International Conference on Hydroinformatics

2014

# Monitoring Spatiotemporal Total Organic Carbon Concentrations In Lake Mead With Integrated Data Fusion And Mining (IDFM) Technique

Sanaz Imen

Ni-Bin Chang

Y. Jeffrey Yang

# MONITORING SPATIOTEMPORAL TOTAL ORGANIC CARBON CONCENTRATIONS IN LAKE MEAD WITH INTEGRATED DATA FUSION AND MINING (IDFM) TECHNIQUE

SANAZ IMEN (1), NI-BIN CHANG (2), Y.JEFFREY YANG (3)
*(1), (2): Department of civil, environmental, and construction engineering, University of Central Florida, Orlando, FL, USA*
*(3): Water Supply and Water Resources Division, Water Quality Management Branch, ORD/NRMRL, U.S. EPA, Cincinnati, OH, USA*

Forest fires, soil erosion, and land use changes in watersheds nearby Lake Mead and inflows from Las Vegas Wash into the lake are considered as sources of the lake's water quality impairment. These conditions result in higher concentration of Total Organic Carbon (TOC). TOC in contact with Chlorine which is often used for disinfection purposes of drinking water supply causes the formation of trihalomethanes (THMs). THM is one of the toxic carcinogens controlled by the EPA's disinfection by-product rule. As a result of the threat posed to drinking water used by the 25 million people downstream, recreational area, and wildlife habitat of Lake Mead, it is necessary to develop a method for near real-time monitoring of TOC in this area. Monitoring through a limited number of ground-based monitoring stations on a weekly/monthly basis is insufficient to capture both spatial and temporal variations of water quality changes. In this study, remote sensing technology with the aid of data fusion and mining techniques provides us with information about the spatiotemporal distribution of TOC for the entire lake on a daily basis. A data fusion method was applied to bridge the gap of poor 250/500m spatial resolution for the land bands of Moderate Resolution Imaging Spectroradiometer (MODIS) imageries with the 30 m enhanced spatial resolution of Landsat's imageries which suffers from long overpass of 16 days. Consequently, Integrated Data Fusion and Mining (IDFM) techniques produce synthetic fused images of MODIS and Landsat satellites with both high spatial and temporal resolution to create near real time TOC distribution maps and lead to sustainable water quality management with the aid of IDFM in Lake Mead watershed.

**Keywords:** Remote Sensing, Data Fusion, Data Mining, Total Organic Carbon, Lake Mead.

## INTRODUCTION

Disinfection-By-Products (DBPs) are formed when Natural Organic Materials (NOMs) in surface water come into contact with chlorine. Exposure to DBPs for an extended period or at high levels increases the risk of developing bladder and colon cancer. One of the most important types of DBPs detected in drinking water is trihalomethanes (THMs), Nikolaou et al. [1]. In 1979, standard for THMs under the safe drinking water act was instituted by the EPA at 100 µg.L$^{-1}$; in 1998 it was reduced to 80 µg.L$^{-1}$, Simpson and Hayes [2]. To detect THMs, the most commonly used parameters are total organic carbon (TOC), dissolved organic carbon (DOC), and UV absorbance at 254nm wavelength.

EPA [3] indicated that THM level is probable to exceed 100 µg.L$^{-1}$ in case that levels of TOC are greater than 4 mg.L$^{-1}$ and the residence time in the network is 2-3 days. Moreover, if TOC level is detected between 2 and 4 mg.L$^{-1}$, more assessment is required to determine the effect of TOC concentration on formation of THMs. It is suggested that prior to performing primary disinfection, the levels of TOC be less than 2 mg.L$^{-1}$.

At present, the ground truth TOC measurement is limited in space and time. However, satellites provide us with precise and timely images. Remote sensing algorithms indicate spatial distribution of TOC for the surface water using surface reflectance band data detected by sensors. To find the relationship between surface reflectance band data and TOC, we need to have information about the spectral reflectance peaks of TOC on the electromagnetic spectrum, Chang and Vannah [4].

There have been only a few studies that were focused on finding the spectral reflectance peaks of TOC, although there have been several studies focused on finding a relationship between spectral reflectance band data and other water quality parameters such as Chromophoric Dissolved Organic Matter (CDOM) which are significantly related to TOC. Chang et al. [5] indicated that the first, second, and third bands of synthetic fused images of MODIS and Landsat, with spectral ranges between 459-900 nm, had the most contribution in determining TOC concentration. The earlier studies focused on relationship between COD and spectral reflectance indicate a broad peak on the reflectance spectrum at 550-570 nm and 650-700 nm for COD, Menken et al. [6];Arenz et al. [7]; Vertucci and Linkens [8]; Ficek et al. [9]. Comparing the CDOM spectral peaks with MODIS Terra and Landsat 5 TM/7 ETM+ band widths shows that band 2 of Landsat and band 4 of MODIS catch the peak of 550-570 nm. However, for peak of 650-700, only band 3 of Landsat captures the peak and band 1 of MODIS is not able to catch the end of the peak.

Although MODIS has high temporal resolution, it suffers from the low spatial resolution. This issue can be solved by developing a method to fuse MODIS images with Landsat images which has high spatial resolution of 30 m, even if it has lengthy revisit time of 16 days. Therefore, the main objective of the current study is applying the Integrated Data Fusion and Mining (IDFM) techniques developed by Chang and Vannah [4] to produce near real time TOC concentration maps to access the near real time monitoring system for drinking water in Lake Mead.

## Study area

Lake Mead, the largest reservoir in United States with maximum capacity of 28 million acre-feet (35 km$^3$) was selected for this study. It is located on the border of Nevada and Arizona (Figure 1). This lake supplies drinking water for more than 25 million people, irrigation water of more than 2.5 million acres, and recreation opportunities for more than 8 million people. Moreover, it provides electricity for major cities such as Las Vegas, Phoenix, Los Angles, Tucson, and San Diego. Furthermore, it is considered as an important habitat for the federally listed endangered fish and wildlife species.

According to water quality report of Las Vegas Valley water district in 2013, testing the water quality of treated water supply in Las Vegas valley water district distribution system in 2012 indicated increase in maximum concentration of THMLs to 84 mg.L$^{-1}$, which is higher than the maximum concentration limit (MCL) of this parameter (80 mg.L$^{-1}$). Based on the revised EPA regulation in 2012, maximum level greater than MCL is allowable as long as neither running annual average nor locational running annual average exceeds MCL. Therefore, this issue necessitates monitoring of TOC level in Lake Mead, since this parameter in contact with Chlorine, which is often used for disinfection purposes of drinking water supply, causes the formation of THMs.

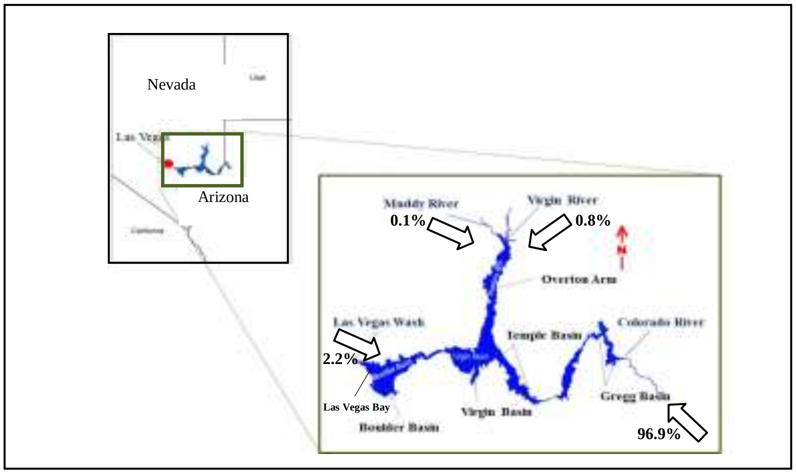

Figure1. Location of Lake Mead and percentages of inflow into the lake

MODIS Terra and Landsat 5 TM/7 ETM+ were processed in ArcGIS. The Spatial and Temporal Adaptive Reflectance Fusion Model (STAR-FM) were applied to fuse Landsat and MODIS images. As a next step, from fused images the spectral reflectance values of different bands at the location of ground-truth data for the sampling days were extracted, and applied as inputs into machine-learning program. The artificial neural network toolbox in MATLAB was used to determine the linkage between the spectral reflectance values of different bands and ground-truth data. Finally, the TOC concentration maps were developed.

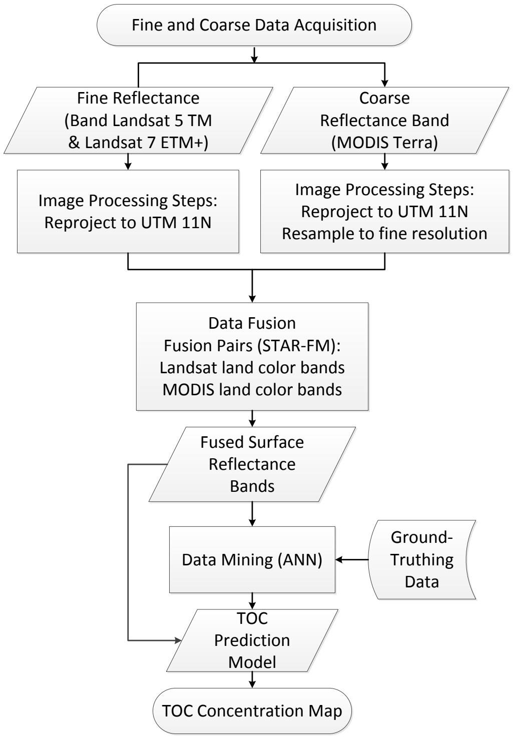
Fine and Coarse Data Acquisition
Fine Reflectance (Band Landsat 5 TM & Landsat 7 ETM+)
Coarse Reflectance Band (MODIS Terra)
Image Processing Steps: Reproject to UTM 11N
Image Processing Steps: Reproject to UTM 11N Resample to fine resolution
Data Fusion Fusion Pairs (STAR-FM): Landsat land color bands MODIS land color bands
Fused Surface Reflectance Bands
Data Mining (ANN)
Ground-Truthing Data
TOC Prediction Model
TOC Concentration Map

Figure 2. Analytical framework of study

## 1. Data acquisition

1250 ground-truth data were collected from 26 sampling stations of four different agencies including: 1) southern Nevada water system, 2) city of Las Vegas (water pollution control facility), 3) southern Nevada water system-city of Las Vegas, 4) southern Nevada water system-Clark county water reclamation district.

MODIS Terra surface reflectance images were obtained from USGS MODIS reprojection tool web interface (MRTWeb) (https://mrtweb.cr.usgs.gov/). It should be mentioned that for MODIS images, all cloud free data were downloaded consistent with the collection dates of ground-truth data. In addition, Landsat 5TM and Landsat 7ETM+ images were obtained from the United States Geological Survey (USGS) site (http://earthexplorer.usgs.gov/). All data collected for the period of 2000-2013. In Table 1, number of downloaded images, and numbers of land surface reflectance bands are presented.

Table 1. Number of download satellite images

| Satellite | Band | No. of Downloaded Images |
|---|---|---|
| MODIS Terra | 1-4, 6-7 | 870 |
| Landsat 5TM | 1-5,7 | 164 |
| Landsat 7 ETM+ | 1-5,7 | 217 |

## 2. Image processing

Since the STAR-FM algorithms requires the same size of the bit-depth and spatial resolution of image pairs, the Landsat bit-depth was increased to 16-bit integer and MODIS images were resample to 30 m resolution. In addition, all images were reprojected to UTM zone 11N. All MODIS data are pre-processed at a level 2 basis. This includes the radiometrically calibrated data that were atmospherically corrected for aerosols and scattering. In addition, Landsat data came processed on a level-2 basis with full atmospheric corrections completed using MODIS 6S radiative transfer code.

## 3. Data Fusion

Although MODIS sensor can provide us with daily images, it has a coarse spatial resolution (250/500m). This issue can be solved by using Landsat images which has fine spatial resolution of 30m. But it has a lengthy revisit time of 16 days. Therefore, it is necessary to find a method to combine these sensors to get synthetic images which have both high spatial and temporal resolution. For this purpose, the STAR-FM algorithm were applied in this study. This algorithm was developed by the National Aeronautics and Space Administration (NASA) based on pixel level data fusion. The high temporal resolution feature of MODIS is fused by high spatial resolution feature of Landsat through STAR-FM to simulate fused images with reasonable spatial and temporal resolution features.

Available MODIS and Landsat images of Lake Mead from 6 May 2001 to 14 May 2001 are shown in Figure 3. As you can see in this figure, Landsat images were only available for two days during the selected time period. Therefore, daily MODIS images and Landsat images of these two days were applied to create fused images. For instance, to fill a gap of DOY 127, pre-condition synthetic image was generated by three images A, B, and J. Then, three images (i.e. B, I, and S) were applied to generate the post –condition synthetic image. Therefore, a single synthetic image (image k) was produced using the created post- and pre-condition images. It should be mentioned that only cloud free images can be applied in STAR-FM algorithm. Therefore, STAR-FM algorithm was not capable to generate fused images for cloudy days (i.e. DOY 132, 133).

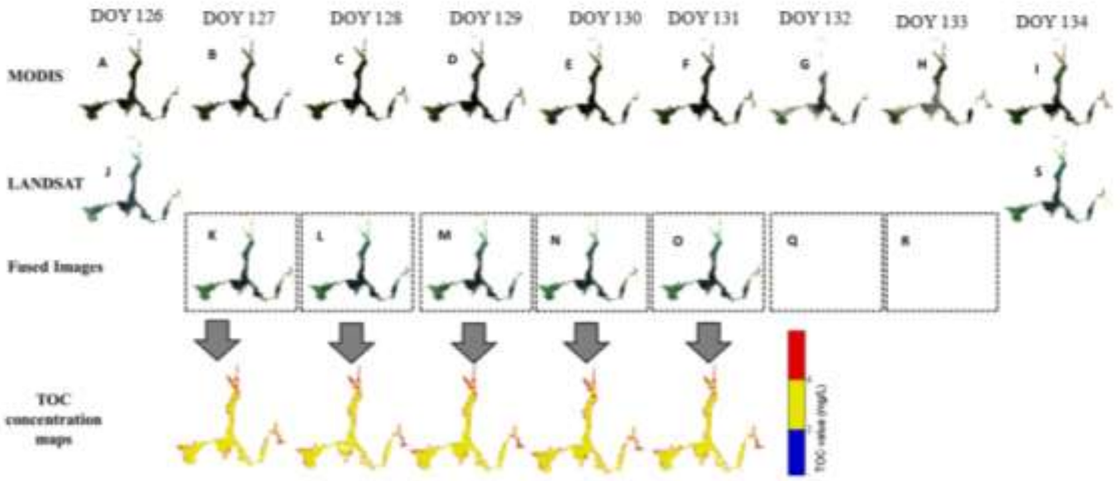

Figure 3. This figure indicates the ability of IDFM technique in gap-filling of Landsat images and predicting daily TOC concentration maps for cloud-free days. The top row is daily MODIS images from 6 May 2001 (DOY 126) to 14 May 2001(DOY 134), and the second row shows the available Landsat images for Lake Mead during the selected period. Third and the last rows show fused images and daily TOC concentration maps respectively which were created by surface reflectance value of fused images and ANN model.

## 4. Machine-Learning

Artificial Neural Network (ANN) was applied to identify the relationship between TOC and surface reflectance values. For this purpose, surface reflectance values extracted from synthetic images of only ground-truth dates were applied to train ANN model. 60% and 30% of ground-truth data were applied to train and validate the ANN model, respectively. In addition, 10% of ground-truth data were used to test the model. The MATLAB ANN toolbox with feed forward neural network and back propagation algorithm was applied. The selected ANN has two-layer with 30 hidden neurons.

Comparison between the predicted and observed TOC value is shown in Figure 4. Based on this figure, the squared correlation coefficient between observed and predicted TOC is equal to 50%. In addition, time series of observed and predicted TOC and their annual average values were compared in Figure 5. According to this figure, predicted TOC values follow the same trend as observed data. However, the model was not able to model peaks. The high concentration values of observed TOC were detected in August and October 2007 at Virgin and Muddy Rivers which may be caused by previous forest fires in this region ( Figures 4 and 5).

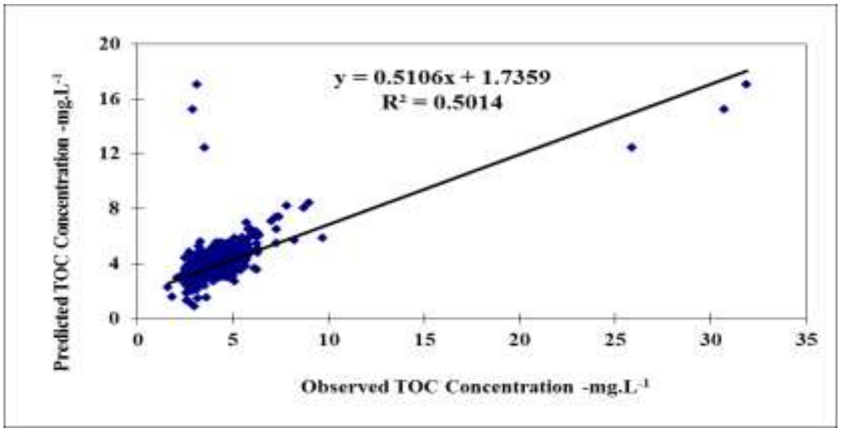

Figure 4. Correlation between observed and predicted TOC values

The error between observed and predicted values was calculated by root mean square error (RMSE) based on the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_{pi} - x_{oi})^2}{N}} \qquad (2)$$

Where $X_{oi}$ is observed values and $X_{pi}$ is predicted values, and N is number of ground truth data. RMSE close to zero shows the least difference between observed and predicted values. In this study, calculated RMSE was equal to 1.09.

## 5. Mapping TOC concentration

Unlike Genetic Programming model, it is not possible to get an explicit equation from ANN. The selected ANN, which provides us with highest squared correlation coefficient between TOC and reflectance values was applied to create TOC concentration maps. Using fused

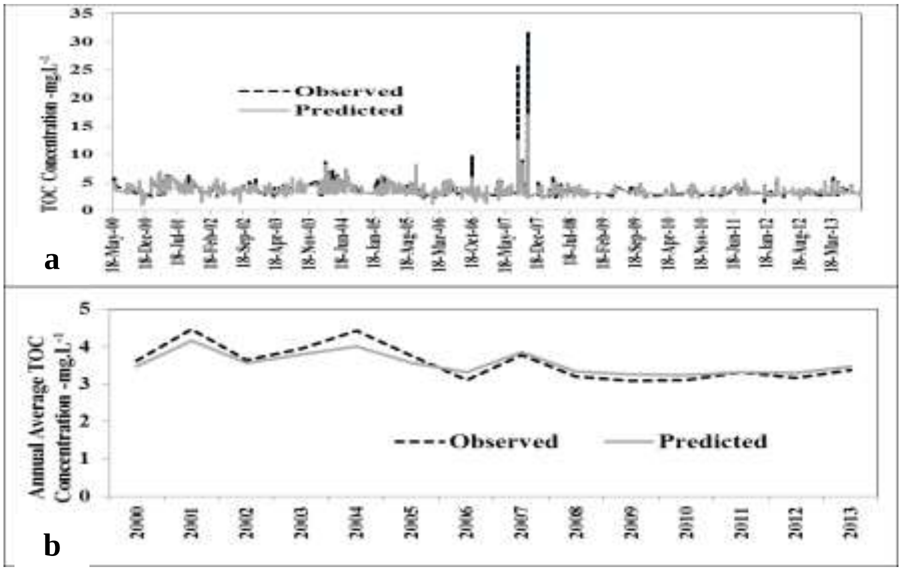

Figure 5. a) time series plot comparing the observed TOC values with predicted TOC; b) time series plot comparing the observed and predicted annual average TOC values

spectral reflectance values as inputs into the selected ANN model, corresponding TOC values were determined. This procedure was done pixel by pixel until we got the TOC concentration values of all pixels inside the Lake. Daily TOC concentration maps from 7 May 2011 to 11 May 2011 are shown in Figure 3. Based on this figure, TOC concentration values were classified into three different categories (i.e. blue: good (<2), yellow: fair (between 2 and 4), red: poor (>4)) defined by EPA [3]. Each pixel of TOC concentration maps has a same resolution as Landsat images (30×30 m). In all developed TOC concentration maps in figure 3, TOC concentration value is uniform through the lake and it varies between 2-4 mg.L$^{-1}$.

To assess the performance of IDFM method, TOC concentration maps were compared with ground truth data of corresponding dates. For instance, five ground truth data collected from five sites (i.e. CR346.4, LVB2.7, LVB6.7, LWLVB, and LWLVB-B) are available on 8 May 2001. Ground truthing concentration and location were shown in Table 2. Based on this table, the observed TOC values at sampling locations located in Boulder Basin and Las Vegas Bay (near mid channel) on 8 May 2001 are about 3 mg.L$^{-1}$. From figure 3, it can be seen that the TOC values at the same location have been predicted in the range of 2-4 mg.L$^{-1}$. Furthermore, the observed TOC values at sampling locations located in Las Vegas Bay and Las Vegas Wash are about 6 mg.L$^{-1}$ which are in consistent with predicted TOC ranges (>4 mg.L$^{-1}$) at the same location. From Figure 3, the highest concentration of TOC was observed at the inflows of the Muddy River, the Virgin River, the Colorado River, and the Las Vegas wash into Lake Mead. This issue might be caused by recorded forest fires, soil erosion, and land use changes upstream of Muddy and Virgin rivers, runoff from Las Vegas metropolitan area and wastewater treatment facilities, and effects of ongoing drought on Colorado River.

## Results and discussion

The integration of data fusion and data mining techniques were applied to develop a near real time TOC monitoring system. ANN model was used as a primary part of data mining technique, and the spectral reflectance values were applied to train the model. The squared correlation coefficient between predicted and observed TOC concentration values, determined by ANN model, was about 50%. In addition, the error between the observed and predicted concentration values showed the ability of model in predicting the trend and its inability in predicting the high concentration values. To create TOC concentration maps for the specific date, fused spectral reflectance values of different bands of each pixel were used as inputs into ANN model

Table 2. Ground truthing concentration and location on 8 May 2001 (location of Las Vegas Wash , Las Vegas Bay, and Boulder Basin are shown in Figure 1) values was about one which shows the moderate precision of prediction. Comparing trend of observed and predicted TOC

| Site | TOC (mg.L$^{-1}$) | Location |
|---|---|---|
| CR346.4 | 3.4 | Boulder Basin, close to Hoover Dam |
| LVB2.7 | 6.3 | Las Vegas Bay-near Las Vegas Wash |
| LVB6.7 | 2.8 | Las Vegas Bay-near mid channel |
| LWLVB | 6.2 | Las Vegas Wash |
| LWLVB-B | 6.2 | Las Vegas Wash |

to determine the corresponding TOC values. Comparing the results of TOC concentration maps with ground-truth data at the same day shows that the accuracy of model in predicting the TOC values. It should be mentioned that TOC concentration maps were only prepared for cloud-free days, and the model is not capable to produce TOC concentration maps for cloudy days. These near real time TOC concentration maps can be considered as an applicable tool in developing a strategy to reduce THM which is posed as a threat to drinking water of 25 million people, recreationl area, and wildlife habitat of Lake Mead.

**Acknowledgments**

The authors acknowledge the help from Ms. Susan Holmes in SNWA for providing us with lower quality river water quality database. This study is funded through EPA contract WA 1-52. Support from the USEPA is gratefully acknowledged.

## REFERENCES

[1] Nikolaou A.D., Kostopoulou M.N., Lekkas T.D., "Organic by-products of drinking water chlorination", *Glob Nest: the Int J.*, Vol.1, No. 3, (1999), pp. 143-156.

[2] Simpson K.L., Hayes K.P., "Drinking water disinfection by-products: an Australian perspective", Wat. Res., Vol. 32, No. 5, (1998), pp. 1522-1528.

[3] EPA Office of Environmental Enforcement, "EPA drinking water guidance on disinfection by-products advice note No.4 version 2. Disinfection by-products in drinking water", ISBN 978-1-84095-444-9, p.27.

[4] Chang N.B., Vannah B., "Monitoring the total organic carbon concentrations in a lake with the integrated data fusion and machine-learning (IDFM) technique", SPIE Proceedings, Vol. 8513, (2012), 851307-1.

[5] Chang N.B., Vannah B.W., Yang Y.J., Elovitz M., "Integrated data fusion and mining techniques for total organic carbon concentrations in lake", Integrated Journal of Remote Sensing, 35(3), pp. 1064-1093.

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

[7] Arenz, R., Lewis, W., and Saunders, J., "Determination of Chlorophyll and Dissolved Organic Carbon from Reflectance Data for Colorado Reservoirs", *International Journal of Remote Sensing*, Vol. 17, No. 8, (1996), pp. 1547-1566.

[8] Vertucci, F., Likens G.E., "Spectral Reflectance and Water Quality of Adirondack Mountain Region Lakes", *Limnology and Oceanography*, Vol. 34, No.8, (1989). Pp.1656-1672.

[9] Ficek D., Zapadka T., Dera J., "Remote Sensing Reflectance of Pomeranian Lakes and the Baltic", *Oceanologia*, Vol. 53, No. 4, (2011), pp. 959-970.