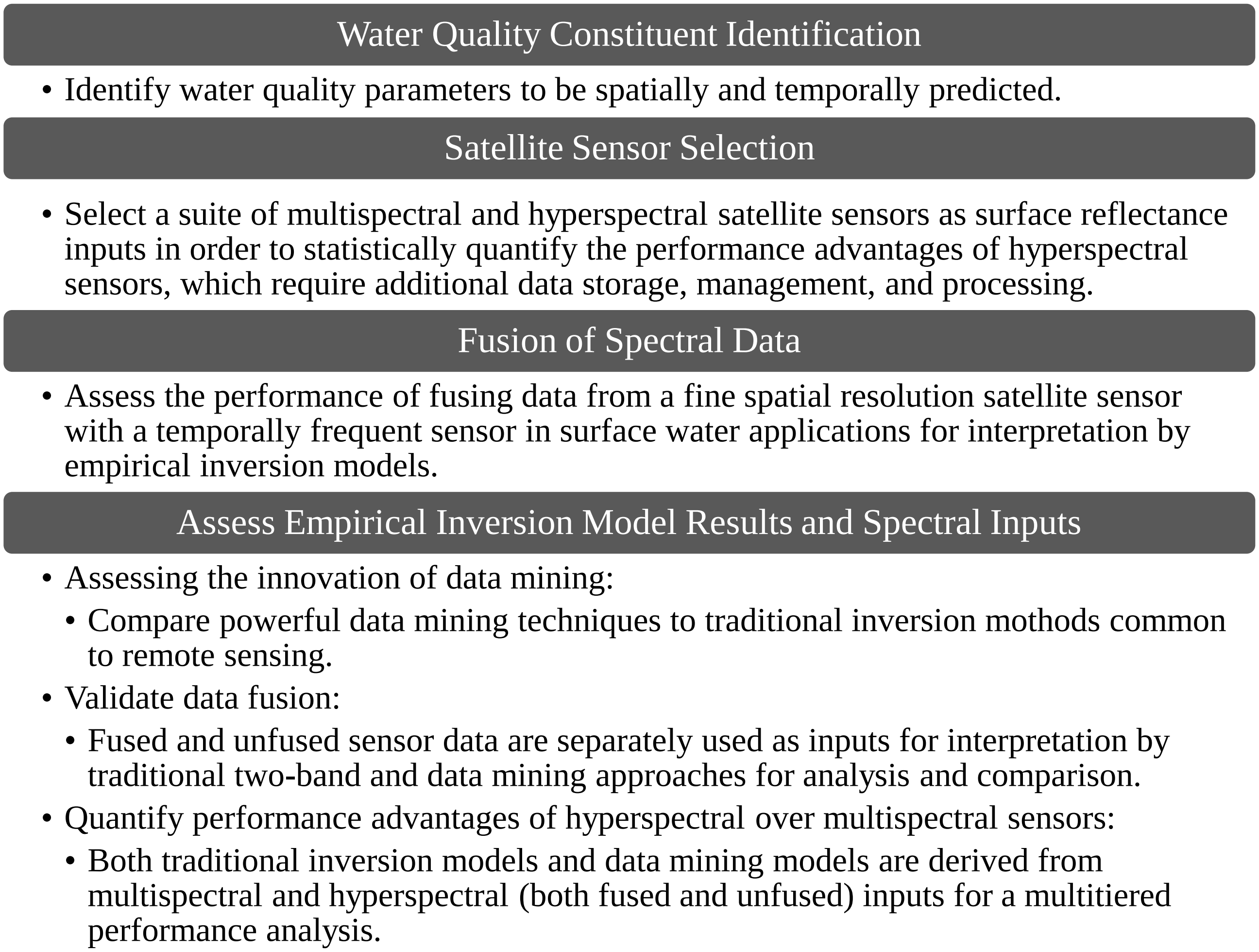
1.1. Research Objectives

The question remains for how to develop a cost-effective, daily spatiotemporal monitoring system of microcystin and TOC concentrations in a water body to achieve a successful early warning system. The overall goal of this research was to develop a robust framework for monitoring water quality constituents on a daily basis in inland water bodies of various sizes. This robust and innovative technological development for determining concentrations and distributions of potentially hazardous water quality constituents would have strong benefits for the following users: 1.) drinking water treatment plants, 2.) commercial fisherman, 3.) recreational users, and 4.) municipal decision makers assessing total maximum daily loads (TMDL). The resulting methodology designed for this task is the Integrated Data Fusion and Mining (IDFM) technique, and 3 specific research objectives were formulated:

1. apply data or sensor fusion techniques, in which high spatial resolution satellite sensors are fused with high temporal resolution sensors into a single data stream possessing enhanced spatial and temporal properties;
2. develop advanced forecasting models that link the data mining techniques with ground-truth databases and fused spectral inputs for TOC and microcystin prediction;
3. apply the modeling technologies in both large and small inland water bodies to evaluate these techniques in developing microcystin and TOC concentration maps for a functional early warning system.

1.2. Thesis Organization

This thesis is organized into two (2) standalone chapters featuring applications of the IDFM technique. Thesis organization from a research objective point of view is visually depicted in Figure 1.1:



Water Quality Constituent Identification

- Identify water quality parameters to be spatially and temporally predicted.

Satellite Sensor Selection

- Select a suite of multispectral and hyperspectral satellite sensors as surface reflectance inputs in order to statistically quantify the performance advantages of hyperspectral sensors, which require additional data storage, management, and processing.

Fusion of Spectral Data

- Assess the performance of fusing data from a fine spatial resolution satellite sensor with a temporally frequent sensor in surface water applications for interpretation by empirical inversion models.

Assess Empirical Inversion Model Results and Spectral Inputs

- Assessing the innovation of data mining:
  - Compare powerful data mining techniques to traditional inversion mothods common to remote sensing.
- Validate data fusion:
  - Fused and unfused sensor data are separately used as inputs for interpretation by traditional two-band and data mining approaches for analysis and comparison.
- Quantify performance advantages of hyperspectral over multispectral sensors:
  - Both traditional inversion models and data mining models are derived from multispectral and hyperspectral (both fused and unfused) inputs for a multitiered performance analysis.

Figure 1.1: Thesis organization from a research objective point of view.

Chapter 2 investigates the use of IDFM techniques for predicting TOC spatiotemporal distributions in a small lake. Harsha Lake is located in the East Fork of the Little Miami River in Clermont County, Ohio, and it covers an area of 8.739 km^2. The objectives of the case study featuring the small lake were: 1.) quantify the advantages of using multispectral sensors versus hyperspectral sensors and 2.) assess the performance of a suite of traditional inversion models against innovative and complex data mining techniques for TOC prediction. The multispectral data set was provided by the Landsat TM/TM+ and Terra MODIS sensors, while the hyperspectral data set was supplied by Envisat MERIS. A traditional two-band ratio inversion model was tested against a genetic programming model. Concentration maps depicting the spatiotemporal distribution of TOC in Harsha Lake were created using the most successful model. Lastly, seasonal TOC maps were generated to visually identify seasonal trends for TOC levels and distribution within the lake.

Chapter 3 applies a similar approach using IDFM for predicting microcystin concentrations and distributions in Lake Erie. This large lake is subject to frequent and expansive algal blooms each summer, and the goal is to delineate between toxic and nontoxic algal blooms covering the lake. Once more, the objective is to assess the performance of IDFM; however, this time it is applied for microcystin prediction in a large lake. The specific objectives are: 1.) quantify the advantages of using multispectral sensors versus hyperspectral sensors and 2.) assess the performance of a suite of traditional inversion models against innovative and complex data mining techniques for microcystin prediction. The multispectral data set was provided by the Landsat TM/TM+ and Terra MODIS sensors, while the hyperspectral data set was supplied by Envisat MERIS. A genetic programming model was tested against two traditional models, a two-band ratio model and a two-band spectral slope model. The model exhibiting the highest performance was used to generate microcystin concentration maps for both the multispectral and hyperspectral spectral inputs.

1.3. Study Limitations

Cloud cover is a primary limitation in both applications of IDFM in this study. Detection of electromagnetic waves reflecting off the water's surface by the sensors relies on the study site being void of clouds. Both Harsha Lake and Lake Erie are subject to heavy cloud cover throughout the year, which would prevent the early warning system from functioning when the study site is visually obscured. The physical reason for this limitation is attributed to the portion of the electromagnetic spectrum detected by the sensors. The visible and infrared frequencies reflected off the water are unable to pierce through thick fog, clouds, and smoke. One possible workaround is to use spectral sensors mounted on airplanes for monitoring sites, instead of space-borne sensors. While this method is more costly, it is a reliable method of data collection on days when the view of the water body is obstructed from space. Alternatively, the inclusion of microwave sensors capable of piercing through cloud cover is a possibility to be explored. Similarly, a second limiting scenario posed by clouds is when the site is only partially exposed to cloud cover. This makes for less accurate fused images by the STAR-FM data fusion algorithm, because the obscured areas cannot be used for fusion.

The data mining methods used in this study are all used to develop empirical models. Often, these models are data heavy. As a result, these models rely on a robust set of ground-truth data that accurately depict all water quality conditions experienced in the lake throughout the year. Furthermore, surface reflectance observations should be available for a comprehensive range of concentrations for the water quality parameter. Otherwise, the model may not accurately predict concentrations outside of this range.

It should be understood that objects, pollution plumes, and algal blooms smaller than the spatial resolution of the satellite sensor may not be observable. For example, the MODIS sensor used in this study has a spatial resolution of 250 m for bands 1 and 2. The spectral signatures of all objects at the water's surface within a 250 m box will be averaged into the observed reflectance for that pixel. If an algal bloom encompasses the majority of the pixel, then the spectral reflectance signature of the bloom will dominate. However, a mix of small toxic blooms and pristine water in the 250 m search box will result in a lower microcystin concentration being predicted, since the poor resolution of the satellite prevents it from delineating between small algal blooms and clean water. Wind is also an influence that can alter the estimation of microcystin. The sheer forces from wind cause algal blooms at the surface to mix with the water below. Consequently, the bloom becomes vertically dispersed throughout the water column, and the spectral reflectance from the bloom at the water's surface is diminished.

CHAPTER 2: INTEGRATED DATA FUSION AND MINING TECHNIQUES FOR MONITORING TOTAL ORGANIC CARBON CONCENTRATIONS IN A LAKE

2.1. Introduction

Total organic carbon (TOC) is a gross water quality parameter comprised of particulate, colloidal and dissolved organic matter, which can include floating vegetative and animal matter and volatile organic matter (GEAS 1994). Disinfection byproducts (DBPs) are formed when organic matter reacts with oxidants, such as free chlorine, in drinking water disinfection processes. Disinfection byproducts in finished water of a water treatment plant often include trihalomethantes (THMs) and haloacetic acids (HAAs). The surface water treatment rules managed by the United States Environmental Protection Agency (US EPA) require water disinfection in treatment plants in order to protect the consumers from microbiological contaminants, and at the same time restrict the concentration of disinfection byproducts such as THMs and HAAs, below the levels that are harmful to human health (USEPA 1998).

Manual water sampling and laboratory TOC measurement of source water is the widely used practice, which has a substantial cost and a significant time delay (O’Connor 2004). For large TOC changes in source water, the time delay and a lack of information on TOC spatial distribution can impede the preparation of operational and engineering adjustment in drinking water treatment. This is the impetus to develop the remote sensing methods in this study for an accurate TOC estimation on a daily basis. The capabilities of the Integrated Data Fusion and Mining (IDFM) techniques may offer daily monitoring for examining water quality conditions to present a series of near-real-time TOC estimations as an early warning system around the water intake and associated water body.

Satellites’ sensors can detect the surface reflectance emissions with medium to high resolution images of selected light spectrums, which can be used to estimate the level of TOC concentrations present in water using inversion models and machine learning techniques (Smith and Baker 1978, Stramski *et al.* 1999, Stedmon *et al.* 2000, Ohmori *et al.* 2010). The individual Landsat and MODIS sensors have their own stand-alone disadvantages for CDOM detection in water, such as long revisit times, low spatial resolution, and inability to sense the majority of the reflected light at the CDOM peaks. One of the possible solutions to overcome this barrier is to incorporate the best qualities of MODIS and Landsat data streams into a composite image with both high temporal and spatial resolution. This was done through the use of data fusion algorithms, as discussed in this paper.

The question remains for how to develop a cost-effective, daily spatiotemporal monitoring system of TOC concentrations in water body to fulfill its early warning functionality. The objective of this study is thus to provide a near-real-time monitoring system measuring the spatiotemporal distributions of TOC concentrations in a lake through the use of the proposed IDFM techniques. In this paper, we wish to explore: 1) the feasibility of determining the spatiotemporal distributions of TOC within a lake using fused image reflectance bands and machine learning algorithms, 2) if there is a justifiable advancement in using fused images from Landsat and MODIS to provide more accurate estimation of TOC concentrations than that done by using MODIS or Landsat images alone, 3) how the performance of genetic programming (GP) models based on a machine learning algorithm compares to a traditional two-band ratio model, and 4) the spectral bands that achieved the most frequent use among the GP models.

2.2. Background Information

2.2.1. Total Organic Carbon Sources and Effects

TOC is the measure of organic molecules of carbon in water, and it is the sum of dissolved organic carbon (DOC) and particulate organic carbon (POC). It serves as a key water quality parameter in lakes and reservoirs, due to its effect on pH, redox reactions, bioavailability of metals, and the sorption capacity of suspended solids with regards to hydrophobic organic chemicals (Thurman 1985; Parks and Baker 1997). TOC is introduced to surface waters from both natural and anthropogenic sources. Humic substances and degraded vegetation and animal matter are naturally occurring sources of TOC (Thurman 1985; GEAS 1994; Bayram *et al.* 2011). Anthropogenic sources include fertilizers captured by stormwater runoff and irrigation return flows, pesticides, surfactants, and solvents from sewage treatment plants (Visco *et al.* 2005). While TOC by itself is not the direct cause of human health issues in drinking water treatment, its presence during disinfection operations generates DBPs that are hazardous to humans when consumed.

As is required by the EPA, drinking water treatment plants use disinfection to protect consumers from pathogens. When TOC has not been removed from the source water prior to disinfection using chlorinated agents, the formation of DBPs, specifically THMs and HAAs, occurs. Knowledge of TOC concentrations in the source water enables treatment plant operators to alter the treatment strategy to limit DBP generation. This can include but is not limited to adjusting coagulant feed rates or altering the pH for heightened TOC removal through coagulation (TCEQ 2002). In addition, TOC removal in most drinking water treatment plants may be facilitated with the following processes: coagulation, granular activated carbon adsorption, membrane filtration, and ion exchange. These processes, however, have varying removal efficiencies and vastly different operation and maintenance costs.

Monitoring TOC can be an expensive and time consuming process. Purchasing equipment for on-site measurement of TOC costs between $20 000 and $30 000 USD, while using an outside laboratory for analysis takes 2 to 4 weeks to obtain results (TCEQ 2002). This is a key motivation for using the IDFM method to predict TOC concentrations. IDFM rapidly generates a TOC concentration map for the entire lake; meaning, a TOC concentration value is assigned to every pixel comprising the image of the lake. Another benefit is that the TOC map generation comes at no cost when using freely available data from Landsat and MODIS. This is also an advantage to water treatment plants that are required by law to routinely sample the source water to characterize TOC content. Instead of having to sample 3 weeks in advance to have TOC data available for water quality reports at the end of the month, IDFM enables the treatment plant to account for TOC variations for all cloud-free days throughout the month. Niche applications of IDFM include predicting whether TOC levels have increased after heavy rains and aiding in decision making processes when water treatment plants have multiple reservoirs to obtain source water from.

TOC export into water bodies is dependent upon two main aspects: the type of TOC source and the hydrological transport of TOC from the source (Agren *et al.* 2008). The typical source of TOC is the upper layers of the soil matrix, due to the higher levels of organic matter present at the top while decreasing downward into the soil profile (Thurman 1985). Thus, boreal regions can be a large contributor to TOC in nearby surface waters (Agren *et al.* 2008). Another primary TOC source is wetlands, even though their coverage may only account for a small portion of the lake's watershed (Dosseky and Bertsch 1994). Runoff controls the transport of TOC from terrestrial sites into surface waters. This provides insight into seasonal TOC variations, since rainfall intensity and frequency changes between seasons. During the transition from winter to spring in areas with snowfall, TOC concentrations have been known to drop due to dilution from snowmelt (Parks and Baker 1997). A number of studies conclude that the highest levels of TOC in runoff from boreal areas occur during the late summer and fall as a result of the increased rainfall and temperature (Heikkinen 1989, Naden and McDonald 1989, Ivarsson and Jansson 1994, Scott *et al.* 1998, Agren *et al.* 2008). The first strong rainfall events late in the summer yield high concentrations of TOC, due to microbial interactions with organic matter in the upper soil matrix (Parks and Baker, Agren *et al.* 2008). This is because the formation of soluble organic matter resulting from oxidation and microbial activity increases with temperature (Ivarsson and Jansson 1994, Christ and David 1996, Scott *et al.* 1998). Thus, the soluble organic carbon builds up in the soil, until it is washed away and carried into surface waters during the wet season in late summer and autumn (Scott *et al.* 1998). For this reason, TOC in runoff leading to surface waters is lower during the winter and spring (Argen *et al.* 2008). In conclusion, TOC concentrations will always increase with runoff, but the concentration of TOC present in the runoff is subjected to seasonal temperature fluctuations, exhibiting a buildup of TOC in the soils during the summer and autumn seasons.

2.2.2. Remote Sensing of Organic Carbon

In a remote sensing study conducted by Smith and Baker (1978), particulate organic carbon had an effect on remotely sensed chlorophyll-a readings from phytoplankton. In more recent years, the practice of organic carbon detection through remote sensing has been proven effective by a number of studies (Stramski *et al.* 1999, Stedmon *et al.* 2000, Ohmori *et al.* 2010). This study expands upon the preliminary efforts detailed in Chang and Vannah (2012) for remote TOC retrieval using IDFM. The major differences and advancements carried out in this study are as follows: 1.) the previous study was a brief foray into remote TOC retrieval to assess the feasibility of using IDFM on a small inland lake, 2.) this expanded study approaches the problem of remote TOC prediction using genetic programming (GP) as an alternative retrieval algorithm to artificial neural networks (ANNs), 3.) to justify the use of more complex and computationally intensive machine learning techniques, a performance comparison between the GP model and a traditional two-band ratio model is carried out, 4.) an extended background review has been presented for deepened discussion of the results, and 5.) more lucid explanations of the methodological processes comprising IDFM are given.

TOC can be identified based on its unique spectral response and differentiated from other compounds in the water column. This is because every substance gives off a different spectral signature or pattern, some of which can overlap. For any given substance, the measured reflectance will vary throughout the length of the electromagnetic spectrum, since reflectance is a function of wavelength. When a substance is especially reflective to a specific wavelength of light, a peak will be incurred on the spectral signature graph, because the majority of light at this wavelength rebounded off the object. For this study, the spectral reflectance peaks of interest for TOC are based on published experiments measuring the peaks of chromophoric dissolved organic matter (CDOM). CDOM is the light absorbing fraction of dissolved organic carbon. In a series of three different case studies, the spectral reflectances associated with varying levels of CDOM were measured in over 38 different lakes and reservoirs. The observed spectral peaks are detailed in Table 2.1:

Table 2.1: CDOM spectral peaks as determined from case study analysis.

CDOM Spectral Peaks (nm)	Sampling Locations	Sampling Instrument	Source
570	25	Spectron Engineering SE-590 spectrometer	Menken *et al.*(2005)
550, 571, 670, 710	8	Scanning spectrophotometer	Arenz *et al.*(*199*5)
560, 650—700	5+	Spectron Instrument CE395 spectroradiometer	Vertucci (1989)

It is imperative that satellites with sensors that detect light at bandwidths corresponding to reflectance peaks be chosen when determining TOC concentrations based upon surface reflectance. The ability for MODIS and Landsat to detect the main spectral peaks of CDOM is shown in Figure 2.1:

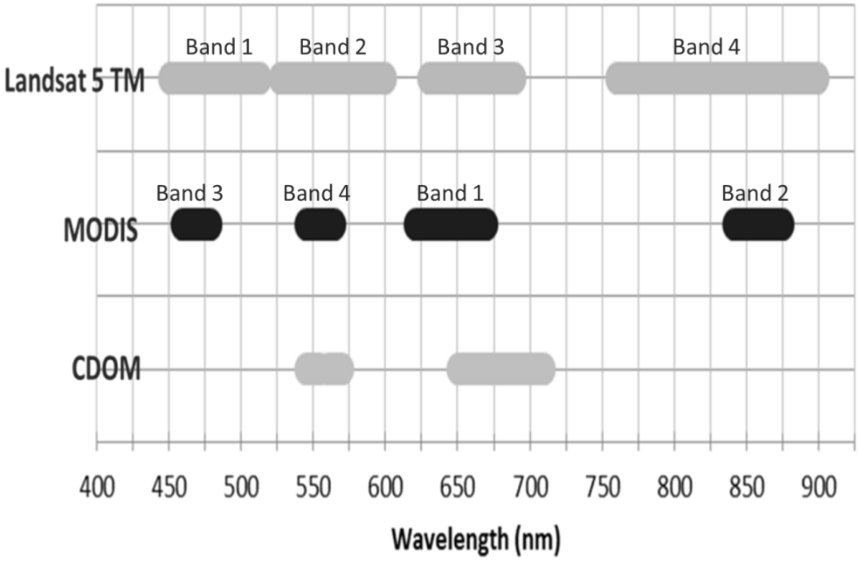


Figure 2.1: CDOM spectral reflectance peaks and the associated MODIS and Landsat sensor capabilities.

Two main spectral peaks for CDOM are marked in gray in Figure 2.. In order for the MODIS and Landsat to determine TOC concentrations, the bands for these satellites must overlap with the CDOM spectral peaks. For the first CDOM peak, it can be seen that the Landsat satellite is able to detect the entirety of the peak, whereas the MODIS satellite can detect the majority of the reflected light caused by CDOM. However, for the second CDOM peak, Landsat can detect the first two-thirds of the CDOM peak, whereas MODIS only the first half. Thus, Landsat not only contributes enhanced spatial information to the fused image, but the third band provides additional spectrally-relevant data for TOC estimation.

According to previously mentioned studies, observed spectral reflectance emissions were linked to organic carbon concentrations in the water. Smith and Baker (1978) and Stramski *et al.* (1999) performed this for particulate organic carbon, and Stedmon *et al.* (2000) conducted a similar analysis for CDOM. In an effort to measure TOC (the dissolved and particulate forms of organic carbon), Ohmori *et al.* (2010) used a spectroradiometer (Opt Research Inc., HSR-8100) to capture the spectral signature of the water body and relate it to TOC. While Ohmori's study is titled "Feasibility Study of TOC and C/N Ratio Estimation from Multispectral Remote Sensing Data," the spectral signatures were obtained at ground level from a spectroradiometer (e.g., C/N ratio stands for carbon to nitrogen ratio). Then, the spectral signatures were manipulated to simulate the bands for the SGLI sensor onboard the GCOM-C satellite. This study expands on the progression of TOC estimation in multiple ways. First, actual remote sensing data from Landsat and MODIS are used. Thus, the inversion and machine learning models are exposed to data that has been radiometrically and geometrically corrected, instead of being retrieved with a handheld spectroradiometer. Secondly, both a traditional two-band ratio model and a GP model are evaluated for TOC prediction by a comparative way. Lastly, the use of standalone MODIS imagery and fused satellite data as inputs for training and validation for the GP model are statistically assessed. Data fusion enhanced the Landsat and MODIS input data spatially and temporally, and the authors hypothesize that this will yield a more precise early warning message than if only MODIS were used for TOC detection spatially.

2.2.3. Data Fusion

Fused images are the algorithmic fusion of the spectral, temporal, or spatial properties of two or more images into a composite or synthetic image possessing the characteristics of the input images (Genderen and Pohl 1994). The fusion of data streams into a single image has the potential to increase the reliability of the data, and displays more of an object's defining attributes at once (Pohl and Genderen 1998). Such benefits can lead to more informed decision making (Hall 1992). There are a number of data fusion techniques available, and selecting an algorithm to apply depends upon the type of output data required for the application, the accuracy of the fused data, and the characteristics of the input data streams that the user would like to fuse. Data fusion techniques are classified into three groups according to the level at which the processing takes place (Pohl and Genderen 1998) including: 1) pixel level, 2) feature level, and 3) decision level. Pixel level image fusion refers to the fusion of the measured physical attributes of the data, prior to significant processing, as shown in Figure 2.2:

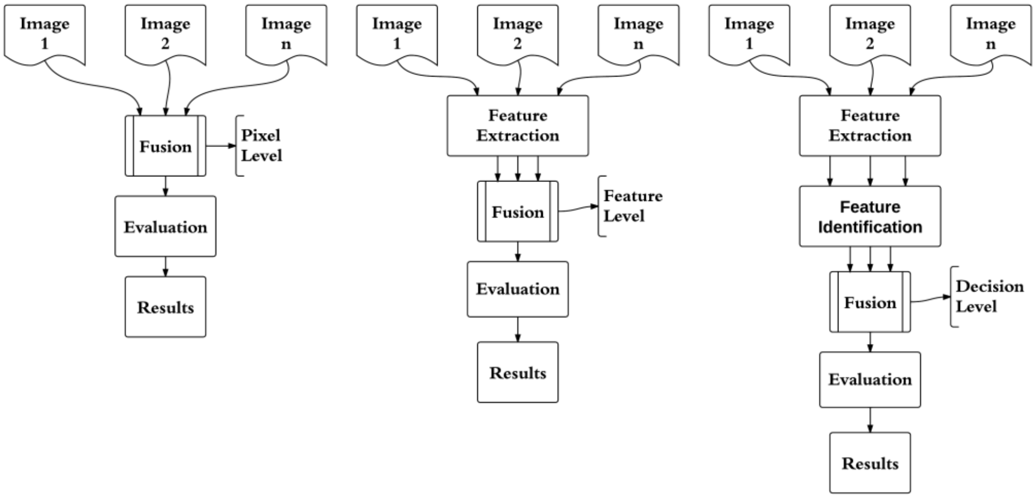


Figure 2.2: Image fusion processing methodologies for pixel level fusion (left), feature level fusion (center), and decision level (right) fusion (adapted from Pohl and Genderen 1998).

Preprocessing steps typically entail radiometric correction, resampling, and reprojection, by which measurement errors are corrected and compatibility is assured between data streams. Image fusion at the feature level takes the measured input data and extracts objects using segmentation procedures (Pohl and Genderen 1998). The classification of objects can be a function of their shape, location, pixel value, and extent (Mangolini 1994). Classified objects from the data streams are then fused in preparation for assessment via statistical methods or Artificial Neural Networks (ANN) (Pohl and Genderen 1998). Decision level fusion, also called interpretation level fusion, takes feature level fusion a step farther by processing the classified data in order to glean additional information, which is then fused according to user-defined decision rules (Shen 1990). Some fusion techniques and their associated fusion level are detailed in Table 2.2:

Table 2.2: Fusion techniques and their associated fusion level.

Fusion Technique	Fusion Level	Enhanced Properties	Source
Color Composites (RGB)	Pixel	Spectral	Pohl and Genderen (1998)
Intensity-Hue-Saturation (HIS)	Pixel	Spectral	Pohl and Genderen (1998)
Principal Component Analysis	Pixel	Spatial	Pohl and Genderen (1998)
Wavelets	Pixel	Spatial	Pohl and Genderen (1998)
High Pass Filtering (HPF)	Pixel	Spatial	Pohl and Genderen (1998)
STAR-FM	Pixel	Spatial/Temporal	Gao *et al.* (2006)
Bayesian Inference	Decision/ Feature	Spatial/Temporal	Robin *et al.* (2005)
Dempster-Shafer	Decision/ Feature	Spatial/Temporal	Yeng *et al.* (2006)

The STAR-FM algorithm performs fusion at the pixel level based upon spectral similarities. It was selected in this study to produce a fused image of enhanced spatial, spectral, and temporal properties, thereby allowing for accurate prediction of TOC concentrations in the water body on a daily basis, assuming cloud-cover is not blocking the area of interest. The spectral reflectance value for each pixel of the MODIS image is a conglomeration of the surface reflectance from each object in the 250 by 250 m area for MOD09GQ and 500 by 500 m for MOD09GA. Alternatively, spectral reflectance values provided by the Landsat image are an average of objects contained within a 30 by 30 m pixel. The extremely coarse spatial resolution of MODIS is the primary justification for performing fusion with Landsat. Integrating the enhanced spatial resolution of Landsat into the fusion process is beneficial for small water bodies due to the supplementary spatial detail. With regard to the fusion process, the STAR-FM algorithm relates the changes between an input MODIS image and another MODIS image taken on the day the synthetic image is being generated on. These changes are then applied to a Landsat image taken on the same date as the MODIS image, which produces the predicted or synthetic Landsat image (Gao *et al.* 2006). The computational time requirement for fusing Landsat and MODIS images for a lake of this size was under 5 seconds for the computer used in this experiment (computer specifications: Intel® Core™ i7-3720QM CPU at 2.6 GHz, 8 192 MB RAM, and 500 GB hard drive). Fusion needs to be performed prior to generating TOC predictions, and processing time is of the essence in any early warning system. After the machine learning model is trained and validated using the fused imagery, it does not need to be retrained, except for periodic updates to reflect the evolution of the water quality characteristics of the water body over time. One caution in the method is noted in fusing the data streams from Landsat and MODIS. Landsat band 1 does not have the same spectral range of MODIS band 1. Instead, band 1 of Landsat corresponds to Band 3 of MODIS and so on. Table 2.3 details the proper band combinations for the fusion of MODIS and Landsat:

Table 2.3: Landsat 5/7 and Terra MODIS band comparisons. Matching the bands for fusion is based upon corresponding band centers, instead of fusion the same band number. For example, Landsat band 1 is fused with MODIS band 3, since they detail similar portions of the electromagnetic spectrum.

Landsat 5 TM Band	Landsat Bandwidth (nm)	Terra MODIS Band	MODIS Bandwidth (nm)
1	450–520	3	459–479
2	520–600	4	545–565
3	630–690	1	620–670
4	760–900	2	841–876
5	1550–1750	6	1628–1652
7	2080–2350	7	2105–2155

As the STAR-FM program translates through the matrix of pixels in the Landsat and MODIS images, it may select a central pixel every few steps and reassign the pixel’s value based upon candidate pixels that are both near the central pixel and spectrally similar. The candidate pixels are then filtered out if they exhibit more change over time than the central pixel, or if the spectral features of the candidate pixel are greater than the difference between the spectral features of the central pixel in the Landsat and MODIS images (Gao *et al.* 2006). Thus, the surface reflectance of the central pixel will be generated from the group of selected candidate pixels. However, the surface reflectance of the candidate pixels is not simply the average value of all surface reflectance values involved. A weighted average is applied based upon how likely each of the selected candidate pixels could represent the central pixel. Higher weighting factors are assigned if the candidate pixel is spectrally and temporally similar to the central pixel, in addition to its geometric distance from the central pixel (Gao *et al.* 2006). Through this entire process, the synthetic Landsat image is generated based on the input candidate pixels in the MODIS image taken during the desired prediction date.

2.2.4. Machine Learning and Genetic Programming

A number of different machine learning techniques can be applied to identify patterns and perform classification or regression within a data set. The IDFM technique uses machine learning to determine the complex relationships between independent variables, specifically surface reflectance, and the concentration of a water quality parameter. The machine learning algorithms can be classified as supervised, unsupervised, and semi-supervised learning algorithms. The first type is supervised learning in which labeled training and validation data sets are analyzed to develop a function linking input data to the target data. Unsupervised learning is most closely related to data mining techniques, since the data has not been preprocessed, and the learning algorithm makes discoveries for clustered data based upon inherent similarities. Lastly, semi-supervised learning uses both unprocessed and processed input data to generate a function explaining the relationship between the input and target values. Notable machine learning techniques with regression capabilities include GP, ANN, and Support

Vector Machining (SVM) (Doerffer and Schiller 2007, Chen *et al.* 2008, Ioanna *et al.* 2011, Song *et al.* 2012, Chang *et al.* 2013, Nieto *et al.* 2013).

GP is a machine learning or data mining method that is one of a class of techniques called "evolutionary computing algorithms" based on the Darwin principles. These algorithms solve problems by mimicking the natural evolutionary processes (Goldberg 1989, Davis 1991). GP can decode system behaviors based on empirical data for symbolic regression in an unsupervised learning fashion and examine observation data sets using association, path analysis, classification, clustering, and forecasting in the context of data mining via all dimensions of the machine learning efforts (Seifert 2004). This is very useful as the user does not need to specify a solution procedure or have prior knowledge of the relationship between the model inputs and the objective. Holland (1975) first developed genetic algorithms (GA), which are the basis of evolutionary computing, and Koza (1992) advanced evolutionary computing by developing the GP techniques that are commonly used today.

The first step of GP is to *initialize* the population by creating a number of programs randomly. The larger the population, the greater the ability to accurately model the problem, yet this requires more computational time and computer memory (Francone 1998). Programs in the population are *evaluated* in order to rank their fitness. If a program meets the minimum error criteria set by the user, then the GP process is complete. However, if the stop criterion has not been achieved, it is necessary to start the next iteration to create an improved generation of programs (Francone 1998, Nieto *et al.* 2013). The new generation is formed by applying 3 principal *search* or *genetic operators* to the better fitting programs in order to replace programs with poor performance. The principal genetic operators are as follows:

- Reproduction: programs of good fit are copied without change into the new generation.
- Crossover/Recombination: Crossover exchanges instructions or nodes within best fit programs to develop a new program. There are three specific types of crossover that can be applied when generating a new program (Engelbercht 2007):
    - Asexual: only one program is used to create the new program
    - Sexual: two programs are used to create the new program
    - Multi-recombination: nodes from more than two programs are used to develop the new program
- Mutation: Random changes are made to the best fit programs, which results in the formation of a new program, and thereby, promote genetic diversity within a population. It should be noted that mutation is applied probabilistically to all programs of best fit, including those that have been selected for crossover.

The Discipulus® software used in this study has the same methodological flow, and the GP algorithm is detailed in the following steps (Francone 1998):

1. Initialize the population with a default starting size of 500.
2. A tournament is run using 4 randomly selected programs out of the population. Based upon their fitness, 2 programs are retained and the other 2 are removed.
3. Principal genetic operations are applied to the winning programs to produce 2 children to replace the losers. The specifics of this process are detailed below:
    a. Copy the 2 winners
    b. Using the default crossover frequency (50%), crossover the copies of the winners
    c. Using the default mutation frequency (95%), mutate one of the copies from step 3a
    d. Repeat step 3c for the other copy produced in step 3a
4. Replace the losers with the offspring produced in step 3
5. Repeat steps 2 through 4 until the run is terminated

Discipulus® uses an instruction set composed of mathematical operations (additional, subtraction, multiplication, division, trigonometry, exponents, and arithmetic), data transfer, stack rotation, and conditions. This means, each instruction can be represented by a mathematical equation or logic function, resulting in a white-box model. If only the mathematical operations are used as instruction sets, then they can be solved in order to generate a single equation.

A primary advantage of GP is that the length of a program is only limited by the memory capabilities of the computer or software. This allows for programs to grow and evolve without constraint until the stop criterion has been achieved. Additionally, little knowledge of the relationships between the input and target values is required when using this supervised learning technique. Unconstrained program size is also a disadvantage of GP, since lengthy programs burden the computer's resources and take a significant amount of time to develop (Francone 1998, Nieto *et al.* 2013). This is because a single program can start with a handful of operations and grow to a candidate solution that is comprised of thousands of operations; yet, the increase in performance can be negligible (Luke 2000). This is known as code bloat in GP. It is a severe issue when dealing with large, complex problems, and developing a GP model under these conditions turns into a race against time to obtain an optimal solution before the search procedure is unduly hampered with code bloat (Luke 2000, Liu *et al.* 2007). For example, in a study by Luke (2000) for evaluating team strategies in the Robocup Soccer Server (Kitano *et al.* 1995), their specific problem was expected to require a full year of evolution time due to its size and complexity, but after taking steps to minimize code bloating, their solution time was cut down to several months. Even though GP develops a white-box model, the complexity of lengthy solutions, which can muddled with code bloat, complicates drawing clear conclusions based upon the white-box model.

2.3. Methods and Materials

2.3.1. Study Site

The William H. Harsha Lake (Figure 2.3) is located in Clermont County, Ohio, roughly 40 km (25 miles) east of Cincinnati, and covers an area of 8 739 360 m^2 (2 160 acres).

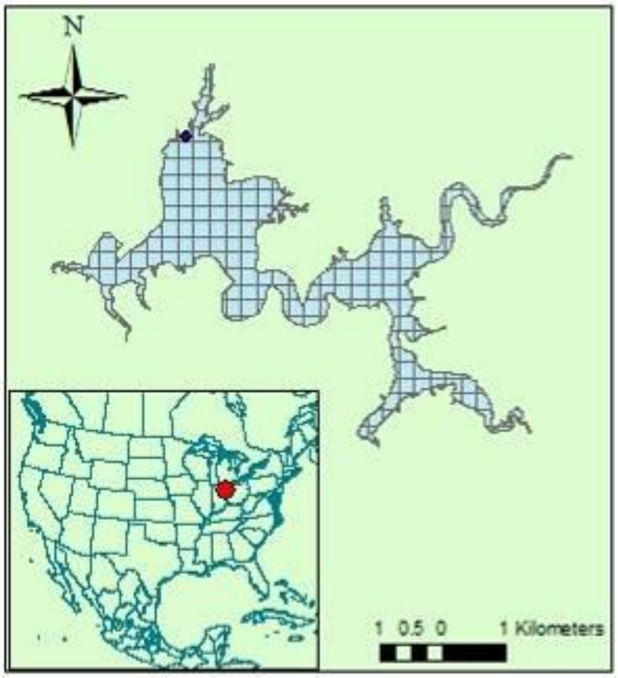


Figure 2.3: The William H. Harsha Lake is located in Ohio, USA. The intake to the water treatment plant is marked with a black dot shown near the northwest corner of the lake.

Since its impoundment, the lake has prevented over $77 million in flood damages, and it has generated $32.7 million in visitor expenditures (USACE, n.d.). This valuable community resource also serves as a surface water intake for the Bob McEwen surface water treatment plant that has a design capacity of 37 600 m^3·day^{-1} (10 MGD). The water quality fluctuates seasonally as detailed in Table 2.4:

Table 2.4: Harsha Lake water quality characterization (Green *et al.* 2010).

Raw Water Characterization		
Parameter/Dose	Range	Average
Turbidity (ntu)	3.0–38.7	9.1
UV-254 (cm^{-1})	0.153–0.231	0.18
pH	7.1–7.86	7.6
Temperature (°C)	11.5–17.0	-
Alkalinity ($mg \cdot L^{-1}$ as $CaCO_3$)	98–108	103
Total Organic Carbon ($mg \cdot L^{-1}$)	5.6–5.9	5.7
Total Manganese ($\mu g \cdot L^{-1}$)	17–618	120

2.3.2. Methodology

The IDFM procedural steps undertaken in this study are detailed in Figure 2.4:

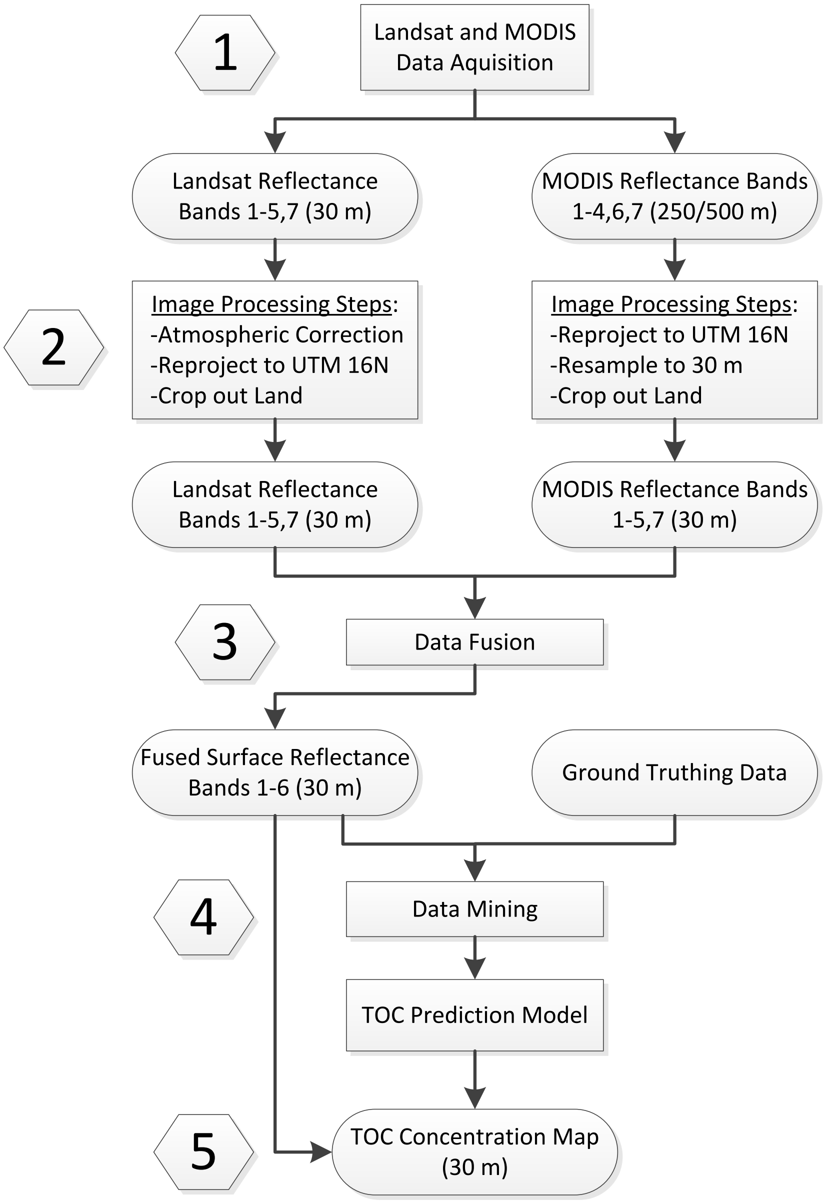

1
Landsat and MODIS Data Aquisition
Landsat Reflectance Bands 1-5,7 (30 m)
MODIS Reflectance Bands 1-4,6,7 (250/500 m)
2
Image Processing Steps:
-Atmospheric Correction
-Reproject to UTM 16N
-Crop out Land
Image Processing Steps:
-Reproject to UTM 16N
-Resample to 30 m
-Crop out Land
Landsat Reflectance Bands 1-5,7 (30 m)
MODIS Reflectance Bands 1-5,7 (30 m)
3
Data Fusion
Fused Surface Reflectance Bands 1-6 (30 m)
Ground Truthing Data
4
Data Mining
TOC Prediction Model
5
TOC Concentration Map (30 m)

Figure 2.4: Methodological flowchart in the TOC concentration retrieval procedures.

The chart is split into five main steps. Step one pertains to acquiring the Landsat and MODIS swath path images containing Harsha Lake. The second step involves the procedures required to prepare the images for fusion. Data fusion procedures are encompassed in step three; this study used the Spatial and Temporal Adaptive Reflectance Fusion Model (STAR-FM), although other candidate pixel selection and fusion algorithms can be used. The fourth step involves assimilating the ground-truth data and fused image band data into a machine learning algorithm. In this study, Discipulus® was used to perform the GP modeling analysis for the estimation of TOC concentrations via a nonlinear equation to be developed in terms of relevant fused images (bands). Lastly, step five specifically relates the GP model to the ground-truth data to train and validate the TOC concentration maps based on the band values embedded in the fused images. The essential steps are described in the following sections.

2.3.2.1. Data Acquisition [Figure 2.4; Step 1]

The ground-truth data for TOC in Harsha Lake were collected by the Army Corps of Engineers during 2008 and 2009 and by the US EPA from 2010 to 2012. The date for each ground-truth sample is shown in Table 2.5:

Table 2.5: Ground-truth acquisition dates. Grey images correspond to data used to train the GP model and the data in the blue cells were designated for model validation.

Date	TOC (mg/L)	Date	TOC (mg/L)	Date	TOC (mg/L)
08/20/2008	4.50	06/16/2010	6.90	07/27/2011	6.20
5/18/2009	8.36	06/16/2010	7.20	07/27/2011	7.70
6/29/2009	9.12	06/16/2010	7.20	07/28/2011	8.50
7/6/2009	8.27	9/7/2010	5.92	07/28/2011	11.00
7/13/2009	8.66	9/8/2010	5.50	07/28/2011	10.00
7/21/2009	7.77	09/08/2010	5.70	8/1/2011	5.92
8/3/2009	7.75	09/08/2010	6.70	8/22/2011	5.37
8/17/2009	7.65	09/08/2010	5.60	08/23/2011	14.00
8/31/2009	7.33	9/9/2010	5.60	08/23/2011	12.00
9/14/2009	7.73	09/09/2010	5.80	08/24/2011	11.00
9/28/2009	6.45	3/2/2011	6.14	08/24/2011	12.00
10/5/2009	6.33	3/17/2011	6.91	08/24/2011	6.30
10/26/2009	5.82	4/13/2011	7.35	8/29/2011	5.49
4/12/2010	8.23	5/23/2011	7.01	10/5/2011	5.38
4/19/2010	8.96	6/1/2011	7.05	11/2/2011	5.45
5/3/2010	7.26	6/7/2011	7.57	11/17/2011	5.99
5/24/2010	7.88	6/13/2011	6.98	05/24/2012	5.70
6/14/2010	6.72	6/21/2011	5.98	6/13/2012	4.30
06/15/2010	7.20	7/5/2011	6.34		
06/15/2010	6.90	7/11/2011	7.03		

Reflectance data for Harsha Lake were collected from the Landsat 5 TM, Landsat 7 ETM+, and Terra MODIS satellites. A comparison between these satellites is presented in Table 2.6:

Table 2.6: Satellite products utilized in this study.

Satellite Sensor	Product Selection	Spatial Resolution	Temporal Resolution	Bands Used
Terra MODIS	Surface Reflectance (MOD09GA)	250/500 m	Daily	1–4,6,7
Landsat 5 TM Landsat 7 ETM+	Surface Reflectance	30 m	16 Days	1–5,7

MODIS Terra images were obtained from the online Data Pool overseen by the NASA Land Processes Distributed Active Archive Center (LP DAAC), United States Geological Survey (USGS), and Earth Resources Observation and Science (EROS) Center, located at Souix Falls, South Dakota. The USGS also provided the Landsat imagery that was used for this study; however, these images were obtained from the Global Visualization Viewer, which is maintained by the LP DAAC, USGS, and EROS Center. Dates for downloading MODIS imagery were based on two criteria: 1.) each ground-truth date must be cloud-free (see Table 2.5) and 2.) remotely acquired on the same day as all Landsat imagery used in this study. As is described in greater detail in the data fusion section below, a Landsat image is required before and/or after each of the ground-truth dates. If a cloud free Landsat image was not available within two 16 day revisit cycles of the ground-truth date, then the ground-truth date was not used, since there was insufficient information for data fusion. For example, the first ground-truth sample was taken on August 20, 2008. As a result, a MODIS image on this date, as well as Landsat and MODIS images on the 16th of August and 1st of September were acquired. For the last ground-truth sample taken on June 13, 2012, a MODIS image on this data, in addition to Landsat and MODIS images on the 31st of May and the 16th of June were downloaded.

2.3.2.2. Image Processing and Preparation [Figure 2.4; Step 2]

The acquired MODIS data are at a level-2G basis, where the data have been radiometrically calibrated and atmospherically corrected to account for scattering and aerosols (Vermote *et al.* 2011). The Landsat data is on a level-1T basis, with radiometric and geometric corrections (USGS n.d.). As denoted by step 2 of Figure 2.4, ArcGIS, mapping and spatial analysis software, was used to process the images in preparation for the data fusion process. It was necessary to perform the following actions on the Landsat images: 1) perform atmospheric correction using MODIS 6S radiative transfer code using the LEDAPS toolbox supplied by NASA for this operation; 2) carry out reprojection to the Universal Transverse Mercator (UTM) zone 16 North; and 3) crop-out land data from around Harsha Lake. Besides, the following steps were taken to process the MODIS images: 1) perform reprojection to the UTM zone 16 North; 2) carry out resampling to a 30 m spatial resolution, and 3) crop-out land data from around Harsha Lake.

Processing of the images consists of two essential categories, namely: 1) modifying the images to have the same projection, pixel size, and scale in order to fuse them, and 2) preparing the images to increase fusion accuracy by cropping out the land and narrow portion of the lake. In the first processing category, images of disparate geographic map projections cannot be accurately compared. Therefore, UTM 16 North projection was used applied to all Harsha Lake images to ensure the same viewing angle. Next, only the MODIS images came pre-processed to adjust for backscattering effects of the atmosphere. MODIS/6S radiative transfer code was applied to algorithmically correct the pixel values to generate more accurate surface reflectance values. Resampling of the MODIS imagery to the resolution of the Landsat images was required since the STAR-FM data fusion algorithm compares images on a pixel by pixel basis; thus, each image needs to have the same number of pixels, rows, and columns. Landsat and MODIS surface reflectance products store the reflectance values at different scales. The Landsat product stores the surface reflectances on a scale from 0 to 255 (USGS n.d.), and the MODIS data is ranges from -100 to 16 000 (Vermote *et al.* 2011). In order for STAR-FM to compare pixel values, each of the images needs to have the same scale. The LEDAPS Processing Toolbox uses the MODIS/6S radiative transfer approach (Vermote *et al.* 1997) to atmospherically correct the

Landsat surface reflectance bands (Masek *et al.* 2006). This is the same procedure used to correct MODIS images during level 2 processing. Therefore after applying this toolbox to the Landsat images, both Landsat and MODIS data are stored as signed 16-bit integers that are scaled from -100 to 16 000.

The second element of image processing was to crop out the land and narrow portions of the lake. This is shown in Figure 2.5:

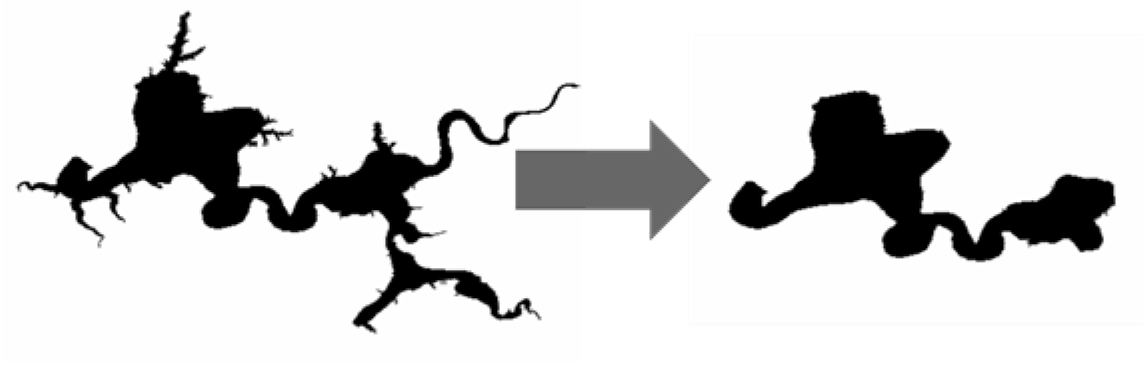


Figure 2.5: Cropping out narrow channels of the Lake to reduce land surface reflectance contamination.

The rationale behind this approach is to reduce the potential for surrounding land to contaminate the fused image. The contamination may occur during the data fusion process, when the STAR-FM algorithm searches through neighboring pixels. In order to limit the search to just the lake, the land has been removed. Additionally, narrow areas of the lake were blacked out; when water channels are smaller than MODIS’ 250/500 m resolution, it is likely that a part of the land surface is averaged into the pixel value representing the reflectance of the water causing contamination.

2.3.2.3. Data Fusion [Figure 2.4; Step 3]

Fused images corresponding to the ground-truth dates were developed for every day a ground-truth sample was collected with two notable limitations. The first is that the area of interest must be free of cloud-cover, since neither MODIS nor Landsat observe frequencies that pierce through clouds. Secondly, both near and off-nadir viewing angles of the lake were used, since this is representative of how the IDFM technique would be applied in the field, instead of a best case scenario featuring only near-nadir images of the lake. For each ground-truth observation shown in Table 2.5, a MODIS image was acquired. For fusion, Landsat and MODIS images taken before or after the ground-truth date are required. The accuracy of the synthetic image can be increased by using Landsat and MODIS images taken both before and after the ground-truth date (Gao *et al.* 2006). Generation of the fused image is illustrated in Figure 2.6 with this study site used as an example:

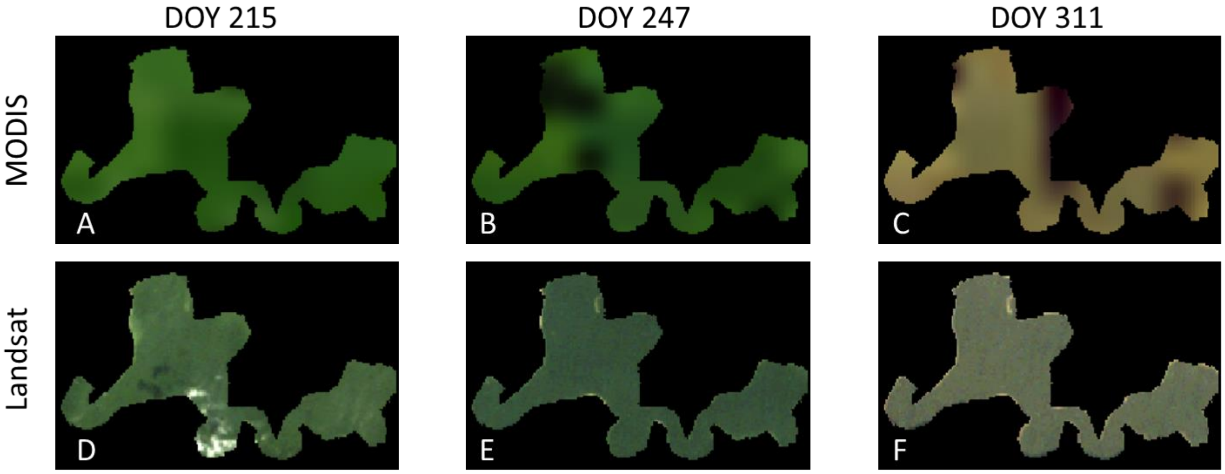


Figure 2.6: A Test of the STAR-FM Algorithm for gap-filling using images of Harsha Lake from the noted days in 2009.

The top row in Figure 2.6 is three coarse MODIS images (A, B, C) and the bottom row is the corresponding fine Landsat images (D, E, F). The individual image pairs A and D, B and E, and C and F correspond to satellite images captured on the same date. The three MODIS images and the two Landsat images (D and F) on the two adjacent dates were used in the IDFM process. Image E is the actual Landsat image taken, and the synthetic image will be compared to this.

In reconstruction of past events for gap-filling purposes, a total of five images, three MODIS and two Landsat, should be used to increase accuracy of the output image (Gao *et al.* 2006). This provides the algorithm with a set of pre and post conditions. Using A, D, and B a synthetic image based upon pre-conditions is created. Next, images C, F, and B are used to create the post-condition synthetic image. The two synthetic Landsat images based upon pre and post conditions are used to generate a single synthetic image. The synthetic image fills the data gap between images D and F. The resulting fused product is compared to the actual Landsat image shown in Figure 2.7:

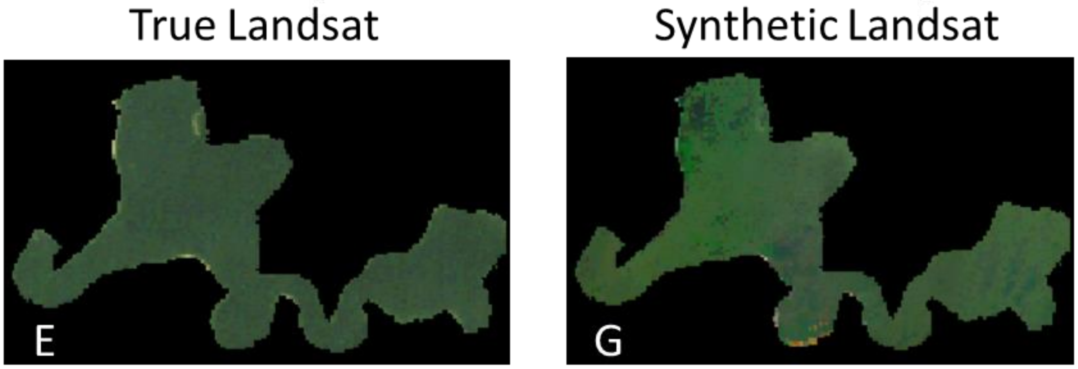


Figure 2.7: A Comparison between the true Landsat (E) and synthetic Landsat product (G).

To assess the performance of STAR-FM, the correlation between the actual (E) and fused (F) Landsat true-color images shown in Figure 2.7 is calculated. The true-color Landsat image was created using bands 3 (red), 2 (green), and 1 (blue). The true-color synthetic image was formed using the RGB bands of the fused product. Note that the fused image uses bands 3 and 1 of Landsat and MODIS for red, 2 and 4 for green, and 1 and 3 for blue. The correlation coefficient the two RGB images in Figure 2.7 is 0.7301 and the coefficient of determination is 0.5330.

The STAR-FM algorithm is effective at filling in data gaps when pre and post conditions are available, but what about using the algorithm for near-real-time monitoring? Post conditions will obviously not be available when monitoring images at the present time, yet the STAR-FM algorithm process can be manipulated for such applications. Recall that a synthetic image was generated for both pre and post condition cases; and, then, it was combined to form the final synthetic image. In a near-real-time monitoring situation, the pre-condition image will be the final image. Using only pre-condition images (A, D, and B) to predict E is less accurate, and the resulting synthetic image yields a coefficient of determination of 0.4147 when compared to the true image E. When using images B, C, and F to predict E, a coefficient of determination of 0.7482 was yielded. This is quite similar to using both pre and post conditions, but it goes to show the amount of variability that can be incurred when only using one set of conditions.

2.3.2.4. Machine Learning and Data Mining [Figure 2.4; Step 4]

The GP model used in this study was developed using the Discipulus® software package, created by Francone (1998). Discipulus® is designed to sort through GP models using supervised machine learning techniques and determine 30 of the best models based on the fitness of the training and validation data. The arithmetic operations selected for training the GP model were addition, subtraction, division, multiplication, trigonometric functions, square root, and absolute value. The recommended values of 95% and 50% were used for mutation and crossover frequency (Francone 1998). Lastly, the initial program size was set to 80 Mb with a max program size of 256 Mb. If an accurate GP model cannot be developed for the provided training and validation data sets, then the max program size is gradually increased. This enhances the explanatory power of the model by increasing the number of mathematical operations that can be used to relate surface reflectance to TOC. This is especially useful for complex, nonlinear relationships between the independent and dependent variables, but larger program sizes increase the potential for over fitting the model and it can take longer to solve (Francone 1998).

After the program has finished creating models, the 30 models with the best overall performance are saved and analyzed. An advantage of Discipulus® is the capability to adapt current models with new ground-truth data (Francone 1998). As additional TOC samples are collected, this allows for the model to be updated to reflect hydrological and anthropogenic changes over time. Since Discipulus® ranks the models based on the average fitness between the training and calibration data sets, it is necessary to discern whether the high average fitness is due to over-fitting of either the validation or calibration data set. Thus, the model that yields high fitness values for both calibration and validation is selected for the GP model used in this study. The GP model is presented as a series of mathematical or logic operations that must be applied in sequence to the fused surface reflectance band data in order to generate a TOC concentration value. Furthermore, since this study proposes the development of an early warning system for water treatment plant operators, it is imperative that the selected model be capable of predicting peak TOC concentrations in the lake. This ensures that the plant operators are able to observe and track plumes of TOC in the lake that are in the vicinity of the treatment plant's source water intake. With this knowledge, the treatment operations can be adjusted to minimize the production of disinfection byproducts. The GP model was tested against a traditional two-band model (Vincent 2004) which was solved through a linear regression model in Matlab using band ratios instead of individual bands as explanatory variables. The generic form of the two-band ratio model is shown in Eq. 2.1:

$$C_{TOC} = A*\lambda_1/\lambda_2 + B \tag{2.1}$$

where C_{TOC} is the concentration of TOC, A is the slope, λ_1 is the wavelength of the first band, λ_2 is the wavelength of the second band in the ratio, and B is the intercept. The same training and calibration data sets used for creating the GP model were employed to train and calibrate the two-band model.

The training set was allotted 67% of the ground-truth data, which corresponds to 39 of the grey colored cells in Table 2.5, to aid the training and calibration procedure. The remaining 33% or 19 ground-truth observations were used for validation. Determining which observations were selected for calibration and validation data sets is based purely on the measured concentration without regard to the temporal aspects. First, the observation data was sorted from low to high values, and then, 67% of the low, medium, and high concentrations were allotted to the training and calibration data set. Then the remainder was used for validation. It is necessary to ensure that both the calibration and validation data sets are exposed to the widest range of TOC values available to increase the accuracy of the model's prediction capabilities at extremes. After training, the validation stage confirms whether the model is well suited for calculating TOC concentrations by checking the model's performance using the validation data set. For example, a model may aptly predict a peak TOC concentration when using the training data set, but this may be a result of over fitting or chance if these results cannot be achieved when attempting to predict a peak value in the validation data set. The final choice of a model must be based on the correct prediction capabilities when both the training and validation data sets may exemplify good fitness (Francone 1998). To gain a level of general understanding of the TOC characteristics in the lake, an analysis of the ground-truth data is provided in Table 2.7:

Table 2.7: Ground-truth data analysis.

Parameter	Value
Ground-truth Time Period	2008-2012
Number of Ground-truth Samples	58
Average TOC Value ($mg \cdot L^{-1}$)	7.2
Maximum TOC Value ($mg \cdot L^{-1}$)	14.0
Minimum TOC Value ($mg \cdot L^{-1}$)	4.3
Sample Standard Deviation ($mg \cdot L^{-1}$)	1.9

2.3.2.5. Concentration Map Generation [Figure 2.4; Step 5]

The GP model translates the surface reflectance values to TOC concentrations and maps the TOC concentrations throughout Harsha Lake finally. Each pixel of the fused images represents a 30 by 30 meter square of the lake, characterizing the surface reflectance at the bandwidths specified in Table 2.3. The TOC concentration at each pixel can be obtained by using the GP regression equation that is highly nonlinear in terms of surface reflectance values. This estimation process is then repeated for each of the pixels making up the lake map. As determined by the GP model, certain bandwidths may have a stronger explanatory power in the determination of the TOC concentration. It was also likely that in some iterations of the GP model, not all bandwidths were used in the determination of the TOC concentration.

2.3.3. Statistical Indices for Assessment and Model Selection

When developing multiple models, methods for comparison must be established to evaluate and rank the models. In general, a model is accurate if the predicted values closely match the observed values. For this study, model performance was analyzed using four statistical indices. The indices include the root mean square error (RMSE), ratio of the standard deviations (CO), mean percent of error (PE), and the square of the Pearson product moment correlation coefficient (RSQ). The RMSE and CO are given by Eqs. 2.2 and 2.3, respectively:

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(y_{p_i}-y_{oi})}{N}} \tag{2.2}$$

$$CO=\sqrt{\frac{\sum_{i=1}^{N}(y_{pi}-\bar{y}_{pi})^2}{\sum_{i=1}^{N}(y_{oi}-\bar{y}_{oi})^2}} \tag{2.3}$$

where y_{pi} is the i^th^ predicted value and y_{oi} is the i^th^ observed value; n is the number of samples taken; $\bar{y}_{pi}$ is the arithmetic mean of the i^th^ predicted value; and $\bar{y}_{oi}$ is the arithmetic mean of the observed value. Ideally, the RMSE should be zero, and a model exhibits ideal performance when the CO equals 1. The remaining two statistical indices for judging the estimation accuracy are PE and RSQ:

$$PE=\frac{\sum_{i=1}^{N}[\frac{y_{pi}-y_{oi}}{y_{oi}}\cdot 100\%]}{N} \tag{2.4}$$

$$RSQ = R^2 = [\frac{\sum_{i=1}^{N}(y_{oi} - \bar{y}_{oi})(y_{pi} - \bar{y}_{pi})}{\sqrt{\sum_{i=1}^{N}(y_{oi} - \bar{y}_{oi})^2 \sum_{i=1}^{N}(y_{pi} - \bar{y}_{pi})^2}})]^2 \tag{2.5}$$

Model performance ranks well when the PE approaches zero, and the RSQ should ideally equal 1 for a good model.

2.4. Results

The GP model as an integral part of the IDFM algorithm reconstructed the TOC variations over various dates based on the surface reflectance bands of the fused images. To substantiate the advantage of the data fusion process for the purpose of comparison, a second model was also developed by simply using MODIS surface reflectance. The same data splitting techniques described in Step 4 of the Methodology section were applied for calibrating the MODIS-based GP model. This addresses the science question aimed at determining whether the MODIS or fused MODIS-Landsat will be better inputs for predicting TOC. Both the fused and MODIS images were separately processed for 75 runs in Discipulus®. A single run is characterized by the Discipulus® building a model until the maximum size of 256 Mb has been reached. Thus, both the MODIS and fused GP models were appropriated equal amounts of computational effort. The correlation between the predicted and observed TOC concentration for the MODIS surface reflectance training and validation data sets are presented in Figure 2.8. The coefficients of determination for the training data and validation data sets are 0.6836 and 0.4570. These results indicate a strong correlation for the training data set, and a moderate correlation for the validation data set. Examination of Figure 2.8 shows that the model performs well when predicting concentrations within one standard deviation of the average TOC value. Yet, the model under-predicts higher TOC concentration values, since a number of predicted TOC values fall below the 45 degree line past 9 mg·L⁻¹. The corresponding correlation plot for the GP model using the fused data as inputs is shown in Figure 2.9:

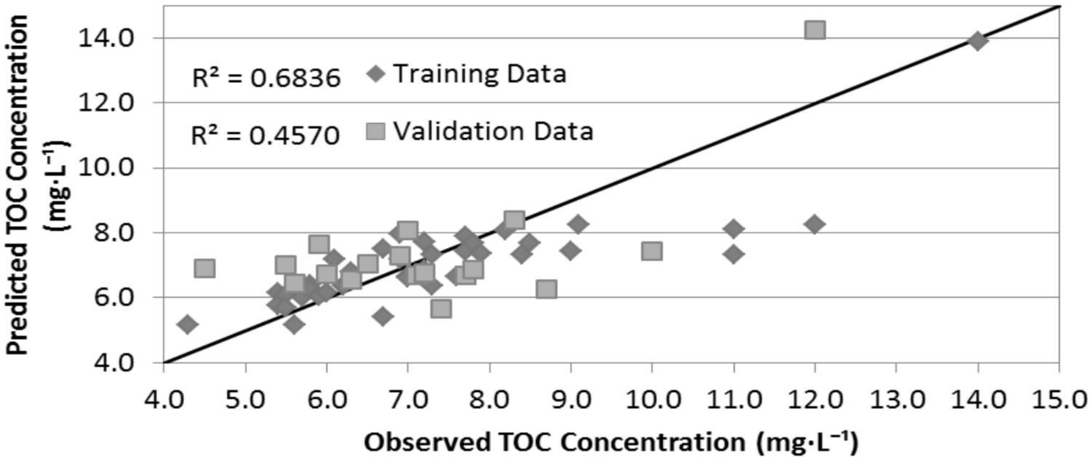


Figure 2.8: Correlation between predicted vs. observed TOC concentrations formulated using the MODIS surface reflectance as inputs to the GP model.

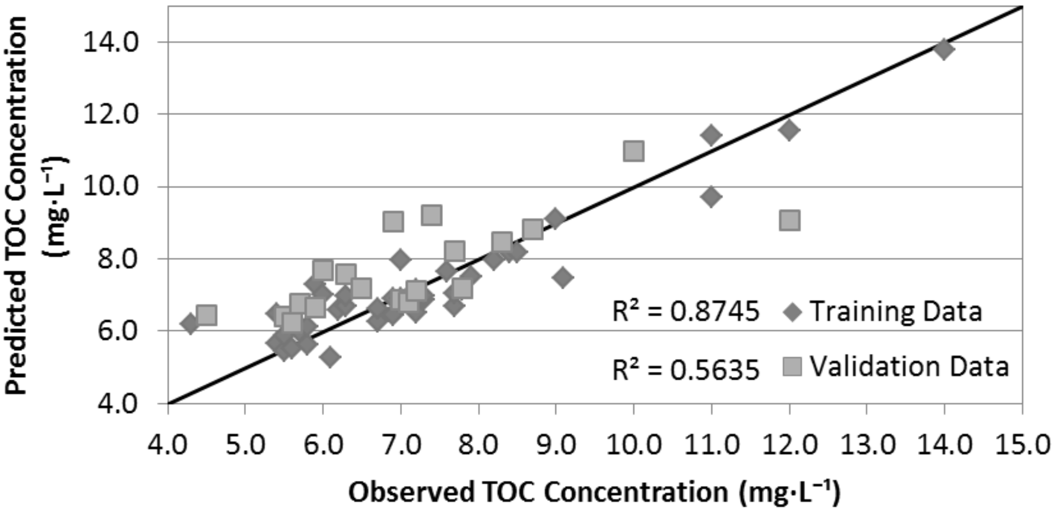


Figure 2.9: Correlation between estimated vs. observed TOC concentrations formulated using the fused image surface reflectance as inputs to the fusion-based GP model.

The fusion-based GP model yields coefficients of determination of 0.8745 and 0.5635 for the training and validation data sets. These values entail that a strong relationship exists between the predicted and observed values. Unlike the MODIS-based GP model that under-predicted TOC values past 9 mg·L⁻¹, the fusion-based model excels at predicting peak concentrations without a bias toward under or over-prediction .

A more lucid comparison of the MODIS-based GP model in Figure 2.8 and the MODIS-based GP model of Figure 2.9 is presented by plotting the predicted and observed results as a time series in Figure 2.10:

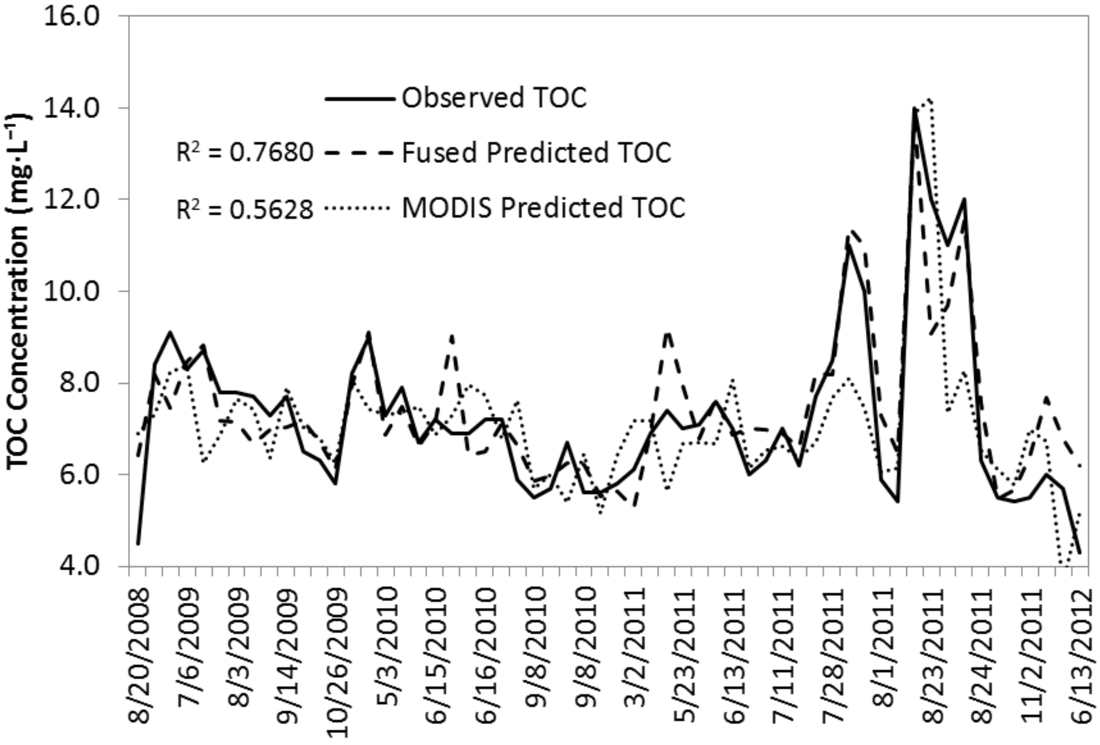


Figure 2.10: Time series plots comparing the predicted TOC values from the MODIS and Fused GP Models. The predicted TOC values based on the MODIS surface reflectance share similar accuracy to the TOC values with predicted ones using the fused surface reflectance; however, the GP model using the fused image surface reflectance as inputs excelled at predicting the peak TOC values.

Through visual examination of Figure 2.10, it can be observed that both GP models exhibit moderate to strong performance, since the predicted values (dotted lines) replicate the temporal trends of observed TOC values in the lake. In comparing the two models, the fused images as inputs to the GP model exhibit more accurate estimation of TOC above 9 mg·L⁻¹. Further analysis is difficult to draw by visual examination alone.

The four statistical indices presented in the methodology section are used to quantify the relative performance between the MODIS-based and fusion-based GP models, as well as a traditional two-band ratio model. The resulting two-band model is shown in Eq. 2.6:

$$C_{TOC} = -0.04630 * v_3/v_5 + 7.2087 \quad (2.6)$$

where C_{TOC} = concentration of TOC. Comparison using the four statistical indices is presented in Table 2.8:

Table 2.8: Observed vs. predicted TOC values and indices of accuracy for the traditional two-band model and the GP models created from MODIS and fused images.

Metric	2-Band	MODIS GP	Fusion-based GP
TOC Obs mean (mg·L⁻¹)	7.276	7.276	7.276
TOC Pred Mean (mg·L⁻¹)	7.276	7.085	7.433
Percent Difference of the Means (%)	0.000	2.629	-2.166
Root Mean Square Error (mg·L⁻¹)	1.716	1.248	0.900
Ratio of St. Dev.	0.394	0.851	0.855
Mean Percent Error (%)	5.046	-0.448	3.921
Square of the Pearson Product Moment Correlation Coefficient	0.1974	0.5628	0.7680

As shown in Table 2.8, the mean predicted values of the two-band model equaled the observed mean values. However, a more detailed analysis of the predicted values indicates poor performance. The RMSE of 1.716 mg·L^{-1} and the CO of 0.394 are far from the ideal values of 0 mg·L^{-1} and 1. The PE of 5.046% is relatively low. But this is likely due to the fact that the linear regression two-band model yielded predicted values close to the observed average of 7.276 mg·L^{-1}, and the range of TOC values in the lake were typically between 6 and 9 mg·L^{-1}. This is why it is important to use multiple methods to assess prediction quality. Lastly, the R^2 value for this model was 0.1974, which means that only 19.74% of the variation is explained by the two-band model. In conclusion, the traditional two-band model exhibited poor TOC prediction capabilities across all statistical indices, thus, necessitating a more powerful modeling technique.

Next, the performance of the GP models is presented. The average values of estimated TOC concentrations based on the MODIS-based and fusion-based-based GP models are within the range 2.629% and -2.166% of measured TOC concentrations in the lake, respectively. However, more systematic analyses are required to examine the relative performance of the GP model based on MODIS only data vs. fused (MODIS-Landsat) data. RSME values of 1.248 and 0.900 mg·L^{-1} for the MODIS-based and fusion-based GP models are reasonably close to the minimum error of zero, with the fusion-based GP model exhibiting higher accuracy. A CO value close to 1 indicates the estimated values are close to the observed values. Both GP model performed quite well with the ratios of 0.851 and 0.855, respectively, as opposed to the two-band model. With regard to the PE, both models are under 5%, although the MODIS-based GP model outperforms the fusion-based GP model by 3.433% with an actual value of -0.488. In principle, the higher the levels of accuracy, the closer the PE value is to zero. The negative value indicates that the estimated TOC concentrations are lower than the observed counterparts. In terms of RSQ, accurate results are depicted as the R^2 value approaches 1. The values achieved by these two models are 0.5628 and 0.7680, which corresponds to a positive correlation between the estimated and observed values. The MODIS-based GP model accounts for a little over half of the variation, whereas the fusion-based GP model successfully explains over 75% of the variation.

The fused-based GP model with the aid of data fusion developed to fit the observed data curve in Figure 2.10 is explicitly depicted in Appendix A. Variables v_0, v_1, v_2, v_3, v_4, and v_5 characterize the band data of the fused images (Table 2.9). As previously noted in Table 2.3, the Landsat and MODIS images were fused in accordance with their bandwidths, not their band numbers. Among the 30 best candidate models being generated, the frequency of use for each bandwidth is shown in Table 2.9:

Table 2.9: Frequency of use for bands in the TOC GP model.

Variable	Band Number	Bandwidth (nm)	MODIS-based GP Frequency of Use (%)	Fusion-based GP Frequency of Use (%)
v_0	MODIS Band 3 Landsat Band 1	459–479 450–520	67	97
v_1	MODIS Band 4 Landsat Band 2	545–565 520–600	100	100
v_2	MODIS Band 1 Landsat Band 3	620–670 630–690	80	97
v_3	MODIS Band 2 Landsat Band 4	841–876 760–900	70	93
v_4	MODIS Band 6 Landsat Band 5	1628–1652 1550–1750	87	80
v_5	MODIS Band 7 Landsat Band 7	2105–2155 2080–2350	63	70

The frequency of use describes how often a variable (a specific bandwidth) was used among the 30 best GP models developed. 100% frequency of use means that the variable was used to compute TOC in all 30 models. In the context of the fusion-based GP model, it can be seen that

v_1 was used in all of the programs, while v_0 and v_2 were determined to be of significance in the 67 to 97 & and 80 to 97% of the programs. When comparing this to the MODIS-based model, it is noticeable that although v_1 was also heavily used by the program, then the similarities between band usage deviates over the subsequent bands.

2.4.1. TOC map, seasonal changes, and limitations

For the purpose of applications, model predictability for TOC concentrations can be further assessed by reconstruction of the TOC distributions in Harsha Lake. First, the fusion-based GP model with the solution procedure listed as a series of equations in Appendix A may yield the TOC concentration map for the dates of missing high-resolution Landsat imaginaries. As an example, the fusion-based GP model developed above was applied based on the proposed data fusion process using surface reflectance values collected from mostly cloud-free days in June and July of 2011. Fused images and derived TOC concentration maps are shown in Figure 2.11 for the 158th, 164th, and 191st day of year in 2011 that were missing high-resolution Landsat imageries:

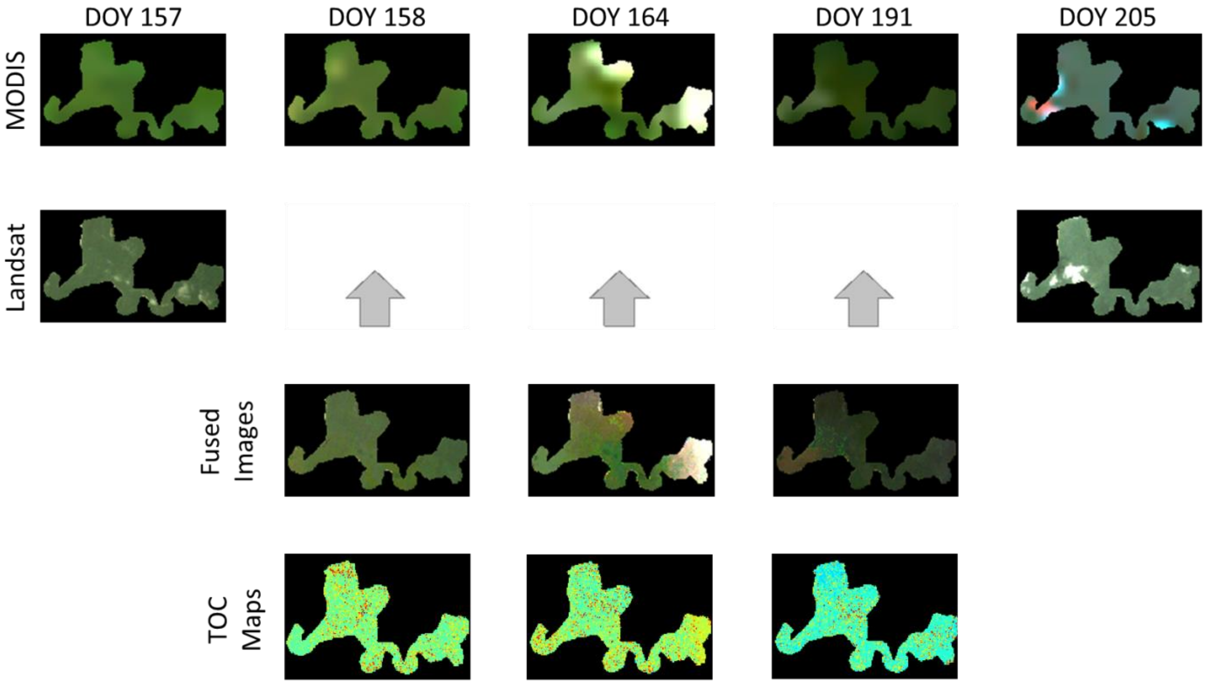


Figure 2.11: Data fusion and TOC map generation for June-July 2011. This represents both the gap filling capabilities of the IDFM technique and the ability to predict TOC concentrations for cloud free regions of the water body. The top row is the daily MODIS images, and the second row is comprised of 2 Landsat images. STAR-FM is used to create high resolution synthetic images for ground-truth dates. Surface reflectance values at sampling locations in the fused images are used to train the genetic programming model for TOC map generation featuring the lake.

MODIS images are available for each of these days, and the Landsat images serve as a pre and post reference for conditions on the lake. Using STAR-FM, the high spatial and temporal fused images fill in the gaps left by Landsat on the 158th, 164th, and 191st day of year. The 4th row consists of the concentration maps that are generated to detail the spatiotemporal variations and concentration of TOC on the lake. A closer inspection of the TOC maps indicates no significant spatial variations in concentration levels during individual dates of the 2011 summer season with the predicted concentrations falling into a narrow range of 5 and 9 mg·L^{-1} with few extreme values. Closer analysis of a TOC concentration map is provided for on 24-May, 2010 (Figure 2.12):

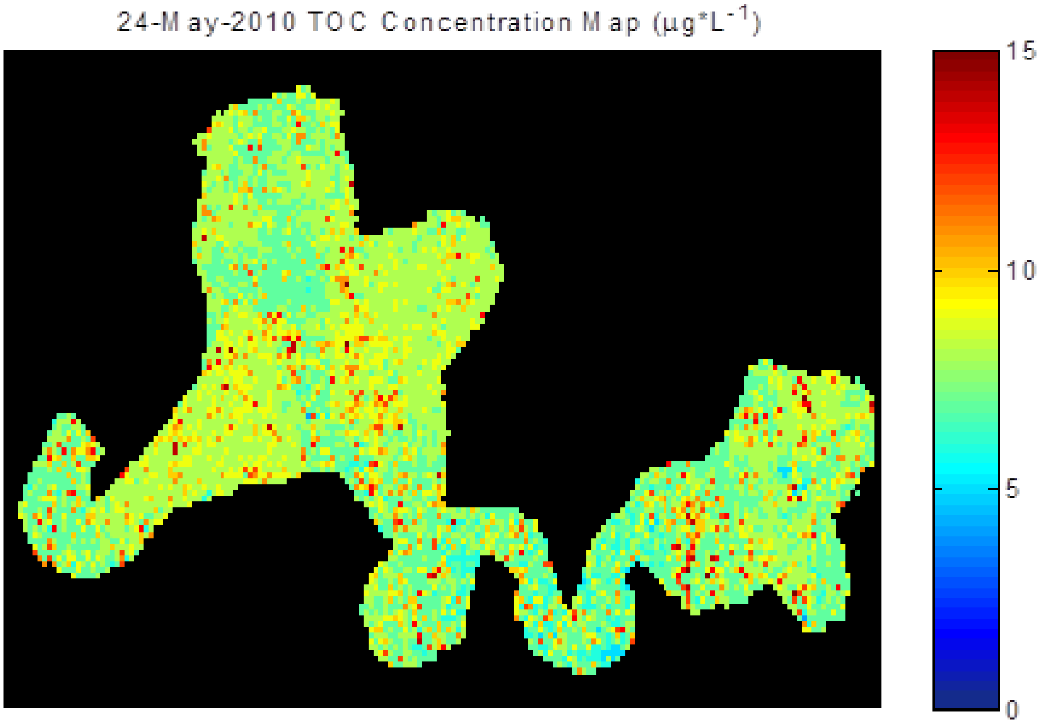
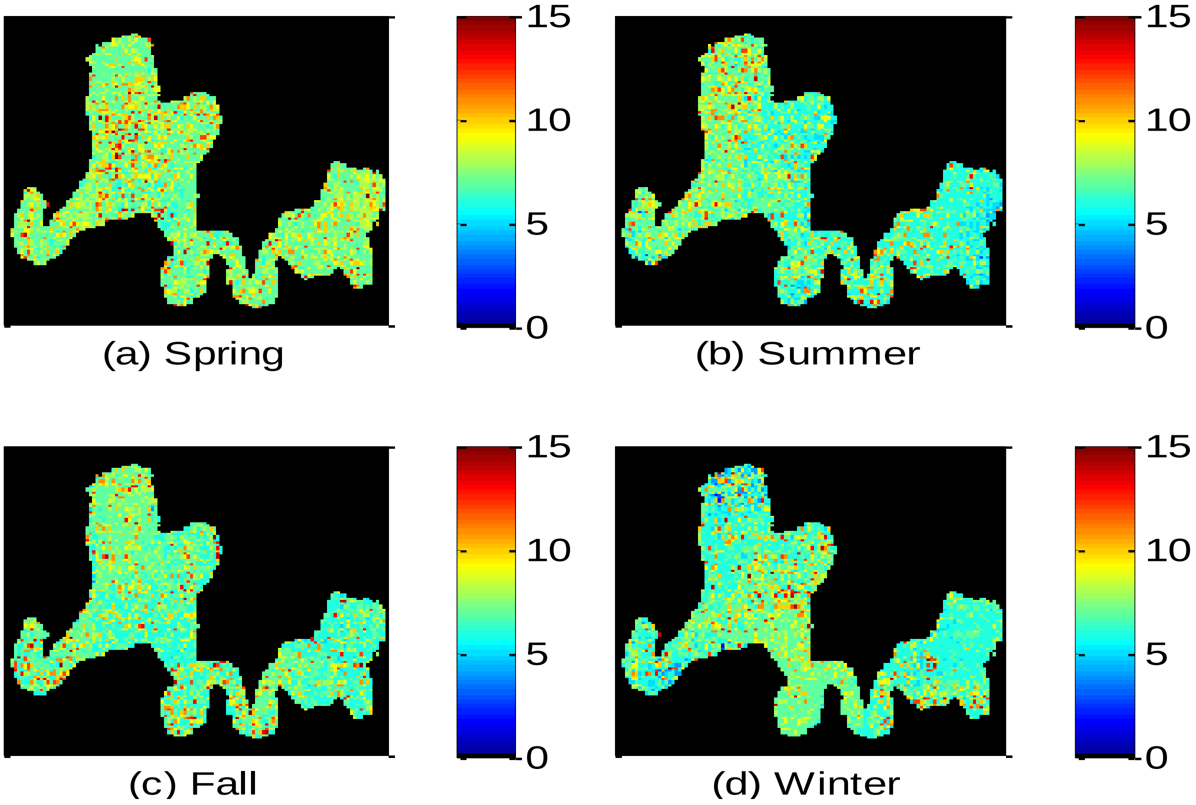


Figure 2.12: TOC concentration map for Harsha Lake on 24-May-2010.

The concentrations identified in this image range are as low as 5 mg·L^{-1} and as high as 15 mg·L^{-1}. The average concentrations range between 5.5 mg·L^{-1} and 9 mg·L^{-1}. While the TOC concentration is mostly uniform through the lake, the GP model has identified patches of extreme TOC levels occurred in the lake. Streams and tributaries feeding into the lake do not exhibit an influx or reduction of TOC at these interfaces.

Seasonal TOC maps provide a visual representation into the dynamics of TOC throughout the year. The seasonal TOC maps (Figure 2.13) were generated by grouping the images into seasons (Table 2.10) and averaging the predicted TOC values during the season. The average TOC value in the winter, spring, summer, and fall were 7.2 mg·L^{-1}, 8.0 mg·L^{-1}, 7.2 mg·L^{-1}, and 7.4 mg·L^{-1}. In the spring, TOC concentrations remained high at 7 to 9 mg·L^{-1} and homogenous throughout the lake. The TOC level in summer and fall decreased to a range of 6 to 8 mg·L^{-1}, and the TOC remained highest toward the western side of the lake. In winter, there appears to be a transition in which higher TOC concentrations began to occur in the middle portion of the lake.



(a) Spring

(b) Summer

(c) Fall

(d) Winter

Figure 2.13: Seasonal average TOC concentration maps predicted by IDFM.

Table 2.10: Separation of the sampling data by season.

Spring		Summer			Fall	Winter
05/18/09	04/13/11	08/20/2008	09/14/09	07/28/2011	09/28/09	03/02/11
04/12/10	05/23/11	06/29/09	09/07/10	08/01/11	10/05/09	03/17/11
04/19/10	06/01/11	07/06/09	09/08/2010	08/22/11	10/26/09	
05/03/10	06/07/11	07/13/09	09/09/2010	08/23/2011	10/05/11	
05/24/10	06/13/11	07/21/09	06/21/11	08/24/2011	11/02/11	
06/14/10	05/24/2012	08/03/09	07/05/11	08/29/11	11/17/11	
06/15/10	06/13/2012	08/17/09	07/11/11			
06/16/2010		08/31/09	07/27/2011			

2.5. Discussion

2.5.1. Spatiotemporal TOC Variations

The first scientific question addresses the spatiotemporal variations in TOC calculated by the fusion-based GP model. Temporal variations in TOC were examined by averaging the predicted TOC values in the spring, winter, summer, and fall seasons (Table 2.10). The average TOC value in the winter, spring, summer, and fall were 7.2 mg·L^{-1}, 8.0 mg·L^{-1}, 7.2 mg·L^{-1}, and 7.4 mg·L^{-1}. Previous studies have reported that TOC values were lowest during the winter (Agren *et al.* 2008; Bayram *et al.* 2011). Our study likewise found that that winter had one of the lowest TOC values, due to low temperatures reducing microbial activity and runoff. Spring yielded the highest TOC value in the study, and summer yielded one of the lowest. This was largely unexpected, since Bayram *et al.* (2011) reported that spring TOC values were lower than fall and late summer. Agren *et al.* 2008 similarly found that spring TOC values peaked slightly, yet late summer and fall TOC values were still higher due to increased temperature and precipitation. Possible explanations can be attributed to differences in a number of variables that factor into TOC generation and export, such as regional climate influences, municipal discharge, land usage, and soil types. The fall TOC value was higher than the summer and winter, which coincides with the findings of Bayram *et al.* 2011.

Figure 2.12 is an example of spatial variations of TOC within the lake, particularly for tracking TOC plumes. The accuracy of these plumes cannot be completely verified due to the narrow range of available ground-truth TOC concentrations that were used to train the model. The majority of the *in-situ* values used in training were from 6 to 9 $mg \cdot L^{-1}$. Nevertheless, from Figure 2.10, the fusion-based GP model was able to accurately predict peak TOC values at 11, 12, and 14 $mg \cdot L^{-1}$, which does showcase the potential for plume tracking. For this reason, the current fusion-based GP model may not be thoroughly trained to detect events that are significantly higher or lower than the narrow range of the sample data. Ideally, a wider range of TOC *in-situ* data would have been provided for the study, as well as samples and the location of a major TOC plume detected the lake. However, this could still be explored in a future study due to the flexibility of the IDFM technique and the Discipulus® GP software. Discipulus allows for GP models to be updated as new *in-situ* data is obtained, and IDFM can use the new GP model and previously derived fused images to immediately develop updated TOC concentration maps. Thus, periodic sampling of source water allows a water treatment plant to acclimate their model to accurately predict TOC fluctuations due to anthropogenic influences, climate change, and weather. This ensures the TOC prediction capabilities of the GP model stays relevant as water quality conditions change on a long-term scale.

2.5.2. Impact of Data Fusion

The second scientific question in this study aimed to assess whether a GP model would have better performance when using fused imagery (Landsat and MODIS) as opposed to using

MODIS imagery only. Data fusion was used to create a daily synthetic image with the spatial resolution of Landsat (30 m), whereas daily MODIS images have a 250 to 500 m resolution depending on the bands used. The enhanced spatial resolution is a significant advantage when monitoring water quality in a small lake. This provides reduces the amount of surface reflectance from the land contaminating shoreline pixels, since the fused image offers delineation between the land, the shoreline, and open water. The finer spatial resolution can detect pollution events on the water's surface that are only 30 m in size, whereas MODIS would require the plume to be 250 to 500 m. This is crucial for detecting plumes coming toward source water intakes of water treatment plants.

A comparison of Figs. 8 and 9 showed that MODIS-based GP model and the fusion-based GP model had little error and show negligible bias when predicting TOC concentrations below 9 $mg \cdot L^{-1}$. However, the MODIS-based GP model exhibited an underestimation bias for TOC concentrations at and above 9 $mg \cdot L^{-1}$, which was not observed in the fusion-based GP model. Possible explanations for this are based on sensor limitations and possible constraints imposed while solving the GP model. First, band 1 of the MODIS sensor does not capture the entire CDOM spectral feature between 650 and 710 nm. MODIS band 1 is centered at 645 nm and its upper range is 670 nm, while the fused image benefits from additional spectral data of Landsat band 3, which is centered at 660 nm with an upper range of 690 nm. Another reason for the bias of the MODIS-based GP model could be due to the constraints imposed while training the GP model. The best GP model was selected after 75 runs, and this may have not been enough time for the MODIS-based GP model to decode and explain the relationship at higher TOC concentrations. Since both models were afforded the same training time, this indicates that the fusion-based GP model is a better input given how it trained better in the same number of runs. Statistical analysis presented in Table 2.8 supports the claim that the fusion-based GP model noticeably outperformed the MODIS-based GP model for RMSE, CO, and R^2.

2.5.3. GP versus Two-Band Ratio Inversion Modeling

The third scientific question aimed to compare a traditional two-band model to a GP model. Per the results in Table 2.8, both GP models yielded more accurate RMSE, CO, PE, and R^2 values. Two-band models are generally analytically derived based on knowledge of which band contains a telltale spectral feature for TOC, then dividing by another band to reduce systematic noise, backscattering, or reflectance contamination from other water quality parameters. This can be effective in case I waters, where the surface reflectance is the sum of clean water and a low number of water quality constituents. This method is less effective in case II waters, in which the spectral reflectance is a product of numerous water quality constituents. On the other hand, GP is suited for decomposing complex relationships without prior knowledge or input. This is especially handy for developing an empirical model specifically tuned to the unique water quality characteristics and trends of the lake in question.

2.5.4. GP Model – Identifying Important Spectral Bands

The frequency of use explains which bands the GP model found most useful in explaining the relationship between surface reflectance and TOC concentration. This discovery is the topic of the fourth scientific question. Band data frequency of use is beneficial in determining which satellites have the bands necessary to monitor the lake, as well as limiting the amount of band data to be downloaded and stored. The frequency of use for each band is given in Table 2.9. Overall, v_1 and v_2 would be commonly used in most occasions for TOC estimation, since these bands correspond with the spectral reflectance peaks for TOC occurring at 550 and 675 nm as seen in Figure 2.. The MODIS-based model still prioritizes the use of the v_2 band in 80% of the best programs, yet v_4 was given a slightly higher significance, as it has been used in 87% of the programs. By analyzing Figure 2. once more, a possible explanation can be observed for why the fusion-based model uses v_2 in 97% of the best models, while it is only used 80% of the time in the MODIS based models. It can be seen that Landsat is capable of detecting more of the CDOM peak occurring at v_2 (Band 3 in Figure 2.) than MODIS. This is another advantage of using data fusion, since the additional spectral information from Landsat band 3 has been integrated with MODIS band 1 to form a single synthetic image. Lastly, in the fusion-based GP model, the variables v_4 and v_5 were used in the least priority, which implies that there is not a strong relation between the TOC concentration in the water and TOC's reflectance at these wavelengths.

2.6. Conclusions

Real-time knowledge of TOC distribution in source water can help treatment operation to minimize the byproduct generation. Yet how to fuse fusion-based images to achieve essential resolutions spatially and temporally by an optimal way requires screening multiple inverse modeling in a timely fashion. Using the MODIS and Landsat data streams, the STAR-FM algorithm generated accurate fused images with high temporal and spatial resolutions. With the IDFM method, the fusion-based GP model was able to fuse different band data to estimate and reconstruct TOC concentrations in Harsha Lake for dates of no high-resolution Landsat imageries. The calibration and validation plots of the fusion-based GP model had R^2 values of 0.5635 and 0.8745, respectively, which excelled beyond that of the GP model developed simply using MODIS surface reflectance data. Overall analysis of the GP model showed that the data from the first, second, and third fused bands contributed the most in determining the TOC concentrations upon the lake. These bands correspond with spectral ranges between 459-900 nm, of which TOC has two spectral peaks around 550 and 675 nm. Upon assessing the model for accuracy testing using the RMSE, ratio of standard deviations, PE, and square of the Pearson product moment correlation coefficient, it was observed that the fusion-based GP model yielded low error for the specific set of fused image input data.

The IDFM technique proved reliable in estimating TOC concentrations spatially and temporally. However, there still exist difficulties that need to be overcome. The GP model was still not sensitive enough when it encountered the peak values of TOC within the lake. The model robustness should be improved through the collection of a large amount of ground-truth data, which will allow for accurate event-based detection event. Furthermore, such a near-real-time monitoring system using Landsat and MODIS imageries is impractical in areas with significant cloud clover. To resolve the deficiency, the integration of band data from microwave satellite sensors capable of penetrating clouds may become necessary. More inverse modeling tools with regression capabilities, such as GP, ANN, and SVM, may be compared further to improve the estimation accuracy in the future.

2.7. Acknowledgements

Support from the USEPA National Risk Management Research Laboratory is gratefully acknowledged. The research described herein has been subjected to the Agency’s peer and administrative review and has been approved for external publication. Any opinions expressed in this paper are those of the author(s) and do not necessarily reflect the views of the Agency,

therefore, no official endorsement should be inferred. Any mention of trade names or commercial products does not constitute endorsement or recommendation for use.

2.8. References

Agren, A., Jansson, M., Ivarsson, H., Biship, K., and Seibert, J., 2008, Seasonal and runoff-related changes in total organic carbon concentrations in the River Ore, Northern Sweden. *Aquatic Sciences*, vol. 70, no. 1, pp. 21-29.

Arenz, R., Lewis, W., and Saunders III, J., 1995, Determination of Chlorophyll and Dissolved Organic Carbon from Reflectance Data for Colorado Reservoirs. *International Journal of Remote Sensing*, vol. 17, no. 8, pp. 1547—1566.

Bayram, A., Onsoy, H., Akinci, G., and Bulut, V., 2011, Variation of total organic carbon content along the stream harsit, Eastern Black Sea Basin, Turkey. *Environmental Monitoring and Assessment*, vol. 182, pp. 85-95.

Chang, N. and Vannah, B., 2012, Monitoring total organic carbon concentrations in a lake with the integrated data fusion and machine learning (IDFM) technique. *Proc. SPIE* 8513, Remote Sensing and Modeling of Ecosystems for Sustainability IX, 851307.

Chang, N., Xuan, Z., Yang, Y., 2013, Exploring spatiotemporal patterns of phosphorus concentrations in a coastal bay with MODIS images and machine learning models. *Remote Sensing of Environment*, vol. 134, pp. 100-110.

Chen, L., Tan, C., Kao., S., and Wang, T., 2008, Improvement of Remote monitoring on water quality in a subtropical reservoir by incorporating grammatical evolution with parallel genetic algorithms into satellite imagery. *Water Research*, vol. 42, pp. 296—306.

Davis, L., 1991, *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York.

Doerffer, R., and Schiller, H., 2007, The MERIS Case 2 water algorithm. *International Journal of Remote Sensing*, vol. 28, pp. 517-535.

Dosskey, M., and Bertsch, 1994, Forest sources and pathways of organic matter transport to a blackwater stream: a hydrologic approach. *Biochemistry*, vol. 24, pp. 1-19.

Engelbercht, A., 2007, *Computational Intelligence: An Introduction*, Wiley, New York.

Francone, D.F., 1998, *Discipulus Software Owner's Manual, version 3.0 DRAFT*, Machine Learning Technologies, Inc., Colorado.

Gao, F., Masek, J., Schwaller, M., and Hall, F., 2006, On the Blending of Landsat and MODIS Surface Reflectance: Predicting Daily Landsat Surface Reflectance. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 8, pp. 2207—2218.

Genderen, J., and Pohl, C., 1994, Image Fusion: Issues, Techniques, and Applications. Intelligent Image Fusion. in 1994 *Proc. EARSel Workshop*, Strasbourg, France, pp. 18-26.

General Electric Analytical Systems (GEAS), 1994. USEPA FAQs for TOC Monitoring. *Water & Process Technologies Analytical Instruments*, rev. B, Boulder, Colorado.

Goldberg, D., 1989, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, New York.

Green, R., Heiser, E., Korategere, V., and Shank, D., 2010, *Meshing Treatment Objectives, Water Quality Goals and Regulatory Requirements into a Plant Expansion Project*, Clermont County Water Resources Department.

Hall, D., 1992, *Mathematical Techniques in Multisensor Data Fusion*. Norwood: Artech House, Inc.

Heikkinen, 1989, Organic carbon transport in undisturbed boreal humic river in northern Finland. *Archiv Fur Hydrobiologie*, vol. 117, pp. 1-19.

Holland, J.M., 1975, *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press.

Ioannou, I., Gilerson, A., Gross, G., Moshary, F., and Ahmed, S., 2011, Neutral Network approach to retrieve the inherent optical properties of the ocean from observations of MODIS. *Applied Optics*, vol. 50, pp. 3168-3186.

Ivarsson, H., and Jansson, M., 1994, Temporal variations of organic carbon in River Ore, Northern Sweden. *Verhandlugen, International Vereinigung fur Theoretische und Angewandte Limnologie*, vol. 25, pp. 1522-1525.

Kintao, H., Asada, M., Kuniyoshi, Y., Noda, I., and Osawa, E., 1995, Robocup: The robot world cup initiative. In *Proceedings of the Workshop on Entertainment and AI/ALife, International Joint Conferences in Artificial Intelligence*.

Koza, J.R., 1992, *Genetic Programming: On the Programming of Computers by Means of natural Selection*. Cambridge, MA: MIT Press.

Liu, L., Cai, H., Ying, M., and Le, J., 2007, RLGP: An Efficient Method to Avoid Code Bloating on Genetic Programming. International Conference on Mechatronics and Automation, August 5-7, 2007, Harbin, China.

Luke, S., 2000, Issues in Scaling Genetic Programming: Breeding Strategies, Tree Generation, and Code Bloat, doctoral dissertation, University of Maryland.

Mangolini, M., 1994, Apport de le fusion d'images satellitaires multicapteurs au niveau pixel en teledetection et photo-interpretation, M.S. thesis, Univ. of Nice—Sophia Antipolis, France.

Masek, J., Vermote, E., Saleous, N., Wolfe, R., Hall, F., Huemmrich, F., Gao, F., Kutler, J., and Lim, T., 2006, A Lansat Surface reflectance data set for North America, 1990-2000. Geoscience and Remote Sensing Letters, vol. 3, pp. 68-72.

Menken, K., Brozonik, P., and Bauer, M., 2005, *Influence of Chlorophyll and Colored Dissolved Organic Matter (CDOM) on Lake Reflectance Spectra: Implications for Measuring Lake Properties by Remote Sensing*. University of Minnesota, MN.

Naden, P., and McDonald, A., 1989, Statistical Modeling of Water Colour in the Uplands: The Upper Nidd Catchment 1979-1987. *Environmental Pollution*, vol. 60, pp. 141-163.

Nieto, P., Fernandez, J., Juez, F., Lasheras, F., and Muniz, C., 2013, Hybrid modeling based on support vector regression with genetic algorithms in forcasting cyanotixins presense in the Trasona reservoir Northern Spain. *Environmental Research*, vol. 122, pp. 1-10.

O'Connor, J., 2004, *Removal of Total Organic Carbon*. H2O'C Engineering.

Ohmori, Y., Kozu, T., Shimomai, T., Seto, K., and Sampei, Y., 2010, Feasibility Study of TOC and C/N Ratio Estimation from Multispectral Remote Sensing Data. *Remote Sensing of Spatial Information Science*, vol. 38, pt. 8.

Parks, S., and Baker, L., 1997, Sources and Transport of Organic Carbon in an Arizona River-Reservoir System. *Water Research*, vol. 31, pp. 1751-1759.

Pohl, C., and Genderen, J., 1998, Multisensor Image Fusion in Remote Sensing: Concepts, Methods, and Applications. *International Journal of Remote Sensing*, vol. 19, no. 5, pp. 823—854.

Robin, A., Hegarat, S., and Moisan, L., 2005, A Multiscale Multitemporal Land Cover Classification Method Using a Bayesian Approach. *Image and Signal Processing for Remote Sensing*, vol. 11.

Scott, M., Jones, M., Woof, C., and Tipping, T., 1998, Concentrations and fluxes of dissolved organic carbon in drainage water from an upland peat system. *Environmental International*, vol. 24, pp. 537-546.

Seifert, J., 2004, Data Mining: An Overview. *CRS Report for Congress*.

Shen, S., 1990, Summary of Types of Data Fusion Methods Utilized in Workshop Papers. in *Multisource Data Integration in Remote Sensing, Proc. Of Workshop*, Maryland, pp. 145—149.

Smith, R. and Baker, K., 1978, The Bio-optical State of Ocean Waters and Remote Sensing. *Limnology and Oceanography*, vol. 23, no. 2, pp. 247—259.

Song, K., Li, L., Tedesco, L., Li, S., Clercin, N., Hall, B., Li, Z., and Shi, K., 2012, Hyperspectral determination of eutrophication for a water supply source via genetic algorithm-partial least squares (GA-PLS) modeling. *Science of the Total Environment*, vol. 426, pp. 220-232.

Stedmon, C., Markager, S., and Kass, H., 2000, Optical Properties and Signatures of Chrolophoric Dissovled Organic Matter in Danish Coastal Waters. *Estuarine, Coastal and Shelf Science*, vol. 51, no. 2, pp. 267—278.

Stramski, D., Reynolds, R., Kahru, M., and Mitchell, B., 1999, Estimation of Particulate Organic Carbon in the Ocean from Satellite Remote Sensing. *Science*, vol. 285.

Texas Commission on Environmental Quality (TCEQ), 2002, *Total Organic Carbon (TOC) Guidance Manual*. Water Supply Devision, Austin, Texas.

Thurman, E., 1985, *Organic Geochemistry of Organic Waters.* Martinus Nijhoff/Dr. W. Junk Publishers, Dordrecht, The Netherlands.

United States Army Corps of Engineers (USACE). *William H. Harsha Lake: Benefits* [Online]. Available: http://www.lrl.usace.army.mil/whl/

United States Environmental Protection Agency (USEPA), 1998, National Primary Drinking Water Regulations: Disinfectants and Disinfection Byproducts. *Federal Register: Rules and Regulations*, vol. 68, no. 241, Boulder, Colorado.

United States Geological Survey (USGS). *Landsat Processing Details* [Online]. Available: http://landsat.usgs.gov/Landsat_Processing_Details.php

Vermote, E., 1997, Atmospheric correction of visible to middle-infrared EOS-MODIS data over land surfaces: Background, operational algorithm, and validation. *Journal of Geophysical Remote Sensing*, vol. 35, pp. 1179-1188.

Vermote, E., Kotchenova, S., and Ray, J., 2011, *MODIS Surface Reflectance User's Guide. v1.3*, MODIS Landsat Surface Reflectance Science Computing Facility.

Vertucci, F., 1989, Spectral Reflectance and Water Quality of Adirondack Mountain Region Lakes. *Limnology and Oceanography*, vol. 34, no.8, pp. 1656-1672.

Vincent, R., Xiaoming, Q., McKay, R., Miner, J., Czajkowski, K., Savino, J., and Bridgeman, T., 2004, Phycocyanin detection from Landsat TM data for mapping cyanobacterial blooms in Lake Erie. *Remote Sensing of the Environment*, vol. 89, pp. 381-392.

Visco, G., Campanella, L., and Nobili, V., 2005, Organic carbons and TOC in waters: an overview of the international norm for its measurements. *Microchemical Journal*, vol. 79, pp. 185-191.

Zeng, Y., Zhang, J., and Genderen, J., 2006, Comparison and Analysis of Remote Sensing Data Fusion Techniques at Feature and Decision Levels. *From Pixels to Processes*.

CHAPTER 3: COMPARATIVE SENSOR FUSION BETWEEN HYPERSPECTRAL AND MULTISPECTRAL REMOTE SENSING DATA FOR MONITORING MICROCYSTIN DISTRIBUTIONS IN LAKE ERIE

3.1. Introduction

Human population growth and agricultural use have led to the increase in eutrophic conditions in surface waters. The subsequent influx of nutrients has fueled cyanobacteria-dominated algal blooms in polluted waters in many parts of the globe (WHO 1999). Blooms containing toxins that negatively impact human health and the environment are referred to as harmful algal blooms (HABs). Not only can HABs form and spread rapidly, but wind and water currents will mobilize the blooms (Lekki *et al.* 2009). The dynamic movement of the HABs requires constant monitoring and forecasting, due to the threat posed to humans recreating on the lake, commercial fishing operations, and water treatment facilities. The predominant species of cyanobacteria that produce cyanotoxins are *Microcystis aeruginsa*, *Microcystis viridis*, *Aphanizomenon flos-aquqe*, and *Anabaena*. While there are a variety of cyanotoxins, microcystin is the main toxin produced (WHO 1999, Hitzfield 2000). The aberrant toxicity of microcystin can lead to liver cancer, liver failure, and even death (Toivola *et al.* 1994, WHO 1999). To ensure the protection of human health from microcystin exposure, it is necessary to develop a reliable method for the near real-time prediction of microcystin within hazardous algal blooms.

Satellites can provide medium to high resolution images of selected light spectrums. Using the detected surface reflectance emissions, predictions of microcystin concentrations are possible. The theoretical basis for this claim lies behind the fact that every substance gives off a unique spectral signature. As a substance is exposed to different portions of the electromagnetic spectrum, it will reflect a certain percentage of the light. The percentage of reflectance can be plotted as a function of wavelength to clearly display which frequencies the substance has an affinity for absorbing and reflecting. The unique curve that is produced is known as a spectral signature. The substance will have the defining spectral peaks and troughs, almost like a fingerprint for that object, where much of the radiation has either been reflected or absorbed. The intensity of the reflectance at different wavelengths can be then used to determine the amount of the substance present in the water. However, the relationship between reflectance and concentration is highly nonlinear for certain substances. As a result, effective data mining techniques must be applied to accurately predict the concentration for an observed spectral response. In this paper, we demonstrate the utility, technical difficulties, as well as data mining approaches for near real-time monitoring of microcystin concentrations in Lake Erie.

3.2. Literature Review

The prediction of microcystin concentrations in a lake poses a unique problem, since 95% of the microcystin is contained within healthy *Microcystis* cells (Jones and Orr 1994). It is not until death or induced rupture of the cell wall that the toxin is released. Thus, in order to generate an accurate estimate of microcystin concentration, it is necessary to establish a relationship between microcystin and other substances present in the water. These substances will serve as indicators of microcystin concentration. Chlorophyll-a levels in *Microcystis* blooms are related to the amount of microcystin present (WHO 1999, Rogalus and Watzin 2008, Rinta-Kanto *et al.* 2009). Since *Microcystis* is a bacterium that uses photosynthesis for energy production, it is reasonable to conclude that high concentrations of *Microcystis* can be linked with elevated chlorophyll-a levels. In a study by Budd *et al.* (2001), algal blooms were detected and tracked using AVHRR and Landsat Thematic Mapper (TM) images to determine chlorophyll-a concentrations in the lake. Their study established that it is possible to use surface reflectance data to detect and track algal blooms based upon chlorophyll-a levels, and Wynne *et al.* (2008) expanded the depth of this study by using the surface reflectance of chlorophyll-a to specifically predict *Microcystis* blooms, instead of algal blooms in general. It was discovered that *Microcystis* blooms can be distinguished from other cyanobacteria blooms through close analysis of the detected surface reflectance at 681 nm (Ganf *et al.* 1989). Studies by Mole *et al.* (1997) and Ha *et al.* (2009), had similar findings for chlorophyll-a as an indicator for microcystin in algae blooms that have stabilized, and reached the late exponential growth phase and stationary phase. In summary, chlorophyll-a is a reliable indicator of microcystin for *Microcystis* HABs that are no longer in the peak of the exponential growth phase.

Phycocyanin is a pigment that all cyanobacteria contain (WHO 1999), and it has been shown that phycocyanin concentrations share a positive correlation with microcystin levels (Rinta-Kanto *et al.* 2009). In a study by Vincent *et al.* (2004). Landsat TM images in the visible and infrared spectral bands were used to generate algorithms to predict phycocyanin concentrations with 73.8% to 77.4% accuracy. Thus, the surface reflectance of phycocyanin, chlorophyll-a, and *Microcystis* are suitable indicators for the prediction of microcystin levels in a lake. The surface reflectance curves for chlorophyll-a and phycocyanin in surface waters peak at 525 nm, 625 nm, 680 nm, and 720 nm, and these spectral peaks are aptly captured by Landsat and MODIS satellites (multispectral fusion pair), as well as MERIS and MODIS (hyperspectral fusion pair). However, a significant drawback of Landsat and MERIS is their 16 and 3 day revisit times. Daily revisit time of the MODIS sensor can fill in the data gaps through the use of data fusion. However, MODIS alone cannot be used as a substitute because of its poor 250/500 m spatial resolution for the land bands, which is outclassed by Landsat's 30 m resolution, and its 1,000 m spatial resolution for the ocean bands, which is enhanced by MERIS' 300 m resolution. We propose an ultimate solution by fusing Landsat and MODIS (MODIS' land bands) or MERIS and MODIS (MODIS' ocean color bands) pairwise to generate a synthetic image with both enhanced spatial and temporal resolutions. Such a synthetic image can enable near real-time monitoring of microcystin concentrations, creating seasonal maps, and populating a database with information on spatial occurrence and its timing of HABs in the lake and general movement patterns. The information provides water treatment, fishing operations, and areal residents with the knowledge required in decision-making.

Having presented the rational for fusing the selected satellites, the next consideration is given to the intercomparisons between multispectral hyperspectral remote sensing data. Multispectral sensors collect a smaller number of noncontiguous, wide spectral bands (less than 20) (Belokon *et al.* 1997, Pabich 2002, Shippert n.d., Bianco n.d.); they typically offer enhanced spatial resolution. Hyperspectral sensors, on the other hand, provide greater spectral solution by capturing a greater number of spectral bands with a bandwidth of 10 - 20 nm. Since the emergency in the 1970s and 80s, hyperspectral remote sensing techniques have advanced significantly with the development of the Compact Airborne Spectrographic Imager (CASI) in 1978 and the proposal of the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) to the National Aeronautics and Space Administration (NASA) in 1983. Due to the small width of each band, hyperspectral sensors have a worse lower signal-to-noise ratio than multispectral scanners; the latter collect more photons per band, thus are able to lessen the impact of the noise (Chang *et al.* 2004).

Multispectral sensors aptly function in the open ocean, from which a few, select bands can be used to monitor water quality constituents (O'Reilly *et al.* 1998). Case 2 water bodies (such as the one examined in this study) exhibit significantly more optical complexity than the open ocean, and the algorithms utilizing multispectral data products have diminished performance in these waters (Hu *et al.* 2000, Lee and Carder 2002). The benefit of hyperspectral sensors is the number of additional bands they provide at a finer bandwidth. The added bands with more narrow bandwidths more accurately depict the spectral reflectance curve of the water body, as is depicted in Figure 3.1:

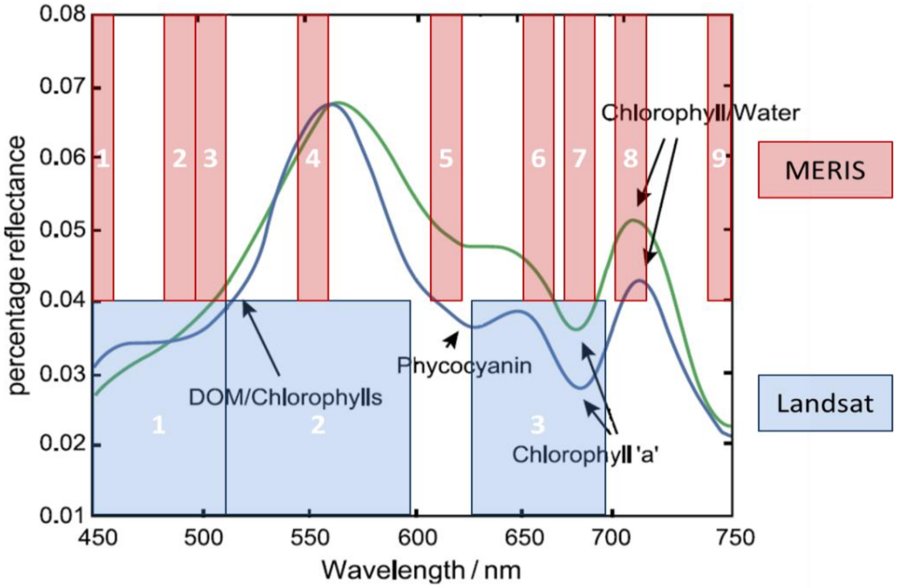


Figure 3.1: Comparing the bandwidths between multispectral (Landsat) bands and hyperspectral (MERIS) bands (adapted from Vincent *et al.* 2004).

From Figure 3.1, one can see that the defining peaks and troughs for Landsat are smoothed out when the total reflectance is averaged for the bands; therefore, the resulting band is unable to reveal detailed reflectance information that may be necessary for differentiating between water quality constituents and characterizing the species within a phytoplankton bloom. In comparison, MERIS captures the unique spectral features (peaks and troughs). When combining the multi-spectral Landsat and hyperspectral MERIS, one may be able to take the advantage of each satellite imagery techniques and address the aforementioned short-comings. The application of this hypothesis for ocean color remote sensing has been reported by Lubac *et al.* (2008), which concluded that both multispectral and hyperspectral data can be used to quantify phytoplankton blooms, yet an enhanced hyperspectral resolution provides superior quantitative assessment and monitoring of phytoplankton blooms. Through this approach, the superior detail of hyperspectral information introduces more degrees of freedom, and allows for optical models and algorithms of higher explanatory power to quantify the nonlinear relationships between surface reflectance and concentrations, more accurately classifying species and concentrations of water quality constituents, and enhancing the determination of inherent optical properties that vary with depth (Chang *et al.* 2004, Torrecilla *et al.* 2009).

The goal of this study was to develop the Integrated Data Fusion and Mining (IDFM) technique of combing multispectral and hyperspectral data, and further to quantify the performance for providing near real-time monitoring of the spatiotemporal distributions of microcystin in a lake . In addition to real-time monitoring, seasonal maps of microcystin were retrieved to assess HABs spatial distribution throughout the year. In this paper, we wish to explore: 1.) the feasibility in predicting microcystin concentrations in a lake using the IDFM technique, 2.) which of the fused band combinations are most useful in determining microcystin concentrations in a case-2 inland water body, and 3.) whether hyperspectral data products provide a significant advantage over multispectral data products for microcystin prediction.

3.3. Methodology

3.3.1. Study Site

Lake Erie is one of the five Great Lakes located in North America. Together, the lakes make up the largest supply of fresh water in the world. The lakes provide drinking water for over 40 million Americans, in addition to 56 billion gallons per day withdrawn from the lakes for industrial, agricultural, and municipal use (Lekki *et al.* 2009). Each summer, the Great Lakes are threatened by Microcystis blooms, yet the blooms in western Lake Erie are the most severe and contain levels of microcystin that are not suitable for drinking water. Throughout the 2000's, Microcystis blooms have increased in frequency and severity (Bridgeman 2005, Ouelette *et al.* 2006).

3.3.2. Satellites Used and In Situ Data

Surface reflectance data utilized in this study were obtained from Landsat TM, MERIS and MODIS sensors. MERIS is a hyperspectral sensor with moderation 300 m resolution, and has a 3-day revisit time for sites near the equator. Landsat offers superior spatial resolution at 30 m; however, the spectral resolution is much poorer than MERIS. The revisit time of Landsat is significantly longer at 16 days. In developing the near real-time monitoring system, daily satellite images of the area of interest are required for which data fusion techniques are used to fill in the data gaps in MERIS and Landsat by using the MODIS sensor of daily revisit time. The spatial, temporal, and spectral resolutions of these two satellites central to this technical approach are compared in Table 3.1:

Table 3.1: Spatial, temporal, and spectral properties of the satellite sensors used in this study. The band centers shared between the satellites have been aligned in the table. Band combinations that occur on the same row are suitable candidates for spectral fusion.

Parameters	Hyperspectral Sensor Pair		Multispectral Sensor Pair	
	MERIS	MODIS TERRA (ocean bands)	Landsat TM	MODIS TERRA (land bands)
Product	MER_FR_2P	MODOCL2	LT5	MODO9
Spatial Resolution	300 m	1000 m	30 m	250/500 m
Temporal Resolution	1-3 days	1 day	16 days	1 day
Band Number: Band Center ± Band Width (nm)	1: 412 ± 10	8: 413 ± 15	1: 485 ± 35	3: 469 ± 10
	2: 443 ± 10	9: 443 ± 10	2: 570 ± 40	4: 555 ± 10
	3: 490 ± 10	10: 488 ± 10	3: 660 ± 30	1: 645 ± 25
	4: 510 ± 10		4: 840 ± 60	2: 859 ± 18
		11: 531 ± 10	5: 1650 ± 100	6: 1640 ± 12
	5: 560 ± 10	12: 551 ± 10	7: 2090 ± 130	7: 2130 ± 25
	6: 620 ± 10			
	7: 665 ± 10	13: 667 ± 10		
	8: 681 ± 10	14: 678 ± 10		
	9: 708 ± 10			
	10: 753 ± 10	15: 748 ± 10		
	11: 760 ± 10			
	12: 779 ± 10			
	13: 865 ± 10	16: 869 ± 15		

As shown in the table, MERIS is fused with the ocean color bands of MODIS. The pixel level data fusion of STARFM requires input images to be spectrally similar. Accordingly, Landsat TM were fused with the land bands of MODIS.

The National Oceanic and Atmospheric Administration (NOAA) is the sole provider of the *in situ* data for microcystin concentration. NOAA collects surface water samples in western Lake Erie, when probable blooms are identified based on their analysis of satellite images. Samples are taken at the surface to provide surface microcystin concentrations that coincide with the surface reflectance observed in satellite data products. ELISA techniques are used to quantify total microcystin concentration. In total, 44 microcystin measurements were made from 2009 to 2011 available for ground-truth (Table 3.2). These data only include those with sampling locations free of cloud-cover, land aerosol contamination, and significant suspended sediment levels in the corresponding satellite images.

Table 3.2: Ground-truth samples were taken at various sites in western Lake Erie on these days.

	Jun.	Jul.	Aug.	Sep.
2009		7,14		
2010	28	26	2,16,30	2
2011		12	11	14

3.3.3. Methodology

The IDFM technique for the prediction of microcystin is shown in Figure 3.2. It is designed to fuse satellite data streams and apply machine-learning algorithms to derive a working model relating the data streams to the desired output parameter. For this study, Landsat, MERIS, and MODIS surface reflectance imagery serve as the data streams, and the estimated concentration of the toxin microcystin is the desired output parameter. Data mining techniques are applied to incorporate data into a single image for analysis by machine learning techniques, which create prediction models for near real-time monitoring and data gap-filling applications.

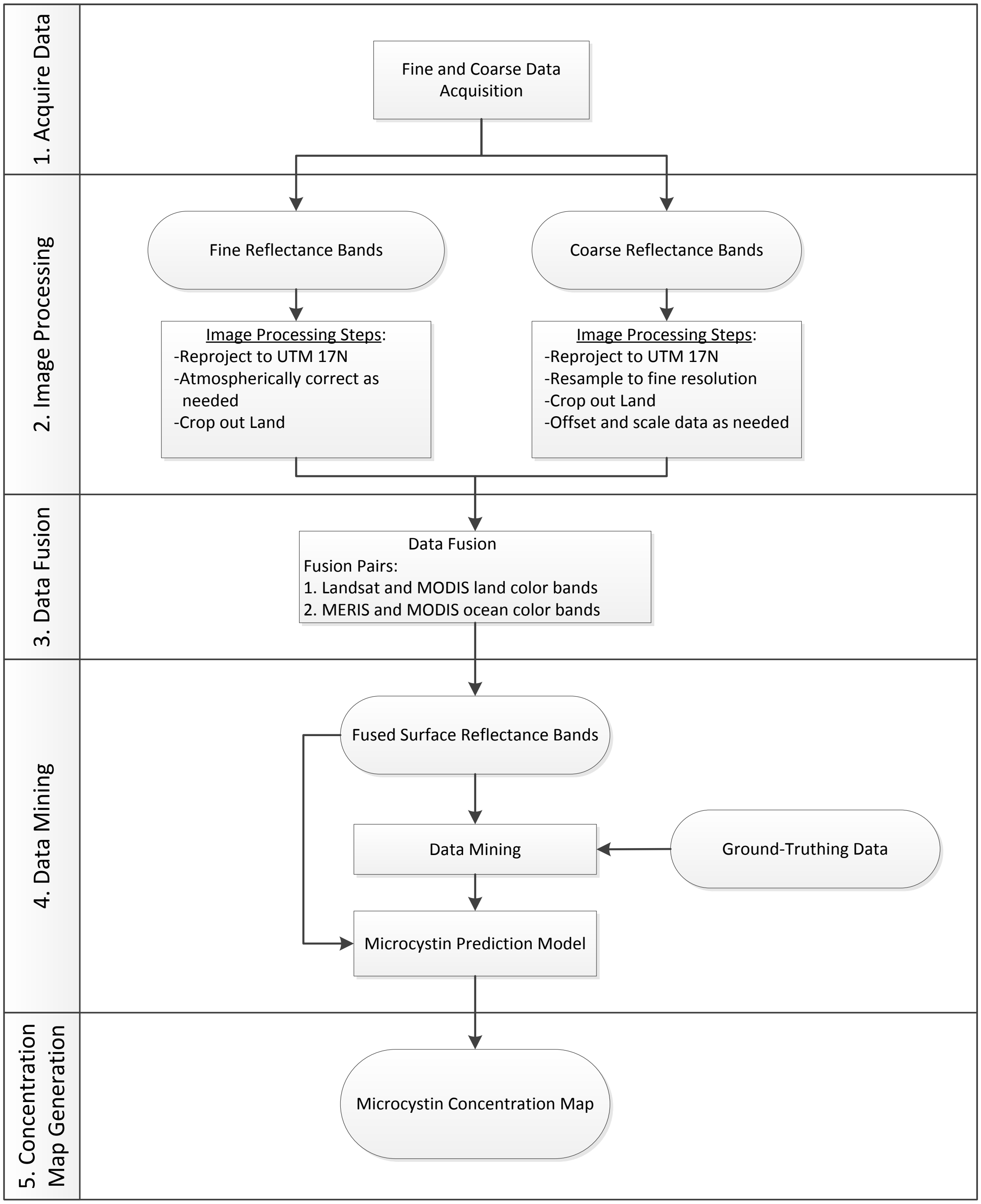

1. Acquire Data
2. Image Processing
3. Data Fusion
4. Data Mining
5. Concentration Map Generation
Fine and Coarse Data Acquisition
Fine Reflectance Bands
Coarse Reflectance Bands
Image Processing Steps:
-Reproject to UTM 17N
-Atmospherically correct as needed
-Crop out Land
Image Processing Steps:
-Reproject to UTM 17N
-Resample to fine resolution
-Crop out Land
-Offset and scale data as needed
Data Fusion
Fusion Pairs:
1. Landsat and MODIS land color bands
2. MERIS and MODIS ocean color bands
Fused Surface Reflectance Bands
Data Mining
Ground-Truthing Data
Microcystin Prediction Model
Microcystin Concentration Map

Figure 3.2: Methodological flowchart for the IDFM procedure using hyperspectral or multispectral data

The IDFM technique consists of five main steps. Step one is the acquisition of the surface reflectance data from MERIS and MODIS. The second step formats the images for fusion followed by the application of data fusion techniques and algorithms. This study employed the Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM) algorithm to fuse the MODIS (ocean bands) and MERIS pair, and also the MODIS (land bands) and Landsat pair. In Step four, the synthetic images and ground-truth data were used as inputs for data mining. A genetic programing (GP) model was trained in Discipulus to create an explicit, nonlinear equation relating the fused band data to the ground-truth data. Lastly, the fifth step uses the GP model created in step four to compute microcystin concentration maps using the fused band data generated in Step Three (Figure 3.2).

3.3.3.1. Data Acquisition [Figure 3.2; Step 1]

Surface reflectance data for Lake Erie were collected from the ENVISAT MERIS, Terra MODIS satellite, and Landsat TM sensors. Level 2, ocean-band images for 2009-2011 from the Terra MODIS satellite were downloaded from the online repository through the NASA Ocean Color Web. Since multiple ground-truth samples were taken at different locations on the same day, the MODIS images were inspected for cloud cover at each of the locations. The level 2 image was downloaded as an HDF-EOS image, only when at least one location was not obstructed by cloud cover. The same criterion was applied to the rest of the satellite data acquired. Additionally, Level 2, land-band images for Terra MODIS were downloaded from the online repository overseen by the NASA Land Processes Distributed Active Archive Center (LP DAAC), United States Geological Survey (USGS), and Earth Resources Observation and Science Center (EROS). Level 2, full resolution MERIS data was obtained through the European Space Agency (ESA). Failure of the MERIS sensor in 2012 prevented usage of ground-truth data during this time period. Lastly, the Landsat TM data was obtained through the USGS by way of the Global visualization Viewer, which is maintained by LP DAAC, USGS, and EROS.

3.3.3.2. Image Processing [Figure 3.2; Step 2]

The MODIS images were preprocessed at a level 2 basis. This includes the radiometrically calibrated data that were atmospherically corrected for aerosols and scattering (Vermote *et al.* 2011). Full resolution MERIS data came processed on a level 2 basis, with radiometric, geometric, and atmospheric corrections (ESA 2006). Landsat data came processed on a level-1T basis with radiometric and geometric corrections (USGS, n.d.). Because the fusion process requires input image pairs to have the same bit-depth and spatial resolution, the input images were processed in ArcGIS, a mapping and spatial analysis software. Specifically, MERIS images were processed by:

- Reproject to the Universal Transverse Mercator (UTM) zone 17 North
- Crop out land data surrounding Lake Erie

The MODIS ocean-band images were processed in a similar manner:

- Reproject to UTM zone 17 North
- Resampling to the 300-m spatial resolution to match those of MERIS
- Land data was cropped out from around Lake Erie
- Surface reflectance values were recalculated to using the same offset and scaling applied to MERIS data

The MODIS land-band images were processed in a similar manner:

- Reproject to UTM zone 17 North
- Resampling to 30 m spatial resolution to match Landsat
- Land data was cropped out from around Lake Erie
- Surface reflectance values were recalculated to using the same offset and scaling applied to MERIS data

Landsat images were processed as follows:

- Atmospherically correct Landsat images using the LEDAPS Processing software available through NASA
- Reproject to UTM zone 17 North
- Land data was cropped out from around Lake Erie

The image processing consists of three categories of actions: 1) modification of the geometric projections, pixel size, bit depth, and scale in order to fuse them properly, 2) atmospheric correction, and 3) preparation of the image to increase the accuracy of the fused image by removing land contamination from the original images. With regard to the first category, the images need to have the same geographic projection and scaling prior to fusion. Otherwise data for the same pixels between the satellite pairs becomes incomparable, because they represent different swaths of land and have differently scaled values. In this study, all images were projected to the UTM 17N. Additionally, the MODIS ocean-band data were resampled to match the resolution of MERIS and Landsat. In color modification, the MODIS ocean color band surface reflectance values were scaled upward to match those of MERIS; Because of the dimensionless integer values in MERIS, integer values are required for input into STARFM. This ensures that both input images have the same number of pixels, enabling the pixel by pixel comparison techniques used by STARFM. Initially, the Landsat and MODIS data are not the same bit depth, but atmospheric correction using the LEDAPS Processing tool scales the Landsat values from 0—255 to -100—16,000 (Vermote *et al.* 2002, Masek *et al.* 2005). This is because the same MODIS 6S radiative transfer techniques are applied to correct the Landsat data.

For the other two categories, atmospheric correction is needed to remove the scattering effects of the atmosphere from the raw data, thus, producing surface reflectance instead of top of atmosphere radiance. The last category of processing was performed on all images with the purpose to mask the pixel values for the land surrounding Lake Erie. This step is required to prevent fusing land pixel values with surface water values during processing with the STARFM algorithm.

3.3.3.3. Data Fusion [Figure 3.2; Stage 3]

A fused image is created by the algorithmic fusion of the spectral, temporal, and spatial properties of two or more images (Genderen and Pohl 1994). The resulting synthetic or fused image has all the characteristics of the input images, and incorporates object's defining attributes into a single image with a potential to increase the reliability of the data (Pohl and Genderen 1998). Fusion of spatial and temporal properties on a pixel by pixel basis was required for this study based on the STARFM algorithm by NASA. For this study, the algorithm was used to fill in data gaps caused by the 1-3 day revisit time of MERIS using MODIS ocean color bands, and the 16 day revisit time of Landsat using MODIS land color bands. The Landsat and MERIS images are of higher quality than MODIS, but they are sparse in time. As a result, MODIS data are used to capture temporal changes during the periods of data gaps. The overall workflow of the STARFM algorithm is detailed in Figure 3.3:

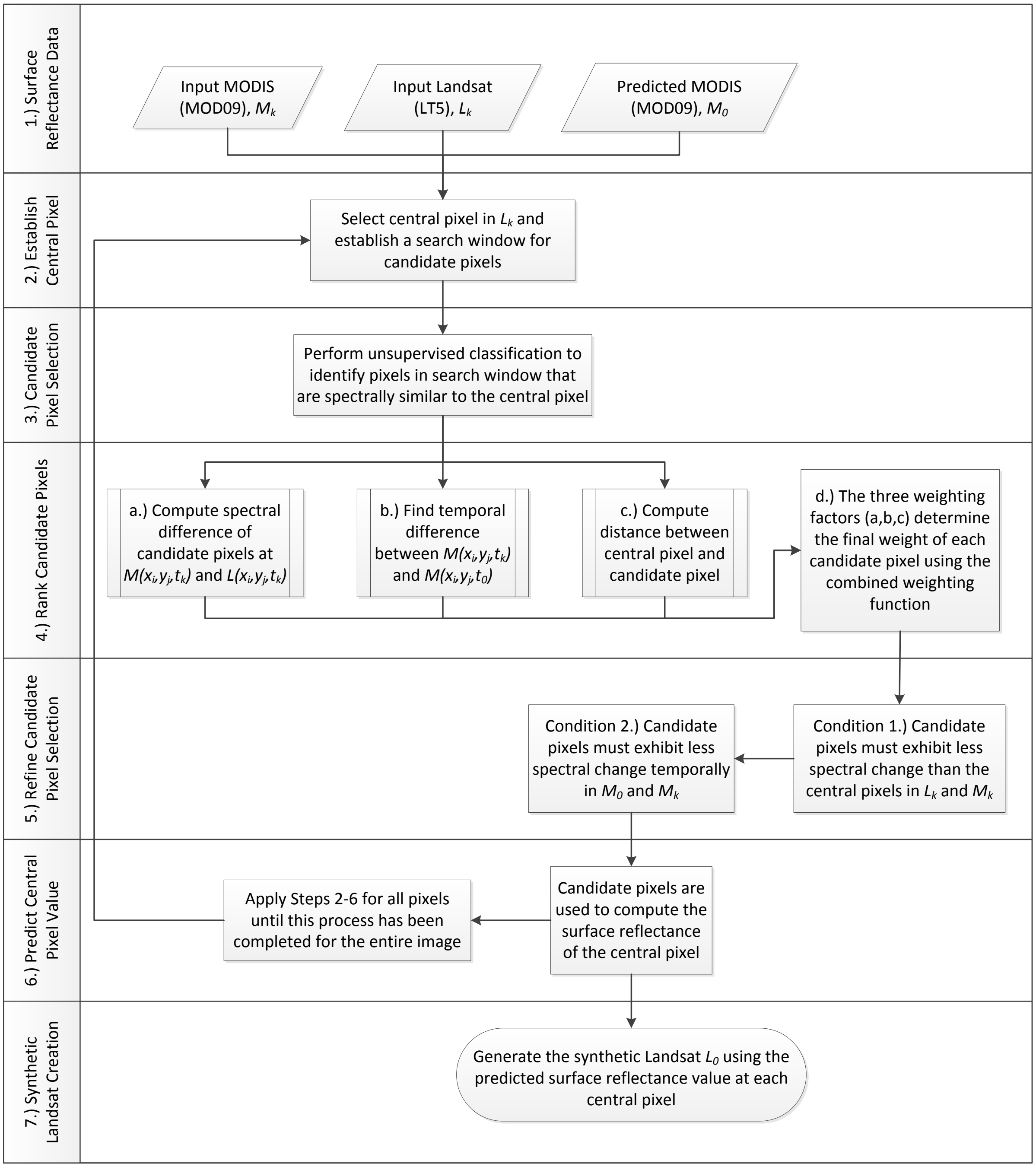

1.) Surface Reflectance Data
Input MODIS (MOD09), M_k
Input Landsat (LT5), L_k
Predicted MODIS (MOD09), M_0
2.) Establish Central Pixel
Select central pixel in L_k and establish a search window for candidate pixels
3.) Candidate Pixel Selection
Perform unsupervised classification to identify pixels in search window that are spectrally similar to the central pixel
4.) Rank Candidate Pixels
a.) Compute spectral difference of candidate pixels at $M(x_i,y_j,t_k)$ and $L(x_i,y_j,t_k)$
b.) Find temporal difference between $M(x_i,y_j,t_k)$ and $M(x_i,y_j,t_0)$
c.) Compute distance between central pixel and candidate pixel
d.) The three weighting factors (a,b,c) determine the final weight of each candidate pixel using the combined weighting function
5.) Refine Candidate Pixel Selection
Condition 1.) Candidate pixels must exhibit less spectral change than the central pixels in L_k and M_k
Condition 2.) Candidate pixels must exhibit less spectral change temporally in M_0 and M_k
6.) Predict Central Pixel Value
Candidate pixels are used to compute the surface reflectance of the central pixel
Apply Steps 2-6 for all pixels until this process has been completed for the entire image
7.) Synthetic Landsat Creation
Generate the synthetic Landsat L_0 using the predicted surface reflectance value at each central pixel

Figure 3.3: Procedural flow for the STARFM algorithm as shown for the MODIS and Landsat fusion pair. The same process is applied when using MERIS (substitute for Landsat) and MODIS ocean color products (Gao *et al.* 2006).

The methodology described here is data fusion using the Landsat and MODIS images as the input pair; the same approach applies for the MERIS (fine spatial resolution image) and MODIS pair. In order to generate a synthetic Landsat image L_0 (the higher spatial resolution image) to fill in the data gap at time t_0, a MODIS image M_0 from that day is required. This MODIS image is referred to as the predicted MODIS in following discussions. Next, a temporally analogous MODIS M_k and Landsat L_k image pair are required from before or after the prediction date t_0. The input image pair (M_k and L_k) obtained for date t_k serve as a boundary condition detailing how the area of interest looked prior to or after t_0. The dates (t_0 and t_k) of the input and predicted images should be as close as possible, because the chance for significant spectral change between the two images increases with time. In the case of the Landsat input image, it would preferable be from the next 16 day revisit cycle. Acquiring these images is step 1.

Step 2 involves selecting a central pixel from the L_k image. This is a sequential process, which starts with the first pixel in the image and then moves to the second pixel, etc. During each round of iteration, the central pixel value in L_k, M_k, and M_0 is the same. Step 3 identifies the candidate pixels which will be used to predict the reflectance of the central pixel in L_0. Unsupervised classification is performed on pixels that fall within the search box encompassing the central pixel; the user defines the size of the search box. Pixels sharing the same classification type as the central pixel are then selected as candidate pixels. Step 4 is designed to rank the candidate pixels based on three criteria that determined how much they are related to the central pixel. From Step 4a, the spectral differences between the candidate pixels in the input MODIS and Landsat are computed, as shown in Eq. 3.1 (Geo *et al.* 2006):

$$S_{ijk} = \left|L\left(x_i, y_j, t_k\right) - M(x_i, y_j, t_k)\right| \tag{3.1}$$

where x_i corresponds to a row in the image, y_j denotes a specific column, $L(x_i,y_j,t_k)$ refers to a specific pixel in the input Landsat image, and $\mathrm{M}(x_i,y_j,t_k)$ refers to a specific pixel in the input

MODIS image. Recall that the spatial resolution of MODIS can be up to 500 m, compared to the 30 m of Landsat. As a result, the spectra of objects within the 500 m are averaged for the MODIS pixel. This step determines how well the spectral reflectance of the MODIS pixel compares to the Landsat pixel value. If the two have minimal spectral differences, a small value for S_{ijk} will be computed and a high weighting will be assigned to that candidate pixel. Step 4b compares the spectral changes that occur temporally in the input and predicted MODIS images, as detailed in Eq. 3.2 (Geo *et al.* 2006):

$$T_{ijk} = |M(x_i, y_j, t_k) - M(x_i, y_j, t_0)| \tag{3.2}$$

where M(x_i,y_j,t_0) refers to a specific pixel in the predicted MODIS image. A large value of T_{ijk} indicates that there has been significant change in the water quality at this candidate pixel, and it is assigned a lower weighting. Step 4c follows the basic logic that candidate pixels closer to the central pixel should receive a higher weighting, as shown in Eq. 3.3 (Geo *et al.* 2006):

$$D_{ijk} = 1.0 + \frac{\sqrt{(x_{w/2}-x_i)^2+(y_{w/2}-y_j)^2}}{A} \tag{3.3}$$

where w is the user defined search box side length, $x_{w/2}$ is row of the central pixel, $y_{w/2}$ is the column of the central pixel, and A is a constant relating the importance of the spatial distance D_{ijk} to the spectral S_{ijk} and temporal T_{ijk} distances. Candidate and central pixels that are close together will likely exhibit similar spectral changes over time, whereas a candidate pixel farther from the central pixel is less spatially similar and it receives a lower weighting. Step 4d combines individual ranking criteria (D_{ijk}, S_{ijk} and T_{ijk}) to form an overall weighting factor for each candidate pixel. This is accomplished in Eqs. 3.4 and 3.5 (Geo *et al.* 2006):

$$C_{ijk} = \ln(S_{ijk} * B + 1) * \ln(T_{ijk} * B + 1) * D_{ijk} \tag{3.4}$$

$$W_{ijk} = \frac{\left(\frac{1}{C_{ijk}}\right)}{\sum_{i=1}^{w}\sum_{j=1}^{w}\sum_{k=1}^{n}\left(\frac{1}{C_{ijk}}\right)} \tag{3.5}$$

where *B* is a scale factor and *W_{ijk}* is the combined weighting factor. The value for B is 10,000 when using LEDAPS reflectance products, and a value of 54,645 was used for the MODIS ocean color and MERIS pair, since a scaling factor of $1.83*10^5$ is applied to MERIS products to store them as an integer.

Step 5 further refines the selection of candidate pixels based on two conditions shown in Eqs. 3.6 and 3.7 (Geo *et al.* 2006):

$$S_{ijk} < |L(x_{w/2}, y_{w/2}, t_k) - M(x_{w/2}, y_{w/2}, t_k)| \tag{3.6}$$

$$T_{ijk} < |M(x_{w/2}, y_{w/2}, t_k) - M(x_{w/2}, y_{w/2}, t_0)| \tag{3.7}$$

Eq. 3.6 requires that candidate pixels in the input image pair exhibit less spectral change than the central pixels, and Eq. 3.7 requires that candidate pixels in the input and predicted MODIS images show less temporal change than the central pixels. Otherwise, the pixel is considered a “worse neighboring pixel” and it is not used for the predicting the surface reflectance of the central pixel in the synthetic image. Now that a suitable subset of candidate pixels have been related to the central pixel, the predicted surface reflectance for the central pixel in the synthetic image is performed, as shown in Step 6. Steps 2-6 are repeated for all of the pixels. The entire process of summed up in Eq. 3.8 (Geo *et al.* 2006):

$$L(t_0) = \sum_{i=1}^{w}\sum_{j=1}^{w}\sum_{k=1}^{n} W_{ijk} * \left[L(x_i, y_j, t_k) - M(x_i, y_j, t_k) + M(x_i, y_j, t_0)\right] \tag{3.8}$$

where *$L(t_0)$* is the synthetic Landsat image formed using spatial information from the fine spatial resolution Landsat image and the temporal changes from the coarse MODIS images. It should be noted that pixels containing clouds cannot be used for the fusion process, and they must be masked out. Furthermore, if there is a great deal of change between the predicted date and a boundary image or one of the boundary images exhibits a significant temporal difference, then the fused results will be less accurate. Lastly, this explanation only showed one pair of input images being used to generate the synthetic image. If an additional input pair is used the results can be improved (Geo *et al.* 2006). Think of this as providing the algorithm with a set of both pre and post conditions, instead of just one. This study provided the algorithm with both pre and post condition, as long as they were cloud-free and taken within 2 revisit cycles of the prediction date.

3.3.3.4. Data Mining [Figure 3.2; Step 4]

The IDFM technique permits the use of numerous machine learning techniques to derive an explicit equation or black box model relating the fused surface reflectance data to the ground-truth observations. Notable data mining and machine learning algorithms include Genetic Programming (GP), Artificial Neural Networks (ANN), ANN and Adaptive Resonance Theory, Constrained Optimization Techniques, Adaptive Dynamic K-means, Principal Component Analysis, and Support Vector Machines (SVM). The GP model for this study was created using the Discipulus software package. The user provides the software with inputs and outputs, which are used to train and calibrate the model. During training, the accuracy of a model is determined using least-squares. Discipulus identifies 30 of the best programs, and the model exhibiting the highest fitness is usually selected (Francone 1998).

The GP models were compared against a traditional two-band model, which was solved through a linear regression model using band ratios instead of individual bands as explanatory variables (Vincent *et al.* 2004). The generic setup for a two-band model is shown in Eq. 3.9:

$$C_{MS} = a * \frac{Rrs(\lambda_1)}{Rrs(\lambda_2)} + b \tag{3.9}$$

where *Rrs(λ)* is the atmospherically corrected surface reflectance at the band center λ. The coefficients a and b denote the slope and intercept obtained through regression. Additionally, a spectral slope two-band model was included in the analysis (Dash *et al.* 2011). The spectral slope is calculated using Eq. 3.10:

$$Slope = \frac{R_{rs}(\lambda_1) - R_{rs}(\lambda_2)}{|\lambda_1 - \lambda_2|} \tag{3.10}$$

A non-linear exponential fit was used to determine the spectral slope coefficients relating the exponential increase of absorption with wavelength for chlorophyll and phycocyanin. For both of the models, band combinations were compared to determine the two bands possessing high correlation with microcystin and PC estimation (both indicators of *Microcystis*). The same training and calibration data sets used for creating the GP models were employed to train and calibrate the two-band model.

3.3.3.5. Concentration Map Generation [Figure 3.2; Step 5]

Microcystin concentration maps for western Lake Erie are generated by applying the GP model to the fused data product created in step 3. For each pixel of the fused image, there are six surface reflectance values, one corresponding to each band from MODIS and MERIS. For this study, these band values are used as variables in the explicit equation created from the GP model. As determined by the GP model, certain band values will share a strong relationship in the determination of the microcystin concentration, while others may offer weak explanatory power. Thus, the GP model uses the fused surface reflectance values of the pixel to predict the microcystin concentration at that location. After this process is applied to the entire lake, a clear depiction of microcystin blooms is available. Analysis of these maps can lead to the discovery of yearly problem spots, factors that contribute to microcystin generation, and probable directions of travel for the blooms.

3.4. Results and Discussion

3.4.1. Method Reliability

An IDFM-based early warning system for quantifying toxin levels in algal blooms using satellite remote sensing data depends upon two primary constituents for success: 1.) accurate surface reflectance data of the water body and 2.) a reliable algorithm for predicting microcystin concentration. This section will quantify the advantages of using data-heavy hyperspectral products (MERIS and MODIS ocean color bands) over multispectral products (Landsat and MODIS land bands) using traditional two band inversion models and more computationally-intensive GP models. As detailed in Table 3.2, 44 ground-truth samples were used to train and calibrate the models. 60% of the input data was used to train the models, and the remaining 40% was used to validate the performance of the model. The method for splitting up the data into training and calibration sets is as follows: 1) order the ground-truth values from low to high 2) in an alternating manner, assign data to the training and validation data sets. This procedure exposes the models to the same range of microcystin concentration values during training and validation.

While the traditional two-band models will always yield the same coefficients when solved using regression techniques, the equation and performance of each GP model will vary during different runs. This is a result of the random starting weights and the fundamental methodology used during model creation. To lucidly depict the variation and average performance of the GP models, Discipulus was used to train 5 models for both multispectral and hyperspectral inputs. The coefficient of determination, time required to computing each model, and the run number of each model are detailed in Table 3.3:

Table 3.3: Statistical comparison between GP models created using fused multispectral and hyperspectral data sets.

	Model Number	1	2	3	4	5	**AVG**
Fused Multispectral GP Models	R2	0.8425	0.7931	0.7683	0.8449	0.8344	**0.8166**
	Run Time (s)	194	272	246	92	5	**162**
	Run Number	34	43	40	26	4	**29**
Fused Hyperspectral GP Models	R2	0.8243	0.9270	0.8847	0.9177	0.8879	**0.8883**
	Run Time (s)	211	450	437	932	389	**484**
	Run Number	30	50	48	71	43	**48**

The GP models using fused hyperspectral data products took 322 seconds longer to solve on average, yet the resulting coefficient of determination was 0.8883, which is 0.0717 greater than the coefficient of determination derived from fused multispectral data products. The multispectral solutions had shorter run times, since they stopped improving earlier on in the model development. The greater coefficients of determination for the hyperspectral GP models is attributed to the finer band widths (refer to Figure 3.1), which allows for telltale peaks and troughs of chlorophyll-a and phycocyanin (indicators of microcystin) to be more readily identified. An interesting observation is made by analyzing the run times for the multispectral GP models. The 5th model was derived in a mere 5 seconds, while the next closest solution was obtained in 92 seconds. The order of magnitude difference is rare, yet it is due to randomly generated starting weights and randomly selected input data that are used to initiate model formulation. In this case, the program stumbled upon an excellent combination of parameters for determining the relationship between surface reflectance and microcystin concentration.

Verifying that the model has appropriately related the band data from the fused images to the microcystin ground-truth data is accomplished through a least squares analysis between the observed and the predicted microcystin values. The best model is selected based on the coefficient of determination, fitness level achieved, and a visual confirmation that the model can accurately identify peak microcystin values. Identification of high microcystin values is imperative for an early warning system. Based on these criteria, the fourth fused multispectral GP model and the second fused hyperspectral GP model from Table 3.3 were selected for further analysis. The predictive capabilities of 3 GP models developed from pure MERIS (A & D), fused multispectral data (B & E), and fused hyperspectral data (C & F) are presented in Figure 3.4:

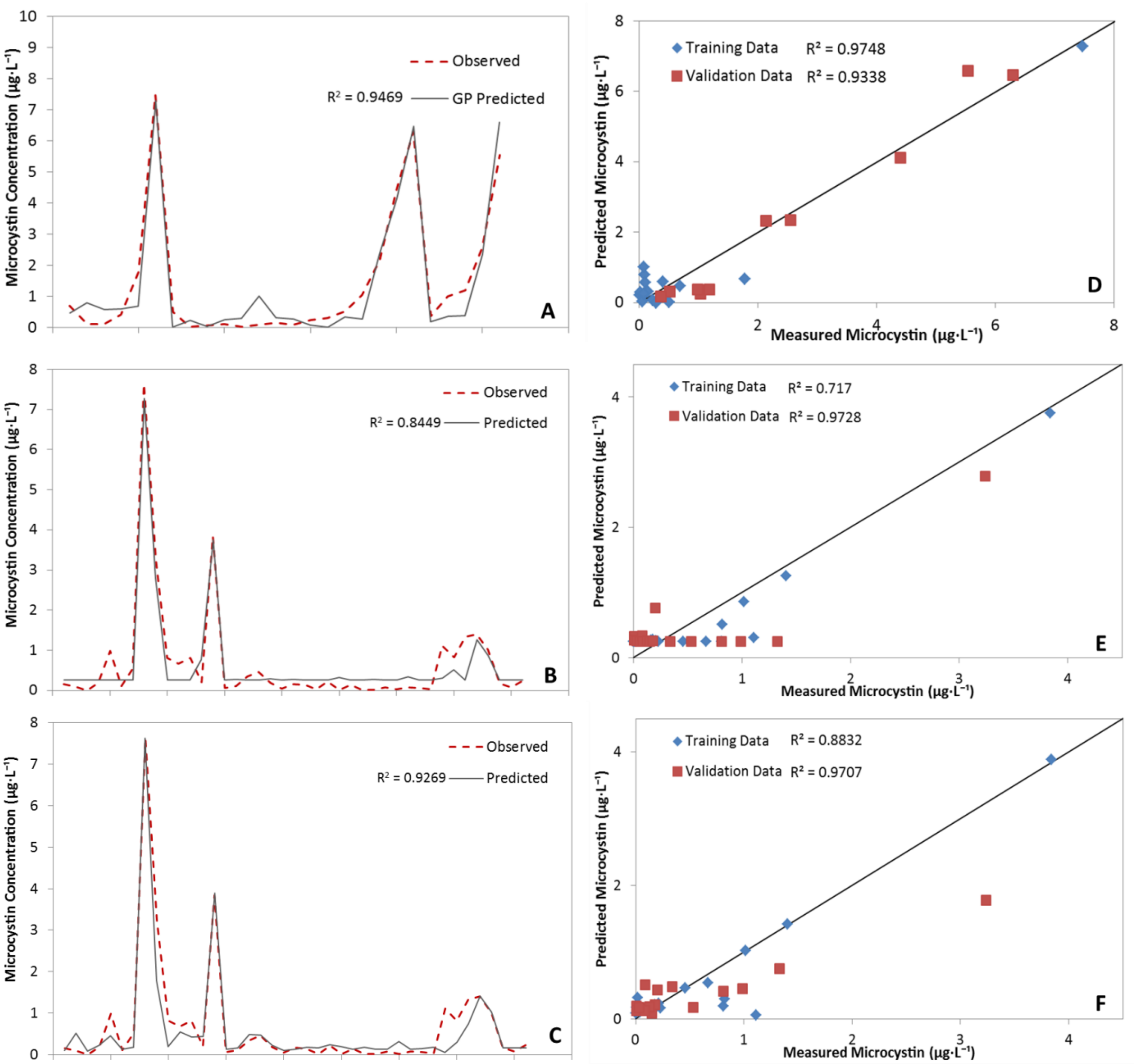


Figure 3.4: Time series plots in the left column exhibit the predictive performance of a pure MERIS GP model (A), a fused multispectral GP model (Landsat and MODIS land bands) (B), and a fused hyperspectral GP model (MERIS and MODIS ocean color bands) (C). As can be seen in images A, B, and C, the models aptly predict peak microcystin values; however, the ability to predict low microcystin values varies between the models. In A, it can be seen that the predicted values at low concentrations show mediocre correlation with the observed values. From image B, the model has a horizontal line for predicting observed values below 0.3 µg·L⁻¹. The hyperspectral GP model (C) having the best success at estimating the microcystin at low concentrations. For images in the right column (D, E, and F), the predicted microcystin values have been plotted against the observed values to accentuate any biases that are present in the predicted values.

The MERIS GP model (A & D) served as a reference to compare the fused GP models (the two GP models derived using the fused spectral data as inputs) to. Only 26 data points were available to generate the MERIS model, compared to 41 data points for the fused models; yet the pure MERIS model actually had the best performance with a coefficient of determination equal to 0.9469 and admirable capacity at predicting peak values. More data points were available for the fused models, due to the additional information provided by fusion with MODIS. The MERIS GP model obviously requires less computational power and data storage requirements to develop, since data from only one satellite sensor is necessary. But, before the fused GP models are characterized as underperforming, their inherent advantages should be discussed. The revisit time of MERIS is up to 3 days in length, which leaves sizeable data gaps. The fused GP models are more reliable and provide a better early warning system because they are able to provide medium to high resolution data for generating microcystin concentration maps on a daily basis. It should be noted that the MODIS ocean color bands are capable of providing similar spectral data on a daily basis; however, the spatial resolution is over 3 times coarser (1000 m) than the fused hyperspectral product (300 m). In summary, the drawback of additional computing power and storage capacity required to formulate the fused hyperspectral GP models is outweighed by the benefits of daily, 300 m concentration maps with comparable performance.

In analyzing the multispectral (B & E) and hyperspectral (C & F) performance, the GP model created from the fused hyperspectral data yields a better coefficient of determination of 0.9269 compared to 0.8449. Both of the models are capable of predicting peak microcystin values, which is a necessary function for delineating between a harmful algal bloom laced with microcystin and an algal bloom comprised of nontoxic algal species. The predictive capabilities of the fused hyperspectral GP model excelled at predicting microcystin concentrations less than 1 $\mu g \cdot L^{-1}$, while the multispectral model simply flatlines in this region as detailed in Figure 3.4B.

The difference in predictive power in this region is also identified when comparing images E and F. As seen by the horizontal set of data points in image E, the fused multispectral GP model consistently underestimates low microcystin values. The fused hyperspectral GP model shown by image F has the advantage of more accurately predicting low microcystin values, which allows for the identification of HAB formation at an early stage. As a result, these areas can be more closely monitored for continued HAB formation. This is also useful for assessing water quality in environmentally sensitive areas. The near real-time early warning system with more accurate microcystin prediction at all concentrations is paramount, and the GP model formulated from fused hyperspectral data successfully achieved this.

3.4.2. Model Predictability

To compare predictability between the GP models and traditional inversion methods, a two band ratio model and a spectral slope model were used (Vincent *et al.* 2004, Dash *et al.* 2011). The ideal bands for the two band models were found by testing all possible band combinations and choosing which yielded the highest coefficient of correlation and fitness. For the GP models, all of the bands were supplied as inputs to Discipulus, and the program determined the bands that shared a relationship with the microcystin concentration. Comparing the two band models along with the GP models was done using 4 statistical indices: the root mean square error (RMSE), ratio of standard deviations (CO), mean percent error (PE), and the square of the Pearson product moment correlation coefficient (RSQ = R^2). The results are presented in Table 3.4 with special attention to the computational time required to solve the models:

Table 3.4: GP and two-band models using multispectral and hyperspectral surface reflectance input data are evaluated using 4 indices of accuracy. Bolded values indicate the two models exhibiting the highest performance in the assigned statistical category. The computational time is the amount of seconds required to generate the model. As expected, machine learning methods took longer to solve than regression techniques. The fused hyperspectral input provided the most accurate results overall.

	Fused Multispectral Input*			**Fused Hyperspectral Input****		
	Two-Band Ratio Model	**Spectral Slope Model**	**GP Model**	**Two-Band Ratio Model**	**Spectral Slope Model**	**GP Model**
Observed Microcystin Mean (μg·L^{-1})	0.6718	0.6718	0.6718	0.6718	0.6718	0.6718
Predicted Microcystin Mean (μg·L^{-1})	2.226	0.1792	0.6360	1.008	0.3571	0.5936
Root Mean Square Error (μg·L^{-1})	1.348	1.340	**0.3451**	1.356	0.7583	**0.3530**
Ratio of St. Dev.	0.8270	0.1238	**0.6787**	0.5540	0.5589	**0.6837**
Mean Percent Error (%)	87.57	**5.251**	38.07	61.87	**2.177**	25.01
Square of the Pearson Product Moment Correlation Coefficient	0.02393	0.09625	**0.8449**	0.2710	0.7062	**0.9269**
Computational Time (Seconds)	< 1	< 1	92	< 1	< 1	450

*Fused Multispectral Input Pair: Landsat and MODIS land bands

**Fused Hyperspectral Input Pair: MERIS and MODIS ocean color bands

The observed microcystin mean values are the same for each of the models, since they share the same set of ground-truth data. The next point of detail is that the traditional two-band models performed worse than the spectral slope and GP models. The spectral slope models performed poorly when using multispectral input values (R^2 = 0.09625), but this model performed significantly better for the hyperspectral surface reflectance inputs (R^2 = 0.7062). This is likely due to the quality of the hyperspectral data. As previously mentioned, the multispectral data covers a wide portion of the electromagnetic for each band. This often leads to spectral peaks and troughs becoming averaged with nearby data; thus, losing its shape and detail. Hyperspectral data is better suited for the spectral slope model for microcystin prediction, since the defining features in the spectral reflectance curves for chlorophyll-a and phycocyanin are captured. As a result, the clearly delineated peak or trough values produce a lucid response when analyzed with the spectral slope model. The next comparison is focused on the GP models. The hyperspectral GP model underestimates the mean observed microcystin value by 0.0782 μg·L⁻¹, while the multispectral GP model is much closer to the mean with an average underestimation of 0.0338 μg·L⁻¹. Recall from Figure 3.4B that the multispectral model often overpredicted the microcystin values at low concentrations, yet it underpredicts values close to 1 μg·L⁻¹. The hyperspectral model matched the observed microcystin values more closely, but this model underestimated concentrations more often than it overestimated concentrations. The GP model using hyperspectral data yielded an RMSE value of 0.3530 μg·L⁻¹, which is slightly worse than the 0.3451 μg·L⁻¹ obtained by the multispectral GP model. Both of these minor shortcomings for the hyperspectral GP model are compensated for by the improved performance with regard to the CO, PE, and R^2 values. The multispectral and hyperspectral GP models had CO values of 0.6787 and 0.6837, which are close to the ideal value of 1. The hyperspectral GP model yielded a PE of 25%, while the multispectral GP model had 13.06% higher error. With regards to the R^2 values, both models exhibited strong statistical significance and a positive correlation with values of 0.8449 for the multispectral GP model and 0.9269 for the hyperspectral GP model. In conclusion, the GP models are better suited for determining the complex, nonlinear relationship between microcystin and surface reflectance, and hyperspectral surface reflectance inputs yielded more accurate results than multispectral surface reflectance inputs.

Analysis of the computational time required to derive the models provides interesting theoretical insights. The drawback for using machine learning techniques is the time required to retrieve and formulate the nonlinear models. For the purpose of comparison, the traditional models were also solved using regression techniques. Computational time required for solving the multispectral and hyperspectral GP models are 92 and 450 seconds respectively. The hyperspectral spectral slope model actually yielded reasonable predictions for microcystin. Machine learning techniques typically take longer to solve when provided with a significant amount of ground-truth data. This would be a slight advantage for the hyperspectral spectral slope model, since it could serve as a way to quickly assess water quality in an area, while the primary GP model is trained. Of course, the GP model only needs to be trained once (with periodic updates in the future) for determining the relationship between a water body's unique surface reflectance characteristics and the microcystin levels.

Additional insights can be gleaned by analyzing the spectral bands that were used to create each of the models. The bands used to train the two-band models are provided in Table 3.5:

Table 3.5: Spectral band centers with the highest performance for the traditional two-band models.

Model Type	Band Centers (nm)	R^2
Multispectral Two-Band Ratio	570 & 840	0.02393
Multispectral Spectral Slope	570 & 660	0.09625
Hyperspectral Two-Band Ratio	560 & 681	0.2710
Hyperspectral Spectral Slope	665 & 681	0.7062

The band centers most used by the traditional two-band models fall in the range of 560-570 nm and 660-681 nm. This corresponds with the spectral features observed in Figure 3.1. Chlorophyll produces a distinct reflectance dip or trough in the range 660-681 nm, and it is clear why this wavelength would exhibit a strong relationship with microcystin prediction. Next, the GP models are analyzed for the frequency of use for each of the bands. The frequency of use is how often a specific band was used in the 30 best programs created in Discipulus. If a band was used in every program, it would have 100% frequency of use, and it would likely share a high correlation between surface reflectance and microcystin. The frequency of use for the variables identified in the GP models is presented in Table 3.6:

Table 3.6: Frequency of use for the band centers used as spectral inputs for the multispectral and hyperspectral GP models. The top 3 bands for each sensor type have been bolded.

Fused Multispectral Input		Fused Hyperspectral Input	
Band Center (nm)	**Frequency of Use (%)**	**Band Center (nm)**	**Frequency of Use (%)**
477	67	**412.5**	**83**
562.5	53	**443**	**80**
652.5	**97**	489	57
849.5	57	555.5	27
1645	**80**	666	7
2010	**70**	**689.5**	**100**

The fused multispectral and hyperspectral do not share many of the same band centers, so a direct comparison cannot be made between the two sensor types, since they observed different portions of the electromagnetic spectrum. Nevertheless, an independent analysis of the frequency of use can be offered for each sensor type. The fused multispectral GP model favored band center at 652.5, 1645, and 2010 nm. This corresponds to bands 3, 5, and 7 of Landsat and the MOIDS land color bands 1, 6, and 7. Comparison to Figure 3.1 shows that the wide band center at 652.5 nm averages spectral features unique to phycocyanin and chlorophyll-a (both indicators of microcystin). What is interesting is the strong emphasis placed on the shortwave infrared bands for predicting microcystin. The fused hyperspectral GP model frequency used bands centered at 412.5, 443, and 689.5 nm. Chlorophyll-a and phycocyanin both have low reflectance at the first two band centers, which is a possible explanation for delineating between these parameters and other water quality parameters in this range. The band center at 689.5 nm was also used in the traditional two-band models, as it directly corresponds with a strong reflectance trough caused by chlorophyll-a in the water. While phycocyanin and chlorophyll-a serve as strong microcystin indicators, variations in optical complexity, such as heavy suspended solid levels commonly induced by storm events in the Maumee Bay region, may be necessary to identify more abstract indicators, such as HAB growth rate (tied to microcystin production (Wynne *et al.* 2008)) and weather patterns, which may limit light levels.

3.4.3. Microcystin Maps

Using the GP model derived from the fused band data, maps of the microcystin concentration throughout Lake Erie can be reconstructed to allow for the assessment of blooms during the summer. As a result, detailed information on *Microcystis* bloom proliferation and transportation can be identified, and subsequently used to identify probable problem spots that require close monitoring during the summer. To illustrate this, microcystin map generated from a fused hyperspectral GP model and a fused multispectral GP model are shown in Figure 3.5, and they are compared to a false color image of the algal bloom occurring on the same day:

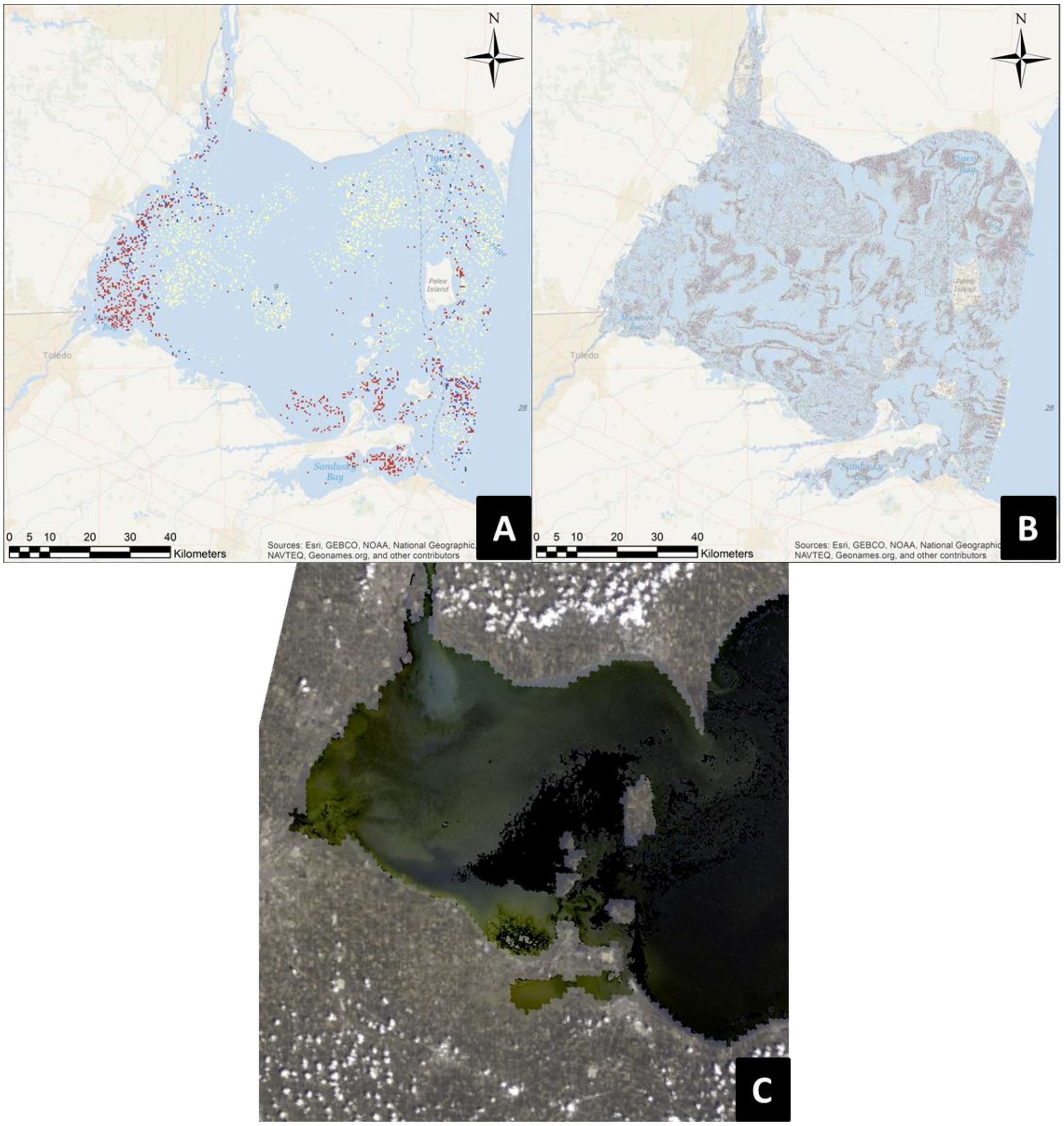

N
Pelee Island
Toledo
Sandusky Bay
0 5 10 20 30 40
Kilometers
Sources: Esri, GEBCO, NOAA, National Geographic, NAVTEQ, Geonames.org, and other contributors
A
N
0 5 10 20 30 40
Kilometers
Sources: Esri, GEBCO, NOAA, National Geographic, NAVTEQ, Geonames.org, and other contributors
B
C

Figure 3.5: The concentration maps were generated using the hyperspectral GP model (A) and multispectral GP model (B). The false color image of western Lake Erie is presented on the bottom (C). Large algal blooms spawning out of the Maumee and Sandusky Bays on July 26, 2010 are seen as dark green, while the sediment is a pale white in (C). Dark red spots in (A) and (B) denote areas of high microcystin concentration that pose a health threat, while yellow spots indicate low to medium concentrations. The 30 m spatial resolution of the multispectral image provides more detailed outlines, while the coarser (300 m) hyperspectral resolution predicts microcystin concentrations in locations that more closely align with HAB presence.

The concentration map from the fused hyperspectral data (A) is much less detailed, due to the 300 m resolution. The apparent advantage is that the predicted medium and high concentrations of microcystin align with the green HABs observed in the false color image (C). The less accurate concentration map derived from multispectral data (B) appears to exaggerate microcystin concentrations throughout the lake. However, the enhanced detail provided by the 30 m resolution provides insight into the benefits that a hyperspectral satellite with fine spatial resolution would yield. This is evidenced by the apparent currents and potential bloom delineations seen in B.

3.5. Conclusion

STARFM was able to accurately fuse the both hyperspectral (MERIS and MODIS ocean color bands) and multispectral (Landsat and MODIS land bands) image pairs to generate synthetic images possessing both moderate spatial and temporal resolution. The synthetic images contain more data than a single image from either satellite, and the fusion method is used to fill in data gaps from the lengthy revisit times of MERIS and Landsat. In comparing traditional two-band models to more complex GP models, it was observed that the GP models required longer training times, yet they offered higher explanatory power in relating microcystin to surface reflectance. Next, it was shown that the fused hyperspectral GP model excelled over the fused multispectral GP model for microcystin prediction. This was quantified using 4 statistical indices. The fused multispectral GP model yielded more desirable mean prediction errors and RMSE values of 0.0358 $\mu g \cdot L^{-1}$ and 0.3451 $\mu g \cdot L^{-1}$, compared to 0.0782 $\mu g \cdot L^{-1}$ and 0.3530 $\mu g \cdot L^{-1}$. The fused hyperspectral GP model ranked the highest when evaluated with the CO, PE, and R^2 statistical indices, achieving values of 0.6837, 25.01 %, and 0.9269, compared to 0.6787, 38.07 %, and 0.8449. While the fused hyperspectral GP model required the longest training time of 450 s, it had the highest explanatory power microcystin prediction and the fulfillment of an early warning system.

One limiting factor to the ground-truth data is that the majority of the samples correspond to fixed points that were sampled when HABs on the lake were observed. Ideally, sampling would have been carried out on a daily basis starting from when the HAB formed and stopping after it dissipated. This would provide a representative idea on when microcystin began to form within the HAB, and daily tracking of the HAB with corresponding microcystin samples could corroborate the success of such a map, since a time series of maps would lucidly depict HAB mobility. The second limitation is that many of the ground-truth points were below 1 $\mu g \cdot L^{-1}$. Even though the GP models successfully predicted peak microcystin concentrations, a larger and more diverse data set would improve the predictability of the models at low and peak values. With the recent failure of the MERIS sensor, this work would be further explored following the upcoming launches of the Sentinel multi-satellite project.

3.6. Acknowledgement

This work is a part of US EPA Water Resources Adaptation Program (WRAP) funded through EPA contract EP-C-11-006. The manuscript has been subjected to the administrative review by EPA and has been approved for external publication. Any opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the Agency; therefore, no official endorsement should be inferred. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

3.7. References

Belokon, W., Emmons, M., Fowler, W., Gilson, B., Hernandez, G., Jonson, A., Keister, M., McMillan, J., Noderer, M., Tullos, E., and White, K., (1997), *Multispectral Imagery Reference Guide*. Gogicon Geodynamics, Inc., Fairfax Virginia.

Bridgeman, T., (2005), University of Toledo Lake Erie Center – Water Quality Monitoring in Western Lake Erie and Maumee Bay.

Budd, J., Beeton, A., Stumpf, R., Culver, D., and Kerfoot, W, (2001), "Satellite observations of Microcystis blooms in western Lake Erie," *Verh. Internat. Verein. Limnol.*, vol. 27, pp. 3787-3793.

Chang, G., Mahoney, K., Briggs-whitemire, B., Kohler, D., Mobley, C., Lewies, M., Moline, M., Boss, E., Kim, M., Philpot, W., and Dickey, T., (2004), *The New Age of Hyperspectral Oceanography*, Oceanography.

Dash, P., Walker, N., Mishra, D., Hu, C., Pinckney, J., and D'Sa, E, (2011), "Estimation of cyanobacterial pigments in a freshwater lake using OCM satellite data," Remote Sensing of the Environment, vol. 115, pp. 3409—3423.

Del Bianco, A., Serafino, G., and Spock, G., (n.d.) *An Introduction to Spectral Imaging*. Carinthian Tech Research GmbH, Europastrasse 4 A-9524 Village St. Magdalen, Austria.

European Space Agency (ESA), (2006), *MERIS Product Handbook Issue 2.1*.

Francone, D., (1998), *Discipulus Software Owner's Manual, version 3.0 DRAFT*, Machine Learning Technologies, Inc., Colorado.

Ganf, G., Oliver, R., and Walsby, A., (1989), “Optical properties of gas-vacuolate cells and colonies of *Microcystis* in relation to light attenuation in a turbid, stratified reservoir (Mount Bold Reservoir, South Australia),” *Australian Journal of Freshwater Research*, vol. 40, pp. 595-611.

Gao, F., Masek, J., Schwaller, M., and Hall, F., (2006), “On the Blending of Landsat and MODIS Surface Reflectance: Predicting Daily Landsat Surface Reflectance,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 8, pp. 2207—2218.

Genderen, J.and Pohl, C., “Image Fusion: Issues, Techniques, and Applications. Intelligent Image Fusion,” in 1994 *Proc. EARSel Workshop,* Strasbourg, France, 1994, pp. 18-26.

Ha, J., Hidaki, T., and Tsuno, H., (2009), “Analysis of Factors Affecting the Ratio of Microcystin to Chlorophyll-a in Cyanobacterial Blooms Using Real-Time Polymerase Chain Reaction,” *Environmental Toxicology,* DOI 10.1002, pp. 21-28.

Hitzfield, B., Hoger, S., and Dietrich, D., (2000), “Cyanobacterial toxins: Removal during drinking water treatment and human risk assessment,” *Environmental Health Perspectives*, 108, pp. 113-122.

Hu, C., Carder, K., and Muller-Karger, F., (2000), “Atomospheric Correction of SeaWIFS Imagery over turbid Coastal Waters: A Practical Method,” *Remote Sensing of the Environment*, vol. 74, iss. 2, pp. 195—206.

Jones, G. and Orr, P., (1994), “Release and degradation of microcystin following algaecide treatment of a microcystis aeruginosa bloom in a recreational lake, as 248 determined by HPLC and protein phosphate inhibition assay,” *Water Resources*, vol. 28, pp. 871-876.

Lee, Z. and Carder, K., (2002), "Effect of Spectral band numbers on the retrieval of water column and bottom peroperties from ocean color data," *Applied Optics,* vol. 41, pp. 2191—2201.

Lekki, J., Anderson, R., Nguyen, Q., and Demers, J., (2009), "Development of Hyperspectral Remote Sensing Capability for Early Detection and Monitoring of Harmful Algal Blooms (HABs) in the Great Lakes," AIAA Aerospace Conference, Seattle, Washington, April 6-9, 2009.

Lubac, B., Loisel, H., Guiselin, N., Astoreca, R., Artigas, L., and Meriax, X, (2008), "Hyperspectral and multispectral ocean color inversions to detect Phaeocystis globosa blooms in coastal waters," Journal of Geophysical Research, vol. 113.

Masek, J., Vermote, E., Saleous, N., Wolfe, R., Hall, F., Huemmrich, F., Gao, F., Kutler, J., and Lim, T., (2005), "A Landsat surface reflectance data set for North America, 1990-2000," *IEEE Geoscience and Remote Sensing Letter,* vol. 3, no. 1, pp. 69-72.

Mole, J., Chow, C., Drikas, M., Burch, M., (1997), "The influence of cultural media on growth and toxin production of the cyanobacterium *Microcystis aeruginosa* Kutz Emend Elenkin," Paper presented at the 13th annual conference of the Australian Society for Psychology and Aquatic Botany, Hobart, January 1997.

O'Reilly, J., Maritorena, S., Mitchell, B., Siegel, D., carder, D., Garver, s., kahru, M., and McClain, C., (1998), "Ocean color chlorophyll algorithms for SeaWiFS, *Journal of Geophysical Research,* vol. 103, no. 24, pp. 937—953.

Ouelette, A., Handy, and Wilhelm, S., (2006), “Toxic Microcystis is widespread in Lake Erie: PCR detection of toxin genes and molecular characterization of associated cyanobacterial communities,” *Microbiology Ecology*, vol. 51, pp. 154-165.

Pabich, P., (2002), *Hyperspectral Imagery: Warfighting Through a Different Set of Eyes*. Air University, Maxwell Air Force Base, Alabama.

Pohl, C. and Genderen, J., (1998), “Multisensor Image Fusion in Remote Sensing: Concepts, Methods, and Applications,” *International Journal of Remote Sensing*, vol. 19, no. 5, pp. 823—854.

Rinta-Kanto, J., Konopko, E., DeBruyn, J., Bourbonniere, R., Boyer, G., and Wilhelm, S., (2009), “Lake Erie Microcystis: Relationship between microcystin production, dynamics of genotypes and environmental parameters in a large lake,” *Harmful Algae*, vol. 8, pp. 665-673.

Rogalus, M. and Watzin M., (2008), “Evaluation of Sampling and Screening Techniques for Tiered Monitoring of Toxic Cyanobacteria in Lakes,” *Harmful Algae,* vol. 7, pp. 504-514.

Shippert, P., (n.d.), *Introduction to Hyperspectral Image Analysis*. Research Systems, Inc.

Toivola, D., Eriksson, J., and Brautigan, D., (1994), “Identification of protein phosphate 2A as the primary target for microcystin-LR in rat liver homogenates,” *FEBS Letters*, vol. 344, pp. 175-180.

Torrecilla, E., Piera, J., and Vilaseca, M., (2009), *Derivative analystis of hyperspectral oceanographic data, Advances of Geoscience and Remote sensing,* Gary Jedlovec (Ed.), ISBN: 978-953-307-005-6, InTech.

United States Geological Survey (USGS). Landsat Processing Details [Online]. Available: http://landsat.usgs.gov/Landsat_Processing_Details.php

Vermote, E., Kotchenova, S., and Ray, J., (2011), *MODIS Surface Reflectance User's Guide. v1.3*, MODIS Landsat Surface Reflectance Science Computing Facility.

Vermote, E., Saleous, N., and Justice, C., (2002), "Atmospheric correction of MODIS data in the visible to middle infrared: First results," *Remote Sensing of the Environment,* vol. 83, no. 1, pp. 97-111.

Vincent, R., Xiaoming, Q., McKay, R., Miner, J., Czajkowski, K., Savino, J., and Bridgeman, T., (2004), "Phycocyanin detection from Landsat TM data for mapping cyanobacterial blooms in Lake Erie," *Remote Sensing of the Environment*, vol. 89, pp. 381-392.

World Health Organization (WHO), 1999, Toxic Cyanobacteria in Water: A guide to their public health consequences, monitoring and management, E & FN Spon, London, England.

Wynne, T., Stumpf, R., Tomlinson, M., Warner, R., Tester, P., Dyble, J., and Fahnenstiel G., (2008), "Relating spectral shape to cyanobacterial blooms in the Laurentian Great Lakes," *International Journal of Remote Sensing*, vol. 29, no. 12, pp. 366-3672.

CHAPTER 4: GENERAL CONCLUSIONS AND RECOMMENDATIONS

4.1. Conclusions

Results demonstrated that the use of IDFM was suitable for an early warning system for monitoring TOC and microcystin concentrations in both small and large lakes. STAR-FM as a data fusion technique functions well for the pixel level fusion between Landsat and MODIS land bands, as well as MERIS and MODIS ocean color bands. A correlation analysis from Figure 2.7, showed that there was a moderate to strong correlation (R^2 = 0.5330) between the observed and predicted synthetic image, when using a pre- and post-condition image for STAR-FM. When using only a pre-condition image the R^2 value dropped to 0.4147, and when using a post-condition image the R^2 value was 0.7482. This indicates the degree of variability involved when forecasting TOC and microcystin values with the early warning system.

In both studies, the GP models outperformed the traditional two-band ratio and two-band slope inversion models at predicting TOC and microcystin. In the study for TOC prediction, the traditional two-band model yielded an R^2 value of 0.1974, and the fusion-based GP model had an R^2 value of 0.7680. For microcystin prediction, the two-band and GP models had R^2 values of 0.2710 and 0.9269. Two-band analytical models are generally applied for case 1 waters, and they fail to yield accurate predictions for water quality constituents in case 2 waters. The explanatory power of the GP models excels in identifying the complex relationship between surface reflectance and water quality parameters in case 2 waters. This is due to the problem solving approach used by a GP. Given the time and computational power, a GP model can be made to explain any relationship.

The last comparison of interest is between the use of multispectral and hyperspectral band data as inputs to IDFM. It is well known that hyperspectral inputs will outperform multispectral inputs, but there is a cost for using hyperspectral data. First, multispectral data is more widely available from repositories, while hyperspectral data may come at a cost or require a proposal to be submitted with the agency the satellite belongs to. Secondly, hyperspectral data has more bands than multispectral data, which increases handling and storage requirements. Lastly, the number of hyperspectral sensors is quite limited, and a sensor malfunction or downtime would render the early warning system unusable. Thus, it is important to quantify the advantage gained by using hyperspectral data over multispectral data. From Table 3.4, the multispectral and hyperspectral two-band ratio models yielded R^2 values of 0.02393 and 0.2710. While the hyperspectral model constitutes an obvious improvement, both exhibit weak correlations. Using these inputs for the spectral slope model confirms the sheer advantage of applying hyperspectral data for this analytical model. The R^2 value for the multispectral spectral slope model was 0.09625, while the R^2 value for the hyperspectral input was 0.7062. The detailed band data from the hyperspectral data set highlighted telltale spectral features that could be used to identify chlorophyll-a, a primary indicator of microcystin. The differences in performance for the multispectral and hyperspectral GP models were less pronounced with R^2 values of 0.8449 and 0.9269. The computational time for training the multispectral GP model was 92 seconds, while the hyperspectral GP model required an extra 358 seconds. When monitoring water quality parameters that can have significant impacts on the environment and human health, it is worth the additional training time and processing power for a noticeable gain in estimation accuracy.

4.2. Recommendations

A functional early warning system is a vital asset for decision makers, if and only if it is functional under all weather and complex water quality conditions. The current application of IDFM in the above case studies suffered from two primary limitations: 1.) it requires cloud-free surface reflectance data acquired during the day and 2.) during training and calibration, the model must be exposed to a thorough ground-truth data set depicting all possible water quality conditions. The suite of satellites for obtaining surface reflectance data utilized passive sensors that were sensitive to the visible, near infrared, and infrared portions of the electromagnetic spectrum. Passive sensors operate by detecting solar radiation from the sun that has reflected off the target object. As a result, the sensor can only function during daylight hours. To further enhance the early warning system, the use of active sensors should be explored. Active sensors emit their own electromagnetic waves and observe the signal that is reflected back; this enables surface reflectance readings to be obtained at night. Synthetic aperture radar (SAR) is a key example of this. Microwaves from a SAR sensor have a secondary and potentially more important advantage. Microwaves are capable of piercing through cloud cover. If a sensor is unable to observe the land or water hidden below cloud-cover, then the early warning system is rendered inoperable for that day. This is a significant problem for satellites like Landsat, which have lengthy revisit times. If the site is covered with clouds during the 2 times a month that Landsat passes over the area, then an entire month of high resolution spatial data is unavailable. And, it is crucial for STARFM to have recent high resolution spatial data to generate accurate synthetic images each day. Integrating SAR or other sensors capable of piercing through cloud cover is a needed development to enrich IDFM.

The second limitation faced by empirically derived models, such as GP and ANN, is that the predictions made by the models are only accurate for the range of ground-truth data used during training and validation. If a range of surface reflectance values indicative of rare peak concentrations of TOC or microcystin are fed into the model for prediction, then the model cannot be guaranteed to output that a high concentration exists at this location. Model accuracy is only assured for the range of conditions it has been exposed to. And, decision makers depend on the model's ability to outline dangerous blooms of TOC or microcystin. This issue can be remedied with time due to the suite of data mining techniques used in IDFM. The GP and ANN models can be recalibrated to take new ground-truth data into consideration. Thus, a periodic sampling routine throughout the year to gain additional ground-truth data will build a robust and reliable model over time. Periodic sampling of the surface water, also recalibrates the model to account for water quality changes in the surface water that gradually occur over long timespans.

Measuring the spatiotemporal distribution of water quality parameters is quicker and more economical when conducted by way of remote sensing instead of manual sampling. It is a challenging task to train models to function in various types of optically complex surface waters for the prediction of differing water quality constituents. However, with time and the ever-expanding lineup of satellite sensors with increased spatial, spectral, and temporal resolution these relationships can undoubtedly be deciphered. IDFM establishes a powerful yet flexible framework capable of adapting to the new sensors and inversion models applied for predicting water quality parameters in the constantly evolving conditions in surface waters throughout the globe.

APPENDIX A: CHAPTER 2 GENETIC PROGRAMMING SOLUTION

The resulting GP algorithm for TOC prediction using the fused band data is given in the columns below. TOC is predicted in units of $mg \cdot L^{-1}$ by completing each of the calculations shown below starting with the first column. The variables *f0*, *f1*, and *f2* are all initially 0. The variables *v0*, *v1*, *v2*, *v3*, *v4*, and *v5* correspond to the surface reflectance values given in Table 2.9.

```
f0 = f0+v1;
f0 = f0-v2;
f0 = f0-v3;
f0 = f0+f0;
f0 = sin(f0);
f0 = f0+f0;
f2 = f2+f0;
f0 = f0+v1;
f0 = f0+f0;
f0 = f0+f0;
f0 = sqrt(f0);
f0 = f0-1.45;
f0 = cos(f0);
f2 = f2+f0;
f0 = f0*f1;
f0 = cos(f0);
f0 = f0+f0;
f0 = sqrt(f0);
f0 = cos(f0);
```

```
f1 = f1+f0;
f0 = f0/f0;
f0 = f0+f0;
f0 = sqrt(f0);
f0 = f0+f0;
f1 = f1+f0;
f0 = sin(f0);
f0 = f0/v0;
f0 = f0+v1;
f0 = f0-v2;
f0 = f0-v3;
f0 = f0+f0;
f0 = sin(f0);
f0 = f0+f0;
f2 = f2+f0;
f0 = cos(f0);
f2 = f2+f0;
f0 = f0 - f0;
f0 = f0+v1;
```

```
f0 = f0-v2;
f0 = f0+v1;
f0 = f0-v2;
f0 = f0+v1;
f0 = f0+f0;
f0 = sin(f0);
f0 = f0+f0;
f2 = f2+f0;
f0 = f0+v1;
f0 = f0-f1;
f0 = f0+f0;
f0 = f0+f0;
f0 = sqrt(f0);
f0 = f0+v0;
f0 = cos(f0);
f2 = f2+f0;
f0 = f0+v1;
f0 = f0+f0;
f0 = sin(f0);
```

```
f0 = cos(f0);
f2 = f2+f0;
f0 = f0-f0;
f0 = f0+v1;
f0 = f0+f0;
f0 = sqrt(f0);
f0 = f0*v4;
f0 = abs(f0);
f0 = f0/1.25;
f0 = f0/1.25;
f0 = f0+v0;
f0 = f0*0.43;
f0 = f0-f1;
f0 = f0+f0;
f0 = f0+f0;
f0 = sqrt(f0);
f0 = f0-1.45;
f0 = cos(f0);
f2 = f2+f0;
f0 = f0+v1;
f0 = f0-v2;
f0 = f0-v3;
f0 = f0+f0;
f0 = sin(f0);
f0 = f0+f0;
f2 = f2+f0;
f0 = f0+v1;
f0 = f0-f1;
f0 = f0+f0;
f0 = f0+f0;
f0 = sqrt(f0);
f0 = f0-1.45;
f0 = cos(f0);
f2 = f2+f0;
f0 = f0/f0;
f0 = f0+f2;
f0 = f0+f0;
f0 = f0+v1;
f0 = f0+v1;
f0 = f0+f0;
f0 = sin(f0);
f0 = f0+f0;
f2 = f2+f0;
f0 = f0+v1;
f0 = f0+v1;
f0 = f0/0.92;
f0 = f0+v1;
f0 = f0+0.13;
f0 = f0-1.45;
f0 = cos(f0);
f2 = f2+f0;
f0 = f0/f0;
f0 = f0+v0;
f0 = f0*0.43;
f0 = f0-f1;
f0 = f0+f0;
f0 = f0+f0;
f0 = sqrt(f0);
f0 = f0-1.45;
f0 = cos(f0);
f2 = f2+f0;
f0 = f0+v1;
f0 = f0-v2;
f0 = f0-v3;
f0 = f0+f0;
f0 = sin(f0);
f0 = f0+f0;
f2 = f2+f0;
f0 = f0+v1;
f0 = f0+f0;
f0 = sqrt(f0);
f0 = f0+f0;
```

```
f0 = sqrt(f0);
f0 = f0-1.45;
f0 = cos(f0);
f2 = f2+f0;
f0 = f0+v1;
f0 = sqrt(f0);
f0 = f0+f2;
f0 = f0+f0;
f0 = sin(f0);
f2 = f2-f0;
f0 = cos(f0);
f2 = f2+f0;
f0 = f0-f0;
f0 = f0+v0;
f0 = f0+v1;
f0 = f0-v2;
f0 = f0+v1;
f0 = f0+f0;
f0 = sin(f0);
f0 = f0+f0;
f2 = f2+f0;
f0 = f0+v1;
f0 = f0+f0;
f0 = f0+f0;
```

```
f0 = sqrt(f0);
f0 = f0-1.45;
f0 = cos(f0);
f2 = f2+f0;
f0 = f0+f0;
f2 = f2*f0;
f0 = f0+0.139;
f0 = f0+v1;
f0 = f0+v1;
f0 = f0+f0;
f0 = sin(f0);
f0 = abs(f0);
f2 = f2+f0;
f0 = f0-f0;
f0 = f0+v1;
f0 = f0-v2;
f0 = f0+v1;
f0 = f0-v2;
f0 = f0+v1;
f0 = f0+f0;
f0 = sin(f0);
f0 = f0+f0;
f2 = f2+f0;
f0 = cos(f0);
```

```
f0 = f0+v2;
f0 = f0+f0;
f0 = sqrt(f0);
f0 = f0-v0;
f0 = f0+f0;
f0 = sin(f0);
f0 = abs(f0);
f2 = f2+f0;
f0 = f0+v1;
f0 = f0+f0;
f0 = f0+f0;
f0 = sqrt(f0);
f0 = f0-1.45;
f0 = cos(f0);
f2 = f2+f0;
f0 = f0-f0;
f0 = f0-v2;
f0 = f0+v1;
f0 = f0+f0;
f0 = sin(f0);
f0 = f0+f0;
f2 = f2*f0;
f0 = f0+f2;
f0 = f0*f0;
```

```
f0 = f0+v1;

f0 = cos(f0);

f0 = f0+v1;

f0 = f0+v2;

f0 = sin(f0);

f0 = f0+f0;

f2 = f2+f0;

f0 = f0+f2;

f0 = f0+f2;

f0 = f0*f0;

f0 = f0+v1;

f0 = f0-v2;

f0 = f0+v1;

f0 = f0/1.25;

f0 = sqrt(f0);

f0 = f0+f0;

f0 = sqrt(f0);
```

APPENDIX B: MATERIALS UNDER REVIEW

The contents for chapters 2 and 3 have been submitted for review as follows:

- The content in chapter 2 has been submitted for publication as: Chang, N.B., Vannah, B., Yang, J., and Elovitz, M., “Integrated Data Fusion and Mining Techniques for Monitoring Total Organic Carbon Concentrations in a Lake”, International Journal of Remote Sensing, in review, June. 2013.
- The content in chapter 3 has been submitted for publication as: Chang, N.B., Vannah, B., and Yang, J., “Comparative Sensor Fusion Between Hyperspectral and Multispectral Remote Sensing Data for Monitoring Microcystin Distribution in Lake Erie,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, in review, Oct. 2013.