APPLICATIONS OF SOFT COMPUTING AND STATISTICAL METHODS IN WATER RESOURCES MANAGEMENT

By

Yaseen A. Hamaamin

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biosystems Engineering-Doctor of Philosophy

2014

ABSTRACT

APPLICATIONS OF SOFT COMPUTING AND STATISTICAL METHODS IN WATER RESOURCES MANAGEMENT

By

Yaseen A. Hamaamin

Water resource management is the development and use of different techniques for water system planning, development, and operation to overcome problems related to quality and quantity of water. With the increase of pressures on water resources, namely anthropogenic activity and climate change, the ability to accurately predict extreme conditions continues to be a challenge to decision makers and watershed managers. The objectives of this study were to analyze and test the ability of new modeling techniques to find robust and cost-effective models for sustainable water resource managements in both water quantity and quality fields.

Water quantity: Stream networks are the blood vessels of terrestrial and aquatic life in a watershed. Therefore, flow decreases during dry seasons and can directly impact the sustainability of ecosystem health. Index flow is the criterion that determines the minimum flow rate, which maintains and protects stream aquatic ecosystems. Therefore, this index was chosen to describe the impacts of water withdrawals on stream ecosystem health. Having the knowledge and the ability to precisely determine water withdrawals within a watershed using index flow is essential for decision makers and watershed managers. In the water quantity part of this study, various new modeling techniques were tested to find more robust approach(s) for estimating the index flow for ungaged streams in the State of Michigan. Four different techniques, linear regression, fuzzy regression, fuzzy expert, and adaptive neuro-fuzzy inference system (ANFIS),

were evaluated using a 10-fold cross validation method. Results of the study showed that the fuzzy expert (Mamdani) model was the most robust technique for modeling index flow.

Water quality: Sediment is considered the largest surface water pollutant by volume, which needs to be addressed through surface water quality planning and managements. In the planning process, different management scenarios have to be evaluated by watershed managers and stakeholders, which require multiple water quality parameter forecasting and estimation. Physically based models are considered good techniques for sediment estimations; however, they require a large number of parameters and massive calculations, especially during different management scenario evaluations. For the simulation process, the use of new cost-effective modeling approaches to reproduce the results obtained from a physically based (input intensive) models will save time and calculation efforts. In the water quality part of this study, two fusion or blend methods were created to model the sediment load for the Saginaw River Watershed. ANFIS and Bayesian Regression models were tested to find the best alternative(s) to a calibrated physically based model (Soil and Water Assessment Tool - SWAT). For these two models, four different method-types were considered and tested, namely *General*, *Temporal*, *Spatial* and *Spatiotemporal*. Both techniques, Bayesian *Spatiotemporal* and ANFIS *Spatial* models were revealed as good alternatives to the SWAT model for sediment estimations at the watershed scale (global level). However, at the subbasin scale (local level), both Bayesian and ANFIS techniques showed satisfactory results for about 50% of the total of 155 subbasins in the watershed. Transformation of sediment data improved the forecasting capability of both ANFIS and Bayesian techniques even though sediment data still had a bimodal distribution after the transformation.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor Dr. Amir Pouyan Nejadhashemi for his endless support, guidance and inspiration during every stage of my PhD program. He has always been generous with his valuable suggestions and directions from the beginning stages of my research.

I would also like to thank my committee members Dr. Steven Safferman, Dr. Timothy Harrigan, and Dr. Alexandra Sasha Kravchenko for their helpful direction and suggestions as I moved from a research idea to an accomplished study.

I am very grateful to my family as well for their never-ending love, patience, and continuous support while I pursued this degree.

My gratitude is also extended to my fellow students and graduates Zhen, Giri, Umesh, Sean, Emily, Mathew, Melissa, Fariborz, Georgina, Edwin, Andy, Matt, and Brad, for their support and friendship. It would have been a lonely lab without them.

Finally, I would like to give special thanks to everyone who contributed to this study for generously sharing his/her time and ideas.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

A	Temporal dependence covariance matrix
AND	Fuzzy operator that uses the minimum
ANFIS	Adaptive neuro-fuzzy inference system
ANN	Artificial neural networks
b_i	Parameters of a regression equation
$\hat{\beta}_j$	Posterior mean estimates for a Bayesian regression model
CAR	Conditional Autoregressive
CLIPS	C-Language Integrated Production System
DEM	Digital elevation model
D	Spatial dependence covariance matrix
DIC_4	Deviance Information Criterion
$\overline{D(\theta)}$	The posterior expected value of the joint deviance
δ^2	Measures the unexplained variation of the nugget effect
E	Expected value
ε	Residual of a regression model
F-IND	A software for the development of multivariable indices with a fuzzy approach
FIS	Fuzzy inference system
f_i	Function between different input variables of a fuzzy model
$f_{all,i}$	Sum of different functions of a fuzzy model
ø	The temporal dependence of a Bayesian model

G	The adjacency matrix
genfis1	MATLAB’s fuzzy MF generation function which uses grid partitioning
genfis2	MATLAB’s fuzzy MF generation function which uses clustering
γ	The spatial dependence of a Bayesian model
HRU	Hydrologic response unit
H-Trans	High transmissivity
HUC	Hydrologic unit code
IQ_{50}	Index flow
I	Identity matrix
IY_{50}	Index water yield
$\surd IY_{50}$	Square root of index water yield
Km	Kilometer
Km^2	Square Kilometer
L-Trans	Low transmissivity
M	The diagonal matrix
M_a (x)	The membership value (membership degree) for the element x of the fuzzy set a
MF	Membership function
MDEQ	Michigan Department of Environmental Quality
MIRIS	Michigan Resource Information System
M-Trans	Medium transmissivity
MSE	Mean squared error
NASA	National Aeronautics and Space Administration

NASS	National Agricultural Statistics Service
NHST	The null hypothesis significant test
N	Number of observation of a data set
NHD	National Hydrography Dataset
NRCS	Natural Resources Conservation Service
NSE	Nash-Sutcliffe efficiency
OR	Fuzzy operator that uses the maximum
PBIAS	Percent bias
P	Probability
p_i	Potential of a data point as a cluster center
p_i^*	A point with highest potential to be a cluster center
Precip	Precipitation
Q	The adjacency matrix
$Q_{7,\ 2}$	Average flow that is expected to be exceeded in 1 out of 2 years
$Q_{7,\ 10}$	Average flow that is expected to be exceeded on average in 9 out of 10 years
R	Real numbers
R^2	Coefficient of determination
r_a	Influential radius to define the neighborhood of a cluster of a data
r_b	Cluster radius which defines a neighborhood that has reductions in potential
RMSE	Root mean squared of errors
RSR	Ratio of the root mean square error to the standard deviation of measured data

ρ	System parameters
s	Spatial variability
STATSGO	State Soil Geographic Database
SWAT	Soil and Water Assessment Tool
σ^2	Variance of a data set
σ	Standard deviation
t	Temporal variability
T	Time serious intervals
τ^2	The variation due to geographic regions
Θ	Different operations between fuzzy subsets
u	Spatial random effect
⋃	Union mathematical operations
USDA	United States department of Agriculture
$\hat{u}_s$	Posterior mean estimate for spatial random effect in the Bayesian model
ARS	Agricultural Research Service
USGS	United States Geological Survey
v	Temporal random effect
$\hat{v}_t$	Posterior mean estimate for temporal random effect in the Bayesian model
W_i	Fuzzy weight factor (firing strength) for fuzzy rules
W_i	Normalized weight factor (normalized firing strength)
$W_{implicated}$	Product of W_i with membership value of the output variable

$W_{output,i}$	Aggregated vale of different $W_{implicated}$
y	Observed response variable
$\bar{y}$	Average of observed values
$\hat{y}$	Predicted response variable
$\hat{\bar{y}}$	Average of predicted values
$\hat{y}_{st}$	Spatiotemporal model of a variable
x_i	Predictor variables
X	A universe of sets
$\otimes$	Kronecker product operation between two matrices
ξ	Predefined factor which controls the number of cluster centers

1 INTRODUCTION

Water is necessary for life and human activities, and it must be conserved to protect the environment and natural resources. Population growth, agriculture, industry, urbanization, and climate change are the main causes for scarcity and deterioration of freshwater resources. Billions of people have no access to safe drinking water which results in 3800 children deaths worldwide every day. (Benedini and Tsakiris, 2013).

Water resource engineering is the analysis and design of water systems to manage the quantity, quality, and allocation of water to meet the needs of both humans and the ecosystem (Chin, 2013). The evaluation criteria for water systems has changed overtime, starting a century ago with a single objective criterion (safe drinking water), and recently has changed to multi objective criteria (sustainability) (Loucks and Beek, 2005). Sustainable management allows society to meets its current needs while accommodating the needs of future generations (Brundtland, 1987). Sustainability refers to renewability, resilience, and recoverability. Renewability is the rate of recharge and replacement of a resource (Harmancioglu et al., 2012):

- The renewability of a resource can be maintained by limiting the rate of use below the renewable level.
- Resilience is the ability of a resource to resist stress without permanent damage; here the system should not be overused to prevent surpassing the resilience limit of the resource.
- Recoverability is the time period required for the system to retrieve after an impact; therefore, the frequency and rate of impacts to a resource should be limited to prevent exceeding the recovery cycle rate.

Typically, a sustainable water resource development starts with information collection and is followed by planning, decision making, financing, and finally executing the planned management; preserving the sustainability of water resources and the environment (Johnson, 2008).

Monitoring stations can collect data on the quality and quantity of water in a watershed; however, no monitoring system can produce data in space and time to predict outcomes from different management plans. Therefore, to produce the required predictions, models are needed. Models can be used as complementary tools to monitored data from a watershed for management purposes. Watershed models deal with the most complex problems in water resource managements such as spatial and temporal variability (Benedini and Tsakiris, 2013). However, these models are input intensive and difficult to operate. The objective of this study was to test the ability of alternative techniques to improve the ability of predictive models for sustainable water resource managements.

This dissertation consists of two studies, which are the application of new soft computing techniques in both water quantity and quality areas.

1.1 FIRST STUDY

Low flow is the flow rate of a stream during the dry season and is considered an important criterion for water quality and quantity managements (Smakhtin, 2001 and Lee and Kim, 2008). Low flow is essential for many applications including drinking and agricultural water supply, wastewater treatment plants effluent dilution, navigation, hydropower plant operation, water withdrawals, power plant cooling facilities, and ecosystem health (Kroll et al., 2004 and WMO, 2008). The most common way to measure low flow is the $Q_{7,10}$ method, which is defined as the observed lowest 7-day average flow that is expected to be exceeded 9 out of 10 years (Eng and Milly, 2007). Since the flow rate of streams has a direct impact on aquatic organisms, the degree to which impacts can be seen are most prevalent during the dry season; and can also directly impact the sustainability of the ecosystem's health (Bubb et al., 2002; Arthington et al., 2006 and Koster et al., 2010). However, the low flow index has been found to be an ineffective criterion to protect aquatic ecosystems (Pyrce, 2004). In this regard index flow has been defined as the as the minimum monthly median flow during the months of July, August or September (Michigan's Public Act -33, 2006).

Due to the installation and operation costs of stream flow monitoring stations, it is not feasible to have a station for each stream to calculate the index flow. The common approach is to build a mathematical model to estimate index flows for ungaged sites using climatic and basin characteristics of the stream. Currently, a regression model developed by Hamilton et al. (2008) is used to estimate stream index flows for Michigan. However, the model was not validated because all the data was used for model generation and the regression technique is sensitive to extreme data points. In addition, the regression model uses a high number of explanatory variables, which requires a longer time for data collection and processing. The main goal of this study is to discover a robust and improved model to estimate index flows for ungaged streams in the State of Michigan.

The specific objectives are:

1- Identify the best method between the four techniques, regression, fuzzy regression, fuzzy expert, and adaptive neuro-fuzzy inference system (ANFIS) to predict index flows for ungaged streams in the State of Michigan.

2- Prevent over fitting during calibration and validation of models using the 10-fold cross validation technique.

3- Select the best model among the 10 models created in the previous section, for each technique.

4- Analyze the index flow prediction techniques through graphical analysis such as distribution and residual plots.

5- Compare the best index flow prediction method with the conventional regression method.

1.2 SECOND STUDY

Suspended sediment, in streams acts as a mobile substrate to transport many pollutants, which degrade the aquatic wildlife habitat, elevate water treatment costs, and reduce storage capacity of reservoirs (Ritter and Shirmohammadi, 2001 and Gunawardana et al., 2011). In a watershed, sediment loads vary both temporally and spatially resulting in a non-linear and complex relationship (Rezapour et al., 2010 and Kumar et al., 2012). Due to this complexity, regression and rating curves have been found to be ineffective techniques for sediment predictions (Kisi, 2004 and Bhattacharya and Solomatine, 2005). However, physically based models (such as SWAT) can predict sediment load satisfactorily, but these models are input intensive, need an expert operator, and require massive calculations making the evaluation process tedious and time consuming (Tayfur et al., 2003 and Bhattacharya, 2005). Consequently, these drawbacks of conventional sediment modeling approaches ultimately raise the demand for new effective modeling techniques (Cigizoglu and Kisi, 2006; Rezapour et al., 2010 and Kumar et al., 2012).

Soft computing techniques, including fuzzy logic and artificial neural networks (ANN), involve approximate functions connecting system state variables (input, internal, and output) to solve non-linear complex problems (Rai and Mathur, 2008 and Kisi et al., 2009). Probabilistic reasoning which inserts expert knowledge to the probability theory was added lately to the list of soft computing techniques as an effective tool in handle uncertainty (Huang et al., 2010). These characteristics made the soft computing techniques a viable alternative to the conventional methods. The main goal of this study is to discover the applicability of the soft computing techniques to estimate sediment load for a large and diverse watershed.

The specific objectives are:

1. Identify new cost-effective, fast modeling technique(s) to replace a calibrated and validated physically based watershed model.
2. Test four different method-types (*General*, *Temporal*, *Spatial* and *Spatiotemporal),* to find more appropriate set of data for the predicting sediment loads.
3. Analyze and evaluate the performance of the new modeling techniques at the watershed level (global model).
4. Analyze and evaluate the prediction of the new modeling techniques at the subbasins level (local model).
5. Compare the global and local performance the new modeling techniques.

2 LITERATURE REVIEW

2.1 WATER RESOURCES MANAGEMENTS

Water resources management is the development of a wide range of tools and techniques for water systems planning, development, and operation to overcome problems related to quality and quantity of water. Therefore, a holistic approach to watershed management requires integrating socio-economic, environmental, political, and engineering considerations to minimize adverse impacts on both the environment and society (Jain and Singh, 2003).

Freshwater represents about 3 percent of all the water that exists on earth, however from this small portion, the ready to use surface freshwater in streams and lakes is no more than 0.007 percent. Furthermore, this little amount of freshwater is not evenly distributed as required, some places have too much while others and too little, causing flood and drought events (Loucks and Beek, 2005). Water bodies are exposed to pollution, and due to population growth, per capita availability of freshwater has dropped to more than half during the last 30 years, furthermore climate change makes these conditions more severe. In order to reduce the severity of these extreme conditions, enhance the quality of water and also protect aquatic ecosystems, a viable water use plan and management is required (Loucks and Beek, 2005).

2.1.1 Water Systems Performance Criteria

The performance criteria for water systems have changed overtime. It started a century ago with a single objective criterion which was safe drinking water, followed by multipurpose economic development, environmental, and ecosystem protection, and recently sustainability (Loucks and Beek, 2005). Sustainable water management is using water in a way that the needs of both society and ecosystems are met (to the possible amount) now and in the future. A sustainable solution for a water management problem starts with defining the problem. Afterward tools are required to examine different scenarios and finally derived information and results are reported to stakeholders and decision makers to select the best sustainable solution (Loucks and Beek, 2005).

2.1.2 Management Parameters Estimation and Calculation

Two main approaches are available for water resources management scenarios evaluation and parameter calculation, direct measurement and indirect measurement.

Direct measurement of environmental parameters is a more realistic and correct way for measurement because it represents the physical required parameter values. However most of the time direct measurements (monitoring) are expensive, takes much time, and are constrained to current condition. Therefore, direct measurement or monitoring has been found to be ineffective for large-scale parameter estimations. On the other hand, indirect measurement (modeling) has been found to be effective for large-scale parameter estimations because it cost less, needs shorter time to draw conclusions, and can be used to examine both current and future conditions (Ozturk, 2001; Li et al. 2006 and Wang, 2008).

2.2 WATER RESOURCES MODELS

For the purpose of parameters calculation, models are used for estimation rather than observation, to reduce time and cost. To create a model, previously observed output data and other related input variables are required. The related input variables measurements must be easier and cheaper to calculate or measure compared to the measured response variable. For the required problem a suitable modeling technique and predictor variables have to be found from background knowledge and literature. Modelers may try many techniques and evaluate their performances to select the best prediction model (Wang et al., 2008).

Water resources models are challenging because of the complexity of the problem and the limitation of our current knowledge about the behavior and interaction of physical, biological, and social parameters. In order to develop a model, initially a conceptual model of the problem is required which is a non-quantitative representation of the system. Expressing the conceptual model mathematically results in a mathematical model. Finally, through calibration and validation processes, the accuracy of the model has to be tested (Loucks and Beek, 2005). Different types of modeling approaches are available for water resources modeling, such as hydraulic and mathematical models. The hydraulic (physical) models represent the water system with a smaller sized model to analyze and measure the required parameters. Mathematical models represent systems with numerical values which can be adopted to quantify different parameters and their components. Recently due the development of more powerful computers and software programs, mathematical models have become more popular compared to the other types of models (Benedini and Tsakiris, 2013).

Two main mathematical modeling approaches are available for water systems modeling, hard models (e.g. physically based and statistical models) and soft computing models (e.g. fuzzy logic and ANN) (Huang et al., 2010).

2.2.1 Physically Based Models -Hard Models

Hard computing is a precise and quantitative procedure that tries to achieve absolute optimization, which treats the data sets as a crisp value. For a hard model, such as physically based model, exact solutions are required which makes it expensive, takes longer, and, sometimes, exact solution may not be available (Li et al., 1998 and Huang et al., 2010).

2.2.2 Statistical Models

Inferential statistics generate predictions and comparisons about population (larger amounts of data) using information collected from a sample (smaller in size) of that population (Hon, 2013). Thus, an inferential statistical model defines a mathematical relationship between two sets of variables, the predictors and the output to be used for predictions and estimations (Lyman and Longnecker, 2010).

In the null hypothesis significant test (NHST), which is also known as frequentist inference, the probability of the data can be estimated depending on the trueness of the null hypothesis, In general, the problems can be stated as: what is the probability of the data, given the hypothesis (Hoff, 2009 and Lyman and Longnecker, 2010). In general, this type of model has been found to be ineffective for nonlinear, complex systems such as water resources systems models (Kisi, 2004 and Bhattacharya and Solomatine, 2005 and Hamaamin et al., 2013).

2.2.3 Soft Computing Models

Soft computing is an automated intelligent estimation to solve complex problems by accepting good enough results, which may not be solvable by hard computing. Soft computing provides approximate solutions with low cost of calculation and is faster compared to hard models. It can solve complex problems approximately if exact solution is not available or too complex to analytically determine. During the last three decades different soft modeling methods have been developed which are the integration of biological structures and computing techniques such as artificial neural networks (ANN), genetic algorithms, fuzzy logic, and fusion methods (Li et al., 1998; Pomerantsev and Rodionova, 2005; Nayak et al., 2005 and Huang et al., 2010).

With increases in pressure on water resources, along with growing attention to climate changes and anthropogenic activity, the ability to accurately predict water resources conditions continue to be invaluable (Smakhtin, 2001). However conventional methods may not remain cost-effective or robust compared to alternative soft computing methods (Huang et al., 2010).

2.2.3.1 Artificial Neural Networks

Artificial neural networks (ANN) system updates its structure based on external and internal information that feed to the system network during the learning phase. This mimics the working of the human brain neurons network which learns from experience without any previous knowledge or information on the system (Jain and Martin, 1998 and Nagy et al., 2002).

2.2.3.2 Genetic Algorithms

Genetic algorithms use the concept of Darwin's theory of evolution, which is an optimization heuristic robust tool to search for global optimum solutions without being stuck in local optimum points. The technique inspired by evolutionary biology which depends loosely on the natural evolution principles known as mutation and recombination. The algorithm selects individuals from the population randomly and uses them as parents to produce the next generation children (Mitchell, 1999 and Huang et al., 2010).

2.2.3.3 Bayesian Methods

Bayesian is probabilistic reasoning method, which combines skills and beliefs from the obtained information with the probability theory to handle the uncertainty of the inference system (Huang et al., 2010). In contrary to NHST systems, Bayesian systems assume that the model parameters are random. In other words, it adds expert knowledge by revising the probability and distribution of parameters depending on the obtained data (Hoff, 2009 and Kruschke, 2010). To model the relationship between parameters (ρ) and variables (y), Bayes theorem calculates the probability distribution of the parameters given the data as shown in Equation 2-1 (the posterior is proportional to the prior times the likelihood) (Hoff, 2009 and Kruschke, 2010).

$$P(\rho \mid y) = \frac{P(\rho)P(y \mid \rho)}{P(y)} \qquad (2\text{-}1)$$

Where, *P* is the probability.

Several applications of Bayesian methods have been recorded lately as new successful modeling techniques for complex nonlinear water resource systems (Schmelter et al. 2011 and 2012 and Leisenring and Moradkhani, 2012).

2.2.3.4 Fusion Methods

Fusion or hybrid methods are advancement in soft computing which combines two different soft computing techniques to improve system performance and offsets the disadvantages of one by the advantages of the other method. ANFIS, which is combination of fuzzy logic with ANN, is one of the successful examples of hybrid soft computing method that has a wide use in different modeling fields (Jain and Martin, 1998 and Huang et al., 2010).

2.2.3.5 Fuzzy Logic

Fuzzy logic models uses approximate reasoning (IF-THEN) rules, which works like human decision making “rule of thumb.” The final product is a graded statement rather than strict results such as true or false (Jain and Martin, 1998 and Sen, 2010).

2.3 APPLICATION OF FUZZY LOGIC IN WATER RESOURCES

2.3.1 Application of Fuzzy Logic in Water Quantity Modeling

In recent years, many applications of fuzzy logic modeling have entered into the hydrological and water resources domain as an alternative and powerful method for modeling. For example, fuzzy computing techniques have been extensively used for flow rate and flood modeling (Nayak et al., 2005; Deka and Chandramouli, 2005; Toprak, 2009; Al-Zu’bi et al., 2010 and Nayak, 2010). Altunkaynak et al. (2005) used Sugeno fuzzy modeling to predict urban water consumption amounts and had fewer errors compared to the auto regressive models observed. Mahabir et al. (2003) found considerably superior water supply forecasting for a fuzzy logic system compared to the regression equations. Fuzzy models were also observed to perform better for flow estimation at poorly gaged sites by Toprak et al. (2009). In addition, ANFIS hybrid models have been used for different applications of studies including flood estimation, missed rainfall data, water level predictions, and predicting evaporation (Chau et al., 2005; Mahabir et al., 2006; Kisi, 2006 ; Shu and Quarda, 2008; Tzimopoulos et al., 2008; Keskin et al., 2009; Cobaner, 2011 and Sanikhani and Kisi, 2012). Further studies have shown similar results with fuzzy models being a suitable alternative for modeling non-linear systems such as rainfall and runoff (Hundecha et al., 2001; Jacquin and Shamseldin, 2006 and Ghalhary et al., 2009).

2.3.2 Application of Fuzzy Logic in Water Quality Modeling:

Fuzzy logic is a powerful modeling technique and a better alternative to many former methods for river suspended sediment modeling using different fuzzy inference systems such as ANFIS fuzzy hybrid techniques (Kisi et al., 2006; Rajaee et al., 2010; Kisi et al., 2009 and Kisi, 2010). Azamathulla et al. (2012) found that the ANFIS was the best technique in predicting sediment transport in storm sewer pipes comparing to the regression model. In addition, fuzzy logic and ANFIS inference systems are being employed successfully to control wastewater treatment processes in comparison to the traditional control systems (Carrasco et al., 2004; Abdou et al., 2008 and Fan and Boshnakov, 2010). The superior performance and capabilities of the fuzzy logic water quality index in dealing with non-linear, complex, and uncertain systems is assessed through a comparison with other water quality indices such as the water quality index suggested by U.S. National Sanitation Foundation (Nasiri et al., 2007; Lermontov et al., 2009 and Bai et al., 2009). Mah (2011) used ANFIS successfully to assess the health and functionality of wetlands. El-Sebakhy et al. (2007) found ANFIS as a better predictor for bacteria growth comparing to regression and multilayer neural networks models.

2.4 HISTORY AND CONCEPT OF FUZZY LOGIC

2.4.1 Historical Root of Fuzzy Logic and Crisp Logic

Historically the fuzzy logic idea originated from Buddhism. Buddha's philosophy was based on the idea that the world is made up of contradictions in such a way that nearly everything has some of its opposite, for example, a thing can be X and not-X at the same time (kosko, 1993). Conflicting to Buddha, Aristotle believed that the world is made up of opposites, such as hot and cold, active and passive, and male and female. Everything has to be either X or not-X, this idea resulted in crisp logic (classical logic) (Kosko, 1993).

2.4.2 Fuzzy Logic Concept

The idea, and the name “fuzzy”, was introduced by Zadeh in 1965, the pioneer and the developer of fuzzy logic after publication of his very first fuzzy logic paper (Zadeh, 1965). A decade after, the Mamdani fuzzy inference system is the first fuzzy control systems presented, which make the fuzzy concepts famous and popular afterwards (Mamdani, 1974; Mamdani and Assilian, 1975). Now, fuzzy logic is being used in many science and technology fields for modeling or controlling (Kisi, 2006). A word that indicates numerical values is a fuzzy word, which can be defined as linguistic variable. Fuzzy words examples include rainfall, runoff, temperature, discharge, force, porosity, acceleration, and permeability. Examples of non-fuzzy words are lake, sea, water, cow, confusion.

2.4.3 Advantages of Fuzzy Logic Tools

Fuzzy logic has a wide field of applications; it can find certain conclusions from imprecise, approximate and vague information, which resembles human decision making (Sen, 2010). When using a fuzzy logic approach for modeling, a sound knowledge of the physical processes which occurs in the system is not necessity. For fuzzy logic models only qualitative relationship between system inputs (cause) and outputs (effect) are required. Therefore describing the mathematical relationship among the system parameters is not necessary. In a physically based model of a complex system, some parameters may not be included explicitly in the model due to the lack of understanding of the relationship between them; those parameters can be included as fuzzy rule arguments easily in a fuzzy modeling process (Hundecha et. al., 2001).

A membership function (MF) is the function that defines belongings (association value) of elements to the set. The difference between fuzzy logic and probabilistic approaches is that fuzzy membership functions express the outcomes as a possibility instead of a likelihood of a probability density function. Accordingly, during modeling processes any non-possible scenario can be easily canceled by removing the related fuzzy rule (Mckone and Deshpande, 2005). In contrast to the other statistical approaches, fuzzy logic method performance is not dependent on the volume of available data because their outcomes depend on linguistic rules. In addition, for a small data set, outliers do not have a large influence on the outcome of the fuzzy model (Mahabir, 2003).

2.4.4 Fuzzy Numbers

A general mathematical definition of a fuzzy set is a set of ordered pairs of numbers as shown in equation 2-2.

$$d = \{(x, M_d(x)) \mid x \in R\} \tag{2-2}$$

Where d is the fuzzy set. M_d (x) is the membership value (membership degree) for the element x of the fuzzy set, and R is real numbers **(**Zimmermann, 2001).

2.4.5 Classical Sets versus Fuzzy Sets

A collection of elements in a universe is called a “set.” Elements are related to the set by a MF. In order to understand fuzzy sets, a description of weaknesses and problems related to traditional crisp sets is also required.

Fuzzy sets were introduced by Zadeh (1965) for the first time to represent vagueness and uncertainties in classical crisp sets. Fuzzy sets are generalization of crisp sets. Crisp sets always have unique function to describe the belonging of elements of the sets which always looks like a rectangle over the range of data points. On the other hand fuzzy sets have infinite functions to describe the belonging of the elements to the set (Sen, 2010).

In classical sets (crisp sets), there is no uncertainty of the boundary locations (Figure 2.1). In other words, each element either belongs to the set with degree of membership *M*(*x*) of "1" or does not belong to the set with degree of membership of "0". Therefore, crisp sets have no ambiguity in their membership. The statement of Figure 2.1 can be expressed mathematically as in equation 2-3, where a and b are boundaries of the crisp set:

$$M(x) = \begin{cases} 1 \;\; if \;\; a \le x \le b \\ otherwise \end{cases} \tag{2-3}$$

Fuzzy sets, on the other hand, are defined by their vague properties; hence the boundaries are specified ambiguously. The degree of membership of a fuzzy set will be valued between "1" with fully belonging to the set and gradually changes depending on the type of the selected MF to " 0" with none belonging to the set, a triangular MF shown in figure 2.2 (Sivanandam et al., 2007 and Sen, 2010). Fuzzy set elements can also be members of other fuzzy sets in the same universe (Ross, 2004).

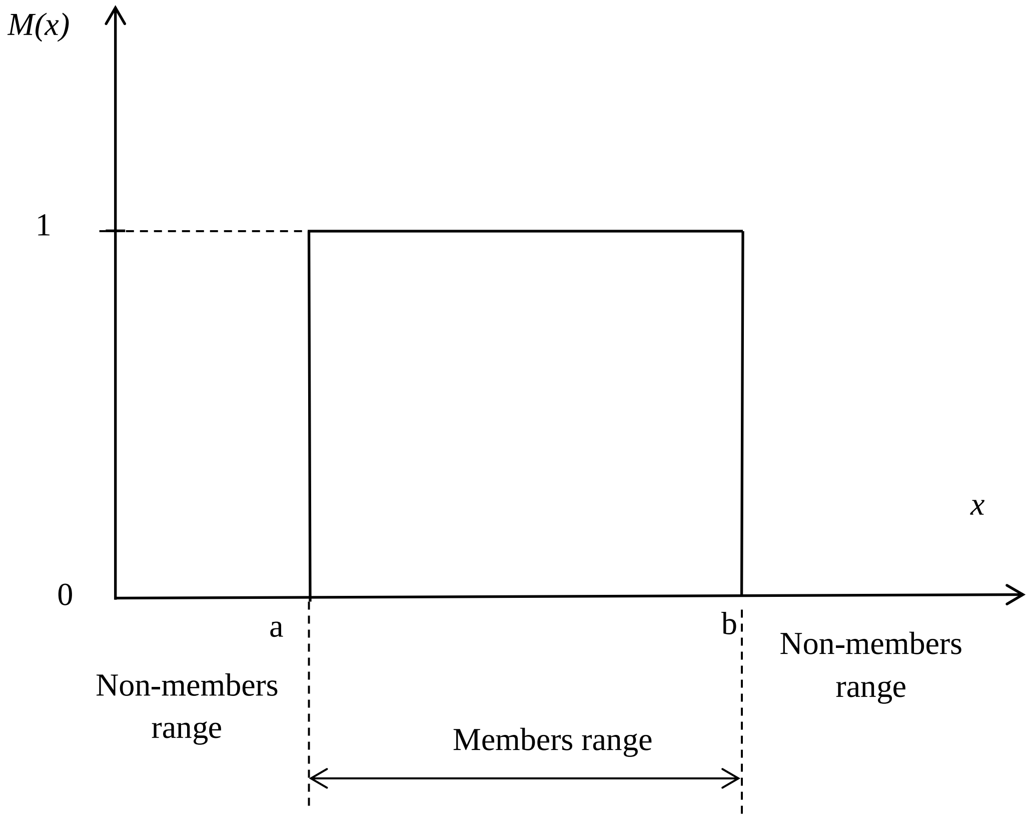

Figure 2.1. Crisp set MF

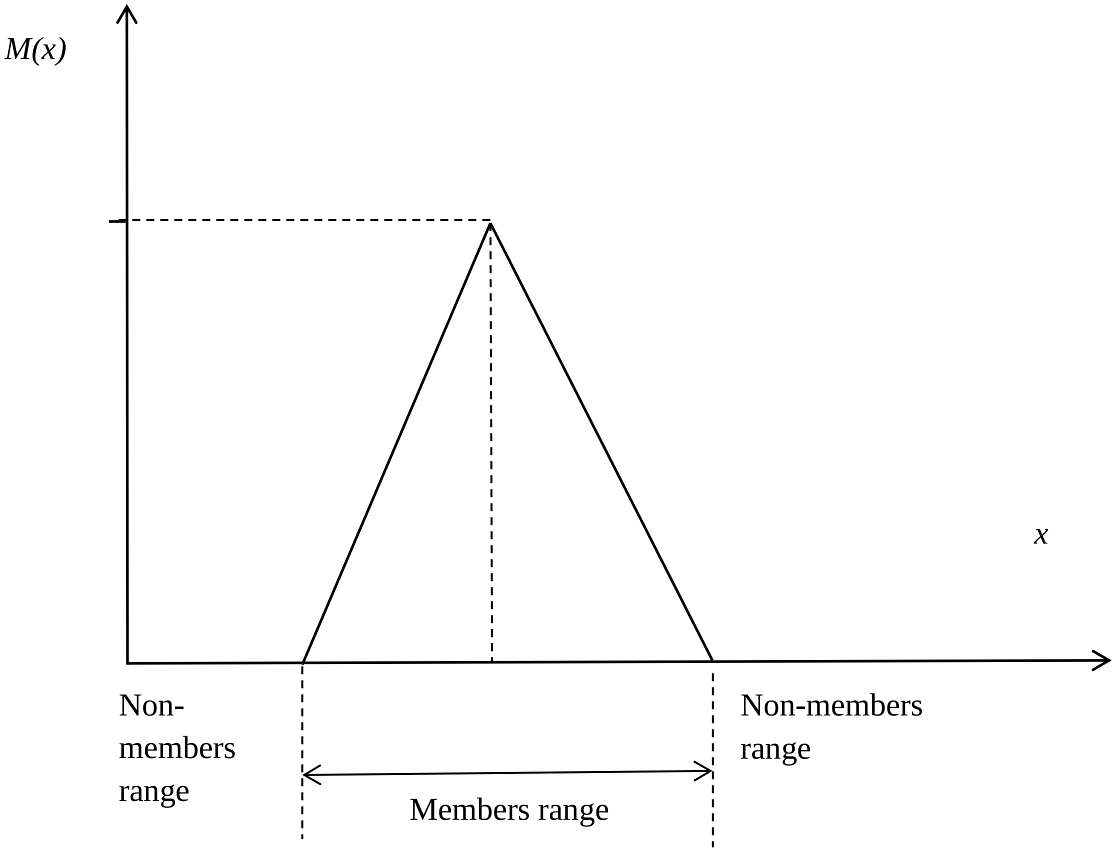

Figure 2.2. Fuzzy set MF

2.4.6 Logical Operations on Fuzzy Sets

The most common operations applicable for both crisp and fuzzy sets are complement, equality, intersection, and union. However a large number of other operations can be applied to fuzzy sets which are not applicable to crisp sets (Ross, 2004; Klir and Yuan, 1995).

2.4.6.1 Complement

The complement $\overline{d}$ of a fuzzy set *d* in the universe *X* is defined for all elements $x \in X$ as in equation 2-4.

$$\overline{d}(x) = 1 - d(x) \tag{2-4}$$

2.4.6.2 Equality

Two sets *d* and *e* are equal if and only if their MFs mappings are equal as in equation 2-5 over universal set of *X*.

$$M_d(x) = M_e(x) \tag{2-5}$$

Where $M_d(x)$ and $M_e(x)$ are membership degrees from both fuzzy sets *d* and *e* respectively.

2.4.6.3 Intersection

The intersection (AND-ing) of two fuzzy sets *d* and *e* ($d \cap e$) is defined for all $x \in X$ by equation 2-6.

$$(d \cap e)(x) = \min[M_d(x), M_e(x)] \tag{2-6}$$

Where “min” denotes the minimum operator, minimum operator is the lower value of membership degree between *d* and *e* to be selected. In linguistic logic (fuzzy rules) the “AND” operator is used for minimum operator.

2.4.6.4 Union

The union (OR-ing) of two fuzzy sets *d* and *e* ($d \cup e$) is defined for all $x \in X$ as in equation 2-7.

$$(d \cup e)(x) = \max[M_d(x), M_e(x)] \tag{2-7}$$

Where “max” denotes the maximum operator, which is the higher value of the two membership degrees *d* and *e* in numerical calculation. In linguistic logic this maximum is expressed with the “OR” operator in building fuzzy rules.

2.5 FUZZY MEMBERSHIP FUNCTION

A MF applies a fuzzy value (degree of membership) to the elements of a set from complete exclusion (0) to absolute inclusion (1), Figure 2.3. The membership value describes the degree of belonging of the element in the fuzzy set. The change from belonging to not belonging to the set is gradual, which gives a good way to handle the uncertainty and overcomes a main flaw of crisp sets (Al-Muhseen, 2009). MF can be functional or graphical and the shape or type of MF is an important criterion that has to be considered in fuzzification process.

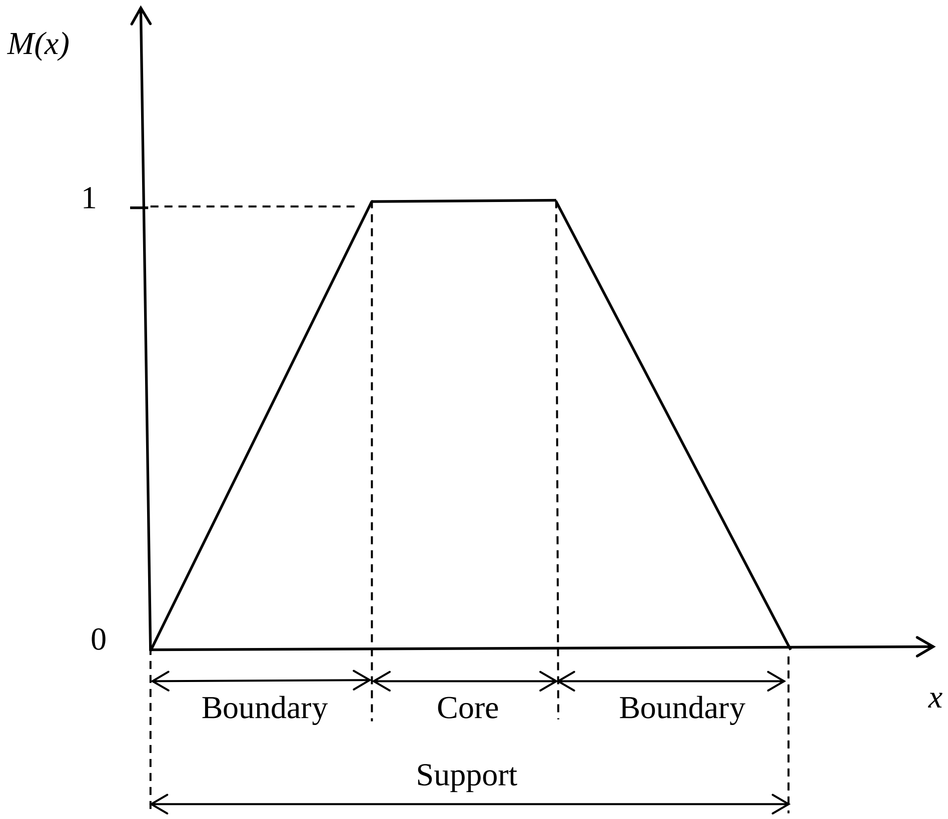

Figure 2.3. Properties of membership function.

Sivanandam et al. (2007) defined MF by three properties, core, support and boundary. For the trapezoidal MF shown in Figure 2.3, the core is that portion or region of data set with a a MF of “1”. Support is the overall region that has membership value more than “0”. Boundary is the region with membership more than “0” and less than “1” excluding the region with exact value of “1”.

2.5.1 Generating MFs

Sen (2010) addressed the most common MF as being triangular shape; however, other types are also applicable. The following criteria can be considered to generate MFs and an example of generating three triangular MFs of *l, m* and *h* for variable *x* is shown in Figure 2.4:

- In general, same MF shapes are used to represent data of a variable.
- Membership functions are consistently distributed over the variable range.
- Adjacent MFs cross each other at membership values of 0.5.
- Input and output variables may have the same or different shapes of MFs.

MFs can be developed using two main approaches, subjective and objective groupings.

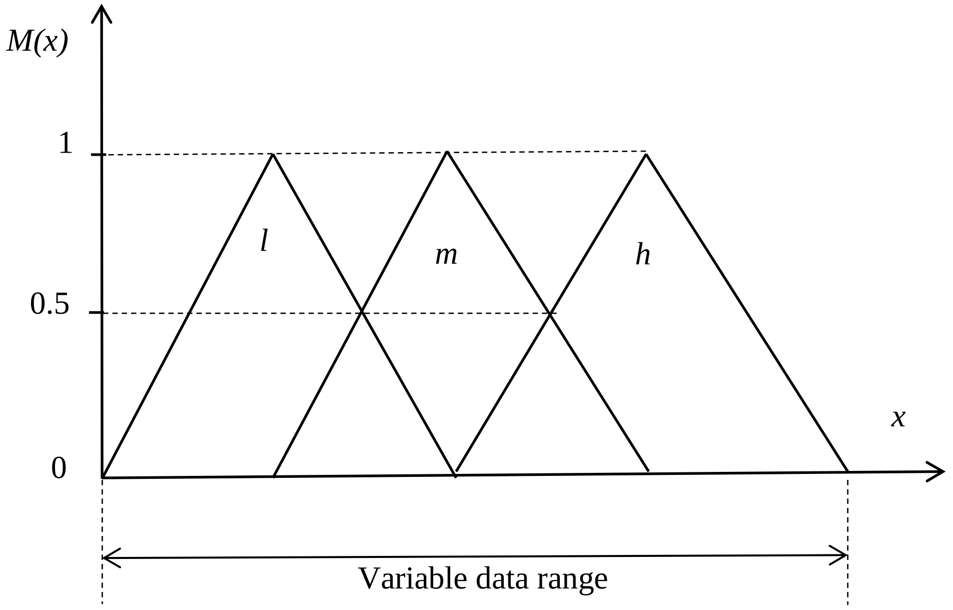

Figure 2.4. Generated triangular MFs for a variable

2.5.2 Subjective Grouping

Representing data by MFs is challenging because most of the time the chosen type and number of MFs are not representing the actual data distribution for input and output variables. An easy and approximate way for building MFs is to plot data available for each input variable separately against the output variable as shown in Figure 2.5. From the scatter plots membership functions can be derived or assumed by visual interpretations. Each cluster subset of data points can be assumed as a subjective MF and the range of each subset and its overlap with the other clusters can be determined. The value at the center of each group represents the membership degree of "1" for that MF (Sen, 2010 and Cobaner 2011).

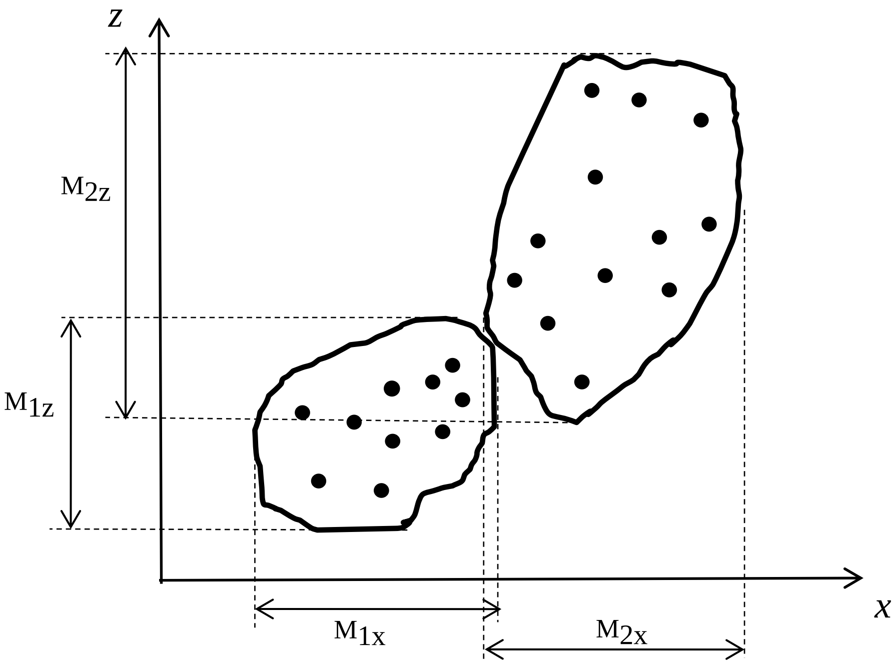

Figure 2.5. Allocation of membership functions from graphical plotting

2.5.3 Objective Grouping

In this method the data range is divided into different clusters depending on the resemblance between data points using clustering algorithm, Figure 2.6 .The center of the cluster has membership degree of "1". Each point has deviation depending on its distance from the center, which refers to the point membership degree to that MF.

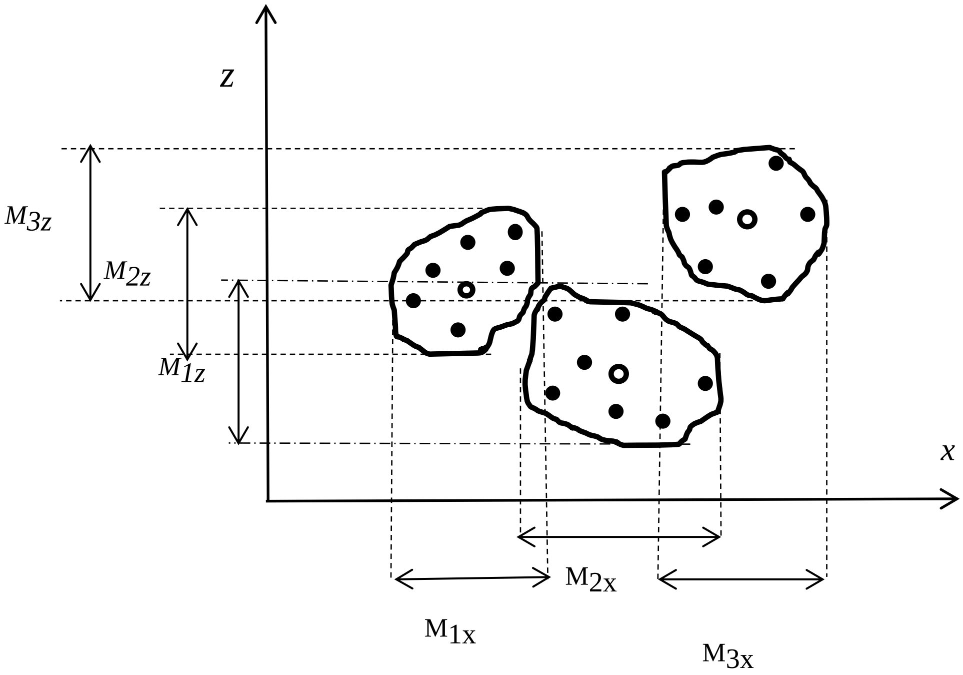

Figure 2.6. Allocation of MFs from cluster analysis

2.6 TYPES OF MFs

MFs can be chosen either arbitrarily or manually based on the modeler’s experience. Membership functions created by two different users might be different depending on their perspectives and knowledge. Membership function can be designed or created automatically by using learning methods such as ANFIS. Membership functions can take any shape or function such as triangular, trapezoidal, sigmoid, piecewise-linear, Gaussian, two-piece Gaussian, and bell-shaped. Following are the most used MFs.

2.6.1 Triangular MF

The simplest forms of MF and it can be defined using three parameters (a, b and c) as shown in Figure 2.7. Equation 2-8 shows the mathematical expression of a triangular MF.

$$M(x) = \begin{Bmatrix} 0 & if\ x \le a \\ \dfrac{x-a}{b-a} & if\ a \le x \le b \\ \dfrac{c-x}{c-b} & if\ b \le x \le c \\ 0 & if\ x \ge c \end{Bmatrix} \qquad (2\text{-}8)$$

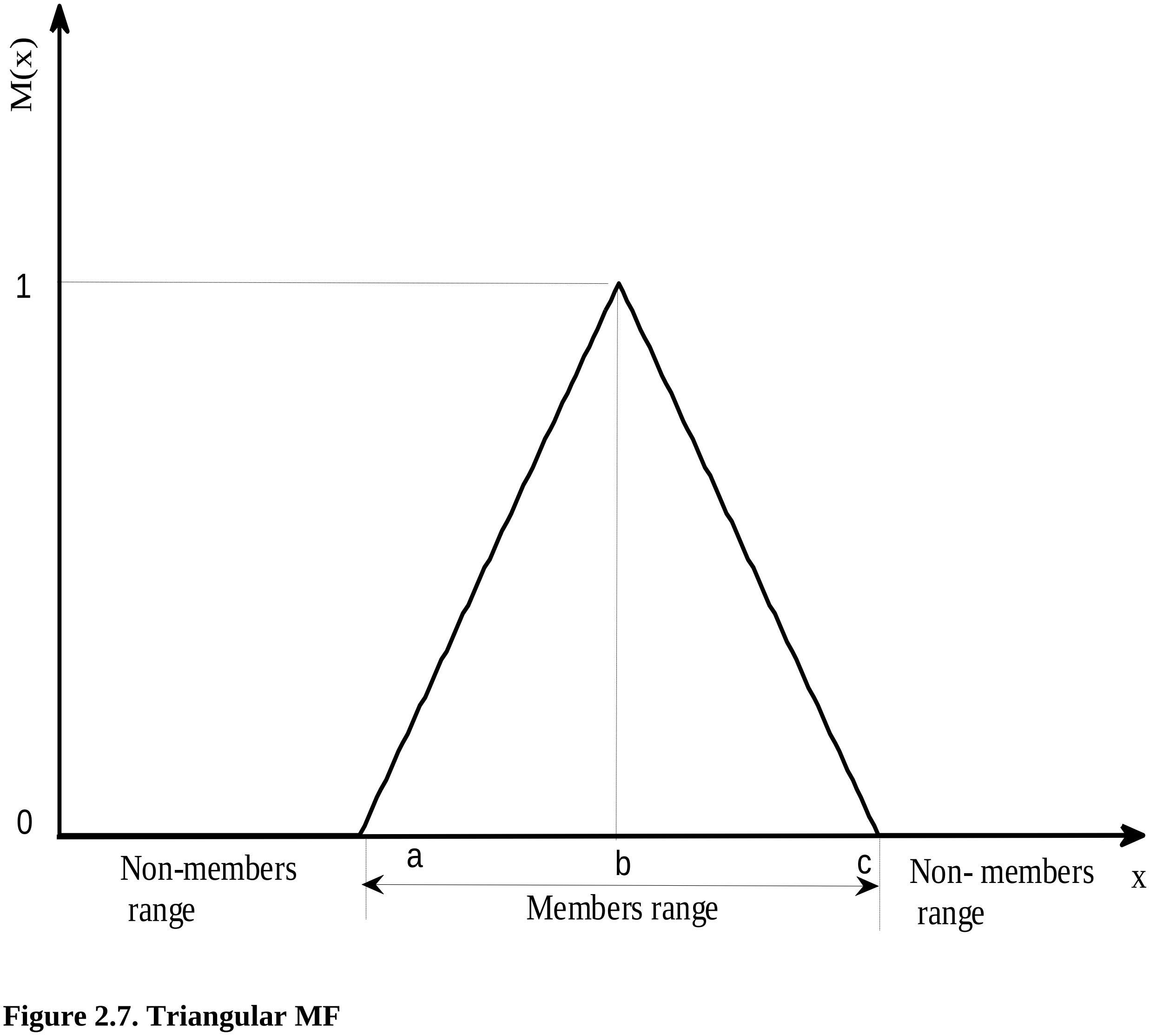

Figure 2.7. Triangular MF

2.6.2 Trapezoidal MF

Trapezoidal MFs (Figure 2.8) can be defined by four parameters (a, b, c and d). The function can be expressed mathematically as in equation 2-9.

$$M(x)=\left\{\begin{array}{ll} 0 & if\ \ x\le a \\ \dfrac{x-a}{b-a} & if\ \ a\le x\le b \\ 1 & if\, b\le x\le c \\ \dfrac{d-x}{d-c} & if\ \ c\le x\le d \\ 0 & if\ \ x\ge d \end{array}\right\} \qquad (2\text{-}9)$$

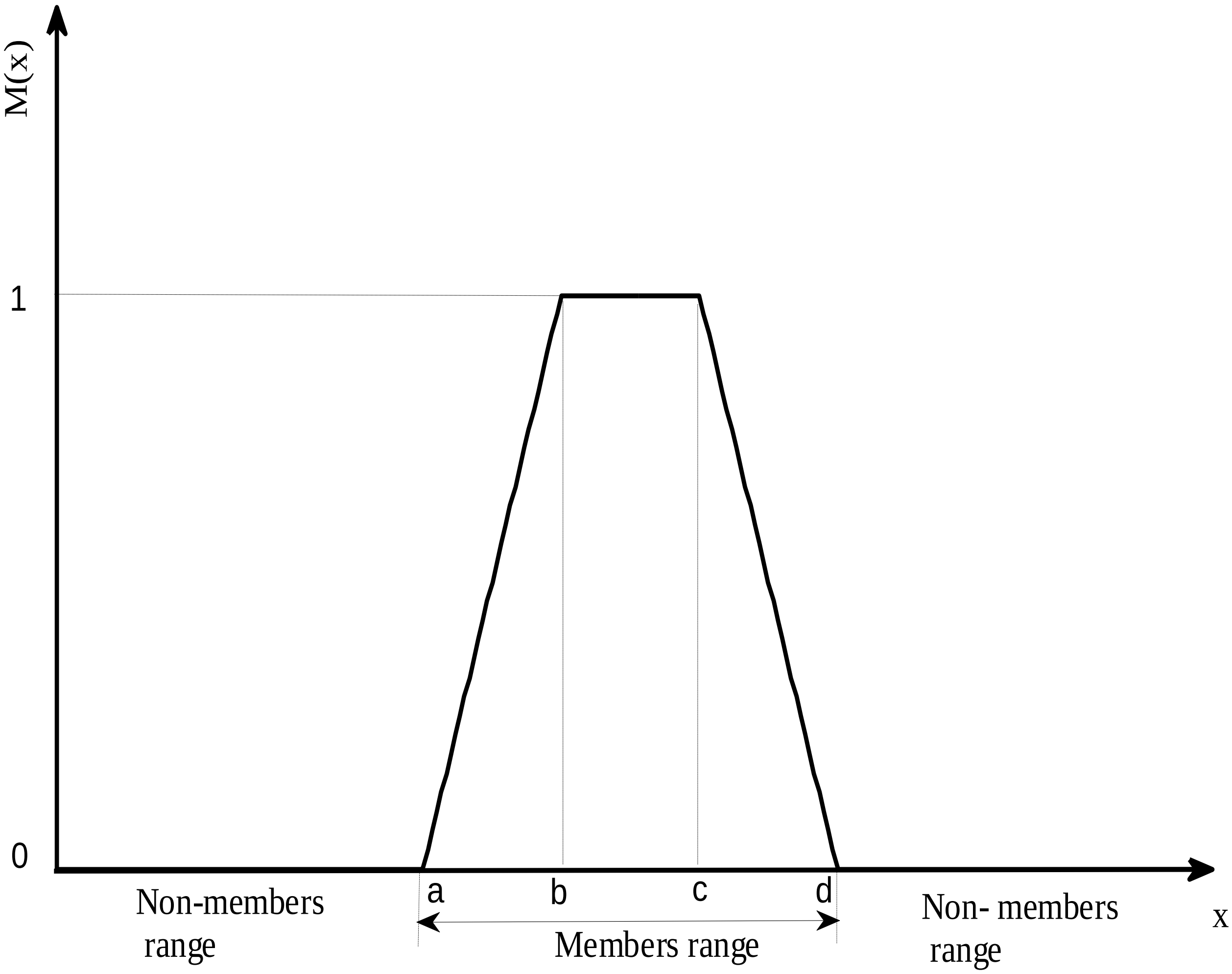

Figure 2.8. Trapezoidal MF

2.6.3 Gaussian MF

Gaussian MF is shown in Figure 2.9. The function can be defined using two parameters, the spread or standard deviation (σ) of values around the mean of values (center of the curve, c) where the function has value of (1). In general Gaussian MF can be represented mathematically as in equation 2-10.

$$M(x) = e^{-\frac{(x-c)^2}{2\sigma^2}} \tag{2-10}$$

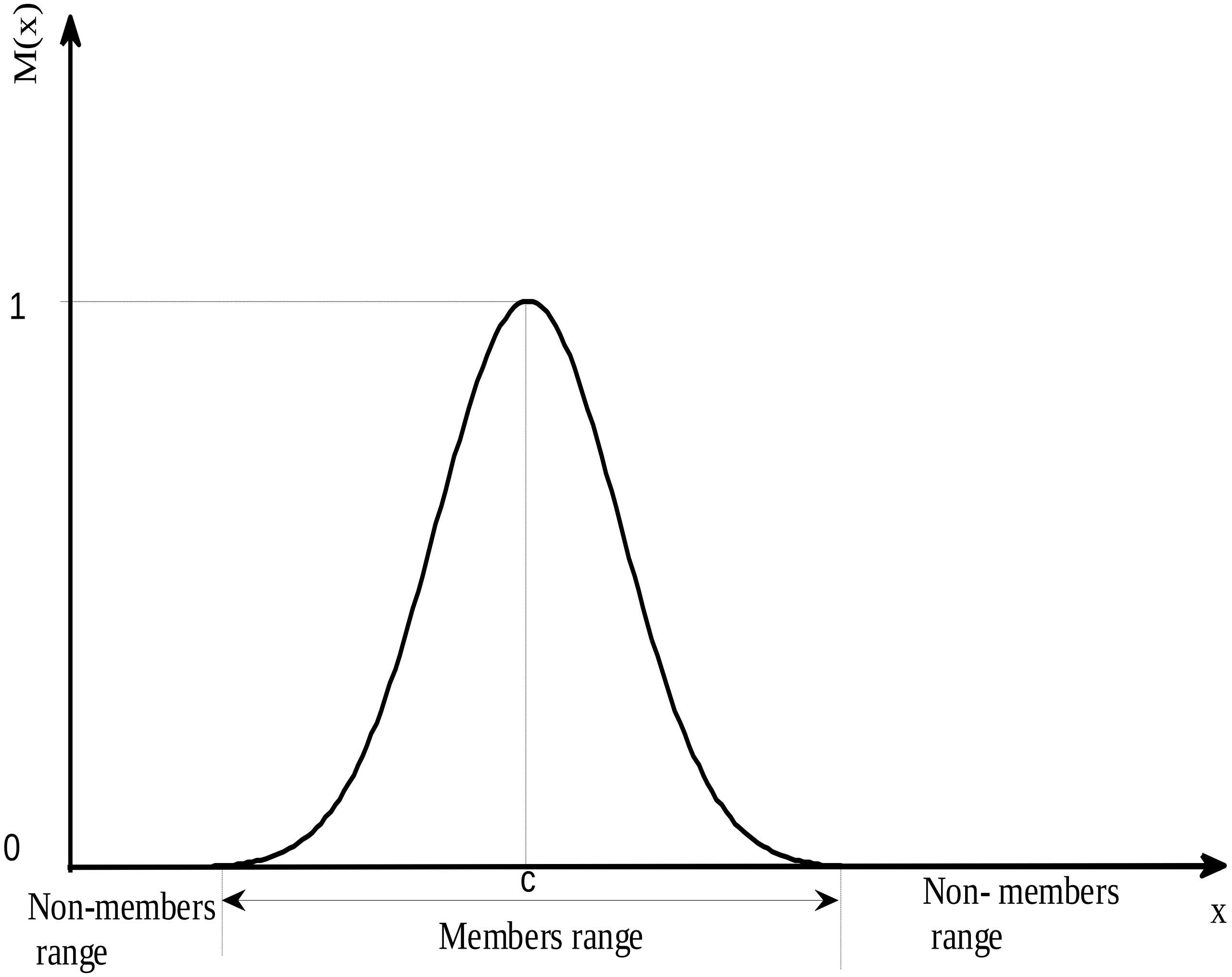

Figure 2.9. Gaussian MF

2.6.4 Bell-shaped MF

Bell shaped curve MF (Figure 2.10) can be defined using three parameters (a, b, and c) as in equation 2-11.

$$M(x) = \frac{1}{1 + \left|\frac{x-c}{a}\right|^{2b}} \tag{2-11}$$

Where c is the center of the curve, and a and b are defining the shape of the curve.

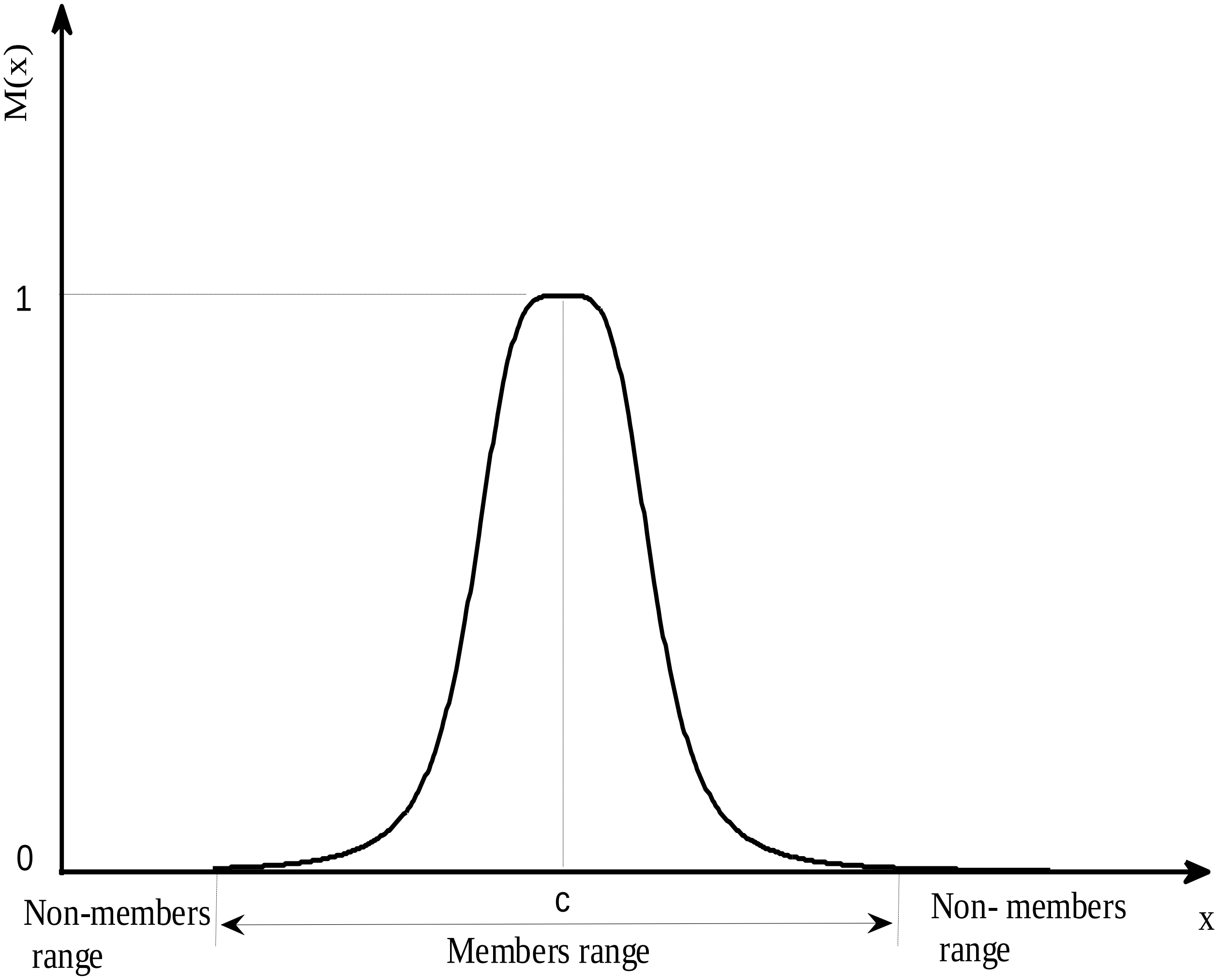

Figure 2.10. Bell shaped MF

2.7 FUZZY INFERENCE SYSTEMS

Fuzzy modeling generally is a non-linear mapping function from input space to the output space which applies a inexplicit reasoning tool. The general functional blocks for a fuzzy system are shown in Figure 2.11. In building a fuzzy expert system the crucial steps are fuzzification and building rules. These steps can be designed by two available approaches. The first approach is by using information from expert or background knowledge about the problem. The second approach is using machine learning such as neural network, ANFIS and genetic algorithms to learn about MFs and fuzzy rules. The first method does not use past history of the available data to learn from, it uses built-up information and experience by experts in that filed. The second method learns from the past data available and uses that information for future predictions. Moreover the first method of building a fuzzy model which is known as fuzzy reasoning, works well when not enough observed data is available while the behavior of the system is known from background knowledge about the problem. The second method of automatic learning such as ANFIS is useful when the background knowledge about the structure and behavior of the system is not available (Jang, 1993; Nayak et al., 2005 and Sivanandam et al. 2007).

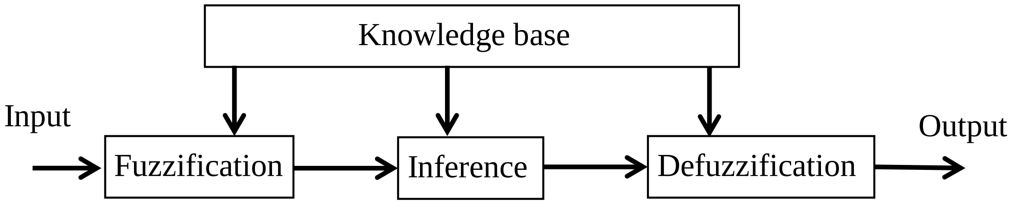

Figure 2.11. Functional blocks of a fuzzy inference system

2.8 STEPS IN FUZZY LOGIC MODELING

Fuzzy reasoning describes the behavior of a system based on available data, by generating mechanisms of the system and its processes based on inputs (antecedents) and outputs (consequents). Fuzzy reasoning also considers expert knowledge that provides insights about the system (Sen, 2010). The following are the five steps for a successful application of a fuzzy model.

2.8.1 Fuzzification

Fuzzification is the first step of fuzzy modeling that includes building MFs which are curves that define how each point of input and output variables is mapped to a fuzzy degree of membership between 0 and 1. Each variable is divided to a number of subsets randomly or, depending on background knowledge, automatically as in ANFIS. In this stage input variables data points mapped into MFs and a fuzzified value(s) obtained to be used for the next stage of calculation. In general the type of function used for the MF is an arbitrary process by testing different types of functions and finding the most appropriate one, which is better represent the variability of the data points. Usually a suitable linguistic name is used for a set of MFs of a fuzzy set, such as low, medium and high or cold, moderate, warm and hot. (Toprak et al., 2009). Eldin et al. (2004) suggested the most well-known MFs for real life data are triangular, trapezoidal, and Gaussian functions.

2.8.2 Inference Using IF-THEN Rules

IF-THEN rules can be built based on expert knowledge, available literature or learning from the data set, the IF part of each rule can be calculated in this step. The inference step (rules product) assigns one or more input fuzzy MF (subset) to an output variable subset in each rule.

Mostly, each input variable data point falls within two membership functions. In cases of more than one input variable, each of these membership values is used in a different rule with the membership value of another variable. Using different operators between input variables membership values results in a single fuzzy membership value as an output of this step of calculation to be used in the next step of calculation. The "IF" part of fuzzy rules is related to input variables membership functions and is called premise or antecedent; the "THEN" part, which is related to the output variable membership functions and is called the conclusion or consequent part of the rule (Cobaner, 2011).The number of fuzzy IF-THEN rules in a system depends on the number of membership functions and the number of input variables. In general, for a system, the number of fuzzy IF-THEN rules equals the sum of product of multiplying each input variable by the number of its MFs, however sometimes some rules can be eliminated due to their physical unavailability.

2.8.3 Implication of Fuzzy Rules

The implication stage calculates the THEN part of each fuzzy rule using the result of the previous step. Each fuzzy rule assigns input(s) membership function(s) to a single output fuzzy membership function. The single result of the input part of each rule is used to calculate membership value or to fire the output membership function.

2.8.4 Aggregation of Different Rules Outputs

Aggregation is the process of collecting all the possible outputs from inputs mapped values. From the previous stage different outputs are obtained from each rule, this step collects all these different output membership values (Toprak et al., 2009). Sen (2010) considered the implication and aggregation process as a single step of creating a fuzzy model, which results in four steps of fuzzy inference calculation.

2.8.5 Defuzzification

The aggregated result appears as a fuzzy value, and therefore it is necessary to return (defuzzify) this fuzzy value to a normal crisp value. Defuzzification can be done using different techniques such as centroid, bisector, last of maxima, middle of maxima, and smallest of maxima (Cobaner, 2011). However some fuzzy inference systems such as the Sugeno system have no fuzzy membership functions for the output variable like the Mamdani inference system. An equation such as least square is used to connect input to the output variables. In this case, the defuzzification step is not required because the output value is already crisp number.

2.9 TYPES OF FUZZY INFERENCE SYSTEMS

2.9.1 Mamdani Fuzzy Inference Systems

Mamdani fuzzy inference system involves all steps in fuzzy modeling in which both inputs and outputs for the system are fuzzy numbers (Mamdani, 1974; Mamdani and Assilian, 1975). Guney and Sarikaya (2009) presented a Mamdani structure as shown in Figure 2.12 using all five fuzzy calculation layers: fuzzy layer, product layer, implication layer, aggregation layer, and defuzzify layer. In Figure 2.12 circular nodes in a layer are fixed which are not changing during the calculation while rectangular nodes are adaptive. An example for two inputs variables *x* and *y* and the output variable *z* illustrated in the figure, the input variables are subdivided into three membership functions (M_1x, M_2x, M_3x) and (M_1y, M_2y, M_3y) and the output into nine membership functions (M_1z, M_2z… M_9z) for *x, y* and *z*. Nine membership functions are assumed for the output variable, a membership function for each rule. However, we can use any number of output membership functions (more than one) according to the type of the problem or expert knowledge. Therefore the number of membership functions is arbitrary and can be changed to a suitable number "more than one" for both inputs and output variables. However the only constrain here is a high number of rules consequences a relatively high number of membership functions which may create noisy calculation. The possible number of rules in Mamdani system depends on the number of input variables and their number of MFs. In this example, for two input variables *x* and *y,* each with three membership functions, the number of rules equals 3^2 results in 9 rules as shown in 2-12. However, the number of rules can be reduced if some combination rule scenarios are not possible.

Mamdani fuzzy inference system consists of the following layers of calculation.

In layer one (fuzzy layer), the crisp input values are converted to fuzzy values using the MFs. A suitable number and shape of the MF depends on the type of the available data, which ultimately dictate the fuzzy values (Eldin et al., 2004).

$$\left.\begin{array}{l}
\text{Rule (1) IF } (x \text{ is } M_1x) \text{ and } (y \text{ is } M_1y) \text{ THEN } (z \text{ is } M_1z) \\
\text{Rule (2): IF } (x \text{ is } M_1x) \text{ and } (y \text{ is } M_2y) \text{ THEN } (z \text{ is } M_2z) \\
\text{Rule (3): IF } (x \text{ is } \mathrm{M}_1x) \text{ and } (y \text{ is } M_3y) \text{ THEN } (z \text{ is } M_3z) \\
\text{Rule (4): IF } (x \text{ is } M_2x) \text{ and } (y \text{ is } M_1y) \text{ THEN } (z \text{ is } M_4z) \\
\text{Rule (5): IF } (x \text{ is } M_2x) \text{ and } (y \text{ is } M_2y) \text{ THEN } (z \text{ is } M_5z) \\
\text{Rule (6): IF } (x \text{ is } M_2x) \text{ and } (y \text{ is } M_3y) \text{ THEN } (z \text{ is } M_6z) \\
\text{Rule (7): IF } (x \text{ is } M_3x) \text{ and } (y \text{ is } M_1y) \text{ THEN } (z \text{ is } M_7z) \\
\text{Rule (8): IF } (x \text{ is } M_3x) \text{ and } (y \text{ is } M_2y) \text{ THEN } (z \text{ is } M_8z) \\
\text{Rule (9): IF } (x \text{ is } M_3x) \text{ and } (y \text{ is } M_3y) \text{ THEN } (z \text{ is } M_9z)
\end{array}\right\} \qquad (2\text{-}12)$$

Layer two (product layer) calculates the weighting factor W_i or firing strength of each rule shown in equation 2-13. In the “IF” part of fuzzy rules the product of different fuzzy membership values from the previous layer have to be calculated using different operations. This process continues with a single fuzzy number (weight factor) from each rule to fire the output membership functions in the next layer. As an example, in fuzzy rule 1 for variable x we have M_1x and for y variable we have M_1y. The combined product of these two fuzzy numbers results in one final fuzzy number w_1 to fire the output membership function which is M_1z in the next layer. Most common operations used are “and” which apply “minimum”, and “or” operation which apply “maximum” during the calculation.

$$\left.\begin{array}{l} W_1 = M_1 x \Theta M_1 y \\ W_2 = M_2 x \Theta M_2 y \\ . \\ . \\ . \\ W_9 = M_9 x \Theta M_9 y \end{array}\right\} \qquad (2\text{-}13)$$

Where, Θ is the desired fuzzy operation.

Layer three (implication layer) is represented by $W_{implicated,}$ as shown in the set of equations 2-14, is calculated using the fuzzy weight factor W_i from the previous layer and the output membership function. Note that during the mapping processes more than one rule product may be obtained for each input data point, which results in more than one $W_{implicated}$. Consequently, for each input variable a number of fuzzy values can be obtained as the end product of operation of the weight factor with the output membership function for each rule.

$$\left.\begin{array}{l} W_{implicated\,1} = W_1 M_1 z \\ W_{implicated\,2} = W_2 M_2 z \\ . \\ . \\ . \\ W_{implicated\,9} = W_9 M_9 z \end{array}\right\} \qquad (2\text{-}14)$$

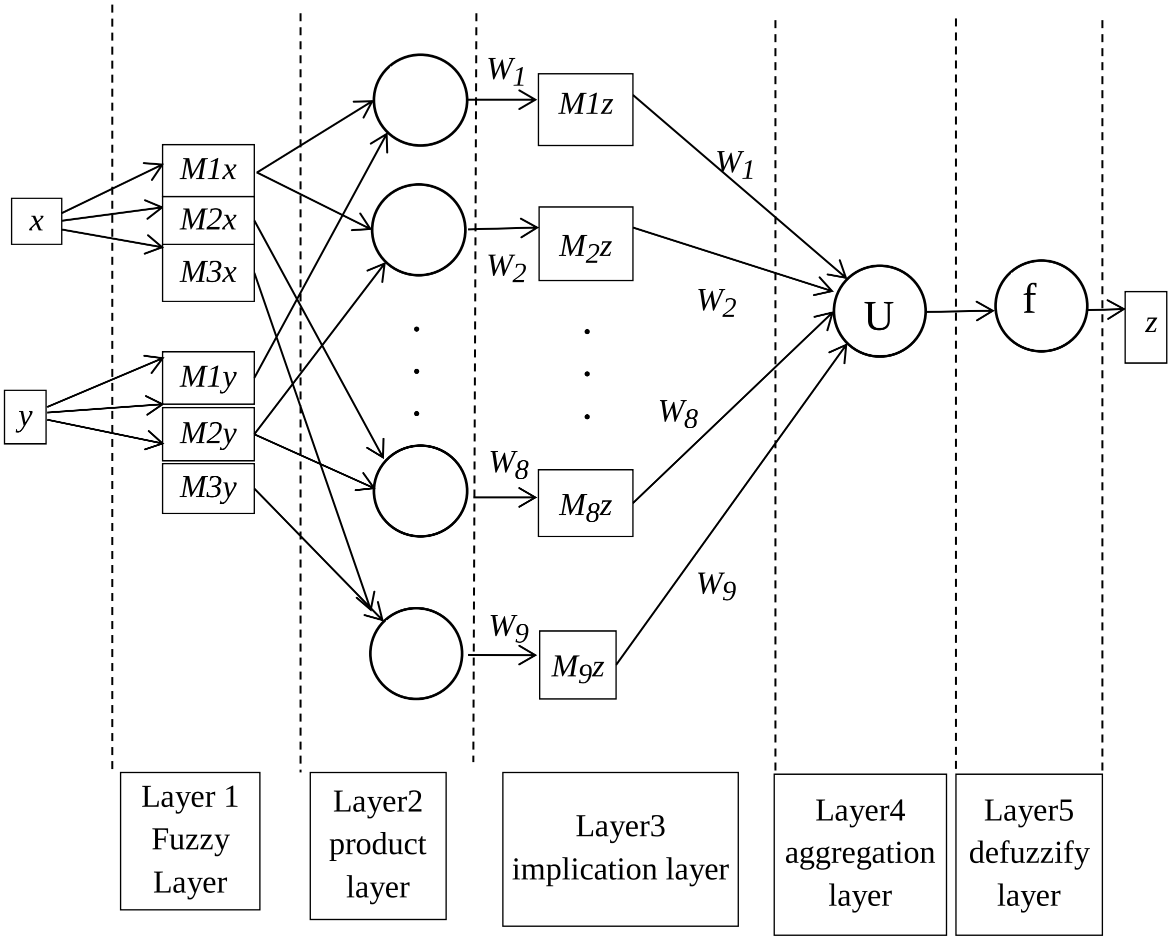

Figure 2.12. Architecture of Mamdani method

Layer four (aggregation layer) calculates the overall output $W_{implicated}$ functions through an aggregation process using different operation algebraic such as "sum and maximum" between different $W_{implicated}$ functions as in equation 2-15.

$$W_{output,i} = \bigcup_{k=1}^{9} W_{implicated,k} \tag{2-15}$$

Where $\bigcup$ is desired aggregation operation, k is the number of $W_{implicated}$ output functions.

Layer five (defuzzify layer) transfers the fuzzy aggregated output value from the layer 4 to a crisp value which is known as defuzzification using different functions such as area centroid, bisector or medium. Among them the most commonly used is the centroid, which is the center of the area of different output membership functions for an input (Guney and Sarikaya, 2009).

2.9.2 Sugeno Fuzzy Inference Systems

The Takagi and Sugeno (1985) method is similar to Mamdani and includes all procedures described above but the last step. This method has no explicit defuzzification calculation. In the Takgi-Sugeno approach which is mostly known as Sugeno, the consequent part is no longer fuzzy rather than a simple linear mathematical function. The structure of Sugeno system is the same as ANFIS structure as discussed in the next section.

2.9.3 Adaptive Neuro-Fuzzy Inference System

ANFIS is a hybrid Sugeno-type fuzzy system with neural network learning capabilities presented by (Jang, 1993). ANFIS membership function parameters are tuned (adjusted) automatically using either a back propagation algorithm or combination of back propagation with least squares method. This technique allows fuzzy systems to learn from the data (Sen, 2010; Yazdi and Pourreza, 2010).

ANFIS is a multi-layer network with internal hidden layers. The nodes in each layer connected to the other nodes through directional links. Each node has a function with adjustable or fixed parameters. In the training process the final value of the parameters is determined when the difference between network output and desired outputs reaches the minimum. Sugeno inference IF-THEN rules which have consequent part as crisp functions make the technique computationally efficient (Jang, 1993 andCobaner, 2011).

Figure 2.13 demonstrates an ANFIS structure network for two input variables *x* and *y* and output *z* based on first degree Sugeno with a set of four IF-THEN rules, are shown in 2-16.

$$\left.\begin{array}{l}\text{Rule (1): IF } x \text{ is } M_1x \text{ and } y \text{ is } M_1y \text{ THEN } f_1{=}z = p_1x+q_1y+r_1\\ \text{Rule (2): IF } x \text{ is } M_2x \text{ and } y \text{ is } M_2y \text{ THEN } f_2{=}z = p_2x+q_2y+r_2\\ \text{Rule (3): IF } x \text{ is } M_3x \text{ and } y \text{ is } M_3y \text{ THEN } f_3{=}z = p_3x+q_3y+r_3\\ \text{Rule (4): IF } x \text{ is } M_4x \text{ and } y \text{ is } M_4y \text{ THEN } f_4{=}z = p_4x+q_4y+r_4\end{array}\right\} \qquad (2\text{-}16)$$

Where M_i *x* and M_i *y* are *x* and *y* membership functions, respectively, *pi*, *qi*, and *ri* are first degree least square equation for *f* relating input variables *x*, and *y* to output variable *z*. The *f* function degree may take other degrees such as zero or two depending on the type of problem (Jang, 1993 and MathWorks, 2010).

The ANFIS structure as in Figure 2.13 has five layers of calculations. Each layer has nodes of the same function types; circle nodes indicate fixed nodes, while rectangles indicate adaptive nodes which are change during the calculations.

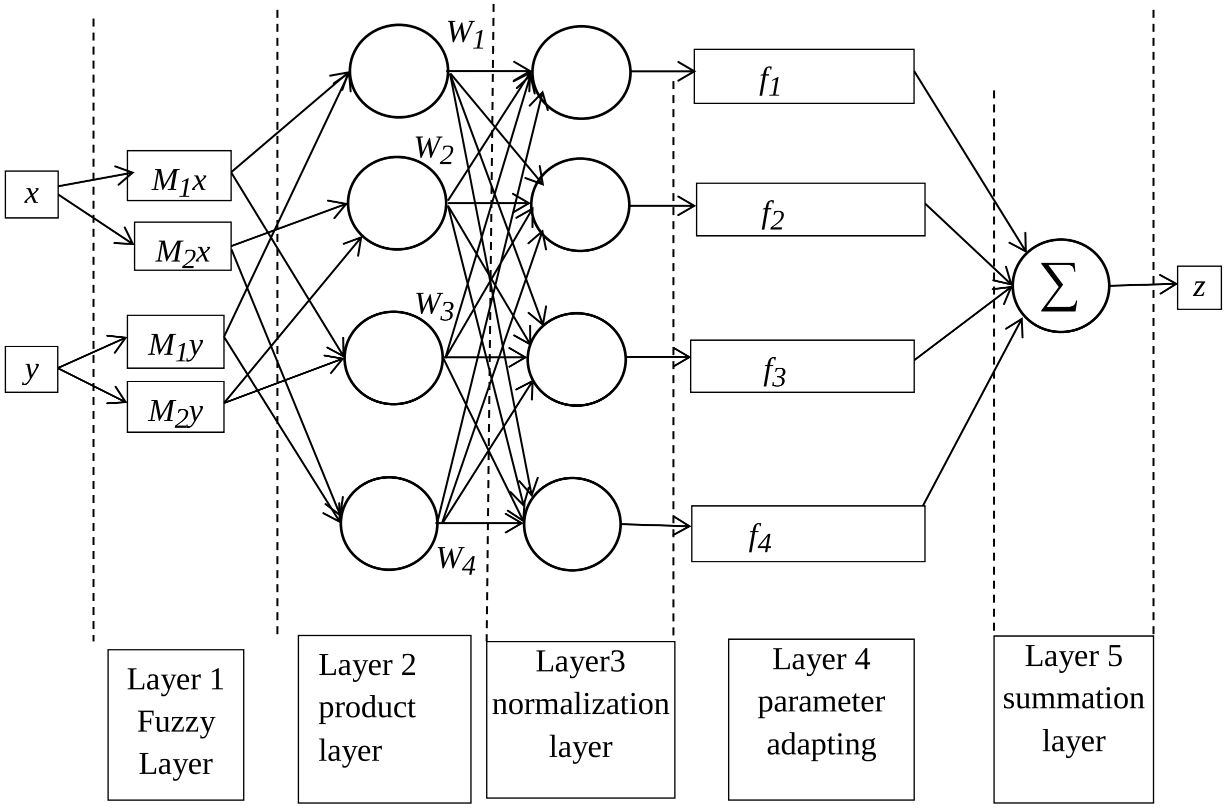

Figure 2.13. ANFIS structure network

Layer 1 (fuzzy layer): this layer consists of a number of nodes, for variable *x* and *y with* the result is expressed as in equation 2-17 and 2-18 respectively.

$$first\ layer\ output\ for(x) = M_i(x) \quad (2\text{-}17)$$

$$first\ layer\ output\ for(y) = M_i(y) \quad (2\text{-}18)$$

Where *x and y* is input to the node *i* with linguistic name M_ix and M_iy which represent the membership function using either grid partition or a cluster of data points in sub-cluster function types.

Layer 2 (product layer): in this layer, incoming signals are summed and sent to the next layer using an AND function (AND function multiples the signals) as in equation 2-19:

$$W_i = M_i(x) \times M_i(y) \tag{2-19}$$

Where W_i is the firing strength function or the product sum of the signals, i=1, 2, 3, 4.

Layer 3 (normalization layer): for each node the normalized firing strength W_i is calculated using all the results *wi* from different rules as in equation 2-20.

$$W_i = \frac{W_i}{W_1 + W_2 + W_3 + W_4} \tag{2-20}$$

Layer 4 (parameter adapting layer): this layer multiplies the normalized factor W_i with the parameters of the least square function f_i to obtain adapted parameters of the function so as to reduce the error of mapping, as shown in equation 2-21.

$$W_i f_i = W_i \; p_i x + q_i y + r_i \tag{2-21}$$

Where, W_i is the output for the layer 3 and f_i is a liner function between input variables *x* and *y* with the equation parameters p_i, q_i and r_i for Sugeno system.

Layer 5 (summation layer): in this layer the overall coming signals are summed to form the final output $f_{all,i}$ as in equation 2-22 which is the adapted final form of function used to calculate the output *z*.

$$f_{all,i} = \quad w_i f_i \tag{2-22}$$

ANFIS uses ANN to train fuzzy parameters and find the best results depending on minimizing the amount of error. During the calculation, the ANFIS software tunes two different types of parameters, linear and non-linear. The non-linear parameters are called premise parameters and define the fuzzy membership functions, depending on the type of the function. For example a triangular membership function can be defined with three parameters, while trapezoidal membership function needs four parameters. Thus the type of the function affects the number of modifiable parameters. The linear parameters are called consequent parameters and represent least square equation parameters (p, q,*and* r) which relate the x and y inputs to z output (MathWorks, 2010).

The learning process is started with selecting the number and type of function which gives value to the nonlinear parameters. This will plug training data to the final overall equation $f_{all,i}$ which can be used to find linear parameters. Each learning process consists of a number of epochs, and each epoch consists of two phases of calculation forward and backward. Thus the consequent parameters are identified in the forward pass of the hybrid learning process. After calculating the estimated values of output using the least square equation, the estimated values are compared to the real data to calculate the error of the model. The error is compared to the predefined desired value of error and it propagates in a backward pass to adapt the premised parameters using a gradient descent method. This process continues until the desired error is reached. In other words, the process continues as a trial and error, each time different numbers of membership functions are created and the one with the minimum possible error will be finally selected (Jang, 1993; Kisi et al., 2009).

MATLAB fuzzy logic toolbox used two different methods to divide the input variable space which are grid partition and clustering (MathWorks, 2010).

2.9.3.1 Grid Partitioning Method

The grid partition method considers membership functions according to data subset groups. The input divides into different spaces using axis-paralleled method in which each grid represents a fuzzy membership function. During the training, least square parameters are calculated depending on predefined numbers and the type of membership functions. Finally, the equation is used to find fuzzy prior parameters by minimizing the difference between the real data and predicted one using artificial neural network techniques.

2.9.3.2 Subtractive Clustering Method

Data points are clustered depending on predefined radius of clusters. In subtractive clustering method each data point is an assumed cluster center. The potential of data point x_i as a cluster center is calculated according to equation 2-23.

$$p_i = \quad {}^{N}_{j=1} \; e^{-\psi\|x_i - x_j\|^2} \qquad (2\text{-}23)$$

Where, $\psi = \frac{4}{r_a^2}$, r_a is a positive constant predefined influential radius to define the neighborhood. Data points out of this range have little influence on the potential while those points with large number of neighbor have more potential score. After calculating the potential of all points, the first cluster center x_{j1} is chosen as point with highest potential value of p_i^*.

Next, the potential p_i of each data point x_i is revised with respect to the first cluster center potential using equation 2-24:

$$p_i = p_i - p_i^* \; e^{-\beta\|x_i - x_{j1}\|^2} \qquad (2\text{-}24)$$

Where, $\beta = \frac{4}{r_b^2}$, r_b is a positive constant, which defines a neighborhood that has measurable reductions in potential. Therefore, the data points close to the first cluster center will have significantly reduced potential. To avoid obtaining cluster centers close to each other, r_b must be greater than r_a. Cobaner (2011) used r_b equal to $1.5\ r_a$, while Sanihkani and Kisi (2012) suggested r_b equal to *1.25* r_a .After revising the potential of all points, the point with the highest remaining potential are selected as the second cluster center. In the same manner the potential of all points revised with the second cluster center, the process will continue to p_k^* until a threshold value reached compared to the first cluster center calculated using equation 2-25.

$$\frac{p_k^*}{p_i^*} < \xi \qquad (2\text{-}25)$$

Where ξ is the predefined factor which controls the number of cluster centers by the effect of both r_a and r_b values (Nayak et al., 2005). Overall, the influential radius (r_a) controls the number of clusters. A small r_a results in higher number of clusters, consequently higher number of rules in the fuzzy system. Sanihkani and Kisi (2012) suggested selecting the value of r_a between 0 and 1, while Samhouri et al. (2009) suggested good value of r_a between 0.2 and 0.5.

2.9.3.3 Limitations of ANFIS Systems

During the training processes to prevent over fitting of ANFIS models, the number of adaptive parameters (linear and non-linear) must not exceed the number of data points. Also high number of input variables and membership functions are results in large number of rules which required high number of data points during the training process. Cobaner (2011); Sanikhani and Kisi (2012) limited the maximum number of input variables to 6 to prevent large numbers of rules and large adapting parameters in the ANFIS models which results in a noisy calculations and also required large number of data points which is not possible most of the time.

2.10 FUZZY COMPUTER SOFTWARE TOOLS

Many fuzzy logic software tools have been created; the most popular software's are the followings:

2.10.1 MATLAB

MATLAB fuzzy logic toolbox is the most popular specialized collection of files for fuzzy logic modeling, which supports design and analysis of Mamdani, Sugeno, ANFIS, and other fuzzy based systems. The toolbox supports all phases of the fuzzy process, including development, design, simulation, research, and real-time implementation (MathWorks, 2010).

2.10.2 F-IND

This is a software framework implemented in Java which is used for the development of multivariable indices with a fuzzy approach. The framework is suitable for ecological systems and it can be used for evaluation of vulnerability, quality, sustainability, magnitude, impact or any further ecosystem property (Marchini et al., 2009).

2.10.3 Fuzzy CLIPS

Fuzzy CLIPS is an extension of CLIPS (C-Language Integrated Production System) expert system developed by the National Aeronautics and Space Administration (NASA). It can accept exact, fuzzy and combined resourcing allowing fuzzy and normal terms to be combined in the rules of any expert system (Ibrahim, 2004).

2.10.4 Linguistic Fuzzy Logic Controller

Linguistic Fuzzy Logic Controller is provided by the Institute of Research and Application of Fuzzy Modeling, University of Ostrava, Czech Republic. This software has two different basic inference approaches. The first method is based on the interpretation of the IF-THEN rules that described logical implications linguistically and the second is the interpolation of the unknown functions (Ibrahim, 2004).

2.11 SUMMARY

Table 2.1 shows the type of fuzzy model used and the area of application for successful fuzzy logic models in water resources domain.

Table 2.1. Summary of application of fuzzy logic in water resources

Area of application	Author	Type of fuzzy model
Nitrogen leaching	Haberlandt et al. (2002)	Mamdani
Water quality index	Mujumdar and Sasikumar (2002)	Mamdani
Sediment in rivers	Tayfur et al. (2003)	Mamdani
Rainfall- runoff	Mahabir et al. (2003)	Mamdani
Waste water treatment	Carrasco et al. (2004)	Mamdani
Flow rate	Toprak et al. (2009)	Mamdani
Drought	Al-Mohseen (2009)	Mamdani
Flow in constructed canals	Toprak (2009)	Mamdani
Rainfall forecasting	Ghalhary et al. (2009)	Mamdani
Water quality in rivers	Bai et al.,(2009)	Mamdani
Water level predicting	Alvisi, et al. (2006)	Mamdani, Sugeno
Irrigation water quality	Mirabbasi, et al. (2008)	Mamdani, Sugeno
Dissolved oxygen	Altunkaynak et al. (2005b)	Sugeno
City water demand	Altunkaynak et al. (2005a)	Sugeno
Flood forecasting	Nayak (2010)	Sugeno
River flow rate	Al-Zu'bi et al. (2010)	Sugeno
Extrapolate rainfall data	Tzimopoulos et al. (2008)	Sugeno, ANFIS
River ice breakup	Mahabir et al. (2006)	ANFIS
Evaporation predicting	Kisi (2006)	ANFIS
Bacteria growth model	El-Sebakhy et al. (2007)	ANFIS
Sediment in rivers	Kisi et al. (2009)	ANFIS
Wetlands health model	Mah (2011)	ANFIS
Flow rate	Sanikhani and Kisi (2012)	ANFIS
Nitrate in groundwater	Mousavi et al.,(2011)	Fuzzy regression
Monthly flow rate	Sadatinejad et al. (2009)	Fuzzy regression
Index flow	Hamaamin et al. (2013)	Mamdani, ANFIS and fuzzy regression

2.12 CHALLENGES AND RECOMMENDATIONS FOR FUTURE WORKS

The following are challenges and recommendations from different fuzzy logic modelers to be considered for future water resource works.

1- Alternative methods such as boosting or genetic algorithms for emerging issues in water resources such as climate changes, and extreme events (Hamaamin et al., 2013).

2- Hybrid (fusion) methods between physically based models and soft models may result in a good combined method for water resources managements, in which we realize the advantages of two totally different techniques (Solomatine et al., 2008).

3- Correct estimation of monthly suspended sediment is difficult, which requires improvement. Building a rule based sediment model depending on data from different watersheds is an attempt to build a model to cover broader area and to strengthen the use of the model (Kisi et al., 2009).

4- Artificial Intelligence models are not performing well for the cases where the space-time distributions are one of the variables. It will be a good effort if spatial and time variables added to the model implicitly to be applicable in different time to different watersheds (Tayfur et al., 2003).

5- To reduce calculation costs, some variables may be included in other variables implicitly to. For example the stream and time coefficients (stream type and season) can be used as variables to contain the effect of many variables implicitly. Snowmelt and evapotranspiration can be included in time coefficient, while basin area, infiltration, slope, and other physical conditions for the stream can be included in a variable such as stream coefficient. (Toprak et al., 2009).

6- Fuzzy logic models that are depending on outputs from physically based systems such as SWIM and SWAT. Using more than one modeling system outputs from a different area and using different software may reduce model bias, and the resulted fuzzy model will be applicable for a wider area (Haberlandt et al., 2002).

7- In fuzzy logic methodology, unlike statistical methods, there are no restricting assumptions such as normality, linearity, and independence of residuals. It may perform better if some conditions also assumed for example normality of the data sets prior the modeling process (Altunkaynak et al., 2005).

8- Most of fuzzy logic modelers, suggest using up to six input variables for model processing. Higher number of variables can be tested with a lower number of rules and membership functions in an attempt to reduce the noise in the problem due to high-unknown parameters (Sanikhani and Kisi, 2012).

9- Fuzzy logic models perform better when the physical phenomenon considered and synthesized is limited number in variables and IF-THEN rules. Fuzzy models accuracy increase to a certain number of rules, beyond that number the accuracy decrease with increasing the number of rules in the model. Finding a relationship of how accuracy depends on the number of rules and variables may be a good future research topic (Alvisi, et al., 2006).

10- The nature of fuzzy logic models provides the ideal platform to explore the likely effects of climate change on hydrologic events. For an adaptive model under global climate change impacts, it is hard to meet the new climate changes using conventional parameters based on the local-scale data for models predictions. For a basin model, data collection range may extend to a wider area outside the watershed to include the effect of neighbor watersheds or global data regarding the climate change (Mahabir et al., 2006, Chen and Chang, 2010).

3 INTRODUCTION TO METHODOLOGY AND RESULTS

This dissertation consists of two research papers covering water resource modeling in both water quantity and quality areas. The first paper covers water quantity and models the index flows for ungaged streams using observed flow data from other gaged streams. In this study, more robust and cost-effective techniques are used as a substitution for conventional techniques. The second paper covers water quality and models sediment loads in a watershed using a fusion and Bayesian modes. This method makes repetitive sediment estimation more effective, especially during the evaluations of different management scenarios by stakeholders and watershed managers.

The title of the first paper is “Application of Fuzzy Logic Techniques in Estimating the Regional Index Flow for Michigan” The objective of the study was to test and find more robust technique(s) for estimating the index flows for ungaged streams for the State of Michigan. The index flow is critical for watershed managements in limiting withdrawals and protecting stream ecosystems. The performance of the four different modeling methods, multiple linear regression, fuzzy regression, fuzzy expert, and ANFIS, were tested to select the best prediction method. Results of this study showed that the fuzzy expert model was the best method to use to estimate index flows in Michigan. Also the fuzzy expert method uses less number of predictors (four) compared to the state adapted regression model, which uses six predictors.

The title of the second paper is “Evaluation of Neuro-Fuzzy and Bayesian Techniques in Estimating Sediment Loads”. The objective of this study was to find more effective modeling techniques for sediment estimation in a watershed; especially during the evaluation of different management scenarios. In this study, the applicability of a fusion method in modeling sediment was tested by comparing the ANFIS and Bayesian methods to a physically based model called soil and water assessment tool (SWAT). First, a calibrated and validated SWAT model was used to estimate sediment loads in the Saginaw Watershed. Afterward ANFIS and Bayesian Regression techniques were tested to reproduce sediment results with a fewer number of parameters and a smaller number of calculations. Results of the study showed that both techniques were able to estimate sediment loads at the watershed level successfully. Also the applicability of the model was tested at the subbasin level where each method produced acceptable results for about half the number of the subbasins. The ANFIS method was able to produce satisfactory results for upstream subbasins while the Bayesian method was able to produce satisfactory results for downstream subbasins.

4 APPLICATION OF FUZZY LOGIC TECHNIQUES IN ESTIMATING THE REGIONAL INDEX FLOW FOR MICHIGAN

4.1 ABSTRACT

Knowledge of flow dynamics within streams is not only helpful in determining applications, such as water withdrawals, but it also is a driving force in a watershed for other components such as nutrients, dissolved oxygen, and ecological health. With increases in pressure on water resources, along with growing attention to climate changes and anthropogenic activity, the ability to accurately predict extreme conditions continues to be invaluable to decision makers and watershed managers. The objective of this research was to explore new concepts and develop robust techniques for estimating the index water yield and subsequently the index flow for ungaged streams for the state of Michigan. In this study, four different modeling methods (linear regression, fuzzy regression, fuzzy expert, and adaptive neuro-fuzzy (ANFIS) were evaluated using 10-fold cross validation technique. Results showed that the fuzzy logic expert method was the best and most robust predicting model. Further evaluation revealed that the ANFIS model preserves many attributes of the data set, at the expense of over-fitting and generating unexplained values such as outliers.

4.2 INTRODUCTION

Low flow is a seasonal phenomenon and an integral component of the flow regime for any river or stream (Lee and Kim, 2008). Therefore, knowledge and understanding about the low flow in streams is important for water resources planning and management. There are many applications where low flow data is essential, including catchment abstraction management, water supply for agricultural productions, in-stream flow determination, effluent dilution estimation, navigation, and hydropower plant operation (WMO, 2008). Low flow information is also used to determine waste-load allocations, manage basin withdrawals plan, and design cooling-plant facilities (Kroll et al., 2004). In addition, low flow data may also be used to determine characteristics of the groundwater or the health of the stream ecosystem and may reflect the geology of the region, including the degree of urbanization (Hejazi and Moglen, 2007). As a result of these numerous uses, many definitions for low flow exist. Some definitions consider criteria such as flow rate during the dry season; while in erratic and intermittent semi-arid flow regimes, it is the length of time between flood events (Smakhtin, 2001; Eslamian et al., 2010; Mandal and Cunnane, 2009). However, it is important to distinguish low flow from drought, which is a natural event resulting from unusually low precipitation for an extended period of time (Lee and Kim, 2008).

Many techniques have been developed for determining low flow, all of which require adequate observed streamflow records from gaged catchments. One of the most common techniques is the flow duration curve. In this method, low flow frequency curves define low flow for a certain stream as the lowest average reoccurring flow period (Smakhtin, 2001). In the USA, the two widely used techniques are $Q_{7,\ 10}$ and $Q_{7,\ 2}$. The $Q_{\ 7,\ 10}$ method is defined as the observed lowest 7-day average flow that is expected to be exceeded on average in 9 out of 10 years. $Q_{7,\ 2}$ is similarly denoted, however this is defined based on the estimated lowest 7-day average flow that is expected to be exceeded in 1 out of 2 years (Eng and Milly, 2007).

In cases where observed streamflow records are not available, regional models are developed based on catchment physiographic characteristics, spatial interpolations, and synthetic flow time series. Perhaps the most widely used technique in regional low flow estimation at ungaged sites is the regression model (Smakhtin, 2001). The regression model illustrates a relationship between dependent low flow variables and independent catchment and climatic variables. To generate a good regression model for predicting low flow in a stream, a somewhat large amount of observed flow data is required from surrounding gaged sites that are hydrologically similar (Castiglioni et al, 2009). The streamflow data used should represent natural flow conditions in the catchments and must be unaffected by regulation from reservoirs and local flow diversions; otherwise the results will be misleading (Martin et al., 2010). Consequently, regression models can be created for gaged sites and transferred to ungaged sites (Smakhtin, 2001).

4.2.1 Index Flow

With flow being the driving force in a watershed, it influences temperature, sediments, nutrients, and several other stream conditions. This is further reflected in impacts to the ecology of the systems, including ecological processes and aquatic organisms.

Flow can also have a more direct effect on the aquatic organisms, such as its influence on habitat availability, wash out, and other stressors. These impacts can be especially relevant during dry seasons and can directly impact the sustainability of ecosystem health (Bubb et al., 2002; Arthington et al., 2006; Koster et al., 2010). Meanwhile, there are concerns about the application of low flow as a design or index flow to protect aquatic ecosystems (Pyrce, 2004). In this regard, the Michigan Legislature has passed the first state law to regulate water withdrawals, while protecting streams' ability to support characteristic fish populations. According to Michigan's Public Act -33-(2006), the index flow is the median flow for the lowest flow month of the flow regime. For the sites with long-term streamflow data, the index flow (IQ_{50}) is defined as the minimum of median flows during the months of July, August or September. This is the period of greatest stress on flow regimes because of lower flow rates and high temperatures. For ungaged sites, Hamilton et al. (2008) developed a linear regression model for computing the index flows in Michigan using all unregulated sites having 10 or more years of record (147 stations). However, the developed method could not be validated since information of all flow stations within Michigan was used for model development.

Since index flow variability is mostly related to the upstream drainage area, the index water yield was chosen as a response variable. Hamilton et al. (2008) developed a multiple-linear regression for estimating index flow for Michigan by relating the index flow to climate and basing characteristics. Their model based on the assumption that errors or differences between measured and estimated index flow values are normally distributed, which is not possible without transforming the response variable by taking the square root of the index water yield variable . In Hamilton's model, the index water yield (IY_{50}) was calculated as index flow divided by the drainage area above the point of interest. However, since IY_{50} was not normally distributed, a square root transformation was applied. They identified six explanatory variables in the regression model including the percentage of the basin classified as having low and high ground-water transmissivities, the percentage of the basin classified in the A and D soil groups, the percentage of the basin where land cover is classified as forest, and average annual precipitation.

4.2.2 Fuzzy Logic and Hydrologic Applications

The fuzzy logic approach is an approximate reasoning method for dealing with uncertainties in modeling. Fuzzy reasoning is a way to interpret system behavior with interpolations based on limited available numerical data, by generating mechanism of the system and its processes based on inputs (causes, antecedents) and outputs (results, consequents). Along with data and information about the input and output variables, fuzzy reasoning also considers expert knowledge and insights about the system (Sen, 2010). The difference between fuzzy logic and probabilistic approaches is that fuzzy membership functions express the outcomes as a possibility instead of a likelihood of probability density functions (Mckone and Deshpande, 2005). In contrast to the other statistical approaches, FL methods performance is not highly dependent on the volume of available data, since their outcomes depend on linguistic rules. In addition, for the small data sets, outliers in the data do not have a large influence on the fuzzy logic model (Mahabir, 2003).

Fuzzy subsets (membership functions), "IF...THEN..." rules and reasoning concepts are the basis of the soft computing known as Fuzzy inference System. Meanwhile, there are also a variety FIS approaches, including the Mamdani (Mamdani, 1974), Takagi-Sugeno (Takagi and Sugeno, 1985), and adaptive neuro-fuzzy inference systems (ANFIS) approaches (Sen, 2010). Mamdani FIS involves both fuzzy inputs and outputs. The general fuzzy process used in Mamdani has four steps (fuzzification, inference, composition, and defuzzification). Fuzzification is the first step and includes building membership functions, which are curves that define how each input point is mapped to a membership value (or degree of membership) between 0 and 1. The inference step assigns one fuzzy subset to each output variable for each rule. This is where the “IF...THEN..." rules are determined. This is followed by composition, where fuzzy subsets assigned to each output variable are combined to form a single fuzzy subset for each output variable. Finally, defuzzification converts the fuzzy output set to a crisp number (Sen, 2010). The Takgi-Sugeno FIS method is similar to Mamdani and involves all the above mentioned steps except the last step. With the Sugeno approach, the consequent part is no longer fuzzy and rather built based on simple linear mathematical functions (Sen, 2010). Unlike the other methods, ANFIS is a hybrid Sugeno-type fuzzy system with neural–network learning capabilities. ANFIS membership function parameters are tuned (adjusted) automatically using either a back propagation algorithm or combination of back propagation with least squares method. This technique allows fuzzy systems to learn from the data they are modeling (MathWorks, 2010). In recent years, many applications of FL modeling have entered into hydrological and water resources domain as an alternative and powerful method for modeling. For example, fuzzy computing technique has been extensively used for flow rate and flood modeling (Nayak et al., 2005; Deka and Chandramouli, 2005;   Toprak, 2009; Al-zu'bi.et al., 2010 and Nayak, 2010). Altunkaynak et al. (2005) used Sugeno fuzzy modeling to predict urban water consumption amounts and observed less errors compared to auto regressive models. Mahabir et al. (2003) found considerably superior water supply forecasting results for a FL system compared to the regression equations. Fuzzy models were also observed to perform better for flow estimation at poorly gaged sites by Toprak et al. (2009). In addition, hybrid neuro-fuzzy models have been used for different applications of study including flood estimation, missed rainfall data, water level predictions, and predicting evaporation (Chau et al., 2005; Mahabir et al., 2006; Kisi, 2006; Shu and Quarda, 2008; Tzimopoulos et al., 2008; 2004, Keskin et al., 2009). Further studies have shown similar results with fuzzy models being demonstrated as a suitable alternative for modeling the non-linear relationship between rainfall and runoff (Hundecha et al., 2001; Jacquin and Shamseldin, 2006; Ghalhary et al., 2009).

4.2.3 Research Objectives

With increases in pressure on our water resources, along with growing attention to climate changes and anthropogenic activity, the ability to accurately predict low-flow conditions continues to be invaluable (Smakhtin, 2001). Meanwhile, conventional methods may not always be cost-effective or robust and there is growing interest in alternative soft computing methods such as fuzzy logic and artificial neural networks (Huang et al., 2010). The overall objective of this study was to develop a more robust technique(s) for estimating the index water yield and subsequently the index flow for ungaged streams in the state of Michigan. The study aims at building, evaluating, and comparing performance of four different modeling methods, multiple linear regression, fuzzy regression, fuzzy expert, and adaptive neuro-fuzzy (ANFIS) systems, to select the best predicting method.

4.3 MATERIALS AND METHODS

4.3.1 Description of Study Area

This study focuses on the state of Michigan (Figure 4.1). The state is made up of two peninsulas that have a total land area of 150,504 km^2 (Michigan Library, 2006). The Upper Peninsula is heavily forested with a large portion of its lands made up of bedrock just below the surface. The Lower Peninsula is covered by glacial drift, with more sandy soils and forested areas to the north. Land use and soils in the lower portion of the Lower Peninsula are mixed, with high percentage of land dedicated to agricultural production and urban use. Michigan is bordered by four of the five Great Lakes (Lake Superior, Lake Michigan, Lake Erie, and Lake Huron) as well as Lake Saint Clair. Not only is it surrounded by an abundance of water, but it consists of roughly 3,380 km^2 of inland waters as well (Michigan Library, 2006). Among the inland waters, there are 58,500 km of rivers and streams, making up 63 major watersheds. The northern portions of the state are home to several trout streams that are sustained by their base flow and are characterized by their colder temperatures (NORS, 1999).

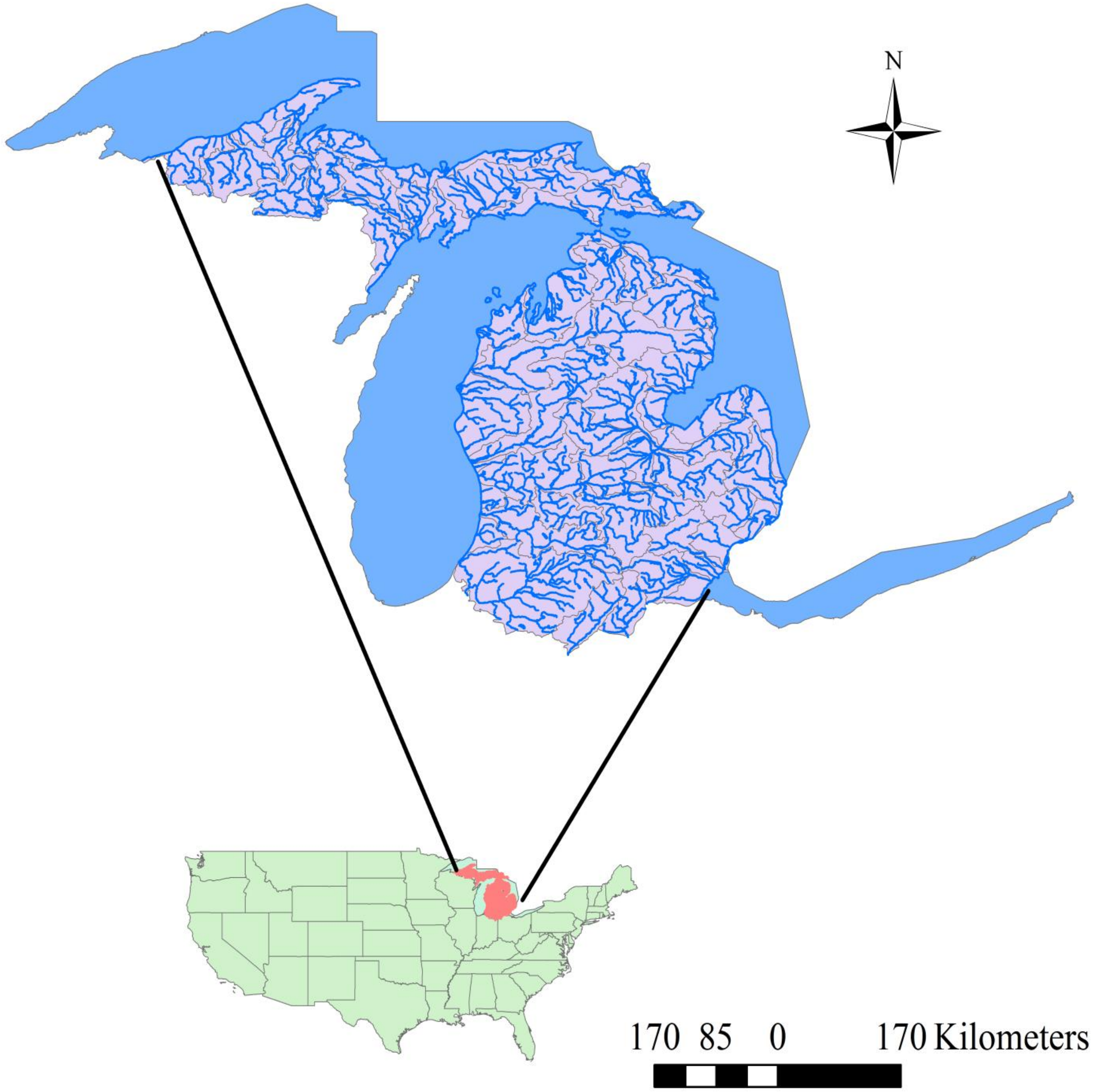

Figure 4.1. State of Michigan. For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

4.3.2 Data Used

Stream flow data from Michigan USGS gaging stations were used in this study to obtain index flows. Applicable stations were those with at least 10 years of continuous record data. A water year is the 12-month period from October 1 to September 30 and is identified by the calendar year in which it ends. Station selection for the study is also depended on characteristics of daily flow estimates. Stations were unacceptable if flow estimates were thought to be meaningfully affected by regulation through water withdrawals, diversion, and natural storage in lakes or retention in regulated surface-water bodies. In other words, stations were rejected if they were suspected to significantly mask the hydrologic response from precipitation. Under the above constraints, 147 stream flow gaging stations were selected for inclusion in the analyses. Stations data records ranged from 11 to 91 years in length, with an average of 40.2 years. The first water year of record used in the analysis was 1901 and the last one was 2005 (Hamilton et al., 2008).

Basin and climatic characteristics considered physically and statistically related to the index water yield were used as input variables to the models and also gathered for this study. For the variable selection process 12 hydrologically significance explanatory variables were considered. These variables included H-Trans, M-Trans, and L-Trans (indicating the percentage of the basin area with high, medium, and low aquifer transmissivity, respectively); forest (indicating the percentage of basin area covered with forest); A-Soils, B-Soils, C-Soils, and D-Soils (indicating the percent of basin areas classified as hydrologic soil group A, B, C, and D, respectively); RCN (indicating the runoff curve number for the basin area); Precip (indicating the basin average annual precipitation); and Snowfall (indicating the snowfall depths in the basin). A stepwise regression technique was performed on the explanatory variables starting with the most significant variables for the model. Six variables (H-Trans, L-Trans, Forest, A-Soils, D-Soils, and Precip) were selected based on the criteria including, but not to limited, the explanation of a significant amount of the variability in the output, low overall estimation error, and a 5%

significance level for individual parameters. A more detailed explanation of this process can be found in Hamilton et al. (2008). In the following section, each parameter will be discussed in more details.

The first variable, transmissivity, is the capacity of an aquifer to transmit water horizontally. The transmissivity of an aquifer is expressed in units of length per day multiplied by its saturated thickness. The Michigan Department of Information Technology (2005) classified the surface geologic deposits of the state into 10 land systems, each with a different class of aquifer transmissivity. A median estimated aquifer transmissivity of 67.2 m^2/day was classified as low, 187.7 m^2/day as medium, and 351.2 m^2/day as high. As explained earlier, two basin aquifer classifications, low and high transmissivity, were used as input variables in the models.

Land-use and land-cover characteristics affect the rate at which water infiltrates into the soil, resulting in water either draining to the groundwater system or flowing overland to a nearby stream. The Michigan Resource Information System (MIRIS, 1978) contains land-use and land-cover data for the state of Michigan in 30 m^2 area grids. The spatial distribution of forest lands was determined for each basin using MIRIS (MIRIS, 1978). The percentage of the basin area considered as forested was then used as the stations' forest input variable for the model.

The average annual precipitation was used for each basin with a station. According to Michigan Climatological Resources Program (2004), average annual precipitation ranges from about 72.4 cm/year in the northeastern part of the Lower Peninsula to about 96.5 cm/year in southeastern part of the Lower Peninsula. Precipitation ranges from about 81.3 cm/year in the eastern part of the Upper Peninsula to 88.9 cm/year in the far western part. Normal annual snowfall ranged from 101.6 cm in southeastern Lower Peninsula to 558.8 cm in the northwestern Upper Peninsula.

As for soils data, hydrologic soil groups A and D were used as defined by the U.S. Department of Agriculture Natural Resources Conservation Service (NRCS) (NRCS, 2007). One input variable, the percent of soil from group A, is characterized by a low runoff potential when thoroughly wet; water is transmitted freely through the soil. Group A-soils typically have less than 10 percent clay and more than 90 percent sand or gravel. Another input variable is the percent of soil from group D-soil, which is characterized by a high runoff potential when thoroughly wet. Water movement through the soil is restricted or very restricted. Group D-soils typically have greater than 40 percent clay and less than 50 percent sand.

4.3.3 Modeling Methods

As it was discussed earlier, the overall goal of this study was to develop a more robust technique(s) for estimating the index water yield and, subsequently, the index flow for ungaged streams for the state of Michigan. In this section, four different modeling methods, multiple linear regression, fuzzy regression, fuzzy expert, and adaptive neuro-fuzzy (ANFIS) systems, will be explained in more details.

4.3.3.1 Existing Regression Method

The first method used was multiple linear regressions. The regression model was executed through a stepwise process by Hamilton et al. (2008). They presented a regression model with six parameters to estimate the flow index 'square root of yield' for the Michigan area using all the 147 USGS stations' data points. The six parameters, low-transmissivity (*L-Trans*), high-transmissivity (*H-Trans*), *forest* cover (*Forest*), *precipitation (Precip)*, *A-soil* type (*A-Soils*), and *D-soil* type (*D-Soils*) were used as in equation 4-1:

$$\sqrt{}IY_{50} = -0.541982 + -0.00136258\ L\text{-}Trans + 0.00204796\cdot\ H\text{-}Trans + 0.00402190\ Forest + 0.0236424\ Precip + 0.00225536\ A\text{-}Sois + 0.00162107 D\text{-}Soils \qquad (4\text{-}1)$$

Where, $\sqrt{}IY_{50}$ is the predicted square root of water yield.

4.3.3.2 Fuzzy Regression Method

The second method, fuzzy regression, is somewhat similar to classical regression; however, it is applicable for determining regression relations from fuzzy initial data. In classical regression, one or more data points may be influenced by unknown factors and can greatly affect the outcome and accuracy of the model. However, fuzzy methods incorporated with regression addresses these concerns (Gu et al., 2006). If a variable y is related linearly to other variables $x_{1,}$ x_2 ... x_k, the regression model of these variables is shown in equation 4-2:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k \qquad (4\text{-}2)$$

Where, b_0, b_1, b_2 ... b_k are k +1 regression parameters that have yet to be determined. When the k +1 parameters are determined on the basis of observed values, the regression model will be resolved accordingly and y will be predicted by independent variables x_1, x_2 ... $x_{k.}$. Note that $\hat{y}$ expresses predicted values, while y expresses observed values (also called dependent variable). If a fuzzy domain is added, discrete data points will have a reduced effect on the fitness result and the concentrated data points will have an enhanced effect on the fitness (Gu et al., 2006). A normal distribution membership function is defined in equation 4-3.

$$ui = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\bar{y})^2}{2\sigma^2}} \qquad (4\text{-}3)$$

Where, y is the observed data points, $\bar{y}$ is the average of the observed data points and σ^2 is the variance (Gu et al., 2006). Marza and Seyyedi (2009) presented a group of equations to obtain fuzzy parameters. According to the input variables $x_{1,}\ x_{2,} ...,x_k$, a set of equations (equations 4-4) can be produced to find the parameters of the regression equation, which are b_1, $b_{2,} ...,b_k$.

$$\left\{\begin{matrix} s_{11}b_1 + s_{12}b_2 + ... + s_{1k}b_k = s_{1y} \\ s_{21}b_1 + s_{22}b_2 + ... + s_{2k}b_k = s_{2y} \\ . \\ . \\ . \\ s_{k1}b_1 + s_{k2}b_2 + ... + s_{kk}b_k = s_{ky} \end{matrix}\right\} \tag{4-4}$$

Where, both i and j= 1,2,…,k.

$$s_{ij} = \sum u \sum ux_i x_j - \sum ux_i \sum ux_j \tag{4-5}$$

$$s_{iy} = \sum u \sum ux_i y - \sum ux_i \sum uy \tag{4-6}$$

By solving these equations (in this study there are six equations), values of b_1, b_2... b_6 are obtained and, finally, the value of b_0 can be obtained by the equation 4-7.

$$b_0 = \frac{\sum uy}{\sum u} - b_1 \frac{\sum ux_1}{\sum u} - b_2 \frac{\sum ux_2}{\sum u} - ... - b_k \frac{\sum ux_k}{\sum u} \qquad (4\text{-}7)$$

4.3.3.3 Fuzzy Expert System

The fuzzy logic approach can be especially effective when working with highly variable and uncertain data, as well as complex and non-linear processes (Adriaenssens et al., 2004). In comparison with common linear regression methods, fuzzy expert system requires fewer independent variables and can often be easily interpreted based on their linguistic nature. In order to reduce model complexity, different combinations of variables were tested to find the most effective. Out of all six variables, the four variables that were selected for input into the fuzzy expert system, based on their autocorrelation coefficients, were: *H-Trans*, *Forest*, *Precip*, and *A-Soils*. As stated by Mahabir et al. (2003) the number of IF-THEN rules depends on the number of input variables and the number of membership functions for each variable to cover all possible combination of all input variables membership functions. Accordingly, in this model, we need 3^4 or 81 rules to cover all possible combinations of the four input variable numbers and three MFs for each of them resulted in best model predictions. Consequently with these four variables, 81 IF-THEN rules were created for the model. In a rule-based model, the relationship between input variables and the results is simple. For example, if *H-Trans* is medium, *Forest* cover is low, *Precip* is high, and *A-Soils* is low, then square root of yield is medium. Each variable was made up of three membership functions: two-side trapezoidal for high and low and one middle triangle for the medium. For example, Figures 4.2 and 4.3 show the membership functions for *H-Trans* and square root of water yield (output) respectively. MATLAB's Fuzzy Logic Toolbox (version 2.1) was used to perform the fuzzy modeling. The equally weighted variables were combined into rules using the concept of 'AND'. Implication was achieved with the minimum function and aggregation was performed with the maximum function. Finally, the centroid method was applied as a defuzzification of the output membership functions to determine a crisp set.

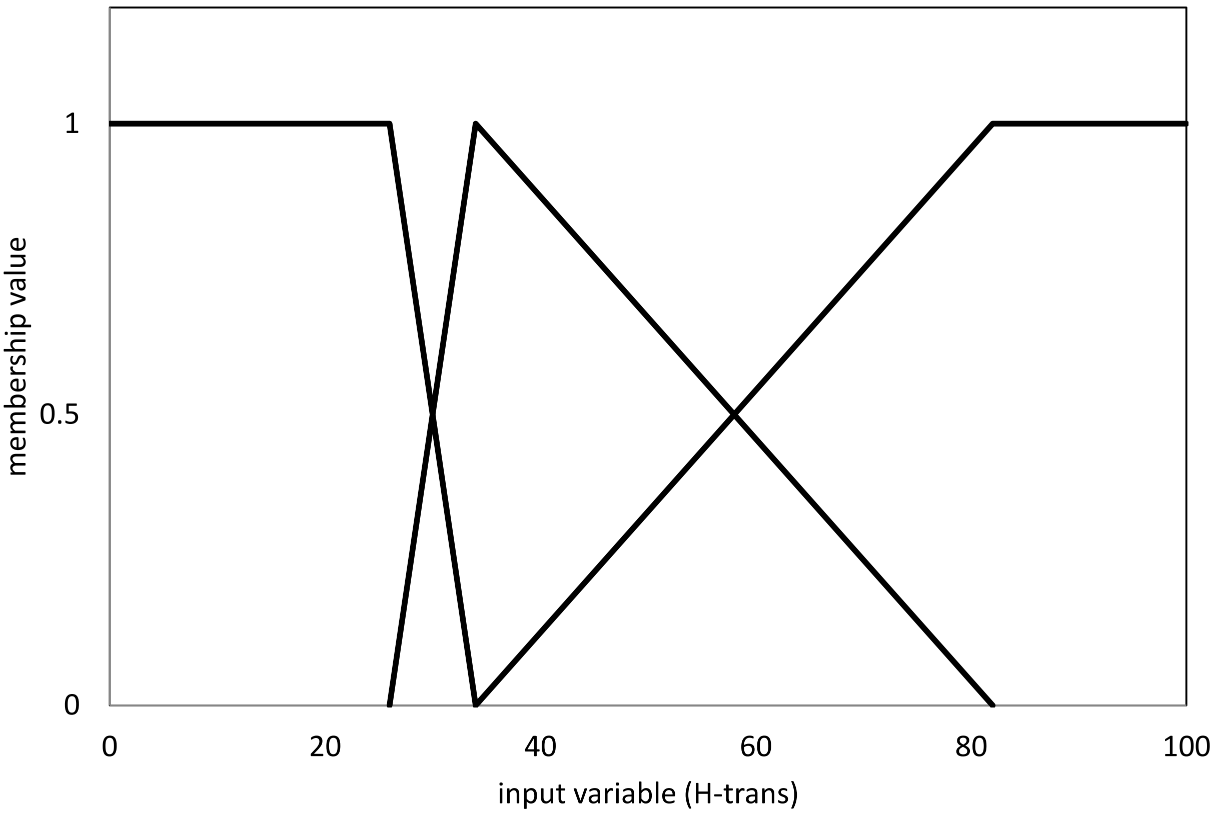

Figure 4.2. Input variable (H-transmissivity) membership function

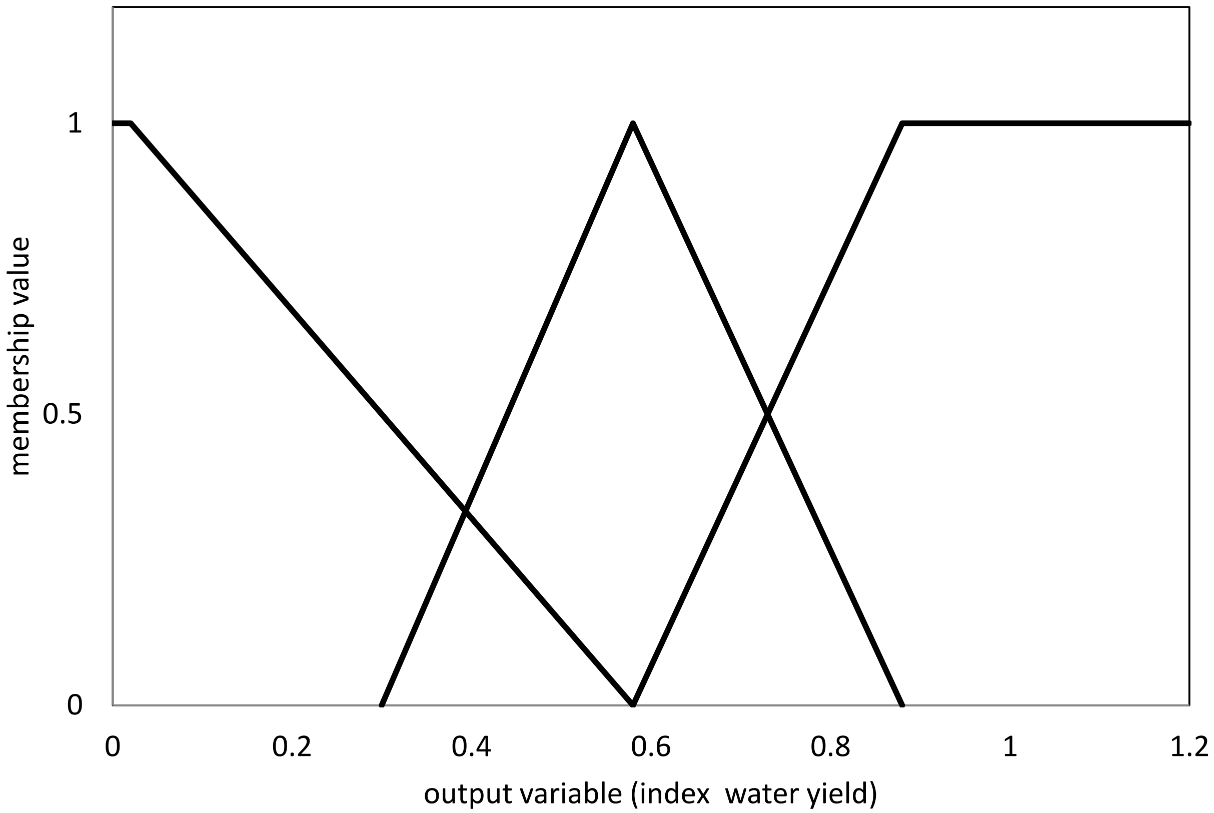

Figure 4.3. Output variable (square root of index water yield) membership function

4.3.3.4 ANFIS

The final method employed is a fusion technique known ANFIS, which is adaptive neuro-fuzzy inference system. In this method, artificial neural networks (ANN) can be employed in connection with fuzzy logic to identify and determine a models rule base. Through repetitive evaluations, ANNs can incrementally change the influence of the rules on model results. Although this is a useful method, there is a need to avoid over and under-training (Mahabir, 2006). In the fuzzy expert system, the rule numbers and structures are essentially predetermined by arbitrarily chosen fixed membership functions. ANFIS learning techniques provide a method for the fuzzy modeling procedure to learn information about a data set. The MATLAB's Fuzzy Logic Toolbox function ANFIS computes the membership function parameters that are the most beneficial for the associated fuzzy inference system to track the given input/output data. This allows the toolbox to create a membership function automatically. The previously selected four input variables (*H-Trans*, *Forest, Precip,* and *A-Soils*) were used once again for this model. Because the number of training data points are limited and need to exceed the number of modifiable parameters, two membership functions were used for each of the four independents variables. The two types of data points, training and testing, were loaded into MATLAB's *ANFIS* command in the Fuzzy Logic Toolbox. The models were built on 90% of the data points, the training portion of data, and checked on the remaining 10%, the testing part of data. The final model for each data set was chosen based on which model had the least amount of errors for both the training and testing sets. Finally, the number of iterations was limited to a point where error was minimized. The root mean square of errors calculated in MATLAB is illustrated in Figure 4.4. Iteration number 15 represents the lowest error for both functions (the diamonds represent the testing errors whereas the asterisks represent the training errors). In order to obtain best estimations for the output the model has been tuned by changing the shape and parameters of each input variables membership functions.

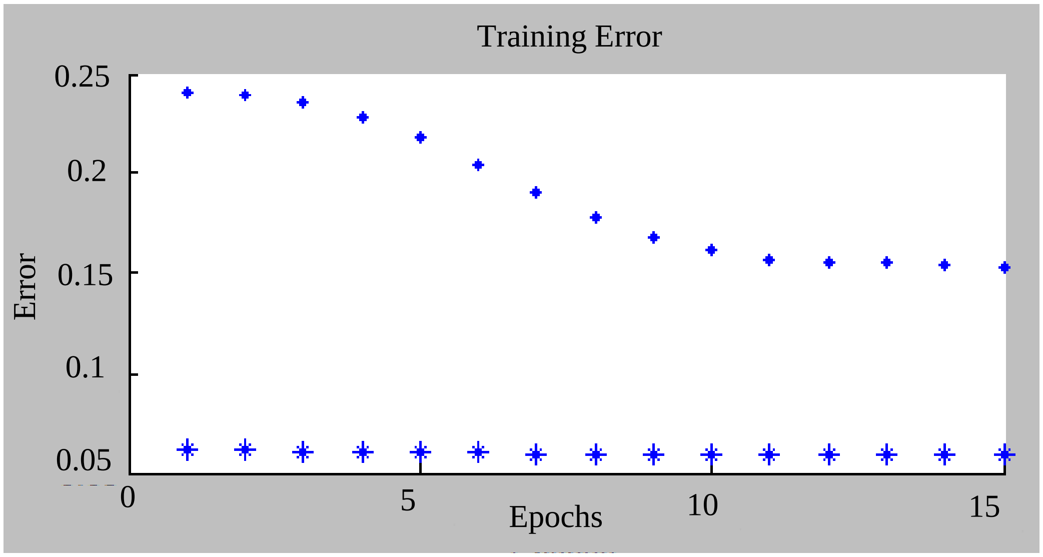

Figure 4.4. MATLAB's plot of training and testing errors

4.3.3.5 Cross-Validation

To further validate the models, choose the best method, and address the concern of overfitting, cross validation was performed for each of the above-mentioned methods. Overfitting is usually observed in a complex model when the model does not capture the underlying relationship.. Cross-validation is a statistical method of evaluating and comparing learning algorithms by splitting data into two segments: one used to train a model and the other used to validate (test) the model. One version of cross-validation is k-fold cross-validation. K-fold cross-validation is an effective method of model validation when it is not possible for the researcher to collect new data (Mahmood and Khan, 2009). The 10-fold cross-validation (k = 10) is the most common technique used in data mining and machine learning. In 10-fold cross-validation, the data is first partitioned into 10 equally (or nearly equally) sized parts or folds. Then 10 iterations of training and validation are performed. Within each of the iterations, a different fold of the data is held-out for validation, while the remaining nine folds are used for calibration. Both the training set and the test set affect the performance measure. The training set affects the measurement indirectly through the learning algorithm whereas the composition of the test set has a direct impact on the performance measure (Bouckaert, 2003). The available data points were broken into 10 sections of 10% each through random selection in the R program (version 2.13.0). The mean square of error (*MSE* which is the expected value of the square of the deviation of the predicted values from the observed values) for each case was then calculated (equation 4-6), where E is the expected value (average). After obtaining the *MSE* of the test set for each fold, the average *MSE* was then determined. The method with the smallest average *MSE* was selected as the best method. The coefficient of determination (R^2) also used (equation 4-7) to evaluate the quality of the models (Lyman and Longnecker, 2010).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \qquad (4\text{-}8)$$

$$R^2 = \left[\frac{\Sigma_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \hat{\bar{y}})}{\sqrt{\Sigma_{i=1}^{N}(y_i - \bar{y})^2(\hat{y}_i - \hat{\bar{y}})^2}}\right]^2 \qquad (4\text{-}9)$$

Where, *N* is the number of observations, y and y are the observed and predicted values respectively. y and y are the averages of observed and predicted values respectively.

The next step is to identify the best single model. The common approach is to build a model from all the data based on the best method identified earlier. However, this can lead to overfitting and misleading results with no model validation. In order to address this problem, the best model in each method, created during 10-fold cross validation (using 90% of the data), was identified and used to predict $\surd IY_{50}$ and IQ_{50} on each of the 10 folds. The goal is to demonstrate that the best model for each method will not always generate the most reliable results, even though it might have higher R^2 and lower *MSE*.

4.4 RESULTS AND DISCUSSION

4.4.1 Regression Method

The regression models were created and applied to the 10 fold sets in order to measure their overall performances (Table 4.1). The *MSE*'s based on the test sets that were obtained from the models ranged from 0.0000784 to 0.0003779, and had an average value of 0.0001704. The R^2 values observed ranged from 0.43 to 0.89, with an average of 0.70. However, these test results varied from that of the training results based on 90% of the data. Half of the models showed better performance during validation based (test set) on R^2, while the other half showed a poorer performance. Meanwhile, the validation *MSE's* showed more favorable values than that of the calibration (training set) in 70% of the models. However, the performance of combined training and validation models were very close and around 0.72 for R^2 and 0.00016 for MSE, which provide the overall estimate of error.

Table 4.1. Results from the 10-fold cross-validation for the regression model

Model created/tested with fold #	90% of data Training Set		10% of data Test Set		100% of data Combined Set	
	R^2	*MSE* $(m^3/s)/km^2$	R^2	*MSE* $(m^3/s)/km^2$	R^2	*MSE* $(m^3/s)/km^2$
1	0.722	0.000162	0.680	0.0001422	0.719	0.0001601
2	0.713	0.0001669	0.794	0.0000913	0.720	0.0001597
3	0.713	0.0001665	0.763	0.0000940	0.720	0.0001596
4	0.746	0.0001417	0.587	0.0003302	0.719	0.000161
5	0.722	0.0001603	0.691	0.0001597	0.719	0.0001603
6	0.707	0.0001626	0.766	0.0001415	0.719	0.0001605
7	0.692	0.0001666	0.886	0.0001015	0.719	0.000160
8	0.758	0.0001355	0.426	0.0003779	0.719	0.0001603
9	0.704	0.0001690	0.844	0.0000784	0.720	0.0001597
10	0.729	0.0001578	0.587	0.0001872	0.718	0.0001608
Average	0.721	0.0001704	0.702	0.0001704	0.719	0.0001602

4.4.2 Fuzzy Regression Method

The fuzzy regression method results (Table 4.2) showed *MSE's* among the test sets ranging from 0.0001123 to 0.0004217, with an average *MSE* of 0.0002164. These values varied from the training *MSE'*s, with five out of the 10 models being higher and the other five showing lower values. The validation R^2*'s* recorded ranged from 0.38 to 0.84, with an average of 0.67. Unlike the classical regression approach, the validation based on R^2 showed better performance than of the training in majority (70%) of the folds. However, the two models showing the worst validation (0.38 and 0.48), showed the highest R^2 in the training data (0.73 and 0.70). This can be attributed to overfitting in some folds and proves the benefits of 10-fold validation to spot these types of issues with each method.

Table 4.2. Results from the 10-fold cross-validation for the fuzzy regression model

Model created/tested with fold #	90% of data Training Set		10% of data Test Set		100% of data Combined Set	
	R^2	*MSE* $(m^3/s)/km^2$	R^2	*MSE* $(m^3/s)/km^2$	R^2	*MSE* $(m^3/s)/km^2$
1	0.663	0.0002137	0.695	0.0001434	0.664	0.000207
2	0.660	0.0004053	0.681	0.0002237	0.664	0.000388
3	0.670	0.0002150	0.736	0.0001300	0.679	0.0002070
4	0.736	0.0001652	0.571	0.0002929	0.709	0.0001782
5	0.684	0.0002041	0.738	0.0001123	0.688	0.0001947
6	0.674	0.0002009	0.835	0.0002188	0.691	0.0002027
7	0.658	0.0002050	0.802	0.0002240	0.678	0.0002070
8	0.733	0.0001712	0.377	0.0004217	0.693	0.0001968
9	0.670	0.0002081	0.773	0.0001448	0.683	0.0002016
10	0.703	0.0001910	0.470	0.0002516	0.685	0.0001972
Average	0.685	0.0002180	0.668	0.0002164	0.683	0.0002181

4.4.3 Fuzzy Expert System

The validation R^2 and *MSE* for fuzzy expert models (Table 4.3) ranged from 0.64 to 0.89 and 0.000051 to 0.00034, respectively. The average validation R^2 for the test sets was 0.77, while the average validation *MSE* was 0.0001665. Overall, the validation performance was not as good as the training performance on 60% of the cases based on both calculated R^2 and *MSE* values. The models performance based on the training data showed an averaged R^2 of 0.83 and an average *MSE* of 0.0001296. Overall, the performance of the models based on the 10-folds showed larger R^2's and lower *MSE*'s than the other methods. These better results show that the fuzzy models are better capturing the non-linear relationships between the variables while preventing the overfitting.

Table 4.3. Results from the 10-fold cross-validation for the fuzzy expert model

Model created/tested with fold #	90% of data Training Set		10% of data Test Set		100% of data Combined Set	
	R^2	*MSE* $(m^3/s)/km^2$	R^2	*MSE* $(m^3/s)/km^2$	R^2	*MSE* $(m^3/s)/km^2$
1	0.818	0.0001332	0.709	0.0001419	0.812	0.0001341
2	0.827	0.0001342	0.836	0.0000867	0.814	0.0001301
3	0.820	0.0001390	0.883	0.0000510	0.823	0.0001310
4	0.838	0.0001167	0.744	0.0002296	0.824	0.0001282
5	0.838	0.0001177	0.705	0.0001340	0.829	0.0001193
6	0.825	0.0001391	0.733	0.0002662	0.819	0.0001520
7	0.804	0.0001380	0.895	0.0001050	0.814	0.0001350
8	0.827	0.0001211	0.675	0.0002429	0.810	0.0001335
9	0.823	0.0001251	0.861	0.0000675	0.825	0.0001192
10	0.833	0.0001312	0.640	0.0003404	0.809	0.0001526
Average	0.825	0.0001296	0.768	0.0001665	0.818	0.0001335

4.4.4 ANFIS

In Table 4.4, the results from the ANFIS method are presented, including both R^2 and *MSE*. The validation R^2 values ranged from 0.57 to 0.88, having an average of 0.75. In seven out of the 10 models, these values were higher than that of their corresponding training values. The three models with the worst validation (R^2) had three of the highest calibration values, one of which was 0.93. This once again demonstrates the benefits of the 10-fold cross validation technique in preventing overfitting. *MSE* values, based on test sets, ranged from 0.000098 to 0.000302, having an average of 0.0001684. Unlike R^2, the validation *MSE* values showed a different story with only four of the 10 models showing better performance during validation. Overall, although the ANFIS model showed promising results compared to regression and fuzzy regression, on average it did not outperform the fuzzy expert model and the overall estimate of error based on combined set is also larger than the fuzzy expert model.

Table 4.4. Results from the 10-fold cross-validation for the ANFIS model

Model	90% of data Training Set		10% of data Test Set		100% of data Combined Set	
created/tested with fold #	R^2	*MSE* $(m^3/s)/km^2$	R^2	*MSE* $(m^3/s)/km^2$	R^2	*MSE* $(m^3/s)/km^2$
1	0.537	0.0003912	0.778	0.0001626	0.552	0.0003694
2	0.850	0.0000870	0.877	0.0001404	0.840	0.0000921
3	0.856	0.000084	0.652	0.0001920	0.836	0.000094
4	0.932	0.0000380	0.688	0.0002556	0.896	0.0000601
5	0.723	0.0001598	0.767	0.0001077	0.729	0.0001545
6	0.583	0.0002391	0.742	0.0001551	0.609	0.0002306
7	0.760	0.0001300	0.843	0.0001440	0.769	0.0001320
8	0.764	0.0001327	0.565	0.0003021	0.738	0.0001500
9	0.711	0.0001651	0.756	0.0001261	0.717	0.0001612
10	0.653	0.0002039	0.805	0.0000983	0.666	0.0001931
Average	0.737	0.0001631	0.747	0.0001684	0.735	0.0001637

4.4.5 Selecting the Best Method

The summary of different method's prediction capabilities of water yield (√IY_{50}) is presented in Table 4.5. According to the results shown in this table, for the fuzzy expert model the average R^2 was the highest with a value of 0.77 and the average *MSE* was the lowest with at 0.000167. These values indicate that the fuzzy expert system was able to accurately predict the square root of water yield and out-perform the other modeling methods and did so only using four variables (*H-Trans, Forest, A-Soils, and Precip)* comparing to six variables that were used in regression and fuzzy regression methods. The nonlinear nature of the fuzzy expert model allows the elimination of two less related variables (*L-Trans* and *D-Soils*) while the remaining four variables were successfully represented by both physiographical and climatic characteristics of the region.

Although the ANFIS method still out-performed the regression and fuzzy regression methods with the same number of variables, it was not able to predict better than the fuzzy expert method. This is illustrated with a R^2 and *MSE* of 0.75 and 0.0001684, respectively. This may be explained by lack of sufficient available data points for training the ANFIS method successfully. In addition, in order to keep the number of training data points below the number of modifiable parameters during the training process in MATLAB, only two membership functions were built into the model (even though one can introduce more than two MFs). This limited number of membership functions may not fully capture the hidden variably of the data set. These results may also be attributed to the overfitting of the data, providing poorer performance during validation. This kind of problem can be avoided if larger dataset is used during the training of the models.

Out of the four methods, fuzzy regression performed the worst, having the lowest average R^2 and the highest average *MSE* among the validation sets. Although the idea behind fuzzy regression method is to reduce the effect of outliers on the fitness, but this outcome did not appear in modeling th**e** $\sqrt{IY_{50}}$.

Overall, it is important to keep in mind that the main goal of statistical analysis here is to provide the repeatable prediction that can be generalized for the region. Based on the above argument, the fuzzy expert method is the best method to be generalized for the region.

Table 4.5. Average performance evaluations for all methods in predicting$\sqrt{IY_{50}}$ based on 10-Fold

Method	Average R^2	Average MSE [(m^3/s)/km^2]
Regression	0.70	0.0001704
Fuzzy Regression	0.67	0.0002164
Fuzzy Expert	**0.77**	**0.0001670**
ANFIS	0.75	0.0001684

4.4.6 Selecting the Best Model for each Method

As it was shown in Tables 4.1 to 4.4 for all of the four methods, 10 different models were built based on the 10-folds. In this section, each model within each method was evaluated using all of the 10 test sets and once again an average R^2 and MSE were calculated. The model, which on average has the lowest MSE between the 10 test sets, was selected as the best local model for that method. Table 4.6 shows the results of average R^2 and MSE of the 10 test sets for all 10 models for each method. In addition, the best model in each of the four methods is bold in table 4.6. The best model chosen for the fuzzy expert method (highlighted) exhibited an average R^2 of 0.83 and MSE of 0.000118 across all the data folds. Although the best models and their performance reflect the results from the best method analysis, there is one substantial difference. From the data in Table 4.6, one can conclude that the best local model of ANFIS performed better than the fuzzy expert, explaining an additional 7% of the variation in $\sqrt{IY_{50}}$. In addition, the MSE was approximately half of that of fuzzy expert. However, it is important to recognize that this is the results for only one (the best) of the 10 ANFIS models built in the 10 fold cross validation process. Other models built based on other folds, however, can perform worse than the fuzzy expert system, which is one of the motivations for the 10-fold cross validation. Therefore, if the concern is to develop the best model for the dataset in hand, the recommended method is the ANFIS. However, for developing the regional model, the best method is the fuzzy expert system. In the next step, performances of the best local models were further evaluated using all data points (Table 4.7). As expected, the results from this analysis show the best ANFIS model outperforms other best local models.

Table 4.6. Average performance of all 10 models for the best model predicting$\sqrt{IY_{50}}$

	Regression		Fuzzy Regression		Fuzzy Expert		ANFIS	
	R^2	*MSE* $(m^3/s)/km^2$	R^2	*MSE* $(m^3/s)/km^2$	R^2	*MSE* $(m^3/s)/km^2$	R^2	*MSE* $(m^3/s)/km^2$
1	0.719	0.0001590	0.664	0.0002057	0.821	0.0001331	0.570	0.0003658
2	0.719	0.0001586	0.664	0.0003854	0.835	0.0001292	0.834	0.0000922
3	**0.719**	**0.0001580**	0.679	0.0002060	0.828	0.0001300	0.822	0.0000950
4	0.718	0.0001599	**0.708**	**0.0001769**	0.830	0.0001269	**0.897**	**0.0000596**
5	0.718	0.0001592	0.687	0.0001933	0.833	0.0001181	0.733	0.0001535
6	0.716	0.0001593	0.690	0.0002013	0.818	0.0001507	0.613	0.0002298
7	0.718	0.0001612	0.677	0.0002084	0.818	0.0001371	0.774	0.0001325
8	0.717	0.0001591	0.693	0.0001954	0.810	0.0001323	0.742	0.0001489
9	0.719	0.0001586	0.681	0.0002002	**0.834**	**0.0001180**	0.727	0.0001598
10	0.717	0.0001596	0.683	0.0001958	0.815	0.0001509	0.672	0.0001919

Table 4.7. Performance of the best models for predicting$\sqrt{IY_{50}}$ based on all data

Method	R^2	*MSE* $[(m^3/s)/km^2]$
Regression	0.72	0.000158
Fuzzy Regression	0.71	0.000177
Fuzzy Expert	0.83	0.000118
ANFIS	0.90	0.000060

4.4.7 Relating the Index Water Yield Results to the Index Flow

As it was discussed earlier, the overall objective of this study is to develop a more robust technique(s) for estimating the index flow for ungaged streams for the state of Michigan. Previous efforts resulted in developing multiple models that estimate the square root of index water yield. Therefore, in order to estimate the index flow from squared root of water yield, the predicted values must be squared and multiplied by the area of the basin upstream of the streamflow station. This allows for comparison with the observed values from the streamflow gaging station. Knowing how the model performs in predicting these values is ultimately what is important. The prediction of index flows enables decision makers and others to directly relate watershed characteristics to water withdrawals limitations, ecological health benchmarks, and other flow related issues. Table 4-8 shows the results of the best performing index flow model for the four different methods. In this table the advantages of fuzzy systems over the regression model is more clear, with fuzzy methods providing higher R^2 and lower *MSE* values for all methods. This even includes fuzzy regression techniques. As briefly mentioned in previous sections, the performance of fuzzy regression is worse than that of regression for predicting when predicting $\sqrt{IY_{50}}$. However, when considering the actual index flows and not the transformed measures ($\sqrt{IY_{50}}$), the effect of larger outliers are more prevalent and have more of an influence on the performance of the models. Meanwhile, unlike in the above predictions of $\sqrt{IY_{50}}$, the fuzzy expert system showed the highest R^2 value of 0.98 for predicting the index flow. The best *MSE* however is achieved through the ANFIS model. The higher value of R^2 in the fuzzy logic model along with the lower value of *MSE* in the ANFIS model demonstrates the importance of the use of more than one evaluation technique to select the best predicting method.

However, based on the results obtained, it is expected that the fuzzy expert system provides better predictability at regional scale while, ANFIS model perform better within the sample dataset.

Table 4.8. Performance of the best models for predicting IQ_{50} based on all data

Method	R^2	MSE (m^3/s)
Regression	0.92	3.557
Fuzzy Regression	0.95	3.151
Fuzzy Expert	0.98	2.429
ANFIS	0.97	1.407

4.4.8 Graphical Comparison of Model Predictions versus Observed for Square Root of Index Water Yield

It is imperative to identify the range and distribution of sample data points graphically in order to analyze the relationship. This also depends on the true distribution of the population and whether the model data reflects this. If this is not the fact, modeling and computing the relationship may not be robust or accurate (Ayyub and McCuen, 2003). Results and comparisons among modeling methods are also demonstrated through plots in Figures 4.5 through 4.8. These plots represent the relation between predicted $\sqrt{IY_{50}}$ (the estimated by each method) and observed $\sqrt{IY_{50}}$ (basis on long-term streamflow-gaging station records). These figures also represent the trend line (solid line) along with the 95% confidence intervals (dashed line). On the top and right sides of each figure, two histograms were provided showing the distribution of the observed and predicted $\sqrt{IY_{50}}$ values, respectively. Overall, the ranges of observed data are fairly well distributed between 0 and 0.1196, as is shown in top histograms. Meanwhile, the ranges for predicted values for regression, fuzzy regression, fuzzy expert, and ANFIS models are 0.0097 to

0.1045, 0.0192 to 0.0930, 0.0209 to 0.0910, and 0.0107 to 0.1192, respectively. Between all of these methods, the ANFIS model preserved the range closest to observed range. Regarding the means and standard deviations, the ANFIS model's predictions have similar values to that of the observed data (0.0551 and 0.0239, respectively). However, standard deviations were lower than the observed value for the rest of the models. Regarding histograms of the predicted distributions, both classical regression and fuzzy expert models show deviations from having a normal distribution, while fuzzy regression and ANFIS models to a large extent preserve the normal distribution.

Overall, model prediction behavior is best described by the trend lines for residuals. As it was shown in Figures 4.5 through 4.8, no model has perfect prediction, as demonstrated by a non-zero intercept value and deviation from the 45° line. Among the models, the regression and fuzzy regression models show a very similar trend, while a higher level of variation can be seen around the regression model. As mentioned earlier, the advantage of fuzzy regression is the ability of the method to reduce the effects of outliers on the model, which is not revealed by the plots. Meanwhile, the fuzzy expert and ANFIS model show a smaller level of variation to the trend lines except a single outlier in the ANFIS model. Figures 4.9 through 4.12 show the residual plots of $\sqrt{IY_{50}}$ predictions using the different models. These graphs were created to display how well the model predictability matched the observed data. In this type of graph, a horizontal trend line with zero intercept represents a perfect prediction. Based on these criteria, ANFIS and fuzzy expert can be considered the better models. In general, all models were over-predicting. Meanwhile, the ranges of residual values for regression, fuzzy regression, fuzzy expert, and ANFIS models are 0.0857, 0.0852, 0.0570, and 0.0680, respectively. Therefore, the fuzzy expert model shows the lowest range in variation of residuals. The lowest standard deviation was observed for the ANFIS model (0.0077), followed by fuzzy expert (0.0102), regression (0.0127), and fuzzy regression (0.0.0133). However, the fuzzy expert model is the only model to reduce the effects of outliers and show no outlying residuals.

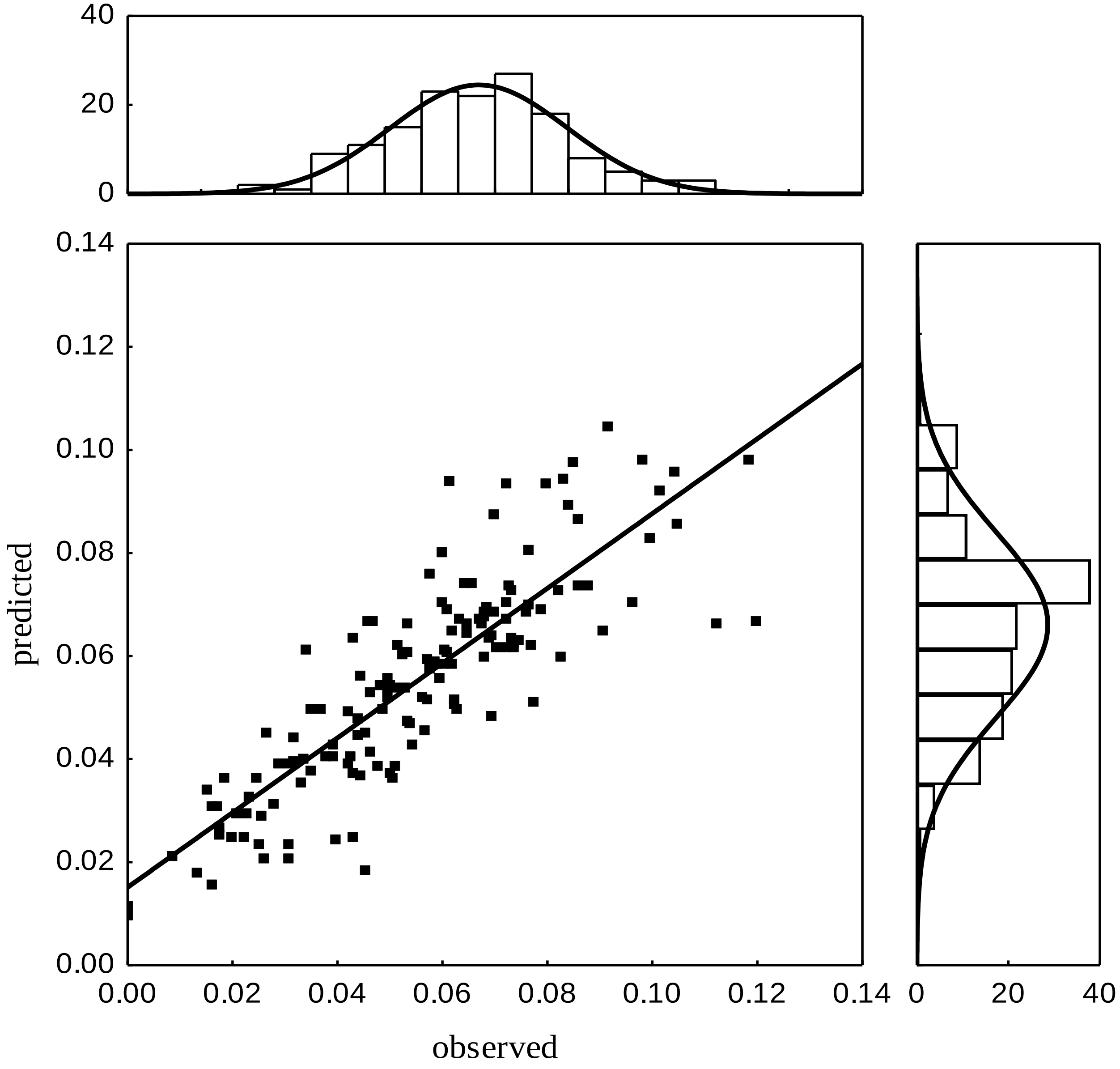

Figure 4.5. Relation between observed and predicted $\sqrt{IY_{50}}$ by the best regression model [(m^3/s)/km^2]

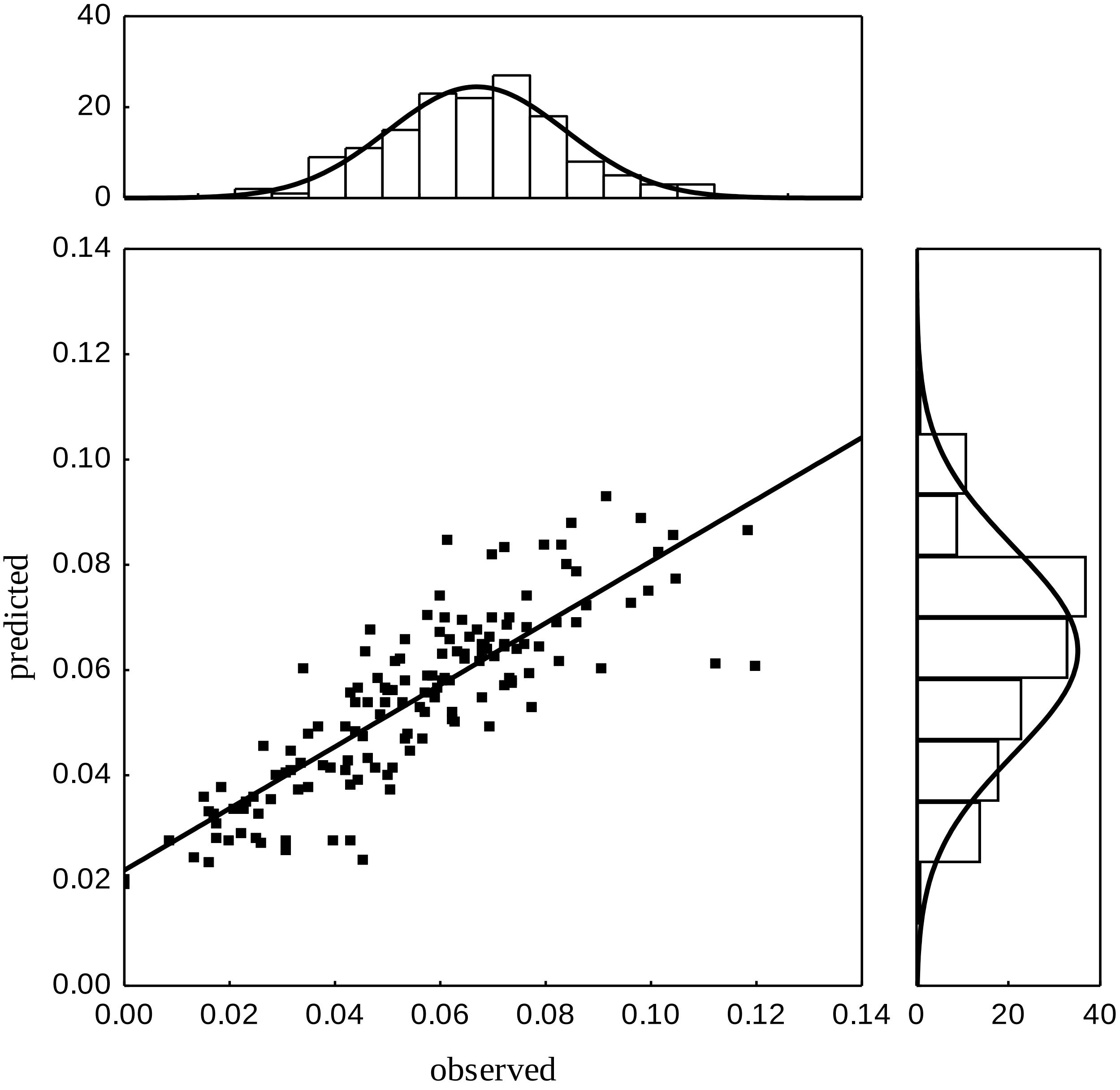

Figure 4.6. Relation between observed and predicted $\sqrt{IY_{50}}$ by the best fuzzy regression model [(m^3/s)/km^2]

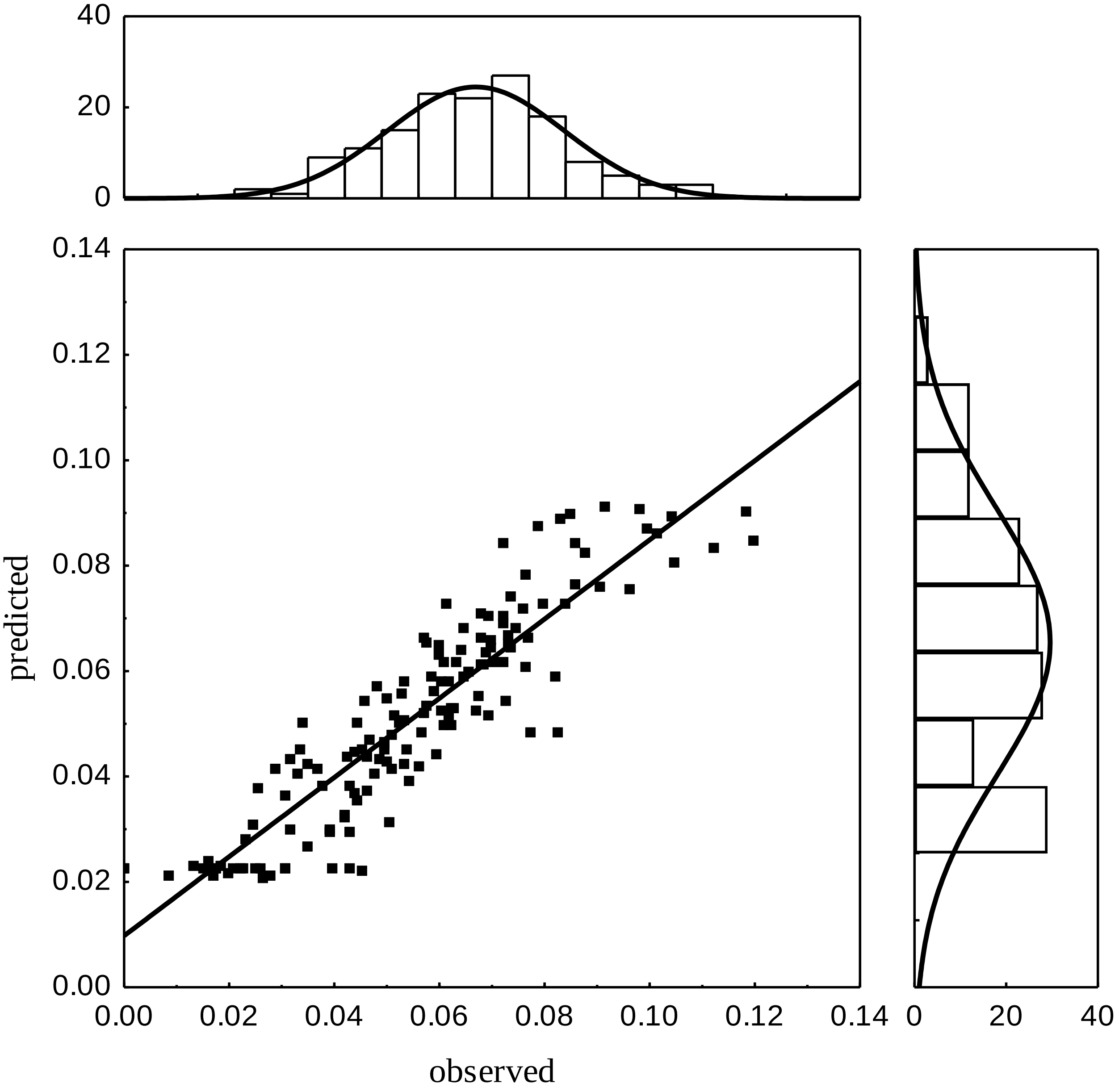

Figure 4.7. Relation between observed and predicted $\sqrt{IY_{50}}$ by the best fuzzy expert model $[(m^3/s)/km^2]$

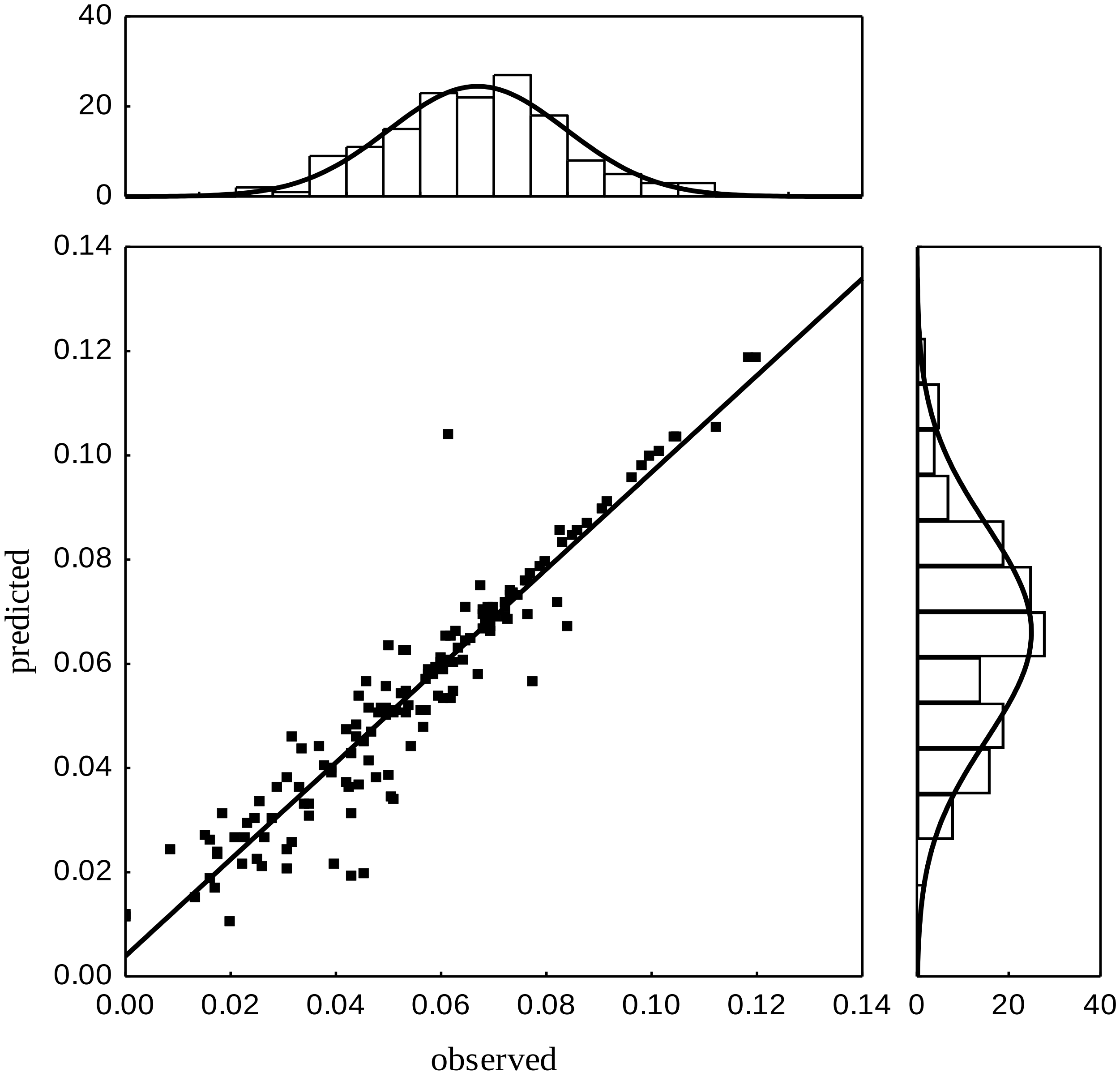

Figure 4.8. Relation between observed and predicted $\sqrt{IY_{50}}$ by the best ANFIS model $[(m^3/s)/km^2]$

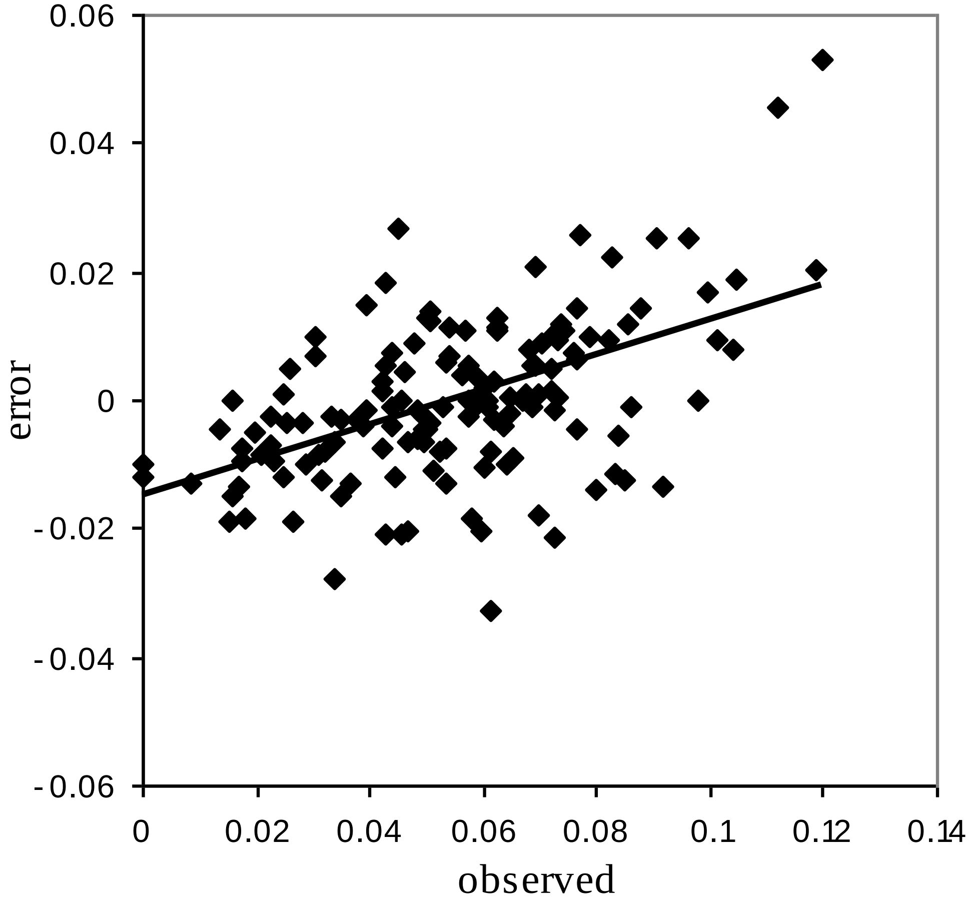

Figure 4.9. Residual plot of √*IY₅₀* predictions using the regression model [(m^3/s)/km^2]

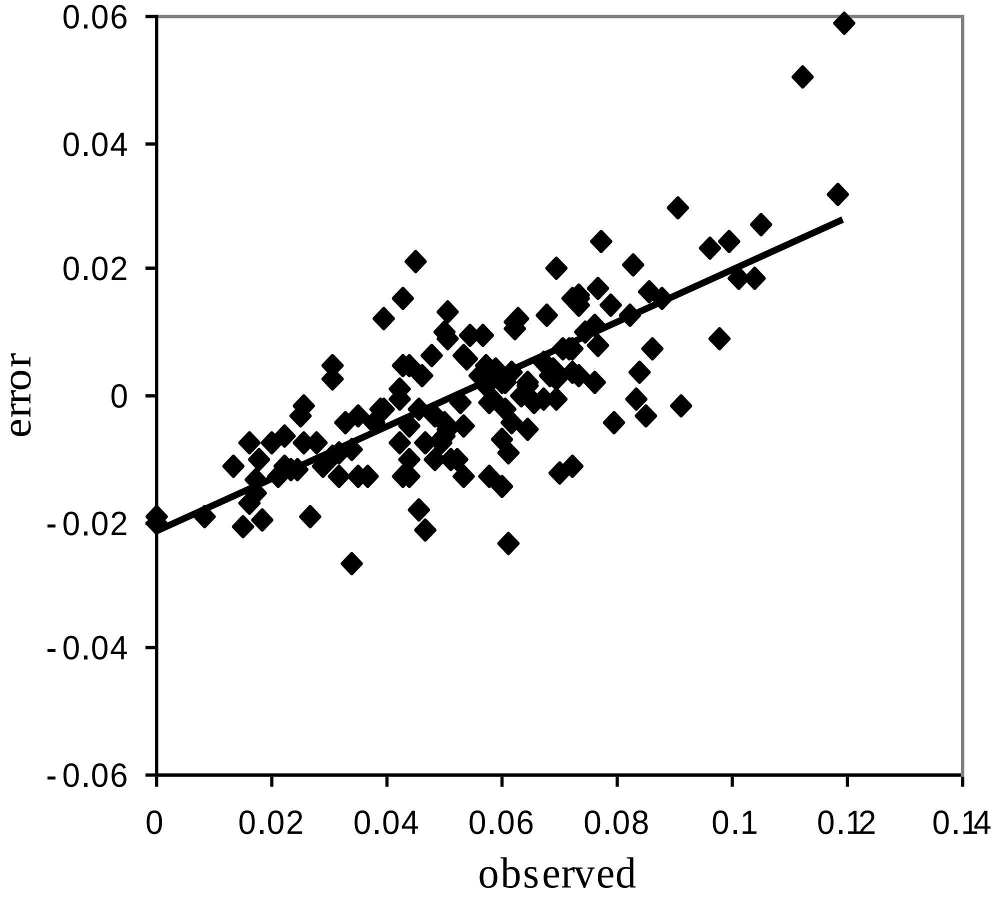

Figure 4.10. Residual plot of √*IY*$_{50}$ predictions using fuzzy regression model [(m^3/s)/ Km^2]

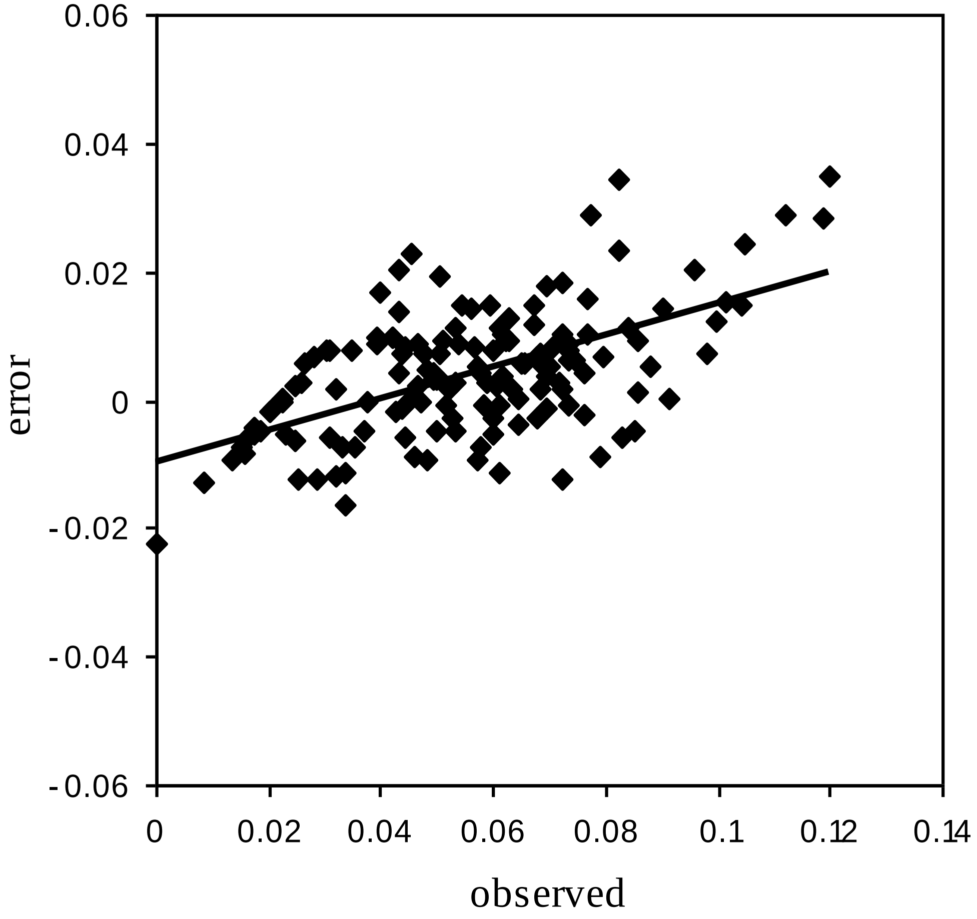

Figure 4.11. Residual plot of $\sqrt{IY_{50}}$ predictions using fuzzy expert model [(m^3/s)/km^2]

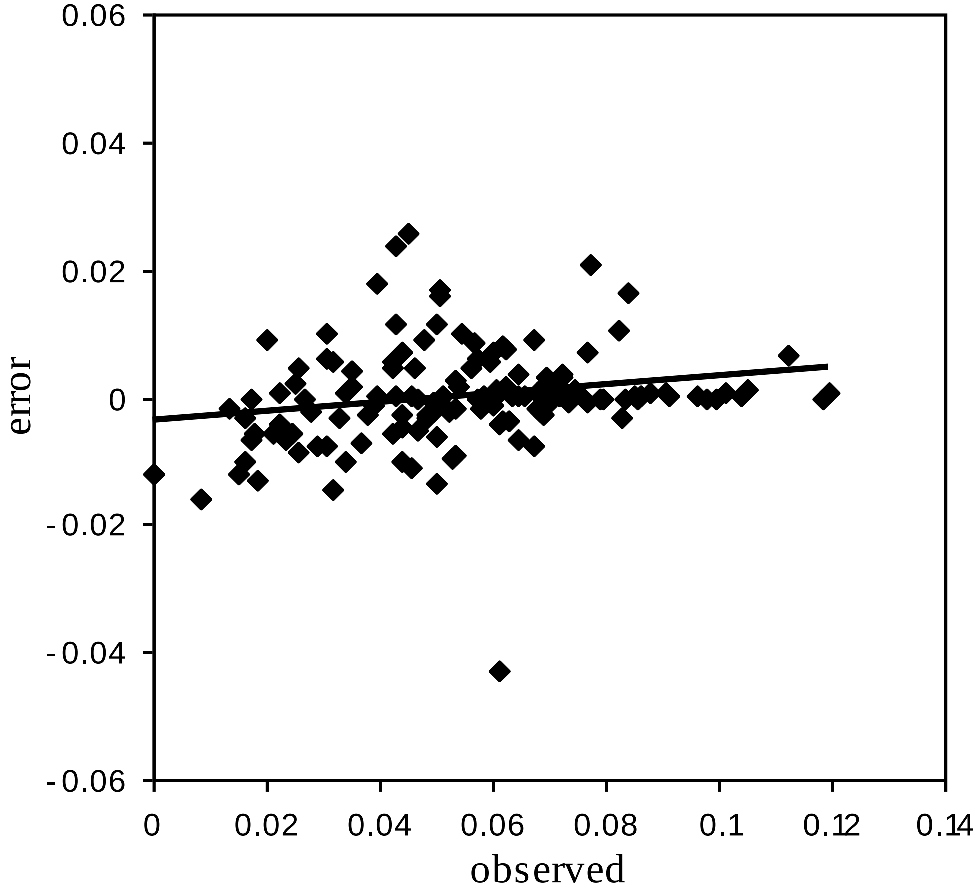

Figure 4.12. Residual plot of √IY_{50} predictions using ANFIS model [(m^3/s)/km^2]

4.4.9 Graphical Comparison of Model Predictions versus Observed for the Index Flow

In this section, the capabilities of the model predictions for the index flow were evaluated. Figures 4.13 through 4.16 represent the relation between predicted IQ_{50} (the estimated by each method) and observed IQ_{50} (basis on long-term streamflow-gaging station records).

Overall, all models performed very well, with R^2 values above 0.9. The range of observed data is fairly well distributed between 0 and 52.386 (m^3/s). Meanwhile, the ranges for predicted values for regression, fuzzy regression, fuzzy expert, and ANFIS models are 0.0018 to 38.048, 0.0021 to 39.129, 0.0013 to 44.236, and 0.0003 to 50.817 (m^3/s), respectively. Between all of these methods, the ANFIS model preserved the range closest to observed range, while the regression and fuzzy regression models did not. This result indicates the regression models are biased toward lower index flow values. Regarding the means and standard deviations, the ANFIS model's predictions have similar values to that of the observed data (3.288 and 6.446, respectively). However, standard deviations were lower than the observed value for the rest of the models.

Figures 4.17 through 4.20 show the residual plots of IQ_{50} predictions using the different models. As described above, a perfect model prediction can be demonstrated by a zero intercepting horizontal trend line of the residuals. Based on these criteria, both ANFIS and fuzzy expert can be considered the better models. In general, all models were over-predicting. Meanwhile, the ranges of residual values for regression, fuzzy regression, fuzzy expert, and ANFIS models are 22.758, 16.744, 10.071, and 18.523 (m^3/s), respectively. Therefore, the fuzzy expert model shows the lowest range in variation of residuals. However, the ANFIS model has the lowest residual sum (0.068). The lowest standard deviation was observed for the ANFIS model (1.161), followed by fuzzy expert (1.444), fuzzy regression (1.728), and regression (1.897). However, the fuzzy expert model is the only model showing no outlying residuals.

The result of this section indicate that the ANFIS model tries to preserve many attributes of the data set, possibly, at the expense of over-fitting and generating unexplained values such as outliers. However, no surprise was observed for the fuzzy expert model.

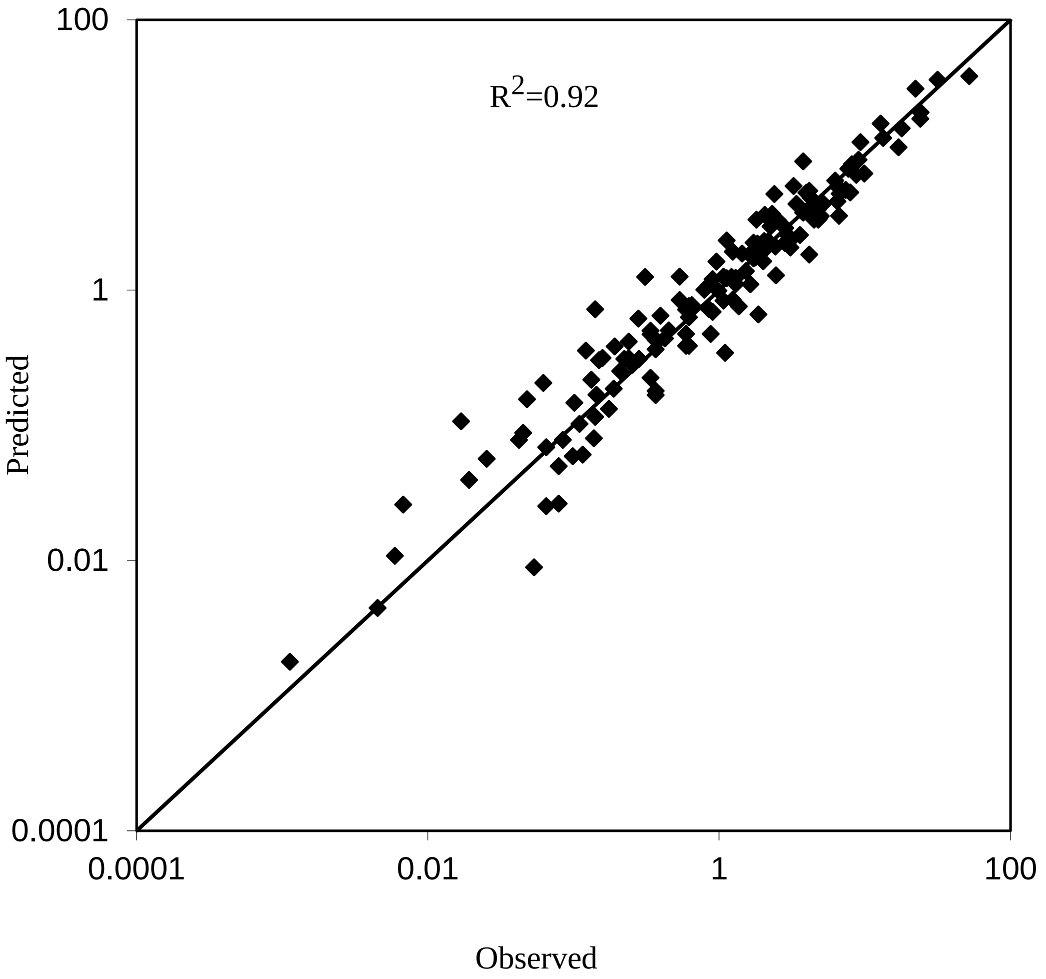

Figure 4.13. Relation between observed and predicted index flow *IQ_{50}* by the regression model (m^3/s)

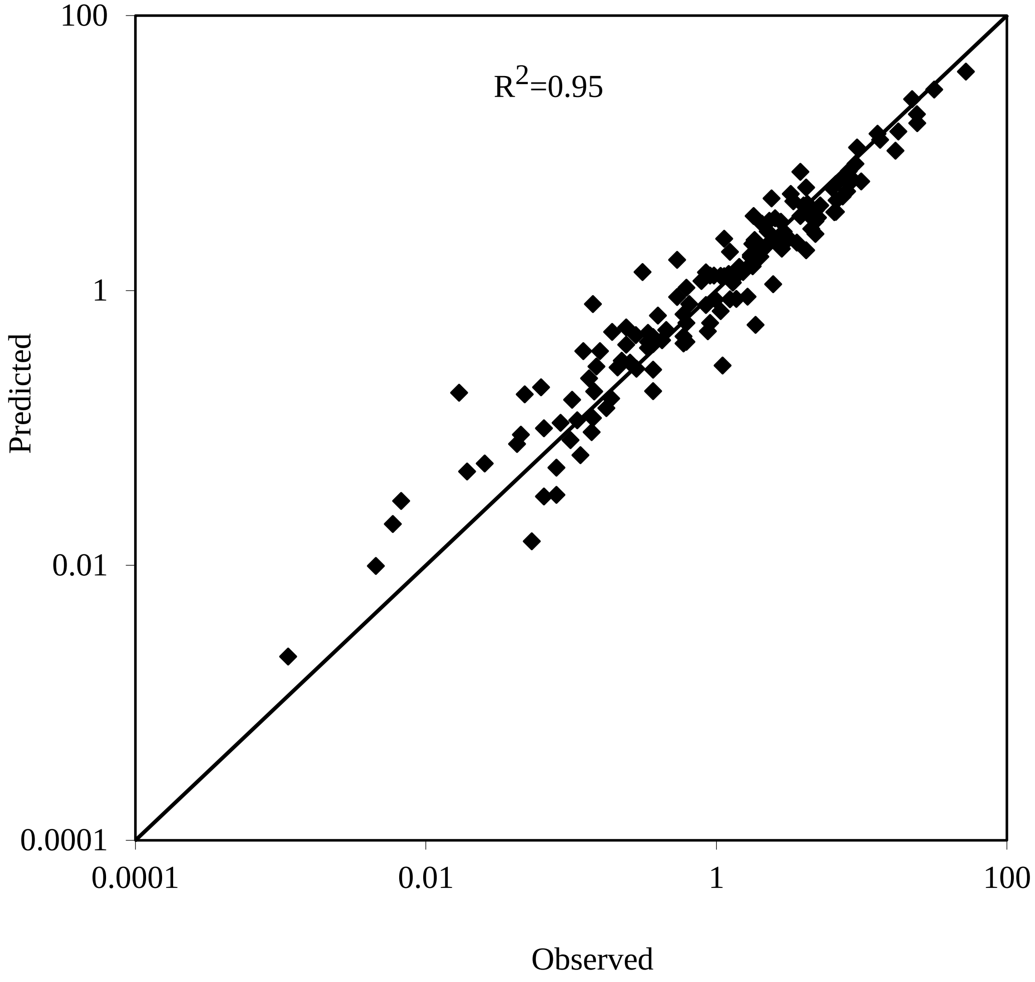

Figure 4.14. Relation between observed and predicted index flow IQ_{50} by the fuzzy regression model (m^3/s)

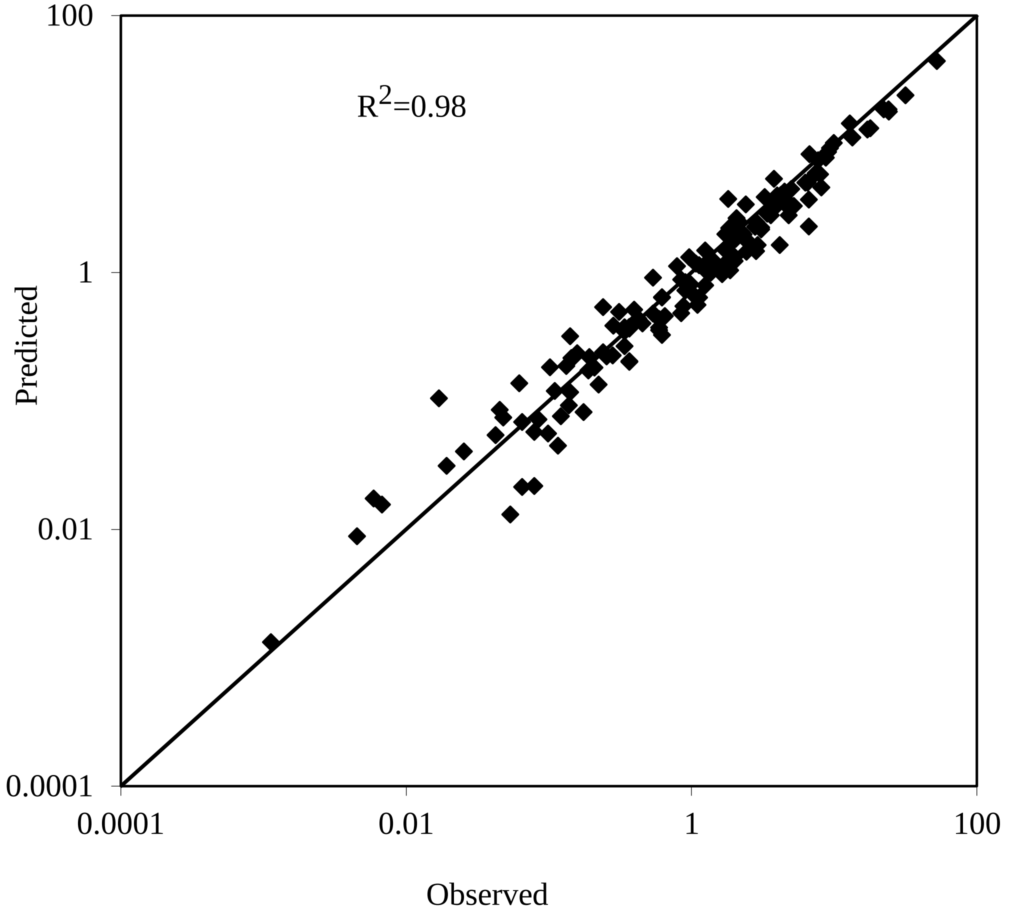

Figure 4.15. Relation between observed and predicted index flow *IQ_{50}* by the best fuzzy expert model (m^3/s)

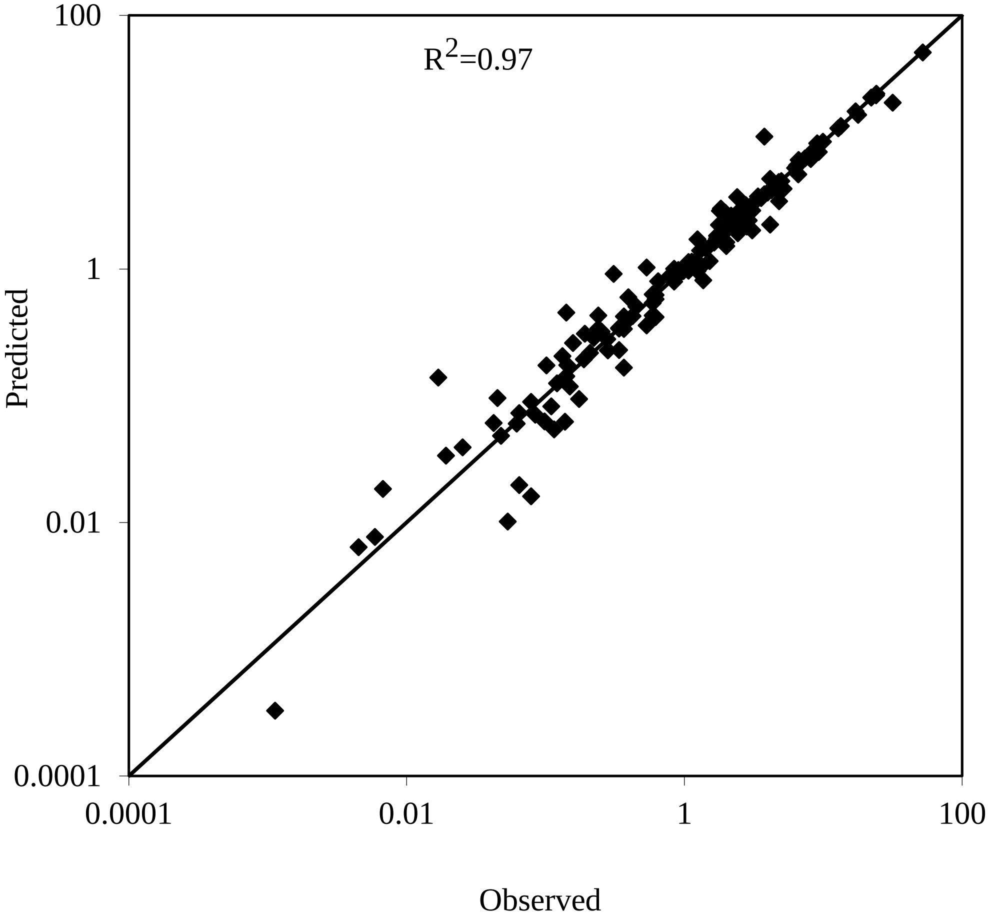

Figure 4.16. Relation between observed and predicted index flow *IQ_{50}* by the best ANFIS model (m^3/s)

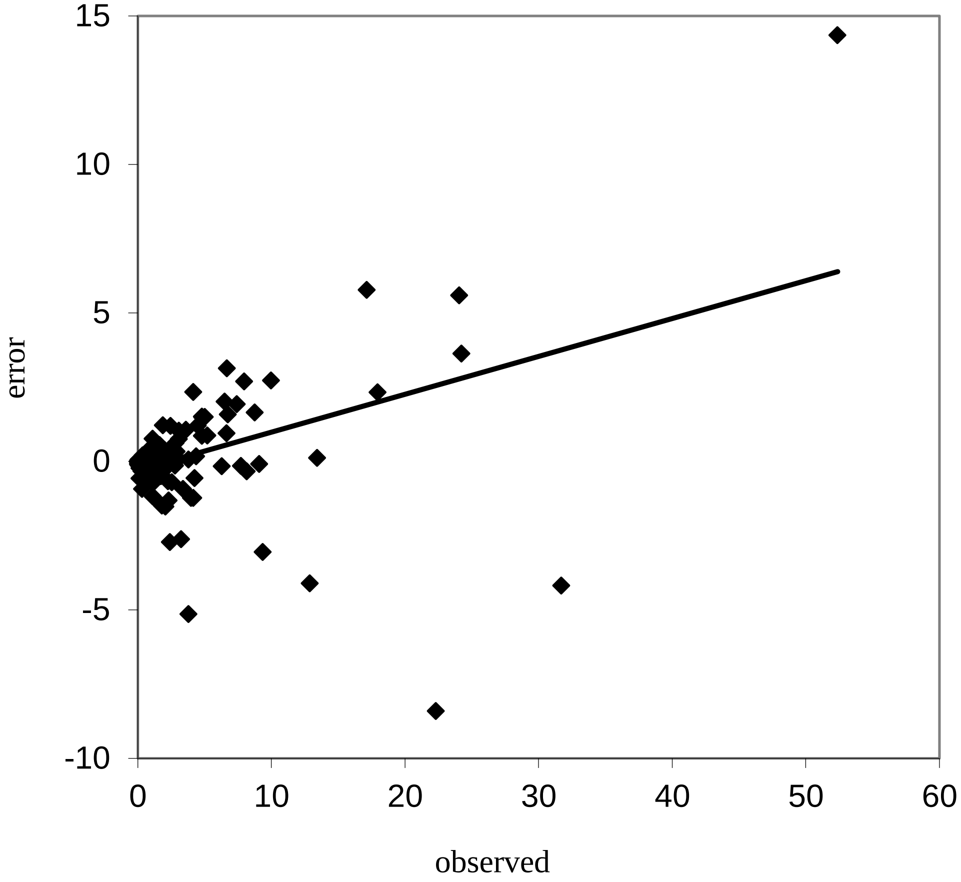

Figure 4.17. Residual plot of *IQ_{50}* predictions using regression model (m^3/s)

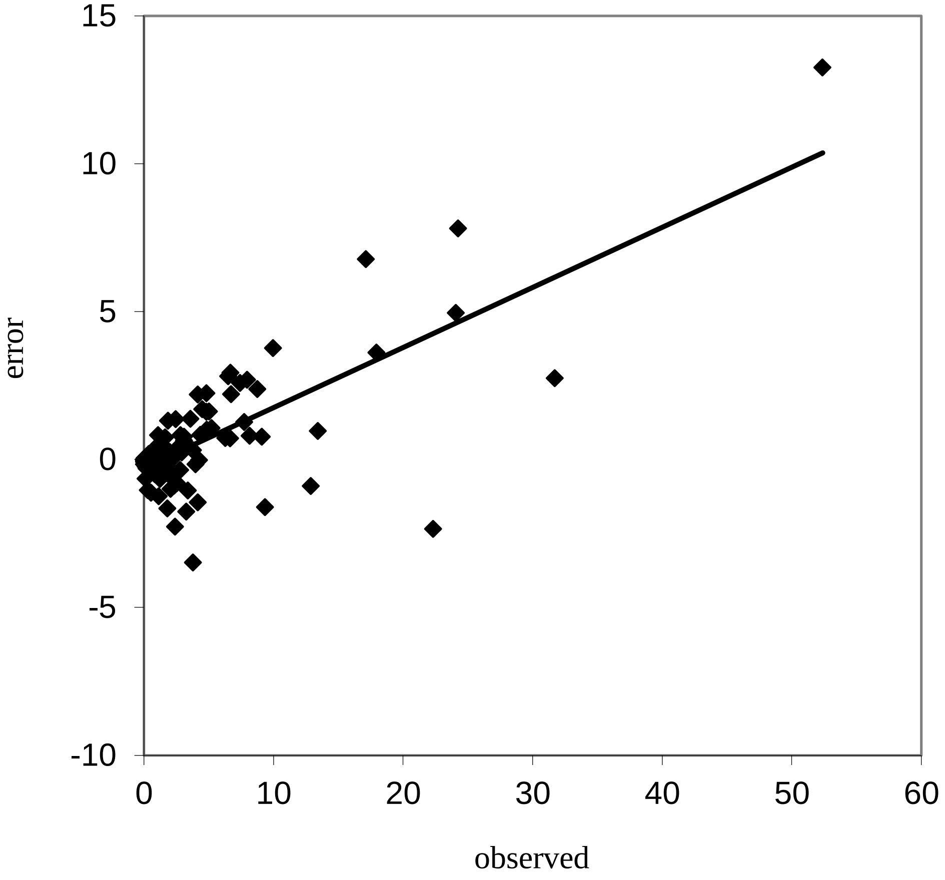

Figure 4.18. Residual plot of *IQ_{50}* predictions using fuzzy regression model (m^3/s)

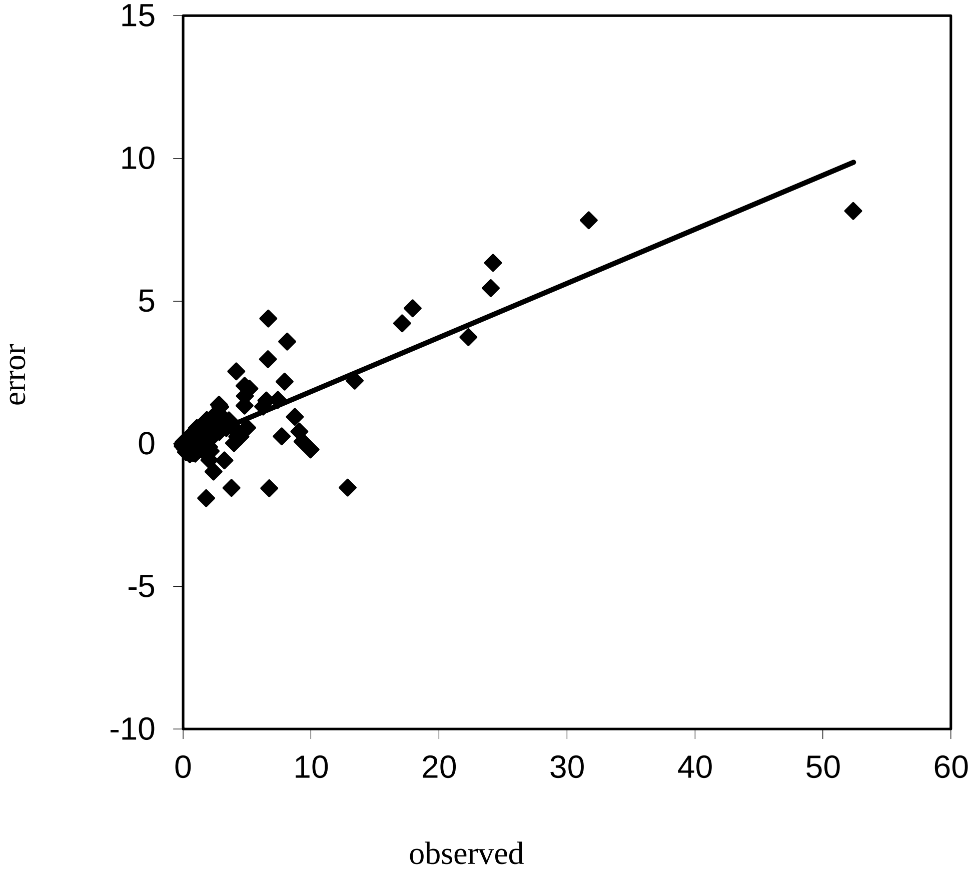

Figure 4.19. Residual plot of *IQ_{50}* predictions using fuzzy expert model (m^3/s)

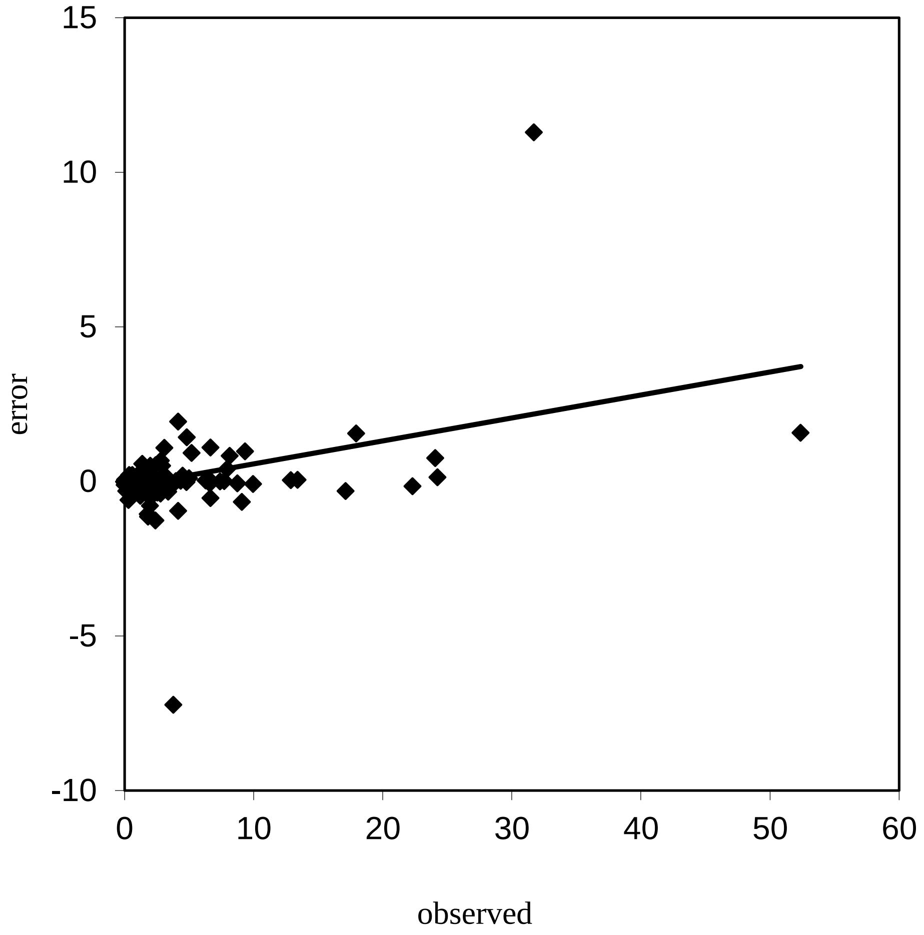

Figure 4.20. Residual plot of *IQ_{50}* predictions using ANFIS model (m^3/s)

4.4.10 Using All Data Points to Build and Compare Models

Based on the above discussion, the fuzzy expert method was selected as the best method among the studied methods. In the final stage of this study, the performance of the fuzzy expert model was compared with the regression model, which was originally developed by Hamilton et al (2008) for the state of Michigan. The difference between this section of study and previous sections is that all data points were used to develop a new fuzzy expert model. Tables 4.9 and 4.10 provide the performance values in predicting $\sqrt{IY_{50}}$ and IQ_{50}. Overall, the fuzzy expert model outperformed the regression model in the prediction of both $\sqrt{IY_{50}}$ and IQ_{50}, as shown by having a higher R^2 and lower *MSE* values. In addition, it is clear from Figures 4.21 through 4.24 that the fuzzy expert model clearly reduced the effect of outliers and therefore is a more robust model for predicting the index flow values for the state. However, the presence of outliers is very clear in the regression model, which can be resulted in flaw predictions. In addition, for the regression model, both the slope and absolute intercept residual values are higher than the expert model. This evidences are also indicate that the fuzzy expert model is the more reliable model. Comparing Figures 4.21 to 4.22 and 4.23to 4.24 also show more scattered values around the trend lines for the regression model. As a matter of fact, the residual ranges are about 50% and 160% wider for the regression models than the fuzzy expert models.

Table 4.9. Performance of regression and fuzzy expert models in predicting$\sqrt{IY_{50}}$, built using all data

Method	Average(R^2)	Average (*MSE*) [(m^3/s)/km^2]
Regression	0.72	0.0001595
Fuzzy Expert	0.83	0.0001190

Table 4.10. Performance of regression and fuzzy expert models in predicting IQ_{50}, built using all data

Model	Average(R^2)	Average (*MSE*) (m^3/s)
Regression	0.92	3.482
Fuzzy expert	0.98	2.434

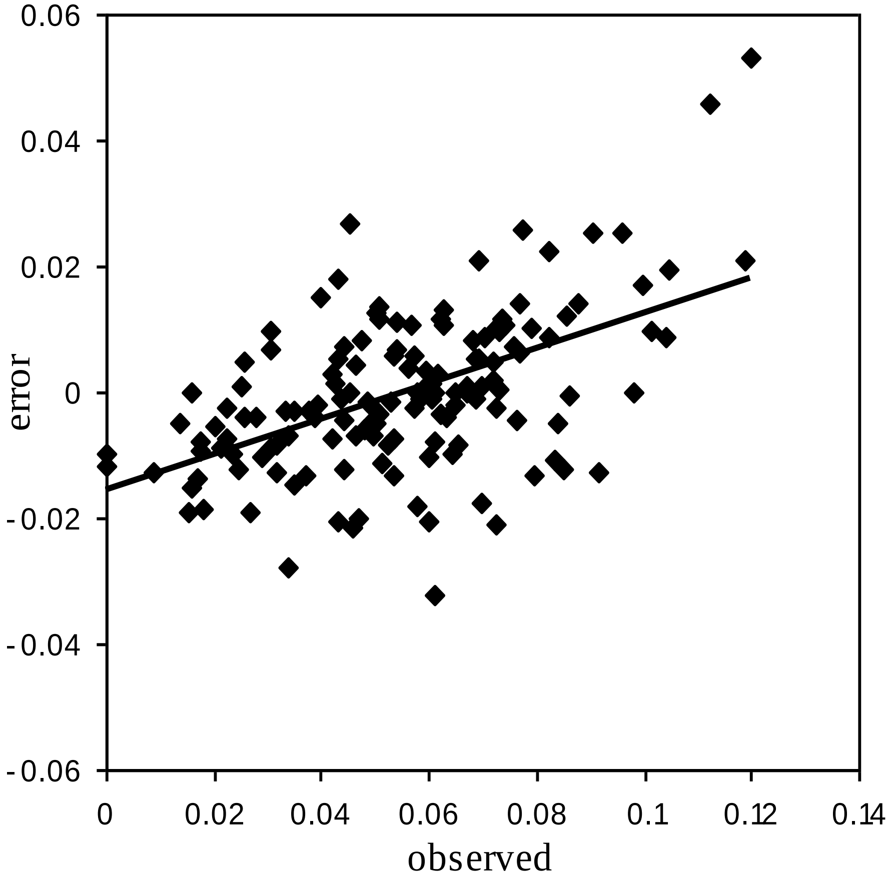

Figure 4.21. Residual plot of √*IY*$_{50}$ predictions using regression using all data points model [(m^3/s)/km^2]

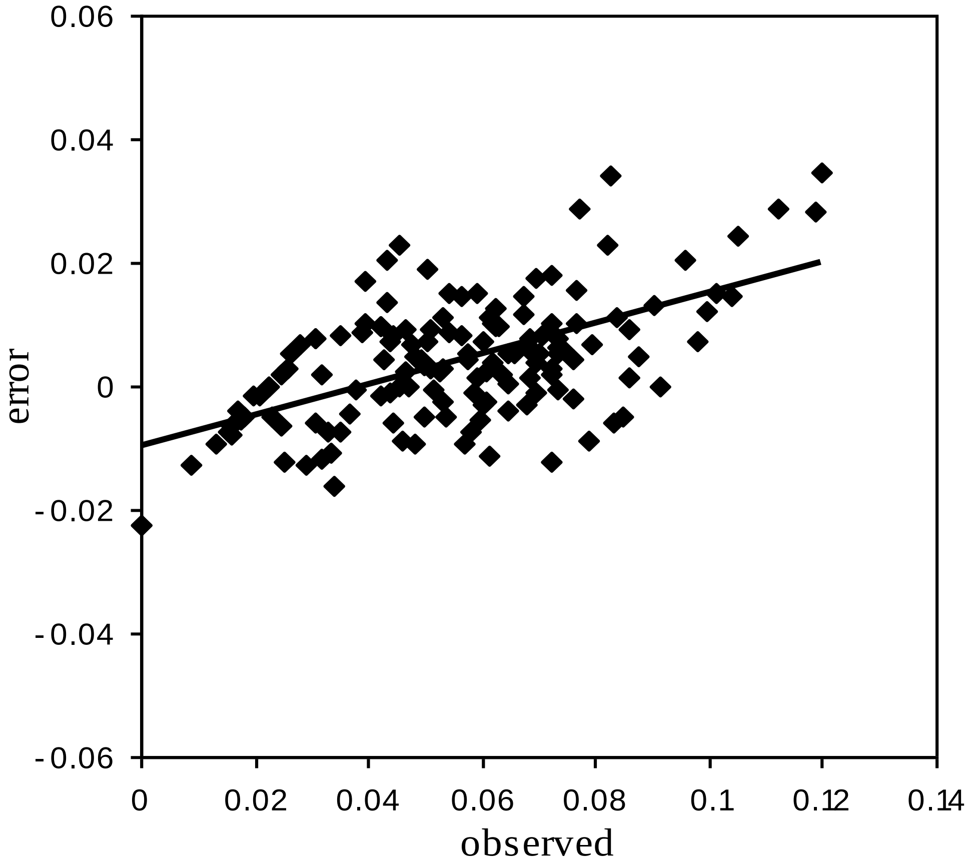

Figure 4.22. Residual plot of $\sqrt{IY_{50}}$ predictions using fuzzy expert fuzzy expert model [(m^3/s)/km^2]

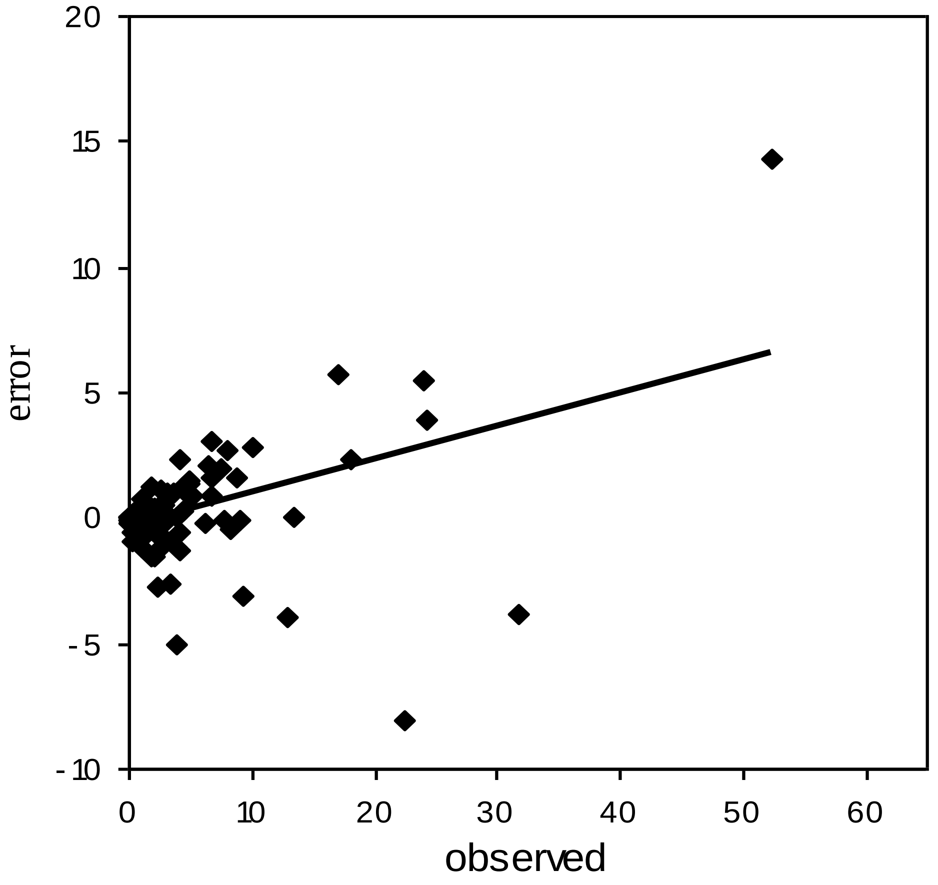

Figure 4.23. Residual plot of *IQ_{50}* predictions using regression model (m^3/s)

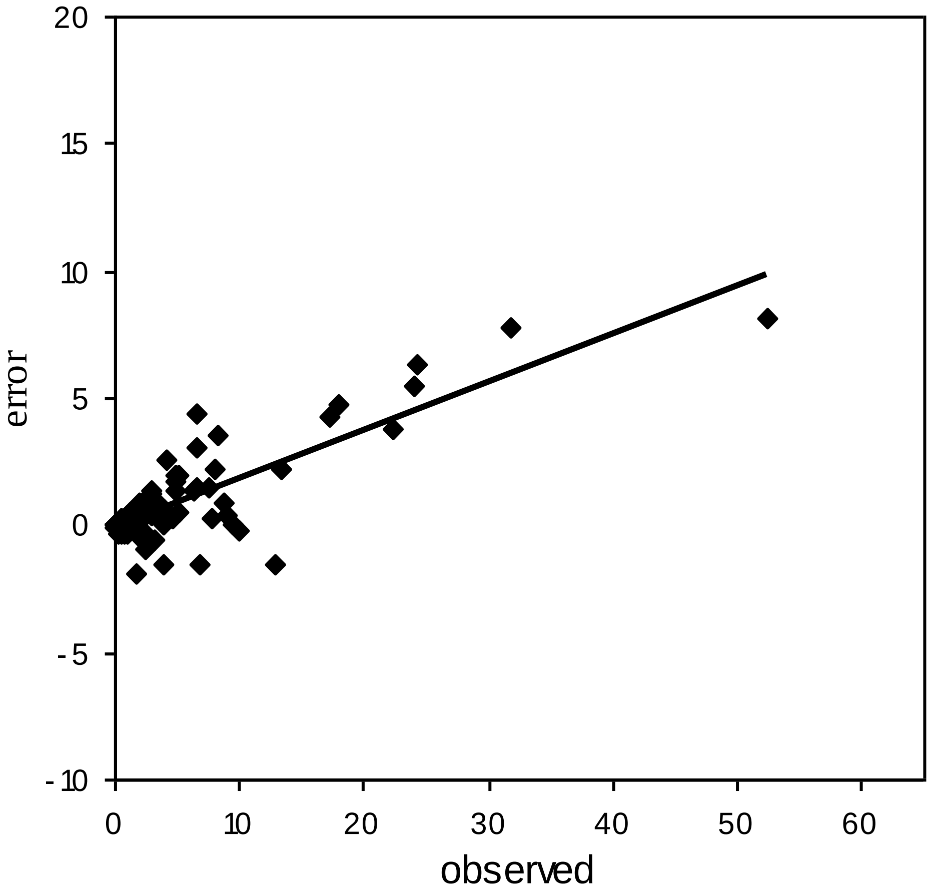

Figure 4.24. Residual plot of IQ_{50} predictions using fuzzy expert model (m^3/s)

4.5 CONCLUSION

Having the knowledge and the ability to accurately predict low and index flows within a watershed is essential to decision makers and watershed managers. With conventional methods not always being cost-effective (number of variables) and robust, there is growing interest in alternative soft computing methods such as fuzzy logic and artificial neural networks (Huang et al., 2010). In this study, the capabilities of different fuzzy techniques along with the a regression model developed by Hamilton et al. (2008) were evaluated and compared to develop a more robust model in predicting index flows for the State of Michigan. Three different fuzzy methods were used including fuzzy regression, fuzzy expert, and ANFIS. To further choose the best method and address the concern of overfitting, cross validation was performed for each of the above-mentioned methods. Results through the 10-fold cross validation demonstrated that the fuzzy expert method was the best with in predicting $\sqrt{IY_{50}}$ with an average R^2 of 0.77 and an average *MSE* of 0.0001670.

In the next step, the best model for each method was identified and compared. Overall, the ANFIS model outperformed other models with an average R^2 of 0.90, *MSE* of 0.0000601 for $\sqrt{IY_{50}}$, R^2 of 0.97 and *MSE* of 1.407 for IQ_{50}. In addition, in the ANFIS model, many attributes of the data points such as mean and standard deviation were preserved. However, results of the graphical analysis of predictor versus observed and residuals versus observed revealed that the ANFIS model is prone to overfitting and generating unexplained values such as outliers. Meanwhile, the fuzzy expert model provides the repeatable prediction that can be generalized for the region is very robust and reliable manner.

Among the studied methods for predicting the index water yield, the performance of regression and fuzzy regression models were lower with an average R^2 of 0.72 and 0.71 and an average *MSE* of 0.000158 and 0.000177, respectively. In fact, out of the four methods, fuzzy regression performed the worst, having the lowest average R^2 and the highest average *MSE* among the validation sets. Although the idea behind fuzzy regression method is to reduce the effect of outliers on the fitness (least square) method results and increase the effect of concentrated data points on the model, it does not seem to be as effective as other methods in the prediction of $\sqrt{IY_{50}}$ but it performed better than the regression method in predicting the IQ_{50}.

Over all the combination of statistical and graphical analysis confirmed the advantages of 10-fold analysis in selecting the best model while avoiding overfitting or generating unexplained values such as outliers. Finally, results of this study showed potential for fuzzy logic as powerful and robust technique compared to conventional techniques in predicting hydrologic variables such as index flow and water yield using fewer variables. Further study should explore alternative methods such as boosting algorithms for emerging issues in water resources such as climate changes, and extreme events.

5 EVALUATION OF NEURO-FUZZY AND BAYESIAN TECHNIQUES IN ESTIMATING SEDIMENT LOADS

5.1 ABSTRACT

Sediment is the largest surface water pollutant by volume, and should be considered for the surface water planning and management. Different management scenario evaluations require multiple in-stream sediment forecasts and estimations. Physically-based models are considered effective techniques for sediment estimation; nevertheless they require a large number of parameters and intensive calculations. This study aims to enhance sediment predicting techniques by using efficient fusion modeling techniques for sediment forecasting and future scenario evaluations by watershed managers and stakeholders.

The ability of new modeling and cost-effective approaches were tested to reproduce the results obtained from a physically-based model for estimating sediment loads in a watershed. Adaptive neuro-fuzzy inference system (ANFIS) and Bayesian Regression models were tested to find the best alternative to a calibrated and validated Soil and Water Assessment Tool (SWAT) model to predict sediment loads in the Saginaw River watershed. For both methods, four different method-types were tested, namely *General*, *Temporal*, *Spatial* and *Spatiotemporal*.

Results of the study showed that both methods can be used as good alternatives to the SWAT model at the global level for watershed estimations. The best sediment replicating models, the Bayesian *Spatiotemporal* and ANFIS *Spatial*, produced results with Nash-Sutcliffe model efficiency values of 0.95 and 0.94, respectively. For the subbasin level, Bayesian and ANFIS techniques showed satisfactory results for 84 and 77 subbasins, respectively, out of 155 subbasins in the watershed. Box-Cox transformation of sediment load values made the use of the Bayesian model feasible and improved the prediction of the ANFIS models. However, sediment

data exhibited a bimodal distribution after transformation, making the modeling process challenging and complex. Finally, the results of this study confirmed the importance of neighboring conditions for sediment load estimations; spatial relationships are formed when sediments from upstream ultimately make their way downstream.

5.2 INTRODUCTION

Sediment can be defined as particles of sand, clay, silt, and other substances that have been or are being transported by flowing water or wind from the erosion of soil or the decomposition of organisms and point sources such as wastewater discharges (EPA, 2013). Sediment is estimated to be the largest water pollutant by volume, 700 times more prevalent than sewage (Nova, 1988). The amount of sediment being transported by a river is important to the design and management of water resource mitigation projects (Kisi, 2008, Cobaner et al., 2009). The significant direct and indirect effects of sediments can be seen with respect to attached pollutants, channel navigability, reservoir filling, hydroelectric-equipment longevity, aquatic habitat, sustainability of biological resources, soil and water conservation planning, and water quality (Sivakumar, 2006; Kisi, 2008, Kisi and Shiri, 2012).

Sediment erosion (physical land degradation) has both on-site and off-site costs. The on-site cost is associated with the decrease of agricultural products due to loss of soil organic carbon, ultimately leading to a global yearly cost of more than $ 2 billion (Braun et al., 2012).. The off-site cost of sediment erosion is the deposition of large amounts of eroded soil in streams, lakes, and other ecosystems. The worldwide annual cost of the siltation of water reservoirs is roughly $18.5 billion (Braun et al., 2012).

Direct measurement of the sediment load in a stream is the most accurate and reliable measurement method. However, this method requires a good sampling technique, which can be very expensive and time consuming (Ozturk et al., 2001; Wang et al., 2008). Due to these limitations, researchers use modeling approaches to assess sediment transport as an acceptable balance between the cost and accuracy of acquiring simulated sediment data (Li et al., 2006).

Sediment loads depend on climatic and basin characteristics that vary both temporally and spatially due to non-linear and complex relationships (Rezapour et al. 2010, Kumar et al. 2012). Furthermore, these complexities make it challenging to express an accurate transport process through a deterministic mathematical equation or physically based equation when trying to model sediments. Additionally, empirical models can be applied only to the cases that they have been developed for and often give unreliable results for other locations (Tayfur et al., 2003; Bhattacharya, 2005). In addition, a physically-based model such as Soil and Water Assessment Tool (SWAT) requires understanding and expression of all related processes within the system. Also, the application of the physically-based computer simulation can be time consuming and often problematic due the use of idealized sedimentation components, which may not accurately represent real-world conditions (Cigizoglu and Kisi, 2006). Another widely used method in estimating the sediment load being transported by a river is through the use of sediment rating curves. A sediment rating curve is a regression model (in the form of power equation) between the sediment and river discharges, which are also limited in its ability to model this nonlinear process (Jain, 2001 and Kisi, 2004). Similarly, other regression methods (e.g. least square regression) have been found to be ineffective approaches for sediment estimations (Bhattacharya and Solomatine, 2005). Overall, it has been documented that there are numerous drawbacks to conventional sediment modeling approaches, which ultimately increases the demand for more powerful nonlinear modeling approaches (Cigizoglu and Kisi, 2006; Rezapour et al., 2010, Kumar et al., 2012).

Soft computing techniques, including fuzzy logic and artificial neural networks, involve approximate functions connecting system state variables (input, internal, and output) to solve non-linear complex problems (Cigizoglu and Kisi, 2006). Bayesian is a soft computing technique that applies probabilistic reasoning in a form of expert knowledge to handle uncertainty (Huang et al., 2010). These methods have been observed to provide reliable results using less time and calculation efforts than conventional methods. Additionally, adequate knowledge of the physics and natural processes are not necessary (Kisi, 2005; Bhattacharya and Solomatine, 2005 and Cigizoglu and Kisi, 2006). Artificial neural networks (ANN) consist of inter-connected nodes which work similar to neurons in the human brain (Jain, 2001). Adaptive neuro-fuzzy inference system (ANFIS) is a hybrid of ANN with fuzzy logic, and addresses some of the disadvantages of ANN models with the advantages of fuzzy logic methods. Meanwhile, both methods have the ability to handle arbitrary, non-linear relationships between the variables of complex systems (Rai and Mathur, 2008, Kisi et al., 2009).

Lately the applications of ANN and fuzzy logic have been becoming more and more common in water resource engineering to model the many complex issues successfully compared to other conventional techniques (Kisi and Shiri, 2012). Jain (2001) found ANN was a powerful modeling technique to model stage-discharge-sediment concentration relations when compared to the conventional sediment rating curve. Tayfur et al. (2003) also obtained reasonable results using slope and rainfall data to model sediment load through ANN and fuzzy logic models. Meanwhile, Kisi (2005) observed more successful results from an ANFIS model when estimating suspended sediment from streamflow and antecedent sediment data compared to estimates from sediment rating curves and multi-linear regression methods. Similarly, Kisi et al. (2006) found that a fuzzy logic model outperformed the sediment rating curve model in estimating sediment concentration from stream flow and sediment data. Others, including Wang et al. (2008), Nayak and Jain (2011), and Kisi and Shiri (2012) also determined regression techniques were outperformed by ANN and/or ANFIS techniques in modeling sediment concentrations. Studies show that the ANFIS techniques can outperform those of just ANN. For example, Kisi et al. (2009) found that their ANFIS model was superior to ANN and sediment rating curves techniques at modeling sediments from monthly flow and sediment data. Similarly, Cobaner et al. (2009) concluded that ANFIS sediment models performed better than ANN and sediment rating curve models using hydro-meteorological data to model the sediment concentration in a stream.

More recently, researchers have been using Bayesian approaches to model complex systems successfully by accommodating uncertainty in parameters and their predictions (Kruschke, 2010). Bayesian techniques, which incorporate expert knowledge, are robust alternatives to common statistical methods (Schmelter et al., 2012). Contrary to statistical inference, the Bayesian approach assumes that the model parameters are random (Hoff, 2009). Mount and Stott (2008) assessed the Bayesian technique as a novel approach for modeling suspended sediment, compared to traditional statistical approaches. Schmelter et al. (2011 and 2012) developed Bayesian models to predict sediment loads using prior knowledge to enhance the probability framework. Leisenring and Moradkhani (2012) found that adding a Bayesian technique to a non-linear regression model significantly improves the prediction ability of the sediment model in a watershed.

All of the aforementioned models and techniques were only used for sediment estimation at one or a few points in the watershed. The objectives of this study are to analyze, compare, and test the performance of new soft computing techniques namely Bayesian and neuro-fuzzy models (ANFIS) at the watershed scale for both global and local estimations of sediment loads. These may be used as alternative techniques to a physically based model namely SWAT. Overall, this study is an effort of enhancing sediment modeling techniques and making them more efficient through the use of both a physically-based model and soft computing approaches to save computational time and efforts during the future simulations by watershed managers and stakeholders.

5.3 MATERIALS AND METHODS

5.3.1 Study Area

This study focused on the Saginaw River watershed (Figure 5.1), which is Michigan's largest six-digit hydrologic unit code (HUC) watershed, with an area of 22,556 km^2. The Saginaw watershed is located in the east central portion of Michigan's Lower Peninsula (approximately 15% of Michigan's total land area) and drains into Lake Huron. This HUC-6 watershed is made up of six HUC-8 watersheds; the Tittabawassee, Pine, Shiawassee, Flint, Cass, and Saginaw. The watershed is one of Michigan's most diverse areas consisting of 45% forest, 38% agricultural and pastures, 11% water and wetlands, and urban in the remaining area. However, several resource concerns exist within the watershed. High amounts of soil erosion, excessive nutrients (phosphorus and nitrogen), and contaminated sediments are considered main problems in the watershed (EPA, 2012).

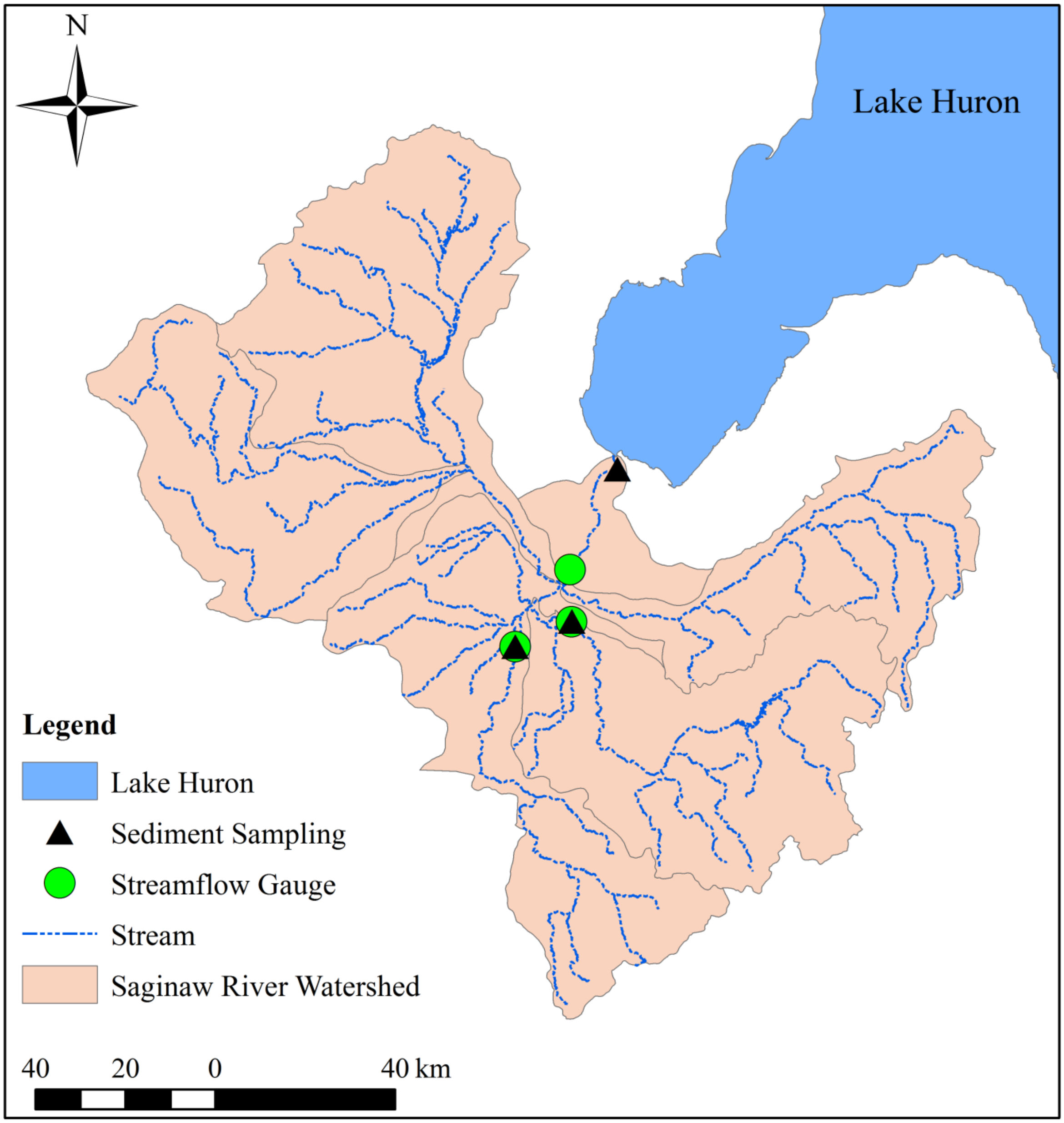

Figure 5.1. Saginaw Watershed

5.3.2 SWAT Model Setup

In order to compare the applicability of ANFIS and Bayesian technique capabilities to estimate sediment load, results from a calibrated SWAT model were used. SWAT is one of the most widely used spatially explicit watershed models (Arnold et al., 1998; Gassman et al., 2007), and was developed by US Department of Agriculture – Agricultural Research Service (USDA-ARS) Temple Texas. This model can be used to simulate flow, erosion, nutrient, and sediment loadings and, ultimately, provide insight for decision makers and watershed managers (Neitsch et al., 2005).

In general, physical-based models such as SWAT are input intensive requiring detailed soil, climate, land use, and topography data. Soil data was downloaded from the State Soil Geographic Database (STATSGO), which contains both physical and chemical properties of soil (Muttiah and Wurbs, 2002). Meanwhile, land use data, including crop-specific classifications, was acquired from the 2008 Cropland Data Layer assembled by the USDA National Agricultural Statistics Service (NASS, 2008). The topographic data was obtained from the Better Assessment Science Integrating point and nonpoint sources (BASINS 4.0) program in the form of a digital elevation model (DEM). Stream network data was obtained using the United States Geological Survey (USGS) National Hydrography Dataset (NHD). Based on the stream network (NHD) and elevation data (DEM), the watershed was divided into 155 subbasins.

As previously mentioned, climate data was also needed to reliably simulate the processes within the watershed. Nineteen years (1990-2008) of observed daily temperature and precipitation measures were obtained from the National Climatic Data Center for 15 temperature and 19 precipitation stations within the watershed. The remaining required meteorological data, such as solar radiation, wind speed, and relative humidity, were acquired using the SWAT weather generator program built into the model.

Additionally, the model was tailored to the local conditions by including specific management operations, crop rotations, and timelines that are commonly used in this watershed. A detailed description of these custom modifications regarding crop rotations and schedules can be found in Love and Nejadhashemi (2011). Hydrologic response units or HRUs, which are basic land units that represent an area with specific dominating land use, soils, and slopes, were also defined. For this study, the HRUs thresholds values were defined as 20, 10, and 20 percentages for land use, soil type, and slope, respectively. It is at the level of these HRUs where the model does several calculations and simulations including sediment loads, nutrient losses, and a hydrologic balance.

5.3.2.1 SWAT Calibration and Validation

Once the SWAT model was built and run with a two-year warm up period, both calibration and validation for sediment and stream flow was performed with a daily time step to fine tune the model and evaluate its performance and predictive capabilities. Calibration was executed through the testing of the parameters that were identified through the sensitivity analysis.

Actual observed data (six years) was used to compare with the model output and perform the calibration process. Flow data was obtained from three US Geological Survey (USGS) gauging stations (04145000, 04149000, and 04157000), and sediment data was downloaded from Michigan Department of Environmental Quality (MDEQ) for stations 730023 (Shiawassee), 730285 (Flint), and 90177 (Saginaw) (Figure 5.1). Flow and sediment stations for Shiawassee and Flint are at the same location. However, the station number for flow is different than sediment as a two data were collected by different agencies. The SWAT model was calibrated for three years (2002 to 2004) and then validated for the next three years (2005 to 2007) to evaluate the model reliability.

5.3.2.2 SWAT Model Outputs

With the model calibrated and validated, it was then run for 19 years (1990-2008), with the first two years being a warm up period for the model. With 17 years of modeled data at each of the 155 sub-watersheds, a total of 2635 data points were obtained for the whole watershed. For each data point, total annual sediment loads and average flow rates were acquired from the simulation. In addition, other modeling variables including basin physical parameters and hydro-meteorological data were gathered from the SWAT simulation for use in the next step to examine whether ANFIS and Bayesian techniques are capable of modeling sediment load both at the watershed and subbasin scales with fewer numbers of parameters. The SWAT output variables had to be checked for skewness, which violates the Bayesian model assumption. Under this condition, variable transformation can be used to eliminate the skewness and other distributional features that complicate the analysis. A Box-Cox power transformation on a variable is a useful method to make a normally distributed set possible when the distribution of the dependent variable is unknown (Kutner et al., 2004).

5.3.3 ANFIS Method

5.3.3.1 ANFIS Modeling Technique

The next step of this study was to use the outputs obtained from the SWAT model to evaluate the ANFIS model performance in estimating sediment load. ANFIS is an effective example of a hybrid method in which ANN and fuzzy logic inference systems are combined. ANFIS uses ANN learning ability to draw fuzzy rules, perform fuzzification, inference, and defuzzification of the system (Thipparat, 2012). The ANFIS system uses an adaptive learning algorithm of either back propagation (gradient descent) or its combination (hybrid) with least square methods to update the system parameters until an acceptable level of error is reached (Jang, 1993 and Cobaner, 2011). Depending on the predefined level of accuracy, the number of iterations can be fixed prior the training (calibration) of the model. Each iteration includes two passes of calculations, forward and backward. The forward pass fixes the antecedent parameters and the consequent parameters are determined as least square estimation parameters. In the backward pass, the consequent parameters are fixed and the errors are propagating backwards to update the antecedent parameters and to lower the error in the prediction process (Rai and Mathur, 2008; Kisi et al., 2009 and Bianconi, 2010). During the iteration process, the system error trend plot is monitored and the iteration is terminated when the error is minimized to prevent overfitting of the calibrated model (Thipparat, 2012).

5.3.3.2 ANFIS Fuzzy Subsets

ANFIS system training starts with rule generation and dividing the input data set into fuzzy subsets using either grid partitioning or subtractive clustering methods, which are defined as *genfis1* and *genfis2* membership functions generating algorithms in MATLAB. The function *genfis1* uses grid partitioning, which generates the rules by enumerating all possible combinations of membership functions and will lead to an exponential increase in number of rules depending on the number of variables and the membership function numbers. The *genfis1* is not useful for input data with many dimensions or number of variables. The second type of function, *genfis2*, uses subtractive clustering to produce scatter partitions, which generate one rule for each cluster of the output data. This makes the second function useful for problems with a high number of variables. However, Cobaner (2011) and Sanikhani and Kisi (2012) recommended not using a large number of independent variables for ANFIS models to avoid extensive calculation time due to the exponential increase in the number of new rules required during the training process. In *genfis2* the value of a cluster radius controls the number of rules (a small radius results in a high number of clusters and an accordingly high number of rules and vice versa) that affect the complexity of the problem. The best values for a radius of influence are usually between 0.2 and 0.5 (MathWorks, 2010)

For this study a MATLAB program code was written that allows the fuzzy logic toolbox to use the sub-clustering method (*genfis2*) for membership functions and rule generation. The two types of data sets, training and testing, were loaded into MATLAB (version 7.12.0). The models were built using tenfold cross validation on 90% of the data (the training portion of the data) and checked on the remaining 10% (the testing portion of the data). The final model for each data set was chosen based on which model had the least amount of error. A Gaussian membership function was used for all input variables and linear parameters were selected for the output membership function (first degree least square function). The employed radius of influence ranged from 0.2 to 0.5 with increments of 0.01, and the model with the best performance was selected. Meanwhile, the number of iterations was limited to a point at which prediction error was minimized.

5.3.3.3 Selecting Best Predictor Variables and ANFIS Inputs

Variable selection started with analysis of the SWAT sediment model output and other related variables that could be used to estimate sediment. For each subbasin, the drainage area, land use, and soil-types for the upstream areas were calculated. Additional variable sets were also considered to account for spatial and temporal viabilities of sediment delivery including time period, geographical coordinates of the subbasin, and precipitation. Within each subbasin, land uses were classified as urban, forest, water, and agriculture; and soils were classified according the Hydrologic Soil Group into A, B, C, and D.

Overall, a set of 13 different variables were examined. As the sediment data was highly skewed, a Box-Cox transformation was performed on the data with estimated optimal λ= 0.12581 to yield highest normality. Although the use of normally distributed data is not required in the ANFIS technique, the transformed sediment data significantly improved model prediction. Cobaner (2011) and Sanikhani and Kisi (2012) stated that the use of numerous input variables for fuzzy logic model setup can result in a large number of rules and adapted parameters in the models, which ultimately increase the noise level in the calculations. Therefore, due to the limitation of the ANFIS technique, not all variables were used. Among the thirteen variables, different sets with seven input variables were selected. In addition, the size of the upstream area and annul precipitation depth were included in each set due to their sensitivity to sediment load. Multiple combinations of the above mentioned variables were considered to develop four different models-types; *General*, *Spatial*, *Temporal*, and *Spatiotemporal*. For the *General* (non-random) model, the set of variables included total upstream area, precipitation, three types of land use, and two soil-types. This technique led to 20 possible variable combinations that were further tested to select the best variable combination to model sediment load in the watershed. The *Spatial* model considered the use of spatial information in combination with the previously used variables to estimate sediments in the watershed. For this model, the set of variables included total upstream area, precipitation, latitude, longitude, two types of land use, and one soil-type. This variable arrangement resulted in 24 possible combinations that were tested during the modeling process to select the best set describing the sediment estimation for the watershed. The *Temporal* model included time as a predictor to estimate sediments. The set of variables included total upstream area, precipitation, time period (year), three types of land use, and one soil-type were used. This resulted in 16 possible variable combinations that were examined to estimate sediments in the watershed. The *Spatiotemporal* model included both spatial and temporal variables along with the other predictor variables. For the *Spatiotemporal* model, the set of variables included upstream total area, precipitation, latitude, longitude, time period, one type of land use, and one soil-type. For this method-type 16 possible sets of variable combinations were tested.

5.3.3.4 Calibrating the ANFIS Models

Ten-fold cross validation was used to select the best set of predictors for each of the four sediment method-types while avoiding overfitting. The ten-fold cross validation process started with randomly partitioning data into ten-folds. Each time, nine-folds of data were used to calibrate (train) a model and the remaining one-fold was used to validate (test) the trained model. Finally for each set of variables, ten different models were generated with the calibration data and tested on the validation data. By employing this technique all the parts of the dataset are eventually used for both calibration and testing (Bouckaert, 2003 and Mahmood and Khan, 2009). Afterward, for each method-type, the set of variable combinations with the best average performance for the test sets were selected. In the next step, the model with the best average performance (the highest coefficient of determination and the lowest root mean square error) on all ten test sets was selected as the best model to be used for sediment load estimations (Hamaamin et al., 2013). Finally the selected best ANFIS models were evaluated both at the global (watershed outlet) and local (155 subbasins) levels.

5.3.4 Bayesian Technique

For the Bayesian technique, the four method-types of *General* (non-random), *Spatial*, *Temporal*, and *Spatiotemporal* were considered. The models variables including precipitation, total upstream area, three land uses (urban, forest, and agricultural), type-A soil, type-B soil, and type-C soil, neighborhood matrix, and year-specific random effects were used.

5.3.4.1 Bayesian Model Specification

For a spatiotemporal regression model (Equation 5- 1), the response variable y_{st} is observed for $s = 1, 2 \ldots N$ geographical region and $t = 1, 2 \ldots T$ time points (years), the model of the observed values can be expressed as following:

$$\hat{y}_{\mathrm{st}} = \sum_{j=1}^{p} X_{stj} \beta_j + u_s + v_t + \varepsilon_{st} \tag{5-1}$$

With *p*-covariates (can be intercept), we assume stationary over both spatial and temporal space and assume residual $\varepsilon_{st} \sim N(0,\delta^2)$, spatial random effect $u = (u_1, u_2,\ldots, u_N)'$ and temporal random effect $v = (v_1, v_2,\ldots, v_T)'$

$$u \sim N(0,\tau^2 D(\gamma)) \tag{5-2}$$

$$v \sim N(0,\sigma^2 A(\phi)) \tag{5-3}$$

Assuming an autoregressive (AR) model with order 1 for temporal dependence with covariance matrix $A(\phi) = [\phi^{n-m}]_{\ 1 \le m,n \le T}$ and a Conditional Autoregressive (CAR) model for spatial dependence with the covariance matrix $D\ (\gamma) = (I\text{-}\gamma\ MG)^{-1}\ M$.

In the expression, *G* is the adjacency matrix with $g_{ij} = 1$ if site *i* and site *j* are neighbors and 0 otherwise, with $g_{ii} \equiv 0$ for $1 \le i, j \le N$. For each site *i*, define $g_{i+} \equiv \sum_{j=1}^{N} g_{ij}$ to be the sum that represents the total number of neighbors of site *i*. *M* is the diagonal matrix with diagonal entries $m_{ii}=1/g_{i+}$.

In this model specification, δ^2 measures the unexplained variation of the nugget effect (unexplained variation) and τ^2 measures the variation due to geographic regions and, γ measures the spatial dependence. A γ that is far from 0 indicates strong spatial dependence. Furthermore, σ^2 measures the variation due to the time with ø measuring the temporal dependence. A ø that is close to 1 (-1) indicates strong positive (negative) correlation between any observation and its next value in time.

By stacking data in the order of time and then geographical regions, we can write the model in a canonical form (equation 5-4) of a mixed-effects model

$$y = X\beta + Z_u u + Z_v v + \varepsilon \qquad (5\text{-}4)$$

Where the design matrix for the random effects is

$Z_u = I_N \otimes 1_T \quad and \quad Z_v = 1_N \otimes I_T$, which are the Kronecker product between identity matrix and column vector with all entries of 1. The Kronecker product is an operation between two matrices that results in a block matrix. The following flat or diffused priors were selected:

$$\pi(\beta_j) \propto 1 \qquad (5\text{-}5)$$

$$\pi(\gamma) = Unirorm(\lambda_N^{-1}, \lambda_1^{-1}) \qquad (5\text{-}6)$$

$$\pi(\phi) = Uniform(-1,1) \qquad (5\text{-}7)$$

$$\pi(\delta^2) = igamma(a_\delta, b_\delta) \qquad (5\text{-}8)$$

$$\pi(\tau^2) = igamma(a_\tau, b_\tau) \qquad (5\text{-}9)$$

$$\pi(\sigma^2) = igamma(a_\sigma, b_\sigma) \qquad (5\text{-}10)$$

Where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$ are ordered eigenvalues of *MG*.

It is required to have $\gamma \in (\lambda_N^{-1}, 1)$ to ensure positive definiteness of *D (γ)*. The upper limit of the interval is 1 since the row sum of *MG* is 1 and $\lambda_N^{-1} < 0$ since the trace of *MG* is zero. We chose the shape parameters $a_\delta = a_\tau = 2$ and scale parameters $b_\delta = b_\tau = 0.01$ for the inverse gamma distribution (*igamma*), which yields a rather dispersed prior density to involve less subjectivity and make the estimators more data-driven.

5.3.4.2 Implementation of Bayesian Models

The parameters under our proposed Bayesian model can be estimated using Gibbs Sampler (Gelfand and Smith, 1990). The full conditional distributions of fixed-effects and random-effects are shown in equations 5-11 through 5-13:

$$\pi(\beta\,|\,...) = N(\mu_{\beta}, \Sigma_{\beta})\begin{cases}\Sigma_{\beta} = \delta^2 (X^{'}X)^{-1} \\ \mu_{\beta} = (X^{'}X)^{-1} X^{'}(y - Z_u u - Z_v v)\end{cases} \tag{5-11}$$

$$\pi(u\,|\,...) = N(\mu_u, \Sigma_u)\begin{cases}\Sigma_u = (\delta^{-2} T I_N + \tau^{-2} D(\gamma)^{-1})^{-1} \\ \mu_u = \Sigma_u Z_u^{'}(y - X\beta - Z_v v)/\delta^2\end{cases} \tag{5-12}$$

$$\pi(v\,|\,...) = N(\mu_v, \Sigma_v)\begin{cases}\Sigma_v = (\delta^{-2} N I_T + \sigma^{-2} A(\phi)^{-1})^{-1} \\ \mu_v = \Sigma_v Z_v^{'}(y - X\beta - Z_u u)/\delta^2\end{cases} \tag{5-13}$$

The conditional distributions of variance components and spatiotemporal dependence are (Gelfand and Smith, 1990):

$$\pi(\delta^2\,|\,...) = igamma\left(a_{\delta} + NT/2, b_{\delta} + \varepsilon^{'}\varepsilon/2\right) \tag{5-14}$$

Where, $\varepsilon = \eta - X\theta - Z_u u - Z_v v$

$$\pi(\tau^2\,|\,...) = igamma\left(a_{\tau} + N/2, b_{\tau} + u^{'}D(\gamma)^{-1}u/2\right) \tag{5-15}$$

$$\pi(\sigma^2\,|\,...) = igamma\left(a_{\sigma} + T/2, b_{\sigma} + v^{'}A(\phi)^{-1}v/2\right) \tag{5-16}$$

$$\pi(\gamma\,|\,...) \propto \left|D(\gamma)\right|^{-1/2} \exp\{\gamma u^{'}Gu/(2\tau^2)\}.I(\gamma \in (\lambda_N^{-1}, \lambda_1^{-1})) \tag{5-17}$$

$$\pi(\phi\,|\,...) \propto \left|A(\phi)\right|^{-1/2} \exp\{-v^{'}A(\phi)^{-1}v/(2\sigma^2)\}.I(\phi \in (-1,1)) \tag{5-18}$$

For the spatiotemporal dependence parameters with full conditional distributions (equations 5-17 and 5-18) that are not well-known densities, the Griddy-Gibbs sampler was implemented (Ritter and Tanner, 1992).

5.3.4.3 Bayesian Models Comparisons

The proposed model can be partially implemented with spatial components only, temporal components only or with both spatiotemporal components masked to evaluate the significance of each component. To compare the Bayesian models, we use Deviance Information Criterion (DIC) for the mixed-effects model. The DIC_4 based on complete likelihood as in equation 5-19 (Celeux et al., 2006).

$$\begin{aligned}\text{DIC}_4 &= -4\text{E}_{\theta,\alpha}[\log f(y,\alpha \mid \theta) \mid y] + 2\text{E}_{\alpha}[\log f(y,\alpha \mid \text{E}_{\theta}[\theta \mid y,\alpha]) \mid y] \\ &\overset{\Delta}{=} -4\text{E}_1 + 2\text{E}_2\end{aligned} \tag{5-19}$$

Where, $\text{E}_{\theta}[\theta \mid y,\alpha]$ can be evaluated by sampling θ for each posterior sample of the random effects α and the obtain mean.

The posterior expected value of the joint deviance was reported as following in equation 5-20 (Celeux et al., 2006):

$$\overline{D(\theta)} = -2\text{E}_1 \tag{5-20}$$

Also model dimensionality was expressed as following in equation 5-21(Celeux et al., 2006):

$$p_{D4} = \overline{D(\theta)} + 2\text{E}_2 \tag{5-21}$$

Therefore, a smaller P_{D4} indicates a simpler model. Generally a smaller DIC_4 indicates better predictive power.

In the full spatiotemporal model, α consists of two components u and v. To assess the model fit, given the L posterior samples of parameters: $(\beta_j^{(l)}, u_s^{(l)}, v_t^{(l)}), where, l = 1,2,...L.$

The fitted value $\hat{y}_{st}$ can be calculated as in equations 5-22 and 5-23 (Celeux et al., 2006):

$$\hat{y}_{st} = \frac{1}{L}\sum_{l=1}^{L}\left(\sum_{j=1}^{p} X_{stj}\beta_j^{(l)} + u_s^{(l)} + v_t^{(l)}\right) \tag{5-22}$$

$$\hat{y}_{st} = \sum_{j=1}^{p} X_{stj}\hat{\beta}_j + \hat{u}_s + \hat{v}_t \tag{5-23}$$

Where, $\hat{\beta}_j, \hat{u}_s$ and $\hat{v}_t$ are posterior mean estimates.

5.3.5 Methods Evaluation Criteria

In this study the predictive accuracy of the ANFIS and Bayesian sediment models was evaluated using five different criteria; the coefficient of determination (R^2), the root mean square of errors (RMSE), the ratio of the root mean square error to the standard deviation of measured data (RSR), Nash-Sutcliffe model efficiency (NSE) coefficient, and the percent bias (PBIAS).

- ***Coefficient of determination*** (equation 5-24) is the square of the Pearson product moment correlation coefficient that shows the degree to which two variables are related, and it ranges between zero and one. A predictive model perfectly explains the variance of the actual data set with R^2 = 1 (Lyman and Longnecker, 2010). According to Arnold et al. (2012), R^2 > 0.5 represents satisfactory model performance.

$$R^2 = \left[\frac{\Sigma_{s=1}^{N}\Sigma_{t=1}^{T}(y_{st}-\bar{y})(\hat{y}_{st}-\hat{\bar{y}})}{\sqrt{\Sigma_{s=1}^{N}\Sigma_{t=1}^{T}(y_{st}-\bar{y})^2(\hat{y}_{st}-\hat{\bar{y}})^2}} \right]^2 \tag{5-24}$$

Where y_{st} and $\hat{y}_{st}$ are observed and predicted values for *s*-th subbasin and *t*-th year, respectively, with *N* represents the total number of subbasins and *T* represents period of the study, $\bar{y}$ and $\hat{\bar{y}}$ are averages of observed and predicted values for *s*-th subbasin and *t*-th year, respectively.

- ***Root mean square of error*** (equation 5-25) shows the goodness of a model with relatively high values of data points.

$$RMSE = \sqrt{\frac{\Sigma_{s=1}^{N}\Sigma_{t=1}^{T}(y_{st}-\hat{y}_{st})^2}{NT}} \tag{5-25}$$

The value of RMSE = 0 describes best fit between predicted and observed values (Lyman and Longnecker, 2010; Nayak and Jain 2011).

- ***Ratio of the root mean square error to the standard deviation of measured data*** (equation 5-26) standardizes the RMSE by using the observations of the standard deviation. A RSR of zero indicates perfect prediction.

$$RSR = \frac{RMSE}{\sigma} = \frac{\sqrt{\Sigma_{s=1}^{N}\Sigma_{t=1}^{T}(y_{st}-\hat{y}_{st})^2}}{\sqrt{\Sigma_{s=1}^{N}\Sigma_{t=1}^{T}(y_{st}-\bar{y})^2}} \tag{5-26}$$

Where σ is the standard deviation of the observed values (Moriasi et al., 2007). According to Moriasi et al. (2007), RSR > 0.7 represents unsatisfactory model performance.

- ***Nash-Sutcliffe model efficiency coefficient*** (equation 5-27) determines the relative magnitude of the residual variance compared to the measured data variance, NSE =1 is the optimal value for best prediction (Moriasi et al., 2007). According to Moriasi et al. (2007), NSE > 0.5 represents satisfactory model performance.

$$NSE = 1 - \frac{\Sigma_{s=1}^{N}\Sigma_{t=1}^{T}(y_{st} - \hat{y}_{st})^2}{\Sigma_{s=1}^{N}\Sigma_{t=1}^{T}(y_{st} - \bar{y})^2} \tag{5-27}$$

- ***Percent bias*** (equation 5-28) measures the average tendency of the predicted data to be larger or smaller than their observed counterparts. The optimal value of PBIAS is zero, positive values indicate underestimation bias, and negative values indicate model overestimation bias. In the case of sediment, PBIAS > ±55 is considered as unsatisfactory (Moriasi et al., 2007).

$$PBIAS = \frac{\Sigma_{s=1}^{N}\Sigma_{t=1}^{T}(y_{st} - \hat{y}_{st})\times 100}{\Sigma_{s=1}^{N}\Sigma_{t=1}^{T}y_{st}} \tag{5-28}$$

5.4 RESULTS AND DISCUSSION

5.4.1 SWAT Model Results

The model was calibrated on daily time steps for both flow and sediment satisfactorily at all three locations within the watershed, based on guidelines provided by Moriasi et al. (2007) using R^2, RMSE, NSE, PBIAS and RSR. The calibration, validation, and overall combined calibration/validation results are presented in Table 5.1.

Table 5.1. Saginaw River Watershed calibration and validation results

Station ID	Constituent	Statistic	Calibration	Validation	Overall
04145000	Flow	R^2	0.81	0.76	0.79
		RMSE	8.21	9.58	8.92
		NSE	0.63	0.55	0.59
		PBIAS	5.38	3.51	4.30
		RSR	0.61	0.67	0.64
730023	Sediment	R^2	0.89	0.80	0.82
		RMSE	6.12	7.42	6.83
		NSE	0.66	0.58	0.61
		PBIAS	32.10	1.66	16.57
		RSR	0.59	0.65	0.62
04149000	Flow	R^2	0.77	0.73	0.75
		RMSE	17.42	17.38	17.40
		NSE	0.57	0.53	0.55
		PBIAS	26.62	-0.96	11.28
		RSR	0.66	0.68	0.67
730285	Sediment	R^2	0.85	0.69	0.78
		RMSE	33.38	43.96	39.28
		NSE	0.68	0.20	0.47
		PBIAS	9.36	-27.98	-7.08
		RSR	0.57	0.90	0.73
04157000	Flow	R^2	0.86	0.83	0.84
		RMSE	75.62	76.55	74.69
		NSE	0.69	0.66	0.67
		PBIAS	28.85	16.62	22.07
		RSR	0.56	0.58	0.57
090177	Sediment	R^2	0.77	0.84	0.80
		RMSE	845.15	1110.56	989.42
		NSE	0.61	0.58	0.59
		PBIAS	3.67	36.42	24.01
		RSR	0.62	0.65	0.64

5.4.2 ANFIS Technique

5.4.2.1 Selecting Best set of Predictor Variables

All data points obtained from the 155 subbasins over 17 years of model simulations (2635 data points) were used for the global models development. The hypothesis is that the best global model has the capability to estimate sediment load at individual subbasin level. For the *General* method-type, 20 sets (combination of seven predictors) were tested to find the best predictor variables. The best predictor variables for the sediment model were total upstream area, precipitations, urban land use, forest land use, agricultural land use, type A-soil, and type B-soil. For the *Spatial* method-type, 24 sets of predictors were tested. The best predictors were total upstream area, precipitations, forest land use, agricultural land use, latitude, longitude, and type B-soil. For the *Temporal* method-type, 16 different sets of predictors were tested. The best variables predicting sediment were total upstream area, precipitations, forest land use, agricultural land use, water land use, time (year) and type B-soil. For the *Spatiotemporal* method-type, 16 sets of predictors were tested to find best predictor set of variables which is consisting of 7 variables. The best predictors were total upstream area, precipitations, agricultural land use, latitude, longitude, time, and type B-soil. The average calibration (90% of data) and validation (remain 10% of data) of 10-fold results for the best set of variables for each method-type are shown in Table 5.2. Overall, 2356 models have been tested during the calibration process of the ANFIS technique,76 different sets of 7 variables have been created for the ANFIS method to find the best predictor variable for each method-type. For each set, 31 models have been tested with different cluster radius of influence starting from 0.2 and expanding to 0.5 with increments of 0.01.

Based on the results of this section, the best ANFIS method-type for estimating the global sediment load is the *Spatial* model with the average validation results of highest R^2 and NSE (0.92) and the lowest RMSE (8140 ton). However, in order to compare the ANFIS and Bayesian models, the best model under the four method-types should be identified first, which is the subject of the next section.

Table 5.2. Average calibration and validation results for ANFIS individual method-types

Method-type	Calibration					Validation				
	R^2	RMSE (ton)	RSR	NSE	PBIAS	R^2	RMSE (ton)	RSR	NSE	PBIAS
General	0.87	10623	0.36	0.87	10.54	0.81	13377	0.45	0.78	9.14
Spatial	0.93	8090	0.27	0.92	3.88	0.93	8140	0.28	0.92	5.05
Temporal	0.83	10032	0.42	0.82	11.53	0.84	12046	0.41	0.83	15.16
Spatio-temporal	0.84	10960	0.40	0.84	9.33	0.82	12776	0.44	0.79	9.86

5.4.2.2 Selecting the Best ANFIS model for Each Method-type

In the previous section, the best combinations of variables for the model development under each method-type were identified. In this section, all ten models developed under each method-type were tested against the 10 sets of validations data points. The model with the highest average NSE and R^2 and the lowest average RMSE was selected as the best model under each method-type (Hamaamin et al., 2013). For all method-types, the average results of all test sets for the best model using the best predicting variables are shown in Table 5.3.

Table 5.3. Statistical average for all test sets for the best ANFIS models under each method-type

Method-type	R^2	RMSE (ton)	RSR	NSE	PBIAS
General	0.89	10104	0.34	0.88	9.16
Spatial	0.94	7459	0.25	0.94	4.00
Temporal	0.85	11585	0.39	0.84	13.98
Spatiotemporal	0.85	10739	0.39	0.84	7.23

As expected from the result of the previous section, the best performance model was a *Spatial* model, while the worst was a *Temporal* model.

5.4.2.3 ANFIS Global Estimations for Each Method-type

The global estimation of sediment load for the entire watershed using the best model for each method-type section is shown in Table 5.4. The *Spatial* model shows the best results for the global estimation of sediment, followed by the *General*, *Spatiotemporal,* and *Temporal*. However, all ANFIS method-types predictions can be categorized as “very good” according to Moriasi et al. (2007) criteria for the model performance.

Table 5.4. ANFIS global estimation for each method-type using the best estimation model

Method-type	R^2	RMSE (ton)	RSR	NSE	PBIAS
General	0.89	9795	0.33	0.89	-1.49
Spatial	0.94	7509	0.25	0.94	4.04
Temporal	0.83	12287	0.42	0.83	12.89
Spatiotemporal	0.84	11907	0.40	0.84	8.10

5.4.3 Bayesian Technique

Similar to the ANFIS technique, in this section sediment load was estimated using the four method-types (*General*, *Spatial*, *Temporal* and *Spatiotemporal*). The *General* model was generated without any of the temporal and spatial random effectiveness’s. For the *Temporal* method-type the spatial random effectiveness was not included, for the *Spatial* method-type the temporal random effectiveness was not included. For the *Spatiotemporal* model both temporal and spatial random effectiveness's were included. The models are implemented by running three Monte Carlo Markov Chains, each with 16,000 iterations. For the full *Spatiotemporal* model, 16,000 iterations were done with around 3 minutes of elapsed time. The convergence is well committed for the first 15,000 iterations, and the last 1000 samples for each chain are used as posterior samples. We also fit the *Spatial*-only version for the model in equation 5-23 by excluding temporal random effect *v*, *Temporal*-only version by excluding spatial random effect *u*, and *General* (non-random) effects model by ruling out both *u* and *v* from the model, which reduces the model to ordinary regressions. The parameter estimations and model assessment are summarized in Table 5.5, which shows the posterior inference for sediment load (tons), with posterior mean (95% highest probability density set) from the total 3000 posterior samples. The spatial component has two parameters: τ^2 the variability and γ the dependence. The temporal component has two parameters: σ^2 the variability and ϕ the dependence.

According to the DIC_4 values (Table 5.5) the best model for the sediment load estimation is the *Spatiotemporal* model. The *Spatial* model comes as the second best model and then the *Temporal* and finally the *General* model with high DIC_4 values indicating it as the worst model for the estimations.

In the next step, each method-type has been used separately to estimate sediment for the whole watershed considering for each method-type the predefined random, spatial, and temporal effects. Table 5.6 shows the sediment load estimation results for the different Bayesian method-types for the entire watershed using back-transformed data. As it shown in Table 5.6, the best

Bayesian model to estimate sediment is the *Spatiotemporal* model (R^2 and NSE = 0.95) followed by the *Spatial* model (R^2 and NSE = 0.95). Both models' performances are categorized as "very good" (Moriasi et al., 2007). However, the *General* and *Temporal* models have fallen into the unsatisfactory performance rating.

Table 5.5. The parameter estimations and model assessment for Bayesian technique

Parameter	*General*	*Spatial*	*Temporal*	*Spatiotemporal*
Intercept	12.68(12.48,12.88)	10.38(8.45,10.18) ∪ (10.73,11.91)	12.68(12.46,12.88)	12.32(10.64,11.66) ∪ (14.35,14.98)
Area (km^2)	3.83(3.62,4.04)	4.32(3.37,5.10)	3.83(3.62,4.03)	4.15(3.64,4.58)
Urban	0.04(-0.25,0.31)	-0.55(-1.25,0.07)	0.04(-0.23,0.30)	0.34(-0.28,0.95)
Forest	0.37(-0.01,0.74)	0.60(-0.94,2.00)	0.36(-0.03,0.76)	0.61(-0.45,1.83)
Agriculture	1.90(1.47,2.34)	2.43(0.54,1.45)	1.90(1.45,2.37)	2.53(0.68,0.78)
A-soil	-5.92(-7.41,-4.36)	-3.44(-7.81,0.94)	-5.93(-7.43,-4.43)	-0.16(-4.37,5.09)
B-soil	-2.63(-4.17,-0.99)	-0.90(-5.67,4.11)	-2.64(-4.23,-1.04)	2.92(-1.37,8.26)
C-soil	-2.40(-3.56,-1.24)	-0.92(-4.31,3.06)	-2.41(-3.58,-1.26)	1.67(-1.76,5.74)
Precip.(mm)	0.59(0.39,0.79)	0.67(0.63,0.71)	0.59(0.38,0.79)	0.58(0.52,0.64)
Residual δ^2	27.7(26.23,29.24)	1.15(1.09,1.21)	27.65(26.2,29.1)	0.91(0.86,0.96)
Spatial τ^2	0.00	98.93(77.8,122.3)	0.00	99.8(76.97,123.35)
Temporal σ^2	0.00	0.00	0.01(0.00,0.4)	0.24(0.09,0.42)
Spatial γ	0.00	0.93(0.83,1.00)	0.00	0.90(0.80,0.99)
Temporal ø	0.00	0.00	-0.11(-1.07,0.87)	-0.02(-0.42,0.40)
$D(\theta)$	16127	8721	16076	8138
pD_4	9.81	11.27	10.59	13.44
DIC_4	16137	8733	16087	**8151**

Table 5.6. Bayesian global estimations for each method-type using best estimation model

Method-type	R^2	RMSE (ton)	RSR	NSE	PBIAS
General	0.52	21913	0.74	0.45	28.36
Spatial	0.93	8025	0.27	0.93	5.27
Temporal	0.53	21709	0.73	0.46	28.28
Spatiotemporal	0.95	6517	0.22	0.95	4.62

5.4.4 Global Application of ANFIS and Bayesian Best Models at Watershed Scale

As shown in Table 4, all method-types developed based on the ANFIS technique performed satisfactory at the global scale, while two method-types developed based on the Bayesian technique are satisfactory (Table 5.6). However, the overall best prediction between all method-types is the Bayesian *Spatiotemporal* method-type, with the highest value of NSE = 0.95. Meanwhile, the best ANFIS model has a NSE = 0.94 for the *Spatial* method-type. Despite the fact that the NSE values between these two techniques are very close, the data are more scattered around the 45 degree line in the ANFIS model compare to the Bayesian model (Figures 5.2 and 5.3). The ANFIS model slightly over-predicted in 50.1% of the data points, while the Bayesian model over-predicted in 51% of data points. The range of error for the ANFIS model was from 46,490 ton/year to 50,568 ton/year with median and average of errors in prediction of 0.02 ton/year and 612 ton/year, respectively. The range of error for the Bayesian model was from 30,390 ton to 63,232 ton with median and average of errors in prediction of 0.13 ton and 699 ton, respectively.

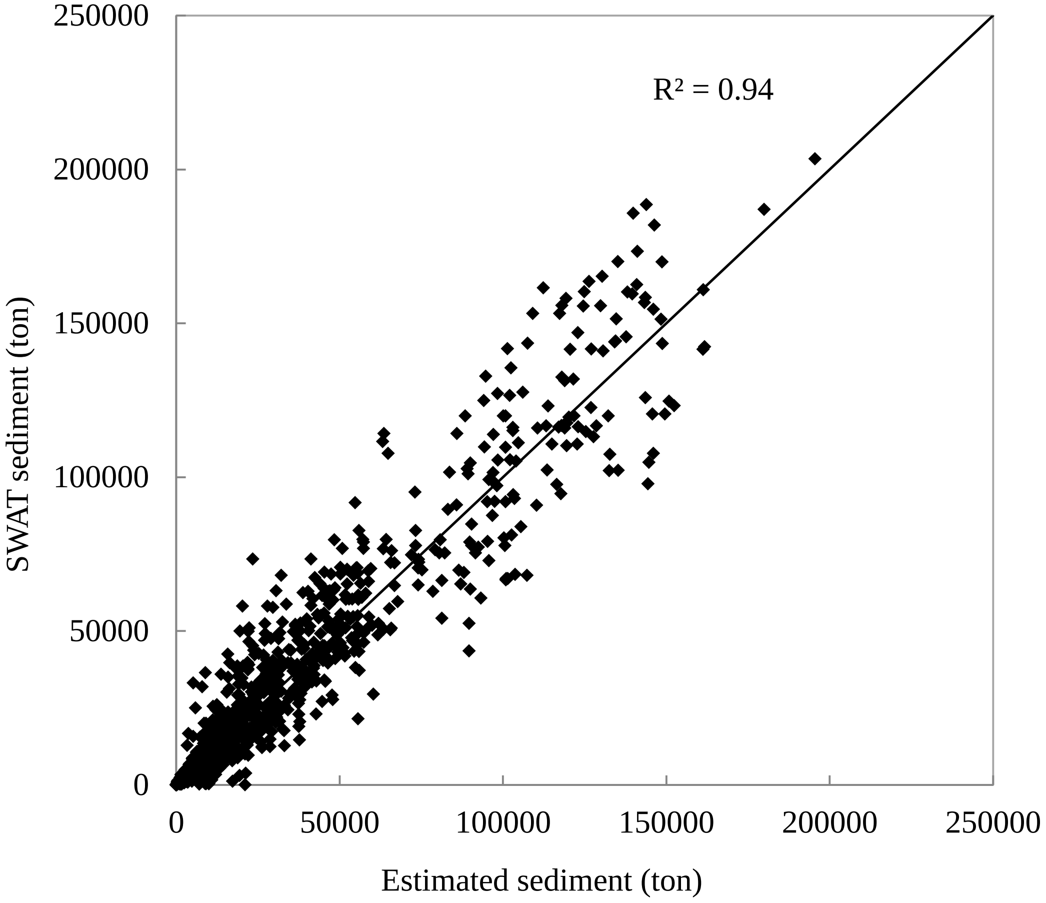

Figure 5.2. ANFIS *Spatial* model sediment estimations

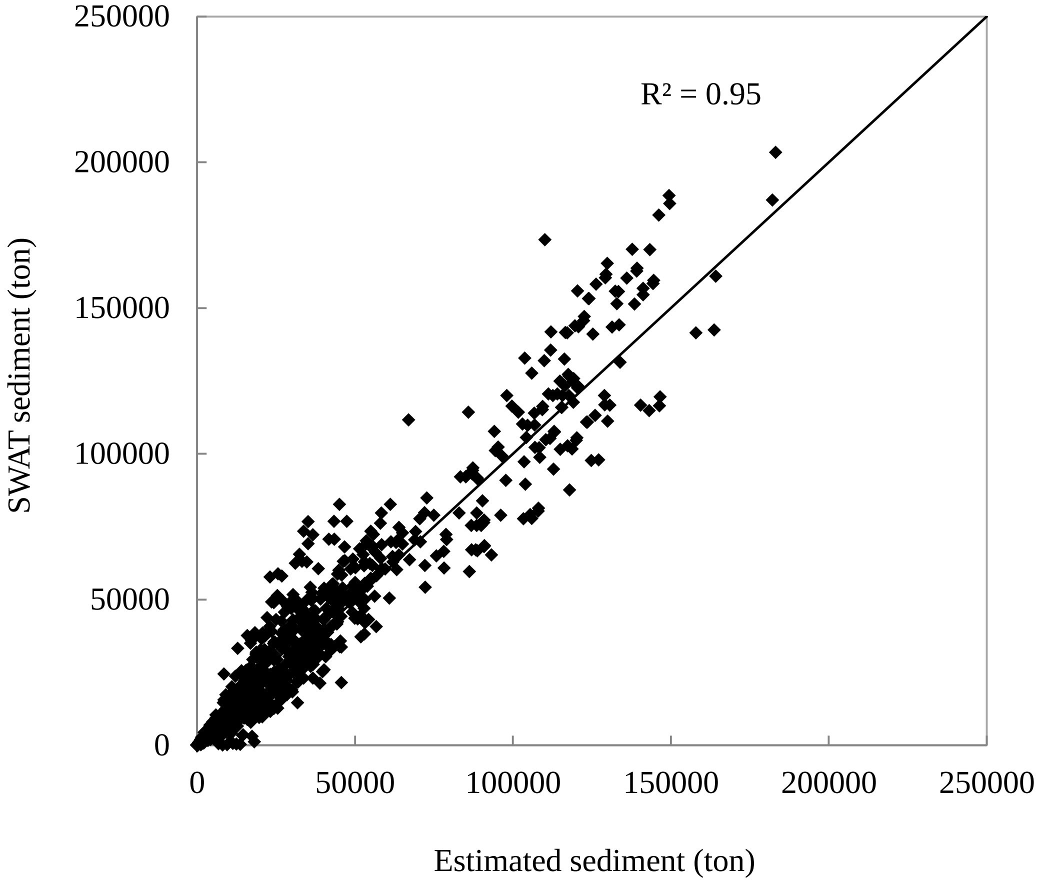

Figure 5.3. Bayesian *Spatiotemporal* model sediment estimations

5.4.5 Local Application of ANFIS and Bayesian Best Models at Subbasin Scale

In this section, the best models of both ANFIS and Bayesian methods, within each method-type, were used to estimate sediment load at each subbasin to further test their accuracy at a finer resolution. Each model was applied locally on the 155 subbasins with 17 data points (years) for each subbasin to test their performance at the subbasins level. Table 7 shows the number of subbasins with satisfactory results (NSE ≥ 0.5) for each method-type according to Moriasi et al. (2007) guidelines. The Bayesian technique showed the highest number of subbasins with satisfactory results. The Bayesian best method-type (*Spatiotemporal*) provided satisfactory results for 84 subbasins out of the total 155 subbasins, while the ANFIS best method-type (*Spatial*) was successful for only 77 subbasins. Figures 5.4 and 5.5 show the subbasins with satisfactory results for ANFIS and Bayesian techniques, respectively. There are 42 subbasins common between the two methods with satisfactory results of NSE ≥ 0.5. The ANFIS technique has satisfactory results for 35 subbasins, which are not captured by the Bayesian technique. On the other hand Bayesian technique has 42 subbasins which are not captured by the ANFIS technique. Figure 5.6 shows the subbasins in Saginaw watershed with combined satisfactory results from both methods. Additionally, there are 35 subbasins out of the total 155 subbasins which are not captured by any of the two models.

Table 5.7. Performances of all method-types for both ANFIS and Bayesian techniques on subbasins

Method-type	Technique used	No. of subbasins with Nash ≥0.5
General	ANFIS Bayesian	38 4
Spatial	ANFIS Bayesian	77 48
Temporal	ANFIS Bayesian	10 6
Spatiotemporal	ANFIS Bayesian	15 84

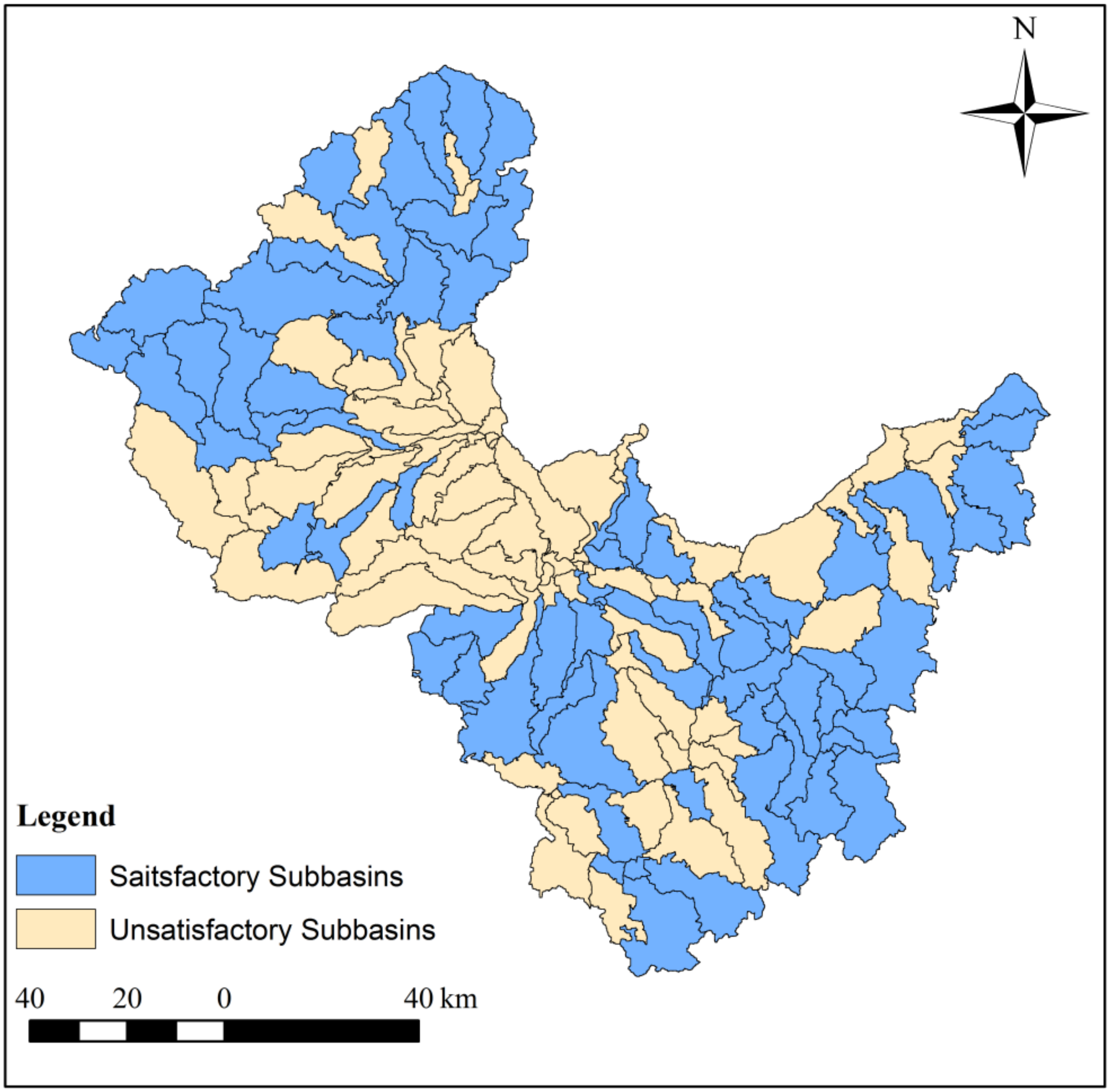

Figure 5.4. Subbasins with satisfactory results from ANFIS technique

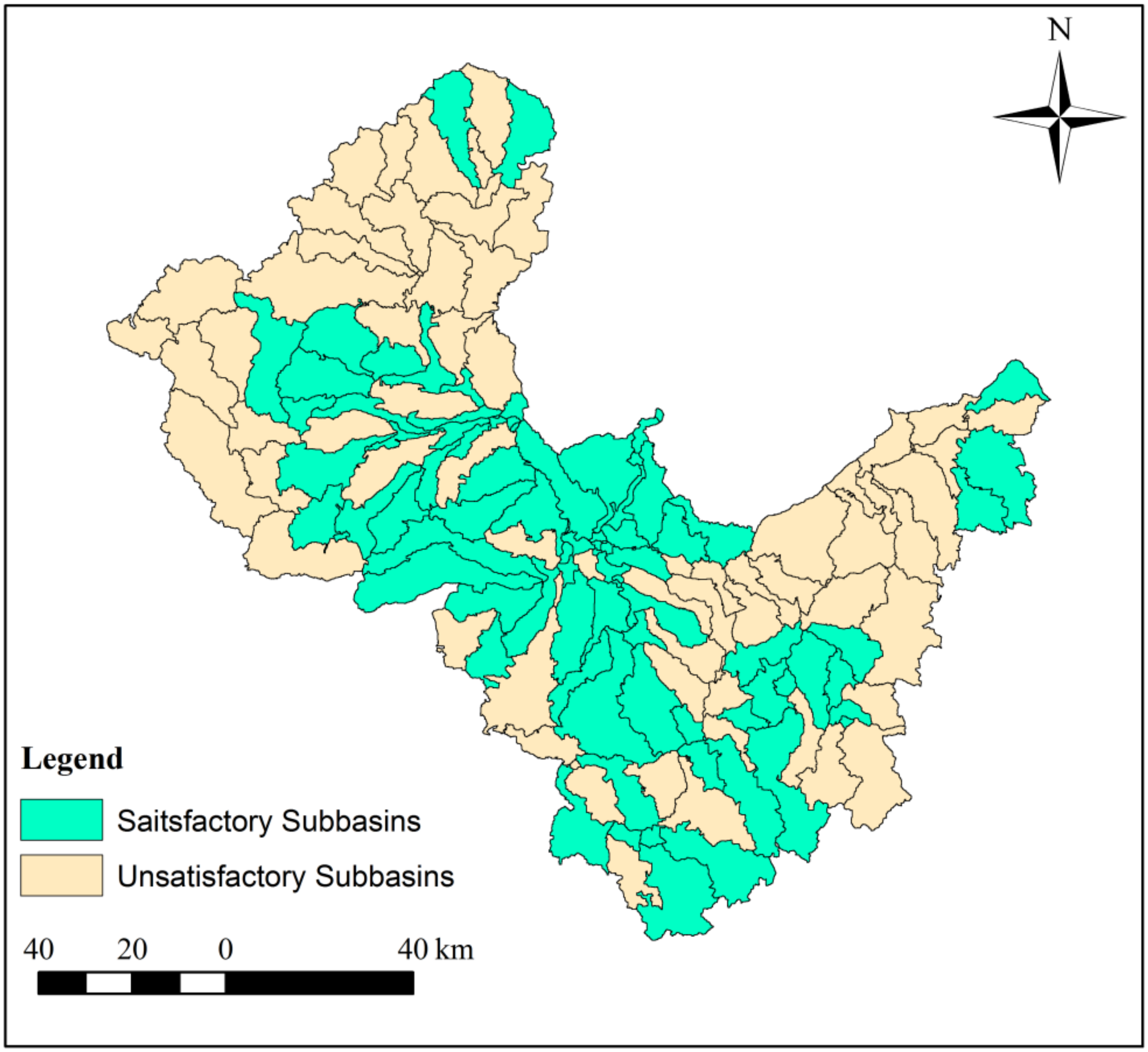

Figure 5.5. Subbasins with satisfactory results from Bayesian technique

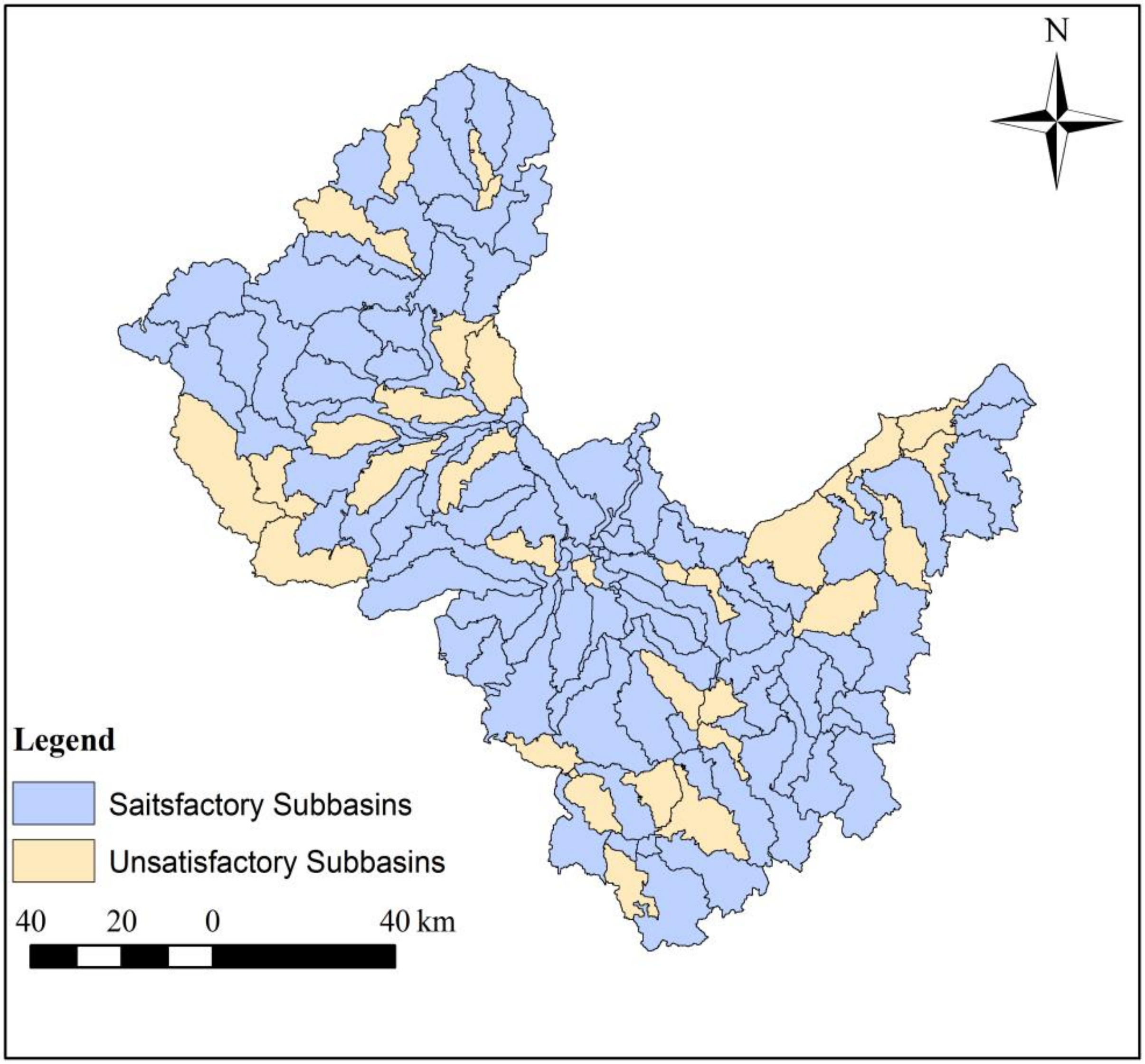

Figure 5.6. Subbasins with satisfactory results from either Bayesian or ANFIS techniques

In general, as it is shown in Figures 5.4 and 5.5, upstream subbasins are mostly covered by ANFIS technique while Bayesian technique covered the downstream subbasins. A series of t-tests were performed to assess whether there was a relationship between model performance and subbasin characteristics. The results from Table 5.8 suggest that for the subbasins which the ANFIS model performed satisfactory annual sediment loads (Figure 5.8), flow rates (Figure A9), drainage sizes (Figure A8), and agricultural areas (Figure A1) are significantly lower than the rest of the watersheds but higher in forest areas (Figure A2). On the other hand, the subbasins for which the Bayesian model produced satisfactory (Table 5.9) annual sediment loads (Figure 5.8), flow rates (Figure A9), drainage areas (Figure A8), agricultural areas (Figure A1), and proportion of type B-soil (Figure A5) are significantly higher but portion of the drainage area in forest (Figure A2) and type soil-A (Figure A4). As discussed earlier, those subbasins were mostly located in downstream regions of the watershed (Figure 5.5), which are expected to have higher sediment loads, flow rates, and drainage areas. Annual sediment loads, flow rates, and upstream drainage areas in the subbasins with satisfactory results from the Bayesian model were about 4 times larger than the respective parameters in the subbasins with unsatisfactory results. Results from Table 5.10 show that the subbasins where neither the ANFIS nor the Bayesian model produced satisfactory results had significantly lower sediment loads (Figure 5.8), lower flow rates, and lower upstream areas (Figure 5.8). Those subbasins were mostly scattered throughout of the watershed (Figure 5.6).

Table 5.8. Evaluation of subbasin characteristics versus ANFIS model performance using the t-test

Input Parameters	Mean values		*p*-value	Percent Difference
	Satisfactory	Unsatisfactory		
Annual sediment load (ton)	12560	17710	0.00	-29.08
Flow rate (m^3/s)	6.4	13.2	0.00	-51.17
Sub-basin area (km^2)	870.5	1864.9	0.01	-53.32
Annual precipitation (mm)	815	809	0.09	0.74
Urban area	0.03	0.02	0.21	64.30
Forest area	0.58	0.46	0.02	24.98
Water area	0.09	0.11	0.18	-25.33
Agricultural area	0.31	0.41	0.04	-23.94
Soil type A	0.26	0.28	0.36	-5.78
Soil type B	0.56	0.57	0.40	-2.11
Soil type C	0.16	0.15	0.29	12.53
Soil type D	0.02	0.01	0.10	129.30

Table 5.9. Evaluation of subbasin characteristics versus Bayesian model performance using the t-test

Input Parameters	Mean values		*p*-value	Percent Difference
	Satisfactory	Unsatisfactory		
Annual sediment load (ton)	22950	5757	0.00	298.64
Flow rate (m^3/s)	15	4	0.00	270.40
Sub-basin area (km^2)	2108.5	478.7	0.00	340.50
Annual precipitation (mm)	810	814	0.31	-0.57
Urban area	0.03	0.02	0.38	85.01
Forest area	0.41	0.64	0.00	-35.59
Water area	0.08	0.13	0.16	-37.10
Agricultural area	0.48	0.22	0.00	121.28
Soil type A	0.20	0.35	0.00	-43.57
Soil type B	0.63	0.49	0.01	26.33
Soil type C	0.17	0.14	0.38	21.20
Soil type D	0.01	0.02	0.40	-43.16

Table 5.10. Evaluation of ANFIS or Bayesian models performance versus subbasin characteristics using the t-test

Input Parameters	Mean values		*p*-value	Percent Difference
	Satisfactory	Unsatisfactory		
Annual sediment load (ton)	17544	6947	0.00	152.5
Flow rate (m^3/s)	11.2	5.2	0.00	116.8
Sub-basin area (km^2)	1600.1	577.4	0.00	177.2
Annual precipitation (mm)	813	810	0.58	0.4
Urban area	0.03	0.00	0.01	663.1
Forest area	0.51	0.55	0.60	-6.4
Water area	0.08	0.16	0.13	-46.0
Agricultural area	0.38	0.30	0.24	27.0
Soil type A	0.24	0.34	0.06	-28.5
Soil type B	0.59	0.49	0.10	18.8
Soil type C	0.15	0.15	1.00	0.1
Soil type D	0.01	0.01	0.54	52.0

The inability of both techniques to predict sediment load in all the subbasins may be due to heterogeneity in subbasin properties in which some variables are significantly different at local levels. During calibration, a global model with the lowest error will be selected as the best model and, therefore, it is expected that the model does not perform satisfactory in areas in which physiographical characteristics are unique. In general, the ANFIS model performed well in subbasins that generate lower than average sediment loads while the Bayesian model acts opposite. From another point of view, if the model can capture large values, it may not be able to capture small values and vice versa. The complexity of the erosion and sediment transport problems and spatially and temporally heterogeneity of watershed characteristics creates an unprecedented condition for the artificial intelligence and statistically- based models to compete with the physically based models such as SWAT.

5.4.6 The Gap between Global and Local Model Performances

The gap between the global and local models performance can be explained by investigating the statistical and spatial distributions of the sediment data shown in Figures 5.7 and 5.8. Figure 5.7 shows the relative frequency histogram and density function for the Box-Cox transformed sediment data while Figure 5.8 represents the spatial distribution of sediment load within each subbasin's river. Two different distributions can be realized from the global data distribution, which is bimodality. Any observations from one normal component will appear as outliers for the other component, which results in a large variability in the output data. This is especially true when dealing with large number of observation points, 2635, which makes the denominator for calculating the global NSE equation very large and hence the NSE approaches 1 easily (which is considered as perfect fit). This is however not the case at the local level (subbasins) which has sample size of only17 points. This bimodality for the mean level of sediment load is well captured by the intercept from the optimal Bayesian *Spatiotemporal* model. On the other hand, from the output of both *Spatial* and *Spatiotemporal* models, the variation due to subbasins (τ^2) is greatly larger than the variation due to random errors (δ^2). This is further investigated by comparing Figures 5.2 and 5.3 with Figures 5.9 and 5.10 for both ANFIS and Bayesian techniques, respectively. At the global level (Figures 5.2 and 5.6) the pairs of observed and fitted values are relatively close to the reference line, which indicates consistency, while at the local level (Figures 5.9 and 5.10); the pairs are relatively more spread around the line. Overall, the best ANFIS and Bayesian models are able to capture the higher variability (spatial) that is dominant at the global level (watershed) but are not able to capture the smaller size of variability (temporal) that are dominant at the local level (subbasin).

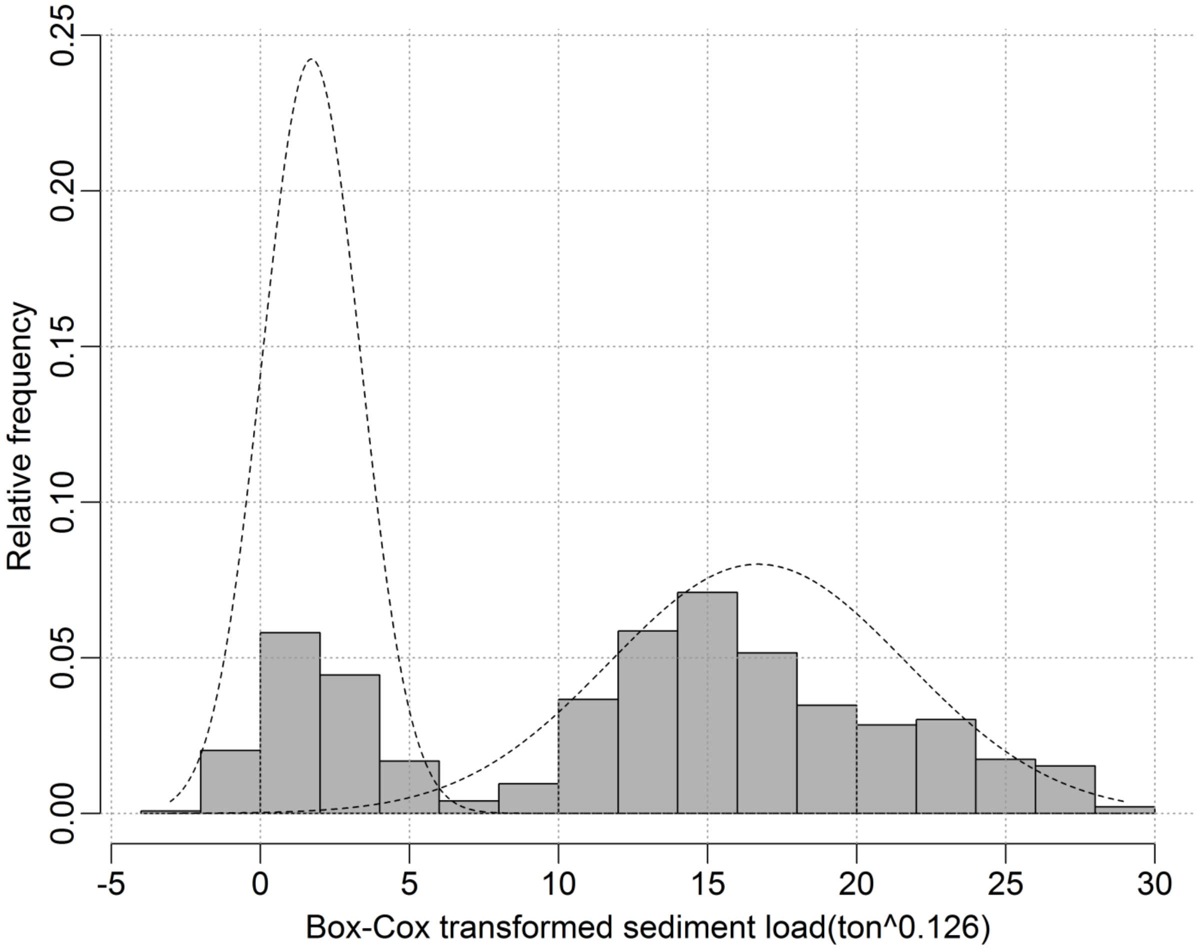

Figure 5.7. Relative frequency histogram and density function for sediment load transformed data (ton$^{0.126}$)

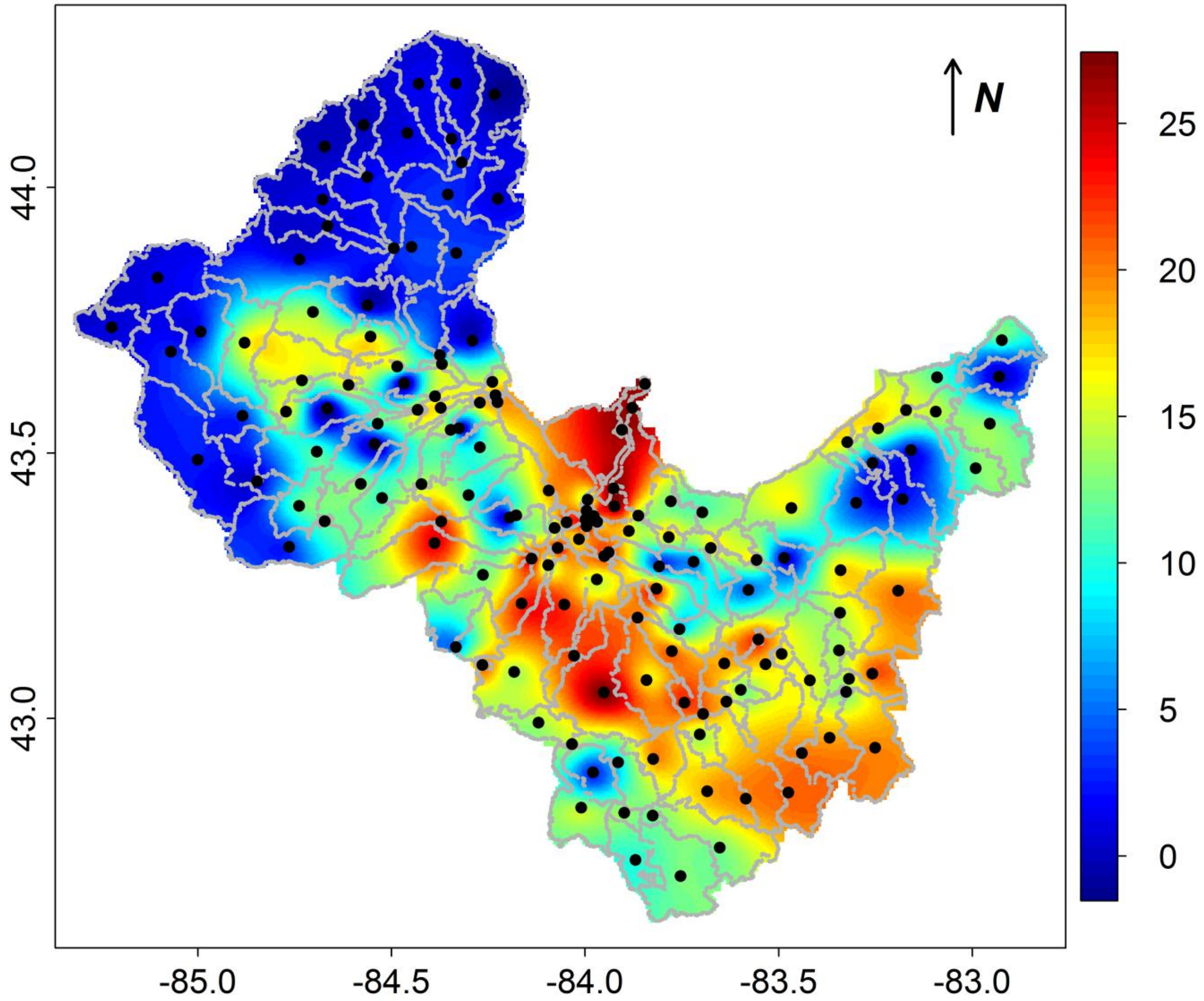

Figure 5.8. Spatial plot of SWAT sediment transformed output variable (ton$^{0.126}$)

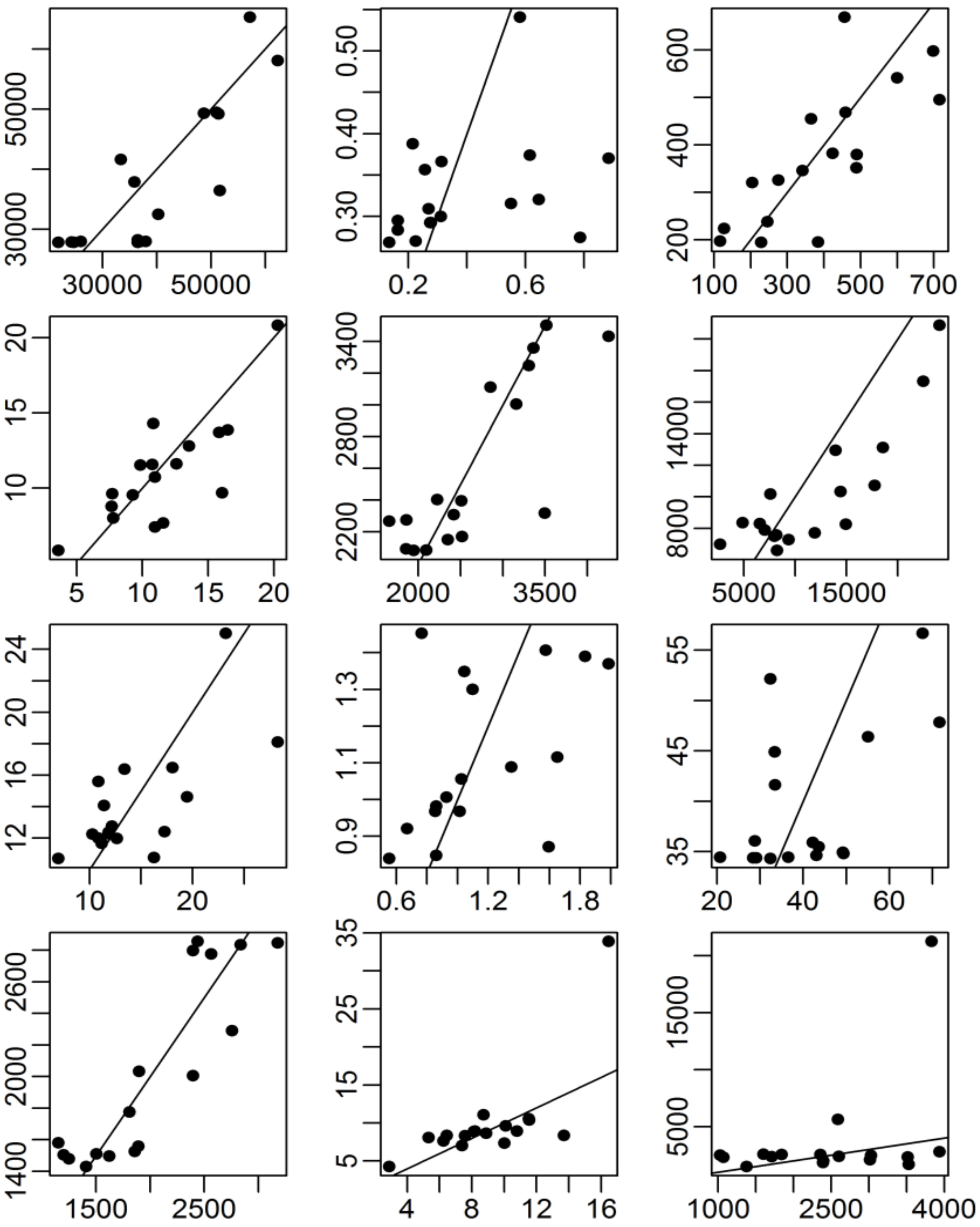

Figure 5.9. ANFIS estimated sediment load versus observed values for randomly selected 12 subbasins for illustration, each with T = 17 samples. The reference line for all panels has intercept 0 and slope 1

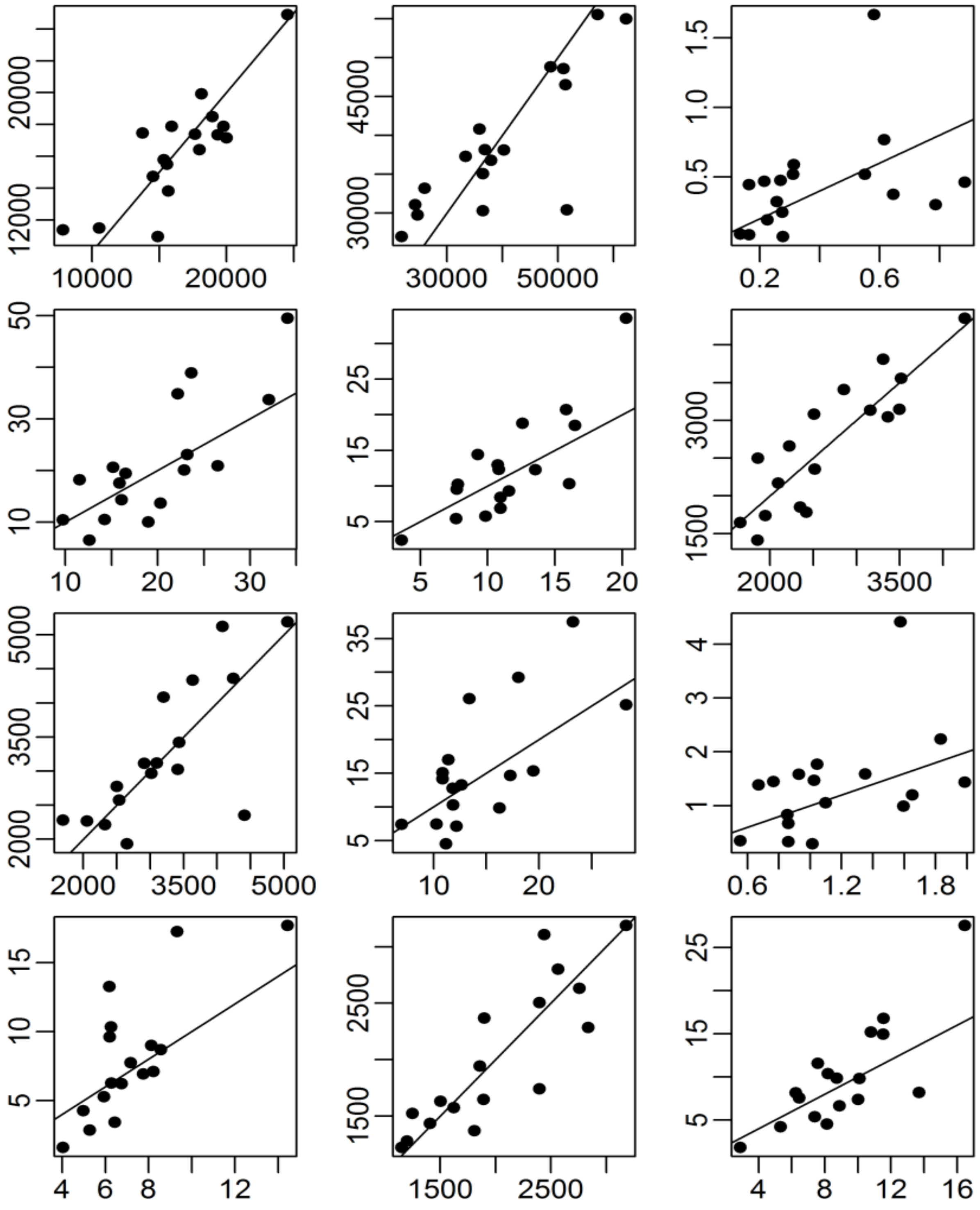

Figure 5.10. Bayesian estimated sediment load versus observed values for randomly selected 12 subbasins for illustration, each with T = 17 samples. The reference line for all panels has intercept 0 and slope 1

5.5 CONCLUSION

Having an effective and inexpensive predicting model in terms of balancing calculation time and model accuracy is essential for effective watershed management and decision making. New soft computing techniques such as ANFIS and Bayesian models can save time, money, and calculation efforts compared to a more complex physically based hydrologic models such as SWAT. The large numbers of parameters needed for the SWAT model are reduced to seven parameters in the ANFIS model and ten parameters in the Bayesian model. This work is an effort to use the advantages of new techniques to build a cost-effective fusion models for future uses as a web application tool or decision support system for non-expert usage; compared to physically-based models, which require more time for model setup and data collection, calculations, and the need for an expert operator.

Results of this work showed that Bayesian and ANFIS models can be used to estimate sediment loads for the Saginaw Watershed successfully for global estimations with less calculation and parameter requirements. All the ANFIS method-types were in the very good range of estimation, while only two method-types of Bayesian technique were in the very good range of estimation. Overall, the best sediment estimation was obtained from the Bayesian *Spatiotemporal* model at both global and local levels. The Bayesian *Spatiotemporal* model, with NSE=0.95, showed better sediment load estimations compared to the ANFIS model with NSE =0.94. At the subbasin level, higher number satisfactory results were obtained from the Bayesian technique (84) compared to the ANFIS technique (77). However, the ANFIS model used fewer parameters, seven compared to the Bayesian model, which used ten.

This research suggested using transformation techniques such as Box-Cox to improve the predictive abilities of both ANFIS and Bayesian techniques even though it is not a requirement for the data preparation phase of the ANFIS model.

Both Bayesian and ANFIS techniques, attested to the importance of spatial data for sediment modeling process by having the best results with models including spatial data in the predictor variables. This confirms the influence of neighboring conditions can improve the models' predictability, which is originated from the fact that sediment, like flow, moves from upstream to downstream leading to a spatial relationship.

Results of the study confirmed that the ANFIS technique has better prediction in the upstream subbasins where the sediment loads are relatively lower than the average, while the Bayesian technique has better prediction in the downstream subbasins where the sediment loads are relatively higher than the average. This performance can, to some extent, be explained by the bimodal nature of sediment loads in this watershed.

Sediment data is affected by both spatial and temporal random variability. If a predictive model can capture higher variability such as spatial variance at large scale, this limits the model's performance at smaller scales, such as temporal variance. This was confirmed by both techniques, which they have the best prediction ability at the watershed level while their prediction ability reduced at subbasins level.

Finally, this work is a trial to combine two different types of models (physical model with soft/statistical model) in a fusion technique to save time and calculation efforts during future watershed management scenario evaluations. Because of the complexity of erosion and sediment transport within a watershed, future works should consider combining more than two techniques, which may result in a more robust modeling system to enhance the ability of sediment load estimations.

6 CONCLUSIONS

For a sustainable water resource planning and management, forecasting and prediction of water quantity and quality parameters is required. In this regard, mathematical and statistical sciences play an important role. The main objectives of this study were to analyze and test the ability of fuzzy logic and Bayesian techniques to find more accurate, robust, and cost-effective models for sustainable water resource managements concerning water quantity and quality. The following can be concluded from the results of this study:

Conclusions for the first paper titled "Application of Fuzzy Logic Techniques in Estimating the Regional Index Flow for Michigan":

- Having the knowledge and the ability of accurately determine water withdrawals within a watershed is essential for decision makers and watershed managers.
- Inaccurate index flow estimation could damages streams ecosystems due to excess water withdrawals.
- This study showed the potential of fuzzy logic models as a cost-effective (using less number of variables) and more accurate alternative modeling technique to the regression model, which is currently used by Michigan's Water Withdrawal Assessment Tool to estimate stream index flows.
- The ANFIS model over-fitted at the expense of having the closest mean and standard deviation of the observed values, resulting in a few outliers.

- Despite the fact that the fuzzy expert model ranked second based on performance criteria (R^2 and MSE) for the square root of index water yield, it provided repeatable prediction that can be generalized for the study area and other regions as well.
- Results showed the advantages of the 10-fold cross validation technique in selecting the best model.

Conclusions for the second paper titled "Evaluation of Neuro-Fuzzy and Bayesian Techniques in Estimating Sediment Loads":

- Soft computing techniques such as ANFIS and Bayesian regression can replicate global model estimation based on a calibrated physically based model (such as SWAT) using significantly less input variables.
- In replicating a physically based model with soft computing techniques, both spatial and temporal variables should be included in the Bayesian regression model, while spatial variables are sufficient for the ANFIS model.
- Despite the fact that normality is not a requirement in fuzzy methods, input data transformation to achieve normality improved ANFIS model prediction.
- The application of global soft computing techniques such as ANFIS and Bayesian regression at the local level was generally not satisfactory. This indicates that the global model was unsuccessful in capturing inherent relationships between watershed characteristics and sediment load at the local scale.

7 RECOMMENDATIONS FOR FUTURE RESEARCH

Based on the outcomes from this study, the following are recommended for future studies:

- We propose to test the predictability of fuzzy expert models for index and low flow in different physiographical regions.
- While water resources are complex systems and have many interdependent parameters, further studies should explore alternative methods, such as boosting algorithms, for emerging issues in water resources such as climate changes and extreme events. Boosting can enhance the learning ability of the inherent system to improve prediction.
- Because of the complexity of sediment generation and transport phenomena within a watershed, future works should consider combining additional methods to improve model prediction accuracy at the local level. Combining more than two techniques is expected to increase prediction power to the estimation fused technique.
- For better sediment prediction, we recommend using more predictor variables in future works including some variables that better describe human activities, such as agricultural and construction practices, which produce time-dependent results and outliers in the sediment load.
- For future sediment predictive models, we recommend performing statistical tests to check the multimodality of the sediment data, and separate data on the basis of different clusters and make unique models for each cluster to capture both high and low sediment loads.

APPENDIX

Additional Materials to Section 5 Titled "Evaluation of Neuro-Fuzzy and Bayesian Techniques in Estimating Sediment Loads"

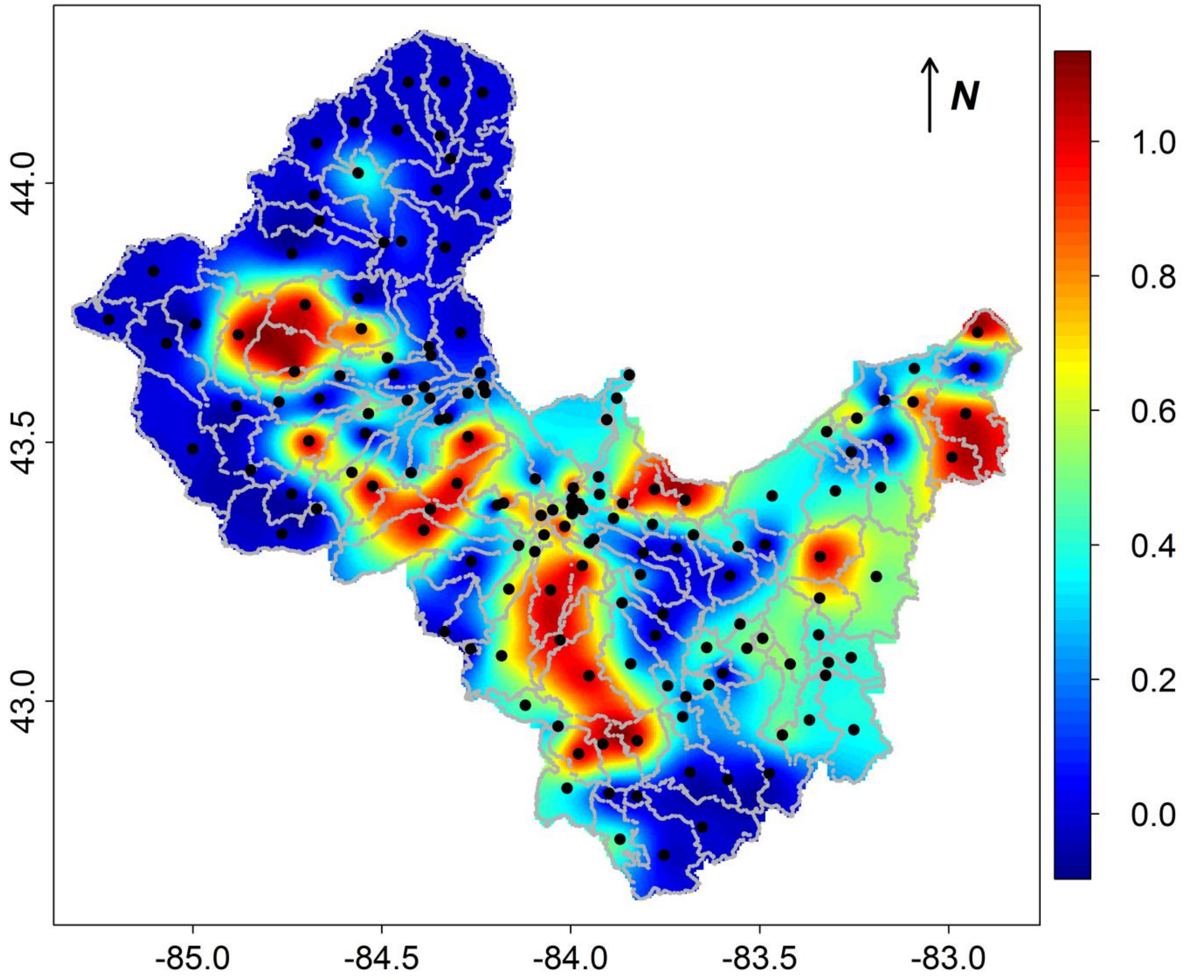

Figure A1. Spatial plot of agricultural area predictor variable (agricultural area shown as a fraction of the subbasins area, 0 represents no agricultural area, and 1 represents the 100% of area is agricultural)

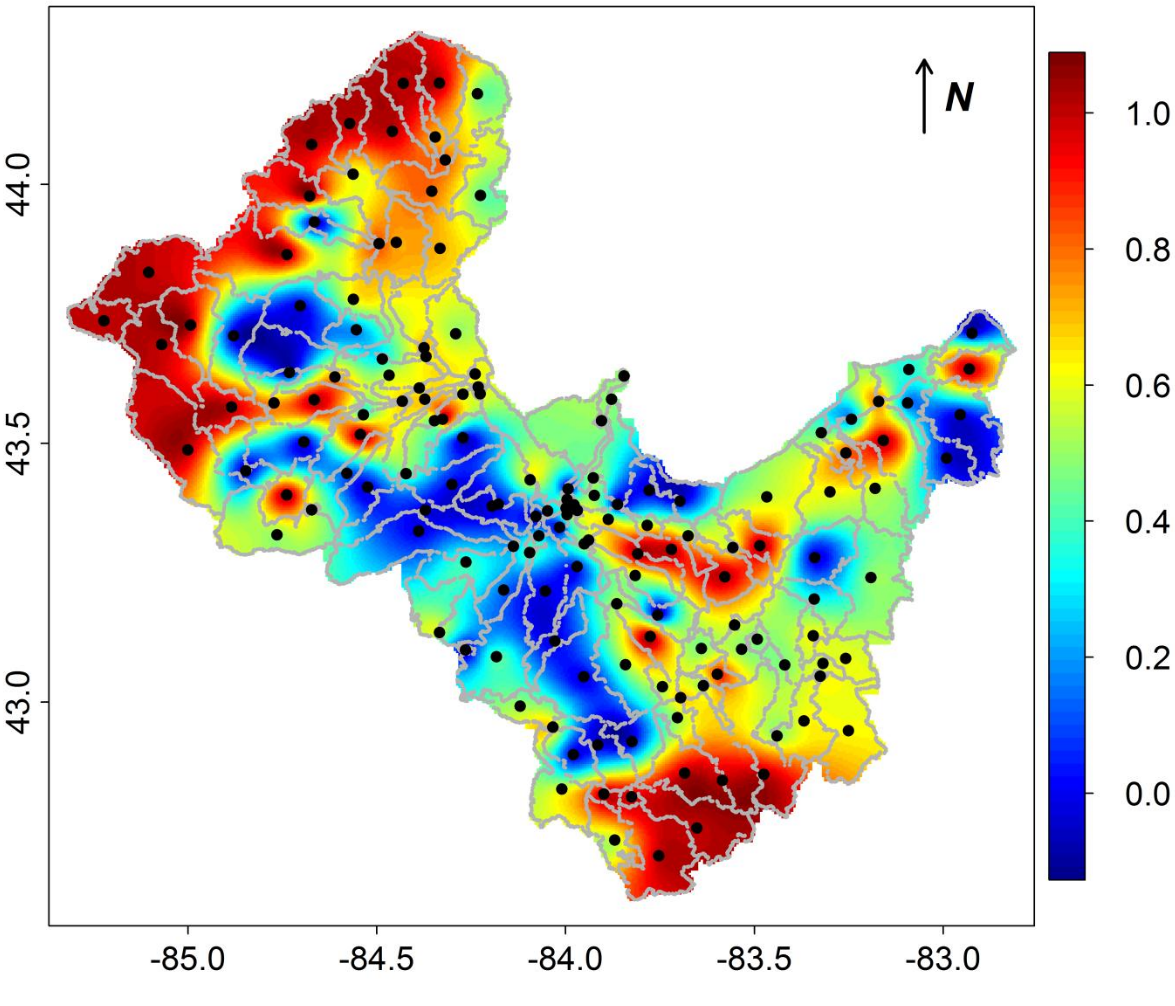

Figure A2. Spatial plot of forest area predictor variable (forest area shown as a fraction of the subbasins area, 0 represents no forest area, and 1 represents the 100% of area is forest)

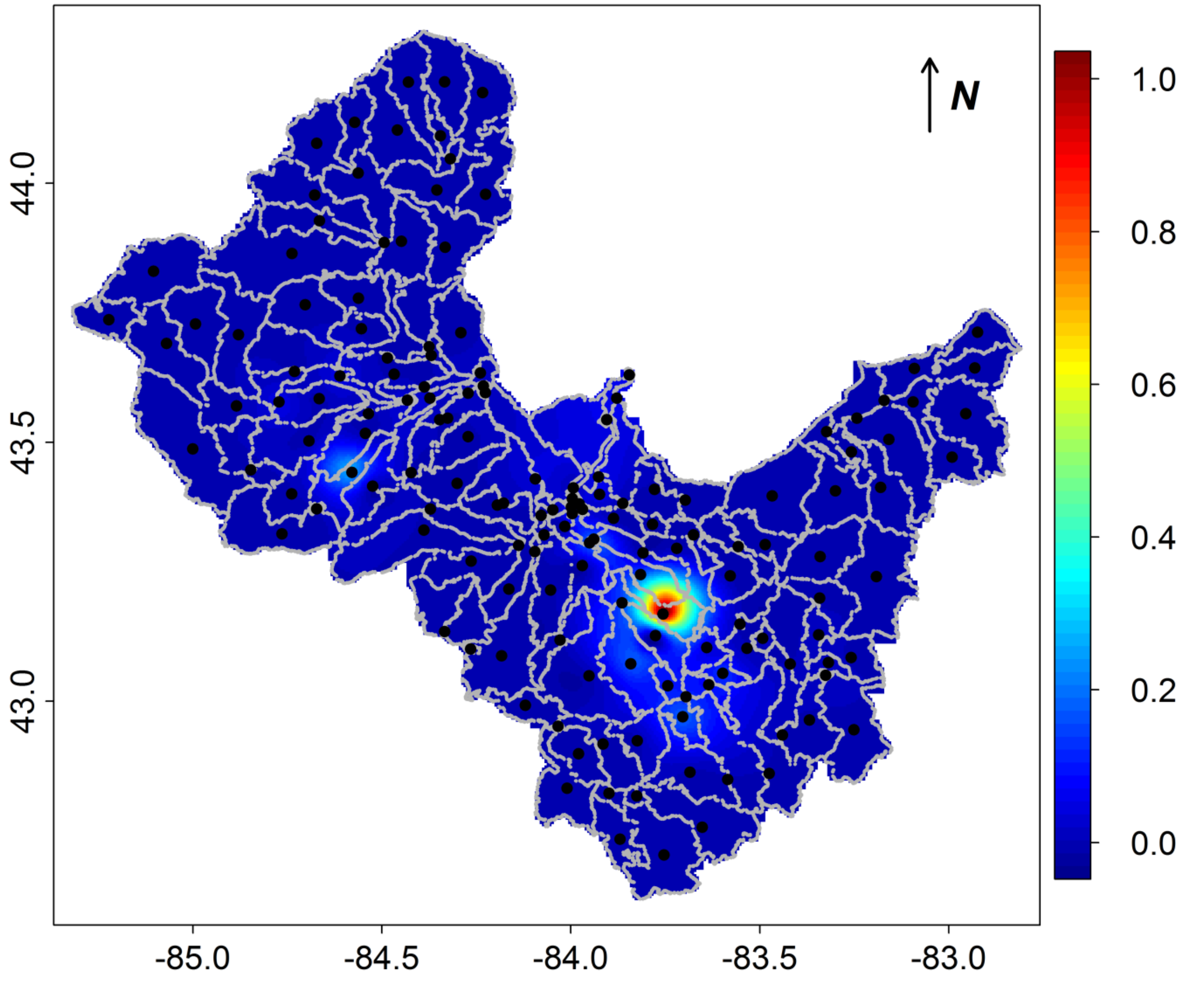

Figure A3. Spatial plot of urban area predictor variable (urban area shown as a fraction of the subbasins area, 0 represents no urban area, and 1 represents the 100% of area is urban)

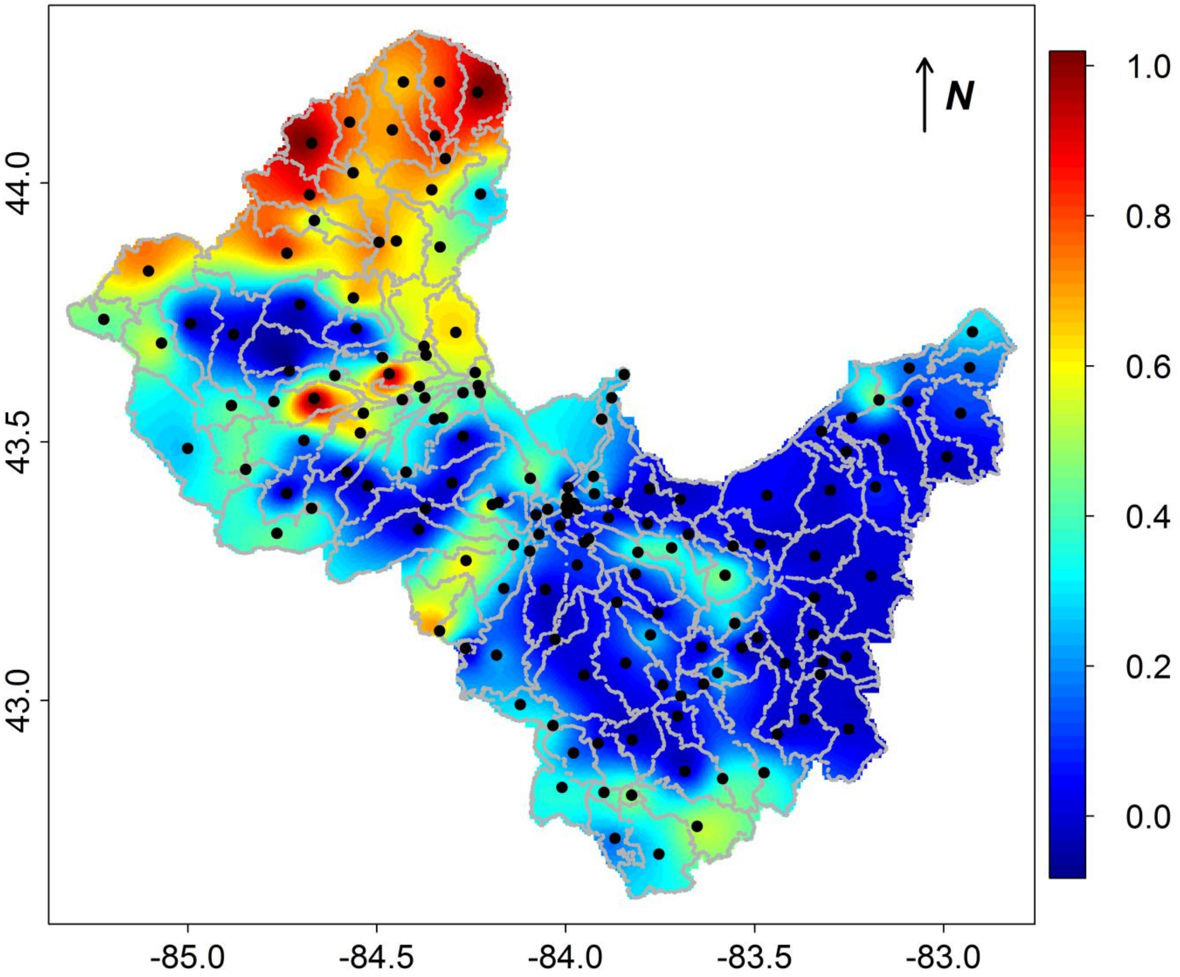

Figure A4. Spatial plot of type A-soil type predictor variable (A-soil area shown as a fraction of the subbasins area, 0 represents no A-soil area, and 1 represents the 100% of area is A-soil)

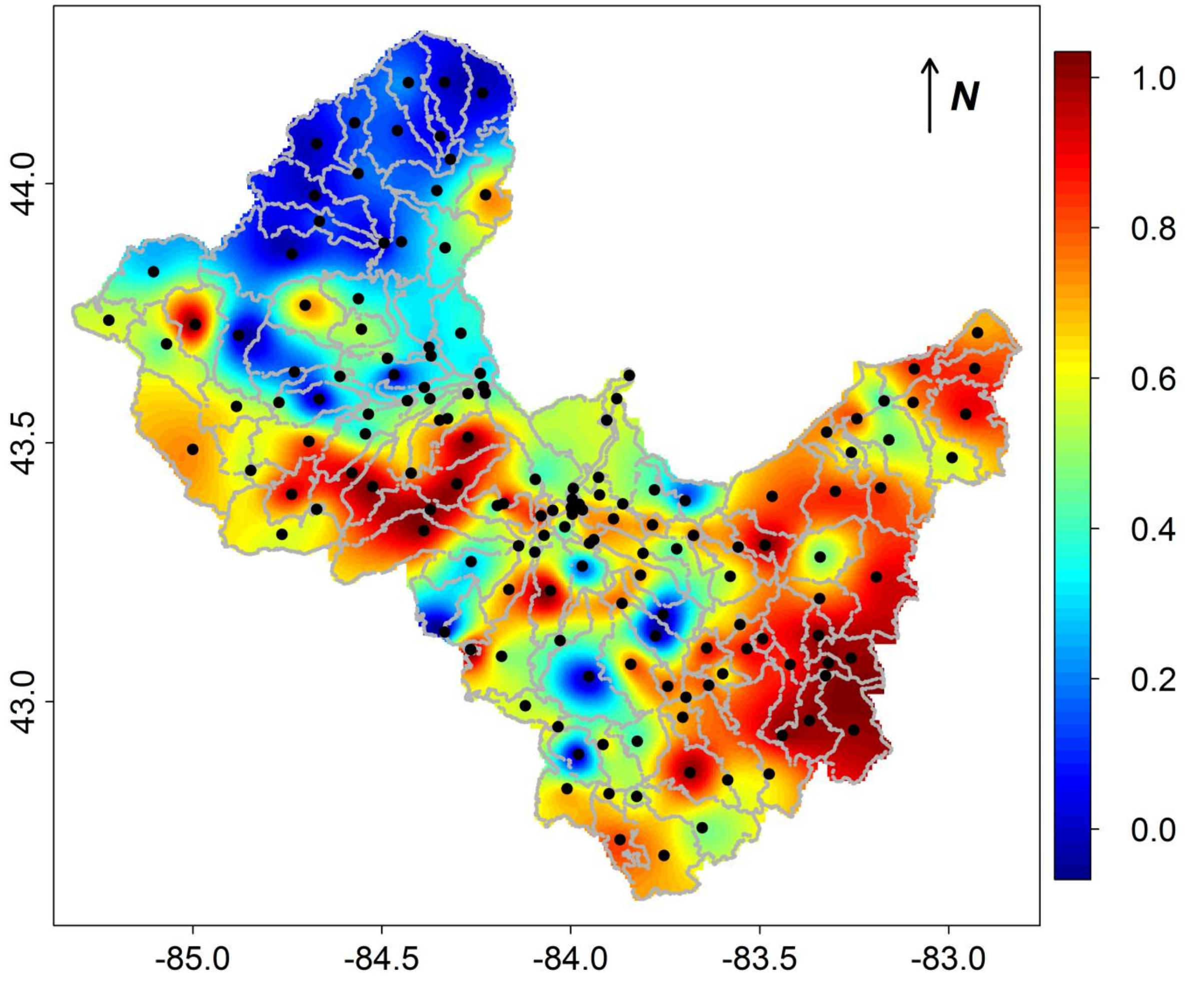

Figure A5. Spatial plot of type B-soil type predictor variable (B-soil area shown as a fraction of the subbasins area, 0 represents no B-soil area, and 1 represents the 100% of area is B-soil)

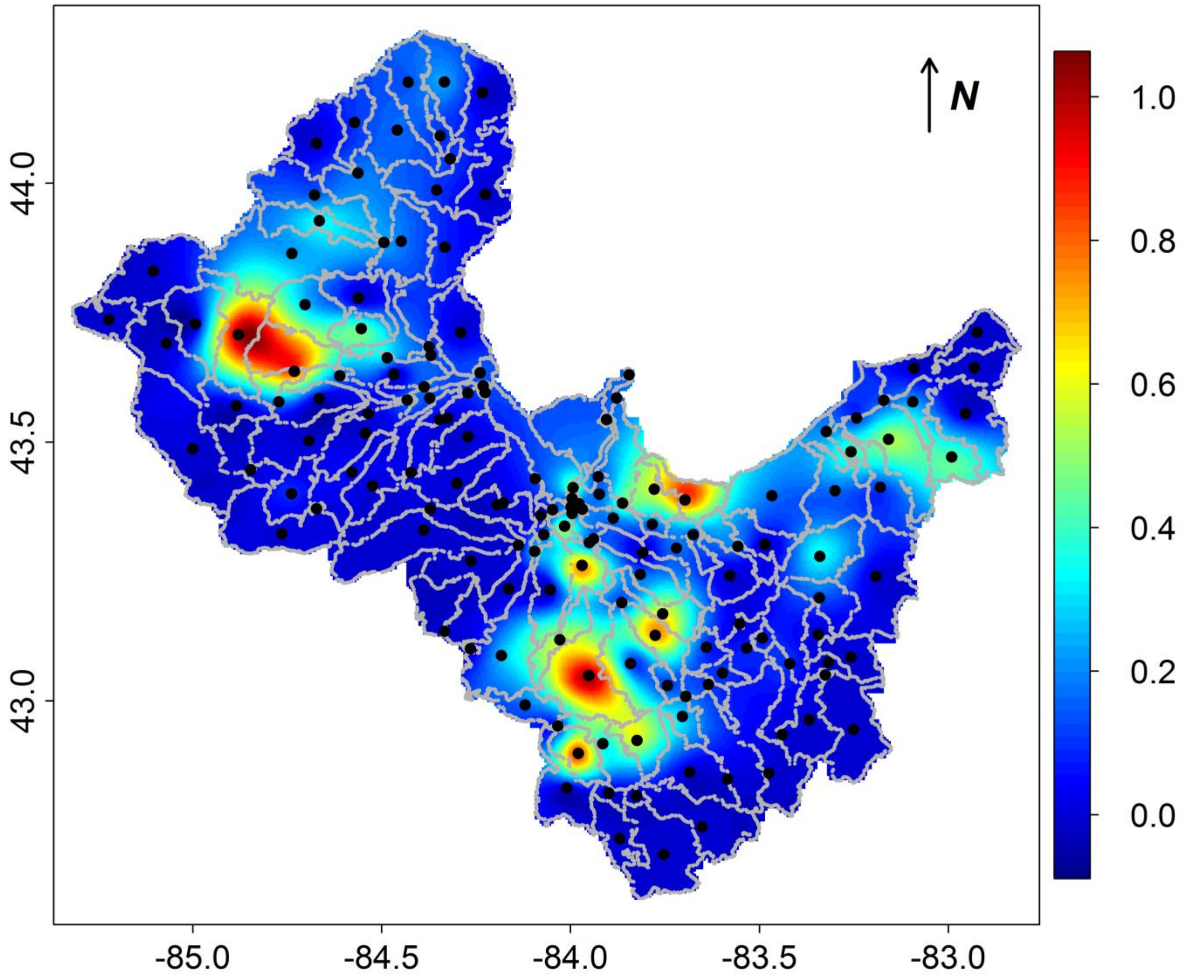

Figure A6. Spatial plot of type C-soil type predictor variable

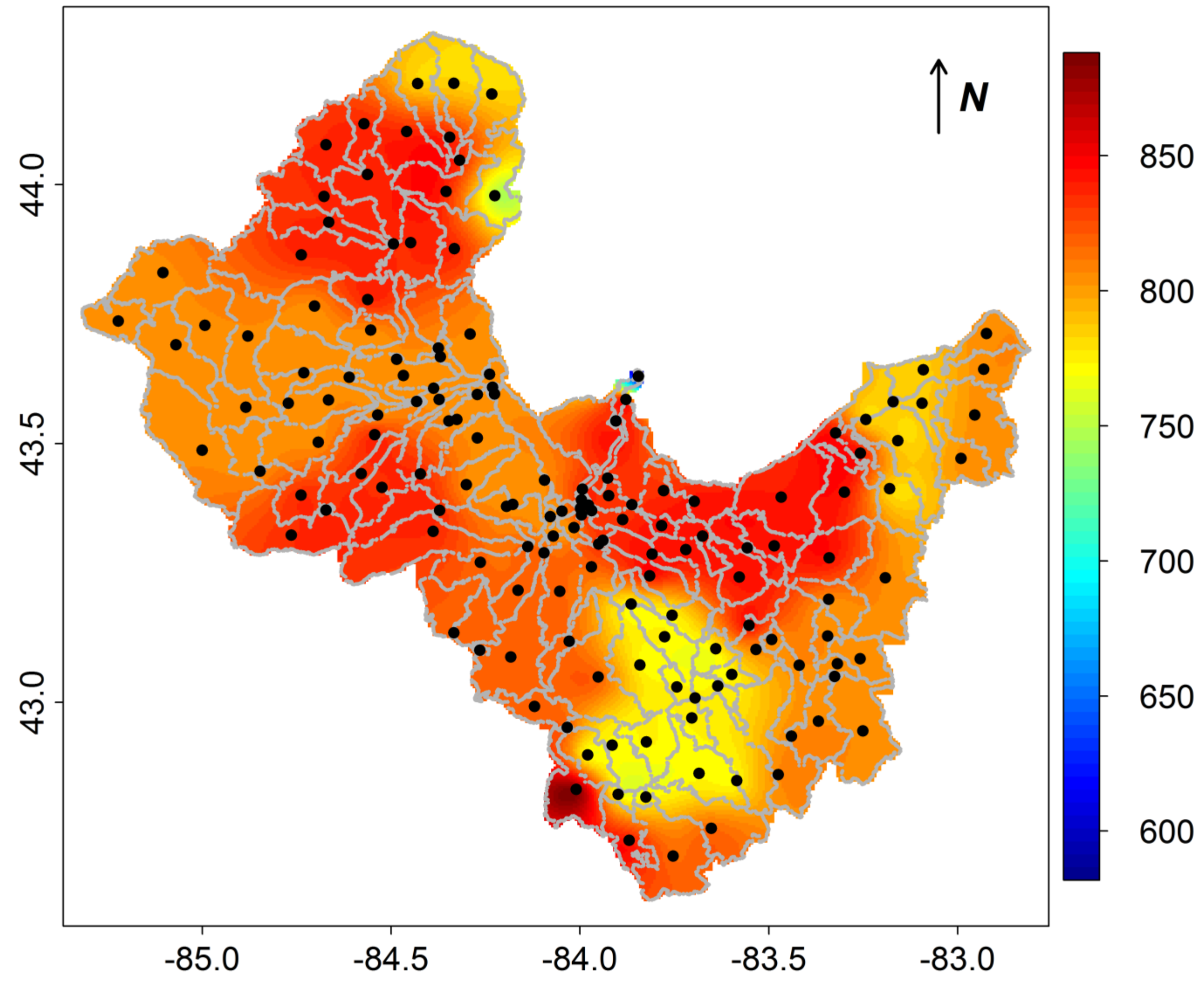

Figure A7. Spatial plot of precipitation predictor variable (mm/year)

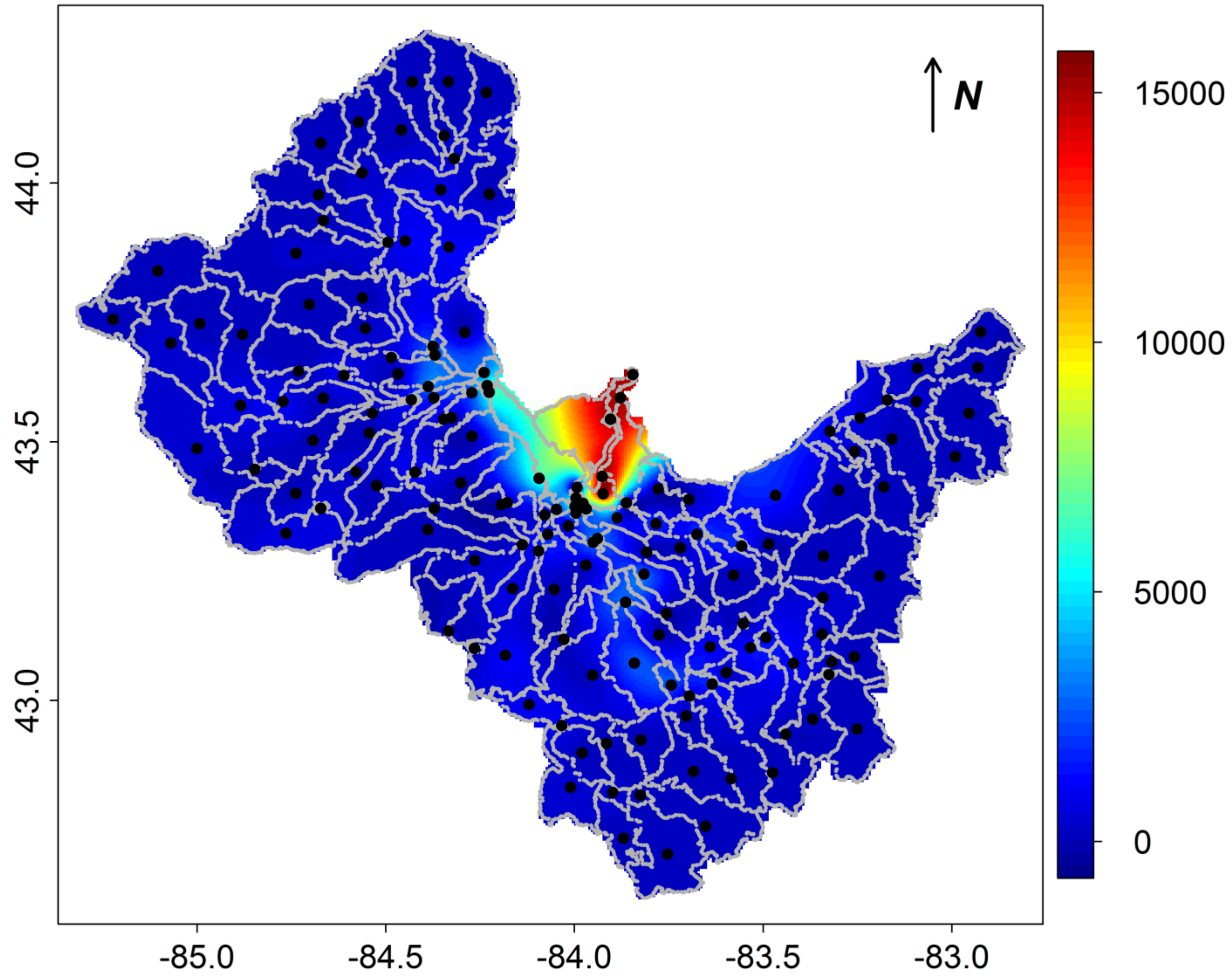

Figure A8. Spatial plot of total area predictor variable (km^2)

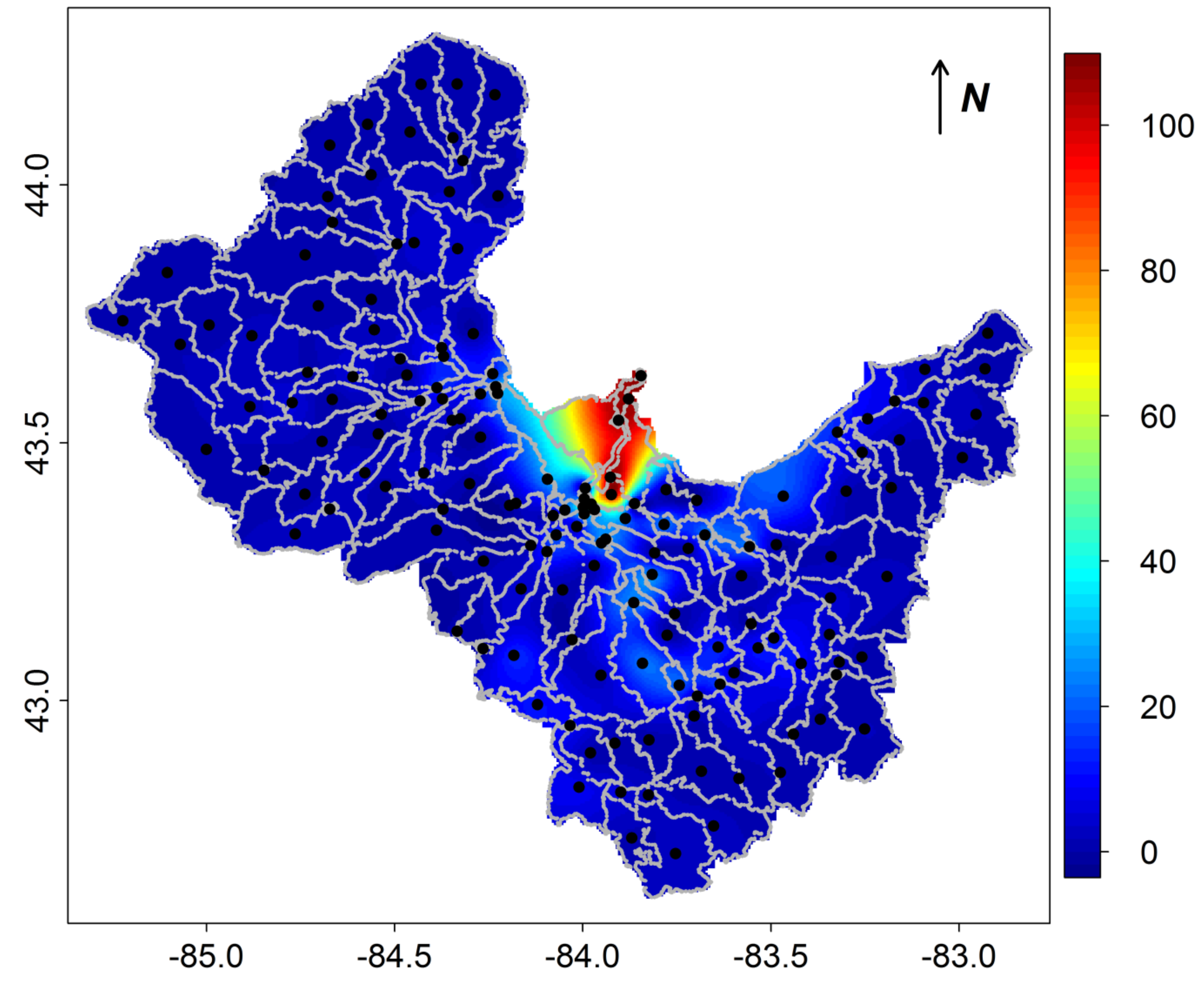

Figure A9. Spatial plot of flow rate (m^3/s)

REFERENCES

REFERENCES

Adriaenssens, V., Baets, B. D., Goethals, P. L., & Pauw, N. D. (2004). Fuzzy rule-based models for decision support in ecosystem management. *Science of the Total Environment*, *319*(1), 1-12.

Alvisi, S., Mascellani, G., Franchini, M., & Bardossy, A. (2006). Water level forecasting through fuzzy logic and artificial neural network approaches. *Hydrology and Earth System Sciences*, 10, 1–17.

Altunkaynak, A., Özger, M., & Çakmakci, M. (2005). Water consumption prediction of Istanbul city by using fuzzy logic approach. *Water Resources Management*, *19*(5), 641-654.

Al-Zu'bi, Y., Sheta, A., & Al-Zu'bi, J. (2010). Nile River flow forecasting based Takagi-Sugeno fuzzy model. *Journal of Applied Sciences*, *10*(4), 284-290.

Arnold, J.G., Srinivasan, R., Muttiah, R.S., &Williams, J.R.(1998). Large area hydrologic model development and assessment part 1: model development. *Journal of the American Water Resources Association*, 34, 73, 89.

Arnold, J. G., Moriasi, D. N., Gassman, P. W., Abbaspour, K. C., White, M. J., Srinivasan, R., ... & Jha, M. K. (2012). SWAT: Model use, calibration, and validation. *Transactions of the ASABE*, *55*(4), 1491-1508.

Arthington, A. H., Bunn, S. E., Poff, N. L., & Naiman, R. J. (2006). The challenge of providing environmental flow rules to sustain river ecosystems. *Ecological Applications*, 16(4), 1311-1318.

Ayyub, B. M., & McCuen, R. (2011). *Probability, statistics, and reliability for engineers and scientists*. Taylor & Francis US.

Azamathulla, H.Md., Ab. Ghani, A., &Fei, S.F. (2012). ANFIS-based approach for predicting sediment transport in clean sewer. *Applied Soft Computing*, 12, 1227–1230.

Bai V, R., Bouwmeester, R., &Mohan, S. (2009). Fuzzy Logic Water Quality Index and Importance of Water Quality Parameters. *Air, Soil and Water Research*, 2, 51–59.

Benedini, M., & Tsakiris, G. (2013). *Water quality modelling for rivers and streams*. Springer.

Bhattacharya, B., Price, &Solomatine, D.P. (2005). Data-driven modelling in the context of sediment transport. *Physics and Chemistry of the Earth*; 30, 297–302.

Bianconi, A., Zuben, C.J.V., Serapaiao, A.B.S., &Govone, J.S. (2010). Artificial neural networks: A novel approach to analyzing the nutritional ecology of a blowfly species, Chrysomya megacephala. *Journal of Insect Science*, 10, (58).

Bouckaert, R. R. (2003). Choosing between two learning algorithms based on calibrated tests. *In Proc. 20th Intl. Conf. on Machine Learning* (ICML-2003) Menlo Park, Cal.: AAAI Press.

Braun, J.V., Gerber, N., Mirzabaev, A., &Nkonya, E. (2012). The Economics of Land Degradation, *An Issue Paper for Global Soil Week.Center for Development Research, University of Bonn*; October 10.

Brundtland, G. (1987). Our common future: *Report of the 1987 World Commission on Environment and Development*.

Bubb, D.H., Lucas, M. C., Thom, T. J., & Rycroft, P. (2002). The potential use of PIT telemetry for identifying and tracking crayfish in their natural environment. *Hydrobiologia*, 483(1-3), 225–230.

Carrasco, E.F. Rodrıguez, J., Punal, A., Roca, E.,   &Lema, J.M.(2004). Diagnosis of acidification states in an anaerobic wastewater treatment plant using a fuzzy-based expert system. *Control Engineering Practice*, 12, 59–64.

Castiglioni, S., Castellarin, A., & Montanari, A. (2009). Prediction of low-flow indices in ungauged basins through physiographical space-based interpolation. *Journal of Hydrology*, 378(3-4), 272–280.

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M. (2006) Deviance information criteria for missing data models. Bayesian Analysis, 1, (4): 651-673.

Chau, K.W., Wu, C. L., & Li, Y. S. (2005). Comparison of Several Flood Forecasting Models in Yangtze River. *Journal of Hydrologic Engineering*, 10(6), 485-491.

Chen, H.W., &Chang, N.B. (2010). Using fuzzy operators to address the complexity in decision making of water resources redistribution in two neighboring river basins. *Advances in Water Resources*, 33, 652–666.

Chin, D. A. (2013). Water-resources engineering. *Pearson Education*, Inc.Upper Saddle River, New Jersey 07458.

Cigizoglu, H.K., Kisi, O. (2006). Methods to improve the neural network performance in suspended sediment estimation. *Journal of Hydrology*, 317,221–238.

Celeux, G., Forbes, F., Robert, C. P., &Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1, (4), 651-673.

Cobaner, M., Unal, B., &Kisi, O. (2009). Suspended sediment concentration estimation by an adaptive neuro-fuzzy and neural network approaches using hydro-meteorological data. *Journal of Hydrology*, 367,52–61.

Cobaner, M. (2011). Evapotranspiration estimation by two different neuro-fuzzy inference systems. *Journal of Hydrology*, 398,292−302.

Deka, P., & Chandramouli,V. (2005). Fuzzy Neural Network Model for Hydrologic Flow Routing. *Journal of Hydrologic Engineering*, 10(4),302-314.

Eldin, A. A., van den Berg, J., & Wagenaar, R. (2004, March). A fuzzy reasoning scheme for context sharing decision making. In *Proceedings of the 6th international conference on Electronic commerce* (pp. 371-375). ACM.

El-Sebakhy, E. A., Adem, I.R.S., &Khaeruzzaman, Y. (2007). Neuro-Fuzzy Systems Modeling Tools for Bacterial Growth. IEEE, *International Conference on Computer Systems and Applications, Amman, Jordan*, 374-380.

Eng, K., & Milly, P. C. D. (2007). Relating low-flow characteristics to the base flow recession time constant at partial record stream gauges. *Water Resour. Res.*, 43 (W01201), 1-8.

EPA, (2012). United States Environmental Protection Agencies, Great Lakes areas of concerns. http://www.epa.gov/grtlakes/aoc/saginaw-river/index.html. Accessed 2013.

EPA, (2013). United States Environmental Protection Agencies. Glossary and Acronyms. http://water.epa.gov/polwaste/sediments/cs/glossary.cfm. Accessed 2013.

Eslamian, S., Ghasemizadeh, M., Biabanaki, M., & Talebizadeh, M. (2010). A Principal Component Regression Method for Estimating Low Flow Index. *Water Resour Manage.*, 24(11),2553–2566.

Fan, L., &Boshnakov, K. (2010). Fuzzy Logic based Dissolved Oxygen Control for SBR Wastewater treatment Process. *8th World Congress on Intelligent Control and Automation*, July 6-9, Jinan, China.

Gassman, P.W., Reyes, M.R., Green, C.H., Arnold, J.G. (2007). The Soil and Water Assessment Tool: historical development, applications and future research directions. *Transaction of the ASABE*, 50, 1211-1250.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, *85*(410), 398-409.

Ghalhary, G.A.F, Baygi, M.M., & Nokhandan, H. (2009). Annual Rainfall Forecasting by Using Mamdani Fuzzy Inference System. *Research Journal of Environmental Science*, 3 (4), 400-413.

Gu, X., Song, G., & Xiao, L. (2006, November). Design of a Fuzzy Decision-making Model and Its Application to Software Functional Size Measurement. *In Computational Intelligence for Modelling, Control and Automation, 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on* (pp. 199-199). IEEE.

Gunawardana, C., Goonetilleke, A., Egodawatta, P., Dawes, L., & Kokot, S. (2011). Role of solids in heavy metals buildup on urban road surfaces. *Journal of Environmental Engineering, 138*(4), 490-498.

Guney, K., sarikaya, N. (2009). Comparison of Mamdani and Sugeno Fuzzy Inference System Models for resonant Frequency Calculation of Rectangular Microstrip Antennas. *Progress In Electromagnetics Research B*, 12,81–104.

Haberlandt, U., Krysanova, V., &Bardossy, A. (2002). Assessment of nitrogen leaching from arable land in large river basins Part II: regionalisation using fuzzy rule based modeling. *Ecological Modelling*, 150, 277–294.

Hamaamin, Y.A., Nejadhashemi, A.P, & Einheuser, M.D. (2013). Application of Fuzzy Logic Techniques in Estimating the Regional Index Flow for Michigan. *Transactions of the ASABE*, 56(1),103-115.

Hamilton, D. A., Sorrell, R. C., & Holtschlag, D. J. (2008). *A regression model for computing index flows describing the median flow for the summer month of lowest flow in Michigan*. US Department of the Interior, US Geological Survey.

Harmancioglu, N. B., Barbaros, F., & Cetinkaya, C. P. (2012). Sustainability issues in water management. *Water Resources Management*, 1-25.

Hejazi M.I., & Moglen, G. E. (2007). Regression-based approach to low flow prediction in the Maryland Piedmont region under joint climate and land use change. *Hydrol. Process.*, 21(14),1793–1801.

Hoff, P.D. (2009). A first Course in Bayesian Statistical Methods. *Springer Science + Business Media*, LLC, 233 Spring Street, New York, NY10013, USA.

Hon, K. An Introduction to Statistics, http://www.artofproblemsolving.com/LaTeX/ Examples/statistics_firstfive.pdf, accessed, 2013.

Huang, Y., Lan, Y., Thomson ,S. J., Fang ,A., Hoffmann ,W. C., & Lacey ,R. E.(2010). Development of soft computing and applications in agricultural and biological engineering. *Computers and Electronics in Agriculture*, 71 (2),107–127.

Hundecha, Y., Bardossy, A., & Theisen, H.W. (2001). Development of a fuzzy logic-based rainfall-runoff model. *Hydrological Sciences*, 46(3), 363-376.

Ibrahim, A.M. (2004). *Fuzzy Logic for Embedded systems applications*. Elsevier Science, Burlington, MA.

Jacquin, A.P., & Shamseldin, A. Y.(2006). Development of rainfall–runoff models using Takagi–Sugeno fuzzy inference systems. *Journal of Hydrology*, 329(1-2), 154– 173.

Jain, L.C., &Martin, N.M. (1998). Fusion of Neural Networks, Fuzzy Systems and Genetic Algorithms: Industrial Applications. CRC Press LLC, Boca Raton, FL.

Jain, S.K. (2001). Development of Integrated sediment Rating Curves Using ANNs. Journal of Hydraulic Engineering, 127(1), 30-37.

Jain, S.K., &Singh, V.P. (2003). *Water Recourse Planning and Management*. Elsevier, Antony Rowe Ltd, Eastboume, Netherland.

Jang J.S.R. (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE Trans Syst Man Cybern*, 23 (3),665–685.

Johnson, L. E. (2008). *Geographic information systems in water resources engineering*. CRC Press.

Keskin, M.R., Terzi, O., & Taylan, D. (2004). Fuzzy logic model approaches to daily pan evaporation estimation in western Turkey. *Hydrological Sciences Journal*, 49(6), 1001-1010.

Keskin, M.E., Terzi, Ö., & Taylan ,D.(2009). Estimating daily pan evaporation using adaptive neural-based fuzzy inference system. Theor. Appl. Climatol., 98(1-2), 79–87.

Kisi, O. (2004). Daily suspended sediment modelling using a fuzzy differential evolution approach. *Hydrological Sciences Journal*, 49(1) February, 183-197.

Kisi, O. (2005). Suspended sediment estimation using neuro-fuzzy and neural network approaches. *Hydrological Sciences Journal*, 50(4), 683-696.

Kisi, O., Karahan, M.E., & Sen, Z. (2006). River suspended sediment modeling using a fuzzy logic approach. Hydrological Processes; 20: 4351–4362.

Kisi, O.(2006). Daily pan evaporation modeling using a neuro-fuzzy computing technique. *Journal of Hydrology*, 329(3-4), 636– 646.

Kisi, O. (2008). Constructing neural network sediment estimation models using a data-driven algorithm. *Mathematics and Computers in Simulation*, 79, 94–103.

Kisi, O., Haktanir, T., Ardiclioglu, M., Ozturk, O., Yalcin, E., &Uludag, S. (2009). Adaptive neuro-fuzzy computing technique for suspended sediment estimation. Advances in Engineering Software, 40, 438−444.

Kisi, O. (2010). River suspended sediment concentration modeling using a neural differential evolution approach. *Journal of Hydrology*, 389, 227–235.

Kisi, O., &Shiri, J. (2012). River suspended sediment estimation by climatic variables implication: Comparative study among soft computing techniques. *Computers & Geosciences*, 43, 73–82.

Klir, G.J., Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic Theory and Applications*. Prentice Hall P T R, Upper Saddle River, New Jersey.

Kosko, B. (1993). *Fuzzy Thinking*. Hyperion, New York.

Koster, B.S., Bunn, S.E., Mackay, S.J., Poff, N.L., Naiman, R.J., & Lake, P.S. (2010). The use of Bayesian networks to guide investments in flow and catchment restoration for impaired river ecosystems. *Freshwater Biology*, 55(1), 243–260.

Kroll, C., Luz ,J., Allen ,B., R. & Vogel ,M.(2004). Developing a Watershed Characteristics Database to Improve Low Streamflow Prediction. *J. of Hydrologic Engineering*, 9(2), 116-125.

Kruschke, J. (2010). *Doing Bayesian Data Analysis: A Tutorial Introduction with R*. Academic Press.

Kumar, A.R.S., Ojha, C.S.P., Goyal, M.K., Singh, R.D., &Swamee, P.K. (2012). Modeling of Suspended Sediment Concentration at Kasol in India Using ANN, Fuzzy Logic, and Decision Tree Algorithms. *Journal of Hydraulic Engineering @ASCE* / March, 394-404.

Lee, K. S., & Kim, S.U. (2008). Identification of uncertainty in low flow frequency analysis using Bayesian MCMC method. *Hydrol. Process*., 22(12), 1949-1964.

Leisenring, M., &Moradkhani, H. (2012). Analyzing the uncertainty of suspended sediment load prediction using sequential data assimilation *Journal of Hydrology*, 468–469, 268–282.

Lermontov, A., Yokoyama, L., Lermontov, M., &Machado, M.A.S. (2009). River quality analysis using fuzzy water quality index: Ribeira do Iguape river watershed, Brazil. *Ecological Indicators*, 9,1188–1197.

Li, X., Ruan, D., van der Wal, A.J. (1998) Discussion on soft computing at FLINS 96. *International Journal of Intelligent Systems*, 13(2-3), 287- 300.

Li, z., Zhang, Y.K., Schilling, K., &Skopec, M. (2006) Cokriging estimation of daily suspended sediment loads. *Journal of Hydrology*, 327, 389– 398.

Love, B. & Nejadhashemi, A.P. (2011). Water quality impact assessment of large-scale biofuel crops expansion in agricultural regions of Michigan. *Biomass and Bioenergy*, 35, 2200–2216.

Loucks, D. P., Van Beek, E., Stedinger, J. R., Dijkman, J. P., & Villars, M. T. (2005). *Water resources systems planning and management: an introduction to methods, models and applications*. Paris: UNESCO.

Lyman, O.R., & Longnecker, M. (2010). An Introduction to Statistical Methods and Data Analysis, 6th edition, *BROOKS/COLE Cengage Learning*.

Mah, D.Y.S. (2011). Conservation of Sarawak peat swamp in an urban landscape by fuzzy inference system. *Reg. Environ Change*, 11,307–310.

Mahabir, C., Hicks, F. E. & Fayek, A. R. (2003). Application of fuzzy logic to forecast seasonal runoff. *Hydrol. Process*. , 17(18), 3749–3762.

Mahabir, C., Hicks, F., & Fayek, A. R. (2006). Neuro-fuzzy river ice breakup forecasting system. *Cold Regions Science and Technology*, 46 (2), 100–112.

Mahmood, Z., & Khan, S. (2009). On the Use of K-Fold Cross-Validation to Choose Cutoff Values and Assess the Performance of Predictive Models in Stepwise Regression. *The International Journal of Biostatistics*, 5(1), Article 25.

Mamdani, E.H. (1974). Application of fuzzy algorithms for simple dynamic plant. *Pro.IEE*., 121(12), 1585-1588.

Mamdani, E.H., &Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *Int J Man Mach Stud*, 7(1),1–13.

Mandal, U., & Cunnane, C. (2009). Low-flow Prediction for Ungauged River Cathchemnts in Ireland. Irish National Hydrology Seminar, 33-48.

Marchini, A., Facchinetti, T., &Mistri, M.(2009). F-IND: A framework to design fuzzy indices of environmental conditions. *Ecological indicators*, 9, 485-496.

Martin, G. R., & Arihood, L. D. (2010). Methods for Estimating Selected Low-Flow Frequency Statistics for Unregulated Streams in Kentucky. *U.S. Geological Survey,* Scientific Investigations Report 2010–5217.

Marza, V., & Seyyedi, M.A. (2009). Fuzzy Multiple Regression Model for
Estimating Software Development Time. *International Journal of Engineering Business Management*, 1(2), 79-82.

MathWorks, (2010). MATLAB Fuzzy Logic Toolbox User's Guide2, *The MathWorks, Inc.*

Mckone, T.E., & Deshpande, A. W. (2005). Can Fuzzy LOGIC Bring Complex Environmental Problems into Focus. *Environmental science and technology*, Jan 15.

Michigan Climatological Resources Program (2004). Michigan climatological normals, 1971–2000: East *Lansing, Michigan, State Climatologist's Office*, Michigan State University, digital media.

Michigan Department of Information Technology (2005). Glacial deposits estimated transmissivity geographic theme: *Lansing, Michigan. Center for Geographic Information.* http://www.mcgi.state.mi.us/mgdl/?rel=ext&action=sext. Accessed 22 September 2011.

Michigan Library (2006). Michigan in Brief—Information about the state of Michigan: *Lansing, Michigan*. www.michigan.gov/hal/0,1607,7-160-15481_20826_20829-56001--,00.html. Accessed 22 September 2011.

Michigan's Public Act -33, (2006). Michigan Public Acts of 2006, *Act No. 33: 93rd Legislature, Regular Session of 2006*. http://www.legislature.mi.gov/documents/2005-2006/publicact/pdf/2006-PA-0033.pdf. Accessed 22 September 2011.

MIRIS, (1978). Michigan Resource Information System Landuse/cover Polygon—Geographic data and metadata: Lansing, Michigan. Michigan Department of Natural Resources. www.mcgi.state.mi.us/mgdl/. Accessed 22 September 2011.

Mishra, A., Hata ,T., Abdelhadi ,A. W., Tada , & Tanakamaru ,A. H.(2003). Recession flow analysis of the Blue Nile River. Hydrol. Process., 17(14),2825–2835.

Mitchell, M., (1999). *An Introduction to Genetic Algorithms*. A Bradford Book The MIT Press, Massachusetts Institute of Technology.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., &Veith, T.L. (2007). Model Evaluations Guidelines for Systematic quantification of accuracy in Watershed Simulations. *Transactions of the ASABE*, 50(3), 885−900.

Mount, N., &Stott, T. (2008). A discrete Bayesian network to investigate suspended sediment concentrations in an Alpine proglacial zone. *Hydrol. Process*. 22, 3772−3784.

Nagy, H. M., Watanabe, K., &Hirano, M. (2002). Prediction of Sediment Load Concentration in Rivers using Artificial Neural Network Model. *Journal of Hydraulic Engineering*, 128 (6), 588-595.

Nasiri, F., Maqsood, I., Huang, G., &Fuller, N. (2007). Water Quality Index: A Fuzzy River-Pollution Decision Support Expert System. *Journal of Water Resources Planning and Management*,133(2), 95-105.

NASS (2008) National Agricultural Statistics Service: CropScape-Cropland Data Layer. http://nassgeodata.gmu.edu/CropScape/. Accessed 2013.

Nayak P.C., Sudheer, K. P. , & Ramasastri, K. S. (2005). Fuzzy computing based rainfall–runoff model for real time flood forecasting. *Hydrol. Process*., 19(4) ,955–968.

Nayak, P.C. (2010). Explaining Internal Behavior in a Fuzzy If-Then Rule-Based Flood-Forecasting Model. *Journal of Hydrologic Engineering*, 15(1), 20-28.

Nayak, C.P., &Jain, S.K. (2011). Modelling runoff and sediment rate using a neuro-fuzzy technique. *Water Management*, 164(WM4),201-209.

Neitsch, S. L., Arnold, J. G., Kiniry J. R., & Williams J. R. (2005). Soil and Water Assessment Tool Theoretical Documentation, Version 2005. Temple, Tex.

Nova, Nova Scotia Department of the Environment (1988). Erosion and Sedimentation Control. Environmental Assessment Division, Canada. ISBN 0-88871-116-6.

NORS (1999).*The National Organization for Rivers*. (http://www.nationalrivers.org/states/mi-view.htm). Accessed July 10, 2010.

NRCS (2007). Natural Resources Conservation Service. Hydrologic soil groups, Chapter 7 of Part 630. *Hydrology of National engineering handbook*, http://policy.nrcs.usda.gov/media/pdf/H_210_630_7.pdf. Accessed 22 September 2011.

Ozturk, F., Apaydin, H., &Walling, D.F. (2001). Suspended Sediment Loads Through Flood Events for Streams of Sakarya River Basin. *Turk J Engin Environ Science*, 25, 643- 650.

Pyrce, R. (2004). Hydrological low flow indices and their uses. *Watershed Science Centre,(WSC) Report*, (04-2004).

Pomerantsev, A.L., &Rodionova, O.Y., (2005). Hard and soft methods for prediction of antioxidants’ activity based on the DSC measurements. *Chemometrics and Intelligent Laboratory Systems*, 79, 73 – 83.

Rai, R.K., Mathur, B.S. (2008). Event-based Sediment Yield Modeling using Artificial Neural Network. *Water Resour Manage*, 22,423–441.

Rajaee, T.(2010). Wavelet and Neuro-fuzzy Conjunction Approach for Suspended Sediment Prediction. *Clean – Soil, Air, Water*, 38 (3), 275 – 288.

Rezapour, O.M., Shui, L.T., &Ahmad, D.B. (2010). Review of Artificial Neural Network Model for Suspended Sediment Estimation. *Australian Journal of Basic and Applied Sciences*, 4(8), 3347-3353.

Ritter, C., Tanner, M. A. (1992). Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler. *Journal of the American Statistical Association*, 87(419).

Ritter, W.F., Shirmohammadi, A.(2001). Agricultural nonpoint source pollution: Watershed management and hydrology. *CRC Press,* Boca Raton, Florida.

Ross, T. J. (2004). *Fuzzy logic with engineering applications*. John Wiley & Sons.

Samhouri, M., Abu-Ghoush, M., Yaseen, E., & Herald, T. (2009). Fuzzy clustering-based modeling of surface interactions and emulsions of selected whey protein concentrate combined to ι-carrageenan and gum arabic solutions. *Journal of Food Engineering*, 91(1), 10-17.

Sanikhani, H., Kisi,O. (2012). River Flow Estimation and Forecasting by Using Two Different Adaptive Neuro-Fuzzy Approaches. *Water Resour Manage*, 26,1715–1729.

Santhi, C., Allen, P.M., Muttiah, R.S., Arnold, J.G., & Tuppad, P. (2008). Regional estimation of base flow for the conterminous United States by hydrologic landscape regions. Journal of Hydrology, 351(1-2), 139– 153.

Schmelter, M.L., Hooten, M.B., &Stevens, D.K. (2011). Bayesian sediment transport model for unisize bed load. *WATER RESOURCES RESEARCH*, 47, W11514.

Schmelter, M.L., Erwin, S.O., &Wilcock, P.R. (2012) Accounting for uncertainty in cumulative sediment transport using Bayesian statistics. *Geomorphology*, 175–176 , 1–13.

Sen, Z. (2010). *Fuzzy logic and hydrological modeling*. Boca Raton, FL: CRC Press.

Shu, C., and Ouarda, T.B.M.J. (2008). Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system. *Journal of Hydrology*, 349(1-2), 31–43.

Sivakumar, B. (2006). Suspended sediment load estimation and the problem of inadequate data sampling: a fractal view. *Earth Surf. Process. Landforms*, 31, 414–427.

Sivanandam, S. N., Sumathi, S., &Deepa, S. N. (2007). Introduction of Fuzzy Logic using MATLAB. *Springer-Verlag*, Berlin Heidelberg.

Smakhtin, V.U. (2001). Low flow hydrology: a review. *Journal of Hydrology*, 240(1-2), 147–186.

Solomatine, D., See, L.M., &Abrahart, R.J. (2008). Practical Hydroinformatics, Computational Intelligence and Technological Developments in Water Applications Chapter 2: Data-Driven Modelling: Concepts, Approaches and Experiences. *Springer-Verlag*, Berlin Heidelberg.

STATSGO (1995). USDA. State soil geographic data base. Data use information http://www. nrcs.usda.gov/technical/techtools/statsgo_db.pdf 1995.

Takagi, T., & Sugeno, M. (1985). Fuzzy Identification of Systems and Its Applications to Modeling and Control. *IEEE Transactions on Systems*, Man, and Cybernetics, 15(1), 116-132.

Tayfur, G., Ozdemir, S., &Singh, V.P. (2003). Fuzzy logic algorithm for runoff-induced sediment transport from bare soil surfaces. *Advances in Water Resources*, 26, 1249–1256.

Thipparat, T. (2012). Application of Adaptive Neuro Fuzzy, Fuzzy Logic - Algorithms, Techniques and Implementations-Ch.6, Prof. Elmer Dadios (Ed.), ISBN: 978-953-51-0393-6, InTech.

Toprak, Z.F. (2009). Flow Discharge Modeling in Open Canals Using a New Fuzzy Modeling Technique (SMRGT). Clean, 37 (9), 742 – 752.

Toprak, Z. F., Eris ,E., Agiralioglu ,N., Cigizoglu ,H. K., Yilmaz ,L., Aksoy ,H., Coskun ,H. G., Andic ,G., &Alganci ,U.(2009) . Modeling Monthly Mean Flow in a Poorly Gauged Basin by Fuzzy Logic. *Clean*, 37 (7), 555 – 564.

Tzimopoulos, C., Mpallas ,L., & Evangelides ,C.(2008). Fuzzy Model Comparison to Extrapolate Rainfall Data. *Journal of Environmental Science and Technology*, 1(4), 214-224.

Ulke, A., Tayfur, G., &Ozkul, S. (2009). Predicting Suspended Sediment Loads and Missing Data for Gediz River, Turkey. *Journal of Hydrologic Engineering*, 14(9), 954-965.

Wang, Y., Traore, S., &Kerh, T. (2008). Monitoring Event-Based Suspended Sediment Concentration by Artificial Neural Network Models. *WSEAS TRANSACTIONS on COMPUTERS*, 7(5),359-368.

World Metrological Organization (2008). Manual on low-flow estimation and Prediction. *WMO-No.1029*, Operational Hydrology Report No.50.

Yazdi, H.S., &Pourreza, R. (2010). Unsupervised adaptive neural-fuzzy inference system for solving differential equations. Applied Soft Computing, 10, 267–275.

Zadeh, L.A. (1965). Fuzzy sets. *Inform Control*, 8, 338–353.

Zimmermann, H.J. (2001). Fuzzy Sets Theory and its Applications. *Kluwer Academic Publishers Group*, 3300 AH Dordrecht, The Netherlands.