Summer 2011

Water quality modeling and rainfall estimation: a data driven approach

Evan Phillips Roz
University of Iowa

WATER QUALITY MODELING AND RAINFALL ESTIMATION: A DATA DRIVEN APPROACH

by

Evan Phillips Roz

A thesis submitted in partial fulfillment
of the requirements for the Master of Science
degree in Industrial Engineering
in the Graduate College of
The University of Iowa

July 2011

Thesis Supervisor: Professor Andrew Kusiak

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

MASTER'S THESIS

This is to certify that the Master's thesis of

Evan Phillips Roz

has been approved by the Examining Committee
for the thesis requirement for the Master of Science degree
in Industrial Engineering at the July 2011 graduation.

Thesis Committee: _____
Andrew Kusiak, Thesis Supervisor

_____
Yong Chen

_____
Marian Muste

To My Family

Stay hungry, stay foolish.

Steve Jobs

ACKNOWLEDGMENTS

I would like to thank my graduate advisor, Professor Andrew Kusiak, for offering me the opportunity to work in his laboratory. The attention to detail that he instilled onto his understudies will surely benefit us in our future work. I would also like to acknowledge Professors Marian Muste, Jerry Schnoor, David Bennett, and Nandita Basu as well as my colleagues of the Cyber-enabled Discovery and Innovation (CDI) team. The CDI team meetings and discussions were instrumental for inspiring new research ideas and provided me the guidance throughout my two year study with the group..

I would like to thank Professors Yong Chen and Professor Marian Muste for serving on my Thesis Committee. I am also grateful for the financial support from the National Science Foundation, who provided the financial support which made my research opportunity possible.

I thank all the members of the Intelligent Systems Laboratory (ISL) who have worked with me and provided me with advice.

Without the absolute support of my family to continue further studies, my work here at the University of Iowa would not be possible. The unconditional support I received from my parents, Jonathan Roz and Patricia Phillips, my brother Ethan Manning and sister-in-law Denise Marcello Manning, and my closest friends from Albany, NY helped to keep my spirit strong, and provided me with moral support throughout my time in Iowa.

ABSTRACT

Water is vital to man and its quality it a serious topic of concern. Addressing sustainability issues requires new understanding of water quality and water transport. Past research in hydrology has focused primarily on physics-based models to explain hydrological transport and water quality processes. The widespread use of in situ hydrological instrumentation has provided researchers a wealth of data to use for analysis and therefore use of data mining for data-driven modeling is warranted. In fact, this relatively new field of hydroinformatics makes use of the vast data collection and communication networks that are prevalent in the field of hydrology.

In this Thesis, a data-driven approach for analyzing water quality is introduced. Improvements in the data collection of information system allow collection of large volumes of data. Although improvements in data collection systems have given researchers sufficient information about various systems, they must be used in conjunction with novel data-mining algorithms to build models and recognize patterns in large data sets. Since the mid 1990's, data mining has been successful used for model extraction and describing various phenomena of interest.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1.

INTRODUCTION

The availability of quality water is a concern, and human-environment interactions still leave much to be understood. Knowledge about water transport, quality, and quantity awaits further discovery. Water quality has high variance from location to location and time to time, due to its sensitivity to both chemistry (i.e. nutrient loading), and transport (i.e. stream flow). Both human activity such as the application of fertilizers and land management practices, and meteorology play a strong role in water quality.

Accurate water quality prediction would provide us with a better understanding of the human influence on aquatic life and provide knowledge for intelligent decision making in regards to ecological conservation. In the past, physics and chemistry-based models were used to model water quality and the transport of nutrients. Due to sources of error in measurement, misunderstanding of hydrological systems, and errors in modeling building/approximation, the results of such models leaves much for improvement. Data driven techniques can eliminate some of these sources of error because they do not require a strong physical understanding of the system to be modeled. Data is used *directly* for model building, not for validation of a theoretical physical concept.

1.1. Physics-based modeling approaches in water quality

Numerous efforts to model water quality have been made by research communities such as the United States Geological Survey (USGS) and the United States Department of Agriculture (USDA). For example, the Soil Water Assessment Tool (SWAT) model was developed by the USDA's Agricultural Research Service (ARS) to predict the impact of land management practices on water and agriculture resources [2]. The Environmental Protection Agency (EPA) and the United States Department of Agriculture (USDA) are active in building physical models to tackle water quality estimation and prediction. Some of the EPA's products are AQUATOX, a freshwater ecosystem model, CORMIX, a hydrological mixing model, and QUAL2K, river and stream water quality models. It is difficult to separate water quality from hydrology, as it is as strongly impacted by water transport as it is water chemistry. For this reason, water quality models are often derived from the classical Navier-Stokes equations of fluid dynamics [3,4].

With the recent deployment of in situ instrumentation in rivers, streams, and creeks nationwide, as well as real-time data reporting via satellite communication technology, a wealth of data is available that had never before in the past. Data mining can utilize this vast base of data for pattern recognition and machine learning, so as to make accurate predictions.

1.2. Data-driven modeling approaches in water quality

Data mining makes models from the "ground up" rather than using the traditional top-down approach of its physics-based counterpart. As data-driven models are derived directly from the data, their accuracy is unparalleled by physics-based models. Several data-driven methods have been used, such as fuzzy logic [5, 6] for lake eutrophication modeling, support vector machine regression for hydrological model approximation [7] and river discharge modeling [8], with the MLP, also known as the neural network (NN) being the most preferred.

Previous work has applied MLPs to both water quality and water quantity. Palani *et al.* (2008) applied neural networks (NNs) to model seawater temperature, salinity, and chlorophyll (Chl-a) concentrations at time *t* and also forecast seawater temperature one week ahead [9]. Sérodes *et al*. (2001) forecasted the residual chlorine in drinking water of the city of Sainte-Foy, using the parameters chlorine dosage, residual chlorine concentrations, temperature, and flow rate [10]. Sahoo used regression analysis, NNs, and Chaotic Non-Linear Dynamic Models to forecast stream water temperature [11]. The ANN has also been used to approximate the output of a hydrological model, the Soil and Water Assessment Tool (SWAT) [12].

Data driven approaches have also been applied to water quantity and flood prediction. Choy and Chan used a support vector neural network (SVNN) was used to predict the discharge of the Fuji River in Japan, so as to provide an early warning for inhabitants in the case of a heavy rain event [8]. Damle and Yalcin (2007), in their study of a St. Louis gauging station on the Mississippi River focused on daily discharge values

from 1933 to 2003 for discharge prediction [13]. Dibike and Solomatine forecasted river flow and modeled the looped rating curve by way of NNs [14]. Willby *et al.* attempted to extract knowledge of the physical system from NNs built to model rainfall-runoff [15].

In order to achieve high accuracy water quantity estimation, high spatiotemporal resolution precipitation data is highly desirable. There have been a few efforts to utilize data-driven modeling for precipitation estimation via NEXRAD radar data. There have been fewer attempts to make this link between radar data and tipping bucket data with data-driven techniques. Feed forward neural network (FFNN) have applied for rainfall estimation using radar reflectivity and rain gauge data [16,17]. Trafalis *et al*. considered some different parameters, such as wind speed and bandwidth to complement reflectivity, but with unimproved results. The best performing models in the study all had MSE's less than 0.1mm/hr [18]. Liu *et al*. (built a recursive NN with a radial basis function (RBF) that would continuously update its training data set with time. The architecture of such neural networks has also been experimented with improved results [19].

1.3. The multilayer perceptron (MLP)

As the algorithm used throughout this Thesis is the multilayer perceptron (MLP), otherwise known as neural network (NN) or artificial neural network (ANN), an in depth algorithm description is justified. It has found widespread success in many areas other than hydrology due to its ability to model noisy data and usefulness for both classification and regression. This section should provide insight to one of the machine learning algorithms that has been so widely labeled a "black box" model.

1.3.1. MLP overview

The MLPs applied in this research are feed forward backwardly propagating neural networks. The MLP's structure consists of nodes in an input layer, a hidden layer(s), and an output layer. The concept was biologically inspired to represent the human brain's ability to process in parallel, to learn from experience, and to be highly connective and modifiable. The brain also operates via supervised learning, or the ability to train itself and learn from past experiences. The brain has the ability both to feed connections forward, near sensory input, and feed connections backwards near sensory input. These connections are mimicked by the NN with the use of loops. Feed forward NNs do not have loops, while in a looping, or recurrent NN, information is fed back from an output node to an input node [20-22]. In both categories of NNs, each input/output parameter is assigned a node in its respective input/output layer.

1.3.2. The MLP structure and algorithm

Figure 1.1 is a diagram of a single perceptron with two inputs**,** and a simple binary output. The inputs are multiplied by their respective weights and the products are summed at the junction. If the sum at the junction is greater than the threshold (Θ), the perceptron "fires." In the binary example, firing means outputting a "1." Equation (1.1) describes the summation that occurs at the node.

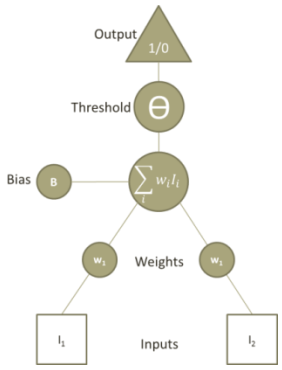

Figure 1.1. Perceptron

$$y_j = f\left(\sum_{i=1}^{m} x_i w_{ji} - b\right) \qquad (1.1)$$

Where y_j is the output of the j^{th} node, m is the number of inputs to the j^{th} node, x is the input value, w is the input weight, and b is a bias factor.

After each element in the data set, the weights for the inputs are updated, based on error. If the target value was achieved, the weights remain unchanged. Equation (3) describes how the neural network updates the j^{th} weight in the i^{th} layer.

$$w'_{ji} = w_{ji} + \alpha\delta\frac{df_1(e)}{de} \qquad (1.2)$$

Where α is the learning rate, δ is the error attributed to the node and f is the activation function.

It is this recalculating of the weights that allows the neural network to "learn" a dataset. Stopping criteria is user defined, usually by limiting the number of epochs, or cycles through the data set, the model continues. The original perceptron was developed by Rosenblatt (1958) in at the Cornell Aeronautical Laboratory**,** but the observation was made that the single layer perceptron was only capable of learning when the data set was linearly separable, such as modeling the XOR gate [23]. However, after further development, multiple perceptrons was placed in layers (see figure 1.2), and the simple stepwise activation function was replaced with a continuous and differentiable sigmoidal one, so that its outputs could be continuous. The resulting structure of the perceptron when put into layers, can be seen below in the Figure 2 which is an MLP schematic with two hidden layers and 15 nodes. An example of the new sigmoidal activation function for continuous MLPs, in this case the logistic function, is show in equation 1.1.

$$f(x) = \frac{1}{1 + e^{-z}} \qquad (1.3)$$

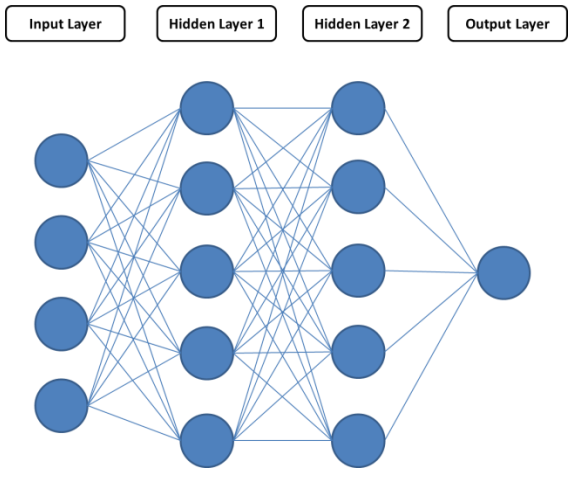

Figure 1.2. Multilayer perceptron

The optimal structure of a NN still remains a trial-and-error process, but there are several rules of thumb that previous researchers have found useful. For example, Tarassenko (1998) states that the number of samples in the training set should be greater than the number of synaptic weights in the network, and according to Hecht-Nielsen (1987) the number of hidden nodes, *M*, in a single hidden layer model NN is between *I* and 2*I* +1, where *I* is the number of input nodes [24, 25]. Data-mining software, such as Statistica or WEKA can be a useful tool for testing multiple NN structures to find optimal results [26].

CHAPTER 2.

TWO DATA MINING APPROACHES FOR FILLING MISSING DATA

2.1. Introduction

Dissolved oxygen (DO) concentration serves as a benchmark for measuring the ecological health of aquatic systems. Environmentalists, namely the Department of Natural Resources (DNR) have established a minimum DO concentration value of $5mgL^{-1}$ as a minimum threshold for water quality [38]. Hydrological data is notoriously erroneous and often contains missing values, as the instruments used in data collection are outdoors and must withstand the elements, making them susceptible to debris and organic matter which may cause instrument failure.

There have been several models developed to explain the processes of dissolved oxygen (DO), a primary water quality indicator. The most popular of which being based on the classical Streeter and Phelps (1925) equations to simulate changes in DO concentrations over distance. These changes in DO are due to the complex and highly nonlinear deoxygenation and reaeration systems that are active in aquatic environments [1].

Water quality data of 20-minute resolution obtained from the Clear Creek basin of Johnson County, Iowa is used to build multilayer perceptron networks for the modeling and forecasting the dissolved oxygen concentrations at a downstream gaging station. Two methods for estimating missing water quality data, a condition that has troubled many scientists in the field of hydrology, are considered. Thus, two types of models are

built for data estimation; (Type-1) a model that uses *other* water quality data to estimate dissolved oxygen concentration and (Type-2) another that utilizes time series data mining techniques and dissolved oxygen memory parameters for current dissolved oxygen estimation.

Aside from targeting such a meaningful water quality assessment parameter, this chapter acts as an application of data mining for left-handed data estimation, which is generalizable for water quality measurements other than just DO. In other words, a left-handed data estimate is one that acts on only past, or historical, data, rather than a considering two-sided approach that uses future data (i.e. hindcasting) [39]. Such a left-handed model is compatible with hydrological models that run in real-time. An accurate Type-1 model, as described in the abstract section, may substitute a DO sensor altogether or provide information where DO data is not available, but other water quality data is.

2.2. Data description

DO concentration is known to be dependent on several factors, such as temperature, pH time of day, solar radiation, and as mentioned above, the presence of aquatic biotic. It is noted in the literature that these factors are correlated [40]. For example, solar radiation is strongly related to aquatic diurnal cycles, which both affect temperature, and aquatic life is sensitive to the pH of a water body. The deployment of in situ water quality instrumentation has given scientists the ability to observe water chemistry phenomena at a sub-daily time scale. The observation of the diel cycles of DO concentration and dissolved inorganic carbon are two products of this new capability

[41]. Chapra and Di Toro (1991) found that decreased DO concentrations often happen at night in aquatic environments because of the respiration processes of large phytoplankton communities [42]. Also, DO and pH are inversely related. High DO concentrations are associated with acid mixed liquor and nitrification while low DO is associated with high pH and a non-nitrifying, or even denitrifying system state [43]. Finally, as no solar radiation data available for the study area, the hour of the day will be used as an input parameter for this study. Obviously, higher solar radiation values are observed during the midday hours.

Nearly six months of water quality data, from 3/22/2006 to 9/12/2006, was collected from the HIS from the South Amana gauge location in the Clear Creek basin in Johnson County, Iowa (lat/lon N31.736 W91.931). A map of the study area is displayed in Figure 2.3, where each square is approximately 1 km^2.

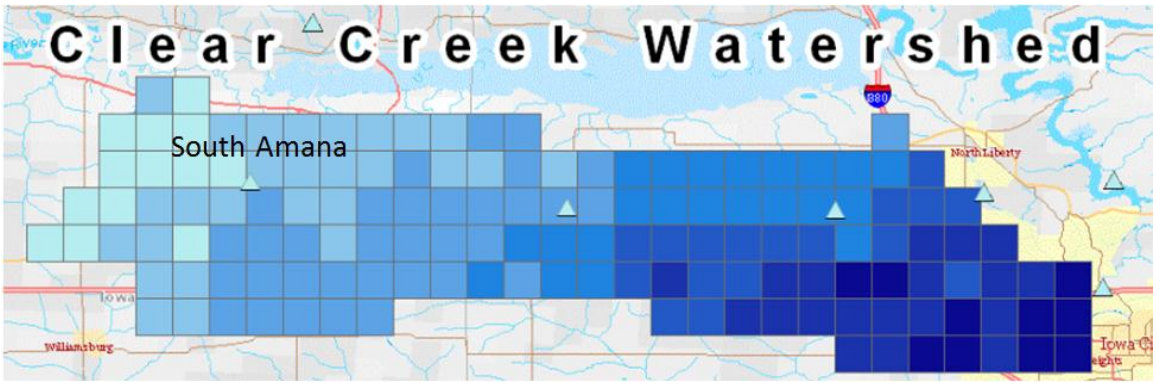

Figure 2.1 Map of the Clear Creek Digital Watershed

The sampling frequency of the data set was 3 observations per hour and included DO (mg/L), temperature (degrees C), pH, specific conductivity (mS/cm), and turbidity (NTU). Two hours (6 time steps) of memory for each parameter was used and considered a new feature. This raised the datasets dimensionality from 5 to 35. Memory parameters act to improve the results of a dynamic system by providing the model information on the rate of change of its features, instead of the model only considering a simple "snapshot." Using model memory is somewhat analogous to a time derivative in physics-based models. The table below provides a complete list of the input variables and the target variables in this study as well as their respective units of measurement, where mS/cm is microSiemens per centimeter, deg C is degrees Celsius, NTU is nephelometric turbidity units, and mg/L is milligrams per liter.

Table 2.1. Data description statistics

	Dissolved oxygen (mg/L)	Temperature (deg C)	pH	Specific conductivity (mS/cm)	Turbidity (NTU)
Min	0.20	1.19	6.39	0.15	0.00
Max	17.57	27.23	8.71	0.83	3000.00
Mean	8.36	13.29	7.35	0.58	80.18
Std. Dev.	3.23	5.52	0.36	0.03	312.63

The following table lists the correlation coefficients (ρ) of each water quality parameter with dissolved oxygen.

Table 2.2. Water quality correlation coefficient matrix

Water quality parameter	ρ_{DO}
Specific conductivity	-0.38
pH	0.31
Temperature	-0.18
Turbidity	0.11

As evidenced in the above figures and chart, there exists a negative correlation between temperature and dissolved oxygen concentration, and specific conductivity and dissolved oxygen concentration. This is in agreement with Colt (1983) which states that temperature decreases cause an increase in the saturation concentration of DO [43].

2.3 Data preprocessing

Data sets often contain missing values, outliers, and features that are not useful to the model. Useless features may be contain redundant information with others or are not well correlated to the target feature. While table 2.2 provides some insight as to the relationship between input and output, further parameter selection will be accomplished with the use of wrapper algorithms. In this chapter there were five durations where the instruments were taken offline, two of which lasted nearly 3 weeks, or 1331 instances. In total 3298 instances of missing data were removed, leaving 9252 instances for analysis.

The input data was then linearly normalized so that the feature space would be symmetrical. If the data is not normalized, the feature space will take the shape of the features with the largest range which will skew the results. These values will have a bigger influence on the model even if they are not a stronger predictor. The data was

normalized in a linear fashion, by first subtracting the minimum value for each instance, and then dividing this difference by the range feature's range. A mathematical description of the linear normalization is provided below in equation (2.1).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (2.1)$$

2.4. Initial parameter selection

For the time series data mining case, 8 hours, or 32 time steps of memory are considered. Thirty-two memory parameters were considered for each of the water quality measurements. Wrapper feature selection algorithms are utilized with heuristic search methods to find the optimal subset, or combination of input parameters, that provide the best results. Table 2.1 shows the list of memory features chosen.

Table 2.3. Initial input variables for Type-1 DO modeling

Time step	Input variables	Time step	Input variables
t-0	Hour of day	*t*-3	Hour of day
	Specific conductivity (mS/cm)		Specific conductivity (mS/cm)
	Temperature (deg C)		Temperature (deg C)
	pH		pH
	Turbidity (NTU)		Turbidity (NTU)
t-1	Hour of day	*t*-5	Hour of day
	Specific conductivity (mS/cm)		Specific conductivity (mS/cm)
	Temperature (deg C)		Temperature (deg C)
	pH		pH
	Turbidity (NTU)		Turbidity (NTU)
t-2	Hour of day	*t*-6	Hour of day
	Specific conductivity (mS/cm)		Specific conductivity (mS/cm)
	Temperature (deg C)		Temperature (deg C)
	pH		pH
	Turbidity (NTU)		Turbidity (NTU)
t-3	Hour of day		
	Specific conductivity (mS/cm)		
	Temperature (deg C)		
	pH		
	Turbidity (NTU)		

2.5. Parameter selection algorithm

As mentioned above, the wrapper search method is much more computationally expensive than the correlation-based filter methods described in Chapter 2, especially for long or highly dimensional data sets. For this reason, five percent of the total instances (361/9220) were chosen at random and without replacement for processing in the wrapper algorithms. The genetic search algorithm as described in Chapter 1 is used, and compared with the best first search algorithm.

The best first search uses the method of greedy hill-climbing to find an optimal subset. Each subset is trained with the user selected modeling algorithm, customary for wrapper search approaches. The accuracy of each model (RMSE) is used as the metric for determining the optimality of the subset. Once the accuracy is determined, the search continues onward by adding a new feature to the current subset, backtracking if necessary. The backtracking facility ensures that if a newly added feature does not improve results of the previous set of selected features, it will "backtrack," or return to the previous subset that showed better accuracy. From this return point, the search continues by adding a different feature to the original subset, and analyzing this new subset's performance. The algorithm stops after a user-defined number of forward tracking instances result in no model improvement. The search method can be run backwards (starting with the entire data set and removing features) or forwards (starting with a single feature and adding features) [26]. The search algorithm was defined to run forward and terminate after backtracking 5 times.

The results of the feature selection for both the Type-1 and Type-2 DO modeling are shown below in Table 3 and Table 5.

Table 2.4 Type-1 DO modeling wrapper feature selection results

Best first search		Genetic algorithm search	
Temperature *t*-0		Temperature *t*-0	Hour of day *t*-3
Specific conductivity *t*-0		pH *t*-0	Temperature *t*-3
Hour of day *t*-1		Turbidity *t*-0	pH *t*-3
Temperature *t*-1		Hour of day *t*-1	Turbidity *t*-3
pH *t*-1		Temperature *t*-1	Hour of day *t*-5
Specific conductivity *t*-1		pH *t*-1	Hour of day *t*-6
pH *t*-2		Specific conductivity *t*-1	Temperature *t*-6
Temperature *t*-3		Hour of day *t*-2	pH *t*-6
Specific conductivity *t*-3		pH *t*-2	Specific conductivity *t*-6
pH *t*-3		Specific conductivity *t*-2	Turbidity *t*-6
Temperature *t*-6		pH *t*-3	
		Specific conductivity *t*-3	
		Turbidity *t*-3	

Table 2.5. Type-2 DO modeling wrapper feature selection results

Best first	Genetic search
Dissolved oxygen *t*-1	Dissolved oxygen *t*-1
Dissolved oxygen *t*-8	Dissolved oxygen *t*-9
Dissolved oxygen *t*-31	Dissolved oxygen *t*-10
	Dissolved oxygen *t*-18
	Dissolved oxygen *t*-26
	Dissolved oxygen *t*-27
	Dissolved oxygen *t*-28

2.6. Model training/testing

Following Tan *et al.* (2006), 2/3 of the dataset was used for training, and 1/3 for testing, which is a commonly used ratio to balance generalizability with accuracy [21]. The networks were tested in their ability to model the DO concentration at the South Amana.

For both the DO modeling and forecasting, the same accuracy metrics are considered. The metrics chosen to are the mean absolute error (MAE) and relative absolute error (RE), whose mathematical representations are shown in equations (2)-(5).

$$\mathrm{AE} = \mathrm{v(target)} - \mathrm{v(predicted)} \qquad (2.2)$$

$$\mathrm{RAE} = \frac{\mathrm{AE}}{\mathrm{v(target)}} \qquad (2.3)$$

$$\mathrm{MRAE} = \frac{\sum_{i=1}^{N} \mathrm{RAE_i}}{N} \qquad (2.4)$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} \mathrm{AE_i}}{N} \qquad (2.5)$$

2.6.1. Type-1 DO modeling

Using Statistica's "Automatic Network Search," 100 MLP's were generated with random attributes, such as learning rate, momentum, number of hidden layers, and number of nodes. The activation functions tried in the neurons were the identity, logistic, tanh, and exponential functions. The top 5 performing MLPs were retrained (tuned) and their results are shown in Tables 2.6 and 2.7.

Table 2.6. Type-1 modeling wrapper-best first search derived MLPs

Network structure	Training corr.	Test corr.	Training RAE	Test RAE	Hidden activation	Output activation
MLP 11-13-1	0.933	0.917	0.566	0.857	Tanh	Logistic
MLP 11-13-1	0.936	0.900	0.632	1.028	Tanh	Logistic
MLP 11-6-1	0.868	0.853	1.273	1.363	Tanh	Logistic
MLP 11-6-1	0.906	0.877	0.923	1.231	Tanh	Tanh
MLP 11-6-1	0.910	0.885	0.888	1.168	Logistic	Tanh

Table 2.7. Type-1 modeling wrapper-genetic search derived MLPs

Network structure	Training corr.	Test corr.	Training RAE	Test RAE	Hidden activation	Output activation
MLP 23-11-1	0.939	0.923	0.612	0.783	Tanh	Logistic
MLP 23-13-1	0.931	0.912	0.688	0.913	Exponential	Logistic
MLP 23-8-1	0.938	0.928	0.626	0.736	Logistic	Tanh
MLP 23-12-1	0.913	0.897	0.837	1.052	Tanh	Exponential
MLP 23-6-1	0.912	0.893	0.865	1.092	Tanh	Identity

2.6.2. Type-2 DO modeling

The following tables show the network structure of the MLPs derived from both feature selection algorithms, along with their respective training and testing performance, and their hidden activation and output activation functions.

Table 2.8. Wrapper-genetic search derived NNs

Network structure	Training corr.	Test corr.	Training RAE	Test RAE	Hidden activation	Output activation
MLP 7-3-1	0.995	0.997	0.053	0.029	Exponential	Exponential
MLP 7-12-1	0.997	0.997	0.035	0.028	Logistic	Tanh
MLP 7-7-1	0.996	0.997	0.039	0.029	Tanh	Identity
MLP 7-10-1	0.996	0.997	0.037	0.033	Logistic	Tanh
MLP 7-7-1	0.996	0.997	0.038	0.028	Logistic	Identity

Table 2.9. Wrapper-best first search derived NNs

Network structure	Training corr.	Test corr.	Training RAE	Test RAE	Hidden activation	Output activation
MLP 3-3-1	0.996	0.997	0.039	0.028	Logistic	Identity
MLP 3-8-1	0.997	0.998	0.033	0.026	Exponential	Identity
MLP 3-7-1	0.997	0.998	0.033	0.026	Logistic	Identity
MLP 3-9-1	0.995	0.998	0.052	0.026	Logistic	Logistic
MLP 3-7-1	0.996	0.998	0.038	0.026	Exponential	Identity

It is apparent from the above tables that the MLPs derived from the genetic search method were comparable to those selected by the best first search algorithm. It is also apparent that the results from the type-2 modeling are far superior to those of the type-1. This is intuitive because the type-2 modeling used past values of the target variable as input. For the remainder of this study, the top performing MLP from each wrapper algorithm of the type-2 modeling is considered for iterative forecasting.

2.7. Dissolved oxygen concentration forecasting

An obvious extension of modeling DO concentration is to make a DO forecast. This is done by simply iterating the Type-2 DO modeling result successively and updating the memory parameters accordingly [44]. For example, if the oldest memory parameter is *t*-28, as in the MLPs derived from the genetic search algorithm, then on the 29^{th} iterative forecast, the model will be running entirely on model output data, rather than actual observed data.

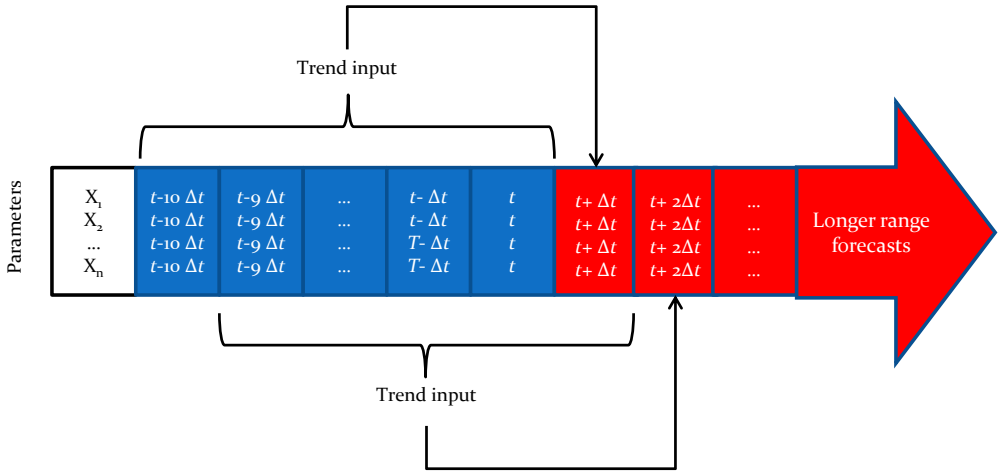

Figure 2.2. Sliding learning window schematic

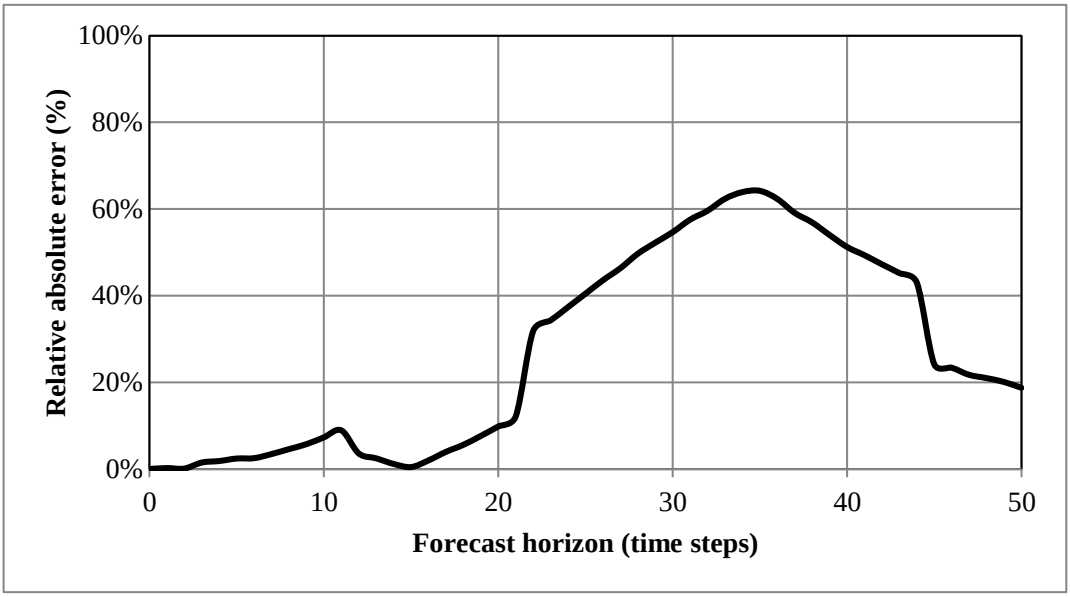

Figure 2.3. MLP derived from genetic search feature selection algorithm

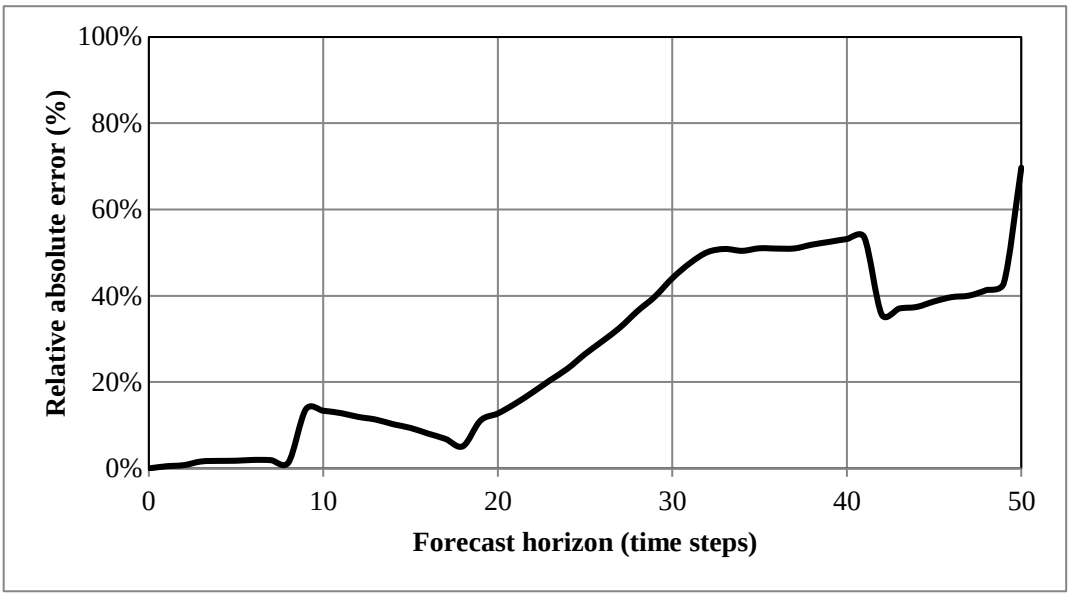

Figure 2.4. MLP derived from best first search feature selection algorithm

It can be noted that the iterative forecasts for both models show good accuracy (less than 20% error) through twenty time steps.

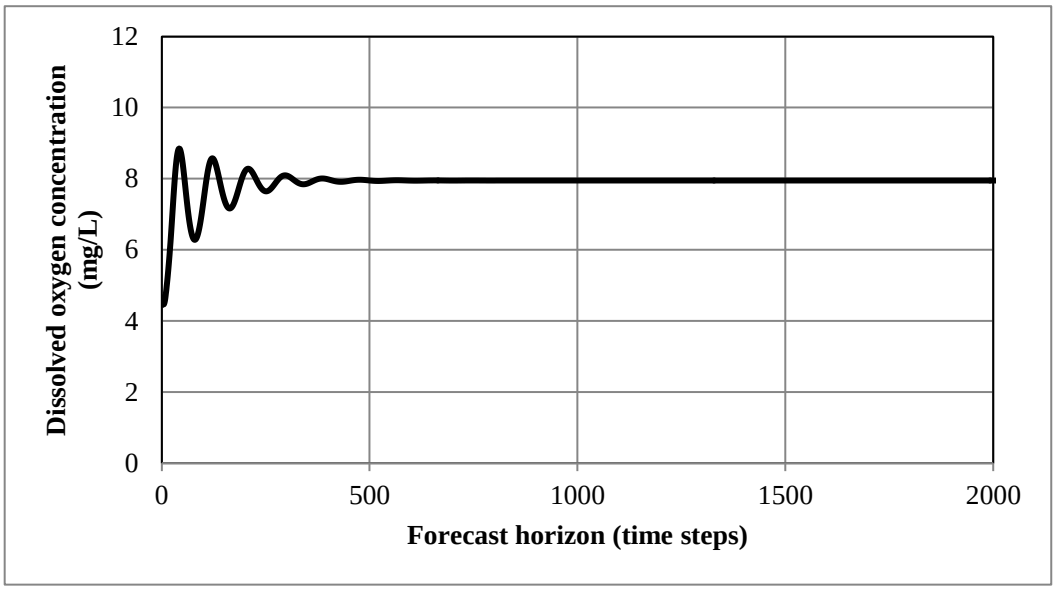

Figure 2.5. MLP derived from genetic search feature selection algorithm

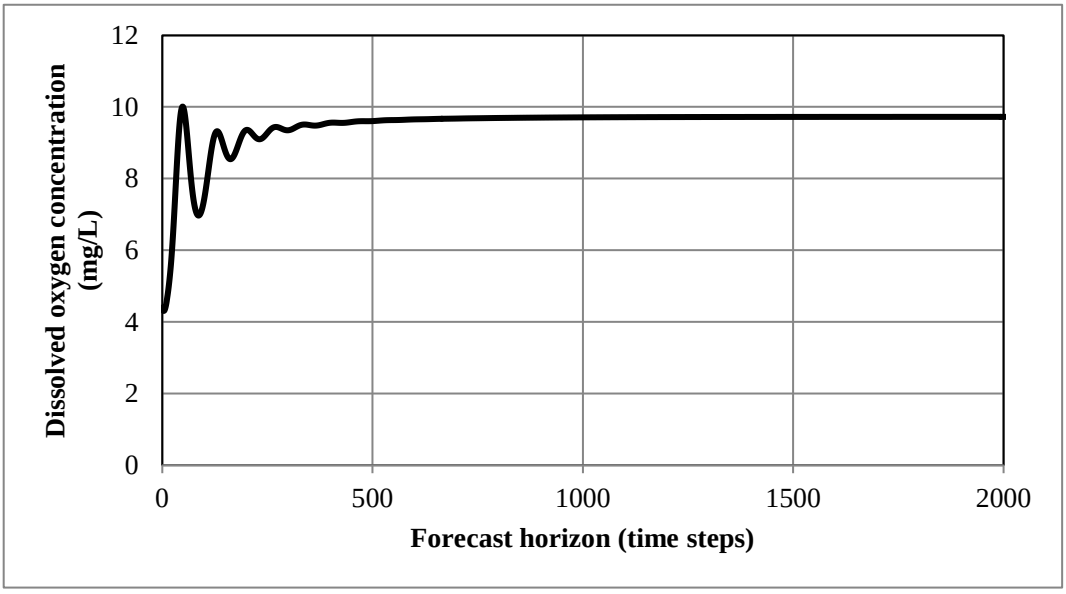

Figure 2.6. MLP derived from best first search feature selection algorithm

Table 2.10. MLP forecast performance through 8900 time steps

MLP model	ME (mg/L)	MAE (mg/L)	RE
Genetic search derived	-0.593	2.725	0.523
Best first search derived	1.138	2.707	0.633

Table 10 is complimentary to the charts displayed in Figures 2.3 and 2.4. Figure 2.3 shows the model's forecast converging to ~8mg/L, and Figure 2.4 shows its respective model's forecast converging to ~10mg/L. From Table 2.10 the ME of the genetic search MLP and best first MLP was -0.593 mg/L and 1.133 mg/L, respectively. In other words, the first model tended to underestimate the DO concentration while the second overestimated.

2.8. Conclusion

Two data driven techniques for DO modeling (1) using concurrent remaining water quality measurements and (2) using previous DO observations (time series data mining), were exercised in this chapter. The genetic search method showed slight dominance over the best first search method for selecting the optimal features in the Type-1 modeling case, but the two algorithms had comparable results in the Type-2 case.

The Type-2 DO modeling technique outperformed the Type-1 technique and was used to make longer range iterative forecasts. These forecasts showed accuracy up to 20 time steps with a relative absolute error within 10%. Both forecasting models converged over time (around 500 time steps) to a value close to the average DO concentration for the period.

In the case where future data is not possible and data must be estimated with only past observations, these two methods may be applied. One method uses other concurrent and past water quality parameters to estimate a water quality value that is either missing or not measured at a given location. The other method uses memory values of the desired parameter for modeling and forecasting using time series data mining.

CHAPTER 3.

TURBIDITY FORECAST WITH DATA DRIVEN MODELING

3.1. Introduction

Turbidity is one of the basic measures of water quality. Cloudy water would not be consumed for obvious reasons. Besides helping to determine potable water, turbidity has a significant impact on ecology. Suspended particles block the passing of light through water, limiting the ability of photosynthetic life, and those creatures which feed on such organisms, to flourish. Carnivorous predators have trouble locating food in murky waters. Extreme turbidity values disrupt fish respiration and may lead to their extinction. Finally, recreational activities such as fishing and swimming are negatively impacted. Furthermore, a water quality forecast provides water treatment plants advanced notice so they may make operational adjustments so as to conserve energy [27].

In this chapter, data mining is utilized to predict water quality, which continues to be a modeling challenge, as it is characterized by nonlinear and non-stationary properties. Five months of real-time data from Clear Creek, an Iowa River tributary, was collected to model turbidity at a downstream location using MLP algorithms. To achieve the best prediction accuracy, correlation analysis and parameter selection algorithms are used to select the most suitable inputs. The resulting neural network model is compared to a linear regression as a demonstration of the neural network's ability to interpret the nonlinear dynamics of water quality.

3.2. Data description

The Consortium of Universities for the Advancement of Hydrologic Science **(***CUAHSI***)** is a National Science Foundation (NSF) supported organization for the development of infrastructure and services of hydrologic science, executed at the university level. A product of the CUAHSI at the University of Iowa is the Clear Creek Digital Watershed (CCDW). The CCDW's Hydrological Information System (HIS) provides various data (i.e., water quality, discharge, and NEXRAD) from several locations within the CCW. Five months (5/12/2009 to 9/24/2009) of 15-minute water quality and discharge data was selected for study on the basis of completeness. This time series was also ideal because 15-minute discharge data is available. Prior to 12/2/2008, the discharge measurements were recorded every 30 minutes. The dataset considered in this research contains turbidity observations, measured in Nephelometric Turbidity Units (NTU), and discharge observations, measured in cubic meters per second (CUMEC), from two locations located approximately 13 km from one another. The two gauging locations are at Oxford, IA and Coralville, IA, both on Clear Creek, and can be viewed below in Figure 3.1. Table 3.1 provides some statistics of the data set.

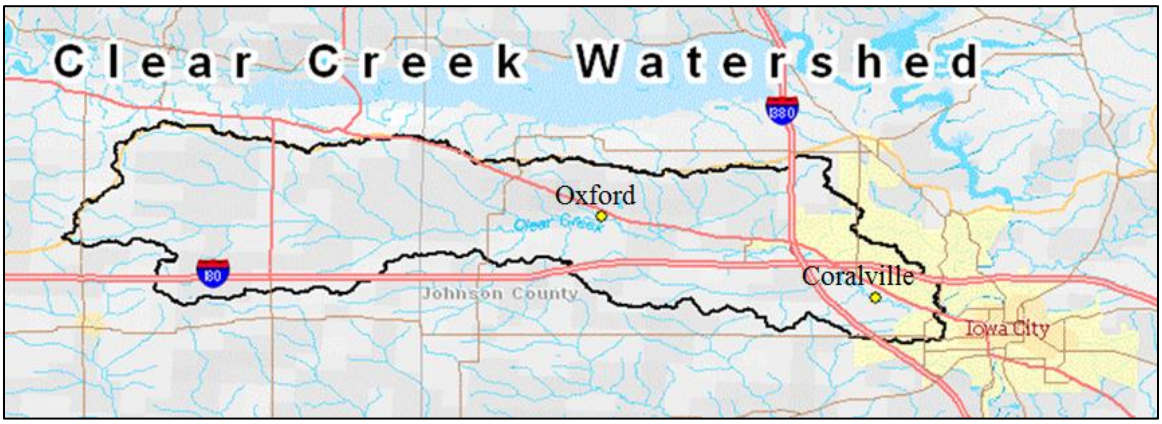

Figure 3.1. Map of Clear Creek study area

Table 3.1. Data statistics

	Coralville turbidity (NTU)	Oxford turbidity (NTU)	Oxford discharge (CUMEC)	Coralvile discharge (CUMEC)
Min	0.00	0.00	24.00	27.00
Max	1590.00	1584.00	4000.00	3510.00
Standard deviation	185.74	262.48	369.77	361.90

3.3. Data preprocessing

The original data set was from 5/12/2009 to 10/23/2009. However, there were many gaps in the time series for various parameters and even changes in the temporal resolution of the discharge data. As the analysis is to include memory parameters, or past values, it is essential that missing data be accounted for and assigned a no data indicator (NaN) rather than simply omitted from the time series. If the empty data points are not accounted for, the model may errantly select a memory parameter for more than one time

stamp in the past. Missing data was removed pairwise. After missing data removal, the number of elements in the data set was reduced from 15718 to 7527.

The input data is also linearly normalized from 0 to 1. The purpose of this is to prevent parameters with wide ranges from dominating the model. Essentially this reshapes the feature space so as to be more spherical, rather than skewed in the direction of the parameter with the largest domain.

3.4. Initial parameter selection

Memory parameters provide a sense of the process' local rate of change at a given instant. Modeling without memory parameters is equivalent to predicting the position of an object in motion using a snapshot rather than a flip book. Downstream turbidity (target variable) memory values were *not* used as input parameters so as to create a regression based solely on independent variables for both modeling and prediction. However, upstream turbidity was used as an input parameter.

Like the work of Palani *et al*. (2008) involving chlorine concentrations, the correlation between upstream and downstream turbidity was key information in selecting input parameters [9]. The lag, or travel time, between the Oxford (upstream) and Coralville (downstream) observed discharge and turbidity was essential in defining the bounds of the model's memory. Figure 3.2 shows the turbidity time series of the two gauges. One may notice a slight lag, with Oxford (upstream) turbidity peaks preceding Coralville (downstream) peaks. Figure 2 is a normalized time series of Oxford discharge and Coralville turbidity that show a similar lag. Due to the likeness of the two figures,

the observation can be made qualitatively that turbidity and discharge at a given site are strongly correlated. For a quantitative comparison, a cross correlation is conducted.

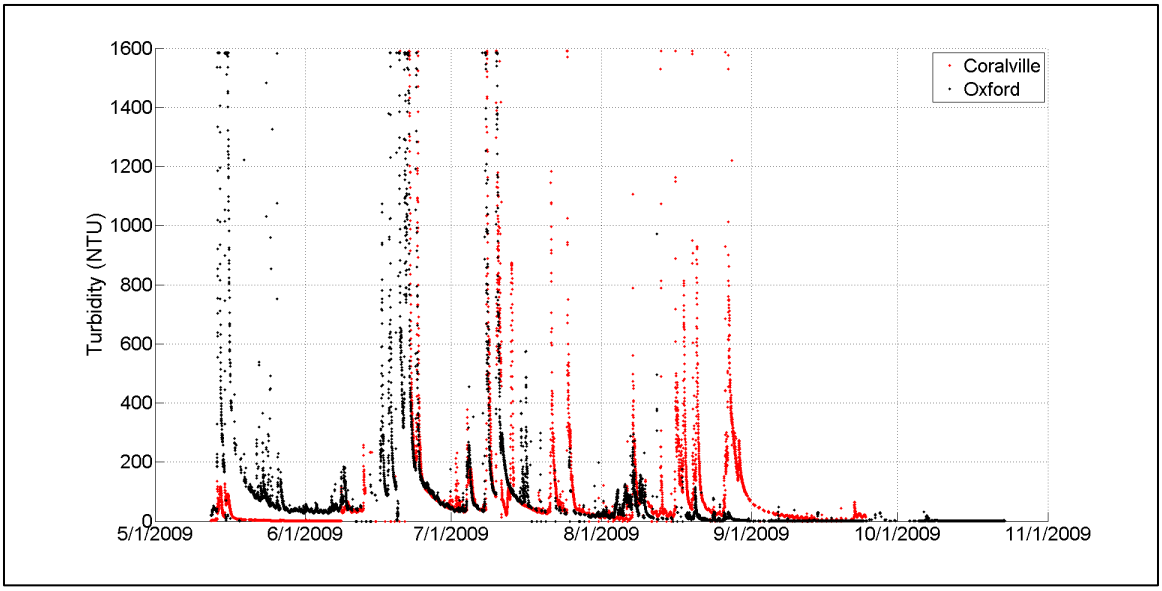

Figure 3.2. Oxford and Coralville turbidity time series

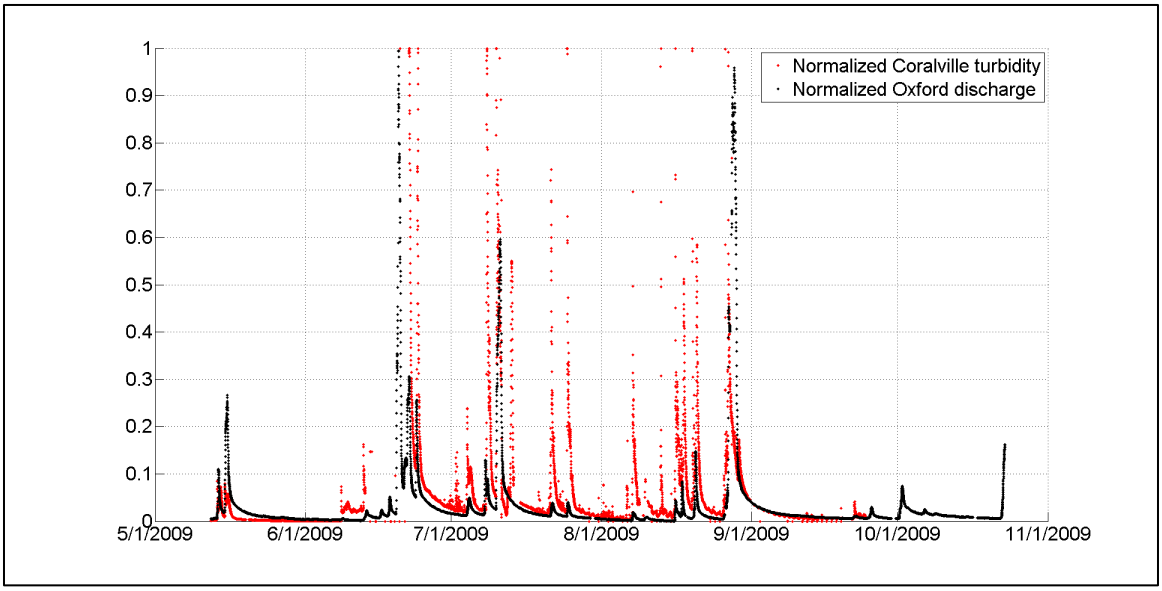

Figure 3.3. Normalized Coralville turbidity and normalized Oxford discharge

Figures 3.4-3.7 are graphs of the cross-correlations of Oxford turbidity and Coralville turbidity, Oxford discharge and Coralville discharge, and Coralville discharge and Coralville turbidity. The cross correlation equation can be seen below, in equation 3.1, Where X_n and Y_n are jointly stationary random processes and E $\{\cdot\}$ is the expected value operator [28].

$$R_{XY}(m) = E\{X_{n+m}Y_n^*\} = E\{X_nY_{n-m}^*\} \quad (3.1)$$

The purpose of these three charts is to better visualize and analyze the lag time between the two gauging stations. After solving the normalized cross correlation equation, and locating the peaks of Figures 3.4-3.7, the maximum cross correlation value is at *t*+12 for Oxford turbidity-Coralville turbidity, *t*+24 for Oxford discharge-Coralville turbidity, and *t*+0 for Coralville discharge-Coralville turbidity. As each time step is 15 minutes for this data set, these three value scan be expressed as 3 hours, 6 hours, and 0 hours, respectively.

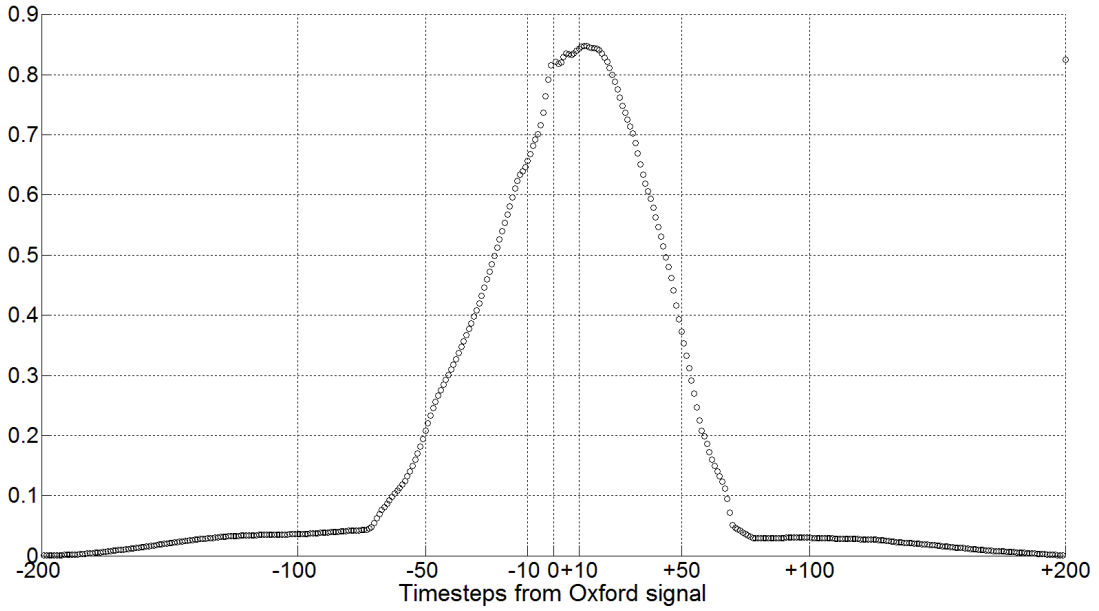

Figure 3.4. Oxford turbidity-Coralville turbidity cross-correlation

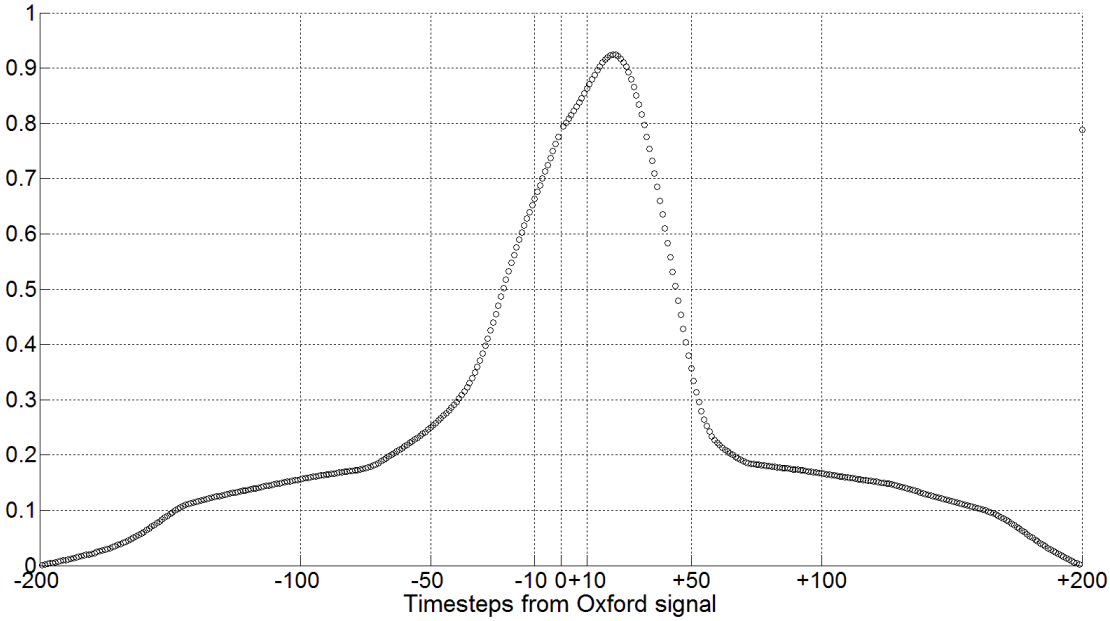

Figure 3.5. Oxford discharge-Coralville discharge cross-correlation

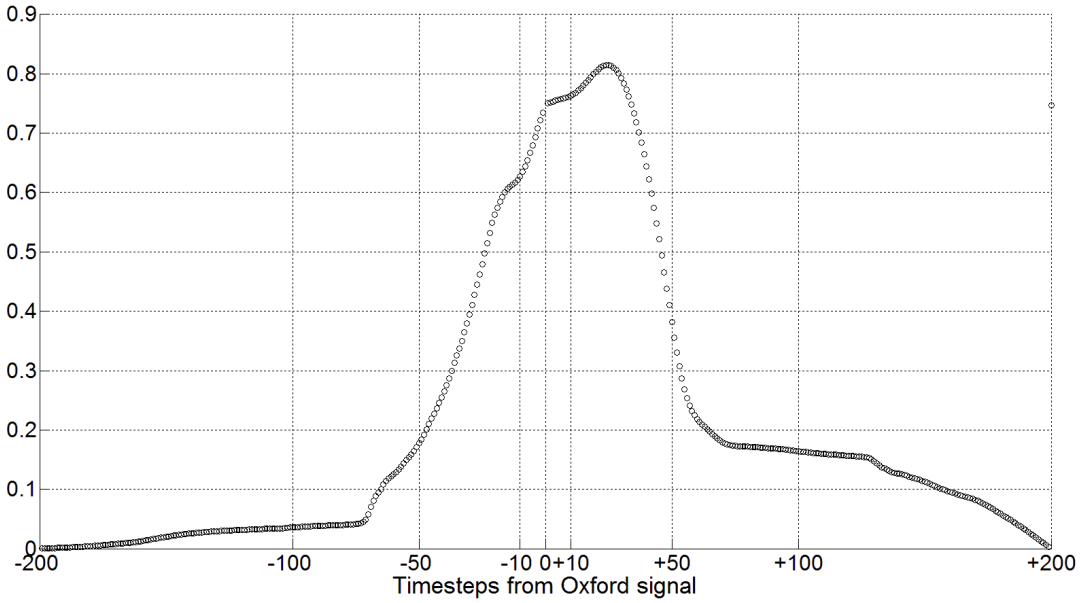

Figure 3.6. Oxford discharge-Coralville turbidity cross-correlation

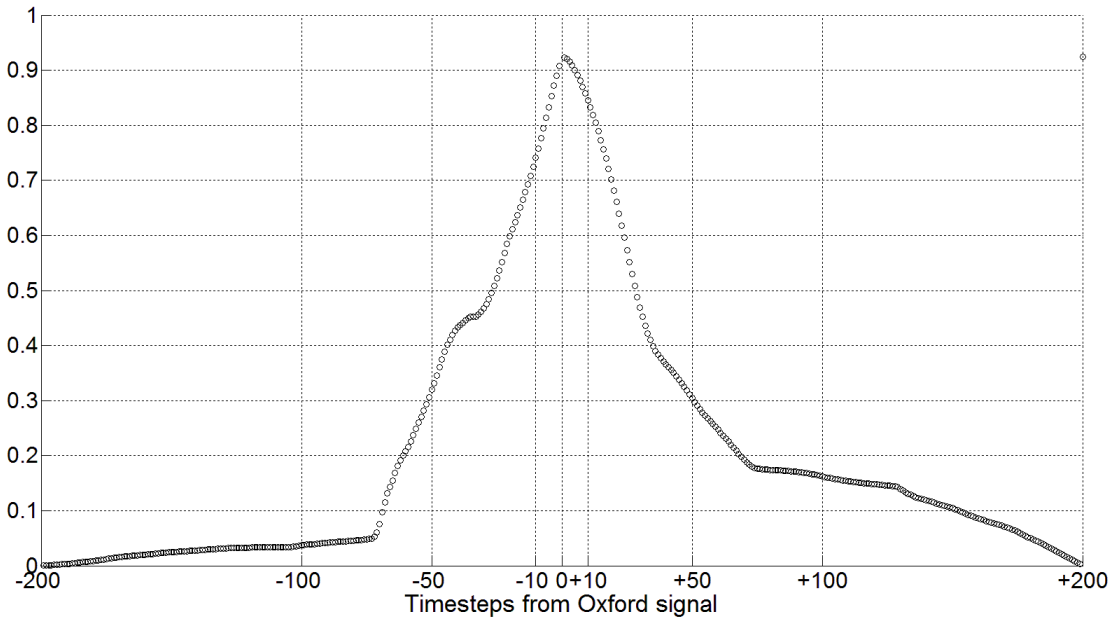

Figure 3.7. Coralville discharge-Coralville turbidity cross-correlation

3.5. Parameter selection algorithm

To further reduce the dimensionality of the model a genetic search algorithm was used to select the optimal subset of input parameters for model building. While the parameter selection algorithms are much better than correlation-based parameter selection, their computational demand restricts the dimensionality of the data set used for input. For this reason, the correlation filtering procedure of the previous section was used.

The genetic search algorithm optimizes the subset of input parameters through the evolutionary processes of crossover, mutation, and selection. A population of random subsets is created, and then trained/tested with a multilayer perceptron. Once the accuracy is calculated (RMSE), those individuals (subsets) whom show the highest degree of accuracy, survive, and are randomly paired up for crossover. During crossover, the pairs of individuals will swap a portion of their subset with one another. This introduces diversity to the set of solutions. Mutation occurs at random in the new generation, also to introduce diversity to the population. During mutation, one of input parameters is exchanged with another. The offspring's fitness is calculated (RMSE), and the process is iterated a user-defined number of times.

In this chapter the initial population was set to 20, the crossover probability to 0.6, mutation probability to 0.033, and number of generations to 50, as per Hall *et. al* [26]. The equations for root mean squared error (RMSE) and mean squared error (MSE) are found in the equations (3.2) and (3.3) [28]. Table 3.2 and Figure 3.8 display the parameters chosen by the genetic search algorithm and its convergence behavior throughout 50 generations.

$$\mathrm{RMSE}(\theta_1, \theta_2) = \sqrt{MSE(\theta_1, \theta_2)} \quad (3.2)$$

$$MSE = \frac{\sum_{i=1}^{n}(x_{1,i} - x_{2,i})^2}{n} \quad (3.3)$$

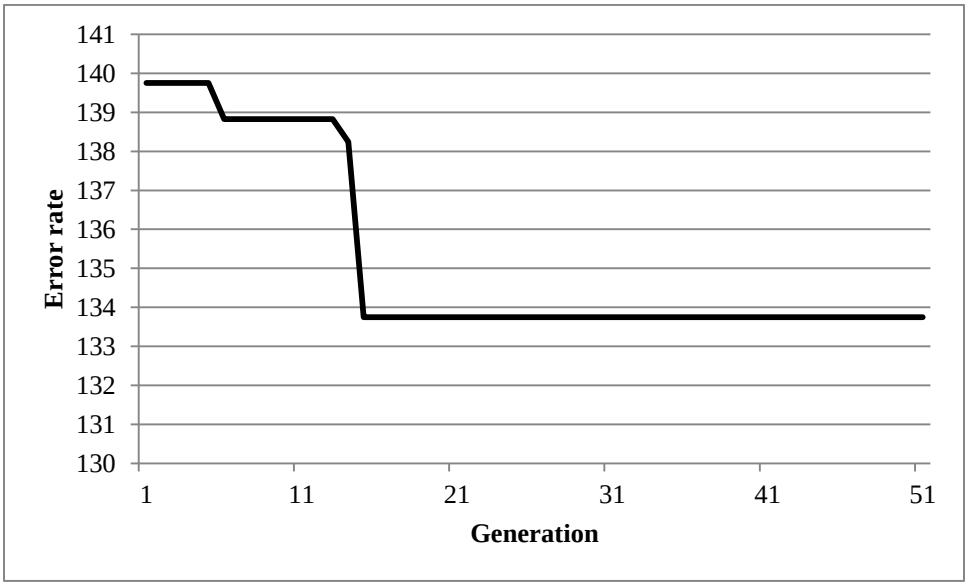

Figure 3.8. Genetic search error rate convergence through 50 generations

Table 3.2. Genetic search parameter selection results

Coralville Q *t*-2	Oxford Q *t*-25
Oxford Q *t*-18	Oxford Q *t*-27
Oxford Q *t*-20	Oxford Q *t*-28
Oxford Q *t*-22	Oxford T *t*-11
Oxford Q *t*-23	

Where Q is discharge and T is turbidity.

It can be seen that the error rate (RMSE) makes decrease between generations 10 and 15, and then converges to around 134 for the rest of the run time. After the correlation based preprocessing of section 3.2 and this genetic search parameter selection, the number of parameters is reduced to only 9.

3.6. Algorithm training and testing

The data set, now consisting of 7527 elements was split into three sections; one held 70% of the elements and was used for testing, and the two others, both contained 15% tuning and testing purposes. The data was allocated to these three subsets in a random fashion so as not to over train, and thus over fit, the model to one particular period in the time series. The data-mining software randomly built 20 networks with various numbers of nodes, hidden layers, learning rates, and activation functions, all within user defined thresholds. The five top performing NNs from the initial twenty were retrained (tuned) and tested.

3.6.1. Metrics for comparison

The metrics chosen to measure performance are the correlation coefficient (ρ) mean absolute error (MAE), relative error (RE), whose mathematical representations are shown in equations (3.1)-(3.4).

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{3.4}$$

$$\mathrm{AE} = \mathrm{v(target)} - \mathrm{v(predicted)} \tag{3.5}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} \mathrm{AE_i}}{N} \tag{3.6}$$

$$\mathrm{RE} = \frac{\mathrm{AE}}{\mathrm{v(target)}} \tag{3.7}$$

N is the number of elements, or length of the data set, *v(target)* is the observed turbidity measurement, and *v(predicted)* is the predicted turbidity of the model.

3.7. Results

The results of the top five MLPs for turbidity modeling are visible below in table 3.4, and their details in Table 3.5. Corr. coeff is the correlation coefficient between the models turbidity prediction and the target turbidity, ME is mean error, MAE.

Table 3.3. Downstream turbidity modeling results

Algorithm	Corr. coeff.	ME	MAE	RE
1.MLP 9-8-1	0.874	-0.209	39.121	0.428
2.MLP 9-11-1	0.881	-0.973	38.851	0.425
3.MLP 9-13-1	0.900	-3.750	38.065	0.417
4.MLP 9-12-1	0.851	-0.330	41.998	0.460
5.MLP 9-13-1	0.885	-5.116	40.417	0.442
Ensemble	0.909	-2.076	34.171	0.374

Table 3.4. MLP details

Network structure	Hidden activation	Output activation
MLP 9-8-1	Tanh	Tanh
MLP 9-11-1	Tanh	Identity
MLP 9-13-1	Tanh	Logistic
MLP 9-12-1	Exponential	Tanh
MLP 9-13-1	Logistic	Exponential

All the MLPs showed similar accuracy, which makes for a stable ensemble. The number of hidden nodes ranged from 8 to 13 (center hyphenated number under "Network structure" column). The most common hidden activation functions were the hyperbolic tangent with the exponential and logistic functions also appearing. The results of the training and testing were similar in accuracy. This is consistent training and testing performance is usually attributed to models that do not over fit [36].

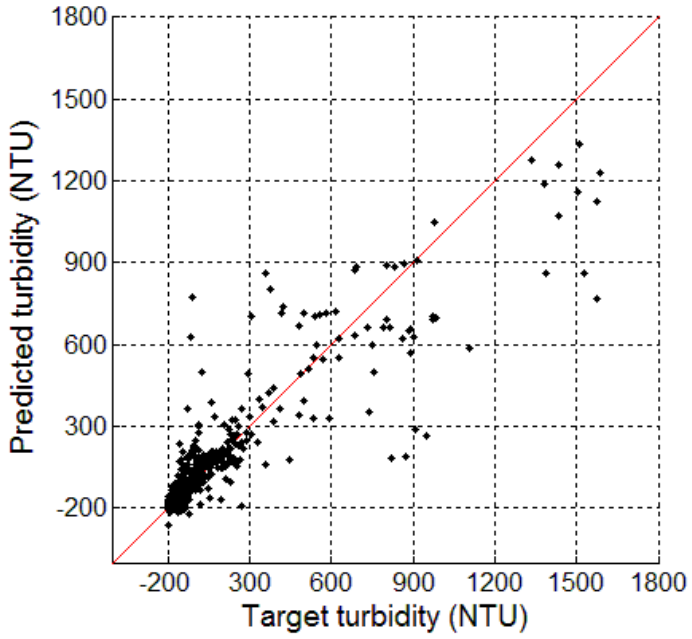

Figure 3.9. Ensemble predicted downstream turbidity scatter plot

Table 3.5. Turbidity modeling sensitivity analysis

Model	OX Q *t*-18	CO Q q *t*-2	OX Q *t*-28	OX Q *t*-20	OX Q *t*-22	OX Q *t*-27	OX Q *t*-25	OX Q *t*-23	OX T *t*-11
MLP 9-6-1	11.61	17.16	2.38	7.90	7.14	2.98	6.82	1.13	1.49
MLP 9-12-1	10.50	24.40	2.68	2.68	5.48	2.59	1.29	2.24	2.13
MLP 9-9-1	1.51	18.62	1.17	1.20	1.19	2.84	1.22	1.22	1.78
MLP 9-7-1	20.65	18.60	10.13	19.50	4.90	5.79	15.50	10.39	1.49
MLP 9-11-1	203.77	26.49	83.56	29.51	33.15	26.86	12.67	16.26	2.21

Where CO is Coralville and OX is Oxford.

A sensitivity analysis was conducted so as to give some insight on which parameters were most influential to the model. Based on the above figure it is apparent that the Oxford discharge at t-18 was most influential to the model, seconded by the Coralville discharge at t-2.

3.7.1. Comparison with ordinary least squares regression (OLR)

The MLP's ability to model nonlinear and non-stationary phenomena is evidenced when compared to an ordinary linear regression (OLR) model. Although the MLP only had mild success in modeling the downstream turbidity, when considering the high variance of turbidity, and its difficulty in modeling due to its nonlinearity, the results are impressive when compared to the OLR. The OLR was set to step through the attributes removing the one with the smallest standardized coefficient until no improvement was observed in the estimate of the error given by the Akaike information criterion (AIC), which is an effective criterion for estimating the relative support of a model [17]. Also, collinear parameters were removed in model building. The equation for the AIC is provided below.

$$\mathrm{AIC} = 2\mathrm{k} - 2\ln(\mathrm{L}) \qquad (3.8)$$

Where *k* is the number of parameters in the model and *L* is the estimated

The results of the OLR can be found in table below.

Table 3.6. Ordinary linear regression results

Correlation coefficient	0.556
Mean absolute error	71.614
Root mean squared error	159.229
Relative absolute error	0.710

Comparing the result of the above table with the results from Table 3.4., it is obvious that the MLP shows dominance over the OLR and is better able to interpret the nonlinearities of the hydrological system.

3.8. Conclusion

Data-mining algorithms were used to model downstream turbidity of the Clear Creak tributary of the Iowa River and obtain high quality downstream turbidity predictions. The extent of the model's memory was user defined by correlation analysis, and then justified with a standard parameter selection algorithm, the genetic search. The performances of twenty neural networks were evaluated and the top five retained for the building of an ensemble model. The ensemble accuracy is reported algorithm reported accuracy greater than 95%.

CHAPTER 4.

PRECIPITATION ESTIMATION WITH DATA DRIVEN MODELING

4.1. Introduction

The connection between radar data and tipping bucket precipitation has been a topic of interest in the hydrological and meteorological community for a decade and is motivated by the necessity for higher resolution precipitation for hydrological model input. In this paper, a series of multilayer perceptrons (MLPs) trained with next generation radar (NEXRAD) and rain gauge data for precipitation estimation at Oxford, IA. The resulting MLPs have MAEs of less than 0.1mm/hr. The vision of the author is to develop this model, which links rain gauge and radar data, to produce a system of "virtual tipping buckets" (VTBs) that benefit from the accuracy of physical tipping bucket rain gauges, and the spatiotemporal resolution of NEXRAD system technology. The system of VTBs has been developed to serve as input to the Soil and Water Assessment Tool (SWAT) hydrological model [Neitsch SWAT].

The high spatiotemporal resolution of next generation radar (NEXRAD) makes it a useful instrument for precipitation estimation. NEXRAD-II data are the three meteorological base data quantities: reflectivity, mean radial velocity, and spectrum width. NEXRAD-III data are derived from various algorithms for processing NEXRAD-II data to produce numerous meteorological analysis products, such as storm velocity, one hour precipitation total, storm total precipitation, digital mesocyclone detection, digital precipitation array, wind profiles, and vertical integrated liquid content [45].

Radar data has sources of error which could be mitigated by the aid of a secondary system, such as a rain gauge. Blockage by mountains and hilly terrain, confusion with flocks of birds and swarms of insects, and signal attenuation are all problematic to radar observations. In fact, a field of radar studies, called radar ornithology, uses radar system to study the migratory habits of birds. Rain gauges measure rather than estimate precipitation and are thus deemed as the most truthful account of rainfall available.

However, rain gauges provide mere point measurements, and their values may be different from those at another gauge only a few kilometers away. It is common, especially during the convective season when the atmosphere is often unstable, for very high precipitation rates to be measured at one location, and none at another. Should the two technologies be melded together, that is NEXRAD and tipping bucket rain gauge, the strengths of both systems could be utilized.

The aim of this chapter is to use NEXRAD-II reflectivity data from a weather station in Davenport, IA and tipping bucket rain gauge data from South Amana and Iowa City, IA to train multilayer perceptron (MLP) for precipitation estimation at a rain gauge in Oxford, IA. The resulting model is then compared with the National Oceanic and Atmospheric Association's (NOAA) algorithm for converting reflectivity data to hourly precipitation, a NEXRAD-III product. The robustness of this model is tested at two other tipping bucket locations, South Amana and Iowa City. This model could then be used to provide the SWAT hydrological model with rainfall data of a 5 minutely observation frequency and a spatial resolution of 1 km^2. Currently, the SWAT uses three tipping buckets within the ~250 km^2 basin that report rain rates at 15 minute intervals.

4.2. Radar precipitation estimation (Z-R conversion)

The most common conversion (Z-R) of reflectivity to precipitation rate takes the following relationship:

$$Z = a \cdot R^b \qquad (4.1)$$

Where Z is the reflectivity, R is the precipitation rate, and a and b are constants from empirical studies (calibration). Typically, the values used for a and b are 200 and 1.6 respectively. The National Oceanic and Atmospheric Association (NOAA) has its own algorithms for estimating rainfall based on the relationship described in equation (4.1). NOAA's radar-based estimation of rainfall at the tipping buckets is downloaded can be found in NEXRAD-III products.

4.3. Data acquisition

Two types of data were collected for the building of the MLP in this study, (1) radar reflectivity data and (2) tipping bucket precipitation data. A third data set, the NOAA hourly rain fall total, was collected for comparison with the developed model. Although other work has considered using reflectivity bandwidth and horizontal wind velocity [46-48] in their models, their experimental results conclude that reflectivity is the only useful input.

4.3.1. Doppler WSR-88D radar

The National Weather Service's (NWS) Next Generation Radar (NEXRAD) system is comprised of 137 radar sites in the contiguous United States, each of with is

equipped with Doppler WSR-88D radar capable of reporting high resolution data and making a full 360 degree scan every 5 minutes, with has a range of ~230km and a spatial resolution of about 1km by 1km (Baer, 1991). The weather station used in this study is located in Davenport, IA (KDVN), which is approximately 150 km from the tipping bucket locations. Reflectivity was collected from four altitudes above ground level (AGL), 1km, 2km, 3km, and 3km. As both the intensity and altitude of the reflectivity values are required to describe the shape of the approaching storm, it is necessary to provide data from multiple levels [49]. This is also consistent with the literature [10-13].

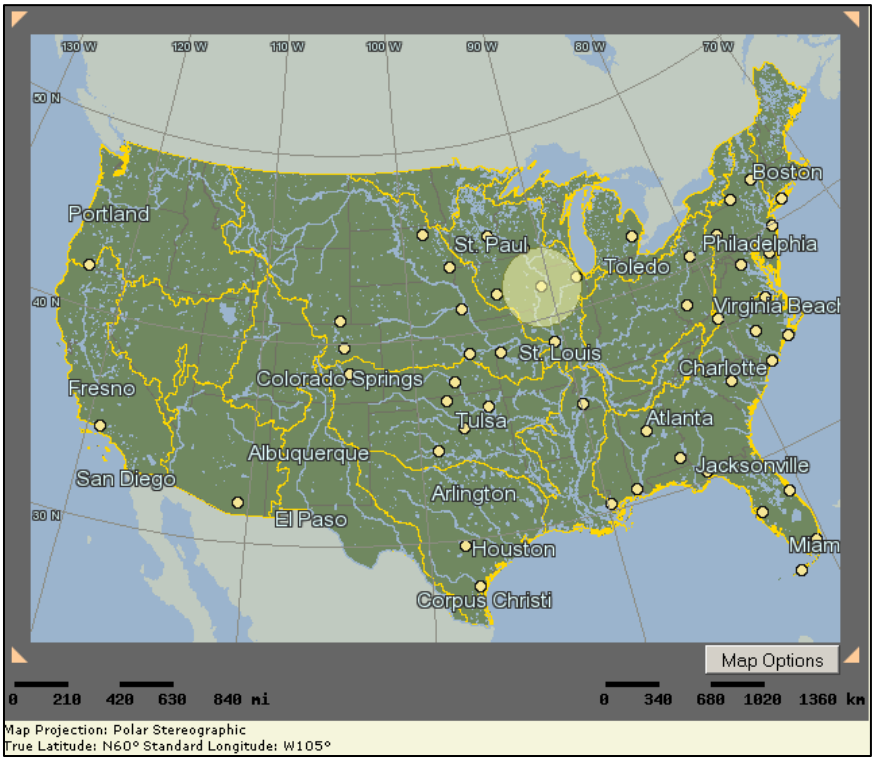

Figure 4.1. Hydro-NEXRAD image of KDVN radar coverage

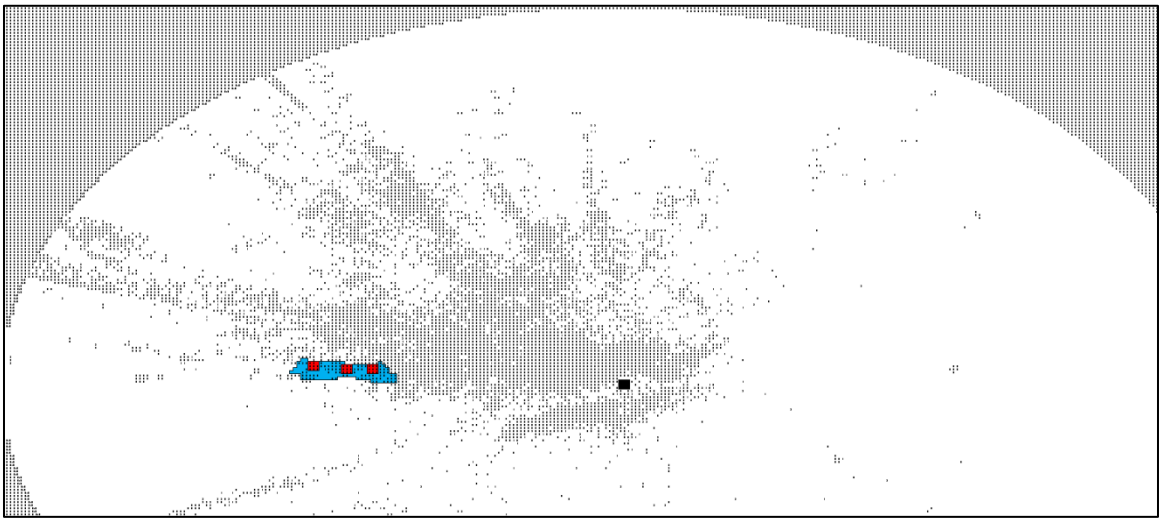

Figure 4.2. NEXRAD reflectivity raster with Clear Creek superimposed

4.3.2 Dual tipping bucket rain gauge

The rain gauge sites make part of the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Clear Creek Digital Watershed (CCDW). They are located at South Amana, Oxford, and Iowa City in Johnson County, IA, about 110-130 kilometers from the KDVN radar site. Each instrument is equipped with dual buckets for quality checking purposes and redundancy. It records precipitation rate in 0.0001 mm/hr, every 15 minutes. The gauges are taken offline during the winter months.

4.4. Preprocessing

Preprocessing data is a crucial step of the data mining process. Outliers, missing data, and unreliable or low quality data all need to be considered before analysis. The

NEXRAD data was ordered from the hydro NEXRAD site and downloaded via an FTP connection. A script was written in Matlab to select the closest grid points that corresponded with the Oxford tipping bucket location. Nine grid points were selected about the tipping bucket location, in agreement with Liu, Chandrasekar, and Xu (2001) [19]. This is to provide some margin for error in the GPS mapping of the tipping buckets and gridding of the KDVN radar raster map. Also, rain does not fall straight down but may be advected horizontally. Finally, The NEXRAD data was collected at 5-min intervals, which is inconsistent with the temporal resolution of the tipping bucket, reported every 15-min. This issue was simply dealt with by averaging the three radar observations made within each tipping bucket observation. Also, the tipping bucket values were recorded to the 0.0001 mm/hr, which seemed excessively precise. These values were rounded to the nearest 0.01 mm/hr for modeling purposes.

The time series considered was from April 1, 2007 to September 30, 2007 and was formatted to 15-min resolution, for a total of 50,792 data points. Figure 3 shows the location of the three tipping buckets in the Clear Creek basin and the radar grid superimposed.

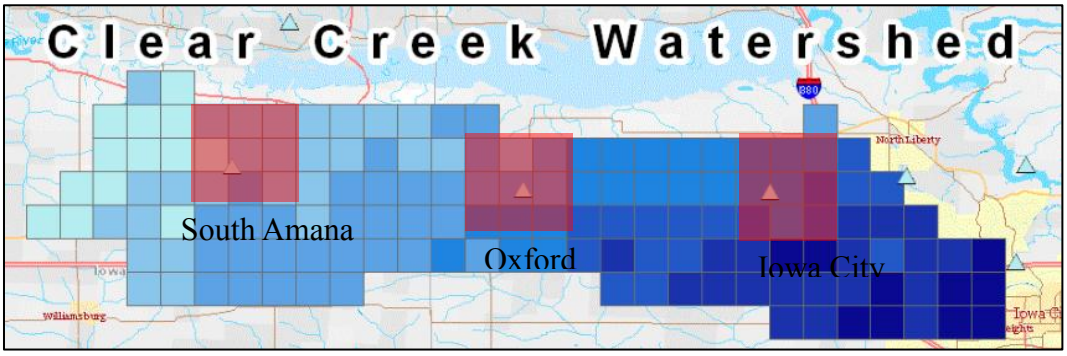

Figure 4.3. Tipping bucket locations

4.5. Parameter selection

Consistent with the previous chapters, the data has been plotted on scatter plots for better visualization of the input parameters relationship with the target variable. Figures 4.4-4.8 display scatter plots of the mean of the tipping buckets versus the Oxford tipping bucket, and the reflectivity values of the center most grid location at each altitude versus the Oxford tipping bucket.

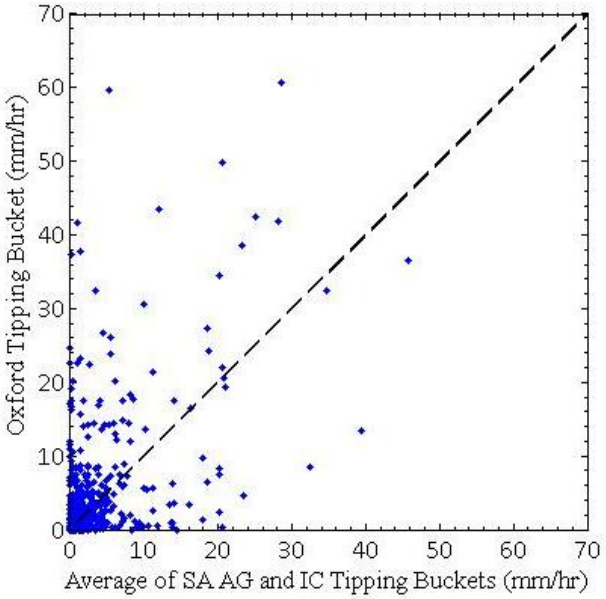

Figure 4.4 Tipping bucket average versus Oxford tipping bucket scatterplot

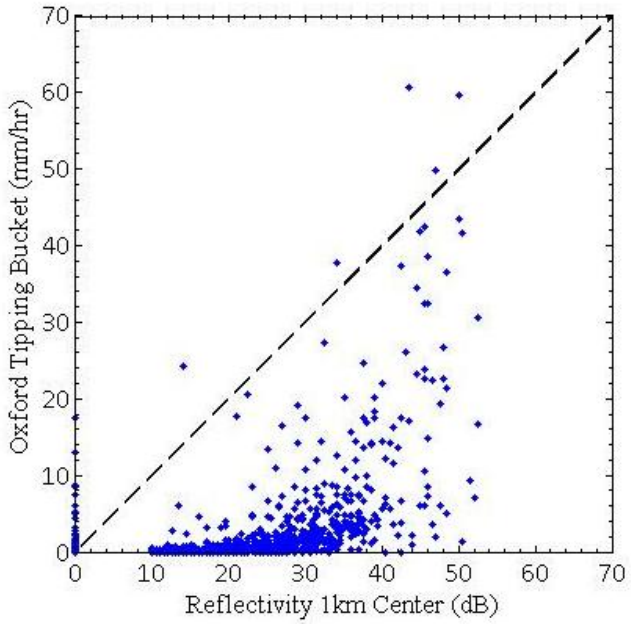

Figure 4.5. Reflectivity at 1km versus Oxford tipping bucket scatterplot

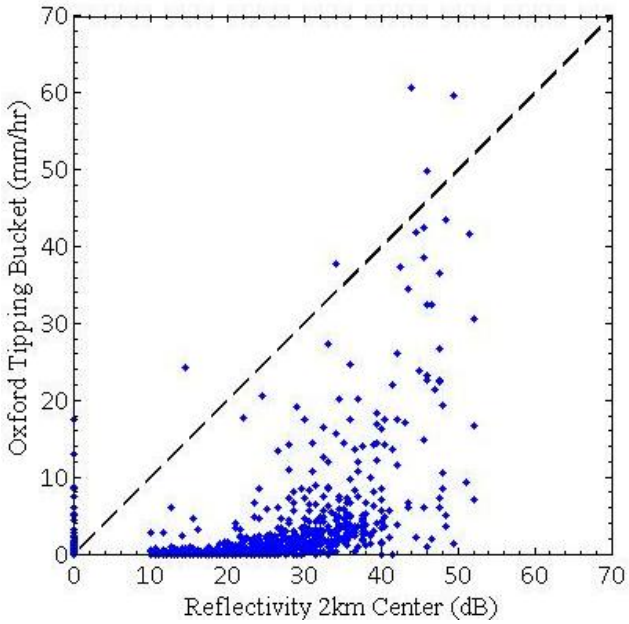

Figure 4.6. Reflectivity at 2km Oxford tipping bucket scatterplot

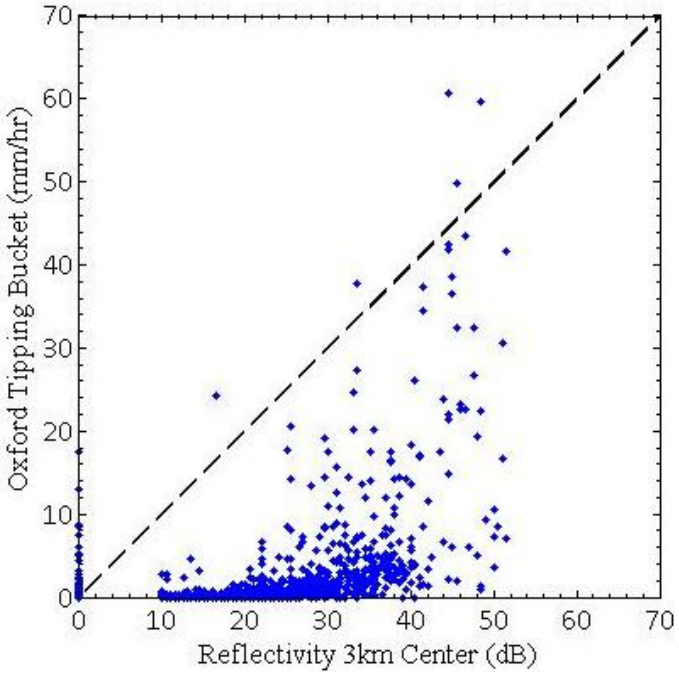

Figure 4.7 Reflectivity at 3km versus Oxford tipping bucket scatterplot

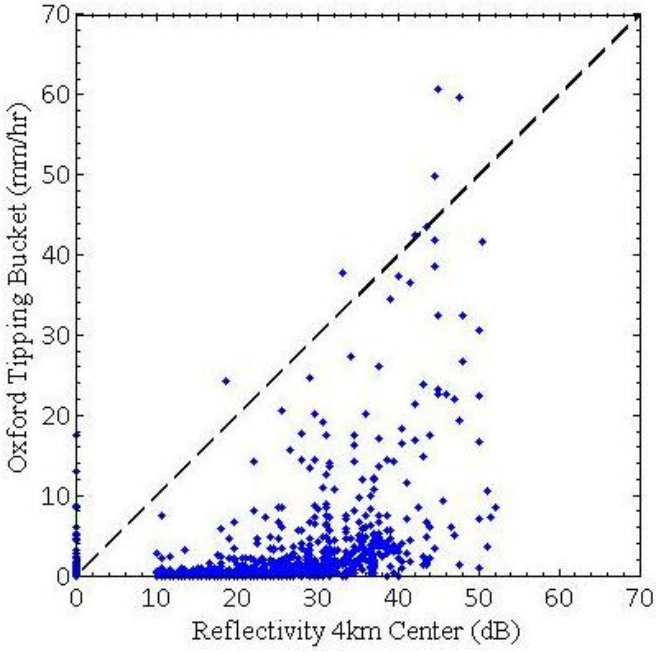

Figure 4.8. Reflectivity at 4km versus Oxford tipping bucket scatterplot

A more detailed, quantitative representation of the correlation between the different reflectivity parameters and the Oxford tipping bucket can be found below in Tables 4.1-4.3.

Table 4.1. Reflectivity-tipping bucket correlation by height-direction

Feature	ρ	Feature	ρ	Feature	ρ	Feature	ρ
1km NW	0.51	2km NW	0.51	3km NW	0.51	4km NW	0.45
1km N	0.38	2km N	0.38	3km N	0.50	4km N	0.44
1km NE	0.39	2km NE	0.39	3km NE	0.50	4km NE	0.44
1km W	0.53	2km W	0.53	3km W	0.53	4km W	0.47
1km C	0.52	2km C	0.52	3km C	0.53	4km C	0.47
1km E	0.51	2km E	0.52	3km E	0.53	4km E	0.46
1km SW	0.53	2km SW	0.53	3km SW	0.53	4km SW	0.47
1km S	0.52	2km S	0.52	3km S	0.53	4km S	0.47
1km SE	0.51	2km SE	0.52	3km SE	0.53	4km SE	0.47

Table 4.2. Reflectivity-tipping bucket correlation by height

Altitude	Corr. coeff.
4km	0.52
3km	0.52
2km	0.51
1km	0.51

Table 4.3. Reflectivity-tipping bucket correlation by direction

Direction	ρ
W	0.53
SW	0.53
C	0.53
S	0.53
E	0.52
SE	0.52
NW	0.51
NE	0.40
N	0.40

There appears to be a very mild trend in both altitude and location in reflectivity and correlation with the Oxford tipping bucket. While correlation measures the strength of the linear relationship, nonlinear relationships may exist in the data set. Heuristic feature selection algorithms are often used in the field of computational intelligence to find optimal subsets for modeling nonlinear phenomenon. The feature selection algorithms selected are the best first and genetic search algorithms, as in the previous chapters. These algorithms are "wrapped" within the MLP algorithm to find the parameters in the data that result in the best model, based on the selected metrics described in section 2.3. In other words, both these algorithms employ a heuristic approach to training and testing data subsets in search of a local optimum.

Table 4.4 shows the results of the feature selection for the genetic search and best first search discussed earlier in this Thesis. Figure 4.9 shows the convergence of the genetic search algorithm through 100 generations, where the value at generation 1 is the, and generation 101 is the error rate of the most fit individual of the final population.

Although the error continues to drop throughout the entire 100 generations,

convergence is observed after 20 generations at an error rate of 0.0230. The most fit individual, as measured by error rate, at the last generation is selected for further analysis.

Table 4.4. Wrapper-genetic search feature selection results

Mean (SA and IC tipping bucket)
North-1km
Southeast-1km
East-2km
Southwest-2km
Northwest-3km
Central-3km
East-3km
Northwest-3km

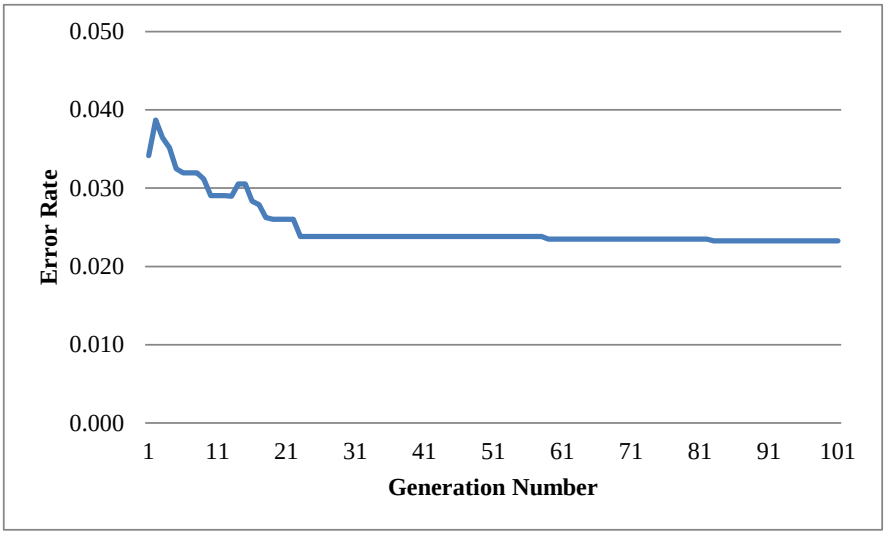

Figure 4.9. Genetic search convergence through 100 generations

4.6. Model training/testing

In training and testing of a data-driven model, there is always a balance between accuracy and overfitting, or lack of generalizability, of the model. Especially for the purpose of this research, which is to establish a model that can be used at other tipping bucket locations, generalizability is of great importance. Following Tan *et al.* (2006), 2/3 of the dataset was used for training, and 1/3 for testing, which is a common split to balance generalizability with accuracy [21]. The networks were tested for predicting the rainfall rate (mm/hr) at the Oxford tipping bucket.

Using Statistica's "Automatic Network Search" option 100 MLP's were generated with random attributes. Some of these characteristics were learning rate, momentum, number of hidden layers, and number of nodes. The activation functions tried in the neurons were the identity, logistic, tanh, and exponential functions. The top 20 performing MLPs were retrained (tuned).

4.7. Metrics for algorithm evaluation

The metrics chosen to measure performance are the mean error (ME) and mean absolute error (MAE) of the models, whose mathematical representations are shown in equations (4.2)-(4.6).

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \qquad (4.2)$$

$$\mathrm{AE} = \mathrm{v(target)} - \mathrm{v(predicted)} \qquad (4.3)$$

$$RE = \frac{AE}{v(target)} \quad (4.4)$$

$$MRAE = \frac{\sum_{i=1}^{N} RAE_i}{N} \quad (4.5)$$

$$MAE = \frac{\sum_{i=1}^{N} AE_i}{N} \quad (4.6)$$

Where N is the number of elements, or length of the data set, RR(target) is the observed precipitation rate measurement, and RR(predicted) is the output of the model.

4.8. Post processing

In some instances the MLP output negative precipitation values, which are infeasible. These data were post processed by simply changing them to "zero" rainfall values. Furthermore, the MLP outputs were rounded from 0.00001 mm/hr, which is not only of higher precision than the tipping bucket rain gauge, but also seemed excessively precise. The output was rounded to the nearest 0.01mm/hr.

4.9. Results

The results of the model building can be found in Table 1, which shows the results of the five top performing MLPs based on the evaluation metrics described earlier. The ensemble model, which is an average of the 5 MLPs, is also included. Figure 1 shows the MLP ensemble model estimated precipitation rate versus the observed rain rate at Oxford.

Table 4.5. MLP performance

Model	Test Corr.	Test ME	Test % error	Std. dev of error
MLP 9-11-1	0.894	-0.012	0.167	0.125
MLP 9-4-1	0.894	-0.009	0.168	0.130
MLP 9-13-1	0.897	-0.007	0.145	0.115
MLP 9-5-1	0.895	-0.012	0.251	0.261
MLP 9-5-1	0.904	-0.008	0.103	0.094

The MLPs shows good accuracy in estimating the magnitude of rainfall rate, and few false positives. The two prominent instances of high error were checked. Upon reviewing these data points, it is found that all three tipping bucket values recorded precipitation, but the radar data reported a clear sky, or reflectivity values of zero. These points are most likely outliers.

4.9.1. Comparison with NEXRAD-III Z-R conversion

As many data was required for the building of the VTB model at the Oxford location, there was not many data points (less than 2000) left for testing. For this reason the scatter plot of Figure 4.8 seems emptier. However, in the next section of this thesis, where the robustness of the VTB model is tested at two new locations, there are many more data points as none were required for retraining at these new locations

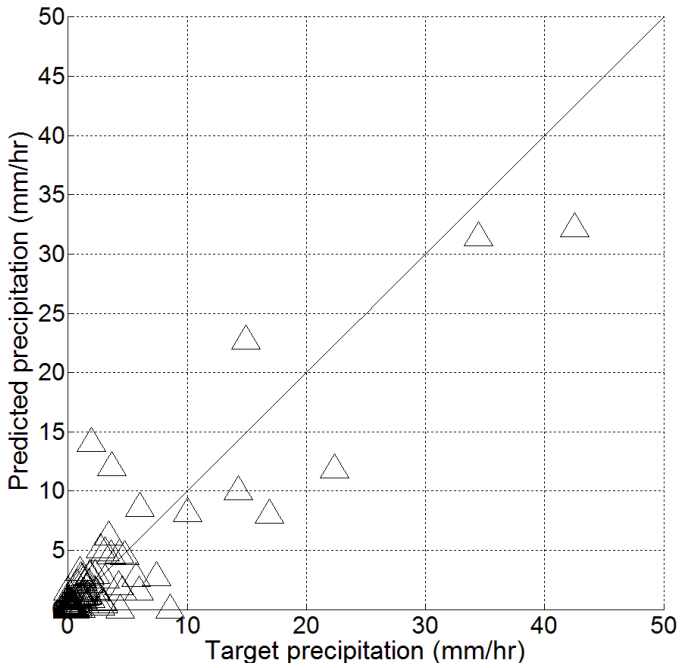

Figure 4.10. VTB scatter plot

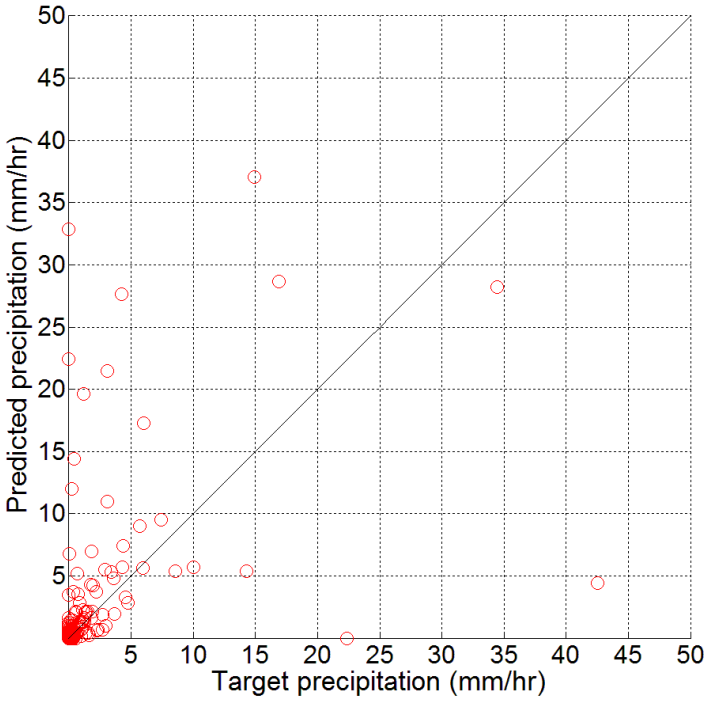

Figure 4.11. NEXRAD-III scatter plot

Table 4.6. VTB vs. NEXRAD-III

	Corr. coeff.	ME (mm/hr)	MAE (mm/hr)	RE
NEXRAD-III	0.507	0.084	0.201	0.550
VTB	0.908	-0.019	0.085	0.555

The above two figures (4.8, 4.9) show that the VTB is better able to estimate higher precipitation values than the NEXRAD-III product. It appears the NEXRAD tends to underestimate the actual precipitation value. This notion is confirmed in the above table, as the mean error (ME) of the NEXRAD product is much greater than the underestimate of the VTB.

Table 4.7. NEXRAD-III confusion matrix

		Predicted	
		Rain	No rain
Target	Rain	103	57
	No rain	103	1730

Table 4.8. VTB confusion matrix

		Predicted	
		Rain	No rain
Target	Rain	95	66
	No rain	89	1735

As an alternative way to compare the two methods a confusion matrix was built. In this case, true/false values correspond to it raining/not raining. Both the NEXRAD product and VTB perform relatively the same here. The dominance of the VTB over the NEXRAD is not due to the NEXRAD misclassifying rain events, but inaccurately estimating the precipitation rate.

4.9.2. Robustness of VTB model

With the model built with respect to the Oxford rain gauge, it is of interest to test the model at the two other rain gauge locations at South Amana and Iowa City. The radar data needs to be acquired for these two locations as before, and missing data taken care of. When testing the VTB at the South Amana location, the mean tipping bucket parameter will be the average of the Oxford and Iowa City buckets, and likewise for the testing at the Iowa City rain gauge. If the model proves to be robust at new locations, it can be situated throughout the basin with high confidence in its accuracy. The figures and tables below prove that the model is truly robust.

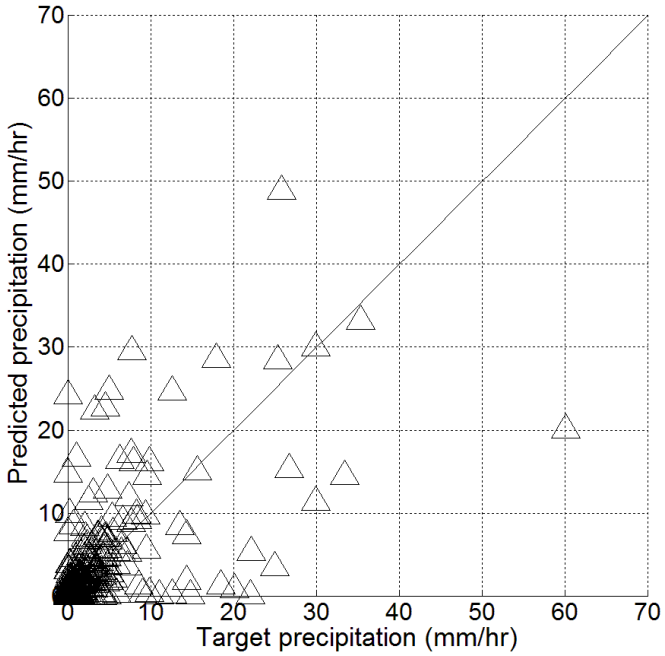

Figure 4.12. VTB at South Amana location

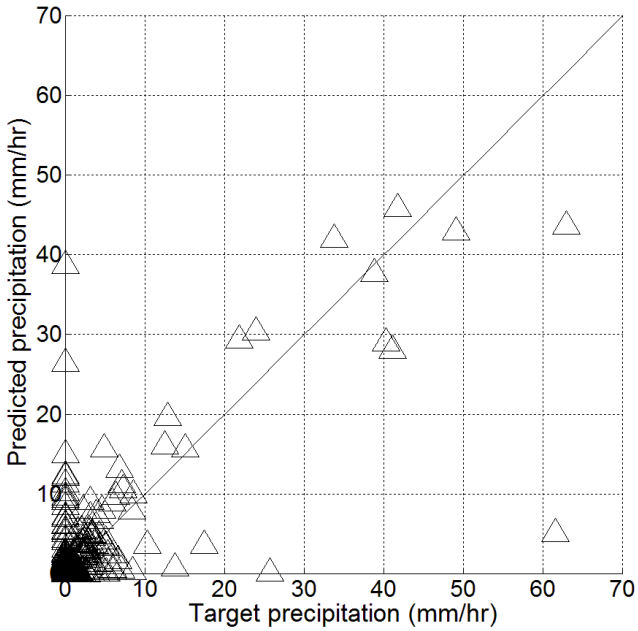

Figure 4.13. VTB at Iowa City location

Table 4.9. VTB results at South Amana and Oxford

	Corr coeff.	ME (mm/hr)	MAE (mm/hr)
South Amana TB	0.6818	0.001	0.1235
Iowa City TB	0.7614	0.0271	0.1123

Table 4.10. VTB confusion matrix at South Amana

		Predicted	
		Rain	No rain
Target	Rain	353	205
	No rain	279	6496

Table 4.11. VTB confusion matrix at Iowa City

		Predicted	
		Rain	No rain

Target	Rain	238	126
	No rain	640	6330

4.9.3. Introduction of VTB in the SWAT model

To test the effectiveness of the VTB, the same 5 months of data (5/1/2007-9/30/2007) was input into the SWAT model at three arbitrarily chosen locations. Again, the VTBs seen in Figure 4.14 as dark green triangles, do not require any actual instrumentation. They simply consider the KDVN radar reflectivity data and the three surrounding tipping buckets as input. They are capable of making estimation the precipitation at their locale with the accuracy illustrated in the above figures.

As the temporal resolution of the VTB (15-minutely) and the resolution of the SWAT model (daily) do not agree, a monthly water balance over the entire watershed was considered. The results are seen graphically and in tabular form in Figure 4.15 and Table 4.12.

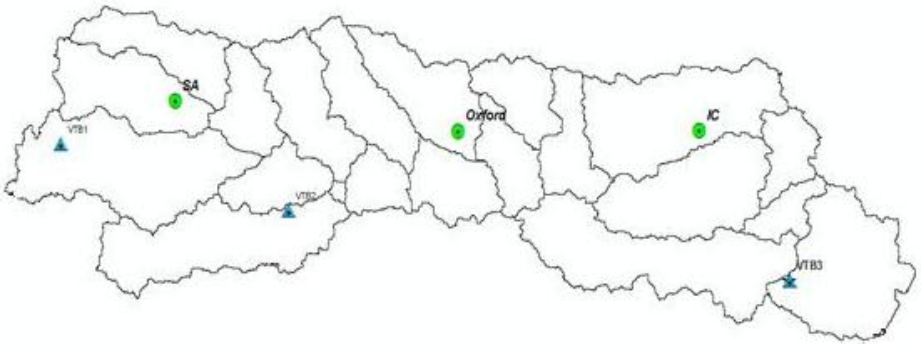

Figure 4.14 VTB and actual tipping bucket locations in Clear Creek watershed

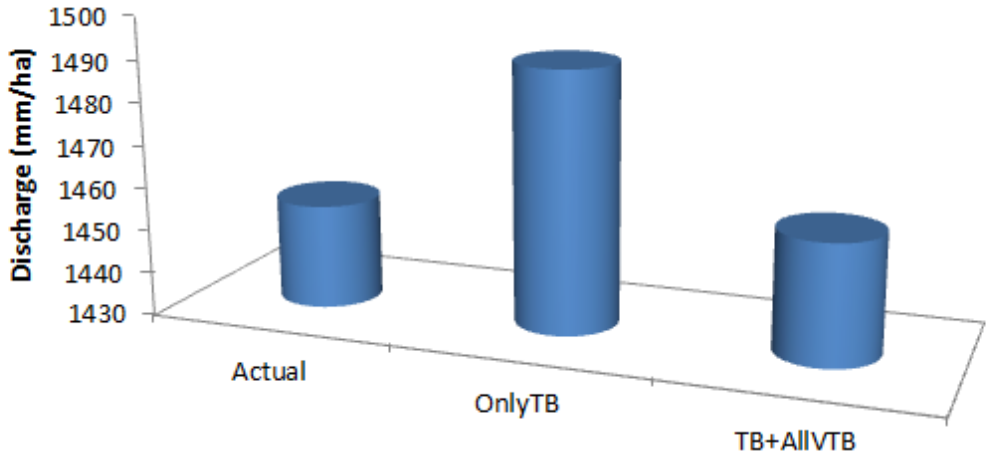

Figure 4.15 SWAT water balance results with VTB input

Table 4.12. SWAT water balance results with VTB input

Measured water balance (mm/ha)	1454.51
Only TB simulated water balance (mm/ha)	1490.91
TB VTBs simulated water balance (mm/ha)	1458.03

4.10. Discussion

The MLP is capable of estimating the rainfall rate at the ground, verified by the tipping bucket rain gauge. The two instances of high error in MLP mentioned above in the results section were reviewed. It is found that all three tipping bucket values recorded precipitation, but the radar data reported a clear sky, or reflectivity values of zero. Based on the agreement between the tipping buckets, these radar data is probably erroneous at these two points.

In general, the MLP outperforms the Z-R conversion technique used by the NEXRAD-III. The error that arises from the Z-R conversion is that it depends on the size of the rain droplets, which is different for every type of rainfall event. Rainfall droplet size may vary between storms or even within a single storm itself [48-53]. It is apparent

that the complexity of the MLP makes this data driven model more capable of modeling rainfall than the NOAA algorithm.

The wrapper feature selection algorithm with the genetic search algorithm selected both tipping bucket and NEXRAD data to be included in the model. It can be deduced that the two data sets (NEXRAD and rain gauge) both provide useful input for the model, and the strengths of both data systems, high accuracy rain gauge data, and locality specific radar data, are utilized.

4.11. Conclusion

This chapter describes the development of a multilayer perceptron trained with NEXRAD-II reflectivity data from a weather station in Davenport, IA and tipping bucket rain gauge data from South Amana and Iowa City, IA. As the model could be synced with real time radar and tipping bucket data to provide rainfall estimation in remote areas where no instrumentation exists, the resulting models have been named virtual tipping buckets (VTBs). The motivation for a system of VTBs is to provide higher resolution precipitation input for hydrological models. The VTB model was compared with the National Oceanic and Atmospheric Association's (NOAA) algorithm for converting reflectivity data to hourly precipitation, which it outperformed.

CHAPTER 5.

CONCLUSION

This Thesis explores some practical applications of data mining techniques, evolutionary computation, and heuristic search methods in the field of hydrology. Data sets considered for study included water quality parameters, discharge, radar reflectivity, and tipping bucket. Statistical analysis, in particular correlation-based analysis, was used for the selection of input parameters for the modeling challenges tackles throughout this work.

Chapter 1 provided background information and a literature review of past applications of data mining in hydrology, as well as an introduction to the multilayer perceptron (MLP), which was extensively applied throughout this Thesis.

The Second Chapter proved data mining and the MLPs competence at making a prediction at a different spatial location. In this data driven model, turbidity at a downstream location was predicted with turbidity and discharge data from an upstream location, as well as the discharge data from the downstream location. Turbidity is a particularly difficult water quality parameter to predict due to its erratic and fluctuating behavior. The MLP model derived in this chapter makes a turbidity prediction at a gauge 13 km downstream with an error less than does so with a mean absolute error less than 40 NTU.

Chapter 3 considered multiple water quality parameters, and provided a methodology toward a very practical use of data mining; data gap filling. Two methods for filling messing data are presented. One, called Type-1 modeling, considered

complimentary water quality parameters, to predict dissolved oxygen concentrations. These other water quality parameters were measured concurrently with dissolved oxygen. This method may be useful at a location that is missing a dissolved oxygen sensor, but has other water quality sensors (i.e. temperature, pH, specific conductivity, etc.). The second methodology introduced utilizes time series data mining, or the use of historical dissolved oxygen values to predict the current dissolved oxygen concentration. Both of these methods were used to model the current dissolved oxygen concentration, and alsot o make a short term forecast. The behavior of the model is analyzed when making longer term forecast, as well.

Chapter 4 combines tipping bucket data and Next Generation Radar data to build a virtual tipping bucket (VTB) model, by way of MLP. The VTB shows superior accuracy to the NEXRAD-III product for 15-minute total precipitation, which is used in some flood forecasting models. The model's robustness is analyzed as it is tested outside of its training domain, at two other locations along the Iowa River's Clear Creek. The results show that the model is highly robust. Finally, the increased spatiotemporal resolution provided by the VTB input shows to have improved the water budget results of the SWAT model.

Future research will focus primarily on the VTB implementation and experimentation within the SWAT model. The apparent usefulness of such high spatiotemporal resolution precipitation data to hydrological models, namely, flood forecasting models, makes this an exciting area of research. Some topics that will be studied in the future are (1) optimal siting of VTBs for best hydrological modeling results, (2) further robustness testing, such as testing the VTBs performance in other

regions farther away with different terrain properties, (3) the VTB's dependency on nearby actual tipping buckets for accurate prediction, and (4) determining the SWAT model's sensitivity to VTB locations, which may provide insight to the soil or hill slope properties of these locations.

REFERENCES

[1] J.V. Loperfido, "High-frequency sensing of clear creek water quality: mechanisms of dissolved oxygen and turbidity dynamics, and nutrient transport." *Doctoral dissertation,* The University of Iowa, Dept of Civ. Eng., 2009.

[2] R.A. Park, J.S. Clough, and M.C. Wellman, Park, R.A., Clough, J.S., Wellman, "AQUATOX: Modeling environmental fate and ecological effects in aquatic ecosystems," *Ecological Modelling*, Vol. 213, No. 1, pp. 1-15, 2008.

[3] R.L. Doneker and G.H. Jirka, "CORMIX user manual: a hydrodynamic mixing zone model and decision support system for pollutant discharges into surface waters", *EPA-823-K-07-001*, 2007.

[4] QUAL2K: A model for river and stream water quality, center for exposure assessment modeling (CEAM), *U.S. Environmental Protection Agency*, 2003.

[5] Q. Chen and A. Mynett, "Integration of data mining techniques and heuristic knowledge in fuzzy logic modelling of eutrophication in Taihu Lake." *Ecological Modelling*, 162, pp. 55–67, 2003.

[6] I. Westerberg, J.L. Guerrero, J. Seibert, K.J. Beven, and S. Halldin, "Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras," I, Vol. 25, No. 1, pp. 603-613, 2011.

[7] S.L Neitsch, A.G. Arnold, J.R. Kiniry, and J.R. Williams, "Soil and water assessment tool theoretic documentation, version 2005." *Grassland, Soil and Water Research Laboratory*, Agricultural Research Service, Temple, Texas, 2005.

[8] K.Y. Choy and C.W. Chan, "Modelling of river discharges using neural networks derived from support vector regression," *Fuzzy Systems*, FUZZ '03. The 12th IEEE International Conference on, 2003.

[9] S. Palani, S.Y. Liong and P. Tkalich, "An ANN application for water quality forecasting," *Marine Pollution Bulletin*, 56(9), 1586-1597, 2008.

[10] J. B. Sérodes, M. J. Rodriguez, and A. Ponton, "Chlorcast: a methodology for developing decision-making tools for chlorine disinfection control," *Environmental Modelling & Software*, Vol. 16, No. 1, pp. 53-62, 2000.

[11] G.B. Sahoo, S.G. Schlawdowa, and J.E. Reuter, "Forecasting stream water temperature using regression analysis, artificial neural network, and chaotic non-linear dynamic models," *Journal of Hydrology*, Vol. 378, pp. 325–342, 2009.

[12] X. Zhang, S. Raghavan, and M.V. Liew, “Approximating SWAT model using artificial neural network and support vector machine," *Journal of the American Water Resources Association*, Vol. 45, No. 2, pp. 460–474, 2009.

[13] C. Damle and A. Yalcin, “Flood prediction using time series data mining,” *Journal of Hydrology*, Vol. 333, No. 2-4, pp. 305-316, 2007.

[14] Y.B. Dibike and D. P. Solomatine, (). “River flow forecasting using artificial neural networks,” *Physics and Chemistry of the Earth (B*), 26, pp. 1-7, 2000.

[15] R.L Wilby, R.J. Abrahart, and C.W. Dawson, “Detection of conceptual model rainfall–runoff processes inside an artificial neural network,” *Hydrological Science Journal*, Vol. 48, No. 2, pp. 163–181, 2003.

[16] G. Tiron and S. Gosav, “The July 2008 rainfall estimation from barnova wsr-98 d radar using artificial neural network,” *Romanian Reports in Physics*, 62, pp. 305–313, 2009.

[17] R. Teschl, W. Randeu, and F. Teschl, "Improving weather radar estimates of rainfall using feed-forward neural networks," *Neural Networks*, Vol. 20, pp. 519–527, 2007.

[18] T.B. Trafalis, M.B. Richman, A. White, and B. Santosa, “Data mining techniques for improved WSR-88D rainfall estimation,” *Computers & Industrial Engineering*, Vol. 33, pp. 775–786, 2002.

[19] H. Liu, V. Chandrasekar, and G. Xu, "An adaptive neural network scheme for radar rainfall estimation from WSR-88D Observations," *Journal of Applied Meteorology*, Vol. 30, pp. 2038-2050, 2001.

[20] R. Chattamvelli, *Data Mining Methods*. Alpha Science International Ltd., Oxford, U.K., 2009.

[21] P.N Tan, M. Steinback, and V. Kumar, *Introduction to Data Mining*, Pearson Education/Addison-Wesley, Reading, Mass, 2006.

[22] T. Masters, *Advanced Algorithms for Neural Networks. A C++ Sourcebook*. John Wiley and Sons, New York, 1993.

[23] F. Rosenblatt, “The Perceptron: a Probabilistic Model for Information Storage and Organization in the Brain.” *Psychological Review*, Vol. 65, No. 6, pp. 386-308, 1958.

[24] L. Tarassenko, *A Guide to Neural Computing Applications,* Arnold Publishers, London, UK, 1998.

[25] R. Hecht-Nielsen, "Kolmogorov's mapping neural network existence theorem," *Proceedings of 1st IEEE International Joint Conference of Neural Networks*, Institute of Electrical and Electronics Engineers, NewYork, NY, pp. 11-13, 1987.

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The WEKA data mining software: an update." *SIGKDD Explorations*, Vol. 11, No. 1, pp. 10-18, 2009.

[27] J.T. Kuo, Y.Y. Wang, and W.S. Lung, "A hybrid neural–genetic algorithm for reservoir water quality management," *Water Research,* Vol. 30, No. 1, pp. 1367-1376, 2006.

[28] G. Casella, R. Berger, "*Statistical inference, 2nd edition*," Pacific Grove, CA, Duxbury Press; 1990.

[29] A.D. Back, and T.P. Trappenberg, "Input variable selection using independent component analysis," *International Joint Conference on Neural Networks*, Vol. 2, No. 1, pp. 989-992, 1999.

[30] H.R. Maier and G.C. Dandy, "Determining inputs for neural network models of multivariate time series," *Microcomputers in Civil Engineering – Journal of Computer-Aided Civil and Infrastructure Engineering,* Vol. 12, No. 5, pp. 353-368, 1997.

[31] K. Kira, and L.A. Rendell, "A practical approach to feature selection," *Proceedings of the Ninth International Workshop on Machine Learning*, Vol. 1, No. 1, pp.239-256, 1992.

[32] T.M. Fernando, H.R. Maier and G.C. Dandy Fernando, "Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach," *Journal of Hydrology,* Vol. 367, No. 3, pp.165-176, 2009.

[33] M. Robnik-Sikonja and I. Kononenko, "An adaptation of relief for attribute estimation in regression," *Fourteenth International Conference on Machine Learning*, Nashville, TN, 296-304, 1997.

[34] I. Kononenko, "Estimating attributes: analysis and extensions of relief." *European Conference on Machine Learning*, pp. 171-182, 1994.

[35] A. Kusiak, M. Li, and Z. Zhang, "A data-driven approach for steam load prediction in buildings," *Applied Energy,* Vol. 87, No. 1, pp. 925-933, 2010.

[36] M. Kantardzic, “Data mining: concepts, models, methods & algorithms.” *Journal of Comput. Inf. Sci. Eng.*, Vol. 5, No. 1, pp. 393, 2003.

[37] H. Bozdogan, “Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions,” *Psychometrika*, Vol. 52, No 3, pp. 345-370, 1987.

[38] “IA DNR: water qualitystandards,” Iowa Department of Natural Resources, 2007.

[39] I. Leetmaa and M. Ji, “Operational hindcasting of the tropical Pacific,” *Dynamics of Atmospheres and Oceans*, Vol.13, No. 3-4, pp. 465-490, 1989.

[40] I. Murray, J.W. Parsons, and K. Robinson, “Inter-relationships between nitrogen balance, pH and dissolved oxygen in an oxidation ditch treating farm animal waste,” *Water Research*, Vol. 9, pp. 25-30, 1973.

[41] S.R. Parker, C.H. Gammons, S.R. Poulson, M.D. DeGrandpre, C.L. Weyer, M.G. Smith, J.N Babcock, and Y. Oba,“Diel behavior of stable isotopes of dissolved oxygen and dissolved inorganic carbon in rivers over a range of trophic conditions, and in a mesocosm experiment,” *Chemical Geology*, Vol. 269, No. 1, pp. 22-32, 2010.

[42] S.C. Chapra, and D.M. Di Toro, “Delta Method for Estimating Primary Production, Respiration, and Reaeration an Streams,” *J. Environ. Eng*., Vol. 117, No. 5, 630-655, 1991.

[43] J. Colt, “Computation of Dissolved Gas Concentrations in Water as Functions of Temperature, Salinity, and Pressure”, *American Fisheries Society*, Bethesda, MD, 1983.

[44] T.G. Barbounis, J.B. Theocharis, M.C. Alexiadis, and P.S. Dokopoulos, “Long-term wind speed and power forecasting using local recurrent neural network models,” *IEEE Transactions On Energy Conversion*, Vol. 21, No. 1, 2006.

[45] V.E. Baer, “The Transition from the Present Radar Dissemination System to the NEXRAD Information Dissemination Service (NIDS).” *American Meteorological Society Bulletin*, Vol. 72, No. 1, pp. 29-33, 1991.

[46] J.A. Smith, and W.F. Krajewski, “A modeling study of rainfall rate-reflectivity relationsihps,” *Water Resour. Res*., Vol. 29, pp. 2505–2513, 1993.

[47] M. Steiner, R.A. Houze, and S.E. Yuter, “Climatological characterization of three-dimensional storm structure from operational radar and rain gauge data,*” J. Appl. Meteorol*. Vol. 33, pp. 1978–2007, 1995.

[48] L.J. Battan, *Radar observation of the atmosphere*, III. Univ. of Chicago Press, Chicago, 1973.

[49] P.M Austin, "Relation Between Measured Radar Reflectivity and Surface Rainfall," *Mon. Weather Rev*., Vol. 115, pp. 1053–1070, 1987.

[50] A Tokay and D.A. Short, "Evidence from tropical raindrop spectra of the origin of rain from stratiform versus convective clouds," J. *Appl. Meteorol*. Vol. 35, pp. 355–371, 1996.

[51] M. Steiner, J.A. Smith, S.J. Burges, C.V Alonso, and R.W. Darden, "Effect of bias adjustment and rain gauge data quality control on radar rainfall estimation," *Water Resour. Res*., Vol. 35, No. 8, pp. 2387–2503, 1999.

[52] S. Chumchean, A. Sharma, and A. Seed, "Radar rainfall error variance and its impact on radar rainfall calibration." *Physics and Chemistry of the Earth*, Vol. 28, pp. 27–39, 2008.