DATA MINING, GIS AND REMOTE SENSING: APPLICATION IN WETLAND HYDROLOGICAL INVESTIGATION

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

By
Bo Zhang, M.S, M.A.S

****

The Ohio State University
2008

Dissertation Committee: Approved by
Professor Franklin W. Schwartz, Adviser
Professor Carolyn J. Merry
Professor C. K. Shum
Assistant Professor Ningchuan Xiao

__________________________________
Adviser

Environmental Science Graduate Program
ABSTRACT

The Prairie Pothole Region in the United States contains millions of seasonal, semi-permanent, or permanent lakes and wetlands that typically range in size from 0.1 to 10 hectares. These lakes and wetlands are vulnerable to climate change, especially in our study area of South Dakota, in which a period of deluge following a sharp drought considerably expanded the areal extent of prairie pothole lakes during the last decade of the twentieth century. Preliminary estimates of lake areas, determined using Landsat 5 and 7 images, had appreciable errors especially for the smallest of these lakes. We developed a new sub-pixel approach integrated with a Classification and Regression Tree (CART) model using a Geographical Information System (GIS) to quantify mixed water pixels along lake boundaries to improve the area estimations for pothole lakes. Errors in estimated area were typically 10 percent or less for lakes greater than 1 hectare in size.

An analysis of lakes in our study area using GIS and remote sensing technologies demonstrates how total areas and numbers of lakes and wetlands in different sizes changed with the transition from drought to deluge. Small lakes exhibited a distinct seasonal variation in contrast to large lakes that tended to follow longer trends more broadly. The total areas and numbers of small lakes and wetlands are mostly related with the 6-month evaportranspiration (ET) variation, while the variables of large lakes are highly correlated with the mean Palmer Drought Severity Index (PDSI) of a 48-month
time period. We also examine the response of a complex lake/wetland system to variations in climate. The focus is on lakes and wetlands within the Prairie Coteau Region, which is part of the larger Prairie Pothole region of the Central Plains of North America. Information on lake size was enumerated from satellite images and aerial photos and yielded power-law relationships for different hydrological conditions. Of particular interest was a recent drought and deluge sequence, 1988-1992 and 1993-1998. Results showed that the pothole lakes followed well-defined power laws that changed annually and interannually as a function of climate. The power laws for spring seasons in years 1987, 1990, 1992, 1997, and 2002 yielded a relatively constant slope. However, slopes changed with time within each year. The lines produced from Landsat images and aerial photos indicated scale independence for lakes with a size from 100 m$^2$ to more than 40,000 m$^2$. This fractal tendency and aerial photos taken in 1939/7/29 provides an approach to reconstructing the distribution of pothole lakes back to 1939, the end of the “Dustbowl” drought. The study shows that smaller lakes are profoundly affected seasonally by the strength of the spring snow melt and evapotranspiration. Larger lakes are influenced more slowly by longer term periods of drought and deluge.

Using the TOPEX radar altimeter for land cover studies has been of great interest due to the TOPEX near global coverage and its consistent availability of waveform data for about one and a half decades from 1992 to 2005. However, the complexity of the TOPEX Sensor Data Records (SDRs) makes the Classification of land cover using particularly difficult. In this study, regression tree and artificial neural networks as the most powerful algorithms in data mining are investigated for water proportion assessment over Lake of
the Woods area using TOPEX SDR waveform data. Results demonstrate that these data
mining technologies have provided insight into identifying water proportion from the
TOPEX radar waveforms, with predicted errors controlled in a reasonable range. The
distinct tailing pattern of radar echoes from water plays an important role in water ratio
regression.
Dedicated to my parents
I want to express my gratitude to Dr. Franklin W. Schwartz for guiding all my research projects during my study at Ohio State, as well as his financial aids within the past 5 years. Dr. Schwartz, who accepted me as one of his Ph.D. students almost six years ago and gave me lots of opportunities in finding my own research interests. There is no doubt that I would not have finished this dissertation without his guidance and financial support. I have to give part of the credits to Dr. Schwartz for the uniqueness and innovation in my dissertation. I also would also like to thank Dr. Shum and Dr. Ibaraki from School of Earth Science, Dr. Merry from Civil Engineering department, and Dr. Xiao from Geography department serving as my committee members. They all provide me with their expert knowledge and passions on the Ph.D research. I would also like to thank my family, who always have faith in me and give me constant support and encouragement throughout the time of my study. I will dedicate this dissertation to them in memory of what they have done for me.
VITA

4 September 1977.................................Born – Xinjiang, China

2000 ..................................................B.S., Environmental Engineering, Tsinghua University

2002 ..................................................M.S., Institute of Nuclear Energy, Tsinghua University

2006 ..................................................M.A.S., Statistics, The Ohio State University

2002 - 2007.................................Graduate Research Assistant, School of Earth Sciences, The Ohio State University

2007 - 2008.................................Data Mining Specialist, The National Drought Mitigation Center

PUBLICATIONS


FIELDS OF STUDY

Major Field: Environmental Science
TABLE OF CONTENTS

Abstract .......................................................................................................................... ii
Dedication ......................................................................................................................... v
Acknowledgments .......................................................................................................... vi
Vita .................................................................................................................................... vii
List of Tables .................................................................................................................. x
List of Figures ................................................................................................................ xi

Chapters:

1. Introduction ................................................................................................................. 1

2. A CART Based Sub-pixel Method to Map Spatial and Temporal Patterns of Prairie Pothole Lakes with Climatic Variability......................................................... 5
   2.1. Introduction ............................................................................................................ 5
   2.2. Methodology ......................................................................................................... 7
       2.2.1. Description of the Study Area ........................................................................ 7
       2.2.2. Image Processing and Classification ............................................................... 9
       2.2.3. Water Fraction Classification by CART ......................................................... 11
       2.2.4. Validation of CART model ............................................................................. 13
   2.3. Results and Discussion ......................................................................................... 20
       2.3.1. Assessment of Unsupervised Classification .................................................. 20
       2.3.2. Accuracy of area estimation ........................................................................... 21
   2.4. Conclusion ............................................................................................................. 25

3. Spatial and Temporal Variations of Prairie Pothole Lakes in Connection with Climate Variability ......................................................................................................... 26
   3.1. Introduction............................................................................................................. 26
   3.2. Methodology ......................................................................................................... 28
       3.2.1. Description of the study area ........................................................................... 28
       3.2.2. Landsat image processing and water proportion regression ......................... 31
       3.2.3. Meteorological Data and Drought Indices ...................................................... 32
   3.3. Results and Discussion ......................................................................................... 33
       3.3.1. Temporal Changes of Lakes and Wetlands ..................................................... 34
       3.3.2. Drought to deluge transition .......................................................................... 38
       3.3.3. Connections to climatic variability ................................................................. 40
   3.4. Conclusions ............................................................................................................ 52
4. Systematics in the Size Structure of Prairie Pothole Lakes through Drought And Deluge ................................................................. 54

4.1. Introduction ....................................................................................... 54
4.2. Methodology .................................................................................. 56
   4.2.1. Description of the study area ................................................. 56
   4.2.2. GIS and remote sensing approaches to evaluate pothole lakes .... 60
   4.2.3. Hierarchical Bayesian linear model ......................................... 64
4.3. Results ............................................................................................ 66
   4.3.1. Power-law lines of potholes in different time ..................... 66
   4.3.2. Parameters of power-law lines from Bayesian analysis ......... 70
   4.3.3. Power-law in small scale and reconstruction of dust bowl scenario... 77
4.4. Discussions .................................................................................. 83
4.5. Conclusions ................................................................................ 88

5. Data Mining In Topex SDR Data: A Case Study In Monitoring Inland Water Zones ........................................................................... 90

5.1 Introduction .................................................................................... 90
5.2 Methodology .................................................................................. 93
   5.2.1 TOPEX data description ....................................................... 93
   5.2.2 Data mining algorithms ........................................................ 99
      5.2.2.1 Regression Tree ....................................................... 100
      5.2.2.2 Neural Networks ..................................................... 102
5.3 Results ............................................................................................ 103
   5.3.1 Regression Tree ...................................................................... 103
   5.3.2 Neural Networks .................................................................. 110
   5.3.3 Comparison of ensemble models of regression tree and neural networks .............................................................. 115
5.4 Discussions ................................................................................ 115
5.5 Conclusions ................................................................................ 122

6. Conclusions ..................................................................................... 124

7. Bibliography ................................................................................. 129
<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Relative Importance Values of CART Model</td>
<td>14</td>
</tr>
<tr>
<td>2.2 Results of Splitting Training and Validating Dataset by CART</td>
<td>16</td>
</tr>
<tr>
<td>2.3 Accuracy of unsupervised classification by ISODATA algorithm</td>
<td>21</td>
</tr>
<tr>
<td>3.1 List of Landsat images</td>
<td>31</td>
</tr>
<tr>
<td>3.2 Expanded water areas in 1997 from 1990 within different areas</td>
<td>39</td>
</tr>
<tr>
<td>5.1 Results of regression trees</td>
<td>109</td>
</tr>
<tr>
<td>5.2 selected variables of regression trees</td>
<td>110</td>
</tr>
<tr>
<td>5.3 Statistics with a neural networks based on a three-fold cross-validation process</td>
<td>114</td>
</tr>
<tr>
<td>5.4 Comparison between the ensemble trees and the ensemble neural networks</td>
<td>118</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Study area of the Prairie Coteau Region of South Dakota</td>
</tr>
<tr>
<td>2.2</td>
<td>Tree structure of CART model</td>
</tr>
<tr>
<td>2.3</td>
<td>Rules of sub-pixel classifier for classifying mixed water pixels</td>
</tr>
<tr>
<td>2.4</td>
<td>Sample image classified by water fraction values</td>
</tr>
<tr>
<td>2.5</td>
<td>Relative errors of lake area estimation</td>
</tr>
<tr>
<td>3.1</td>
<td>Level Fluctuation and precipitation plots during the study period</td>
</tr>
<tr>
<td>3.2</td>
<td>Plot of total area of lakes in different sizes vs. date</td>
</tr>
<tr>
<td>3.3</td>
<td>The number of lakes plotted vs. times by different lake sizes (small lakes, large lakes and overall lakes)</td>
</tr>
<tr>
<td>3.4</td>
<td>Comparison of lake occurrences between 1990 (dry year) and 1997 (wet year)</td>
</tr>
<tr>
<td>3.5</td>
<td>Correlations between areas and numbers of lakes and wetlands and PCP, ET, and PDSIs in certain time periods</td>
</tr>
<tr>
<td>3.6</td>
<td>Linear regressions between total areas and numbers of lakes and meteorological variables</td>
</tr>
<tr>
<td>4.1</td>
<td>Monthly precipitation data from the Andover station in Day County, South Dakota from 1986 – 2003</td>
</tr>
<tr>
<td>4.2</td>
<td>Pothole lakes presented on a Landsat image and the DOQQ at different resolutions</td>
</tr>
<tr>
<td>4.3</td>
<td>Power-law lines of pothole lakes from 1987 – 2002. (a) Inter-annual patterns (b)</td>
</tr>
</tbody>
</table>
Seasonal patterns ................................................................. 69

4.4 Posterior distributions generated by MCMC chains for parameters of the power-law line on 4/23/1987. ‘σ’ is the square root of inverse value of ‘τ’, which refers to the square root of the variance ......................................................... 71

4.5 Box-plots of distributions of slopes and intercepts of annual power-law lines simulated by Bayesian linear models.................................................. 74

4.6 Box-plots of distributions of slope and intercept of seasonal power-law lines simulated by Bayesian linear models ........................................... 76

4.7 Comparison of lakes from aerial photos taken in 1939 (a) and 2003 (b)......... 79

4.8 (a) Power-law lines from Landsat and DOQQ (b) comparisons of power-law lines among year 1939, 1990, 1992 and 1997......................................................... 82

4.9 Conceptual model of power-law relationships for pothole-lake under different hydrologic conditions ................................................................. 86

5.1 Samples of TOPEX waveforms over land surface with different water ratios within the footprint. Panel a is 100%, Panel b is 75%, Panel c is 25% and Panel d is 5%..... 95

5.2. Illustrative TOPEX waveforms show how the return signal behavior changes in transitioning from water to land along the ground track. On the image, water is masked ................................................................. 96

5.3. Map of the sample area, Lake of the Woods, and ground track samples of the Topex waveform data. ................................................................. 98

5.4. Error functions of regression trees from 3-fold cross-validation ............ 107
5.5. Structures of regression trees from 3-fold cross-validation .......................... 108

5.6. Error functions of neural networks from 3-fold cross-validation .................. 113

5.7. Residual plots vs. predicted water ratio of the ensemble trees (a) and the ensemble neural networks (b) .......................................................... 117

5.8. Predicted errors from the ensemble tree model over the study site .............. 120

5.9. TOPEX footprint over Devils lake (left panel) and its radar waveform (right panel) ........................................................................................................ 122
CHAPTER 1

INTRODUCTION

This study explores new applications of data mining, geographical information system (GIS), remote sensing technologies and radar altimeter data in the observation and analysis of lake/wetland systems over large areas. Data from the Landsat TM Sensor, aerial photography and the TOPEX radar altimeter are used for hydrologic investigations on lakes and wetlands of the Prairie Pothole Region (PPR) of North America. This region is of particular concern in the study because the PPR contains millions of lakes and wetlands of all sizes, and supports much of the waterfowl in North America as breeding and brood-rearing habitat (Batt et al., 1989). However, significant variability in climate creates seasonal and interannual cycles ranging between extreme droughts and deluges that considerably impact lake and wetland sizes. Moreover, this region is faced with anthropocentric pressure from agricultural activities including damages from sediments, and pollution from fertilizers and pesticides (Martin and Hartman, 1987). Understanding the hydrology of the PPR region is essential water resources for maintaining the ecological functions and protecting economic benefits. Unfortunately, due to the large scale, geological complexity, and lack of hydrological surveys, most of the lakes especially those small ones were barely monitored. Therefore, remote sensing
technologies play an important role to provide continuous spatial and temporal measurements of wetland areas, and are widely applied in these hydrological studies.

In this study, data from the Landsat imagery sensor and aerial photos are utilized to develop and evaluate new approaches to monitoring lakes and wetlands in this large area. The water proportion of land surface is quantified by the analysis of Landsat imagery and Digital Orthophoto Quarter-Quadrangle (DOQQ) to estimate the water abundance on the ground surface. The results can also be applied to training and validating data mining algorithms to classify surface water information from waveform data of the TOPEX radar altimeter. Combined with the climate data, the study eventually explores the climatic function to the cycles of transition between drought and deluge and construct a conceptual model was constructed to predict the long-term variation of wetlands and lakes of the PPR under the climate control. Climate indices and meteorological data for this transitional period are correlated to regional changes of lakes and wetlands to discover the function of climatic variability. The correlation between the climatic data and the variation of lakes and wetlands of the PPR are assessed to develop predictive models for future changes.

Although imagery data (e.g., Landsat/ASTER) are widely applied in evaluating changes in lakes and wetlands area (Ozesmi and Bauer 2002; Schmugge et al., 2002), cloud coverage can prohibit large-scale space/time syntheses. Radar altimetry with relatively short measurement cycles and reduced atmospheric influence is a method that can provide well-constrained time windows to assess these changes. Hence, technologies
using radar waveforms are explored to produce a practical approach to predict the content of water abundance on the ground surface. But the complexity of returned radar waveforms containing information on the distribution of ground features such as water bodies present a challenge in its applications to predict land-surface characteristics of the study sites. Extremely difficult theoretical problems, which are caused by the complexity of the scattering and attenuation of radar energy, may make this simulation approach intractable. To solve this, various data mining technologies including statistical regression methods, tree classifier, and neural networks, are carried out to evaluate water conditions on the land surface to correlate characteristics of the waveform data.

In this study, the first objective is to construct an accurate approach to evaluate the changes of prairie potholes by climate variability using the Landsat satellite imagery data, which is presented in Chapter 2. In Chapter 2, a Classification and Regression Tree (CART) based sub-pixel method that can accurately estimate the areas of wetlands even for very small ones is developed and assessed. The second objective is to discover how climate variability controls the changes of these potholes in number and size, which is presented in Chapter 3. Chapter 3 shows the relationships between these changes and climatic indices and meteorological data. Chapter 4 discusses unique characteristics of the size distribution of wetlands in fractal. Based on some findings of consistence across multiple years under extremely different hydrological scenarios, wetlands are evaluated over the whole region for the “Dustbowl” period according to very limited aerial photos in 1939. The third objective is to construct a practical approach using the TOPEX radar altimeter to monitor wetlands and lakes. Chapter 5 shows data mining algorithms such as
regression tree and neural networks provide great insight into assessing water proportion of ground tracks of the TOPEX, which shows a promise to broaden the application of TOPEX from measuring the elevation to imaging land covers. In summary, the ultimate goal is to combine data mining, GIS and remote sensing technologies to evaluate the impact of climate on changes of prairie potholes. This study presents a better and more detailed understanding about how these wetlands response to climate change in contrast to traditional hydrological surveys.
2.1 Introduction

The Prairie Pothole Region (PPR) of North America has millions of lakes and wetlands. These surface-water bodies exhibit tremendous hydrological variability on annual and interannual time scales because of marked variability in climate. Understanding the regional behavior of these lakes and wetlands is important because their hydrologic and ecologic functions are significant at regional and even continental scales. Area-wise, the PPR constitutes only 10% of the total waterfowl breeding area in North America, but produces more than 50% of all breeding ducks and more than 60% of continental mallards (Guntenspergen, et al., 2000; Batt, et al., 1989). Moreover, wetlands in the PPR serve to reduce the severity of floods (McAllister et al., 2000), to attenuate contaminant loading and transport from fertilizer and pesticide, and to sequester carbon.

Our broad interest with the lakes is in understanding their response to episodes of drought and deluge. The severe drought from 1988-92 in South Dakota was the second worst of 20th century, and led to significant hydrologic impacts. This drought was followed by the
most significant wet period of the century (Winter and Rosenberry, 1998). Lake and wetland flooding occurred to an extent not seen perhaps since the turn of 1900.

However, the large extent of the PPR and its geological complexity limit the possibility for detailed hydrological surveys. Fortunately GIS and remote sensing approaches may be useful for studies of wetland hydrology since the 1970s (Lunetta and Balogh, 1999; Ozesmi and Bauer1, 2002; Pietroniro and Prowse, 2002; Sawaya et al., 2003; Schmugge et al., 2002; Stewart et al., 1980). Satellite sensors significantly improve the observations of hydrological conditions, by providing long term data records over large areas. Hence remotely sensed data have been broadly applied in hydrological studies of the PPR (e.g., Todhunter and Rundquist, 2004; Work and Gilmer, 1976).

Satellite-based approaches are useful in the study of prairie pothole lakes to estimate spatial changes as a function of climatic variability. Not surprisingly, the archive of Landsat data from 1985 to 2002 is particularly useful. However, the moderate spatial resolution (30 m) impacts the assessment of lakes and wetlands, especially when small pothole lakes are represented by just a few pixels. The problems of resolution are of particular concern in the PPR where more than 80% of the wetlands are smaller than 0.8 ha (Johnson and Higgins, 1997). Those small ponds are manifested by a few mixed water/land pixels and a few pure water pixels. It is apparent that traditional classification methods using moderate Landsat data underestimated the areal extent of water bodies because mixed water/land pixels along boundaries are not properly included in the classification. For our applications then, approaches are required that are capable of
interpreting the composition of different classes within a single pixel.

The goal of this chapter is to describe a new sub-pixel method to estimate the water area of small water bodies integrated with GIS, represented by perhaps several Landsat pixels. We illustrate this approach in an assessment of the hydrologic behavior of lakes and wetlands on the Prairie Coteau of South Dakota.

2.2. Methodology

2.2.1 Description of the study area

The study area is located on the northern tip of the Prairie Coteau area of northeastern South Dakota. During the Wisconsin deglaciation, stagnation of the ice on this upland area produced an extremely hummocky topography and ultimately thousands of close-basin lakes and potholes. The map of the study area (Figure 2.1) shows the large recreational lakes also found there. The most important of these large lakes are those of the Waubay Lakes chain. Available stage hydrographs for several of these large lakes show the dramatic change in water levels through the drought-deluge transition. No similar data exist for the smaller pothole lakes. Most of the lakes and wetlands occur in closed basins essentially disconnected from river systems. With the exception of interlake transfers, the Waubay Lakes system is also closed hydrologically.

This region has a subhumid to subarid continental climate with short hot summers, long cold winters, low levels of precipitation, and high evaporation. The mean annual
precipitation ranges from 50 to 56 cm (20-22 in) and temperature ranges from 1°F in January to 85°F in July.

Figure 2.1. Study area of the Prairie Coteau Region of South Dakota.

2.2.2 Image processing and classification

A Landsat image scene, which was acquired on 7/24/2003 together with a DOQQ (Digital Orthophoto Quarter-Quadrangle) created on 7/04/2003, were used as data samples or learning datasets to construct the Classification and Regression Tree (CART) scheme. The image was calibrated and processed as the first step in an overall assessment. Image processing to covert digital value to absolute reflectance of each pixel was accomplished
to diminish the variability in solar irradiance (Chander and Markham, 2003; Landsat Project Science Office website). Furthermore, general atmospheric conditions are assumed for each image based on the meteorological conditions on the dates when the images were obtained.

In our study, the first procedure is classification to identify water pixels according to their spectral response in remotely sensed images, and then, summing all connected water pixels provide area estimation for a lake. A variety of classification approaches exists such as unsupervised classification and supervised classification (Jensen, 1996). These methods work well for small homogenous areas mainly with uniform ground characteristics. Several approaches have been implemented in delineating water boundaries on Landsat images, for example, density slicing of TM band 5 and supervised maximum likelihood classification (Frazier and Page, 2000; Manavalan et al., 1993). These popular hard classifiers, however, produce poor estimates for small ponds (Frazier and Page, 2000), because they overlooked partial water pixels mixed with other land features along the boundaries of small ponds.

In recent years, methods such as sub-pixel classification and fuzzy classification have improved the classification accuracy in extracting certain ground features. Typically, such approaches are based on complicated models and machine learning technologies, including linear or nonlinear spectral mixture models, neural networks and decision trees. For example, neural network and decision trees have been applied to forest, vegetation, and urban area classification with somewhat improved accuracies as compared to
traditional methods (e.g., Atkinson and Tate, 2000; Brown de Colstoun et al., 2003; DeFries and Chan, 2000; Pal and Mather, 2003; Tatem, et al., 2002). In particular, decision tree approaches have the capability of increasing classification accuracy significantly without invoking a prerequisite Gaussian assumption by using the maximum likelihood method. Decision tree approaches are also easier to interpret, as compared to black-box model approaches like neural networks. Reported comparison among different classification methods, including decision tree, maximum likelihood and fuzzy c-means clustering indicates that maximum likelihood produces the highest accuracy for pure pixels, while decision tree methods provide the greatest accuracy for mixed pixels (Xu et al., 2005).

In our study, two classification steps were implemented for pure water pixels and mixed water pixels, respectively. First, each calibrated image was classified using an unsupervised ISODATA (Iterative Self-Organizing Data Analysis Technique) clustering algorithm (Tou and Gonzalez, 1974) to provide seven classes – pure water, wetlands, agricultural land, range land, bare land, forest and build-up land according to existing investigations of land cover and the scheme recommended by the USGS. The ISODATA algorithm was chosen because of its simplicity and robustness as well as its ease of incorporation into other software packages. Water pixels classified were treated as pure water pixels and were preserved from other pixels by a masking process. Then mixed water pixels were identified as pixels adjacent to pure water pixels, assuming that lake boundaries are located within one pixel of pure water pixels, based on the common shapes of pothole lakes within the study area. Finally, a Classification and Regression
Tree (CART) based method to quantify the extent to which water is represented in mixed pixels along the boundary of each lake is adopted. The method we developed here incorporated both pure water pixels and mixed water pixels in an area estimation called the sub-pixel method, in contrast to the unsupervised method counting only pure water pixels.

2.2.3 Water fraction classification by CART

To construct the CART scheme, a total of 488 pixels were randomly selected (299 pixels for data training and 189 pixels for validating) from the sample Landsat image using the ArcGIS software. The DOQQ within a 20-day time window was used as the ground truth reference. In effect, for the lake-boundary pixels, the training data link the Landsat spectral data for each pixel to an exact estimate of how much water is present in each pixel, as indicated from the DOQQ. The validating data are represented as a set of new pixels used to verify and refine the CART model by comparison with independent estimates from the DOQQ.

One of the considerations in using a tree-based approach is to determine how many nodes should be included in the tree. A training data set was classified to determine the relative error associated with different numbers of nodes. After splitting the training dataset in various trials, a tree was selected with the smallest relative error (the ratio of the variance from the fitted tree model to the original sample variance) of 0.45 and only eight terminal nodes (Figure 2.2). The eight terminal nodes indicate that eight classes of water fraction
values assigned to the training dataset are appropriate. In addition, the small size of the tree also avoids potential overfitting, which in larger trees is usually minimized by a pruning process (Pal and Mather, 2003).

Figure 2.2 shows the structure of the CART model for the mixed water pixel classification. In Figure 2.2, the 299 mixed pixels in the training dataset are first separated by their reflectance values in band 4. Pixels with a reflectance less than or equal to 0.2378 are assigned to node 2 and others go to node 3. Then each node is continuously split until the terminal nodes are reached. The terminal nodes provide the final classes for each of the pixels.

* Avg - the average water fraction of all the pixels within this node; N - the total number of pixels assigned to this node; RE_B4 - the reflectance of band 4.

Figure 2.2 Tree structure of CART model.
2.2.4 Validation of CART model

The splitting procedures for constructing the CART model also include a feature selection process, which provides measures of relative importance (Table 2.1). Relative importance is calculated as $\chi^2$ statistics and quantifies the weights of the attributes (bands of Landsat images) of the data in the classification. Table 2.1 shows that Bands 4, 5, 7, 1 are sequentially the most important in classifying mixed water pixels with relative importance values 100.0, 72.3, 59.5, and 48.1, respectively. This result further indicates that bands 4, 5 and 1 are sufficient to classify the training data set. Because band 7 is highly correlated with bands 4 and 5, band 7 was not included in the model. Thus, the CART model involves bands 4, 5 and 1. The relative importance values are in the same order as an earlier study (Xu, et al., 2005) showing that band 4 is the most important in classifying mixed pixels. It also follows then that band 4 is most useful to delineate water boundaries in land cover classification (Jensen, 1996).

<table>
<thead>
<tr>
<th>Band</th>
<th>Relative Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band 4</td>
<td>100.000</td>
</tr>
<tr>
<td>Band 5</td>
<td>72.308</td>
</tr>
<tr>
<td>Band 7</td>
<td>59.478</td>
</tr>
<tr>
<td>Band 1</td>
<td>48.149</td>
</tr>
<tr>
<td>Band 2</td>
<td>14.602</td>
</tr>
<tr>
<td>Band 3</td>
<td>3.112</td>
</tr>
</tbody>
</table>

Table 2.1 Relative Importance Values of CART Model

The performance of this CART model is evaluated by applying the tree to the validating
dataset. Pixels in the validating dataset are independently sampled from the training dataset. The goodness of fit of the model to the testing dataset validates the CART model in the classification and implies no severe overfitting problems needed to be considered. Table 2.2 presents the classification results for both training and validating datasets. For each terminal node in the table, the mean water fraction values between the training data and the validating data are very close, with the largest difference of 0.138 in terminal node 5. The standard errors are relatively small and stable, ranging from 0.141 to 0.178. The standard deviations of both the training and validating datasets are reasonably constant. Table 2.2 shows that there is a good fit with the validating dataset and this validates the tree model in our mixed water pixel classification. However, the errors evidently grow as the mean water fraction values in the terminal nodes decrease. In other words, smaller quantities of water within a pixel are associated with greater uncertainties in the estimate. Therefore, pixels with water fraction values less than 0.5 are not considered in the following analysis to avoid additional errors.
<table>
<thead>
<tr>
<th>Terminal Node</th>
<th>Dataset</th>
<th>Count</th>
<th>Mean</th>
<th>StdDev</th>
<th>StdError</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Training</td>
<td>40</td>
<td>0.927</td>
<td>0.074</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Validating</td>
<td>24</td>
<td>0.861</td>
<td>0.135</td>
<td>0.150</td>
</tr>
<tr>
<td>2</td>
<td>Training</td>
<td>64</td>
<td>0.821</td>
<td>0.135</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Validating</td>
<td>56</td>
<td>0.832</td>
<td>0.140</td>
<td>0.141</td>
</tr>
<tr>
<td>3</td>
<td>Training</td>
<td>9</td>
<td>0.607</td>
<td>0.117</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Validating</td>
<td>13</td>
<td>0.555</td>
<td>0.162</td>
<td>0.170</td>
</tr>
<tr>
<td>4</td>
<td>Training</td>
<td>44</td>
<td>0.764</td>
<td>0.140</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Validating</td>
<td>19</td>
<td>0.711</td>
<td>0.154</td>
<td>0.163</td>
</tr>
<tr>
<td>5</td>
<td>Training</td>
<td>6</td>
<td>0.582</td>
<td>0.109</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Validating</td>
<td>3</td>
<td>0.444</td>
<td>0.113</td>
<td>0.178</td>
</tr>
<tr>
<td>6</td>
<td>Training</td>
<td>39</td>
<td>0.626</td>
<td>0.170</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Validating</td>
<td>17</td>
<td>0.536</td>
<td>0.142</td>
<td>0.168</td>
</tr>
<tr>
<td>7</td>
<td>Training</td>
<td>91</td>
<td>0.520</td>
<td>0.146</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Validating</td>
<td>57</td>
<td>0.434</td>
<td>0.127</td>
<td>0.154</td>
</tr>
<tr>
<td>8</td>
<td>Training</td>
<td>6</td>
<td>0.689</td>
<td>0.218</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Validating</td>
<td>0</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

Note: * no data.

Table 2.2 Results of Splitting Training and Validating Dataset by CART Model

Notice also that no validating data were assigned to terminal node 8. Actually, terminal node 8 is abnormal compared with terminal node 7. Terminal node 8 has a higher average value of water content and a greater reflectance in band 4 than those in terminal node 7. This result contradicts the general principle that the lower reflectance in band 4 implies the higher water fraction. This unusual result is caused by specific errors in training dataset sampling.

The standard error associated with terminal node 2 has the smallest value of 0.141 with the second largest number of pixels of 56 in the validating data set. Terminal node 5 has
the largest standard error of 0.178 with the smallest number of pixels of 3. This result implies that large sample size of pixels could reduce the standard error and increasing the sample size would improve the classification model.

According to these analyses, we improved the CART model (shown in Figure 2.3) to refine the scheme and to correct potential errors and applied this model to all images. We redefined the water fraction of all mixed pixels into six classes (less than 0.5, 0.5, 0.6, 0.75, 0.82, and 0.90) that were related to the eight terminal nodes of the original CART model. The nodes of the previous CART model having similar mean values are regrouped into one class of water fraction. Moreover, pixels with a water fraction value smaller than 0.5 are ignored to reduce errors and uncertainties in the classification. To fix the issue of terminal node 8 in the previous CART model, terminal node 7 and terminal node 8 are merged as one class with the same attributes as their parent node the internal node 7. These modifications simplify the structure of the classifier and make it more reasonable and robust than the original CART model.
Figure 2.3 Rules of sub-pixel classifier for classifying mixed water pixels.

The rules of the modified CART model in Figure 2.3 can be illustrated by an example of water fraction class of 0.6: if a pixel has a reflectance value of band 4 larger than 0.2378, a reflectance value of band 5 less than or equal to 0.1208, and a reflectance value of band 1 less than or equal to 0.1436, this pixel will be assigned to the class with a water fraction value equals to 0.6.
Figure 2.4 shows a sample of images classified by water fraction values. It is clear that most of the areas of lakes and wetlands are covered, if both pure and mixed water pixels are counted in. Otherwise, severe biases would be introduced with only the pure water pixels.

2.3. Results and discussion

2.3.1 Assessment of unsupervised classification

Because a land-cover classification is integral to our approach, an assessment is needed
of the accuracy of the ISODATA algorithm in the unsupervised classification. Accuracy in the assignment of each class was determined by evaluating the classification of 71 randomly created control points. Table 2.3 shows that the simple unsupervised classification method is successful in identifying pure water pixels with an approximate 100% accuracy. Among all the other land-cover classes, forest has the lowest accuracy because forests are sparsely scattered within this region in relatively small patches, which are commonly mixed together within other classes, such as built-up land, range land and wetlands. Range land and agricultural land are both major classes that are also difficult to differentiate from each other in the Landsat images. Because the water class is of primary concern, the spotty performance in the other classifications is not a problem to our study.

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Reference Totals</th>
<th>Classified Totals</th>
<th>Correct Number</th>
<th>Producers Accuracy</th>
<th>Users Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agricultural land</td>
<td>10</td>
<td>22</td>
<td>9</td>
<td>90.0%</td>
<td>40.9%</td>
</tr>
<tr>
<td>Bare land</td>
<td>7</td>
<td>9</td>
<td>5</td>
<td>71.4%</td>
<td>55.6%</td>
</tr>
<tr>
<td>Build-up land</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Forest</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>33.3%</td>
<td>20.0%</td>
</tr>
<tr>
<td>Range land</td>
<td>31</td>
<td>17</td>
<td>14</td>
<td>45.2%</td>
<td>82.4%</td>
</tr>
<tr>
<td>Water</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Wetland</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>50.0%</td>
<td>75.0%</td>
</tr>
<tr>
<td>Total</td>
<td>71</td>
<td>71</td>
<td>46</td>
<td>64.8%</td>
<td>64.8%</td>
</tr>
</tbody>
</table>

Table 2.3 Accuracy of unsupervised classification by isodata algorithm

2.3.2 Accuracy of area estimation

This section presents an analysis of the improvement in lake-area estimates from Landsat.
images provided by the sub-pixel method outlined above. The basis for comparison is a conventional unsupervised classification (ISODATA) approach that is capable of identifying pure water pixels. This analysis involves 42 lakes of different sizes that are randomly selected. For each lake, a highly accurate measurement of area is determined from the high-resolution DOQQ, which provides the ground truth reference. For each of the test lakes selected from the study area in South Dakota, areas are estimated using (1) the Landsat image with only the unsupervised classification method, and (2) the classification scheme improved by the sub-pixel approach. Relative errors of area estimation in the two different Landsat-based methods are calculated as ratios by the following equation. For each lake, relative errors provide a measure of accuracy of the methods.

\[
\text{Relative error} = \left( \frac{\text{Area estimated by Landsat} - \text{Area estimated by DOQQ}}{\text{Area estimated by DOQQ}} \right)
\]

The results of this analysis are shown in Figure 2.5 (a), where the relative errors are plotted versus actual lake areas estimated by using DOQQ. As expected, errors of estimation increase with a decrease in the lake areas. However, area estimates based on the sub-pixel method have remarkably smaller errors as compared to the unsupervised classification. Indeed, the sub-pixel approach works well for quite small lakes. For example, even lakes about 0.5 ha (or about 5 Landsat TM 30-m pixels) can be estimated with errors of about 20% or less. The estimates of area for lakes larger than 1 ha (or about 11 Landsat TM 30-m pixels) are less than 10%.
Another advantage of the sub-pixel method is that the errors are generally unbiased with errors almost symmetrically distributed around 0, as shown by the histogram plots in Figure 2.5 (b).
Figure 2.5 (b). The relative errors from the sub-pixel method have an approximately normal distribution with a mean of -0.04 and a standard deviation of 0.12, while the distribution of the relative errors from the unsupervised method has a mean of -0.31 and a standard deviation of 0.23. This result indicates an unbiased estimation of lake areas from sub-pixel method as compared to the strongly skewed distribution from the unsupervised method. In addition, the standard deviation for the new method is only about half of that from the unsupervised method, indicating that the method is robust even when lake sizes are quite small.

The skewness of underestimation from the unsupervised method is produced by a mixture of water and sub-emergent plants or grass within the pixels along the boundaries of prairie lakes and wetlands. Because vegetation has a considerably high reflectance in Band 4 and Band 5 of Landsat images, even a small ratio of vegetation would create a high DN in these bands, which cause the mixed water pixels to be misclassified into other categories.

Some general errors of both methods remain, which can be attributed to one or several of the following possibilities. There exists a 20-day time difference between when the DOQQ and the Landsat image were collected. Small lakes, in particular, might exhibit some changes over this period. A second possible source of error might be related to errors in the sub-pixel classification because of the limited sample size of pixels and uncertainties in the sampling process. The third possibility is the registration of the DOQQ and Landsat images. A slight geometric distortion could also create small errors.
in the estimates of water fraction. On balance, the Landsat images can provide a much better area estimation if the sub-pixel approach is adopted.

2.4. Conclusion:

A sub-pixel based scheme for analyzing Landsat data showed good promise in substantially improving estimates of lake areas, especially for lakes as small as within 1 ha. This improvement in accuracy comes about by including mixed water pixels along lake boundaries.
3.1 Introduction:

Lakes and wetlands in the Prairie Pothole Region (PPR) of North America occur in glaciated depressions commonly ranging in size from 0.1 ha to 10 ha. These pothole wetlands represent only about 10% of the total waterfowl habitat of North America, but contribute more than 50% of the migratory duck population [Guntenspergen et al., 2000; Batt et al., 1989]. These wetlands are also important for the roles they play in flood control, contaminant attenuation, and carbon sequestration [McAllister et al., 2000; Lorah et al., 1997; Pant et al., 2003]. Typically, these basins are closed hydrologically and hence, are particularly vulnerable to climate variability [Covich et al., 1997; Johnson et al., 2005].

The climate across much of the PPR is extremely variable with cyclical swings from periods of extreme drought to deluge. A case in point is the Prairie Coteau Region (PCR) in northeastern South Dakota, which experienced cycles of extreme climate change at the end of the 20th century. The severe drought from 1988 to 1992 is considered the second
worst of the century. It was followed almost immediately by the most significant wet period of the century beginning in 1993 and extending through 1997 [Winter and Rosenberry, 1998]. Lakes and wetlands flooded to an extent not seen perhaps since the turn of the 19th century. These dramatic changes profoundly impacted the hydrology, ecology, biology, and geochemistry of the lakes and wetlands in the region [e.g., Covich et al., 1997; Conly and Van Der Kamp, 2001; Kruse et al., 2003; Winter and LaBaugh, 2003; Seabloom et al., 1998]. Fluctuation of water levels altered flow paths between surface and ground water [Johnson et al., 2004], changed salinity of water in some wetlands [Gorham et al., 1983; LaBaugh et al. 1996], modified vegetation patterns by reestablishing emergent species during dry periods accompanied with low water levels [Mulhouse and Galatowitsch, 2003; van der Valk, 2005], and affected food supply and nesting patterns of waterfowl [Swanson, 1988; Swanson and Duebbert, 1989; Covich et al., 1997].

There have been a variety of studies focusing on the hydrology of lakes and wetlands in the Northern Great Plains. The most important are long-term studies at the Cottonwood Lake area by the U.S. Geological Survey [e.g., Winter and Rosenberry, 1995]. Other studies have examined wetlands and lakes in a broader context [e.g. Winter and Rosenberry, 1998; Euliss and Mushet, 1996; Johnson et al., 2004]. This collection of studies has contributed a fundamental understanding of hydrologic and ecological functions of wetlands and pothole lakes in PPR. Where our study here differs from the previous work is a focus on the lake/wetland complex encompassing tens of thousands of lakes, considering temporal changes over large spatial scales. The variables of interest in
our study such as numbers and density of lakes and wetlands, can be directly related to the duck population [Pietz et al., 2000; Sorenson et al., 1998; Austin et al., 2001; Niemuth and Solberg, 2003].

The goal of this paper is to understand how climate variability influences the occurrence of surface water on the Prairie Coteau of South Dakota both temporally and spatially. More specifically, we use a Geographical Information System (GIS) together with Landsat imagery to analyze the spatial and temporal changes in water area and numbers of lakes of a given size over the latter part of the 20th century. Regression modeling linked the hydrologic characteristics of lakes and wetlands to the variability in climate as represented by meteorological variables and indices, such as average precipitation (PCP), average evapotranspiration (ET), and the Palmer Drought Severity Index (PDSI).

3.2 Methodology

3.2.1 Description of the study area

The study site is described in Chapter 2 and shown by Figure 2.1. The map shows the huge recreational lakes also found in the area. The large recreational lakes in the area belong to the Waubay Lakes chain. Stage hydrographs for several of these lakes show the dramatic change in water levels through the drought-deluge transition (Panel (a) of Figure 3.1). No similar data exist for the smaller pothole lakes. With the exception of interlake transfers, the Waubay Lakes system is also closed hydrologically.
This region has a subhumid to subarid continental climate with short hot summers, long cold winters, low levels of precipitation, and high evaporation. The mean annual precipitation ranges from 50 to 56 cm (20-22 in) and temperature ranges from 1°F in January to 85°F in July. Winter and annual precipitation data are shown in Panel (b) of Figure 3.1 for the Waubay Natl Wild Life station. The data summary clearly shows the swing from extreme drought conditions in 1988-92 to extremely wet conditions, in 1995-1998.
* Winter season include the November and December of the previous year and the January and February of the current year.

Figure 3.1. (a) Lake Level Fluctuation during the study period. (b) Annual and winter precipitation plots of the Waubay Natl Wild Life Station.
3.2.2 Landsat image processing and water proportion regression

There is sufficient Landsat TM and Landsat ETM+ coverage of the Prairie Coteau Region to examine the expansion of wetlands and lakes in this region following the drought of 1988-1992. Fourteen Landsat scenes were selected (Table 3.1) to provide areal information of wetlands and lakes from 1985 to 2003. Images were calibrated and processed as the first step in an overall assessment. Image processing was performed covert digital value to absolute reflectance of each pixel to minimize the variability in solar irradiance (Chander and Markham, 2003; Landsat Project Science Office website). Furthermore, the general atmospheric conditions are assumed for each image based on the meteorological conditions on the dates when images were obtained. The sub-pixel method described in Chapter 2 were applied to each image to provide estimates of total area and number of wetlands.

<table>
<thead>
<tr>
<th>Satellite Sensor</th>
<th>Image Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landsat-5 TM</td>
<td>5/19/1985</td>
</tr>
<tr>
<td>Landsat-5 TM</td>
<td>4/23/1987</td>
</tr>
<tr>
<td>Landsat-5 TM</td>
<td>4/15/1990</td>
</tr>
<tr>
<td>Landsat-5 TM</td>
<td>8/5/1990</td>
</tr>
<tr>
<td>Landsat-5 TM</td>
<td>5/6/1992</td>
</tr>
<tr>
<td>Landsat-5 TM</td>
<td>9/17/1994</td>
</tr>
<tr>
<td>Landsat-5 TM</td>
<td>8/19/1995</td>
</tr>
<tr>
<td>Landsat-7 ETM+</td>
<td>5/4/1997</td>
</tr>
<tr>
<td>Landsat-7 ETM+</td>
<td>9/25/2000</td>
</tr>
<tr>
<td>Landsat-7 ETM+</td>
<td>6/16/2001</td>
</tr>
<tr>
<td>Landsat-7 ETM+</td>
<td>8/27/2001</td>
</tr>
<tr>
<td>Landsat-7 ETM+</td>
<td>5/18/2002</td>
</tr>
<tr>
<td>Landsat-7 ETM+</td>
<td>7/29/2002</td>
</tr>
<tr>
<td>Landsat-7 ETM+</td>
<td>9/15/2002</td>
</tr>
</tbody>
</table>

Table 3.1. List of Landsat images.
### 3.2.3 Meteorological Data and Drought Indices

To explore the relationships between spatial and temporal changes of lakes and wetlands and climatic variability, meteorological data and indexes, such as average precipitation (PCP) and average temperature (TMP) together with the Palmer Drought Severity Index (PDSI), were obtained from National Climatic Data Center (NCDC) database ([http://www.ncdc.noaa.gov/oa/climate/climatedata.html](http://www.ncdc.noaa.gov/oa/climate/climatedata.html)) and were used as climatic variables.

Given that the PCR is underlain by low permeability glacial deposits, the main source of water to lakes and wetlands is from snowmelt, early spring precipitation, and an occasional heavy summer rainstorm. Water is lost by evapotranspiration during the hot, dry summers. Thus, precipitation is a direct measure of potential inflow, while temperature is a proxy for evapotranspiration (ET). In our study, potential ET is calculated by the Thornthwaite equation (Gray et al, 1970) as:

\[
E_T = 1.62 \left( \frac{10T_{ai}}{I} \right)^a \\
I = \sum_{i=1}^{12} \left( \frac{T_{ai}}{5} \right)^{1.5} \\
a = 0.492 + 0.0179I - 0.0000771I^2 + 0.00000675I^3
\]

Where:

- \( E_T \) is the potential evapotranspiration in cm/month;
- \( T_{ai} \) is the mean monthly air temperature in °C for month \( i \);
- \( I \) is the annual heat index, and
- \( a \) is constant.

The PDSI is a longer term (months) measure of dryness based on precipitation and
temperature, as well as soil moisture as a drought index [Palmer, 1965]. A large positive PDSI value (>4.0) indicates an extreme of wet weather. A large negative PDSI (<-4.0) points to extremely dry conditions. In general, PDSI finds important use as a comprehensive indicator of drought for understanding hydrology of lakes and wetlands.

We also averaged each climatic variable within specific time periods of 1 month, 3 months, 6 months, 12 months, 24 months, 36 months, 48 months, and 60 months, exactly before each date when a Landsat image was obtained, to identify the time factor of these variables to lake changes observed on the Landsat imagery.

3.3. Results and Discussions

The period (1985-2002) is noteworthy in that it captures a swing from extreme drought to deluge. The water areas and the numbers of pothole lakes as a function of time are estimated at a number of times through this period of extreme climatic variability. For this analysis, the lakes and wetlands are classified according to their sizes as follows: small lakes having areas less than 1 ha (about 12 pixels), large lakes having areas greater than 9 ha (100 pixels), and medium lakes whose sizes are between small and large lakes. These metrics are estimated for each of the size groups, as well as all the lakes together, and are presented in Figures 3.2 and 3.3.

3.3.1 Temporal Changes of Lakes and Wetlands

Figure 3.2 is a summary diagram showing the inter-annual variation of total lake areas. It
shows a major decline in total water area from 1987 to a minimum in 1990 of 27,237.7 ha (8/5/1990). This decline in area was followed by a dramatic increase from 1992 to 1997 with the total water area of lakes reaching a maximum 92,625.4 ha (5/4/1997). In effect, the total area represented by lakes more than tripled during the swing from drought to deluge.

Figure 3.2 also shows that more than 80% of the overall water area is reflected by lakes larger than 9 ha. Hence it was variations in the areas of this population of lakes that impacted the water-area totals. The total area represented for the lakes less than 1 ha was small. Clearly, these small lakes behave differently in that there are three minima in total area of around 670 ha, occurring on 8/5/1990, 9/25/2000 and 9/15/2002. The largest total areas of about 2600 ha for small lakes were observed on 5/19/1985, 5/4/1997 and 6/16/2001. The May/June maxima and August/September minima implies that the areas of the small lakes were to an important extent controlled by seasonal climatic patterns.

Using the year 2002 (Figure 3.2) as a demonstration, note how the total areas of large lakes declined slightly from spring to summer and then increased in autumn. However, the total area of the small lakes (smaller 1 ha) decreased by about half from spring to summer and by about half again from summer to fall producing a loss of nearly 3/4 of the water area.

The numbers of lakes in the three size categories are plotted versus time in Figure 3.3. The smallest total number of lakes was 3247 on 8/5/1990 and the largest total was 10264 on 5/4/1997. Overall, the total numbers of lakes tripled from a minimum associated with the drought in 1990 to a maximum associated with the deluge in 1997. This tripling in
numbers of lakes interestingly coincides with a tripling in total area. It is also evident that from 2000 to 2002 that the total numbers of lakes were extremely variable, with a declining overall trend.

What are different with the 2001 and 2002 data, as compared to earlier years, are multiple samples within each year. The seasonal variations are quite remarkable. The numbers of lakes in spring were much greater than that in autumn. Of the lakes present in May/June of 2001 and 2002, about half were gone by late autumn. As might be expected, the loss in numbers of lakes was most evident with the small lakes. For example, in 2002, the number of small lakes declined from 4302 in spring to 849 in autumn. With the large lakes, there was only a slight reduction in the numbers of lakes from 899 in spring to 886 in late summer. Because the small lakes represent the greatest relative fraction of lakes present, the overall changes in lake numbers are dependent on the fluctuations in the smallest numbers of lakes.
Figure 3.2 Plot of total area of lakes in different sizes (small, medium, and large) vs. date.
Figure 3.3. The number of lakes plotted vs. times by different lake sizes (small lakes, large lakes and overall lakes).
Figures 3.2 and 3.3 show that lake size was a key variable determining how water area and numbers of lakes fluctuated as the climate varied. Small lakes with areas less than 1 ha were preferentially reduced in total numbers through long hot summers. In contrast, the numbers of large lakes were much less influenced by seasonal variations and temporary declines in precipitation.

The total surface area of lakes appears to reflect climatic fluctuations over a period of years. Thus, the transition from drought to deluge in the early 1990s produced a dramatic increase in total lake area. The area parameter strongly weights the behavior of large lakes, because the total area of lakes observable with Landsat is mostly represented by the large lakes. In summary then, the total number of lakes appears to be controlled mostly by seasonal variability in climate, while lake area is controlled mostly by interannual variations.

3.3.2 Drought to deluge transition

Figure 3.4 shows a regional comparison of water areas for two extreme hydrological situations. In contrast to the effects of the severe drought on 8/5/1990, both the number and area of lakes expanded significantly in response to much wetter conditions. For example, on 5/4/1997, some big lakes, like the lakes in the Waubay Lake chain, coalesced. Moreover, the majority of new emergent lakes were concentrated within the southwestern region in contrast to the relatively much smaller changes along the eastern edge and northeastern portion of the region. A more detailed analysis together with data
from Digital Elevation Models (DEMs) indicates that the elevation of the southwestern region is much lower with much less steep local relief than other areas. This setting facilitated the expansion of the water area when deluge occurred. As shown in Table 3.2, the densities of expanded water areas within the lower regions are about two to three times more than those within the higher regions. Ducks and other types of waterfowl are more concentrated within the lower areas in the southwestern region, as most of the small potholes emerged in the deluge period.

<table>
<thead>
<tr>
<th>Area</th>
<th>Elevation (m)</th>
<th>Expanded Water Area (ha)</th>
<th>Total Area (ha)</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>406 - 546</td>
<td>22,020.93</td>
<td>89,114.22</td>
<td>0.2471</td>
</tr>
<tr>
<td>2</td>
<td>546 - 557</td>
<td>24,147.99</td>
<td>94,264.65</td>
<td>0.2562</td>
</tr>
<tr>
<td>3</td>
<td>557 - 574</td>
<td>10,559.25</td>
<td>85,339.08</td>
<td>0.1237</td>
</tr>
<tr>
<td>4</td>
<td>574 - 644</td>
<td>6,143.58</td>
<td>81,433.44</td>
<td>0.0754</td>
</tr>
</tbody>
</table>

Table 3.2 Expanded water areas in 1997 from 1990 within different areas
Figure 3.4. Comparison of lake occurrences between 1990 (dry year) and 1997 (wet year)

3.3.3 Connections to climatic variability

Clearly, the small lakes and large lakes responded differently to seasonal and inter-annual fluctuations in climate. This difference in behavior has important implications concerning the concept of hydrologic drought as it applies to prairie pothole lakes that I will explore in this section.
The concept of drought is defined variously to the extent that it can be ambiguous and sometimes confused with aridity [Beaudoin, 2002]. The term drought implies more than simply an ‘absence of rain’. Practically, the concept of drought is associated with operational definitions. Most pertinent to my studies are (i) meteorological droughts – defined by a long-term lack of precipitation, and (ii) hydrologic droughts - defined as a period of reduced surface runoff, decreased streamflow, and reduced ground-water levels [Beaudoin, 2002]. By implication, hydrological droughts would also encompass a reduction in the stage of closed-basin lakes. Both meteorological and hydrological droughts describe a deficiency in the water supply. Other operational definitions reflect a reduction in the demand for water [Beaudoin, 2002].

Because lakes have a complicated relationship with climate, the question arises as to what extent indices like the Palmer Drought Severity Index (PDSI), average precipitation (PCP) or average evapotranspiration (ET) provide useful information on the state of pothole lakes. PDSI measurements are of particular interest because of the direct connection to drought and proxy records of climate, for example, tree rings archives. This question is addressed by statistical analyses that examine correlations of PCPs, ETs, and PDSIs to changes in the sizes and numbers of small and large lakes with time.

The purpose of this analysis is to determine how water areas and numbers of lakes of different sizes depend upon conditions in the months preceding the measurement. The variables PCP and ET measure the water availability through precipitation and the losses
through snowmelt. PDSI is a mixed indicator of drought/deluge, characterizing the departure from normal conditions. The values of these variables are averaged through 1, 3, 6, 12, 24, 36, 48 and 60 months time periods to capture both seasonal and inter-annual variations. It is tested here because of its common use to characterize climate in North America.

Panels (a) and (b) in Figure 3.5 clearly show that both total area and numbers of small lakes are mostly correlated to a 6-month average ET with the largest correlation value of -0.7980 and -0.6810, respectively, and overall, PDSI has weak correlations. Surprisingly, PCPs within 6 months periods are negatively correlated with changes of small lakes. One explanation is that relatively large quantities of precipitation were received during summer, but only slightly contributes to the lake budgets because of the high ET rate. In general, Figure 3.5 suggests that for the small lakes and wetlands, evapotranspiration for the preceding six months determines the sizes and numbers of small lakes and wetlands within the PCR. Note that in the case of ET, averaging periods longer than six months lose their significance because temperature values tend toward a single annual average without significant trends. The small lakes and wetlands with relatively small water volumes and surface areas are being primarily impacted by evapotranspiration as compared to the large lakes.
Figure 3.5. Correlations between areas and numbers of lakes and wetlands and PCP, ET, and PDSIs in certain time periods.
Figure 3.5 Continued

![Graph showing correlation between variables over time]

**c. Total area of large lakes**

![Graph showing correlation between variables over time]

**d. Number of large lakes**
Panel (c) and (d) present the results of comparable correlations analyses for the large lakes. It is obvious that the patterns of correlation for the large lakes are very different than the smaller lakes. The total areas and numbers of large lakes are significantly correlated to long-range averages in PCPs and PSDIs. There are no significant correlations of these parameters with ETs. The highest correlation values are 0.9410 between total water areas and the 60-months average PDSI and 0.9310 between total numbers of large lakes and the 48-months average PDSI. Lake water areas and lake numbers are also strongly correlated with 48 or 60 month averages in precipitation (PCP). Correlations are fairly constant for periods of 48 months and 50 months for total areas and for periods of 36 months to 48 months for numbers of lakes.

The results for the large lakes imply that some condition represented by PCP or PDSI need to be imposed for four or five years (48 to 60 months) before the numbers or water areas of large lakes change. Thus, large lakes should be able to sustain a drought of less than three to four years without undue impact. Short-term changes (i.e., seasonal) in temperature or precipitation have no impact on these lakes. PDSI combines rain and temperature as well as soil moisture, in the calculation. However, when averaged across a 48 months period, the temperature only has a small contribution to PDSI values because of the insignificant annual variation, which makes PDSI relying more on the amount of the rain. This is shown in Panel (c) and (d) that the correlations between large lakes and PSDIs are very close to correlations between large lakes and PCPs, particularly for long periods, such as 48 months and 60 months. Also, the larger a lake is, the longer period of PDSI it is correlated to. This is implied by the number of large lakes is mostly correlated
to the 48 months average PDSI while the total area has the highest correlation with the 60
months average PDSI, because generally the number of lakes is more contributed by
relatively smaller lakes, but the total area is more contributed by larger lakes. Long term
averaged PDSI is a very good indicator of changes of these large lakes for two major
reasons: first, PDSI is such a comprehensive climatic index which incorporates all
fundamental water balance modeling components like rain, evapotranspiration, soil
water contents, that provides overall information of the climate; secondly, PDSI takes
historical meteorological conditions into account, which is more important for large lakes,
as they are more impacted by cumulative effects from the past.

Figure 3.5 shows that 6 month ETs (panel (a) and (b)) and 48 month PSDIs (panel (c) and
(d)), respectively are good predictor variables for total areas and numbers of lakes and
wetlands. I applied linear regression models between these predictor variables and
response variables. To avoid multicollinearity, only one predictor variable, either a six
month ET or a 48 month PDSI is included in each regression model. Figure 3.6 shows
linear regressions and equations relating the total areas and numbers of lakes and
wetlands of different sizes to their most correlated variables (i.e., average ET for the
preceding six months and average PDSI for the preceding 48 month. The linear trends in
Figure 3.6 are all significant, although the regression (in panel b) between numbers of
small lakes and the six-month ET has the smallest $R^2$ of 0.48. There are two obvious
outliers indicated by the dates 4/15/1990 and 6/16/2001. Both these data points lie far off
the line of each regression of small lakes in panel (a) and panel (b) of Figure 3.6. The
outlier appears to be related to a somewhat atypical series of precipitation events. The
4/15/1990 data point is the result of a meager winter snow pack. The outlier of 6/16/2001 was caused by a huge rainfall event that occurred several days before the Landsat pass. The number of water areas of the small lakes are thus extremely sensitive to short-lived climatic fluctuations.
Figure 3.6. Linear regressions between total areas and numbers of lakes with meteorological variables continuation
c. Total area of large lakes

![Graph showing the relationship between Average PDSI of 48 months and Total area of large lakes. The equation is Total area = 26938.54 + 8610.68 * PDSI48. The R^2 value is 0.8869.]

d. Numbers of large lakes

![Graph showing the relationship between Average PDSI of 48 months and Number of large lakes. The equation is Total number = 520.75 - 111.63PDSI. The R^2 value is 0.8340.]

Figure 3.6 Continued
The slope of each regression line describes how the response variable reacts to a change of the predictor variable. For example, for small lakes, a 1 cm increase in 6 month-average ET will reduce the mean total water areas by 107 ha (Figure 3.6a) and will decrease the mean number of lakes by 170 (Figure 3.6b). Decreasing the 48 month average PDSI for large lakes by 1 will reduce the total water area by 8611 ha (Figure 3.6c) and the numbers of lakes by 112 (Figure 3.6d).

The short and long-term behavior of lakes in terms of total water areas and numbers is particularly relevant in other applications. For example, the Prairie Coteau region of South Dakota is particularly productive for ducks, as compared to other areas, such as the Drift Plain and Red River Valley in the United States [Austin, et al., 2001; van der Valk, 2005]. A variety of research [e.g., Sorenson et al. 1998; Austin 2002; Niemuth and Solberg 2003] showed strong associations of waterfowl populations of various species with hydrological conditions (number, density, and water level) of wetlands within the PPR. If it is the case that numbers of small lakes and wetlands hold linear relationships with population of birds as in preceding studies [Sorenson et al., 1998], we can estimate how much this population changed with climate variability. For instance in 2002, 4302 small lakes and wetlands emerged in May, providing excellent nesting and breeding habitats for waterfowl. Small potholes were used as a demonstration because they are more productive to waterfowls by providing a higher density of nutrients and facilitating more emerged and submerged plants with shallow water levels in comparison with large lakes. As it is mostly related to small lakes and wetlands, a 6 month ET is a more
reasonable indictor to duck populations than PDSI that was used in previous studies [Sorenson et al., 1998]. Under the scenario of global warming, shrinkage of small potholes in both areas and numbers will considerably change the population of ducks in this region.

3.4. Conclusion

The usefulness of GIS and remote sensing approaches were demonstrated to estimate total areas and numbers of lakes and wetlands in different sizes for the study area in South Dakota, which was substantially impacted by the combination of drought and deluge. Changes in total areas and numbers from 1985 – 2002 using Landsat images showed systematic changes controlled by climate. The total area are essentially tracking the behavior of large lakes (>9 ha), which appears to be tracking the broad climatic behavior (drought to deluge) and highly associated with the 48-month PDSI. The small lakes appear to be controlled by seasonal variability in climate with the highest correlation with 6 months ET. The emergence of a large number of lakes and wetlands in spring implies the importance of the snowmelt at this time of the year. Areas and numbers appear to be dramatically reduced in late autumn due to the evaporative losses throughout the summer. Lower regions have a higher density of lakes and wetlands and are easier for water to expand from drought to deluge, hence breed most of waterfowl in comparison of higher areas in this area. More work is presently underway to examine the broad statistical structure of the surface water resource in the PPR.
CHAPTER 4
SYSTEMATICS IN THE SIZE STRUCTURE OF PRAIRIE POTHOLE LAKES THROUGH DROUGHT AND DELUGE

4.1. Introduction

Studies have shown that the water areas of lakes and wetlands follow a power law [Rapley et al., 1987; Wetzel, 1990; Birkett and Mason, 1995; Meybeck 1995, Lehner and Doll, 2004]. This systematic size structure was first pointed out by Korcak [1940]. Korcak’s power law showed that areas of geographical objects follow hyperbolic distribution with a power density function. Plotting counts or perimeters versus areas in logarithmic coordinates produces a straight line. For lakes, this idea was tested by Kent and Wong in 1982 for about 2,500 lakes in Canada. Now, power law relationships between counts and water areas of lakes have been demonstrated and applied in global and regional assessments of surface-water resources [Lehner and Doll, 2004; Downing et al., 2006; Daya Sagar, 2007]. Beyond the obvious systematics in the change of the frequency of lakes of different sizes are remarkable coincidences in the slopes of the regression lines for different areas [Downing et al., 2006].

With power-law relationships, complex heterogeneity among lakes can be rationalized by just a few parameters. Fractal theory, which describes the inherent complexity of such
systems, depends on power-laws to a considerable extent. If a power-law relationship exists that is independent of scale, the object or process is fractal. The doubled slope value of a power-law line is defined as fractal dimension, a parameter describing the complexity [Mandelbrot, 1982; Goodchild and Mark, 1987]. The interesting self similarities of fractal objects or processes have proven useful in numerous applications, for example, in hydrology, geomorphology, and terrain modeling [Bishop et al., 1998; Rapley, et al., 1987; Robert and Roy, 1990; Rodriguez-Iturbe and Rinaldo, 1997; Rosso et al., 1991; Sagar and Srinivas, 2002; Sagar and Tien, 2004; Veitzer and Gupta, 2000].

This study examines the possibilities for defining multi-temporal power-law relationships to assess the complex structure and behavior of lakes and wetland systems in the Prairie Pothole Region (PPR) of the Great Plains of the United States and Canada. The PPR contains literally millions of lakes and wetlands, which are sensitive to climate variability and likely to be impacted by global climate change [Larson, 1995; Sorenson et al., 1998]. While much is known about the fundamental hydrology of lakes and wetlands in the PPR [e.g., Winter and Rosenberry, 1998; Euliss and Mushet, 1996; Johnson et al., 2004], the size and complexity of the PPR, coupled with the large numbers of lakes involved make broad regional/temporal assessments difficult. If consistent multi-temporal power-law relationships can be shown to hold, it should be possible to use limited historical data (e.g., climate data and aerial photography) to better understand the regional response of lakes and wetlands to important periods of drought and deluge in the past.
Our work focuses on pothole lakes and wetlands located along the northern tip of the Prairie Coteau in South Dakota, where pothole lakes and wetlands are numerous. Of particular interest is the behavior of the lake/wetland complex through an interesting transition from a period of drought (1988-1992) to one of deluge (1993-1998) and at the end of the Dust Bowl Drought of the 1930s. Although the locus of the Dust Bowl Drought was centered in Kansas and Oklahoma, it is the most important drought of the last century further north as well. The challenge with understanding hydrologic conditions during Dust Bowl times is the lack of data on small lakes and wetlands.

4.2. Methodology

4.2.1 Description of the study area

The PPR of central North America is unique because of the extremely large numbers of small lakes, ponds, and wetlands located there and its important ecological significance. This approximately 750,000 km² region extends from southern Canada, including Alberta, Saskatchewan, and Manitoba to the United States including Montana, North Dakota, South Dakota, Nebraska, Minnesota, and Iowa. The exact numbers of ponds, lakes, and wetlands is unknown but could easily be several million [Swanson, 1995]. Although the PPR represents about 10% of the total breeding and nesting area for waterfowl in North America [Guntenspergen, et al., 2002], it produces more than 50% of all breeding ducks and more than 60% of continental mallards [Batt, et al., 1989].
Pothole lakes, ponds and wetlands are isolated from stream networks. Major inflows of water come from snowmelt and summer precipitation, whereas the greatest loss of water is due to evapotranspiration [Winter, et al., 2001]. With the high rates of evapotranspiration in summer and variable precipitation from year to year, the lakes and wetlands can exhibit significant variability in size.

During the last century, parts of the PPR experienced two extreme droughts, including the Dust Bowl drought of the 1930s and a shorter and more recent one from 1988 to 1992. This latter drought was followed by the most significant wet period of the century beginning in 1993 [Winter and Rosenberry, 1998]. The resulting flooding in ponds, lakes, and wetlands had not been observed, perhaps since the middle of the 19th century. Currently the region is arid again and people are concerned with the possible reoccurrence of Dust Bowl conditions of the 1930s [Andreadis, 2005; Fye 2003; Stahle, 2007].

The 5750 km² study area is located in the PPR at the northern tip of the Prairie Coteau of northeastern South Dakota (Figure 2.1). During the Wisconsin deglaciation, stagnation of the ice on this broad upland area (i.e., Prairie Coteau) produced hummocky topography and ultimately thousands of close-basin pothole lakes and wetlands. Bedrock in this area is Cretaceous shale, and is mantled by more than 120 m of glacial till. The region is important for agriculture. Flat areas are typically used for dry-land farming, while more hummocky land is used as rangeland. In the past years, there have been aggressive
federal programs to take marginal land out of agricultural uses to enhance waterfowl production.

This region has a subhumid to subarid continental climate with short hot summers, long cold winters, low levels of precipitation, and high evaporation. The mean annual rain ranges from 50 to 56 cm (20-22 in) and temperature ranges from -17.2 °C in January to 29.4 °C in July. Figure 4.1 shows precipitation data with rainfall and snow separated from 1986 to 2003. The data are from the Andover station, located in Day County, South Dakota. A relatively small proportion of the annual precipitation (averaging 66.8 cm) occurs as snow. However, snowfall amounts can be quite variable (Figure 4.1), as 1994 and 1997 had the largest amounts, which, for example, were significantly greater than those in 1990 and 1992 (the drought year). Summer rainfall is quite variable from month to month in different years. However, from year to year, this variability in amount was not as distinctly different as found with the snowfall.
Figure 4.1 Monthly precipitation data from the Andover station in Day County, South Dakota from 1986 – 2003.
4.2.2 GIS and remote sensing approaches to evaluate pothole lakes

Most previous studies of lakes and wetlands in the PPR commonly rely on careful observations of conditions for a few clusters of pothole lakes and wetlands. This study examines the systematics of the nearly 10,000 pothole lakes within the study area. Given the large number of lakes involved, it was necessary to rely on remote sensing technologies, which are compatible with the scope of the observational problem.

Since the 1970s, aerial photos and satellite images have been used for mapping lake areas [Lunetta and Balogh, 1999; Ozesmi and Bauer1, 2002; Pietroniro and Prowse, 2002; Sawaya et al., 2003; Schmugge et al., 2002; Stewart et al., 1980]. These extensive data archives when coupled with processing capabilities of a Geographical Information System (GIS) provide a powerful approach for examining how lake surface areas change with time.

Landsat imagery was used to provide historical data on the occurrence of pothole lakes for various different times from 1987 to 2002. Using ArcGIS, lake areas were determined automatically by advanced image classification and image processing methods, described in detail by Chapter 2 and Zhang [2007]. Because of the moderate spatial resolution of about 30 m, we were limited with Landsat images to studying pothole lakes that were larger than 0.1 ha.
DOQO with about 1 m resolution were available for parts of the study area and were also used in ArcGIS to provide area estimates for smaller water bodies. Small subareas within the larger study area were specified to provide a similar density of pothole lakes. For each sampling site, all the potholes lakes were delineated manually on the corresponding DOQQ and the area was calculated for each water body with GIS tools.

Figure 4.2 provides examples of the two different types of imagery. Panel (a) is a classified water area image from a Landsat scene. The area covered by water is estimated by summing all water pixels for each lake. Panel (b) shows examples of manually delineated pothole lakes on a DOQQ for a smaller area. Using imagery of different resolution provides water areas for pothole lakes of differing sizes. Estimates from Landsat images (e.g., Panel a) encompass lakes ranging in size from Waubay Lake (the largest) to lakes larger than several Landsat pixels. Estimates from DOQQ (e.g., Panel b) provide area estimates for lakes smaller than one Landsat pixel (about 900 m²), but larger than 100 m². This lower threshold helped to avoid errors because lakes smaller than 100 m² could not be properly identified on DOQQs.

The large data base of lake areas developed from the Landsat and DOQQ imagery provided the basis for our assessments. It turned out that power-law functions were extremely useful in this respect. In general, one power-law line was produced from one image. Because DOQQ imagery provided only one-time coverage in 2003, there is at least one pair of Landsat and DOQQ images to study the full spectrum of lake sizes.
Figure 4.2 Pothole lakes presented on a Landsat image and the DOQQ at different resolutions.
4.2.3 Hierarchical Bayesian linear model

After water-area estimates for the pothole lakes were obtained from each Landsat or DOQQ image, a power-law was developed by plotting numbers of lakes of a given size versus area on a log-log scale. Commonly, statistical models are used to estimate parameters (slopes and intercepts) of regression lines. The most common technique is a linear regression model, which provides least-squares estimates of parameters. This statistical approach is simple to use under Gauss-Markov assumptions of normality with 0 mean and a constant variance for random errors. However, this assumption was not satisfied by the data in this study. An alternative Bayesian hierarchical linear regression model was used. Unlike least-squares models, which produce best estimators for unknown fixed model parameters, Bayesian models generate posterior distribution samples of model parameters directly, based on known data. Thus, they produce a more intuitive interpretation of confidence intervals of model parameters, which are of particular interest in this study.

The first step with the Bayesian hierarchical linear regression model is to take all lakes identified from a scene and classify them into bins according to their areas. A bin contains a collection of lakes of similar size. All the bins have a uniform width, for example, one Landsat pixel or 100 m². The central value of a bin provides information on the approximate size of those lakes. For example, a bin with a central value of 1,050 m² and a width of 100 m² would contain lakes with areas ranging from 1,001 m² to 1,100 m².
The hierarchical Bayesian linear model is described as following:

for \( i \)th bin:

\[
\log \text{Count}[i] \sim N(\mu[i], 1/\tau) \\
\mu[i] = \alpha + \beta \times \log \text{Bin}[i] \\
\alpha \sim N(0, 10^6) \\
\beta \sim N(0, 10^6) \\
\tau \sim \text{Gamma}(0.001, 0.001)
\]

where \( \text{Count}[i] \) is the count of lakes within the \( i \)th bin; \( \text{Bin[i]} \) is the central value of the \( i \)th bin.

Initially, the model defines noninformative prior distributions that the logarithm value of the count of pothole lakes in the \( i \)th bin has a normal distribution with a mean of \( \mu[i] \) and a variance of \( 1/\tau \) with \( \tau \) following a gamma distribution. The value of \( \mu[i] \) can be calculated by linear equation: \( \mu[i] = \alpha + \beta \times \log \text{Bin}[i] \), where \( \alpha \) (intercept) and \( \beta \) (slope) have independent identical normal distributions with a mean of 0 and a variance of \( 10^6 \). The parameters of prior distributions \( 10^6 \) and 0.001 of \( \alpha \), \( \beta \), and \( \tau \) were selected to ensure that these distributions approximate to a uniform distribution and proper (the integration of the density function across the whole universe equals to 1) to guarantee that the posterior distributions are convergent and proper.

A Markov chain Monte Carlo (MCMC) algorithm and Gibbs sampler \([Gelman et al., 2003]\) was employed to generate posterior distributions of \( \alpha \) and \( \beta \):

\[
p(\alpha \mid \log \text{Count}[i], \log \text{Bin}(i)) \quad \text{and} \quad p(\beta \mid \log \text{Count}[i], \log \text{Bin}(i))
\]

The posterior
distributions on two parameters of the power-law line are used in this study to describe the behavior of water areas of pothole lakes in each image.

4.3. Results

4.3.1 Power-law lines of potholes for different times

To validate the power law in describing the size structure of this prairie-lake complex, a series of Landsat images were used to produce power-law lines at different times through the span of significant hydrological variability from 1987 to 2002. Four power-laws were generated for years 1990, 1992, 1997, and 2002, as shown in Panel (a) of Figure 4.3. The linear trends in lake count as a function of area with logarithm coordinates are remarkably consistent and validate the power-law relationship. Panel (a) also provides visual evidence of the impact of the drought to deluge transition on the structure of the pothole-lake system. For instance, in the 10 pixel bin (containing all pothole lakes with areas larger than 10 pixels, but smaller or equal to 11 pixels, 0.9 ~ 1 ha in the figure), the counts of these lakes increased from 300 on 5/06/1992 to 620 on 5/04/1997.

The Landsat images used to produce the lines in Panel (a) were all selected from late April to early May. The structure of the pothole-lake system is examined at about the same time each year to make them seasonally comparable. Prairie pothole lakes can exhibit distinct variability in size within the same year [Zhang, 2007]. In Panel (a), the lines for 1997 and 1992 are the upper and lower bounds, respectively, of these power-law
lines during the late spring season. This range reflects the variability due to the drought of 1988-1992 and a deluge reaching a maximum in 1997.

Another important feature of these lines is that they all are nearly parallel. This result implies that the size structure of the pothole-lake system remained constant, even though climatic conditions were significantly different. In other words, the counts of lakes and wetlands of different sizes maintained their ratio of size abundances through the different years.

Panel (b) in Figure 4.3 displays the power-law relationships at different times of the years for 1990, 2001, and 2002. Although a linear power-law relationship still holds, the lines are not parallel at different times in the same year. Lines from summer or fall in a given year are lower and flatter than from the corresponding spring season. This consistent pattern of change in the power-law through the year suggests that the areas of small wetlands and lakes are being impacted by evapotranspiration during the hot summer season when vegetation is flourishing.

In Figure 4.3, the linear trends are only significant for lakes having an area less than $10^2$ pixels (9 ha). The variance increase for the larger lakes is due to the relatively smaller number of lakes in certain size classes (bins) as lake sizes become larger. A regression line for the large lakes alone would be essentially flat. Another possible explanation is that the constant bin width is distorted or more compressed in logarithm coordinates for the larger lakes. The departure of the smallest lakes from the fitted lines is caused by the
coarse resolution of Landsat images with pixels with an area of about 30×30 m². The area estimations have more uncertainties and errors when integrating smaller numbers of pixels for the smaller lakes.

Figure 4.3 Power-law lines of pothole lakes from 1990 – 2002. (a) Inter-annual patterns (b) Seasonal patterns.
4.3.2 Parameters of power-law lines from Bayesian analysis

The convergence of MCMC simulations for posterior distributions of parameters are assessed to validate the Bayesian models. The time-sequence plots for parameters generated by MCMC are shown in Figure 4.4 for the case of the power-law line for 4/23/1987. The trace plots for parameters of power-law lines on other dates are similar as the line for 4/23/1987. They are essentially similar to the results displayed here. Figure 4.4 shows that the simulated samples of three parameters of the model, $\alpha$ (intercept), $\beta$ (slope), and $\tau$ (inverse of variance) all converged quickly, within the first 300 iterations. Although the samples of posterior distributions for $\alpha$ and $\beta$, which were drawn from MCMC were slightly wavy in the series, an assumption of constancy is valid, if the tiny variances of both distributions are taken into account.
Figure 4.4 Posterior distributions generated by MCMC chains for parameters of the power-law line on 4/23/1987. ‘$\sigma$’ is the square root of the inverse value of ‘$\tau$’, which refers to the square root of the variance.
Once the convergence of samples simulated by MCMC for parameters in the Bayesian model has been achieved, the posterior distributions can be summarized for parameter estimations. To remove uncertainties produced at the beginning iterations from initial values, we discarded the first 2000 samples. We compared \( \alpha \)'s (intercept) and \( \beta \)'s (slope) of different power-law lines inter-annually (Figure 4.5) and inner-annually by box plots of their posterior distributions(Figure 4.6). For each distribution, the box plot gives the upper and lower bounds of the 95% confidence interval shown as the top and bottom line segments. The rectangle inside this interval shows samples within 25% (lower bound of the rectangle) and 75% (upper bound of the rectangle) quintiles of the distribution, and the solid dark line segment represents the median of the distribution.

Figure 4.5 shows how intercepts and slopes of power-law lines in late spring changed under different climatological conditions for different years. The median values of intercepts ranged from 4.12 to 4.46, and the median values of slopes differed from -1.59 to -1.80. The confidence intervals shown by box-plots allowed us to interpret differences in these parameters. If the two 95% confidence intervals overlap each other, it means that these two parameters cannot be differentiated at the 5% significance level. With the two extreme years 5/6/1992 (drought) and 5/4/1997 (deluge), the intercepts of the power-law lines were different at the 5% significance level. but the slopes were almost identical. Under the assumption that the two lines were parallel with a difference of about 0.3 in intercepts for 5/6/1992 and 5/4/1997, we can conclude that the number of lakes on 5/4/1997 within each bin was consistently about two \( (10^{0.3}) \) times the number observed.
on 5/6/1992. Examination of Figure 4.1 indicates that 1992 winter was an extremely poor snow year, while 1997 had large snow accumulations.

In Figure 4.5, the power-law line for 4/15/1990 had the largest negative slope as compared to the smallest value of 5/18/2002. Only these two results are different from each other at the 5% significance level, because the 95% confidence intervals do not overlap. We think that the differences in these slopes might be caused by the inability to maintain a close time correspondence. Even though these Landsat images were selected to be time synchronous in a given year, seasonal changes could be influential during the time gap from 4/15 to 5/18 for these two lines. Notwithstanding the slight differences among these two values, Figure 4.5 indicates that these power-law slopes for late spring were surprisingly similar with a slope value about -1.7 despite the variability in climate through this period. It is likely that the tendency for snowmelt to maximize the number and surface areas of smaller lakes during early summer pushes the slopes toward the higher end of their natural range, providing relatively consistent slopes. The numbers of lakes analyzed are obviously too few to develop a more definitive explanation.
Figure 4.5 Box-plots of distributions of slopes and intercepts of annual power-law lines simulated by Bayesian linear models.

The sensitivity of lake surface areas to variability in climate implied by Figure 4.5 is shown convincingly in Figure 4.6 by boxplots of distributions of slope and intercept of
power-law lines in different seasons. It is noteworthy that for each of the three years, the lines for spring had higher intercepts and higher slopes than subsequent lines in each year. Furthermore, these parameters are different at the 5% significance level with most of the intervals on box-plots in Figure 4.6 far apart from each other. The highest intercept value of 4.40 and the highest slope value of -1.80 came from 4/15/1990, compared with the lowest intercept of 3.12 and the lowest slope value of -1.24 on 8/5/1990. As indicated by Figure 4.3b, the change in slope from April to August 1990 is due to preferential reductions in the size of the smaller lakes.

This pattern of variation shows a seasonal influence on power laws generated from spring through summer. The structure of the pothole lake system changes both in terms of the numbers of lakes of a given size and the size relationships within the entire family of lakes. Generally, with the temperature and evapotranspiration rate rising dramatically from spring to summer, the area of smaller lakes declined to a much greater extent than larger lakes. These changes cause the power-law line to move downward and become flatter. Given these results, care should be exercised in using a single power law to estimate how smaller lakes are responding, given these seasonal effects.
Figure 4.6 Box-plots of distributions of slope and intercept of seasonal power law lines simulated by Bayesian linear models.
In summary, the parametric Bayesian analysis with MCMC sampling provides a means to analyze how power law lines of distributions of potholes lakes behaved annually and interannually. Indications are that the relative size structure of the lakes was similar in the springtime but changed through the year. Therefore, it is important in developing power laws on lake size distributions to account for seasonal biases.

4.3.3 Power-law in small scale and reconstruction of the Dust Bowl scenario

The most famous North American drought of the 20th Century was the “Dust Bowl” of the 1930s. This drought devastated the agricultural economy of the Great Plains for about a decade with health and social impacts lingering for years afterwards. Researchers have suggested that the abnormal deficits in precipitation and high temperatures were related to extreme anomalies of Pacific Ocean sea surface temperatures (SSTs) [Schubert et al. 2004; McCabe, et al. 2004]. Laird, et al. [1996] considered the drought of the 1930s as unremarkable in relation to others of greater intensity and frequency before AD 1200 in the Great Plains.

While some hydrological and meteorological records are available for the early 1900s in South Dakota, the fate of the pothole lakes during the Dust Bowl drought is not well known. We investigated the impacts of drought in the Prairie Coteau region using a series of aerial photos taken on 7/29/1939. Figure 4.7 provides illustrative examples of comparable segments of the digitized 1939 imagery with the DOQQ. Thus, it is possible to compare the water distribution in potholes near the end of the Dust Bowl drought with
more recent flooded conditions reflected in the DOQQ created on 7/04/2003. The contrast between the areal extent of pothole lakes in 2003 and that in 1939 is striking. Some large pothole lakes were completely dried up in 1939.

The two sets of aerial photographs also provide a basis for testing power-law relationships between the numbers of potholes and their areas, albeit on a much smaller scale. Unlike the previous processing, lakes on the two sets of aerial imagery were analyzed by manually outlining water areas and by calculating areas with GIS tools. We took the next step of extrapolating power-law relationships from aerial photos in a small scale study area to the Prairie Coteau region more generally. This step, however, requires convincing evidence that such an extrapolation is reasonable.

The analysis starts with the 2003 DOQQ to develop a power-law on the distribution for lake areas at a small scale. All lakes within a 59 km² test area were analyzed using the DOQQ from 7/04/2003 and GIS tools. This test area (Figure 4.2) was randomly selected within our Prairie Coteau study area to provide samples of lake with a water area ranging from 100 m² to 10000 m². The power-law relationship developed from the DOQQ was plotted together with that estimated previously from the Landsat image of 7/29/2002 (Figure 4.8a).
Figure 4.7 Comparison of lakes from aerial photos taken in 1939 (a) and 2003 (b).
The size of the test region analyzed by the DOQQ is much smaller than the region studied with Landsat. With many fewer lakes in the test area, the two power-laws plot at different places on Figure 4.8a. The DOQQ power law is defined by solid circles (bottom left, Figure 4.8a) versus that from Landsat defined by solid triangles. It is, however, possible to normalize the area of coverage of the DOQQ to the same area (4,365 km²) as the Landsat image. This normalization effectively moves the DOQQ-derived line vertically (open circles, Figure 4.8a). Remarkably, with the normalization the two line segments appear to coalesce, essentially describing a single power law.

This result suggests that areas of prairie pothole lakes in the Prairie Coteau region obey the fractal properties observed for lakes in other places. We are able to validate a power-law relationship for a spectrum of lakes ranging in area from 100 m² to more than 40,000 m². For small patches of surface water less than 100 m² in area, we lose the ability to evaluate areas from aerial photos.

This systematic behavior of pothole lakes in the Prairie Coteau region provide a logical basis for interpreting the structure of the lake complex in 1939 from a small collection of aerial photographs of that period. It is this idea that we apply to interpret the likely structure of the pothole lake systems of the Prairie Coteau region in 1939 at the end of the Dust Bowl drought.

A second 90 km² test area was developed with the aerial photos of 1939 to completely cover the area defined by the DOQQ from 2003. The water area of lakes and wetlands greater than 100 m² of potholes was measured using the 1939 photography to define the
structure of the pothole-lake system. To provide a regional comparison, the 1939 relationship was normalized by \( \log \left( \frac{4365}{90} \right) \) above the original, which equals to 1.69 in logarithm value of counts. The result of this normalization process is shown in Panel (b) of Figure 4.8.

Our previous analyses provide justification to extrapolate this curve to larger lakes, which yields the estimated power-law for the overall region for 1939. To provide context for the 1939 result, the power-law relationships are plotted for 1992 (drought) and 1997 (deluge) estimated from Landsat images. This single plot thus represents the distributions of lake areas at the end of the three most significant hydrological extremes of the last century. This estimated power-law line for 1939 is comparable to the other lines in terms of slope.
Figure 4.8 (a) Power-law lines from Landsat and DOQQ (b) comparisons of power-law lines among year 1939, 1990, 1992 and 1997.
The structure of large lakes > 10,000 m² lakes in 1939 stands in stark contrast to the second worse drought of the century (lines for 1990, and 1992), and the greatest deluge (line for 1997). Figure 4.8b suggests the extreme variability in the abundances of lakes of a given size for these different times. For instance in 1939, there were about seven lakes with an area of 10,000 m². During the drought represented by data from 1990 and 1992, there were 15 and 25 lakes of this size, respectively. During the subsequent deluge in 1997, there were about 50 lakes of this size.

The smaller lakes (e.g., 100 to 1,000 m²) show different patterns of variability. Interestingly, the drought of 1990 provided a great impact on these smaller lakes than was evident with the 1939 results.

4.4 Discussions

The pothole-lake system of the Prairie Coteau has an area-frequency relationship that is similar to that observed for combined lakes around the world [Lehner and Doll, 2004; Downing et al., 2006] or for a regional collection of lakes [e.g., Kent and Wong, 1982]. The fractal dimension of the combined DOQQ and Landsat data set, which ranges over more than three orders-of magnitude (Figure 4.8a), is about 2.5 – 2.8 as compared to 2.0-2.7 [Goodchild, 1988], 2.1 for natural lakes [Downing et al., 2006], and 2.8 [Daya Sagar, 2007]. In this respect then, our study confirms previous observations pointing to systematics in the size structure of lakes.
Where our study begins to distinguish itself from others is in terms of temporal trends. The Landsat data archive shows how the structure of the lake complex changes with short-term and long-term climate effects. For any snapshot in time, a single power-law relationship holds for the lake complex. This relationship, however, changes in time as the pothole-lake system responds dynamically to changes in precipitation and evapotranspiration.

Figure 4.9 is a conceptual model of how the power-law relationships for the pothole-lake complex vary as a function of changing hydrologic conditions. For illustrative purposes a single line (e.g., hypothetical June) is shown there. The key to understanding the power-law behavior is to realize that in general the small lakes respond rapidly to short-term seasonal climatic cycles and the large lakes respond much more slowly to multi-year trends in drought and deluge. Effectively then, the small-lake portion of the power law (left part of line Figure 4.9) moves up and down independently and more rapidly, as compared to the large-lake portion of the line.

In any given year, the typical cycle of water excess from snowmelt in spring, which is followed by evapotranspiration through hot, dry summers with sporadic rains, reduces the area of small lakes in just a few months. With large stored water volumes, the annual cycle has less impact with the large lakes. Differences in the hydrologic resiliency in the smaller versus the larger lakes means that the left side of the power-law line moves up and down significantly, while the right end stays in about the same place. With a wet spring or previous wet fall, the left end of any power-law line would start up at the left
(Figure 4.9). It declines rapidly through summer and fall, as small lakes decline in response to evapotranspiration. Thus, much of the variability in the power-law parameters (slope, intercept value) is due to the sensitivity in the size of smaller lakes to variability in snowmelt runoff and the combined effects of summer rain plus evapotranspiration.

Figure 4.9 Conceptual model of power-law relationships for pothole-lake system under different hydrologic conditions.

The end of the power-law line representing larger lakes responds more slowly to broad-scale cycles of drought and deluge because large stored water volumes attenuate the short-term seasonal variability. The Dust Bowl drought over a decade eventually depleted the large lakes, resulting in a decline in the right end of the power-law line. In the case of the shorter drought from 1988-1992, the impact on the large lakes was evident, but less
substantial. The deluge from 1993 to 1998 raised the large-lake end of the line. Overall, there is less variability in the position of the power-law line toward the large-lake portion of the curve. The shaded region on Figure 4.9 shows a fan-shaped envelope that would encompass all the power law lines determined for our study area.

These results have several important implications. Because of natural seasonal fluctuations and longer term cycles of drought and deluge, no single power-law relationship describes the pothole lake complex in this region. For example, slope is maximized by a wet spring in the midst of a long-term drought or minimized by a dry summer in the midst of a multi-year deluge. Also, for lakes, the concept of droughts/deluges as extreme climatic deviations from long-term averages is probably most applicable to large lakes. Small lakes swing naturally from periods of water excess to periods of water deficiency each year, which may or may not be related to the larger scale climatic influences. Thus, wildlife, like ducks, adapted to small water bodies may be less affected by multi-year drought and deluge than one might expect because the small lakes can respond rapidly to a modest single seasonal snowmelt, single heavy rain, or a somewhat drier season in run of wet years.

Many power-law relationships exhibit validity over many orders of magnitude. As suggested by Figure 4.8, the combination of aerial photography and Landsat observations for this study area provides observations from the largest lakes present (40,000 m²) to lakes about two and one-half orders of magnitude smaller (100 m²). Thus, the observed range of validity of the power law is one lake of 40,000 m² and perhaps 25,000 lakes with
an area of 100 m². If the study area was much larger, it would provide the possibility that
the single lake, the upper end, could be much larger.

With higher-resolution observational data, one might conceivably extend the power-law
to millions of small water bodies in spring. Again, there should be a practical limit
provided by the observational resolution. With an excess of water available to the surface,
the power law should be valid down to the scale of pores. Interestingly, this lower limit
would have to change as evaporation through days and weeks preferentially removed the
smallest bodies of stored water. The law in a dry summer might then only be valid to
10 m² depressions with smaller bodies having been dried up. We have theoretical studies
underway to examine the practical lower limit of the validity of power laws under
different settings.

4.5. Conclusions

This study has shown that power-law relations are useful in understanding the behavior
of several tens-of-thousands of lakes, with footprints as small as that of a family home to
as large as a small city. In the study area of South Dakota, lake areas at various times of
the year and through a number of years can always be described by a power law.
Significant annual variability in water excesses and deficiencies, as well as longer term
cycles of drought and deluge, is reflected in variability of the power-law parameters.
Tremendous differences in the volume of water stored in the different sized lakes means
that small lakes are mostly being impacted by short-term hydrological events, and large lakes are mostly being impacted by a different set of longer-term events.

There are obvious natural limits to the range of validity of a power law. For this region, the upper limit is a single lake having an area of about 40,000 m$^2$. The lower limit remains elusive, likely to be influenced by particular conditions of water on the land surface. The practical limit based on our observational data is about 25,000 lakes with a surface area of 100 m$^2$.

The next step in examining the PPR more broadly is to put together observations from climatologically and topographically diverse regions. This kind of analysis could form the basis for assessing hydrologic response of lakes through time.
CHAPTER 5
DATA MINING IN TOPEX SDR DATA: A CASE STUDY IN MONITORING INLAND WATER ZONES

5.1 Introduction

The TOPEX/POSEIDON (T/P) radar altimeter satellite mission was launched in October 1992 by the National Aeronautics and Space Administration (NASA) and the French Space Agency Centre National d’Etudes Spatiales (CNES). TOPEX provides for improved monitoring of ocean dynamics by accurately measuring global sea surface topography every 10 days with a Root Mean Square Error (RMSE) of about 2.7 cm [Zieger et al., 1991 and Fu and Cazenave, 2001]. The satellite carries a dual-frequency radar altimeter, C-band (5.3 GHz) and Ku-band (13.6 GHz). It operates at an orbit of 1334 ± 60 km mean altitude with a mean inclination of 66°. The TOPEX altimeter provides the range R (distance between the satellite and the surface of Earth) from the sensor to the sea surface based on the overall travel time of microwave, transmitted from the sensor towards the sea surface and then reflected back to the sensor.

Although the TOPEX radar altimeter was initially designed for studies of Sea Surface Height (SSH), the massive data with wide coverage from latitude 66 °S to 66 °N and the short repeat cycle of 10 days for about 15 years yields information useful for other
studies. For example, recent studies [Birkett, 1998, Birkett, 2000 and Campos et al., 2001] demonstrated the significant potential of the TOPEX altimeter in measuring the stage of inland lakes and wetlands. Besides the direct water surface measurements, TOPEX data can also be applied in examining features of land surfaces. For instance, Papa et al. [2002] used the TOPEX Altimeter $\sigma_0$ (backscattering coefficient) data to successfully estimate the depth of the snowpack over the Northern Great Plains of the United States. They produced the first continental $\sigma_0$ maps from the TOPEX altimeter and discovered interesting patterns of $\sigma_0$ values corresponding to major land-cover types globally. In addition, the TOPEX altimeter also provides great insight into rain rate estimation [Varma et al., 1999]. Hence, there are incentives to explore capabilities of the TOPEX altimeter further, especially with respect to imaging land-cover features. One of the most important hydrological applications is to estimate the proportion of water within TOPEX footprints over land surfaces. If this application is feasible, TOPEX data could be widely applied in applications to lake hydrology, land change estimation, and flood assessment, etc.

Estimating water ratios using data from the TOPEX altimeter is a difficult problem. Several major factors complicate this process. First, single values of $\sigma_0$ within a footprint are not sufficient to identify different land cover types. Second, because returned radar waveforms contain a complicated mix of information, the waveform data are inherently too complex to utilized in a simple manner, for example, as is the case with image-based data, like Landsat. Finally, differences in the pattern of radar reflectivity with various land-cover types, and interactions of signals with the terrain topography changes the
pattern of radar signals received by the sensor. This last problem of surface roughness, for example, is much less of a concern for oceans because water is not only an excellent reflector, but also ocean waves often can be idealized in terms of Gaussian distributions [Fu and Cazenave, 2001]. Thus, theoretical models for oceans are generally intractable over land surfaces. Moreover, sizes of ground footprints of the TOPEX altimeter are variable and depend on terrain.

This chapter provides an examination of modern data mining algorithms to estimate the fraction of the land surface that is comprised of water. Data mining approaches have proven extremely useful in discovering meaningful patterns or results from massive and complex datasets. For example, machine learning algorithms, such as Decision Tree and Neural Networks, have been successfully used in classification and regression with imagery [e.g., DeFries and Chan, 2000; Tatem et al., 2002; and Pal and Mather, 2003]. However, such data mining approaches are not commonly used in the analysis of TOPEX waveform data to discover the inherent nonlinear relationships. We explore applications of Regression Tree and Neural Networks in a data mining application to estimate water ratio within areas of TOPEX footprints over water and land surfaces. The specific objective is to test the possibilities of using TOPEX waveform data in land cover differentiation. Moreover, these algorithms also provide a first step to extend the application of TOPEX data for monitoring patterns as land change and understanding the hydrology of inland lakes.
5.2 Methodology

5.2.1 TOPEX data description

Data from the TOPEX altimeter is used to generate two data products: Geophysical Data Records (GDRs) and the Sensor Data Records (SDRs) in a frequency of 10 Hz. Measurements are made about every 600 m along the ground track [Remy et al., 1996 and Kruizinga, 1997]. The GDRs include the basic altimetry data for most applications in geodetics and ocean topography studies. The most important attribute of GDRs is $\sigma_0$, which is a measure of the proportion of the magnitude of reflected signals per unit solid angle to the incident waves. The SDRs contain detailed information characterizing the radar waveforms. Continuous echoes are discretized into 64 bins with different time intervals (6.25 ns or 12.5 ns) [Rodriguez and Martin, 1994]. Consequently, each waveform is represented by 64 digital numbers (DNs) in a time series. Each DN is a normalized value of reflected radar signal power within the corresponding time interval.

In general, a waveform consists of three components, the thermal noise, the leading edge, and the trailing edge [Kruizinga, 1997]. Over oceans, a returned waveform has a sharp, stable narrow peak marking the return of a strong signal from the surface of the ocean (Figure 5.1a). Over land, the radar signal from the satellite is attenuated by the ground surface, providing weak power waveforms [Birkett, 1998]. When water bodies become discrete, the returned waveforms broaden and typically contain multiple peaks, caused by reflection from different objects on the ground (Figure 5.1b,c, and d) [Koblinsky et al., 1993]. Because water is an excellent reflector of radar waves, a waveform collected over
the continents should provide information of the proportion of the land surface represented by water, even though variability in land slope and roughness within the observed footprint may complicate the analyses. Waveform data like peakiness from T/P are therefore expected to be useful in providing information on water abundance on ground tracks [Birkett, 1998].

![Figure 5.1](image1.png)

**Figure 5.1** Samples of TOPEX waveforms (10 Hz) over land surface with different water ratios within the footprint. Panel (a) is 100%, Panel (b) is 75%, Panel (c) is 25% and Panel (d) is 5%.

![Figure 5.2](image2.png)

**Figure 5.2** shows how the character of the returned signals change when the ground track moves from water to land. Interestingly, waveforms start to change over the water area.
about 6 km far from land. At this range, a portion of the footprint is responding to the presence of land. Measurements collected even closer to land along the ground track reflect a loss in the power of reflected radar signals (areas under waveform curves in Figure 5.2), particularly at the tailing portion that decays much faster compared to water signals.

Figure 5.2. Illustrative TOPEX waveforms show how the return signal behavior changes in transitioning from water to land along the ground track. On the image, water is masked.
In this study, more than 800 waveform frames on pass as 169 and 178 in the period of the TOPEX SDRs were sampled over Lake of the Woods, including Minnesota in the United States, and Ontario and Manitoba in Canada (Figure 5.3). This area was selected because (i) the land cover there is typical of agricultural and wetland areas elsewhere; (ii) Lake of the Woods is large enough to provide waveform samples of open water far from land, as well as nearby land; (iii) TOPEX has two independent ground tracks across this region, providing a large number of data samples; and (iv) the terrain of this area is fairly flat, which reduces the interference of reflected signals due to local roughness of the land surface. Waveforms were selected from cycles 294, 295, and 296 acquired in September, 2002. The timing was designed to coincide with a Landsat image, which was acquired on September 16, 2002. The late summer timeframe also avoided snow and ice conditions, which begin to develop after October and that can change the radar response.

For each set of waveforms, a circle footprint with a 6 km radius over the ground was identified on the Landsat image and water proportion in the footprint was determined using the Geographical Information System (GIS), as described in Chapter 1.
Figure 5.3. Map of the sample area, Lake of the Woods, and ground track samples (10 Hz data) of the TOPEX waveform data.
Data mining algorithms are trained based on the relationship between radar signals and known water ratio for a TOPEX measurement. The ultimate goal is eventually to estimate the water ratio in the footprint from its corresponding 10 Hz waveform. More specifically, in regression models the input variables are the 64 DNs from each frame of the waveform. The corresponding target variable is the water ratio value of each footprint estimated using GIS and the Landsat image that matches the TOPEX footprint on the ground tracks.

5.2.2 Data mining algorithms

Data mining describes a process of using predictive models to extract useful patterns from massive datasets. In recent years, this technology has developed and been widely applied in concert with the improvement in computation performance. In contrast to traditional statistical modeling, data mining algorithms attempt to develop patterns between input and target variables without parametric assumptions. Data mining algorithms are also particularly helpful in solving nonlinear problems for which linear and polynomial models are incapable of producing good predictions. In general, three steps are required to construct data mining algorithms: training, validating, and testing. Typically, the original dataset is divided into three - a training dataset, a validating dataset, and a testing dataset. Predictive models are first constructed from the training dataset. These models can be refined according to their performance with the validating dataset; and finally, the testing dataset is used to produce an assessment of the models. For a case where a large number of observations are not available, a k-fold cross-validation procedure [Ripley, 1996; Breiman et al., 1984] is deployed allowing the whole
dataset to be utilized more completely. In a k-fold cross-validation procedure, the original dataset is divided into k independent subsets, and k validations are processed to construct the best model. For each of the validations, only one subset is selected for data validation and testing, and the other k-1 subsets provide the training data necessary to build an initial model. Final predictions are made by averaging the k models created from k validations to provide the best model. Hence, all the data samples are used in training and validating procedures. In our study, because of the limited waveform sample size, k-fold cross-validation is employed to make a sufficient use of these samples in model development.

Several modeling approaches are available to estimate the proportion of water within a footprint using the reflected radar signals from TOPEX. This study tested the most popular data mining algorithms, Regression Tree and Neural Networks, to determine which was most suited to the radar data. A k-fold cross-validation procedure was used because the dataset is not large enough.

5.2.2.1 Regression Tree

In a Regression Tree, the target variables are continuous. This tree differs from a so-called Decision Tree where the target values are discrete. The Regression Tree is a recursive binary splitting algorithm that conceptually relies on a tree-like structure to minimize sample variance of the target variable by assigning samples into terminal nodes.
(leaves) along with similar target variable values [Breiman et al., 1984]. The structure of a Regression Tree model can be demonstrated as following chart:

First, the model starts with all training data samples at Node 0. All of the observational data are partitioned into either Node 1 or Node 2 based on the criteria of $X_j$ as shown in the chart. $X_j$ ($1 \leq j \leq M$) is one of the attributes or variables of data samples, assuming each data sample is an M dimensional vector, where M is the total number of data. The criteria for separation is determined by computational trials to minimize the sample variance in the target variable Y. Second, the data samples in either node are continuously split, as with Node 0, again to maximally reduce the variance of Y until terminal leaves are reached. The process is stopped at the leaves where further partitioning is incapable of reducing variance of Y further. The significance is assessed by a p value of an F test of reduction of variance of Y, as generally performed by other regression models. Thus, the leaves in a Regression Tree represent final nodes. A predicted target variable value will be assigned to each data observation within a leaf by averaging target variable values of all data samples within this terminal node.
5.2.2.2 Neural Networks

Artificial neural networks are algorithms that mimic a biological neural network structure. They are particularly helpful in solving nonlinear problems. The algorithms are implemented by linearly combining nonlinear functions of input variables to produce predictions of the target variable [Ripley, 1996]. A structure of the most commonly used type of neural networks, Multilayer Perceptron (MLP) neural networks with one hidden layer and 5 hidden nodes is given by:

For this MLP neural network, predicted target values of $Y$ are given as:

$$
g^{-1}(E(Y)) = w_0 + w_1 H_1 + w_2 H_2 + w_3 H_3 + w_4 H_4 + w_5 H_5
$$

$$
H_{i=10,5} = \tanh(w_{i0} + w_{i1} X_1 + w_{i2} X_2 + ... + w_{im} X_M)
$$

(5.1)
where $Y$ – target variable (Probability of water)

$X_{j=1..M}$ – variables or attributes of data

$w$ – weights or parameters of MLP model

$H_{i=1..S}$ – values of hidden nodes

$tanh(X_1, X_2, ..., X_M)$ – hyperbolic tangent activation functions

$g^{-1}(Y)$ – link function of target variables such as logit()

MLP neural networks use sigmoid activation functions, such as the hyperbolic tangent function of input variables, and thus can approximately simulate any continuous function if sufficient hidden nodes are used [Ripley, 1996]. $g^{-1}()$ is called the link function of the $Y$ variable, if $Y$ follows a distribution in the exponential family other than a normal distribution. In this study, $g^{-1}()$ is LOGIT($Y$), with $Y$, the water ratio value, modeled as an expected value of a binomial distributions. In this study, each Landsat pixel within the TOPEX footprint is classified as either water or land.

5.3 Results

5.3.1. Regression Tree

Three regression trees were constructed from the cross-validation as RT294 (using cycle 294 as validation and the other two cycles as training), RT 295, and RT 296. The error functions for both training and validation are shown in Figure 5.4. Using RT294 created in the 1st-fold cross-validation as an example (Figure 5.4a), the tree was continuously developed to 31 leaves, which achieved a small training error of about 0.014 but a very large validation error $> 0.10$. These results indicate a problem of severe overfitting at this point. Pruning was applied to find the optimal tree satisfying my model decision criteria,
namely the minimization of the validation error. The dashed line (Figure 5.4a) shows that the smallest validation error, approximately 0.071 is associated with a tree having eight leaves. However, both the training and validation errors hardly change as the tree grows from four leaves to eight leaves. Thus, the smaller tree with only four leaves is essentially the same as the optimal tree with eight leaves. To provide for the most robust regression tree, this simpler version with four leaves is determined as the model from this validation run. The structure of this model is shown in panel (a) of Figure 5.5. Also shown in Figure 5.5 are two other trees built from runs RT295 (Panel b) and RT 296 (Panel c). Figure 5.5 also shows rules applied to the nodes of these three regression trees. Using RT294 as an example, if a waveform frame has a DN53 value less than 30.5, and a DN57 value less than 16.5, this frame is assigned into the left most leaf of RT294 (Figure 5.5a) with a water ratio value estimated as 0.136. For each leaf shown in Figure 5.5, two statistics (number of observations and average water ratio value) are presented for both training and validation datasets. In fact, the training average water ratio value is also the model prediction, and the validation value for the average water ratio value gives a sense of the goodness of the model fit. Generally, a small difference between training and validation average values indicates goodness and consistency in model performance across different datasets. In Figure 5.5, most of the leaves have comparable average water ratios in the training and validation datasets. An exception is the largest gap of 0.23 with a leaf of RT295. However, both these cases have a small number of samples, 14 and 6 from the training and validation data sets, respectively. Overall, there is no prominent overfitting problem with any of the regression trees.
The goodness of fit for these regression trees is shown in Table 5.1. Each of the trees achieves a very high training $R^2$ value above 0.70, which means the model fits training data very well and explains more than 70% of the data variance in water proportion. The training average least squared errors are controlled within a 0.05 level. The validation data yield similar results to training data, except for RT294. The $R^2$ value for validation data is only 0.53, significantly smaller than the training $R^2$ value of 0.78, and the validating average square error is 0.0709, much higher than 0.05. However, in general, these regression trees performed well in terms of the fit of both training and validation datasets. With RT294, the inconsistency between training and validation results is compromised by averaging the three trees in an ensemble model.
Figure 5.4. Error functions of regression trees from 3-fold cross-validation
* “N” indicates the number of data samples; “Average” indicates the average water ratio value, where the left and right numerical values are respectively for training data and validating data.

Figure 5.5. Structures of regression trees from 3-fold cross-validation
<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT294</td>
<td>Train</td>
<td>N</td>
<td>509</td>
</tr>
<tr>
<td></td>
<td>Train</td>
<td>Ave Sq Err</td>
<td>0.0336</td>
</tr>
<tr>
<td></td>
<td>Train</td>
<td>R Squared</td>
<td>0.7830</td>
</tr>
<tr>
<td></td>
<td>Valid</td>
<td>N</td>
<td>238</td>
</tr>
<tr>
<td></td>
<td>Valid</td>
<td>Ave Sq Err</td>
<td>0.0709</td>
</tr>
<tr>
<td></td>
<td>Valid</td>
<td>R Squared</td>
<td>0.5272</td>
</tr>
<tr>
<td>RT295</td>
<td>Train</td>
<td>N</td>
<td>497</td>
</tr>
<tr>
<td></td>
<td>Train</td>
<td>Ave Sq Err</td>
<td>0.0429</td>
</tr>
<tr>
<td></td>
<td>Train</td>
<td>R Squared</td>
<td>0.7169</td>
</tr>
<tr>
<td></td>
<td>Valid</td>
<td>N</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>Valid</td>
<td>Ave Sq Err</td>
<td>0.0477</td>
</tr>
<tr>
<td></td>
<td>Valid</td>
<td>R Squared</td>
<td>0.6952</td>
</tr>
<tr>
<td>RT296</td>
<td>Train</td>
<td>N</td>
<td>488</td>
</tr>
<tr>
<td></td>
<td>Train</td>
<td>Ave Sq Err</td>
<td>0.0422</td>
</tr>
<tr>
<td></td>
<td>Train</td>
<td>R Squared</td>
<td>0.7248</td>
</tr>
<tr>
<td></td>
<td>Valid</td>
<td>N</td>
<td>259</td>
</tr>
<tr>
<td></td>
<td>Valid</td>
<td>Ave Sq Err</td>
<td>0.0356</td>
</tr>
<tr>
<td></td>
<td>Valid</td>
<td>R Squared</td>
<td>0.7682</td>
</tr>
</tbody>
</table>

* "N" means the number of data samples.

Table 5.1 Results of regression trees

One of the significant advantages of a tree model is the automatic variable selection process, which otherwise is conducted independently using other algorithms, such as neural networks. Figure 5.5 implies that only four variables are involved in each regression tree model regardless the 64 dimensions of the original input data space. In addition to the considerable dimension reduction in the input variables, the tree algorithm also computes an importance value for each of the input variables and ranks them to measure the usefulness of each variable in building the model. Importance values of variables selected with the regression trees (Table 5.2) are actually $\chi^2$ statistics normalized to 1. The $\chi^2$ test is conducted similar to the linear regression model for each variable to determine the significance of adding the variable into the model. Based on the
results from Table 5.2, DN53 and DN51 exhibit importance values significantly greater than other variables, and are the most important variables in the regression tree models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Variable</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT294</td>
<td>DN53</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>DN62</td>
<td>0.2383</td>
</tr>
<tr>
<td></td>
<td>DN57</td>
<td>0.2156</td>
</tr>
<tr>
<td></td>
<td>DN18</td>
<td>0.0532</td>
</tr>
<tr>
<td>RT295</td>
<td>DN51</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>DN57</td>
<td>0.3083</td>
</tr>
<tr>
<td></td>
<td>DN56</td>
<td>0.2514</td>
</tr>
<tr>
<td></td>
<td>DN50</td>
<td>0.1004</td>
</tr>
<tr>
<td>RT296</td>
<td>DN51</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>DN57</td>
<td>0.3265</td>
</tr>
<tr>
<td></td>
<td>DN62</td>
<td>0.1991</td>
</tr>
<tr>
<td></td>
<td>DN54</td>
<td>0.1823</td>
</tr>
</tbody>
</table>

Table 5.2 selected variables of regression trees

5.3.2. Neural Networks

A three-fold cross-validation method defined by three cycles - 294, 295 and 296 as before - was employed to incorporate all the data samples in training and validation of neural networks. Unlike applying regression trees directly to the data, a variable selection process is conducted in order to reduce the data dimension and computation efforts excessively costly for weight optimization before fitting the neural networks. After selection, only DNs, which are significant in the $\chi^2$ test with the target variable – water ratio - are retained as inputs in neural networks.
Three neural networks were built from the three-fold cross-validation procedure. In each validation process, the neural network started with weights assigned by initial values. A Quasi-Newton optimization algorithm [Ripley, 1996] was used to update the weights iteratively to obtain the minimum of the error function. This algorithm is one of the most robust with the fastest rate of convergence in artificial neural computation, using only first derivatives instead of second partial derivatives of error surfaces.

To avoid model overfitting due to noise in the training data, iterations were terminated when the model reached the minimum error function of validating data. The neural network with weights computed in the last iteration was selected as a fitting model. Figure 5.6 shows examples of the training and validation error functions for the three-fold cross-validation. In the first plot, the error functions are from first-fold cross-validation using cycle 294 as validating and the other two cycles as training. Although the training error function was minimized with iterations continuously running, the smallest validating error (less than 0.1) was obtained at the 12th iteration. Therefore, the neural network constructed at the 12th iteration was selected as a fitted model in the first-fold cross-validation.
Figure 5.6. Error functions of neural networks from 3-fold cross-validation
The statistics of fitting three neural networks are displayed in Table 5.3. The final prediction errors range overall from 0.05 to 0.07. The neural networks produce very accurate predictions of water ratio values, even though some of the outliers are shown in both the training and validation processes by Maximum Absolute Errors. Also notice that the goodness of fit is slightly dependent on data splitting. For example, the largest Root of Average Squared Error (Root ASE) with the training data is 0.2062 by the second-fold validation, whereas the first-fold validation produced the largest validation Root ASE of 0.2484.

<table>
<thead>
<tr>
<th>Model</th>
<th>Fit Statistic</th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN294</td>
<td>Root Average Squared Error</td>
<td>0.2012</td>
<td>0.2484</td>
</tr>
<tr>
<td></td>
<td>Error Function</td>
<td>-110.7383</td>
<td>15.2114</td>
</tr>
<tr>
<td></td>
<td>Maximum Absolute Error</td>
<td>0.7163</td>
<td>0.6418</td>
</tr>
<tr>
<td></td>
<td>Final Prediction Error</td>
<td>0.0620</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>Akaike's Information Criterion</td>
<td>-7.4766</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>Schwarz's Baysian Criterion</td>
<td>445.3954</td>
<td>.</td>
</tr>
<tr>
<td>NN295</td>
<td>Root Average Squared Error</td>
<td>0.2062</td>
<td>0.1981</td>
</tr>
<tr>
<td></td>
<td>Error Function</td>
<td>-104.9095</td>
<td>-60.1769</td>
</tr>
<tr>
<td></td>
<td>Maximum Absolute Error</td>
<td>0.6759</td>
<td>0.8636</td>
</tr>
<tr>
<td></td>
<td>Final Prediction Error</td>
<td>0.0687</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>Akaike's Information Criterion</td>
<td>24.1810</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>Schwarz's Baysian Criterion</td>
<td>516.5860</td>
<td>.</td>
</tr>
<tr>
<td>NN296</td>
<td>Root Average Squared Error</td>
<td>0.1849</td>
<td>0.1639</td>
</tr>
<tr>
<td></td>
<td>Error Function</td>
<td>-176.0879</td>
<td>-114.5490</td>
</tr>
<tr>
<td></td>
<td>Maximum Absolute Error</td>
<td>0.6868</td>
<td>0.6626</td>
</tr>
<tr>
<td></td>
<td>Final Prediction Error</td>
<td>0.0534</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>Akaike's Information Criterion</td>
<td>-138.1758</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>Schwarz's Baysian Criterion</td>
<td>310.1880</td>
<td>.</td>
</tr>
</tbody>
</table>

Table 5.3 Statistics with a neural networks model based on a three-fold cross-validation procedure.

5.3.3. Comparison of ensemble models of regression tree and neural networks
Because the manner in which the original data is split into training and validation subsets has an effect on model building, I created an ensemble model. It combines three individual models from the three-fold cross-validation as the overall regression model for both the regression tree and neural networks algorithms. Predicted values of water ratio from these three individual models were averaged in the ensemble model as the final predictions. This averaging procedure uses all of the data records in building the final predictive models and makes the model more robust in terms of problems of data splitting and noise related to overfitting of particular subsets. The fitting errors for ensemble models are shown in Table 5.4. Here, an ensemble model provides a comprised value by averaging predictions from three validations, which is close to the optimal one among those individual models. From Table 5.4, the ensemble model for each algorithm produced good predictions of water ratio with Root ASE smaller than 0.20. The values of Root ASE of the two ensemble models are very similar with differences less than 0.01. The neural networks provide better predictions, but are not significantly better as compared to the regression tree approach.

<table>
<thead>
<tr>
<th>Ensemble model</th>
<th>Train: Root ASE</th>
<th>Valid: Root ASE</th>
<th>Test: Root ASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural networks</td>
<td>0.1984</td>
<td>0.1609</td>
<td>0.1862</td>
</tr>
<tr>
<td>Regression tree</td>
<td>0.2097</td>
<td>0.1621</td>
<td>0.1945</td>
</tr>
</tbody>
</table>

Table 5.4 Comparison between the ensemble trees and the ensemble neural networks

To examine the performance of models in detail by relating predictions to their real values, residuals from ensemble models were plotted against the predicted values (Figure 5.7). These residuals are differences between model predictions and the real water ratios.
estimated from Landsat images. From the figure, prediction errors are controlled within
the ± 0.2 range for footprints where water ratios are greater than 0.6. Residuals are more
scattered for smaller water ratios. More outliers are present beyond the 0.2 bounds when
predictions are between 0.2 - 0.6, which means errors are larger for such predictions.

(a) Ensemble Regression Trees

![Graph showing residual plots vs. predicted water ratio for ensemble regression trees.]

(b) Ensemble Neural Networks

![Graph showing residual plots vs. predicted water ratio for ensemble neural networks.]

Figure 5.7 Residual plots vs. predicted water ratio of the (a) ensemble trees and the (b)
ensemble neural networks
5.4 Discussion

Neural networks perform well as a regression algorithm, especially for interpreting water ratio as a complex, nonlinear function of radar signals. Predictions using neural networks typically are more accurate than other methods used to explore and interpret these data, because neural networks use combinations of nonlinear function surfaces of predictors to approximate the target variable function surface. But one of the biggest drawbacks with neural networks is that the global minimum of the error function surface (the gap between the true target function surface and the predicted target function surface) is not guaranteed, despite using different optimization algorithms (like Quasi-Newton, Back-Propagation). Generally, the gradient descent process to minimize the error function surface is trapped into local minima. In this case, the best set of weights from cross-validation might not be generalized for the overall data. Data splitting is another factor producing inconsistency in the results from the neural networks. As shown in Table 5.3, predictive errors vary with different fold validation runs. This variability across different data combinations is another evidence that global optimization is still a challenge to artificial neural computation. Therefore, if computation cost, model simplicity, and ease in explanation are taken into account, regression tree outperforms neural networks by providing very close predictions without demanding extra efforts in optimization and interpretation.

Unlike the black-box neural networks, regression tree can be intuitively explained by its structure. As shown in Table 5.2, input variables DN51, DN53, DN56 and DN57 are
most related to water ratio predictions based on $\chi^2$ statistics. The waveform plots in Figure 5.1 indicate that the first peaks in the radar echoes come out at around the 25th bin. These may be useful in estimating the elevation of water surfaces, but are not very useful in determining the water ratio. The tails of radar echoes from DN51 to DN60 represent decayed signals reflected back to the satellite. The flatness and high reflectivity of the water surface could make echoes have tailings decaying very smoothly and slowly from the first spike, producing significantly high signals in the tails. In contrast, reflected waveforms from land surface have very low tailing parts in spite of multiple peaks, because the interference and scattering related to the roughness of the land surface dramatically weaken the later segments of reflected signals. Therefore, the trailing portions of waveforms provide information useful in distinguishing between land and water.

The errors from the ensemble tree model are plotted over the study area in Minnesota (Figure 5.8). It shows that most of the absolute errors over the lake area are less than 0.05, but are enlarged to more than 0.25 over land. The predictions over the boundary of the lake are deteriorated by errors worse than 0.5. This indicates that predicted water ratio values are more accurate for footprints dominated by water and less accurate for a footprint containing a mixture with land. This is because the mixture with water and land complicated echoes back to the TOPEX sensor and makes pattern recognition difficult. This can be explained in two different ways. First, radar waves are reflected much better from water than land. Accordingly, waveforms returning from water contain more
signals than noises, while echoes from land surfaces most likely were either weakened or scattered by rough surfaces or vegetation, making it much more difficult to separate noise and signals. Secondly, on board TOPEX data processing was initially designed for ocean
surfaces, the calibration of this process had not taken into account the echoes from land. As a result, more uncertainties were associated with land data.

In summary, the TOPEX SDR data are capable of estimating water extent within the corresponding footprints by the aid of data mining algorithms. The importance is strengthened in hydrological studies of inland lakes and wetlands, such as the Great Lakes and the Amazon basin, particularly when retrospective investigations during the last decade are needed and other remotely sensed data are absent. With the relatively large uncertainties in predictions for areas with less than 60% water, this application currently has limitations for small wetlands. In addition, under more complicated conditions where terrain changes sharply or water areas are extremely discretely distributed, such as the Prairie Pothole Region of North America, further investigations are necessary to use the TOPEX SDRs to monitor inland water hydrology or land changes. Because under such situations, water areas on the ground are not large enough to reflect radar signals back to the satellite to create significant tailing parts of the TOPEX SDRs. This is shown by Figure 5.9. Even over relatively large lakes (like Devils Lake in North Dakota), the radar signal reflected back from the ground surface is still insignificant in the tailings. For small wetlands and lakes, the vegetation and slopes and relief of the land surface may make this recognition worse.
5.5 Conclusions

The TOPEX SDRs data are shown to be useful in monitoring inland water zones by using data mining predictive models to estimate the water extent of the TOPEX footprints. Both neural networks and regression tree algorithms perform well in predicting water ratio values from the tailing portion of corresponding radar waveforms. In conclusion, regression tree is selected as the final solution based on its simplicity in computation and interpretation, as well as the similar predictions compared with those from neural works. The errors are well controlled within a range of -0.2 to 0.2 when water ratio is greater than 0.6, however, there erros are amplified over land areas where water proportion is extremely low. The mixture of land and water within a TOPEX footprint make it particularly difficult to estimate the water ratio especially when the tailing part of the waveform is unrecognizable. This prohibits a broad application to monitor small wetlands and lakes without further exploration.
CHAPTER 6

CONCLUSIONS

Data mining, GIS, and remote sensing can play a critical role in monitoring wetland hydrology and spatial and temporal changes by providing a broad view of a large region in multiple time windows. Combined with multidisciplinary approaches including CART, GIS and remote sensing, small wetlands can also be well recognized and assessed, which is very useful in the Prairie Pothole Region of North America. The error of area estimation for lakes greater than 0.5 ha is controlled to less than 20%, and smaller than 10% when lakes are greater than 1 ha.

Prairie potholes are very important wetlands and unique ecosystems that are valuable in many aspects. Climate is vital with respect to maintaining the hydrology of these wetlands. However, lakes and wetlands in different sizes react differently to climate with regard to the time duration. Based on the accurate evaluation of total area and count using data mining, GIS and remote sensing, wetlands and lakes within the study site – the Prairie Coteau Region of South Dakota - exhibited significant variations during the transition from the extreme drought in 1990 to the greatest deluge in 1997 dominated by climate. Small wetlands were mostly impacted by the six-month average ET, whereas large lakes were more related to the four-year period PDSI. The correlation coefficient of
the area and number of large lakes to the 48-month PDSI is greater than 0.93, indicating that this comprehensive climatic index is an excellent predictor for changes of large lakes. However, small lakes are more ephemeral, and the highest correlation coefficients are with ET (0.68). Results also show that the largest lakes are mostly related to the longer duration of PDSI, and reversely, the smaller lakes are mostly related to the shorter period of ET. The emergence of a large number of lakes and wetlands in the spring implies the importance of the snowmelt at this time of the year. Area and number appear to be dramatically reduced in late autumn due to the evaporative losses throughout the summer. Consequently, under the large scenario of global warming, the population of waterfowl will be severely influenced because these small wetlands are more important to hosting ducks and birds.

The size distributions of lakes and wetlands within the region greatly follow power-law function. Even under distinct hydrological conditions in different years, based on the simulation outcomes of Bayesian models, the power-law relationship still maintained a constant slope in similar seasons. This provides a great opportunity to assess lakes and wetlands within the region during the “Dust Bowl” period from the limited number of aerial photos taken in 1939. Comparisons among year 1939, 1990, 1992, and 1997 reveal the extremely different situations of lakes and wetlands controlled by climate variability. For those lakes larger than 1 ha, the number of lakes and wetlands in 1939 is only 1/2 to 1/3 of those in 1990 to 1992, and less than 1/8 of that in 1997. Seasonal variations of power-law lines in lake size distributions also show different rate of change of lakes and wetlands for different sizes. Small wetlands are mostly influenced by high ET rates in
summer and the numbers of lakes and wetlands considerably decreased in August, this resulted in flat distributions of those lakes, indicated that the large lakes are not as much impacted and the numbers are fairly constant.

Over 15 years of operation, the TOPEX radar altimeter has provided a tremendously huge volume of data of GDRs as well as SDRs with a global coverage. By the aid of modern data mining technologys, these SDRs have potential in quantifying the water content of the TOPEX footprint over the land surface. This significantly broadens the applications of the TOPEX radar altimeter to various studies, such as land cover and land change, inland water hydrology, and flood monitoring. Two regression algorithms, regression tree and neural networks, were tested by a k-fold cross-validation method in water ratio estimation, and the regression tree is selected based on overall performance with regard to computation cost, model simplicity and prediction accuracy. Using regression tree as predictive models, the water ratio estimation can be accurate with errors within ± 20% for areas containing 60% of water or more. However, to produce more accurate predictions of water ratio for footprints where land is the dominant component is still a big challenge requiring further exploration, not only in improving data mining algorithms, but also in including terrain models and land cover classification.

In general, this study thoroughly explores the changes of prairie potholes in area and count during the extreme hydrological variations within the last century. The study also provided a means to discover their relationships with climate variability by the aid of data mining, GIS and remote sensing. We also demonstrate the promising potential of the
TOPEX waveforms with respect to inland water assessment. This research work is a significant contribution in such fields.

Our future research plan based on current findings can include: (1) collect more satellite data involving diverse sensors and platforms, such as SPOT, QUICKBIRD, MODIS, to enhance the remote sensing research capability in both spatial and spectral aspects, and focus on the most recent time periods to track the latest changes of lakes in the PPR; (2) the image classification method will be improved by considering many other factors, like atmosphere, location, and more specific ground features, to refine our sub-pixel method to increase the accuracy; (3) more meteorological data and climatic indices, such as Standardized Precipitation Index (SPI) and Palmer Hydrological Drought Index (PHDI), will be included into the relationships to the numbers and areas of lakes, and hydrological models will be developed to simulate such changes using geological and meteorological data; and (4) more testing sites at different scales will be sampled to enhance the patterns of power-law relationships in lake size distribution. Imagery data from different platforms within multiple time frames will be used to strengthen the findings of the consistency and variation of the power-law linear functions regarding annual and seasonal changes respectively. To overcome the problems caused by the complexity of TOPEX SDRs from ground cover with a small water fraction value, terrain information such as DEMs, as well as land cover, will be included as predictors in the water ratio regression model to extend the TOPEX water ratio prediction into wider areas. Our final goal is to apply this methodology to the Prairie Pothole wetlands and boreal forest areas in North America.
BIBLIOGRAPHY


Conly, F. M. and G. Van Der Kamp 2001. Monitoring the hydrology of Canadian prairie wetlands to detect the effects of climate change and land use changes, Environmental Monitoring and Assessment 67: 195–215,


Kruizinga, G. “Validation and application of satellite radar altimetry,” PhD Dissertation, Univ. of Texas at Austin, December, 1997


Lehner, B. and P. Doll (2004), Development and validation of a global database of lakes, reservoirs and wetlands


Pietz, P. J., G. L. Krapu, D. A. Buhl, and D. A. Brandt, 2000 Effects of water conditions on clutch size, egg volume, and hatchling mass of mallards and gadwalls in the prairie pothole region, The Condor 102:936–940


North Dakota in the Perspective of longer term measured and proxy hydrological records, *Climate Change*, 40 (2): 189-209.


