

**APPLICATION OF A SPATIALLY REFERENCED WATER QUALITY MODEL
TO PREDICT *E. coli* FLUX IN TWO TEXAS RIVER BASINS**

A Thesis

by

DEEPTI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

August 2008

Major Subject: Biological and Agricultural Engineering

**APPLICATION OF A SPATIALLY REFERENCED WATER QUALITY MODEL
TO PREDICT *E. coli* FLUX IN TWO TEXAS RIVER BASINS**

A Thesis

by

DEEPTI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Raghupathy Karthikeyan
Committee Members,	Patricia Smith
	Emily Zechman
Head of Department,	Gerald Riskowski

August 2008

Major Subject: Biological and Agricultural Engineering

ABSTRACT

Application of a Spatially Referenced Water Quality Model to Predict *E. coli* Flux in
Two Texas River Basins. (August 2008)

Deepti, B.Tech., Punjab Agricultural University, Ludhiana, India;

M.Tech., Indian Institute of Technology, Kharagpur, India

Chair of Advisory Committee: Dr. Raghupathy Karthikeyan

Water quality models are applied to assess the various processes affecting the concentrations of contaminants in a watershed. SPATIALLY Referenced Regression On Watershed attributes (SPARROW) is a nonlinear regression based approach to predict the fate and transport of contaminants in river basins. In this research SPARROW was applied to the Guadalupe and San Antonio River Basins of Texas to assess *E. coli* contamination. Since SPARROW relies on the measured records of concentrations of contaminants collected at monitoring stations for the prediction, the effect of the locations and selections of the monitoring stations was analyzed. The results of SPARROW application were studied in detail to evaluate the contribution from the statistically significant sources. For verification of SPARROW application, results were compared to 303 (d) list of Clean Water Act, 2000. Further, a methodology to maintain the monitoring records of the highly contaminated areas in the watersheds was explored with the application of the genetic algorithm. In this study, the importance of the available scale and details of explanatory variables (sources, land-water delivery and reservoir/ stream attenuation factors) in predicting the water quality processes were also analyzed. The effect of uncertainty in the monitored records on SPARROW application was discussed. The application of SPARROW and genetic algorithm were explored to design a monitoring network for the study area. The results of this study show that SPARROW model can be used successfully to predict the pathogen contamination of rivers. Also, SPARROW can be applied to design the monitoring network for the basins.

DEDICATION

This work is dedicated to the efforts of my mother Mrs. Ushendra Jaiswal and memories of my late father Raj Kumar Jaiswal.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Raghupathy Karthikeyan, and my committee members, Dr. Patricia Smith and Dr. Emily Zechman, for their guidance and support throughout the course of this research.

Thanks also go to Dr. Meghna Babbar-Sebens who assisted me with the technical details of my research. I am thankful to my friends Aarin, Bailey, Kendra and Vinay and the faculty and staff of Biological and Agricultural Engineering, for making my time at Texas A&M University a great experience. I would like to acknowledge the Department of Biological and Agricultural Engineering for providing a graduate teaching assistantship to support my graduate studies in this great university. I want to extend my gratitude to the Texas Water Resources Institute, which provided the Mills Scholarship and USGS Research Grant to partially support this research. I am also thankful to the Office of Graduate Studies for providing a travel grant to present this research at a conference.

TABLE OF CONTENTS

	Page
ABSTRACT	iii
DEDICATION	iv
ACKNOWLEDGEMENTS	v
TABLE OF CONTENTS	vi
CHAPTER	
I INTRODUCTION	1
1.1 Pathogen Problem	1
1.2 Modeling Approaches for Water Quality Assessment.....	2
1.3 Spatially Referenced Regression Model	3
1.4 Advantages and Disadvantages of SPARROW Modeling Approach ..	4
1.5 Applications of Genetic Algorithm in Water Quality Models.....	5
1.6 Scope of the Research.....	6
1.7 Objectives of the Research.....	6
II PREDICTING THE FATE AND TRANSPORT OF <i>E. coli</i> IN TWO TEXAS RIVER BASINS USING A SPATIALLY REFERENCED REGRESSION MODEL	7
2.1 Introduction.....	7
2.2 Methodology	9
2.2.1 Model Details.....	9
2.2.2 Selection of the Best Model	11
2.2.3 Study Area	13
2.2.4 Data Sources	13
2.3 Results and Discussion	18
2.3.1 Results of Three Prediction Models	19
2.3.2 Coefficients Estimation of the Best Model.....	21
2.3.3 Prediction of <i>E. coli</i> Flux Using Model III	22
2.4 Conclusions.....	25

CHAPTER	Page
III OPTIMIZATION OF WATER QUALITY MONITORING NETWORK TO MONITOR <i>E. coli</i> USING A SPATIALLY REFERENCED WATER QUALITY MODEL AND GENETIC ALGORITHM.....	27
3.1 Introduction.....	27
3.2 Methodology	30
3.2.1 Study Area and Data Sources	30
3.2.2 Genetic Algorithm.....	31
3.2.3 Representation, Objectives and Constraints	32
3.2.4 GA Application	35
3.3 Results and Discussion	36
3.4 Conclusions.....	39
IV SUMMARY AND CONCLUSIONS.....	40
4.1 Summary	40
4.2 Conclusions	40
4.3 Scope for Future Research	42
REFERENCES.....	43
APPENDIX A FIGURES.....	48
APPENDIX B TABLES	69
VITA	75

CHAPTER I INTRODUCTION

1.1 Pathogen Problem

Pathogenic contamination affects designated water uses such as recreation, public water supplies, aquifer protection, fish, shellfish and wildlife protection and propagation. Animals (wildlife, pets and cattle), storm-water runoff, combined sewer and sanitary sewer overflows and wastewater treatment plant effluents are the important sources of pathogenic microorganisms. Watershed attributes such as precipitation, temperature, soil properties (permeability, pH, salinity and available nutrients), land use characteristics and sunlight affect the land-water transport of pathogens in surface water (Ferguson et al., 2003).

In accordance with Section 305(b) of the Clean Water Act (CWA), Texas Commission on Environmental Quality (TCEQ) prepares the 303(d) list of impaired waterbodies and streams (TCEQ, 2006a). In the summary of the 2006 303(d) list, 41.7 % of the streams in Texas were contaminated by bacteria. There was a 31% increase in impaired waterbodies observed since the last assessment and the largest increase (about 60%) was observed in contamination from bacteria. Pathogenic contamination of streams and waterbodies has affected recreational activities considerably. Section 305(b) estimates a \$630 million loss of recreational revenue in Texas is estimated because of high concentrations of pathogens in ocean shorelines and waterbodies. The estimated annual cost of CWA program implementation for Texas is \$11 million (TCEQ, 2006b). The CWA program requires the state to develop pollutant specific Total Maximum Daily Load (TMDL) for the 303(d) listed waterbodies.

This thesis follows the style of *Transactions of the ASABE*.

1.2 Modeling Approaches for Water Quality Assessment

Water quality models are frequently used to predict the concentration of contaminants in stream networks because only a limited number of monitoring stations can be established in a basin because of the involved cost of monitoring. These models have been widely used in TMDL development and implementation processes. A selected water quality model should be thorough, robust and accurate in describing the watersheds' processes and prediction. It should capture the spatial and temporal variability of watershed attributes and include details according to the scale of application. It should also be able to predict the contaminant concentration at unmonitored locations of the watershed (Letcher et al., 2004). The water quality models can be defined as mechanistic or statistical based on the model development. In mechanistic models it is assumed that a complex and comprehensive description of model parameters results in accurate predictions. A macroscopic-scale mechanistic model requires time, effort and large amount of data for calibration and validation. Data at the required resolution are rarely available and the output accuracy of the model is limited. Uncertainty in mechanistic models results in the analysis of watershed processes using statistical, stochastic or heuristic methods.

A statistical or empirical model predicts water quality by interpolation of available data with or without actually including the important mechanisms. These models involve judgment in parameter specification but not in their actual significance and value estimation. Artificial Neural Networks (ANN), kriging, linear or nonlinear regression are some common statistical modeling approaches. These models can be applied to large watersheds or river basins with diverse features. Here, the error terms associated with model predictions can be easily quantified. The disadvantages of statistical models include poor representation and little explanation of physical processes and lack of spatial details in output. Thus, desired details, current knowledge and available data become important constraints to select a water quality model (Reckhow, 1994).

1.3 Spatially Referenced Regression Model

A Spatially Referenced Nonlinear Regression Model on Watershed Attributes (SPARROW) has been applied to predict the fluxes and concentrations of contaminants in unmonitored stream reaches and to track the sources of these contaminants (Alexander et al., 2002). This model can be characterized as a hybrid approach of mechanistic and statistical models. SPARROW identifies every stream-reach as a basic unit to spatially distribute contaminant sources, delivery and attenuation factors to represent the stream network in detail. Simple empirical relations can be defined for the fate and transport of specific contaminants in the watershed. The SPARROW model is based on a least squares fit of mean annual flux of a contaminant with explanatory variables such as land use, population, fertilizer application, precipitation and soil permeability. The mean annual flux for every monitoring station is calculated from the recorded concentration of contaminant and stream-flow data with a detrending model such as LOADEST or FLUXMASTER (Alexander et al., 2002; McMahon et al., 2003).

In addition to the regression approach, SPARROW can include mass balance constraints and nonlinear and spatially distributed parameters. Preston and Brakebill (1999) have reported that the results of SPARROW improved with high resolution spatial data. Nutrient assessment of Chesapeake Bay using this model improved significantly by including the coastal details available in the National Hydrography Dataset (1:24,000) as compared to earlier available hydrography data of coarse scale (1:500,000). The minimum size of the study area for the application of SPARROW is constrained by the number of monitoring stations available in the basin (Smith et al., 1997). Schwarz et al. (2007) have concluded that the model complexity (number of significant parameters) can be increased by including more monitoring stations and monitored data. They recommended at least 20 monitoring stations to include the regional variability in the application of SPARROW. Therefore, minimum error in prediction is related to the length of monitoring records and quality and scale of explanatory variables.

1.4 Advantages and Disadvantages of the SPARROW Modeling Approach

The SPARROW model output includes the concentration of contaminants for every stream-reach. It is easy to identify the reaches with some uncertainty for which the water quality standards are not satisfied. A National Research Council (NRC) assessment committee has recommended the SPARROW model in a Bayesian framework or in conjunction with other models to make TMDL decisions (McMahon et al., 2003). SPARROW also includes the percentage contribution of contaminant load from various sources in a river basin. This information along with statistical estimates of involved error can be useful for load allocation in developing a TMDL program (Moore et al., 2004).

In nonlinear regression, parameter coefficients are assessed for all observations. However, extreme values of a response variable (mean annual flux of contaminant) may have an effect on the values of coefficients. Bootstrap analysis helps to determine the effect of extreme values on the coefficients. However, regional patterns of the residuals of mean annual flux cannot be detected using this analysis. The spatial patterns can only be recognized from the localized clusters of same-sign residuals. By including the autocorrelation due to explanatory variables, the model can be significantly improved (McMahon et al., 2003). To address the model and data errors in SPARROW, Qian et al. (2005) applied the Markov Chain Monte Carlo method to estimate the posterior distribution because the least square method did not predict the parameters for every watershed or spatial autocorrelation in error terms. They applied the State Space (STSP) model to adjust the errors due to an upstream monitoring station to a downstream station and used a Conditional Autoregressive (CAR) term to fix the problem of arbitrary spatial correlation. Based on Bayes factor and Deviance Information Criterion (DIC), the SPARROW model performance was observed to improve for predicting the nutrients in North Carolina (Qian et al., 2005).

A regression based model may not be able to include all important factors and their mutual interactions. Currently, SPARROW application depends on the estimate of mean annual flux measured from concentrations of contaminants and streamflow data (Moore et al., 2004). Large errors in determination of mean annual flux for the contaminants resulting from short time periods and irregularly observed records can affect the model output. Unfortunately, *E. coli* measurements are not frequently carried out and estimation of mean annual *E. coli* flux using SPARROW may result in large errors. Further, monitoring data may have associated uncertainty from sample collection, handling and laboratory analysis of the contaminant and from streamflow measurement (Harmel et al., 2006). Moriasi et al. (2007) have given guidelines for the determination of goodness-of-fit and the accuracy of prediction of a model. To quantify the effects of measurement error on model prediction, different modifications in error values can be applied on the basis of probability distribution and moment information of the data (Harmel and Smith, 2007).

1.5 Applications of Genetic Algorithm in Water Quality Modeling

Genetic Algorithm (GA), a heuristic technique for optimizing, is based on Darwin's evolution concept of natural selection. In the water quality discipline, GA has been applied for model calibration and sensitivity analysis and load allocation with uncertainty analysis (Savic and Khu, 2005; Srivastava et al., 2002; Yandamuri et al., 2006). Ning and Chang (2004) have applied a fuzzy weighing approach to form a single objective problem from multiobjectives based on water quality compliance, spatial influence and proximity of monitoring site to human population in a GA application to expand an existing monitoring network. Using GA, Park et al. (2006) have designed a monitoring network based on river basin representation, water-quality standard compliance and pollution sources supervision. They found that existing monitoring-network design in the study area needs significant improvement to convert it into an optimum network for the river basin located in Korea.

1.6 Scope of the Research

The SPARROW model has been applied for the assessment of nutrients (Alexander et al., 2002; McMahon et al., 2003; Preston and Brakebill, 1999; Smith et al., 1997). However, so far SPARROW model has not been applied to predict *E. coli* flux in river basins. Pathogenic contaminant assessment is subjected to large uncertainties because of scarcity of monitored data. Uncertainty in mean annual flux estimation and selection of monitoring stations for model input can result for errors in SPARROW prediction. The issues related to the effects of uncertainty in inputs on the SPARROW application should be clearly addressed before applying this model in TMDL development and implementation processes. SPARROW can also be explored to design monitoring networks by observing the improvement in its predictions with simulation of different sampling locations and time gaps between two consecutive measurements (Smith et al., 1997). An optimum design of monitoring stations can also be selected from the existing monitoring network based on accuracy and complexity of SPARROW by integrating it with a GA.

1.7 Objectives of the Research

The overall objective was to quantify the uncertainties in the prediction of *E. coli* flux using the SPARROW model.

The specific objectives of the proposed study were to:

1. Study the effect of monitoring station selection on the prediction of *E. coli* flux in the Guadalupe and San Antonio River Basins using SPARROW on the basis of statistical indices.
2. Optimize the number and location of monitoring stations in the Guadalupe and San Antonio River Basins using a genetic algorithm and SPARROW.

Each objective is discussed in the next two chapters followed by a chapter to present the conclusions of this study.

CHAPTER II

PREDICTING THE FATE AND TRANSPORT OF *E. coli* IN TWO TEXAS RIVER BASINS USING A SPATIALLY REFERENCED REGRESSION MODEL

2.1 Introduction

Section 303(d) of the Clean Water Act (CWA) requires each state to list water bodies impaired for their designated uses (public water supplies, aquifer protection, and propagation of fish, shellfish and wildlife) and to develop pollutant specific Total Maximum Daily Loads (TMDL) for watershed protection. Texas Commission on Environmental Quality (TCEQ, 2006a) has reported pathogens to be the major cause for contamination in Texas water bodies. Animals (wildlife, pets and cattle), storm-water runoff, combined sewer and sanitary sewer overflows and wastewater treatment plant effluents are some of the main sources of pathogenic microorganisms.

Water quality models are frequently used to predict the distribution of pathogens in unmonitored stream networks because only a limited number of monitoring stations can be established in a basin because of the monitoring cost. These models have been widely used in the TMDL development and implementation processes. A good water quality model should be comprehensive in describing the hydrological processes, and should be accurate in predicting the contaminant load. It should capture the spatial and temporal variability of watershed attributes and should include the affect of various environmental factors relevant to the scale of processes being modeled (Letcher et al., 2004). The spatial features of the hydrological systems can be incorporated in the models using Geographic Information Systems (GIS) techniques. For example, models, such as Soil and Water Assessment Tool (SWAT), Chemicals, Runoff and Erosion from Agricultural Management Systems (CREAMS) and Agricultural NonPoint Source (AGNPS) pollution model, etc., use complex mechanistic relationships within GIS-based frameworks for predicting the spatial and temporal distribution and loadings of contamination from point and nonpoint pollution sources (Lo and Yeung, 2002). In comparison to these traditional purely mechanistic water quality models, Spatially

Referenced Nonlinear Regression Model on Watershed Attributes (SPARROW) is a regional-level GIS-based water quality model that overcomes the limitations of using overly complex mechanistic relationships when data quantity and quality are compromised. It uses a hybrid statistical and process-based approach for predicting fluxes and sources of contaminants in a river basin (Alexander et al., 2002). SPARROW identifies every stream-reach as a basic unit to spatially distribute the contaminant sources, delivery, and attenuation factors. It is based on a least-square fitting of nonlinear relationships between the dependant variable (mean annual flux of a contaminant) and various explanatory spatial variables (such as land use, population, fertilizer application, precipitation and soil permeability).

The mean annual flux of each monitoring station is first estimated from monitored data using a rating curve model, such as those implemented by the regression tools LOADEST and FLUXMASTER (Alexander et al., 2002; McMahon et al., 2003; Schwarz et al., 1997). The mean annual fluxes are then used to estimate the parameters of the SPARROW model. Therefore, the accuracy of the SPARROW model predictions relies on the accuracy of the rating curve model's predictions of the mean annual fluxes (Moore et al., 2004). Large errors in determination of mean annual flux for the contaminants due to short time periods and irregularly observed records can affect the overall performance of the SPARROW model. Apart from the accuracies in estimated mean annual fluxes at individual monitoring stations, the total number of monitoring stations selected to fit a SPARROW model also affects the ability of the model to detect the effect of various explanatory factors on the stream loads. Unfortunately, *E. coli* monitoring is not frequently carried out and estimation of mean annual *E. coli* flux may result in large predictive errors. Further, the monitoring data may also be affected by uncertainty in sample collection, uncertainty due to handling and laboratory analysis of the contaminant, and uncertainty in streamflow measurements (Harmel et al., 2006). This uncertainty accompanied by scarcity in monitoring data can pose limitations in the applicability of SPARROW in estimating loads and sources of *E. coli* in river basins. In

this research, SPARROW is explored as a useful tool to estimate the ‘most statistically significant’ sources based on the available quantity and quality of *E. coli* data, without delving into overly complex traditional water quality models.

The Guadalupe and San Antonio River Basins in Texas have been going subjected to severe changes of land use due to increase in the population and industrialization over several decades. Recently, many waterbodies in this region have been enlisted in 303(d) list for pathogen contamination. The objectives of this research were to assess the pathogenic contamination in the area by applying the SPARROW model and to analyze the impact of monitoring station selection on the model prediction. Model results of three sets of monitoring stations selected based on the standard error in the mean annual flux estimation in FLUXMASTER, were compared. The final model was selected as the most accurate by comparing the statistical indices. The selected model was described in detail. To the best of our knowledge, this is the first application of SPARROW to predict *E. coli* fluxes in a river basin.

2.2 Methodology

2.2.1 Model Details

In SPARROW, simple empirical relations based on understanding of the contaminant processes can be defined for a watershed as shown in Figure 2.1. Monitoring station A is located upstream to B which is the end node of *i* stream. All the streams (*j*) below monitoring station A which contribute to B from the set of streams *J*(*i*). The mathematical equation for SPARROW application can be defined (McMahon et al., 2003) as:

$$L_i = \sum_{n=1}^N \sum_{j \in J(i)} \beta_n S_{n,j} e^{(-\alpha Z_j)} H_{i,j}^S H_{i,j}^R \varepsilon_i \quad (2.1)$$

where L_i is the contaminant concentration or flux in reach *i*; N is the total number of sources; $J(i)$ are the upstream reach segments and *i* stream except the streams contributing to the upstream monitoring station (A); β_n is the source coefficient for

source n ; $S_{n,j}$ is the contaminant from source n in drainage to reach j ; α is the estimated land delivery coefficient from the watershed attribute Z_j (e.g., permeability, precipitation etc.); $H_{i,j}^s$ is the fraction of contaminant present in waterbody j that is transported to waterbody i from streams; $H_{i,j}^R$ is the fraction of the contaminant present in the waterbody j that is transported to waterbody i through lakes and reservoirs and ε_i is the multiplicative error assumed to have a Gaussian distribution across separate sub-basins defined by the intervening drainage areas between monitoring stations. The term $\beta_n S_{n,j}$ quantifies the production of contaminant from a source n in the watershed of waterbody j . The exponential factor of the equation is the contaminant delivered through Z factor in waterbody j from source n . The exponential term is equal to unity for point sources which enter without any land-water delivery processes to the waterbody j . For the upstream monitored fluxes also, the land-water delivery factor is unity. Thus, the load leaving a monitored stream is the sum of the load transported from the upstream network and the load generated and delivered from the incremental watershed of that stream. The L_i (mean annual flux) is determined for every monitoring station from the daily streamflow and periodically monitored water quality data with the application of a detrending model FLUXMASTER.

The fraction of the contaminant transported to waterbody i after stream decay is a function of stream channel properties and can be quantified as:

$$H_{i,j}^s = \prod_m \exp(-k_{sm} L_{i,j,m}) \quad (2.2)$$

where k_{sm} is the first order loss coefficient, m is the number of discrete flow classes and $L_{i,j,m}$ is the length of the stream channel between waterbodies j and i in flow class m . The relation is based on a theoretical mass balance to describe the contaminant transport in streams. The loss coefficients are the estimation of contaminant losses per unit stream length of stream sizes, categorized according to the mean annual flow of the streams. In general, the contaminant loss depends on its contact and exchange with benthic

sediments. The loss decreases with increase in flow-depth, leading to low rate of contamination loss or storage for large sized streams.

The fraction of the contaminant in waterbody j delivered to waterbody i as a function of lakes and reservoirs is calculated as:

$$H_{i,j}^R = \prod_m \exp(-k_r q_{i,j,l}^{-1}) \quad (2.3)$$

where k_r is an estimated first order loss rate (apparent settling velocity). $q_{i,j,l}^{-1}$ is the reciprocal of areal hydraulic load (the ratio of outflow discharge to the water surface area of lakes and reservoirs) and l denotes lakes and reservoirs located between the waterbodies j and i. Apparent settling velocity is the function of areal hydraulic load and measures the contaminant removal or addition along with the water displacement or velocity in the reservoir.

The model residuals are the difference between the observed and predicted contaminant fluxes of the selected set of monitoring stations. The model calibration is based on a least square estimation (minimum sum of the square of residuals). The residuals are checked that they satisfy regression assumptions of independent and identically distributed residuals. The statistical significance of explanatory variables is evaluated according to the standard *t* test statistics which is based on the mean and standard error of the estimated coefficients. The flowchart shown in Figure 2.2 represents the SPARROW application as an integrated model of a flux estimation model (FLUXMASTER) and the attributes of streams and sources.

2.2.2 Selection of the Best Model

The accuracy of the regression based models can be estimated by comparing the R^2 and mean of the square errors if number of observations (monitoring stations) and estimated parameters are same. But to compare the models with different selections of monitoring stations R^2 and mean of the square errors are not enough to select the model which can

accurately describe the contaminant fate and transport. Based on parsimony, complexity, and the efficiency of the model, various selection criteria are used to select the best model. Some of these criteria are discussed below.

The Akaike Information Criterion (AIC) is used to compare different regression models based on their model complexity (more model parameters) or accuracy (minimum error term). AIC is defined for a model with n number of observations and p parameters as (Rasch, 1995):

$$\text{AIC} = n \left[\ln \left(\frac{(n-p)\sigma^2}{n} \right) \right] + \frac{n(n+p)}{n-p-2} \quad (2.4)$$

where σ^2 is the variance of normally distributed residuals. The AIC is always a positive number and the minimum value is desired for the best model.

Nash- Sutcliffe Efficiency (NSE) is applied for indicating the variation of residuals with respect to the deviation of observed data from their mean. The coefficient can vary from $-\infty$ to 1. The mathematical equation is given as (Moriassi et al., 2007):

$$\text{NSE} = 1 - \frac{\sum_{i=1}^n (Y_i^{obs} - Y_i^{pred})^2}{\sum_{i=1}^n (Y_i^{obs} - Y_{mean}^{obs})^2} \quad (2.5)$$

where Y_i^{obs} is i^{th} value of observed variable from FLUXMASTER, Y_i^{pred} is i^{th} value of the predicted variable from SPARROW and Y_{mean}^{obs} is the mean of the observed values. For a model, the value of NSE is desired to be 1 that implies residual variance is 0. NSE less than 0 indicates that the mean model (Y_i^{pred} as a function of only Y_{mean}^{obs}) would be just as good as the predicting model.

Percent bias (PBIAS) measures the bias of a model towards over (positive) or under (negative) estimation.

$$\text{PBIAS} = \frac{\sum_{i=1}^n (Y_i^{obs} - Y_i^{pred})}{\sum_{i=1}^n Y_i^{obs}} \times 100 \quad (2.6)$$

These indices are useful to select the most efficient and accurate model with no or minimum bias.

2.2.3 Study Area

In this study, the SPARROW model was applied to assess water contamination due to *E. coli*, an indicator of pathogenic contamination, in the Guadalupe and San Antonio River Basins of Texas. The spatial extent of the study area (29380 km²) is from longitude 30°18'44"N to 28°22'2"S and latitude 99°42'31"W to 96°47'10"E. The study area includes a major metropolitan area (San Antonio), an unconfined aquifer (Edwards Aquifer) which is the main source of water supply to San Antonio and forest and pasture as major land uses (55.4% and 28.0% of total land, respectively) (Figure 2.3).

2.2.4 Data Sources

Water Quality Data

The mean annual fluxes for monitoring stations are calculated by taking averages of estimated daily contaminant loads predicted by the Statistical Analysis Software (SAS) based FLUXMASTER model. FLUXMASTER uses continuous daily observed streamflow data and discontinuous water quality concentration data to predict a nonlinear regression curve that can be, in turn, used to predict continuous daily contaminant loads. The monitored data of pathogens (*E. coli* and fecal coliform) were taken from the Guadalupe Blanco River Authority (GBRA, 2007) and San Antonio River Authority (SARA, 2007). For some monitoring stations only fecal coliform or both fecal coliform and *E. coli* records were available for a few years. The records that had observations for both *E. coli* and fecal coliform were used to fit a linear regression equation that provided relationship between fecal coliform concentrations and *E. coli* concentrations, with a coefficient of determination (R^2) equal to 0.9. The concentration records of fecal coliform were converted to equivalent *E. coli* concentration (CFU/100ml) based on the following estimated regression relation:

$$E. coli = 0.49 \times \text{fecal coliform} + 30.91 \quad (2.7)$$

Pathogen records available from 49 GBRA stations and 33 SARA water quality monitoring stations were used for this study. The locations of the monitoring stations, based on longitudes and latitudes, on the stream network were obtained from GBRA and SARA data sources. These locations were verified and corrected from road and geographical maps of counties. Topography and Google maps were also useful to determine the exact locations of monitoring stations. The daily streamflow data were obtained from U.S. Geological Survey (USGS) monitoring stations. The locations of the USGS monitoring stations were available from National Hydrography Dataset (NHD) Plus (described in next section). Every water quality monitoring station was associated with a USGS monitoring station based on the ratio of drainage areas of the two gages, which ranged from 0.75 to 1.25.

The seasonal and other biases were removed by detrending the flow to a common base year 2005. The model with all possible coefficients that relate measured contaminant concentration to streamflow and temporal variations is given as (Smith et al., 1997):

$$\ln(l) = \beta_0 + \beta_1 t + \beta_2 \sin(2\pi t) + \beta_3 \cos(2\pi t) + \beta_4 \ln(q) + \beta_5 [\ln(q)]^2 + \theta \quad (2.8)$$

where l is the instantaneous concentration of the contaminant, t is decimal time, q is instantaneous discharge and $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ and β_5 are the regression coefficients of intercept, seasonal trend and natural-logarithmic terms. These coefficients can be zero or nonzero based on the estimated trend of a monitoring station. The term θ is the sampling and model error assumed to be independent and identically distributed. The estimates of mean annual flux of the base year corresponded to the average conditions over the period 1987- 2007. A total of 16 detrending models were applied based on the different combinations of seasonal and lag trends of water quality and streamflow records. For every monitoring station, the model with the minimum sum of the squares of the error was selected. Since a short time period of the monitoring record may result in large standard errors, the minimum time period for a monitoring record was fixed to two years. The mean annual flux (L) was calculated using the following equation:

$$L = \frac{1}{365} \sum_{i=1}^{365} \exp[\beta_0 + \beta_1 t_i + \beta_2 \sin(2\pi t_i) + \beta_3 \cos(2\pi t_i) + \beta_4 \ln(q_i) + \beta_5 [\ln(q_i)]^2] V_f \quad (2.9)$$

where q_i is average of the daily stream flow values for the i th day of the years over the 1987-2007 period, t_i is decimal time for the i th day of the base year, and V_f is the minimum variance bias correction factor. Variance bias correction factor is applied to remove the collinearity in the equation.

When applying SPARROW model in a national water quality study, the maximum ratio of standard error (SE) to mean annual flux of contaminant in LOADEST (a detrending model similar to FLUXMASTER) was limited to 0.2 for selecting the monitoring stations (Smith et al., 1997). In the current study, due to short time period of records and the insufficient number of observations for pathogen concentration the ratios were observed to be greater than 0.2 for most of monitoring stations. To explore the effect of variability in monitoring station locations on model complexity and accuracy, three sets of monitoring stations with 56, 35 and 21 monitoring stations were chosen based on varying values of the maximum permissible SE to flux ratio as 10, 1 and 0.6, respectively (Figure 2.4).

Watershed Attributes

In our study, NHD Plus, available from U.S. Environmental Protection Agency (USEPA, 2005) was used to obtain watershed spatial characteristics. NHD Plus is a combination of four datasets: NHD, National Elevation Dataset (NED), National Watershed Boundary Dataset (WBD) and National Land Cover Dataset (NLCD). NHD is available at 1:100,000 scale. There are some database files with value added attributes (VAA) related to reaches and catchments in this dataset. All the files and layers are connected with a primary key (USEPA, 2005).

The VAA include the details of streams (slope, length, velocity and flow), watersheds

(area and percentage land use) and climate (mean annual temperature and precipitation). Because of various available features such as stream level, link-node traversal, hydrologic sequence, terminal identifier and level path identifier, this database can be easily explored and utilized. The linking and corrections of stream network geometry and flow direction have made watershed delineation, characterization and the complex model simulations considerably easier than the original NHD. The study areas (Guadalupe and San Antonio River Basins) used in this study lie in the 12c hydrologic region of NHD Plus dataset, and has 5167 watersheds (Mean area = 5.42 km² ranging from 0.001 to 99.69 km²).

Land Uses as Sources

The 1992 NLCD data, included in NHD Plus, was replaced with 2001 land use data, available from Multi Resolution Land Characteristics (MRLC) Consortium (USGS, 2001). In this study, the forest area, equal to 16288 km² (55.4% of the study area), was defined as the sum of deciduous forest, evergreen forest, mixed forest, dwarf scrub and shrub/scrub land classes. The pasture land included the land classes: pasture/ hay and grasslands/ herbaceous, and covered an area of 8241 km² (28.0% of the study area). The wetland area (woody wetlands and emergent herbaceous wetlands) in the region equaled to 853 km² (2.9% of the study area). The urban land use included the following land use classes based on the urban development intensity- open space (LU21), low intensity (LU22), medium intensity (LU23) and high intensity (LU24). The sum of all urban land uses was 2334 km² (7.8% of the study area). The agricultural land use (1438 km²; 4.9% of the study area) comprised of the spatial area used for cultivated crops. The Figure 2.5 shows the land distribution in the Guadalupe and San Antonio River Basins.

Other Sources

Fecal material excreta of warm blooded animals and effluents from wastewater treatment plants are the major sources of *E. coli*. The Census of Agriculture from the National Agricultural Statistical Survey (NASS) (USDA, 2002) provides the number of cattle for

every county of Texas. The Guadalupe and San Antonio River Basins include 25 counties. The total population of cattle (711,330) was spatially distributed on the pasture lands of the study area. The human population (U.S. Bureau of census, 2000) was spatially distributed on the urban areas of the watersheds in the study area. The effluent discharge (USEPA, 2007) from wastewater treatment plants and their spatial locations (TCEQ, 2007) were also included in the model as *E. coli* sources. A total of 73 wastewater outfalls exist in the study area and the effluent discharge varied from 0.0125 to 83 million gallons per day (47.35 to 314,190 m³ per day). Since data for pathogen concentration in the effluent discharges was not available, the model parameter $S_{n,j}$ (Equation 2.1) for wastewater point source was defined as discharge (m³yr⁻¹) instead of *E. coli* load (CFU per unit year) in our research.

Land-Water Delivery Factors

The important factors for land-water delivery of *E. coli* are climatic factors (precipitation and temperature), soil characteristics (permeability, drainage and soil type), reach slope and drainage density. In NHD Plus, mean annual temperature and precipitation (from 1961 to 1990) is derived from the PRISM (Oregon State University) as the attributes of watersheds. The mean annual temperature and precipitation of all the watersheds varied from 17.15°C to 21.57°C and 694.86 to 1004.97 mm, respectively. Soil characteristics are available in the STATE Soil GeOgraphic database (STATSGO) at resolution 1:250,000. The processed STATSGO soil data was taken from Pennsylvania State University's Center for Environmental Informatics database (C.E.I., 2007). Since surface water contamination is affected by the permeability of only the top layer, the average soil permeability of only the two top layers, i.e., 0-5 cm and 5-10 cm was included in the model. In MRLC, different subclasses of the urban land use are defined based on the percent impermeable area. The impermeable area of every watershed was calculated as the average impermeability for these land use classes. This impermeable area was reduced from the total watershed area to estimate the average soil permeability for the watershed. The permeability ranged from 0.00 to 30.23 cm h⁻¹ for different watersheds in

the study area. Drainage density ranging from 0.01 to 54.00 km⁻¹ was calculated as the ratio of the length of stream to incremental watershed area. The stream slope was obtained from NHD plus and varied from 0.00 to 0.15 for the study area.

Streams and Reservoirs Attenuation Factors

There were 1188 waterbodies located as reservoirs, lakes and soil conservation sites. To account for the reservoir attenuation factor, areal hydraulic load (for the study area, range 0.0 to 9818.4 m yr⁻¹) was used. The stream/ reach decay factor depends on the streamflow. The flow varied for this study area from 4.2 x 10⁻⁶ to 124.73 m³ s⁻¹. The streams were divided into three classes namely small (flow less than or equal to 0.02 m³ s⁻¹), medium (0.02 to 0.13 m³ s⁻¹) and large (greater than 0.13 m³ s⁻¹) based on a quantile distribution of all the streamflow values in the region. Reach travel time was calculated as the ratio of length and mean annual velocity of the stream, and its value ranged from 2.92x10⁻⁴ to 0.92 days.

Geographic Information System (GIS) was used to spatially distribute all land uses, contaminant sources, land-water delivery and attenuation factors at the watershed scale. The attributes were connected with streams using the common identification number of a stream. The response variable (mean annual fluxes at the monitoring stations) was linked with the corresponding stream on which the monitoring station was located. A SAS input data file of all these details was used to run the SPARROW model.

2.3 Results and Discussion

The SPARROW model was applied on three sets of water quality monitoring stations 56, 35 and 21, selected based on the standard error to mean annual flux ratio from FLUXMASTER results. The significance of the calibrated model parameters is a function of number of monitoring stations, as well as the amount of variability and collinearity in the explanatory factors. Since the human population and urban land were collinear variables, only one out of these was included at a time during model

calibration. Similarly only one of the possible indicators of livestock sources, i.e., cattle population and pasture land, was selected for the model calibration.

2.3.1 Results of Three Prediction Models

The Table 2.1 shows the coefficients and p-value (in parenthesis) of the parameters and statistics of three fitted models (I, II and III) based on monitoring data from 56, 35 and 21 monitoring stations respectively. A p-value is the probability that the null model could, by random chance, produce a coefficient value as extreme as or more extreme than the observed value. The value shows the statistical significance level of the estimated coefficient and a low p-value is desired as evidence against the null hypothesis (Weisberg, 2005). Point sources contributing to *E. coli* contamination were included in the first two models (Models I and II), but with only moderate statistical significance (due to moderately high p-values). Model I and II had monitoring stations close to these point sources with high permitted flows, but was not able to detect any trend between point source permitted flows and mean annual fluxes. This can be attributed to two possible reasons: (1) using permitted flows instead of the actual concentrations of *E. coli* in flows from wastewater treatment plants, because the concentration data was not available. (2) there are large number of point sources spatially distributed throughout the study area, discharging relatively low flows (Figure 2.6). Point sources contributing to *E. coli* contamination were excluded from Model III. This occurred because monitoring stations located downstream from point sources with high permitted flows were excluded from the final set of monitoring stations used by Model III (Figure 2.4 and Figure 2.6), due to the more stringent selection criteria based on the standard error to mean annual flux ratio. This led to the model not being able to detect the effect of loadings from crucial urban point sources in the San Antonio area and south-east region near the Gulf of Mexico.

Two important land uses, urban and pasture appeared as statistically significant contributors of *E. coli* in Model I, while in the Model II, influence of these sources

decreased considerably. Model coefficients values for sources changed from 5.57 (Model I) to 2.22 (Model II) for urban areas and 20.58 (Model I) to 9.30 (Model II) for pasture areas, when the number and locations of monitoring stations were changed. In the Model III, any factor (number of cattle or pasture land) related to livestock contribution was also not included. The forest land use was not a source of *E. coli* in the Model I, but was a highly significant source in the Model III ($p = 0.06$). The urban land use was also included as a significant source in Model III ($p = 0.12$), though the coefficient related to sources from this land use had the lowest value of 0.94. These different levels of significances and exclusions of various nonpoint sources of *E. coli* can be contributed to the locations and number of the monitoring stations and the differences in mean annual fluxes of included stations used in the calibration of the SPARROW model. It should be noted that the manure-applied agricultural lands were not included as *E. coli* contributing sources. This is mainly because there is no information available about such land uses in the study area. Most of the monitoring stations used for Model III are located in the North of the study area where forest is a major land use (Figure 2.4). This could have caused the Model III to detect the significant effect of sources related to the forest land use and not detect any significant effect of the sources related to pasture land use where there are hardly any monitoring stations. Also, though the Model III used fewer number of monitoring stations located in the urban land use, the high mean annual fluxes monitored at these monitoring stations led to the detection of the urban land use as a significant source of *E. coli*.

Rainfall, a land-water delivery factor entered only in the Model I and III but with only moderate significance. The rainfall was assumed to affect the land-water delivery positively by increasing the storm flow and decreasing the travel time. The rainfall might not be a significant factor due to inaccuracy or lack of spatial variability in the dataset at the model application scale. However, all the models included temperature as a highly significant delivery factor (Table 2.1). Among the stream attenuation factors, only the coefficient for medium-sized streams entered in all the three models significantly. This

could be because of the ideal combination of long travel time and more benthic contact for water in the medium-sized streams, which provides favorable conditions for the decay of *E. coli*.

Model III with only 21 stations (standard error to flux ratio less than or equal to 0.6) explained maximum ($R^2 = 0.85$) variability in mean annual flux due to the source, land-water delivery and stream/ reservoir attenuation factors. In the earlier applications of SPARROW model, R^2 varied from 0.88 to 0.97 (Alexander et al., 2002; McMahon et al., 2003; Smith et al., 1997). The better explanation of variability in these studies might be due to high density of monitoring stations in the watersheds and long records of monitored water quality data available for nitrogen and phosphorus. In Table 2.2 PBIAS was observed to be positive for all the three models indicating the overestimation of the mean annual *E. coli* flux in the model predictions. Among all three models, NSE was the highest and PBIAS and AIC were the lowest for the Model III. Considering all the model selection criteria (AIC, NSE, PBIAS and model output statistics: SSE, MSE, RMSE and R^2), Model III was selected as the most appropriate model for predicting the *E. coli* flux and concentrations in the study area (Table 2.1 and Table 2.2). For the Model III, the Equation 2.1 can be written as:

$$Li = \sum_{n=1}^N \sum_{j \in J(i)} (0.73S_{\text{Forest}} + 0.94S_{\text{Urban}}) \exp(-4.54Z_{\text{Rainfall}} - 2.41Z_{\text{Temperature}} + 0.08Z_{\text{Permeability}}) \prod \exp(-24.58L_{i,j,2}) \prod \exp(19.81q_{i,j,1}^{-1}) \epsilon_i \quad (2.10)$$

2.3.2 Coefficients Estimation of the Best Model

In Model III predictions, the p-values of areal hydraulic load, permeability and rainfall are quite large. This implies that there are high probabilities of the null hypothesis, parameter coefficients = 0 of areal hydraulic load, permeability and rainfall, being true for these variables. These factors were removed from the Model III and the model was recalibrated based on only the remaining variables. Uncertainty assessments for the parameters of the final recalibrated model were made via bootstrapping. Bootstrapping is

a data-based statistical inference method used for estimating the sampling distribution of an estimator by sampling from the original samples of observations. It can be used to estimate robust estimates of standard errors and confidence intervals of various population parameters, such as the mean, median, correlation coefficient, regression coefficient, etc. (Fox, 2002). In our study, 250 bootstrap samples were used to estimate the 90% confidence interval of coefficients with significance level at 15% level in parametric regression. Table 2.3 shows the results of bootstrap analysis. In bootstrap analysis, the multiple sets of coefficient estimates are generated either from the resampled data (nonparametric bootstrap) or from the normally-distributed randomly generated coefficient vectors (parametric bootstrap) (Fox, 2002). In Table 2.3, the average values of parameters for parametric bootstrap are similar to the parametric coefficients estimated by recalibrating the SPARROW model on the basis of only the most significant parameters in Model III. The average values estimated by nonparametric bootstrap are, however, larger than those estimated via recalibration (Table 2.3). The nonparametric bootstrap technique is, however, more suitable for robust estimation of uncertainty in parameters, especially when small size datasets require the need to reject the assumption of normality (Fox, 2002).

2.3.3 Prediction of *E. coli* Flux Using Model III

Figures 2.7 and 2.8 show the incremental and delivery *E. coli* yield in the study area. The Figure 2.9 shows the relationship between natural logarithm of estimated mean annual flux of *E. coli* using FLUXMASTER (Observed flux, input for SPARROW) and natural logarithm of the predicted flux using the Model III. Though positive PBIAS (Table 2.2) reflected that all the SPARROW models have overestimated the flux, in Figure 2.9 the selected model underestimated for the monitoring stations with large *E. coli* flux values since the observed *E. coli* fluxes (estimated by FLUXMASTER) are less than the predicted fluxes (predicted by the final SPARROW Model III). The R^2 of the regression line for the estimates in Figure 2.9 is 0.85 (Table 2.2).

In Figure 2.10, the logarithmic residual values of the monitoring stations are shown at their locations in sub-basins. The logarithmic residuals indicate the over and under prediction of the regions in study area. The negative values indicate the over prediction whereas positive values of residuals shows the under prediction. So, the spatial trend of predictions can be observed. It is clear that the monitoring stations located in San Antonio River Basin and downstream to the San Antonio metropolitan area show the trend towards under prediction. These are the areas with large *E. coli* fluxes (Figure 2.7). The *E. coli* fluxes of these monitoring stations have been underestimated in Figure 2.9. The underestimation of flux may be due to the unaccountability of point sources or underestimation of the influence of urban area on regional water quality. In the Guadalupe River Basin the model predictions were both under and overestimated.

Incremental and Delivered E. coli Yields

Incremental *E. coli* yield using the selected SPARROW model is defined as the amount of *E. coli* generated locally in a watershed independent of upstream load. Delivered yield is the share of the total *E. coli* yield of a watershed that will reach the Gulf of Mexico (the end of stream network) after in-stream and reservoir losses. Incremental yield is an indicator of pollution production in a watershed whereas delivered yield gives the idea of how much contaminant sources located in a particular watershed can influence the water quality of the region. Figure 2.7 shows that the south-west region of the study area has the highest incremental yield of *E. coli* compared to other spatial regions. Some of these watersheds also delivered large share of incremental yield of *E. coli* to downstream watersheds (Figure 2.8). The locations of the highly contaminated watersheds predicted by the model are closer to the downstream end of the river network (Figure 2.8). So, there is less opportunity for decay of *E. coli* before reaching the shorelines in this region. This will ultimately result in contaminated shores.

Land Use Contribution of E. coli Contamination

Percent contribution of *E. coli* delivered from the two significant sources identified by

Model III, urban and forest land uses, helps us compare the spatial distribution of *E. coli* load resulting from these two sources in river basins. Figure 2.11 shows the *E. coli* load contribution from urban and forest land uses in the subbasins of the study area. The urban land use (Figure 2.5) is a major contributor of pathogens only in the upper San Antonio River subbasin ('12100301' subbasin; Figure 2.9). The lower southwest section of the San Antonio river subbasin, which has diverse land uses (Figure 2.5), has been predicted as the region with high incremental yield of *E. coli* (Figure 2.7). The major land use affecting the water quality in the southern subbasins is forest land use. In spite of low density forested areas in these subbasins (Figure 2.5), *E. coli* loadings per unit area is relatively high (Figure 2.11). Results from this study clearly show that source-specific best management practices (BMPs) should be implemented at a subbasin-level to address the *E. coli* contamination in Guadalupe and San Antonio River Basins.

Impaired Streams due to Pathogens

According to Texas water quality standards (TCEQ, 2000a), the geometric mean concentration for recreational use of water is 126 colonies per 100 ml (with physical contact) and 605 colonies per 100 ml (without contact). Figure 2.12 shows the predicted *E. coli* concentration using Model III for major streams with flow greater than $0.13 \text{ m}^3 \text{ s}^{-1}$. A vast majority of streams especially in the south-west of San Antonio River Basin have the concentrations above the *E. coli* standards. These streams should be carefully monitored for the impairment due to *E. coli*. The spatial location of impaired streams of Guadalupe and San Antonio River Basins have also been listed in the 303(d) list provided by TCEQ (2000b). In their monitoring process (TCEQ, 2000b) approximately 2617.6 km stream length was observed for the water quality violations and 1143.1 km stream length was found to be impaired (Figure 2.13). The 67% of these impaired monitored streams were located in San Antonio River Basin alone, and about 82.3% of the impaired streams in San Antonio River Basin were contaminated by high concentration of pathogens. Overall in Guadalupe and San Antonio River Basins, 72% (824.3 km of stream length) of the impaired streams were listed due to pathogen

contamination. Contamination of streams predicted by SPARROW model (Figure 2.12) and listed by TCEQ (Figure 2.13) were compared qualitatively and it was found that almost all of the streams listed in the 303(d) list for impairment due to high pathogen concentrations have also been successfully detected by Model III as impaired streams.

2.4 Conclusions

In this study, we have demonstrated the advantages of using spatially referenced statistical relationships along with parsimonious mechanistic relationships to simulate fate and transport of *E. coli* in river basins and identify the major sources of pathogen contamination. Without delving into complex mechanistic water quality models, the SPARROW model effectively simulated the incremental yield and delivery of *E. coli* in these river basins. The final selected model was able to explain the variability in mean annual flux due to the different sources, land-water delivery factors and stream/ reservoir attenuation factors with a R^2 of 0.85. The major sources of *E. coli* contamination identified in these river basins forest and urban land use, which implies that the BMPs for the protection of watersheds from pathogens should focus on the sources specific to these land uses. With the application of SPARROW, major contributing sources at watersheds or subbasins level can be identified for the implementations of BMPs.

Since point sources were not included as a significant source in any of the final models, it can be concluded that the available scale and details of the explanatory variables affect their statistical significance in the SPARROW model. The lack of long historical records of monitored *E. coli* in the Guadalupe and San Antonio River Basins resulted in large standard error in mean annual flux estimation in this application of SPARROW. In spite of the challenges posed by data scarcity and details the selected model has successfully identified almost all of the 824.3 km of stream length listed in the 303(d) list for impairment by pathogens. Thus, SPARROW model can be used as a prediction tool to identify impaired streams due to bacteria in river basins.

In our research, the effects of number and the locations of monitoring stations on the SPARROW model accuracy and complexity were also analyzed. The selection of monitoring stations in SPARROW is very critical to include the important factors affecting the regional water quality. The criterion used for selecting the most appropriate set of monitoring stations, on one hand, ensured that more accurate rating curve models were used to estimate the mean annual flux inputs for SPARROW; however, on the other hand, it precluded many critical monitoring stations located at the area with high concentration of *E. coli* from the Model III. Due to insufficient representation of highly contaminated regions in the study area, the Model III underestimated *E. coli* flux for monitoring stations with the large values of mean annual flux (Figure 2.9). The biased predictions will further affect the outcomes of the TMDLs and Watershed Protection Plans (WPP). So, the final selection of monitoring stations should be made carefully especially for a relatively small study area with limited number of monitoring stations and for the contaminants with the limited monitoring records.

CHAPTER III
OPTIMIZATION OF WATER QUALITY MONITORING NETWORK TO
MONITOR *E. coli* USING A SPATIALLY REFERENCED WATER QUALITY
MODEL AND GENETIC ALGORITHM

3.1 Introduction

The monitoring network for a river system is designed to provide the information of water quantity and quality. The water-quality monitoring stations are located at critical locations for the surveillance of waterbodies from pollution sources. Water resources utility and management programs such as Total Maximum Daily Loads (TMDL) development and Watershed Protection Plans (WPP) also require the systematic monitoring of waterbodies (Park et al., 2006). The spatial and temporal characteristics of water quality determine the locations of monitoring stations in a watershed (USEPA, 2002). To design a water quality monitoring network, the considerations of water uses and contamination levels are important. Implementation of TMDLs requires the assessment of contaminant loads from point and nonpoint sources and possible reduction in the load from these sources being delivered to streams. The monitoring network is subjected to objectives and constraints related to the cost of monitoring and trends of regional water quality (USEPA, 2002). Thus an optimization process is required to design a monitoring network.

Genetic Algorithm (GA) is a heuristic optimization technique based on Darwin's theory of natural selection (Holland 1979). GA is a robust technique to obtain near-optimal solutions in the decision space beginning with a randomly-chosen initial solution set. The solution space is explored and exploited by applying genetic operators such as crossover, mutation and selection methods. For water quality modeling applications, GA has been applied to calibrate and perform sensitivity analysis on models and to allocate the contaminant load with uncertainty (Savic and Khu, 2005; Srivastava et al., 2002; Yandamuri et al., 2006). Icaga (2005) compared the results of GA to a previous application of dynamic programming for reducing the number of monitoring stations in a

basin. It was observed that the performance of GA can vary widely and depends on initial population size, crossover and mutation rates. Park et al. (2006) applied a single objective GA by aggregating multiple objectives with normalized weights based on the river basin representation, water-quality standard compliances and pollution sources supervision. They found that the existing monitoring-network design required significant improvements for converting it into an optimum network. Reed et al. (2000) discussed practical methodologies to implement an efficient single objective GA to design a groundwater quality monitoring network. The application of multiobjective algorithm to minimize the cost and error in estimation of concentration of contaminants by reducing the number of groundwater monitoring stations was further explored (Reed et al., 2001).

All of the above applications have been implemented to design monitoring networks or allocate loads for nonpathogenic contaminations. Pathogenic contamination has become a dominant water quality issue in the U.S. in recent years, largely because of increasing population, failing septic system and nonpoint pollution from forests and pasture land uses. There are only limited water-quality records available for pathogens as compared to traditionally observed water quality parameters such as nutrients. A modeling approach to develop TMDLs, using the monitored water-quality data as input, is a widely accepted approach to assess the load from nonpoint and point pollution sources for pathogens. Due to the scarcity of monitored data, pathogenic load assessment is associated with large uncertainty in general (Dorner et al., 2006). For optimum performance of a water-quality model, therefore, input uncertainty in the monitored records should be minimized.

A Spatially Referenced Nonlinear Regression Model on Watershed Attributes (SPARROW) is a water quality model to predict fluxes and concentrations and to track the sources of the contaminants (McMahon et al., 2003; Smith et al., 1997). The model relates the monitored water quality records to spatially referenced contaminant sources and land-water delivery factors on the basin attributes (stream or watershed). These

factors are the regional characteristics which can affect the increment, decay and delivery of load in a stream network. Since pathogen concentrations are monitored monthly or randomly (e.g., during or after storm events), the concentration data is analyzed with daily-monitored streamflow records by applying a load-estimator model (FLUXMASTER) to estimate the mean annual flux. The mean annual fluxes of the water-quality monitoring stations serve as the observations of response variable for the SPARROW model. The flowchart in Figure 2.2 shows the SPARROW application as an integrated model of FLUXMASTER and the streams, watershed and source characteristics.

Smith et al. (1997) have recommended the application of SPARROW in monitoring network design by observing the improvement in its predictions with simulation of different sampling locations. The uncertainty in the mean annual flux estimation due to scarcity of monitored data and monitoring stations can result in large errors in SPARROW prediction. An optimum set of monitoring stations can be selected from the existing monitoring network based on an accurate and simple application of SPARROW with minimum correlations among the explanatory variables (sources, land –water delivery and stream/ reservoir attenuation factors) for the selected monitoring stations.

In our study, the optimum water quality monitoring networks were selected to assess *E. coli* loads with minimum uncertainty for two major river basins (Guadalupe and San Antonio) of Texas. A GA was applied to select the monitoring networks from the FLUXMASTER assessment of mean annual flux with adequate spatial variation. The multiple objectives of the optimization problem were defined to include the maximum number of monitoring stations with large values of mean annual flux and least uncertainty and to minimize the cost of monitoring. Constraints related to the monitoring of critical locations were included in a multiobjective optimization problem. From the optimum sets of monitoring networks, the best monitoring network was selected based

on the minimum root mean square error and the simplified description of the land–water processes in SPARROW.

In the following section (Section 3.2), SPARROW application on the Guadalupe and San Antonio River Basins to assess pathogen contamination, and the GA application, including a short description of GA parameters and optimum solutions, is given. In the last section (Section 3.3), results of GA and SPARROW applications are presented.

3.2 Methodology

3.2.1 Study Area and Data Sources

The SPARROW model was applied to assess contamination due to *E. coli*, an indicator of pathogenic contamination, in the Guadalupe and San Antonio River Basins of Texas. The spatial extent of the study area (area 29380 km²) is from longitude 30°18'44"N to 28°22'2"S and latitude 99°42'31"W to 96°47'10"E. The study area includes a metropolitan area (San Antonio), an unconfined aquifer (Edwards Aquifer) and forest and pasture as major land use (55.4 % and 28.0% of total land uses respectively). The watershed and water body's attributes such as land use, average temperature and precipitation, reach slope and velocity and reservoir area distributed on reach basis was obtained from National Hydrography Dataset (NHD) Plus (USEPA, 2005). Monitored records of *E. coli* concentrations at the stations located on the Guadalupe and San Antonio Rivers were obtained from the Guadalupe Blanco River Authority (GBRA) and San Antonio River Authority (SARA) (GBRA, 2007; SARA, 2007). The daily stream-flow data at stream gages was available from USGS (NWIS, 2007). The effluent discharge (USEPA, 2007) from wastewater treatment plants and their spatial locations (TCEQ, 2007) were included as probable *E. coli* sources. Since the concentration data of contaminants in the effluent was not available, the permitted flows from wastewater treatment plants were considered in the model. There are many point sources spatially distributed throughout the study area, discharging relatively low flows. So, only the wastewater treatment plants with discharge greater than two million gallon per day were

included. Soil permeability values were derived from State Soil Geographic Database (STATSGO) soil data (C.E.I., 2007). Since the size of reach affects the decay of pathogens, the reaches were divided into three categories; small, medium and large, on a quantile basis for the reach decay factors. Figure 3.2 shows the location of monitoring stations in the sub-basins of Guadalupe and San Antonio basins. The major streams include reaches with flow greater than $0.13\text{m}^3\text{ s}^{-1}$.

The watershed attributes along with the spatially distributed factors were associated with the corresponding streams. To reduce the effect of irregular monitoring and the short time period of records on the water quality assessment, an initial set of monitoring stations was selected on the basis of standard error to mean annual flux ratio in FLUXMASTER application. Only 56 out of 72 monitoring stations were selected and taken as inputs for GA application. Schwarz et al. (2007) recommended at least 20 monitoring stations for the application of SPARROW to include the spatial variability. The model error in the SPARROW predictions is related to the quality and scale of explanatory variables. Applying combinatorial mathematics, there are 7.86×10^{14} possibilities to select 20 stations from the 56 monitoring stations with equal probability for every monitoring station. Thus, the application of GA is justified to find the optimum sets of input for the SPARROW model.

3.2.2 Genetic Algorithm

To initialize the GA, a random set of solutions (called the population) are coded into various formats (binary, real, integers etc.) and are called chromosomes. Every chromosome consists of genes which provide information on the solution attributes. The population of solutions (or designs) in the solution space are evaluated to obtain their fitness values. The fitness value is an indicator of how good a solution is with respect to the problem's objective functions and constraints. The variation operators, mutation and recombination, are applied to create a diverse set of solutions, called children, from the existing solutions, or parents. Recombination or crossover includes the interaction

between two or more parents whereas mutation is the outcome of a random change in the chromosome of a parent. The children are tested for their fitness and selected using a specific selection operator as parents in a new generation. The mutation operator helps in exploring the new solutions from unexplored regions in solution space and the selection operation assists in ensuring an overall improvement in the mean quality of solutions in the next generation. The newly created generation is treated as parents until the current solution set converges to the best possible solution of the problem before a predefined termination condition (Eiben and Smith, 2003). Figure 3.3 represents the flowchart of a GA application.

3.2.3 Representation, Objectives and Constraints

Uncertainty in prediction of mean annual flux by regressions tools, such as FLUXMASTER, can occur due to poor data quality and quantity that arise from limitations in the monitoring process, such as unaccounted trends or irregular monitoring. To minimize the effect of the uncertainty in FLUXMASTER on the uncertainty in SPARROW, uncertainty was considered as one of the objectives to select the monitoring stations. The objectives and constraints were defined in the mathematical form for GA application. A solution was represented by a binary string of 56 genes to represent decisions at 56 locations in the stream network. The decision to include a station in the network is binary in nature where 0 implies ‘not selected’ and 1 ‘selected’. The variable Y_i represents the decision to include a monitoring station at location i ($i = [1, 56]$).

1. To minimize the cost of monitoring, the number of monitoring stations was kept as small as possible without compromising with water-quality standard. This helped in maintaining long duration records for sufficient number of monitoring stations instead of many redundant monitoring records for shorter durations.

$$\text{Min} \sum_{i=1}^n Y_i$$

2. The average of logarithmic of mean annual fluxes for all the monitoring stations was maximized to ensure that the set of monitoring stations selected for calibrating the SPARROW model included the monitoring stations from high risk areas,

$$Max \frac{\sum_{i=1}^n Y_i \log(m_i)}{\sum_{i=1}^n Y_i}$$

where m_i is the predicted mean annual flux from FLUXMASTER output.

3. To keep the uncertainty in mean annual fluxes of the monitoring stations reasonable, the standard error to flux ratio (SF_i) for the monitoring stations was minimized:

$$Min \sum_{i=1}^n SF_i Y_i$$

The following are the constraints to find the optimal solutions:

1. The total number of monitoring stations to represent the variability of water quality in the region was greater than 20.

$$\sum_{i=1}^n Y_i \geq 20$$

2. Monitoring stations located at places which have ecological or hydrological importance and require continuous attention (monitoring), must be maintained in the network- Y_i will be 1 if i represents such a location. Considering the contamination of surface water, the Edwards Aquifer contributing zone has an impact on the recharge of the aquifer and there are four monitoring stations located in this zone. In San Antonio there are six monitoring stations. At least two stations in the Edwards Aquifer region and three stations in San Antonio were retained in the final solution. For this a penalty was imposed by making the value of the third objective impossibly large if a set of monitoring stations violated this constraint.

Since SPARROW is a regression based model, its results are affected by exceptionally high or low values of mean annual flux. The outliers can result in inaccurate understanding of the watershed processes and can result in ignoring actual sources of contaminants in the model results. The root mean square error of the SPARROW application for a set of monitoring stations is desired to be the minimized to predict water quality. The SPARROW model was applied to the optimum solutions obtained after the application of GA. The objectives and constraints for selecting the optimum network for SPARROW application are shown in Figure 3.4.

The application of a GA involves the selection of parameters for genetic operations to obtain the optimal solution efficiently (Grefenstette, 1986). To solve a multiobjective problem, two different strategies can be applied either by aggregating the weighted objectives to form a single objective problem or by finding the multiple solutions on a Pareto front to generate the best alternatives. The first method provides the leverage to solve the problem as a single objective, but assigning weights can be challenging for most problems. The second method requires solving the problem for all the objectives to obtain the non-dominated or Pareto set of solutions. Non-dominated sets of solutions consist of feasible optimal solutions. Non-dominated solutions sacrifice in one objective(s) to achieve gain in the other objective(s) of a problem. This provides the flexibility of different possible options to make a final decision. The dimension of Pareto front is equal to the number of objectives in the problem. It is desired that the Pareto front should provide a uniformly distributed and complete spectrum of the problem including the extreme ends of the objective functions. Based on the second method, various multiobjective algorithms such as Vector Evaluated Genetic Algorithm (VEGA), Strength Pareto Evolutionary Algorithm (SPEA), Nondominated Sorted Genetic Algorithm (NSGA) and their modifications are available (Konak et al., 2006; Zitzler et al., 2000). These algorithms are designed on the basis of dominance rank, count, distance, or their combination to distinguish between the dominated and non-dominated solutions. A member's dominance rank and count are defined on the basis of the number

of members dominating the member and the number of members being dominated by the member. The distance is a measure of the well distributed Pareto front.

3.2.4 GA Application

The population size for this application of GA was selected to be 100. After trying different number of generations for the convergence of the population, the number of generations was kept as 30. Elites are the members of a population which go directly to the next generation and the size of the elite set specifies the number of members that is guaranteed to survive into the next generation. The size of elite set was 20 and the crossover was two points binary crossover with a probability of 0.6 whereas mutation was uniformly distributed with the probability of 0.05. In two point crossover operator, two positions in parents strings are chosen at random and new offspring are formed by swapping the element values between parents. The uniformly distributed operator is applied by choosing an element in the original member and this element is changed to a random value between the upper and lower bounds defined for the elements. The MultiObjective Evolutionary Algorithm (MOEA) (Sarker et al., 2002) was adapted to accommodate GA operators and was applied to obtain the Pareto sets of monitoring network. The steps of modified MOEA, as applied for the selection of monitoring stations, are shown in Figure 3.4. In this algorithm, non-dominated solutions from the previous generation enter as elites into the current generation and the rest of the population are chosen among the dominated solutions using tournament selection based on all objectives. These solutions undergo crossover and mutation operators. If there were more non-dominated solutions than the size of elite set, elites were selected based on the neighborhood distance. The neighborhood distance $D(x)$ for a member x is the sum of the distance from m nearest solutions on the same Pareto front for the problem of m objectives.

$$D(x) = \frac{\min \sum_{j=1}^m \{ \|x - x_j\| : x, x_j \in \{x_1, \dots, x_{nd}\} \}}{m} \quad (3.1)$$

In this problem, m is equal to three and nd is the number of nondominated members. The

process continued till the last generation. The GA application was repeated for ten times to compare the results. Finally, the SPARROW model was applied on the Pareto solutions.

3.3 Results and Discussion

From the modified MOEA application, four sets of monitoring stations were obtained as the alternatives. The alternatives (A, B, C and D) and their corresponding objective values are listed in the Table 3.1. In all the alternatives, the number of selected monitoring stations in the optimized sets varies from 20 to 30, which are significantly less than the existing number of monitoring stations (56) in the basin. Reducing the monitoring stations to this extent can result in a significant decrease in the cost of monitoring in the study area. By maintaining the appropriate records at the selected monitoring stations, the contamination of the study area might be predicted more effectively as compared to establishing new monitoring stations with short duration of record.

All the four sets of selected monitoring stations (A, B, C and D) are shown in Figure 3.5. The majority of the monitoring stations are common to all the alternatives. The SPARROW model was applied to all these alternative sets of monitoring stations. The coefficients for the sources, land –water delivery and stream/ reservoir attenuation factors were analyzed for the correlations. Finally, selected models for these alternative sets and corresponding statistical indices, discussed in Chapter II, are represented in Table 3.2 and 3.1 respectively. The objective values are reported for the alternatives and the selected model (Model III) of Chapter II (Table 3.1). In Table 3.1, alternative B has the minimum number of monitoring stations (Objective 1), and lowest average flux assessment (Objective 2) and uncertainty in flux (Objective 3) among all the alternatives. If cost reduction is a major criterion this alternative appears to be the best option. Alternative D has the highest number of monitoring stations, the highest average flux assessment and uncertainty in the flux (Table 3.1). For Model III the average of

logarithmic mean annual flux values of the monitoring stations (Objective 2) is lowest among all other alternatives. This implies that the model does not include many monitoring stations with high mean annual flux values. On the other hand, statistical indices are not very good for any of the alternatives. The value of AIC is large for the alternatives. This implies that the complex representation of the processes in these models is not able to predict the phenomena accurately. NSE is close to 0 for all alternatives except the alternative C. The biases may be caused due to the inclusion of some extremely high values of flux. Percent bias is minimum for the alternative C (3.20%), reflecting that this model neither over nor under estimate the mean annual flux.

In Table 3.2, the p-values of the parameter coefficients are listed in parenthesis. The Model A, B, C and D correspond to the sets of the monitoring stations A, B, C and D, respectively. The root mean square error is the least for Model C whereas the Model A is the least complex model with minimum degree of freedom (five). All the models included urban and pasture land uses as sources of *E. coli*. The forest land use, which had been included as a significant source of *E. coli* in the selected model in Chapter II, appeared as a non-significant source in Model C. This model also includes point sources as a significant contributor of pathogens; none of the other models have included this source of *E. coli* with any statistical significance (p-value less than 25%). Thus alternative C can be considered as a good alternative based on the fitted models, objective values and statistical indices (Table 3.1 and 3.2).

Model C is the most accurate model among the models fitted for the alternative selections of monitoring stations due to the minimum error term and maximum coefficient of determination. The selected monitoring stations in Model C are shown in Figure 3.6. This model selected the majority of monitoring stations in the highly contaminated and critical areas (Edwards Aquifer and San Antonio) of the study area, those require the rigorous monitoring. In Model C, coefficient and statistical significance of temperature, a land –water delivery factor, differed considerably from the other

models (Table 3.2). Reach slope which has not been included in any of the other models, has been included in Model C. Rainfall has entered only in Model B as a negative land-water delivery factor. Thus, the selection of monitoring stations has significant effect on the representation of water quality processes occurring in the study area.

The selected monitoring stations in Model III and Model C are shown in Figure 3.7. It can be observed that the monitoring stations selected in Model III are located in cluster whereas in Model C, the monitoring stations are distributed throughout the study area. The Model III and Model C are compared in Table 3.3. In Model III, forest and urban land use are the significant sources of *E. coli* whereas in Model C, only point sources are significant sources and the coefficients for forest and urban land are very small and insignificant. Based on the statistical indices, Model III appears to be a better model since AIC is less and NSE is more than that of Model C. At the same time, considering regression error statistics (RMSE and R^2), Model C is better than Model III. In Figure 3.7, the observed and predicted logarithmic mean annual flux values are plotted. This model does not show any systematic trend in the mean annual flux towards under or overestimation. This has been reflected in percent bias statistics (Table 3.1).

In the earlier application of SPARROW, monitoring stations were selected based on the standard error (Smith et al., 1997). The selection of monitoring stations based on only standard error to flux ratio were very different from the current selections (Figure 2.4, Chapter II). These sets of monitoring stations did not include sufficient numbers of monitoring stations from the San Antonio River Basin (Chapter II) whereas all the alternative sets of monitoring stations generated by GA have adequately represented the San Antonio River Basin. From the previous application of SPARROW for *E. coli* assessment in this study area, it had been concluded that San Antonio River Basin is more contaminated and requires more rigorous monitoring (Chapter II). Here the selection of monitoring stations based on the maximum mean annual flux (Objective 2)

ensures the inclusion of the monitoring stations with large value of flux for SPARROW application.

3.4 Conclusions

The location of the monitoring stations plays a crucial role in representing the water quality scenario in a basin. Especially for the pathogen contamination due to the limited availability of the monitored concentrations of *E. coli*, uncertainty in the observed mean annual flux is large. With the help of a multiobjective algorithm the monitoring stations can be selected for the application of a water quality model such as SPARROW. These selection criteria also form the basis to design the water quality monitoring network or to improve an existing network. This will help in accurate prediction of contamination in river basins.

CHAPTER IV

SUMMARY AND CONCLUSIONS

4.1 Summary

The Guadalupe and San Antonio River Basins in Texas have undergone land use changes for several decades. High concentrations of pathogens have been observed in several waterbodies in these river basins. In the past, various water quality models have been applied to study impairment due to bacteria and other nutrients in watersheds. SPARROW, a new modeling approach, relying on the nonlinear regression to assess pathogen contamination was applied in this research to study the fate and transport of *E. coli* in the Guadalupe and San Antonio River Basins. This model requires spatially distributed watershed and reaches attributes and the mean annual *E. coli* flux calculated using the monitored water quality and streamflow data at various monitoring stations located in the river basins. Thus the assessment of sources, land-water delivery and attenuation factors in SPARROW model depends on the quality and quantity of monitored records of water quality and streamflow. In our study, the effect of the selection of the monitoring stations on the SPARROW prediction has been analyzed. The selection of monitoring stations is the function of the uncertainty in the mean annual flux assessment, high flux values and adequate representation of the study area.

4.2 Conclusions

- The SPARROW model was used successfully to predict the pathogen contamination in the Guadalupe and San Antonio River Basins. However, the associated error with flux determination was relatively large due to the limited availability of monitored data on *E. coli* concentration in waterbodies in the study area. It was noted that while applying the SPARROW model to contaminants with limited data availability, uncertainty in the response variable, mean annual flux, will be a major issue. The selections of monitoring stations based on the standard error in FLUXMASTER results have influenced the coefficients significantly in the assessment of different sources, land-water delivery factors and stream/ reservoir attenuation factors in the

SPARROW models. The upper and lower San Antonio River Subbasins were observed to be more contaminated because of relatively large contribution from forest and urban land uses. The selected model based on various statistical indices was able to predict almost all of the streams listed in the Clean Water Acts, 303 (d) list for impairment due to bacteria (Chapter II).

- The effects of the number of monitoring stations and their locations on the prediction of the SPARROW model were also analyzed. A model with few numbers of monitoring stations resulted in accurate assessment of the *E. coli* fluxes but ignored some of important sources while a model with large number of monitoring stations resulted in the complex description of the water quality processes in the study area (Chapter II).
- Not only is the monitored water quality data crucial for SPARROW application, but also the scale and details of the explanatory variables (sources, land water delivery and attenuation factors). In the earlier application of SPARROW, Preston and Brakebill (1999) observed significant improvement in the nutrient assessment of Chesapeake Bay by including the coastal details available in National Hydrography Dataset (1:24,000) as compared to stream dataset available at 1:500,000. In this study point sources were found not to contribute to pathogen contamination. This may be attributed to the unavailability of the actual *E. coli* concentration data in wastewater effluents (Chapter II).
- It was realized that selection of monitoring stations based on standard error to mean annual flux ratio does not insure the adequate representation of the monitoring stations at highly contaminated or sensitive places. In this study, we designed the application of a GA to select the monitoring stations based on maximum detection of the pathogen with minimum uncertainty and number of the monitoring stations. Among these alternative sets of the monitoring stations, the optimum set of monitoring stations was selected based on SPARROW results. This provided a methodology to optimize the existing monitoring network for the contaminant assessment in Guadalupe and San Antonio River Basins (Chapter III).

4.3 Scope for Future Research

The water quality model SPARROW relies on monitored data to assess the fate and transport of contaminants in watersheds. FLUXMASTER and SPARROW can also be explored to simulate the effect of time intervals between the measurements of pathogen concentrations based on yearly trend in water quality of a region. Based on the incremental and delivered contaminant yield predicted by SPARROW, highly contaminated areas in a watershed can be detected. This information can be further used for TMDL or WPP development.

REFERENCES

- Alexander, R. B., H. E. Alexander, U. Shankar, and G. B. McBride. 2002. Estimating the sources and transport of nutrients in the Waikato River Basin, New Zealand. *Water Resources Research* 38(12): 1-23.
- Center for Environmental Informatics Database (C.E.I.). 2007. Pennsylvania State University. Available at: <http://dbwww.essc.psu.edu/>. Accessed on 15 August 2007.
- Eiben, A.E., and J.E. Smith. 2003. Introduction to evolutionary computing. Natural Computing Series, Springer, New York, pp. 36-51.
- Ferguson, C., A. M. Husman, N. Altavilla, D. Deere, and N. Ashbolt. 2003. Fate and transport of surface water pathogens in watersheds. *Critical Reviews in Environment Science and Technology* 33(3): 299-361.
- Fox, J. 2002. Bootstrapping regression models, Appendix to an R and S-Plus companion to applied regression. Available at: <http://cran.r-project.org/doc/contrib/Fox-Companion/appendix.html>. Accessed on 03 January 2008.
- Grefenstette, J. J. 1986. Optimization of control parameters for genetic algorithm, IEEE Transactions on Systems. *Man and Cybernetics* 16(1): 122-128.
- Guadalupe and Blanco River Authority (GBRA). 2007. Water quality data collection. Available at: <http://www.gbra.org/>. Accessed on 15 July 2007.
- Harmel, R. D., and P. K. Smith. 2007. Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling. *Journal of Hydrology* 337: 326-336.
- Harmel, R. D., R. J. Cooper, R. M. Slade, R. L. Haney, and J. G. Arnold. 2006. Cumulative uncertainty in measured streamflow and water quality data for small watersheds. *Transactions of the ASABE* 49(3): 689-701.
- Icaga, Y. 2005. Genetic algorithm usage in water quality monitoring networks optimization in Gediz (Turkey). *Environmental Monitoring and Assessment* 108: 261-277.

- Konak, A., W. C. David, and A. E. Smith. 2006. Multiobjective optimization using genetic algorithms: a tutorial. *Reliability Engineering and System Safety* 91: 992-1007.
- Letcher, R. A., A. J. Jakeman, and B. F. W. Croke. 2004. Model development for integrated assessment of water allocation options. *Water Resources Research* 40: 1-15.
- Lo, C. P., and A. K. W. Yeung. 2002. *Concepts and techniques in Geographic Information Systems*. Prentice Hall in Geographic Information System, Prentice Hall, New Jersey, pp. 365-367.
- McMahon, G., R. B. Alexander, and S. S. Qian. 2003. Support of Total Maximum Daily Load programs using spatially referenced regression model. *Journal of Water Resources Planning and Management* 129(4): 315-329.
- Moore, R. B., C. M. Johnston, K. W. Robinson, and R. D. Jeffery. 2004. Estimation of total nitrogen and phosphorus in New England streams using spatially referenced regression models. *Scientific investigation report 2004-5012*, U.S. Geological Survey, New Hampshire.
- Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Binger, R. D. Harmel, and T. L. Veith. 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE* 50(3): 885-900.
- National Water Information System (NWIS). 2007. USGS real-time water data for the nation. Available at: <http://waterdata.usgs.gov/nwis/rt>. Accessed on 08 July 2007.
- Ning, S., and N. Chang. 2004. Optimal expansion of water quality monitoring network by fuzzy optimization approach. *Environmental Monitoring and Assessment* 91: 145-170.
- Park, S., J. H. Choi, S. Wang, and S. S. Park. 2006. Design of a water quality monitoring network in a large river system using the genetic algorithm. *Ecological Modelling* 199: 289-297.
- Preston, S. D., and J. W. Brakebill. 1999. Application of spatially referenced regression modeling for the evaluation of total nitrogen loading in the Chesapeake Bay

- watershed. *Water-resources investigation report 99-4054*, U.S. Geological Survey, Maryland.
- Qian, S. S., K. H. Reckhow, J. Zhai, and G. McMahon. 2005. Nonlinear regression modeling of nutrient loads in streams: a Bayesian approach. *Water Resources Research* 41: 1-10.
- Rasch, D. 1995. The robustness against parameter variation of exact locally optimum designs in nonlinear regression- a case study. *Computational Statistics and Data Analysis* 20: 441-453.
- Reckhow, K. H. 1994. Water quality simulation modeling and uncertainty analysis for risk assessment and decision making. *Ecological Modelling* 72: 1-20.
- Reed, P., B. Minsker, and D. E. Goldberg. 2000. Designing a competent simple genetic algorithm for search and optimization. *Water Resources Research* 36(12): 3757-3761.
- Reed, P., B. Minsker, and D. E. Goldberg. 2001. A multiobjective approach to cost effective long-term groundwater monitoring using an elitist nondominated sorted genetic algorithm with historical data. *Journal of Hydroinformatics* 3(2): 71-89.
- San Antonio River Authority (SARA). 2007. SAR basin data. Available at: <http://www.sara-tx.org/>. Accessed on 08 July 2007.
- Sarker, R., K. H. Liang, and C. Newton. 2002. A new multi-objective evolutionary algorithm. *European Journal of Operational Research* 140: 12-23.
- Savic, D., and S. T. Khu. 2005. *Evolutionary computation in hydrological sciences*. Encyclopedia of Hydrological Sciences, John Wiley and Sons Ltd, New York.
- Schwarz, G. E., A. B. Hoos, R. B. Alexander, and R. A. Smith. 2007. The SPARROW surface water-quality model: theory, application and user documentation. Available at: <http://usgs.er.gov/sparrow/sparrow-mod/>. Accessed on 01 January 2007.
- Smith, R. A., G. E. Schwarz, and R. B. Alexander. 1997. Regional interpretation of water-quality monitoring data. *Water Resources Research* 33(12): 2781-2798.

- Srivastava, P., J. M. Hamlett, and P. D. Robillard. 2002. Watershed optimization of best management practices using AnnAGNPS and a genetic algorithm. *Water Resources Research* 38(3): 1-14.
- Texas Commission of Environmental Quality (TCEQ). 2000a.: Texas surface water quality standards 307. Chapter 307. Available at:
<http://www.tceq.state.tx.us/assets/public/legal/rules/rules/pdflib/307%60.pdf>.
Accessed on 08 March 2008.
- TCEQ. 2000b. TCEQ hydrology layers- stream segments 2000. Available at:
<http://www.tceq.state.tx.us/gis/hydro.html>. Accessed on 08 April 2008.
- TCEQ. 2006a. Executive summary, Texas water quality inventory and 303 (d) list. Available at:
http://www.tceq.state.tx.us/compliance/monitoring/water/quality/data/wqm/305_303.html#y2006. Accessed on 21 October 2007.
- TCEQ. 2006b. Benefits and costs of surface water quality programs, Texas water quality inventory and 303 (d) list. Available at:
http://www.tceq.state.tx.us/assets/public/compliance/monops/water/06twqi/2006_cobstbene.pdf. Accessed on 21 October 2007.
- TCEQ. 2007. Site layers- permitted wastewater outfalls. Available at:
http://www.tceq.state.tx.us/nav/data/swq_data.html. Accessed on 01 July 2007.
- U.S. Bureau of census. 2000. Census 2000 Tiger/ line data. Available at:
<http://www.esri.com/data/download/census2000tigerline/index.html>. Accessed on 30 July 2007.
- U.S. Department of Agriculture (USDA). 2002. National Agricultural Statistics Service- Data and statistics. Available at: http://www.nass.usda.gov/Data_and_Statistics/. Accessed on 20 July 2007.
- U.S. Environmental Protection Agency (USEPA). 2002. Consolidated Assessment and listing methodology: toward a compendium of best practices. Available at: http://www.epa.gov/owow/monitoring/calm/calm_ch11.pdf. Accessed on 18 May 2008.

- USEPA. 2005. National Hydrography Dataset (NHD) Plus. Available at:
<http://www.horizon-systems.com/nhdplus/>. Accessed on 02 April 2007.
- USEPA. 2007. Water discharge permits (PCS). Available at:
http://www.epa.gov/enviro/html/pcs/pcs_query_java.html. Accessed on 01 July 2007.
- U.S. Geological Survey (USGS). 2001. Multi Resolution Land Characteristics (MRLC) consortium. Available at: www.mrlc.gov/. Accessed on 15 September 2007.
- Weisberg, S. 2005. *Applied linear regression*. Wiley Series in Probability and Statistics, Hoboken, New Jersey, pp. 19-40.
- Yandamuri, S. R. M., K. Srinivasan, and S. M. Bhallamudi. 2006. Multiobjective optimal waste load allocation models for rivers using nondominated sorting genetic algorithm-II. *Journal of Water Resources Planning and Management* 132(3):133-143.
- Zitzler, E., K. Deb, and L. Thiele 2000. Comparison of multiobjective evolutionary algorithms: empirical results. *Evolutionary Computation* 8(2):173–195.

APPENDIX A

FIGURES

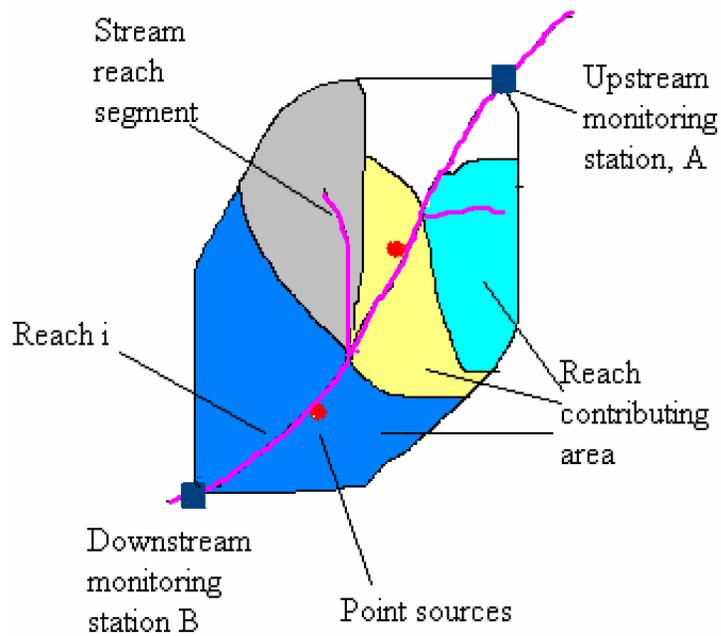


Figure 2.1. A schematic representation of a simulated watershed for SPARROW model application.

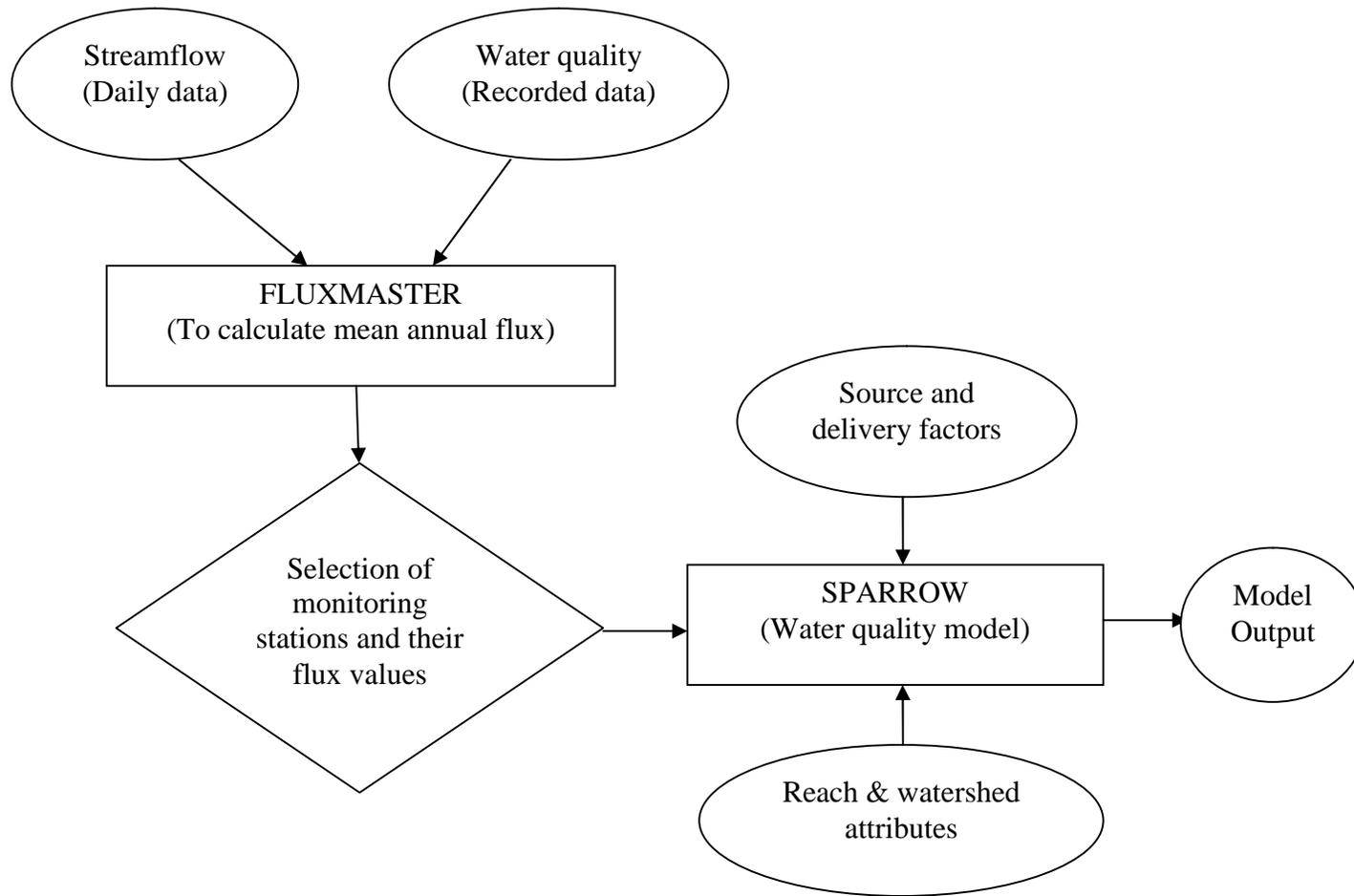


Figure 2.2. Flowchart of SPARROW application.

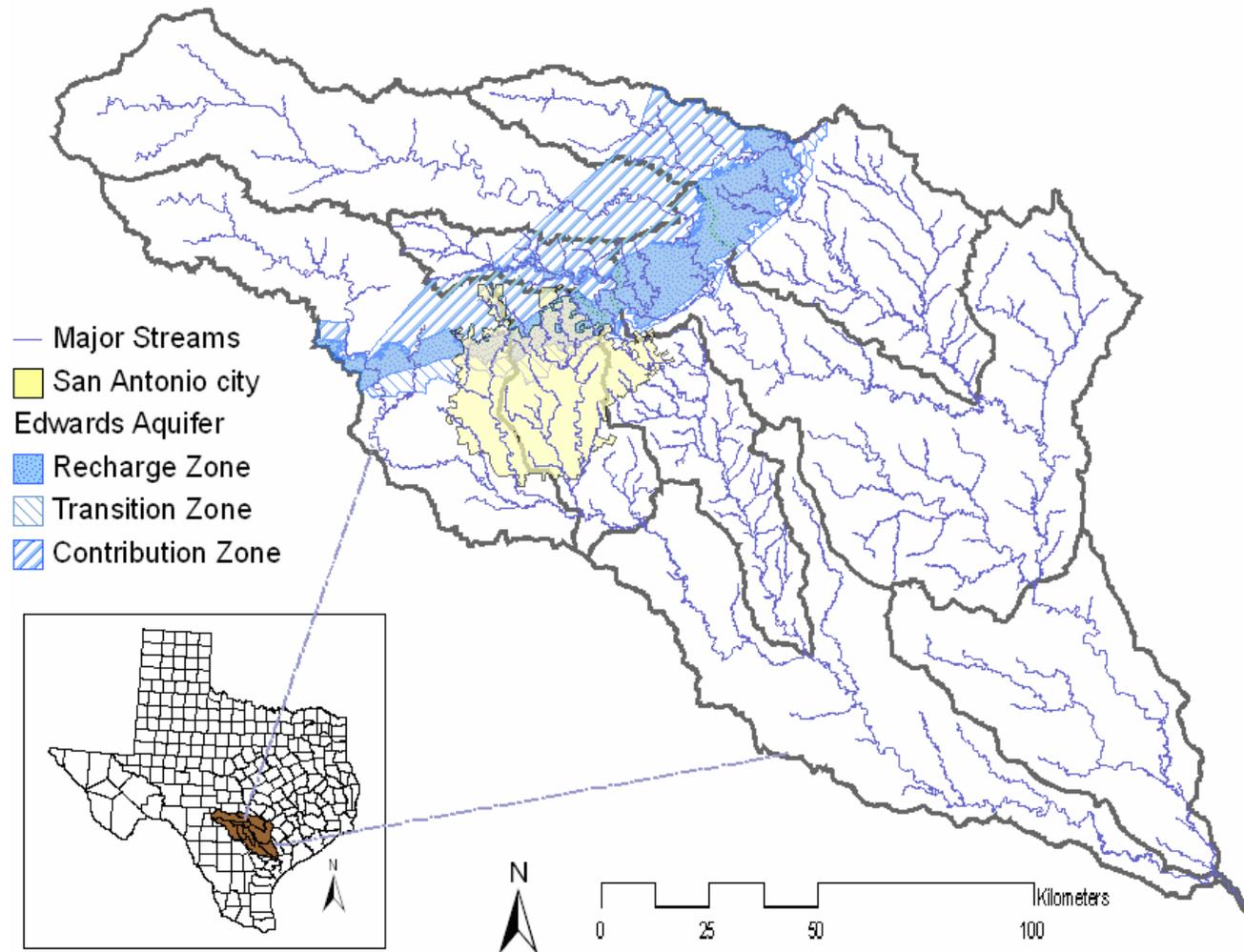


Figure 2.3. Guadalupe and San Antonio River Basins in Texas.

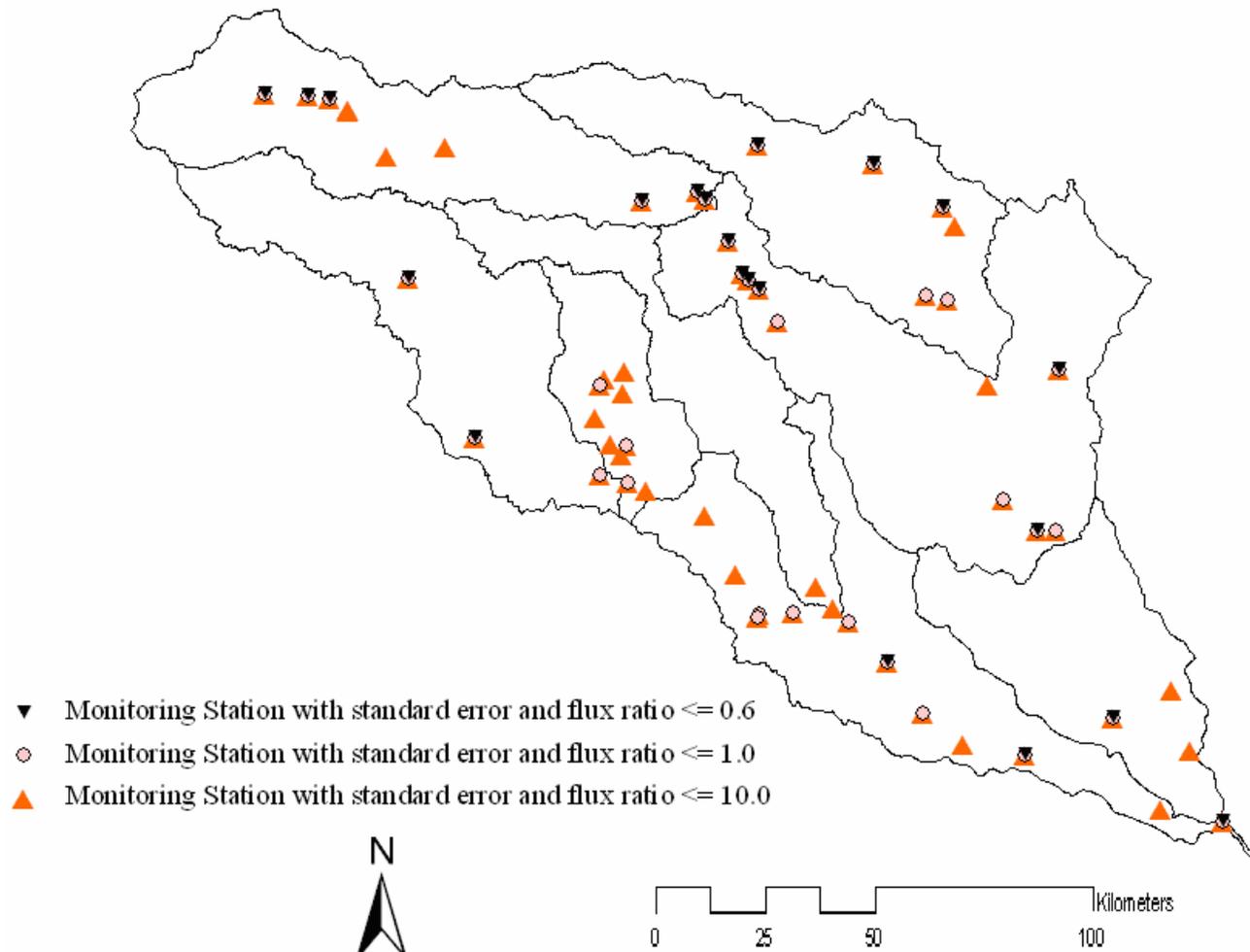


Figure 2.4. Locations of selected monitoring stations in the Guadalupe and San Antonio River Basins.

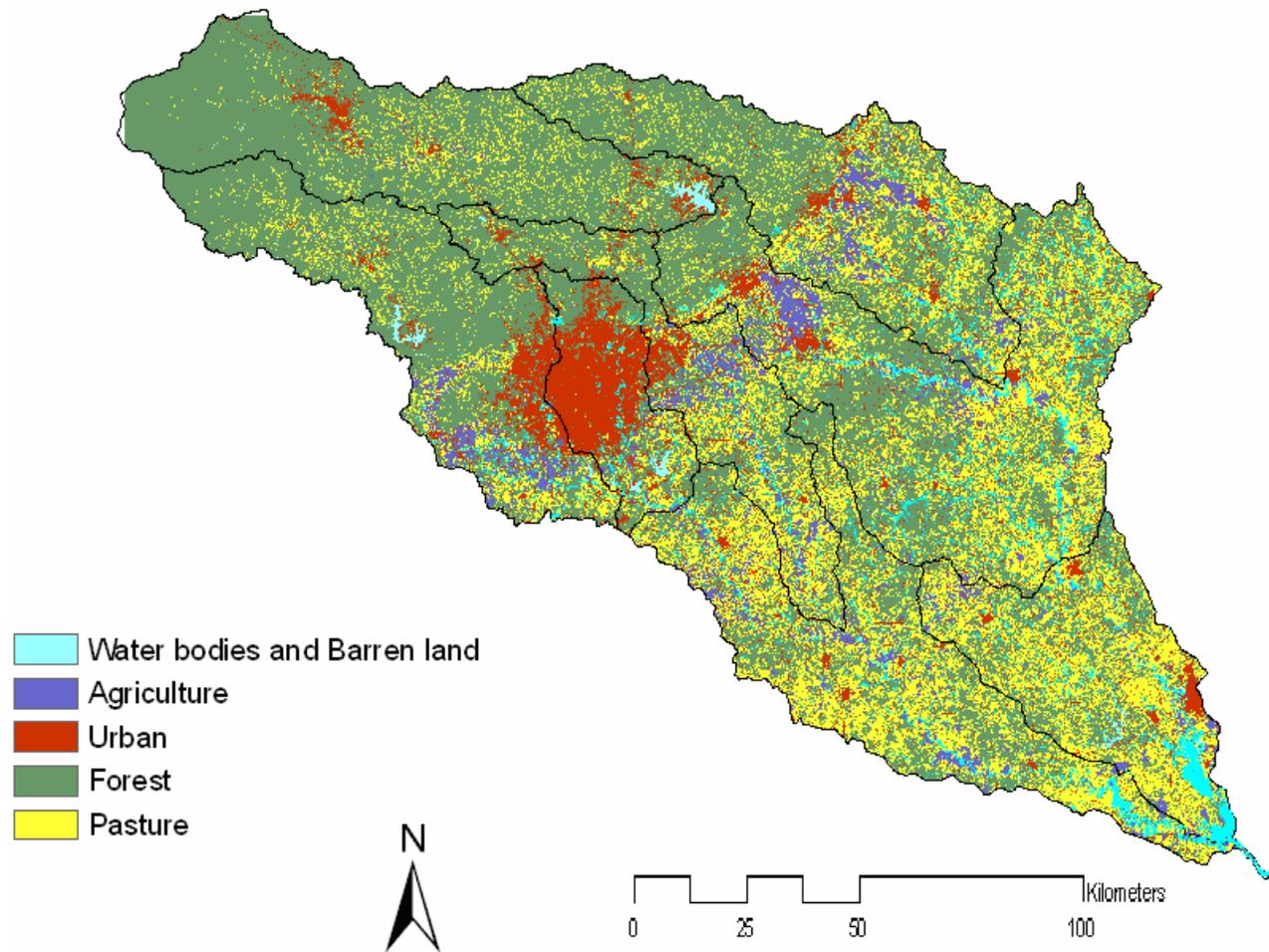


Figure 2.5. Different land use in the study area.

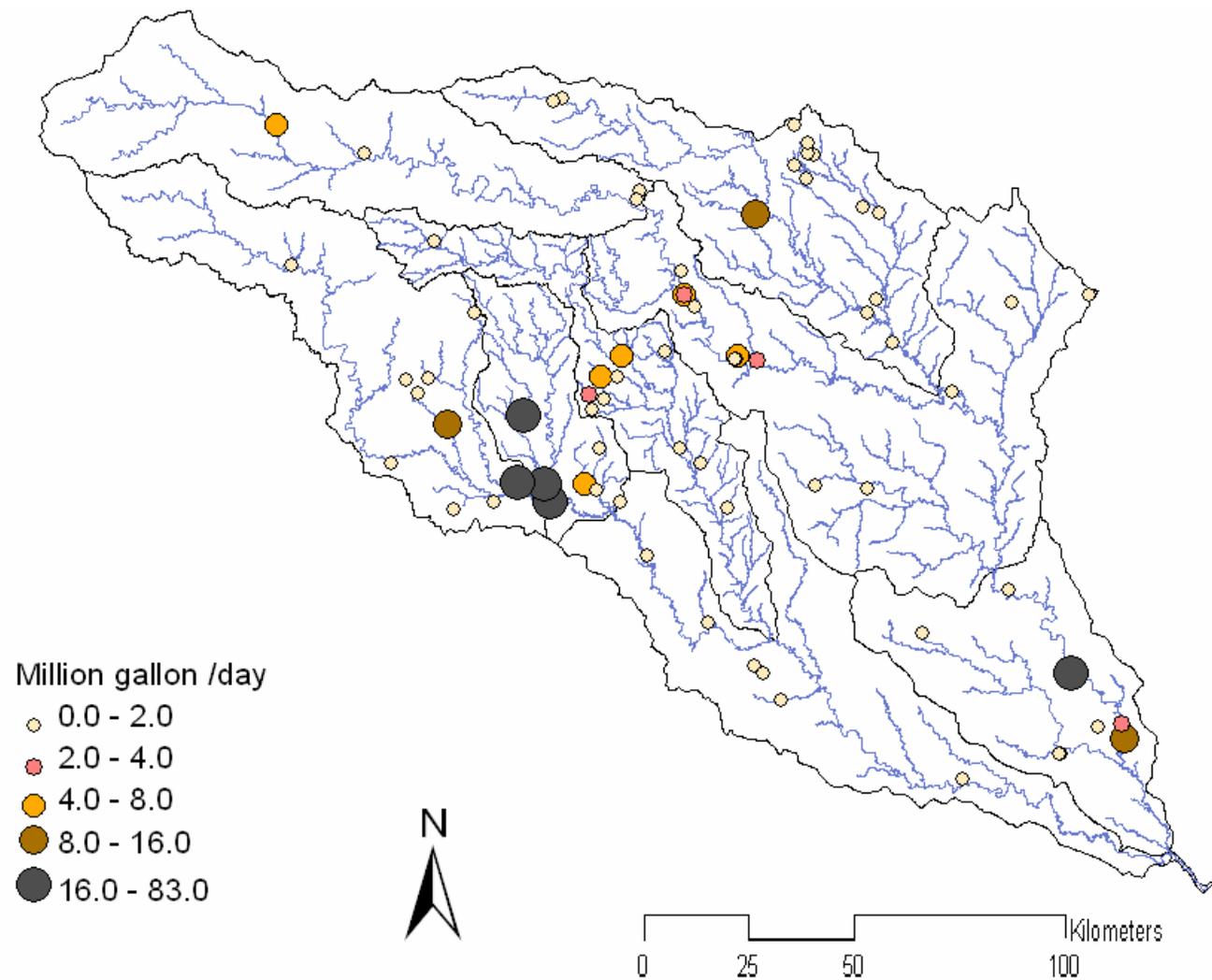


Figure 2.6. Discharge from point sources in million gallons per day.

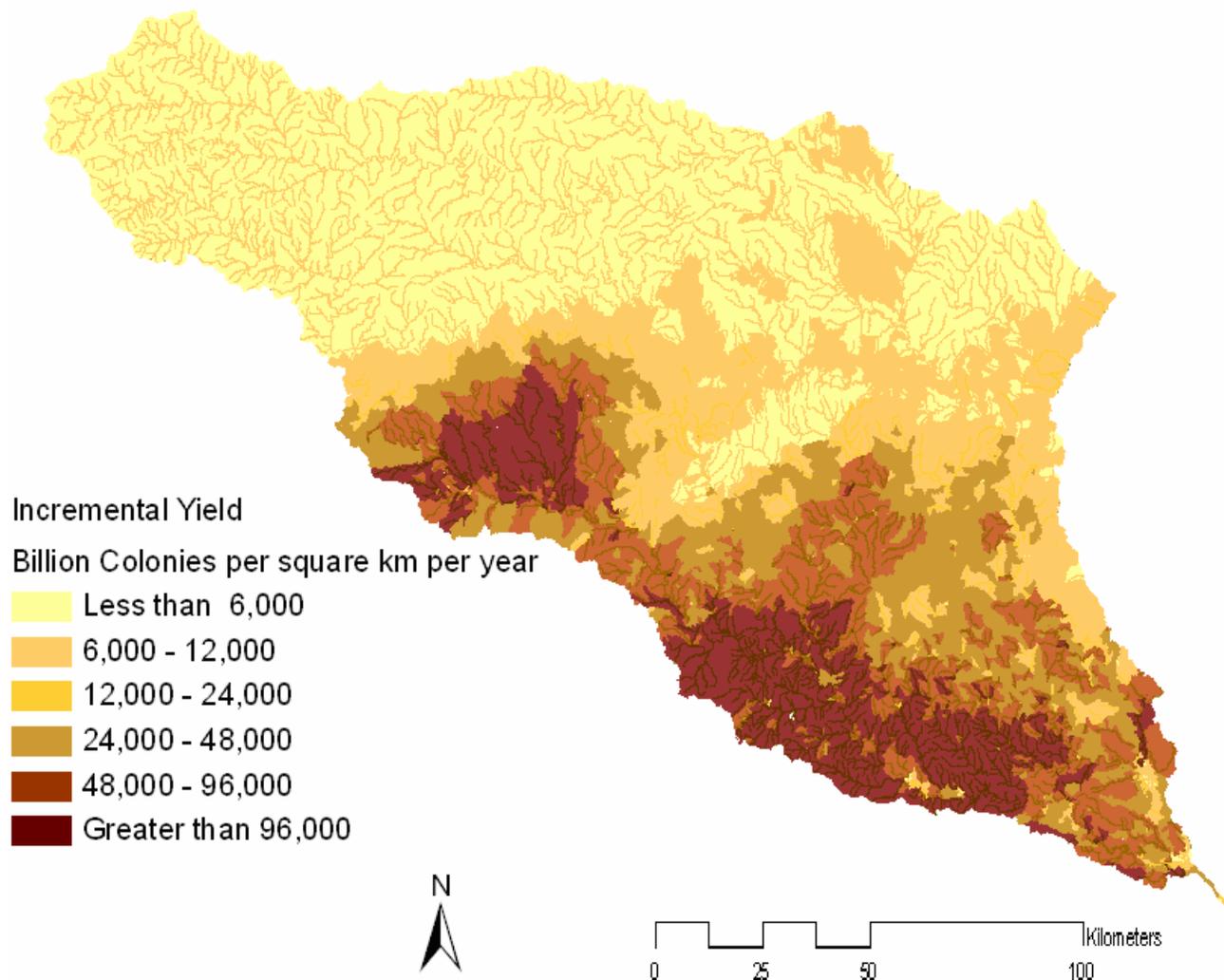


Figure 2.7. Spatial distribution of incremental *E. coli* flux of all the watersheds in the study area using Model III.

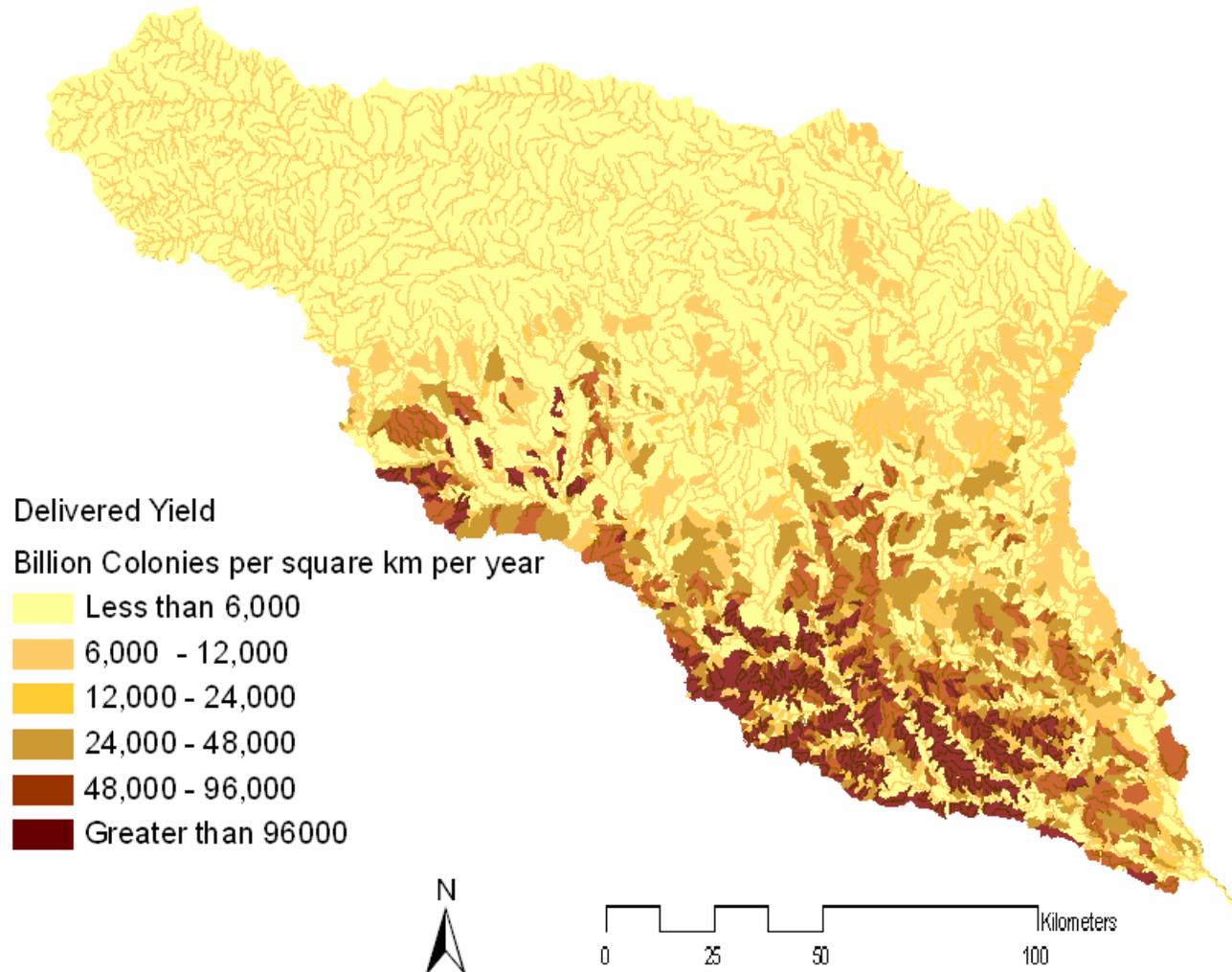


Figure 2.8. Spatial distribution of delivery of *E. coli* flux of all the watersheds in the study area using Model III.

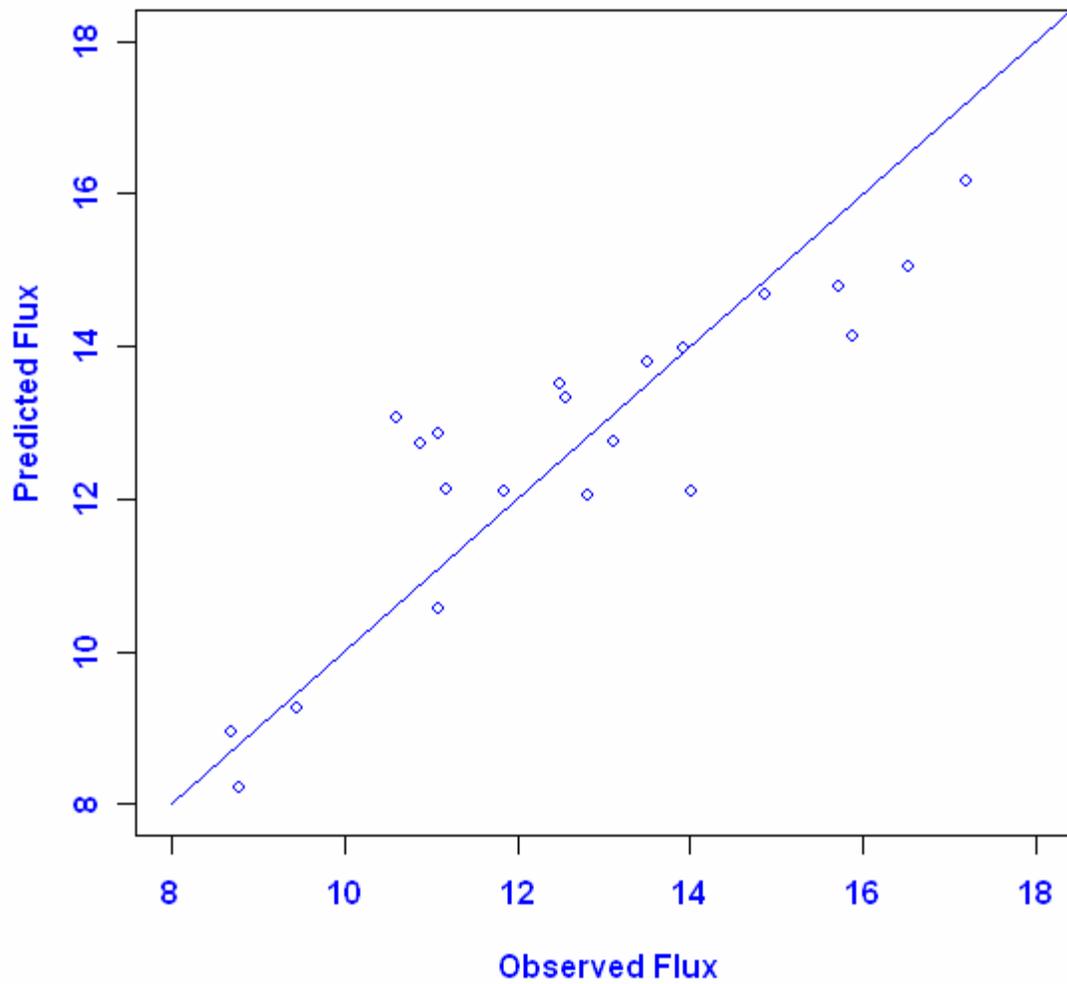


Figure 2.9. Relationship between the Natural Logarithm of observed (estimated mean annual flux in FLUXMASTER) and the predicted *E. coli* flux (SPARROW results) for Model III.

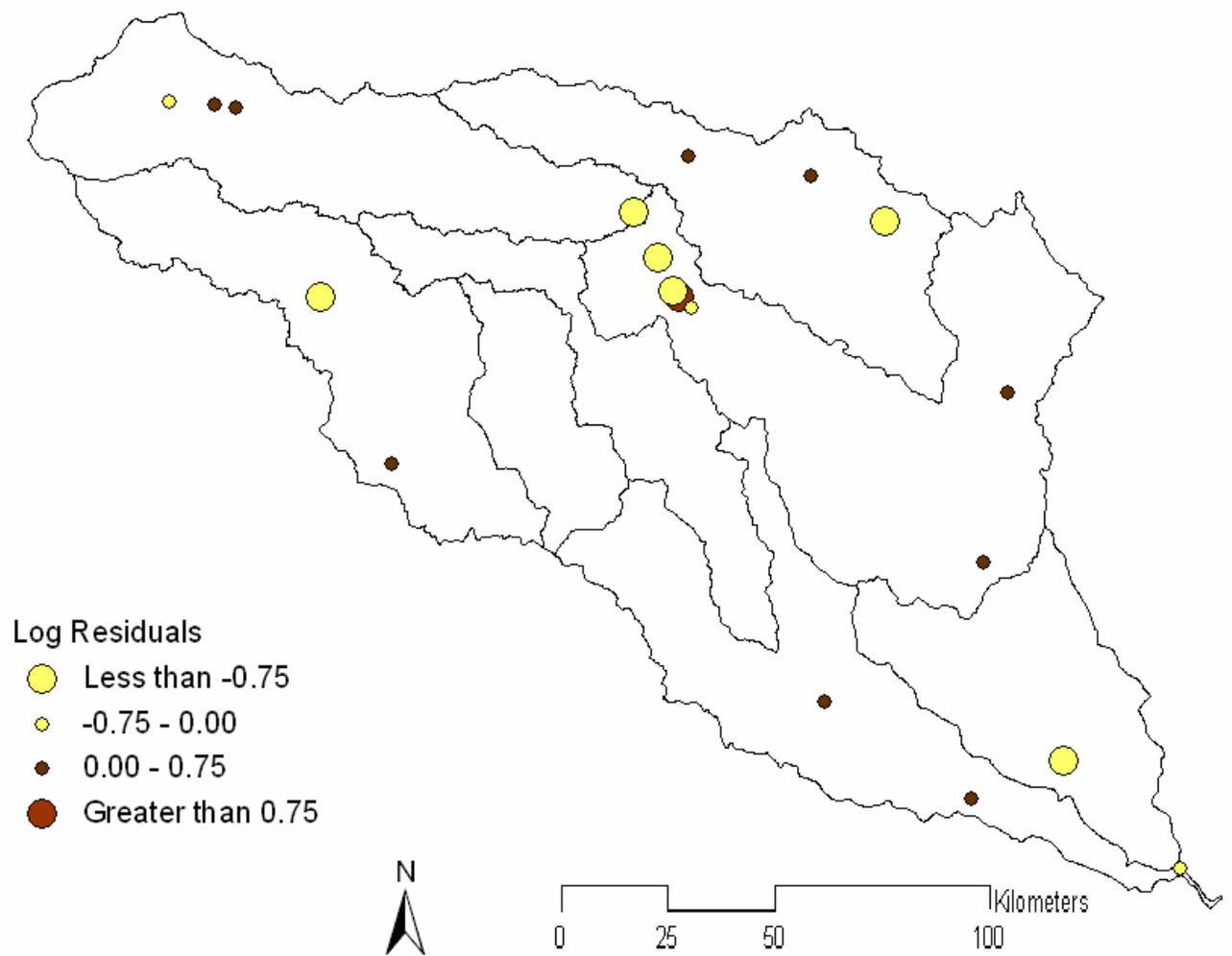


Figure 2.10. Log residuals of predicted *E. coli* flux at the location of monitoring stations for Model III.

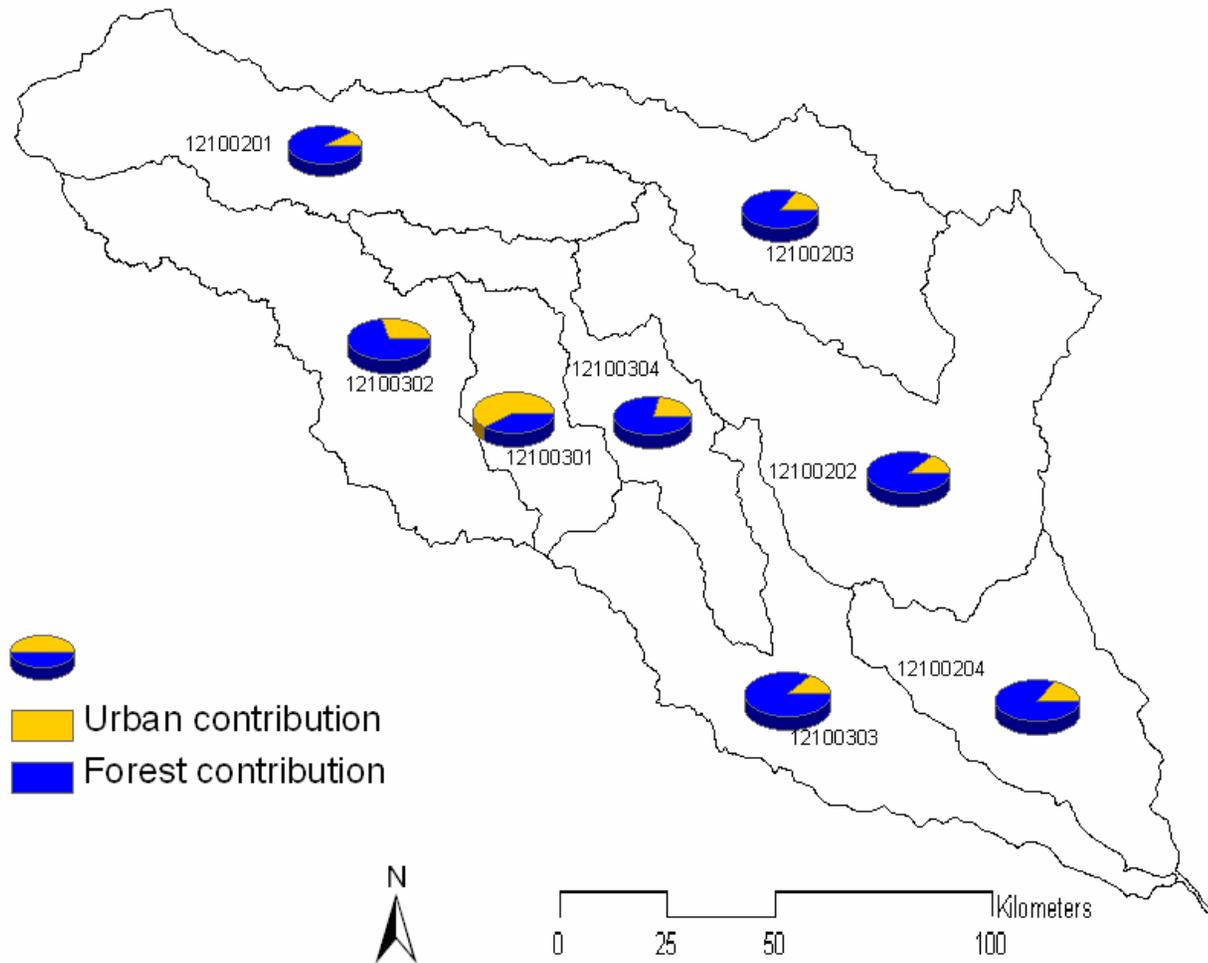


Figure 2.11. *E. coli* load contribution from the significant sources of Guadalupe and San Antonio River Subbasins, as predicted by Model III.

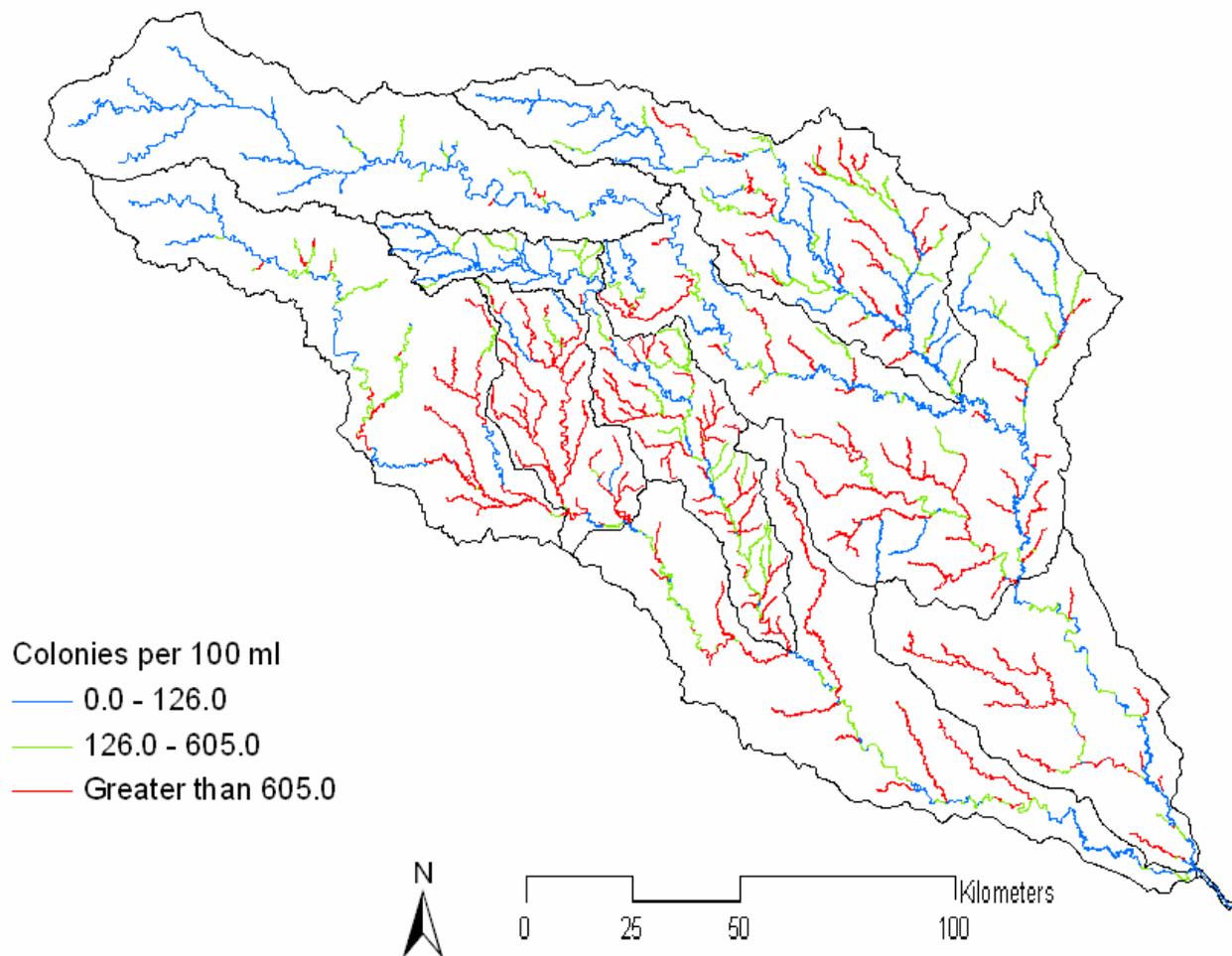


Figure 2.12. Predicted *E. coli* concentration in the major streams (flow greater than $0.13 \text{ m}^3 \text{ s}^{-1}$) in Guadalupe and San Antonio River Basins.

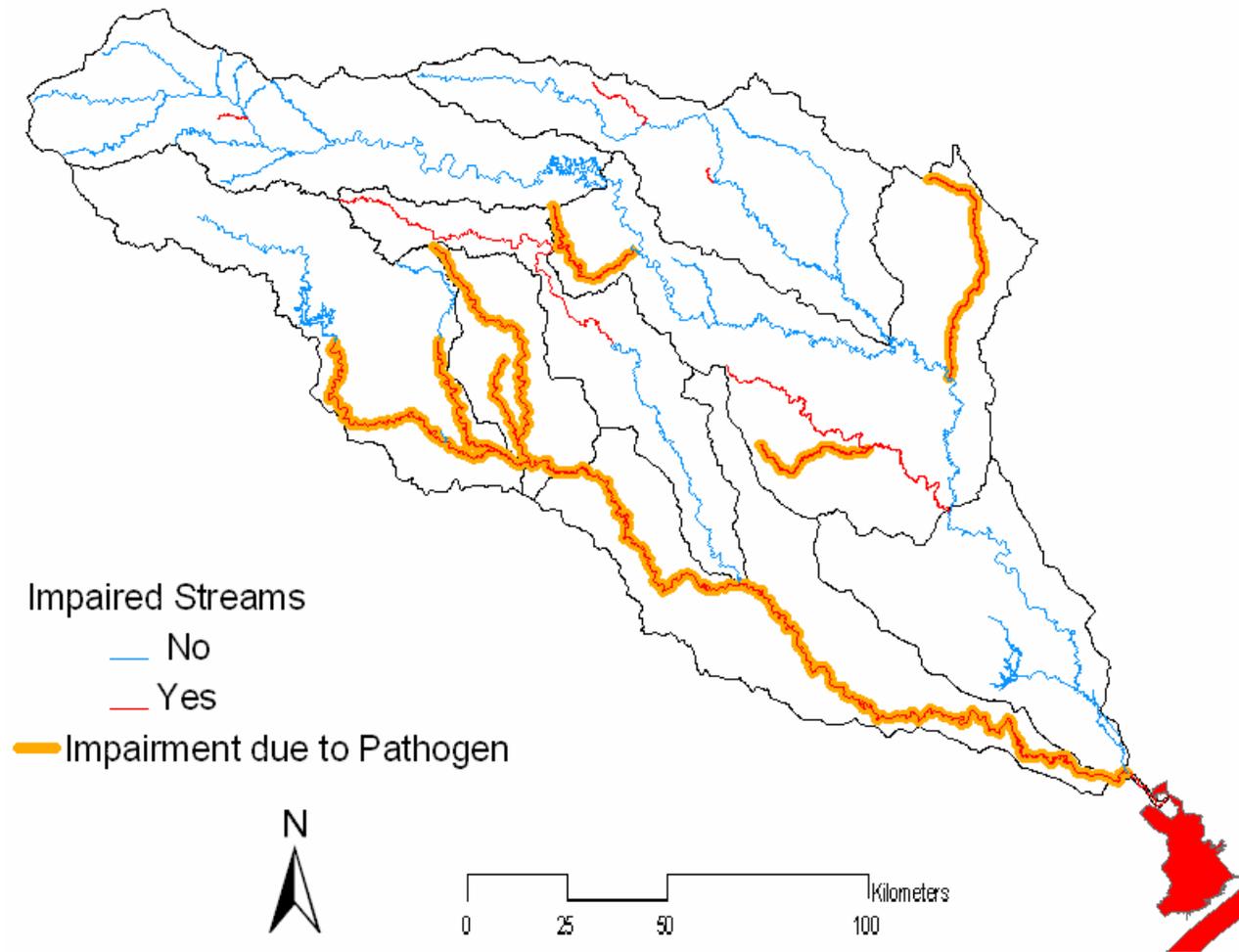


Figure 2.13. Monitored streams impaired due to pathogen listed in 303 (d) list of year 2000 (Source: TCEQ, 2000b).

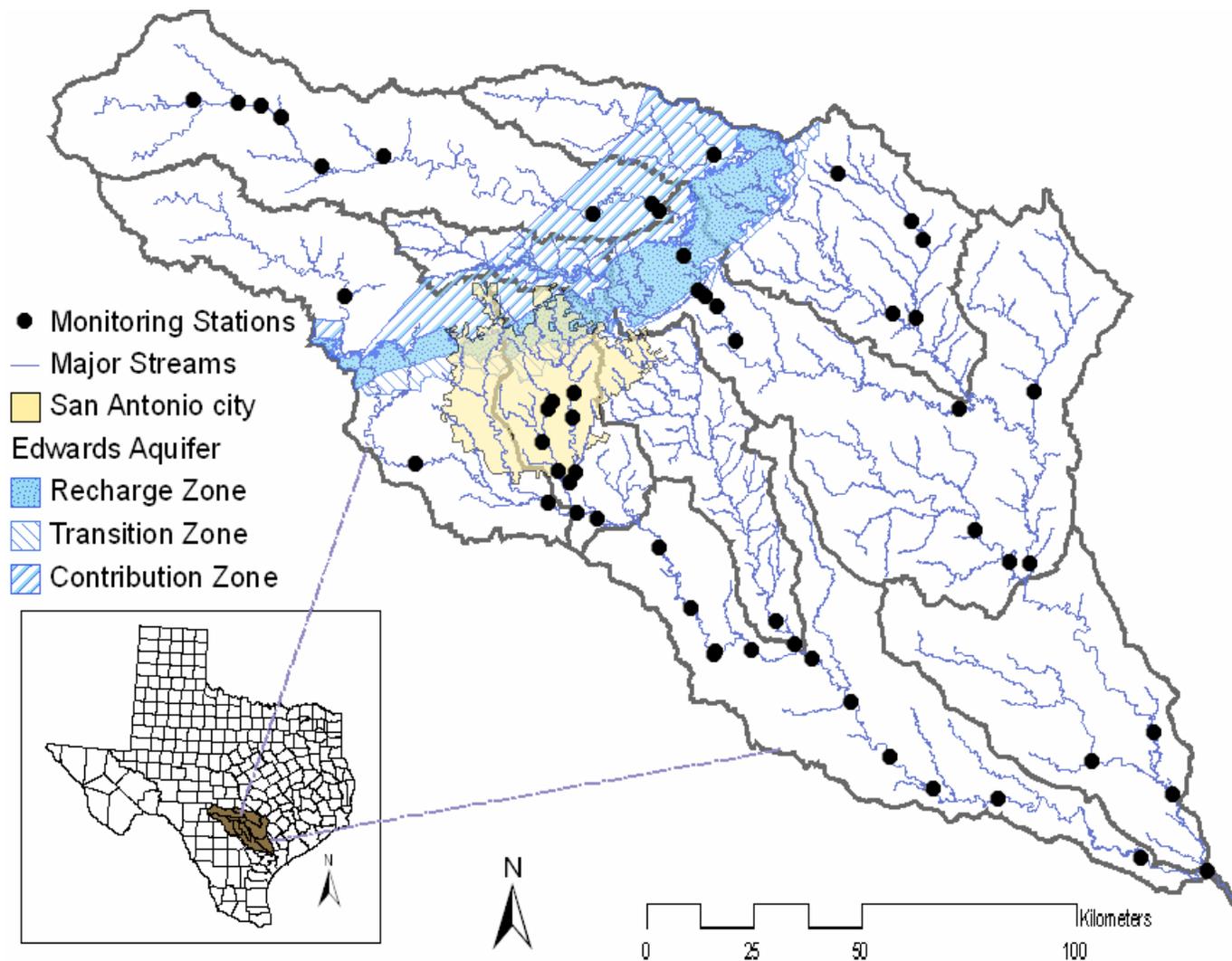


Figure 3.1. The locations of monitoring stations in the Guadalupe and San Antonio River Basins.

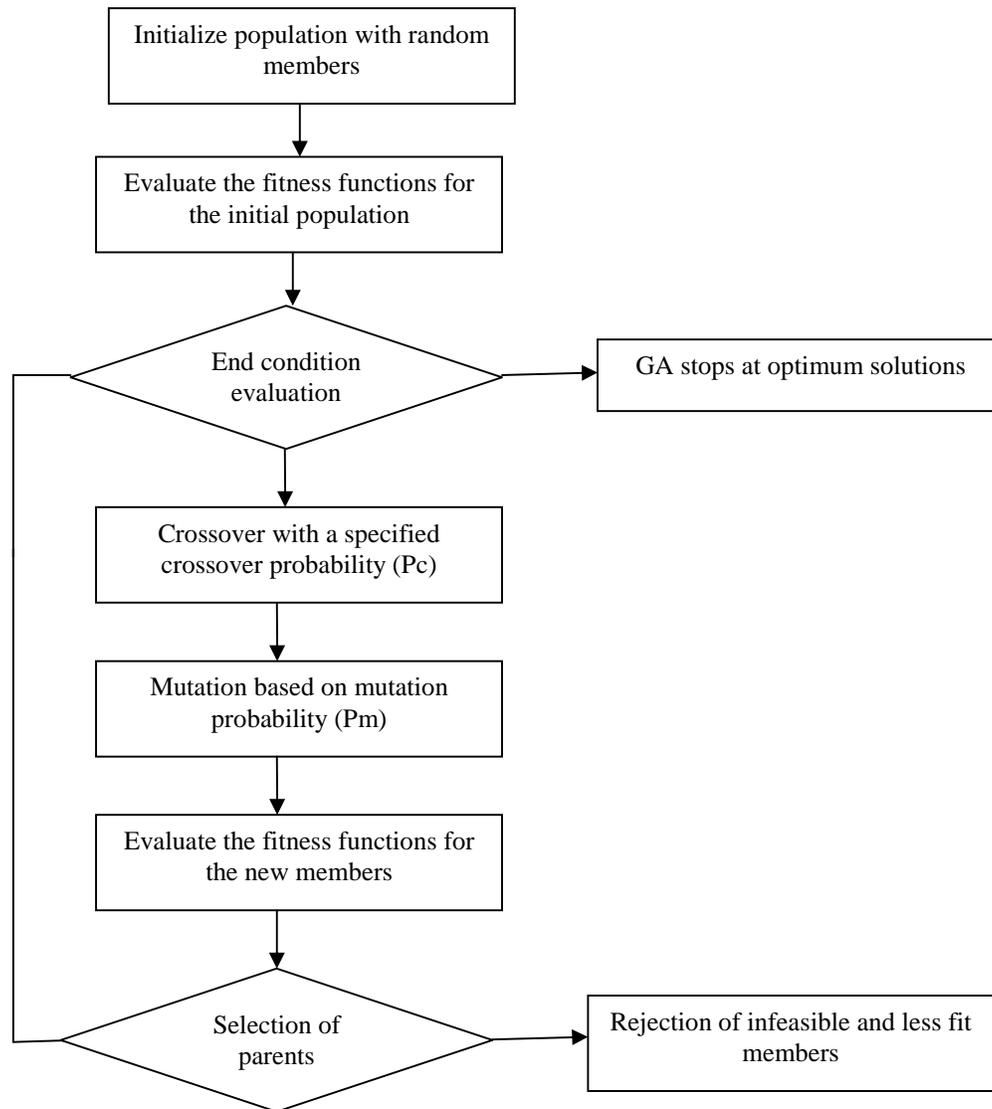


Figure 3.2. Flowchart of application of genetic algorithm.

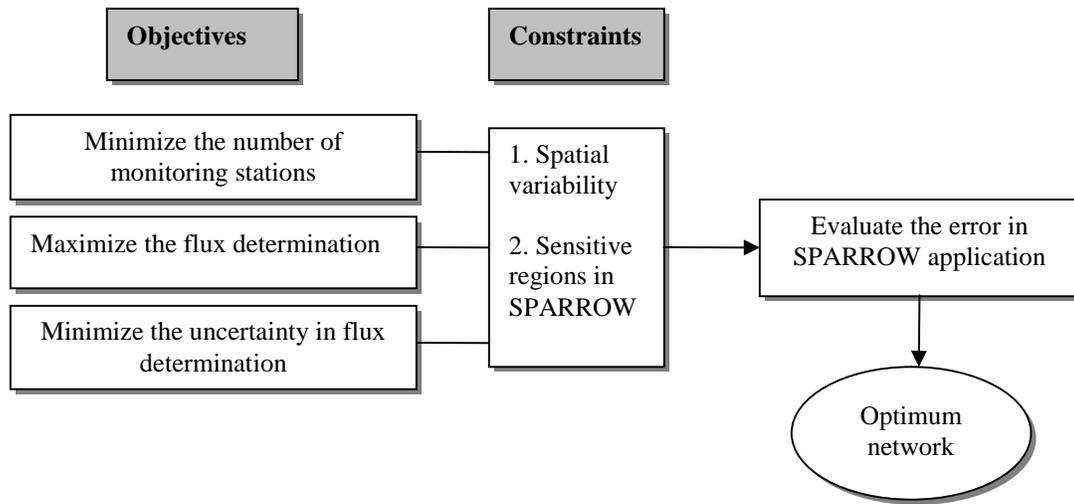


Figure 3.3. Objectives and constraints for multiobjective problem.

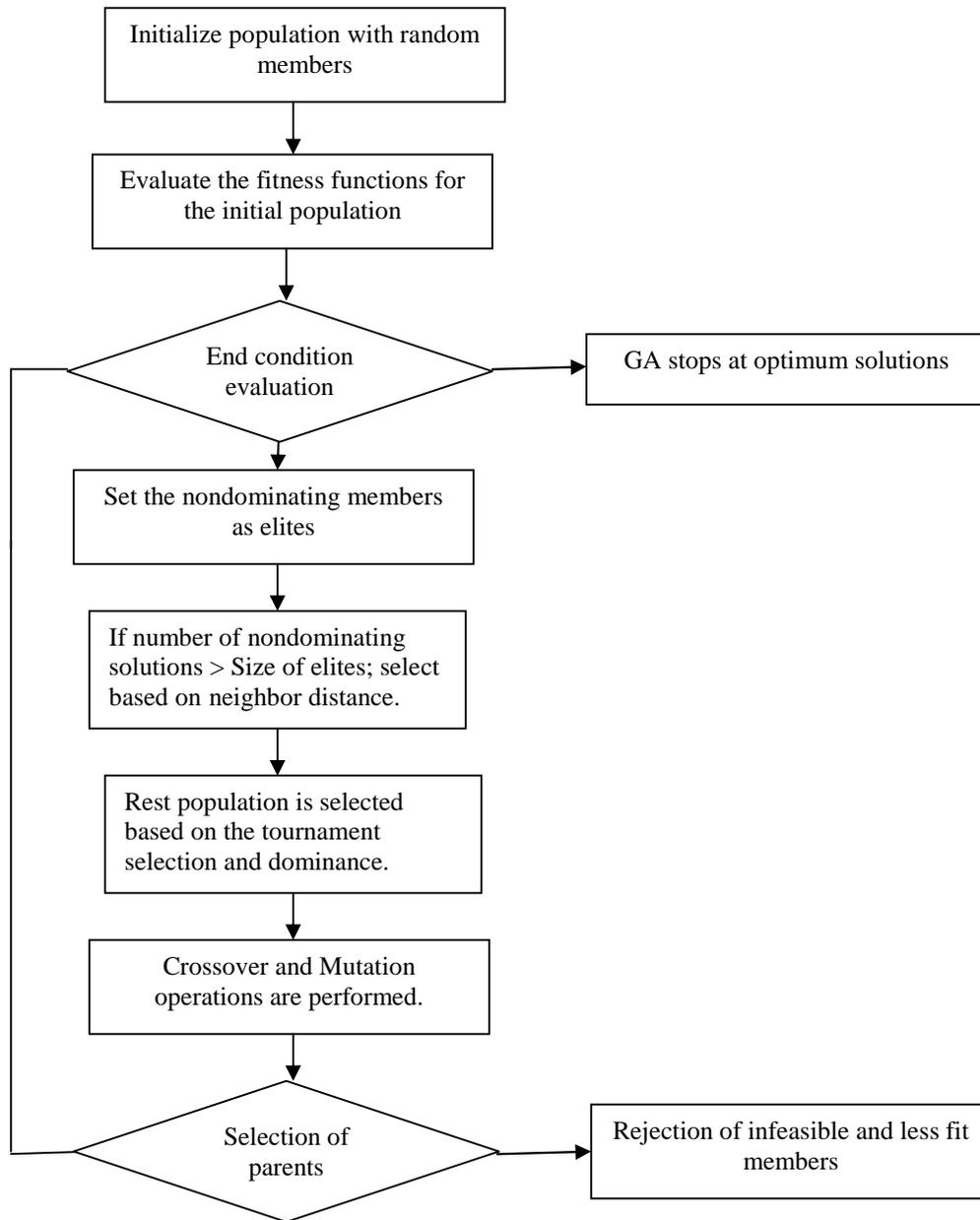


Figure 3.4. Flowchart of MultiObjective Genetic Algorithm (MOGA).

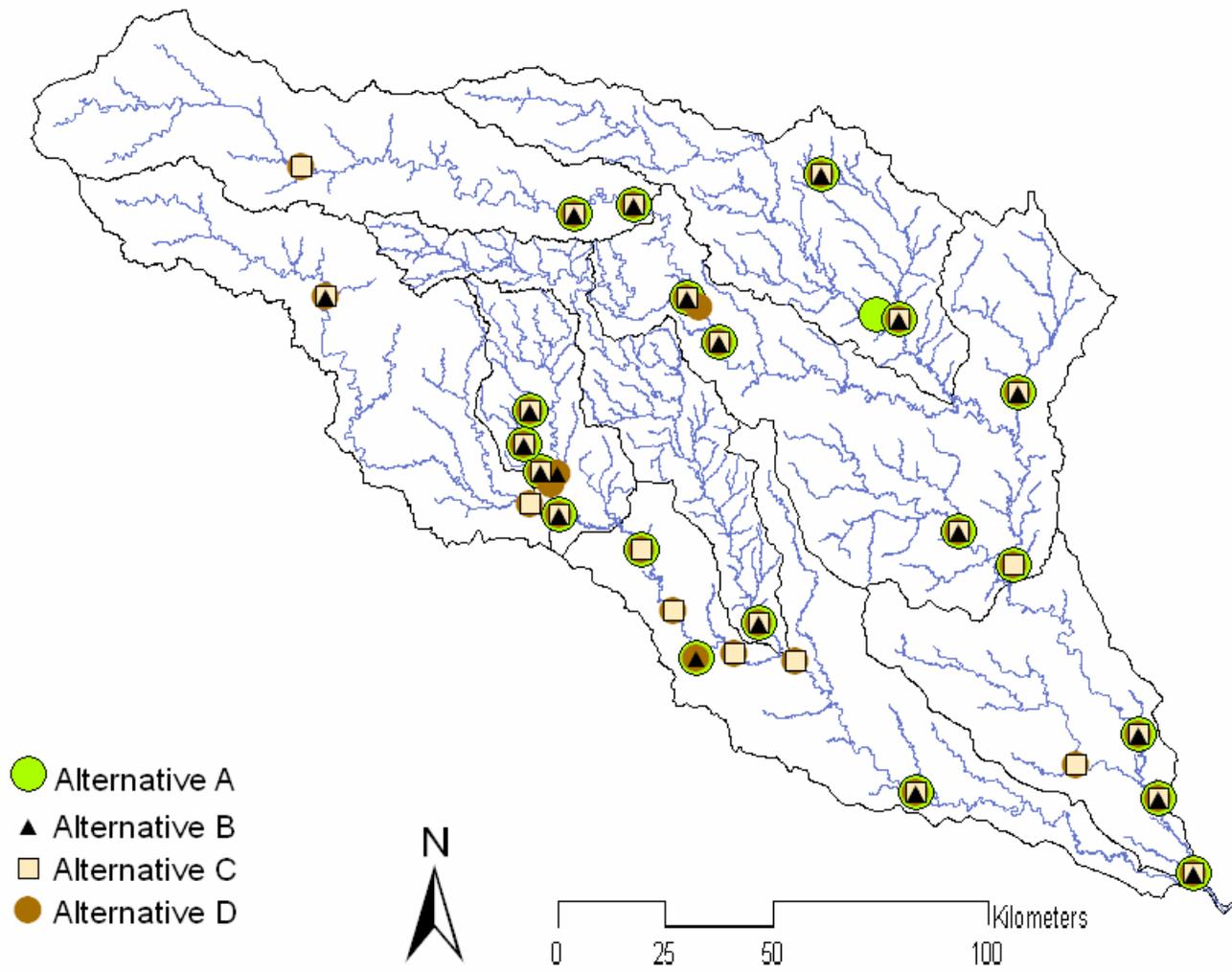


Figure 3.5. Four alternative sets of monitoring stations obtained after the application of MOGA.

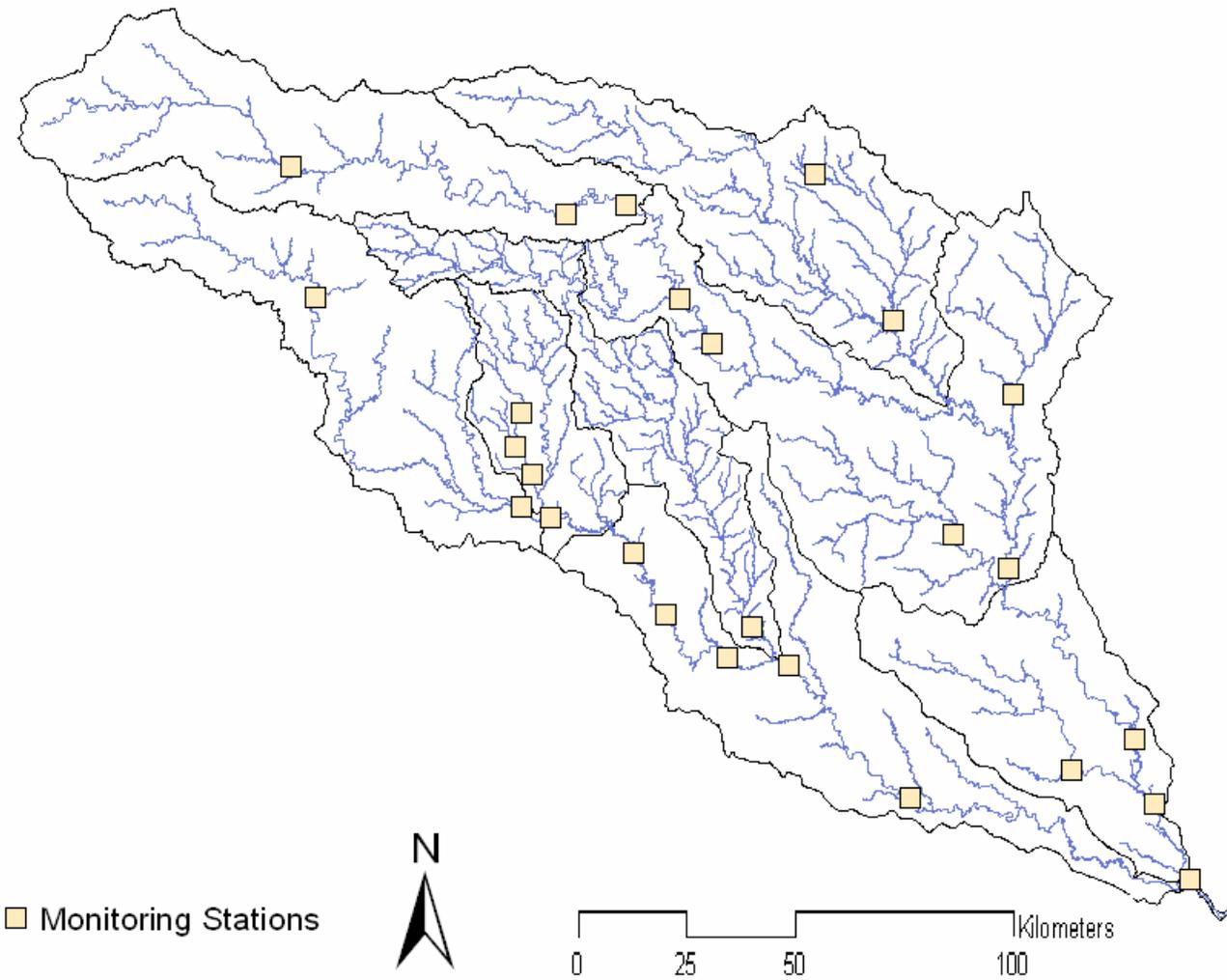


Figure 3.6. Selected set of monitoring stations (alternative C).

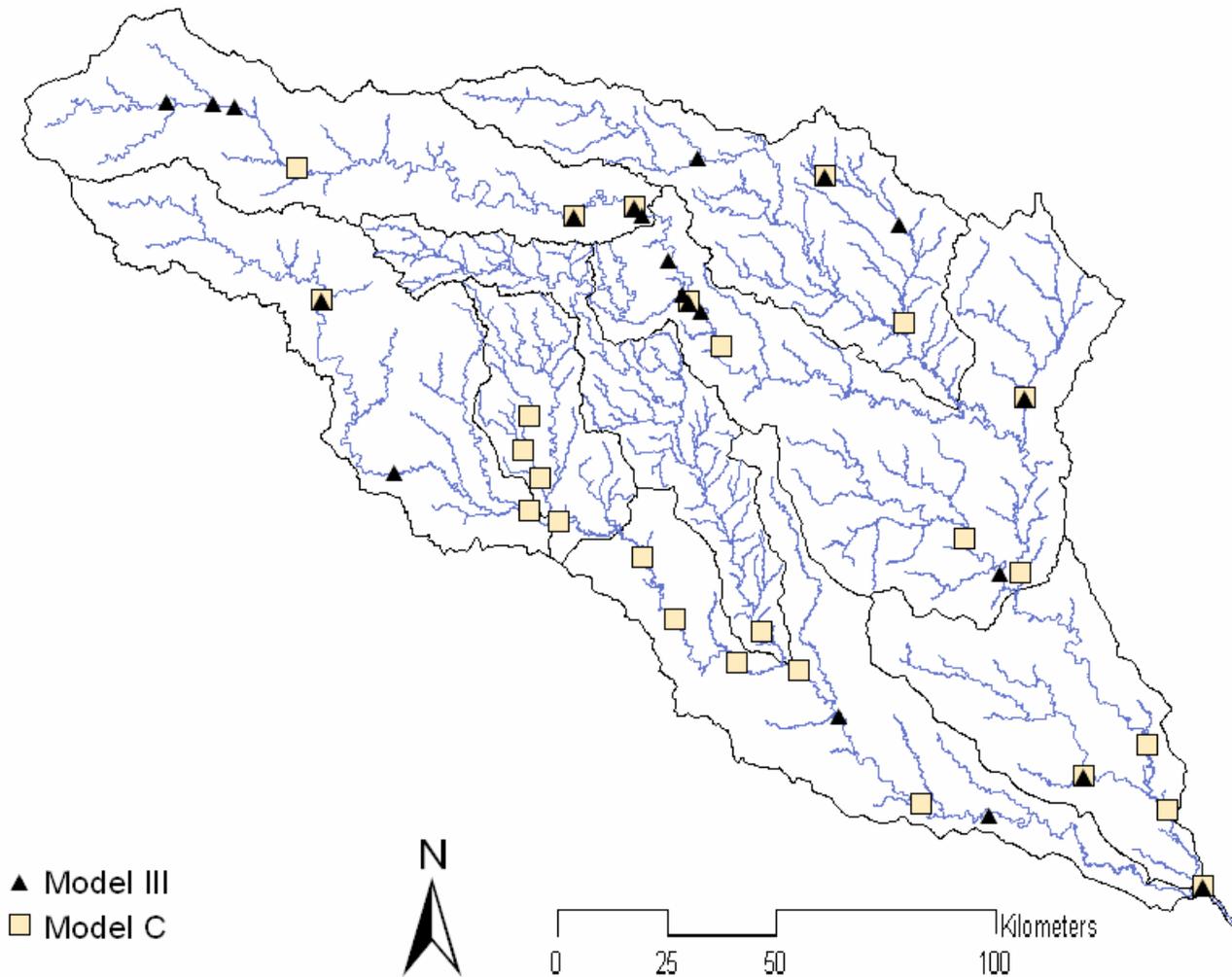


Figure 3.7. Locations of monitoring stations selected in Model C and Model III (Chapter II)

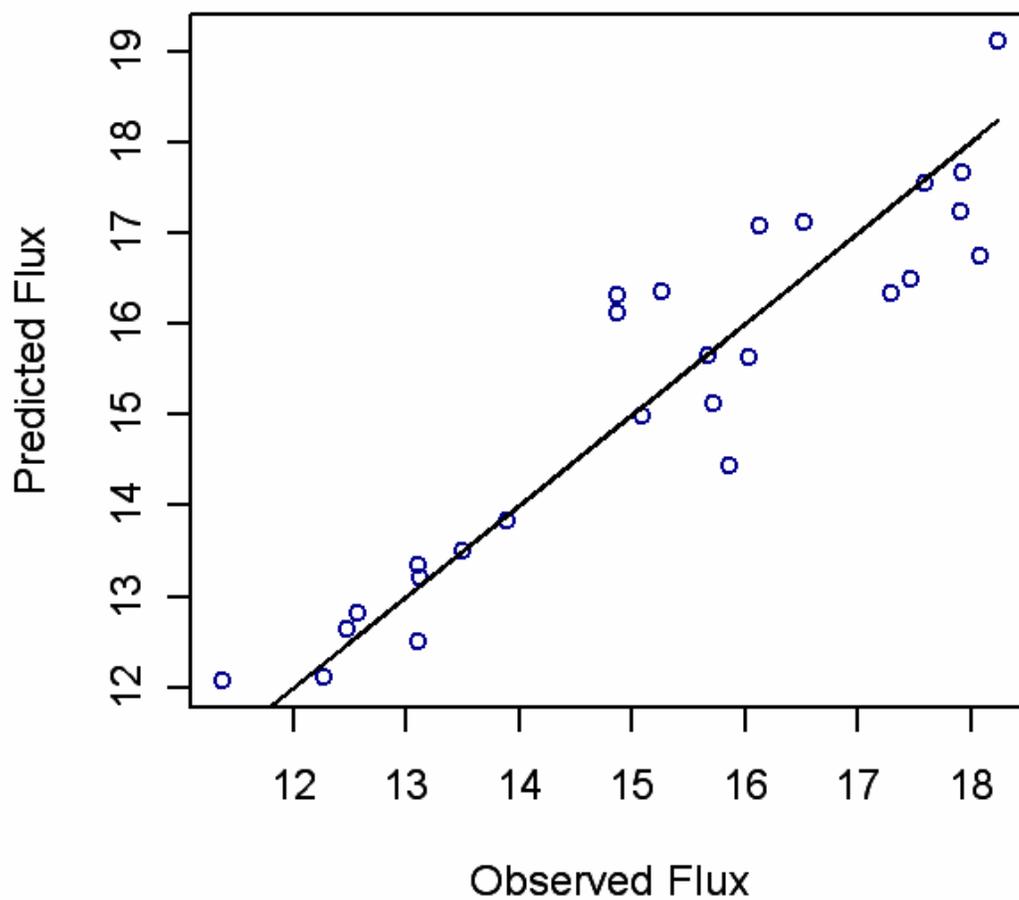


Figure 3.8. Relationship between the Natural Logarithm of observed (estimated mean annual flux in FLUXMASTER) and the predicted *E. coli* flux (SPARROW results) for Model C.

APPENDIX B
TABLES

Table 2.1. Coefficients and p-values (in parenthesis) of parameters and error statistics of the models for the selected sets of monitoring stations (for the comparison of p-values, level of significance is 0.15).

Model	I	II	III
Number of monitoring stations	56	35	21
Standard error/flux ratio <=	10	1	0.6
Sources			
Point sources flow (cubic meter yr ⁻¹)	0.03 (0.48)	0.03 (0.48)	
Pasture land (m ²)	20.58 (0.08)	9.30 (0.14)	
Forest land (m ²)		1.20 (0.26)	0.73 (0.06)
Urban land (m ²)	5.57 (0.13)	2.22 (0.20)	0.94 (0.12)
Delivery Factors			
Rainfall (m)	1.90 (0.37)		4.59 (0.24)
Temperature (°C)	1.17 (9.6 x10 ⁻⁶)	1.69 (1.9 x10 ⁻⁵)	2.41(4.1 x10 ⁻⁶)
Drainage density (km ⁻¹)	2.70 (2.3 x10 ⁻⁴)	2.58 (6.6 x10 ⁻³)	
Permeability (cm hr ⁻¹)	-0.01 (0.89)	-0.09 (0.23)	-0.08 (0.36)
Reservoir/Stream Decay Factors			
Areal hydraulic load (m yr ⁻¹)	58.06 (0.42)	36.49 (0.38)	19.81 (0.41)
Medium sized stream (0.02 < flow <= 0.13 m ³ s ⁻¹)	14.76(2.8 x10 ⁻³)	14.29(4.8 x10 ⁻³)	24.58(4.5x10 ⁻⁴)
Sum of Square Error (SSE)	119.20	46.71	18.27
Mean Square of Error (MSE)	2.54	1.79	1.30
Root Mean Square Error (RMSE)	1.59	1.34	1.14
Coefficient of determination (R²)	0.67	0.80	0.85

Table 2.2. Statistical indices (AIC, NSE and PBIAS) for all three models.

Model	I	II	III
Number of monitoring stations	56	35	21
Standard error/flux ratio <=	10	1	0.6
AIC	338.30	188.30	101.50
NSE	0.00	0.37	0.88
PBIAS (%)	58.00	44.10	28.00

Table 2.3. Parameter estimation for Model III using Bootstrap analysis.

		Nonparametric Bootstrap			Parametric Bootstrap		
		Bootstrap coefficient	Upper 90% confidence interval	Lower 90% confidence interval	Bootstrap coefficient	Upper 90% confidence interval	Lower 90% confidence interval
Model III with 21 monitoring stations	Parametric coefficients of Model III after recalibration						
Forest land (m ²)	0.34	0.49	0.53	0.45	0.35	0.37	0.33
Urban land (m ²)	0.52	0.62	0.67	0.57	0.53	0.57	0.50
Temperature (°C)	2.00	2.03	2.06	2.00	2.03	2.06	2.00
Medium sized stream (0.02 < flow <= 0.13 m ³ s ⁻¹)	17.65	20.03	20.81	19.25	17.68	18.19	17.16

Table 3.1. All objectives and statistical indices for alternatives and Model III.

	A	B	C	D	III
Number of monitoring Stations	21	20	26	30	21
Average of logarithmic mean annual fluxes for the selected monitoring stations	18.13	17.13	21.37	25.82	5.50
Sum of standard error to mean annual flux ratio for the selected monitoring stations	16.14	14.89	19.01	22.61	7.7
Akaike Information Criteria (AIC)	874.20	840.40	824.41	1249.20	101.50
Nash- Sutcliffe Efficiency (NSE)	-0.05	0.00	0.57	-0.03	0.88
Percent Bias (PBIAS)	52.70	67.00	3.20	79.13	28.00

Table 3.2. Coefficients and p-values (in parentheses) of the sources, land-water delivery and stream/ reservoir attenuation factors for four sets of monitoring stations (for the comparison of p-values, level of significance is 0.25).

Model	A	B	C	D
Selected monitoring stations	21	20	26	30
Sources				
Point sources discharge (cubic meter yr ⁻¹)			0.49 (0.22)	0.48 (0.42)
Pasture land (m ²)	13.98 (0.13)	7.29 (0.19)	0.19 (0.32)	36.74 (0.13)
Forest land (m ²)			1.41x10 ⁻⁵ (0.50)	
Urban land (m ²)	3.52 (0.22)	3.23 (0.25)	4.2x10 ⁻³ (0.32)	9.48 (0.26)
Delivery Factors				
Rainfall (m)		-5.03 (0.20)		
Temperature (°C)	1.75 (3.9x10 ⁻³)	1.74 (3.9x10 ⁻³)	-0.26 (0.23)	1.09 (0.02)
Drainage density (km ⁻¹)	2.19 (.01)	1.60 (0.13)	3.03 (2.3x10 ⁻⁵)	2.71 (3.6x10 ⁻³)
Permeability (cm hr ⁻¹)		0.14 (0.30)		-0.01 (0.89)
Reach slope (%)			-12.20 (6.5x10 ⁻³)	
Reservoir/Stream Decay Factors				
Areal hydraulic load (m yr ⁻¹)			30.18 (0.39)	218.85 (0.47)
Medium sized stream (0.02 < flow <= 0.13 m ³ s ⁻¹)	13.38 (.02)	14.99 (0.02)	13.52 (1.1x10 ⁻³)	17.12 (0.02)
Sum of Square Error	30.45	26.03	14.67	49.87
Root Mean Square Error (RMSE)	1.34	1.41	0.93	1.50
Coefficient of Determination	0.62	0.72	0.86	0.67
Degree of Freedom	5	7	9	8

Table 3.3. Coefficients and p-values (in parentheses) of the sources, land-water delivery and stream/ reservoir attenuation factors for the Model III and Model C. (for the comparison of p-values, level of significance is 0.25).

Model	III	C
Selected monitoring stations	21	26
Sources		
Point sources discharge ($\text{m}^3 \text{ yr}^{-1}$)		0.49 (0.22)
Pasture land (m^2)		0.19 (0.32)
Forest land (m^2)	0.73 (0.06)	1.41×10^{-5} (0.50)
Urban land (m^2)	0.94 (0.12)	4.2×10^{-3} (0.32)
Delivery Factors		
Rainfall (m)	4.59 (0.24)	
Temperature ($^{\circ}\text{C}$)	2.41 (4.1×10^{-6})	-0.26 (0.23)
Drainage density (km^{-1})		$3.03 (2.3 \times 10^{-5})$
Permeability (cm h^{-1})	-0.08 (0.36)	
Reach slope (%)		$-12.20 (6.5 \times 10^{-3})$
Reservoir/ Stream Decay Factors		
Areal hydraulic load (m yr^{-1})		30.18 (0.39)
Medium sized stream ($0.02 < \text{flow} \leq 0.13 \text{ m}^3 \text{ s}^{-1}$)	13.38 (0.02)	$13.52 (1.1 \times 10^{-3})$
Sum of Square Error	18.27	14.67
Root Mean Square Error (RMSE)	1.14	0.93
Coefficient of Determination	0.85	0.86
Degree of Freedom	6	9

VITA

Name: Deepti

Address: 7721 Acrocomia Dr, Hanover, MD 21076.

E-mail: deepti202jais@gmail.com

Education:

- B.Tech., Agricultural Engineering, Punjab Agricultural University, Ludhiana, India, 2002, GPA 7.7/10.0.
- M.Tech., Agricultural Engineering, Indian Institute of Technology (I.I.T.), Kharagpur, India, 2004, GPA 9.2/10.0.
- M.S., Biological and Agricultural Engineering, Texas A&M University, College Station, U.S., GPA 4.0/ 4.0.

Professional Experience:

- Project Executive, Self Reliant Initiative through Joint Action (SRIJAN) 2004-2006.

Honors and Awards:

- Texas Water Resources Institute Mills Scholarship 2007-08.
- Scholarship for Study Abroad, K C Mahindra Scholarship Trust, 2006.
- GATE (Graduate Aptitude Test in Engineering) Scholarship, 2002-2004.
- Merit Scholarship, Punjab Agricultural University, 1998-2002.

Grants:

- U.S. Geological Survey Research Grant 2008-09.
- Graduate Student Research and Presentation Grant, Spring, 2008.

Certification:

- Graduate Certificate in Geographic Information Systems (GIS), 2008.
- Graduate Teaching Assistant Certificate, 2008.