COMPUTATIONAL APPROACHES FOR SATELLITE DATA ANALYSIS FOR CHLOROPHYLL CONCENTRATION: CHALLENGES AND INITIAL RESULTS

By

RESHMA CHOWDARY PASUMARTHI

(Under the Direction of Lakshmish Ramaswamy)

ABSTRACT

Both environmental monitoring and the assessment of risks to the ecosystem play a significant role in maintaining environmental sustainability. Among several impacts, toxins produced by cyanobacteria in water affect aquatic plants, animals and human beings, they can grow faster with high availability of nutrients and warm temperatures. Based on public reports, the National Wildlife Federation noted in a report that cyanobacterial harmful algal blooms are common, with 21 states of the U.S reporting blooms at 147 locations between May and September 2013 [1]. In our research, we have studied the process to extract “chlorophyll a” concentration from lakes in Georgia using scenes captured by MERIS satellite; also studied the process for extracting physical parameters Land Usage Land Cover (LULC), Normalized Difference Vegetation Index (NDVI), and Palmer Drought Severity Index (PDSI) data for lakes in Georgia. We have also studied lakes, which have similar trend with respect to “chlorophyll a” concentration from 2002 to 2012, and impact of physical parameters for change in concentration by performing machine learning analysis. Our research seeks to explore the challenges and

approaches in the extraction of data from satellite scenes and to apply data analytics to environmental monitoring.

INDEX WORDS: Environmental Monitoring, Environmental Sustainability, toxins, cyanobacteria, Chlorophyll a, machine learning

COMPUTATIONAL APPROACHES FOR SATELLITE DATA ANALYSIS FOR CHLOROPHYLL CONCENTRATION: CHALLENGES AND INITIAL RESULTS

by

RESHMA CHOWDARY PASUMARTHI

B. Tech, SHRI VISHNU ENGINEERING COLLEGE FOR WOMEN, INDIA, 2010

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2015

COMPUTATIONAL APPROACHES FOR SATELLITE DATA ANALYSIS FOR CHLOROPHYLL CONCENTRATION: CHALLENGES AND INITIAL RESULTS

by

RESHMA CHOWDARY PASUMARTHI

Major Professor:	Lakshmish Ramaswamy
Committee:	Deepak Mishra
	Ismailcem Budak Arpinar

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2015

DEDICATION

I would like to dedicate this work to my husband Subhash Pasumarthi and my family for their unconditional love and support.

ACKNOWLEDGEMENTS

I would like to express deep gratitude to my major advisor Dr. Lakshmish Ramaswamy who believed in me and allowed me to be part of this project that will help society. His guidance, motivation and support has been vital throughout my research and also, in writing my Thesis.

I would like to extend my warm thanks to Dr. Deepak Mishra for being part of my Thesis committee and providing guidance and support throughout my research.

I would like to extend my warm thanks to Dr. Budak Arpinar for being part of my Thesis committee and supporting me throughout the M.S program.

Special thanks to my team members Vinay Kumar and Benjamin Page for their help, assistance and support throughout this project.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

In this chapter, we introduce cyanobacteria and their effects on the ecosystem; also field monitoring and remote sensing techniques used to detect cyanobacteria concentration. We have also briefly discussed major challenges associated with cyanobacteria data extraction, how machine learning techniques can be used for the analysis and our contributions on these lines.

1.1 Introduction

Cyanobacteria commonly known as blue-green (BG) algae are important class of phytoplankton [3]. Abundant growth of cyanobacteria in aquatic systems creates problems due to their capacity to produce toxins which are known as cyanotoxins [3]. These cyanotoxins include neurotoxic, hepatotoxic, genotoxic, inflammatory, microcystins and cytotoxic agents [5]. Among these agents microcystins are the most potent and commonly encountered [5]. Cyanobacteria are the only bacteria that contain chlorophyll-a. The blue green algae contain the pigments phycoerythrin and phycocyanin [9]. Cyanobacteria and their cyanotoxins are unregulated contaminants [2]. Mass populations of cyanobacteria are described as “blooms”; and most of the blooms are toxic species [6]. Cyanotoxins can be found in water bodies used for drinking, aqua culture, crop irrigation, and recreation [2]. Algal blooms degrade the quality of the lakes and reservoirs by forming surface scums, inducing unpleasant taste and odor in the drinking water and causes effects to human health [4].

Cyanobacteria blooms can normally be unappealing, which are concentrated along the shorelines where they are encountered frequently by the public [2]. The main reasons for growth of cyanobacteria blooms are increased nutrients level, increase in temperatures where warm surfaces favor cyanobacteria growth, and changes in land use practices such as increase in agricultural growth tends to increase in nutrients flow to water bodies [2].

Cyanobacteria toxins are also present in treated drinking water supplies when cyanobacteria blooms occur in their sources [6]. Increased population and depletion of ground water resulted in increase of usage of surface water as a water source, both from rivers/lakes and reservoirs [6]. Cyanobacterial toxins cause livestock poisoning, which has been extensively reported in America, Europe and Australia [6]. Livestock are vulnerable to cyanobacteria poisoning because they tend to drink toxic water in ponds and lakes in farms [6]. Human poisoning also occurs by toxins entering the food chain through shellfish, mussels, oysters or scallops, which are generally consumed by humans [6].

Cyanobacterial blooms have been documented across US, and numerous states have issued health advisories or closed the recreational areas due to potential risks caused by them [2]. A few states have toxins monitoring programs, while others conduct event-based responses and some provide public education focused on human and animal protection from toxin exposure [2]. While there are different techniques followed for detecting cyanobacteria blooms, the typical approach is collecting samples and testing them in laboratory and satellite remote sensing.

There are many challenges faced by different states in US in the development of field based monitoring programs because they are insufficient to provide timely warnings of cyanobacteria bloom development across large geographic areas [2]. Field monitoring program is difficult because estimating cyanobacteria blooms is time consuming and labor intensive,

involving water sample collection, laboratory analysis, and the visual identification and enumeration [2]. To overcome these challenges, satellite remote sensing has been recommended to provide more reliable information about the cyanobacterial blooms [7]. Remote sensing based techniques are used to detect and map the cyanobacterial blooms from space [8]. There are a variety of approaches and methods developed for identifying the blooms and scums, estimation of phycocyanin or detecting the presence of phycocyanin, and chlorophyll a concentration [2]. Various satellites have been used such as Landsat, MODIS and MERIS [2]. For remote sensing techniques, phycocyanin (PC) which is a characteristic photosynthetic pigment in cyanobacteria is used as proxy to detect blue green algae [25].

Algal blooms cause critically stressful conditions in aquatic ecosystems so, predicting their occurrence is very important [10]. Various modelling techniques are applied to analyze fresh water ecosystems but they are generally very complex; machine learning techniques are efficient methods to deal with complex datasets such as, the long term time series data that arise in ecology in comparison with traditional techniques [10].

1.2 Challenges in cyanobacteria data extraction and analysis

There are different techniques used for cyanobacteria detection to identify the concentration and their impact to the ecosystem. However, there are some challenges associated with remote sensing when used for detection. Some of the challenges include,

1. The satellite scenes captured using remote sensing has noise associated with them because of cloud cover.
2. The physical parameters such as, land coverage land usage (LULC), palmer drought severity index (PDSI), and normalized difference vegetation index(NDVI) which has

effect on cyanobacteria growth. The physical parameter which has more affect to the change in cyanobacteria concentration is not known for Georgia lakes and reservoirs.

3. Georgia has many lakes and reservoirs' so deploying sensors in every water body to detect cyanobacteria bloom is not practical and can be costly.

1.3 Major Contribution

In this research, we have extracted LULC which gives information related to land coverage around lakes such as urban, forest, crop coverage and others, NDVI gives information related to live green vegetation and PDSI data which is about temperature and precipitation information from different data resources for 492 lakes of Georgia. We have also extracted Normalized Difference Chlorophyll Index (NDCI) which gives information related to chlorophyll concentration of lakes for years 2002 to 2012 from scenes captured by MERIS satellite by eliminating noise induced by cloud cover [24]. Our study shows the impact of physical parameters for change in chlorophyll concentration. We also performed clustering on lakes and identified lakes which have same trend with respect to chlorophyll-a concentration for years 2010 and 2011. All lakes in a single cluster will have same trend so deploying sensor in one lake for each cluster helps analyzing other lakes in the same cluster which reduces cost. Our study states that physical parameters have impact on phytoplankton growth and also as drought conditions increase chlorophyll a concentration in lakes also increases compared to LCLU and NDVI parameters.

The rest of the thesis is organized as follows:

CHAPTER 2 describes different remote sensing techniques, different satellites for environmental monitoring and their characteristics, tools used to process the satellite images, machine learning

algorithm used for data analysis and also overview of cyanoTracker project.

CHAPTER 3 describes the related work.

CHAPTER 4 gives an overview of system architecture and process used to extract LULC, NDVI, PDSI and chlorophyll a concentration data.

CHAPTER 5 gives empirical study done to cluster lakes of Georgia for years 2010 and 2011 based on chlorophyll concentration and identifying the impact of physical parameters on chlorophyll concentration.

CHAPTER 6 gives the conclusion of the analysis we performed.

CHAPTER 2

BACKGROUND

In this chapter, we have briefly described remote sensing and multispectral remote sensing, which play an important role in capturing cyanobacteria and physical parameters data. we also described satellites whose scenes are used to extract data in our analysis. We have described different tools and file formats, which are used for processing satellite images. We also described clustering technique used for data analysis and gave an overview of cyanoTracker project, which is a research project initiated by researchers at University of Georgia.

2.1 Remote Sensing

Remote sensing refers to the activities of monitoring or observing something from far away distance or remote places. In this process sensors are not in direct contact with the objects or places it captures. Electromagnetic radiations are used as a carrier for information from the object to the capturer [11]. Using remote sensing we get image or a scene, which contains data captured for each scene. The analysis and processing of data need to be done in order to extract useful information out of satellite images [11]. To avoid noise induced by cloud cover and others while capturing, image correction should be done. Remote sensing can be done using satellites that are placed in the orbit for different purposes like improving communication, earth observation, navigation, weather and many others [11]. Passive and Active remote sensing are two different kinds of remote sensing techniques.

Passive remote sensing data is captured based on natural radiation or reflection where reflected sunlight is the common source of radiation which is used in capturing images [11]. Film photography, infrared, and radiometers are based on passive remote sensing. Active remote sensing emits energy in order to capture objects where sensors detect and measure the reflection from the target [11].

The traditional monitoring methods for ecosystem consist of collection of field samples, laboratory analysis, and manual cell counts so, these methods are time consuming, labor intensive and costly [3]. To improve the monitoring technique, remote sensing is proved to be valuable [3]. Using satellite remote sensing data over large areas are gathered quickly and economically [12].

There are different active remote sensing techniques available, among them we are more interested in multi spectral remote sensing because the MERIS satellite from which we extracted chlorophyll-a concentration data for lakes of Georgia using downloaded images developed with this technology.

2.2 Multispectral Remote Sensing

Multispectral airborne and satellite sensing have been employed for gathering data in fields of agriculture and food production, geology, oil, mineral, geography and urban to non-urban localities [12]. Multispectral remote sensing systems use parallel sensor arrays that detect radiation in small number of broad wavelength bands [12]. Most multispectral satellites captures three to ten spectral bands [12]. This technique allows for the discriminations of different types of vegetation, rocks and soils, clear and turbid water, and selected man made materials [12].

2.3 MEdium Resolution Imaging Spectrometer (MERIS)

MERIS is one of the instruments on board of the European Space Agency (ESA) satellite ENVISAT. The contributions of MERIS are measurement of photosynthetic potential by detection of phytoplankton, detection of yellow substance, and detection of suspended matter [13]. The primary mission of MERIS is to monitor the ocean color including chlorophyll concentrations for open oceans and coastal areas [13]. It also provides information related to land. There are two level of products are processed using MERIS scenes where level1 products are images resampled on a path-oriented grid, with pixel values are calibrated to match the top of atmosphere radiance. Level2 products are processed to get the geophysical measurements and level1 products are input for the level2 products [13]. MERIS observes earth in 15 spectral bands among them we used two bands, 665nm band that contain information related to Chlorophyll absorption and 708nm band that contain data related to atmosphere correction [13]. We used satellite scenes of MERIS for calculating chlorophyll concentration.

Moderate Resolution Imaging Spectrometer (MODIS) is another satellite whose captured scenes are used in our analysis. MODIS satellite view earth for every 1 to 2 days and acquires data in 36 spectral bands [14]. The data collected helps in understanding the processes occur on the land, in the oceans, and in the lower atmosphere [14]. MODIS has 13 visible and near infrared bands that could be potentially used in aquatic remote sensing [7]. We used MODIS satellite scenes for extracting NDVI data in our research.

Cyanobacteria can be detected near 630nm where there is peak in reflectance spectra of cyanobacteria. MODIS does not provide any information at this spectral region so, we cannot use MODIS for detecting cyanobacteria blooms [7]. Once the satellite images from MERIS are

downloaded we need to process them to measure chlorophyll concentration so, we used ArcGIS and Beam software's.

2.4 ArcGIS and Beam Applications

ArcGIS is a geographic information system used for working with maps and geographic information. It is used for creating maps, analyzing mapped information, compiling geographic data, sharing and discovering geographic information [15]. The file formats supported by ArcGIS and used in our research are shape files, raster files, tiff files.

Shape files are Esri (Environmental Systems Research Institute) vector storage data format for storing the location, shape and attributes of geographic features [16]. Shape file format can spatially describe features as points, line and polygons which represents water wells, rivers and lakes. The raster data type is, any type of digital image represented by reducible and enlargeable grids. It consists of rows and columns of cells, with each cell storing a single value. In raster images each pixel contains color information. Along with color information raster images can also have data related to land usage, temperature or a null value. TIFF is one of the file formats raster data is stored in [17].

BEAM is an open source toolbox and development platform for viewing, analyzing and processing of remote sensing raster data. It was originally developed to process image data from Envisat's optical instruments [18]. Envisat (Environmental Satellite) is the European Space Agency's largest civilian Earth observation satellite put into space [17]. In our research we used BEAM for processing satellite scenes captured by MERIS.

2.5 K Means Clustering

To design a cost effective process for deploying sensors into water bodies of Georgia we used kmeans clustering technique. This technique is used for clustering lakes into groups to identify the lakes which have same trend related to chlorophyll concentration.

Clustering algorithms are presented with a set of data instances that must be grouped according to some notion of similarity [19]. Among clustering formulations that are based on minimizing a formal objective function, the most widely used and studied is k-means clustering [20]. Kmeans clustering is a method commonly used to automatically partition a dataset into k groups [19]. This algorithm works by selecting k initial cluster centers and then iteratively refining them by assigning each instance to its closest cluster center, and cluster center is updated by mean of its instances [19]. The input of the algorithm in our analysis is NDCI 7 classes data of lakes captured. Euclidean distance is used for calculating distance between two instances. The optimal number of clusters k for the input data set should be measured so, we used silhouette coefficient technique which is used to study the separation between two groups and their values ranges from [-1,1]. If the coefficient is closer to -1 the neighboring clusters are dissimilar, if it 0 then they are neither similar nor dissimilar, and if it closer to1 the neighboring clusters are very similar [21]. Silhouette analysis is used to choose the optimal number of clusters for the given dataset [21].

2.6 cyanoTracker Project

As mentioned before toxins released by cyanobacteria are harmful to animals, human and other living organisms. So, in “cyanotracker” [31] project at the University of Georgia, working to implement an early warning system which monitors cyanobacteria blooms in lakes of Georgia.

This project is about implementing a multi cloud framework that integrates community observations, remote sensing measurements, and multimedia analytics for environmental monitoring. The main objectives of this project are design of multi-cloud data monitoring along with event detection strategies and motivating community members to contribute content, designing a cost effective approach for data extraction, segmentation, registration and indexing of image and video data acquired from heterogeneous sources, and designing a cost effective approach for deploying hyperspectral sensors, managing them and analyzing the data collected for cyanobacteria concentration on daily basis.

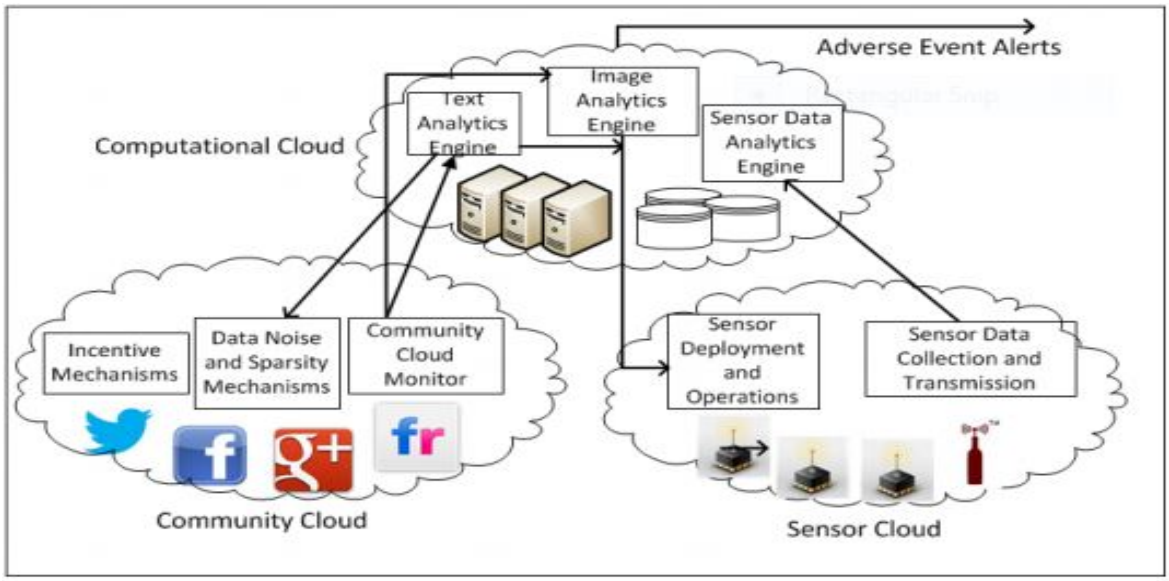


Figure1. cyanoTracker Architecture (cyanotracker.uga.edu)

Figure 1 shows the cyanoTracker architecture [31], the three main components are Text Analytics Engine, Sensor Data Analytics Engine, and Image Analytics Engine. To design a cos effective approach for deploying sensors in lakes of Georgia the data is obtained from remote sensing satellites and analyzed the data for identifying the lakes which have same behavior related to their chlorophyll concentration recorded. This analysis includes land usage, vegetation,

temperature, precipitation, and chlorophyll concentration data from remote sensors, which would be combined with data obtained from social media and other sources. We worked on extracting data from MERIS, MODIS and other data resources, processing the satellite scenes to extract required data by eliminating noise, and clustering lakes into groups based on the chlorophyll concentration recorded.

CHAPTER 3

RELATED WORK

In this chapter, we have described related work and literature survey about different techniques used for cyanobacteria detection. There are several approaches being followed for many years for detecting and analyzing cyanobacteria concentration.

3.1 Traditional Method

There are different approaches for extracting cyanobacteria concentration from water bodies. The initial detection of cyanoHABs rely on visual observation. If we find discoloration of water, fish kill and thick mat like accumulations on the shore are the primary symptoms of the cyanoHABs [2]. In this process samples are collected manually, the samples collected must consist of water, algae, sediments of the water body etc. These samples need to be stored properly in order to avoid exposure to sunlight and other potential damages. The collected samples are then tested in laboratory for toxins that includes certain steps of manual analysis because not all cyanobacteria cells have toxins, it is important to separate the cells without toxins and with toxins [22]. The field monitoring approach which consists of sample collection, laboratory analysis, and detection of phytoplankton which takes few days or weeks [4]. These methods are expensive and time consuming, and often require lengthy process to perform filtration on the samples [4]. Cyanobacteria bloom observed places are inaccessible or unnoticed due to depth of the water or water bodies are spread for long distance and,

hence collecting samples is difficult [22]. There are difficulties associated with developing appropriate sampling designs that address large areas using fewer or affordable resources and detection methods [2]

3.2 Satellite Remote Sensing

To overcome the challenges associated with traditional method new tool is required to develop efficient and effective cyanobacteria monitoring, where satellite remote sensing provides an opportunity [2]. Due to patchy nature of phytoplankton distribution it is difficult to collect samples for large water bodies so remote sensing supports capturing almost or all the surface zones of them [23]. For these techniques, phycocyanin(PC) which is a characteristic photosynthetic pigment in inland blue green algae has been used for detecting the cyanobacteria concentration [3]. Studies are shown that satellite data can detect and quantify cyanobacteria blooms in lakes [2]. This includes variety of methods and approaches including identifying scums and blooms, estimating PC, detecting the presence of PC and cell count concentration [2]. Various satellites have been used such as MERIS, MODIS, Landsat and others [2]. Satellite measurements are useful for detection of phytoplankton, which has cyanobacteria because of the unique spectral characteristics of photosynthetic pigments [3]. The absorption of phycocyanin information is at ~620nm [3]. This band can be used for analyzing and estimating the cyanobacteria concentration. There are five different algorithms proposed for estimating cyanobacteria concentration by using PC absorption feature, which are a semi-empirical baseline algorithm, a single reflectance band ratio algorithm, a nested semi-empirical band ratio algorithm, a new single reflectance band ratio algorithm, and a three band algorithm. Each algorithm has different approach in analyzing the cyanobacteria concentration which use

different reflectance bands [3]. All the above algorithms are developed for areas with latitudes higher than 35^0 but not for lower latitudes [3]. However, results are less accurate because of mixed pixel issues, geometric and radiometric noise, and inaccurate data structure introduces errors in the prediction algorithms. Hyper-spectral sensors are used to capture spatial and temporal data for phytoplankton blooms, which allows capturing cyanobacteria blooms from different aquatic systems [3]. Data from *in situ* is used to identify the relationship between reflectance and PC concentration where 620nm and 650nm bands are used for finding the relationship. A proximal hyper spectral remote sensing algorithm was developed to analyze the spectral reflectance properties of cyanobacteria with changing pigment concentration [3]. This algorithm performs better compared to other algorithms but this model deal with only specific absorption coefficient. Almost all algorithms did a decent job in extracting cyanobacteria with certain limitations, which require more research to develop effective algorithm for extracting cyanobacteria concentration.

CHAPTER 4

SYSTEM ARCHITECTURE AND OVERVIEW

In this chapter, we described research challenges faced while performing the study of cyanobacteria data extraction and analysis. We also described system architecture and data extraction procedure.

4.1 Challenges

While using satellite data for analyzing the impact of physical parameters LULC, NDVI, and PDSI on chlorophyll-a concentration of lakes, we encountered few challenges.

The first challenge is while downloading PDSI data. We were able to extract data at divisional level instead of each lake, the lakes that belong to same division will have same PDSI value for that year and month. This caused data redundancy for lakes which belongs to same division for each year.

The second challenge is related to MERIS data download. Since MERIS satellite spatial resolution is 300m, most of lakes are not captured. Among few lakes which are captured contains less than 5 pixels which are insufficient for that lake to be considered for our analysis because the accuracy of cyanobacteria bloom occurrence is very low. Due to these limitations we are able to perform analysis only on lakes that are large in size and captured by MERIS.

The third challenge is due to cloud cover on lakes, while capturing information from lakes using MERIS cloud covers introduced lot of noise in NDCI data. We extracted scenes with cloud percentage less than 10 but that did not eliminate the noise. To fix this problem we manually

selected lakes that are visible to the naked eye and then tried to extract pixels of those lakes from the scenes. The manual selection did not resolve our issue, hence we filtered pixels whose values are less than -1 from the scenes using ArcGIS and finally extracted the pixel count for the lakes. In this process we could not extract data for many lakes. Initially we started with 492 lakes but we are able to extract NDCI data for <100 lakes for all 10 years.

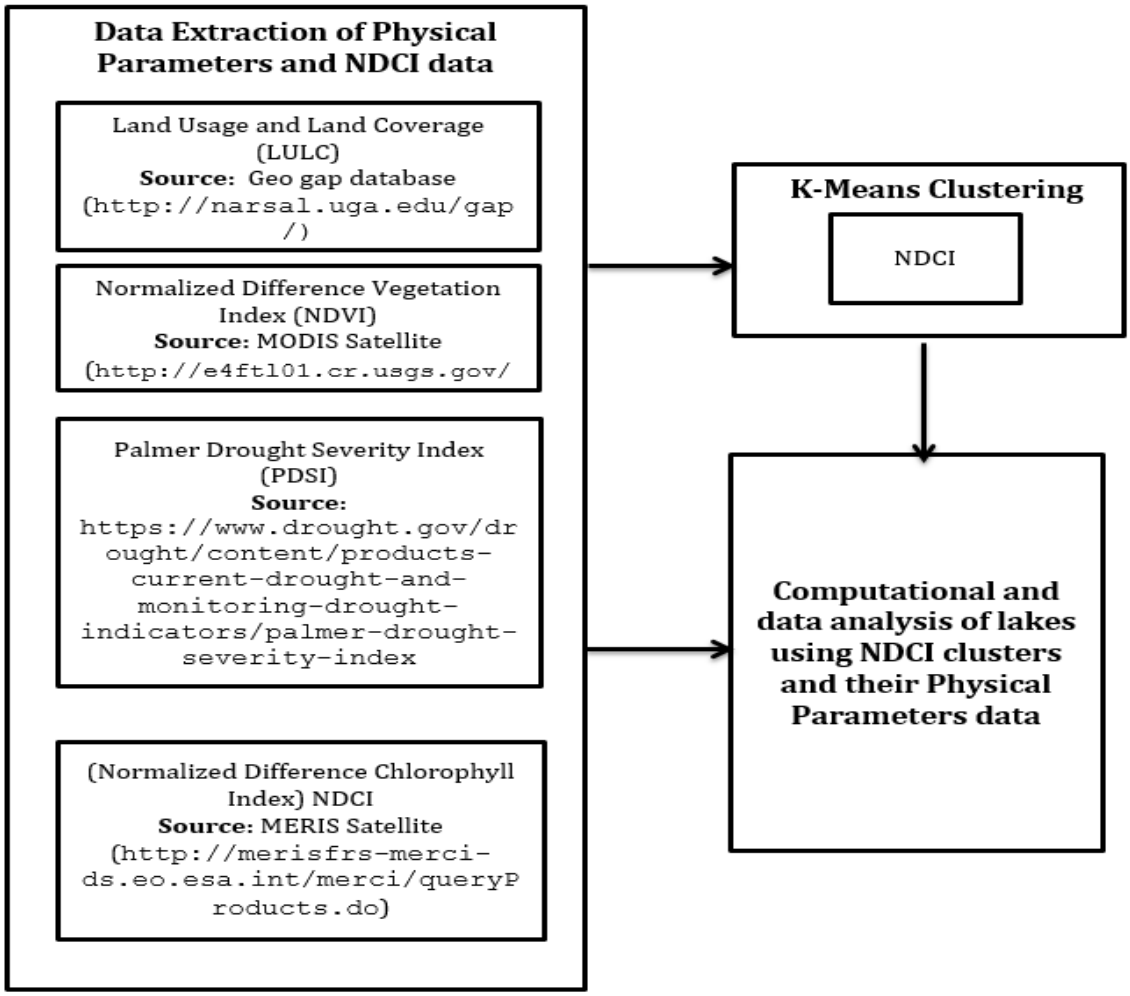


Figure 2. System Architecture and Overview

The figure2 shows the system architecture and overview of this research. It consists of three main components,

- Data Extraction
- K-Means Clustering

- Computational and Data analysis

We extracted physical parameters and NDCI data from different sources for 492 waterbodies of Georgia. There are more than 100,000 waterbodies available in Georgia, among them we selected only 492 lakes and reservoirs because NDCI data is extracted from MERIS satellite whose pixel resolution is 300m. We require waterbodies whose pixel resolution is greater than 300m, therefore we selected waterbodies whose area > 0.25sqkm and extracted data for NDCI, LULC, NDVI and PDSI for time period 2002 April to 2012 March because MERIS satellite is active in that time period.

4.2 Data extraction of Land Usage and Land Cover (LULC)

LULC provide the land coverage and land usage type data such as urban, forest, low urban, agriculture, etc. We downloaded raster files which has LULC data for Georgia from geogap database. Raster files are extracted from satellite scenes captured by LANDSAT and stored in http://narsil.uga.edu/gap database. LULC data is available for entire Georgia but we need to extract data for only 492 water bodies so, downloaded Georgia water bodies shape file from http://nhd.usgs.gov/. The downloaded shape file has 100,000 water bodies but we need only lakes and reservoirs among them so, we filtered water bodies that are of type lakes and reservoirs using ArcGIS filter technique. As mentioned earlier among lakes and reservoirs we selected 492 water bodies based on the area>0.25sqkm using ArcGIS. For each water body among 492 we created a buffer of 5 miles' radius and we extracted LULC data. Adding 5miles radius buffer for each water body is done using ArcGIS that helps in extracting the possible surround conditions for each lake. We used 492 water bodies shape file as mask and extracted LCLU data from each raster file using ArcGIS which gives lake object id which is unique id of lake, 13 land cover type

area surrounded by lakes which are represented with grid codes and their coordinates. We are able to extract data only for years 2002, 2005 and 2008, rest are not available. LCLU data extracted for water bodies are available for different days in a year so we normalized the data by calculating mean for each object id and year, by which we get one entry for each water body per year. Table1 has information about 13 land cover types and their class names based on USGS.

GRID CODE	CLASS NAME
07	Beaches/Dunes/Mud
11	Open Water
22	Low Intensity Urban
24	High Intensity Urban
31	Clearcut/Sparse
34	Rock Outcrop
41	Deciduous Forest
42	Evergreen Forest
43	Mixed Forest
81	Agricultural Land
91	Forest Wetland
92	Coastal Marsh
93	Non-forested Wetland

Table1. Land Cover Types

Figure 3 has information about steps followed to extract the LCLU data for lakes of Georgia.

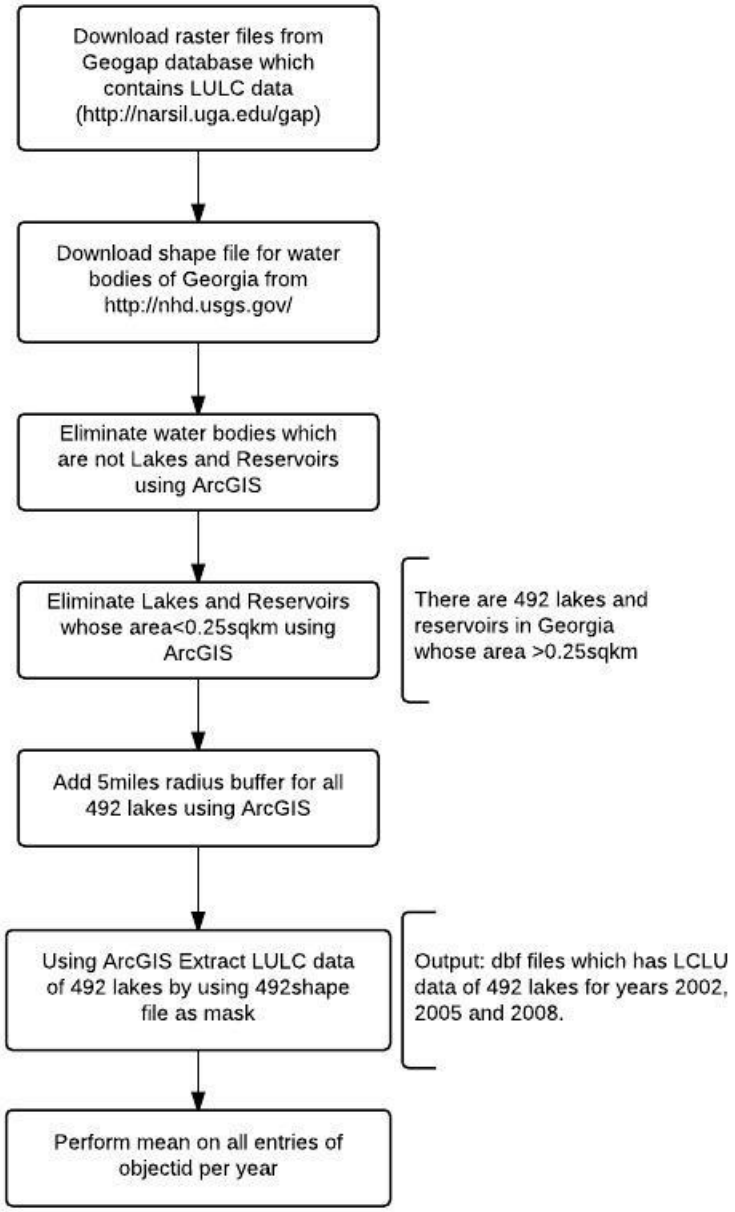


Figure 3. Flowchart of LULC data extraction procedure

4.3 Data extraction of Normalized Difference Vegetation Index (NDVI)

Using NDVI data we get information about whether the area surrounded by each water body has live green vegetation. Dense vegetation will have positive values, cloud and snowfields have negative values. To extract NDVI data we downloaded the satellite scenes which are captured by MODIS satellite from http://e4ft101.cr.usgs.gov website. To download MODIS satellite scenes, we downloaded the R script available online and made necessary changes [27]. The inputs for R script are horizontal (h) and vertical (v) tile information of Georgia and time period we want to download the scenes [28]. Using MODIS, we get different bands information like NDVI, EVI etc. We need to specify which band data we want to download where band=1 is for NDVI. We extracted NDVI for 421 lakes of Georgia because among thirteen-land cover types that are surrounded by each water body GC22, GC24 and GC81 has high impact on algae growth so we selected water bodies which are surrounded by these types. Using this feature reduction, we tried to reduce noise added to our results. The feature reduction process is implemented for 2002, 2005 and 2008 LCLU data, using this process we get three different shape files. These shape files are used as mask and extracted NDVI values using ArcGIS. We assumed there will not be huge difference in LCLU coverage for consecutive years and used the year 2002 shape file as 492 water bodies mask file for 2003 and 2004 satellite scenes, 2005 shape file as mask for 2006 and 2007, 2008 shape file mask for 2009, 2010, 2011 and 2012. From MODIS we get data for every 16 days so we simplified the data by calculating mean for each object id and year.

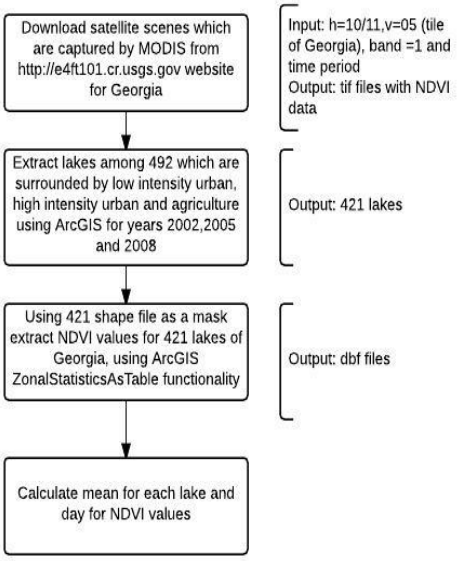


Figure 4. Flowchart of NDVI data extraction procedure

4.4 Data extraction of Palmer Drought Severity Index (PDSI)

PDSI data is measured for estimating drought conditions by using temperature and precipitation data. PDSI data is downloaded from the website which contains data for entire US https://www.drought.gov/drought/content/products-current-drought-and-monitoring-drought indicators/palmer-drought-severity-index. Based on National Climate Data Center, Georgia is divided into 9 climatic divisions where different counties of Georgia are divided into 9 divisions and PDSI data is available for each division. We extracted counties and their divisions manually from the figure 5 and then downloaded zip codes of water bodies using Google maps. Using zip codes of water bodies, we extracted counties of lakes and then mapped the divisions. The downloaded file contains different datasets along with PDSI and contains data for all states of USA, we downloaded data for state code 9 that is Georgia and element code 5 that is PDSI [29]. We downloaded data for years 2002 to 2012 and the data is available monthly. As mentioned

before algae growth is more in summer due to high temperatures so we considered only PDSI data from April to September. Since we have multiple entries for each lake, we simplified them by calculating mean on data from April to September for each year and water body. Table 2 shows different PDSI values and their corresponding drought conditions.

PDSI Values	**Drought Condition**
0 to -0.5	Normal
-0.5 to -1.0	Incipient drought
-1.0 to -2.0	Mild drought
-2.0 to -3.0	Moderate drought
-3.0 to -4.0	Severe drought
>-4.0	Extreme drought

Table 2. PDSI values and their drought condition

Figure 5. Georgia map with 9 divisions by National Climate Data Center

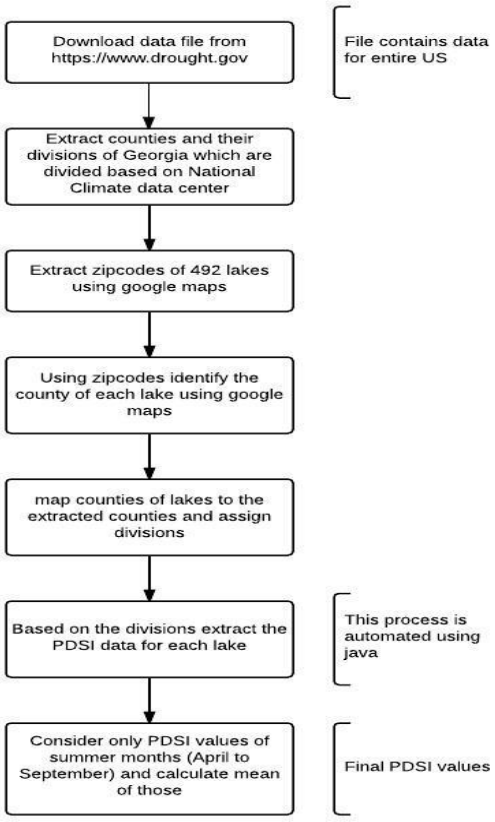


Figure 6. Flow chart of PDSI data extraction procedure

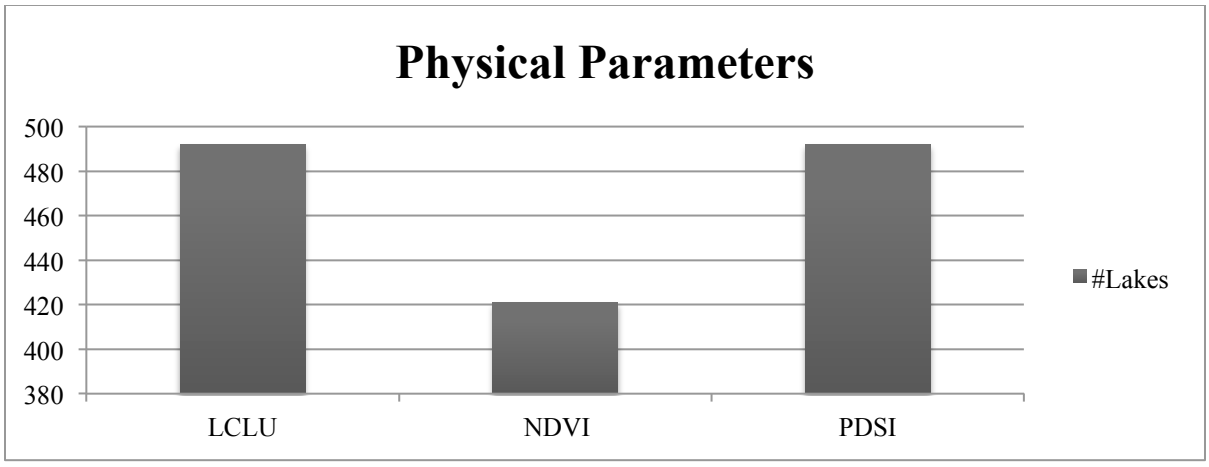


Figure 7. Total no. of lakes for which Physical parameters information is available

4.5 Data extraction of Normalized Difference Chlorophyll Index (NDCI)

NDCI gives chlorophyll a concentration, which may be directly proportional to cyanobacteria concentration. We calculated NDCI data from the satellite scenes captured by MERIS, which was active from 2002 to 2012. We downloaded satellite scenes from the ESA website http://merisfrs-merci-ds.eo.esa.int/merci/query/Products.do. To download data, we need to provide coordinates of Georgia and cloud percentage. Though we downloaded products which are atmospherically corrected at second level we observed cloud cover on the scenes so, we downloaded scenes whose cloud cover is less than 10% to avoid noise introduces while calculating NDCI. We projected the satellite scene to WS84, which involves transforming and rectifying the image into a standard projection. The projected image is exported to tiff using Beam software for further analysis. For each scene there are 48 different products available among them few are captured at different spectral band wavelengths and others are calculated using the captured products. We used algal1 band which contains water bodies that contains algae because we extract chlorophyll concentration data in water bodies which contains algae so, we need only water bodies of Georgia which has algae in them. Algal1 band concentration is derived from the ratio between blue and green signal leaving the water surface and the concentration of the algal pigments.

Using algal1 band as reference we extracted the lakes that are visible for each scene manually. Since MERIS satellite resolutions is 300m and data download is limited to scenes whose cloud percentage is less than 10 most of the water bodies are not captured so, we had to work with only few water bodies. The water bodies that are selected and exported contains negative pixels which are treated as noise and eliminated them using ArcGIS raster calculator. Using modified algal1 as a water mask, we extracted pixel values of each lake from band 665

and band 708. We worked on those two bands because 665 band captured information related to chlorophyll absorption and fluorescence reference and 708 band is for fluorescence reference and atmospheric corrections [30]. Using the pixel values of bands calculated NDCI.

$$NDCI = (Band\ 708 - Band\ 665) / (Band\ 708 + Band\ 665)$$

The NDCI pixel values must be in range -1 to +1 but due to noise while capturing resulted in NDCI pixel exceeds the range so, we eliminated them by performing raster calculator on each scene using ArcGIS. We reclassified NDCI pixels into 7 different classes starts from low concentration to severe bloom based on the pixel range in table 3 [24]. For exporting data from raster file that contains NDCI data to excel, we generated separate shape file for each lake i.e., 492 individual shape files from the single shape file using ArcGIS. Using 492 individual shape files as mask we extracted NDCI data from the raster file. Once we have the pixel counts for each class, water body, year and day we calculated mean for each object id and year to normalize.

NDCI range	**Chl-a range**
<-0.1	<7.5
-0.1 to 0	7.5-16
0 to 0.1	16-25
0.1 to 0.2	25-33
0.2 to 0.4	33-50
0.4 to -0.5	>50
0.5 to 1	Severe bloom

Table 3. NDCI pixel range and Chlorophyll-a concentration

4.6 NDCI Data Extraction Algorithm

For each file and year do

Step 1: Manually select the lakes that are visible from algal1 band

Step 2: Create a mask out of selected lakes as algal1_mask

Step 3: Eliminate negative values from the mask

Step 4: Extract Band7 (665 Band) lakes using algal1_mask

Step 5: Extract Band9 (708 Band) lakes using algal1_mask

Step 6: Eliminate negative values from Band9 and Band7

Step 7: NDCI_nonfilter = (Raster (band_9)-Raster (band_7))/ (Raster (band_9) +Raster (band_7))

Step 8: arcpy.gp. RasterCalculator_sa ("Con (\"%ndci_nonfilter%\"<-1,0,1)", ndci_neg_masked)

Step 9: arcpy.gp. Reclassify_sa (ndci_neg_mask, "Value", "0 NODATA;1 1", ndci_filtered_negative_mask, "DATA")

Step 10: arcpy.gp. ExtractByMask_sa (ndci_nonfilter, ndci_filtered_negative_mask, ndci_filter1)

Step 11: arcpy.gp. RasterCalculator_sa ("Con (\"%ndci_nonfilter%\">1,0,1)", positive)

Step 12: arcpy.gp. Reclassify_sa (ndci_pos_mask, "Value", "0 NODATA;1 1", ndci_filtered_posotive_mask", DATA")

Step 13: arcpy.gp. ExtractByMask_sa (ndci_filter1, ndci_filtered_posotive_mask, ndci)

Step 14: arcpy.gp. Reclassify_sa (ndci_corr, "Value", "-1 -0.10000000000000001 1; -0.1000000000000000 1;0 2;0 0.10000000000000001 3;0.10000000000000001

0.20000000000000001 4;0.20000000000000001 0.40000000000000002

5;0.40000000000000002 0.5 6;0.5 0.99 7", Reclass_ndci, "DATA")

For each 492 lakes shape files do

Step 15: arcpy.gp. RasterCalculator_sa ("\"%ndci%\"", lake)

Step 16: Convert raster file to dbf file

Step 17: Process the file to extract the pixel count

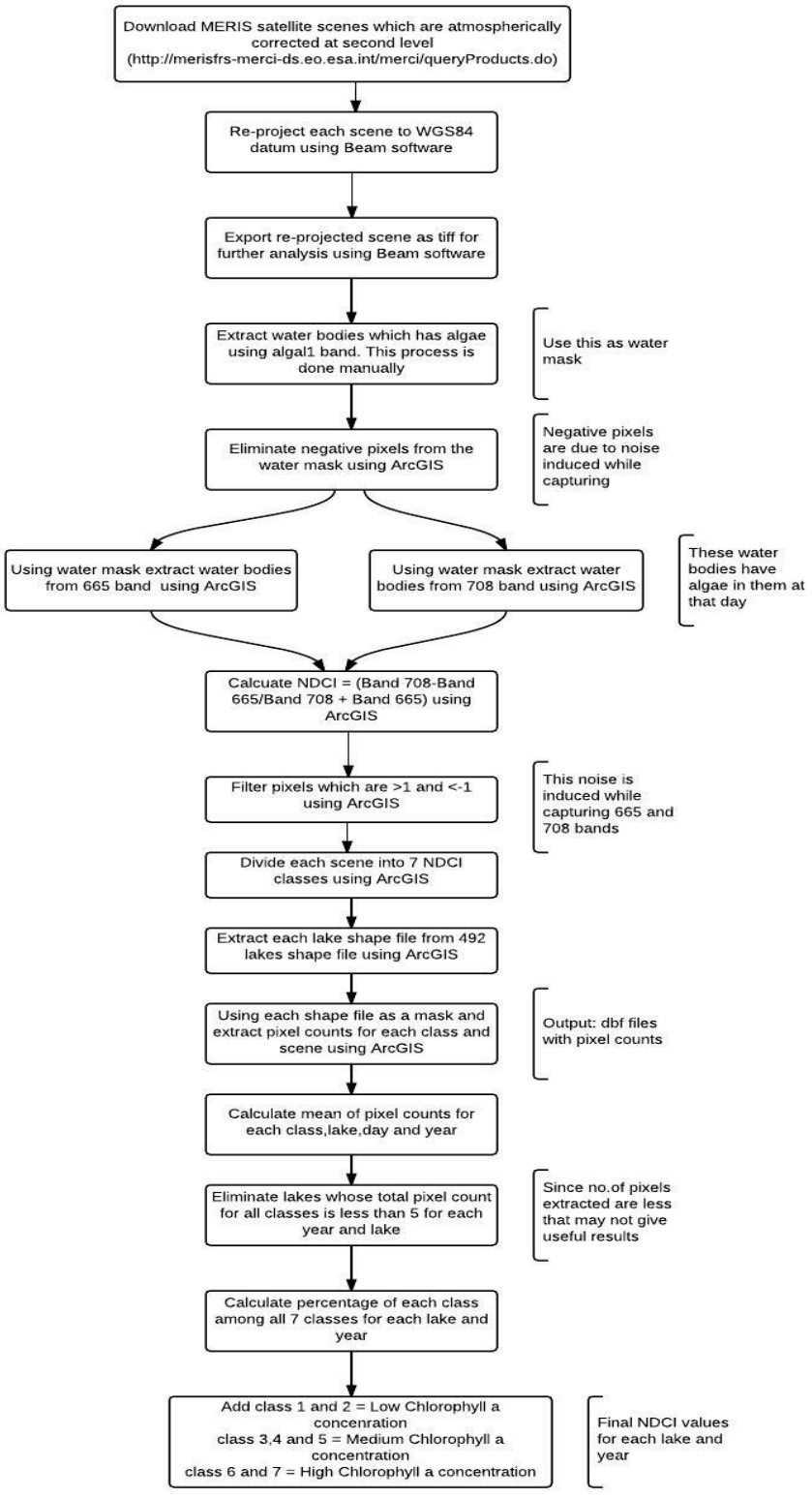


Figure 8. Flowchart of NDCI data extraction procedure

CHAPTER 5

EMPIRICAL STUDY

Our empirical study has three main goals **A.** Performing clustering on NDCI data of year 2010 for lakes of Georgia that are captured using MERIS. **B.** Performing clustering on NDCI data of year 2011 for lakes of Georgia that are captured using MERIS. **C.** Analyzing the clusters of lakes based on NDCI data for years 2010 and 2011 with respect to Physical parameters. There is data available for years 2002 to 2012 but we chose to perform analysis on the data of 2010 and 2011 because MERIS was able to capture maximum number of lakes in these years compared to the other years. As the quality of analysis directly depends on the amount of data available, we restricted our analysis to these two years. As mentioned earlier we downloaded the satellite scenes where cloud cover percentage is less than 10. Though we restricted the cloud cover, we were not able to get as many lakes information as possible because of noise introduced in the data.

There are 53 lakes in total that has NDCI data extracted for years 2010 and 2011. Table 4 has the list of lake id, object id, lake name, coordinates, county, zip code and division/zone information. The lake id is used to represent the lake in further analysis. Lake names for all lakes are not available so we marked them as not available.

Lake id	ObjectID	Lake Name	Ycentroid	Xcentroid	County	Zone
Lake1	16772	Not Available	34.5287	-83.0379	Franklin	3
Lake2	24347	Not Available	33.4465	-81.966	Richmond	6
Lake3	33975	Reservoir 51	32.839	-82.4385	Jefferson	6
Lake4	37656	J. Strom Thurmond Reservoir	33.9799	-82.6104	Wilkes	3
Lake5	42936	Savannah River	34.3933	-82.87	Hart	3
Lake6	45144	Lake Burton	34.8371	-83.5525	Rabun	3

Lake7	49165	Richard B Russell Lake	34.1605	-82.6905	Elbert	3
Lake8	56496	Lake Tobesofkee	32.8377	-83.805	Bibb	5
Lake9	76515	Lake Juliette	33.0452	-83.8016	Monroe	4
Lake10	78307	Cole Reservoir	33.3427	-84.2216	Henry	4
Lake11	83280	Not Available	32.7974	-83.448	Twiggs	5
Lake12	92266	Not Available	32.8694	-83.2207	Wilkinson	5
Lake13	94440	Not Available	33.0628	-83.8249	Monroe	4
Lake14	94519	Not Available	32.8685	-83.3274	Wilkinson	5
Lake15	105217	Not Available	33.368	-83.9985	Henry	4
Lake16	107928	Not Available	32.9056	-83.3004	Wilkinson	4
Lake17	116880	Not Available	32.7824	-83.6141	Bibb	5
Lake18	122801	House Lake	30.8645	-82.195	Charlton	9
Lake19	123616	Not Available	33.0762	-83.8159	Monroe	4
Lake20	132731	Not Available	32.9838	-82.8815	Washington	5
Lake21	136953	Lake Sinclair	33.2266	-83.2641	Putnam	5
Lake22	150600	Jackson Lake	33.3729	-83.8629	Newton	5
Lake23	164064	Not Available	33.041	-82.9034	Washington	5
Lake24	172577	Not Available	32.9109	-83.6539	Bibb	5
Lake25	173019	Not Available	32.6998	-83.5766	Bibb	5
Lake26	186515	Harbins Lake	33.4833	-84.2814	Clayton	4
Lake27	186560	Not Available	32.7054	-83.5711	Bibb	5
Lake28	192980	Perch Lake	30.9394	-82.1674	Charlton	9
Lake29	193055	Lake Oconee	33.4878	-83.2468	Greene	5
Lake30	197809	Blue Ridge Lake	34.8469	-84.2663	Fannin	2
Lake31	209001	Carters Lake	34.6185	-84.6366	Murray	1
Lake32	210641	Nottely Lake	34.9167	-84.0548	Union	2
Lake33	211778	Not Available	34.5996	-84.6789	Murray	1
Lake34	212411	Not Available	32.1531	-85.0432	Harris	4
Lake35	233072	Not Available	33.4294	-85.0567	Heard	4
Lake36	235014	Not Available	31.6104	-84.116	Dougherty	7
Lake37	237515	Lake Acworth	34.0556	-84.6792	Cobb	2
Lake38	245467	West Point Lake	33.0646	-85.1344	Troup	4
Lake39	249472	Chatuge Lake	34.9838	-83.7708	Towns	2
Lake40	251601	Lake Seminole	30.8126	-84.819	Seminole	7
Lake41	254300	Lake Blackshear	31.946	-83.9362	Sumter	7
Lake42	259235	Not Available	30.7689	-84.9511	Seminole	7
Lake43	262177	Not Available	33.4133	-85.053	Heard	4
Lake44	263903	Not Available	32.2493	-84.9175	Stewart	7
Lake45	271963	Lake Oliver	32.5576	-85.0363	Muscogee	4
Lake46	273732	Altoona Lake	34.1402	-84.6433	Cherokee	2
Lake47	276050	Not Available	33.4275	-85.0374	Carroll	4
Lake48	280501	Not Available	33.5423	-84.9485	Carroll	4

Lake49	281848	Walter F George Reservoir	31.8428	-85.0934	Quitman	7
Lake50	282610	Lake Seminole	30.7903	-84.7565	Decatur	7
Lake51	282683	Lake Seminole	30.7781	-84.9132	Seminole	7
Lake52	282815	Lake Sidney Lanier	34.2767	-83.9331	Hall	2
Lake53	282829	Bartlett's Ferry Lake	32.7044	-85.1271	Troup	4

Table 4. Information about lakes captured in years 2010 and 2011

We restricted our analysis to lakes that are captured by MERIS from April to September of every year because cyanobacteria tend to grow in summer with high temperatures. For each lake there are 7 classes of NDCI data available which gives information from low to severe chlorophyll concentration. All 7 classifications are further reduced to 3 as low, medium, and high chlorophyll concentrations to analyze the chlorophyll concentration at different levels. We calculated the percentages of pixels for each class and executed K-Means Clustering algorithm on 53 lakes for 7 classes

5.1 Clustering on chlorophyll concentration data of year 2010

Only 45 lakes chlorophyll concentration data are extracted among 492 lakes in the year 2010 and our analysis focused on these 45 lakes and reservoirs to identify which lakes have same trend related to chlorophyll concentration. The silhouette coefficient determines K value in K-Means Clustering to identify the optimal clusters. We observed optimal dissimilarity for the data set when clustered them into 3 groups and then executed K-Means clustering algorithm with k=3. Three clusters and their data distribution with respect to physical parameters and NDCI values are shown in the below figures: 9 to 14

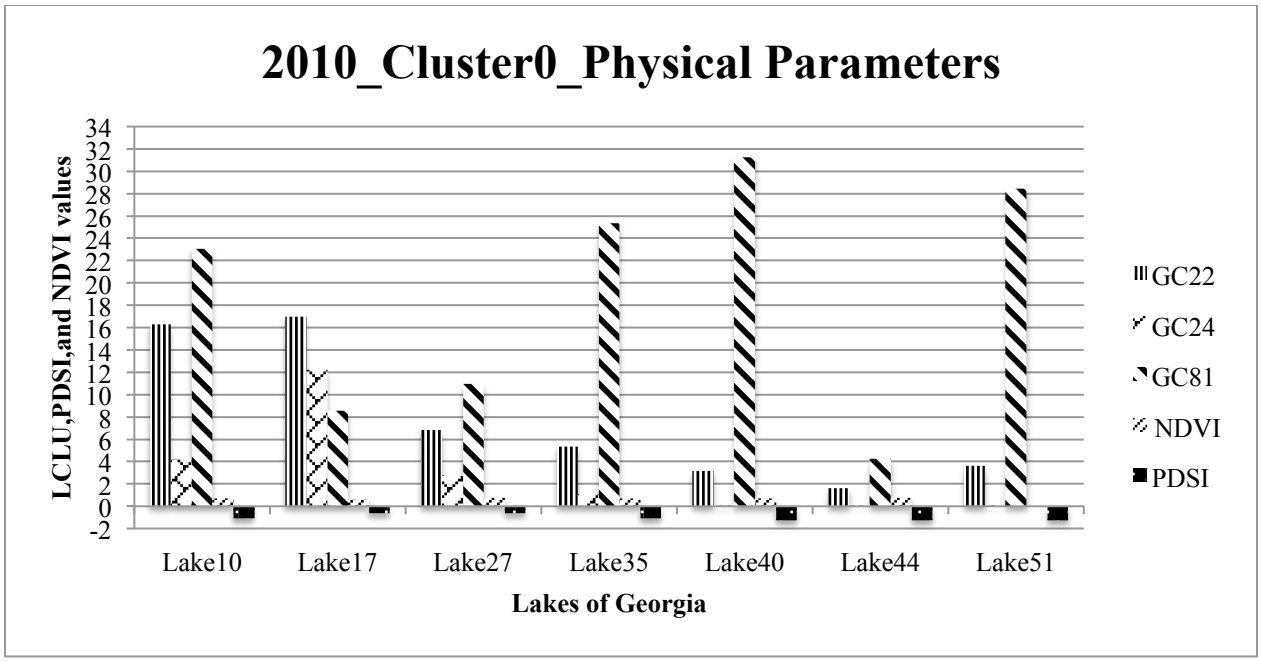


Figure 9. Year 2010 Cluster 0 Physical Parameters data distribution

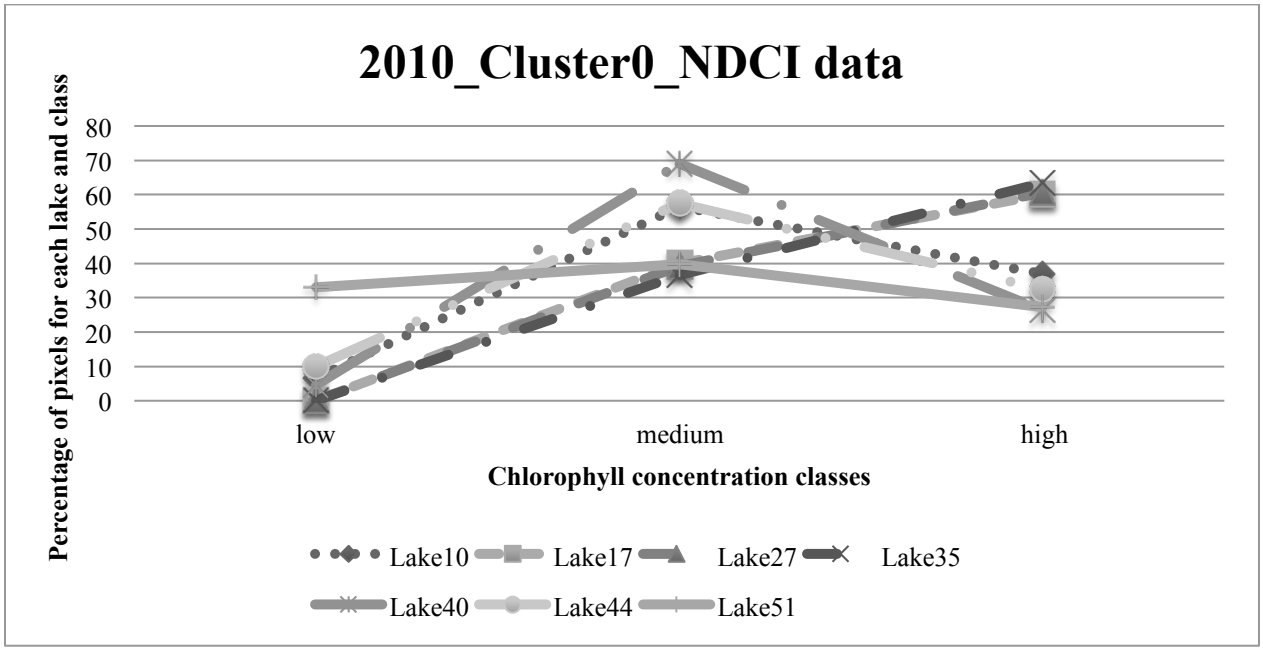
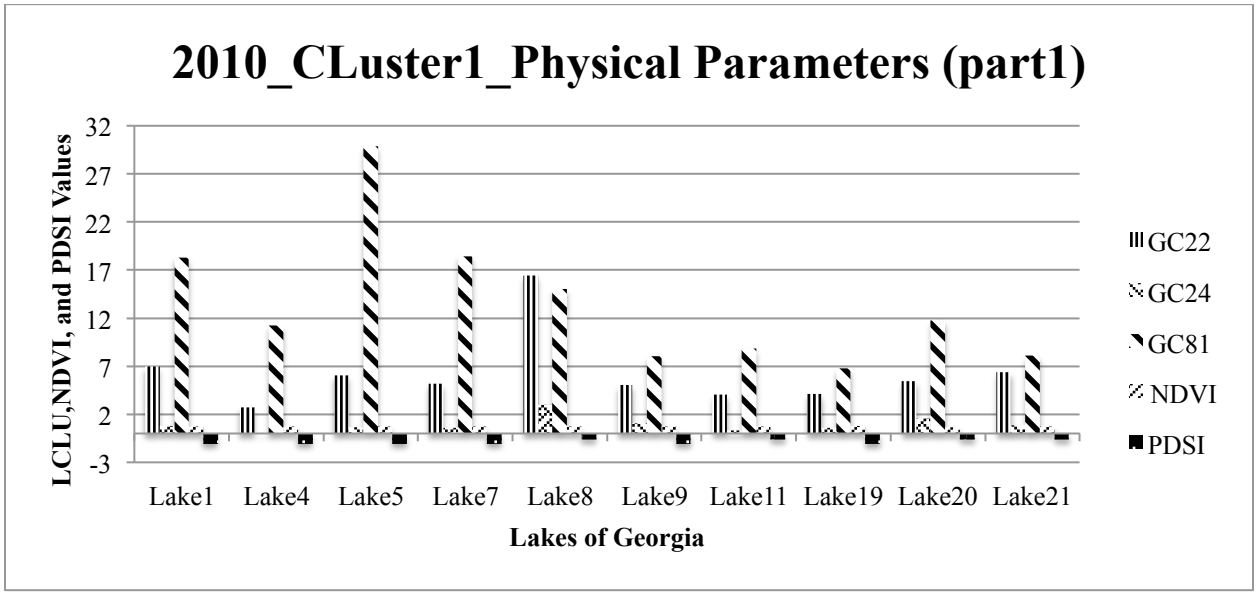
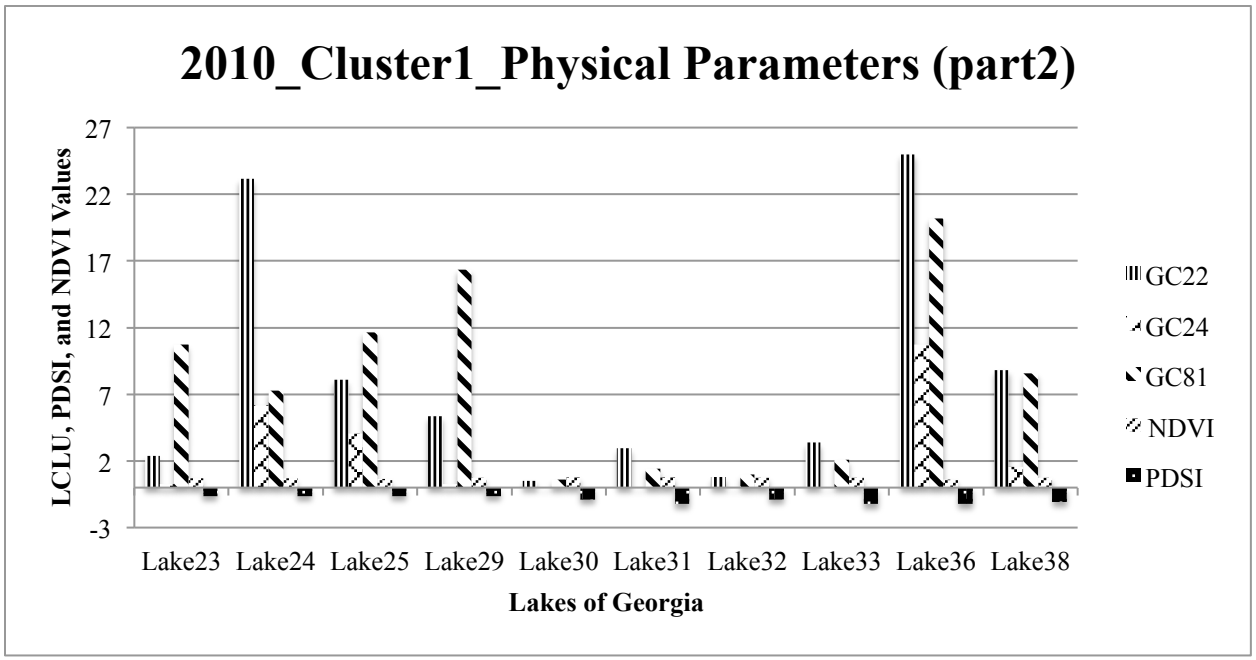


Figure 10. Year 2010 Cluster 0 NDCI data distribution

In cluster0 we observed most of the lakes are high and medium chlorophyll concentrated. In physical parameters data distribution, most of the areas surrounded by these lakes are covered

by agriculture when compared to low urban and high urban. These lakes are in areas where normal drought conditions are observed.





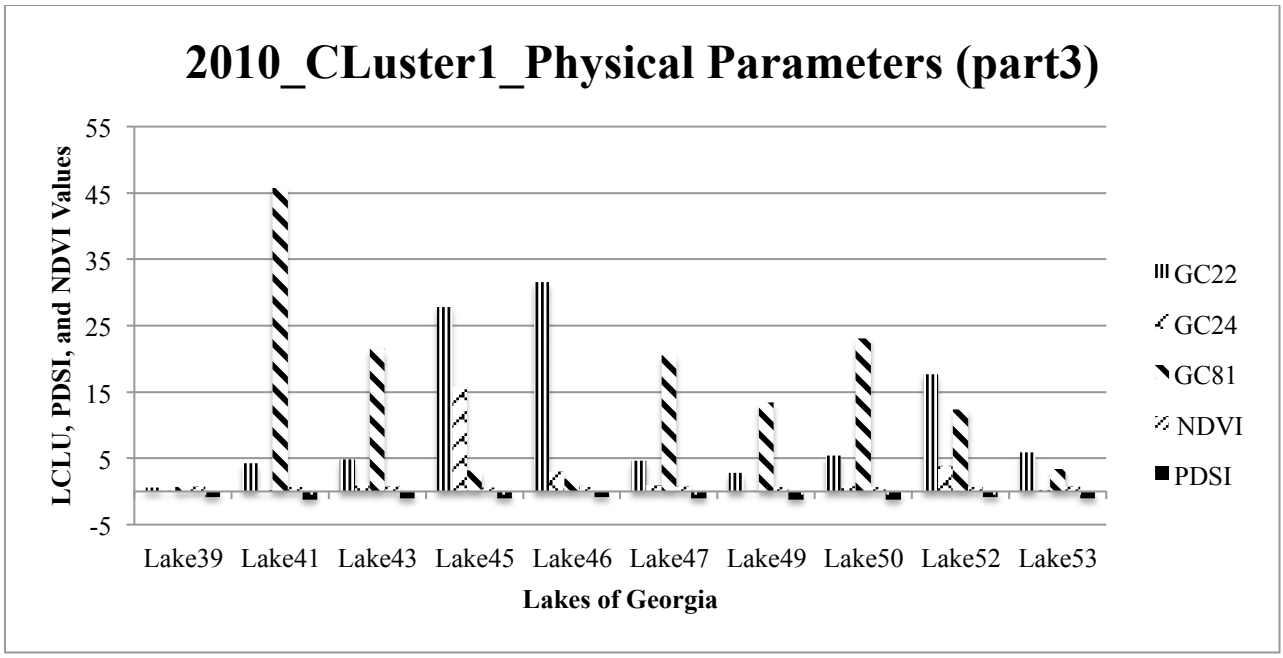


Figure 11. Year 2010 Cluster1 Physical Parameters data distribution

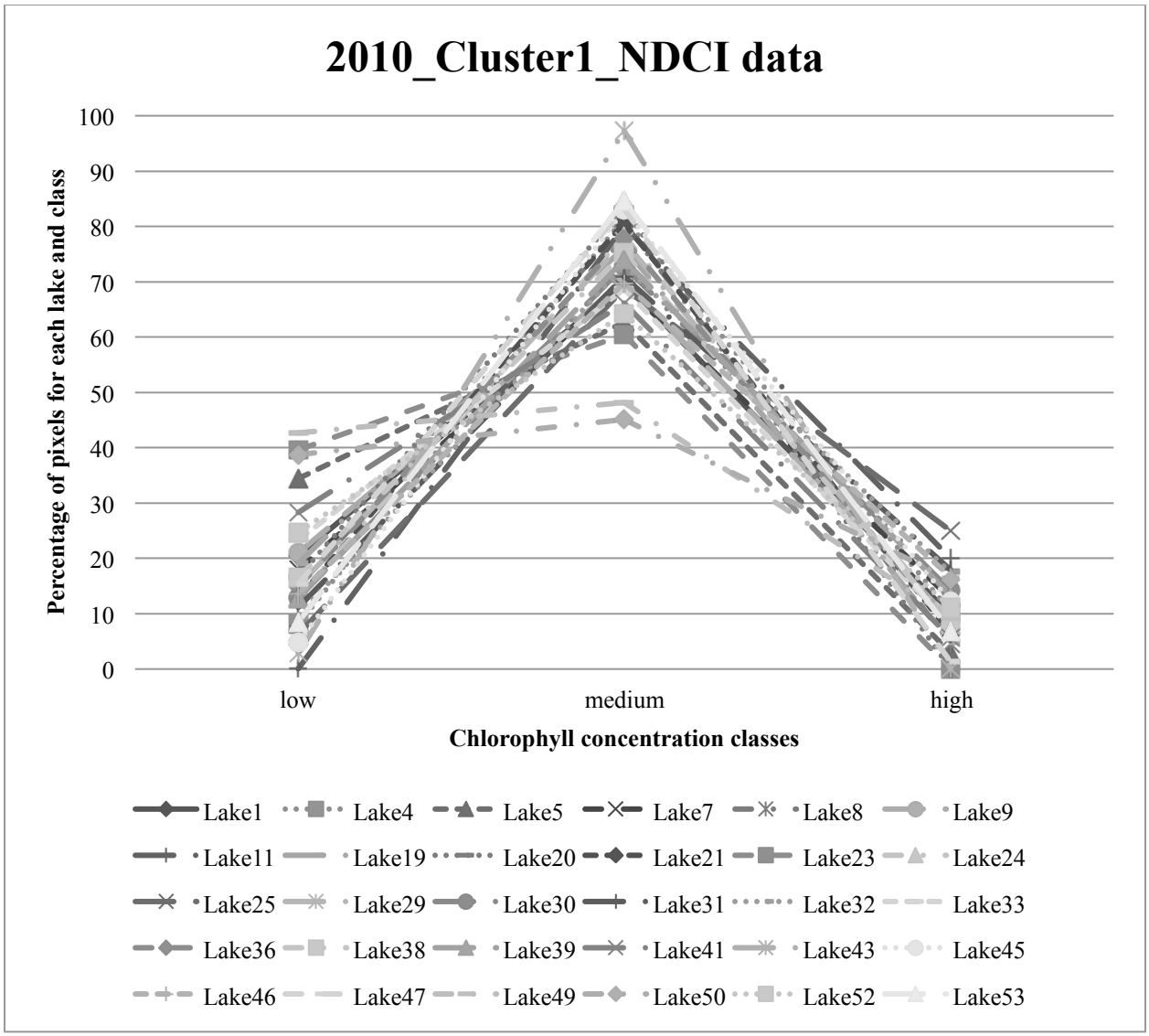


Figure 12. Year 2010 Cluster1 NDCI data distribution

In cluster1 we observed most of the lakes are low and medium chlorophyll concentrated. In the physical parameter data distribution, we observed few lakes are surrounded by high low urban land coverage when compared to agriculture and high urban but for few lakes they are more covered by agriculture land. We observed that lakes are surrounded by normal drought conditions.

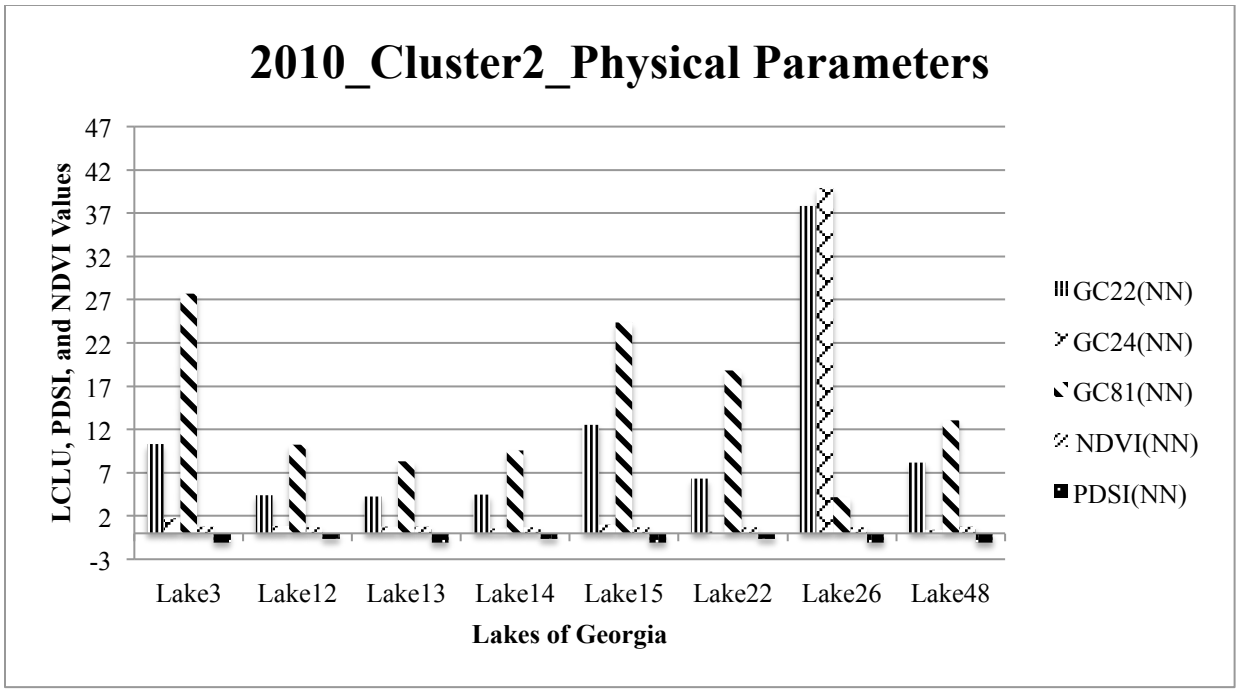


Figure 13. Year 2010 Cluster 2 Physical Parameters data distribution

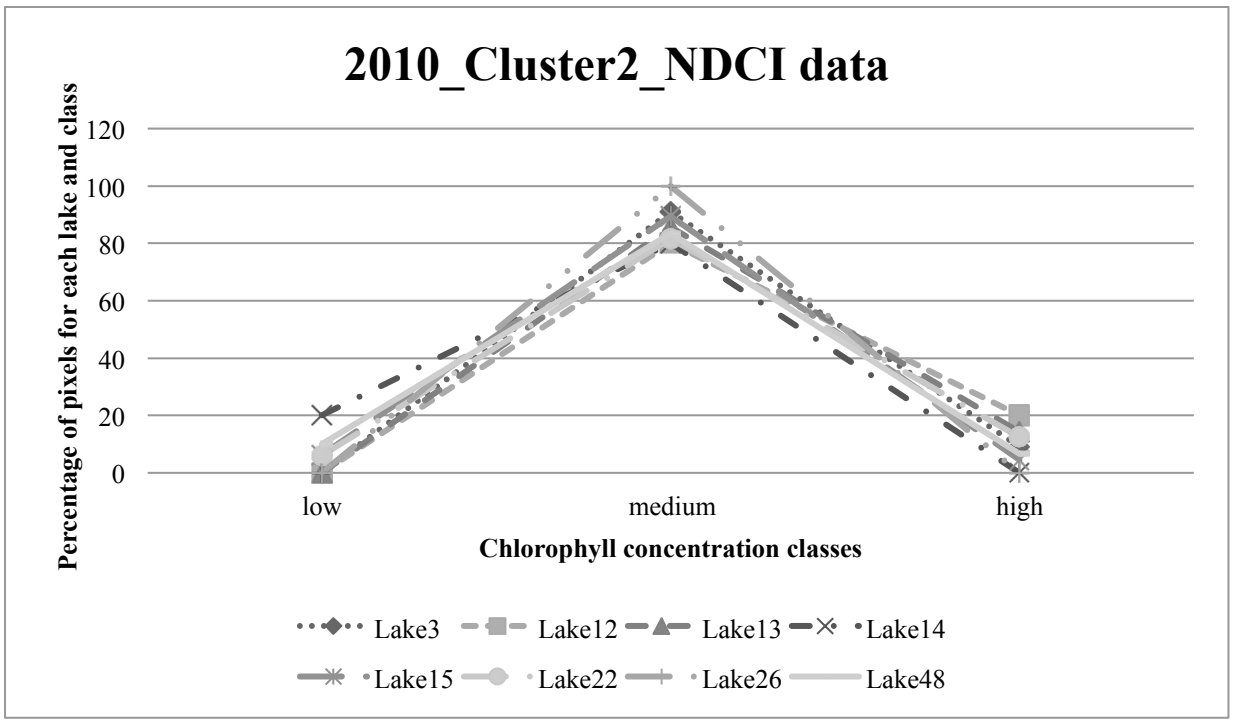


Figure 14. Year 2010 Cluster2 NDCI data distribution

In cluster2 we observed all most all lakes are medium chlorophyll concentrated. In the physical parameter data distribution, we observed few lakes are surrounded by more low urban land and agriculture land coverage when compared to high urban. Among low urban and agricultural land these lakes are more agriculture land covered. We observed that lakes are surrounded by normal drought conditions.

5.2 Clustering on chlorophyll concentration data of year 2011

There are a total of 43 lakes that are captured by MERIS among 492 lakes in year 2011 and we performed analysis only on those lakes.

Silhouette Coefficient determines K value in K-Means Clustering. We executed K-Means clustering algorithm for different k's starting from k=2 and observed maximum dissimilarity when k=6. Six clusters and their data distribution with respect to physical parameters and NDCI distribution is shown in figures 15 to 26

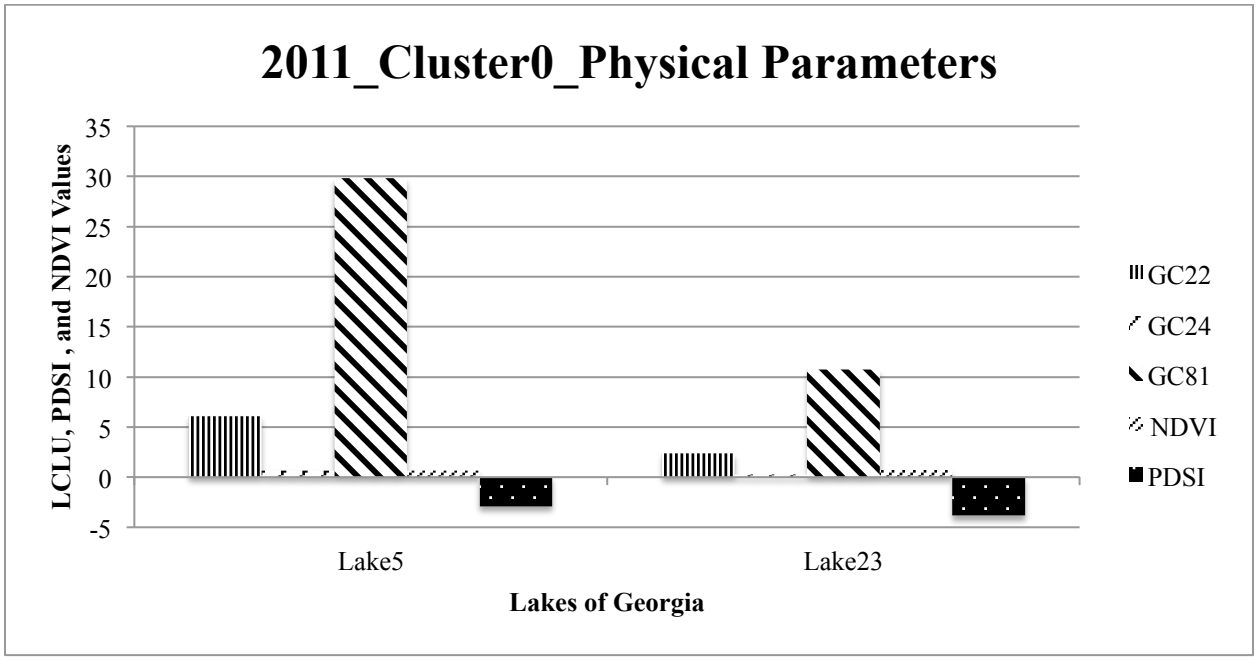


Figure 15. Year 2011 Cluster0 Physical Parameters data distribution

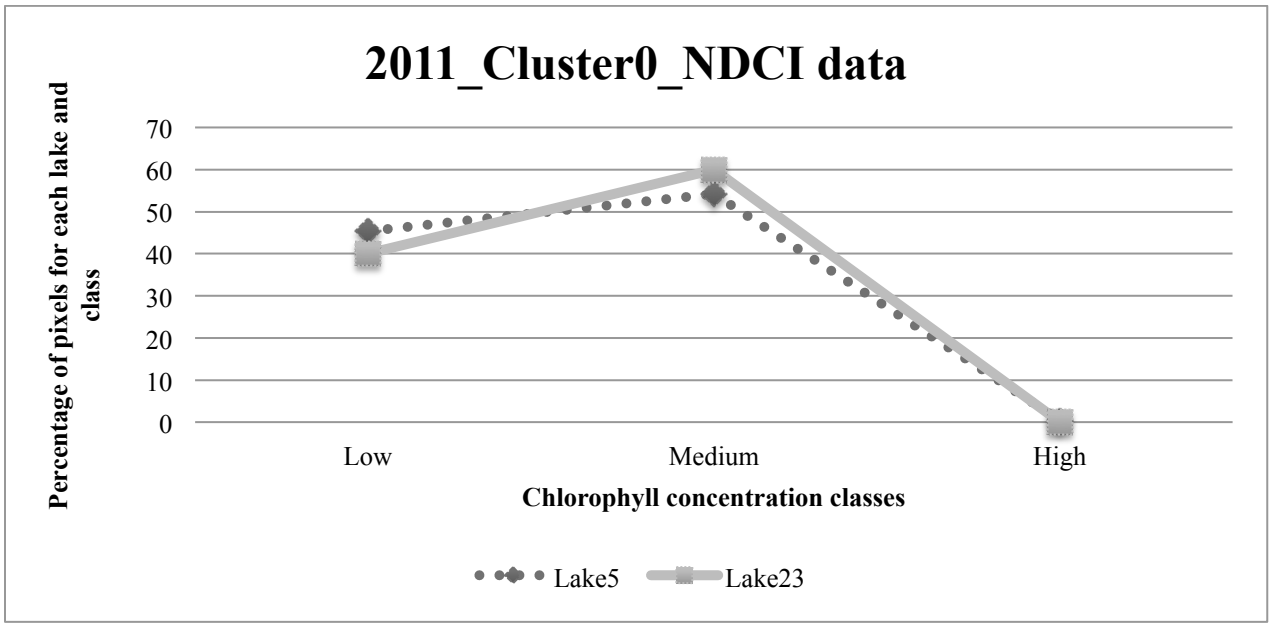


Figure 16. Year 2011 Cluster0 NDCI data distribution

In cluster0 we observed all lakes are low and medium chlorophyll concentrated. In the physical parameter data distribution, we observed lakes are surrounded by more agriculture land coverage when compared to high and low urban. We observed that lakes are surrounded by severe drought conditions.

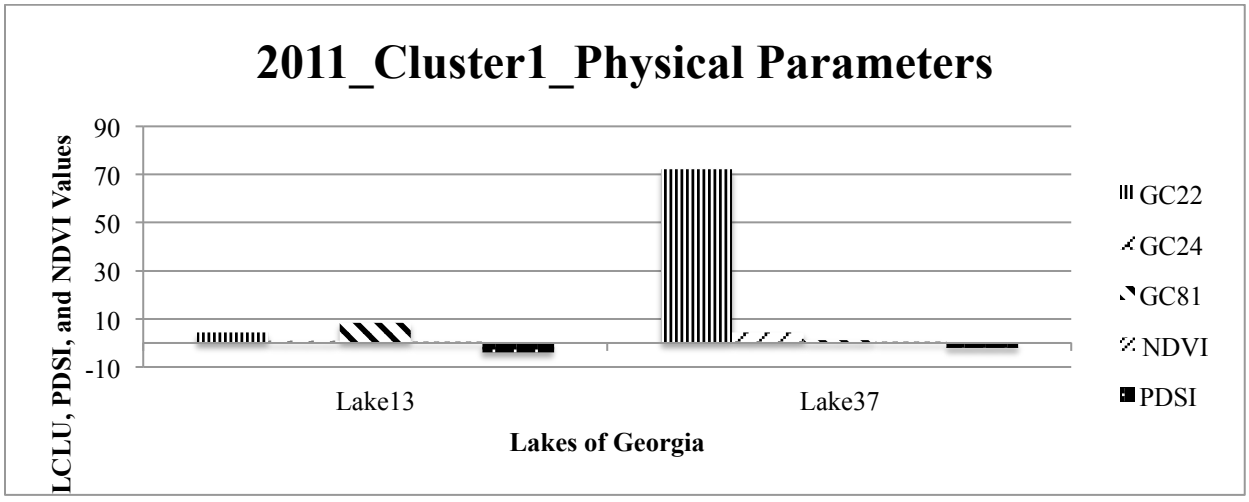


Figure 17. Year 2011 Cluster1 Physical Parameter data distribution

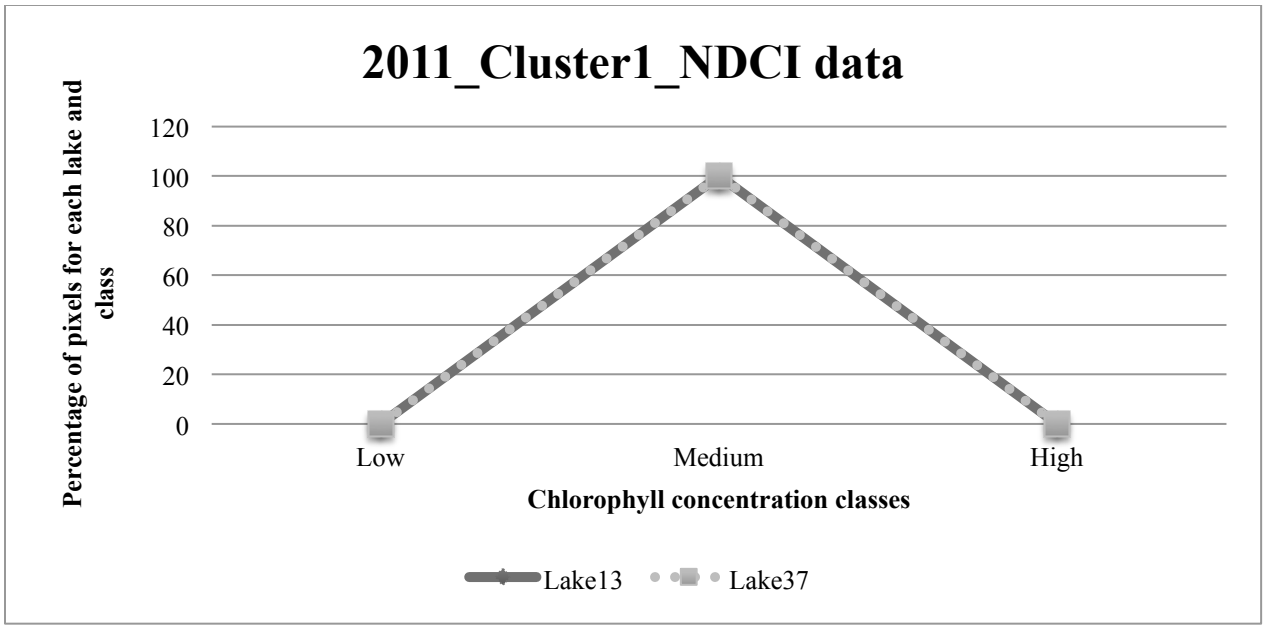


Figure 18. Year 2011 Cluster1 NDCI data distribution

In cluster1 we observed all lakes are medium level chlorophyll concentrated. In the physical parameter data distribution, we observed different distribution of land cover where lake37 is surrounded by low urban area and lake13 has very less coverage of agriculture, low and high urban. We observed that lakes are surrounded by moderate drought conditions.

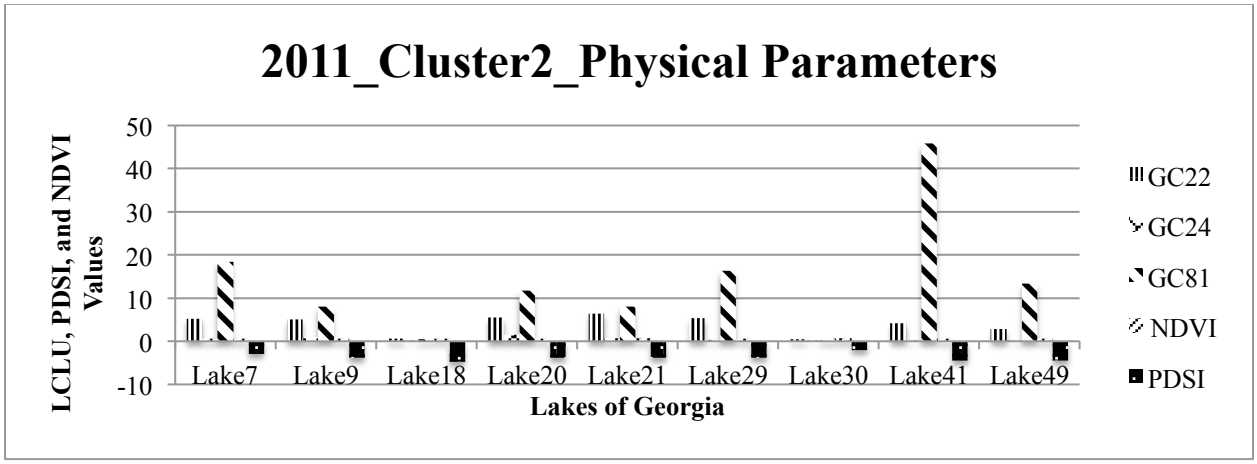


Figure 19. Year 2011 Cluster2 Physical Parameters data distribution

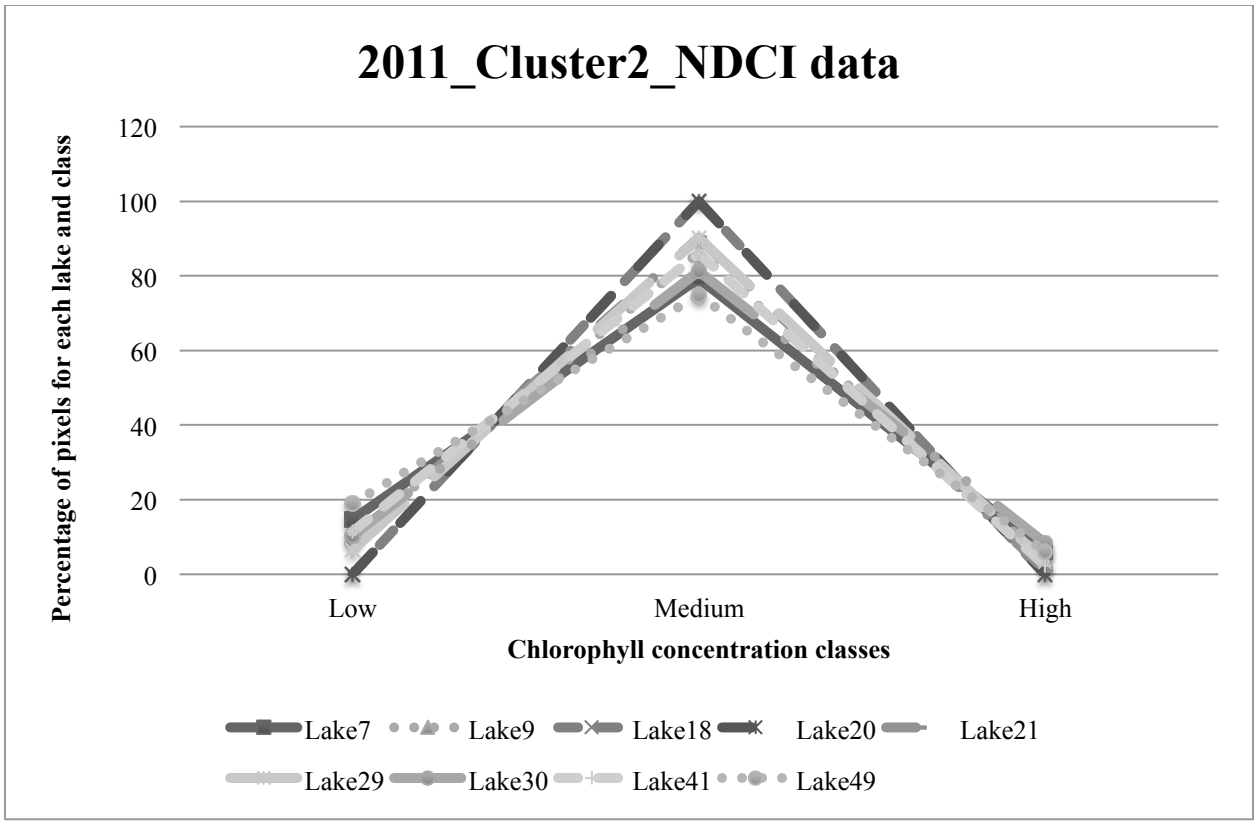
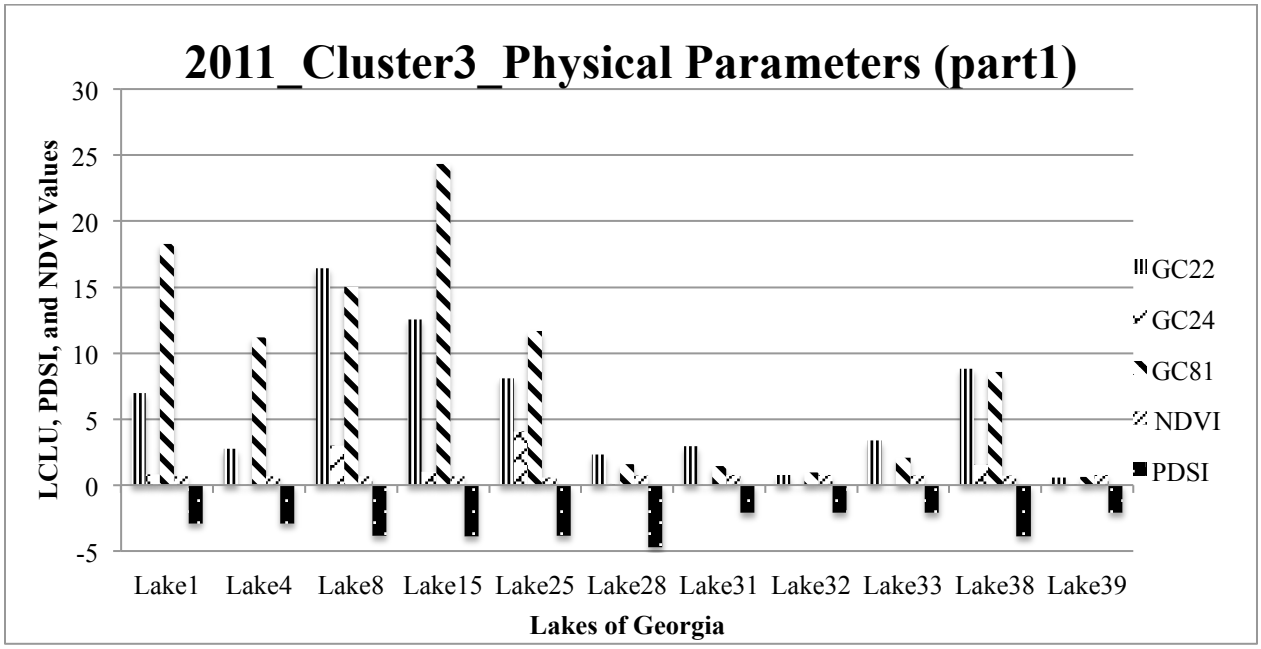


Figure 20. Year 2011 Cluster2 NDCI data distribution

In cluster2 we observed all lakes are high medium chlorophyll concentrated. In the physical parameter data distribution, we observed few lakes are surrounded by more low urban land and agriculture land coverage when compared to high urban. Among low urban and agricultural land, lakes are more agricultural land covered. We observed that lakes are surrounded by severe drought conditions.



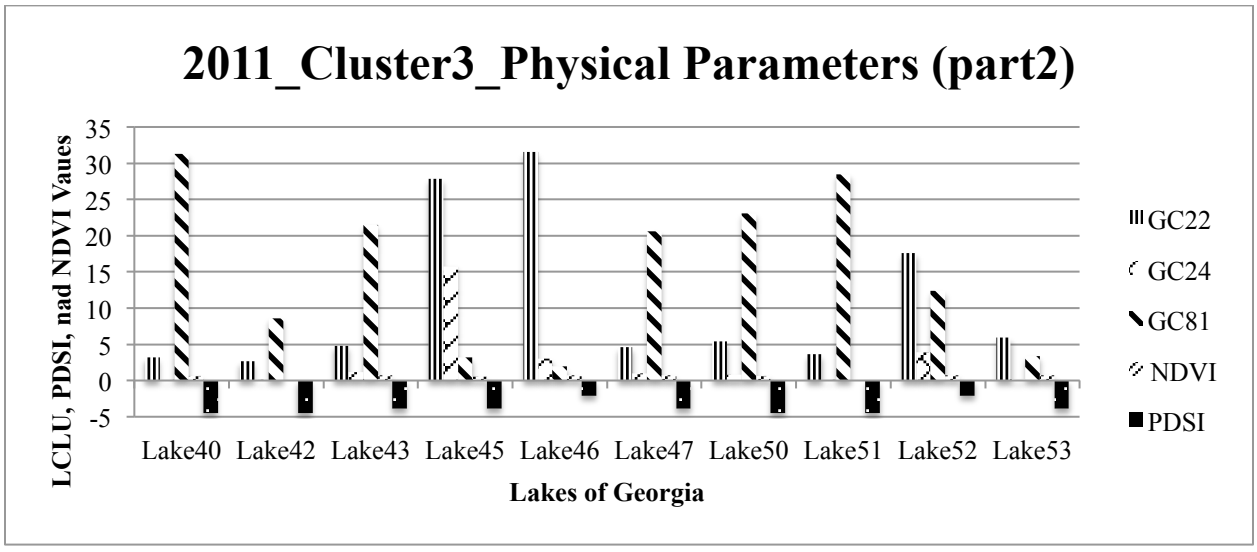


Figure 21. Year 2011 Cluster3 Physical Parameters data distribution

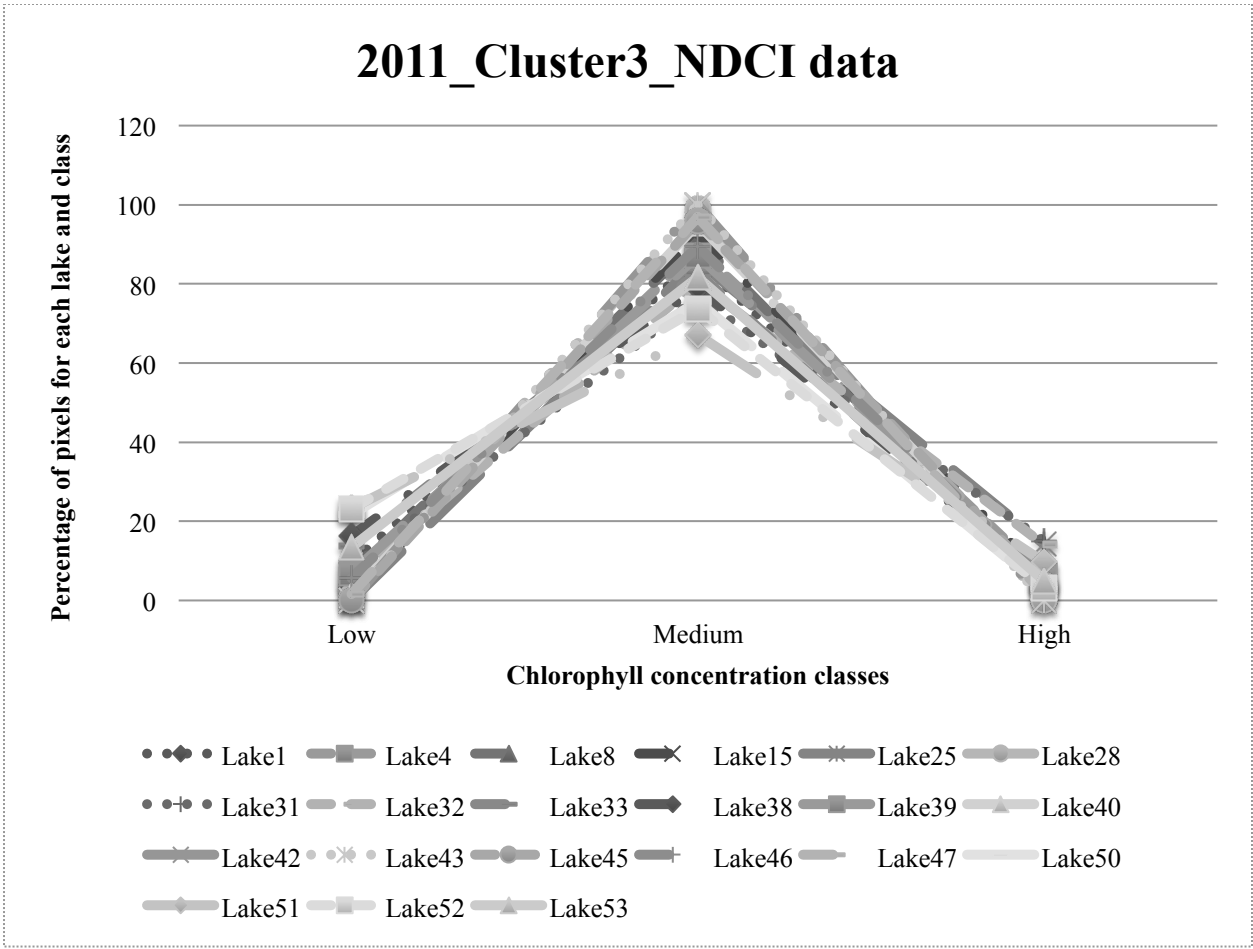


Figure 22. Year 2011 Cluster3 NDCI data distribution

In cluster3 we observed all lakes are high medium chlorophyll concentrated. In the physical parameter data distribution, we observed few lakes are surrounded by more low urban land and agriculture land coverage when compared to high urban. Among low urban and agricultural land, most of the lakes are more agricultural land covered. We observed that most of the lakes are surrounded by severe drought conditions.

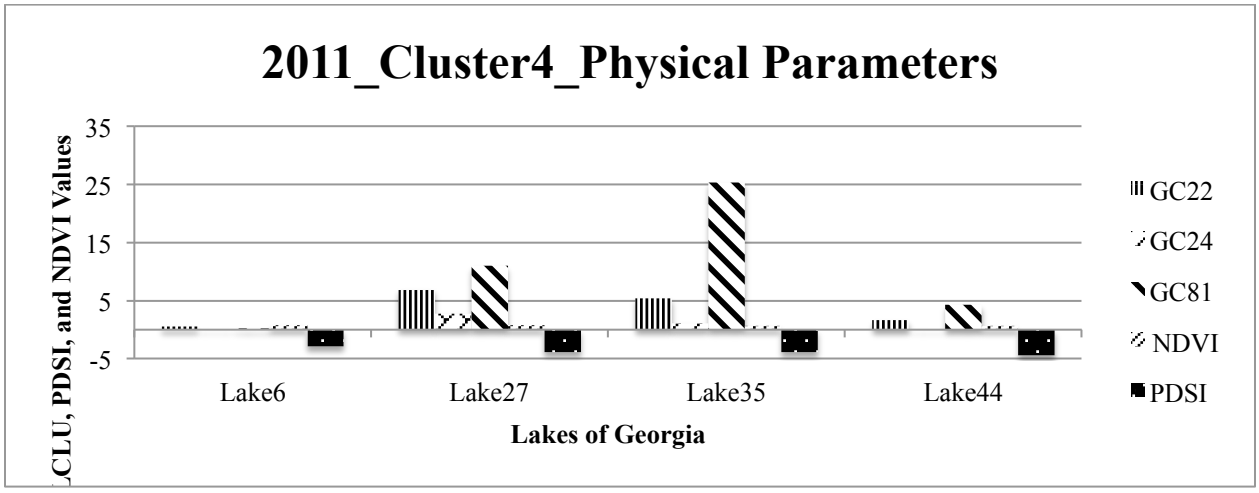


Figure 23. Year 2011 Cluster4 Physical Parameters data distribution

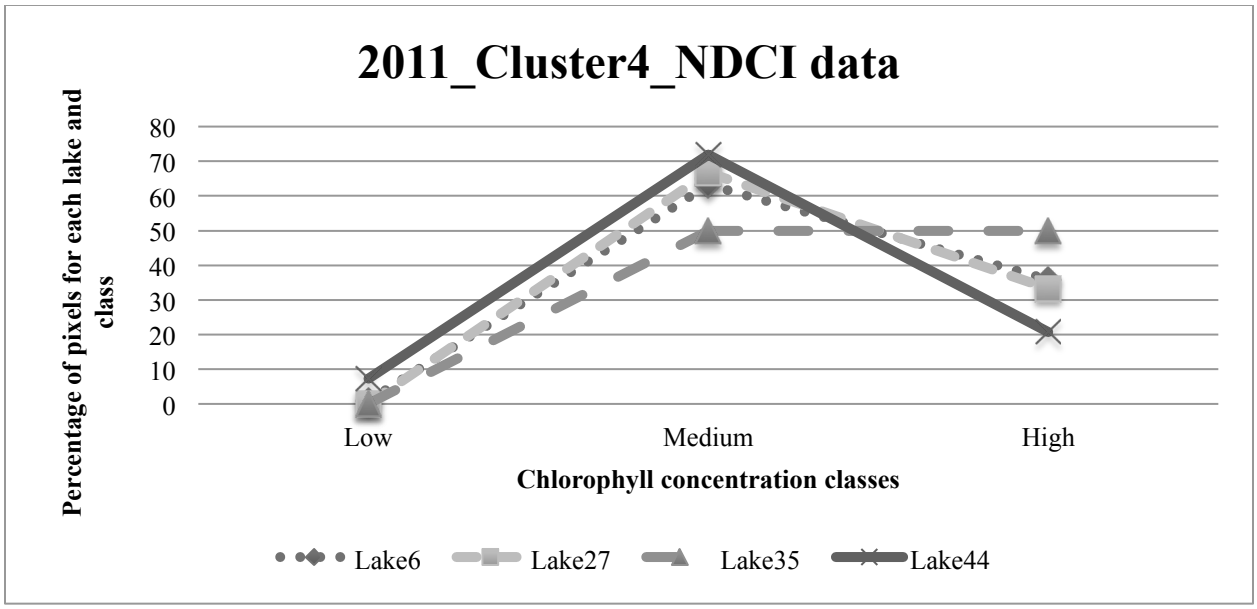


Figure 24. Year 2011 Cluster4 NDCI data distribution

In cluster4 we observed all lakes are high and medium chlorophyll concentrated when compared to low concentration. In the physical parameter data distribution, we observed most of the lakes are surrounded by more low urban land and agriculture land coverage when compared to high urban. Among low urban and agricultural land, those lakes are more agricultural land covered. We observed that most of the lakes are surrounded by moderate drought conditions.

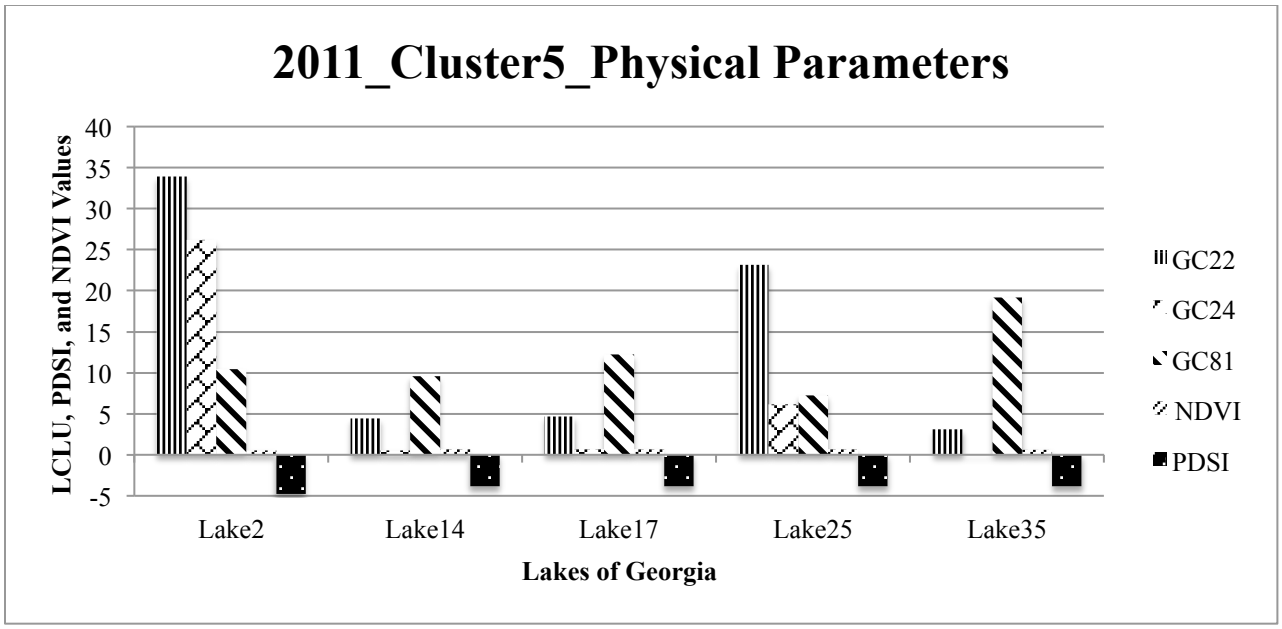


Figure 25. Year 2011 Cluster5 Physical Parameters data distribution

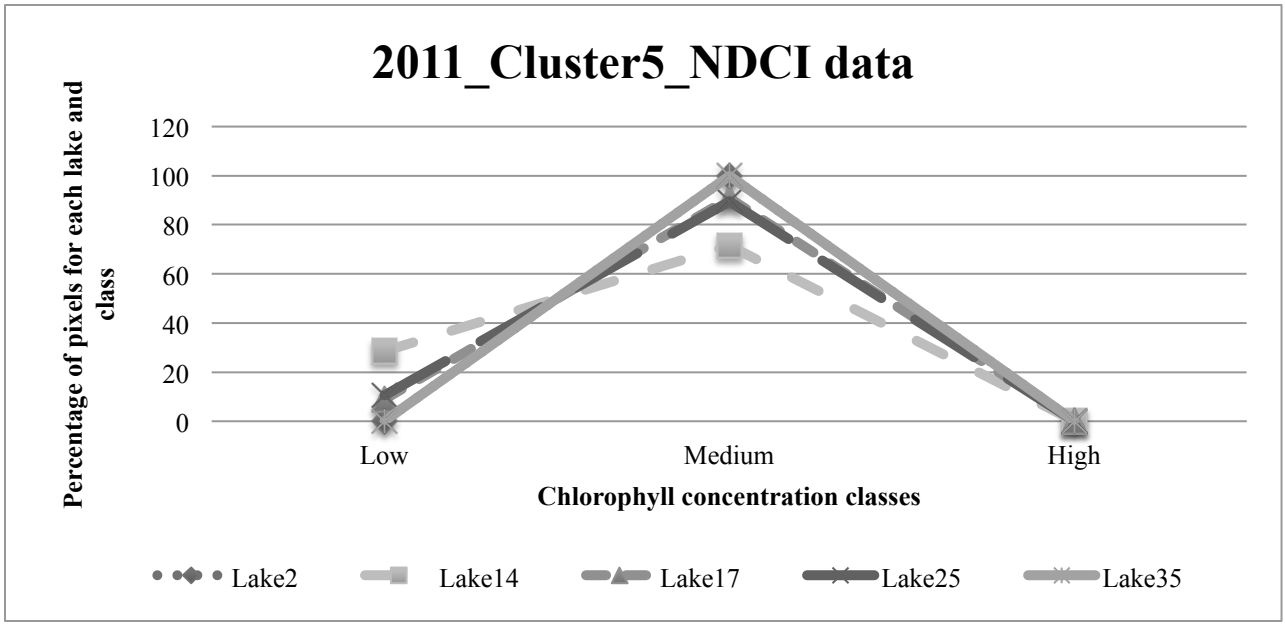


Figure 26. Year 2011 Cluster5 NDCI data distribution

In cluster5 we observed almost all lakes are medium chlorophyll concentrated when compared to low and high concentration. In the physical parameter data distribution, we observed few lakes are surrounded by more low urban land and agriculture land coverage when

compared to high urban. Among low urban and agricultural land, those lakes are covered by more agricultural land but lake2 has more high urban land coverage. We observed that most of the lakes are surrounded by severe drought conditions.

5.3 Computational and Data Analysis for years 2010 and 2011

In 2010 data there are few lakes that have high concentration of chlorophyll and others have medium chlorophyll concentration. In 2011 there are few lakes that are high chlorophyll concentrated and others are medium and low chlorophyll concentrated. In two years there is difference in low and high chlorophyll concentration but most of the lakes have are medium chlorophyll concentrated. To find out which physical parameter has high impact in change related to chlorophyll concentration between these years, we compared and analyzed the data of each physical parameter of all clusters for the lakes that are captured by MERIS in both 2010 and 2011. There are 35 lakes in total which are captured in 2010 and 2011 for which we compared the physical parameters distribution to identify which parameter has more impact towards change in chlorophyll concentration

Land Usage and Land Cover data is same for both the years because we are able to extract data of LULC for years 2002, 2005 and 2008. We do not have specific land cover data for years 2010 and 2011. So, we duplicated the data of 2008 for years 2010 and 2011 assuming that there is no big difference in LULC data. We observed that there is no huge difference between NDVI data of 2011 and 2010 but, we observed severe drought conditions in 2011 compared to 2010. Figures 27 to 31 shows the comparison of physical parameters data in years 2010 and 2011.

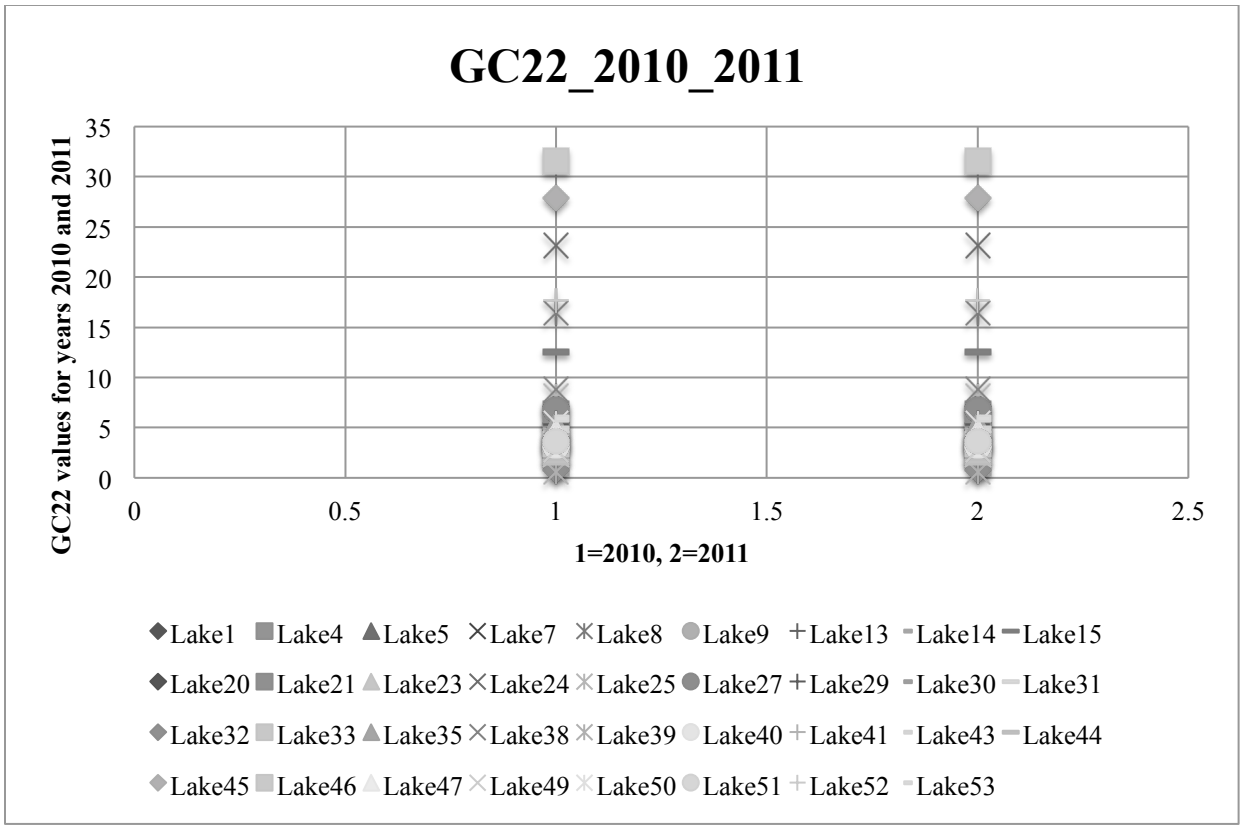


Figure 27. Comparison of GC22 data for years 2010 and 2011

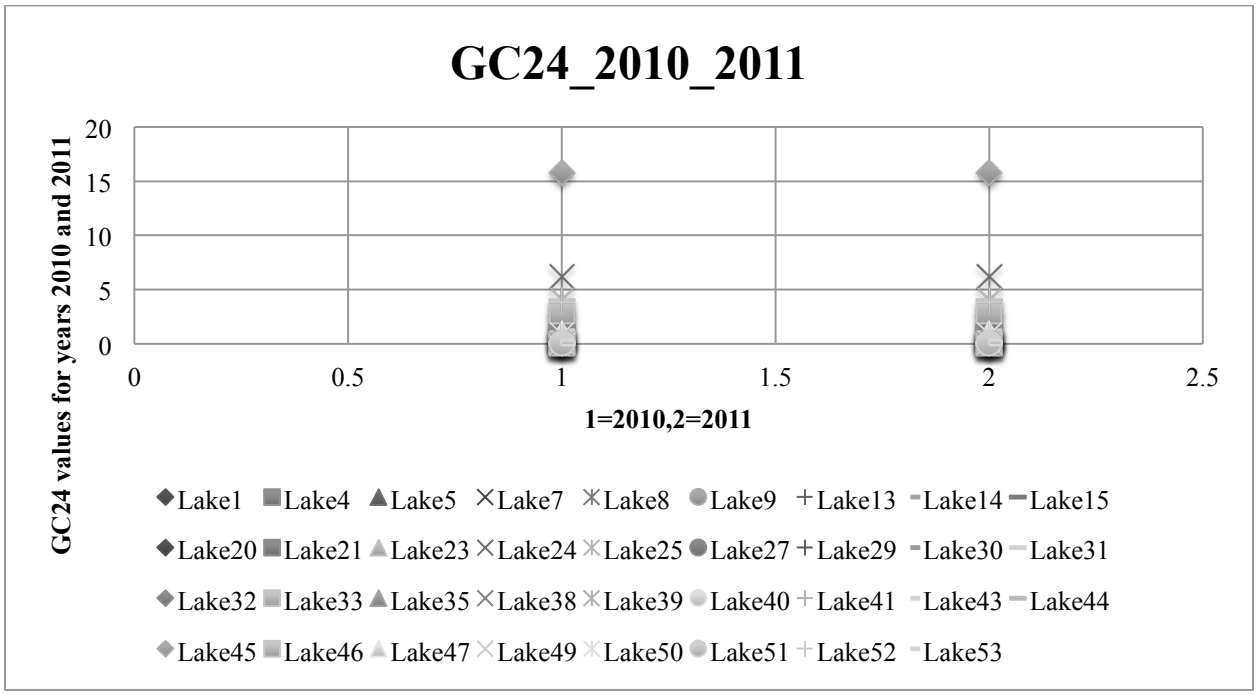


Figure 28. Comparision of GC24 data for years 2010 and 2011

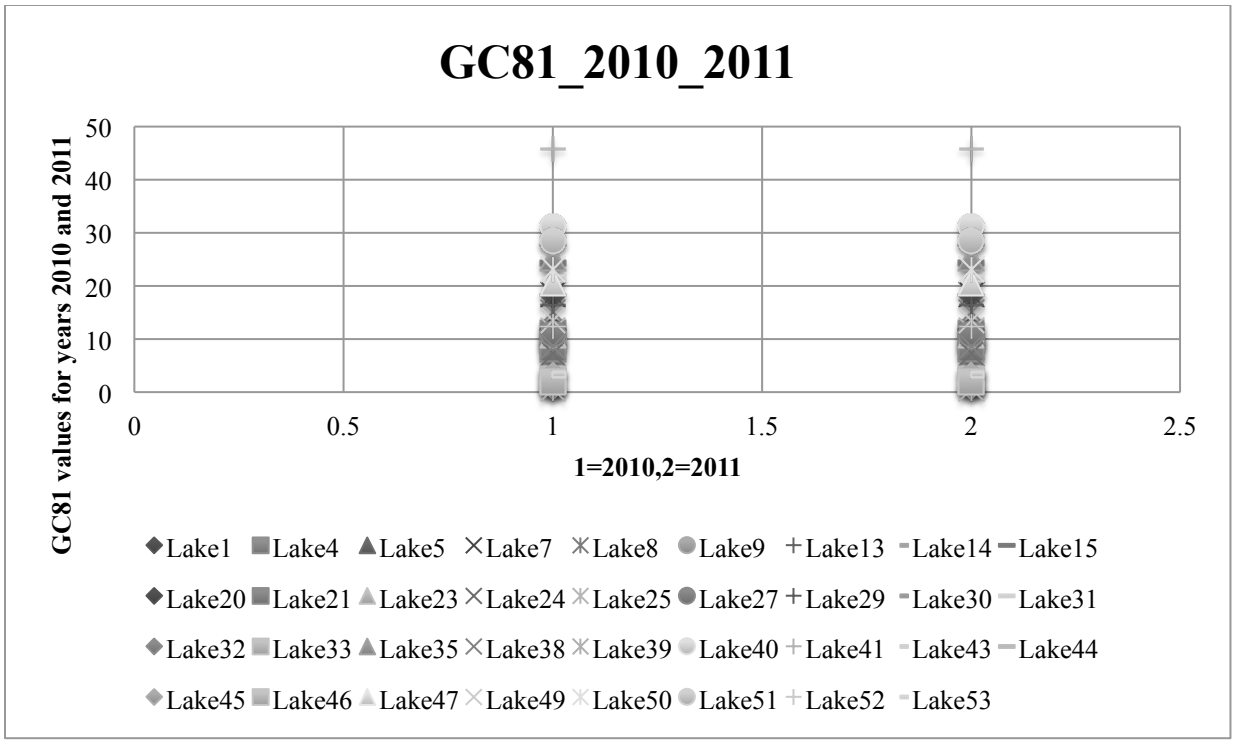


Figure 29. Comparison of GC81 data for years 2010 and 2011

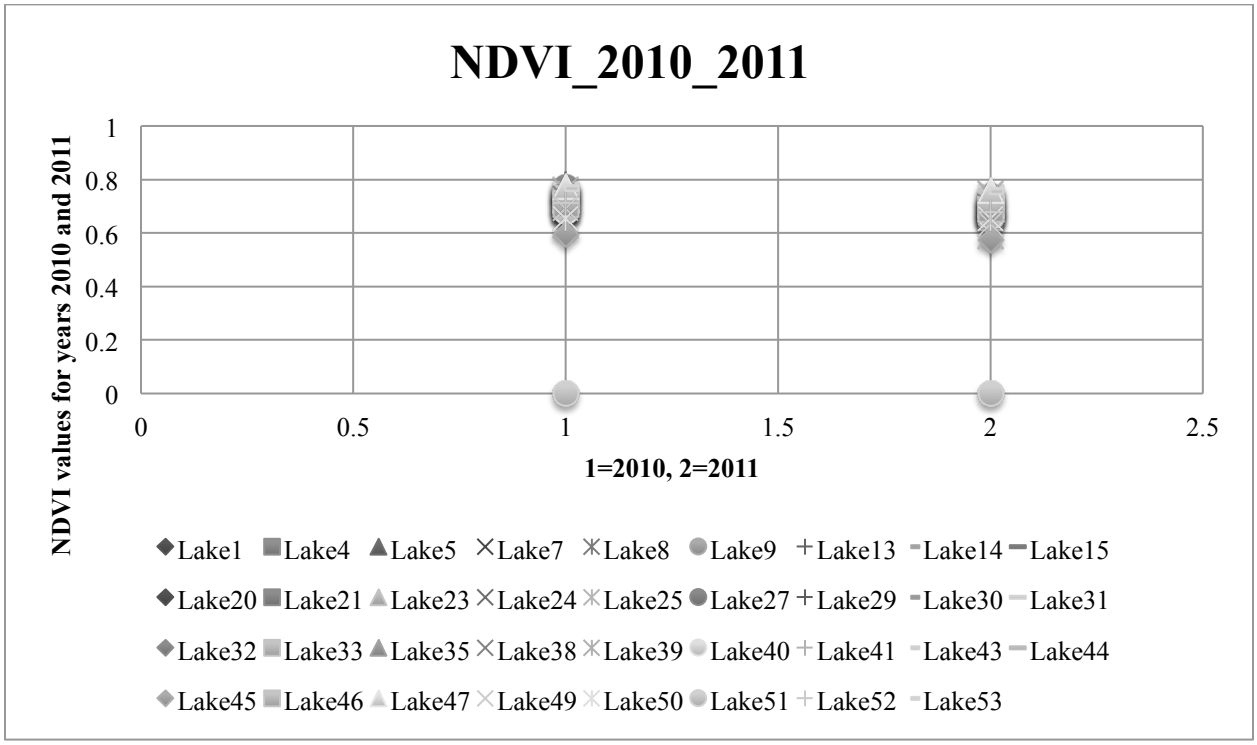


Figure 30. Comparison of NDVI data for years 2010 and 2011

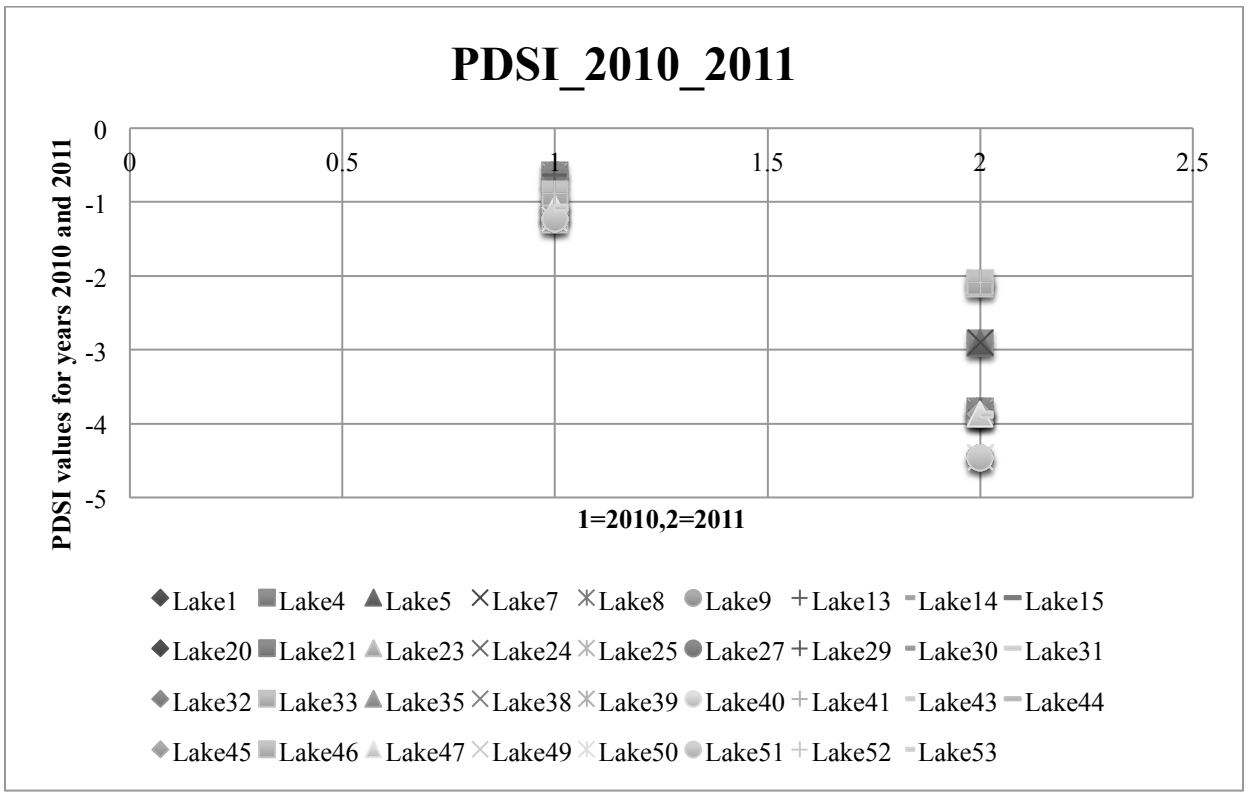
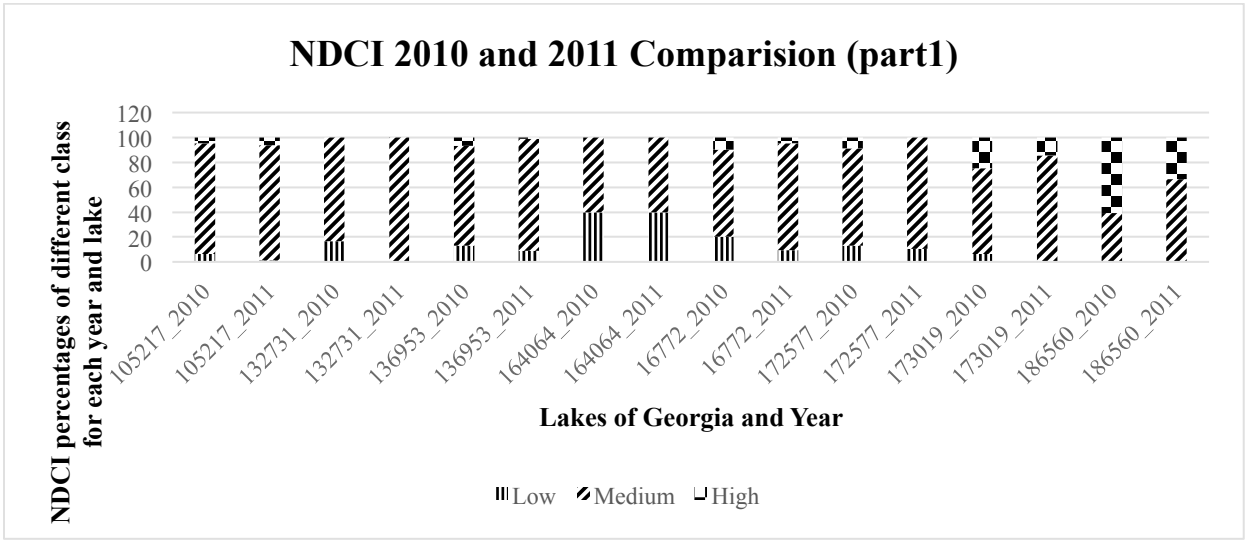
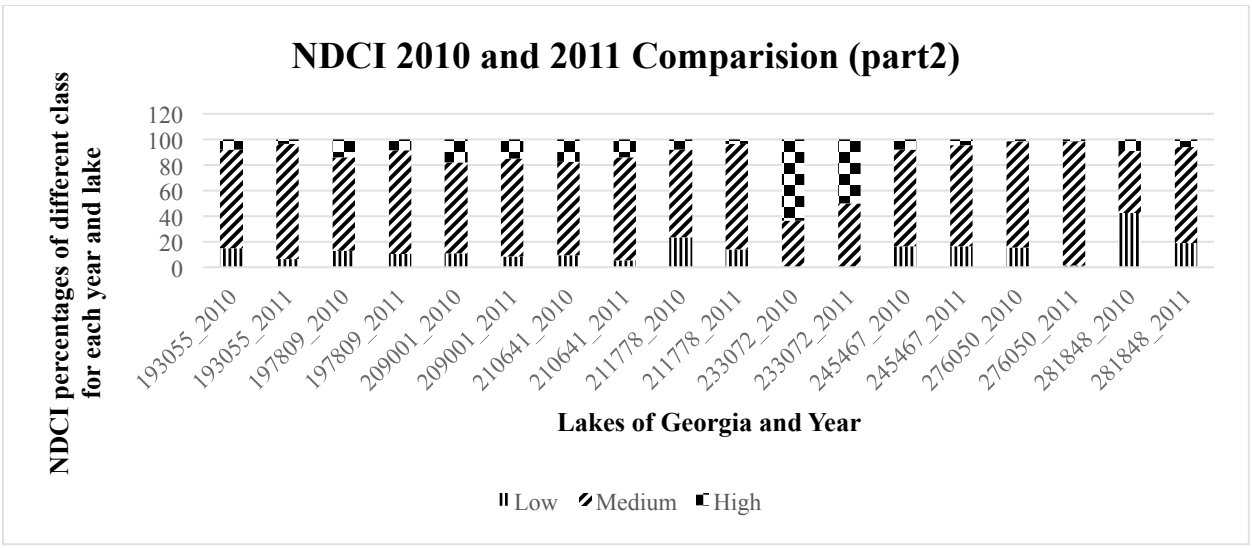
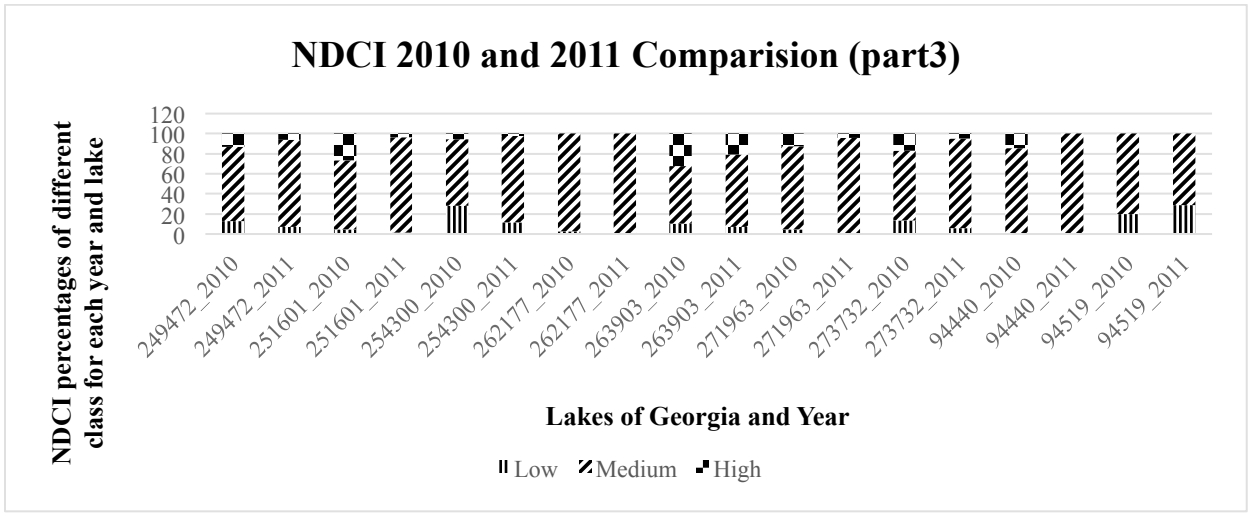


Figure 31. Comparison of PDSI data for years 2010 and 2011

According to the PDSI data, in 2011 there is more drought compared to 2010. Water level reduces due to severe drought and nutrient concentration increases which may cause the increase of chlorophyll concentration in lakes for year 2011.







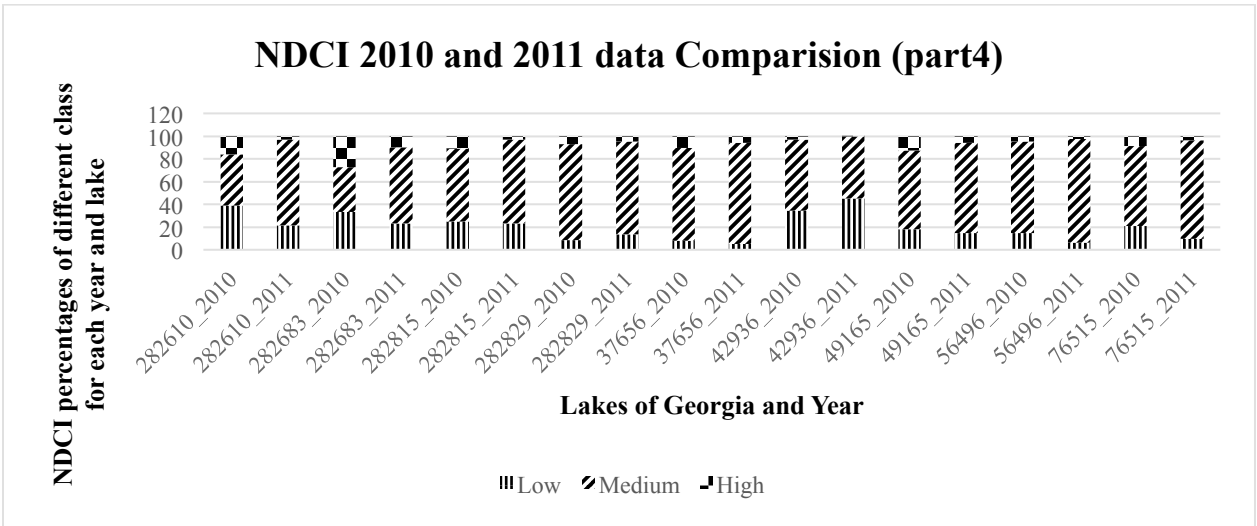


Figure 32. Comparison of NDCI data of lakes for years 2010 and 2011

In figure 32, we show the comparison of NDCI data of all 35 lakes for years 2010 and 2011, we can observe that there is increase in medium chlorophyll concentration from 2010 to 2011 for almost all lakes. As mentioned earlier there is huge difference in PDSI data between 2 years compared to other factors so, drought conditions tend to increase in chlorophyll growth and PDSI data has more indicative of change in chlorophyll concentration.

CHAPTER 6

CONCLUSION

We have performed detailed analysis on how data analytics can be used for detection of chlorophyll-a concentration and also analyzed how physical parameters affect chlorophyll-a concentration. We have used Land Usage and Land Cover, Normalized Difference Vegetation Index, and Palmer Drought Severity Index and Normalized Difference Chlorophyll Index data as metrics for our analysis. We have studied how to extract data from different spectral bands captured by MERIS satellite and also studied about how to process spectral bands pixels for chlorophyll-a concentration with the help of existing techniques. We are able to extract pixel counts for different classes of chlorophyll concentration by eliminating noise by managing the challenges faced due to the inability of MERIS to capture all the lakes of Georgia. Though we successfully extracted data by overcoming the challenges faced due to resolution limitations and noise introduced in MERIS scenes, we could not extract data for all 492 lakes. We limited our analysis to 53 lakes which is very less compared to what we proposed.

We also performed analysis on the NDCI data using K-Means clustering algorithm for years 2010 and 2011 by using silhouette coefficient to determine optimal clusters. Using the clusters formed from K-Means we compared chlorophyll concentrations and physical parameters data to identify the lakes which have similar trend in 2010 and 2011. Based on our analysis PDSI data affects the change in chlorophyll a concentration for lakes of Georgia in years 2010 and 2011.

REFERENCES

[1] From URI: https://www.nwf.org/Wildlife/Threats-to-Wildlife/Pollutants/Algal-Blooms.aspx

[2] Ross S. Lunetta, Blake A. Schaeffer, Richar P. Stumpf, Darryl Keith, Scott A. Jacobs, Mark S. Murphy (2015). Evaluation of cyanobacteria cell count detection derived from MERIS imagery across the eastern USA, Remote Sensing of Environment 157 (2015) 24-34

[3] Igor Ogashawara, Deepak R. Mishra, Sachidananda Mishra, Marcelo P. Curtarelli and Jose L. Stech (2013), A Performance Review of Reflectance Based Algorithms for Predicting Phycocyanin Concentrations in Inland Waters, Remote Sensing ISSN 2072-4292

[4] Kaylan Randolph, Jeff Wilson, Lenora Tedesco, Lin Li, D. Lani Pascual, Emmanuel Soyeux (2008), Hyperspectral remote sensing of cyanobacteria in turbid productive water using optically active pigments, chlorophyll a and phycocyanin, Remote Sensing of Environment 112 (2008) 4009-4019

[5] Andrew N Tyler, Peter D Hunter, Laurence Carvalho, Geoffrey A Codd, J Alex Elliott, Claire A Ferguson, Nick D Hanley, David W Hopkins, Stephen C Maberly, Kathryn J Mearns and E Marion Scott (2009), Strategies for monitoring and managing mass populations of toxic cyanobacteria in recreational waters: a multi-interdisciplinary approach, Environmental Health 2009, 8 (Suppl I): SII doi: 10.1186/1476-069X-8-SI-SII

[6] I. R. Falconer, and A. R. Humpage (2005). Health Risk Assessment of Cyanobacterial (Blue-green Algal) Toxins in Drinking Water, Int. J. Environ. Res. Public Health, 2(1), 43-50

[7] Tiit Kutser, Liisa Metsamaa, Niklas Stormbeck, and Ele Vahtmae (2006), Monitoring cyanobacterial blooms by satellite remote sensing, Estuarine, Coastal and Shelf Science 67 (2006) 303-312

[8] Hunter, Peter D, Tyler, Andrew N, L. Carvalho, Laurence, Codd Geoffrey A., Maberly, Stephen C. (2010), Hyperspectral remote sensing of cyanobacterial pigments as indicators for cell population and toxins in eutrophic lakes, Remote Sensing of Environment, 114.2705-2718. 10.1016/j.rse.2010.06.06

[9] From URI: www.fondriest.com/environmental-measurements/parameters/water-quality/algae-phytoplankton-chlorophyll/

[10] Kim, D. K., Jeong, K. S., McKay, R.I.B., Chon, T.S. and Joo, G.J. (2012), Machine Learning for Predictive Management: Short and Long Term Prediction of Phytoplankton Biomass using Genetic Algorithm Based Recurrent Neural Networks, Int. J. Environ. Res., 6(1): 95-108, Winter 2012, ISSN: 1735-6865

[11] From URI: http://uotechnology.edu.iq/appsciences/Laser/Lacture_laser/thrid_class/Remote_Sensing/3-Remote_Sensing.pdf

[12] M Govendar, K Chetty, and H Bulcock (2007), A review of hyperspectral remote sensing and its application in vegetation and water resource studies, ISSN 0378-4738

[13] From URI: https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/envisat/instruments /meris/design

[14] From URI: http://modis.gsfc.nasa.gov/about/

[15] From URI: https://en.wikipedia.org/wiki/ArcGIS

[16] From URI: https://doc.arcgis.com

[17] From URI: https://en.wikipedia.org/

[18] From URI: http://www.brockmann-consult.de/cms/web/beam

[19] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schroedl (2001), Constrained K-means Clustering with Background Knowledge, Proceedings of the Eighteenth International Conference on Machine Learning, 2001, p. 577-584

[20] Tapas Kanungo, David M. Mount, Nathan S. Nethanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu (2002), An Efficient k-Means Clustering Algorithm: Analysis and Implementation, IEEECS Log Number 111599

[21] From URI: http://scikit-learn.org/

[22] Ingrid Chorus and Jamie Bartram, (1999), Toxic Cyanobacteria in Water: A guide to their public health consequences, monitoring and management ISBN 0-419-23930-8

[23] A.A. Gitelson, Y.Z. Yacobi, D.C. Rundquist, R. Stark, L. Han, and D. Etzion, Remote estimation of chlorophyll concentration in productive waters: Principals, algorithm development and validation.

[24] Sachidananda Mishra, Deepak R. Mishra (2011). Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters, Remote Sensing of Environment 117 (2012) 394-406

[25] Mishra, S., D. R. Mishra, and W. Schluchter (2009). A novel algorithm for predicting phycocyanin concentrations in Cyanobacteria: A proximal hyperspectral remote sensing approach, Remote Sensing, 1, 758-775; doi:10.3390/rs1040758

[26] From URI: http://www.yale.edu/ceo/Documentation/MODIS_data.pdf

[27] From URI: http://r-gis.net/?q=ModisDownload

[28] From URI: http://modis-land.gsfc.nasa.gov/MODLAND_grid.html

[29] From URI: ftp://ftp.ncdc.noaa.gov/pub/data/cirs/drd/divisional.README

[30] From URI: https://wdc.dlr.de/sensors/meris/

[31] Cyanotracker (2015). "Cyanotracker-The University of Georgia." 2015, from http://cyanotracker.uga.edu/