

SURFACE WATER QUALITY ASSESSMENT USING A REMOTE SENSING, GIS, AND MATHEMATICAL MODELLING FRAMEWORK

ESSAM HELMY MAHFOUZ SHARAF EL DIN

May 2018



**TECHNICAL REPORT
NO. 313**

**SURFACE WATER QUALITY
ASSESSMENT USING A REMOTE
SENSING, GIS, AND MATHEMATICAL
MODELLING FRAMEWORK**

Essam Helmy Mahfouz Sharaf El Din

Department of Geodesy and Geomatics Engineering
University of New Brunswick
P.O. Box 4400
Fredericton, N.B.
Canada
E3B 5A3

May 2018

© Essam Helmy Mahfouz Sharaf El Din, 2018

PREFACE

This technical report is a reproduction of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Geodesy and Geomatics Engineering, May 2018. The research was supervised by Dr. Yun Zhang, and funding was provided by the Egyptian Ministry of Higher Education and Scientific Research, Bureau of Cultural & Educational Affairs of Egypt in Canada, and the Canada Research Chairs program.

As with any copyrighted material, permission to reprint or quote extensively from this report must be received from the author. The citation to this work should appear as follows:

Sharaf El Din, Essam Helmy Mahfouz (2018). *Surface Water Quality Assessment Using A Remote Sensing, GIS, and Mathematical Modelling Framework*. Ph.D. dissertation, Department of Geodesy and Geomatics Engineering, Technical Report No. 313, University of New Brunswick, Fredericton, New Brunswick, Canada, 172 pp.

ABSTRACT

The presence of various pollutants in water bodies can lead to the deterioration of both surface water quality and aquatic life. Surface water quality researchers are confronted with significant challenges to properly assess surface water quality in order to provide an appropriate treatment to water bodies in a cost-effective manner. Conventional surface water quality assessment methods are widely performed using laboratory analysis, which are labour intensive, costly, and time consuming. Moreover, these methods can only provide individual concentrations of surface water quality parameters (SWQPs), measured at monitoring stations and shown in a discrete point format, which are difficult for decision-makers to understand without providing the overall patterns of surface water quality.

In contrast, remote sensing has shown significant benefits over conventional methods because of its low cost, spatial continuity, and temporal consistency. Thus, exploring the potential of using remotely sensed data for surface water quality assessment is important for improving the efficiency of surface water quality evaluation and water body treatment.

In order to properly assess surface water quality from satellite imagery, the relationship between satellite multi-spectral data and concentrations of SWQPs should be modelled. Moreover, to make the process accessible to decision-makers, it is important to extract the accurate surface water quality levels from surface water quality raw data. Additionally, to improve the cost effectiveness of surface water body treatment, identifying the major pollution sources (i.e., SWQPs) that negatively influence water bodies is essential.

Therefore, this PhD dissertation focuses on the development of new techniques for (1) estimating the concentrations of both optical and non-optical SWQPs from a recently launched earth observation satellite (i.e., Landsat 8), which is freely available and has the potential to support coastal studies, (2) mapping the complex relationship between satellite multi-spectral signatures and concentrations of SWQPs, (3) simplifying the expression of surface water quality and delineating the accurate levels of surface water quality in water bodies, and (4) classifying the most significant SWQPs that contribute to both spatial and temporal variations of surface water quality.

The outcome of this PhD dissertation proved the feasibility of developing models to retrieve the concentrations of both optical and non-optical SWQPs from satellite imagery with highly accurate estimations. It exhibited the potential of using remote sensing to achieve routine water quality monitoring. Moreover, this research demonstrated the possibility of improving the accuracy of surface water quality level extraction with inexpensive implementation cost. Finally, this research showed the capability of using satellite data to provide continuously updated information about surface water quality, which can support the process of water body treatment and lead to effective savings and proper utilization of surface water resources.

DEDICATION

TO MY MOTHER, FATHER, AND SISTER

TO MY WIFE,

and

TO MY CHILDREN

FOR THEIR SUPPORT, ENCOURAGEMENT, AND INSPIRATION

ACKNOWLEDGEMENTS

I would like to take this opportunity to extend sincere thanks to those people who made this work achievable. First, I would like to thank my supervisor, Prof. Dr. Yun Zhang, for both his guidance and encouragement throughout my PhD program. He provided me many opportunities for teaching, attending conferences, and other academic activities. I am very grateful to Prof. Dr. David Coleman, Prof. Dr. Katy A. Haralampides, Prof. Dr. Ian Church, and Prof. Dr. Quazi K. Hassan for their review of my PhD dissertation. I would like to thank both the United States Geological Survey (USGS) and the Province of New Brunswick for providing the Landsat 8 imagery and the surface water quality ground truth data, respectively.

Sincere thanks go to Prof. Dr. Katy A. Haralampides and Dr. Dennis Connor for their help in the field data collection (i.e., water sampling) and experimental work. I express thanks to the Egyptian Ministry of Higher Education and Scientific Research, Bureau of Cultural & Educational Affairs of Egypt in Canada, and the Canada Research Chair Program for funding this project. I would also like to thank the UNB Writing Center for the valuable comments on my journal papers and dissertation.

I am truly grateful to my mother, my father, and my sister for their support and encouragement from thousands of miles away. Finally, I would like to thank my wife, Dina Dawood, for her unconditional love, support, patience, understanding, and taking care of our children, Malek and Bayan. Your spiritual support is the source of my energy always!

Table of Contents

ABSTRACT	ii
DEDICATION	iiiv
PREFACE	v
ACKNOWLEDGEMENTS	vi
Table of Contents	vii
List of Tables	xiii
List of Figures	xv
List of Symbols, Nomenclature or Abbreviations	xix
Chapter 1: INTRODUCTION	1
1.1 Dissertation Structure	1
1.2 Background	2
1.3 Selected Research Topic	5
1.4 Problem Statement	6
1.4.1 Estimation of the Concentrations of SWQPs from Satellite imagery	6
1.4.2 Mapping the Relationship between Satellite Data and Concentrations of SWQPs	7
1.4.3 Extracting the Accurate Levels of Surface Water Quality Within a Water Body	8
1.4.4 Identifying the Major SWQPs Contributing to Spatio-temporal Surface Water Quality Variations	9
1.5 Research Objectives	10
1.5.1 Estimation of the Concentrations of SWQPs from Satellite imagery	10

1.5.2 Mapping the Relationship between Satellite Data and Concentrations of SWQPs.....	10
1.5.3 Extracting the Accurate Levels of Surface Water Quality Within a Water Body.....	11
1.5.4 Identifying the Major SWQPs Contributing to Spatio-temporal Surface Water Quality Variations	12
1.6 Overview of Each Chapter	12
Chapter 2: ESTIMATION OF BOTH OPTICAL AND NON-OPTICAL SURFACE WATER QUALITY PARAMETERS USING LANDSAT 8 OLI IMAGERY AND STATISTICAL TECHNIQUES	17
Abstract	17
2.1 Introduction	18
2.2 Materials and Methods	23
2.2.1 Selected Study Site	24
2.2.2 Satellite Processing Stage	25
2.2.2.1 Geometric Correction.....	25
2.2.2.2 Radiometric Correction.....	25
2.2.2.3 Atmospheric Correction	26
2.2.2.4 The Water Interface.....	29
2.2.3 Sampling Sites and Laboratory Analysis of SWQPs.....	30
2.2.4 Estimation of Concentrations of SWQPs using the Stepwise Regression Technique.....	32
2.3 Results and Discussion.....	34
2.3.1 Optical and Non-optical Concentrations of SWQPs of Water Samples	35
2.3.2 Relationship between Landsat 8 Satellite Spectral Data and Concentrations of SWQPs.....	37
2.3.3 Estimation and Validation of the Landsat 8-based-SWR Models	40

2.3.4 Landsat 8-based-SWR Spatial Distribution Maps	44
2.4 Conclusion.....	45
Acknowledgements.....	46
REFERENCES	47
Chapter 3: MAPPING CONCENTRATIONS OF SURFACE WATER QUALITY PARAMETERS USING A NOVEL REMOTE SENSING AND ARTIFICIAL INTELLIGENCE FRAMEWORK.....	52
Abstract	52
3.1 Introduction	53
3.2 Artificial Neural Network (ANN) Background	58
3.3 Materials and Methods	60
3.3.1 Remotely Sensed Data	60
3.3.1.1 Study Area.....	60
3.3.1.2 Satellite Processing Steps.....	61
3.3.2 In situ Measurements	62
3.3.3 Mapping Concentrations of SWQPs using the BPNN Algorithm.....	63
3.3.3.1 ANN Input and Output Selection	65
3.3.3.2 ANN Data Division.....	65
3.3.3.3 ANN Architecture Selection	66
3.3.3.4 ANN Structure Selection.....	66
3.3.3.5 ANN Training	67
3.3.3.6 ANN Evaluation.....	68
3.4 Results and Discussion.....	68
3.4.1 Concentration Results of both Optical and Non-optical SWQPs	69
3.4.2 Estimation and Validation of the Landsat 8-based-BPNN Models	70
3.4.2.1 ANN Input and Output Selection	70
3.4.2.2 ANN Data Division.....	71

3.4.2.3 ANN Architecture Selection	72
3.4.2.4 ANN Structure Selection.....	73
3.4.2.5 ANN Training and Evaluation	74
3.4.3 The Landsat 8-based-BPNN Spatial Concentration Maps.....	78
3.4.4 Comparison of Other Model Results	79
3.5 Conclusion.....	81
Acknowledgements.....	82
 Chapter 4: DELINEATING THE ACCURATE PATTERNS OF SURFACE WATER QUALITY BY INTEGRATING LANDSAT 8 OLI IMAGERY, ARTIFICIAL INTELLIGENCE, AND THE WATER QUALITY INDEX.....	87
Abstract	87
4.1 Introduction	88
4.2 Materials and Methods.....	92
4.2.1 Study Area	94
4.2.2 Landsat 8 Image Acquisition and Processing	94
4.2.3 Water Sampling and Laboratory Analysis.....	96
4.2.4 Estimation of Concentrations of SWQPs using the BPNN	98
4.2.5 Applying the CCMEWQI	102
4.3 Results and Discussion.....	105
4.3.1 Concentrations of Optical and Non-optical SWQPs.....	105
4.3.2 Training and Validation of the Proposed ANN	107
4.3.3 Extra Validation of the Developed Approach using Ground Truth Data ..	113
4.3.4 Temporal-spatial Distribution of the Selected SWQPs	118
4.3.5 Delineating the Accurate Levels of Surface Water Quality of the SJR.....	120
4.4 Conclusion.....	122
Acknowledgements.....	123

REFERENCES	123
Chapter 5: ASSESSMENT OF SPATIAL AND TEMPORAL SURFACE WATER QUALITY VARIATIONS USING MULTIVARIATE STATISTICAL TECHNIQUES: A CASE STUDY OF THE SAINT JOHN RIVER, CANADA	127
Abstract	127
5.1 Introduction	128
5.2 Materials and Methods	132
5.2.1 Study Area	132
5.2.2 Water Sampling and Physico-chemical Analysis	133
5.2.3 Multivariate Statistical Techniques.....	135
5.2.3.1 Principal Component Analysis/Factor Analysis (PCA/FA) Technique	135
5.2.3.2 Cluster Analysis (CA) Technique	136
5.2.3.3 Discriminant Analysis (DA) Technique.....	137
5.3 Results and Discussion.....	138
5.3.1 Physico-chemical Analysis of SWQPs	139
5.3.2 Multivariate Statistical Analysis	141
5.3.2.1 Principal Component Analysis/Factor Analysis (PCA/FA) Technique	141
5.3.2.2 Cluster Analysis (CA) Technique	145
5.3.2.3 Discriminant Analysis (DA) Technique.....	147
5.3.2.3.1 Spatial DA.....	147
5.3.2.3.2 Temporal DA	152
5.4 Conclusion.....	156
Acknowledgements.....	157
REFERENCES	157
Chapter 6: SUMMARY AND CONCLUSION	161
6.1 Summary of Research	161
6.2 Achievements of the Research.....	161

6.2.1 Developing the Landsat 8-based-SWR Technique for Estimating Concentrations of Optical and Non-optical SWQPs.....	162
6.2.2 Developing the Landsat 8-based-BPNN Framework for mapping Concentrations of SWQPs	163
6.2.3 Developing the Landsat 8-based-CCMEWQI Technique for Extracting the Accurate Levels of SWQPs	164
6.2.4 Categorizing Spatio-temporal Surface Water Quality Variations Using Multivariate Statistical Techniques.....	165
6.3 Recommendations for Future Work.....	166
Appendix I	168
Appendix II.....	169
Appendix III.....	168
Appendix IV.....	169
Appendix V.....	170
Curriculum Vitae	

List of Tables

Table 2.1 The correlation matrix of both optical and non-optical SWQPs.	37
Table 3.1 Statistics of the concentrations of SWQPs along the study site.	69
Table 3.2 The correlation coefficient (r) matrix of both optical and non-optical SWQPs.	70
Table 3.3 The r values between the Landsat 8 multi-spectral bands and concentrations of SWQPs.....	71
Table 3.4 Statistical measures between the target and actual concentrations of SWQPs using the developed Landsat 8-based-BPNN.	74
Table 3.5 Comparison of the BPNN and SVM statistical results.	81
Table 4.1 The CCME and WHO guidelines for drinking water quality.	103
Table 4.2 Descriptive statistics of the concentrations of SWQPs.....	106
Table 4.3 Correlation coefficient values between the Landsat 8 spectral data and the concentrations of SWQPs.	107
Table 5.1 Statistics of physico-chemical surface water quality parameters (SWQPs)...	139
Table 5.2 The correlation matrix for the measured SWQPs.	140
Table 5.3 The principal components (PCs) along with their respective eigenvalues and the percentage of variance.	143
Table 5.4 The loading values of SWQPs for the significant PCs.	144
Table 5.5 Wilks' lambda and chi-square test for discriminant analysis (DA) of spatial variation in surface water quality across four clusters (groups) of sites.	149
Table 5.6 Structure matrix along with variable scores for DA of Table 5.5.....	149

Table 5.7 Discriminant function coefficients for DA of Table 5.5.....	150
Table 5.8 Classification matrix for DA of Table 5.5.....	151
Table 5.9 Wilks' lambda and chi-square test for DA of temporal variation in surface water quality across four seasons.....	153
Table 5.10 Structure matrix along with variable scores for DA of Table 5.9.....	154
Table 5.11 Discriminant function coefficients for DA of Table 5.9.....	155
Table 5.12 Classification matrix for DA of Table 5.9.....	156

List of Figures

Figure 1.1 Structure of the dissertation.....	2
Figure 2.1 The flowchart of the proposed methodology.....	23
Figure 2.2 The selected study area of the Saint John River (SJR), New Brunswick, Canada (Earth Explorer, 2016)	24
Figure 2.3 (a) The original Landsat 8 satellite sub-scenes and (b) the atmospherically corrected Landsat 8 satellite sub-scenes using the Dark Object Subtraction (DOS) method	27
Figure 2.4 The water interface	29
Figure 2.5 The water sampling locations across the SJR, New Brunswick, Canada.....	30
Figure 2.6 Optical and non-optical concentrations of SWQPs at June 27 th 2015 (a), April 10 th 2016 (b), and May 12 th 2016 (c), respectively.....	35
Figure 2.7 The Landsat 8 estimation models for turbidity (a), TSS (b), COD (c), BOD (d), and DO (e) using the SWR technique based on calibration dataset	39
Figure 2.8 Statistics and accuracy measures between the measured and predicted concentrations of turbidity (a), TSS (b), COD (c), BOD (d), and DO (e) using the SWR technique based on calibration dataset.....	41
Figure 2.9 Statistics and accuracy measures between the measured and predicted concentrations of turbidity (a), TSS (b), COD (c), BOD (d), and DO (e) using the SWR technique based on validation dataset.....	43
Figure 2.10 Spatial concentration maps for turbidity (a), TSS (b), COD (c), BOD (d), and DO (e) generated from the developed Landsat 8-based-SWR approach	44

Figure 3.1 The flowchart of retrieving concentrations of different SWQPs from satellite data by using the proposed Landsat 8-based-BPNN.....	59
Figure 3.2 The selected study area of the SJR, New Brunswick, Canada (Earth Explorer, 2016)	61
Figure 3.3 The Landsat 8 satellite sub-scenes of the study area with sampling locations	63
Figure 3.4 The flowchart of applying the proposed BPNN algorithm	64
Figure 3.5 The architectural design of the proposed ANN.....	73
Figure 3.6 The Graphical fit results of turbidity ((a)(i), (a)(ii), and (a)(iii)), TSS ((b)(i), (b)(ii), and (b)(iii)), COD ((c)(i), (c)(ii), and (c)(iii)), BOD ((d)(i), (d)(ii), and (d)(iii)), and DO ((e)(i), (e)(ii), and (e)(iii)) for training, validation, and testing datasets of the developed Landsat 8-based-BPNN.....	75
Figure 3.7 Training, validation, and testing error curves of turbidity (a), TSS (b), COD (c), BOD (d), and DO (e).....	77
Figure 3.8 Spatial distribution maps of turbidity (a), TSS (b), COD (c), BOD (d), and DO (e) generated from the developed Landsat 8-based-BPNN	79
Figure 4.1 The flowchart of the proposed methodology.....	93
Figure 4.2 The selected study area of the Saint John River (SJR), New Brunswick, Canada (Google Maps, 2016)	94
Figure 4.3 The water profile and the sampling stations.....	97
Figure 4.4 The proposed artificial neural network (ANN) topology	101
Figure 4.5 Scatter plots of observed (measured) vs. modeled (predicted) concentrations of turbidity (a), TSS (b), TS (c), TDS (d), COD (e), BOD (f), DO (g), pH (h), EC (i), and temperature (j) using the training dataset.....	109

Figure 4.6 Scatter plots of observed (measured) vs. modeled (predicted) concentrations of turbidity (a), TSS (b), TS (c), TDS (d), COD (e), BOD (f), DO (g), pH (h), EC (i), and temperature (j) using the testing dataset	110
Figure 4.7 Error surfaces for turbidity (a), TSS (b), TS (c), TDS (d), COD (e), BOD (f), DO (g), pH (h), EC (i), and temperature (j) at the network training and testing phases.	112
Figure 4.8 The 1 st dataset of water samples used for further validation of the developed approach	114
Figure 4.9 The 2 nd dataset of water samples used for further validation of the developed approach	115
Figure 4.10 Scatter plots of observed vs. modeled concentrations of turbidity (a), TDS (b), DO (c), pH (d), EC (e), and temperature (f) using the 1 st dataset	117
Figure 4.11 Scatter plots of observed vs. modeled concentrations of turbidity (a), TDS (b), DO (c), pH (d), EC (e), and temperature (f) using the 2 nd dataset	118
Figure 4.12 Mapping the concentrations of turbidity (a), TSS (b), TS (c), TDS (d), COD (e), BOD (f), DO (g), pH (h), EC (i), and temperature (j) in the selected study area	119
Figure 4.13 Mapping the concentrations of turbidity (a), TSS (b), TS (c), TDS (d), COD (e), BOD (f), DO (g), pH (h), EC (i), and temperature (j) in the selected study area	121
Figure 5.1 The study area of the Saint John River (SJR), New Brunswick, Canada (Google Maps, 2016)	133
Figure 5.2 The collected water sampling stations	134
Figure 5.3 Scree plot of the produced PCs and their respective eigenvalues	142
Figure 5.4 Dendrogram showing hierarchical agglomerative CA of sampling stations.	146
Figure 5.5 Scatter plot for DA of spatial water quality variation across the four groups	148

Figure 5.6 Scatter plot for DA of temporal water quality variation across the four seasons
..... 152

List of Symbols, Nomenclature or Abbreviations

1T	- Terrain corrected
6S	- The second simulation of the satellite signal in the solar spectrum
ANN	- Artificial neural network
ATCOR	- Atmospheric and topographic correction
B	- Blue
BOD	- Biochemical oxygen demand
BPNN	- Back-propagation neural network
C	- Support vector machine penalty coefficient
CA	- Cluster analysis
CB	- Coastal Blue
CC	- Cascade correlation
CCME	- Canadian Council of Ministers of the Environment
CCMEWQI	- Canadian Council of Ministers of the Environment water quality index
COD	- Chemical oxygen demand
DA	- Discriminant analysis
DNs	- Digital numbers
DO	- Dissolved oxygen
DOS	- Dark Object Subtraction
EC	- Electrical conductivity
ETM+	- Enhanced Thematic Mapper Plus
G	- Green

GPS	- Global Positioning System
HELCOM	- Helsinki Commission water quality assessment
LM	- Levenberg-Marquardt
MERIS	- Medium Resolution Imaging Spectrometer
MLP	- Multi-layer perceptron
MODIS	- Moderate Resolution Imaging Spectroradiometer
n	- Number of samples
NDVI	- Normalized difference vegetation index
NDWI	- Normalized difference water index
NIR	- Near-infrared
NSFWQI	- National Sanitation Foundation Water Quality Index
OLI	- Operational Land Imager
OWQI	- Oregon Water Quality Index
PCA/FA	- Principal component analysis/factor analysis
PCs	- Principal components
pH	- Power of hydrogen
p-value	- Significant value
R	- Red
R ²	- Coefficient of determination
r	- Correlation coefficient
RMSE	- Root mean square error
RPD	- Residual prediction deviation
RS	- Remote sensing

SAR	- Synthetic Aperture Radar
SD	- Standard deviation
SeaWiFS	- Sea-viewing Wide Field-of-view Sensor
SNR	- Signal to noise ratio
SVM	- Support vector machine
SWIR1	- Shortwave infrared 1
SWIR2	- Shortwave infrared 2
SWQL	- Surface water quality level
SWQPs	- Surface water quality parameters
TDS	- Total dissolved solids
Temp	- Temperature
TIR1	- Thermal infrared 1
TIR2	- Thermal infrared 2
TM	- Thematic Mapper
TOA	- Top of atmospheric
TS	- Total solids
TSS	- Total suspended solids
Turb	- Turbidity
USGS	- The US Geological Survey
UTM	- The Universal Transverse Mercator
WGS 84	- The World Geodetic System 1984
WHO	- World Health Organization
WQ	- Water quality

- σ^2 - Support vector machine kernel function parameter
- ϵ - width of the insensitive loss function in support vector machine

Chapter 1: INTRODUCTION

This PhD dissertation focuses on the development of new methods that use Landsat 8 satellite data for assessing surface water quality of water bodies. It is an article-based PhD dissertation presented through the following journal papers.

Journal Paper 1 (Peer reviewed):

Sharaf El Din, E., & Zhang, Y. (2017). Estimation of both optical and non-optical surface water quality parameters using Landsat 8 OLI imagery and statistical techniques. *Journal of Applied Remote Sensing*, 11 (4), 046008 (2017), doi: 10.1117/1.JRS.11.046008.

Journal Paper 2 (Peer reviewed):

Sharaf El Din, E., Zhang, Y., & Suliman, A. (2017). Mapping concentrations of surface water quality parameters using a novel remote sensing and artificial intelligence framework. *International Journal of Remote Sensing*, 38 (4), pp. 1023-1042. <http://dx.doi.org/10.1080/01431161.2016.1275056>.

Journal Paper 3 (Peer reviewed):

Sharaf El Din, E., & Zhang, Y. (2018). Delineating the accurate patterns of surface water quality by integrating Landsat 8 OLI imagery, artificial intelligence, and the water quality index. *Remote Sensing of Environment*, under review.

A part of this work has been published in the “*International Archives of the Photogrammetry, Remote Sensing, and Spatial Information Sciences*”, XLII-4/W4, pp. 245-249, <https://doi.org/10.5194/isprs-archives-XLII-4-W4-245-2017>”.

Journal Paper 4 (Peer reviewed):

Sharaf El Din, E., & Zhang, Y. (2018). Assessment of spatio-temporal surface water quality variations using multivariate statistical techniques: a case study of the Saint John River, Canada. *Journal of the American Water Resources Association*, under review.

1.1 Dissertation Structure

This article-based dissertation includes six chapters. Chapter 1 provides the introduction of the research. The next four chapters (Chapter 2 to Chapter 5) present the four peer reviewed journal papers listed above, which are either published or submitted and under review. In each of the four papers, the first author conducted the primary

research, while the second author provided advice on the structure and the remaining authors provided minor input and assistance. Chapter 6 provides the summary and conclusion of this research. **Figure 1.1** illustrates the organization of this dissertation.

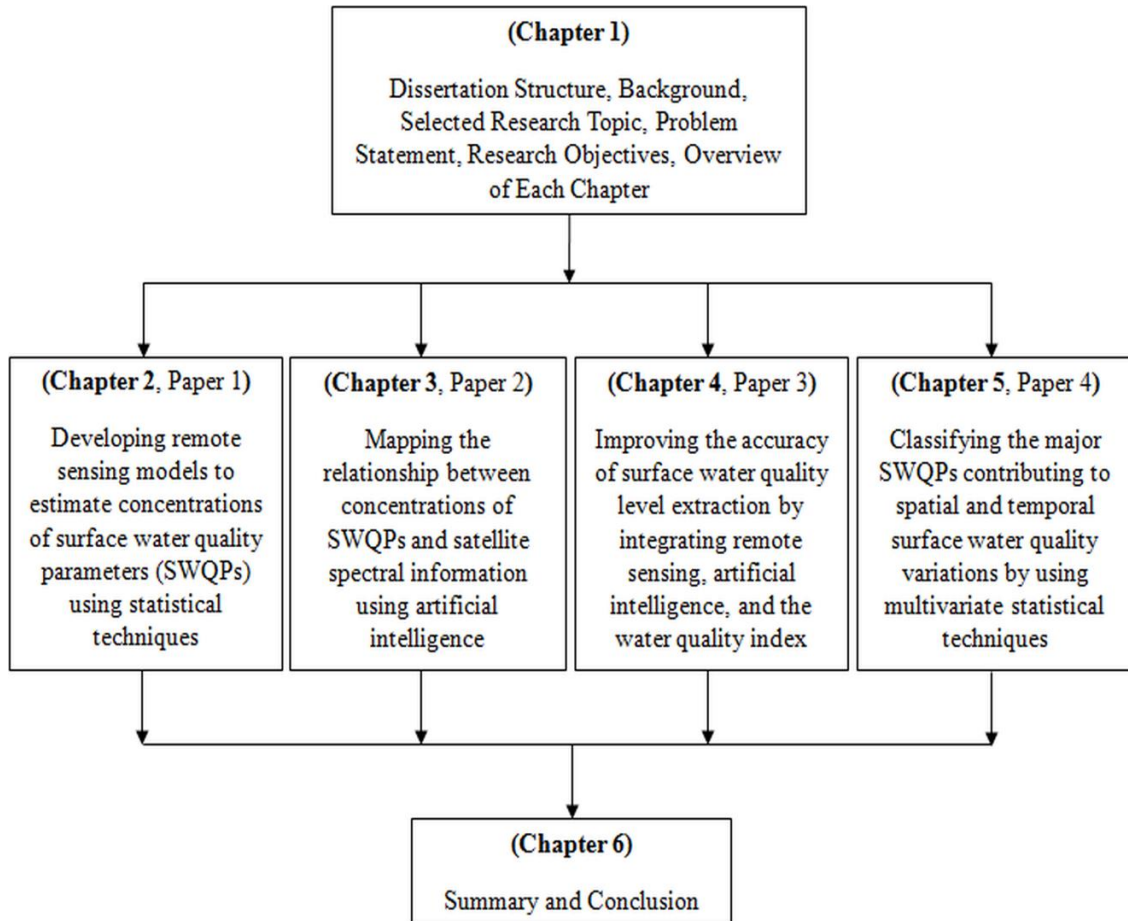


Figure 1.1 Structure of the dissertation

1.2 Background

Surface water quality is the measure of the state of water resources with respect to specific requirements and necessities, such as human needs. It refers to the physical, chemical, and biochemical characteristics of water (CCME, 2001). Surface water quality is very important in maintaining the ecological processes that conserve and support

biodiversity. However, deteriorating surface water quality due to natural (i.e., snow melt, precipitation rate, and sediment transport) and anthropogenic (i.e., urban, industrial, mining, and agricultural activities) processes threatens the stability of the biotic integrity and consequently the aquatic life (Carpenter, Caraco, Correll, Howarth, Sharples, & Smith, 1998; Qadir, Malik, & Husain, 2007).

In the past few decades, the increase of anthropogenic activities, especially in industrial areas, has negatively affected water bodies. The result can be a reduction in water storage capacity or in rivers' ability to support aquatic life. This shortage of water which has increased over the past years is expected to continue in the future (Gaballah, Khalaf, Beckand, & Lopez, 2005). In Canada, like in many countries around the world, the rising demand for safe drinking water directly corresponds to the rapid increase in population and in the economy (CCME, 2001). Thus, providing continuously updated information about surface water quality is indeed essential to help the managers, local administrators, and decision-makers in taking the right action at the right time to protect water bodies (Arseneault, 2008).

Conventional methods of assessing surface water quality of water bodies are limited to a set of in-situ water sampling points and laboratory analysis. These methods are time consuming and cost intensive, and only provide limited information in terms of spatial and temporal surface water quality aspects (Liu, Chin, Gong, & Fu, 2010). In order to properly analyze surface water quality within a water body, spatio-temporal aspects should be considered. Therefore, this dissertation presents research on the exploitation of remotely sensed data for assessing surface water quality and providing both spatial and temporal water quality variations.

Surface water quality assessment using remote sensing imagery is relatively inexpensive and can potentially offer consistent spatial and temporal measurements of surface water quality on a regular basis, which may help identify water bodies with significant surface water quality pollution problems. Remote sensing estimation of surface water quality is based on mapping the relationship between (1) remote sensing multi-spectral signatures and (2) measurements of ground truth data (i.e., concentrations of surface water quality parameters (SWQPs)); however, it is often critical to draw a theoretical expression for this relationship (Zhang, Pulliainen, Koponen, & Hallikainen, 2002).

First, remote sensing sensors are subjected to spatial, spectral, radiometric, and temporal resolution limitations. Spatial and spectral resolutions are often a trade-off with each other because of the sensor design and optical limitations. The data sensitivity (i.e., signal to noise ratio [SNR]) associated with the radiometric resolution can affect the accuracy of retrieving SWQPs (Gower & Borstad, 2004). Moreover, temporal resolution is often a concern for surface water quality assessment particularly for water bodies which are subjected to high dynamic variations. Many satellite sensors with proper spatial resolution, such as Landsat-5 and Landsat-7, were designed mainly for land observation; however, Moderate Resolution Imaging Spectroradiometer (MODIS), Medium Resolution Imaging Spectrometer (MERIS), and Sea-viewing Wide Field-of-view Sensor (SeaWiFS) were designed for ocean color studies, but with a very low spatial resolution (i.e., inappropriate for water bodies with small widths). Hence, selecting the satellite sensor that provides suitable spatial, spectral, radiometric, and temporal resolutions is indeed a critical task in surface water quality studies.

Additionally, a remote sensing study of surface water quality requires multi-spectral data for the surface features, as they would be measured at ground level (Vermote, et al., 1997a). The conversion of the digital numbers (DNs) to the top of atmospheric (TOA) signal then from TOA to the ground level signal is the process of atmospheric correction. Hence, an accurate atmospheric correction is essential for remote sensing applications for surface water quality assessment, since the multi-spectral light signal from water surfaces is much less than the signal from land (Hu, Frank, Serge, & Kendall, 2001).

Second, SWQPs can be broadly classified into two main classes: optical and non-optical SWQPs. Optical SWQPs, such as turbidity and total suspended solids (TSS), are most likely to affect the water colour, the reflected signals, and consequently can be detected by satellite sensors. On the other hand, non-optical SWQPs, such as chemical oxygen demand (COD), biochemical oxygen demand (BOD), dissolved oxygen (DO), total solids (TS), total dissolved solids (TDS), power of hydrogen (pH), electrical conductivity (EC), and surface water temperature are less likely to affect the reflected radiation. Concentrations of both optical and non-optical SWQPs can be measured according to the American Public Health Association (APHA) water and wastewater standards (APHA, 2005).

1.3 Selected Research Topic

Based on the above-mentioned background information, the research topic selected for this dissertation focuses on the assessment of surface water quality by using

satellite imagery. In order to properly assess surface water quality within a water body, it is very important to:

- (1) Estimate the concentrations of both optical and non-optical SWQPs from satellite imagery.
- (2) Map the relationship between satellite multi-spectral data and the measured concentrations of SWQPs.
- (3) Improve the accuracy of surface water quality level extraction from surface water quality raw data (i.e., individual concentrations of SWQPs).
- (4) Classify the most significant SWQPs that negatively affect water bodies and consequently detect both spatial and temporal surface water quality variations.

This will lead to effective savings and proper utilization of water resources (Debels, Figueroa, Urrutia, Barra, & Niell, 2005; Elhatip, Hinis, & Gulghar, 2007; Akbar, Hassan, & Achari, 2011; Natural Resources, 2016). The problems addressed in this dissertation are identified and discussed in the following section.

1.4 Problem Statement

1.4.1 Estimation of the Concentrations of SWQPs from Satellite Imagery

The first challenge is related to quantifying the concentrations of SWQPs from satellite imagery. In literature, remote sensing has been commonly used for retrieving the concentrations of optical SWQPs; however, remote sensing estimation of non-optical SWQPs, such as COD, BOD, DO, pH, and EC, has not yet been performed because they are less likely to affect light signals measured by satellite detectors. However,

concentrations of non-optical SWQPs may be correlated with optical SWQPs, such as turbidity and TSS, which do affect the reflected radiation. In this context, an indirect relationship between satellite spectral information and concentrations of non-optical SWQPs can be assumed (Sharaf El Din, Zhang, & Suliman, 2017a; Sharaf El Din & Zhang, 2017b). Additionally, some of the available research has used remote sensing data provided from the Landsat TM/ETM+ and MODIS; however, these sensors were designed mainly for earth observation and they are not from the recently launched earth observation satellite sensors.

Therefore, the first concern of this dissertation is to address the problem of retrieving the concentrations of both optical and non-optical SWQPs from satellite imagery. The proposed solution aims at exploring an appropriate regression-based technique to estimate both optical and non-optical SWQPs from a recently launched earth observation satellite sensor, which is freely available and has the potential to support coastal studies.

1.4.2 Mapping the Relationship between Satellite Data and Concentrations of SWQPs

The second challenge is related to mapping the relationship between satellite multi-spectral information and concentrations of SWQPs. In literature, mapping this relationship is achievable via regression techniques. Theoretically, the relationship between satellite multi-spectral signatures and the concentrations of SWQPs is too complex, especially in the presence of various pollutants at the same time (Xiang, Huapeng, Xiangyang, Yebao, Xin, & Hua, 2016). Moreover, it is very challenging for

regression techniques to model such a complex relationship (Sharaf El Din, Zhang, & Suliman, 2017a; Sharaf El Din & Zhang, 2017c).

Therefore, the second concern of this dissertation is to address the problem of modelling the concentrations of SWQPs from satellite imagery. The proposed solution aims at developing a novel artificial intelligence (i.e., learning-based) modelling method for mapping concentrations of both optical and non-optical SWQPs by using remotely sensed multi-spectral data.

1.4.3 Extracting the Accurate Levels of Surface Water Quality within a Water Body

The third challenge is related to improving the accuracy of delineating the accurate levels (patterns) of surface water quality. Existing methods of assessing surface water quality are technically detailed and present monitoring data on individual substances (i.e., individual concentrations of SWQPs). The results of these methods are poorly understood by local administrators and decision-makers (Akoteyon, Omotayo, Soladoye, & Olaoye, 2011). Hence, a method, such as the water quality index (WQI), is needed to provide an integrated picture of surface water quality in water bodies. The WQI can support the accurate interpretation of surface water quality; however, it requires a huge number of water samples obtained by physical monitoring of water quality (CCME, 2001). It is very challenging to provide this type of physical monitoring because this process is costly and time consuming. Moreover, the selected WQI may be biased towards reflecting inaccurate surface water quality levels in the absence of a representative database (i.e. water samples).

Therefore, the third concern of this dissertation is to address the problem of delineating the accurate levels of surface water quality within a water body. The proposed solution aims at developing a novel approach which combines remote sensing multi-spectral data, artificial intelligence, and the WQI to extract accurate surface water quality levels to be accessible to decision-makers.

1.4.4 Identifying the Major SWQPs Contributing to Spatio-temporal Surface Water Quality Variations

The fourth challenge is related to categorizing the most significant SWQPs that negatively affect water bodies and contribute to surface water quality variations. Existing methods are based on understanding the relationships between different SWQPs and their relevance to the actual problem being studied. However, due to the redundancy and complexity of relationships between parameters of surface water quality, it is not easy to draw a clear conclusion directly from surface water quality data (Simeonov, Stratis, Samara, Zachariadis, Voutsas, & Anthemidis, 2003; Shrestha & Kazama, 2007).

Therefore, the fourth concern of this dissertation is to address the problem of classifying the major SWQPs and evaluating variations of surface water quality in a cost-effective manner. The proposed solution aims at using multivariate statistical techniques, such as principal component analysis/factor analysis (PCA/FA), cluster analysis (CA), and discriminant analysis (DA), to help in the interpretation of complex surface water quality data to better understand the surface water quality and ecological status of water bodies. Moreover, these techniques can identify the major pollution sources (i.e., SWQPs) contributing to spatio-temporal variations of surface water quality and provide a

valuable tool for reliable management of water resources as well as offering rapid solutions to control pollution problems.

1.5 Research Objectives

The objectives of this research are fourfold in order to solve the four identified limitations and problems in a progressively improving manner mainly in terms of cost, effort, and computational steps. The four main objectives of this dissertation are described in the following subsections.

1.5.1 Estimation of the Concentrations of SWQPs from Satellite Imagery

Retrieving the concentrations of SWQPs by using satellite imagery is critical. Based on a review of the relevant literature, statistical techniques have the potential to quantify the concentrations of SWQPs from space. However, none of the previous studies have attempted to estimate the concentrations of non-optical SWQPs, such as COD, BOD, DO, pH, and EC. Hence, the first objective of this dissertation, which is addressed in Chapter 2, is to develop a remote sensing technique for estimating the concentrations of both optical and non-optical SWQPs using stepwise regression. The ultimate goal of this objective is to demonstrate the performance of the proposed technique in estimating the concentrations of different SWQPs using satellite multi-spectral data.

1.5.2 Mapping the Relationship between Satellite Data and Concentrations of SWQPs

Mapping the relationship between multi-spectral information and concentrations of SWQPs is a very important step to support the assessment of surface water quality in

water bodies. Based on a review of the relevant literature, regression techniques have been used to support the modelling process of SWQPs; however, these techniques may fail in modelling such a complex relationship, especially in highly polluted water bodies. The use of artificial intelligence instead of regression techniques improves the efficiency and the accuracy of modelling complex relationships. Hence, the second objective of this dissertation, which is addressed in Chapter 3, is to develop a novel framework for mapping the concentrations of SWQPs from satellite imagery by using artificial intelligence. The ultimate goal of this objective is to show the effectiveness of the proposed technique in mapping the complex relationship between satellite multi-spectral signatures and the concentrations of SWQPs.

1.5.3 Extracting the Accurate Levels of Surface Water Quality within a Water Body

Delineating the accurate levels of surface water quality has always presented researchers with a great challenge. Based on a review of the relevant literature, the WQI has been used to provide an integrated picture of surface water quality; however, a huge number of water samples obtained by physical monitoring of surface water quality is needed. Hence, the third objective of this dissertation, which is addressed in Chapter 4, is to develop a novel technique that integrates remote sensing, artificial intelligence, and the WQI, for improving the accuracy of surface water quality level (SWQL) extraction with inexpensive implementation cost. The ultimate goal of this objective is to simplify the expression of surface water quality and illustrate the applicability of the proposed technique in extracting accurate surface water quality levels from water quality raw data.

1.5.4 Identifying the Major SWQPs Contributing to Spatio-temporal Surface Water Quality Variations

The existence of various pollutants in water bodies can lead to deterioration of surface water quality and thus raise the cost of water body treatment. To decrease the cost of the treatment process, evaluating surface water quality based on classifying the most significant SWQPs contributing to spatio-temporal variations of surface water quality is very important. Based on a review of the relevant literature, almost all of the available studies have attempted to categorize the parameters that affect water bodies; however, fewer research attempts focused on extracting spatio-temporal patterns of surface water quality. Hence, the fourth objective of this dissertation, which is addressed in Chapter 5, is to develop a cost-effective technique for classifying the major SWQPs in water bodies and detecting both spatial and temporal variations of surface water quality by using multivariate statistical analysis. The ultimate goal of this objective is to explore the usefulness of the proposed technique in assessing surface water quality by finding out the association between samples and parameters and revealing the most significant information which cannot be observed from the raw data.

1.6 Overview of Each Chapter

In relation to the dissertation structure, the four research objectives, identified above, are carried out in four chapters (Chapter 2-5). While the first three objectives are discussed separately in Chapters 2, 3, and 4, respectively, the fourth objective is addressed in Chapter 5.

Chapter 1 is the introduction. It comprises the structure of the dissertation, research background, topic selection, problem statement, objectives of the research, and an overview of each remaining chapter. Chapters 2 to 5 contain the four peer reviewed journal papers representing the main contributions of this PhD dissertation.

- Chapter 2 introduces the research work related to the developed remote sensing technique for quantifying the concentrations of SWQPs using stepwise regression. To the best of our knowledge, this technique is developed for the first time to estimate the concentrations of non-optical SWQPs, such as COD, BOD, DO, pH, and EC, which have not been estimated before by researchers using Landsat data or any other optical instrument.
- Chapter 3 represents the research work regarding a novel technique that can use artificial intelligence (i.e., learning-based modelling method) for mapping the concentrations of SWQPs from satellite imagery. To the best of our knowledge, this technique is the first to map the complex relationship between satellite multi-spectral data and concentrations of SWQPs with highly accurate results, compared to traditional techniques.
- Chapter 4 provides the research work regarding a novel technique that integrates remotely sensed data, artificial intelligence, and the WQI for simplifying the expression of surface water quality and delineating the accurate surface water quality levels (SWQLs) to be accessible to decision-makers. To the best of our knowledge, this technique is developed for the first time to extract the SWQLs with highly accurate results and inexpensive implementation cost.

- Chapter 5 demonstrates the research work related to the developed cost-effective technique for categorizing the most significant SWQPs and evaluating spatio-temporal surface water quality variations by using multivariate statistical techniques, such as PCA/FA, CA, and DA. This technique illustrates the significance use of multivariate statistical techniques for surface water quality assessment and management leading to effective savings and proper utilization of surface water quality resources.

Chapter 6 presents the conclusions. It summarizes the achievements of this research and outlines its drawbacks and limitations. It also presents some recommendations for future research.

REFERENCES

- Akbar, A. T., Hassan, K. Q., & Achari, G. (2011). A Methodology for Clustering Lakes in Alberta on the basis of Water Quality Parameters. *Clean-Soil, Air, Water, 39* (10), pp. 916-924.
- Akoteyon, I., Omotayo, A., Soladoye, O., & Olaoye, H. (2011). Determination of water quality index and suitability of urban river for municipal water supply in Lagos-Nigeria. *Europ. J. Scientific Res., 54* (2), pp. 263-271.
- APHA. (2005). *Standards Methods for the Examination of Water and Wastewater* (21th ed.). American Public Health Association Washington DC, USA.
- Arseneault, D. (2008). *The Road to Canada - Nomination Document for the St. John River, New Brunswick*. Prepared by The St. John River Society with the support of the New Brunswick Department of Natural Resources.
- Carpenter, S. R., Caraco, N. F., Correll, D. L., Howarth, R. W., Sharpley, A. N., & Smith, V. H. (1998). Non-point pollution of surface waters with phosphorus and nitrogen. *Ecological Applications, 83*, pp. 559-568.

- CCME. (2001). *Canadian water quality index 1.0 technical report and user's manual*. Gatineau, QC, Canada: Canadian Environmental Quality Guidelines Water Quality Index Technical Subcommittee.
- Debels, P., Figueroa, R., Urrutia, R., Barra, R., & Niell, X. (2005). Evaluation of water quality in the Chilla' n River (Central Chile) using physicochemical parameters and a modified water quality index. *Environ. Monit. Assess., Vol.110*, pp. 301-322.
- Elhatip, H., Hinis, M. A., & Gulghar, N. (2007). Evaluation of the water quality at Tahtali dam watershed in Izmir, Turkey by means of statistical methodology. *Stochastic Environmental Research and Risk Assessment*, 22, pp. 391-400.
- Gaballah, M., Khalaf, K., Beckand, A., & Lopez, A. (2005). Water Pollution in Relation to Agricultural Activity Impact in Egypt. *Journal of Applied Sciences, Vol. 1, No. 1*, pp. 9-17.
- Gower, J. F., & Borstad, G. A. (2004). On the potential of MODIS and MERIS for imaging chlorophyll fluorescense from space. *International Journal of Remote Sensing*, 25 (7-8), pp. 1459-1464.
- Hu, C., Frank, E. M., Serge, A., & Kendall, L. C. (2001). Atmospheric correction and cross-calibration of LANDSAT-7/ETM+ imagery over aquatic environments: A multiplatform approach using SeaWiFS/MODIS. *Remote Sensing of Environment*, 78 (1-2), pp. 99-107.
- Liu, D., Chin, C., Gong, J., & Fu, D. (2010). Remote Sensing of Chlorophyll-a Concentrations of the Pearl River Estuary from MODIS Land Bands. *International Journal of Remote Sensing*, 31, pp. 4625-4633.
- Natural resources*. (2016). Retrieved from Statistics Canada: <http://www.statcan.gc.ca/>
- Qadir, A., Malik, R. N., & Husain, S. Z. (2007). Spatio-temporal variations in water quality of Nullah Aik-tributary of the river Chenab, Pakistan. *Environmental Monitoring and Assessment*, 140(1-3), pp. 43-59.
- Sharaf El Din, E., Zhang, Y., & Suliman, A. (2017a). Mapping concentrations of surface water quality parameters using a novel remote sensing and artificial intelligence framework. *International Journal of Remote Sensing*, 38 (4), pp. 1023-1042.

- Sharaf El Din, E., & Zhang, Y. (2017b). Statistical estimation of the Saint John River surface water quality using Landsat 8 multi-spectral data. *ASPRS Annual Conference. Proceedings of Imaging & Geospatial Technology Forum (IGTF)*. Baltimore, US.
- Sharaf El Din, E., & Zhang, Y. (2017c). Neural network modelling of the Saint John River sediments and dissolved oxygen content from Landsat OLI imagery. *ASPRS Annual Conference. Proceedings of Imaging & Geospatial Technology Forum (IGTF)*. Baltimore, US.
- Shrestha, S., & Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environmental Modelling & Software*, 22, pp. 464-475.
- Simeonov, V., Stratis, J. A., Samara, C., Zachariadis, G., Voutsas, D., & Anthemidis, A. (2003). Assessment of the surface water quality in Northern Greece. *Water Research*, 37, pp. 4119-4124.
- Vermote, E. F., Saleous, N. E., Justice, C. O., Kaufman, Y. J., Prevette, J. L., Remer, L., et al. (1997a). Atmospheric correction of visible to middle-infrared EOS-MODIS data over land surface: background, operational algorithm and validation. *Journal of Geophysical Research*, 102, pp. 17131-17141.
- Xiang, Y., Huapeng, Y., Xiangyang, L., Yebao, W., Xin, L., & Hua, Z. (2016). Remote-sensing estimation of dissolved inorganic nitrogen concentration in the Bohai Sea using band combinations derived from MODIS data. *International Journal of Remote Sensing*, 37:2, pp. 327-340.
- Zhang, Y. Z., Pulliainen, J. T., Koponen, S. S., & Hallikainen, M. T. (2002). Application of an empirical neural network to surface water quality estimation in the Gulf of Finland using combined optical data and microwave data. *Remote Sensing of Environment*, 81, pp. 327-336.

Chapter 2: ESTIMATION OF BOTH OPTICAL AND NON- OPTICAL SURFACE WATER QUALITY PARAMETERS USING LANDSAT 8 OLI IMAGERY AND STATISTICAL TECHNIQUES¹

Abstract

Surface water quality assessment is widely performed using laboratory analysis, which is costly, labour intensive, and time consuming. In contrast, remote sensing has the potential to assess surface water quality because of its spatial and temporal consistency. It is essential to estimate concentrations of both optical and non-optical surface water quality parameters (SWQPs) on a regular basis from satellite imagery to provide the desired treatment for water bodies. Remote sensing estimation of non-optical SWQPs, such as chemical oxygen demand (COD), biochemical oxygen demand (BOD), and dissolved oxygen (DO), has not yet been performed because they are less likely to affect signals measured by satellite sensors. However, concentrations of non-optical variables can be correlated with optical variables, such as turbidity and total suspended sediments (TSS), which do affect the reflected radiation. In this context, an indirect relationship

¹ This paper has been published in the “*Journal of Applied Remote Sensing (JARS)*”:

Sharaf El Din, E., & Zhang, Y. (2017). Estimation of both optical and non-optical surface water quality parameters using Landsat 8 OLI imagery and statistical techniques. *Journal of Applied Remote Sensing (JARS)*, 11 (4), 046008 (2017), doi: 10.1117/1.JRS.11.046008.

For consistency throughout the dissertation, the format and style of figure captions, table titles, citation of references in the text, and section numbering have been slightly changed (from the original format of the journal in which the paper has been published or is under review) for Chapters 2-5.

between satellite spectral data and concentrations of COD, BOD, and DO can be assumed. Therefore, this research attempts to develop an integrated Landsat 8 band ratios and stepwise regression approach to estimate concentrations of both optical and non-optical SWQPs. Compared to previous studies, significant correlation between the Landsat 8 surface reflectance and concentrations of SWQPs was achieved and the obtained coefficient of determination (R^2) > 0.85 for turbidity, TSS, COD, BOD, and DO. These findings demonstrated the possibility of using our technique to develop models to estimate concentrations of SWQPs, and to generate spatio-temporal maps of SWQPs from Landsat 8 imagery.

2.1 Introduction

The degradation of surface water quality occurs due to the presence of many pollutants generated from agricultural, residential, commercial, and industrial activities. Moreover, climate changes during global warming can cause floods, drought, or even a noticeable increase in infectious diseases, which may degrade water quality (Murdoch, Baron, & Miller, 2000). Furthermore, continuous changes in the weather temperature due to seasonal impacts can negatively affect surface water quality. Due to these variations, monitoring and estimating concentrations of optically and non-optically active surface water quality parameters (SWQPs) on a large scale by exploiting remotely sensed data is essential for providing the targeted treatment to watersheds.

Remote sensing provides significant benefits over conventional water quality monitoring methods, mainly due to the synoptic coverage and temporal consistency of the data. Remote sensing has the potential to estimate the concentrations of SWQPs on

inland waters and estuaries in regions where traditional monitoring methods are either missing or insufficient (Navalgund, Jayaraman, & Roy, 2007). However, most of remote sensing data investigated in the reviewed research were not from the recently launched earth observation satellite sensors, such as the Landsat 8 Operational Land Imager (OLI). Moreover, remote sensing has been widely used for monitoring a few SWQPs, such as turbidity, total suspended sediments (TSS), secchi disk depth, and chlorophyll-a, which have been typically categorized as optical water quality variables (Alparslan, Aydöner, Tufekci, & Tufekci, 2007; He, Chen, Liu, & Chen, 2008; Mancino, Nolè, Urbano, Amato, & Ferrara, 2009; Liu, Chin, Gong, & Fu, 2010; Bresciani, D., D., G., & C., 2011; Yacobi, et al., 2011; Mao, J., D., B., & Q., 2012; Güttler, N., & G., 2013; Krista, et al., 2015; Sharaf El Din & Zhang, 2017b).

Few studies have attempted to estimate the concentrations of non-optical variables such as total phosphorus (Gonca , Aysegul, Ugur, & Kerem, 2008; Bistani, 2009), dissolved inorganic nitrogen (Xiang, et al., 2016), total nitrogen (He, Chen, Liu, & Chen, 2008; Gonca , Aysegul, Ugur, & Kerem, 2008). Moreover, remote sensing estimation of non-optical SWQPs, such as chemical oxygen demand (COD), biochemical oxygen demand (BOD), and dissolved oxygen (DO), has not yet been performed. Thus, it is a challenge to estimate the concentrations of COD, BOD, and DO from space because they are less likely to affect light signals measured by satellite detectors. However, concentrations of non-optical SWQPs may be correlated with optical variables, such as TSS, which have the potential to affect water color, the reflected radiation, and consequently can be detected by satellite sensors (Chen, et al., 2015; Xiang, et al., 2016). Based on these findings, concentrations of turbidity or TSS which are expected to be

highly correlated with spectral data may serve in the estimation of non-optical variables; thus, a correlation between satellite spectral data and COD, BOD, and DO can be assumed. Accordingly, our focus in this research is to estimate concentrations of both optical and non-optical SWQPs from the Landsat 8 data which have been acquired by a recent satellite sensor.

In the literature, estimating concentrations of SWQPs from space is achievable via regression-based techniques. Correlations between ground measured data and spectral bands can be used to develop remote sensing models for the estimation of SWQPs. A summary of previous methods used to estimate concentrations of SWQPs that are being used in this study is provided in the following four paragraphs.

A Landsat-5 TM image over New York Harbour was used to develop regression models to estimate the levels of turbidity (Hellweger, Schlossera, Lalla, & Weissel, 2004). The red band correlated positively with turbidity for areas affected by river runoff with coefficient of determination (R^2) = 0.78. Basically, using the TM red band is appropriate due to the influence of inorganic suspended particles, such as clay, to scattering in this region. Water quality in Reelfoot Lake, Tennessee, USA was evaluated for turbidity and TSS (Wang, Han, Kung, & Van Arsdale, 2006). There was a positive correlation between the Landsat-5 TM green band and turbidity and TSS. The R^2 values were 0.53 and 0.52 for turbidity and TSS, respectively. The reason of correlation with the Landsat-5 TM green band is that the organic macromolecules such as algae and phytoplankton that form TSS particles are higher than inorganic compounds.

The Landsat-7 ETM blue, green, red, and near-infrared bands have been used to estimate concentrations of TSS of the reservoir behind Omerli Dam (Alparslan, Aydöner,

Tufekci, & Tufekci, 2007). Regression analysis was used to develop empirical models using the Landsat-7 ETM data and ground measured SWQPs. The R^2 value for TSS was 0.99. Although this study provides high R^2 value, it lacks causal explanations and cross-temporal applicability.

The Moderate Resolution Imaging Spectroradiometer (MODIS) data were used to estimate concentrations of TSS across Lake Erie. Remote sensing concentration maps were produced for monthly mean distribution of TSS by using MODIS radiance at 748 nm for the period of five years (Binding, Jerome, Bukata, & Booty, 2010). Turbidity in Tampa Bay, Florida was estimated using MODIS band 1 (620-670) nm. It was observed that there was a significant relationship between MODIS band 1 reflectance values and field measurements of turbidity after rainfall events and the obtained R^2 value was 0.76 (Moreno-Madrinan, Al-Hamdan, Rickman, & Muller-Karger, 2010). The main basis of correlation with MODIS red band can be explained by the contribution of TSS particles to scattering in this particular wavelength. To the authors' best knowledge, the estimation of concentrations of COD, BOD, and DO from space has not been performed by researchers.

Based on our literature review findings, simple linear regression of single bands can provide high correlation between satellite data and measured concentrations of turbidity and TSS. The advantages of using single bands in green, red, and near-infrared wavelengths have been confirmed by researchers (Poets, Costa, Da Silva, Silva, & Morais, 2010). However, there is no obvious agreement between the reviewed studies on which bands are the best to predict the concentrations of turbidity and TSS. Moreover, when a water body is seriously polluted, it is difficult to model the complex relationship

between SWQPs and satellite data using statistical techniques based on single bands; thus, artificial neural network can be used to model such complex relationships (Zhang, Pulliainen, Koponen, & Hallikainen, 2002; Sharaf El Din & Zhang, 2017c). Furthermore, while several studies were carried out on significantly polluted areas, these techniques have not in the past been applied to other only slightly polluted water bodies, such as the Saint John River (SJR) in New Brunswick, Canada.

A question was identified regarding the capability of regression-based techniques in the retrieval of the concentrations of both optical and non-optical SWQPs from the Landsat 8 satellite imagery. The Landsat 8 OLI sensor is selected because its multi-spectral bands have been designed to be narrower than the older sensors and new bands, such as the coastal blue (CB), have been added to support coastal studies. Moreover, the proposed regression-based technique is the stepwise regression (SWR) due to its capability of maximizing prediction power using a minimum number of predictor variables, and efficiency in the applications of models' prediction and averaging (Derksen & Keselman, 1992).

The identified objectives of this research are to (1) verify the potential of using Landsat 8 spectral data in water quality studies and (2) develop a Landsat 8-based-SWR technique to estimate concentrations of both optical and non-optical SWQPs with highly accurate results. To the best of our knowledge, the Landsat 8-based-SWR technique is developed for the first time to estimate three non-optical SWQPs, namely COD, BOD, and DO, which have not been estimated before with Landsat data or any other optical instrument.

2.2 Materials and Methods

The flowchart of the proposed methodology used to retrieve concentrations of both optical and non-optical SWQPs from satellite imagery is shown in **Figure 2.1**.

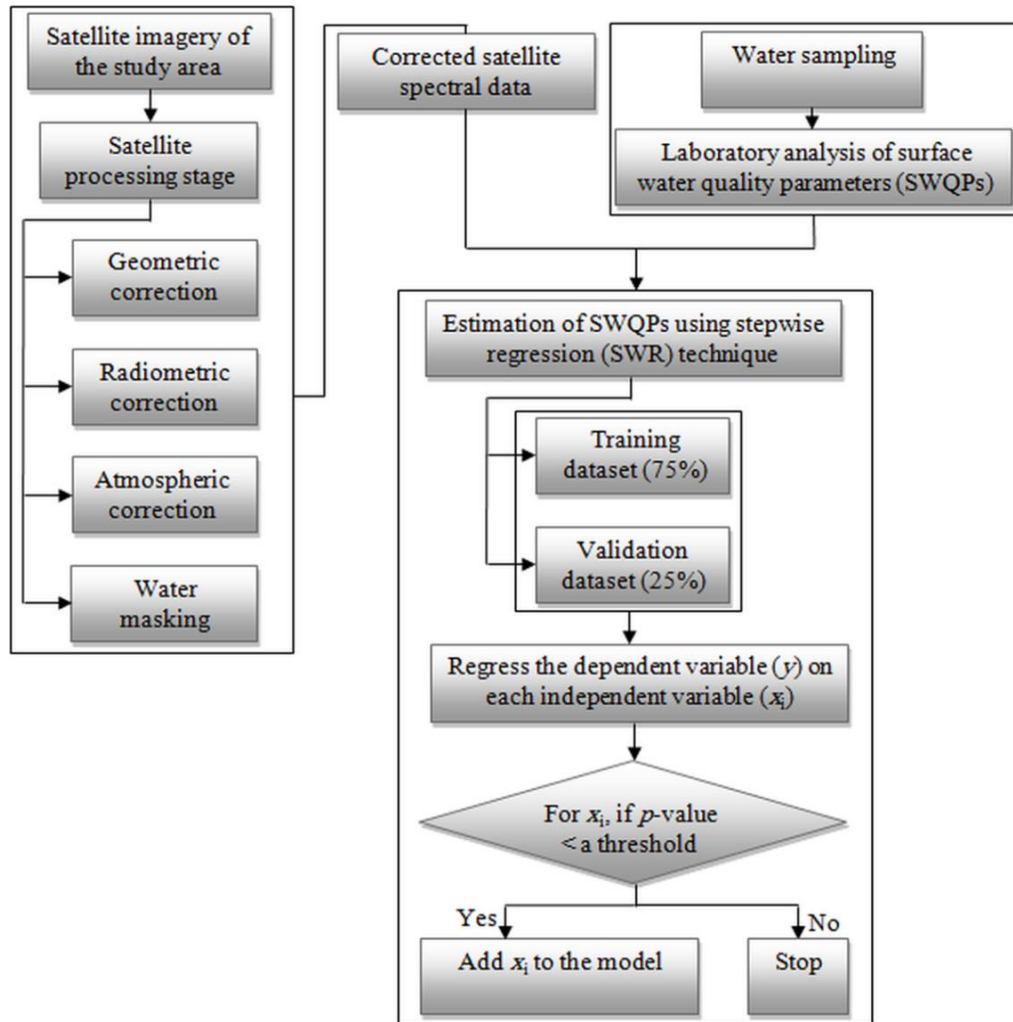


Figure 2.1 The flowchart of the proposed methodology

This section describes in detail the selected study area of the SJR, the Landsat 8 processing stage, water sampling, ground measurements and laboratory analysis, and estimation of the concentrations of optical and non-optical SWQPs using the proposed Landsat 8-based-SWR technique.

2.2.1 Selected Study Site

The SJR is approximately 673 km long with a maximum depth above the Mactaquac Dam of 50 m. Its headwaters are in the State of Maine, but is located principally in the Canadian province of New Brunswick. The selected study site is about 70 km long as shown in **Figure 2.2**.



Figure 2.2 The selected study area of the Saint John River (SJR), New Brunswick, Canada (Earth Explorer, 2016)

The SJR is one of the oldest streams in the Atlantic Ocean Basin (Arseneault, 2008). Peak flows on the SJR occur during the spring season and last several weeks. However, periods of low flow occur during the summer and winter months when the majority of the river is frozen (Arseneault, 2008). Basically, this is the first study to monitor and estimate the concentrations of different SWQPs in the SJR using remotely sensed data.

2.2.2 Satellite Processing Stage

2.2.2.1 Geometric Correction

The three Landsat 8 satellite sub-scenes used in our study were acquired on June 27th, 2015, April 10th, 2016, and May 12th, 2016. Further water samples which were collected in July and August 2016 were used in Chapter 4 and 5. The used images, along with their sampling events, were selected at different seasons to best represent the maximum variation in the concentrations of SWQPs (Arseneault, 2008). The Landsat 8 satellite images are available free of charge at Level 1T (terrain corrected) (Earth Explorer, 2016). The Level 1T data product provides systematic geometric accuracy by incorporating ground control points (GCPs), while also employing a Digital Elevation Model (DEM) for topographic accuracy. GCPs were used to geometrically correct the full landsat 8 scenes to the Universal Transverse Mercator (UTM) projection, World Geodetic System 1984 (WGS 84) datum with a 30 m grid (Earth Explorer, 2016).

2.2.2.2 Radiometric Correction

Normally, the Landsat 8 digital numbers (DNs) are stored in 16 bits unsigned integer format. **Equation (2.1)** is generally used to rescale DN to obtain the top of atmospheric (TOA) reflectance using the radiometric rescaling coefficients of the Landsat 8 data (United States Geological Survey (USGS), 2016). A full computation of the TOA reflectance was performed using PCI *Geomatica* image processing software.

$$\rho^* = M_\rho \times Q_{\text{cal}} + A_\rho \quad (2.1)$$

where ρ^* is the TOA reflectance without correction for the solar zenith angle; M_ρ is the reflectance band-specific multiplicative rescaling factor; Q_{cal} is the quantized and calibrated standard product pixel values (DN); and A_ρ is the reflectance band-specific additive rescaling factor.

Basically, the reflectance obtained from the Landsat 8 data is not corrected for solar zenith angle. This means that the provided reflectance is generally too low and this error increases at high latitudes and in the cold season. The TOA reflectance with a correction for solar zenith angle was performed using **Equation (2.2)** (United States Geological Survey (USGS), 2016). As shown in the Equation below, Landsat-8 TOA is also calculated using Landsat 8 metadata file parameters, such as the spectral radiance at the sensor's aperture, Earth-Sun distance, and mean solar irradiance (United States Geological Survey (USGS), 2016).

$$\rho = \rho^* / \cos(\theta)_{sz} = [\pi \times L_\lambda \times d^2] / [(E_{\text{Sun}\lambda} \times \cos(\theta)_{sz})] \quad (2.2)$$

where ρ is the corrected TOA planetary reflectance and $(\theta)_{sz}$ is the solar zenith angle; L_λ is the spectral radiance at the sensor's aperture; d is the Earth-Sun distance in astronomical units; and $E_{\text{Sun}\lambda}$ is the mean solar irradiance.

2.2.2.3 Atmospheric Correction

The effects of the atmosphere were considered in order to measure the reflectance at the ground. The surface reflectance (ρ_{surface}) is calculated using the Dark Object Subtraction (DOS) method (Chavez, 1988) using **Equations (2.3-2.7)**. As shown in

Figure 2.3, this method can provide accurate results in discriminating and mapping wetland areas (Song, Woodcock, Seto, Lenney, & Macomber, 2001).

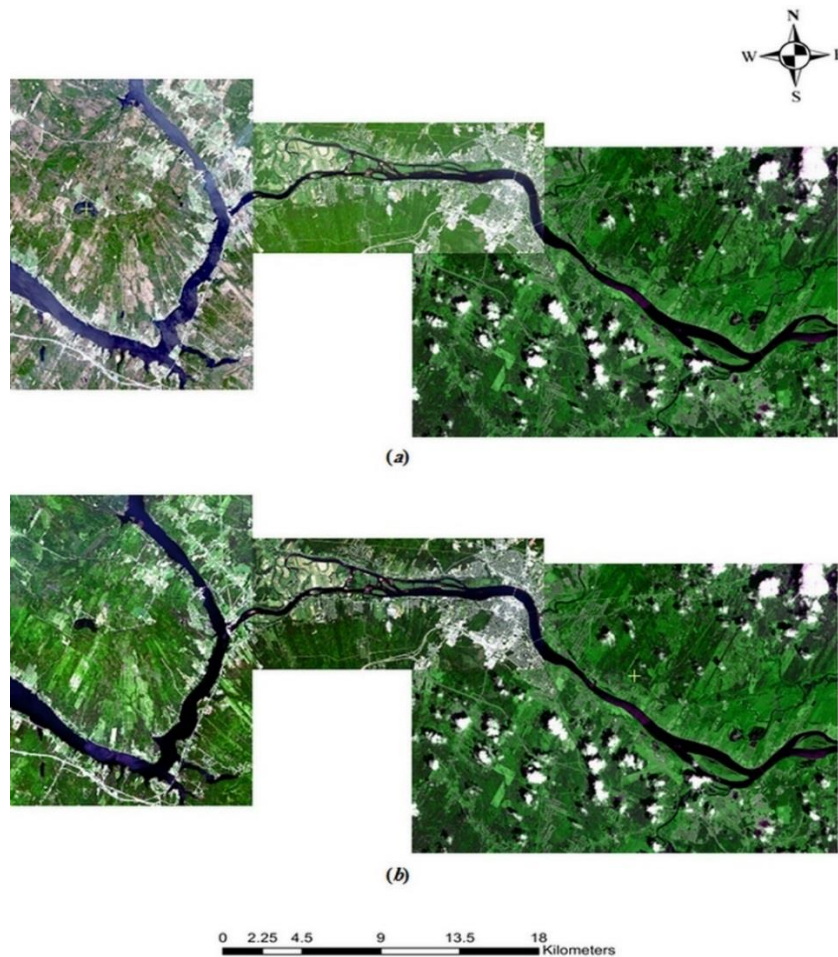


Figure 2.3 (a) The original Landsat 8 satellite sub-scenes and (b) the atmospherically corrected Landsat 8 satellite sub-scenes using the Dark Object Subtraction (DOS) method

Other atmospheric correction methods, such as atmospheric and topographic correction (ATCOR) and second simulation of the satellite signal in the solar spectrum (6S), have been used in remote sensing and digital image processing. However, the main drawback of these methods is that they involve extensive field measurements during each satellite pass. This is unacceptable for a variety of applications and is often impossible, as

when using historical data or when working in very remote or difficult access locations (Pat & Chavez, 1996; Song, Woodcock, Seto, Lenney, & Macomber, 2001). Additionally, Landsat 8 surface reflectance data (i.e., Level 2 product) are available free of charge from USGS; however, the provided data almost always have surface reflectance values > 1 for cloud and snow pixels and < 0 for water and shadow pixels (USGS Landsat 8 Surface Reflectance Product Guide, 2018).

$$\rho_{\text{surface}} = [\pi \times (L_{\lambda} - L_P) \times d^2] / [T_V \times ((E_{\text{Sun}\lambda} \times \cos(\theta)_{\text{sz}}) \times T_Z) + E_{\text{down}}] \quad (2.3)$$

$$L_P = L_{\lambda\text{min}} - L_{\text{DO1\%}} \quad (2.4)$$

$$L_{\lambda\text{min}} = M_L \times DN_{\text{min}} + A_L \quad (2.5)$$

$$L_{\text{DO1\%}} = [0.01 \times T_V \times ((E_{\text{Sun}\lambda} \times \cos(\theta)_{\text{sz}}) \times T_Z) + E_{\text{down}}] / [\pi \times d^2] \quad (2.6)$$

$$E_{\text{Sun}\lambda} = [\pi \times d^2 \times \text{RADIANCE}_{\text{max}}] / [\text{REFLECTANCE}_{\text{max}}] \quad (2.7)$$

where L_{λ} is the spectral radiance at the sensor's aperture; L_P is the path radiance due to atmospheric effects; d is the Earth-Sun distance in astronomical units; T_V is the atmospheric transmittance in the viewing direction; $E_{\text{Sun}\lambda}$ is the mean solar radiation entering to the atmosphere (Landsat 7 Science Data Users Handbook, 2011); T_Z is the atmospheric transmittance in the illumination direction; E_{down} is the downwelling diffuse irradiance; $L_{\lambda\text{min}}$ is the radiance values correspond to the minimum pixel values; $L_{\text{DO1\%}}$ is the radiance of dark object; M_L is the radiance band-specific multiplicative rescaling factor; DN_{min} is the minimum pixel values; and A_L is the radiance band-

specific additive rescaling factor.

2.2.2.4 The Water Interface

To estimate the concentrations of different SWQPs over a specific water body, the water surface should be delineated accurately as shown in **Figure 2.4**. Therefore, instead of using the whole image pixels in the process of mapping the concentrations of SWQPs, only water pixels can be included in this process to accelerate the processing/computational speed of the developed models.

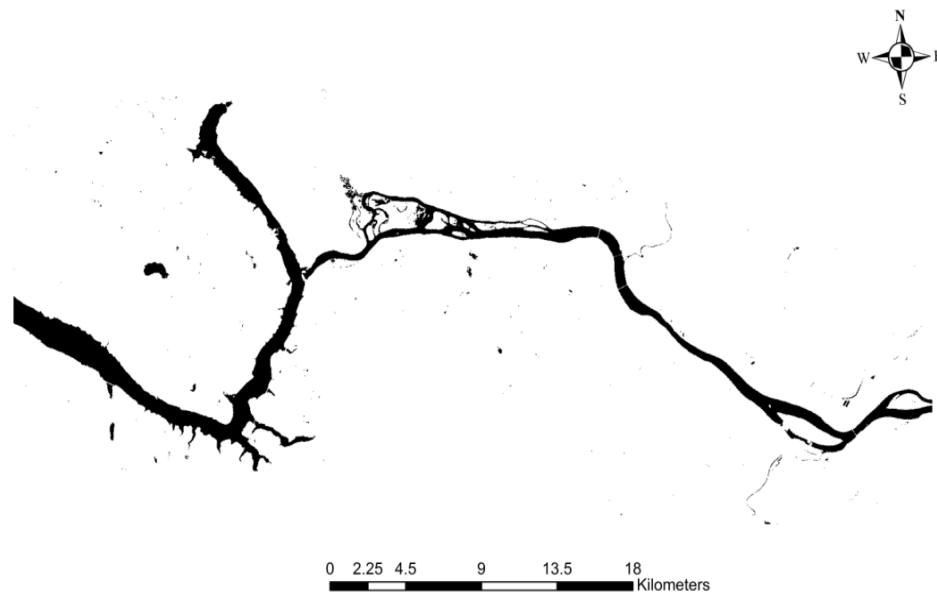


Figure 2.4 The water interface

The water interface was masked using the adjusted Normalized Difference Water Index (NDWI) to separate water and non-water features (Mcfeeters, 1996). The adjusted NDWI is derived by using principles similar to those used to derive the normalized difference vegetation index (NDVI). **Equation (2.8)** was used to calculate the adjusted NDWI and the results of the index ranged from [-1.00 to +1.00]. Water features showed

negative values due to their typically higher reflectance of green band than near-infrared band and accordingly water pixels were directly separated from non-water pixels, which showed positive and zero values.

$$(NDWI) = [(NIR) - (G)] / [(NIR) + (G)] \quad (2.8)$$

where NIR is the near-infrared band; and G is the green band.

2.2.3 Sampling Sites and Laboratory Analysis of Optical and Non-optical SWQPs

Water sampling was performed during three field trips in June 27th, 2015, April 10th, 2016, and May 12th, 2016. Samples were randomly selected and distributed across the entire study area as shown in **Figure 2.5**.

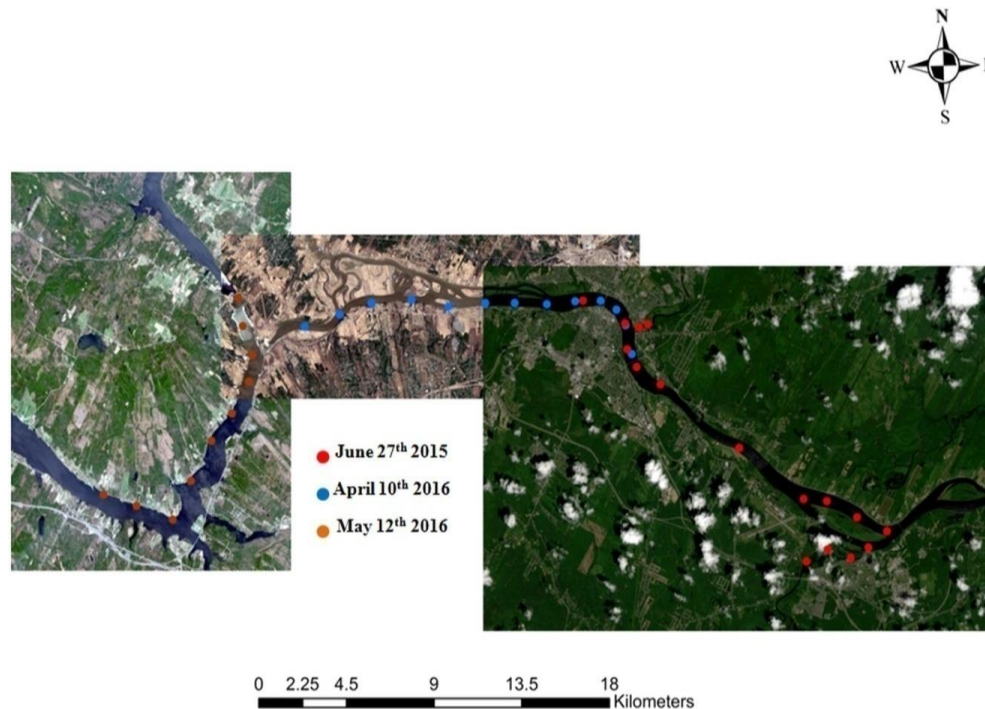


Figure 2.5 The water sampling locations across the SJR, New Brunswick, Canada

Thirty-nine water samples were collected along the selected study area of the SJR and one sample was excluded due to cloud coverage. Coordinates of each sample point were recorded in the field using a handset GPS, GARMIN 76CSx. To determine if a sample size is big enough, a commonly used rule-of-thumb is to use 30 data points or more, especially for parametric statistical methods (Gregory & Dale, 2009; Sitanshu & Archana, 2013). In our study, the number of collected water samples is sufficient compared to other studies. The number of samples (n) was 19 (D'SA & Miller, 2003), $n = 29$ (Floricioiu, Rott, Rott, Dokulil, & Defrancesco, 2004), $n = 33$ (Simis, Peters, & Gonos, 2005), $n = 8$ (Odermatt, Heege, Nieke, Kneubuhler, & Itten, 2008), $n = 23$ (Kratzer, Brockmann, & Moore, 2008), and $n = 36$ (Moses, Gitelson, Berdnikov, & Povazhnyy, 2009a).

Another way to determine the appropriate sample size is to use one of the formulas shown in **Equations (2.9-2.10)** (Lisa, 2016). **Equation (2.9)** can be used if the standard deviation of the outcome of interest (i.e. the selected SWQPs) is known; otherwise, **Equation (2.10)** can be used. At a power (confidence level) of 95%, a desired marginal error of ± 2 units, and an average standard deviation of 5.96 for the measured SWQPs, the obtained z-score is 1.96 and accordingly the calculated number of samples is ($34.12 \approx 35 \pm 2$), which means the number of samples collected in our study is adequate.

$$n = [(z^* \times \sigma_x)/ME]^2 \quad (2.9)$$

$$n = [(z^*/ME)^2 \times p^* \times (1 - p^*)] \quad (2.10)$$

where n is the minimum sample size; z is z-score (value of standard normal distribution

for the desired confidence level); σ_x is the standard deviation of the outcome variable; ME is the desired margin of error (confidence interval (i.e. the maximum allowable deviation or error of the estimate)); and p^* is the proportion of successes in the population (can be obtained from previous similar studies).

In order to carry out this study efficiently, water samples were collected just beneath water surface (i.e., 30 to 50 cm) and around the same time as the satellite sensor overpass (4 hours time difference). Concentrations of turbidity, TSS, COD, BOD, and DO, were measured according to the standard methods for lab examination of water and wastewater of the American Public Health Association (APHA) (APHA, 2005). Turbidity is an optical determination of water clarity and is based on the amount of light scattered by particles in the water column. TSS concentrations are determined by filtering the water sample and weighing the residue left on the filter paper. COD levels are measured as the amount of a specific oxidizing agent that reacts with a sample under controlled conditions and it is expressed as oxygen equivalence. BOD is used to determine the amount of dissolved oxygen needed by aerobic organisms in a water body to break down the organic materials present in the given water sample over 5 days at 20°C temperature. Finally, concentrations of DO are estimated as the level of free (non-compound) oxygen present in a water sample.

2.2.4 Estimation of Concentrations of SWQPs using the Stepwise Regression Technique

Regression analysis is a form of predictive modelling technique which attempts to model the relationship between a dependent and a set of independent variables (i.e.,

predictors). Regression analysis is commonly used for forecasting and time series modelling applications (Derksen & Keselman, 1992). Basically, there are several types of regression techniques, such as linear regression, logistic regression, polynomial regression, ridge regression, and stepwise regression (SWR). The SWR technique is selected as the proposed regression-based technique. The main advantages of using the SWR technique are (1) the ability of managing large amounts of independent variables and tuning the model to choose the best independent variables from the available data, and (2) the computational speed is usually faster than other regression techniques.

Our problem is a good candidate for SWR because (1) the variables are quantitative (i.e. SWQPs are measurable), (2) the independent variables (i.e. surface reflectance values of bands/band ratios) are not highly correlated with each other (i.e. little or no multicollinearity), (3) the errors (difference between observed values and a true value, which is very often the mean value) are normally distributed, and (4) the residuals (difference between observed and predicted values) are independent, as Durbin-Watson was used as a measure of independence and the obtained scores are close to 2.00.

The SWR method selects a sub-set from a list of explanatory (independent) variables and removes and adds variables to the regression model for the purpose of identifying a useful subset of the predictors (Derksen & Keselman, 1992). In this context, the SWR first finds the explanatory variable with the smallest significant value (P -value) to start over. The SWR then tries each of the remaining explanatory variables until it finds the two variables with the smallest P -value. After that, the SWR tries all of them again until it finds the three variables with the smallest P -value, and so on. Generally, the process of adding more variables stops when all of the available variables have been

included or when it is not possible to make a statistically significant improvement in P -value by using any of the variables which have not been yet included.

In our study, the SWR technique was used to model the relationship between the Landsat 8 surface reflectance data and concentrations of optical and non-optical SWQPs. Sampling points were subdivided into two datasets; calibration (75% of all samples) and validation (25% of all samples) to establish and validate the developed models. The performance of the developed models was evaluated by using regression lines' equations, R^2 , root mean square error (RMSE), significant value (P -value), and residual prediction deviation (RPD). The RPD can be used along with R^2 , RMSE, and P -value as an indication of model stability (Nduwamungu, et al., 2009). However, a previous study conducted by Chang, et al., (2001) evaluated the performance of the developed models based only on R^2 and RPD values and three model categories were identified as follows:

- 1st category ($0.80 \leq R^2 \leq 1.00$ and $RPD \geq 2.00$) means accurate prediction.
- 2nd category ($0.50 \leq R^2 < 0.80$ and $1.40 \leq RPD < 2.00$) means satisfactory prediction.
- 3rd category ($R^2 < 0.50$ and $RPD < 1.40$) means unacceptable prediction.

2.3 Results and Discussion

The present study attempts to retrieve concentrations of both optical and non-optical SWQPs from satellite imagery. The main results of this study were divided into (1) concentrations of SWQPs, (2) the relationship between Landsat 8 satellite data and concentrations of SWQPs, (3) estimation and validation of the developed Landsat 8-based-SWR models, and (4) producing Landsat 8-based-SWR spatial distribution maps.

2.3.1 Optical and Non-optical Concentrations of SWQPs of Water Samples

Thirty-nine water samples were collected across the selected study area and analyzed for different SWQPs.

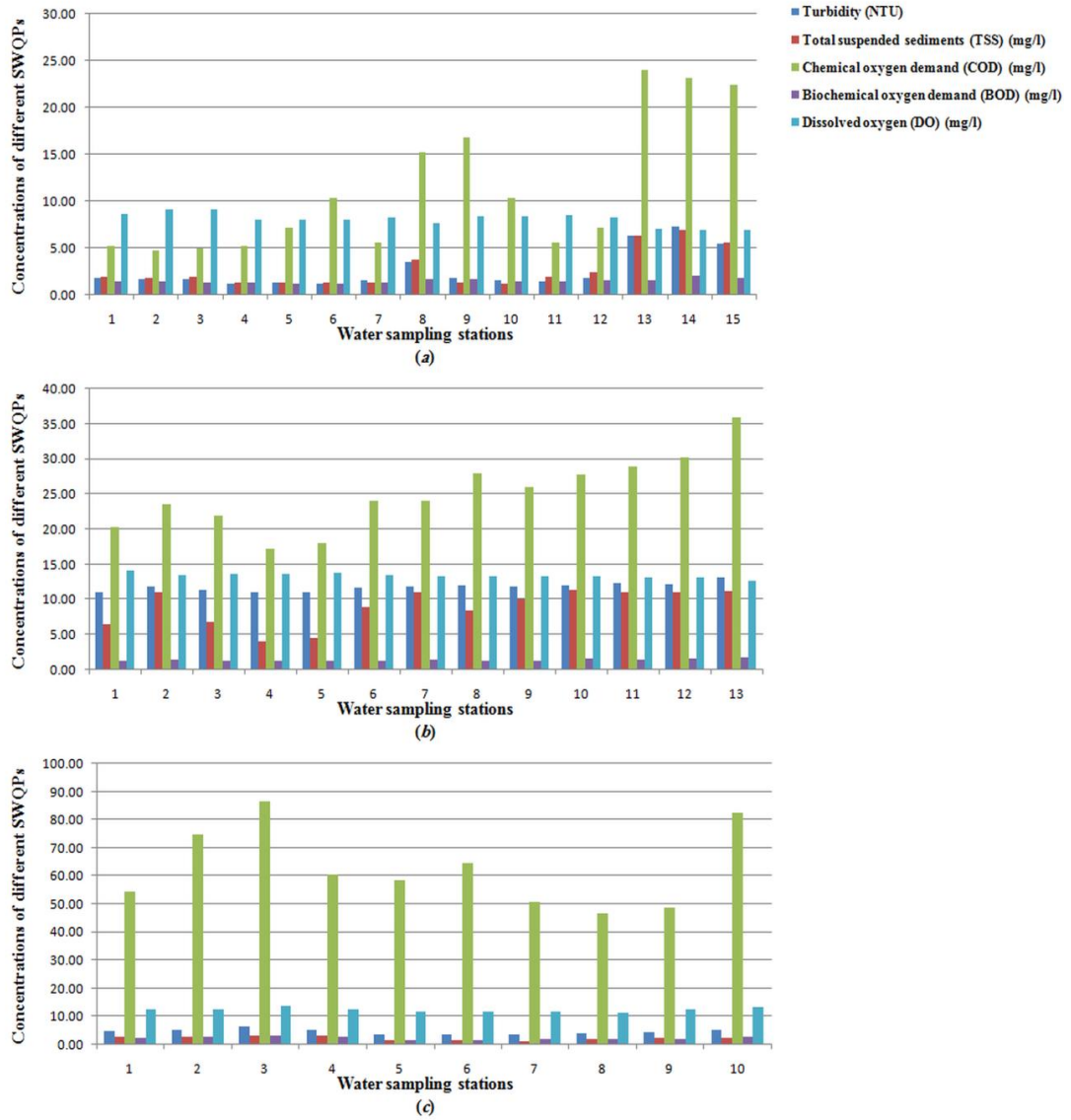


Figure 2.6 Optical and non-optical concentrations of SWQPs at June 27th 2015 (a), April 10th 2016 (b), and May 12th 2016 (c), respectively

The descriptive statistics were measured for turbidity, TSS, COD, BOD, and DO. For the used 38 water samples, as shown in **Figure 2.6**, the concentrations ranged from 1.19 to 13.10 NTU (Nephelometric Turbidity Units) with an average 6.30 NTU, 1.20 to 11.40 mg/l with an average 4.78 mg/l, 4.80 to 86.64 mg/l with an average 29.55 mg/l, 1.21 to 3.25 mg/l with an average 1.70 mg/l, and 6.99 to 14.14 mg/l with an average 11.06 for turbidity, TSS, COD, BOD, and DO, respectively.

As shown in **Figure 2.6**, turbidity and TSS in spring were higher than their concentrations in summer. The main reason is that rainfall and snowmelt are considered as the main contributors to the annual flows of the SJR and consequently wash sediments from agriculture and forestry into the river. On the other hand, the upper part of the selected study area of the SJR has high concentrations of COD and BOD because this region has many industries such as food and paper production located at the SJR shoreline.

The correlation between concentrations of optical and non-optical SWQPs was calculated as shown in **Table 2.1**. The relationship between turbidity and all SWQPs except DO was positively correlated; while, correlation values between DO levels and turbidity, TSS, COD, and BOD were -0.816, -0.824, -0.838, and -0.776, respectively. The non-optical SWQPs are less likely to affect the light signals measured by satellite detectors, and thus they cannot be measured directly by satellite sensors. The only way they can be measured is indirectly by the fact that their concentrations are correlated in some way with optical SWQPs like TSS or turbidity that do affect the signals measured by satellite sensors (Xiang, et al., 2016). Such indirect effects may be site-specific;

however, once the correlation between optical and non-optical SWQPs is found, developing remote sensing estimation models for non-optical SWQPs is guaranteed.

Table 2.1 The correlation matrix of both optical and non-optical SWQPs.

SWQPs	Turbidity	TSS	COD	BOD	DO
Turbidity	1.000	0.976	0.857	0.799	-0.816
TSS	0.976	1.000	0.861	0.850	-0.824
COD	0.857	0.861	1.000	0.895	-0.838
BOD	0.799	0.850	0.895	1.000	-0.776
DO	-0.816	-0.824	-0.838	-0.776	1.000

2.3.2 Relationship between Landsat 8 Satellite Spectral Data and Concentrations of SWQPs

Figure 2.7 indicates that the Landsat 8 surface reflectance data and concentration of the selected SWQPs are significantly correlated with R^2 values exceeded 0.800 and P -value < 0.001 throughout the SJR. Almost all of the Landsat 8 OLI multi-spectral bands, such as blue (B), green (G), red (R), shortwave infrared 1 (SWIR1), and shortwave infrared 2 (SWIR2), significantly contributed to the process of developing accurate models to estimate the concentrations of both optical and non-optical SWQPs in the SJR. Moreover, the new coastal blue (CB) band which was added to the Landsat 8 multi-spectral bands, compared to older sensors, performed very well in developing turbidity, TSS, COD, BOD, and DO estimation models. Furthermore, band ratios were very helpful in estimating concentrations of SWQPs due to the ability to enhance spectral contrast

between different targets, and to remove much of the effect of illumination in the analysis of spectral differences. The criteria of selecting the most statistically significant independent variables (i.e. Landsat 8 spectral bands/band ratios) were performed using the SWR technique, which includes or removes one independent variable at each step, based on the probability of F-to-enter and F-to-remove (Derksen & Keselman, 1992).

Turbidity and TSS concentrations were found to be very sensitive to the Landsat 8 CB, B, and R bands and their band ratios. This result is similar to those of previous studies (Hellweger, Schlossera, Lalla, & Weissel, 2004; Moreno-Madrinan, Al-Hamdan, Rickman, & Muller-Karger, 2010), which showed that turbidity and TSS concentrations were highly correlated to the B and R bands owing to the contribution of inorganic suspended compounds to reflectance in these wavelengths. On the other hand, the CB and SWIR2 bands were very efficient in estimating concentrations of COD and BOD; however, (SWIR1/G) and (CB/B) band ratios were found to be highly correlated with levels of DO.

The main reason of obtaining high correlation between Landsat 8 satellite data and concentrations of different SWQPs is that the Landsat 8 multi-spectral bands were designed to be narrower than older sensors, which can be very helpful in discriminating between fine targets, such as SWQPs in water bodies (United States Geological Survey (USGS), 2016). Moreover, in order to support surface water quality studies, new bands, such as CB, were added to the spectral bands of Landsat 8 data because at this specific wavelength, water pixels tend to reflect all radiation without any scattering. These findings confirm the potential of using Landsat 8 imagery in coastal studies (United States Geological Survey (USGS), 2016).

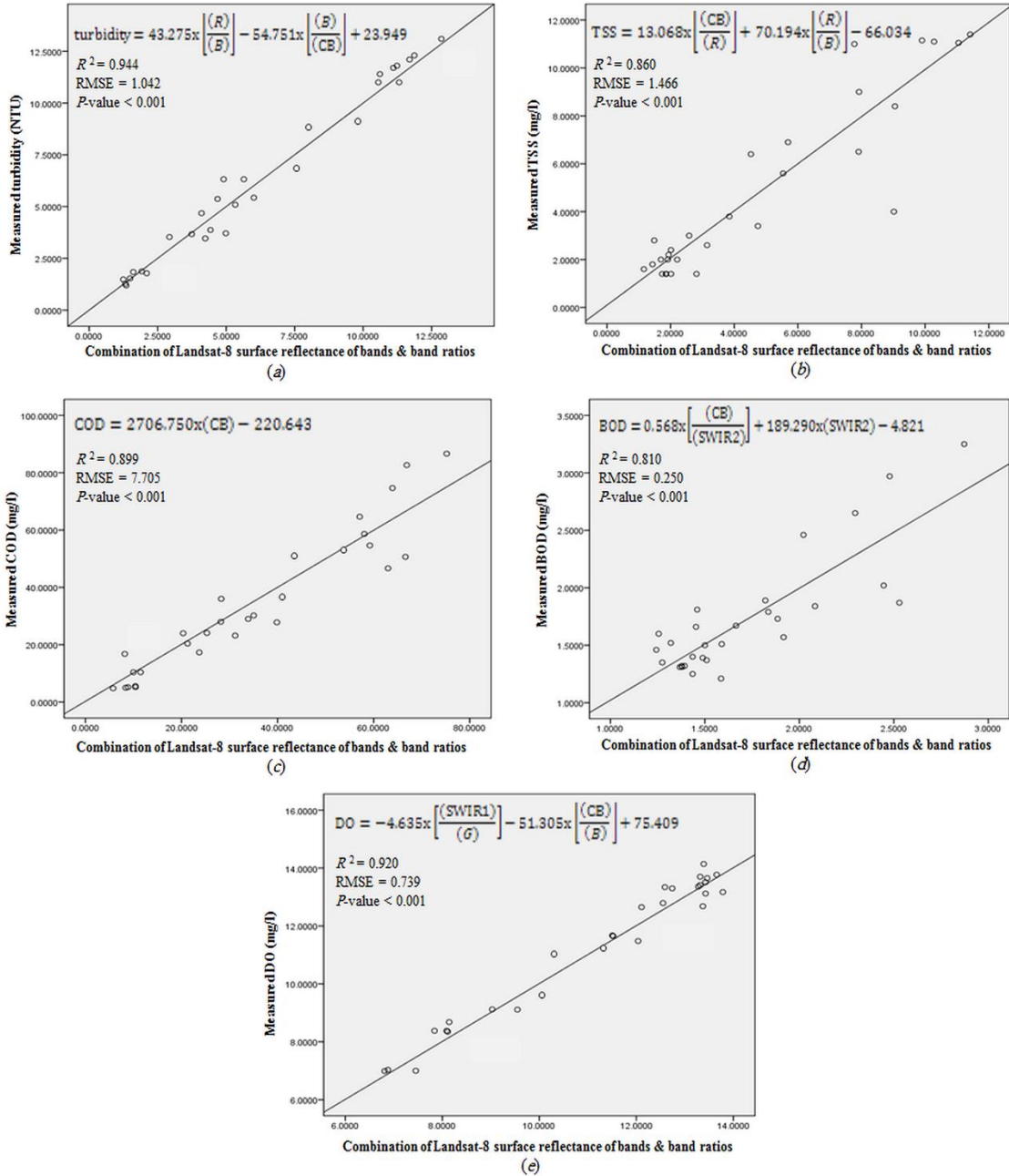


Figure 2.7 The Landsat 8 estimation models for turbidity (a), TSS (b), COD (c), BOD (d), and DO (e) using the SWR technique based on calibration dataset

2.3.3 Estimation and Validation of the Landsat 8-based-SWR Models

The surface reflectance data of the Landsat 8 multi-spectral bands and band ratios were correlated to the measurements of different SWQPs and the relationship between them were calculated using the SWR technique. The Landsat 8-based-SWR estimation models were established based on the calibration dataset and the bands and band ratios that showed the highest correlations were considered in the mathematical model of each water quality variable as shown in **Figure 2.7**. In this context, the best regression models for predicting concentrations of turbidity, TSS, COD, BOD, and DO in the SJR were obtained based on R^2 , RMSE, P -value, and RPD. Moreover, to test the reliability and applicability of the developed Landsat 8-based-SWR models in estimating concentrations of optical and non-optical SWQPs, an independent validation dataset of the remaining water samples was used to validate their performance.

As shown in **Figure 2.8**, concentrations of turbidity and TSS were significantly estimated using the Landsat 8-based-SWR models and the accuracy measures were ($R^2 = 0.965$, RMSE = 0.727 NTU, P -value < 0.001, and RPD = 5.345) and ($R^2 = 0.883$, RMSE = 0.980 mg/l, P -value < 0.001, and RPD = 2.923), respectively. Moreover, the validation models for turbidity and TSS, shown in **Figure 2.9**, remained very stable with ($R^2 = 0.939$, RMSE = 0.784 NTU, P -value < 0.001, and RPD = 4.048) and ($R^2 = 0.891$, RMSE = 0.801 mg/l, P -value < 0.001, and RPD = 3.028), respectively. Similarly, estimation models for COD, BOD, and DO were ($R^2 = 0.886$, RMSE = 7.304 mg/l, P -value < 0.001, and RPD = 2.961), ($R^2 = 0.801$, RMSE = 0.217 mg/l, P -value < 0.001, and RPD = 2.241), and ($R^2 = 0.915$, RMSE = 0.597 mg/l, P -value < 0.001, and RPD = 3.429), respectively.

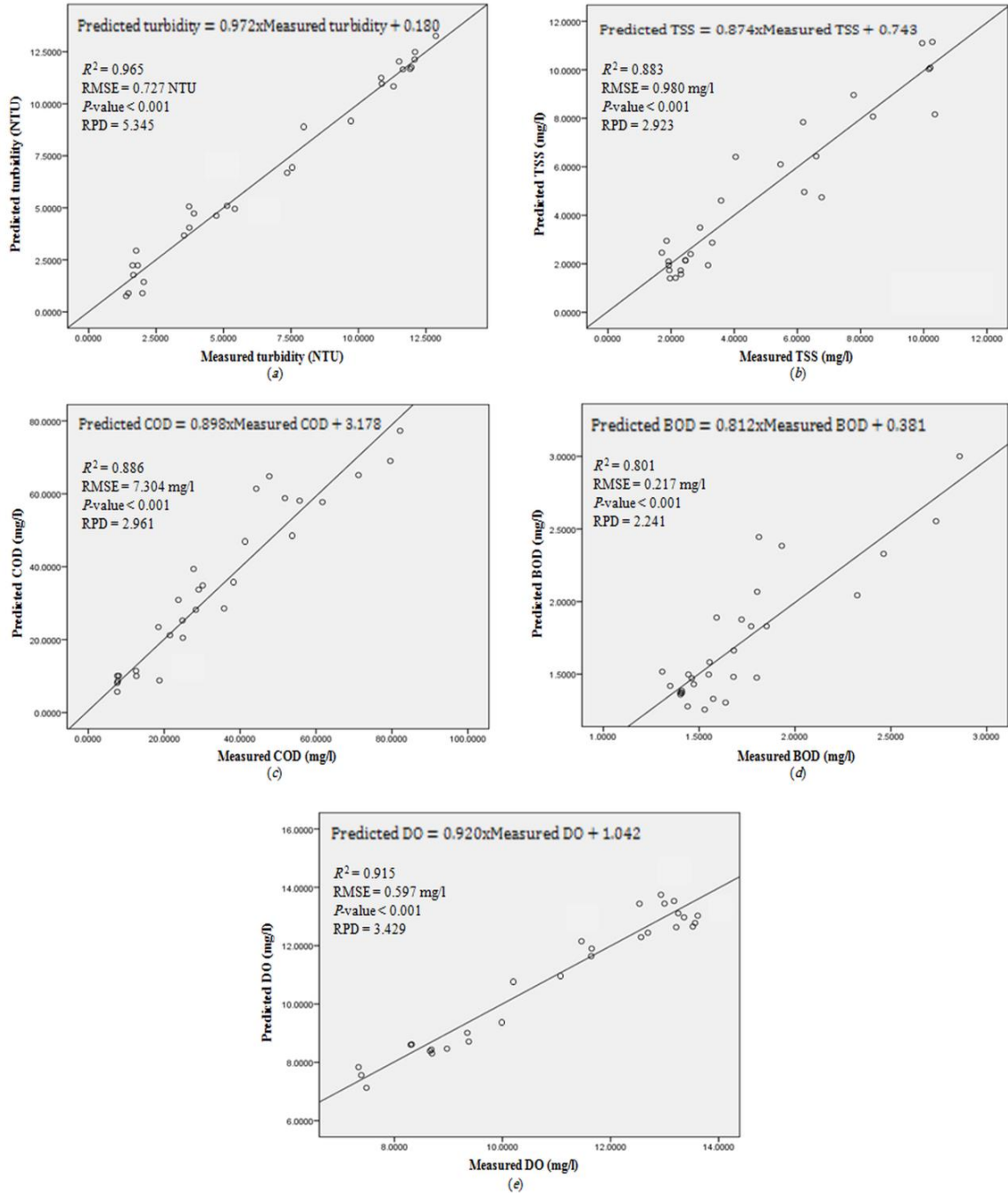


Figure 2.8 Statistics and accuracy measures between the measured and predicted concentrations of turbidity (a), TSS (b), COD (c), BOD (d), and DO (e) using the SWR technique based on calibration dataset

Additionally, the validation models for them were stable with ($R^2 = 0.901$, RMSE = 6.475 mg/l, P -value < 0.001, and RPD = 3.178), ($R^2 = 0.857$, RMSE = 0.180 mg/l, P -value < 0.001, and RPD = 2.644), and ($R^2 = 0.905$, RMSE = 0.730 mg/l, P -value < 0.001, and RPD = 3.244), respectively.

It can be noted that the concentrations of both optical and non-optical SWQPs in the selected study area of the SJR were well established and evaluated using the Landsat 8-based-SWR models. Accordingly, highly accurate results were achieved to retrieve concentrations of optical and non-optical SWQPs from the Landsat 8 satellite data. The main reasons are:

- Water sampling was performed at the same time of Landsat 8 over pass.
- The Landsat 8 satellite data were supposed to be efficient, because their multi-spectral data were designed to be narrower than older sensors, and new bands, such as CB, were added to support water quality monitoring applications.
- The Landsat 8 surface reflectance data were used to represent only the water-leaving reflectance without introducing radiometric or atmospheric distortions.
- Band rationing was found to be a good tool for estimating the concentrations of optical and non-optical SWQPs due to its ability to enhance spectral contrast between different targets, and to remove much of the effect of illumination in the analysis of spectral differences.
- The SWR technique was introduced and implemented because it is capable of maximizing prediction power using a minimum number of independent variables.

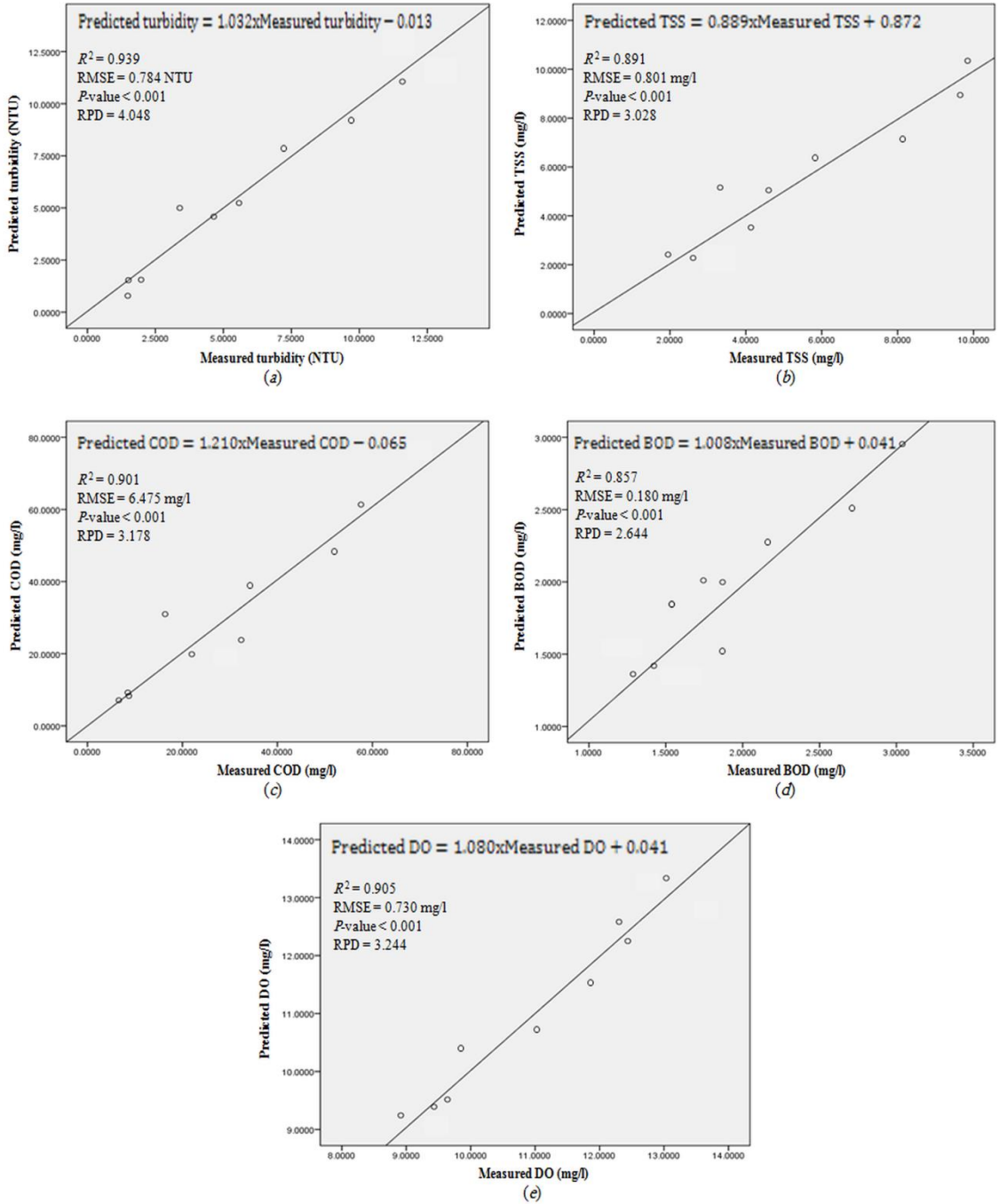


Figure 2.9 Statistics and accuracy measures between the measured and predicted concentrations of turbidity (a), TSS (b), COD (c), BOD (d), and DO (e) using the SWR technique based on validation dataset

2.3.4 Landsat 8-based-SWR Spatial Distribution Maps

As shown in **Figure 2.10**, the developed Landsat 8-based-SWR models were applied to each pixel of the selected study area of the SJR to generate highly accurate spatial concentration maps for turbidity, TSS, COD, BOD, and DO.

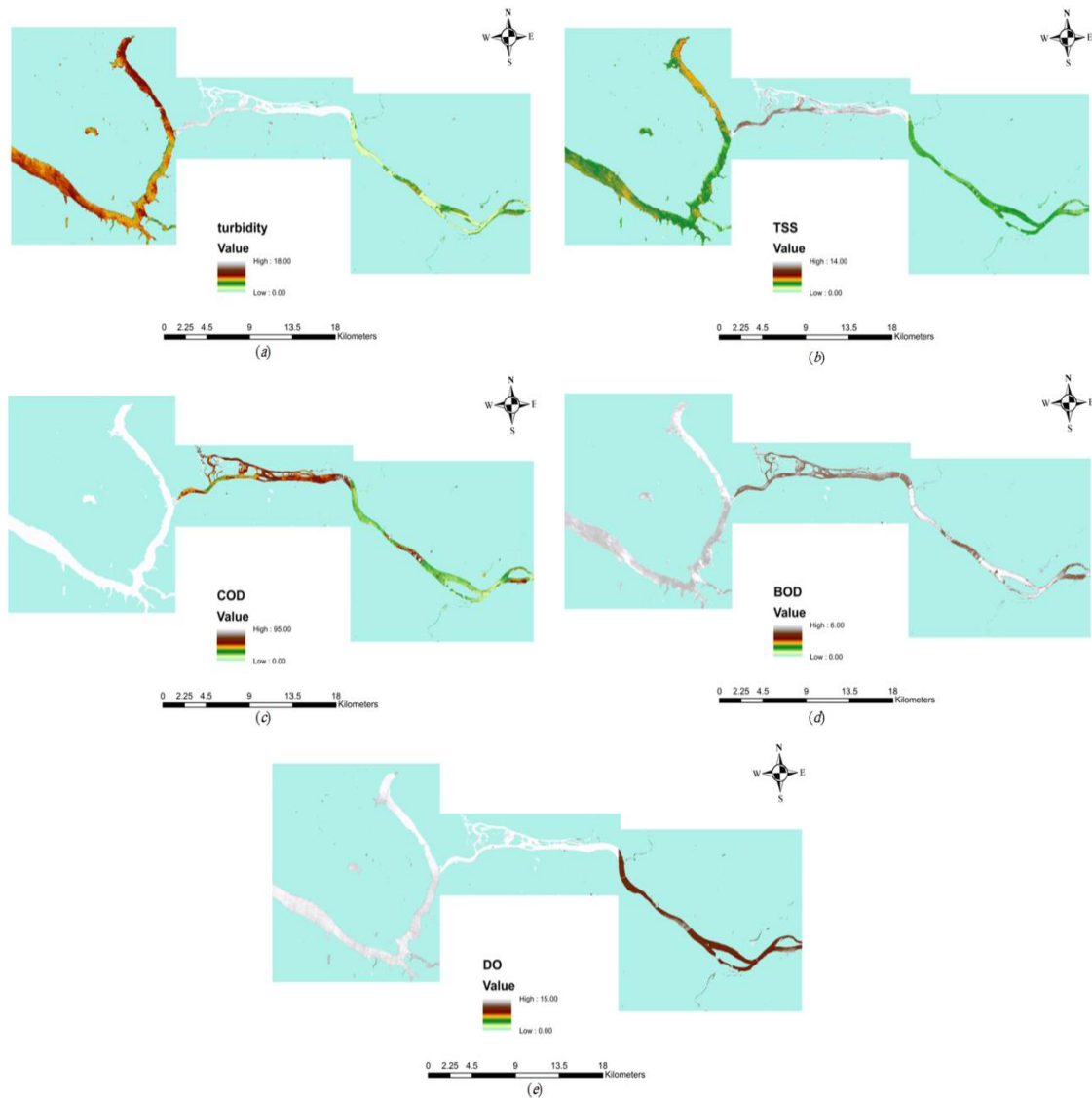


Figure 2.10 Spatial concentration maps for turbidity (a), TSS (b), COD (c), BOD (d), and DO (e) generated from the developed Landsat 8-based-SWR approach

For the entire study area of the SJR, the estimation values of turbidity, TSS, COD, BOD, and DO ranged from 0.900 to 18.000 NTU, 1.100 to 14.000 mg/l, 4.500 to 95.000 mg/l, 2.100 to 6.000 mg/l, and 6.500 to 15.000 mg/l, respectively. From the spatial distribution maps shown in **Figure 2.10**, it can be observed that concentrations of COD and BOD in the upper part of the selected study area of the SJR (i.e. above Mactaquac Dam) were clearly higher than those in the lower part of the river (i.e. below Mactaquac Dam) due to categorizing the upper part as an industrial region. On the other hand, the distribution pattern in concentrations of turbidity, TSS, and DO in the SJR depends mainly on sampling time. Accordingly, concentrations of turbidity, TSS, and DO in spring season were found to be higher than those sampled in summer owing to rainfall, snow melt, and low temperatures.

2.4 Conclusion

The overload of surface water pollutants can negatively affect both water quality and aquatic life. Because of these variations, the estimation of concentrations of optically and non-optically active SWQPs from space is essential to provide both spatial and temporal variability of water quality. Therefore, a remote sensing-based-SWR approach was developed to estimate concentrations of turbidity, TSS, COD, BOD, and DO using the spectral information of the Landsat 8 satellite data.

It was known that regression-based techniques have poor ability to model the complex or unknown relationships (i.e. the relationship between remotely sensed data and non-optical SWQPs which do not have direct optical properties and spectral characteristics). However, in our study, this problem was solved by correlating non-

optical SWQPs with optical variables, such as turbidity or TSS concentrations, which have direct optical properties that can be directly estimated by remote sensing means. After that, indirect relationships between satellite spectral data and concentrations of non-optical SWQPs were generated. As a result, the Landsat 8-based-SWR approach was found to be very efficient in developing highly accurate models to estimate both optical and non-optical SWQPs from space with $R^2 > 85\%$, which is very trustworthy.

Our study is very useful for local administrators, who have to make strict measures to protect surface water quality in their water bodies. In order to extra validate the stability/applicability of the developed Landsat 8-based-SWR approach, further studies, at different seasons, are needed to estimate concentrations of optical and non-optical SWQPs in either the remaining parts of the SJR or other water bodies. Finally, to further improve the accuracy of remote sensing estimation of SWQPs, another mapping tool (i.e. artificial intelligence), which is capable of modelling complex relationship between satellite spectral signatures and concentrations of SWQPs, is needed.

Acknowledgements

This work is jointly supported by the Egyptian Ministry of Higher Education and Scientific Research, Bureau of Cultural & Educational Affairs of Egypt in Canada, and the Canada Research Chair Program. The authors are grateful for the significant comments from the reviewers and the editor for improving this paper. The authors wish to acknowledge the USGS Landsat Archive Center for the Landsat 8 Level 1T imagery. The authors also express thanks to Prof. Dr. Katy Haralampides and Dr. Dennis Connor for their participation in the field work and laboratory analysis.

REFERENCES

- Alparslan, E., Aydoğan, C., Tufekci, V., & Tufekci, H. (2007). Water Quality Assessment at Ömerli Dam Using Remote Sensing Techniques. *Environ. Monit. Assess.*, *135*, pp. 391-398.
- APHA. (2005). *Standards Methods for the Examination of Water and Wastewater* (21th ed.). American Public Health Association Washington DC, USA.
- Arseneault, D. (2008). *The Road to Canada - Nomination Document for the St. John River, New Brunswick*. Prepared by The St. John River Society with the support of the New Brunswick Department of Natural Resources.
- Binding, C. E., Jerome, J. H., Bukata, R. P., & Booty, W. G. (2010). Suspended Particulate Matter in Lake Erie Derived from MODIS Aquatic Colour Imagery. *Int. J. Remote Sens.*, *31*, pp. 5239-5255.
- Bistani, L. F. (2009). *Identifying Total Phosphorus Spectral Signal in a Tropical Estuary Lagoon using a Hyperspectral Sensor and Its Application to Water Quality Modeling*. Doctoral thesis, University of Puerto Rico, Mayagüez Campus, Civil Engineering.
- Bresciani, M., Stroppiana, D., Odermat, D., Morabito, G., & Giardino, C. (2011). Assessing Remotely Sensed Chlorophyll-A for the Implementation of the Water Framework Directive in European Perialpine Lakes. *Sciences Total Environment*, *409* (17), pp. 3083-3091.
- Chang, C. W., Laird, D. W., Mausbach, M. J., & Gonos, H. (2001). Near-Infrared Reflectance Spectroscopy Principal Components Regression Analyses of Soil Properties. *Soil Science Society of America Journal*, *6*, pp. 480-490.
- Chavez, P. S. (1988). An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. *Remote Sensing of Environment*, *24*, pp. 459-479.
- Chen, S. S., Han, L. S., Chen, X. Z., Li, D., Sun, L., & Li, Y. (2015). Estimating Wide Range Total Suspended Solids Concentrations from MODIS 250-M Imageries:

- An Improved Method. *ISPRS Journal of Photogrammetry and Remote Sensing*, 99, pp. 58-69.
- D'SA, E., & Miller, R. (2003). Bio-optical properties in waters influenced by the Mississippi River during low flow conditions. *Remote Sensing of Environment*, 84, pp. 538-549.
- Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45, pp. 265-282.
- Earth Explorer*. (2016). Retrieved from U.S. Geological Survey: <http://earthexplorer.usgs.gov/>
- Floricioiu, D., Rott, H., Rott, E., Dokulil, M., & Defrancesco, C. (2004). Retrieval of limnological parameters of perialpine lakes by means of MERIS data. In *Proceedings of the 2004 Envisat & ERS Symposium (ESA SP-572)*, (pp. pp. 1-5). Salzburg, Austria.
- Gonca , C., Aysegul, T., Ugur, A., & Kerem, C. (2008). Determination of Environmental Quality of a Drinking Water Reservoir by Remote Sensing, GIS and Regression Analysis. *Water Air Soil Pollut*, 194, pp. 275-285.
- Gregory, W. C., & Dale, I. F. (2009). NONPARAMETRIC STATISTICS: AN INTRODUCTION. In *Nonparametric Statistics for Non-Statisticians* (p. p. 2). John Wiley & Sons, Inc.
- Güttler, F. N., Simona, N., & Francis, G. (2013). Turbidity Retrieval and Monitoring of Danube Delta Waters Using Multi-Sensor Optical Remote Sensing Data: An Integrated View from the Delta Plain Lakes to the Western–Northwestern Black Sea Coastal Zone. *Remote Sensing of Environment*, 132, pp. 86-101.
- He, W., Chen, S., Liu, X., & Chen, J. (2008). Water Quality Monitoring in Slightly-Polluted Inland Water Body through Remote Sensing-A Case Study in Guanting Reservoir, Beijing, China. *Frontiers Environment Sciences Engineering China*, 2 (2), pp.163-171.

- Hellweger, F. L., Schlossera, P., Lalla, U., & Weissel, J. K. (2004). Use of Satellite Imagery for Water Quality Studies in New York Harbor, *Estuar. Coast. Shelf Sci*, *61*, pp. 437-448.
- Kratzer, S., Brockmann, C., & Moore, G. (2008). Using MERIS full resolution data to monitor coastal waters – a case study from Himmerfjärden, a fjord-like bay in the northwestern Baltic Sea. *Remote Sensing of Environment*, *112*, pp. 2284-2300.
- Krista, A., Kersti, K., Reiko, R., Petra, P., Elar, A., Jan, P., & Anu, R. (2015). Satellite-based products for monitoring optically complex inland waters in support of EU Water Framework Directive. *International Journal of Remote sensing*, *36*, pp. 4446-4468.
- Lisa, S. (2016) “Power and Sample Size Determination,” Lecture Notes, Boston Univeristy, School of Public Health: http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_power/BS704_Power_print.html.
- Liu, D., Chin, C., Gong, J., & Fu, D. (2010). Remote Sensing of Chlorophyll-a Concentrations of the Pearl River Estuary from MODIS Land Bands. *International Journal of Remote Sensing*, *31*, pp. 4625-4633.
- Mancino, G., Nolè, A., Urbano, V., Amato, M., & Ferrara, A. (2009). Assessing Water Quality by Remote Sensing in Small Lakes: The Case Study of Monticchio Lakes in Southern Italy. *IForest*, pp. 154-161.
- Mao, Z. H., Chen, J. Y., Pan, D. L., Tao, B. Y., & Zhu, Q. K. (2012). A Regional Remote Sensing Algorithm for Total Suspended Matter in the East China Sea. *Remote Sensing of Environment*, *124*, pp. 819-831.
- Mcfeeters, S. K. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *INT. J. REMOTE SENSING*, *17* (7), pp. 1425-1432.
- Moreno-Madrinan, M. J., Al-Hamdan, M. Z., Rickman, D. L., & Muller-Karger, F. E. (2010). Using the Surface Reflectance MODIS Terra Product to Estimate Turbidity in Tampa Bay, Florida. *Int. J. Remote Sens.*, *2*, pp. 2713-2728.

- Moses, W., Gitelson, A., Berdnikov, S., & Povazhnyy, V. (2009a). Satellite estimation of chlorophyll-a concentration using the red and NIR bands of MERIS – the Azov Sea case study. *IEEE Geoscience and Remote Sensing Letters*, 6, pp. 845-849.
- Murdoch, P. S., Baron, J. S., & Miller, T. L. (2000). Potential Effects of Climate Change on Surface-Water Quality in North America. *J. Am. Water Resour. Assoc.*, 36 (2), pp. 347-366.
- NASA (Ed.). (2011). Landsat 7 Science Data Users Handbook Landsat Project Science Office at NASA's Goddard Space Flight Center in Greenbelt.
- Navalgund, R. R., Jayaraman, V., & Roy, P. S. (2007). Remote sensing applications: an overview. *Current Science*, 93, pp. 1747-1766.
- Nduwamungu, C., N., Z., L. E., P., G. F., T., & L., T. (2009). Opportunities For, and Limitations Of, near Infrared Reflectance Spectroscopy Applications in Soil Analysis: A Review. *Canadian Journal of Soil Science*, 89 (5), pp. 531-541.
- Odermatt, D., Heege, T., Nieke, J., Kneubuhler, M., & Itten, K. (2008). Water quality monitoring for Lake Constance with a physically based algorithm for MERIS data. *Sensors*, 8, pp. 4582-4599.
- Pat, S., & Chavez, J. (1996). Image-Based Atmospheric Corrections - Revisited and Improved. *Photogrammetric Engineering & Remote Sensing*, 62 (9), pp. 1025-1036.
- Poets, C., Costa, M. G., Da Silva, J. C., Silva, A. M., & Morais, M. (2010). Remote sensing of water quality parameters over Alqueva Reservoir in the south of Portugal. *International Journal of Remote sensing*, 12, pp. 3373-3388.
- Sharaf El Din, E., & Zhang, Y. (2017b). Statistical estimation of the Saint John River surface water quality using Landsat 8 multi-spectral data. *ASPRS Annual Conference. Proceedings of Imaging & Geospatial Technology Forum (IGTF)*. Baltimore, US.
- Sharaf El Din, E., & Zhang, Y. (2017c). Neural network modelling of the Saint John River sediments and dissolved oxygen content from Landsat OLI imagery. *ASPRS Annual Conference. Proceedings of Imaging & Geospatial Technology Forum (IGTF)*. Baltimore, US.

- Simis, S., Peters, S., & Gonos, H. (2005). Remote sensing of the cyanobacterial pigment phycocyanin in turbid inland water. *Limnology and Oceanography*, *50*, pp. 237-245.
- Sitanshu, S. K., & Archana, R. (2013). IS 30 THE MAGIC NUMBER? ISSUES IN SAMPLE SIZE ESTIMATION. *National Journal of Community Medicine*, *4*(1), pp. 175-179.
- Song, C., Woodcock, C. E., Seto, K. C., Lenney, M. P., & Macomber, S. A. (2001). Classification and change detection using Landsat TM data: when and how to correct atmospheric effects. *Remote Sensing of Environment*, *75*, pp. 230-244.
- United States Geological Survey (USGS). (2016). Retrieved from USGS Landsat 8 Product: http://landsat.usgs.gov/Landsat_8_Using_Product.php.
- United States Geological Survey (USGS) Landsat 8 Surface Reflectance Product Guide. (2018).
- Wang, F., Han, L., Kung, T., & Van Arsdale, R. B. (2006). Applications of Landsat-5 TM Imagery in Assessing and Mapping Water Quality in Reelfoot Lake, Tennessee. *Int. J. Remote Sens.*, *27*, pp. 5269-5283.
- Xiang, Y., Huapeng, Y., Xiangyang, L., Yebao, W., Xin, L., & Hua, Z. (2016). Remote-sensing estimation of dissolved inorganic nitrogen concentration in the Bohai Sea using band combinations derived from MODIS data. *International Journal of Remote Sensing*, *37*:2, pp. 327-340.
- Yacobi, Y. Z., Moses, W. J., Kaganovsky, S., Sulimani, B., Leavitt, B. C., & Gitelson, A. A. (2011). NIRRed Reflectance-Based Algorithms for Chlorophyll-A Estimation in Mesotrophic Inland and Coastal Waters: Lake Kinneret Case Study. *Water Research* *45*, doi:10.1016/j.watres.2011.02.002., pp. 2428-2436.
- Zhang, Y. Z., Pulliainen, J. T., Koponen, S. S., & Hallikainen, M. T. (2002). Application of an empirical neural network to surface water quality estimation in the Gulf of Finland using combined optical data and microwave data. *Remote Sensing of Environment*, *81*, pp. 327-336.

Chapter 3: MAPPING CONCENTRATIONS OF SURFACE WATER QUALITY PARAMETERS USING A NOVEL REMOTE SENSING AND ARTIFICIAL INTELLIGENCE FRAMEWORK²

Abstract

The deterioration of surface water quality occurs due to the presence of various types of pollutants generated from human, agricultural, and industrial activities. Thus, mapping concentrations of different surface water quality parameters (SWQPs), such as turbidity, total suspended solids (TSS), chemical oxygen demand (COD), biological oxygen demand (BOD), and dissolved oxygen (DO), is indeed critical for providing the appropriate treatment to the affected water bodies. Traditionally, concentrations of SWQPs have been measured through intensive field work. Additionally, quite a lot of studies have attempted to retrieve concentrations of SWQPs from satellite images using regression-based methods. However, the relationship between concentrations of SWQPs and satellite spectral data is too complex to be modelled accurately by using regression-based methods. Therefore, our study attempts to develop an artificial intelligence modelling method for mapping concentrations of both optical and non-optical SWQPs. In this context, a remote sensing framework based on the back-propagation neural network (BPNN) is developed for the first time to quantify concentrations of SWQPs from the

² This paper has been published in the “*International Journal of Remote Sensing (IJRS)*”:
Sharaf El Din, E., Zhang, Y., & Suliman, A. (2017). Mapping concentrations of surface water quality parameters using a novel remote sensing and artificial intelligence framework. *International Journal of Remote Sensing*, 38 (4), pp. 1023-1042. <http://dx.doi.org/10.1080/01431161.2016.1275056>.

Landsat 8 satellite imagery. Compared to other methods, such as support vector machine, significant coefficients of determination (R^2) between the Landsat 8 surface reflectance and concentrations of SWQPs were obtained using the developed Landsat 8-based-BPNN models. The resulting R^2 values ≥ 0.93 for turbidity, TSS, COD, BOD, and DO. Indeed, these findings indicate that the developed Landsat 8-based-BPNN framework is capable of developing highly accurate models for retrieving concentrations of different SWQPs from the Landsat 8 imagery.

3.1 Introduction

In the past few decades, the increase of anthropogenic activities, especially in industrial areas, has negatively affected water bodies. Accordingly, the result can be a reduction in water storage capacity or in rivers' ability to support aquatic life. This shortage of water which has increased over the past years is expected to continue in the future (Gaballah, Khalaf, Beckand, & Lopez, 2005). Thus, to help the decision-makers in taking the right action at the right time, the relevant information systems require continuously updated information about water quality (WQ).

WQ changes as water flows through different land-use surfaces (Bolstad & Swank, 1997). These surfaces define the type and amount of surface water quality parameters (SWQPs) of surface water that flows into water bodies (Moss, 1998). The deterioration of surface WQ occurs due to the runoff from the activities on various types of land-use surfaces (e.g., agricultural, residential, commercial, and industrial activities) into water bodies. In addition to that, the climate variations due to the global warming can lead to floods, drought, biodiversity loss, and an increase in the infectious diseases

that can degrade WQ (Murdoch, Baron, & Miller, 2000). Because of these continuous changes in WQ, regular monitoring and estimation of optical and non-optical SWQPs on a large scale is indeed critical for providing the targeted treatment to a specific water body. Thus, remote sensing technology is found to be an appropriate tool for estimating concentrations of SWQPs and potentially offers wide spatial coverage as well as detecting temporal changes.

Bearing that in mind, the relevant research about estimating concentrations of SWQPs from space is reviewed. Accordingly, remote sensing estimation of concentrations of SWQPs, especially optical SWQPs, is achievable via regression-based and learning-based techniques. While most of the available publications were based on exploring regression techniques, relatively fewer research attempts focused on learning-based algorithms. In this context, for instance (He, Chen, Liu, & Chen, 2008; Yang, Liu, Ou, & Yuan, 2011; Nathan, Sarah, Stephen, Narumon, Brian, & Jianguo, 2013; Matias, María, Lucio, & Susana, 2015), (Krista, et al., 2015; Xiang, Huapeng, Xiangyang, Yebao, Xin, & Hua, 2016; Yunlin, Kun, Yongqiang, Xiaohan, & Boqiang, 2016), and (Darryl, Blake, Ross, Richard, Kenneth, & Donald, 2014; Tiit, Krista, Dolly, & Stephan, 2015) have used the Landsat Thematic Mapper (TM), Moderate Resolution Imaging Spectroradiometer (MODIS), and Medium Resolution Imaging Spectrometer (MERIS), respectively, to develop regression models to estimate concentrations of SWQPs.

Theoretically, WQ is complex to have a simple relationship with satellite spectral signatures (Xiang, Huapeng, Xiangyang, Yebao, Xin, & Hua, 2016). Moreover, it is challenging for regression-based techniques to model such a complex relationship between satellite reflectance and concentrations of different SWQPs, especially non-

optical parameters, such as chemical oxygen demand (COD), biological oxygen demand (BOD), and dissolved oxygen (DO). Therefore, our focus in this research is to explore an appropriate learning-based algorithm in mapping both optical and non-optical SWQPs from the Landsat 8 data since these data are freely available and acquired by a recent satellite sensor.

Based on our literature review, we have focused on the competence and performance ability of learning-based techniques to retrieve concentrations of different SWQPs from the Landsat 8 Operational Land Imager (OLI) images. In this article, the selected learning-based technique is the back-propagation neural network (BPNN) since it has been proved in the literature to be successful in the applications of remote sensing image classification and pattern recognition (Suliman & Zhang, 2014). However, almost all of the available research about estimating concentrations of different SWQPs using artificial neural networks (ANNs) is mainly based on two learning-based algorithms: Levenberg-Marquardt (LM), and Cascade Correlation (CC). Consequently, the published work based on these two learning algorithms is reviewed.

The LM learning algorithm has been used to develop empirical models for estimating WQ parameters of chlorophyll, total suspended solids (TSS), turbidity, and secchi disk depth in the Gulf of Finland (Zhang, Pulliainen, Koponen, & Hallikainen, 2002). The inputs of multi-layer perceptron (MLP) network were the digital numbers from the Landsat-5 TM and Synthetic Aperture Radar (SAR) bands, while the outputs were the selected WQ parameters. The determination coefficients (R^2) of the network testing data were 0.84, 0.92, 0.94, and 0.96 for chlorophyll, total suspended solids, turbidity, and secchi disk depth, respectively. In another study, chlorophyll was

investigated using the Landsat-5 TM data across Tucuruí reservoir, Brazil (Ribeiro, Almeida, Rocha, & Krusche, 2008). A model based on the MLP architecture and the radial base function was developed to predict chlorophyll concentrations. The R^2 for chlorophyll testing dataset was 0.92. Another study used the Landsat-5 TM imagery and McCulloch and Pitt's neuron model to quantify chlorophyll, turbidity, and total phosphorus over Kissimmee River in Florida (Yirgalem, 2012). The root mean square error (RMSE) for chlorophyll, turbidity, and phosphorus was below 0.170 mg m^{-3} , 0.500 NTU, and 0.030 mg l^{-1} , respectively, for the validation data. Water samples were gathered with MLP neural network to retrieve suspended sediments from MODIS imagery (Ali, et al., 2013). A robust relationship between MODIS bands 1 and 2 and water samples was established based on a three-layer ANN with six neurons in the hidden layer. The R^2 for suspended sediments was 0.85 for all of the data used.

The CC learning algorithm has been utilized to derive empirical models for estimating and predicting the monthly values of different surface water parameters, such as power of hydrogen (pH) and electrical conductivity (EC), over the Axios River, Greece (Diamantopoulou, Antonopoulos, & Papamichail, 2007). In this study, the ANN training was achieved by using the CC algorithm along with the MLP architecture. The CC algorithm starts the training without any hidden nodes. If the error between the actual output and the targeted output is higher than a defined threshold, it adds one hidden node. This node is connected to all other nodes except the output nodes. The optimal number of the hidden nodes is commonly determined by trial and error. The results showed that the best architecture of the proposed network was composed of one input layer with nine

input variables, one hidden layer with nineteen nodes, and one output layer with one output variable. The R^2 values for both pH and EC were > 0.87 .

Based on what has been reviewed, the use of the LM and CC learning-based algorithms has been suggested in most cases because they are considered as quicker training algorithms. However, the LM learning-based algorithm works well and fast only if the error surface is smooth (i.e. with no local minima); otherwise there is no guarantee to find the global minima (Holger, Ashu, Graeme, & Sudheer, 2010). Also, the CC learning-based algorithm is supposed to be efficient in solving regression problems; however, its propensity to overfit on the training data is considered as a critical disadvantage (Tetko & Villa, 1997).

Additionally, compared to other machine learning-based methods, such as support vector machine (SVM), highly accurate remote sensing estimation models of both optical and non-optical SWQPs can be obtained using the proposed Landsat 8-based-BPNN. As for SVM, it has the defect of parameter selection because of the absence of theoretical guidance (i.e. there is no rule or even a guideline for SVM parameter selection). Moreover, SVM uses quadratic programming to solve the support vector and the process is complex, especially in large-scale applications. Furthermore, the most serious drawback with SVM is the high algorithmic complexity and extensive memory requirements of the required quadratic programming. Additionally, an important practical problem that is not entirely solved is the selection of the optimum kernel function and its corresponding parameters (Valyon & Horvath, 2004).

In contrast, the BPNN learning-based algorithm can result in good generalization when small, large, or even noisy datasets are used (MacKay, 1992). The BPNN can

overcome the defects of slow training speed, poor generalization ability, and low learning accuracy reported in other learning-based techniques. This means the BPNN can not only satisfy the accuracy requirements, but also improve the learning efficiency. Using the BPNN, the validation dataset can be utilized to decide when to stop training in order to avoid overfitting (MacKay, 1992). As an important pattern recognition algorithm, the BPNN is found to be an appropriate tool for WQ assessment, which is a typical pattern recognition problem. Even though the BPNN algorithm has many advantages, the local minima is considered as the most critical problem in the error surface. But, this problem can be solved by choosing an appropriate learning rate to achieve the global minima in the error surface. Therefore, in our study, the BPNN algorithm is proposed to retrieve concentrations of optical and non-optical SWQPs from the Landsat 8 satellite data.

The identified objectives of this research are as follows: (1) to develop a Landsat 8-based-BPNN framework for mapping concentrations of SWQPs from the Landsat 8 satellite data, and (2) to produce a spatial distribution map for each optical and non-optical SWQP over each pixel of the selected study area. To the best of our knowledge, our Landsat 8-based-BPNN framework is the first to map concentrations of SWQPs, especially non-optical parameters, with highly accurate results, compared to regression-based or even other learning-based techniques.

3.2 Artificial Neural Network (ANN) Background

An ANN is a paradigm adapted to mimic the biological neurons using a computing process. The important feature of all ANN types is the adaptive nature, where they learn by examples instead of the use of conventional programming procedures to

solve complex problems (Hinton, 1992; Jain, Mao, & Mohiuddin, 1996). Among various types of ANN architectures, the most widely used type, especially in classification processes, is the MLP network.

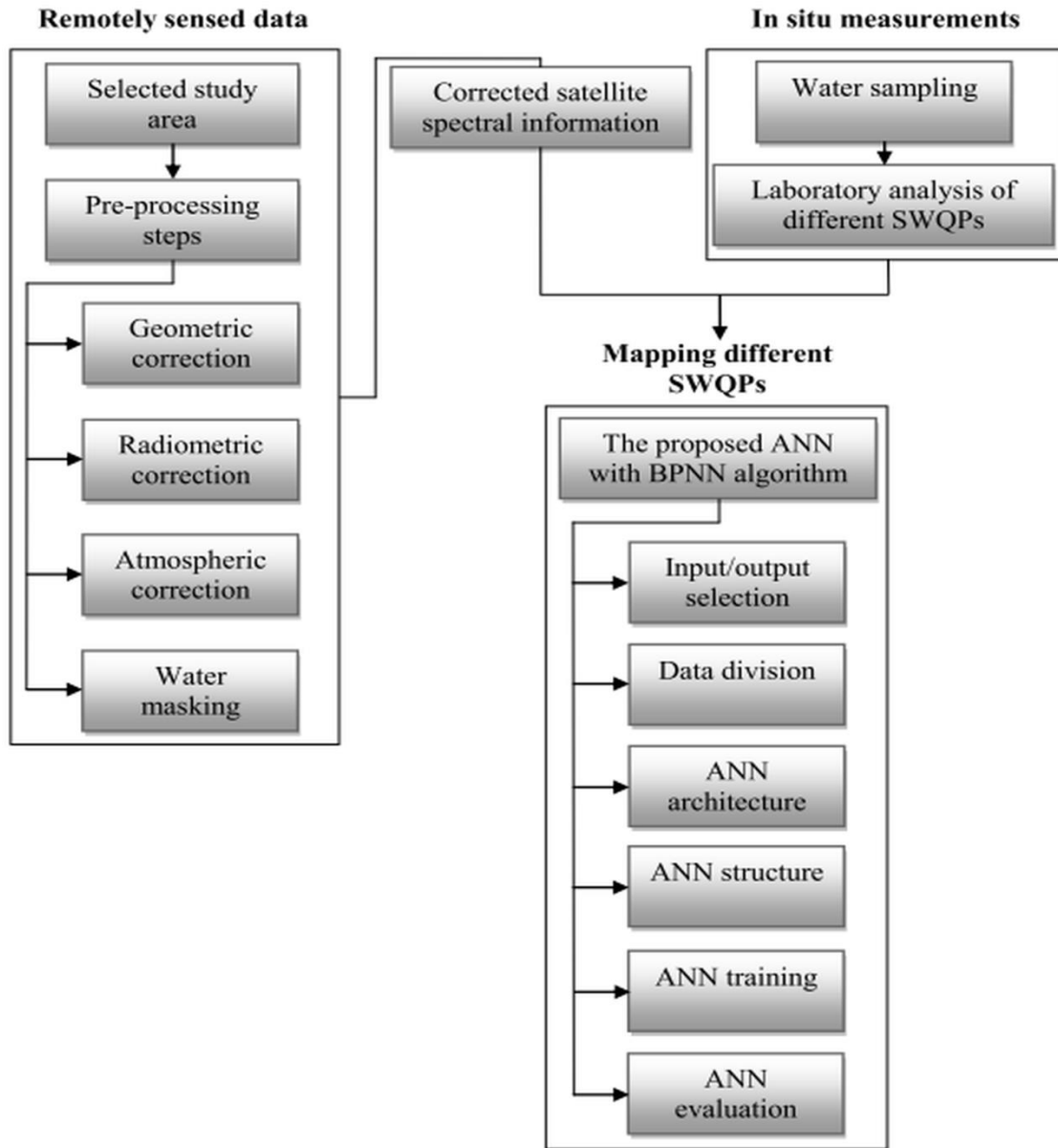


Figure 3.1 The flowchart of retrieving concentrations of different SWQPs from satellite data by using the proposed Landsat 8-based-BPNN

The MLP is organized in layers of computing elements, known as neurons, which are connected between layers via weights. The MLP networks are closely related to statistical models and are the most suited for forecasting applications (Rumelhart, Hinton, & Williams, 1986; Hinton, 1992; Alsmadi, Omar, & Noah, 2009). Basically, one of the most common ANN algorithms, especially in classification and pattern recognition applications, is the BPNN algorithm. More details about the ANN technology and terminology, and the BPNN algorithm are provided in (Hinton, 1992; Suliman & Zhang, 2014).

3.3 Materials and Methods

The method of retrieving concentrations of different SWQPs from satellite data by using the proposed Landsat 8-based-BPNN is flowcharted in **Figure 3.1**. This section is devoted to describing in detail the study area of the Saint John River (SJR), the processing steps of remotely sensed data, the water sampling and laboratory analysis, and mapping concentrations of optical and non-optical SWQPs by using the proposed BPNN algorithm.

3.3.1 Remotely Sensed Data

3.3.1.1 Study Area

The selected study area is a part of the SJR which is approximately 673km long, located principally in the Canadian province of New Brunswick, as shown in **Figure 3.2**. Around 35% and 13% of the SJR watershed is located in the US state of Maine and the

Canadian Province of Quebec, respectively. The remaining 52% of the watershed lies within New Brunswick, covering an area of 4748 km² (Arseneault, 2008).

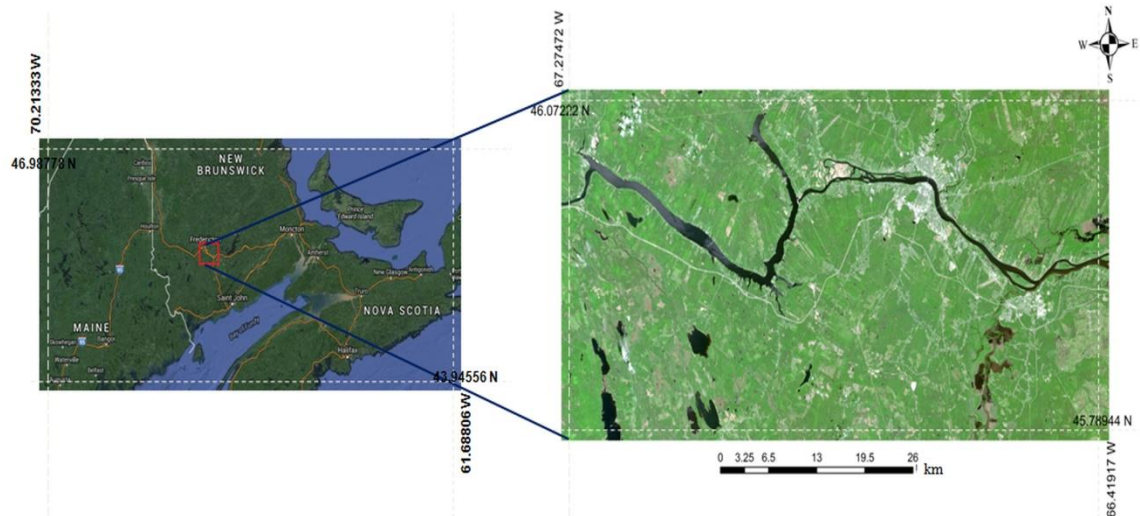


Figure 3.2 The selected study area of the SJR, New Brunswick, Canada (Earth Explorer, 2016)

3.3.1.2 Satellite Processing Steps

The full Landsat 8 scenes are available free of charge at Level 1T (terrain corrected) at Landsat websites maintained by the US Geological Survey (USGS). The three Landsat 8 satellite sub-scenes used in our study were acquired on 27 June 2015, 10 April 2016, and 12 May 2016. Basically, the Level 1T product is a geometrically corrected image and rectified to the Universal Transverse Mercator (UTM) projection, World Geodetic System 1984 (WGS 84) datum. Digital numbers (DNs) of the Landsat 8 satellite images are stored in 16 bits unsigned integer format, and were subsequently corrected to obtain the top of atmospheric (TOA) reflectance using radiometric rescaling coefficients.

In order to remove the effects of the atmosphere, surface reflectance values were calculated using the Dark Object Subtraction (DOS) method (Chavez, 1988). This method is found to be very efficient in discriminating and mapping wetland areas and well accepted by the geospatial community to correct light scattering in remote sensing data (Song, Woodcock, Seto, Lenney, & Macomber, 2001). Other atmospheric correction methods, such as second simulation of the satellite signal in the solar spectrum (6S) and atmospheric and topographic correction (ATCOR), have indeed been used in remote sensing research field. However, the main disadvantage of these methods is that they require extensive field measurements during each satellite pass. This is unacceptable for various applications and is often impossible, as when using historical data or when working in very remote or difficult access locations (Pat & Chavez, 1996).

Finally, to delineate concentrations of different SWQPs over any water body, the water surface was masked using the Normalized Difference Water Index method (Mcfeeters, 1996).

3.3.2 In situ Measurements

In this study, 39 water sample points were randomly selected and distributed over the whole study area during three field trips in 27 June 2015, 10 April 2016, and 12 May 2016, as shown by dots in **Figure 3.3**. One sample was excluded due to cloud coverage. Coordinates of the sample points were recorded in the field through a handset global positioning system (GPS), GARMIN 76CSx. The three sampling events were selected at different seasons (i.e. summer and spring) to best represent the maximum variation in the concentrations of both optical and non-optical SWQPs.

In order to carry out this study efficiently, the representative water samples were collected just beneath the water surface (i.e., 30 to 50 cm) and at the same acquisition time of the full Landsat 8 scenes over the selected study area. At each station, turbidity, TSS, COD, BOD, and DO were measured and analyzed according to the standard methods for lab examination of water and wastewater suggested by the American Public Health Association (APHA) (APHA, 2005).

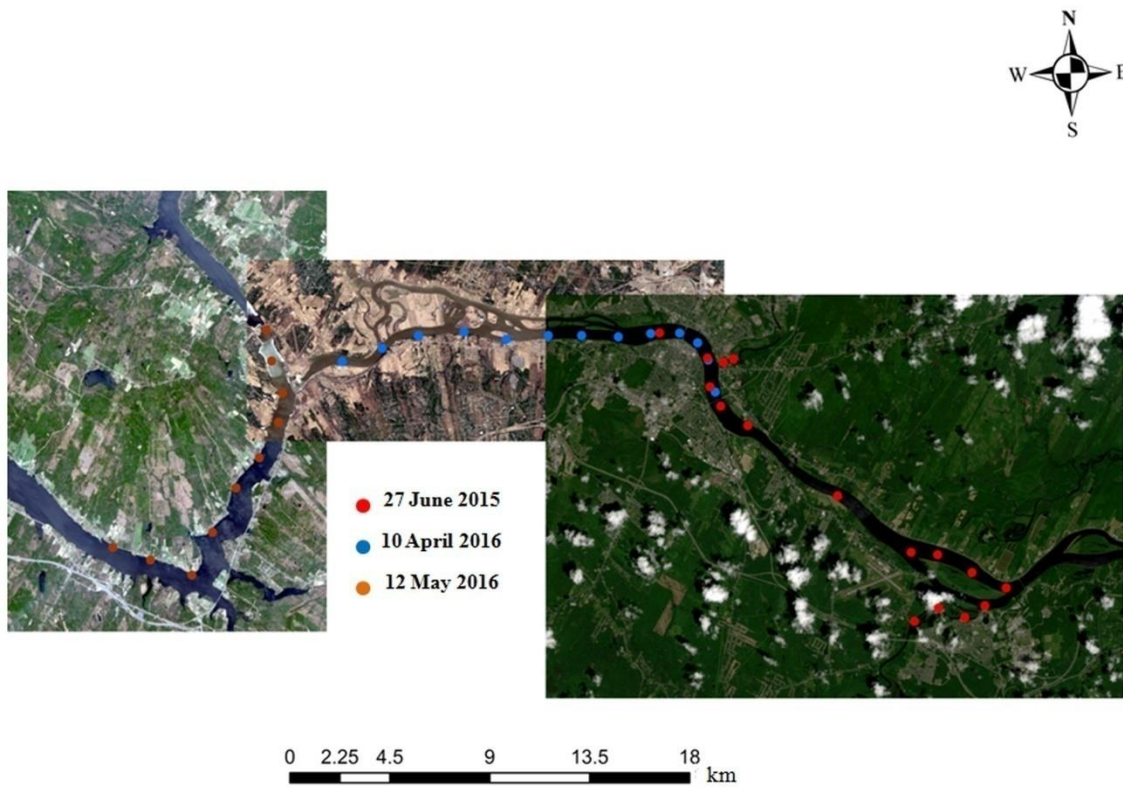


Figure 3.3 The Landsat 8 satellite sub-scenes of the study area with sampling locations

3.3.3 Mapping Concentrations of SWQPs using the BPNN Algorithm

The BPNN is one of the most popular learning-based algorithms utilized in remote sensing applications; however, it is not well known in the WQ research field. The proposed feed-forward BPNN algorithm was adopted, as shown in **Figure 3.4**, to model

the unknown relationship between concentrations of SWQPs and the Landsat 8 surface reflectance information. The main steps of developing the BPNN models are given in the following subsections.

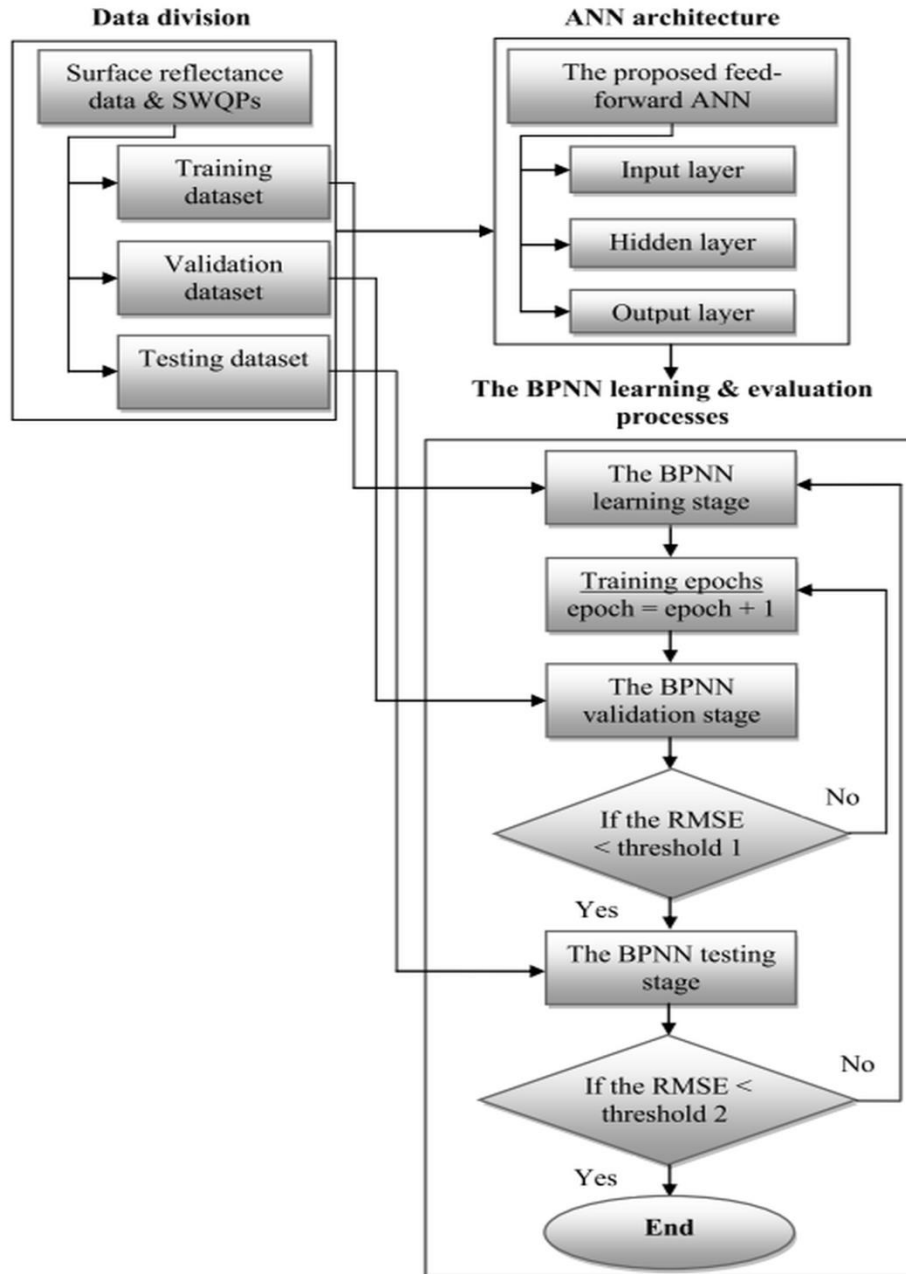


Figure 3.4 The flowchart of applying the proposed BPNN algorithm

3.3.3.1 ANN Input and Output Selection

The first step in the development of the Landsat 8-based-BPNN models is the choice of potential model input variables from the available data and a set of appropriate model outputs. Basically, a number of techniques are available for assessing the significance of the relationship between potential model inputs and outputs. These techniques are mainly subdivided into two basic approaches: model-based and model-free approaches (Maier & Dandy, 2000). The primary disadvantage of the model-based approach is that it is time consuming, as a model structure, training, and evaluation have to be developed for several times before deciding which one is the best.

Consequently, in our study, the model-free approach was utilized and the inputs were selected based on the Landsat 8 multi-spectral information, while SWQP concentrations, one at a time, were selected to form the network outputs. Generally, when a model-free approach is used, a statistical measure of significance is calculated to measure the strength of the relationship between potential model inputs and outputs.

3.3.3.2 ANN Data Division

As part of the Landsat 8-based-BPNN models development process, the available data were normally subdivided into calibration (i.e. training and testing) and validation datasets. The training set was utilized to determine the connection weights, the testing set was used to assess the generalization ability of the trained model, and the validation set was used to decide when to stop training to avoid overfitting. Basically, the methods of dividing the available data into subsets can be divided into random and statistical data division approaches. In WQ studies, the training, testing, and validation datasets should

have the same statistical properties in order to develop the best possible input-output model. In this context, the statistical data division approach was utilized and the available data were subdivided into their respective subsets.

3.3.3.3 ANN Architecture Selection

The neural network architecture determines the overall structure and information flow in ANN models. Thus, it has a significant impact on the functional form of the relationship between model inputs and outputs. Normally, ANN architectures are divided into feed-forward and recurrent networks (Graupe, 2007). The MLP neural network is the most common form of feed-forward model architecture.

In our study, the feed-forward MLP with only three layers was utilized along with a linear aggregation function and a sigmoid function. Basically, the input layer neurons simply passed on the weighted inputs to the hidden and output layer neurons. Additionally, using a sigmoid function can provide the capability of modelling complex relationships between the model inputs and outputs.

3.3.3.4 ANN Structure Selection

The neural network structure, along with the neural network architecture, defines the functional form of the input-output relationship. Determination of an appropriate network structure involves the selection of an appropriate number of hidden neurons and how they process the incoming signals by using a suitable transfer function. The optimal network structure generally creates a balance between the network generalization, processing speed, and complexity.

In our study, the number of hidden neurons was identified by various trials because there is no stable guideline to optimize the number of the hidden neurons. Using too few neurons may lead to underfitting, while using too many may cause overfitting. In order to avoid any overfitting during the training stage, a cross validation procedure was performed by keeping track of the competence of the fitted model. Additionally, a sigmoid function was utilized with the BPNN algorithm because it is differentiable and can provide closer similarity to the biological neuron than do threshold functions (Suliman & Zhang, 2014).

3.3.3.5 ANN Training

The aim of ANN training is to find a set of connection weights that enables the network with a given functional form to best represent the targeted input-output relationship. Generally, ANN training is performed using an appropriate optimization algorithm. The majority of these algorithms can be subdivided into deterministic and stochastic. Deterministic techniques attempt to identify a single parameter vector that minimizes the measured error signal between both the actual and desired outputs. Basically, these methods belong to either local (e.g. BPNN algorithm) or global optimization approaches (e.g. Newton's algorithm).

In our study, the BPNN algorithm was utilized because this method is computationally efficient and can control the learning process by utilizing an appropriate learning rate to achieve the global minimum error. Actually, using too small a value for learning rate may lead to slow learning, while using too large a value may cause instability or poor performance. Additionally, as shown in **Equation (3.1)**, the network

training was assessed at the output layer by using both the actual and desired output signals (Suliman & Zhang, 2014).

$$E = 0.50 \times \sum_k (e_k)^2 = 0.50 \times \sum_k (T_k - O_k)^2 \quad (3.1)$$

where k is the index of the output layer of the network; e_k is the error signal; T_k is the desired output; O_k is the network actual output.

3.3.3.6 ANN Evaluation

In order to determine which network structure is optimal, the performance of a calibrated model is evaluated using statistical criteria. The ANN model performance is usually assessed using a quantitative error metric. In our study, the performance of the developed BPNN models was evaluated based on the coefficient of determination (R^2), root mean square error (RMSE), and significant value (p -value).

3.4 Results and Discussion

Our study aims at estimating concentrations of both optical and non-optical SWQPs from satellite data. To achieve this objective, a methodology based on developing Landsat 8-based-BPNN models was developed to retrieve concentrations of turbidity, TSS, COD, BOD, and DO from the Landsat 8 satellite data. Consequently, a spatial distribution map showing concentrations of each SWQP was generated over the entire study area. The results obtained from this study include (1) concentrations of optical and non-optical SWQPs, (2) estimation and validation of the developed Landsat 8-based-BPNN models, (3) producing a spatial concentration map for each SWQP over

the entire study area, and (4) comparison of other model results, such as support vector machine (SVM). These results are discussed in the following subsections.

3.4.1 Concentration Results of both Optical and Non-optical SWQPs

Thirty-nine water samples were collected over 70 km of the SJR and analyzed for optically and non-optically active SWQPs. The statistics, shown in **Table 3.1**, for turbidity, TSS, COD, BOD, and DO were measured from the collected water samples. The density of samples per km collected in our study is higher than that used in previous studies. For instance, only 11 samples were used by [Yirgalem \(2012\)](#) to capture water quality variables over 37 km of the Kissimmee River.

Table 3.1 Statistics of the concentrations of SWQPs along the study site.

SWQPs	Minimum (Min)	Maximum (Max)	Mean	Standard deviation (SD)
Turbidity (NTU)	1.190	13.100	6.303	4.327
TSS (mg l ⁻¹)	1.200	11.400	4.781	3.617
COD (mg l ⁻¹)	4.800	86.640	29.550	22.803
BOD (mg l ⁻¹)	1.110	3.250	1.707	0.504
DO (mg l ⁻¹)	6.990	14.140	11.062	2.517

The correlation coefficient (r) between the measured parameters was calculated and populated in a matrix form as shown in **Table 3.2**. Based on this correlation matrix, the relationship between turbidity and all SWQPs except DO was positively correlated. On the other hand, the r values between DO levels and turbidity, TSS, COD, and BOD were -0.816, -0.824, -0.838, and -0.776, respectively. Moreover, the relationship between

turbidity and TSS was highly correlated, and this is because suspended sediment is considered as the dominant indicator of turbidity.

Table 3.2 The correlation coefficient (r) matrix of both optical and non-optical SWQPs.

	Turbidity	TSS	COD	BOD	DO
Turbidity	1.000	0.976	0.857	0.799	-0.816
TSS	0.976	1.000	0.861	0.850	-0.824
COD	0.857	0.861	1.000	0.895	-0.838
BOD	0.799	0.850	0.895	1.000	-0.776
DO	-0.816	-0.824	-0.838	-0.776	1.000

3.4.2 Estimation and Validation of the Landsat 8-based-BPNN Developed Models

The main steps of developing the Landsat 8-based-BPNN models, as well as the way the data flow through and the outcomes achieved, are given in the following subsections.

3.4.2.1 ANN Input and Output Selection

The model-free approach was utilized and the r values were used to assess the strength of the relationship between model inputs and outputs. In this context, coastal blue (CB), blue (B), green (G), red (R), near-infrared (NIR), shortwave infrared 1 (SWIR1), and shortwave infrared 2 (SWIR2) multi-spectral bands were selected to form the input layer. As shown in **Table 3.3**, these multi-spectral bands were significantly correlated (i.e. $r \geq 0.50$) with all SWQPs concentrations used in our study. However, the

rest of the Landsat 8 bands such as Cirrus, thermal infrared 1 (TIR1), and thermal infrared 2 (TIR2) were less correlated (i.e. $r < 0.50$) to the selected SWQPs. The reason for achieving less r values between TIR1 and TIR2 bands and SWQPs is that, these bands are mainly designed for detecting surface temperatures; while, Cirrus is commonly used for detecting clouds.

Table 3.3 The r values between the Landsat 8 multi-spectral bands and concentrations of SWQPs.

	Turbidity	TSS	COD	BOD	DO
CB	0.792	0.807	0.752	0.742	-0.761
B	0.605	0.642	0.555	0.549	-0.644
G	0.631	0.668	0.597	0.615	-0.612
R	0.654	0.671	0.605	0.608	-0.590
NIR	0.844	0.887	0.839	0.810	-0.871
SWIR1	0.827	0.799	0.704	0.711	-0.777
SWIR2	0.821	0.810	0.695	0.700	-0.746
Cirrus	0.484	0.432	0.401	0.441	-0.452
TIR1	0.424	0.475	0.411	0.439	-0.408
TIR2	0.438	0.452	0.429	0.433	-0.399

3.4.2.2 ANN Data Division

The statistical data were selected as the proposed data division approach. In coastal studies, all datasets should have the same statistical properties, as much as

possible, to best represent the input-output relationship. Thus, a trial and error procedure was used to divide the available data in such a way that the statistical properties of each subset are close to those of other subsets. The mean, maximum, minimum, and standard deviation were the statistical metrics used to perform this task. Moreover, the proportion of the data to be utilized for training, testing, and validation was selected in advance by the modeller. Sixty percent of water samples (i.e. 22 samples) were utilized for training, while 20% (i.e. 8 samples) were used for testing and the other 20% (i.e. 8 samples) for validation. The data division approach used in our study is quite similar to that of a previous study conducted by [Yirgalem \(2012\)](#).

3.4.2.3 ANN Architecture Selection

The feed-forward MLP was selected as the proposed ANN architecture since it is highly successful in classification and pattern recognition applications. The proposed architecture consisted of three layers (i.e. input, hidden, and output) with a sigmoid activation function which is proved to be sufficient for nonlinear modelling purposes. While the number of the input neurons was selected to be equal to the selected input bands of the Landsat 8 image, the number of the output neurons was selected to be one at a time since we are building an ANN for prediction purposes.

As shown in **Figure 3.5**, seven neurons (i.e. CB, B, G, R, NIR, SWIR1, and SWIR2) were used to form the input layer, while one SWQP at a time was used to outline the output layer. The main basis of using one SWQP at a time is to accelerate the computations of the developed models and to diminish the complexity of the ANN.

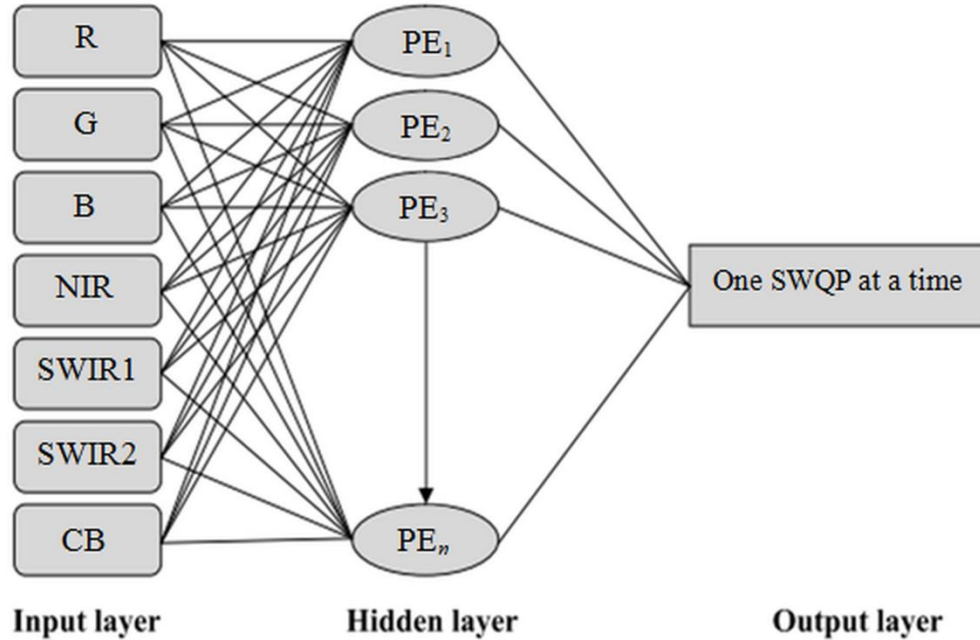


Figure 3.5 The architectural design of the proposed ANN

3.4.2.4 ANN Structure Selection

An appropriate selection of the proper number of hidden neurons besides a suitable transfer function is indeed a critical task. In our study, 20 processing elements (PEs) were experimentally selected to form the hidden layer. The trial and error procedure initially started with two hidden neurons, and then the number of hidden neurons was increased incrementally to thirty. Actually, using less than twenty neurons led to an underfitting problem (i.e. being unable to learn what you want the network to learn), while using more than twenty resulted in slow learning and overfitting problems.

Additionally, a sigmoid transfer function was utilized with the BPNN algorithm because it is differentiable and can provide the powerful capability of modelling the complexity inherent in the system. Once the sigmoid function was used, the model input and output were scaled appropriately to fall within the function limits [0.00 to 1.00] to

avoid the saturation problem in the training stage. The input scaling has been performed using surface reflectance values of the Landsat 8 spectral bands.

3.4.2.5 ANN Training and Evaluation

The BPNN algorithm was utilized as the proposed learning-based algorithm since it is widely used and is found to be very efficient in remote sensing and digital image processing applications. Moreover, the BPNN algorithm was found to be computationally efficient as 1, 4, 2, 1, and 3 seconds were achieved, at the network training phase, for turbidity, TSS, COD, BOD, and DO, respectively. Furthermore, finding the global minima was guaranteed by utilizing an appropriate learning rate. In our study, a learning rate value of 0.01 was adjusted to achieve the minimum error in the error function. Actually, by using a learning rate value beyond 0.005, the ANN computational speed was very slow; however, by using a learning rate value above 0.10, the performance and generalization ability of the proposed ANN were very poor.

Table 3.4 Statistical measures between the target and actual concentrations of SWQPs using the developed Landsat 8-based-BPNN.

SWQPs	R^2 (training)	RMSE (training)	R^2 (validation)	RMSE (validation)	R^2 (testing)	RMSE (testing)
Turbidity (NTU)	1.000	0.305	0.979	0.073	0.991	0.069
TSS (mg l ⁻¹)	0.906	0.092	0.976	0.226	0.933	0.999
COD (mg l ⁻¹)	0.963	0.285	0.918	0.158	0.937	0.877
BOD (mg l ⁻¹)	0.937	0.034	0.941	0.042	0.930	0.076
DO (mg l ⁻¹)	0.985	0.073	0.942	0.188	0.934	0.455

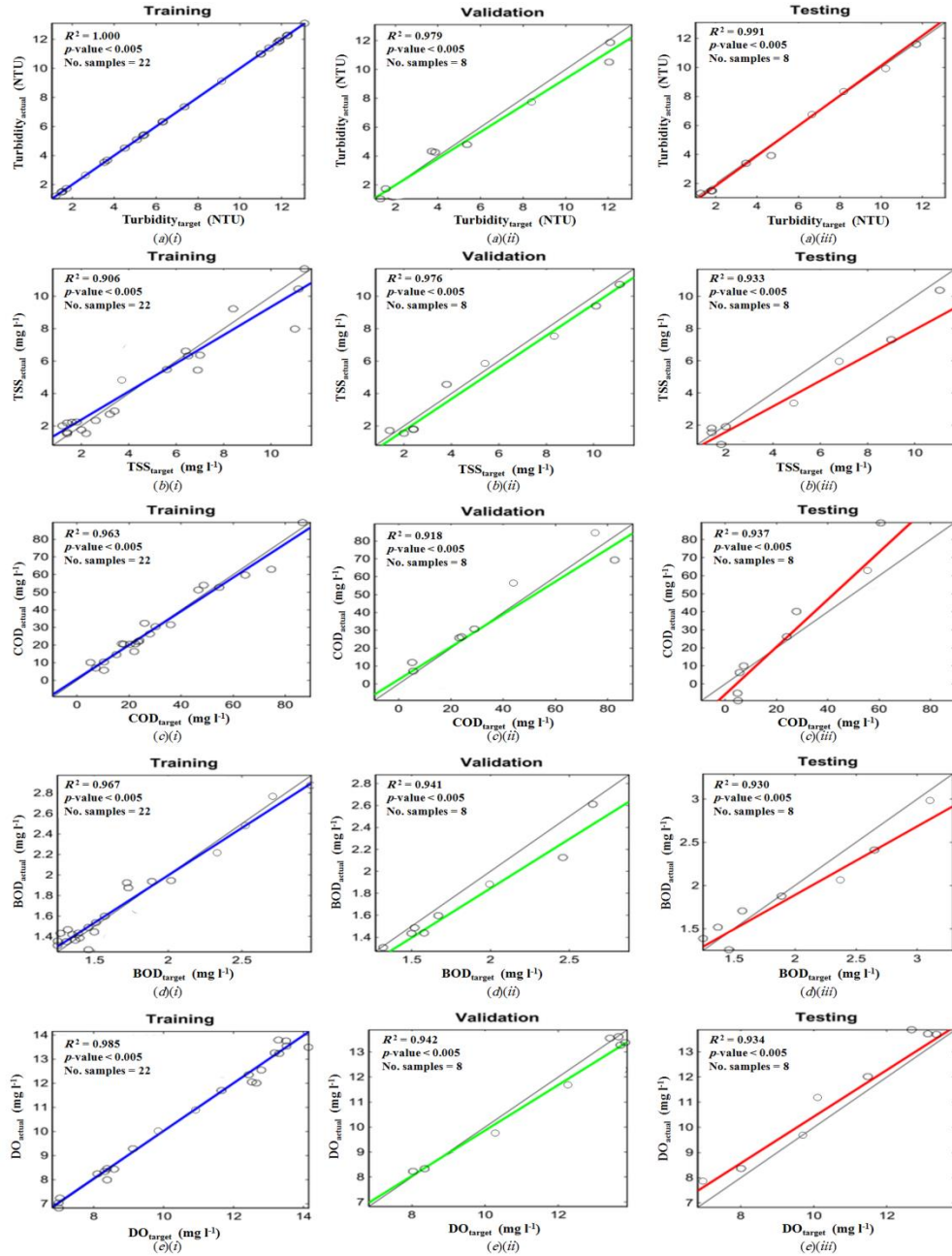


Figure 3.6 The Graphical fit results of turbidity ((a)(i), (a)(ii), and (a)(iii)), TSS ((b)(i), (b)(ii), and (b)(iii)), COD ((c)(i), (c)(ii), and (c)(iii)), BOD ((d)(i), (d)(ii), and (d)(iii)), and DO ((e)(i), (e)(ii), and (e)(iii)) for training, validation, and testing datasets of the

developed Landsat 8-based-BPNN

Using the BPNN algorithm, the RMSEs for turbidity were 0.305, 0.073, and 0.069 NTU for the network training, validation, and testing datasets, respectively. The RMSEs for TSS were 0.092, 0.226, and 0.999 mg l⁻¹ for the network training, validation, and testing datasets, respectively. For the rest of the selected parameters, COD, BOD, and DO, the training, validation, and testing RMSEs were (0.285, 0.158, 0.877 mg l⁻¹), (0.034, 0.042, 0.076 mg l⁻¹), and (0.073, 0.188, 0.455 mg l⁻¹), respectively.

Actually, it is very obvious that the developed Landsat 8-based-BPNN was proved to be very efficient in monitoring and estimating concentrations of different SWQPs, even the non-optically active parameters, with highly acceptable results. As shown in **Table 3.4** and **Figure 3.6**, coefficients of determination were found to be very high ($R^2 \geq 93\%$) at the neural network testing phase along with p -value < 0.005 .

In **Figure 3.6**, the final relationship between the targeted output (concentrations of SWQPs) and the actual output derived from the developed BPNN algorithm was developed in the Matlab environment. The experimental platform was the MATLAB R2014a and concentrations of turbidity, TSS, COD, BOD, and DO were estimated on this platform using an open source code. Accordingly, a Landsat 8-based-BPNN model was generated to predict concentrations of each SWQP individually over each pixel of the selected study area with highly acceptable results.

The threshold values of the validation RMSE were selected to be 0.100, 0.230, 0.160, 0.050, and 0.200 for turbidity, TSS, COD, BOD, and DO, respectively. As shown in **Figure 3.7**, turbidity, TSS, COD, BOD, and DO error curves showed that the training phase has been stopped at epoch 10, 34, 14, 10, and 22, respectively by reaching the stopping point introduced by the validation data set. Visually, it was observed that there

is no further significant improvement in the network performance after realizing the stopping points.

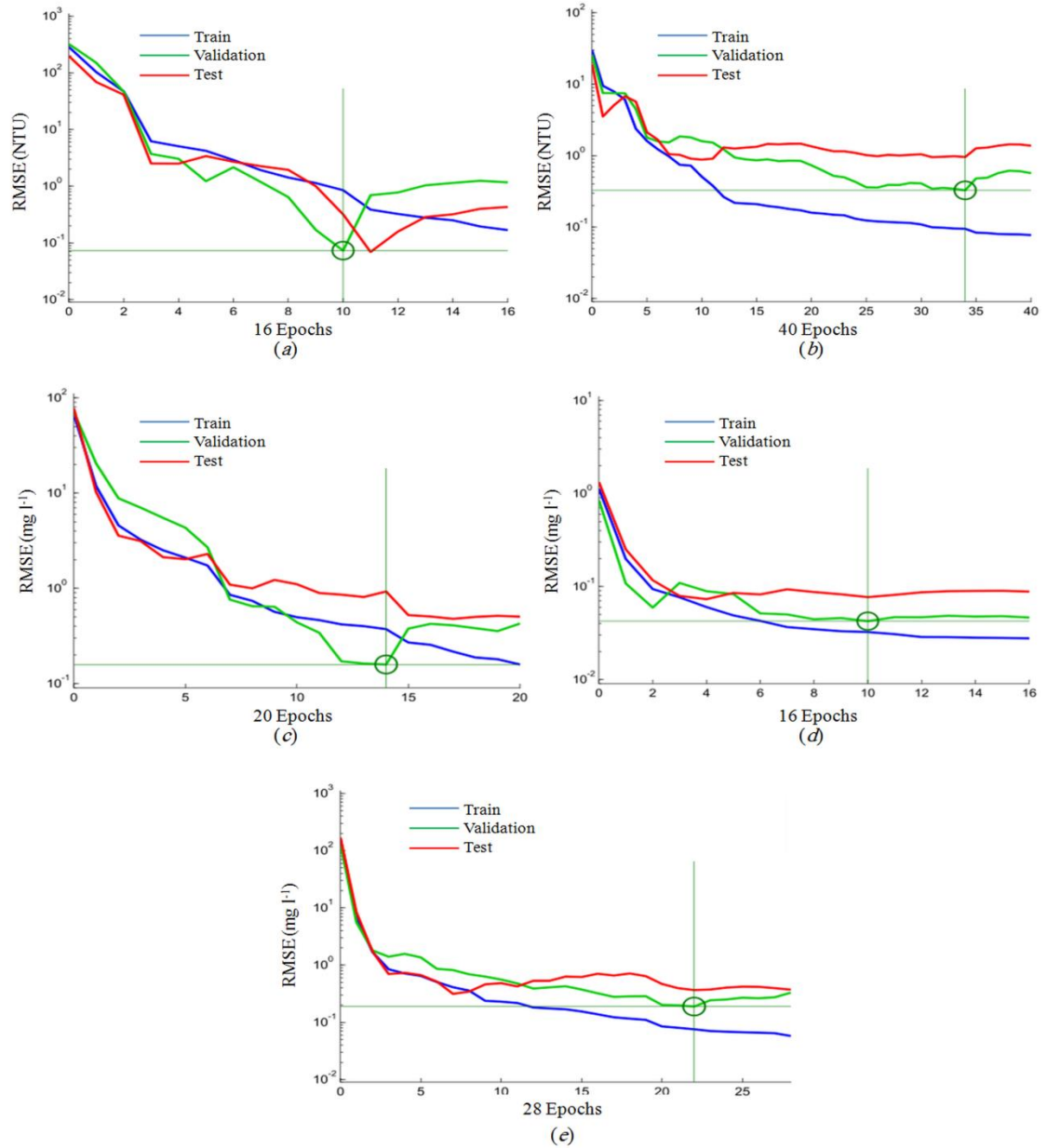


Figure 3.7 Training, validation, and testing error curves of turbidity (a), TSS (b), COD (c), BOD (d), and DO (e)

3.4.3 The Landsat 8-based-BPNN Spatial Concentration Maps

The developed Landsat 8-based-BPNN model for each SWQP was applied to the Landsat 8 satellite data to map concentrations of each optical and non-optical SWQP in the selected study area of the SJR. The whole Landsat 8 surface reflectance data, as an ASCII output from PCI Geomatica, were used pixel by pixel as an input to the developed Landsat 8-based-BPNN models in order to generate spatial concentration maps for turbidity, TSS, COD, BOD, and DO, as shown in **Figure 3.8**.

Overall, the developed Landsat 8-based-BPNN framework could be used to produce highly accurate estimations of optically and non-optically active SWQPs compared to other regression techniques which have been used in various studies such as (He, Chen, Liu, & Chen, 2008; Yang, Liu, Ou, & Yuan, 2011; Nathan, Sarah, Stephen, Narumon, Brian, & Jianguo, 2013; Xiang, Huapeng, Xiangyang, Yebao, Xin, & Hua, 2016). The main basis is that the BPNN algorithm has the capability to generate an appropriate modelling of the unknown, complex, or even non-linear relationship between remotely sensed multi-spectral information and concentrations of different SWQPs without prior knowledge of the parameter relationship.

Compared to other learning-based algorithms utilized in previous studies, such as (Zhang, Pulliainen, Koponen, & Hallikainen, 2002; Diamantopoulou, Antonopoulos, & Papamichail, 2007; Yirgalem, 2012; Ali, et al., 2013), more accurate estimations of concentrations of different SWQPs were obtained by using our novel Landsat 8-based-BPNN framework.

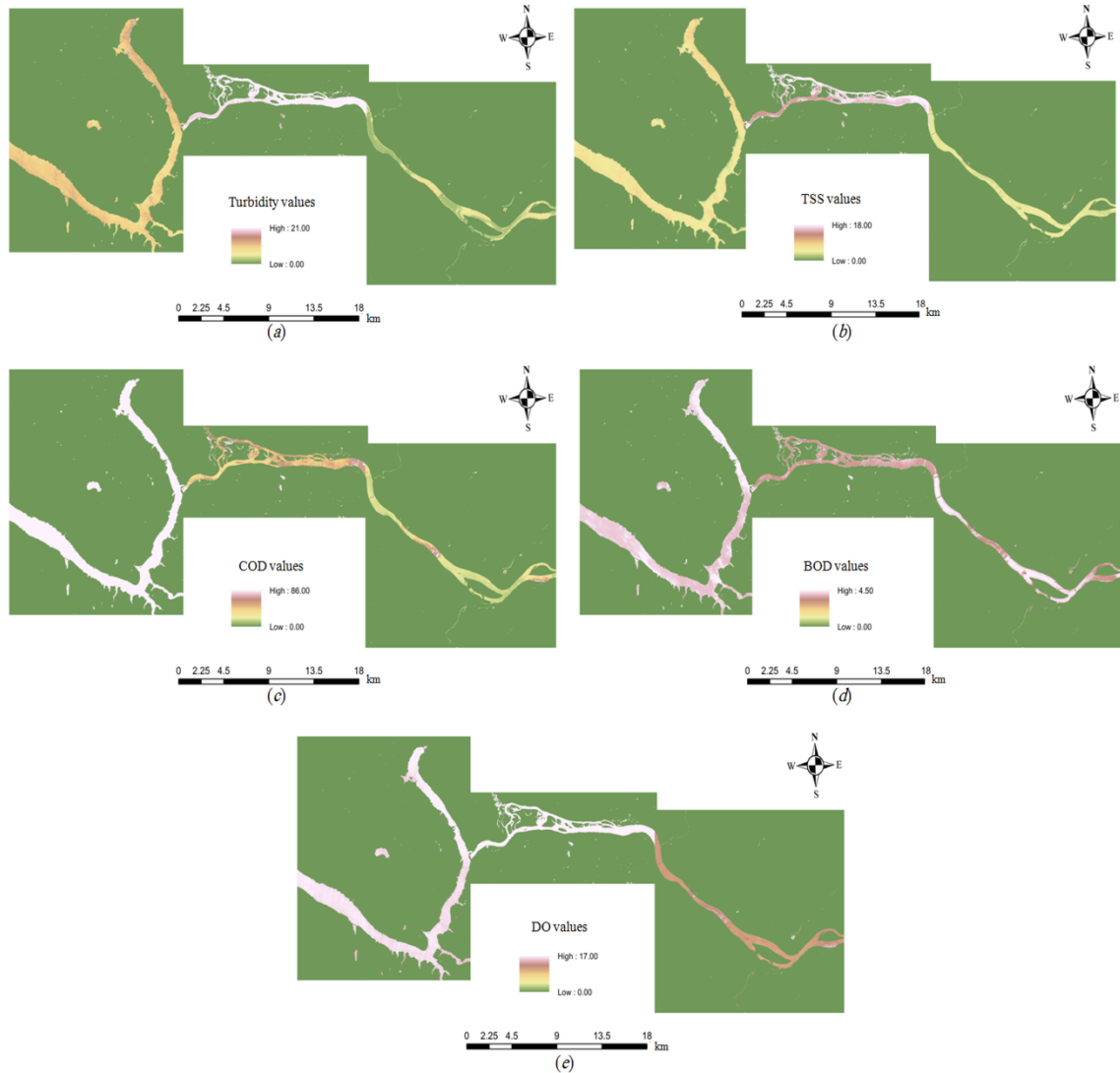


Figure 3.8 Spatial distribution maps of turbidity (a), TSS (b), COD (c), BOD (d), and DO (e) generated from the developed Landsat 8-based-BPNN

3.4.4 Comparison of Other Model Results

The performance of other machine learning-based methods, such as SVM, has been proved to be efficient in remote sensing image classification applications (Liu & Zheng, 2009). Therefore, the addition of a comparison experiment with SVM method was carried

out to justify the performance of the developed Landsat 8-based-BPNN framework.

Basically, the aim of SVM is to produce an input-output regression function by applying a set of high dimensional functions. In our study, the SVM loss function and the kernel function were selected as described by Liu & Zheng (2009). The most critical parameters were the kernel function parameter (σ^2), the penalty coefficient (C), and the width of the insensitive loss function (ϵ). The overall performance of the SVM depends mainly on the interaction of all parameters; however, the individual optimization of each parameter is a very critical task to generate the best input-output regression model. Therefore, the values of $\sigma^2 = 128$, $C = 500$, and $\epsilon = 0.25$ were experimentally selected. Actually, the MATLAB R2014a software package has no LIBSVM toolbox. Therefore, the LIBSVM library was compiled into the MATLAB R2014a environment and then the training and prediction functions were applied. Finally, concentration of both optical and non-optical SWQPs can be retrieved.

As shown in **Table 3.5**, comparing the experimental results of the SVM to the developed Landsat 8-based-BPNN, it can be indicated that the SVM results were not satisfactory because the selection of the model parameters was mainly based on experiments and there is no rule or even a stable guideline for parameter selection. In contrast, the developed Landsat 8-based-BPNN had excellent performance and at the same time, the network complexity was minimized, and the computational speed was greatly accelerated, especially by using one SWQP at a time as the network output. Finally, the nonlinear retrieve system, which was established by the developed Landsat 8-based-BPNN, can provide highly accurate estimation of different SWQP concentrations, and hence can satisfy the demand of WQ monitoring.

Table 3.5 Comparison of the BPNN and SVM statistical results.

SWQPs	BPNN		SVM	
	R^2 (validation)	RMSE (validation)	R^2 (validation)	RMSE (validation)
Turbidity (NTU)	0.979	0.073	0.941	0.118
TSS (mg l ⁻¹)	0.976	0.226	0.930	0.755
COD (mg l ⁻¹)	0.918	0.158	0.895	0.342
BOD (mg l ⁻¹)	0.941	0.042	0.902	0.076
DO (mg l ⁻¹)	0.942	0.188	0.887	0.573

3.5 Conclusion

Water bodies have deteriorated due to the overload of several pollutants such as sediments and nutrients coming from human, agricultural, and industrial activities. These pollutants lead to deterioration of water storage capacity and negatively affect aquatic life and food chain. To overcome these problems, monitoring and estimating optically and non-optically active SWQPs from remotely sensed data is very essential to provide the appropriate treatment at the proper time.

In this study, a Landsat 8-based-BPNN framework was developed to estimate concentrations of SWQPs. It was shown that the Landsat 8 multi-spectral bands can be used to map SWQP concentrations in the study area of the SJR. Moreover, highly accurate Landsat 8-based-BPNN models, with $R^2 \geq 93\%$ at the network testing phase, were obtained to retrieve turbidity, TSS, COD, BOD, and DO concentrations from the Landsat 8 satellite data over the selected study site. Accordingly, these models were used to produce spatial distribution maps for optical and non-optical SWQPs.

In our study, the generated concentration maps can be used to study the evolution of local limnologic processes, which consecutively could be related to the development of WQ in the selected study area. Therefore, this study is very applicable for local administrators who have to make decisions and enact strict measures in order to protect water quality in potable water resources particularly when this resource is indispensable for the citizens who reside in urban centres close to the river.

The future work is to carry out further studies on the SJR at different times of the year. In view of that, water sampling during different seasons is the best way to represent the maximum variation between sampling events. Moreover, it is very helpful to develop generalized models for estimating different SWQPs in the SJR without being dependent on water sampling. Furthermore, in order to properly assess WQ in the SJR, it is essential to delineate the updated water quality status of the SJR and identify the dominant SWQPs that influence water quality variation in the river.

Acknowledgements

This research is supported in part by the Egyptian Ministry of Higher Education, Bureau of Cultural & Educational Affairs of Egypt in Canada, and the Canada Research Chair (CRC) Program. The authors wish to acknowledge the USGS Landsat Archive Centre for the Landsat 8 Level 1T imagery. The authors also thank the UNB environmental research team who provided collaborative work utilized in this research.

REFERENCES

Ali, M., Hossein, A., Seyed, K. A., Jamal, M., Vali, S., Omid, M., et al. (2013). Applying artificial neural networks to estimate suspended sediment concentrations along the

- southern coast of the Caspian Sea using MODIS images. *Arab J Geosci*, 8, pp. 891-901.
- Alsmadi, M., Omar, K., & Noah, S. (2009). Back Propagation Algorithm: The Best Algorithm among the Multi-layer Perceptron Algorithm. *IJCSNS International Journal of Computer Science and Network Security*, 9 (4), pp. 378-83.
- APHA. (2005). *Standards Methods for the Examination of Water and Wastewater* (21th ed.). American Public Health Association Washington DC, USA.
- Arseneault, D. (2008). *The Road to Canada - Nomination Document for the St. John River, New Brunswick*. The St. John River with the support of the New Brunswick Department of Natural Resources.
- Bolstad, P. V., & Swank, W. T. (1997). Cumulative Impacts of Land-Use on Water Quality in a Southern Appalachian Watershed. *J. Am. Water Resour. Assoc.*, 33, pp. 519-534.
- Chavez, P. S. (1988). An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. *Remote Sensing of Environment*, 24, pp. 459-479.
- Darryl, J., Blake, A., Ross, S., Richard, W., Kenneth, R., & Donald, J. (2014). Remote sensing of selected water-quality indicators with the hyperspectral imager for the coastal ocean (HICO) sensor. *International Journal of Remote sensing*, 9, pp. 2927-2962.
- Diamantopoulou, M., Antonopoulos, V., & Papamichail, D. (2007). Cascade correlation artificial neural networks for estimating missing monthly values of water quality parameters in rivers. *Water Resour Manage*, 21, pp. 649-662.
- Earth Explorer*. (2016). Retrieved from U.S. Geological Survey: <http://earthexplorer.usgs.gov/>
- Gaballah, M., Khalaf, K., Beckand, A., & Lopez, A. (2005). Water Pollution in Relation to Agricultural Activity Impact in Egypt. *Journal of Applied Sciences*, Vol. 1, No. 1, pp. 9-17.
- Graupe, D. (2007). Principles of Artificial Neural Networks. *World Scientific*. Singapore, Hackensack, N. J.

- He, W., Chen, S., Liu, X., & Chen, J. (2008). Water Quality Monitoring in Slightly-Polluted Inland Water Body through Remote Sensing-A Case Study in Guanting Reservoir, Beijing, China. *Frontiers Environment Sciences Engineering China*, 2 (2), pp.163-171.
- Hinton, E. G. (1992). How neural networks learn from experience. *Scientific American*, 267 (3), pp. 144-151.
- Holger, R. M., Ashu, J., Graeme, C. D., & Sudheer, K. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environmental Modelling & Software*, 25, pp. 891-909.
- Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial Neural Networks: A Tutorial. *Computer*, 29 (3), pp. 31-44.
- Krista, A., Kersti, K., Reiko, R., Petra, P., Elar, A., Jan, P., et al. (2015). Satellite-based products for monitoring optically complex inland waters in support of EU Water Framework Directive. *International Journal of Remote sensing*, 36, pp. 4446-4468.
- Liu, D., & Zheng, N. (2009). Feature selection and model parameters optimization for SVM based on genetic algorithm. *Computer Application and Software*, vol. 26, no. 1, pp. 30-37.
- MacKay, J. C. (1992). Computation and Neural Systems. *Neural Computation*, 4 (3), pp. 415-447.
- Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software*, 15 (1), pp. 101-124.
- Matias, B., María, C. R., Lucio, P., & Susana, F. (2015). Using multi-temporal Landsat imagery and linear mixed models for assessing water quality parameters in Río Tercero reservoir (Argentina). *Remote Sensing of Environment*, 185, pp. 28-41.
- Mcfeters, S. K. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *INT. J. REMOTE SENSING*, 17 (7), pp. 1425-1432.

- Moss, B. (1998). *Ecology of Fresh Waters: Man and Medium, Past to Future* (3rd ed.). Blackwell Science, Oxford, UK.
- Murdoch, P. S., Baron, J. S., & Miller, T. L. (2000). Potential Effects of Climate Change on Surface-Water Quality in North America. *J. Am. Water Resour. Assoc.*, pp. 347-366.
- Nathan, T., Sarah, H., Stephen, H., Narumon, W., Brian, B., & Jiaguo, Q. (2013). Mapping inland lake water quality across the Lower Peninsula of Michigan using Landsat TM imagery. *International Journal of Remote sensing*, 21, pp. 7607-7624.
- Pat, S., & Chavez, J. (1996). Image-Based Atmospheric Corrections - Revisited and Improved. *Photogrammetric Engineering & Remote Sensing*, 62 (9), pp. 1025-1036.
- Ribeiro, H. M., Almeida, A. C., Rocha, B. R., & Krusche, A. V. (2008). Water Quality Monitoring in Large Reservoirs Using Remote Sensing and Neural Networks. *IEEE LATIN AMERICA TRANSACTIONS, VOL. 6, NO. 5*, pp. 419-423.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. (1986). Learning representations by back propagating errors. *Nature*, 323, pp. 533-536.
- Song, C., Woodcock, C. E., Seto, K. C., Lenney, M. P., & Macomber, S. A. (2001). Classification and change detection using Landsat TM data: when and how to correct atmospheric effects. *Remote Sensing of Environment*, 75, pp. 230-244.
- Suliman, A., & Zhang, Y. (2014). A Review on Back-Propagation Neural Networks in the Application of Remote Sensing Image Classification. *Journal of Earth Science and Engineering*, 5, pp. 52-65.
- Tetko, I., & Villa, A. E. (1997). An enhancement of generalization ability in cascade correlation algorithm by avoidance of overfitting/overtraining problem. *Neural Processing Letters*, no. 6, pp. 43-50.
- Tiit, K., Krista, A., Dolly, N. K., & Stephan, J. K. (2015). Impact of iron associated to organic matter on remote sensing estimates of lake carbon content. *Remote Sensing of Environment*, 156, pp. 109-116.

- Valyon, J., & Horvath, G. (2004). A sparse least squares support vector machine classifier. *International Joint Conference on Neural Networks. 1*, pp. 543-548. IEEE.
- Xiang, Y., Huapeng, Y., Xiangyang, L., Yebao, W., Xin, L., & Hua, Z. (2016). Remote-sensing estimation of dissolved inorganic nitrogen concentration in the Bohai Sea using band combinations derived from MODIS data. *International Journal of Remote Sensing, 37*:2, pp. 327-340.
- Yang, B., Liu, Y. P., Ou, F. P., & Yuan, M. H. (2011). Temporal and Spatial Analysis of COD Concentration in East Dongting Lake by Using of Remotely Sensed Data. *Procedia Environment Sciences, 10*, pp. 2703-2708.
- Yirgalem, C. (2012). Water quality monitoring using remote sensing and an artificial neural network. *Springer Science + Business Media B.V. Water air soil pollut, 223*, pp. 4875-4887.
- Yunlin, Z., Kun, S., Yongqiang, Z., Xiaohan, L., & Boqiang, Q. (2016). Monitoring the river plume induced by heavy rainfall events in large, shallow, Lake Taihu using MODIS 250 m imagery. *Remote Sensing of Environment, 173*, pp. 109-121.
- Zhang, Y. Z., Pulliainen, J. T., Koponen, S. S., & Hallikainen, M. T. (2002). Application of an empirical neural network to surface water quality estimation in the Gulf of Finland using combined optical data and microwave data. *Remote Sensing of Environment, 81*, pp. 327-336.

Chapter 4: DELINEATING THE ACCURATE PATTERNS OF SURFACE WATER QUALITY BY INTEGRATING LANDSAT 8 OLI IMAGERY, ARTIFICIAL INTELLIGENCE, AND THE WATER QUALITY INDEX³

Abstract

Extracting accurate surface water quality levels has always presented researchers with a great challenge. Existing methods of assessing surface water quality are technically detailed and present monitoring data on individual substances; however, the results of these methods are poorly understood by decision-makers. Hence, a method, such as the water quality index (WQI), is needed to provide an integrated picture of surface water quality in water bodies. However, in the absence of a representative database, WQIs may be biased leading to misleading water quality levels. Therefore, we developed a novel approach which combines the Landsat 8 multi-spectral data, the Back-Propagation Neural Network (BPNN), and the Canadian Council of Ministers of the Environment Water Quality Index (CCMEWQI) to extract accurate water quality levels to be accessible to decision-makers. The BPNN was used to develop models to estimate concentrations of surface water quality parameters (SWQPs) from Landsat 8 imagery.

³ This paper is under review in the journal "*Remote Sensing of Environment (RSE)*".

A part of this work has been published in the "*International Archives of the Photogrammetry, Remote Sensing, and Spatial Information Sciences*", XLII-4/W4, pp. 245-249, <https://doi.org/10.5194/isprs-archives-XLII-4-W4-245-2017>".

Then, these models were validated using an independent validation dataset and the results showed that relationship between concentrations of SWQPs and satellite spectral information is highly correlated with coefficient of determination (R^2) > 0.80, which is trustworthy. Moreover, the developed approach was extra validated using ground truth data provided by the Province of New Brunswick, Canada, and the developed models remained very stable with R^2 > 0.75. Finally, the obtained concentrations of SWQPs were used as an input to the CCMEWQI to delineate accurate water quality levels for drinking purposes. Based on the drinking water quality guidelines, the CCMEWQI was observed to be 67 (Fair) and 59 (Marginal) for the lower and middle basins of the Saint John River, respectively. These findings show that our study appeared to be promising in the field of water quality management.

4.1 Introduction

Water is polluted daily due to rapid urbanization, agricultural, and industrial discharge of sewage. Three-fourths of the earth's surface is surrounded by water; but only, 0.40% of it can be used for drinking purposes (Czarra, 2003). This little portion of drinking water is also under tremendous pressure because of the anthropogenic activities that affect surface water quality. Thus, it is of prime importance to extract reliable information on the quality of surface water resources (Singh, Malik, Mohan, & Sinha, 2004).

Existing surface water quality assessment methods are mainly based on the comparison of the experimentally measured surface water quality parameters (SWQPs) with the existing standard values (Debels, Figueroa, Urrutia, Barra, & Niell, 2005). While

this type of assessment is valuable for water quality experts, it is often poorly understood by non-experts, such as decision-makers and the general public. Moreover, in most cases, decision-makers need not be aware of the detailed information of water quality data (Akoteyon, Omotayo, Soladoye, & Olaoye, 2011). Therefore, it is necessary to simplify the expression of water quality and to assess surface water quality in terms of impact on public health and the environment. In this context, evaluating surface water quality based on specified water quality indices (WQIs), which are the most effective tools to extract surface water quality levels of water bodies, is very essential (Bharti & Katyal, 2011).

A WQI is a method based on a numerical expression to identify the level of water quality. It provides a convenient means of summarizing complex water quality data into simplified mathematical numbers, which can be interpreted into text classes (Bordalo, Teixeira, & Wiebe, 2006). WQIs are subdivided into four main categories: Public, Application-specific, Planning, and Statistically-based indices (Jena, Dixit, & Gupta, 2013). The first three categories of WQIs are called “expert-opinion” or “weight-based” approaches. Weights are assigned to SWQPs based on their importance and potential impacts on the water quality. Due to different weights assigned to the same SWQPs by various experts, weight-based approaches become subjective (Horton, 1965). On the other hand, statistically-based WQIs are based on statistical techniques to assess the data, reduce subjectivity, and improve the accuracy of the index. By using statistically-based WQIs, the significance of the major SWQPs in water quality assessment can be identified (Marta, Damià, & Romà, 2010; Akbar, Hassan, & Achari, 2013).

In the relevant literature, very few studies have attempted to extract the overall patterns of water quality via WQIs, such as the National Sanitation Foundation Water

Quality Index (NSFWQI), Oregon Water Quality Index (OWQI), Smith's index, and Helsinki Commission (HELCOM) water quality assessment. Most of the available research is based on two statistically-based WQIs: Overall Index of Pollution (OIP) and the Canadian Council of Ministers of the Environment water quality index (CCMEWQI).

The OIP was used to delineate the levels of water quality in Yamuna River in India by using measurements of turbidity, power of hydrogen (pH), dissolved oxygen (DO), biochemical oxygen demand (BOD), total dissolved solids (TDS), and fluoride (Sargaonkar & Deshpande, 2003). The result of this index is classified as excellent, acceptable, slightly polluted, polluted, and heavily polluted on the basis of the water quality guidelines of India. From 1995 to 1997, water samples were collected from six stations and the average water quality levels were excellent at stations 1 and 3. Stations 2, 5, and 6 were classified as slightly polluted; while station 4 was categorized as polluted.

The CCMEWQI was used to monitor water quality in the Mackenzie River basin of Canada (Lumb, Halliwell, & Sharma, 2006). It was observed that the river is affected by high turbidity and suspended sediment loads. The water quality is mostly rated as marginal (CCMEWQI values range from 43 to 59), when evaluated against the Canadian Council of Ministers of the Environment (CCME) standards. Another study utilized the CCMEWQI for comparative analysis of regional water quality in Canada and it was found to be a good tool for water quality assessment (Rosemond, Duro, & Dubé, 2009). The levels of water quality were calculated annually for each sampling site of data collected monthly. The mean CCMEWQI values ranged from 42.40 to 56.70, which is marginal (i.e. the water quality is frequently threatened or impaired).

Based on the findings of the previous studies, WQIs can support the accurate interpretation of water quality; however, they require a huge number of water samples obtained by physical monitoring of water quality. It is very challenging to provide this type of physical monitoring because this process is costly, labour intensive, and time consuming. Moreover, WQIs may be biased towards reflecting false water quality levels in the absence of a representative database (i.e. water samples). Therefore, the integration of the Landsat 8 multi-spectral data, the back-propagation neural network (BPNN) algorithm, and the CCMEWQI is proposed to extract accurate water quality levels in the selected study area of the Saint John River (SJR), New Brunswick, Canada.

First, five Landsat 8 satellite scenes, acquired in different months (i.e., June 2015, April 2016, May 2016, July 2016, and August 2016), were used along with their water sampling stations to represent the maximum variation in the concentrations of the selected SWQPs. The chemical and physical SWQPs, which were included in the CCMEWQI, are turbidity, total suspended solids (TSS), total solids (TS), total dissolved solids (TDS), chemical oxygen demand (COD), biochemical oxygen demand (BOD), dissolved oxygen (DO), power of hydrogen (pH), electrical conductivity (EC), and water temperature. Turbidity, TSS, TS, and TDS were selected because the major component that can negatively impact water quality and fish population in streams in North America is sediment (Arseneault, 2008). Moreover, due to the high loads of organic pollutants coming from food and paper production industries in the middle basin of the SJR, it is important to measure the levels of COD, BOD, and DO (Sharaf El Din & Zhang, 2017d). Furthermore, pH, EC, and temperature levels were measured due to their direct influence on both drinking water quality and aquatic life.

Then, the BPNN algorithm is selected to generate models to estimate the concentrations of SWQPs by correlating water quality data and Landsat 8 multi-spectral information. The BPNN is proposed because it can lead to good generalization of the network, control the learning process, and achieve the global minimum by adjusting an appropriate learning rate value (Sharaf El Din, Zhang, & Suliman, 2017a).

Finally, the obtained concentrations of SWQPs are used as an input to the CCMEWQI to extract accurate water quality levels. The CCMEWQI is proposed due to its flexibility in the selection of input parameters (i.e. different SWQPs), the capability of minimizing the data volume to a great extent, and simplifying the expression of water quality (CCME, 2001).

The identified objectives of this research are to (1) develop an accurate approach for quantifying concentrations of SWQPs over each pixel of the selected study area by using the BPNN, (2) evaluate the performance and stability of the developed approach using ground truth data (i.e. water quality data) provided by the Province of New Brunswick, and (3) delineate accurate levels of surface water quality by using the CCMEWQI. To the best of our knowledge, the Landsat 8-based-CCMEWQI approach is developed for the first time to extract accurate levels of surface water quality with highly accurate results and inexpensive implementation cost.

4.2 Materials and Methods

The flowchart for delineating accurate water quality levels from satellite imagery by using the proposed Landsat 8-based-CCMEWQI is shown in **Figure 4.1**.

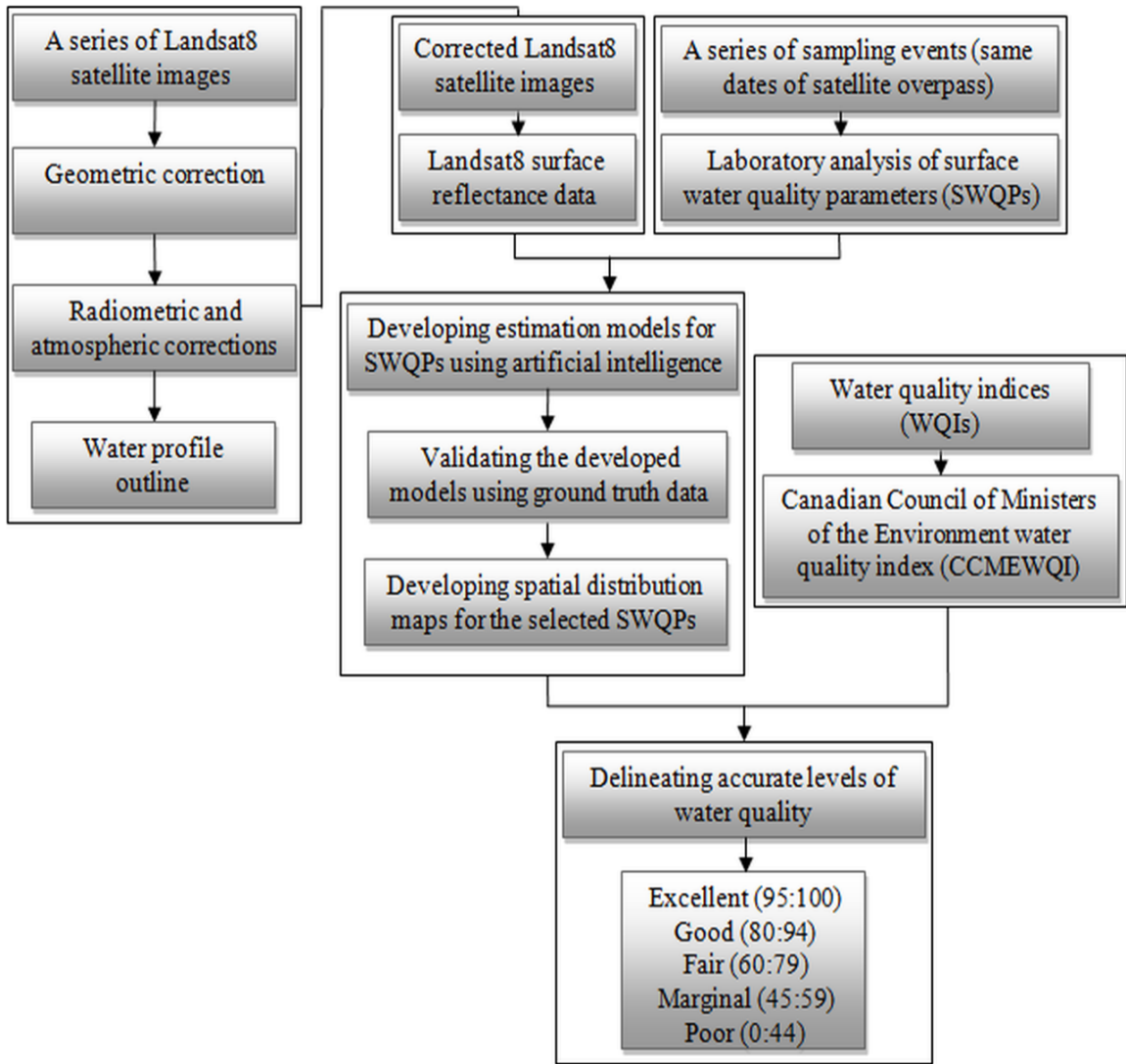


Figure 4.1 The flowchart of the proposed methodology

This section is devoted to describing the selected study area of the SJR, processing steps of the Landsat 8 satellite images, analyzing the collected water samples, developing estimation models for SWQPs, and extracting the exact water quality levels of the SJR.

4.2.1 Study Area

The selected study area covers 130 km of the SJR. As shown in **Figure 4.2**, the study area covers two main parts of the SJR: the lower basin (i.e. below the Mactaquac Dam) and the middle basin (i.e. above the Mactaquac Dam). Compared to the lower basin, the middle basin is more polluted due to the presence of food and paper processing industries (Arseneault, 2008).

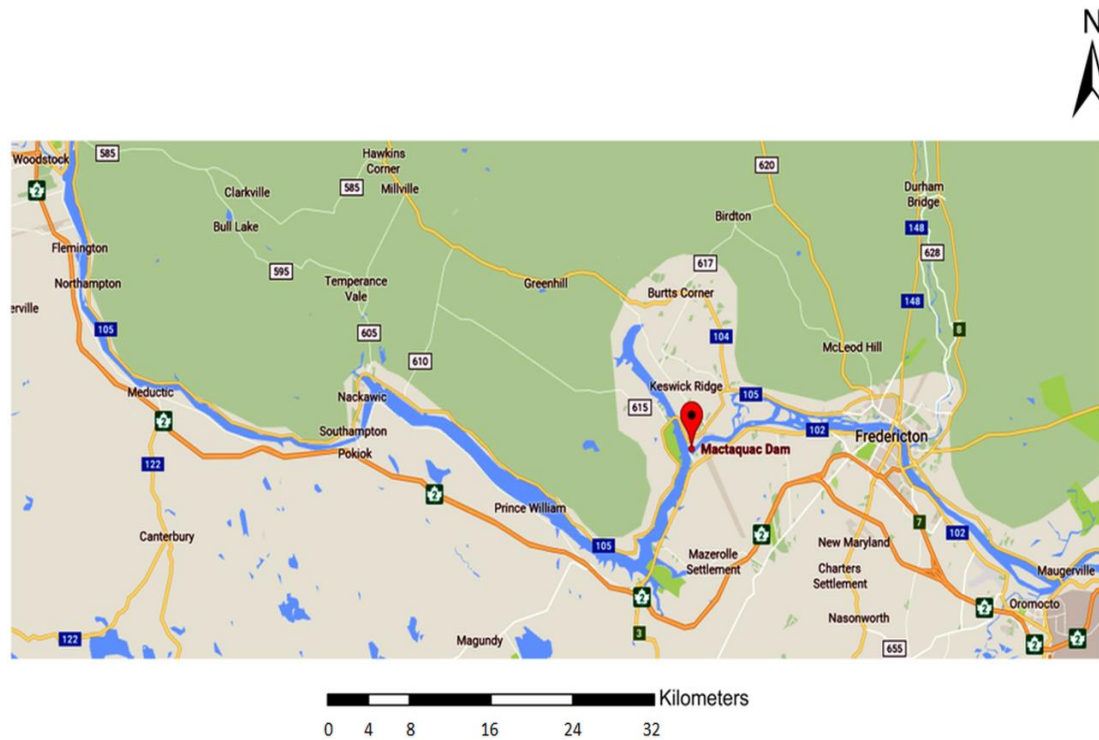


Figure 4.2 The selected study area of the Saint John River (SJR), New Brunswick, Canada (Google Maps, 2016)

4.2.2 Landsat 8 Image Acquisition and Processing

Satellite images with low spatial resolution, such as the Moderate-resolution Imaging Spectroradiometer (MODIS) and the Medium Resolution Imaging Spectrometer

(MERIS), have a larger scale size than the width of the narrow tributaries of the SJR, which causes various mixed pixels, resulting in the low precision estimation of the concentrations of different SWQPs. On the other hand, the Landsat 8 Operational Land Imager (OLI) images have higher spatial resolution (30 m, in the visible spectrum). Compared to the Landsat-5 Thematic Mapper (TM) and the Landsat-7 Enhanced Thematic Mapper Plus (ETM+), the Landsat 8 OLI has enhanced features, which include the addition of three multi-spectral bands (coastal blue visible band, one shortwave infrared band, and one thermal band) (United States Geological Survey (USGS), 2016).

The Landsat 8 OLI sensor uses a pushbroom scanner that enables data acquisition with much better performance in terms of the signal-to-noise ratio (Roy, Wulder, Loveland, & Zhu, 2014). Compared to the previous 8-bit Landsat-7 ETM+ sensor, the Landsat 8 OLI sensor is a 12-bit instrument with a dynamic range of 4096 gray levels. The narrower multi-spectral bands, the higher signal-to-noise ratio, and the higher radiometric resolution demonstrate that the Landsat 8 OLI sensor is less impacted by atmospheric distortions and more sensitive to surface reflectance variations (Roy, Wulder, Loveland, & Zhu, 2014).

In our study, five high-quality Landsat 8 satellite sub-scenes acquired in different months were used to best represent the maximum variation in the concentrations of SWQPs. The satellite images used were acquired on June 27th 2015, April 10th 2016, May 12th 2016, July 22nd 2016, and August 23rd 2016. The Landsat 8 satellite images are available free of charge at Level 1T (terrain corrected) and geometrically corrected to the Universal Transverse Mercator (UTM) projection, World Geodetic System 1984 (WGS 84) datum (Earth Explorer, 2016).

Atmospheric distortions should be eliminated in order to measure the water-leaving reflectance (i.e. surface reflectance). The Dark Object Subtraction (DOS) method was used to calculate the surface reflectance values (Chavez, 1988). This method is well accepted by the geospatial community to correct light scattering in remote sensing data and consequently can provide accurate mapping for wetland areas (Song, Woodcock, Seto, Lenney, & Macomber, 2001). Atmospheric and topographic correction (ATCOR) and second simulation of the satellite signal in the solar spectrum (6S) methods have been used in remote sensing and digital image processing applications. However, the main disadvantage of these two methods is that they entail extensive field and ground measurements during each satellite pass. This is often impossible for several applications when working in very remote or difficult access to locations or when using historical data (Pat & Chavez, 1996).

As shown in **Figure 4.3**, the adjusted normalized difference water index was used to mask the water profile by separating water and non-water features (Mcfeeters, 1996).

4.2.3 Water Sampling and Laboratory Analysis

Sampling was performed during five field trips in June 27th 2015, April 10th 2016, May 12th 2016, July 22nd 2016, and August 23rd 2016. Water samples were randomly distributed across the entire study area, as shown in **Figure 4.3**. Seventy water samples were collected along 130 km of the SJR and four samples were excluded due to cloud coverage. In the field, coordinates of each sample point were recorded using a handset GPS, GARMIN 76CSx.

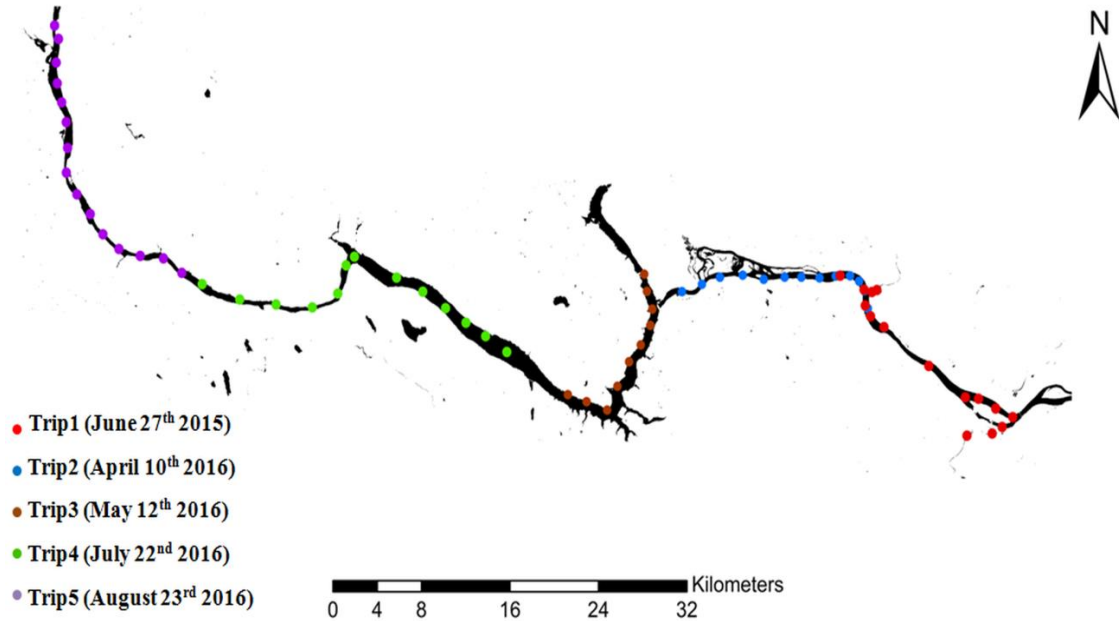


Figure 4.3 The water profile and the sampling stations

Water samples were collected around the same time of each satellite pass (4 hours time span) and just beneath water surface (i.e., 30 to 50 cm). Concentrations of optical and non-optical SWQPs, such as turbidity, total suspended solids (TSS), total solids (TS), total dissolved solids (TDS), chemical oxygen demand (COD), biochemical oxygen demand (BOD), dissolved oxygen (DO), power of hydrogen (pH), electrical conductivity (EC), and water temperature, were measured according to the American Public Health Association (APHA) water and wastewater standards (APHA, 2005).

Turbidity is an optical determination of water clarity and is calculated by measuring the amount of light scattered by suspended particles in the water column. TSS is calculated by filtering the water sample and weighing the residue left on the filter paper. Moreover, TS is determined by evaporating the water sample and weighing the dry residue left, and the difference between TS and TSS represents the TDS. COD is

measured as the amount of a specific oxidizing agent that reacts with a sample under controlled conditions; while BOD refers to the amount of dissolved oxygen consumed by aerobic organisms to break down the organic compounds in five days at 20° Celsius. DO refers to the level of non-compound oxygen present in a water sample. The acidity or alkalinity of a water sample is reported as pH. EC is a measure of how well a water sample transmits an electrical current and it is considered a good indicator of inorganic dissolved solids. Finally, water temperature is a physical property expressing how hot or cold water is. Temperature is an important factor to consider when assessing water quality because it influences other SWQPs and can alter the physical and chemical properties of water.

4.2.4 Estimation of Concentrations of SWQPs using the BPNN

Remote sensing estimation of the optically-active SWQPs (i.e. turbidity, TSS, and chlorophyll), is commonly achieved using regression techniques (Changchun, et al., 2014; Bunkei, Wei, Gongliang, Youichi, Kazuya, & Takehiko, 2015; Shuisen, Liusheng, Xiuzhi, Dan, Lin, & Yong, 2015). However, the relationship between spectral information and concentrations of SWQPs is too complex to be modelled accurately using regression techniques (Zhang, Pulliainen, Koponen, & Hallikainen, 2002). Thus, developing an artificial intelligence modelling method, such as artificial neural network (ANN), for mapping concentrations of SWQPs is essential.

ANN architecture typically comprises three types of neuron layers: an input layer, which contains the independent variables, one or more hidden layers, and an output layer, which contains the dependent variables (Hinton, 1992). One of the most common ANN

algorithms, in digital image processing applications, is the BPNN. The BPNN algorithm can be decomposed into four main steps:

- The feed-forward computation
- The error signal calculation
- Back-propagation of the error to both the output layer and the hidden layer(s)
- Updating the connection weights

Commonly, the inputs of the ANN are the pixel values from satellite spectral bands and they are feed-forwarded into the network towards the hidden layer nodes. As shown in **Equations (4.1-4.2)**, the input values are multiplied by the weights of the connecting nodes, and the values of the hidden layer nodes are computed (Hinton, 1992).

$$z = w^t * x + T \quad (4.1)$$

where z is the linear combination of neuron weighted inputs; w is the input weights vector; x is the input vector; and T is a threshold value.

Normally, the feed-forward computation is divided into two main steps: the first step is to calculate the values of the hidden layer nodes and the second step is to use the obtained values from the hidden layer to calculate the values of the output layer.

$$f(z) = 1/(1 + e^{-z/\theta_0}) \quad (4.2)$$

where θ_0 is the gradient coefficient.

The following step is to calculate the error signal of each node according to **Equation (4.3)**. The actual output of the network is compared to the desired output to determine the error. Once the error is calculated, it will be used for backward propagation and weight adjustment.

$$E = \frac{1}{2} \sum_k (T_k - O_k)^2 \quad (4.3)$$

where E is the error signal; k is the index of the output layer of the network; T_k is the desired output; O_k is the network actual output.

As shown in **Equations (4.4-4.5)**, the gradient descent technique is used with the BPNN algorithm to back-propagate the error and to locate the global minima of the error surface. The error is first back propagated from the output layer to the hidden layer. This is where learning rate can be added to the gradient descent equation ([Sharaf El Din, Zhang, & Suliman, 2017a](#)). Then, the error signal has to be propagated from the hidden layer back to the input layer. The final step is supposed to find out the updated and optimal set of weights, which creates the mapping model that can ideally produce the correct output for the relative input.

$$w'_{kj} = w_{kj} + \eta * \delta_k * O_j \quad (4.4)$$

$$w'_{ji} = w_{ji} + \eta * \delta_j * O_i \quad (4.5)$$

where w'_{kj} is the the updated weight vector for the network connections between the layers indexed by k and j ; w_{kj} is the current weight vector for the network connections

between the layers indexed by k and j ; η is a constant referred to as the learning rate; δ_k is the local gradient of the error function at the layer k ; O_j is the network actual output at the layer j ; w'_{ji} is the the updated weight vector for the network connections between the layers indexed by j and i ; w_{ji} is the current weight vector for the network connections between the layers indexed by j and i ; δ_j is the local gradient of the error function at the layer j ; O_i is the network actual output at the layer i .

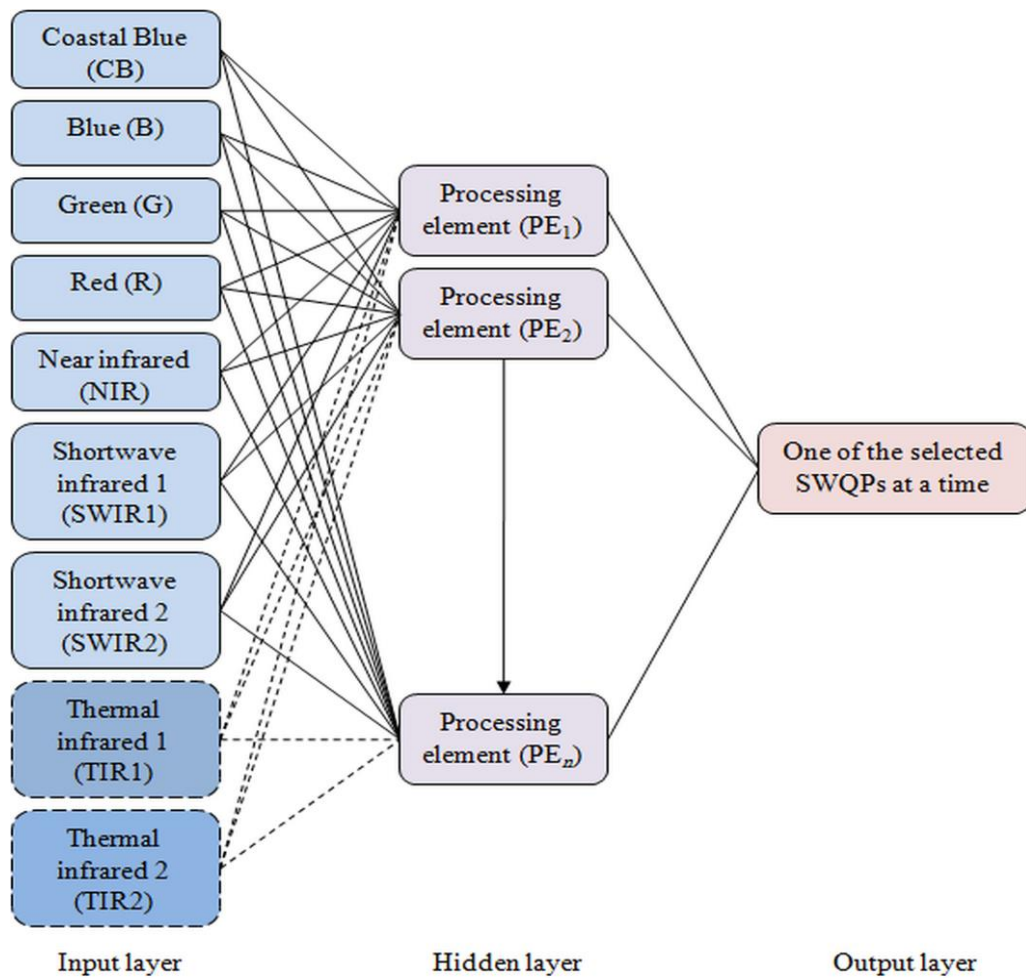


Figure 4.4 The proposed artificial neural network (ANN) topology

In our study, the BPNN algorithm was adopted to model the unknown relationship between the Landsat 8 surface reflectance data and concentrations of turbidity, TSS, TS, TDS, COD, BOD, DO, pH, EC, and water temperature. As shown in **Figure 4.4**, the Landsat 8 multi-spectral bands which show the highest correlation to the selected SWQPs were used to form the input layer. Thermal infrared 1 (TIR1) and thermal infrared 2 (TIR2) bands were used to quantify only surface water temperature because they are mainly designed to detect surface temperatures. While concentrations of SWQPs were selected, one at a time, to compose the output layer, the number of hidden layers and the number of neurons in each hidden layer were experimentally selected.

4.2.5 Applying the CCMEWQI

The CCMEWQI is a method implemented by the Canadian Council of Ministers of the Environment (CCME) for simplifying the extraction of water quality data. It provides meaningful indications of water quality that are very useful to local administrators and managers as well as the general public. As a summary tool, it provides the overall patterns of water quality and is not intended to be a substitute for detailed analysis of water quality data (CCME, 2001). The specific inputs (i.e., the selected SWQPs), objectives, and time period used in the CCMEWQI are not specified and indeed, could vary from region to region, depending on local conditions and issues. A monthly or quarterly monitoring data may be used to reflect water quality levels for a specific period; however, data from different years can be combined, especially when monitoring in certain years are incomplete (CCME, 2001).

As shown in **Table 4.1**, the CCMEWQI can be used to assess water quality relative to its desirable state as defined by drinking water quality objectives (guidelines) given by the CCME.

Table 4.1 The CCME and WHO guidelines for drinking water quality.

Selected surface water quality parameters (SWQPs)	Permissible limits
Turbidity	< 5.00 NTU
Total suspended solids (TSS)	< 25.00 mg/l
Total solids (TS)	< 500.00 mg/l
Total dissolved solids (TDS)	< 500.00 mg/l
Chemical oxygen demand (COD)	< 10.00 mg/l
Biochemical oxygen demand (BOD)	< 3.00 mg/l
Dissolved oxygen (DO)	> 6.50 mg/l
Power of hydrogen (pH)	≥ 6.50 and ≤ 8.50
Electrical conductivity (EC)	< 100 us/cm
Temperature	< 15 Celsius

When the CCME standards are not accessible, the World Health Organization (WHO) recommendations are applied. As shown in **Equations (4.6-4.12)**, the CCMEWQI works by combining three measures of variance (scope, frequency, and amplitude), where these factors are determined on the basis of water quality guidelines according to the specified application (CCME, 2001). The CCMEWQI produces a value within a range from [0 to 100] where zero represents poor water quality and one hundred indicates excellent water quality. The obtained water quality is classified into five

categories, which are Excellent (95-100), Good (80-94), Fair (60-79), Marginal (45-59), and Poor (0-44).

$$CCMEWQI = 100 - ((\sqrt{F1^2 + F2^2 + F3^2})/1.732) \quad (4.6)$$

$$F1 = (\text{Number of failed SWQPs}/\text{Total number of SWQPs}) * 100 \quad (4.7)$$

$$F2 = (\text{Number of failed tests}/\text{Total number of tests}) * 100 \quad (4.8)$$

$$F3 = (\text{nse}/(0.01 * \text{nse} + 0.01)) \quad (4.9)$$

$$\text{nse} = ((\sum_{i=1}^n \text{excursion}_i)/\text{Total number of tests}) \quad (4.10)$$

$$\text{excursion}_i = (\text{Objective}_j/\text{Failed test value}_i) - 1 \quad (4.11)$$

$$\text{excursion}_i = (\text{Failed test value}_i/\text{Objective}_j) - 1 \quad (4.12)$$

where $F1$ (scope) is the percentage of SWQPs where water quality guidelines are not met; $F2$ (frequency), is the percentage of tests that do not meet the objectives; $F3$ (amplitude) shows the amount by which failed tests do not meet the objectives; nse is the normalized sum of excursion; and excursion_i refers to the number of times by which an individual concentration is greater than (or less than, when the objective is a minimum) the objective.

The CCMEWQI can be very useful in tracking water quality changes at a given site over a specific period of time and can also be used to compare directly among sites that employ the same SWQPs and objectives (CCME, 2001). On the other hand, the main drawbacks of using the CCMEWQI include loss of information by combining different SWQPs to obtain a single value (Rosemond, Duro, & Dubé, 2009), loss of interaction between SWQPs, and sensitivity to input parameters (Khan, Paterson, & Khan, 2004).

Finally, the CCMEWQI was not developed to replace the detailed analysis of SWQPs, but rather as a method to help water quality managers and administrators communicate the overall quality of water in a more consistent manner (Sharaf El Din & Zhang, 2017e).

4.3 Results and Discussion

The main results of this study were divided into (1) concentrations of the collected water samples, (2) calibration and validation of the developed BPNN models, (3) spatial distribution of the concentrations of the selected SWQPs in the SJR, and (4) accurate delineation of the accurate levels of surface water quality of the SJR.

4.3.1 Concentrations of Optical and Non-optical SWQPs

Sixty-six water samples were analyzed using standard methods given in APHA, to measure the concentrations of different SWQPs. Twenty-eight samples were collected below the Mactaquac Dam (i.e. the lower basin of the SJR); while, thirty-eight samples were collected above the Mactaquac Dam (i.e. the middle basin of the SJR) in order to represent the maximum variance in sampling concentrations. Water quality is in a better state below the Dam, compared to the area above the Dam, because there is less industry and agriculture, no major dams, and more water flowing into the river (Arseneault, 2008).

As shown in **Table 4.2**, the descriptive statistics were measured for turbidity, TSS, TS, TDS, COD, BOD, DO, pH, EC, and water temperature. The concentrations ranged from 1.19 to 13.10 NTU with an average 4.84 NTU, 0.60 to 11.40 mg/l with an average 3.59 mg/l, 58.00 to 245.00 mg/l with an average 113.92 mg/l, 52.40 to 233.85 mg/l with an average 110.33 mg/l, 4.80 to 86.64 mg/l with an average 27.55 mg/l, 1.21 to 3.25 mg/l with an average 1.75 mg/l, 6.71 to 14.14 mg/l with an average 9.54 mg/l, 6.51

to 8.42 with an average 7.59, 29.50 to 148.90 us/cm with an average 97.09 us/cm, and 5.00 to 23.30 Celsius with an average 15.92 Celsius for turbidity, TSS, TS, TDS, COD, BOD, DO, pH, EC, and temperature, respectively.

Table 4.2 Descriptive statistics of the concentrations of SWQPs.

Optical and non-optical SWQPs	Mean	Minimum	Maximum	Standard deviation
Turbidity (NTU)	4.84	1.19	13.10	3.73
TSS (mg/l)	3.59	0.60	11.40	3.10
TS (mg/l)	113.92	58.00	245.00	42.32
TDS (mg/l)	110.33	52.40	233.85	39.91
COD (mg/l)	27.55	4.80	86.64	19.85
BOD (mg/l)	1.75	1.21	3.25	0.52
DO (mg/l)	9.54	6.71	14.14	2.64
pH	7.59	6.51	8.42	0.33
EC (us/cm)	97.09	29.50	148.90	30.53
Temperature (Celsius)	15.92	5.00	23.30	6.97

Turbidity, TSS, TS, and TDS in spring were higher than their concentrations in summer because snowmelt and rainfall push sediments from agriculture and forestry directly into the river. Alternatively, the middle basin of the SJR has high concentrations of COD and BOD because this region has many industries, such as food and paper processing, located at the SJR shoreline (Sharaf El Din & Zhang, 2017d).

4.3.2 Training and Validation of the Proposed ANN

For appropriate selection of the input layer neurons of the proposed ANN, the Landsat 8 coastal blue (CB), blue (B), green (G), red (R), near infrared (NIR), shortwave infrared 1 (SWIR1), and shortwave infrared 2 (SWIR2) bands were selected to form the input layer for all SWQPS; however in case of surface water temperature, thermal infrared 1 (TIR1), and thermal infrared 2 (TIR2) bands were added to the bands in the input layer.

Table 4.3 Correlation coefficient values between the Landsat 8 spectral data and the concentrations of SWQPs.

	Turbidity	TSS	TS	TDS	COD	BOD	DO	pH	EC	Temperature
CB	0.81	0.79	0.71	0.67	0.69	0.70	-0.66	0.74	0.61	0.83
B	0.59	0.61	0.64	0.68	0.59	0.57	-0.67	0.67	0.58	0.71
G	0.60	0.58	0.55	0.60	0.63	0.59	-0.60	0.59	0.55	0.68
R	0.67	0.69	0.59	0.55	0.53	0.57	-0.54	0.69	0.59	0.73
NIR	0.82	0.85	0.77	0.67	0.71	0.74	-0.71	0.64	0.57	0.65
SWIR1	0.84	0.78	0.73	0.66	0.68	0.67	-0.70	0.57	0.53	0.63
SWIR2	0.79	0.81	0.70	0.62	0.65	0.69	-0.71	0.59	0.56	0.61
Cirrus	0.46	0.40	0.48	0.45	0.41	0.47	-0.49	0.38	0.40	0.45
TIR1	0.35	0.39	0.45	0.44	0.32	0.41	-0.43	0.30	0.42	0.78
TIR2	0.33	0.36	0.44	0.48	0.29	0.37	-0.38	0.28	0.44	0.77

As shown in **Table 4.3**, these multi-spectral bands were significantly correlated (i.e. correlation coefficient ≥ 0.50) to the concentrations of the selected SWQPs used in our study. Additionally, SWQPs were selected one at a time to form the output layer to decrease the ANN complexity and improve the computational speed of the network.

For appropriate data division, a trial and error procedure was used to separate the available water samples in such a way that the statistical properties of the training set are close to those of the testing set. Seventy-five percent of water samples (i.e. 49 samples) were utilized for training the ANN, while twenty-five percent of the collected samples (i.e. 17 samples) were used for testing the performance of the developed BPNN models.

The proposed ANN architecture consisted of three layers with a sigmoid activation function which is differentiable and can provide the powerful capability of modelling complex and nonlinear problems. Selecting the appropriate number of neurons in the hidden layer is a critical task. In our study, 25 neurons were experimentally selected to form the hidden layer. Using a small number of neurons in the hidden layer may lead to an underfitting problem, while using a huge set of hidden neurons may cause overfitting and lead to slow learning.

The BPNN algorithm was used to map the relationship between the Landsat 8 spectral bands and concentrations of SWQPs. This algorithm can result in good generalization when using either large or small datasets (MacKay, 1992). This algorithm is computationally efficient as 4, 5, 8, 12, 22, 21, 10, 4, 18, and 11 seconds were achieved, at the ANN training phase, for turbidity, TSS, TS, TDS, COD, BOD, DO, pH, EC, and temperature, respectively.

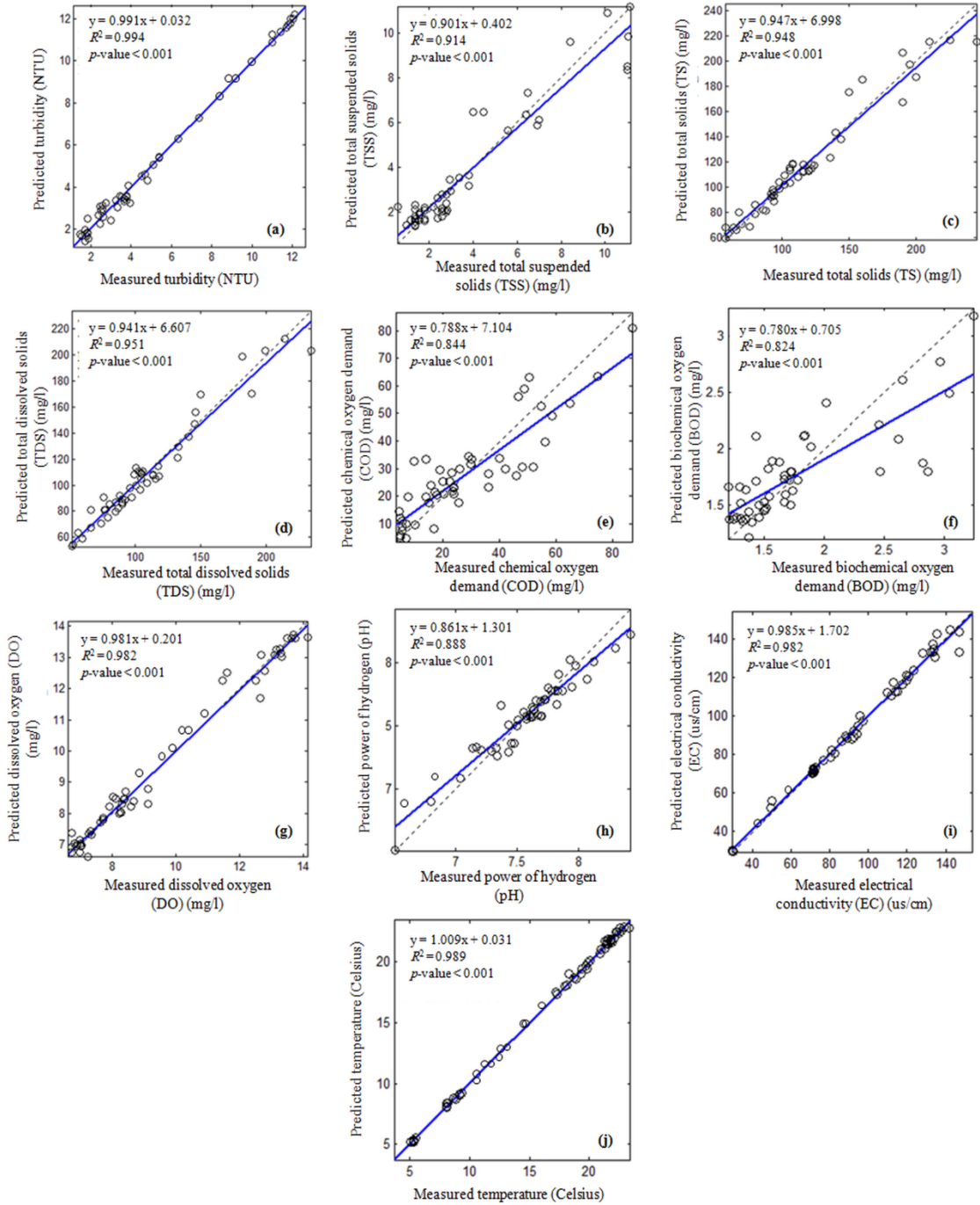


Figure 4.5 Scatter plots of observed (measured) vs. modeled (predicted) concentrations of turbidity (a), TSS (b), TS (c), TDS (d), COD (e), BOD (f), DO (g), pH (h), EC (i), and temperature (j) using the training dataset

Additionally, finding the global minima is guaranteed by utilizing an appropriate learning rate value.

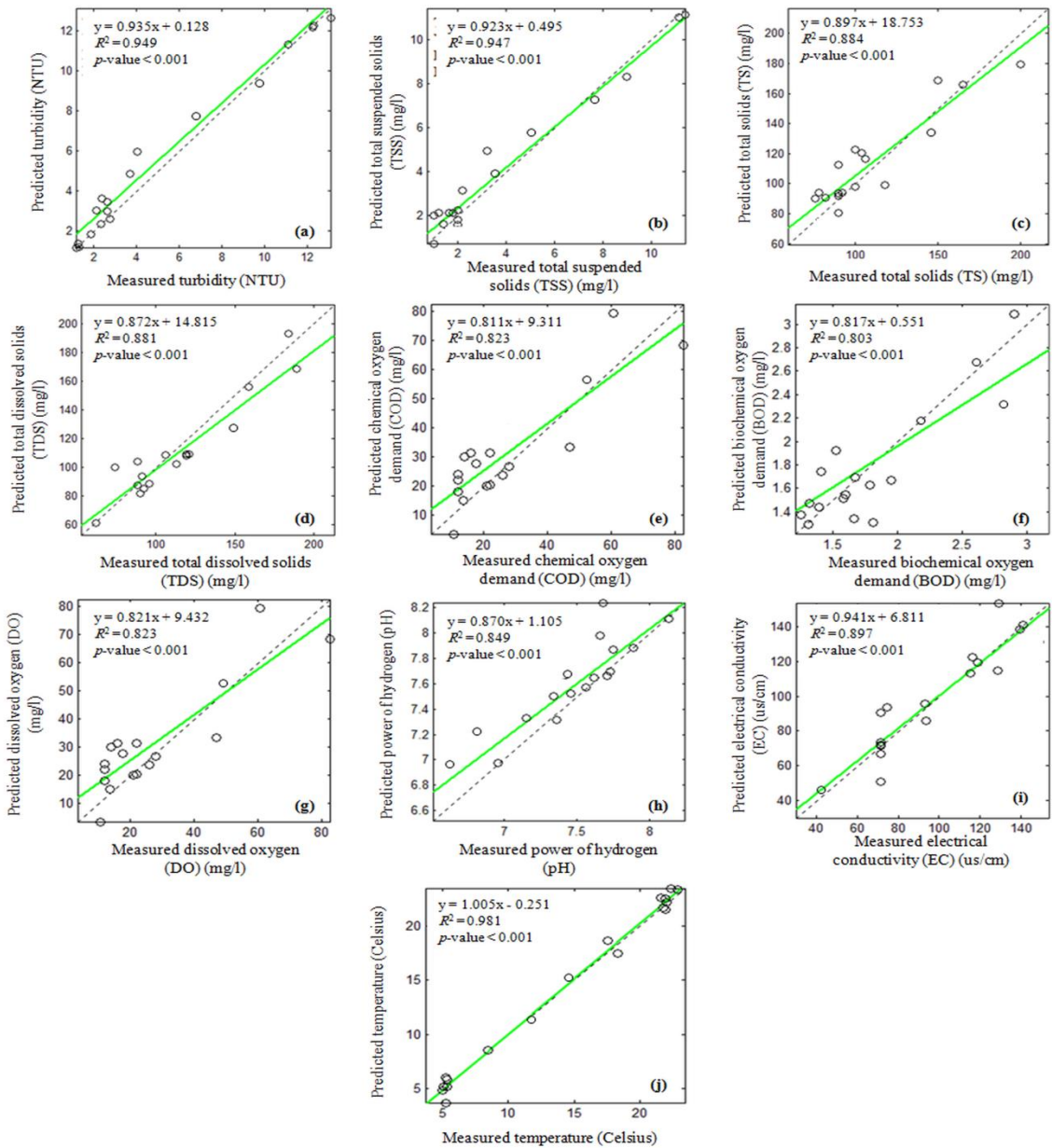


Figure 4.6 Scatter plots of observed (measured) vs. modeled (predicted) concentrations of turbidity (a), TSS (b), TS (c), TDS (d), COD (e), BOD (f), DO (g), pH (h), EC (i), and temperature (j) using the testing dataset

A learning rate value of 0.01 was adjusted to achieve the global minima in the error surface. Using a learning rate beyond the selected value, the system was very slow; however, using a learning rate above the selected value, the generalization ability of the network was very poor.

As shown in **Figure 4.5**, for the whole SWQPs, coefficients of determination were very high ($R^2 \geq 0.824$) at the neural network training phase with p -value < 0.001 . The final relationship between the desired output (i.e. observed concentrations of SWQPs) and the actual output (i.e. derived from the developed network) was developed in the Matlab environment. To test the robustness of the developed BPNN models in the SJR, the testing dataset (i.e. 17 water samples which were not used in the training process) was used to validate their performance. As shown in **Figure 4.6**, for both optical and non-optical SWQPs, $R^2 \geq 0.803$ at the neural network testing phase with p -value < 0.001 . The validation models for turbidity, TSS, TS, TDS, COD, BOD, DO, pH, EC, and temperature remained very stable with $R^2 = 0.949, 0.947, 0.884, 0.881, 0.823, 0.803, 0.823, 0.849, 0.897, \text{ and } 0.981$, respectively.

Figure 4.7 showed that the root mean square errors (RMSEs) were 0.061 NTU, 0.802 mg/l, 0.753 mg/l, 0.522 mg/l, 0.133 mg/l, 0.150 mg/l, 0.121 mg/l, 0.011, 0.021 us/cm, and 0.041 Celsius for turbidity, TSS, TS, TDS, COD, BOD, DO, pH, EC, and temperature, respectively, at the network training phase. Similarly, the RMSEs were 0.557 NTU, 0.654 mg/l, 1.353 mg/l, 1.781 mg/l, 0.112 mg/l, 0.171 mg/l, 0.143 mg/l, 0.451, 0.752 us/cm, and 0.302 Celsius for turbidity, TSS, TS, TDS, COD, BOD, DO, pH, EC, and temperature, respectively, at the network testing phase.

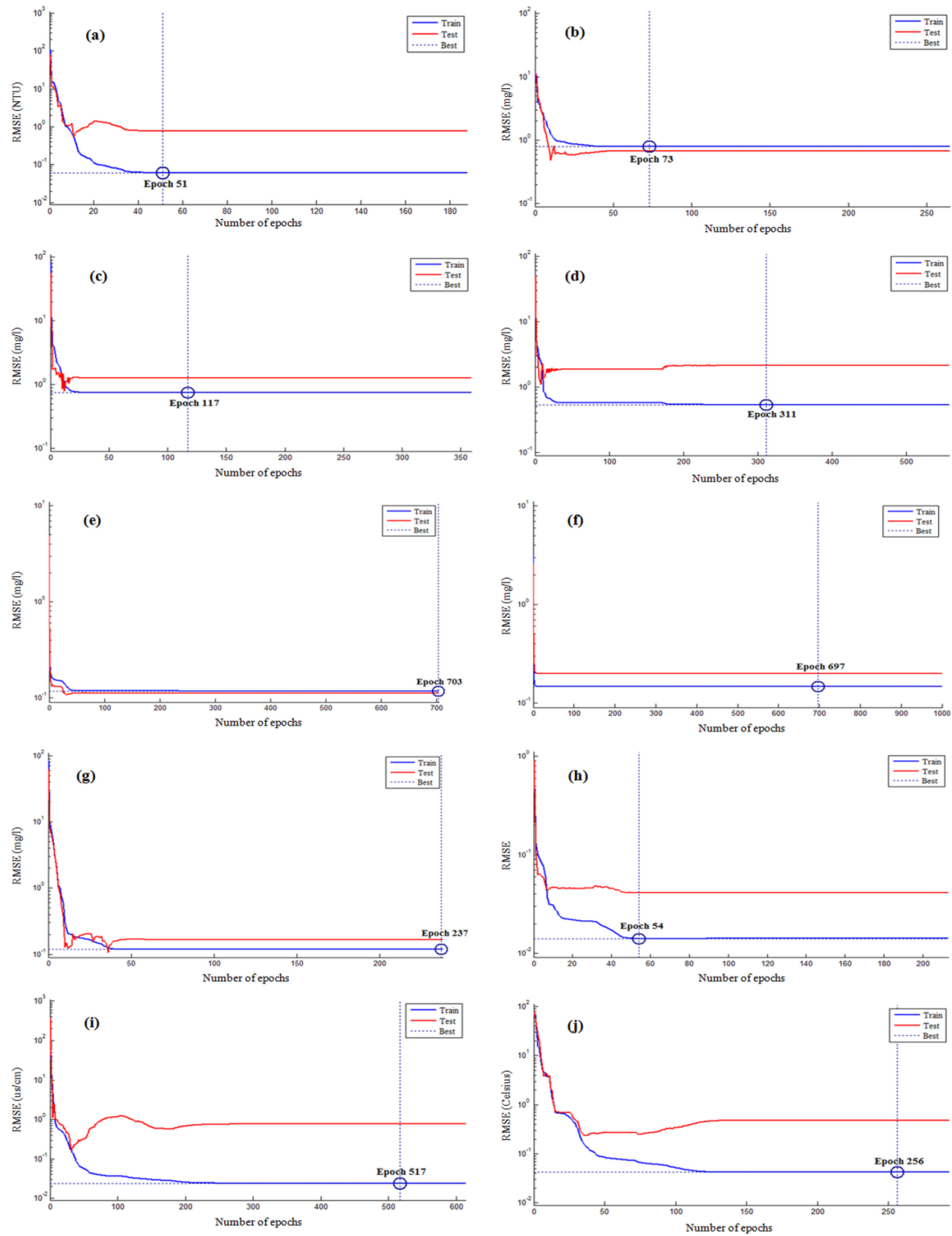


Figure 4.7 Error surfaces for turbidity (a), TSS (b), TS (c), TDS (d), COD (e), BOD (f), DO (g), pH (h), EC (i), and temperature (j) at the network training and testing phases

Moreover, as shown in **Figure 4.7**, turbidity, TSS, TS, TDS, COD, BOD, DO, pH, EC, and temperature error surfaces showed that the training process was stopped at epoch 51, 73, 117, 311, 703, 697, 237, 54, 517, and 256, respectively. Actually, there is no further enhancement in the ANN performance after reaching the stopping points.

Overall, the developed BPNN algorithm was used to produce highly accurate estimations of optical and non-optical SWQPs compared to regression techniques which have been used in previous studies. The main basis is that the BPNN has the potential to map the non-linear relationship between satellite multi-spectral information and concentrations of different SWQPs without prior knowledge of the parameter relationship. Moreover, the BPNN can lead to good generalization, minimizing the complexity, and accelerating the computational speed of the network.

4.3.3 Extra Validation of the Developed Approach using Ground Truth Data

The main purpose of this part is to extra validate the developed approach and the developed BPNN water quality models in order to demonstrate the potential of using these models as a predictive tool in the study of water quality in other parts of the SJR, tributaries of the SJR, and other water bodies in New Brunswick. In this context, two additional sets of ground truth data (i.e. water quality data) in New Brunswick were used to further test and examine the validity and stability of the developed approach. **Figure 4.8** shows the first dataset and the rivers of interest are Saint John, Oromocto, Nashwaak, Keswick, Big Presque, Miramichi, Tobique, Aroostook, and Madawaska; while **Figure 4.9** shows the second set of water samples and the rivers of interest are Croix, Digdeguash, Magaguadavic, Lepreau, Hammond, Kennebecasis, Petitcodiac, Canaan,

Buctouche, Richibucto, and Salmon. The water samples for the first dataset were collected on September 22nd 2015 and September 29th 2015; while the samples of the second set were collected on April 28th 2015 and May 05th 2015.

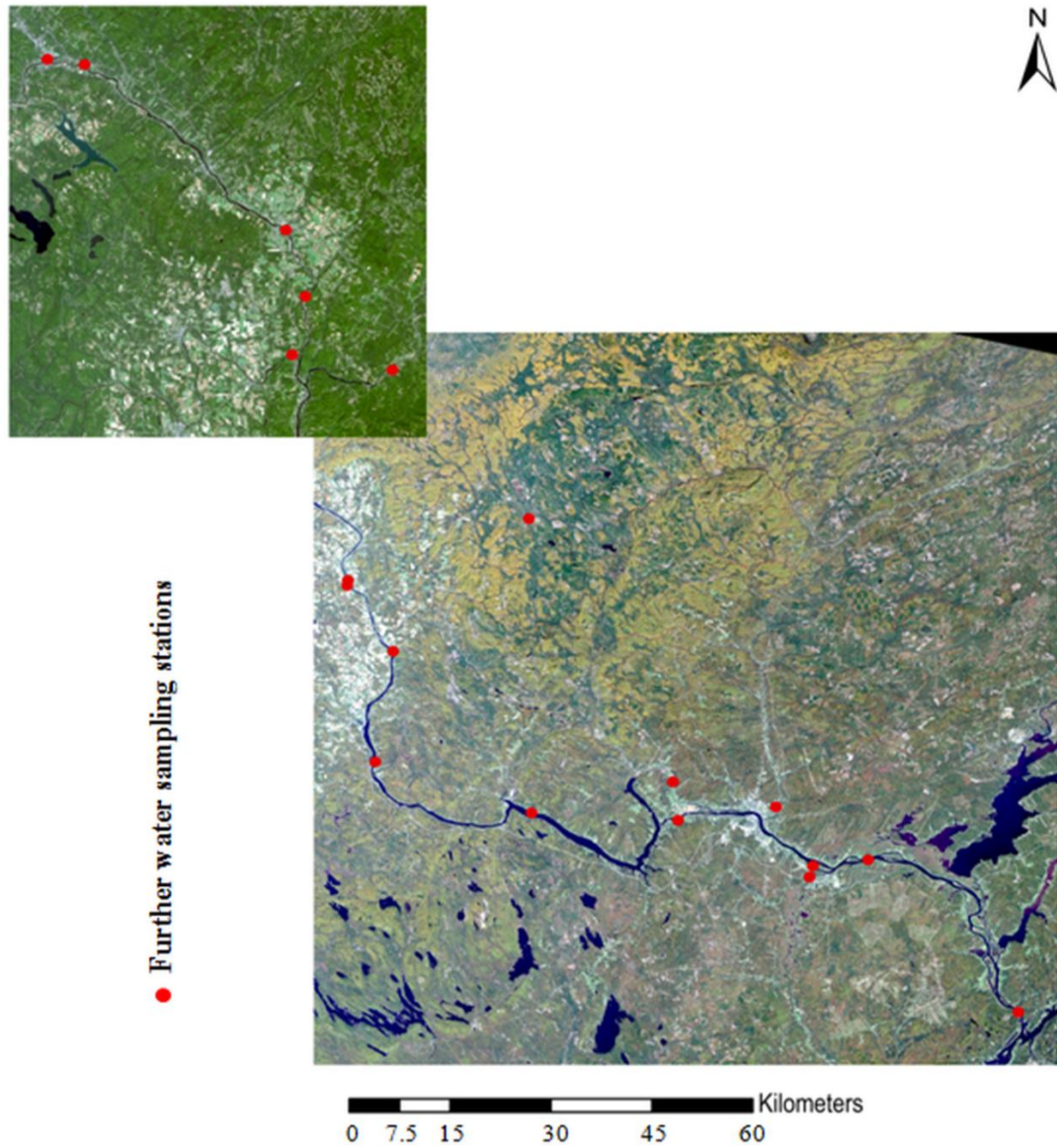


Figure 4.8 The 1st dataset of water samples used for further validation of the developed approach

The collected samples were measured for turbidity, TDS, DO, pH, EC, and temperature. The concentrations of these SWQPs were obtained from the Environment and Local Government Surface Water Quality Data Portal in New Brunswick; however, surface water quality data for TSS, TS, COD, and BOD were not available.

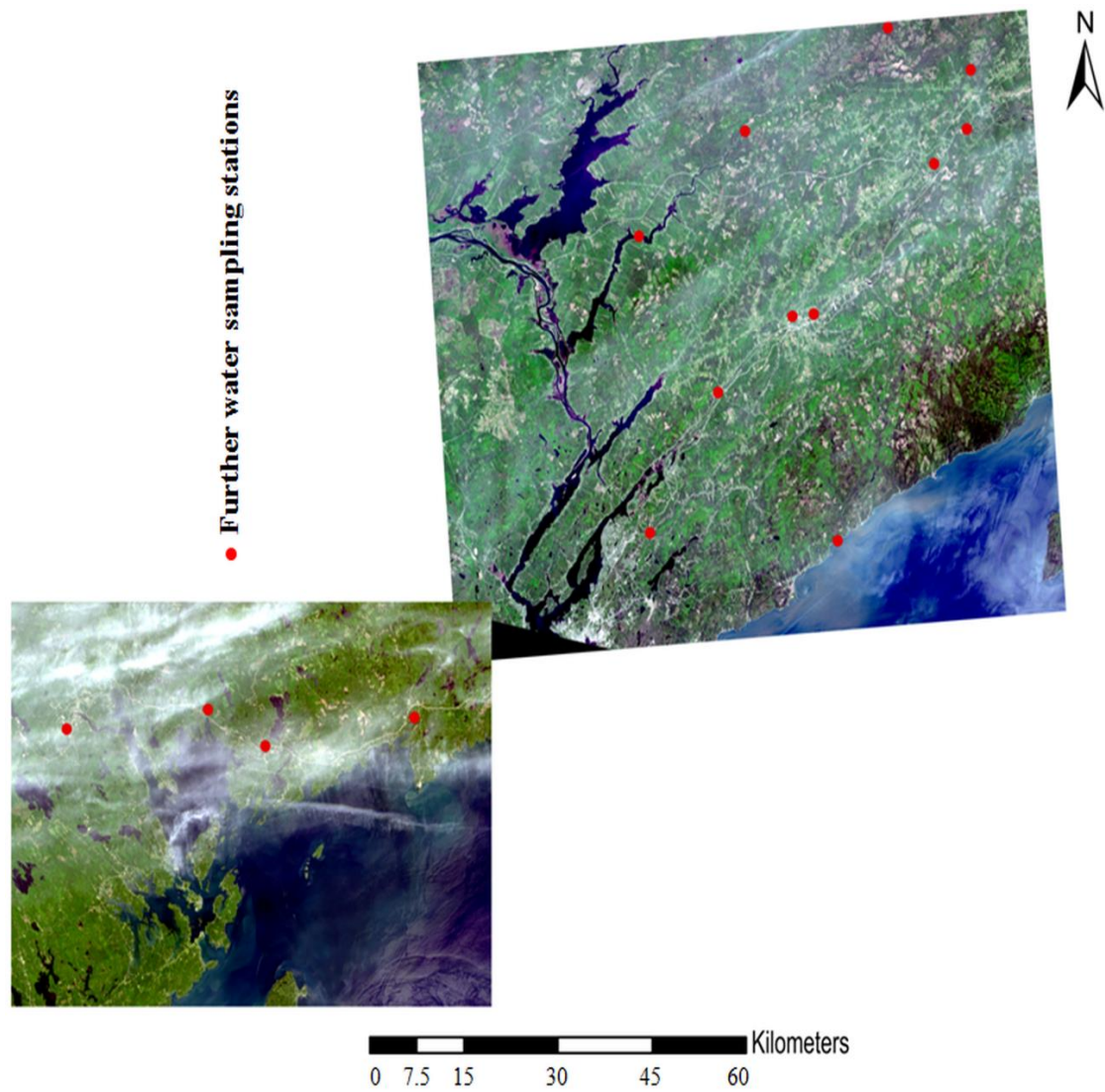


Figure 4.9 The 2nd dataset of water samples used for further validation of the developed approach

We attempted to set the time interval between ground truth data and the corresponding Landsat 8 OLI data to be very small in order to minimize the effects of the temporal difference between them. Therefore, two Landsat 8 OLI images, acquired at September 6th 2015 (top left) and September 15th 2015 (bottom right), were used with the first dataset, as shown in **Figure 4.8**. Moreover, another two Landsat 8 OLI images, acquired at June 4th 2015 (top right) and June 11th 2015 (bottom left), were used with the second dataset, as shown in **Figure 4.9**.

The developed approach was used to predict the concentrations of turbidity, TDS, DO, pH, EC, and temperature in the SJR and its tributaries and other water bodies in New Brunswick by using the two datasets of input data. In order to evaluate the validity of the developed models, the predicted results were compared against the existing ground truth data. As shown in **Figure 4.10**, for the first dataset, the developed models for turbidity, TDS, DO, pH, EC, and temperature were very stable with $R^2 = 0.828, 0.777, 0.792, 0.767, 0.882, \text{ and } 0.781$, respectively. Due to the time interval (i.e. 2 to 3 weeks) which may increase the effects of the temporal difference between field measurements and Landsat 8 multi-spectral information, the results are not higher enough ($R^2 \geq 0.767$) compared to the results obtained from the water samples which have been acquired at the same time of satellite overpass ($R^2 \geq 0.803$).

For the second set of data, the time interval between water sampling and multi-spectral data was around one month, which may lead to a lot of variability of the predicted results. However, as shown in **Figure 4.11**, the developed models for turbidity, TDS, DO, pH, EC, and temperature remained stable with $R^2 = 0.795, 0.759, 0.775, 0.755, 0.832, \text{ and } 0.761$, respectively.

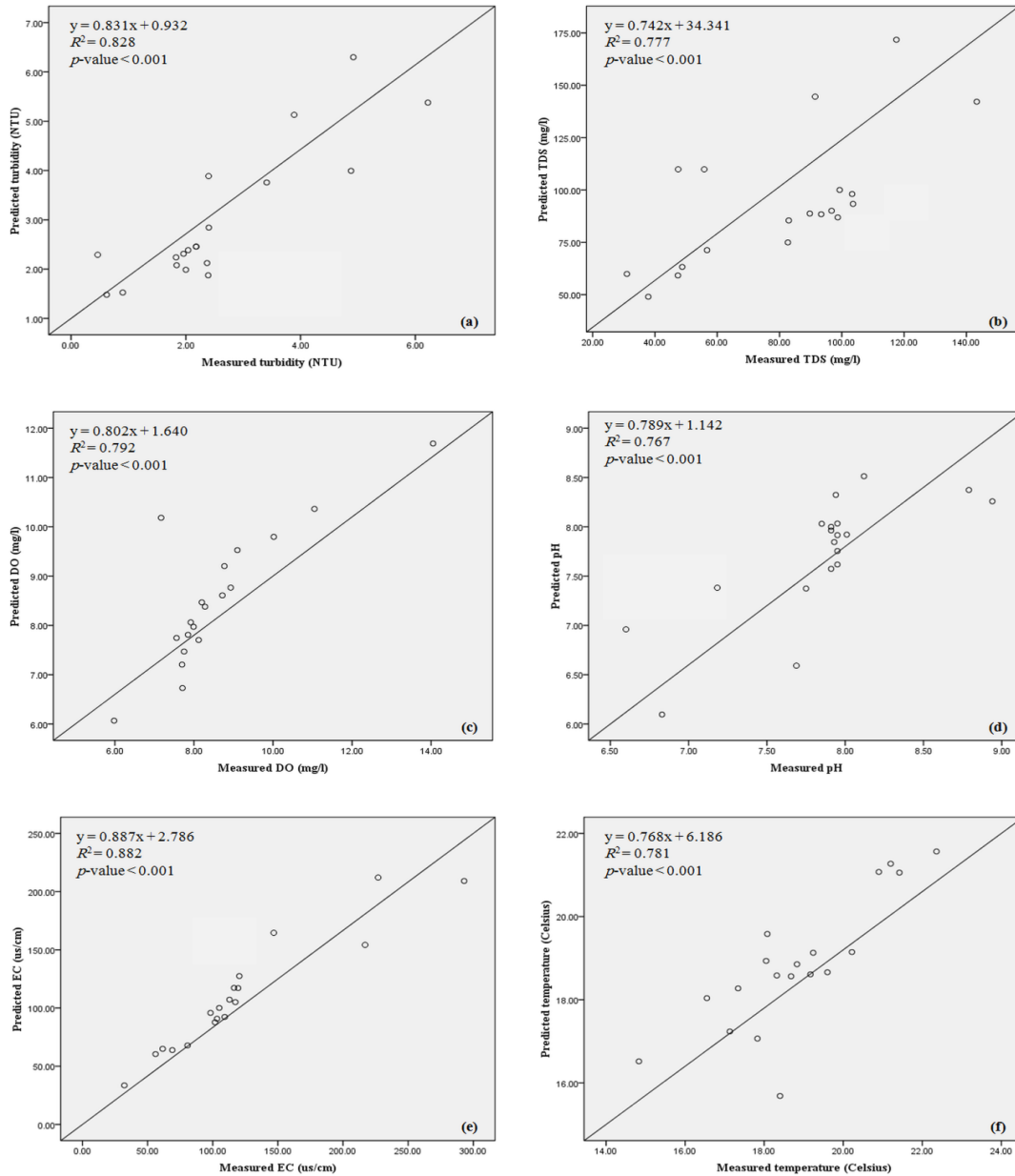


Figure 4.10 Scatter plots of measured vs. predicted concentrations of turbidity (a), TDS (b), DO (c), pH (d), EC (e), and temperature (f) using the 1st dataset

Finally, the results obtained demonstrated the potential of developing generalized models to estimate concentrations of both optical and non-optical SWQPs in the SJR and its tributaries without being dependent on river sampling.

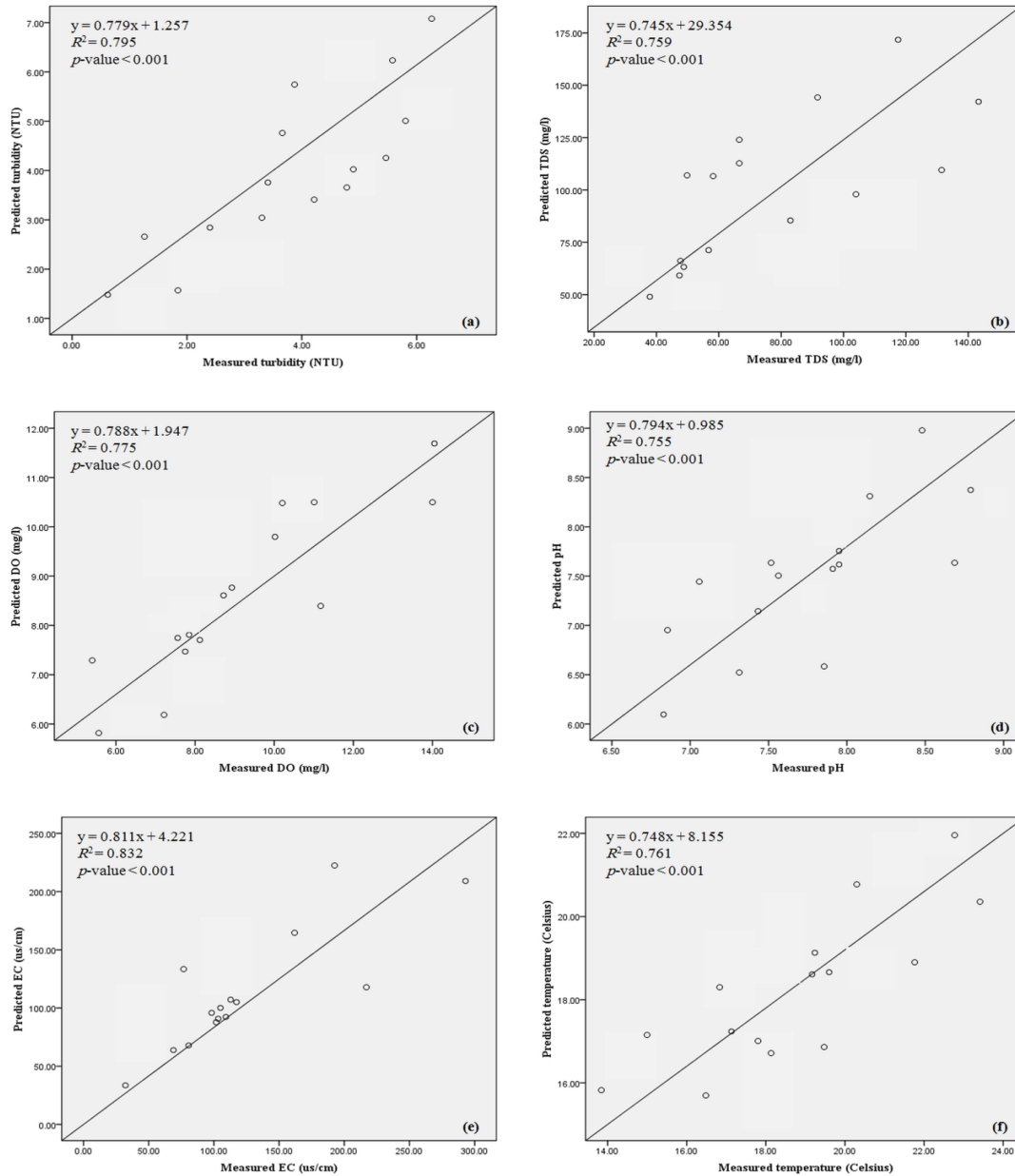


Figure 4.11 Scatter plots of measured vs. predicted concentrations of turbidity (a), TDS (b), DO (c), pH (d), EC (e), and temperature (f) using the 2nd dataset

4.3.4 Spatial Distribution of the Selected SWQPs

Figure 4.12 indicated that the obtained Landsat 8 surface reflectance of water pixels were used as an input to the developed BPNN models in order to generate spatial

distribution maps for turbidity, TSS, TS, TDS, COD, BOD, DO, pH, EC, and water temperature.

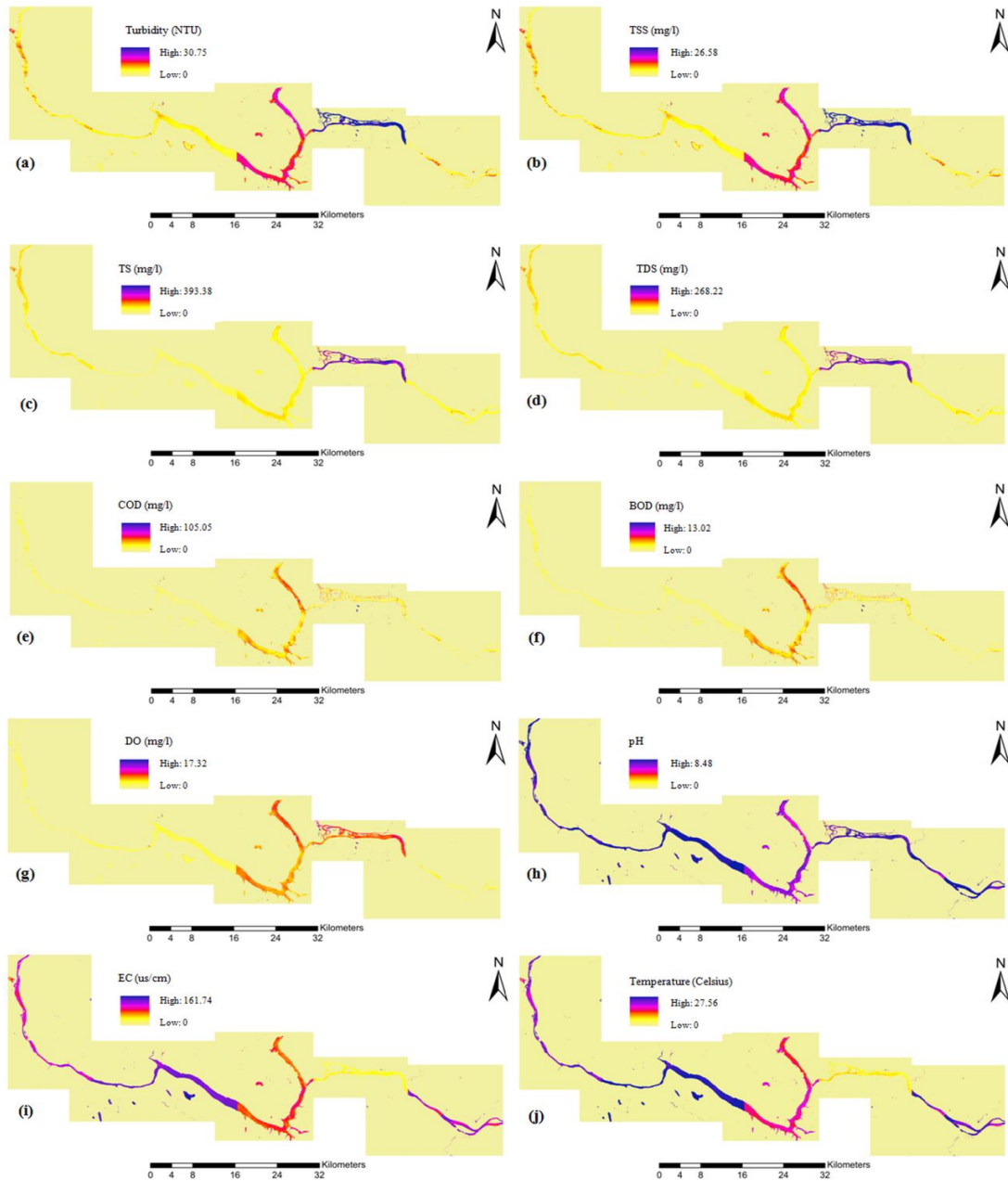


Figure 4.12 Mapping the concentrations of turbidity (a), TSS (b), TS (c), TDS (d), COD (e), BOD (f), DO (g), pH (h), EC (i), and temperature (j) in the selected study area

It can be noted that the concentrations of turbidity, TSS, TS, TDS, and DO in the SJR depend on sampling time. In spring (i.e. April and May), these concentrations were found to be higher than those sampled in summer (i.e. June, July, and August) because snow melt and rainfall may cause soil erosion and consequently push sediments and pollutants from forest and agricultural fields directly into the river. Moreover, concentrations of COD and BOD in the middle basin of the SJR were higher than those in the lower basin of the river due to classifying the middle basin as an industrial area and consequently containing higher levels of organic wastes. Additionally, levels of EC increase as temperature increases. This means that warmer water can hold higher levels of EC than colder waters. High levels of EC can be used as an indication of inorganic dissolved solids and minerals. Accordingly, the presence of free minerals increases the alkalinity of water (i.e. pH levels).

4.3.5 Delineating the Accurate Levels of Surface Water Quality of the SJR

In order to achieve highly accurate estimations of surface water quality levels of the SJR by using the CCMEWQI, the selected study area was subdivided into two main sites: (1) below the Mactaquac Dam and (2) above the Mactaquac Dam. As shown in **Figure 4.13**, twenty-eight water samples were collected below the Mactaquac Dam during trip 1 and trip 2. Instead of using twenty-eight water samples, 47544 water pixels, derived from the developed BPNN with $R^2 \geq 0.803$, were used as an input to the CCMEWQI to extract the accurate water quality level below the dam. Similarly, thirty-eight samples were collected above the Mactaquac Dam during trip 3, 4, and 5. Rather

than using thirty-eight samples, 100606 water pixels were used to delineate the exact water quality level above the dam.

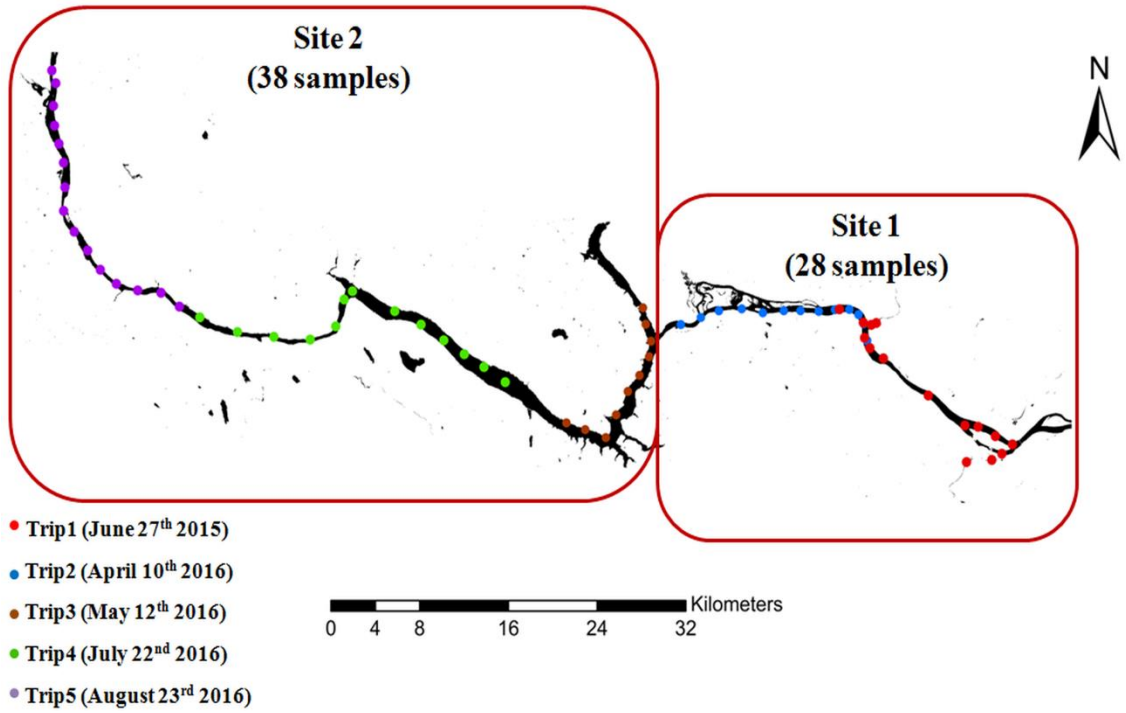


Figure 4.13 Mapping the concentrations of turbidity (a), TSS (b), TS (c), TDS (d), COD (e), BOD (f), DO (g), pH (h), EC (i), and temperature (j) in the selected study area

Based on the result findings, the CCMEWQI calculations were carried out and the concentrations of TS, TDS, and pH were found within the standard limits; however, turbidity, TSS, COD, BOD, DO, EC, and temperature values exceeded the permissible limits given by the CCME and WHO standards for drinking purposes. The obtained CCMEWQI was observed as 67 (Fair) in the lower basin of the SJR, which means the water quality is usually protected but occasionally threatened or impaired. Moreover, the water quality in the middle basin of the SJR was classified as 59 (Marginal), which means the water quality is frequently threatened or impaired. The main reason for

obtaining different levels of water quality at the proposed sites of the SJR is that the lower basin of the river has less industrial and agricultural activity, which may keep this part of the river in a better state than the middle basin of the SJR.

4.4 Conclusion

While traditional methods of assessing water quality are mainly based on comparing the measured concentrations of SWQPs with the existing guidelines, they could not provide the overall trends of water quality to non-experts, such as decision-makers. Hence, the Landsat 8-based-CCMEWQI approach is developed to extract accurate levels of water quality to be accessible to decision-makers. The CCMEWQI was selected because it is capable of minimizing the data volume to a great extent and simplifying the expression of water quality. Moreover, the CCMEWQI is very flexible in selecting input parameters (i.e. physico-chemical SWQPs).

Our approach was validated using two sets of ground truth data (i.e. water quality data) provided by the Environment and Local Government Surface Water Quality Data Portal in New Brunswick. The time interval between the existing ground truth data and the corresponding Landsat 8 multi-spectral data is 2 to 5 weeks, which may cause a lot of deviation of the predicted results. However, our approach remained very stable and the relationship between concentrations of SWQPs and Landsat 8 surface reflectance is correlated with $R^2 > 0.75$.

The results of this study show the potential of generating generalized models to retrieve concentrations of SWQPs from satellite imagery in the SJR, its tributaries, and other water bodies. Additionally, this study is valuable for decision-makers, local

managers, and the general public because the CCMEWQI mechanism gives comparative evaluation of the water quality of sampling sites and summarizes complex water quality data into simplified mathematical numbers, which can be interpreted into text classes, such as excellent, good, fair, marginal, and poor. Finally, further studies are needed to assess water quality on the basis of identifying and classifying the major SWQPs that contribute to water quality variation in the SJR.

Acknowledgements

This research is partly funded by the Egyptian Ministry of Higher Education, Bureau of Cultural & Educational Affairs of Egypt in Canada, and the Canada Research Chair Program. We thank Prof. Dr. Katy Haralampides and Dr. Dennis Connor for their help in field data water sampling and experimental analysis. We express thanks to the USGS for providing the Landsat 8 data product. We also would like to thank the anonymous reviewers for their valuable comments and constructive suggestions that helped improve this manuscript.

REFERENCES

- Akbar, A. T., Hassan, K. Q., & Achari, G. (2013). Clusterization of Surface Water Quality and Its Relation to Climate and Land Use/Cover. *Journal of Environmental Protection*, 39, pp. 333-343.
- Akoteyon, I., Omotayo, A., Soladoye, O., & Olaoye, H. (2011). Determination of water quality index and suitability of urban river for municipal water supply in Lagos-Nigeria. *Europ. J. Scientific Res.*, 54 (2), pp. 263-271.
- APHA. (2005). *Standards Methods for the Examination of Water and Wastewater* (21th ed.). American Public Health Association Washington DC, USA.

- Bharti, N., & Katyal, D. (2011). Water quality indices used for surface water vulnerability assessment. *Int. J. Environ. Sci.*, 2 (1), pp. 154-173.
- Bordalo, A. A., Teixeira, R., & Wiebe, W. J. (2006). A Water Quality Index Applied to an International Shared River Basin: The Case of the Douro River. *Environ. Manage.*, 38, pp. 910-920.
- Bunkei, M., Wei, Y., Gongliang, Y., Youichi, O., Kazuya, Y., & Takehiko, F. (2015). A hybrid algorithm for estimating the chlorophyll-a concentration across different trophic states in Asian inland waters. *ISPRS Journal of Photogrammetry and Remote Sensing*, 102, pp. 28-37.
- CCME. (2001). *Canadian water quality index 1.0 technical report and user's manual, Canadian Environmental Quality Guidelines Water Quality Index Technical Subcommittee, Gatineau, QC, Canada.*
- Changchun, H., Jun, Z., Yunmei, L., Hao, Y., Kun, S., Junsheng, L., . . . Fa, Z. (2014). Assessment of NIR-red algorithms for observation of chlorophyll-a in highly turbid inland waters in China. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, pp. 29-39.
- Chavez, P. S. (1988). An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. *Remote Sensing of Environment*, 24, pp. 459-479.
- Czarra, F. (2003). Fresh Water: Enough for You and Me? *The American Forum for Global Education*, 174, pp. 2-10.
- Debels, P., Figueroa, R., Urrutia, R., Barra, R., & Niell, X. (2005). Evaluation of water quality in the Chilla'n River (Central Chile) using physicochemical parameters and a modified water quality index. *Environ. Monit. Assess.*, 110, pp. 301–322.
- Earth Explorer*. (2016). Retrieved from U.S. Geological Survey: <http://earthexplorer.usgs.gov/>
- Hinton, E. G. (1992). How neural networks learn from experience. *Scientific American*, 267 (3), pp. 144-151.
- Horton, R. K. (1965). An Index Number System for Rating Water Quality. *Journal of the Water Pollution Control Federation*, 37 (3), pp. 300-306.

- Jena, V., Dixit, S., & Gupta, S. (2013). Assessment of Water Quality Index Of Industrial Area Surface Water Samples. *Int. J. Chem. Tech. Res.*, 5 (1), pp. 278-283.
- Khan, A. A., Paterson, R., & Khan, H. (2004). Modification and Application of the Canadian Council of Ministers of the Environment Water Quality Index (CCMEWQI) for the Communication of Drinking Water Quality Data in Newfoundland and Labrador. *Water Quality. Res. J. Can.*, 39, pp. 285-293.
- Lumb, A., Halliwell, D., & Sharma, T. (2006). Application of CCME Water Quality Index to Monitor Water Quality: A Case of the Mackenzie River Basin, Canada. *Environ. Monit. Assess.*, 113, pp. 411-429.
- MacKay, J. C. (1992, May). Computation and Neural Systems. *Neural Computation*, 4 (3), pp. 415-447.
- Marta, T., Damià, B., & Romà, T. (2010). Surface-water-quality indices for the analysis of data generated by automated sampling networks. *TrAC Trends in Analytical Chemistry*, 29 (1), pp. 40-52.
- Mcfeters, S. K. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *INT. J. REMOTE SENSING*, 17 (7), pp. 1425-1432.
- Pat, S., & Chavez, J. (1996). Image-Based Atmospheric Corrections - Revisited and Improved. *Photogrammetric Engineering & Remote Sensing*, 62 (9), pp. 1025-1036.
- Rosemond, S., Duro, D. C., & Dubé, M. (2009). Comparative Analysis of Regional Water Quality in Canada Using the Water Quality Index. *Environ. Monit. Assess.*, 156, pp. 223-240.
- Roy, D. P., Wulder, M., Loveland, T. R., & Zhu, Z. (2014). Landsat 8: Science and product vision for terrestrial global change research. *Remote Sens. Environ.*, 145, pp. 154-172.
- Sargaonkar, A., & Deshpande, V. (2003). Development of an Overall Index of Pollution for Surface Water Based on a General Classification Scheme in Indian Context. *Environ. Monit. Assess.*, 89, pp. 43-67.

- Sharaf El Din, E., Zhang, Y., & Suliman, A. (2017a). Mapping concentrations of surface water quality parameters using a novel remote sensing and artificial intelligence framework. *International Journal of Remote Sensing*, 38 (4), pp. 1023-1042.
- Sharaf El Din, E., & Zhang, Y. (2017d). Estimation of both optical and non-optical surface water quality parameters using Landsat 8 OLI imagery and statistical techniques. *Journal of Applied Remote Sensing*, 11 (4), 046008 (2017), doi: 10.1117/1.JRS.11.046008.
- Sharaf El Din, E., & Zhang, Y. (2017e). Improving the accuracy of extracting surface water quality levels (SWQLs) using remote sensing and artificial neural network: a case study in the Saint John River, Canada. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-4/W4, 245-249, <https://doi.org/10.5194/isprs-archives-XLII-4-W4-245-2017>, 2017.
- Shuisen, C., Liusheng, H., Xiuzhi, C., Dan, L., Lin, S., & Yong, L. (2015). Estimating wide range Total Suspended Solids concentrations from MODIS 250-m imageries: An improved method. *ISPRS Journal of Photogrammetry and Remote Sensing*, 99, pp. 58-69.
- Singh, K. P., Malik, A., Mohan, D., & Sinha, S. (2004). Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India): a case study. *Water Research*, 38, pp. 3980-3992.
- Song, C., Woodcock, C. E., Seto, K. C., Lenney, M. P., & Macomber, S. A. (2001). Classification and change detection using Landsat TM data: when and how to correct atmospheric effects. *Remote Sensing of Environment*, 75, pp. 230-244.
- United States Geological Survey (USGS). (2016). Retrieved from USGS Landsat 8 Product: http://landsat.usgs.gov/Landsat_8_Using_Product.php
- Zhang, Y. Z., Pulliainen, J. T., Koponen, S. S., & Hallikainen, M. T. (2002). Application of an empirical neural network to surface water quality estimation in the Gulf of Finland using combined optical data and microwave data. *Remote Sensing of Environment*, 81, pp. 327-336.

**Chapter 5: ASSESSMENT OF SPATIO-TEMPORAL SURFACE
WATER QUALITY VARIATIONS USING MULTIVARIATE
STATISTICAL TECHNIQUES: A CASE STUDY OF THE SAINT
JOHN RIVER, CANADA⁴**

Abstract

Surface water quality is a worldwide environmental concern due to the presence of both point and non-point sources of pollutants. These pollutants lead to deterioration of surface water quality and consequently raise the cost of water body treatment. To improve the cost effectiveness of the treatment process, assessing surface water quality on the basis of classifying the major surface water quality parameters (SWQPs) that negatively affect water bodies is essential. Therefore, Multivariate Statistical Techniques, such as Principal Component Analysis/Factor Analysis (PCA/FA), Cluster Analysis (CA), and Discriminant Analysis (DA), are proposed to identify the dominant SWQPs and evaluate spatial/temporal water quality variations of the Saint John River (SJR), as the testing water body. The results of PCA/FA showed that turbidity, total suspended solids, chemical oxygen demand, biochemical oxygen demand, and electrical conductivity are the most significant SWQPs contributing to variations in the water quality of the SJR. Moreover, CA and DA indicated a reduction in the dimensionality of our surface water quality data and classified sampling stations based on similarities of

⁴ This paper is under review in the “*Journal of the American Water Resources Association (JAWRA)*”.

water quality characteristics. Our study illustrates the significant use of multivariate statistical techniques for surface water quality assessment and management leading to effective savings and proper utilization of water quality resources.

5.1 Introduction

Surface water quality is generally affected by both natural and anthropogenic processes. Snow melt, precipitation rate, and sediment transport are considered as natural processes, while anthropogenic processes include urban, industrial, and agricultural activities (Carpenter, Caraco, Correll, Howarth, Sharpley, & Smith, 1998; Qadir, Malik, & Husain, 2007). These processes often lead to the degradation of surface water quality by pushing both point and non-point sources of pollutants directly into water bodies. A point source (e.g., industrial discharge) forms a constant polluting source; while a non-point source (e.g., precipitation and snow melting) is a seasonal phenomenon, largely affected by climate changes (Singh, Malik, Mohan, & Sinha, 2004).

Due to these complexities, water quality experts and researchers are confronted with significant challenges to assess surface water quality and consequently provide the appropriate treatment to water bodies in a cost-effective manner (Elhatip, Hinis, & Gulgahar, 2007). In this context, the appropriate treatment of water bodies should be targeted towards the dominant surface water quality parameters (SWQPs) that contribute to both spatial and temporal variations of water quality. This will lead to effective savings and proper utilization of resources in water quality studies (Elhatip, Hinis, & Gulgahar, 2007; Natural resources, 2016). Therefore, multivariate statistical techniques, such as principal component analysis/factor analysis (PCA/FA), cluster analysis (CA), and

discriminant analysis (DA), are proposed to help in the interpretation of complex water quality data to better understand the ecological status of water bodies. Moreover, these techniques can identify the major pollution sources that influence water systems and provide a valuable tool for reliable management of water resources as well as offering rapid solutions to control pollution problems (Vega, Pardo, Barrado, & Deban, 1998; Wunderlin, Diaz, Ame, Pesce, Hued, & Bistoni, 2001; Reghunath, Murthy, & Raghavan, 2002; Simeonov, Stratis, Samara, Zachariadis, Voutsas, & Anthemidis, 2003; Shrestha & Kazama, 2007; Akbar, Hassan, & Achari, 2011; Sharaf El Din & Zhang, 2018).

In the relevant literature, almost all of the available studies have attempted to classify the major parameters that negatively affect water bodies by using multivariate statistical techniques; however, fewer research attempts focused on extracting spatial/temporal patterns of surface water quality.

The PCA technique was used to identify the dominant SWQPs of the Neckar River, Germany based on analyzing ten SWQPs. Four principal components explained 72% of total variance. The overload of phosphorus and nitrogen were responsible for the deterioration of surface water quality in the river (Haag & Westrich, 2002).

PCA was used to assess surface water quality variations along the main stem of the lower St. Johns River, Florida, USA using sixteen physical and chemical SWQPs collected from twenty-two monitoring stations (Ouyang, Nkedi-Kizza, Wu, Shinde, & Huang, 2006). PCA was employed to evaluate the correlation between different SWQPs and to extract the major parameters in the river. The results showed that electrical conductivity and dissolved organic carbon were the most important SWQPs contributing to the river water quality variations.

PCA and CA were used to interpret a large water quality dataset collected from the Songhua River Basin, China (Li, Xu, & Li, 2009). The data set, which contained fourteen SWQPs, was collected from fourteen different sampling sites along the river. Three significant sampling locations (i.e. less polluted, moderately polluted, and highly polluted) were detected by CA and five factors (i.e. organic, inorganic, petrochemical, physiochemical, and heavy metals) were identified by PCA.

PCA and CA were used to monitor variations of surface water quality in Sanya Bay, China (Dong, Zhang, Zhang, Wang, Yang, & Wu, 2010). The water quality associated with one station was impacted by Sanya River and the water quality associated with the rest of sampling stations was influenced by South China Sea. It was concluded that rainfall was responsible for the water quality variations of Sanya Bay.

PCA and CA were employed to detect the major pollutants that affect surface water quality variations at Qiantang River, China (Huang, Wang, Lou, Zhou, & Wu, 2010). Low, moderate, and high pollution zones were identified and classified. Two pollution sources in each of low and moderate pollution zones with 67% and 73% of total variance, respectively, were identified. Moreover, three pollution sources in high pollution zone explained 80% of total variance. Industrial and agricultural activities in addition to urban runoff were considered as the main sources of pollution.

PCA was used for interpretation of a water quality dataset obtained from the River Ganga in Varanasi, India (Mishra, 2010). Sixteen physicochemical SWQPs were measured and analyzed. The dataset was treated using PCA and four Principal components were identified as responsible for explaining 90% of the total surface water quality variance of the dataset.

CA was used to assess variations in the water quality of Euphrates River, Iraq by using sixteen parameters collected from eleven sites (Salah, Turki, & Al-Othman, 2011). CA classified the eleven sampling sites into two groups based on similarities of water quality characteristics. The results of this study showed that water quality data collected in April has higher pollution level related to the other months. This study indicated the usefulness of CA in the interpretation of surface water quality in the selected study area.

PCA was applied to groundwater samples, which were collected from ten sources and analyzed for ten SWQPs (Mahapatra & Mitra, 2012). Four components were used to classify water samples and this process was found to be very helpful for water quality experts and managers to improve data collection and avoid groundwater contamination.

Based on what has been reviewed, the use of multivariate statistical techniques has been suggested in most cases because these techniques can be applied to understand the relationships between different SWQPs and their relevance to the actual problem being studied. Due to the redundancy and complexity of relationships between parameters of water quality, it is not easy to draw a clear conclusion directly from the water quality raw data. Therefore, we need a tool (i.e. multivariate statistical techniques, such as PCA/FA, CA, and DA) that is capable of detecting both spatial and temporal variations of water quality as well as categorizing the dominant SWQPs that influence the water quality of the water body under investigation. The advantages of multivariate statistical techniques include: 1) usefulness in finding the association between samples and parameters and revealing the information which cannot be observed from the raw data, 2) reduction in the complexity of large-scale datasets, 3) identification of the major parameters by reducing large dataset into groups with similar properties, and 4)

efficiency in surface water quality studies (Reghunath, Murthy, & Raghavan, 2002). On the other hand, the disadvantages of multivariate statistical techniques include 1) the presence of same parameters in different principal components (PCs) which may change the interpretation of water quality condition in water bodies and 2) difficulty in finding out the suitable number of clusters (Singh, Malik, Mohan, & Sinha, 2004).

The identified objectives of this research are to: (1) classify the major SWQPs that contribute to surface water quality variations in the Saint John River (SJR), New Brunswick, Canada by using PCA/FA technique, (2) develop multiple levels of clustering to detect the relationship between the collected water samples by using hierarchical agglomerative CA technique, and (3) evaluate both spatial and temporal variations of surface water quality of the selected study area of the SJR by using DA technique. To the best of our knowledge, PCA/FA, CA, and DA were combined for the first time to identify the major pollution sources contributing to surface water quality variations in the SJR with inexpensive implementation cost.

5.2 Materials and Methods

5.2.1 Study Area

The SJR originates principally in the Canadian province of New Brunswick, covering an area of 4748 km². Many tributaries, such as Oromocto, Nashwaak, Keswick, Tobique, Aroostook, and Madawaska, feed the SJR. The river's average width is 750 m and its average depth is 3 m. The SJR is considered to have a cold climate except near the Bay of Fundy coast, which has a maritime climate. The river's mean annual temperature and annual precipitation are 5 °C and 140 cm, respectively (Arseneault, 2008). Moreover,

area. Coordinates of each sample were recorded in the field by using a handset global positioning system (GPS), GARMIN 76CSx. The collection, preservation, and analysis of the collected water samples were carried out as prescribed by the standards given by the American Public Health Association (APHA) (APHA, 2005).

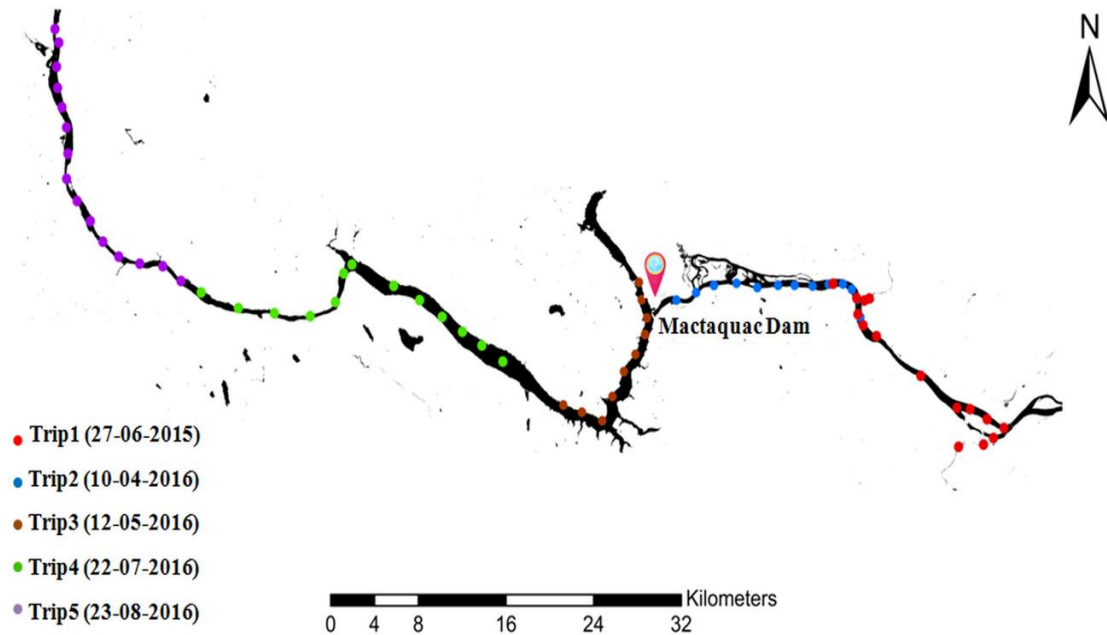


Figure 5.2 The collected water sampling stations

Concentrations of both optical and non-optical SWQPs, such as turbidity (Turb), total suspended solids (TSS), total solids (TS), total dissolved solids (TDS), chemical oxygen demand (COD), biochemical oxygen demand (BOD), dissolved oxygen (DO), power of hydrogen (pH), electrical conductivity (EC), and temperature (Temp), were measured according to the APHA water and wastewater standards. Turb was measured in situ with a portable turbidity-meter which can measure the amount of light scattered by suspended particles in the water column. TSS was determined by filtering each water sample and weighing the residue left on the filter paper. TS was calculated by

evaporating each water sample and weighing the remaining dry residue, while the difference between TS and TSS was used to calculate TDS. COD was estimated by using the closed reflux titrimetric method. BOD was determined by 5-day BOD test at 20 °C. DO levels were measured in situ by using a portable DO-meter. pH, EC, and Temp were tested in the field using a portable pH-meter.

5.2.3 Multivariate Statistical Techniques

Water quality data were subjected to multivariate statistical techniques to extract the parameters which were responsible for spatio-temporal water quality variations in the selected study area. The main concept of employing multivariate statistical techniques, such as PCA/FA, CA, and DA, is provided in the following subsections.

5.2.3.1 Principal Component Analysis/Factor Analysis (PCA/FA) Technique

PCA is a mathematical concept designed to linearly transform the original variables (e.g., SWQPs) into new uncorrelated variables (axes), called principal components (PCs). The new axes lie along the directions of maximum variance. PCA can provide information about the most significant variables within the dataset leading to data reduction with minimum loss of original information (Shrestha & Kazama, 2007). PCs can be expressed as:

$$Z_{ij} = a_{i1} * x_{1j} + a_{i2} * x_{2j} + \dots + a_{im} * x_{mj} \quad (5.1)$$

where Z is the component score; a is the component loading; x is the measured value of a variable (SWQP concentration); i is the component number; j is the sample number; m

is the total number of variables.

FA follows PCA to further reduce the contribution of variables (e.g., SWQPs) with minor significance and to keep only the major variables to simplify even more of the data structure coming from PCA technique. In order to do that, PCs were subjected to varimax rotation to generate new variables, called varifactors (Vega, Pardo, Barrado, & Deban, 1998; Simeonov, Stratis, Samara, Zachariadis, Voutsas, & Anthemidis, 2003). As a result, a small number of variables would usually account for approximately the same amount of information as do the much larger set of original variables. FA can be expressed as:

$$Z_{ji} = a_{f1} * f_{1i} + a_{f2} * f_{2i} + \dots + a_{fm} * f_{mi} + e_{fi} \quad (5.2)$$

where Z is the measured variable; a is the factor loading; f is the factor score; e is the residual term accounting for errors or other sources of variation; i is the sample number; m is the total number of factors.

In our study, the main reasons of using PCA/FA technique are to 1) obtain the major PCs by using a cutoff eigenvalue, 2) determine the loading values for all SWQPs under the major PCs, and 3) diminish the number of the selected SWQPs as much as possible.

5.2.3.2 Cluster Analysis (CA) Technique

CA is a multivariate statistical technique which categorizes entities (e.g., water sampling stations) into distinct groups or clusters based on the characteristics they

possess. K-means and hierarchical clustering are the most common approaches, which can provide intuitive similarity relationships between any sample and the entire dataset. The Euclidean distance can be used to provide the similarity between two samples and a distance can be represented by the difference between analytical values from the samples (McKenna & J.E., 2003).

In our study, hierarchical agglomerative CA was performed on the dataset by means of the Ward's method and squared Euclidean distance as a measure of similarity. The outcome of hierarchical agglomerative CA is visualized by a dendrogram (a tree-like plot), which gives a visual summary of the clusters and their similarity with a dramatic reduction in dimensionality of the original dataset (Shrestha & Kazama, 2007).

The main reasons of conducting hierarchical agglomerative CA technique are to 1) generate multiple levels of clustering to find out the association between sampling stations at different levels, unlike other clustering techniques (e.g., K-means), 2) provide a visual summary of the obtained clusters leading to better understanding of water quality status, and 3) categorize the characteristics of clusters using the dominant parameters.

5.2.3.3 Discriminant Analysis (DA) Technique

DA attempts to describe relationships between two or more pre-specified groups (clusters) of entities based on a set of two or more discriminating variables. DA is used when groups are known *a priori*, unlike in CA. This technique works by deriving one or more linear combinations of discriminator variables, creating a new variable for each function (Singh, Malik, Mohan, & Sinha, 2004). These functions are called “discriminant functions”. The number of discriminant functions possible is either the (number of groups

– 1), or the number of variables, whichever is smaller. The first discriminant function maximizes the differences between groups on that function. The second discriminant function maximizes differences on that function, but must also not be correlated with the first function. This process continues with subsequent functions with the requirement that the new function is not correlated with any of the previous functions (Singh, Malik, Mohan, & Sinha, 2004). Each discriminant function has the general form:

$$D = a + b_1X_1 + b_2X_2 + \dots + b_pX_p \quad (5.3)$$

where D is the discriminant function score (z score); a is the intercept of the regression line; b is the discriminant function coefficient; X is the discriminator variable score; p is the number of discriminator variable.

In our study, the main reasons for using DA technique are to 1) determine the most significant variables associated with differences among the groups and 2) detect both spatial and seasonal variations of surface water quality in the SJR.

5.3 Results and Discussion

Our study aims at classifying the dominant SWQPs that negatively affect water bodies, as well as extracting spatial/temporal water quality variations in the selected study area of the SJR, as the testing water body. The main results obtained from this study include (1) analysis of physico-chemical SWQPs, (2) extraction of the major SWQPs in the SJR using PCA/FA technique, (3) generating multiple levels of clustering using hierarchical agglomerative CA technique, and (4) delineation of both spatial and

seasonal variations of water quality using DA technique. These results are discussed in the following subsections.

5.3.1 Physico-chemical Analysis of SWQPs

The statistical summary of the selected parameters for the water samples was presented in **Table 5.1**.

Table 5.1 Statistics of physico-chemical surface water quality parameters (SWQPs).

Surface water quality parameters (SWQPs)	Mean	Standard deviation
Turbidity (Turb) (NTU)	4.84	3.73
Total suspended solids (TSS) (mg l ⁻¹)	3.59	3.10
Total solids (TS) (mg l ⁻¹)	113.92	42.32
Total dissolved solids (TDS) (mg l ⁻¹)	110.33	39.91
Chemical oxygen demand (COD) (mg l ⁻¹)	27.55	19.85
Biochemical oxygen demand (BOD) (mg l ⁻¹)	1.75	0.52
Dissolved oxygen (DO) (mg l ⁻¹)	9.54	2.64
Power of hydrogen (pH)	7.59	0.33
Electrical conductivity (EC) (us cm ⁻¹)	97.09	30.53
Temperature (Temp) (°C)	15.92	6.97

A total of ten physico-chemical SWQPs (i.e. Turb, TSS, TS, TDS, COD, BOD, DO, pH, EC, and Temp) were analyzed from sixty-six water sampling stations in the SJR by using standard methods given by APHA. Turb levels varied from 1.19 to 13.10 NTU with a mean value of 4.84 NTU. Concentrations of TSS ranged from 0.60 to 11.40 mg l⁻¹

with an average 3.59 mg l⁻¹. While TS varied from 58.00 to 245.00 mg l⁻¹, TDS ranged from 52.40 to 233.85 mg l⁻¹. COD, BOD, DO, pH, EC, and Temp ranged from 4.80 to 86.64 mg l⁻¹ with an average 27.55 mg l⁻¹, 1.21 to 3.25 mg l⁻¹ with an average 1.75 mg l⁻¹, 6.71 to 14.14 mg l⁻¹ with an average 9.54 mg l⁻¹, 6.51 to 8.42 with an average 7.59, 29.50 to 148.90 us cm⁻¹ with an average 97.09 us cm⁻¹, and 5.00 to 23.30 °C with an average 15.92 °C, respectively.

Table 5.2 The correlation matrix for the measured SWQPs.

	Turb	TSS	TS	TDS	COD	BOD	DO	pH	EC	Temp
Turb	1.00	0.92	0.79	0.77	0.69	0.65	-0.77	0.48	0.51	0.80
TSS	0.92	1.00	0.80	0.76	0.65	0.61	-0.62	0.52	0.54	0.67
TS	0.79	0.80	1.00	0.99	0.44	0.39	-0.50	0.34	0.37	0.50
TDS	0.77	0.76	0.99	1.00	0.39	0.31	-0.48	0.29	0.33	0.47
COD	0.69	0.65	0.44	0.39	1.00	0.81	-0.80	0.25	0.22	0.37
BOD	0.65	0.61	0.39	0.31	0.81	1.00	-0.78	0.19	0.21	0.42
DO	-0.77	-0.62	-0.50	-0.48	-0.80	-0.78	1.00	-0.38	-0.58	-0.97
pH	0.48	0.52	0.34	0.29	0.25	0.19	-0.38	1.00	0.66	0.25
EC	0.51	0.54	0.37	0.33	0.22	0.21	-0.58	0.66	1.00	0.64
Temp	0.80	0.67	0.50	0.47	0.37	0.42	-0.97	0.25	0.64	1.00

In April and May (i.e. spring), levels of Turb and TSS were found to be higher than their concentrations in June, July, and August (i.e. summer). The reason is that snow melt and rainfall assigned to spring season can cause soil erosion and consequently wash sediments from agriculture and forestry directly into the SJR (Sharaf El Din, Zhang, & Suliman, 2017a; Sharaf El Din & Zhang, 2017d; Sharaf El Din & Zhang, 2017e).

Additionally, the lower basin of the SJR (i.e. below Mactaquac Dam) has less agricultural, forestry, and industrial activity, which may keep this part of the river in a better state than the middle basin of the river (i.e. above Mactaquac Dam) (Arseneault, 2008).

The correlation coefficient between the measured SWQPs was calculated, as shown in **Table 5.2**. Based on the obtained correlation coefficients, the relationship between Turb and TSS was highly correlated because TSS is commonly used as the main indicator of Turb. Moreover, the relationship between DO levels and the rest of the measured SWQPs (i.e. Turb, TSS, TS, TDS, COD, BOD, pH, EC, and Temp) was negatively correlated. That means once one of these SWQPs increase, DO levels decrease, which may lead to the deterioration of surface water quality and aquatic life.

5.3.2 Multivariate Statistical Analysis

5.3.2.1 Principal Component Analysis/Factor Analysis (PCA/FA) Technique

In order to evaluate the most significant SWQPs in the selected study area of the SJR, the analysis was performed using PCA/FA multivariate statistical technique. The analysis was executed on ten SWQPs for the sixty-six water sampling points in different months (i.e. April, May, June, July, and August) in order to reduce the dimensions of the original water quality dataset and to identify the major factors affecting surface water quality.

PCA extracted a set of PCs along with their corresponding eigenvalues. An eigenvalue provides a measure of the importance of the obtained PCs. The PCs with the

highest eigenvalues are the most significant. Eigenvalues of ≥ 1 are considered significant (Shrestha & Kazama, 2007).

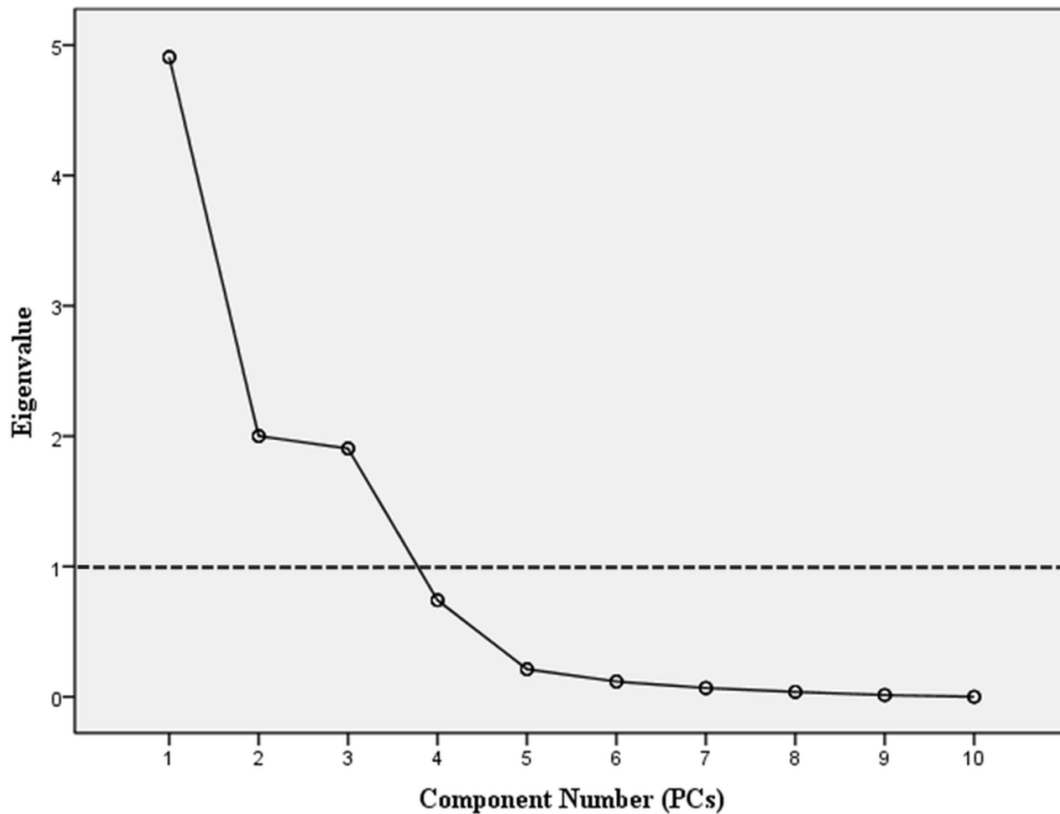


Figure 5.3 Scree plot of the produced PCs and their respective eigenvalues

As shown in the scree plot in **Figure 5.3**, the first three PCs (i.e. PC₁, PC₂, and PC₃) have eigenvalues > 1 and are considered to be the major PCs. These three PCs entirely explained 88.126% of the total variance in the water quality dataset, as shown in **Table 5.3**. PC₁, PC₂, and PC₃ captured approximately 49%, 20%, and 19% of the total variance, respectively.

FA was performed on the extracted PCs by using varimax rotation to improve the interpretation of PCA, as it increased the absolute values of larger loadings and reduced

the absolute values of smaller loadings within each PC. The loading values are classified into three main classes as strong (loading values ≥ 0.75), moderate ($0.75 >$ loading values ≥ 0.50), and weak ($0.50 >$ loading values ≥ 0.40) (Liu, Lin, & Kuo, 2003). Each SWQP with a loading value > 0.75 was considered to be a significant parameter contributing to surface water quality variations in the selected water body. Additionally, SWQPs with loading values less than 0.40 should not be considered due to their minor significance.

Table 5.4 reveals the corresponding loading values for each of the major three PCs.

Table 5.3 The principal components (PCs) along with their respective eigenvalues and the percentage of variance.

Principal components (PCs)	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	4.907	49.071	49.071	4.907	49.071	49.071
2	2.000	20.005	69.076	2.000	20.005	69.076
3	1.905	19.050	88.126	1.905	19.050	88.126
4	0.741	7.407	95.533			
5	0.211	2.116	97.649			
6	0.117	1.170	98.819			
7	0.067	0.673	99.492			
8	0.038	0.375	99.867			
9	0.013	0.132	99.999			
10	0.001	0.001	100.000			

PC₁ revealed that four SWQPs (i.e. Turb, TSS, TS, and TDS) were correlated with each other. Turb, TSS, TS, and TDS were found to be loaded as strong (i.e. > 0.75) with positive values. In PC₁, Turb and TSS are the most significant SWQPs contributing to spatial/temporal variations of surface water quality in the SJR. In particular, the increment of Turb and TSS levels may be associated with the erosion effect during cultivation of soil, natural processes, such as snow melt and rainfall, and the anthropogenic activities, such as agricultural, mining, forestry, and industrial. The generated results were found to be compatible with the results obtained by the New Brunswick Department of Natural Resources (Arseneault, 2008).

Table 5.4 The loading values of SWQPs for the significant PCs.

SWQPs	Significant components (PCs)		
	PC ₁	PC ₂	PC ₃
Turb	0.940	-0.278	
TSS	0.942	-0.288	
TS	0.899	0.229	-0.128
TDS	0.898	0.265	-0.134
COD	0.114		0.968
BOD	-0.112	0.127	0.942
DO	-0.281	-0.413	0.717
pH	0.118	0.734	0.366
EC	0.728	0.917	
Temp	-0.730	0.602	-0.286

The second significant component, PC₂, demonstrated that EC was loaded as strong (i.e. > 0.75) with positive values and was followed by pH with a moderate loading value. In PC₂, EC is considered as the major SWQP responsible for both spatial and seasonal surface water quality variations in the river due to the presence of inorganic dissolved solids coming from irrigation purposes as well as fertilizers and pesticides.

The third dominant component, PC₃, explained that two SWQPs (i.e. COD and BOD) were correlated with each other. COD and BOD were loaded as strong (i.e. > 0.75) with positive values. In PC₃, COD and BOD are the dominant SWQPs contributing to surface water quality variations in the SJR and it can be explained as the industrial effluents from paper and food processing industries along the SJR shoreline, especially in the middle basin of the river.

The results demonstrated that PCA/FA is found to be a cost-effective technique, which can be very useful in surface water quality studies due to its capability of extracting the major pollutants contributing to water quality variations at any water body.

5.3.2.2 Cluster Analysis (CA) Technique

Hierarchical agglomerative CA, which provides multiple levels of clustering, was employed to extract groups of similar water monitoring stations. As a result, it generated a dendrogram, grouping the sixty-six water sampling points into four distinct clusters, by using the Ward's method and squared Euclidean distance as a measure of similarity, as shown in **Figure 5.4**.

Cluster 1 included 28 water sampling stations (i.e. stations 39, 40, 41, ..., to 66). These samples were acquired from the middle basin of the SJR (i.e. above the Mactaquac

Dam) during the fourth trip (July 2016) and fifth trip (August 2016). These sampling stations have higher COD and BOD levels, compared to other sampling points acquired from the lower basin of the SJR, due to the presence of food and paper processing industries along the shoreline of this area.

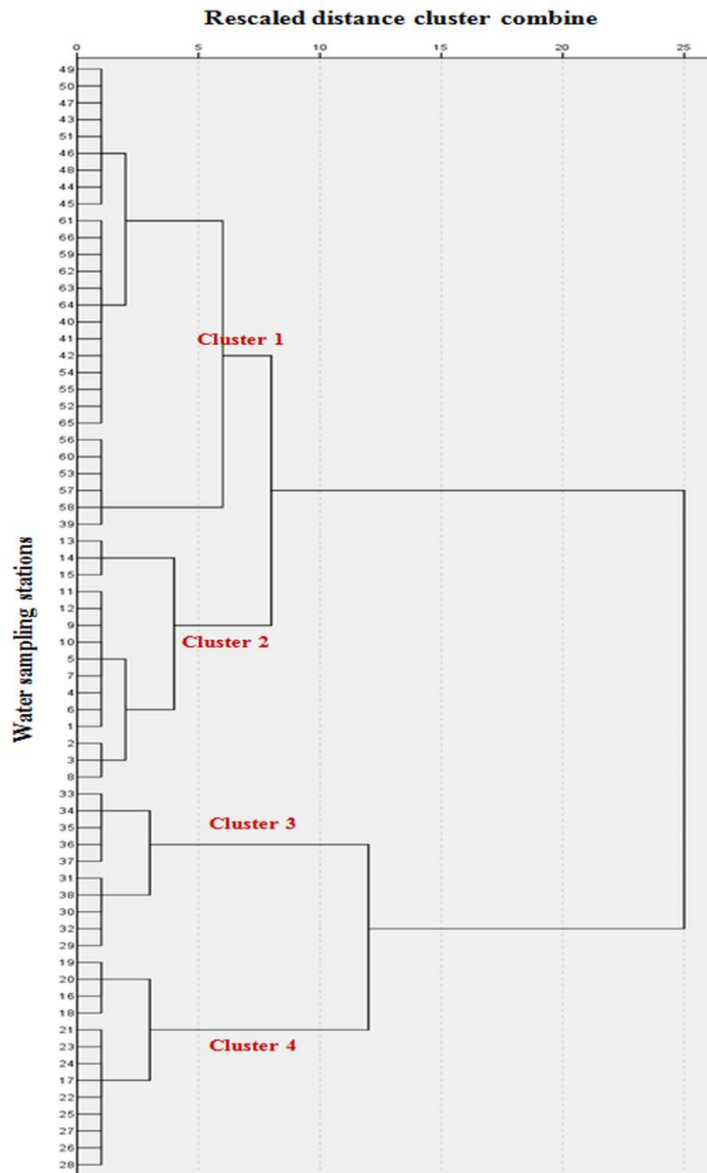


Figure 5.4 Dendrogram showing hierarchical agglomerative CA of sampling stations

On the other hand, cluster 2 included 15 sampling stations (i.e. stations 1, 2, 3, ..., to 15). These samples were acquired from the lower basin of the SJR (i.e. below the Mactaquac Dam) during the first trip (June 2015). The lower basin of the river has less industrial and agricultural activity, which may keep this area of the river less polluted compared to other parts of the SJR.

Cluster 3 included 10 sampling stations (i.e. stations 29, 30, 31, ..., to 38). These samples were acquired in May 2016 (i.e. spring season). These stations receive pollution mostly due to rain fall and snow melt associated with spring. The variation level of Turb and TSS in the SJR is approximately similar during different seasons and the average level of both Turb and TSS is higher in spring as compared to other seasons.

Cluster 4 included 13 sampling points (i.e. stations 16, 17, 18, ..., to 28). These samples were collected in April 2016. Similar to cluster 3, the natural processes, such as rain fall and snow melt, are found to be responsible for increasing the effect of soil erosion, which may raise the levels of both Turb and TSS in the SJR and its tributaries.

These findings were consistent with the results obtained by the New Brunswick Department of Natural Resources (Arseneault, 2008). Moreover, the results indicated that hierarchical agglomerative CA technique is very helpful in surface water quality research studies because of its ability to clearly classify water sampling stations into discrete groups based on their surface water quality characteristics, and consequently to reduce the respective cost in the future monitoring plans.

5.3.2.3 Discriminant Analysis (DA) Technique

5.3.2.3.1 Spatial DA

Spatial variation in surface water quality was further evaluated using DA with groups (clusters) identified by hierarchical agglomerative CA. In this context, the four groups were used as the dependent variables, while all the measured SWQPs (i.e. Turb, TSS, TS, TDS, COD, BOD, DO, pH, EC, and Temp) represented the independent variables. In our study, both standard and stepwise modes of DA were applied, and three discriminant functions were generated. As a result, the identified groups were clearly separated by using the first two discriminant functions (function 1 and function 2), as shown in **Figure 5.5**.

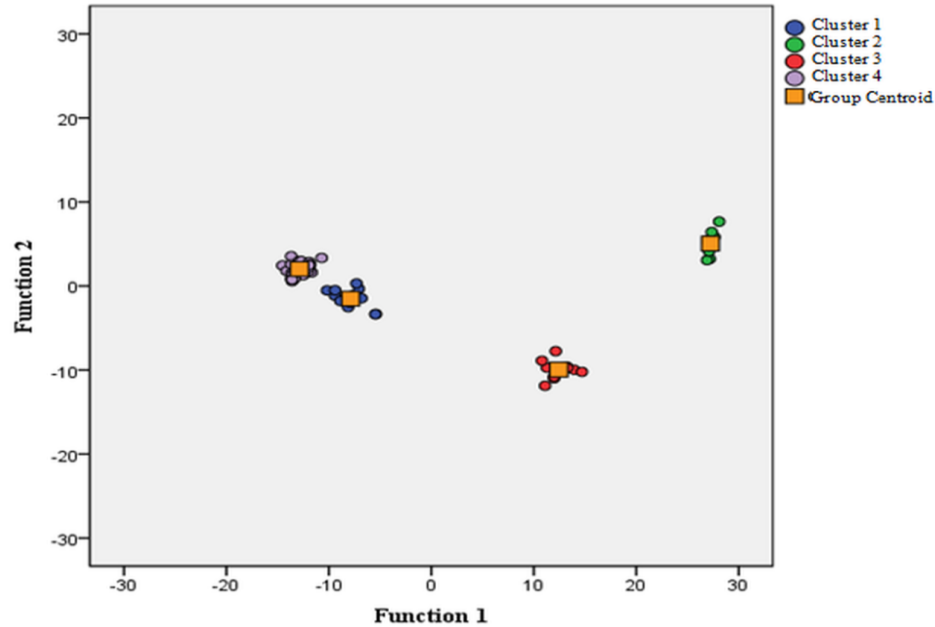


Figure 5.5 Scatter plot for DA of spatial water quality variation across the four groups

As shown in **Table 5.5**, for both standard and stepwise modes, the obtained values of Wilks' lambda and the chi-square for each discriminant function varied from 0.001 to 0.198 and from 97.168 to 618.159, respectively, with p -value < 0.001 , indicating that the spatial DA was reliable and efficient.

Table 5.5 Wilks' lambda and chi-square test for discriminant analysis (DA) of spatial variation in surface water quality across four clusters (groups) of sites.

Mode	Statistical measures	Discriminant function		
		1	2	3
Standard mode	Eigenvalue	270.349	23.739	4.784
	% of variance	90.500	7.900	1.600
	Wilks' Lambda	0.001	0.007	0.173
	Chi-square	618.159	290.360	102.669
	p-value	< 0.001	< 0.001	< 0.001
Stepwise mode	Eigenvalue	246.974	18.877	4.050
	% of variance	91.500	7.000	1.500
	Wilks' Lambda	0.001	0.010	0.198
	Chi-square	607.343	276.544	97.168
	p-value	< 0.001	< 0.001	< 0.001

Table 5.6 Structure matrix along with variable scores for DA of Table 5.5.

SWQPs	Standard mode			Stepwise mode		
	Function 1	Function 2	Function 3	Function 1	Function 2	Function 3
Turb	0.173	0.412	0.033	-0.179	0.387	0.089
TSS	0.090	0.467	-0.118	-0.142	0.413	-0.069
TS	0.086	0.291	0.146	-0.087	0.272	0.212
TDS	0.081	0.285	0.162	-0.077	0.248	0.226
COD	0.332	-0.197	0.281	0.435	-0.227	0.272
BOD	0.374	-0.108	0.185	0.481	-0.044	0.269
DO	0.257	-0.125	0.129	-0.157	-0.133	0.091
pH	-0.013	0.058	0.288	-0.016	0.049	0.290
EC	-0.075	0.101	0.522	0.081	0.091	0.470
Temp	-0.179	0.241	0.113	-0.213	0.296	0.078

Additionally, in standard mode of DA, the obtained three discriminant functions explained 90.50%, 7.90%, and 1.60% of the variance between the groups, respectively. Similarly, in stepwise mode of DA, the explained variance is 91.50%, 7.00%, and 1.50% for the three discriminant functions, respectively.

In **Table 5.6**, for both standard and stepwise modes of DA, SWQPs with variable scores > 0.30 should be identified as the most significant discriminating variables among all the measured SWQPs (Tahir, Quazi, & Gopal, 2011).

Table 5.7 Discriminant function coefficients for DA of Table 5.5.

SWQPs	Standard mode			Stepwise mode		
	Function 1	Function 2	Function 3	Function 1	Function 2	Function 3
Turb	1.055	1.459	0.508	-0.784	1.034	0.495
TSS	-0.140	-0.556	-0.030			
TS	0.008	0.056	-0.023			
TDS	0.017	0.114	-0.104	0.001	0.034	-0.027
COD	-0.002	-0.084	0.091	0.031	-0.111	0.035
BOD	-1.269	-1.044	-1.800			
DO	0.451	0.019	0.438	1.852	0.412	0.005
pH	1.919	-0.320	2.565	-2.604	-0.167	3.593
EC	-0.013	0.018	0.053	0.019	0.032	0.050
Temp	1.657	0.602	0.235			
Constant	5.531	-16.464	-31.620	-8.659	-14.160	-32.548

The first discriminant function categorized both COD and BOD as the major variables, while Turb and TSS variables were classified as the best predictors among all the measured SWQPs in the second discriminant function. Finally, EC was found to be the most important variable in the third discriminant function. These results were in agreement with those obtained in previous subsections for both PCA/FA and CA.

Table 5.8 Classification matrix for DA of Table 5.5.

Mode	Monitoring groups	% correct	Regions assigned by DA			
			Group 1	Group 2	Group 3	Group 4
Standard mode	Group 1	100	15	0	0	0
	Group 2	100	0	13	0	0
	Group 3	100	0	0	10	0
	Group 4	100	0	0	0	28
	Total	100	15	13	10	28
Stepwise mode	Group 1	100	15	0	0	0
	Group 2	100	0	13	0	0
	Group 3	100	0	0	10	0
	Group 4	100	0	0	0	28
	Total	100	15	13	10	28

The discriminant function score (z score) for each discriminant function can be calculated using discriminant function coefficients, provided in **Table 5.7**, for the measured SWQPs. Finally, in both the standard and stepwise modes, the produced classification matrix reached 100% accuracy in the regions assigned by DA, as shown in **Table 5.8**.

5.3.2.3.2 Temporal DA

Temporal variation in surface water quality was also assessed using DA and our surface water quality dataset was subdivided into seasonal groups (i.e. early spring (April 2016), late spring (May 2016), early summer (June 2015 and July 2016), and late summer (August 2016)).

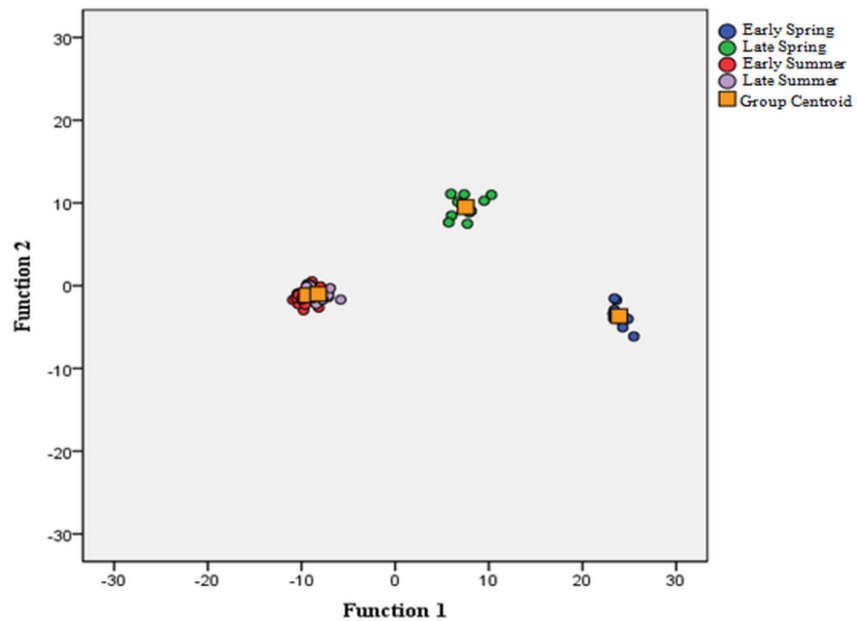


Figure 5.6 Scatter plot for DA of temporal water quality variation across the four seasons

In our study, the four seasonal groups represented the dependent variables, while all the measured SWQPs (i.e. Turb, TSS, TS, TDS, COD, BOD, DO, pH, EC, and Temp) were used as the independent variables. Both standard and stepwise modes of DA were employed, and three discriminant functions were developed. Accordingly, the four seasonal groups were separated by using the first two discriminant functions, as shown in **Figure 5.6**. Furthermore, it is clearly noticeable that both spring and summer seasons

were entirely separated; however, both early and late spring periods were partially separated.

In **Table 5.9**, for both standard and stepwise modes, the values of Wilks' lambda and the chi-square for each discriminant function varied from 0.001 to 0.199 and from 97.839 to 586.553, respectively, with p-value < 0.001, indicating that the temporal DA was valuable. Furthermore, in standard mode of DA, the explained variance is 88.70%, 8.70%, and 2.60% for the three discriminant functions, respectively. Similarly, in stepwise mode of DA, the obtained three discriminant functions explained 86.20%, 9.90%, and 3.90% of the variance between the four seasonal groups, respectively. These results indicated that the first two discriminant functions were sufficient to explain the differences in surface water quality among the four seasonal groups.

Table 5.9 Wilks' lambda and chi-square test for DA of temporal variation in surface water quality across four seasons.

Mode	Statistical measures	Discriminant function		
		1	2	3
Standard mode	Eigenvalue	185.205	18.229	5.317
	% of variance	88.700	8.700	2.600
	Wilks' Lambda	0.001	0.008	0.158
	Chi-square	586.553	280.782	107.832
	p-value	< 0.001	< 0.001	< 0.001
Stepwise mode	Eigenvalue	184.206	18.727	5.820
	% of variance	86.200	9.900	3.900
	Wilks' Lambda	0.001	0.017	0.199
	Chi-square	576.560	270.789	97.839
	p-value	< 0.001	< 0.001	< 0.001

In **Table 5.10**, for both standard and stepwise modes of DA, seven SWQPs, with variable score > 0.30, were identified as the most significant discriminating variables among all the measured SWQPs. The first discriminant function categorized COD, BOD, and DO variables as the best predictors, while both EC and pH were identified as the most significant variables among all the SWQPs in the second discriminant function. Finally, both Turb and TSS are the most important SWQPs in the third discriminant function.

Table 5.10 Structure matrix along with variable scores for DA of Table 5.9.

SWQPs	Standard mode			Stepwise mode		
	Function 1	Function 2	Function 3	Function 1	Function 2	Function 3
Turb	0.221	-0.196	0.437	0.219	0.193	0.560
TSS	0.112	-0.165	0.468	0.111	0.164	0.575
TS	0.112	-0.281	-0.168	0.110	0.284	0.179
TDS	-0.106	-0.274	-0.178	0.125	0.225	-0.162
COD	0.438	0.298	-0.229	0.303	-0.205	0.243
BOD	0.337	-0.147	-0.132	0.325	-0.118	0.199
DO	-0.371	0.293	0.270	-0.357	-0.131	-0.229
pH	-0.003	-0.411	-0.114	-0.011	0.341	0.144
EC	-0.058	-0.541	-0.287	-0.055	0.301	0.272
Temp	-0.269	-0.272	-0.137	-0.259	0.222	0.105

Table 5.11 showed the discriminant function score for each discriminant function and it was calculated using discriminant function coefficients. Finally, as shown in **Table**

5.12, in both the standard and stepwise modes, the generated classification matrix reached 100% accuracy in the regions assigned by DA.

Table 5.11 Discriminant function coefficients for DA of Table 5.9.

SWQPs	Standard mode			Stepwise mode		
	Function 1	Function 2	Function 3	Function 1	Function 2	Function 3
Turb	1.702	-0.926	-0.668	-0.955	1.011	0.595
TSS	-0.379	0.441	0.364	0.005	0.044	-0.016
TS	0.014	-0.059	0.008			
TDS	0.111	-0.023	0.097			
COD	0.018	0.118	-0.042	0.021	-0.121	0.029
BOD	-2.262	-0.081	0.316	1.762	0.522	0.014
DO	0.199	0.169	2.172			
pH	2.546	1.619	0.468	-2.454	-0.237	3.783
EC	0.033	0.013	-0.081	0.022	0.039	0.049
Temp	-1.183	-0.447	0.775			
Constant	-10.645	-1.520	-27.113	-7.558	-15.199	-28.625

The results indicated that DA technique is very helpful in surface water quality research studies because it is able to test the significance of the obtained discriminant functions and to determine the most significant variables associated with differences among both spatial and temporal groups.

Table 5.12 Classification matrix for DA of Table 5.9.

Mode	Monitoring groups	% correct	Regions assigned by DA			
			Early Spring	Late Spring	Early Summer	Late Summer
Standard mode	Early Spring	100	13	0	0	0
	Late Spring	100	0	10	0	0
	Early Summer	100	0	0	28	0
	Late Summer	100	0	0	0	15
	Total	100	13	10	28	15
Stepwise mode	Early Spring	100	13	0	0	0
	Late Spring	100	0	10	0	0
	Early Summer	100	0	0	28	0
	Late Summer	100	0	0	0	15
	Total	100	13	10	28	15

5.4 Conclusion

Due to the overload of both natural and anthropogenic processes, evaluating surface water quality represents a great challenge to researchers. Water body treatment should be directed to SWQPs responsible for spatial/temporal water quality variations. As a result, effective savings and appropriate utilization of resources could be easily achieved. Therefore, in our study, multivariate statistical techniques, such as PCA/FA, hierarchical agglomerative CA, and DA, were used to (1) classify the most significant SWQPs that negatively influence surface water quality in the selected study area of the SJR, (2) minimize the complexity of a water quality dataset to a great extent, and (3) evaluate both spatial and seasonal variations in surface water quality of the SJR.

The main results of our study demonstrated that Turb, TSS, COD, BOD, and EC are the major SWQPs contributing to water quality variations in the river by using PCA/FA technique. Moreover, hierarchical agglomerative CA grouped 66 water sampling stations into four groups (clusters) based on similar water quality characteristics, which means a noticeable reduction in the water quality dataset was achieved. Additionally, DA technique was used to recognize the differences in surface water quality between both the four groups identified by hierarchical agglomerative CA and the four seasonal groups (early spring, late spring, early summer, and late summer).

The future work is to carry out further sampling trips on the SJR, especially in the upper basin of the river, to provide a whole picture of surface water quality in the river. Finally, this study is valuable for local administrators who have to make right decisions to protect surface water quality in their water bodies by using a cost-effective method.

Acknowledgements

This research is supported in part by the Egyptian Ministry of Higher Education, Bureau of Cultural & Educational Affairs of Egypt in Canada, and the Canada Research Chair Program. The authors wish to acknowledge Dr. Katy Haralampides and Dr. Dennis Connor for their help in water sampling and laboratory analysis.

REFERENCES

Akbar, A. T., Hassan, K. Q., & Achari, G. (2011). A Methodology for Clustering Lakes in Alberta on the basis of Water Quality Parameters. *Clean-Soil, Air, Water*, 39 (10), pp. 916-924.

- APHA. (2005). *Standards Methods for the Examination of Water and Wastewater* (21th ed.). American Public Health Association Washington DC, USA.
- Arseneault, D. (2008). *"The Road to Canada - Nomination Document for the St. John River, New Brunswick"*. The St. John River with the support of the New Brunswick Department of Natural Resources.
- Carpenter, S. R., Caraco, N. F., Correll, D. L., Howarth, R. W., Sharpley, A. N., & Smith, V. H. (1998). Non-point pollution of surface waters with phosphorus and nitrogen. *Ecological Applications*, 83, pp.559–568.
- Dong, J., Zhang, Y., Zhang, S., Wang, Y., Yang, Z., & Wu, M. (2010). Identification of Temporal and Spatial Variations of Water Quality in Sanya Bay, China By Three-Way Principal Component Analysis. *Environ. Earth Sci.*, 60, pp.1673-1682.
- Elhatip, H., Hinis, M. A., & Gulgahar, N. (2007). Evaluation of the water quality at Tahtali dam watershed in Izmir, Turkey by means of statistical methodology. *Stochastic Environmental Research and Risk Assessment*, 22, pp. 391-400.
- Google Maps*. (2016). Retrieved from Google Maps: <https://www.google.ca/maps/>
- Haag, I., & Westrich, B. (2002). Processes Governing River Water Quality Identified By Principal Component Analysis. *Hydrol. Process*, 16, pp.3113-3130.
- Huang, F., Wang, X., Lou, L., Zhou, Z., & Wu, J. (2010). Spatial Variation and Source Apportionment of Water Pollution in Qiantang River (China) using Statistical. *Water Res.*, 44, pp.1562-1572.
- Li, Y., Xu, L., & Li, S. (2009). Water quality analysis of the Songhua River Basin using multivariate techniques. *Journal of Water Resource and Protection*, 1(2), pp.110–121.
- Liu, C. W., Lin, K. H., & Kuo, Y. M. (2003). Application of factor analysis in the assessment of groundwater quality in a Blackfoot disease area in Taiwan. *Sci. Total Environ.*, 313(1-3), pp.77–89.
- Mahapatra, S., & Mitra, S. (2012). Managing Land and Water under Changing Climatic Conditions in India: A Critical Perspective. *Journal of Environmental Protection*, 3(9), pp.1054-1062.

- McKenna, J., & J.E. (2003). An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis. *Environmental Modelling & Software*, 18(3), pp.205-220.
- Mishra, A. (2010). Assessment of water quality using principal component analysis: A case study of River Ganges. *Journal of Water Chemistry and Technology*, 32(4), pp.227-234.
- Natural resources*. (2016). Retrieved from Statistics Canada: <http://www.statcan.gc.ca/>
- Ouyang, Y., Nkedi-Kizza, P., Wu, Q. T., Shinde, D., & Huang, C. H. (2006). Assessment of seasonal variations in surface water quality. *Water Research*, 40, pp.3800–3810.
- Qadir, A., Malik, R. N., & Husain, S. Z. (2007). Spatio-temporal variations in water quality of Nullah Aik-tributary of the river Chenab, Pakistan. *Environmental Monitoring and Assessment*, 140(1–3), pp.43–59.
- Reghunath, R., Murthy, T. R., & Raghavan, B. R. (2002). The utility of multivariate statistical techniques in hydrogeochemical studies: An example from Karnataka, India. *Water Research*, 36, pp.2437–2442.
- Salah, E. A., Turki, A. M., & Al-Othman, E. M. (2011). Assessment of water quality of Euphrates River using cluster analysis. *Journal of Environmental Protection*, 3, pp.1629-1633.
- Sharaf El Din, E., Zhang, Y., & Suliman, A. (2017a). Mapping concentrations of surface water quality parameters using a novel remote sensing and artificial intelligence framework. *International Journal of Remote Sensing*, 38 (4), pp. 1023-1042.
- Sharaf El Din, E., & Zhang, Y. (2017d). Estimation of both optical and non-optical surface water quality parameters using Landsat 8 OLI imagery and statistical techniques. *Journal of Applied Remote Sensing*, 11 (4), 046008 (2017), doi: 10.1117/1.JRS.11.046008.
- Sharaf El Din, E., & Zhang, Y. (2017e). Improving the accuracy of extracting surface water quality levels (SWQLs) using remote sensing and artificial neural network: a case study in the Saint John River, Canada. *International Archives of the*

Photogrammetry, Remote Sensing, and Spatial Information Sciences, XLII-4/W4, pp. 245-249, <https://doi.org/10.5194/isprs-archives-XLII-4-W4-245-2017>.

- Sharaf El Din, E., & Zhang, Y. (2018). Application of multivariate statistical techniques in the assessment of surface water quality in the Saint John River, Canada. *UNB Annual Graduate Research Conference (GRC)*. Fredericton, Canada.
- Shrestha, S., & Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environmental Modelling & Software*, 22, pp.464–475.
- Simeonov, V., Stratis, J. A., Samara, C., Zachariadis, G., Voutsas, D., & Anthemidis, A. (2003). Assessment of the surface water quality in Northern Greece. *Water Research*, 37, pp.4119–4124.
- Singh, K. P., Malik, A., Mohan, D., & Sinha, S. (2004). Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India): a case study. *Water Research*, 38, pp.3980-3992.
- Tahir, A. A., Quazi, K. H., & Gopal, A. (2011). A Methodology for Clustering Lakes in Alberta on the basis of Water Quality Parameters. *Clean – Soil, Air, Water*, 39(10), pp.916–924.
- Vega, M., Pardo, R., Barrado, E., & Deban, L. (1998). Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Research*, 32, pp. 3581–3592.
- Wunderlin, D. A., Diaz, M. P., Ame, M. V., Pesce, S. F., Hued, A. C., & Bistoni, M. A. (2001). Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia river basin (Cordoba, Argentina). *Water Research*, 35, pp.2881–2894.

Chapter 6: SUMMARY AND CONCLUSION

This chapter summarizes the research presented in this PhD dissertation. It begins with the summary of each chapter (Chapters 2 to 5). The achievements of this research are then presented. Finally, recommendations for future work are provided.

6.1 Summary of Research

In this dissertation, remote sensing Landsat 8 satellite data were exploited for assessing surface water quality in water bodies. Chapters 2-5 introduced progressively improved methods addressing the four identified challenges associated with the evaluation of surface water quality from satellite imagery. While satellite reflectance data and multiple regression techniques were incorporated effectively in Chapter 2 for estimating concentrations of both optical and non-optical SWQPs, artificial intelligence was successfully exploited in Chapter 3 for mapping the relationship between satellite multi-spectral data and concentrations of SWQPs. Chapter 4 extended the concept of Chapter 3 to improve the accuracy of surface water quality level (SWQL) extraction by integrating satellite data, artificial intelligence, and the water quality index (WQI). Finally, identifying the most significant SWQPs that contribute to spatio-temporal surface water quality variations by using multivariate statistical techniques was demonstrated in Chapter 5.

6.2 Achievements of the Research

A summary of the introduced technologies and achievements in each of the four main chapters is presented in the following subsections.

6.2.1 Developing the Landsat 8-based-SWR Technique for Estimating Concentrations of Optical and Non-optical SWQPs

Chapter 2 introduced a solution for the problems and limitations associated with quantifying the concentrations of SWQPs from satellite imagery. In this context, remote sensing estimation of non-optical SWQPs, such as COD, BOD, DO, pH, and EC, has not yet been performed because these parameters are less likely to affect the reflected radiation measured by satellite sensors. The solution introduced in this chapter is to develop a stepwise regression (SWR) technique to estimate the concentrations of both optical and non-optical SWQPs from the Landsat 8 satellite imagery, which is freely available and has the potential to support surface water quality studies.

The developed Landsat 8-based-SWR technique was generated in three major phases: (1) deriving surface reflectance data (i.e., water leaving reflectance) from Landsat 8 satellite imagery by eliminating radiometric and atmospheric distortions, (2) deriving the actual concentrations of all the measured SWQPs based on the standard methods for lab examination of water and wastewater of the American Public Health Association (APHA), and (3) developing Landsat 8-based-SWR models to estimate the concentrations of the selected SWQPs with accurate results.

To the best of our knowledge, the Landsat 8-based-SWR technique is developed for the first time to estimate the concentrations of three non-optical SWQPs, namely COD, BOD, and DO, which have not been estimated before with Landsat data or any other optical instrument. Compared to previous studies, significant correlation between Landsat 8 surface reflectance data and concentrations of SWQPs was achieved and R^2 values reached high level of accuracy ($R^2 > 0.85$) for turbidity, TSS, COD, BOD, and

DO. These findings are very helpful for local administrators who have to make decisions and enact strict measures in order to protect water quality in potable water resources.

6.2.2 Developing the Landsat 8-based-BPNN Framework for Mapping Concentrations of SWQPs

Chapter 3 replaces regression-based methods by learning-based methods to map the complex relationship between satellite multi-spectral data and concentrations of SWQPs. The problem is that surface water quality is complex to have a simple relationship with satellite multi-spectral signatures and consequently it is challenging for regression-based techniques to model such a complex relationship. Therefore, this chapter introduced the developed Landsat 8-based-backpropagation neural network (BPNN) framework for mapping the concentrations of SWQPs from space.

The novel Landsat 8-based-BPNN framework was generated in three major phases: (1) deriving water leaving reflectance values from Landsat 8 satellite imagery by eliminating radiometric and atmospheric distortions, (2) measuring the actual concentrations of the selected SWQPs based on the standard methods for lab examination of water and wastewater of the APHA, and (3) developing Landsat 8-based-BPNN models for mapping concentrations of SWQPs from Landsat 8 satellite data and consequently providing a spatial distribution map for each optical and non-optical SWQP over each pixel of the selected study area.

To the best of our knowledge, our Landsat 8-based-BPNN framework is the first to map concentrations of different SWQPs, especially the non-optical parameters, with highly accurate results, compared to regression-based or even other learning-based

methods. Compared to previous methods, significant R^2 between Landsat 8 surface reflectance and concentrations of SWQPs were obtained by using the developed framework. The obtained $R^2 \geq 0.93$ for turbidity, TSS, COD, BOD, and DO. These findings demonstrated the feasibility of using the developed framework to generate highly accurate models to map concentrations of SWQPs, and to generate spatio-temporal maps of SWQPs from Landsat 8 imagery.

6.2.3 Developing the Landsat 8-based-CCMEWQI Technique for Extracting the Accurate Levels of SWQPs

Chapter 4 represents an extension to the last introduced solution for mapping concentrations of surface water quality parameters. The problem/challenge of simplifying the expression of surface water quality and improving the accuracy of delineating the accurate SWQLs was addressed in this chapter. The solution introduced in this chapter exploited the concept in Chapter 3 and developed a novel technique that combines remote sensing multi-spectral data, the BPNN algorithm, and the water quality index, introduced by the Canadian Council of Ministers of the Environment (CCMEWQI), to extract accurate surface water quality levels to be accessible to decision-makers.

The developed technique was generated in three major phases: (1) developing an accurate method for mapping the concentrations of SWQPs over each pixel of the selected study area by using the BPNN algorithm, (2) evaluating the performance and stability of the developed method using ground truth data provided by the Province of New Brunswick, Canada, and (3) utilizing all of the obtained concentrations of the selected SWQPs as an input to the CCMEWQI to extract the accurate SWQLs.

To the best of our knowledge, the novel Landsat 8-based-CCMEWQI cost-effective technique is developed for the first time to extract the levels of surface water quality with highly accurate results. This study showed that the CCMEWQI was classified as Fair in the lower basin of the SJR, which means the water quality is usually protected but occasionally threatened or impaired. Moreover, the water quality in the middle basin of the SJR was observed as Marginal, which means the water quality is frequently threatened or impaired. The result findings were found compatible with the results obtained by the New Brunswick Department of Natural Resources; however, they used a huge number of water samples, which is costly and labour intensive.

6.2.4 Categorizing Spatio-temporal Surface Water Quality Variations Using Multivariate Statistical Techniques

Chapter 5 provided a solution for the problems and limitations associated with classifying the major SWQPs that negatively affect water bodies. The problem is that existing methods are mainly focused on understanding the relationship between different SWQPs; however, very few studies have attempted to detect spatio-temporal aspects of surface water quality in water bodies. Moreover, due to the complexity of the relationship between SWQPs, it is not easy to draw a clear conclusion directly from surface water quality data. Therefore, the solution introduced in this chapter is to use the multivariate statistical techniques, such as PCA/FA, CA, and DA, to identify the major SWQPs that contribute to spatio-temporal variations of surface water quality and to help in the interpretation of complex surface water quality data to better understand the surface water quality of water bodies.

The developed technique was generated in three major phases: (1) classifying the major SWQPs contributing to surface water quality variations by using PCA/FA, (2) developing multiple levels of clustering for detecting the relationship between the collected water samples by using hierarchical agglomerative CA, and (3) evaluating both spatial and seasonal surface water quality variations of the study area by using DA.

To the best of our knowledge, PCA/FA, CA, and DA were combined for the first time to categorize the most significant pollution sources contributing to surface water quality variations in the Saint John River (SJR) with inexpensive implementation cost. This study illustrated that turbidity, TSS, COD, BOD, and EC are the major SWQPs contributing to both spatial and temporal variations in the water quality of the SJR. Moreover, the result findings showed a reduction in the dimensionality of surface water quality data by classifying water sampling stations based on similarities of water quality characteristics. Our study demonstrated the significant use of multivariate statistical techniques for surface water quality assessment, which can lead to effective savings and proper utilization of water resources.

6.3 Recommendations for Future Work

Based on the results and contributions discussed in the previous sections, the suggested recommendations for future research are given below:

- The findings of this research are based on field measurements (i.e., water samples) collected during two years (2015 and 2016). Long-term monitoring is very helpful in providing information on surface water quality aspects and trends. Therefore, an intensive monitoring program which would consider more SWQPs

that can reach water bodies, such as total nitrogen and total phosphorus, is recommended.

- This research attempted to retrieve the concentrations of both optical and non-optical SWQPs from Landsat 8 multi-spectral data with highly accurate results. However, using new atmospheric correction methods for obtaining the water leaving reflectance data is recommended to improve the accuracy of SWQP retrieval from Landsat 8 satellite imagery.
- Our study could be extended to provide information about the extinction of different species of fish. In this context, this study could be coupled with eutrophication processes which could result from the overload of the nutrients in the re-suspended sediments. The relationship between suspended sediments and dissolved oxygen demand (DO) is inversely related causing either hypoxia (low DO) or anoxia (No DO) which can lead to the death of different fish species.
- Further research is needed to investigate the optical properties (i.e., absorbance and scattering coefficients) of optical SWQPs, such as turbidity, TSS, chlorophyll-a, and organic constituents to better understand the relationship between surface water quality and the multi-spectral data from remote sensing imagery. This would help to further improve the accuracy of the generated remote sensing models of the optical SWQPs without being dependent on sampling time or even sampling location.

Appendix I

Permission from the “*Journal of Applied Remote Sensing (JARS)*” for the **paper 1**
provided in Chapter 2:



Nicole Harris <nicoleh@spie.org>

Today, 6:56 PM

Essam Helmy Mahfouz Sharaf El Din ↕



Reply all | v

Dear Essam,

Thank you for seeking permission from SPIE to reprint material from our publications. As author, SPIE shares the copyright with you, so you retain the right to reproduce your paper in part or in whole.

Publisher's permission is hereby granted under the following conditions:

(1) the material to be used has appeared in our publication without credit or acknowledgment to another source; and

(2) you credit the original SPIE publication. Include the authors' names, title of paper, volume title, SPIE volume number, and year of publication in your credit statement.

Sincerely,


Nicole Harris
Administrative Editor, SPIE Publications
[1000 20th St.](#)
[Bellingham, WA 98225](#)
+1 360 685 5586 (office)
nicoleh@spie.org

SPIE is the international society for optics and photonics. <http://SPIE.org>




SPIE.

Appendix II


Permission from the “*International Journal of Remote Sensing (IJRS)*” for the **paper 2 provided in Chapter 3:**





Kelly, Andrew <andrew.kelly@tandf.co.uk>


  Reply all | 

Wed 1/17/2018 6:45 AM

To: IJRS-Administrator <IJRS-Administrator@Dundee.ac.uk>;  Essam Helmy Mahfouz Sharaf El Din

Cc: Timothy Warner <tim.warner@mail.wvu.edu> 

 | Action Items



Dear Essam,

Thank you for getting in touch. On publishing your article in *IJRS*, you assigned copyright in your article to Taylor & Francis and retained several rights, which include the following:

The right to include the article in a thesis or dissertation that is not to be published commercially, provided that acknowledgement to prior publication in the Journal is given.

As such, it would be absolutely fine for you to reproduce the paper in your thesis (I assume that it won't be published commercially?). Our standard acknowledgement line is given below. Please include this with the article in your thesis:

Acknowledgement: This <chapter or book> is derived in part from an article published in <Journal> <date of publication> © Informa UK Limited, trading as Taylor & Francis Group, available online: <DOI>

I hope that answers your query and congratulations on obtaining your PhD.

Best wishes,
Andrew

Appendix III

Permission from the “*International Archives of the Photogrammetry, Remote Sensing, and Spatial Information Sciences*” for the **paper 3 provided in Chapter 4:**



fkarimipour <fkarimipour@ut.ac.ir>

Tue 1/16/2018 12:56 PM

To:  Essam Helmy Mahfouz Sharaf El Din ↗



 Reply all | 

Dear Essam

It is to confirm that you can use your paper entitled "IMPROVING THE ACCURACY OF EXTRACTING SURFACE WATER QUALITY LEVELS (SWQLs) USING REMOTE SENSING AND ARTIFICIAL NEURAL NETWORK: A CASE STUDY IN THE SAINT JOHN RIVER, CANADA", which was presented in the Tehran's ISPRS International Conference 2017 and published in the ISPRS Archive, as a part of your PhD dissertation.

Kind regards

Farid Karimipour
Tehran's ISPRS Conference Chair

...

Appendix IV

Proof of submission to “*Remote Sensing of Environment*” for the **paper 3 provided in Chapter 4:**

RSE Submission Confirmation



Remote Sensing of Environment <eesserver@eesmail.elsevier.com>



↻ Reply all | ▾

Today, 10:09 AM

Essam Helmy Mahfouz Sharaf El Din ✉

Dear Essam,

Your submission entitled "Delineating the accurate patterns of surface water quality by integrating Landsat 8 OLI imagery, artificial intelligence, and the water quality index" as Research Paper has been received by Remote Sensing of Environment Journal Office.

You will be able to check on the progress of your paper by logging on to EES as an author. The URL is <https://ees.elsevier.com/rse/>.

Your manuscript will be given a reference number once an Editor has been assigned.

Thank you for submitting your work to Remote Sensing of Environment.

Remote Sensing of Environment

Appendix V

Proof of submission to “*Journal of the American Water Resources Association*” for the **paper 4 provided in Chapter 5:**



Journal of the American Water Resources Association <ciawra.org@manuscriptcentral.com>

  Reply all | v

Mon 04/16/2018, 3:32 PM

Essam Helmy Mahfouz Sharaf El Din v

Dear Mr. Essam Sharaf El Din

Your paper entitled "Assessment of spatio-temporal surface water quality variations using multivariate statistical techniques: a case study of the Saint John River, Canada" has been successfully submitted online and is presently being given full consideration for publication in Journal of the American Water Resources Association.

Your manuscript ID is AWRA-8714-18.

Please mention the above manuscript ID in all future correspondence or when calling the office for questions. If there are any changes in your postal or e-mail address, please log in to ScholarOne Manuscripts at <https://mc.manuscriptcentral.com/> and edit your user information as appropriate.

You can also view the status of your paper at any time by checking your Author Centre after logging in to <https://mc.manuscriptcentral.com/>.

Thank you for submitting your paper to [Journal of the American Water Resources Association](#).

Yours faithfully,

Journal of the American Water Resources Association Editorial Office

Curriculum Vitae

Candidate's full name: Essam Helmy Mahfouz Sharaf El Din

Universities attended (with dates and degrees obtained):

2012: M.Sc., Civil Engineering (Public Works Engineering), Tanta University, Egypt

2008: B.Sc., Civil Engineering, Tanta University, Egypt

Publications:

Peer Reviewed Journal Papers:

- 1) Sharaf El Din, E., Zhang, Y., & Suliman, A. (2017). Mapping concentrations of surface water quality parameters using a novel remote sensing and artificial intelligence framework. *International Journal of Remote Sensing*, 38 (4), pp. 1023-1042. <http://dx.doi.org/10.1080/01431161.2016.1275056>
- 2) Sharaf El Din, E., & Zhang, Y. (2017). Improving the accuracy of extracting surface water quality levels (SWQLs) using remote sensing and artificial neural network: a case study in the Saint John River, Canada. *International Archives of the Photogrammetry, Remote Sensing, and Spatial Information Sciences*, XLII-4/W4, pp. 245-249, <https://doi.org/10.5194/isprs-archives-XLII-4-W4-245-2017>.
- 3) Sharaf El Din, E., & Zhang, Y. (2017). Estimation of both optical and non-optical surface water quality parameters using Landsat 8 OLI imagery and statistical techniques. *Journal of Applied Remote Sensing*, 11 (4), 046008 (2017), doi: 10.1117/1.JRS.11.046008.
- 4) Sharaf El Din, E., & Zhang, Y. (2018). Assessment of spatio-temporal surface water quality variations using multivariate statistical techniques: a case study of the Saint John River, Canada. *Journal of the American Water Resources Association*, under review.
- 5) Sharaf El Din, E., & Zhang, Y. (2018). Delineating the accurate patterns of surface water quality by integrating Landsat 8 OLI imagery, artificial intelligence, and the water quality index. *Remote Sensing of Environment*, under review.

Peer Reviewed Conference Papers:

- 1) Sharaf El Din, E., & Zhang, Y. (2017). Statistical estimation of the Saint John River surface water quality using Landsat 8 multi-spectral data. *ASPRS Annual Conference 2017. Proceedings of Imaging & Geospatial Technology Forum (IGTF). March 12-17, Baltimore, US.*
- 2) Sharaf El Din, E., & Zhang, Y. (2017). Neural network modelling of the Saint John River sediments and dissolved oxygen content from Landsat OLI imagery. *ASPRS Annual Conference 2017. Proceedings of Imaging & Geospatial Technology Forum (IGTF). March 12-17, Baltimore, US.*
- 3) Sharaf El Din, E., & Zhang, Y. (2017). Using remote sensing and artificial intelligence to improve the accuracy of surface water quality level extraction: a case study in the Saint John River, Canada. *ISPRS International Joint Conference 2017. Commission IV, ISPRS WG IV/3. October (07-10), Tehran, Iran.*

Conference Abstracts:

- 1) Sharaf El Din, E., & Zhang, Y. (2018). Application of multivariate statistical techniques in the assessment of surface water quality in the Saint John River, Canada. *UNB Annual Graduate Research Conference (GRC). March 23, Fredericton, Canada.*