TEXAS WATER RESOURCES: VULNERABILITY FROM CONTAMINANTS

A Dissertation

by

DIPANKAR DWIVEDI

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

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Binayak P. Mohanty
Committee Members,	Bruce J. Lesikar
	Raghupathy Karthikeyan
	Yalchin Efendiev
Head of Department,	Stephen W. Searcy

December 2012

Major Subject: Biological and Agricultural Engineering

ABSTRACT

Numerical models of flow and transport are commonly applied for the sustainable management of water resources and for the selection of appropriate remediation techniques. However, these numerical models are not always accurate due to uncertain parameters and the disparity of scales across which observations are made, hydrological processes occur, and modeling is conducted. The modeling framework becomes further complex because hydrologic processes are coupled with chemical and biological processes. This dissertation focuses on the most widespread contaminants of surface and ground water, which are *E. coli* and nitrate, respectively. Therefore, this research investigates the linkages between bio-chemical and hydrologic processes for *E. coli* transport, explores the spatio-temporal variability of nitrate, quantifies uncertainty, and develops models for both *E. coli* and nitrate transport that better characterize these biogeochemical linkages.

A probabilistic framework in the form of Bayesian Neural Networks (BNN) was used to estimate *E. coli* loads in surface streams and was compared with a conventional model LOADEST. This probabilistic framework is crucial when water quality data are scarce, and most models require a large number of mechanistic parameters to estimate *E. coli* concentrations. Results indicate that BNN provides better characterization of *E. coli* at higher loadings. Results also provide the physical, chemical, and biological factors that are critical in the estimation of *E. coli* concentrations in Plum Creek, Texas.

To explore model parameters that control the transport of *E. coli* in the groundwater (GW) and surface water systems, research was conducted in Lake Granbury, Texas. Results highlight the importance of flow regimes and seasonal variability on *E. coli* transport.

To explore the spatio-temporal variability of nitrate across the Trinity and Ogallala aquifers in Texas, an entropy-based method and a numerical study were employed. Results indicate that the overall mean nitrate-N has declined from 1940 to 2008 in the Trinity Aquifer as opposed to an increase in the Ogallala Aquifer. The numerical study results demonstrate the effect of different factors like GW pumping, flow parameters, hydrogeology of the site at multiple spatial scales.

To quantify the uncertainty of nitrate transport in GW, an ensemble Kalman filter was used in combination with the MODFLOW-MT3DMS models. Results indicate that the EnKF notably improves the estimation of nitrate-N concentrations in GW.

A conceptual modeling framework with deterministic physical processes and stochastic bio-chemical processes was devised to independently model *E. coli* and nitrate transport in the subsurface. Results indicate that model structural uncertainty provides useful insights to modeling *E. coli* and nitrate transport.

DEDICATION

This dissertation is dedicated to my parents.

ACKNOWLEDGEMENTS

It is a great pleasure to thank everyone who helped me write my dissertation successfully. I am truly indebted and thankful to my family, advisor, professors, and friends for their encouragement and support during my graduate studies.

I would like to acknowledge Dr. Binayak P. Mohanty for his compassion, patience, and most importantly, his mentorship during my graduate studies at Texas A&M University. His guidance was foremost in providing a wonderful and scholastic experience congruous with my long-term career goals. I am immensely grateful to my doctoral committee members Dr. Raghupathy Karthikeyan, Dr. Bruce J. Lesikar, and Dr. Yalchin Efendiev for their advice, comments, and support during my research work.

I would like to thank the Biological and Agricultural Engineering Department at Texas A&M University. In particular, I would like to thank Dr. Vijay P. Singh, Dr. Cady Engler and the rest of the department for their valuable inputs, discussion, and accessibility. I would also like to extend my gratitude to Dr. Marian Eriksson for enhancing my programming skills. Additionally, I acknowledge the financial support of National Institute for Environmental Health Sciences grant and Texas Water Resources Institute.

I am also thankful to many former and current colleagues, especially Bhavna, Maheshwari, Raghu, Yongchul, Zhenlei, Amor, and Narendra for their assistance and

suggestions. Also, Sonya, Hemant, Ashok, Paritosh, Abhishek, and many others deserve a mention for my possessed graduate experience.

Last, but definitely not the least, gratitude goes to my parents for their unwavering love and patience. I am greatly indebted to my sisters for their enthusiastic support of all my accomplishments.

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iv

ACKNOWLEDGEMENTS v

TABLE OF CONTENTS vii

LIST OF FIGURES xi

LIST OF TABLES xx

CHAPTER

I GENERAL INTRODUCTION 1

1.1 Problem statement 1
1.2 Motivation 3
1.3 Research objectives 3

II ESTIMATING *E. COLI* LOADS IN STREAMS BASED ON VARIOUS PHYSICAL, CHEMICAL, AND BIOLOGICAL FACTORS 7

2.1 Synopsis 7
2.2 Introduction 8
2.3 Study area description and data availability 12
2.4 Methodology 15
2.4.1 Bayesian neural networks 15
2.4.2 Radial basis function (RBF) 17
2.4.3 Load estimator 19
2.4.4 Exhaustive feature selection 20
2.4.5 Model performance 21
2.4.6 Principal component analysis 22
2.4.7 Uncertainty analysis 23
2.5 Results and discussion 24
2.5.1 Identification of the key factors responsible for the *E. coli* loads in Plum Creek 24

Page

CHAPTER

2.5.2 Estimation of *E. coli* loads 29
2.5.3 Uncertainty analysis 33
2.6 Summary and conclusions 36

III *E. COLI* FATE AND TRANSPORT BELOW SUBSURFACE SEPTIC TANKS IN THE LAKE GRANBURY AREA 38

3.1 Synopsis 38
3.2 Introduction 39
3.3 Study area description and data availability 41
3.4 Methodology 42
3.4.1 *E. coli* transport modeling in the unsaturated zone 43
3.4.2 Physical domain setup for the unsaturated zone 44
3.4.3 Sensitivity analysis in the unsaturated zone 47
3.4.4 Physical domain setup for the saturated zone 48
3.4.5 Calibration and sensitivity analysis in the saturated zone 49
3.5 Results 51
3.5.1 *E. coli* transport in the unsaturated zone 52
3.5.2 Sensitivity analysis in the unsaturated zone 58
3.5.3 *E. coli* data and moving average 59
3.5.4 *E. coli* transport in the saturated zone 63
3.5.5 Model calibration and sensitivity analysis in the saturated zone 63
3.5.6 Simulation results of MODFLOW and MT3DMS 64
3.6 Summary and conclusions 66

IV ENTROPY-BASED ANALYSIS FOR SPATIO-TEMPORAL VARIABILITY OF NITRATE IN TEXAS AQUIFERS ACROSS MULTIPLE SCALES 68

4.1 Synopsis 68
4.2 Introduction 69
4.3 Methodology 75
4.3.1 Estimation of variability 75
4.3.2 Entropy 77
4.3.3 Mutual information 79
4.3.4 Normalized risk index 81
4.3.5 Hurst exponent 82
4.3.6 Numerical study 86
4.3.7 Application of the modeling framework 87

Page

CHAPTER

4.3.8 Aquifer descriptions, layers, grid size, model boundary, and parameters....88
4.3.9 The Ogallala Aquifer....88
4.3.10 The Trinity Aquifer....90
4.3.11 On-ground nitrogen loading....92
4.3.12 Model assumptions and limitations....95
4.4 Data....96
4.5 Results and discussion....100
4.5.1 Spatial variability of nitrate-N....100
4.5.2 Spatial variability of Cl^- and mutual information sharing between nitrate-N and Cl^-....108
4.5.3 Trend and persistence of variability of nitrate-N....111
4.5.4 Normalized risk index of nitrate-N....112
4.5.5 Temporal variability of nitrate-N....113
4.6 Summary and conclusions....116

V ADDRESSING UNCERTAINTY IN NITRATE TRANSPORT IN GROUNDWATER USING THE ENSEMBLE KALMAN FILTER....118

5.1 Synopsis....118
5.2 Introduction....119
5.3 Methodology....122
5.3.1 Data assimilation theory and parameter estimation....123
5.3.2 The Kalman filter....124
5.3.3 The ensemble Kalman filter (EnKF)....126
5.3.4 Parameter update using the EnKF....128
5.3.5 Application of the modeling framework....130
5.3.6 Synthetic 2D transient state GW flow model....131
5.3.7 The Ogallala Aquifer GW flow model....132
5.3.8 Numerical study....134
5.4 Results and discussion....134
5.4.1 Parameter optimization using the EnKF algorithm....135
5.4.2 DA algorithm in the synthetic 2D Aquifer....143
5.4.3 DA algorithm in the synthetic 2D Aquifer....146
5.4.4 Root mean square error in the EnKF predictions....148
5.5 Summary and conclusions....151

VI A HYBRID DETERMINISTIC-STOCHASTIC MODELING FRAMEWORK FOR CONTAMINANT TRANSPORT IN THE SUBSURFACE....152

Page

CHAPTER

6.1 Synopsis 152
6.2 Introduction 153
6.3 Methodology 157
6.3.1 Factors affecting nitrogen transport through soils and their mathematical formulations 158
6.3.2 Factors affecting *E. coli* transport through soils and their mathematical formulations 162
6.3.3 Stochastic formulation 164
6.3.4 Model testing 166
6.4 Results and discussion 169
6.4.1 Nitrate transport in one-dimensional soil column 170
6.4.2 Nitrate transport with perturbations to soil temperature 176
6.4.3 Nitrate transport with perturbations to soil pH 178
6.4.4 *E. coli* transport in one-dimensional soil column 181
6.4.5 *E. coli* transport with perturbations to growth and die-off rates 184
6.4.6 *E. coli* transport with perturbations to chemotaxis 186
6.5 Summary and conclusions 187

VII GENERAL CONCLUSIONS 189

REFERENCES 193

LIST OF FIGURES

Page

Figure 1.1: A graphical distribution of water on the Earth and the most ubiquitous contaminants in water resources. 2

Figure 1.2: A schematic representation of process integration in the modeling framework, and analysis of parameter sensitivity and uncertainty in the output. 4

Figure 2.1: The graphical structure of BNN representing cause-and-effect relationship between system variables (water quality parameters) and the *E. coli* loads. The Radial Basis layer is the hidden layer, which uses the transfer function f (thin plate spline), and the linear layer is the output layer, which uses the transfer function f (linear function). The transfer function f establishes a relationship between inputs and outputs, in case of estimation of *E. coli* loads, and thin plate splines work better than other transfer functions (Gaussian or r4 functions). 12

Figure 2.2: Map showing the Guadalupe river basin in east central Texas and the station ID 12645 in Plum Creek (Map modified from http://www.gbra.org/CRP). 13

Figure 2.3: Pareto diagram of principal components shows the percentage explained by each component. The first three components explain 70% of variation in the data set. 25

Figure 2.4: The principal component analysis (PCA) of the dataset is shown above. First principal component is projected against the second principal component. The Euclidean norm (length) reflects the relative importance of different factors. The biplot of the PCA is shown in (A). (B) and (C) demonstrate factors loadings on first and second principal components, respectively. 26

Figure 2.5: Measured and simulated loads of *E. coli* by the BNN model in Plum Creek are shown above. The simulations were tested by three-fold cross validation. Months on X-axis represent the sequence of random splits (not the sequence of a timeseries). 30

Figure 2.6: Measured and simulated loads of *E. coli* by the LOADEST model in Plum Creek are shown above. *E. coli* loads are presented for the three random splits, which were used for three-fold cross validation of the BNN model. Months on X-axis represent the sequence of random splits (not the sequence of a timeseries). 31

Page

Figure 2.7: The cumulative density functions of the observed *E. coli* loads, estimated *E. coli* loads by the BNN, and estimated *E. coli* loads by the LOADEST model are presented above. There are three regions in the figure. Region A signifies that the LOADEST model is able to estimate *E. coli* loads better. In region B and C, the BNN model predicts better, however in region B, both the LOADEST and BNN models overestimate *E. coli* loads. 32

Figure 2.8: Probability distributions of low and high *E. coli* loads of each random split by the BNN model are presented above. PDFs show that uncertainty in the estimation has a small bias but a large variance. These uncertainties mainly stem from flow and other water quality data. 35

Figure 3.1: Plan view of the model domain (m) for (1) MODFLOW/MT3DMS and (2) cross sectional view of the model domain (m) for HYDRUS 2-D along A-A' transect (Map courtesy: Google Earth) 46

Figure 3.2: Precipitation and evapo-transpiration in the Lake Granbury area for 2003 and 2004 demonstrate that June to September have high evapo-transpiration and lower precipitation as compared to October to May (Data: TWDB). 52

Figure 3.3: The breakthrough curves of *E. coli* concentration at observation nodes 1 and 2 demonstrate the impact of scenarios 1 and 2 on *E. coli* transport in the subsurface. Scenario 1 represents summer and scenario 2 represents winter. 53

Figure 3.4: The horizontal (1.5 m below land surface) and vertical profiles (across septic tanks near the Lake) of pressure head demonstrate the impacts of scenario 1 (summer) and 2 (winter) on the extent of the saturated and unsaturated zones. 55

Figure 3.5: The horizontal (1.5 m below land surface) and vertical profiles (across septic tanks near the Lake) of *E. coli* concentration demonstrate the impacts of scenario 1 (summer) and 2 (winter) on *E. coli* transport in the subsurface. 57

Figure 3.6: The (A) autocorrelation (ACF) and (B) partial autocorrelation (PACF) functions of *E. coli* concentrations in the lake describe the seasonality in the observed *E. coli* concentrations. The ACF and PACF demonstrate a repeating pattern after 7 and 10 months. 60

Figure 3.7: The 3-month moving average of *E. coli* concentrations in the canal shows the seasonality in *E. coli* concentrations. 61

Page

Figure 3.8: (A) The Marquardt Lambda with parameter estimation attempts exhibits the performance of PEST during the calibration of MODFLOW/MT3DMS. (B) The change in the sensitivity of parameters with the number of iteration implies that the sensitive parameter is more accurately estimated than other parameters. 62

Figure 3.9: Estimated *E. coli* concentrations from the calibrated MODFLOW and MT3DMS models and observed *E. coli* concentrations in the lake show that 85% observed *E. coli* values fall within the confidence limits of ±σ (standard deviation). Months from July 2002 to January 2006 (0-48) and months from February 2006 to August 2010 (49-98) were used for calibration and validation, respectively. 65

Figure 4.1: Map showing the Trinity Aquifer and its outcrop area. The Trinity Aquifer is a sandstone-carbonate rock aquifer and partly confined. The Trinity Aquifer consists of early Cretaceous age formations of the Trinity Group, from the Red River in North Texas to the Medina River of South-Central Texas. Major Rivers in the study area include: the Red, Trinity, Brazos, Colorado, San Antonio, Guadalupe, and Medina rivers. 73

Figure 4.2: Map showing the Ogallala Aquifer, which is an unconfined aquifer and primarily composed of sand, gravel, clay, and silt deposited during the Tertiary Period. Major Rivers in the study area include: the Canadian, Prairie Dog Town Fork Red River, and Colorado Rivers. The part where the Canadian River wears away demarcates the Ogallala into two separate flow systems referred to as the Northern and Southern High Plains. The Ogallala and the underlying Cretaceous, Jurassic, and Triassic formations are hydrologically connected in many areas. These hydrologic connections also exist between the overlying Quaternary Blackwater Draw Formation wherever present. 74

Figure 4.3: A schematic diagram of the data flow and analyses. Entropy approach (A) was utilized to find the controlling factors at different scales for the variability of nitrate-N. The results were verified by numerical modeling (B). 76

Figure 4.4: Plan view of model domains (km) demonstating wells and boundary conditions (B.C.). The (A) Ogallala and (B) Trinity aquifers were simulated at a discretization of 1 km x 1 km grid size. 84

Figure 4.5: Land use maps for the (A) Trinity and (B) Ogallala aquifers. 93

Page

Figure 4.6: Maps showing (A) average annual atmospheric nitrate-N deposition based on National Atmospheric Deposition Program (NADP) data, (B) spatial distribution of inorganic nitrogen fertilizer application, (C) spatial distribution of organic fertilizer (manure) application, (D) distribution of concentrated animal feeding operations (CAFOs) based on data from TCEQ, TIAER, and USGS and permitted sludge application based on data from TCEQ, and (E) average soil profile organic matter derived from STATSGO database (From *Scanlon et al.*, 2003). 94

Figure 4.7: Map showing average annual precipitation in Texas. Precipitation varies widely across Texas. In the Trinity Aquifer, precipitation varies from a low of 60 cm per year in the west to a high of 110 cm on the eastern portion of the outcrop of the aquifer, and in the Ogallala Aquifer, it varies from a low of 20 cm per year in the west and southwest to a high of 50 cm on the eastern portion. 97

Figure 4.8: The nitrate-N concentrations (1940 to 2008) are shown in the (A) Trinity and (B) Ogallala aquifers. In both the aquifers, three different spatial scales namely fine (2 km×2 km), intermediate (10 km×10 km), and coarse (100 km×100 km) are used to analyze the spatial and temporal variability of nitrate-N. The gridded reconstruction of water quality dataset involved all wells falling in a particular grid (fine, or intermediate, or coarse) as one dataset (or time series). There are 25% and 16% of total wells exceeding 10 mg/L of nitrate-N in the Ogallala and Trinity aquifers, respectively. 98

Figure 4.9: Map shows the contour plots of normalized marginal entropy values of observed nitrate-N in the Trinity and Ogallala aquifers at three spatial scales: fine grid (A and D), intermediate grid (B and E), and coarse grid (C and F), respectively. The higher NME values indicate lower entropy values. 101

Figure 4.10: A numerical study using Visual MODFLOW in the Ogallala Aquifer was conducted to verify the effect of different factors on spatial variability of nitrate-N at multiple scales. Probability distribution functions of the normalized marginal entropy values (%) of nitrate-N in the Ogallala Aquifer are plotted for pumping and hydraulic conductivity scenarios. The results demonstrate that the spatial variability of nitrate-N is controlled by the effect of pumping at the fine grid (A) as compared to intermediate (B) or coarse (C) grids; hydraulic conductivity also plays a key role at the small grid (D) as compared to intermediate (E) or coarse (F) grids. The higher NME values indicate lower entropy values. 103

Page

Figure 4.11: A numerical study using Visual MODFLOW in the Trinity Aquifer was conducted to verify the effect of different factors on spatial variability of nitrate-N at multiple scales. Probability distribution functions of the normalized marginal entropy values (%) of nitrate-N in the Trinity Aquifer are plotted for pumping and hydraulic conductivity scenarios. The results demonstrate that the spatial variability of nitrate-N is controlled by the effect of pumping at the fine grid (A) as compared to intermediate (B) or coarse (C) grids; hydraulic conductivity also plays a key role at the small grid (D) as compared to intermediate (E) or coarse (F) grids. The higher NME values indicate lower entropy values. 104

Figure 4.12: A numerical study using Visual MODFLOW was conducted in the Ogallala Aquifer to verify the effect of different factors on spatial variability of nitrate-N at multiple scales. Probability distribution functions of the normalized marginal entropy values (%) of nitrate-N in the Ogallala Aquifer are plotted for river and layering scenarios. The results demonstrate that the spatial variability of nitrate-N is controlled by the presence of a river at the intermediate grid (B) as compared to fine (A) or coarse (C) grids; geology (layering vs. no layering) plays a key role at the coarse grid (F) as compared to fine (D) or intermediate (E) grids. The higher NME values indicate lower entropy values. 105

Figure 4.13: A numerical study using Visual MODFLOW was conducted in the Trinity Aquifer to verify the effect of different factors on spatial variability of nitrate-N at multiple scales. Probability distribution functions of the normalized marginal entropy values (%) of nitrate-N in the Trinity Aquifer are plotted for river and layering scenarios. The results demonstrate that the spatial variability of nitrate-N is controlled by the presence of a river at the intermediate grid (B) as compared to fine (A) or coarse (C) grids; geology (layering vs. no layering) plays a key role at the coarse grid (F) as compared to fine (D) or intermediate (F) grids. The higher NME values indicate lower entropy values. 106

Figure 4.14: Map shows the contour plots of entropy values of Cl^- in the Trinity and Ogallala aquifers at three spatial scales: fine grid, intermediate grid, and coarse grid, respectively..... 107

Page

Figure 4.15: Map presents the Normalized Mutual Information between nitrate-N and Cl^- in the Trinity and Ogallala aquifers at three spatial scales: fine (A, D), intermediate (B, E), and coarse (C, F) grids, respectively. The green color indicates the codependency of the two variables (nitrate-N and Cl^-). 108

Figure 4.16: The Hurst exponent (H) of nitrate-N in the Trinity Aquifer at the small (A), intermediate (B), and coarse (C) scales, as well as in the Ogallala Aquifer at the small (D), intermediate (E), and coarse (F) scales. The Hurst exponent varies from 0 to 1, 1 being the most persistent, 0.5 being the random changes in the trend, and 0 being the most anti-persistent (or mean reversion) behavior. 110

Figure 4.17: Probability distribution (PDF) of the Hurst exponent of nitrate-N in the (A) Trinity and (B) Ogallala aquifers across different spatial scales (fine, intermediate, and coarse). 111

Figure 4.18: Decadal analysis (mean and standard deviation) of nitrate-N in the Trinity Aquifer across (A) fine, (B) intermediate, (C) coarse scales, and in the Ogallala Aquifer across (D) fine, (E) intermediate, and (F) coarse scales. 114

Figure 4.19: Decadal analysis (% samples having nitrate-N more than 10 mg/ L and NME) of nitrate-N in the Trinity Aquifer across (A) fine, (B) intermediate, (C) coarse scales, and in the Ogallala Aquifer across (D) fine, (E) intermediate, and (F) coarse scales. 115

Figure 5.1: The data assimilation algorithm is presented using the ensemble Kalman filter in a GW and nitrate-N transport modeling framework. The forecast stage consists of generating the initial ensemble of K (recharge, hydraulic conductivity, or decay coefficient individually) fields and the resulting simulated concentration fields. X is the state, K is the Kalman gain, w is the noise, and *d* is the binary matrix. 130

Figure 5.2: Conceptual model domain showing observation well and boundary conditions (B.C.) for the transient GW flow problem..... 131

Figure 5.3: Plan view of the Ogallala Aquifer shows the observation well, forcing, and boundary conditions (B.C.). A larger domain (10,000 m by 10,000 m) was selected to import boundary condition in the problem domain to reduce the edge effect. 132

Figure 5.4: The temporal evolution of the log hydraulic conductivity, first order decay coefficient, and recharge fields for the synthetic and Ogalalla aquifers shows the asymptotic behavior. 136

Page

Figure 5.5: The temporal evolution of the log hydraulic conductivity fields (m/sec) when (A) 2, (B) 10, (C) 50, and (D) 100 measurements of nitrate-N concentrations are assimilated in the synthetic Aquifer. 137

Figure 5.6: The temporal evolution of the log hydraulic conductivity fields (m/sec) when (A) 2, (B) 10, (C) 50, and (D) 100 measurements of nitrate-N concentrations are assimilated in the Ogallala Aquifer. 138

Figure 5.7: The temporal evolution of the log hydraulic conductivity fields (m/sec) when (A) 2, (B) 10, (C) 50, and (D) 100 measurements of nitrate-N concentrations are assimilated in the Ogallala Aquifer. 140

Figure 5.8: The reference recharge and assimilated recharge values for each month (from 1996 to 2005) in the synthetic Aquifer when 100 measurements of nitrate-N concentrations are assimilated. 141

Figure 5.9: The reference recharge and assimilated recharge values for each month (from 1996 to 2005) in the Ogallala Aquifer when 100 measurements of nitrate-N concentrations are assimilated. 142

Figure 5.10: Estimation of hydraulic heads and nitrate-N concentrations in the synthetic 2D problem using the EnKF algorithm with 100 ensembles. Estimation of (A) the ensemble hydraulic heads by updating the first order decay coefficient; (B) the ensemble hydraulic heads by updating the recharge; (C) the ensemble hydraulic heads by updating the hydraulic conductivity; (D) the ensemble nitrate-N concentrations by updating the first order decay coefficient; (E) the ensemble nitrate-N concentrations by updating the recharge; and (F) the ensemble nitrate-N concentrations by updating the hydraulic conductivity. 144

Figure 5.11: Estimation of hydraulic heads and nitrate-N concentrations in the Ogallala Aquifer using the EnKF algorithm with 100 ensembles. Estimation of (A) the ensemble hydraulic heads by updating the first order decay coefficient; (B) the ensemble hydraulic heads by updating the recharge; (C) the ensemble hydraulic heads by updating the hydraulic conductivity; (D) the ensemble nitrate-N concentrations by updating the first order decay coefficient; (E) the ensemble nitrate-N concentrations by updating the recharge; and (F) the ensemble nitrate-N concentrations by updating the hydraulic conductivity. 147

Page

Figure 5.12: Root mean square error (RMSE) in estimating ensemble hydraulic heads in the synthetic 2D problem by updating the (A) first order decay coefficient; (B) recharge; (C) hydraulic conductivity; and the RMSE in estimating ensemble nitrate-N concentrations by updating the (D) first order decay coefficient; (E) recharge; and (F) hydraulic conductivity. 149

Figure 5.13: Root mean square error (RMSE) in estimating ensemble hydraulic heads in the Ogallala Aquifer by updating the (A) first order decay coefficient; (B) recharge; (C) hydraulic conductivity; and the RMSE in estimating ensemble nitrate-N concentrations by updating the (D) first order decay coefficient; (E) recharge; and (F) hydraulic conductivity. 150

Figure 6.1: A conceptual framework showing some key processes for (A) nitrate transport and (B) *E. coli* transport and deposition behavior in the subsurface. 159

Figure 6.2: Nitrate transport with three-species nitrification chain is shown at (A) observation node 1, (B) observation node 2, (C) observation node 3, and (D) observation node 4. Y-axis represents the breakthrough concentration (C/C_0). Solid lines represent the analytical solution, and dotted lines represent the deterministic numerical solution. 171

Figure 6.3: Nitrate transport with three-species nitrification chain is shown at (A) observation node 1, (B) observation node 2, (C) observation node 3, and (D) observation node 4. Y-axis represents the breakthrough concentration (C/C_0). Solid white lines represent the analytical solution, and other lines represent the hybrid deterministic-stochastic numerical solution. 172

Figure 6.4: The probability distribution functions (PDFs) of the root mean square error (RMSE) of predicted (A) ammonium, (B) nitrite, and (C) nitrate concentrations at observation node 4 against the analytical solutions. The PDFs demonstrate the uncertainty in the predictions. 174

Figure 6.5: Nitrate transport with perturbations to soil temperature at (A) observation node 1, (B) observation node 2, (C) observation node 3, and (D) observation node 4. Y-axis represents the breakthrough concentration (C/C_0) 176

Page

Figure 6.6: Nitrate transport with perturbations to soil pH at (A) observation node 1, (B) observation node 2, (C) observation node 3, and (D) observation node 4. Y-axis represents the breakthrough concentration (C/C_0)..... 178

Figure 6.7: Transport and deposition behavior of *E. coli* with attachment-detachment model is shown at observation nodes 1, 2, 3 and 4. Y-axis represents the breakthrough concentration (C/C_0)..... 180

Figure 6.8: The probability distribution functions (PDFs) of the root mean square error (RMSE) of predicted *E. coli* concentrations at A) observation node 1, (B) observation node 2, (C) observation node 3, and (D) observation node 4 against the numerical solution. The PDFs demonstrate the uncertainty in the predictions..... 182

Figure 6.9: Transport and deposition behavior of *E. coli* with perturbations to growth and die-off rates is shown at observation nodes 1, 2, 3 and 4. Y-axis represents the breakthrough concentration (C/C_0)..... 184

Figure 6.10: Transport and deposition behavior of *E. coli* with perturbations to chemotaxis is shown at observation nodes 1, 2, 3 and 4. Y-axis represents the breakthrough concentration (C/C_0). 185

LIST OF TABLES

Page

Table 2.1: Nash-Sutcliffe efficiency (NSE) and normalized mean squared error (NMSE) of the BNN and LOADEST models. 32

Table 3.1: Initial input parameters for *E. coli* transport in the vadose zone [*Bradford et al.*, 2006; *Gelhar et al.*, 1992; *Mace et al.*, 2000; *Pang et al.*, 2004]. 45

Table 3.2: Boundary conditions (BC) for HYDRUS and MODFLOW-MT3DMS models 47

Table 4.1: Hydrogeologic and other properties of the Ogallala and Trinity aquifers used for modeling [*Blandford et al.,* 2003 ; *Long et al.,* 2003; *Dutton et al.*, 2004; *Mace et al.,* 2000]. 85

Table 4.2: Parameters for nitrate-N transport [*Gelhar*, 1992; *Mehta et al.*, 2000; *Burton*, 2007; *Bronson et al*., 2009]. 85

Table 4.3: Normalized risk index (%) in the Trinity and Ogallala aquifers. The NRI values show that the likelihood of hot spots is higher for larger NRI. 113

Table 5.1: Hydrogeologic and other properties of the Ogallala Aquifer used for modeling the Ogallala Aquifer and the synthetic flow problem [*Blandford et al.,* 2003*; Long et al.,* 2003; *Dutton et al.*, 2004]*. 133

Table 5.2: Parameters for nitrate-N transport [obtained from *Gelhar*, 1992; Mehta et al., 2000; Burton, 2007; *Bronson et al*., 2009]. First order reaction rate values were used to initialize the decay coefficient field, and were later updated by the EnKF algorithm. 134

Table 6.1: Description of the one-dimensional soil column. 168

Table 6.2: Initial input parameters for nitrate transport [*van Genuchten*, 1985]..... 168

Table 6.3: Initial input parameters for *E. coli* transport [*Bradford et al.*, 2006]..... 168

CHAPTER I

GENERAL INTRODUCTION

1.1 Problem statement

Protecting the world's freshwater resources requires identification of threats across scales, from global to local. In surface water, microbes have been identified as major contaminants of water resources in USA [*USEPA*, 2006]. In groundwater, nitrate, from both natural and anthropogenic sources, has been identified as the most widespread contaminant [*Nolan et al.*, 2002]. Surface water and groundwater interact directly or indirectly through the vadose zone (Figure 1.1).

The sources of *E. coli* contamination in the surface water include waste water treatment plants, septic systems, wildlife, and livestock such as, cattle, pets, and horses [*Teague et al.*, 2009; *BRA and ECI*, 2010]. While, much of the nitrate contamination of groundwater occurs through the vadose zone, usually from one of the following sources: inorganic nitrate from fertilizers, organic nitrate from animal manure applied to croplands, domestic lawns, golf courses, failing septic systems, domestic wastewater effluents, wastewater from confined animal feeding operations (CAFOs), and leachate from landfills and wastewater treatment lagoons, and feces of range livestock, wildlife, and domestic animals in residential areas [*Gormly* and *Spalding*, 1979; *Keeney*, 1986].

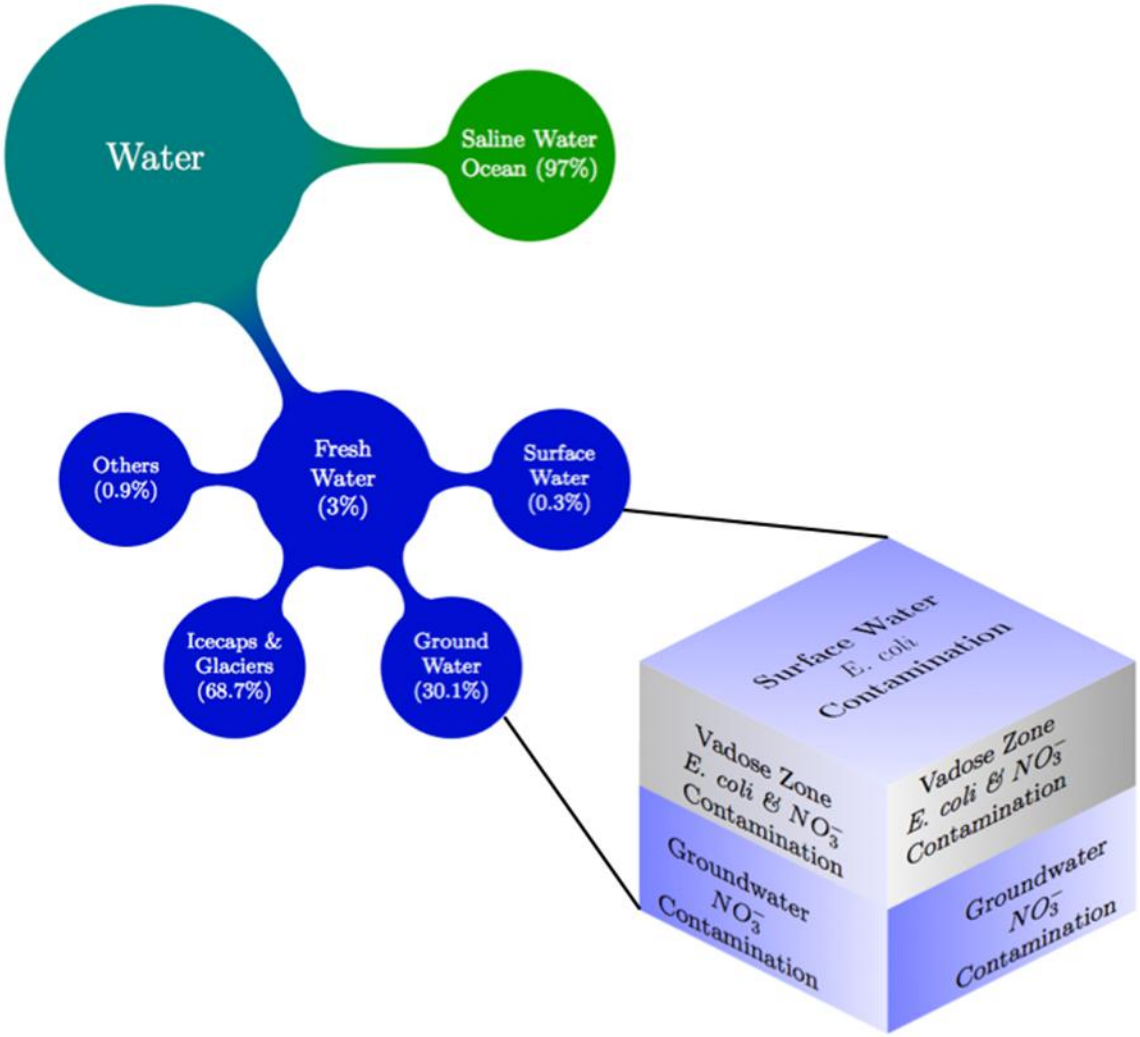

Figure 1.1: A graphical distribution of water on the Earth and the most ubiquitous contaminants in water resources.

1.2 Motivation

Surface water provides 40%, and groundwater provides 60 % of the total water in Texas. Texas has a huge supply of surface waters, including rivers and streams, lakes and reservoirs, springs and wetlands, bays and estuaries, and the Gulf of Mexico. Additionally, there are 9 major and 21 minor aquifers, each with their own characteristics and ability to produce water. Groundwater is used about 80% to irrigate crops. Groundwater supplies are known to decrease in Texas, particularly because of the depletion of the Ogallala Aquifer. In addition, the population of Texas increased over 20% between 2000 and 2010 [*TWDB*, 2012]. With depleting fresh water resources and increasing demand, it is crucial to understand the fate and transport of ubiquitous contaminants in order to mitigate contamination of fresh water resources.

There are uncertainties in the predictions of contaminants; moreover the uncertainty associated with these predictions is magnified due to the spatio-temporal variability of contaminants, unknown parameters, and sparse datasets at large scales (Figure 1.2). For reducing uncertainty, an inherent requirement of contaminant transport modeling is parameter estimation and characterization of uncertainty across scales, and establishing connections between different types of water resources. This information assists environmental managers to design targeted monitoring program, support real-time decision-making.

1.3 Research objectives

The primary objective of this research is to provide a modeling framework by bringing together physical, chemical, biological, and mathematical perspectives to

enable better prediction of fate and transport of contaminants in surface water, groundwater, and vadose zone (Figure 1.2).

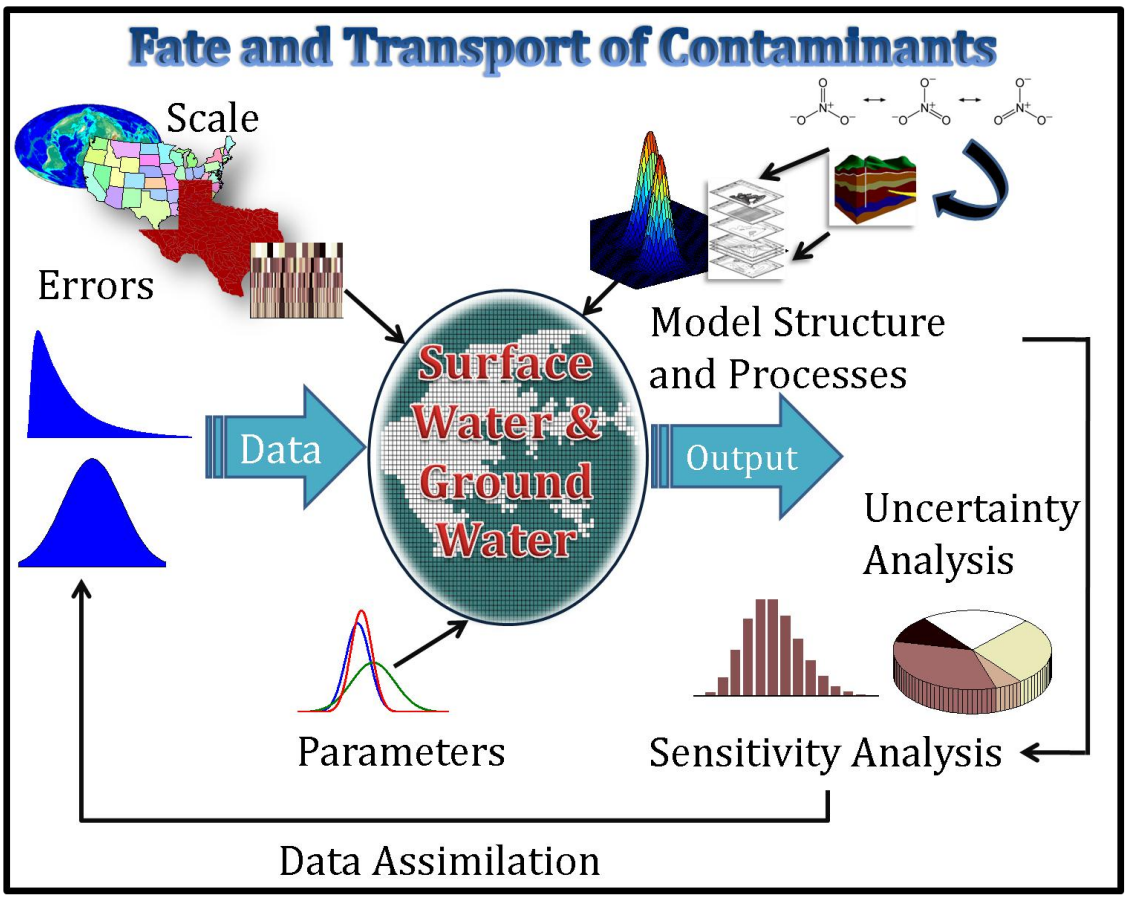

Figure 1.2: A schematic representation of process integration in the modeling framework, and analysis of parameter sensitivity and uncertainty in the output.

The specific objectives of this research were:

1. To identify dominant physical, chemical, and biological factors and reduce uncertainty in predicting *E. coli* loads in surface water.
2. To investigate parameter sensitivity and compare models of *E. coli* transport across surface water and groundwater
3. To investigate the spatio-temporal variability of nitrate, in groundwater.
4. To develop algorithms to reduce uncertatinty in nitrate predictions in groundwater
5. To develop a suitable modeling framework which can improve predictions of *E. coli* and nitrate transport in the subsruface

In Chapter I, the key water quality factors for estimating *E. coli* loads in Plum Creek, Texas are investigated, and *E. coli* loads are then estimated based on these selected key water quality parameters. We also characterize possible uncertainties in the estimation of *E. coli* loads using Bayesian Neural Networks.

We also characterize possible uncertainties in the estimation of *E. coli* loads. Chapter II builds upon this work and explores parameter sensitivity and compares models for improving predictions of *E. coli* loads.

In Chapters II, *E. coli* transport across surface water and groundwater is explored. Chapter II explores parameter sensitivity across unsaturated and saturated zones and investigates the effect of flow regimes and seasonal variability on *E. coli* concentrations.

In Chapter III, the variability of NO_3^--N across spatio-temporal scales in groundwater systems is investigated. We also analyze long-term trend and persistence

associated with disorders in the temporal distribution of NO_3^--N, and explore the complex mechanisms that regulate the spatial variability NO_3^--N in aquifers.

Chapter IV will build upon the previous work and improve groundwater modeling capabilities by reducing uncertainties in NO_3^--N prediction. We will estimate parameters (e.g., hydraulic conductivity, and reaction rate constants, and recharge) at larger scales in order to quantify and reduce the uncertainty of the groundwater flow and solute transport (MODFLOW and MT3DMS) models.

In Chapter V, a novel conceptual modeling framework is developed that accounts for model structural uncertainty and explores the dynamics of *E. coli* and NO_3^--N transport in subsurface. This conceptual framework models physical processes deterministically and treats biological and chemical processes stochastically.

CHAPTER II

ESTIMATING *E. COLI* LOADS IN STREAMS BASED ON VARIOUS PHYSICAL, CHEMICAL, AND BIOLOGICAL FACTORS

2.1 Synopsis

Microbes have been identified as a major contaminant of water resources. *Escherichia coli* (*E. coli*) is a commonly used indicator organism. It is well recognized that the fate of *E. coli* in surface water systems is governed by multiple physical, chemical, and biological factors. The aim of this work is to provide insight into the physical, chemical, and biological factors along with their interactions that are critical in the estimation of *E. coli* loads in surface streams. There are various models to predict *E. coli* loads in streams, but they tend to be system or site specific or overly complex without enhancing our understanding of these factors. Hence, based on available data, a Bayesian Neural Network (BNN) is presented for estimating *E. coli* loads based on physical, chemical, and biological factors in streams. The BNN has the dual advantage of overcoming the absence of quality data (with regards to consistency in data) and determination of mechanistic model parameters by employing a probabilistic framework. This study evaluates whether the BNN model can be an effective alternative tool to mechanistic models for *E. coli* loads estimation in streams. For this purpose, a comparison with a traditional model (LOADEST, USGS) is conducted. The models are compared for estimated *E. coli* loads based on available water quality data in Plum Creek, Texas. All the model efficiency measures suggest that overall *E. coli* loads

estimations by the BNN model are better than the *E. coli* loads estimations by the LOADEST model on all the three occasions (three-fold cross validation). Thirteen factors were used for estimating *E. coli* loads with the exhaustive feature selection technique, which indicated that six of thirteen factors are important for estimating *E. coli* loads. Physical factors included temperature and dissolved oxygen; chemical factors include phosphate and ammonia; biological factors include suspended solids and chlorophyll. The results highlight that the LOADEST model estimates *E. coli* loads better in the lower ranges, whereas the BNN model estimates *E. coli* loads better at the higher ranges. Hence, the BNN model can be used to design targeted monitoring programs and support real-time decision-making.

2.2 Introduction

Microbes have been identified as a major contaminant (13.2% contamination caused by pathogenic microbes of total impaired water body segments) of water resources in USA (*USEPA*, 2006). Common bacterial waterborne pathogens include *Salmonella sp*, *Shigella sp*., few strains of *Escherichia coli* (*E. coli*), *Pseudomonas aeruginosa*, *Aeromonas hydrophila*, *Mycobacteria*, *Helicobacter pylori*, and various others [*Fincher* et al., 2009]. The most widely used indicator organisms are the enteric coliform bacteria, which are Gram-negative bacilli that belong to the family Enterobacteriaceae (e.g., *Klebsiella spp*., *Enterobacter spp*., *Citrobacter spp*., *Escherichia coli*) [*Hipsey* et al., 2008; *Dorner* et al., 2006; *Mead* and *Griffin*, 1998]. The indicator organisms are mostly harmless as compared to the pathogen(s) of concern. However, the indicator organisms are monitored due to the relative ease and lesser cost

involved in their measurements. It is well established that the fate of *E. coli* in surface water systems is governed by multiple physical (e.g., temperature [*Flint*, 1987], chemical (e.g., pH [*Sjogren and Gibson*, 1981], nutrients [*Lessard and Sieburth*, 1983], sulfate [*Robakis* et al., 1983], and nitrate [*Noguchi* et al., 1997]), and biological (Chlorophyll) [*Nevers* and *Whitman*, 2005] factors. The relationship among these factors and *E. coli* loads gets complicated by flow rate [*Whitman* et al., 2004; *McKergow* et al., 2009]. *Vidon et al.* [2008] have reported that *E. coli* loads are significantly higher at high flow than at low flow, whereas *McKergow* et al. [2009] have observed that *E. coli* peak loads always preceded discharge and turbidity peaks (which had similar timings). Therefore, *E. coli* evidently has a nonlinear relationship with the flow rate and the turbidity.

It is important to develop an understanding of the relative importance of these physical, chemical, and biological factors in estimating the survival of *E. coli* in water bodies. However, a direct measurement of *E. coli* fate is not, in general, easy to implement. Therefore, the degree of impairment of a stream is assessed in terms of Total Maximum Daily Load (TMDL). Load duration curves are often used to estimate the reduction of contaminant loads in a watershed, especially in TMDL programs [*Babbar-Sebens* and *Karthikeyan*, 2009]. The load duration curves are measured using the instantaneous ‘‘load’’. The instantaneous ‘‘load’’ passing through a stream cross-section is the product of the flow rate and the constituent concentration.

Various models have been developed that use complex mechanistic and empirical relationships to predict the loads of *E. coli* in surface water systems e.g, Soil

and Water Assessment Tool (SWAT) [*Arnold and Fohrer*, 2005 and *Pachepsky* et al., 2006], Hydrological Simulation Program—Fortran (HSPF) [*Benham* et al. 2006], and a watershed model developed by *Tian* et al. [2002]. However, overly complex mechanistic relationships and requirement of comprehensive descriptions of geometry and capacity of stream, detailed knowledge about sources within the watershed, sedimentation and re-suspension characteristics, and bacteria die-off rates limit the utility of these models. Input parameter approximation and simplification in describing transport processes result in significant uncertainties in *E .coli* loads in streams. Other models have been developed that use statistical modeling framework to predict the loads of *E. coli* in surface water systems. For instance, *Nevers* and *Whitman* (2005) used regression modeling to determine *E. coli* using wave height, lake chlorophyll and turbidity for individual beaches of southern Lake Michigan. Furthermore, *Money* et al. (2009) estimated *E. coli* concentrations using turbidity, where *E. coli* data were not available, to assess fecal contamination along the Raritan River in New Jersey. Different models are relevant for different surface water environments, such as freshwater lakes and reservoirs [*Auer* and *Niehaus*, 1993; *Walker* and *Stedinger*, 1999; *Jin* et al., 2003; *Hipsey* et al., 2008], streams and rivers [*Wilkinson* et al., 1995; *Medema* and *Schijven*, 2001], and estuaries and coastal lagoons [*Steets* and *Holden*, 2003; *McCorquodale* et al., 2004]. It is also difficult for users to confidently implement these models, since models tend to be system or site specific.

In comparison to these mechanistic and statistical models, a Bayesian Neural Network (BNN) provides a Bayesian modeling framework for estimating *E. coli* loads

by utilizing routinely monitored flow rate and water quality data. The input data will comprise of water quality data (physical, chemical, and biological factors) that will provide a functional framework for the BNN. In case of sparse datasets, the Bayesian framework helps in representing input parameters as random variables emphasizing the statistical strength of the available data. Also, the uncertainty in input data sets is reflected through the probabilistic prediction of *E. coli* loads. The graphical structure of the BNN represents a cause-and-effect relationship between system variables (water quality data) and *E. coli* loads, as shown in Figure 2.1. One can use various bases in the formulation of the BNN such as multi-layer perceptron (MLP), radial basis functions (RBF). BNN models with radial basis functions (RBF) have been used in this study as they have an ability to deal with sparse datasets and parameter over-fitting.

The specific objective of this study is to identify the key water quality factors for estimating the *E. coli* loads in streams. Based on identified water quality factors, *E. coli* loads will be estimated in streams along with characterization of possible uncertainties.

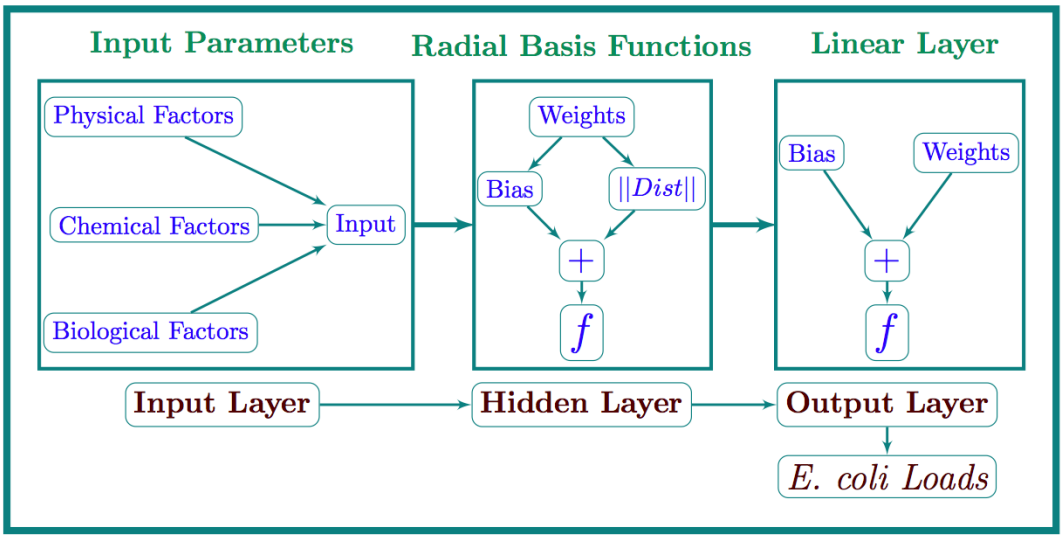

Figure 2.1: The graphical structure of BNN representing cause-and-effect relationship between system variables (water quality parameters) and the *E. coli* loads. The Radial Basis layer is the hidden layer, which uses the transfer function f (thin plate spline), and the linear layer is the output layer, which uses the transfer function f (linear function). The transfer function f establishes a relationship between inputs and outputs, in case of estimation of *E. coli* loads, and thin plate splines work better than other transfer functions (Gaussian or r4 functions).

2.3 Study area description and data availability

This study is conducted at a station (Station ID: 12645; Latitude 29°40'02" and Longitude 97°39'14") in Plum Creek (Figure 2.2), which is monitored by the Texas Commission on Environmental Quality (TCEQ). The Plum Creek watershed is a part of the Guadalupe River basin and is located in east central Texas. It surrounds a drainage area of 1028 km^2 in the counties of Hays, Caldwell, and Travis [*Teague et al.*, 2009]. According to the 2008 Texas water quality inventory and 303(d) list of impaired water bodies, Plum Creek is impaired for bacteria throughout the entire segment [http://www.gbra.org/CRP]. Plum Creek is a shallow, intermittent fifth-order stream. It

is 83 km long and joins the San Marcos River that in turn connects with the Guadalupe River. The watershed has several rapidly growing towns such as Lockhart, Kyle, and Luling. The watershed has a diversified land use from urban to agriculture and oil field activities. The watershed encompasses 38% rangeland, 17% pasture, 11% cultivated cropland, 18% forest, 8% developed land, 6% near riparian forest, and 2% open water and barren land. The landscape is characterized as rolling hills of pasture and cropland surrounded by scrub oak forest. Plum Creek lies in a semi-humid subtropical climate zone and is heavily influenced by its proximity to the Gulf of Mexico [*GBRA*, 2006].

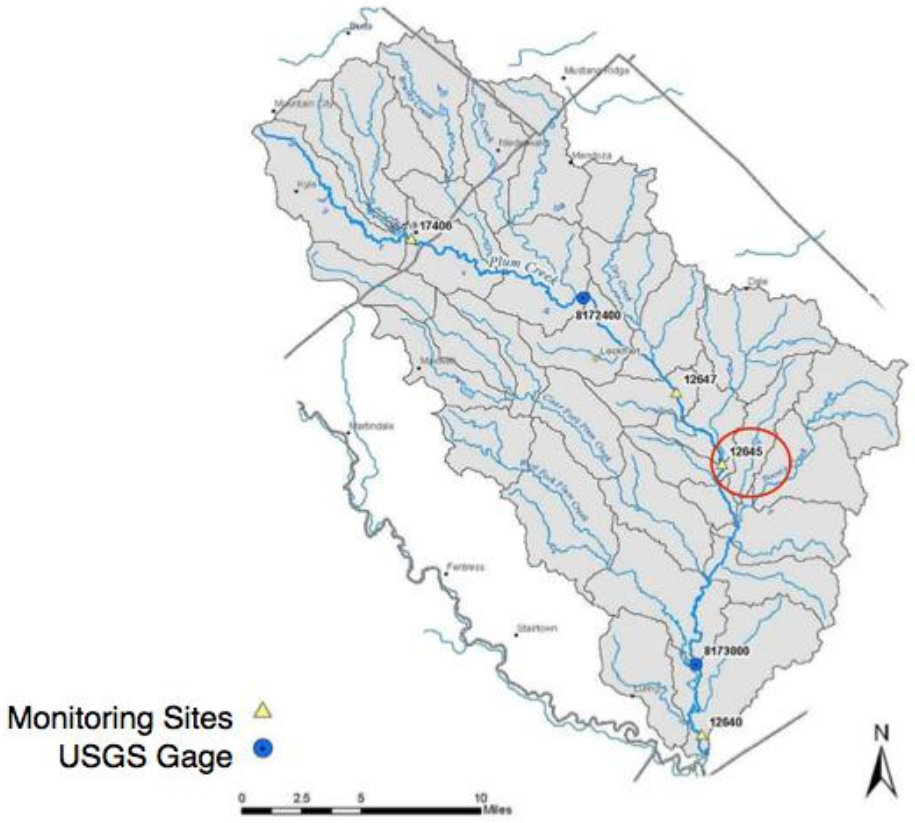

Figure 2.2: Map showing the Guadalupe river basin in east central Texas and the station ID 12645 in Plum Creek (Map modified from http://www.gbra.org/CRP).

Two US Geological Survey (USGS) gage stations are located on Plum Creek to monitor stream flows: one north of Lockhart (Station 08172400) and one near Luling (Station 08173000). Near Lockhart, periods of no flow have occurred almost every year on record. Southern reaches of Plum Creek, particularly south of Lockhart, are fed by a number of small springs and are usually perennial. Based on routine water quality sampling, TCEQ initially listed portions of Plum Creek for bacteria impairment for contact recreation use in 2002. The possible sources of *E. coli* contamination in the creek are cows¸ livestock, wildlife, wastewater treatment plants, septic systems, and pet sources [*Teague et al.*, 2009]. By 2004, bacterial contamination level in Plum Creek was elevated, and it was included in the list of impaired waters of Texas prohibiting wading and swimming. The *E. coli* criteria for designated use of a stream specified in water quality standards (e.g., recreational uses, irrigation, and navigation etc.) require a geometric mean (GM) concentration of *E. coli* less than 126 cfu/100 mL of water with no sample exceeding 235 cfu/100 mL of water. *E. coli* and water quality data at the monitoring sites were available from October 1996 to December 2008 [http://www.gbra.org/CRP]. Water quality data were collected monthly by the TCEQ. The available water quality data include thirteen factors wherein physical factors include turbidity (NTU), temperature (^{o}C), conductivity (μmhos/cm)), and dissolved oxygen (mg/L) (DO); chemical factors include pH, phosphate (mg/L), nitrate-N (mg/L), chloride (mg/L), sulfate (mg/L), total hardness (mg/L), ammonia (mg/L); and biological factors include suspended solids (mg/L) (SS) and chlorophyll (mg/m^3).

2.4 Methodology

In order to identify the key water quality factors responsible for *E. coli* loads in streams, BNN models are run in conjunction with the exhaustive feature selection technique. We use the thirteen physical, chemical, and biological factors described above. The exhaustive feature selection technique provides the best set of water quality factors for estimating *E. coli* loads. A principal component analysis (PCA) is also conducted to get insight into the relative importance of the factors identified by the exhaustive feature selection. These selected factors are subsequently utilized in estimating *E. coli* loads by the BNN model in Plum Creek. The BNN model results are also compared with the LOADEST model. For better decision-making, uncertainty analysis is also conducted. In the subsequent sections, we provide a description of BNN and LOADSEST models, exhaustive feature selection, PCA, and uncertainty analysis.

2.4.1 Bayesian neural networks

The application of the Bayesian learning paradigm to neural networks results in a flexible and powerful nonlinear modeling framework that can be used for regression, density estimation, prediction and classification supporting adaptive decision-making, and accounting for uncertainties [*Andrieu* et al., 2001, *Reckhow*, 1999]. The regression of a target variable Y on an input set of covariates X given the data $D = \{(x_1, y_1), (x_2, y_2), \ldots\}$:

$$y_i = mx_i + \varepsilon_i \tag{2.1}$$

where ε_i are independent and identically distributed (i.i.d.) $N \sim (\mu, \sigma^2)$.

The central process of the Bayesian framework is the calculation of a probability distribution on the unknown parameter (weight) vector w. Prior knowledge that we might have, say for small weights, is updated using the data. These posterior distributions are used in model predictions, with point forecasts given as expectations [*Holmes and Mallick*, 1998]:

$$E[Y|x, D] = \int m(x, w)p(w|D)dw \tag{2.2}$$

where $E[Y|\ x,\ D]$ represents the posterior probability of the parameters of the model m(x, w) given the training data D.

BNN generates a probability distribution of the layer weights, which is dependent on the given input data:

$$P(w|D) = \frac{P(Y|w,X)P(w)}{P(Y|X)} \tag{2.3}$$

where $P(Y|X) = \int P(Y|w, X)P(w)dw$ is the marginal distribution of Y, $P(w)$ is the prior distribution of weights, and $P(Y|w,\ X)$ is the likelihood function [*Gelman* et al., 1995]. Artificial Neural Network (ANN) combined with Monte Carlo Markov Chain (MCMC) generates multiple samples from a continuous target density [*Bates and Campbell, 2001*]. A flat prior can be assumed here, as we do not have any definite prior knowledge of weights [*Sims and Zha, 1998*].

Predictive distribution of Y_{n+1} is given by:

$$P(Y_{n+1}|x_{n+1}, Y, X) = \int P(Y_{n+1}|x_{n+1}, Y, w)P(w|Y, X)dw \tag{2.4}$$

where $n+1$ denotes next realization.

We considered Radial Basis Functions (RBF) architecture in BNN, which has an ability of closely approximating any nonlinear multidimensional mapping [*Ciocoiu*, 2002]. A brief summary of RBF is provided below for completeness.

2.4.2 Radial basis function (RBF)

RBF networks are one of the most commonly used types of feed forward networks. The feed forward neural network is most widely used to solve engineering problems. It is a simple nonlinear model that maps the input vector onto the output vector [*Lanouette* et al., 1999].The architecture of a RBF network consists of three layers: an input layer, a hidden layer, and an output layer. The transformation from input space to hidden unit space is nonlinear, whereas transformation from hidden unit space to output space is linear [*Ciocoiu*, 2002]. During the training stage, a known set of input and output data pairs are delivered to the RBF network to select the centers and compute the output layer weights. The models have radial functions, where each basis is parameterized by a knot or position vector located in the d-dimensional covariate space x. The hidden layer provides a set of functions that constitute an arbitrary basis for the input patterns. The hidden units are known as radial centers and represented by the vectors (C_1; C_2; …, C_h). Conventionally, there are as many basis functions (h) as data points to be approximated with the position vectors set to the data values. The model output $m(x)$ is given by a linear combination of the basis functions response and a low-order polynomial term:

$$m(x) = \sum_{i=1}^{N} w_i \varphi_i\left(\left|\left|x - \mu_i\right|\right|\right) + \sum_{m=0}^{p} a_m \, q_m(x) \qquad (2.5)$$

where $||.||$ denotes a distance metric, usually Euclidean or Mahalanobis, and $q_m(x)$ represents a polynomial of degree *m*. The coefficients *w* and *a* are calculated by least squares where the constraint$\sum_{i=1}^{N} w_i\, q_m(x_i) = 0$ is imposed to ensure the uniqueness of the solution [*Holmes* and *Mallick*, 1998]. Different radial functions (e.g., Gaussian, quadratic, thin plate spline, and inverse quadratic functions) are used for different problems. We used the thin plate spline (TPS) for estimating *E. coli* loads. The TPS is given as:

$$\emptyset(Z) = Z^2 \log(Z) \tag{2.6}$$

where $Z = (||x - \mu_i||)$ (2.7)

RBF networks enlarge the dimensionality of the input data in order to increase the probability that originally nonlinearly separable classes become linearly separable (Cover's theorem) [*Ciocoiu*, 2002]. For modeling a system with limited experimental data, RBF may have an advantage over the other techniques. One of the problems that may occur during neural network training is over-fitting. A frequently used method for improving network generalization is to use an adequate-sized network, which is just large enough to provide an adequate fit [*Cilek and Yilmazer*, 2003]. Over fitting happens when the model has too many degrees of freedom, which is the result of including too many hidden neurons. The neurons in the hidden layer contain transfer functions whose outputs are inversely proportional to the distance from the center of the neuron. With small data sets used in this study, we ensured model accuracy by running multiple simulations by randomizing data sets. We split all the valid data into three randomly distributed groups. Three random sets are selected using randperm function in

MATLAB. Initially, random split 1 is set aside for testing, while the models are parameterized on the basis of random splits 2 and 3. The fitted models are then used to test/predict *E. coli* loads by using input data from the random split 1. Next, random split 2 is set aside for testing, while random splits 1 and 3 are used for training. This pattern is also repeated for the random split 3. We use the same random splits (1, 2, and 3) for estimating *E. coli* loads by using the LOADEST Model.

2.4.3 Load estimator

Load estimator (LOADEST) is a regression-based model for estimating constituent loads in streams and rivers [*Runkel* et al., 2004]. Given a time series of streamflow and constituent concentration (*E. coli*), LOADEST facilitates users in developing a regression model for the estimation of constituent loads [*Cohn*, 2005]. Explanatory variables within the regression model include multiple functions of flow, time, and additional data variables. The developed regression model is then used to estimate loads over a user-specified time interval. Mean loads, standard errors, and 95% confidence intervals are also estimated on a monthly and/or seasonal basis. There are three statistical methods used for calibration and validation (estimation) of LOADEST including Adjusted Maximum Likelihood Estimation (AMLE), Maximum Likelihood Estimation (MLE), and Least Absolute Deviation (LAD). AMLE and MLE are appropriate when the calibration model errors (residuals) are normally distributed, whereas LAD is appropriate when model errors (residuals) are not normally distributed. In our case, calibration model errors are normally distributed, so we used AMLE for estimating *E. coli* loads. The detailed mathematical formulation of LOADEST is

provided elsewhere [*Cohn*, 2005]. In general, total mass loading over an arbitrary time period, τ, is given by:

$$L_{\tau=} \int_0^{\tau} QCdt \tag{2.8}$$

$$\hat{L}_{\tau=} \Delta t \sum_{i=1}^{NP} (\hat{Q}C) = \Delta t \sum_{i=1}^{NP} (\hat{L}) \tag{2.9}$$

where C is concentration [M/L^3], L is the total load [M], Q is instantaneous stream flow [L^3/T], t is time [T], and NP is the number of discrete points in time. The hats on Q, L_{τ}, and L denote the instantaneous values of the respective variables. *E. coli* loads estimated by the LOADEST model are compared with the *E. coli* loads estimated by the BNN model using the key water quality factors. The key water quality factors are identified using the exhaustive feature selection technique.

2.4.4 Exhaustive feature selection

The BNN models are run multiple times with all possible combinations of the thirteen water quality factors ($^{13}C_1$+$^{13}C_2$+….+$^{13}C_{13}$) for estimating *E. coli* loads. The Exhaustive Feature Selection is a technique of selecting a subset of relevant features for building robust models. The brute-force feature selection algorithm is applied to exhaustively evaluate all possible combinations of the input features, and then the best subset is chosen. The exhaustive search's computational cost is prohibitively high, with a considerable danger of overfitting [*Moore* and *Lee*, 94; *Skalak*, 94]. Hence, for avoiding the over-fitting, K-fold (three-fold) cross validation is used in selecting the best subset. The aim of the feature selection is to choose a subset of the set of input features (physical, chemical, and biological factors) so that the subset can predict the output Y (*E. coli* loads) with accuracy akin to the performance of the whole input set X, and with a

reduction of the computational cost. For conducting the exhaustive feature selection, the following steps are outlined:

1. Shuffle the dataset and split into a training set of 2/3rd of the data and a test set of the remaining 1/3rd of the data.
2. Choose all possible combinations of various input variables.
3. Select each subset, and run the BNN model with leave-one-out cross-validation.
4. Store the Nash-Sutcliffe Efficiency (NSE) Coefficients (see section 2.4.5) of each run.
5. Select the feature set which has minimum root mean square error of NSE three-fold validation.

2.4.5 Model performance

We computed the Nash–Sutcliffe efficiency (NSE) and Normalized Mean Squared Error (NMSE) as measures of the model performance.

The Nash–Sutcliffe efficiency (NSE) coefficient is given as:

$$NSE = 1 - \frac{\sum_{\tau=1}^{t}(Q_O^\tau - Q_m^\tau)}{\sum_{\tau=1}^{t}(Q_O^\tau - Q_m^E)} \quad (2.10)$$

The Normalized Mean Squared Error (NMSE) is given as:

$$NMSE = \frac{\sum_{\tau=1}^{t}(Q_O^\tau - Q_m^\tau)}{var(Q_O^\tau)} \quad (2.11)$$

where Q_O^t is observed *E. coli* loads, and Q_m^t is simulated *E. coli* loads at time t. Q_O^E is mean observed *E. coli* loads. $var(Q_O^t)$ denotes the variance of all the observed *E. coli* loads. NSE can range from $-\infty$ to 1. An efficiency of 1 (NSE = 1) corresponds to a perfect match of simulated values to the observed data. An efficiency of 0 (NSE = 0)

demonstrates that the model predictions are as good as the mean of the observed data. In essence, closer the efficiency of the model is to 1, the more accurate is the model. The NMSE of 0 indicates that the model predictions are perfect. The lower the NMSE, the better is the model performance.

The Exhaustive Feature Selection technique in conjunction with the BNN model rendered the best set of factors. In order to assess the relative importance of these factors, principal component analysis (PCA) is done.

2.4.6 Principal component analysis

Principal component analysis (PCA) is a multivariate statistical technique. The transformed features have a descriptive power that is more ordered than the original features. PCA has been applied in describing various aspects of streamflow regimes [*Olden* and *Poff*, 2003], understanding the spatial and temporal changes in water quality [*Bengraıne*, and *Marhaba*, 2003], determination of dominant biogeochemical processes in a contaminated aquifer [*Cazull* and *McGuire*, 2008]. In this study, PCA is used to identify major factors among water quality data that can explain most of the variation of *E. coli* loads.

PCA is an orthogonal linear transformation of the data (e.g., water quality data) to a new coordinate system such that the greatest variance by any projection of the data happens to span on the first quadrant. The principal axis method is used to extract the components, followed by a varimax (orthogonal) rotation with Kaiser Normalization. A detailed description, of how the principal components are calculated, is provided elsewhere [*Jolliffe*, 2002].

2.4.7 Uncertainty analysis

Monte Carlo based statistical techniques – resampling with replacement ("bootstrapping") [*Robert* and *Casella*, 1999] is implemented to estimate the statistical uncertainty in predictions by the BNN and LOADEST models. To explore the uncertainty in the BNN predictions, 10000 realizations of *E. coli* loads are investigated. Bayesian networks are probabilistic models that combine prior distributions of uncertainty with data to yield an updated (posterior) set of distributions [*Helton and Oberkampf*, 2004]. Therefore, inputs are integrated over the weight space of the posterior probability distribution for finding the outputs (i.e., *E. coli* loads) of the networks.

The probability distribution of each output as a random variable is plotted utilizing the kernel density (Parzen window) estimation, which is a non-parametric method [*Silverman*, 1986]. If $x_1, x_2 \ldots x_N$ are samples drawn from the density function of a random variable, then the kernel density approximation of its probability density function is given as:

$$\hat{f}_{h(X)} = \frac{1}{N}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right) \tag{2.12}$$

where K is some kernel and h is a smoothing parameter called the bandwidth. Here, a Gaussian kernel is chosen with mean zero and unit variance:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{2.13}$$

2.5 Results and discussion

The results presented here provide insight into the different physical, chemical, and biological factors that are critical in the estimation of *E. coli* loads in surface streams. The exhaustive feature selection in conjunction with the BNN model (Figure 2.1) identified the best combination of input variables for estimating *E. coli* loads in Plum Creek (Figure 2.2). Out of the thirteen water quality factors, exhaustive feature selection identified six key variables: SS, phosphate, temperature, DO, ammonia, and chlorophyll. In the following section, we will focus on these key variables and their relative importance. Subsequently, we utilize these six key variables using the BNN model for estimating *E. coli* loads in Plum Creek.

2.5.1 Identification of the key factors responsible for the E. coli *loads in Plum Creek*

The exhaustive feature selection identified six factors in estimating *E. coli* loads in Plum Creek namely SS, phosphate, temperature, DO, ammonia, and chlorophyll. To investigate the relative importance of the key factors for *E. coli* loads in streams, a principal component analysis (PCA) was performed as shown in Figures 2.3 and 2.4.

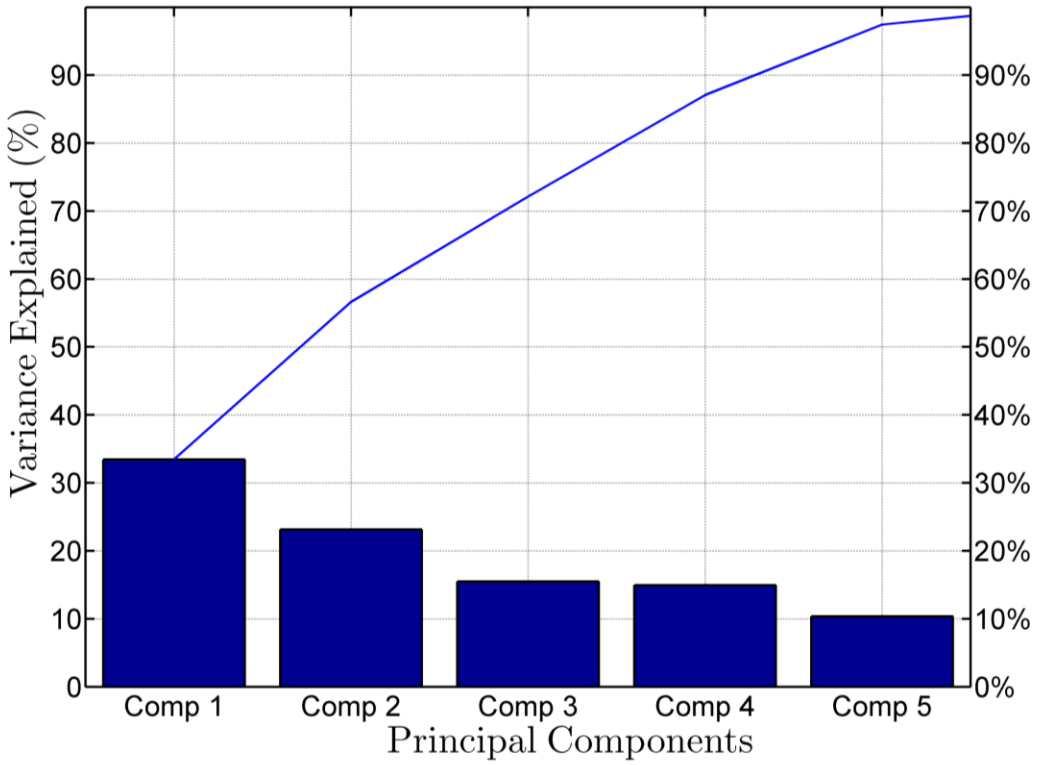

Figure 2.3: Pareto diagram of principal components shows the percentage explained by each component. The first three components explain 70% of variation in the data set.

The PCA explored the relationship among water quality factors such as SS, phosphate, temperature, DO, ammonia, and chlorophyll. The first two components explain 60.0% of the variance; component 1 and component 2 account for 35.6% and 24.4% of the variance, respectively (Figure 2.4). The first principal component (PC) captures the variance of DO and temperature. The second PC captures the variance of SS, phosphate, ammonia, and chlorophyll.

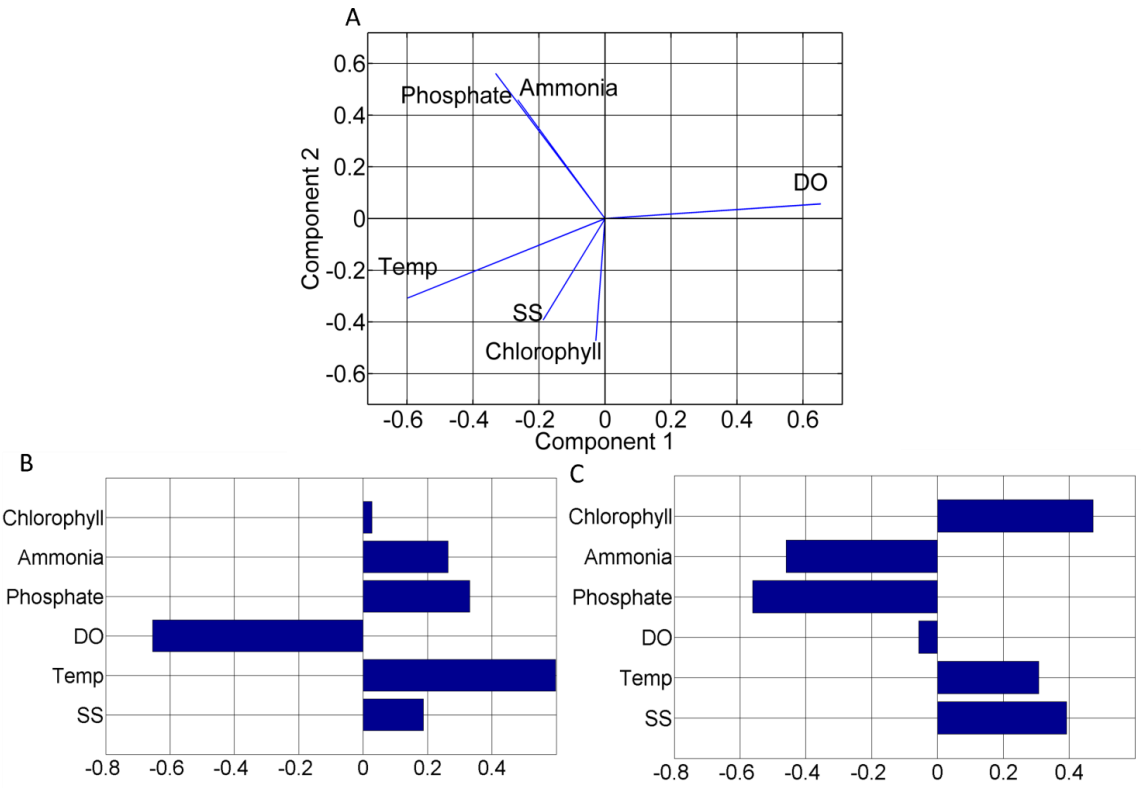

Figure 2.4: The principal component analysis (PCA) of the dataset is shown above. First principal component is projected against the second principal component. The Euclidean norm (length) reflects the relative importance of different factors. The biplot of the PCA is shown in (A). (B) and (C) demonstrate factors loadings on first and second principal components, respectively.

The PCA biplot (Figure 2.4) illustrates a visual interpretation of the factor loadings that result from a bi-cluster system of variables projected onto the first and second PC axes. The biplot tells about the relative positions of the factors, and the angles between the factors give approximate estimates of the correlation among factors; small angles between projected axes imply a high correlation. The direction of axes gives the sign of correlation among factors displayed on the biplot [*Jolliffe*, 2002].

Turning now to the interpretation of the PCs in the present work, the six factors can be divided into three groups. The group 1 includes temperature and DO (physical factors); the group 2 includes phosphate and ammonia (chemical factors); the group 3 includes the SS and the chlorophyll (biological factors). The central idea of this classification is based on the fact that groups of variables often move together, and more than one factor measures the same driving force. The first PC clearly measured physical factors, as DO and temperature have the maximum loadings (Figure 2.4B); moreover, they also have a high negative correlation with each other (almost 180^0 separated in biplot (Figure 2.4A). Therefore, DO and temperature were classified as the group 1. The second PC accounted for the chemical and biological factors (Figure 2.4C). Since, phosphate and ammonia are the dominant factors on the second PC, and phosphate and ammonia also have a high positive correlation with each other (almost overlapping in biplot (Figure 2.4A). For this reason, they were classified as group 2. Similarly, the third PC also accounted for biological factors, as SS and chlorophyll have a medium positive correlation with each other (a small angle between them in biplot Figure 2.4A). Hence, they were grouped together.

The biological tolerance of *E. coli* to different physical, chemical, and biological factors has been well studied, albeit mostly in the laboratory. It has been observed that *E. coli* are sensitive to changes in temperature [*Maeda* et al., 1976; *Berg*, 2004]. The rate of die-off depends on temperature [*Flint* et al., 1987]. Moreover, *E. coli* are anaerobic bacteria, and thus *E. coli* also responds to oxygen gradient. The majority of *E. coli* cannot live in oxygen rich environment [*Berg*, 2004]. This clearly explains the selection of DO and temperature as important physical factors in estimating *E. coli* loads in our study and their negative correlation. Temperature affects positively, whereas DO affects negatively in estimating *E. coli* loads by the BNN model. In the biplot, approximately 180^0 separation of temperature and DO corroborates this behavior of *E. coli* (Figure 2.4A).

Phosphate and ammonia are also found to be important factors in the estimation of *E. coli* loads by the BNN model. This is because phosphate and ammonia act as nutrients or substrates, and the presence of nutrients increases *E. coli* concentrations in streams [*van der Steen* et al., 2000]. These nutrients have significant positive correlation and therefore signify the importance of chemical factors on *E. coli* loads.

In our study, SS and chlorophyll were also important factors in the estimation of *E. coli* loads by the BNN model. In the literature, there is evidence to suggest that high concentrations of chlorophyll and suspended sediments are associated with high *E. coli* concentrations [*Nevers* and *Whitman,* 2005]. However, *Money* et al. (2009) examined the relationship between turbidity and *E. coli* and found a significant correlation between both the parameters. Turbidity indicates high volumes of suspended sediments. SS and

chlorophyll correspond to the biological factors, as they are sources of organic carbon [*Jonge* et al., 1980]. These biological factors were measured by the second PC.

It should be noted that the sign of any PC is completely arbitrary. If every coefficient in a PC has its sign reversed, the variance is unchanged, and so is the orthogonality [*Jolliffe*, 2002]. Therefore, the biplot and loadings only show the relative importance of the factors, they do not demonstrate if a factor is positively or negatively affecting the *E. coli* loads. However, the biplot exhibits how each factor can affect the *E. coli* loads. For example, it is evident from Figure 2.4A that all the factors on the left side of the plot (phosphate, ammonia, temperature, SS, and chlorophyll) are positively associated with *E. coli* loads, whereas the only factor on the right hand side of the plot is DO, and it is negatively associated with *E. coli* loads. This graphic examination further substantiates our findings.

2.5.2 Estimation of E. coli *loads*

In this section, we discuss the discrepancy between simulated and observed *E. coli* loads using the BNN model (using the six key variables) and compare its performance to the LOADEST model. Figure 2.5 and 2.6 show measured and simulated *E. coli* loads in Plum creek using the BNN and the LOADEST models respectively for three random splits.

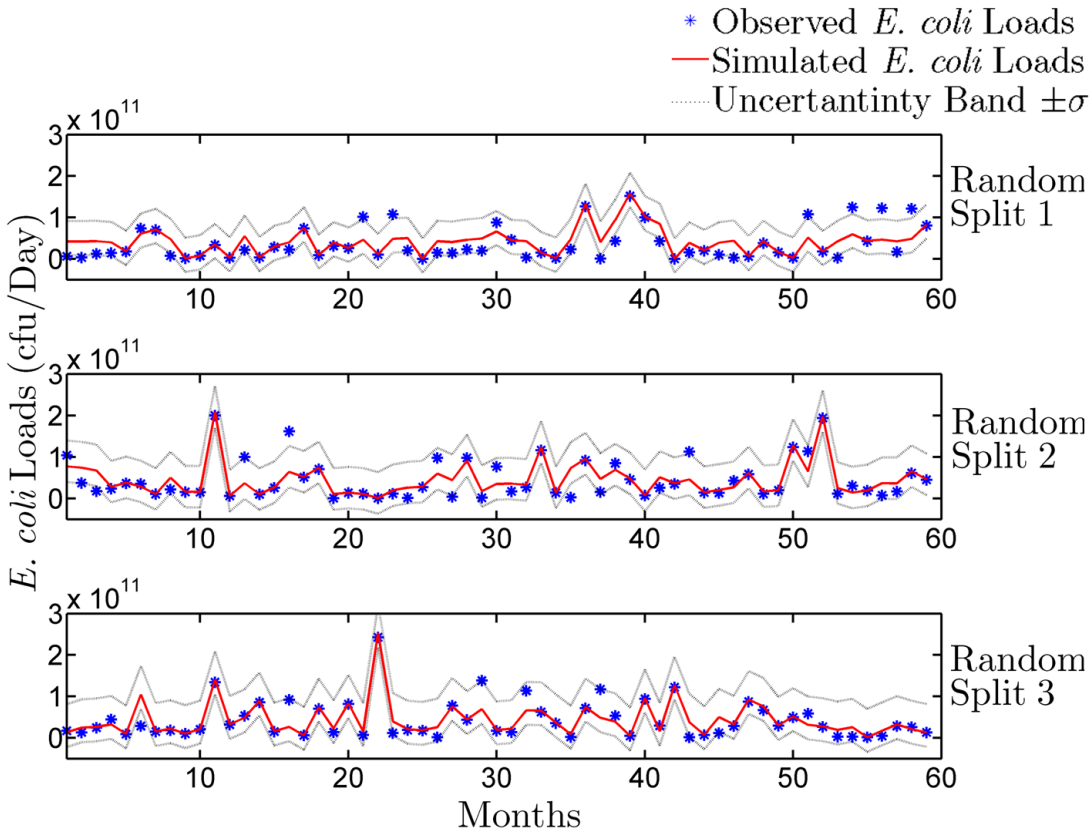

Figure 2.5: Measured and simulated loads of *E. coli* by the BNN model in Plum Creek are shown above. The simulations were tested by three-fold cross validation. Months on X-axis represent the sequence of random splits (not the sequence of a timeseries).

Table 2.1 shows the measures of the models' performance. A three-fold cross validation results show that both modeling approaches (BNN and LOADEST) reproduce observed *E. coli* loads reasonably well, with all NSE values greater than or equal to 0.39 and all NMSE values smaller than or equal to 0.59 (Table 2.1). However, the BNN is able to estimate *E. coli* loads better in all the three random splits (Table 2.1). The uncertainty bands (Figures 2.5 and 2.6) show that the BNN is also able to capture higher *E. coli* loads more accurately than the LOADEST model. This is expected because the

BNN model provides more flexible choices for the functional dependence in estimating *E. coli* loads based on physical, chemical, and biological factors (e.g., SS, phosphate, temperature, DO, ammonia, and chlorophyll), whereas the LOADEST model uses only the *E. coli* and flow data.

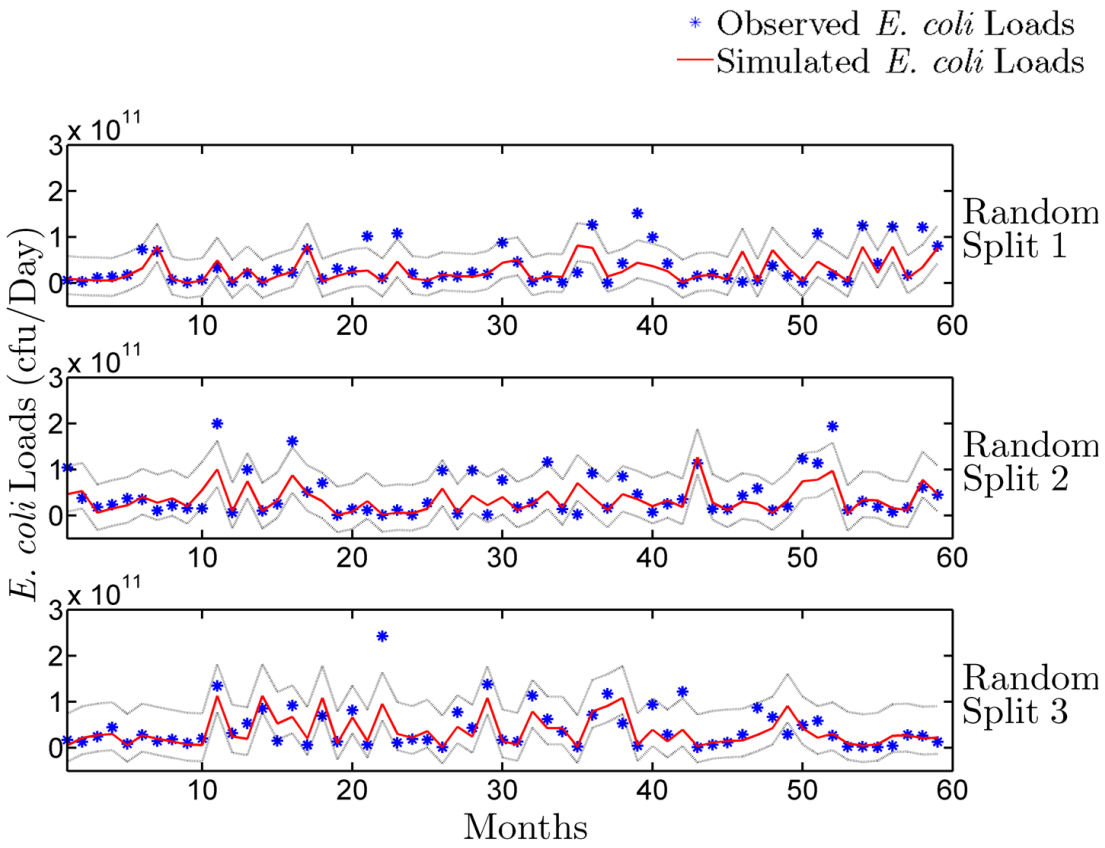

Figure 2.6: Measured and simulated loads of *E. coli* by the LOADEST model in Plum Creek are shown above. *E. coli* loads are presented for the three random splits, which were used for three-fold cross validation of the BNN model. Months on X-axis represent the sequence of random splits (not the sequence of a timeseries).

Table 2.1: Nash-Sutcliffe efficiency (NSE) and normalized mean squared error (NMSE) of the BNN and LOADEST models.

Random splits	Models	NSE	NMSE
Random split 1	BNN	0.48	0.51
	LOADEST	0.39	0.59
Random split 1	BNN	0.69	0.30
	LOADEST	0.55	0.44
Random split 3	BNN	0.75	0.23
	LOADEST	0.52	0.46

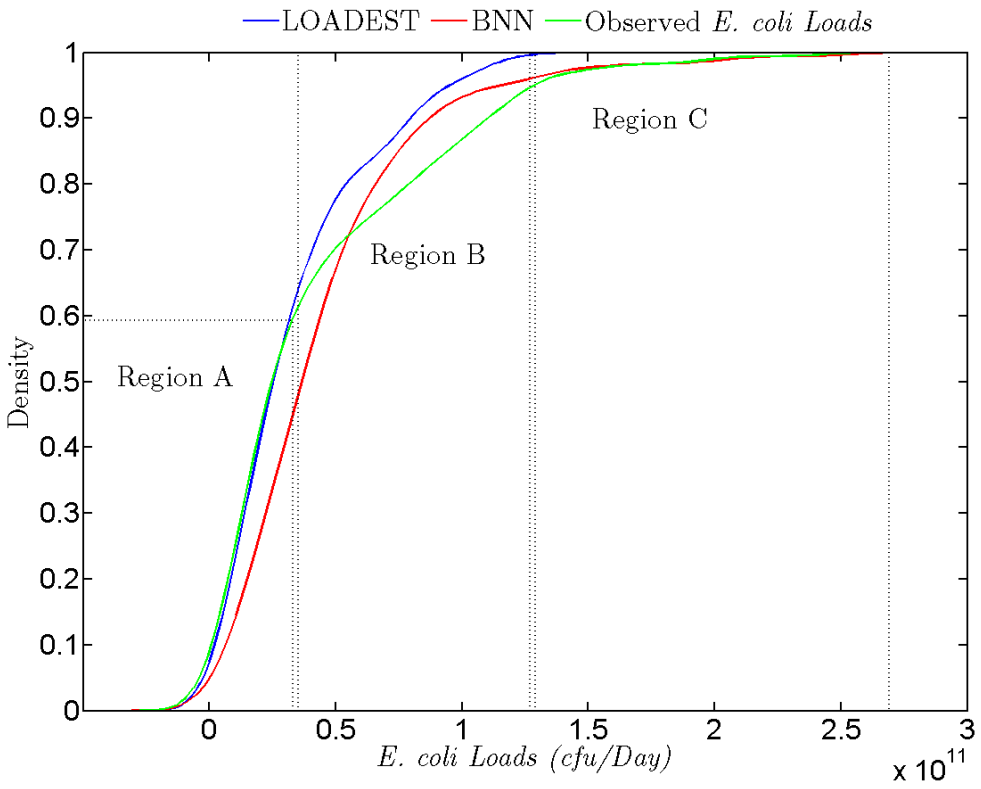

Figure 2.7: The cumulative density functions of the observed *E. coli* loads, estimated *E. coli* loads by the BNN, and estimated *E. coli* loads by the LOADEST model are presented above. There are three regions in the figure. Region A signifies that the LOADEST model is able to estimate *E. coli* loads better. In region B and C, the BNN model predicts better, however in region B, both the LOADEST and BNN models overestimate *E. coli* loads.

Figure 2.7 shows the cumulative distribution functions (CDFs) of the observed, BNN simulated (all the three random splits), and LOADEST simulated (all the three random splits) *E. coli* loads in Plum Creek. The region A signifies the smaller *E. coli* loads (smaller than $0.5x10^{11}$ cfu/Day), which are better estimated by the LOADEST model. The BNN model is underestimating the *E. coli* loads in this region. The region C encompasses the higher *E. coli* loads (greater than $1.5x10^{11}$ cfu/Day), which are better estimated by the BNN model. For best management practices, it is essential to be able to estimate higher *E. coli* loads, and the BNN model is able to estimate values with greater accuracy in this range. The region B ($0.5x10^{11}$ cfu/Day to $1.5x10^{11}$ cfu/Day) constitutes the region with medium loads between regions A and C, and where both the LOADEST and BNN models are overestimating the *E. coli* loads. However, the BNN model is closer to the observed values than the LOADEST model in this region.

2.5.3 Uncertainty analysis

Uncertainty analysis is conducted to further compare the performance of BNN and LOADEST models in estimating *E. coli* loads in Plum Creek. The uncertainty bands ($\pm\sigma$ with 95% confidence) computed using bootstrap samples show that there is more uncertainty for larger loads than smaller loads (Figure 2.5 and 2.6). We believe that this reflects various sources of uncertainties. There is evidence that uncertainties of discrete *E. coli* samples are greater than 30% while the uncertainty in storm water flow measurements average greater than 97% [*McCarthy* et al., 2008]. Therefore, *E. coli* loads will have more uncertainty due to storm events. As high *E. coli* loads are often associated with storm events, the upper limit of the uncertainty band is also wider for

higher loads. These uncertainties in the inputs propagate into larger uncertainties in the output.

Figure 2.8 shows six *E. coli* loads estimated by the BNN model (low and high *E. coli* loads for each random split), and the probability distribution functions (PDFs) of 10,000 realizations for each *E. coli* loads were plotted. As stated previously, Bayesian Neural Networks use a range of weight sets instead of a single set. Each weight gives a realization of *E. coli* loads. The final predicted *E. coli* loads were generated from the average of 10,000 such realizations. The center of mass of a PDF shows the mean of the prediction and spread around the mean shows the uncertainty. It is clear from the Figure 2.8 that *E. coli* loads, estimated by the BNN model, were closer to the centers of the PDFs with high density values (4 of them are >0.4). It should be noted that the BNN model estimates lower *E. coli* loads with a small bias and higher *E. coli* loads with a large bias; however, the performance of the BNN model is better than the LOADEST model for estimating higher *E. coli* loads. Figure 2.8 demonstrates that the *E. coli* loads estimated by the BNN model have a smaller bias but a relatively larger variance. The large variance in the PDFs is due to various uncertainties, which stem mainly from (1) the uncertainties in input data (e.g., flow rate and water quality data); (2) uncertainties in data used for calibration, (e.g., *E. coli* loads). Input data (flow rate and water quality data) and *E. coli* loads have large inherent uncertainties, and these uncertainties cannot be removed from the model predictions in the existing data. However, the advent of newer technologies and careful data collection may help in minimizing these uncertainties in the future. The other source of uncertainties is from model parameters

(weights and biases). These uncertainties are related to the fact that a small bias in the estimation, using a neural network with a training set of fixed size, can only be achieved with a large variance [*Geman et al.*, 1992; *Haykin*, 1996]. This dilemma can be avoided if the training set is made very large, but the total amount of data is limited in our case. However, a possibility of making training sets larger can be plausible in the future.

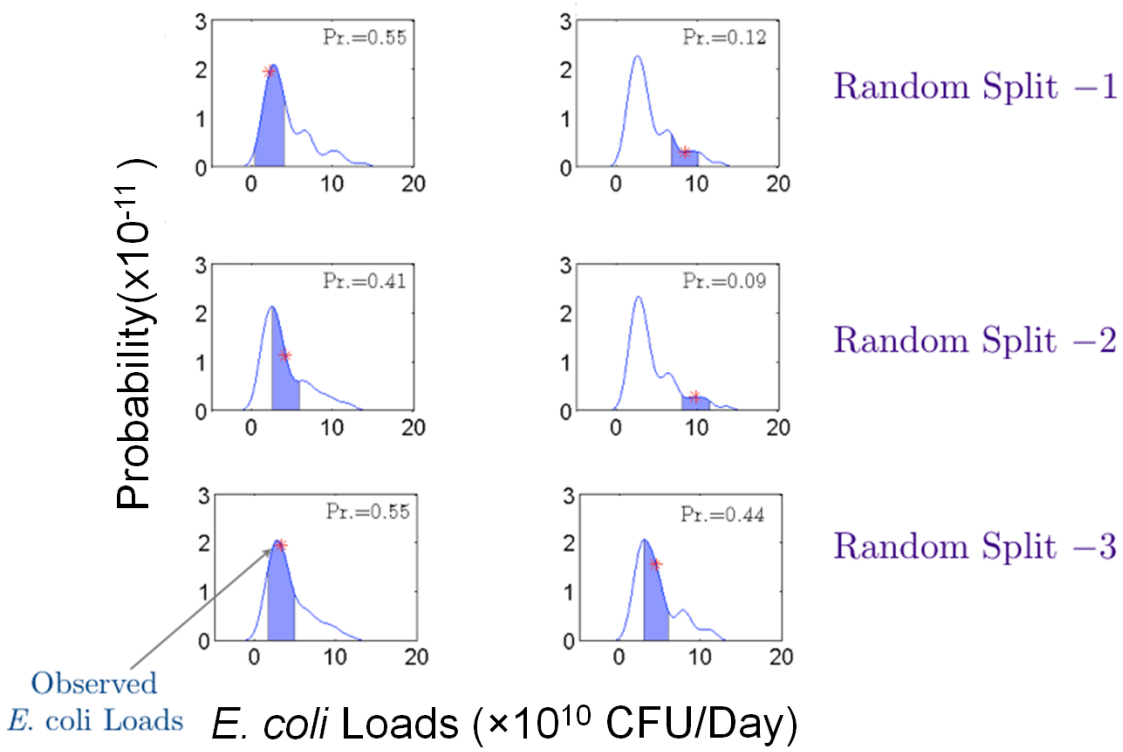

Figure 2.8: Probability distributions of low and high *E. coli* loads of each random split by the BNN model are presented above. PDFs show that uncertainty in the estimation has a small bias but a large variance. These uncertainties mainly stem from flow and other water quality data.

2.6 Summary and conclusions

This study provides a Bayesian Neural Network (BNN) model for *E. coli* prediction in streams. A significant contribution of this paper is in identifying six key variables from a selection of physical, chemical, and biological factors that influence *E. coli* loads in surface streams. An exhaustive feature selection technique used in conjunction with BNN and the principal component analysis (PCA) indicated the importance and correlation among these six variables. Physical factors included temperature and DO; chemical factors include phosphate and ammonia; biological factors include SS and chlorophyll.

The BNN model was then run using these six factors and a comparison with a traditional model (LOADEST) developed by the USGS was also conducted. The inherent differences between the models are the calibration procedures using statistical (LOADEST) versus probabilistic (BNN) framework. The models were compared for estimation of *E. coli* loads based on available water quality data, Nash–Sutcliffe efficiency (NSE), and normalized mean squared error (NMSE). Both the models were evaluated using three-fold cross validation. All the efficiency measures suggest that estimation of *E. coli* loads by the BNN model was better than the LOADEST model on all the occasions during three-fold cross validation. The results also highlight that the LOADEST model estimates *E. coli* loads better in the smaller ranges, whereas the BNN model estimates *E. coli* loads better at the higher ranges. Hence, the BNN model can be useful to decision maker and environmental managers to design targeted monitoring programs and support real-time decision-making.

An uncertainty analysis is also used to compare the predictive powers of the two models. These results suggest that more uncertainty is associated with larger *E. coli* loads, and signify that the major source of uncertainty comes from storm events associated with these loads. The uncertainty analysis also exhibits that the BNN model estimates *E. coli* loads with a small bias but slightly large variance. This can be avoided in the future with the presence of large datasets and improved monitoring programs.

CHAPTER III

E. COLI FATE AND TRANSPORT BELOW SUBSURFACE SEPTIC TANKS IN THE LAKE GRANBURY AREA

3.1 Synopsis

Pathogens are major polluters of water resources in the state of Texas. *Escherichia coli* (*E. coli*) is an indicator microorganism for fecal contamination. *E. coli* contamination of groundwater (GW) and surface water (SW) occurs through the subsurface from onsite wastewater treatment systems (OWTSs). However, *E. coli* transport in the subsurface remains inadequately characterized at the field scale. Therefore, the aim of this research is to explore *E. coli* transport in the vadose zone under dynamic boundary conditions in a linked SW-GW system. Additionally, this study characterizes the impact of flow regimes on *E. coli* transport in the subsurface and explores parameters that control the transport of *E. coli* in the GW-SW system. This study was conducted in Lake Granbury, which is a critical water supply in North Central Texas providing water for over 250,000 people. In order to consider the impact of the vadose zone on *E. coli* migration, HYDRUS model was used to simulate *E. coli* movement in the unsaturated zone. Resulting *E. coli* concentrations from the simulation of HYDRUS were used as concentration boundary conditions in the MODFLOW/MT3DMS model. Results of unsaturated zone modeling manifest the importance of flow regime and seasonal variability on *E. coli* transport in the subsurface. It was found that there is increased mobility of *E. coli* in winter as compared to summer

months. Sensitivity analysis revealed saturated water content, decay coefficient, and dispersivity as important parameters for *E. coli* transport in the vadose zone. Comparatively, saturated zone modeling demonstrates saturated hydraulic conductivity as the most sensitive parameter followed by first order decay coefficient and partition coefficient of *E. coli*.

3.2 Introduction

Water resources are prone to microbial contamination in rural areas, where a large number of onsite wastewater treatment systems (OWTSs) are present. Onsite/decentralized wastewater treatment systems serve approximately 25% of U.S. households and almost 40% new developments [*Lowe and Siegrist*, 2008; *USEPA*, 2005]. The most common OWTS involves a septic tank unit followed by dispersal to a subsurface soil infiltration unit. Onsite systems are one of many known contributors of pathogens and nutrients to groundwater (GW) [*USEPA*, 2002]. *Escherichia coli* (*E. coli*) is a microbe of fecal origin. Therefore, water contamination by various strains of *E. coli* is becoming common in rural areas in the US [*Bradford et al.*, 2006].

The level of *E. coli* contamination to GW depends upon multiple factors such as precipitation pattern, thickness and composition of the vadose zone [*Williams et al.*, 1998], subsurface heterogeneity [*Spalding and Exner*, 1993]. Physical processes controlling the fate and transport of *E. coli* have been studied extensively [*Haznedaroglu et al.*, 2008]. *Smith et al.* (1985) demonstrated that the extent of *E. coli* transport largely depends on soil structure. They further suggested that flow through soil macropores, which bypasses the retentive capacities of the soil matrix, is a common phenomenon.

Gagliardi and Karns (2000) discussed the impact of soil type and method of pathogen delivery on *E. coli* transport in the subsurface. Another study conducted in a karstic aquifer described that the residence time of *E. coli* in the subsurface, which is correlated with the pore velocity, is crucial for GW contamination [*Personne et al.*, 1998]. Furthermore, there is evidence in the literature that the survival and transport of *E. coli* in the subsurface is controlled by various factors, such as climate (e.g. temperature, rainfall), soil type (e.g. porosity), saturated water content [*Federle et al.*, 1986]. Moreover, hydrological interactions between surface water (SW) bodies and GW are also of fundamental concern to the migration of contaminants (such as, *E. coli*) in a linked SW-GW system [e.g., *Townley and Trefry*, 2000; *McMahon et al.*, 1995; *Bethune et al.*, 1996]. However, despite the importance of SW-GW interaction, the characterization of the impact of GW flow patterns, due to dynamic boundary conditions, on fate and transport of *E. coli* near surface water bodies is still lacking. Therefore, the purpose of this research task is to investigate SW-GW interaction near shallow water bodies and to relate the results to *E. coli* transport through the subsurface to the SW.

The fate and transport of *E. coli* in the subsurface poses a twofold challenge to researchers. First, most of the studies to characterize *E. coli* transport in the subsurface involve packed bed column systems. However, there are many differences between real-world scenarios and the sand columns used in the laboratory. Second, the impact of vadose zone flow processes is oversimplified in GW modeling. For example, recharge and concentration boundary conditions are calculated externally without proper

consideration of the vadose zone. Therefore, a study is desired to understand the transport of *E. coli* in the subsurface under dynamic boundary conditions for a linked SW-GW system. For this study, we present a lose coupling between HYDRUS and MODFLOW to simulate *E. coli* migration below the subsurface zone in the Lake Granbury area. HYDRUS is used to produce *E. coli* flux at the water table that becomes a boundary condition for GW flow and transport models—MODFLOW and MT3DMS. HYDRUS can simulate flow and transport in a variably saturated system, however, HYDRUS does not consider head-dependent groundwater discharge [*Niswonger and Prudic*, 2009]. Moreover, we also note the problem of numerical instability that is associated with solutions for Richards' equation despite significantly improved numerical methods. Therefore, MODFLOW and MT3DMS are used for simulating flow and transport of *E. coli* in the GW under dynamic boundary conditions.

The aim of this research is to explore *E. coli* transport in the vadose zone under fluctuating GW table. In addition, this study investigates and uniquely characterizes the impact of flow paths on *E. coli* transport in the vadose zone and explores parameters that control the transport of *E. coli* in the SW-GW system.

3.3 Study area description and data availability

This study is conducted at a station (Segment ID: 1205; Latitude 32°27'40" and Longitude 97°42'53") in the Lake Granbury area, which is monitored by the Brazos River Authority. Lake Granbury is a man-made lake of 35 km^2 within the Middle Brazos-Palo Pinto watershed (USGS Cataloging Unit: 12060201). The watershed is a part of the Brazos River basin and is located in north-central Texas, in Hood County.

Long, hot summers and short mild winters characterize the climate with average temperatures ranging from a low of 1.1 ˚C in January to a high of 35.5 ˚C in July. The average annual precipitation in the watershed is 760 mm. The watershed has a diversified land use from urban to agriculture; about 31 to 40 % of the land is considered prime farmland. The landscape is characterized as rolling hills of pasture and cropland surrounded by deciduous forest. Lake Granbury is an important source of water supply in this area, providing water for over 250,000 people in more than 15 cities. Recent studies by the Brazos River Authority (BRA) and Espey Consultants, Inc. have concluded that some of the Lake Granbury's coves especially shallow bodies of water are contaminated with *E. coli*. The possible sources of *E. coli* contamination in the lake are septic systems, cattle, pets, and wastewater treatment plants [*BRA and ECI*, 2010].

The water quality data in Lake Granbury were collected under Clean River Programs (CRP) by the Brazos River Authority and the Texas Commission on Environmental Quality. *E. coli* (CFU/100 mL) data in Lake Granbury were available from July 2002 to August 2010. *E. coli* data are grab samples collected monthly at 0.3 meters below the surface. Precipitation data were obtained from the National Climatic Data Center (NCDC) [available at *http://www.ncdc.noaa.gov/*], and evapotranspiration data from TexasET Network [available at *http://texaset.tamu.edu/index.php*].

3.4 Methodology

For this study, HYDRUS-3D is used to simulate *E. coli* migration below the subsurface zone in the Lake Granbury area. The HYDRUS-3D software can be used to model 2-dimensional *E. coli* fate and transport in the vadose zone. Therefore, now on,

we will use HYDRUS to indicate 2-D modeling of *E. coli* transport in the vadose zone. HYDRUS produces *E. coli* flux at the water table which becomes a boundary condition for GW flow and transport models—MODFLOW and MT3DMS. Since, the *E. coli* concentrations obtained from HYDRUS are used as concentration boundary condition; we performed forward modeling in the unsaturated zone. However, calibration of MODFLOW and MT3DMS models' is done using the Parameter Estimation (PEST) software to match estimated and observed *E. coli* concentrations. *E. coli* estimation at observation nodes are compared with available monitoring results. In the subsequent sections, we provide a description of the modeling framework in the Lake Granbury area using the unsaturated zone model—HYDRUS and saturated zone models—MODFLOW and MT3DMS. Subsequently, we perform sensitivity analysis to evaluate appropriate range of parameters.

3.4.1 E. coli *transport modeling in the unsaturated zone*

E. coli transport in the subsurface is explored in the Lake Granbury area using the physically-based HYDRUS hydrologic simulation software [*Šimůnek et al.*, 2006]. The HYDRUS software allows the user to analyze fate and transport of *E. coli* through saturated, partially saturated or unsaturated regions with irregular boundaries, and composed of non-uniform soils. In the Lake Granbury area, there are shallow canals, which are hydrologically connected to the lake (Figure 3.1). Two different hydrologic scenarios are investigated and compared. Boundary conditions with respect to water levels in the lake, in the canal, and in the groundwater system are altered and incorporated in the HYDRUS model as follows:

- The first scenario involves the GW level to be lower than the water level in the lake and the canal.
- The second scenario involves the GW level to be higher than the water level in the lake and the canal.

The GW table fluctuates in response to climatic conditions (precipitation and evapo-tranpiration), and this fluctuation imposes a dynamic boundary condition in the linked SW-GW system (Lake Granbury Area). These scenarios allow us to investigate the possible effects of different hydrologic conditions on the transport of *E. coli* in the subsurface.

3.4.2 Physical domain setup for the unsaturated zone

The HYDRUS software package solves the Richards' equation for water flow and advection dispersion equation for solute transport in variably saturated domains using the finite element method. Advective transport in the vadose zone is the major mechanism for *E. coli* transport. Richards' equation (2-D) is used for describing variably-saturated flow:

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z}\left[K(h)\left(\frac{\partial h}{\partial z} + 1\right)\right] + \frac{\partial}{\partial x}\left[K(h)\left(\frac{\partial h}{\partial x}\right)\right] - S \tag{3.1}$$

where t is time [T], z is the vertical coordinate positive upwards [L], x is the lateral coordinate, θ is the water content [L^3L^{-3}], h is the pressure head [L], K is the unsaturated hydraulic conductivity [LT^{-1}], S is a sink term.

The advection dispersion equation (ADE) is used for modeling solute transport:

$$\frac{\partial C}{\partial t} = D_L \frac{\partial^2 C}{\partial x^2} + D_Z \frac{\partial^2 C}{\partial z^2} - v_x \frac{\partial C}{\partial x} - v_z \frac{\partial C}{\partial z} - \frac{B_d}{\theta}\frac{\partial C_s}{\partial t} + \left(\frac{\partial C}{\partial t}\right)_{bio} \tag{3.2}$$

where c is the concentration of *E. coli* in the liquid phase, D_L and D_Z are the dispersion coefficients, v_x and v_Z are the pore water velocities, B_d is the bulk density, C_s is the amount of *E. coli* sorbed, the subscript bio describes the bio-chemical transformation.

The modeling domain is described in Figure 3.1. An initial time step of 0.1 day and minimum and maximum time steps of 0.01 and 5 day are employed. The flow and transport simulations for *E. coli* are performed for a period of 700 days. The simulation period is chosen based on the time required for *E. coli* to reach the peak concentration at the observation nodes. The results are described with respect to observation nodes as shown in Figure 3.1.

Table 3.1: Initial input parameters for *E. coli* transport in the vadose zone [*Bradford et al.*, 2006; *Gelhar et al.*, 1992; *Mace et al.*, 2000; *Pang et al.*, 2004].

Parameter	Value
Longitudinal dispersivity (m)	1
Longitudinal hydraulic conductivity (m/sec)	3.5×10^{-8}
Attachment coefficient (-/sec)	7.73×10^{-5}
Detachment coefficient (-/sec)	5.33×10^{-7}
Unsaturated *E. coli* removal (decay) rate (-/sec)	2.3×10^{-5}
Saturated *E. coli* removal (decay) rate (-/sec)	2.3×10^{-6}
Transmissivity (m^2/sec)	1×10^{-4}
Effective Porosity(-)	0.3
E. coli concentration in septic tanks (CFU/100mL)	1×10^{6}
Effluent Loading (m^3/sec)	1.05×1

We use attachment/detachment model to describe *E. coli* movement in variably saturated soil. We assume monodispersed and time dependent deposition behavior of *E. coli* for this study. Flow and transport parameters used in the HYDRUS model are listed in Table 3.1. Boundary Conditions (BC) for the HYDRUS and MODFLOW-MT3DMS

models are listed in Table 3.2. As this model contains many nonunique input parameters, a sensitivity analysis is used to indicate which of these parameters are critical for describing *E. coli* transport in the subsurface.

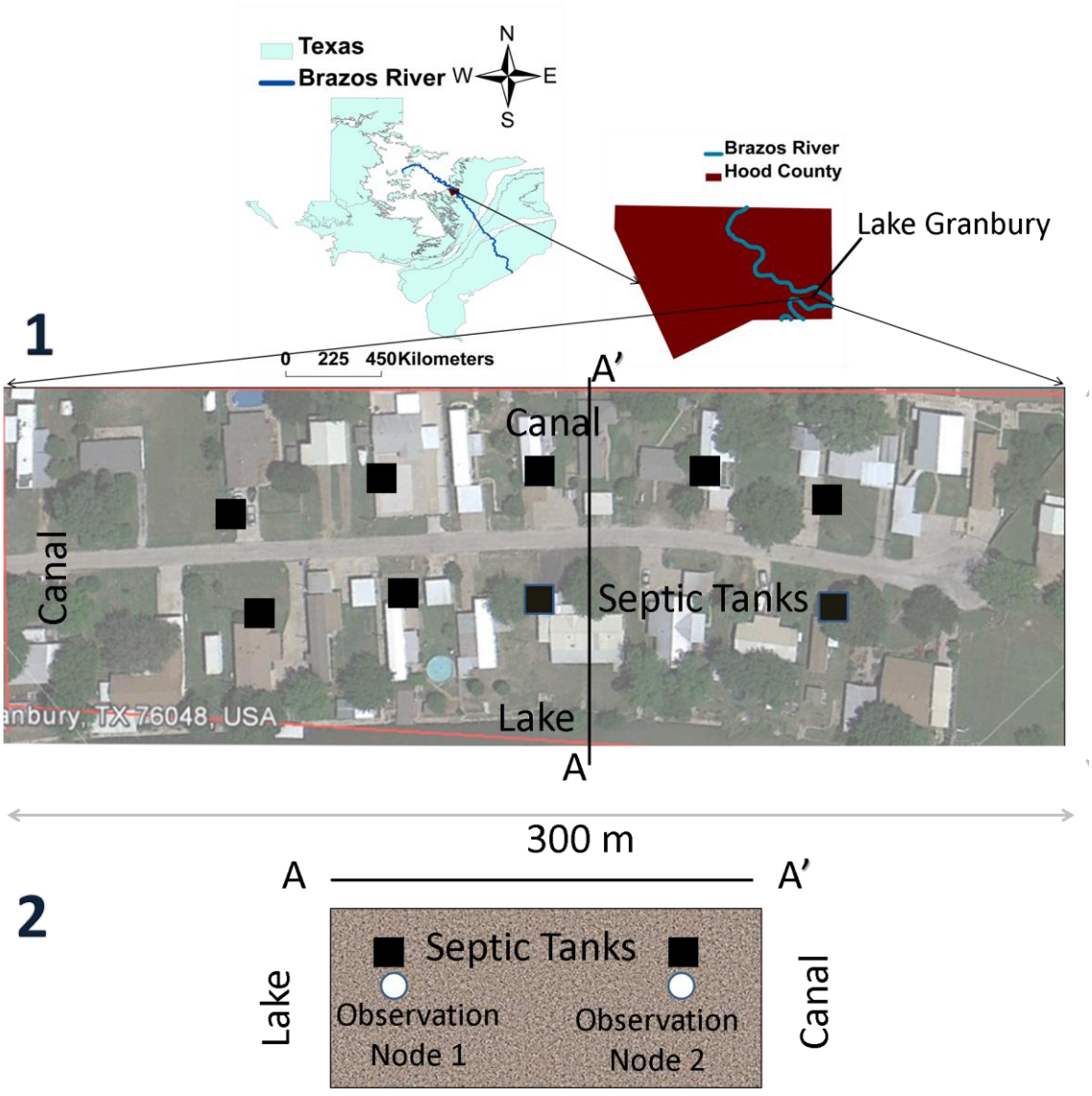

Figure 3.1: Plan view of the model domain (m) for (1) MODFLOW/MT3DMS and (2) cross sectional view of the model domain (m) for HYDRUS 2-D along A-A' transect (Map courtesy: Google Earth).

Table 3.2: Boundary conditions (BC) for HYDRUS and MODFLOW-MT3DMS models

HYDRUS	Boundary Conditions
Hydrologic Upper BC	Atmospheric BC $q(t) = \begin{matrix} 0 \; if \; for \; h(t,z) < 0 \\ q(t,x,z) \end{matrix} for \; h(t,z) \geq 0$
Hydrologic Lower BC	Deep Drainage $q(t) = \begin{matrix} 0 \; if \; for \; h(t,z) < 0 \\ q(t,x,z) \end{matrix} for \; h(t,z) \geq 0$
Hydrologic Lateral BCs	Variable Head $h = h(t)$
Solute Upper BC	No Flux $q_c = 0$
Septic Tanks BC	Constant Flux $q_c = q_{c0}$
Solute Lateral and Lower BCs	Flux Type $q = q_c$(t)
MODFLOW-MT3DMS	**Boundary Conditions**
Hydrologic Upper BC	Atmospheric BC $q(t) = \begin{matrix} 0 \; if \; for \; h(t,z) < 0 \\ q(t,x,z) \end{matrix} for \; h(t,z) \geq 0$
Hydrologic Lower BC	Deep Drainage $q(t) = \begin{matrix} 0 \; if \; for \; h(t,z) < 0 \\ q(t,x,z) \end{matrix} for \; h(t,z) \geq 0$
Lateral BCs	Variable Head $h = h(t)$
Solute Upper BC	Flux Type $q_c = q_c$(t)
Solute Lower BC	Flux Type $q_c = q_c$(t)
Solute Lateral BCs	Flux Type $q_c = q_c$(t)

3.4.3 Sensitivity analysis in the unsaturated zone

The aim of the sensitivity analysis is to estimate appropriate range of flow and transport parameters and ascertain critical values that may lead to optimal solutions. HYDRUS enables two (or three) dimensional flow and solute transport representations

in the unsaturated zone. However, the 2D/3D version of HYDRUS is not equipped with inverse estimation capabilities. Therefore, in this study, sensitivity analysis is carried out by individually varying one-parameter-at-a-time. Each parameter is perturbed by ±30% while keeping the rest of the parameters constant at their assigned values.

3.4.4 Physical domain setup for the saturated zone

E. coli concentrations obtained at the water table, using the optimal parameters in the HYDRUS model for scenarios 1 and 2, were used as the concentration boundary condition (winter and summer respectively) in GW flow and transport models—MODFLOW and MT3DMS.

The MODFLOW [*McDonald and Harbaugh*, 1984], a three-dimensional finite difference model, distributed with a graphical user interface by Waterloo Hydrogeologic (Visual MODFLOW), solves the GW equation for water flow and advection dispersion equation for *E. coli* transport in saturated domains. Advective transport in GW is the major mechanism for *E. coli* transport. GW equation is used for describing saturated flow:

$$\frac{\partial}{\partial x}\left(K_X \frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial h}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z \frac{\partial h}{\partial z}\right) - G = S\frac{\partial h}{\partial t} \tag{3.3}$$

where K is the hydraulic conductivity [LT^{-1}], and h is hydraulic head [L], G is the source or sink term, and *S* is the storativity [L].

For *E. coli* transport, MT3DMS-v5.1 [*Zheng and Wang*, 1999] quantifies advective- dispersive transport process,

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial C}{\partial x_j}\right) - \frac{\partial}{\partial x_i}(v_i C) + \sum R_{reac} \tag{3.4}$$

where C is the total aqueous concentration of solute [ML^{-3}], D_{ij} is the hydrodynamic dispersion coefficient tensor, v_i is the pore-water velocity [LT^{-1}] in direction x_i, and R_{reac} is a source or sink rate due to microbial processes.

The MODFLOW/MT3DMS model domain is described in Figure 3.1. The flow and transport processes are simulated using $1m \times 1m$ uniform grids. The model has one layer, for a total modeling depth of 20 m. Model topography is imported from surveyed elevation data. The top of the model is assigned as a recharge boundary. The lateral limits are defined by physical and hydraulic boundaries (Figure 3.1).

We assume Langmuir adsorption and first order decay models to describe the *E. coli* fate and transport in GW. Flow and transport parameters used for the MODFLOW and MT3DMS models are listed in Table 3.1. The Visual MODFLOW and MT3DMS models are run at steady state, and then calibrated to observed *E. coli* concentrations using PEST software. The calibration is done for a period of 49 months— from July 2002 to January 2006 at observation grid located by the lake. The model is validated for a period of 49 months— from February 2006 to August 2010. Model calibration consists of modifying the flow and transport parameters to minimize the normalized root mean squared (NRMSE) error between estimated and observed *E. coli* concentrations.

3.4.5 Calibration and sensitivity analysis in the saturated zone

PEST is a parameter estimation package, which is model independent. In this study, PEST is utilized to calibrate the saturated zone model parameters such that *E. coli* concentrations match field measurements as closely as possible. PEST implements a particularly robust variant of the Gauss-Marquardt-Levenberg algorithm of nonlinear

parameter estimation [*Marquardt*, 1963]. As a spin-off, PEST also provides linear-based approximations of the uncertainties propagated through parameters that it estimates. The degree of correlation between parameters is also calculated to evaluate the equifinality of the inverse problem. We provide a brief description of the Gauss-Marquardt-Levenberg algorithm of nonlinear parameter estimation used by PEST. Further details can be obtained elsewhere [*WHI*, 2002].

Suppose that the GW model **M,** with the set of parameters b_0, estimates *E. coli* concentrations C_0 as follows:

$$C_0 = M(b_0) \tag{3.5}$$

Concentrations C are generated using another set of parameters b-b_0, where b_0 is in the proximity of b. Taylor's theorem gives:

$$C = C_0 + J(b - b_0) \tag{3.6}$$

where J is the Jacobian matrix of $m \times n$ such that m are the number of observation and n are the first derivatives of one particular observation with respect to each parameter. The objective function is defined using the Gauss-Marquardt-Levenberg algorithm gives:

$$\Phi = (\mathrm{C} - \mathrm{C_0} - \mathrm{J(b - b_0)})^{\mathrm{T}}\mathrm{Q(C - C_0 - J(b - b_0))} \tag{3.7}$$

where Q is an m-dimensional diagonal matrix consisting of squares of weights attached to each observation. The superscript T represents the transpose operation.

The model performance is assessed by normalized root-mean-square error (NRMSE). The Normalized Mean Squared Error (NRMSE) is given as:

$$NRMSE = \frac{\sqrt{E((\hat{C} - C_m)^2)}}{range(C_m)} \tag{3.8}$$

where C_m is measured *E. coli* concentrations, $\hat{C}$ is the estimated *E. coli* concentration at time t, and $range(C_m)$ is the range of measured *E. coli* concentrations. The lower the NRMSE, the better is the model performance.

3.5 Results

This section discusses the key findings from the fate and transport analysis of *E. coli* in the subsurface. In the unsaturated zone, *E. coli* migration was examined at observation nodes 1 and 2 for two year simulations by the HYDRUS. The observation node 1 is located near the lake, whereas the observation node 2 is located near the canal (Figure 3.1). In order to explore important features and processes controlling *E. coli* transport in the subsurface, scenario analysis was conducted based on the different water levels in the lake and canal and GW. Subsequently, resulting *E. coli* concentrations from the simulation of the HYDRUS model were used as concentration boundary conditions in the MODFLOW and MT3DMS models. The results of these simulations were then used to evaluate *E. coli* migration into the lake and canal. An important part of the modeling process was to use field data to calibrate and independently test the predictive capabilities of the MODFLOW and MT3DMS models. The HYDRUS simulations were run from 2003 to 2004, whereas MODFLOW-MT3DMS simulations were run from July 2002 to August 2010. The important results of this study are summarized below.

3.5.1 E. coli *transport in the unsaturated zone*

The GW table fluctuates in response to climatic conditions. For example, in Lake Granbury site, winter (October to May) precipitation is often higher than summer (June to September) precipitation [*Harmel et al.*, 2003] and so GW storage is not fully recharged in summer. Figure 3.2 demonstrates the higher precipitation and lower evapo-transpiration in winter for years 2003 and 2004.

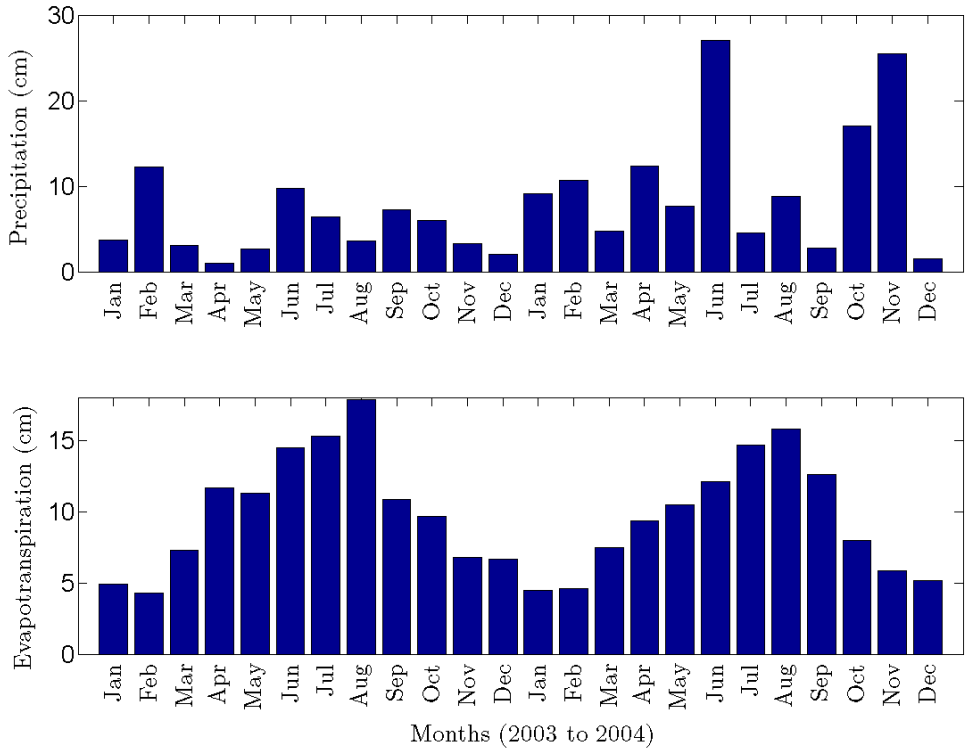

Figure 3.2: Precipitation and evapo-transpiration in the Lake Granbury area for 2003 and 2004 demonstrate that June to September have high evapo-transpiration and lower precipitation as compared to October to May (Data: TWDB).

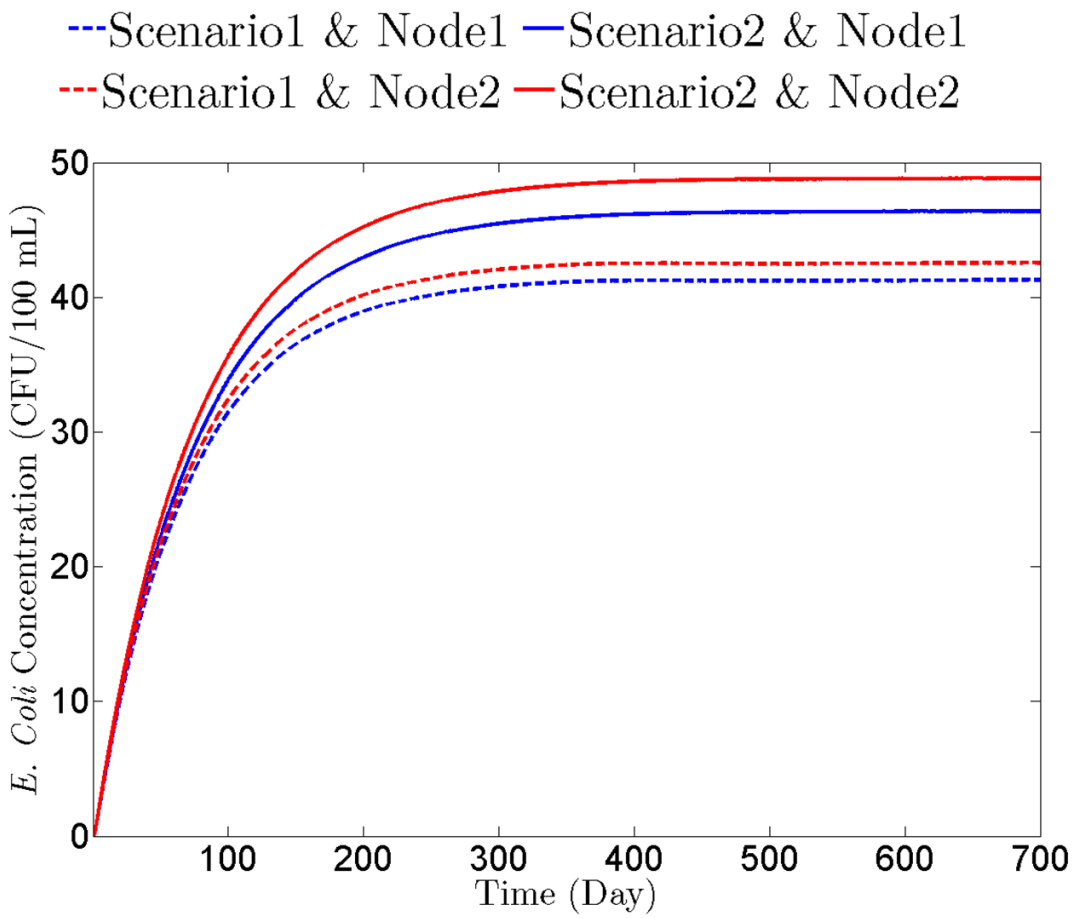

Figure 3.3: The breakthrough curves of *E. coli* concentration at observation nodes 1 and 2 demonstrate the impact of scenarios 1 and 2 on *E. coli* transport in the subsurface. Scenario 1 represents summer and scenario 2 represents winter.

As a result, the water table is lower in the summer period (scenario 1). Therefore, the first scenario, where the GW level is lower than the water level in the lake and the Canal, essentially represents the hydrologic conditions in the summer. Likewise, the second scenario, where the GW level is higher than the water level in the lake and the canal, represents the hydrologic conditions in the winter. Consequently, GW gains water from the lake during summer time; GW loses water to the lake during winter time.

Figure 3.3 describes the characteristics of the *E. coli* breakthrough curves providing information on the peak concentrations of *E. coli* at observation nodes 1 and 2 (Figure 3.1) for scenarios 1 (summer) and 2 (winter). It is evident from the figure that the peak concentrations of *E. coli* are smaller in the case of summer months as compared to winter months for both observation nodes 1 and 2. These results indicate increased mobility of *E. coli* in winter as compared to summer. In addition, Figure 3.3 also demonstrates that higher peak concentrations of *E. coli* are obtained at observation node 2 than observation node 1 for both scenarios 1 (summer) and 2 (winter). This suggests that *E. coli* concentrations migrate towards the canal.

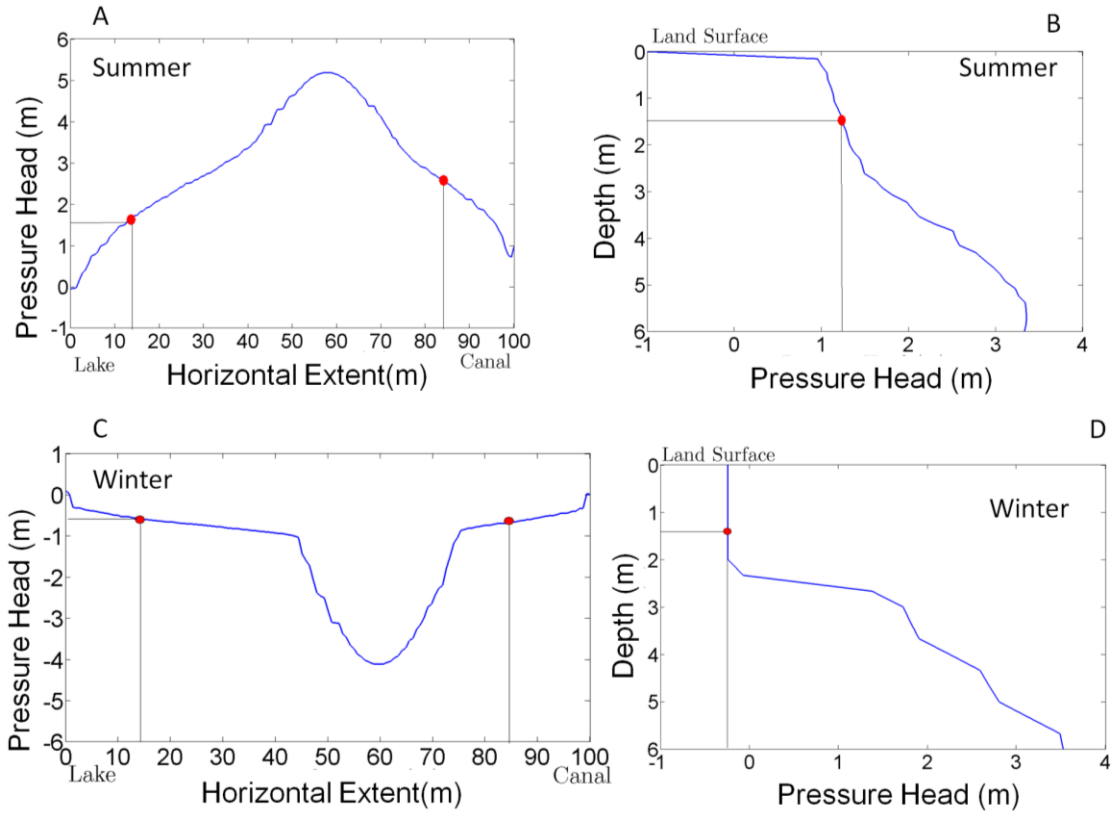

Figure 3.4: The horizontal (1.5 m below land surface) and vertical profiles (across septic tanks near the Lake) of pressure head demonstrate the impacts of scenario 1 (summer) and 2 (winter) on the extent of the saturated and unsaturated zones.

The *E. coli* concentration profiles in vertical and horizontal direction were examined in order to explore the important characteristics of *E. coli* transport in the subsurface. Figure 3.4 shows the change in pressure head profiles during different scenarios (winter and summer).

Figure 3.5 represents the concentration profiles of *E. coli* in the horizontal (1.5 m below land surface) and vertical directions along transect A-A' for scenarios 1 (summer) and 2 (winter). The lake and canal are located at 0 m and 100 m, respectively, on the X-axis. Septic tanks are located at 15 m and 85 m on the X-axis (Figure 3.5).

Figures 3.5 (A and C) again demonstrate higher concentrations near the lake as well as the canal during winter times (scenario 2) as compared to summer times (scenario 1). As GW loses water during winter times (Figure 3.4), there is advective transport towards the lake and canal from GW. Since the major transport modes of *E. coli* in the subsurface are through advective transport [*Jamieson et al.*, 2003], higher *E. coli* concentrations are obtained at observation nodes 1 and 2 during winter times. In addition, winter times (scenario 2) shows small fluctuations in *E. coli* concentrations near septic tanks (Figure 3.5). These fluctuations may be attributed to infiltrating water because there is higher precipitation in the Lake Granbury area in the winter months. The figure also depicts that *E. coli* concentration drops quickly towards the center of the modeling domain (42 to 58 m on the X-axis). This phenomenon—higher *E. coli* concentration near the lake and canal and *E. coli* concentration dropping to zero towards the center of the modeling domain— shows that a GW divide exists in the modeling domain. A GW divide is defined by a hypothetical line on either side of which GW moves in opposite directions. The GW divide occurs because of the shallow water table, which is strongly influenced by surface water flow. The GW divide prevents *E. coli* movement through advective transport towards the center of the modeling domain. As diffusive transport is smaller than advective transport, lower *E. coli* concentrations are observed in this central region.

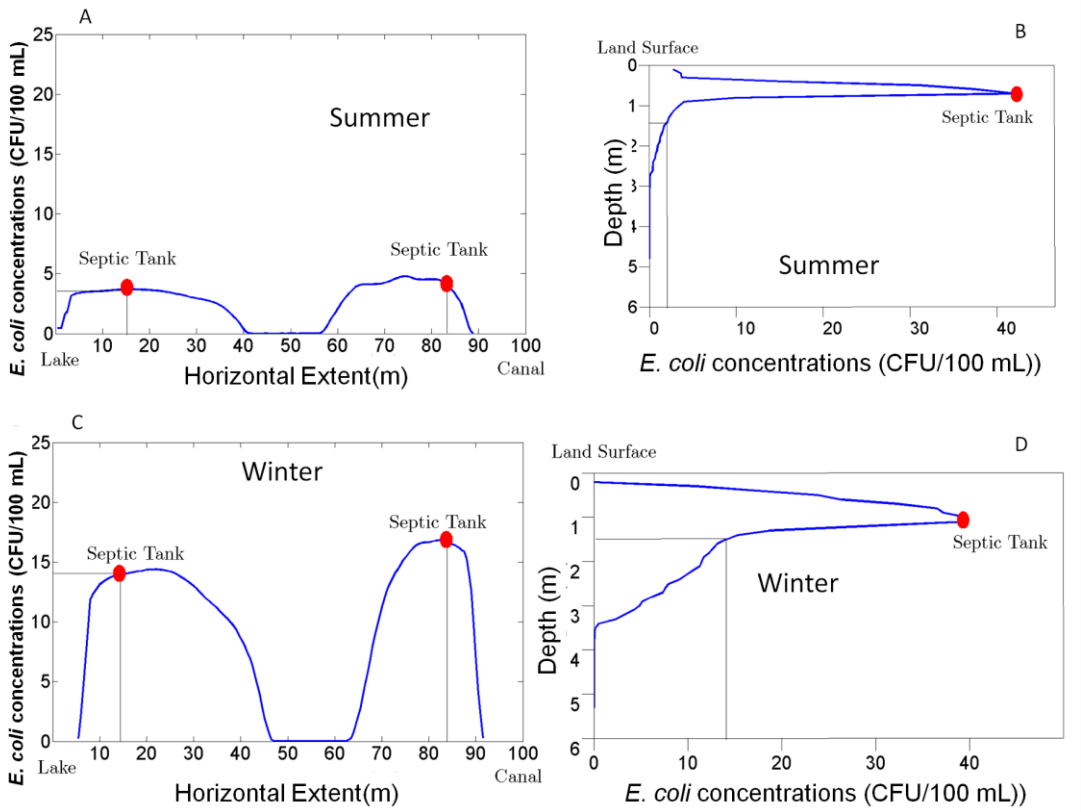

Figure 3.5: The horizontal (1.5 m below land surface) and vertical profiles (across septic tanks near the Lake) of *E. coli* concentration demonstrate the impacts of scenario 1 (summer) and 2 (winter) on *E. coli* transport in the subsurface.

Figures 3.5 (B and D) represent the concentration profile of *E. coli* in the vertical direction (by the septic tanks) for scenarios 1 (summer) and 2 (winter). Figures 3.5 (B and D) manifest little upward retention of *E. coli* for both the scenarios. It is evident from the figure that the soil matrix within a depth of 0.5 m contains 90% of *E. coli* from the source (septic tanks). The figure further demonstrates that *E. coli* concentration gradually changes for scenario 2 (winter) than for scenario 1 (summer), where *E. coli* concentration drops quickly. These results demonstrate the importance of the flow regime and seasonal variability (climatic) on *E. coli* transport in the subsurface.

3.5.2 Sensitivity analysis in the unsaturated zone

Processes for investigating *E. coli* transport at large scales are currently lacking. Thus, these results should be interpreted and treated within the limits of uncertainty of the model and parameters (flow and transport). Nevertheless, the model provides useful insight into the *E. coli* transport processes. The objective of SA was to apportion the uncertainty in the predictions (*E. coli* concentration) to different sources of uncertainty in the flow and transport parameters. One-parameter-at-a-time (OAT) approach was implemented, where each parameter was altered by ±30% while keeping the rest of the parameters constant at their assigned values. It was found that *E. coli* concentration was largely affected by three parameters: saturated water content, decay coefficient, and dispersivity. *E. coli* transport in the vadose zone is contingent upon advection and dispersion. It is known that the saturated water content and transport properties of soils bear strong nonlinear relationships, and water content status in unsaturated porous media directly influences advective fluxes [*Or et al.*, 2007].

As seen in Figures 3.4 and 3.5, dispersive transport is significant for *E. coli* transport in the zones, where advection is not the dominant mode of *E. coli* transport. Decay coefficient reflects the residence time of *E. coli* in soil, and is therefore an important parameter.

3.5.3 E. coli *data and moving average*

As seen in section 3.5.1, seasonal variability greatly influences *E. coli* transport in the subsurface. Therefore, the purpose of presenting autocorrelation function (ACF) and partial autocorrelation function (PACF) of *E. coli* concentrations was to describe the seasonal occurrence of *E. coli* in the lake. The ACF describes the similarity between *E. coli* observations as a function of the time lag between them. The higher ACF values show repeating patterns, such as seasonality. The PACF identifies the extent of the lag in an autoregressive model.

Figure 3.6 demonstrates ACF and PACF of measured *E. coli* in the lake. It is evident from the figure that there are high and low observed *E. coli* values. The ACF and PACF demonstrate a repeating pattern after 7 and 10 months. This is because climatic seasonality has an effect on *E. coli* transport; a repeating pattern of 7 months is because of winter (October to May), and the repeating pattern of 10 months (which occurs after 3 months of winter) is because of summer (June to September). These results corroborate the findings of the scenario analysis on *E. coli* transport in the subsurface.

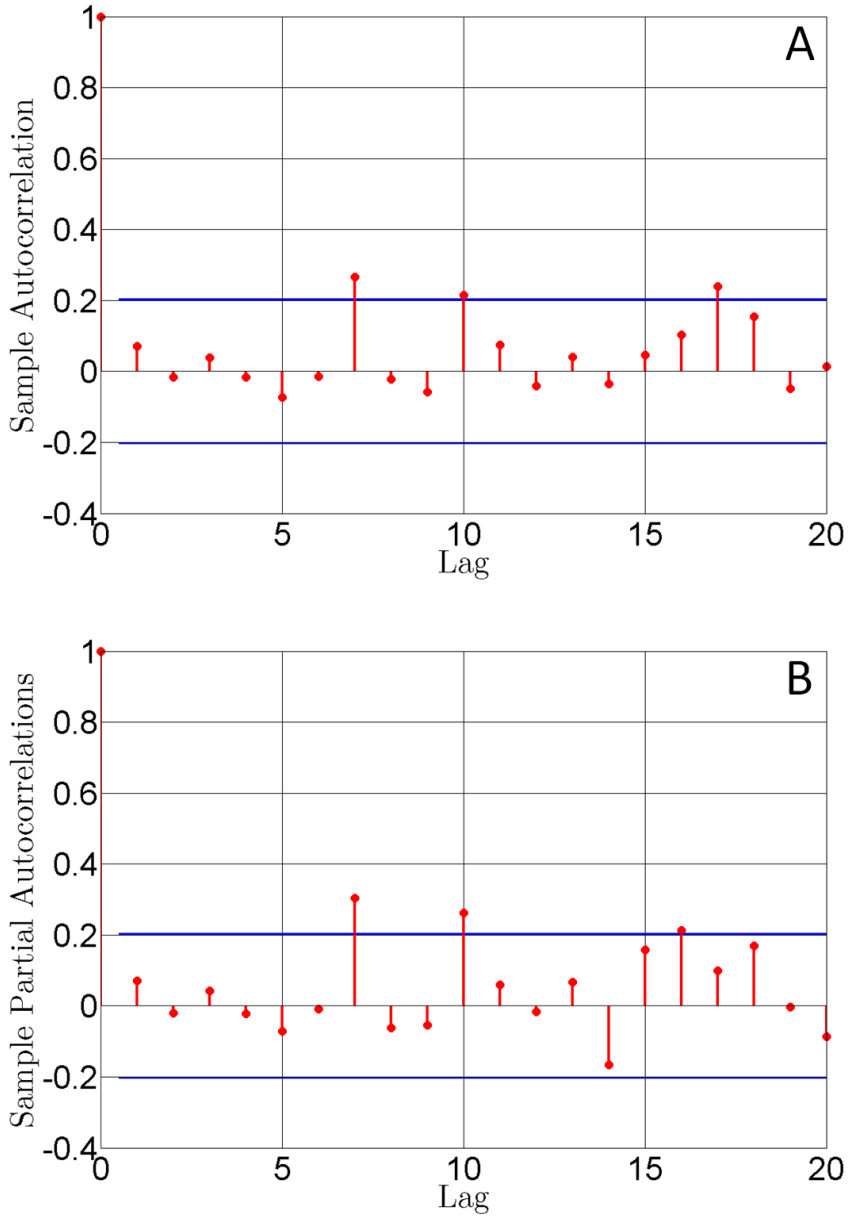

Figure 3.6: The (A) autocorrelation (ACF) and (B) partial autocorrelation (PACF) functions of *E. coli* concentrations in the lake describe the seasonality in the observed *E. coli* concentrations. The ACF and PACF demonstrate a repeating pattern after 7 and 10 months.

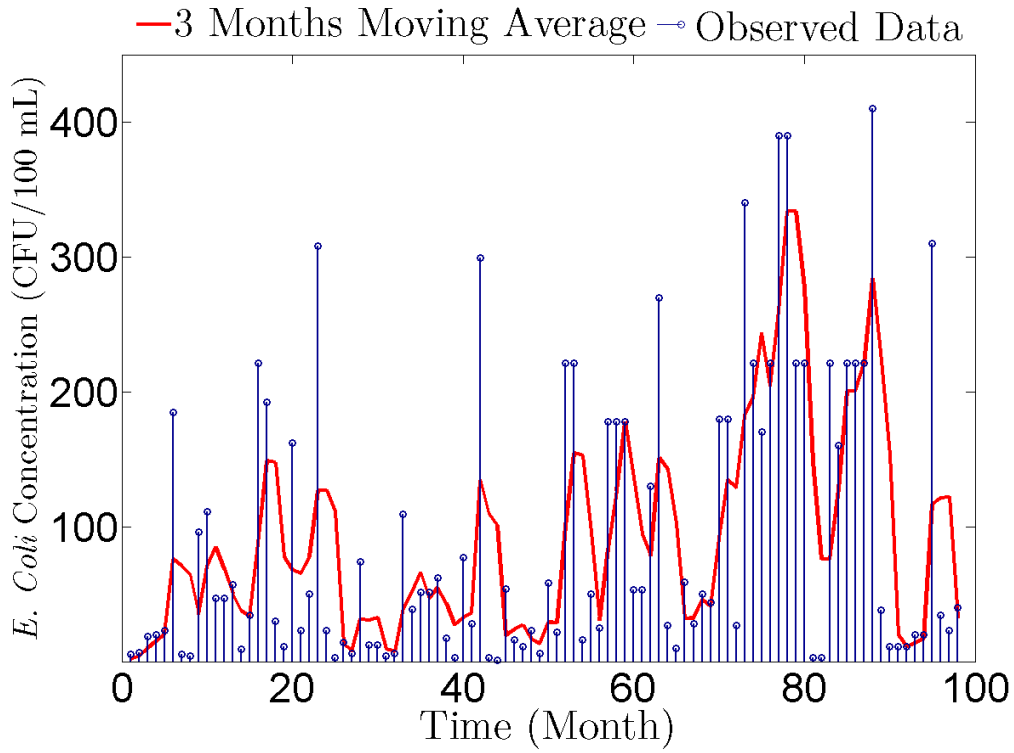

Figure 3.7: The 3-month moving average of *E. coli* concentrations in the canal shows the seasonality in *E. coli* concentrations.

Based on the ACF and PACF analysis, the moving average was selected to smooth out the variability of monthly *E. coli* data (Figures 3.6 and 3.7). 3-month moving average was chosen because this is the time lag between winter and summer months. The 3-month moving average *E. coli* concentrations were used for calibration and validation in the MODFLOW and MT3DMS models. The 3-month moving averages of *E. coli* concentrations are more appropriate than monthly values of observed *E. coli* concentrations, because the presence of *E. coli* in the lake is a continuous phenomenon. Smoothing out the fluctuations allows users to reduce uncertainty in the observed *E. coli* data. The important findings of the saturated zone modeling are summarized below.

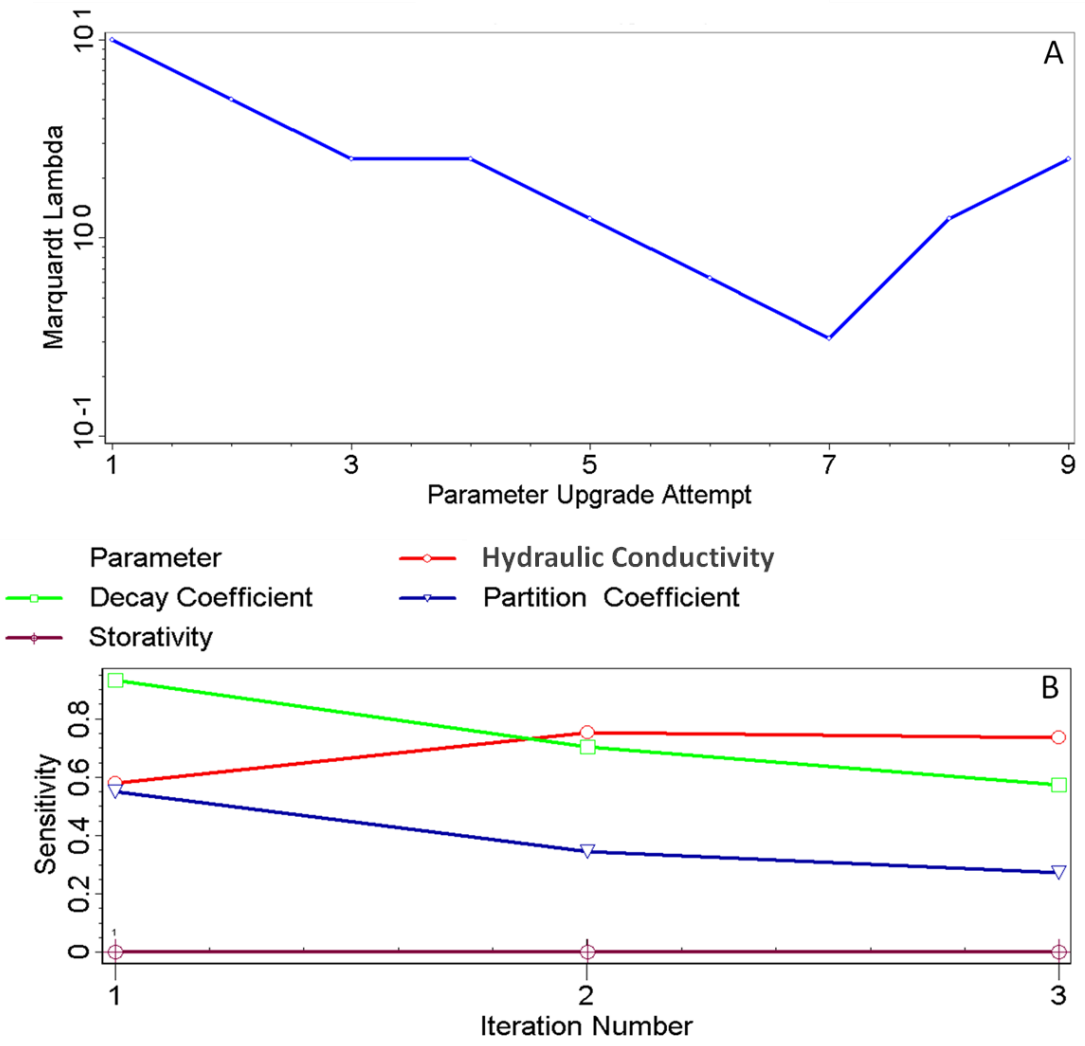

Figure 3.8: (A) The Marquardt Lambda with parameter estimation attempts exhibits the performance of PEST during the calibration of MODFLOW/MT3DMS. (B) The change in the sensitivity of parameters with the number of iteration implies that the sensitive parameter is more accurately estimated than other parameters.

3.5.4 E. coli *transport in the saturated zone*

The aim of the saturated zone modeling was to characterize the subsurface system in its totality using the known hydrogeology of the site. MODFLOW simulated the hydrogeologic regime at the site. MT3DMS simulates advection, dispersion, and first order decay of *E. coli* in GW. PEST software package was employed to calibrate the MODFLOW/MT3DMS models. In this section, we first describe the parameter sensitivity results from the calibration using PEST. Subsequently, we describe results obtained from the calibrated MODFLOW/MT3DMS models.

3.5.5 Model calibration and sensitivity analysis in the saturated zone

Figure 3.8A demonstrates the performance of PEST during the calibration of the saturated zone models. The plot shows the change in the Marquardt Lambda with parameters estimation attempts. The lower values of the Marquardt Lambda indicate that the process of optimization is relatively easy for PEST, while higher values indicate that the process is relatively challenging. In other words, higher values indicate that there is higher uncertainty in the system. It is evident from Figure 3.8A that the value of the Marquardt Lambda decreased after each parameter upgrade attempt (from 1 to 7). This indicates that initial upgrade attempts provided a sufficient reduction in the residual of the objective function (Equation 3.8). PEST increased the Marquardt Lambda with parameter upgrade attempt (from 7 to 9) to obtain optimal parameters in the gradient descent direction.

Figure 3.8B illustrates the change in the sensitivity of different parameters (decay coefficient, storativity, hydraulic conductivity, partition coefficient) with the number of

iteration. It is evident from the figure that saturated hydraulic conductivity is the most sensitive parameter followed by the first order decay coefficient and partition coefficient. The greater sensitivity of hydraulic conductivity implies that this parameter is more accurately determined than other parameters.

3.5.6 Simulation results of MODFLOW and MT3DMS

The calibrated groundwater flow and solute transport model was developed to characterize the Lake Granbury site for possible contamination of the lake from a large population of existing OWTSs. Figure 3.9 shows reasonably good agreement between the *E. coli* concentrations estimated from the MODFLOW/MT3DMS and the observed *E. coli* concentrations in the lake for calibration and validation periods. The normalized mean squared error (NMSE) value is 0.56 (calibration) to 0.59 (validation) from the simulated and observed *E. coli* concentrations. Furthermore, the coefficient of correlation (R) values range from 0.63 (validation) to 0.66 (calibration) from the simulated and observed *E. coli* concentrations. Overall NMSE and R are 0.58 and 0.64, respectively. However, there is uncertainty associated with the estimated/observed *E. coli* concentrations. The confidence interval (CI) band in Figure 3.9 reflects uncertainty in the model output. More than 78% observed *E. coli* values fall within the confidence limits of ±σ (1 standard deviation). Nevertheless, 22% *E. coli* values are not within the limits of the CI, and this mismatch between observed and estimated concentrations of *E. coli* can be a result the assumption that the sole source of *E. coli* contamination is from septic tanks. Therefore, it is believed that as more data become available, these uncertainties will decline. The present modeling effort is best used to provide a

manifestation of the relative contributions of the different processes causing *E. coli* contamination in the lake rather than absolute contributions. The flow paths in the Lake Granbury area suggest that the main body of canal is more at risk of impairment than the lake.

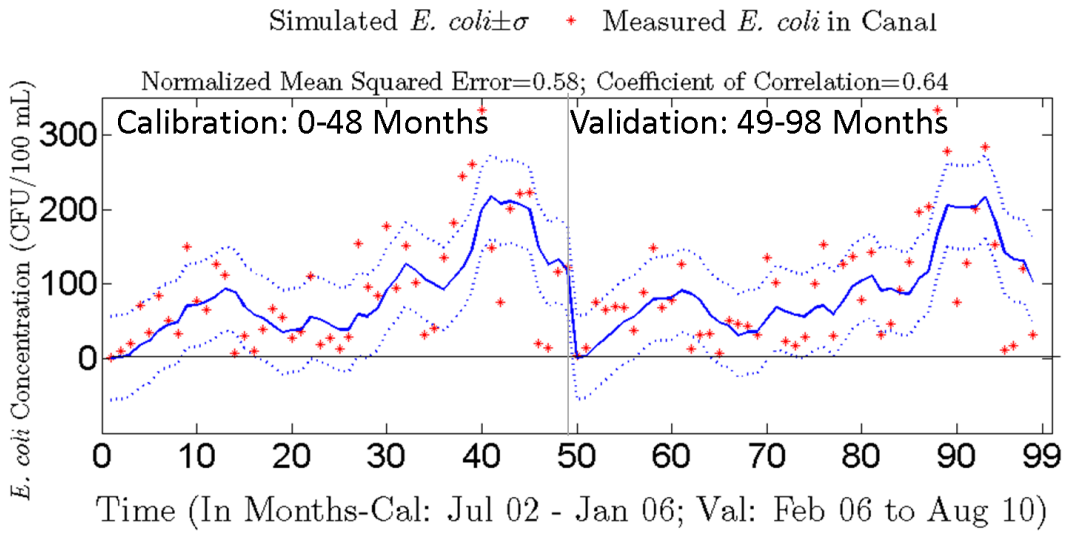

Figure 3.9: Estimated *E. coli* concentrations from the calibrated MODFLOW and MT3DMS models and observed *E. coli* concentrations in the lake show that 85% observed *E. coli* values fall within the confidence limits of ±σ (standard deviation). Months from July 2002 to January 2006 (0-48) and months from February 2006 to August 2010 (49-98) were used for calibration and validation, respectively.

3.6 Summary and conclusions

The major goal of this research was to explore *E. coli* transport in the subsurface by characterizing the vadose zone and groundwater (GW) system. Scenario analysis based on the water level in the lake and canal and GW level was implemented to investigate controls of *E. coli* transport in the subsurface. *E. coli* movement in the unsaturated zone was simulated in HYDRUS. Resulting *E. coli* concentrations from the simulation of the HYDRUS model were used as concentration boundary conditions in the MODFLOW/MT3DMS model. The results of these simulations were then used to evaluate *E. coli* migration into the lake and canal. As a significant contribution, this study demonstrates the impact of flow regime and seasonal variability (precipitation and evapo-transpiration) on *E. coli* transport in the subsurface. The results show increased mobility of *E. coli* in winter months as compared to summer months. It was found that 90% *E. coli* are retained in the soil matrix within a depth of 0.5 m from the source (septic tanks). Saturated water content, decay coefficient, and dispersivity are the top three parameters that produced the most sensitivity to modeled *E. coli* concentration in the vadose zone when compared with the optimal HYDRUS simulation. It was found that saturated hydraulic conductivity is the most sensitive parameter followed by the first order decay coefficient and partition coefficient of *E. coli* in groundwater. The simulated MODFLOW/MT3DMS model produces reasonably good agreement between estimated *E. coli* concentrations and observed *E. coli* concentrations in the lake. Simulated *E. coli* values by the MODFLOW/MT3DMS are able to capture much of the variation (78%) of the measured *E. coli* concentrations regardless of some uncertainties associated with the

observed *E. coli* concentrations. These results are useful to decision makers and environmental managers to design targeted monitoring programs and support real-time decision-making.

CHAPTER IV

ENTROPY-BASED ANALYSIS FOR SPATIO-TEMPORAL VARIABILITY OF NITRATE IN TEXAS AQUIFERS ACROSS MULTIPLE SCALES

4.1 Synopsis

Nitrate is considered to be one of the most prevalent contaminants in groundwater (GW). Because of its high solubility and mobility, it is prone to leaching through soils with infiltrating water. High nitrate-N concentration in GW is a human health concern particularly for infants. NO_3^- in GW shows significant spatio-temporal variability which comes from interaction among multiple geophysical factors such as natural and anthropogenic source availability (land use), environmental forcings such as precipitation history, thickness and composition of the vadose zone, types of aquifers (confined or unconfined), aquifer heterogeneity and geology, etc. The spatial variability of nitrate-N reflects the heterogeneous nature of the major variables that affect nitrate-N in GW, such as uneven application of fertilizers, spatial variability of hydrogeology, different amounts of GW pumping at different locations, and other geo-environmental factors. We developed an entropy based method to describe the spatio-temporal variability of nitrate-N at multiple scales in two different hydrogeologic settings— the Trinity and Ogallala aquifers in Texas at fine (2 km ×2 km), intermediate (10 km×10 km), and coarse (100 km×100 km) scales. The Hurst exponent was used to evaluate the long-term persistence and trend in the variability of nitrate-N. A numerical study using Visual MODFLOW was also conducted to verify the effect of different factors on spatial

variability of nitrate-N at multiple scales. Results demonstrate that the spatial variability of aquifer nitrate-N is controlled by the GW pumping activities, and hydraulic conductivity at the fine scale, the complex interactions between rivers and aquifers at the intermediate scale, and by lithology and geology at the coarse scale. There is maximum variability of nitrate-N contamination at the intermediate scale along with the long-term persistence. In the Trinity Aquifer, overall mean nitrate-N has declined with not much change in the temporal variability over each decade from 1940 to 2008. However, in the Trinity Aquifer, percent samples having nitrate-N > 10 mg/L did not change over time. In the Ogallala Aquifer, at all scales, overall mean nitrate-N has increased because of enhanced use of fertilizer since 1940. However, percent samples having nitrate-N > 10 mg/L has significantly decreased over the last seven decades due to the use of more efficient methods of irrigation. The temporal variability of nitrate-N contamination has increased in the Ogallala Aquifer at all scales since 1940.

4.2 Introduction

Nitrate contamination of groundwater (GW) is a growing concern in the world. In a global context, more than 1.5 billion people, and in the United States, more than 50% of the population (155 million) rely on ground water for their primary source of drinking water [*Alley et al.*, 2002]. In recent times, nitrate (nitrate-N) contamination of GW has posed a severe threat to human health. Despite its critical importance, not enough is known about the long-term spatio-temporal variability of nitrate-N in GW to explain the complex mechanisms that regulate its variability.

The causes of contamination stem from both point sources and nonpoint sources. Various sources of nitrate-N in GW include domestic lawns, golf courses, failing septic systems, domestic and municipal wastewater treatment plants, wastewater treatment lagoons, confined animal feeding operations (CAFOs), landfills, and residential areas. Furthermore, different forms of nitrate-N in GW are inorganic nitrogen (from fertilizers), organic nitrogen (from animal manure applied to croplands), and soil organic nitrogen (from mineralization). [*Gormly* and *Spalding*, 1979; *Keeney*, 1986; *Harter et al.*, 2001; *Green et al.*, 2010]. Subsequently, storage of nitrate in the unsaturated zone occurs and may become a significant source to ground water [*Domagalski et al.*, 2008]. Additionally, airborne nitrogen compounds given off by industry and automobiles are deposited on the land by precipitation [*Nolan et al.*, 2002]. Most of the reactive nitrogen from fertilizer migrates into the atmosphere, rivers, and GW, where it acts as a polluter and poses a hazard to humans [*Nolte et al.*, 2010].

The health risks of nitrate contamination in drinking water above 10 mg per liter of nitrate-N are well documented. Nitrate-N can cause methemoglobinemia in infants ("blue-baby syndrome") and neural tube defects in women during pregnancy in the occurrence of higher nitrate-N intake from dietary, drinking water or other nitrate-N sources [*Brender et al.*, 2004]. New studies suggest that there is a positive association between nitrate-N in drinking water and non-Hodgkin's lymphoma and colorectal cancer in humans [*Lundberg* and *Govoni*, 2004; *Showers et al.*, 2008]. For reducing health risks, cleaning up water contaminated with nitrate-N is an expensive and complicated alternative [*McMahon et al.*, 2008]. Hence, it is crucial to understand the spatio-temporal

variability of nitrate-N and their contributing factors to minimize and mitigate contamination in GW.

Nitrate-N patterns or trends in GW show significant variability in space and time scales [*Williams et al.*, 1998]. This spatio-temporal variability comes from interaction among multiple geophysical factors such as sources of nitrogen loading (land use), precipitation pattern, thickness and composition of the vadose zone [*Williams et al.*, 1998], types of aquifers (confined or unconfined), aquifer heterogeneity and geology [*Spalding and Exner*, 1993]. Spatially varying long response times (years to centuries) of aquifers further contribute towards spatio-temporal variability of nitrate-N contamination. These factors also change across seasons and years across the catchment, and therefore an improved understanding of nitrate-N dynamics across multiple spatio-temporal scales is desired. Several studies were conducted to understand the complexities of nitrate-N contamination in GW. *Williams et al.* [1998] demonstrated different spatio-temporal variations of nitrate-N contamination in isolated and clustered wells, as well as for a large diffuse region of the Sierra Pelona alluvial aquifer. They concluded that nitrate-N patterns are critical to constrain possible contamination sources and transport mechanisms. *Refsgaard et al.* [1999] characterized large scale modeling of GW contamination from nitrate-N leaching, and indicated the limitation posed by the upscaling procedure from point to field scale. There are numerous techniques available for quantifying spatio-temporal variability of a system component. For example, geostatistical methods have been applied to examine spatial variability of infiltration [*Loague et al.*, 1990], arsenic contamination in GW [*Ryker*, 2001; *Winkel et al.*, 2008],

aquifer hydraulic properties [*Carle et al.*, 2006], etc. Geostatistical methods usually require an assumption of stationarity, and the interpolated value may be different from the measurement itself [*Mariethoz et al.*, 2009]. Similarly, a principal component analysis (PCA) and K-means clustering technique were used to analyze the spatial and temporal patterns of droughts [*Santos et al.*, 2010]. The applicability of PCA is limited by certain assumptions as large variances have important structure and linearity. K-means clustering uses the Euclidean norm, which may not be valid in certain cases.

Our aim is to develop an approach for characterizing spatial and temporal variability of nitrate-N at different scales (fine, intermediate, and coarse), and to identify the sources of variability in two different hydrogeologic settings— the Trinity and the Ogallala aquifers in Texas (Figures 4.1 and 4.2). The absence of continuous long-term water quality data for many regions in these aquifers impedes progress toward resolving the spatial complexity of the nitrate-N patterns. For the same reason, the gridded reconstruction of water quality dataset was developed in three spatial scales— fine ($2\ km \times 2\ km$), intermediate ($10\ km \times 10\ km$), and coarse ($100\ km \times 100\ km$) grids.

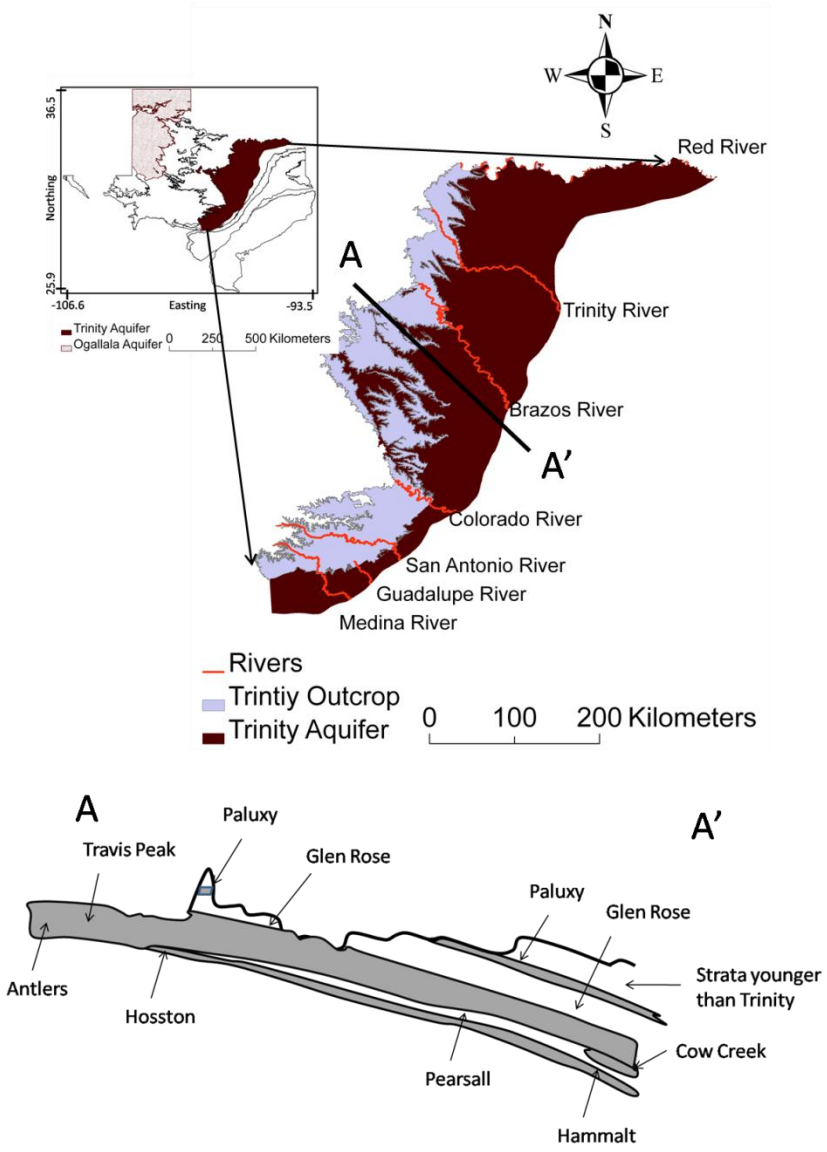

Figure 4.1: Map showing the Trinity Aquifer and its outcrop area. The Trinity Aquifer is a sandstone-carbonate rock aquifer and partly confined. The Trinity Aquifer consists of early Cretaceous age formations of the Trinity Group, from the Red River in North Texas to the Medina River of South-Central Texas. Major Rivers in the study area include: the Red, Trinity, Brazos, Colorado, San Antonio, Guadalupe, and Medina rivers.

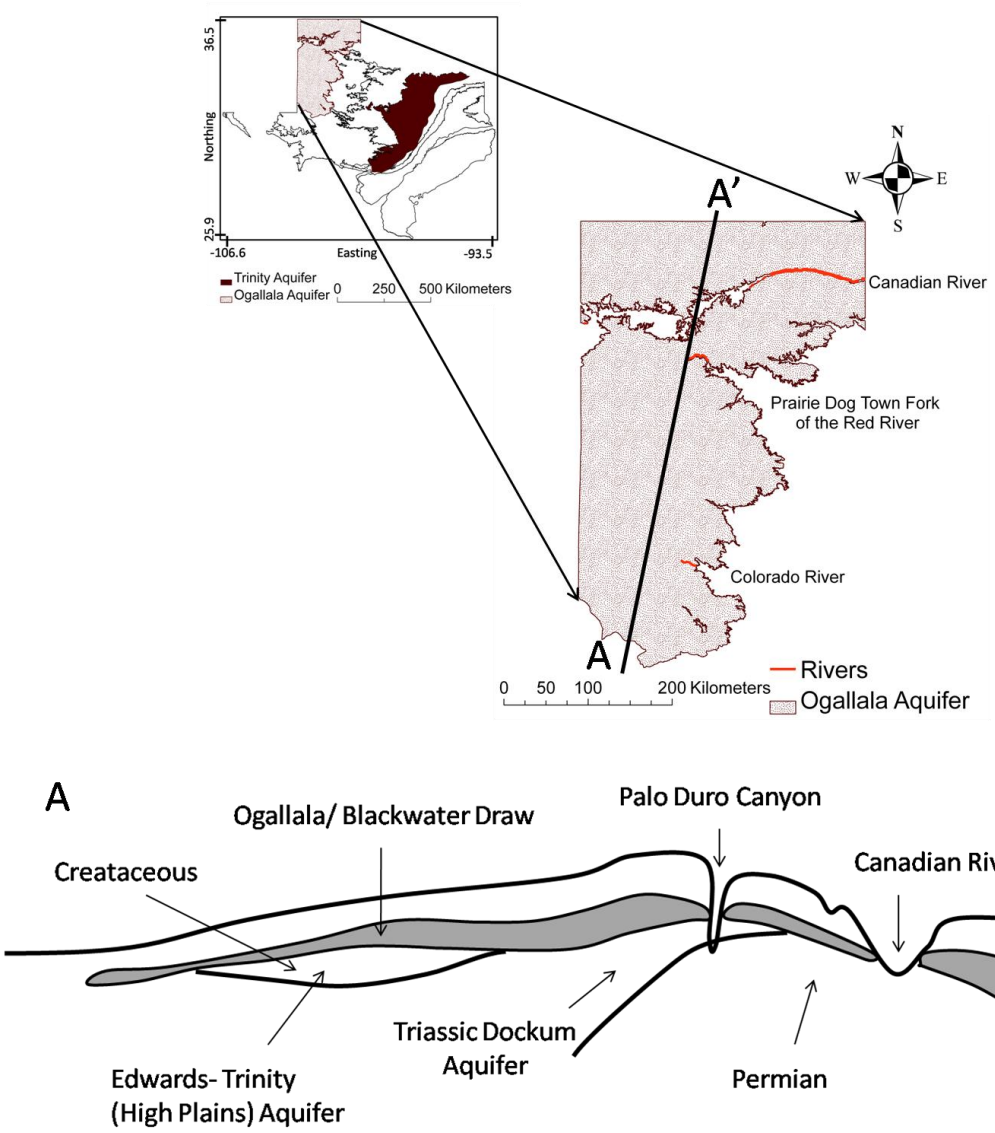

Figure 4.2: Map showing the Ogallala Aquifer, which is an unconfined aquifer and primarily composed of sand, gravel, clay, and silt deposited during the Tertiary Period. Major Rivers in the study area include: the Canadian, Prairie Dog Town Fork Red River, and Colorado Rivers. The part where the Canadian River wears away demarcates the Ogallala into two separate flow systems referred to as the Northern and Southern High Plains. The Ogallala and the underlying Cretaceous, Jurassic, and Triassic formations are hydrologically connected in many areas. These hydrologic connections also exist between the overlying Quaternary Blackwater Draw Formation wherever present.

The specific objectives of this study are:

- To develop an entropy based approach and investigate the spatial variability and corresponding physical controls of nitrate-N associated with fine, intermediate, and coarse scale,
- To investigate the temporal variability and corresponding physical controls of nitrate-N on a decadal scale and to compare disorders within/across decadal scale, and
- To analyze long-term trend and persistence associated with randomness in the distribution of nitrate-N.

4.3 Methodology

In order to analyze the spatial variability of nitrate-N, we utilized an entropy based approach, which is further verified by using numerical modeling and scenario analyses. An information flow schematic of the method is provided in Figure 4.3.

4.3.1 Estimation of variability

Spatial variability is characterized by randomness, periodicity, discontinuity, or due to systematic variation. Within this framework, the terms unpredictable, uncertainty, unstructured, and complex are analogous to variability. Temporal variability indicates randomness, periodicity, or discontinuity in time at a given spatial location. Different descriptive statistics, such as range, mean, standard deviation, and coefficient of variability are widely used for measuring variability in a space or time series.

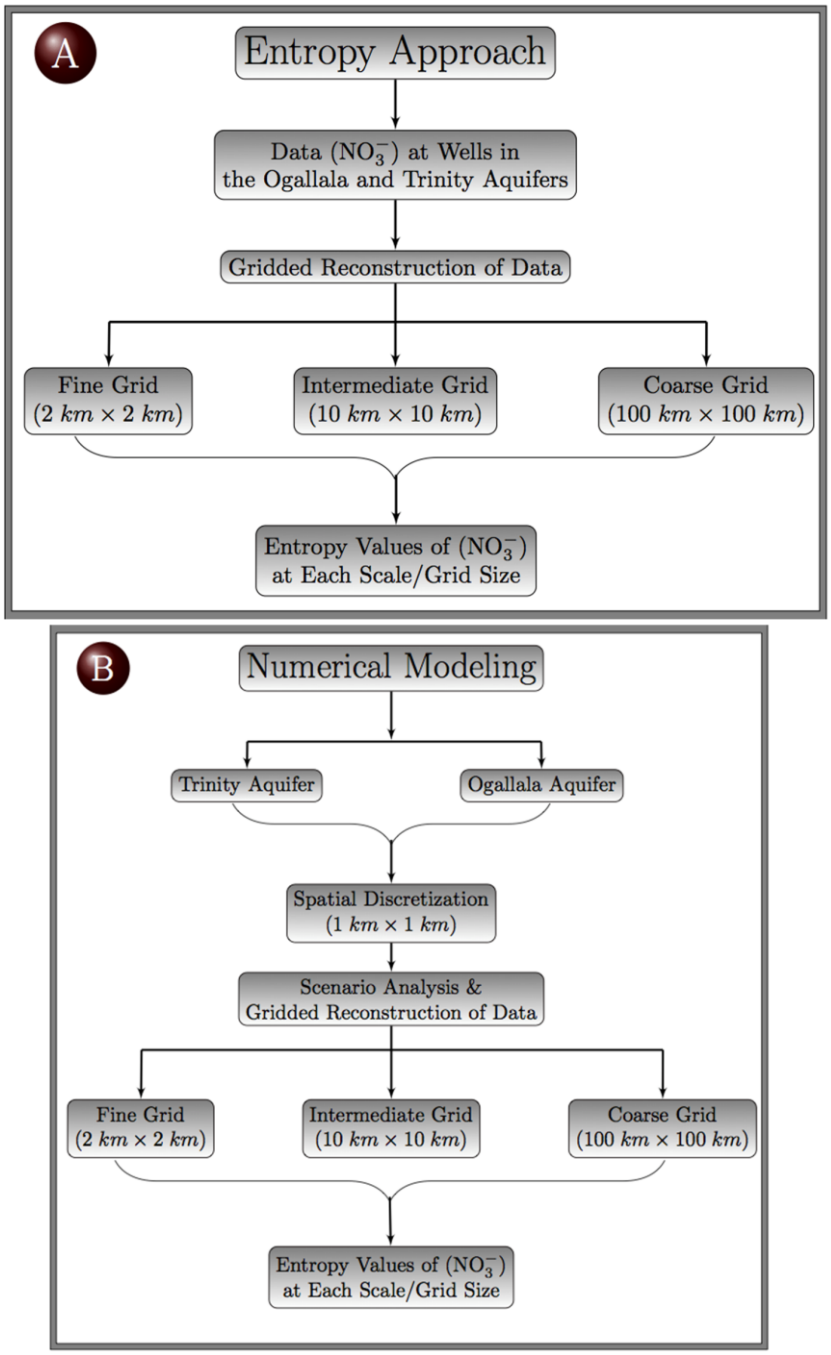

Figure 4.3: A schematic diagram of the data flow and analyses. Entropy approach (A) was utilized to find the controlling factors at different scales for the variability of nitrate-N. The results were verified by numerical modeling (B).

The entropy measure is a probabilistic approach to measure the spatio-temporal variability of a random variable. In information theory, entropy is a measure of the uncertainty associated with a random variable [*Shannon*, 1948]. The term entropy or Shannon entropy usually quantifies bits of the information contained in a dataset. The entropy concept has been widely used in hydrology and other fields and is a well established methodology for measuring spatio-temporal variability and evaluating model performance [*Chapman*, 1986, *Mishra et al.*, 2009; *Sharma et al.*, 2000], application to GW quality monitoring networks [*Mogheir et al.*, 2002], providing guidelines for analysis of time-series data from experiments [*Fraser*, 1989], and cross spectrum analysis of large river flows and the El Nin**õ**-Southern Oscillation [*Khan et al.*, 2006]. We used entropy concept here for analyzing the spatio-temporal variability of nitrate-N in the Ogallala and Trinity aquifers in Texas. The variability obtained through this concept is dimensionless. A brief summary of the entropy concept is provided below for completeness.

4.3.2 Entropy

For a variable with probability density function (PDF) $f_X(X)$, the information entropy is:

$$H = -\int_{-\infty}^{\infty} f_X(x) \log\big(f_X(x)\big)\, dX \tag{4.1}$$

A discrete form of entropy H(x) is given as:

$$H = -\sum_{n=1}^{N} p(X_n) log p(X_n) \tag{4.2}$$

where n denotes a discrete data interval, x_n is an outcome corresponding to interval n, and $p(x_n)$ is the probability of x_n. In a more general mathematical sense, a mapping of m_i

that counts the number of observations that fall into various disjointed categories or bins is desired for calculating the corresponding probabilities. The optimal number of bins can be chosen by using various methods such as Sturges' formula [*Sturges*, 1926], Scott's choice [*Scott*, 1979], Freedman-Diaconis' choice [*Freedman and Diaconis*, 1981], etc. We used Scott's choice for its ability to take into account the integrated squared error. The number of bins using Scott's choice is given as:

$$k = \left\lceil \frac{\max(X) - \min(X)}{h} \right\rceil \tag{4.3}$$

$$h = \frac{3.5\sigma}{N^{1/3}} \tag{4.4}$$

where k is the number of bins, braces indicate the ceiling function, X is the random variable, σ is the sample standard deviation, and N is the number of data points.

The function log (.) can be used with base 2, e, or 10, resulting in information entropy with units of bits, napiers, or decibels, respectively [*Amorocho and Espildor*, 1973; *Mays et al.*, 2002]. We used 2 as the base of the log function in this paper. Information entropy measures the extent to which the probability of each value of a random variable conforms to the noninformative probability distribution of the system, which is the uniform distribution of a finite range in X. Therefore, H reaches its maximum value if all states are equiprobable. In other words, the more informative distribution we have, the lesser will be the disorder and the entropy value. For a certain event, the probability is 1 and the entropy becomes 0. There is no upper bound for entropy, thus for eliminating the effect of different number of data points in datasets and comparing entropies between datasets, we used a normalized measure of entropy, which is given as:

$$H_N = \frac{\max(H)-H}{\max(H)} \times 100; 0 \leq H_N \leq 100 \quad (4.5)$$

where H_N is the normalized marginal entropy (NME) of X, and b is the number of bins. For maximum entropy, all bins will have an equal number of observations; and probability will be log (b). For example, if there are b bins with m observations in each bin, then the entropy will be:

$$H = -\sum_{n=1}^{k} \left(\frac{m}{m \times b}\right) log \left(\frac{m}{m \times b}\right) \quad (4.6)$$

This equation can be simplified to:

$$H = -\sum_{n=1}^{k} \left(\frac{1}{b}\right) log \left(\frac{1}{b}\right) = \log(b) \quad (4.7)$$

Hence, maximum entropy is log of the number of bins (b), and H_N is given by:

$$H_N = \frac{\log(b) - H}{\log(b)} \quad (4.8)$$

The higher the NME, the lower will be the entropy and the variability.

4.3.3 Mutual information

The mutual information (also known as the trans-information, or transfer entropy or in generalized form as the information redundancy) measures the codependency between system variables. Mutual information or transfer entropy has been applied in the study of measuring the directionality and time scale of information flow between pairs of ecohydrological variables [*Ruddell* and *Kumar*, 2009], outlining the pathways through which soil moisture and meteorological phenomena mutually influence one another at local, regional and global scale [*Entekhabi et al.*, 1996, *Eltahir, 1998*], and investigating the spatial variability of soil moisture [*Western* and *Grayson*, 1998]. Others have used

this approach for examining the dependence of complexity of the unsaturated flow process on both geologic heterogeneity and uncertainty due to the flow dynamics [*Mays et al.*, 2002], as well as spatio-temporal variability or randomness in a system induced by complicated local dynamics and large-scale observations [*Vastano* and *Swinney*, 1988].

Mutual information measures the symmetrical statistical information shared between two distributions [*Dembo et al.*, 1991]. Let X and Y be the random variables and I(X, Y) denotes the mutual information between X and Y, and H(X) and H(Y) are the marginal entropies of X and Y respectively.

$$I(X,Y) = \sum_{i=1}^{M} \sum_{j=1}^{N} p(x_i, y_i) log \frac{p(x_i|y_i)}{p(x_i)} \tag{4.9}$$

where x_i and y_j are ith and jth terms in random variables X and Y respectively. Equation 4.9 can be simplified as follows:

$$I(X,Y) = \sum_{i=1}^{M} \sum_{j=1}^{N} p(x_i, y_i) log \frac{p(x_i,y_i)}{p(y_i)p(x_i)} \tag{4.10}$$

Conditional entropy can be defined as the average of entropy of X for each value of Y weighted according to the probability of getting that particular value of Y. Mathematically,

$$H(X|Y) = \sum_{i=1}^{M} \sum_{j=1}^{N} p(x_i, y_i) log p(x_i|y_i) \tag{4.11}$$

Mutual information computes the expectation of the ratio of dependence between the joint distribution of X and Y and what the joint distribution would be if X and Y were independent. Using equations 4.9, 4.10, and 4.11, it can be shown that:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{4.12}$$

where H(X, Y) is joint entropy of X and Y. It is clear from equation 4.12 that I(X, X) =H(X). As I(X, Y) is the mutual information between X and Y, therefore I(X, Y) ≤ min (H(X), H(Y)). I(X, Y) cannot exceed the individual information content of X or Y. There is no upper bound for I(X, Y), so for easier interpretation and comparisons, a normalized version of I(X, Y) that ranges from 0 to 1 is used. The geometric mean of H(X) and H(Y) was used because of the analogy with a normalized inner product in Hilbert space [*Strehl and Ghosh*, 2002]. Thus the normalized mutual Information (NMI) is given as:

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \tag{4.13}$$

where NMI(X, X) =1 and if X and Y are independent then NMI(X, Y) =0.

In the present study, the mutual information between nitrate-N and chloride (Cl^-) concentration in GW was considered.

4.3.4 Normalized risk index

We define a new index the Normalized Risk Index (NRI) for quantifying the distribution of hot spots (nitrate-N > 10 mg/L). The NRI can be extended to any variable depending upon the criterion of hot spots for the variable. For nitrate-N, if there are k spatial locations (well or grid) in an aquifer, there will be k time series datasets. Let N be the total number of cases having nitrate-N > 10 mg/L at a given spatial scale and n_i be the total number of cases for a particular dataset or time series (grid in our case).

$$N = \sum_{i=1}^{k} n_i \tag{4.14}$$

where k is the total number of grids or datasets. Risk Entropy (RE) is given as:

$$RE = \sum_{i=1}^{k} \left(\frac{n_i}{N}\right) \log\left(\frac{n_i}{N}\right) \tag{4.15}$$

where RE measures the density of hot spots in an aquifer. If there is only one time series data set which has all cases (samples having nitrate-N > 10 mg/L), RE will be zero, and if all of the *k* time series datasets have equiprobable number of cases then RE will be maximum i.e. log(k). For eliminating the effect of number data points in datasets and comparing entropies, we used a normalized measure of risk entropy. Normalized Risk Index (NRI) is given as:

$$NRI = \frac{\max(RE) - RE}{\max(RE)} \tag{4.16}$$

or $$NRI = \frac{\max(k) - RE}{\max(k)} \tag{4.17}$$

The range of NRI is between 0 and 1. Therefore, the higher the NRI, the lesser will be the number of hot spots in the aquifer at a particular spatial scale.

4.3.5 Hurst exponent

Since the British hydrologist, H.E. Hurst, who first used the rescaled adjusted range analysis (**R–S** analysis or Hurst Exponent) in hydrology [*Hurst*, 1951], Hurst Exponent is widely used in different fields as an indicator of the irregularity, dependence or persistence such as in analyzing the fractal properties of river networks [*Maritan et al.*, 1996], financial markets for price fluctuations [*Bak et al.*, 1997] and digital signal processing [*Goldberger et al.*, 2002; *Maritan et al.*, 1996]. Hurst constant **H** always lies between 0 and 1, and equals 0.5 for processes that have independent increments e.g. Brownian motion. A Hurst exponent can be used to analyze the trends of nitrate-N in GW. A Hurst exponent value **H** between 0 and 0.5 indicates the ‘‘anti-persistent behavior.” This means that an increase in nitrate-N concentration, in a well for a

particular year, will tend to be followed by a decrease in nitrate-N concentration in the same well next year and vice versa. This behavior is called mean reversion. The intensity of the mean reversion increases as **H** tends to 0. A Hurst exponent **H** between 0.5 and 1 indicates the ‘‘persistent behavior”, that is, the time series has a strong trend, (either increasing or decreasing). The larger the **H** value is, the stronger is the trend. Series of this type is easier to predict than series falling in the other category (mean reversion).

A detailed description of how the Hurst exponent is calculated is provided elsewhere [*Sakalauskienė*, 2003; *http://en.wikipedia.org/wiki/Hurst_exponent*]. A time series of length *N* is divided into shorter time series of length *n*=*N*, *N/2*, *N/4* … The average rescaled range is then calculated for each value of n. To calculate the average rescaled range for each partial time series of length *n*, $X=X_1, X_2,\ldots,X_n$, the following steps are outlined:

- Compute the mean; $m = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Generate a mean adjusted series; $Y_t = X_t - m; for\ t = 1,2,\ldots,n$
- Calculate the cumulative deviate series $Z_t = \sum_{i=1}^{t} Y_i \quad ; for\ t = 1,2,\ldots,n$
- Compute the range R; $R(n) = \max(Z_1, Z_2, \ldots., Z_n) - \min(Z_1, Z_2, \ldots., Z_n)$
- Rescale the range, by dividing the range by the standard deviation *S*.
- Calculate the mean of the rescaled range for all sub-series of length *n*; $(\frac{R}{S})_n = \frac{1}{n}\sum_{i=1}^{n}(\frac{R_i}{S_i})$

- Finally, the value of the Hurst exponent is obtained using an ordinary least square regression with log(n) as the independent variable and $n$log(R/S) as the dependent variable. The gradient of the fit is the estimate of the Hurst exponent.

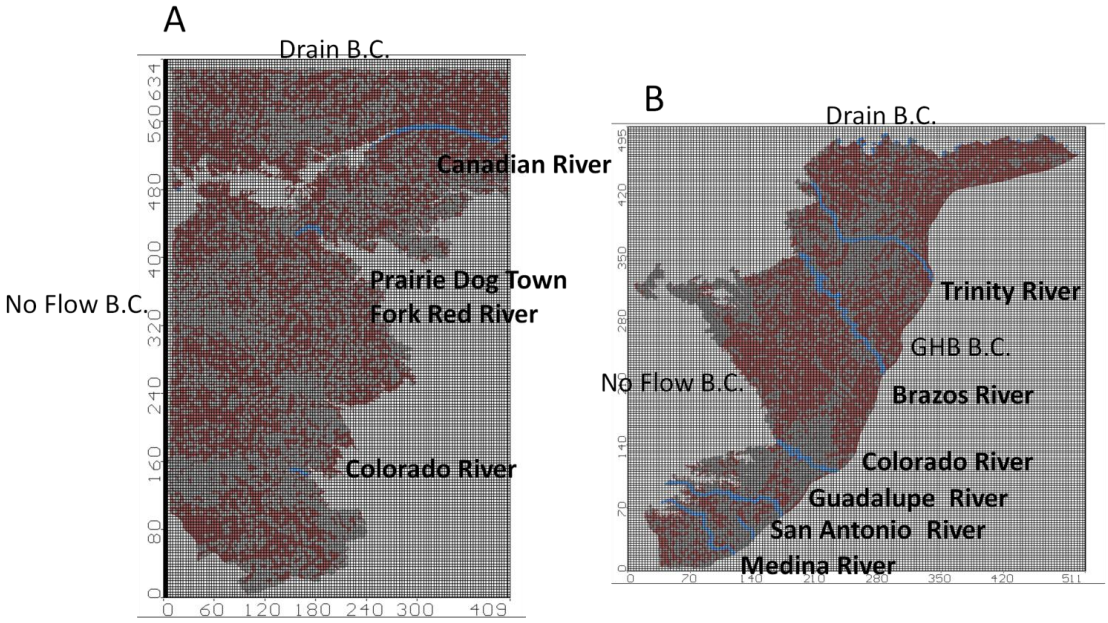

Figure 4.4: Plan view of model domains (km) demonstating wells and boundary conditions (B.C.). The (A) Ogallala and (B) Trinity aquifers were simulated at a discretization of 1 km x 1 km grid size.

Table 4.1: Hydrogeologic and other properties of the Ogallala and Trinity aquifers used for modeling [*Blandford et al.,* 2003 ; *Long et al.,* 2003; *Dutton et al.*, 2004; *Mace et al.,* 2000].

Property	Ogallala Aquifer	Trinity Aquifer
	Parameter range	Parameter range
Hydraulic conductivity in longitudinal direction K_x (m/sec)	7×10^{-7} to 4×10^{-4}	3.5×10^{-8} to 3.5×10^{-3}
Hydraulic conductivity in lateral direction K_y (m/sec)	7×10^{-8} to 4×10^{-5}	3.5×10^{-9} to 3.5×10^{-4}
Hydraulic conductivity in vertical direction K_z (m/sec)	7×10^{-8} to 4×10^{-5}	3.5×10^{-9} to 3.5×10^{-4}
Transmissivity (m^2/sec)	1×10^{-5} to 2×10^{-2}	1×10^{-4} to 6×10^{-2}
Specific storage S_s (1/m)	2.6×10^{-3} to 5.5×10^{-2}	4×10^{-6} to 4×10^{-4}
Specific Yield S_y (-)	0.05 to 0.03	0.1 to 0.3
Effective Porosity (-)	0.2 to 0.3	0.35 to 0.50
Total Porosity (-)	0.2 to 0.3	0.35 to 0.50

Table 4.2: Parameters for nitrate-N transport [*Gelhar*, 1992; *Mehta et al.*, 2000; *Burton*, 2007; *Bronson et al*., 2009].

Model Properties	Values
Longitudinal Dispersivity (m.)	10
Horizontal/Longitudinal Dispersivity (-)	0.1
Vertical/Longitudinal Dispersivity (-)	0.01
Diffusion Coefficient (m^2/day)	5.7×10^{-10}
First Order Reaction Rate for Dissolved Phase (1/sec)	4×10^{-6} to 2×10^{-6}
First Order Reaction Rate for Sorbed Phase (1/sec)	0

4.3.6 Numerical study

We conducted numerical experiments (Figure 4.4) using MODFLOW [*McDonald* and *Harbaugh*, 1984], a three-dimensional finite difference model, with a graphical user interface (Waterloo Hydrogeologic, Visual MODFLOW), to evaluate the effect of different natural and anthropogenic factors on the variability of nitrate-N (e.g., irrigation-pumping, hydraulic conductivities, interaction between the aquifer and river, and geology) at the fine, intermediate, and coarse grids. For completeness, a brief description of numerical model [*McDonald and Harbaugh*, 1984] is given below, and parameters used in simulating the Trinity and Ogallala aquifers are given in Tables 4.1 and 4.2.

The general equation for regional flow of GW derives from the continuity equation,

$$Influx - outflux = -\boldsymbol{\nabla}.\boldsymbol{q} - G = \boldsymbol{S}_S \frac{\partial h}{\partial t} \tag{4.18}$$

where q is flux(m/s), G is the source or sink term, and S_s is the specific storage (m^{-1}). Darcy's law for the flow of GW is given as:

$$q = -K\nabla h \tag{4.19}$$

where K is the hydraulic conductivity, and h is the hydraulic head (m.). The negative sign indicates that GW movement is from high to low water level. Combining equations 4.18 and 4.19 yields the general form of the governing equation for GW flow:

$$\frac{\partial}{\partial x}\left(K_X \frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial h}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z \frac{\partial h}{\partial z}\right) - G = S \frac{\partial h}{\partial t} \tag{4.20}$$

where *S* is the storativity, and it is equal to the saturated thickness of the aquifer times specific storage (S_S).

For nitrate-N transport, MT3DMS-v5.1 [*Zheng and Wang*, 1999] quantifies advective- dispersive transport process, the governing equation for GW flow:

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial C}{\partial x_j}\right) - \frac{\partial}{\partial x_i}(v_i C) + \sum R_{reac} \tag{4.21}$$

where C is the total aqueous concentration of solute (mg/L), D_{ij} is the hydrodynamic dispersion coefficient tensor, v_i is the pore-water velocity in direction x_i, and R_{reac} is a source or sink rate due to chemical reactions.

4.3.7 Application of the modeling framework

The model was run for various scenarios by altering certain features (binary state, for example, typical/normal pumping versus intense pumping), one at a time. In order to evaluate the effects of typical pumping vs. intense pumping, geology (layering vs. no layering), available hydraulic conductivities vs. high hydraulic conductivities, and interactions between GW (aquifer) and surface water (river), the model was run for a series of 48 simulations (24 simulations for the Ogallala Aquifer and 24 simulation for the Trinity Aquifer), varying each of these features. In order to evaluate the effect of intense pumping (50% higher than normal rate), we compared the results for this scenario with that of normal pumping scenario; to evaluate the effect of normal hydraulic conductivities vs. high hydraulic conductivity, we increased the hydraulic conductivity by 50%. For evaluating the effect of geology on the variability of nitrate-N, we carried out simulations with one layer to multiple layers by altering the layer properties (e.g., hydraulic conductivity). For understanding the impact of rivers, we performed GW flow simulations with and without river boundaries. All simulations were

carried out for a 15 year period (1991-2005). The model domains for both the aquifers are shown in Figure 4.4. The streams were represented by the river boundary nodes, when the impact of rivers was evaluated. For temporal discretization, each year was divided into 12 stress periods where each stress period corresponds to one month during which all inputs are constant.

4.3.8 Aquifer descriptions, layers, grid size, model boundary, and parameters

The Texas Water Development Board (TWDB) conducted a study on groundwater availability in Texas. The Groundwater Availability Models (GAMs) for the Trinity and Ogallala aquifers were used for setting up the modeling domain, geology, and other parameters (available at http://www.twdb.state.tx.us/groundwater/models/). A short description of the modeling framework is provided here for completeness.

4.3.9 The Ogallala Aquifer

The Ogallala aquifer is one of the world's largest aquifers, and it covers an area of approximately 450,000 km² in portions of the eight states of South Dakota, Nebraska, Wyoming, Colorado, Kansas, Oklahoma, New Mexico, and Texas [*McGuire*, 2003]. This study critically examines the Ogallala Aquifer with specific focus in the Texas regions. The Ogallala Aquifer, the prime water-bearing entity in the High Plains of Texas, covers an area of approximately 91,815 km² (20% of the aquifer) and provides water to all or parts of 46 counties. Major rivers in the study area include Canadian River, Prairie Dog Town Fork Red River, and Colorado River. The Ogallala is constituted for the most part by sand, gravel, clay, and silt deposited during the Tertiary Period (Figure 4.2).

The significance of the Ogallala Aquifer in the high plains is substantial as the aquifer has long been a principal source of water for agricultural, municipal, and industrial development. The withdrawal of this GW has significantly surpassed the aquifer's rate of natural recharge. Although many communities use the Ogallala Aquifer as their only source of potable water, approximately 30% of the water is used for irrigation in the United States [*McMahon et al.*, 2000]. Previous studies suggest that the Ogallala Aquifer has median nitrate-N concentration exceeding 10.0 mg/l [*Hudak*, 2000].

A basic setup of a single layer model was adopted for the Ogallala Aquifer. We carried out simulations at 1 km x 1 km grids, and then nitrate-N values were computed for each grid. Subsequently, we aggregated the nitrate-N values as a time series in three different grids sizes (fine ($2\ km \times 2\ km$), intermediate ($10\ km \times 10\ km$), and coarse ($100\ km \times 100\ km$) grids) to see the scale effect.

The water-permeated thickness of the Ogallala Formation ranges from a few meters to more than 300 m and is generally greater in the northern plains. The lower boundary of the Ogallala Aquifer was assigned like a no-flow boundary. The top of the model domain was assigned as a recharge boundary, pumping, and return flow from irrigation. Regional recharge rates in the Ogallala aquifer, outside irrigated areas, are generally low (0.1 mm to 45 mm/yr), whereas playa-focused recharge rates are much higher (12 to 218 mm/yr). Irrigated areas also have fairly high recharge rates (15 to 280 mm/yr) [*Scanlon et al*., 2003]. Groundwater discharges from the aquifer through pumping of water wells, and the Ogallala Aquifer has an annual pumpage between

18×10^6 to 12×10^6 cubic m. The outer limits of the model domain were defined by physical and hydraulic boundaries. The eastern and western boundaries of the model domain were defined by the drain and no flow boundaries, and the northern and southern boundaries were maintained as constant head and drain boundaries, respectively. MODFLOW's "river" module was used to denote the interaction of surface and GW along segments of the Canadian River, Colorado River, and Prairie Dog Town Fork Red River. MODFLOW's "general-head boundary" module was used at the southwest side of the model domain between the Canadian River and Prairie Dog Town Fork Red River. Aquifer Properties and model parameters for the Ogallala Aquifer are provided in Table 4.1.

4.3.10 The Trinity Aquifer

The Trinity Group aquifer is an important source of water supply in north-central, central, and southwest-central Texas and covers more than 106,190 km^2 and provides water to all or parts of 52 counties. The Trinity Aquifer consists of early Cretaceous age formations of the Trinity Group, from the Red River in North Texas to the Hill Country of South-Central Texas. Major Rivers in the study area include the Red, Trinity, Brazos, Colorado, San Antonio, Guadalupe, and Medina rivers. The Trinity Group formations encompass the Paluxy, Glen Rose, and Twin Mountains- Travis Peak from the youngest to the oldest. The Antlers Formation, which results from the coalition of the Paluxy and Twin Mountains where the Glen Rose narrows down or almost disappears, is the outcrop part of the Trinity Group (Figure 4.1). Water from the Antlers is mainly used for irrigation on the outcrop area of North and Central Texas. In the

north, the downdip portion of the aquifer is extremely mineralized, and this portion is a source of contamination to drilled wells into the underlying Twin Mountains [*Ashworth*, 1983; *Ashworth and Hopkins*, 1995; *Dennehy et al.*, 2002; *McMahon et al.*, 2006].

The importance for choosing the Trinity Aquifer is because parts of four densely inhabited urban centers: San Antonio, Austin, Fort Worth and Dallas lie directly over the sub-crop of the Trinity Aquifer. The outcrop and sub-crop areas of the Trinity Aquifer provide water for drinking, irrigation and industrial purposes for almost seven million people in this region. The Trinity Aquifer is a vital resource but due to aquifer characteristics and land use, the aquifer has a history of nitrate-N contamination [*Ryder,* 1996].

A basic setup of a seven layer model was adopted for the Trinity Aquifer. These layers correspond to different geologic formations (Figure 4.1). We carried out flow and transport simulations at 1 km x 1 km grids, and then nitrate-N values were computed for each grid. Subsequently, we aggregated the nitrate-N values as a time series in three different grids sizes (fine ($2\ km \times 2\ km$), intermediate ($10\ km \times 10\ km$), and coarse ($100\ km \times 100\ km$) grids) to see the scale effect.

Thickness of the aquifer ranges from a few meters to more than 300 m. The lower boundary of the Trinity Aquifer was defined as a no-flow boundary. The top of the model was assigned as a recharge boundary and pumping. Recharge rates in the Trinity aquifers generally range from 2.5 to 50 mm/yr [*Scanlon et al*., 2003]. Groundwater discharges from the aquifer through pumping of water wells whereas the Trinity Aquifer has an annual pumpage between 12×10^6 to 18×10^6 cubic m. The outer limits were

defined by physical and hydraulic boundaries. The eastern and western boundaries were assigned as no flow and general head boundaries respectively. The northern part of the aquifer was assigned drain boundary conditions. The extreme southern part of the model was assigned with general head boundary conditions for simulating the interaction of the Trinity with the Colorado River. There are some faults in the Trinity model domain, which were simulated using a horizontal flow barrier (HFB) Package. Aquifer Properties and model parameters for the Trinity Aquifer are provided in Table 4.1.

4.3.11 On-ground nitrogen loading

The spatial distribution of the on-ground nitrogen sources were characterized using the National Land Cover Database (NLCD) as prepared by the United States Geological Survey (Figure 4.5). The annual on-ground nitrogen loadings were taken from various sources [*Nolan et al*., 2002; *Scanlon et al*., 2003**;** *Almasria et al.*, 2007; *Bronson et al*., 2009]. The total amount of nitrogen from manure, fertilizer, dairy farm lagoons, and irrigation was calculated based on their application rate (Figure 4.6) and NLCD; the total amount of nitrogen was assigned to the model domain accordingly.

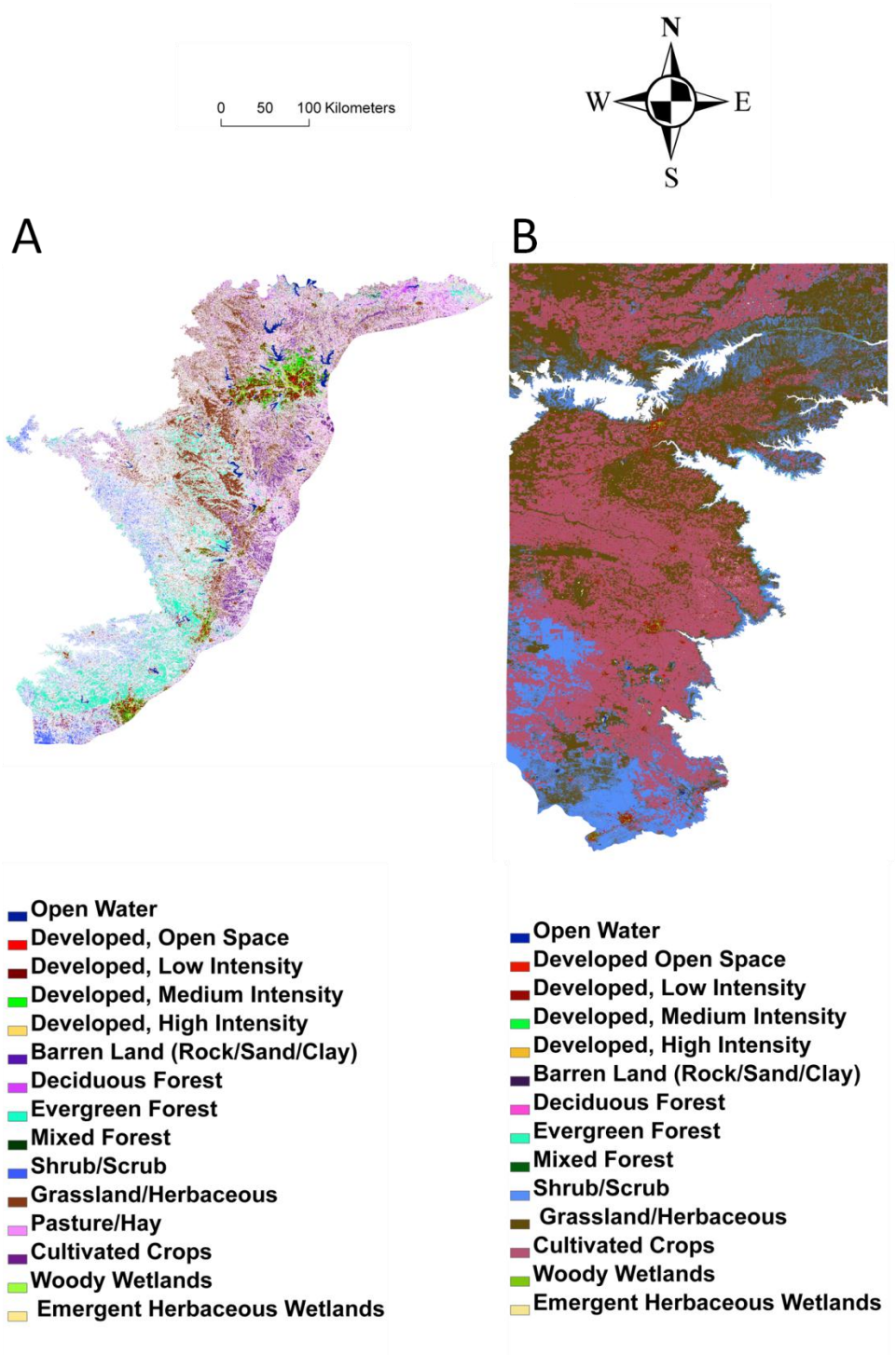

Figure 4.5: Land use maps for the (A) Trinity and (B) Ogallala aquifers.

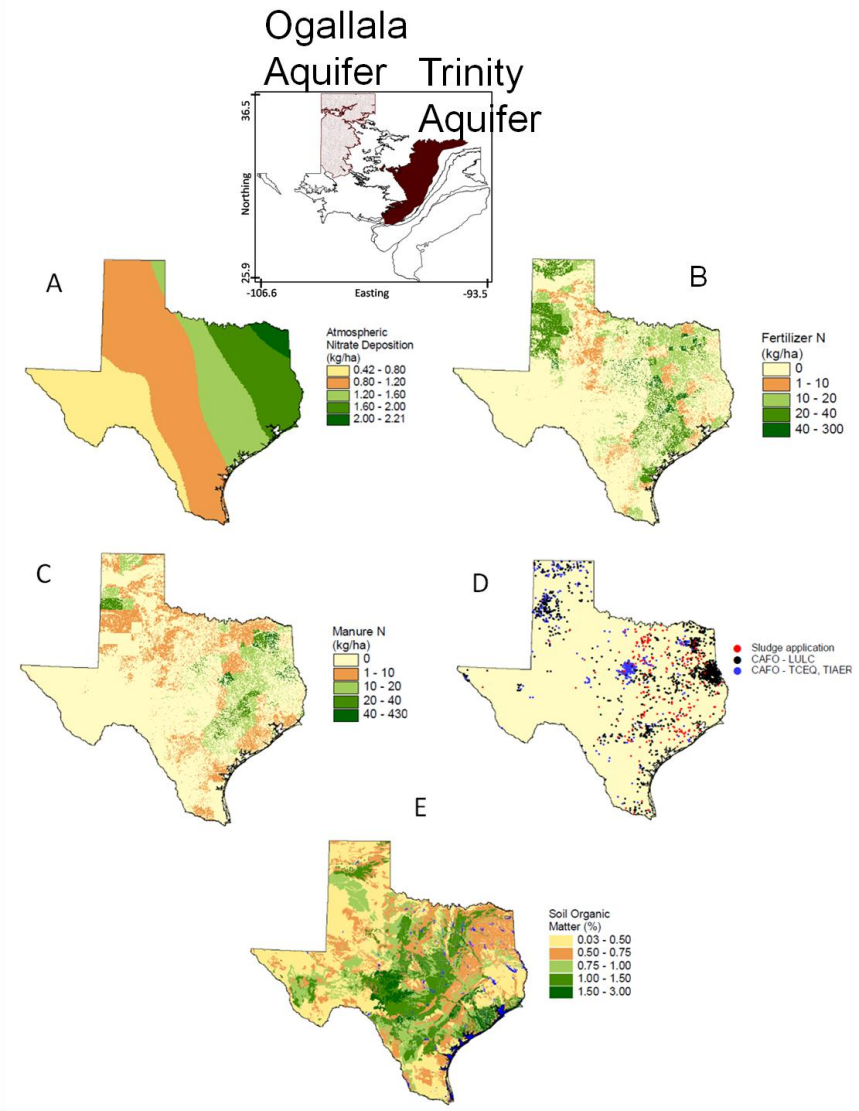

Figure 4.6: Maps showing (A) average annual atmospheric nitrate-N deposition based on National Atmospheric Deposition Program (NADP) data, (B) spatial distribution of inorganic nitrogen fertilizer application, (C) spatial distribution of organic fertilizer (manure) application, (D) distribution of concentrated animal feeding operations (CAFOs) based on data from TCEQ, TIAER, and USGS and permitted sludge application based on data from TCEQ, and (E) average soil profile organic matter derived from STATSGO database (From *Scanlon et al.*, 2003).

For example, the fertilizer application rate was multiplied by the acreage of that crop for nitrogen loading:

$$\sum Nitrogen\ Loading = (NLCD\ CA) \times (R) \quad (4.22)$$

where $NLCD\ CA$ is the NLCD class area and R is the $fertilizer\ application\ rate$. For other NLCD classes and dairy lagoons (wherever present), nitrogen loadings were calculated similarly. The initial conditions were set as the background concentration (mean nitrate-N concentration from 1980 to 1990) of nitrate-N for the entire model domain to predict the ground water nitrate-N.

4.3.12 Model assumptions and limitations

Nitrate concentration measured at a particular well depends upon a number of factors such as irrigation, nutrient loading occurrences, extent of vadose zone, removal time in soil, denitrification and other soil-nitrogen dynamics, precipitation events, well depth below water table, etc. We made following simplifying assumptions due to unavailability of data on certain details of the nitrogen cycle for studying the impact of different factors on the variability of nitrate-N in GW. The model assumes a uniform distribution of nitrogen for a particular land cover. The application of nitrogen on agriculture land will vary depending upon different crops, but we have assumed a uniform distribution. Some model parameters were adopted from the literature (e.g., percentage of nitrogen species in manure and inorganic fertilizers, fertilizer application rate, etc.). Some NLCD classes were merged into one class such as the developed areas with low intensity, medium intensity, and high intensity were grouped as developed areas. We utilized a first-order decay process as a simplified approximation for

denitrification. While the numerical study was an attempt to mimic the real system, but in a large domain like ours, some simplifications were deemed necessary, and some factors were excluded. To predict the concentrations of nitrate-N in monitoring wells, one need all historical data of land use and land covers, besides a detailed parameterization of the aquifer. There are various sources of uncertainty. For example, uncertainty arises from unknown spatial variation of the medium parameters, choice of models, boundary conditions imposed on the flow domain, and concentration of contaminant (nitrate-N) released into the flow system etc [*Smith and Schwartz*, 1980, 1981]. The extents of the aquifers (Ogallala and Trinity) are also large, and aquifer properties and parameters are stochastic in nature. Therefore, it is impossible to predict concentrations of nitrate-N deterministically in all monitoring wells. However, we can predict the trends of concentrations of nitrate-N with a band of uncertainty. Because of these limitations, the scope of this model is limited to estimation of the variability of nitrate-N in the aquifer.

4.4 Data

The land use and soil map were created from digital data (Figure 4.5) acquired from the U.S. Department of Agriculture [available at *http://datagateway.nrcs.usda.gov*]. The water quality (e.g., nitrate-N and Cl^-) and precipitation data (Figure 4.7) were obtained from the Texas Water Development Board (available at *http://www.twdb.state.tx.us*). Data used in this study were from 1940-2008. Geologic information of the study area was procured from the USGS [available at *http://waterdata.usgs.gov*].

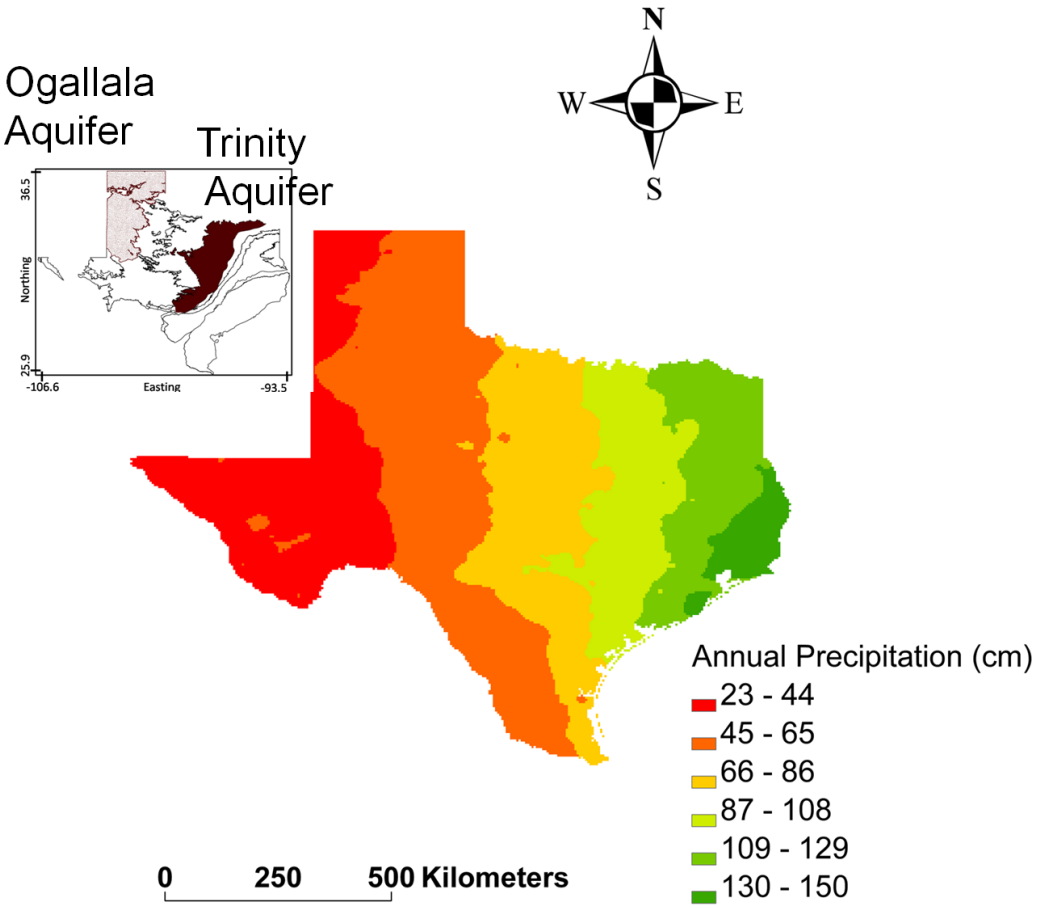

Figure 4.7: Map showing average annual precipitation in Texas. Precipitation varies widely across Texas. In the Trinity Aquifer, precipitation varies from a low of 60 cm per year in the west to a high of 110 cm on the eastern portion of the outcrop of the aquifer, and in the Ogallala Aquifer, it varies from a low of 20 cm per year in the west and southwest to a high of 50 cm on the eastern portion.

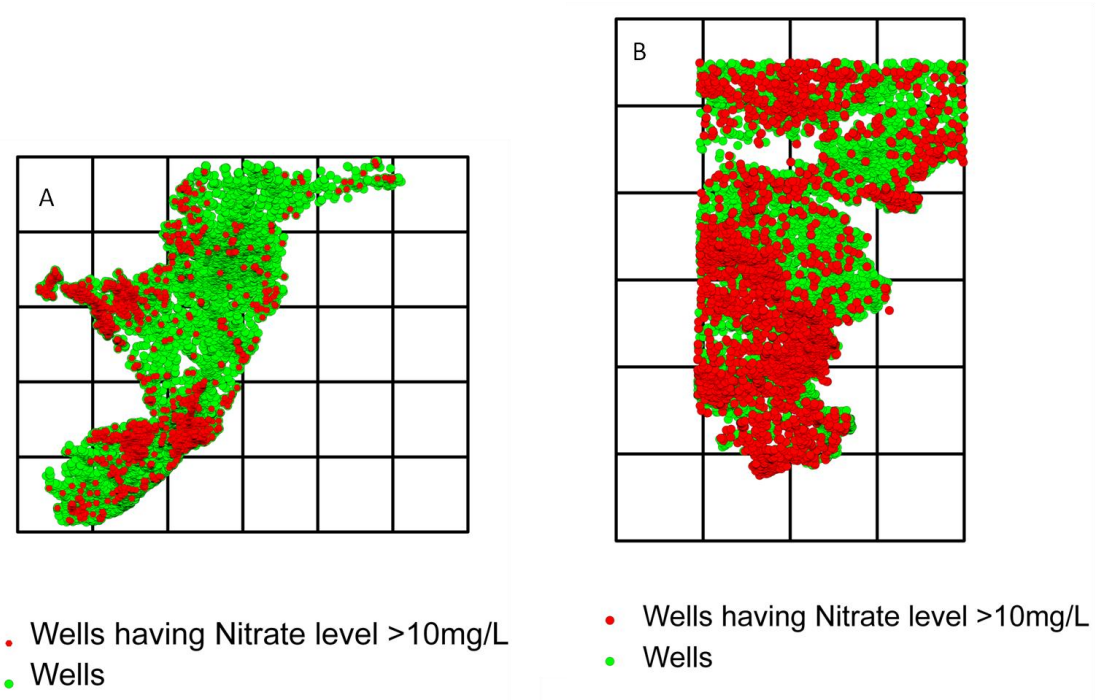

Figure 4.8: The nitrate-N concentrations (1940 to 2008) are shown in the (A) Trinity and (B) Ogallala aquifers. In both the aquifers, three different spatial scales namely fine (2 km×2 km), intermediate (10 km×10 km), and coarse (100 km×100 km) are used to analyze the spatial and temporal variability of nitrate-N. The gridded reconstruction of water quality dataset involved all wells falling in a particular grid (fine, or intermediate, or coarse) as one dataset (or time series). There are 25% and 16% of total wells exceeding 10 mg/L of nitrate-N in the Ogallala and Trinity aquifers, respectively.

The Trinity and Ogallala aquifers were divided into different grid resolutions, as shown in Figure 4.8, fine ($2\ km \times 2\ km$), intermediate ($10\ km \times 10\ km$), and coarse ($100\ km \times 100\ km$) grids. The choice of grids is dependent upon the flow system in the aquifers. An aquifer system can comprise local, intermediate, and regional ground-water flow systems. In a local system of ground-water flow, recharge and discharge areas are adjacent to each other. In an intermediate GW flow system, recharge and discharge areas are separated by one or more topographic highs and lows. In a regional ground-water flow system, recharge areas are along GW divides, and discharge areas lie at the bottom

of major drainage basins [*Toth*, 1963]. Here in the Trinity and Ogallala Aquifers, it was assumed that the local flow system spanning is an area of ($2\ km \times 2\ km$), which is in accordance with the previous studies [*Toth*, 1963; available at *http://pubs.usgs.gov/pp/pp1423-c/pdf/gwflow.pdf*]. The gridded reconstruction of water quality dataset involved all wells within a grid (fine, intermediate, or coarse) as one dataset (or time series). Cl^- along with nitrate-N were chosen for analysis, because Cl^- is a potential indicator of pollution from agricultural practices. The gridded reconstruction was devised because of lack in continuity of data. For example, nitrate-N concentrations in a particular well were available only at five instances from 1940 to 2008, but in some other times, nitrate-N concentrations were available in nearby wells of the well under consideration. Therefore, we combined all wells falling in a particular grid as one time series dataset and analyzed the variability of nitrate-N annually. Each grid represents a time series, and its entropy is calculated as shown in section 4.3.2. For each grid, we have entropy values, and these values are plotted as maps using the surfer software. However, well data are at different depths, but the entropy approach deals with the histogram (distribution) of the data. There are 7866 and 8948 wells in Trinity and Ogallala Aquifers, respectively. In the Ogallala Aquifer, fine grids (FGs), intermediate grids (IGs), and coarse grids (CGs) have on an average 43, 140, and 375 wells. The intermediate and coarse grids have more than 20 wells for 95% of times, whereas fine grids have more than 10 wells for each grid for 90% of times. In the Trinity Aquifer, FGs, IGs, and CGs have on an average 52, 172, and 497 wells. The intermediate and coarse grids have more than 20 wells for 90% of times, whereas fine grids have more

than 10 wells for each grid for 87% of times. As described in section 4.3.2, entropy is a non-parametric approach, and it is a robust measure of variability. Entropy does not change drastically by small changes in data. Therefore, for estimating the variability, we can use the gridded time series.

4.5 Results and discussion

4.5.1 Spatial variability of nitrate-N

Figure 4.9 shows normalized marginal entropy values in the Trinity and Ogallala aquifers at three spatial scales. The Ogallala Aquifer shows comparatively lower NME (or higher variability) than the Trinity Aquifer. The Trinity Aquifer exhibits higher NME at the small scale as compared to intermediate and coarse scales. Similarly, the Ogallala Aquifer also exhibits higher NME at the small scale as compared to intermediate and coarse scales. In both the aquifers, the intermediate scale exhibits the lowest NME. It is evident from Figure 4.9 that NME of nitrate-N is sparse over the small spatial scale in the Trinity and Ogallala aquifers. At the intermediate scale, both the aquifers show variability across different river basins, although the Ogallala Aquifer again displays lower NME (Figures 4.2 and 4.9) than the Trinity Aquifer. in the Ogallala Aquifer, the variability of nitrate-N changes around the Canadian and Colorado rivers, and in the Trinity Aquifer, the variability of nitrate-N is manifested around the Red, Brazos, Colorado, and Guadalupe rivers (Figures 4.1 and 4.2). At the coarse scale, in the Trinity Aquifer, NME exhibit more variability of nitrate-N on the outcrop region (Antlers formation) and in the lower part of the aquifer (Figure 4.9C), close to the northern segment of the Edwards Aquifer. In the Ogallala Aquifer at the coarse scale, NME

values are smaller on the western part of the aquifer, and there are lower NME values of nitrate-N, except in the regions lying between the Canadian and Prairie Dog Town Fork Red River (Figure 4.9F).

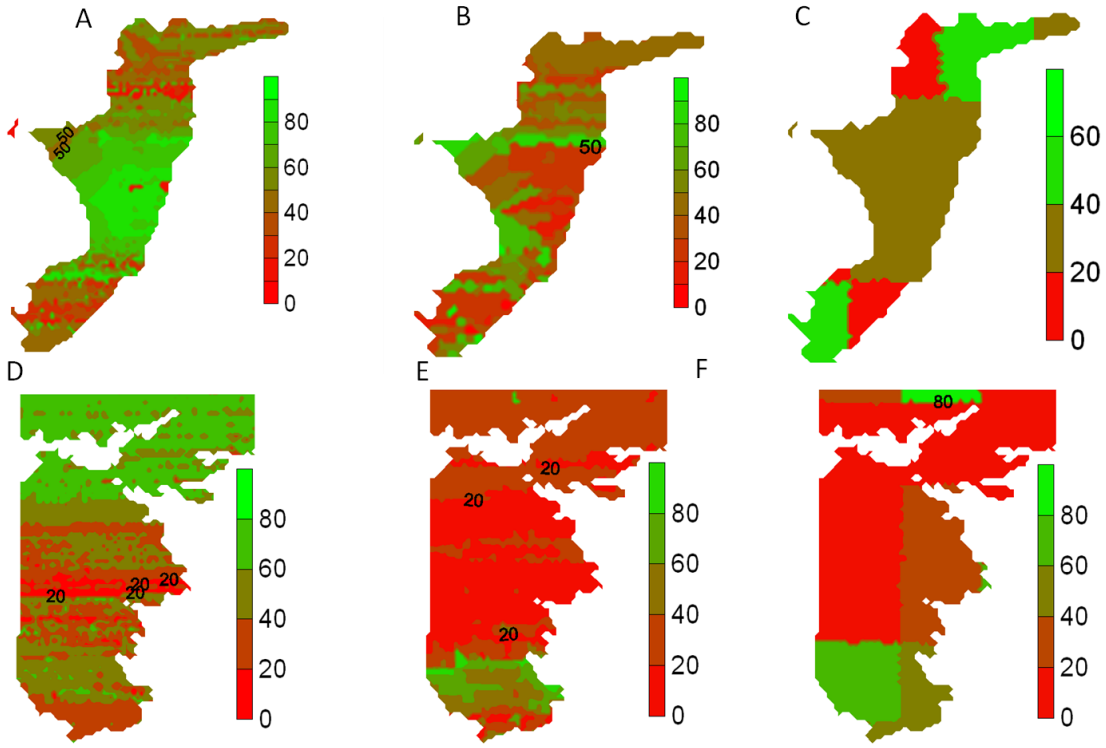

Figure 4.9: Map shows the contour plots of normalized marginal entropy values of observed nitrate-N in the Trinity and Ogallala aquifers at three spatial scales: fine grid (A and D), intermediate grid (B and E), and coarse grid (C and F), respectively. The higher NME values indicate lower entropy values.

The Ogallala Aquifer shows comparatively lower NME than the Trinity Aquifer because of more mixing of nitrate-N at the local scale. The local and intermediate flow systems are more prevalent in the Ogallala Aquifer as most surface drainage in the Ogallala Aquifer is through ephemeral lakes or playas [*Hudak*, 2000]. The flat topography results in mostly internal drainage to thousands of playas. Apart from internal drainage, intense pumping for irrigation and subsequent return flow and higher hydraulic conductivity (sandy aquifer) also contribute towards the higher mixing nitrate-N at the local scale. Pumping and hydraulic conductivity affect NME at the smaller scale as exhibited through the numerical study.

At the intermediate scale, both the aquifers show variability across different river basins (Figures 4.2 and 4.9). Variability of nitrate-N at this scale results from the complex interactions between rivers and aquifers in the study area. Therefore, lower NME values of nitrate-N at the intermediate scale results from the complex interactions between rivers and aquifers in the study area.

At the coarse scale, in the Trinity Aquifer, NME exhibit more variability of nitrate-N on the outcrop region (Antlers formation) and in the lower part of the aquifer (Figure 4.9C), close to the northern segment of the Edwards Aquifer. Lithology and local fracturing and Balcones Fault Zone control the entropy of nitrate-N in the Trinity Aquifer. In the Ogallala Aquifer, there are lower NME values of nitrate-N at the coarse scale, except in the regions lying between the Canadian and Prairie Dog Town Fork Red River and in the northern part (Figure 4.9F). This is because the geology and confining strata of the aquifer determine the degree of nitrate-N levels in GW at the coarse scale.

Figures 4.10 and 4.11 illustrate the effect of pumping and hydraulic conductivity on spatial variability of nitrate-N at three spatial scales using numerical experiments in the Ogallala and Trinity aquifers respectively. Intense pumping and high hydraulic conductivity exhibit lower NME values at the small grid, whereas the NME of nitrate-N has not changed significantly at the intermediate and coarse scales (Figures 4.10 and 4.11).

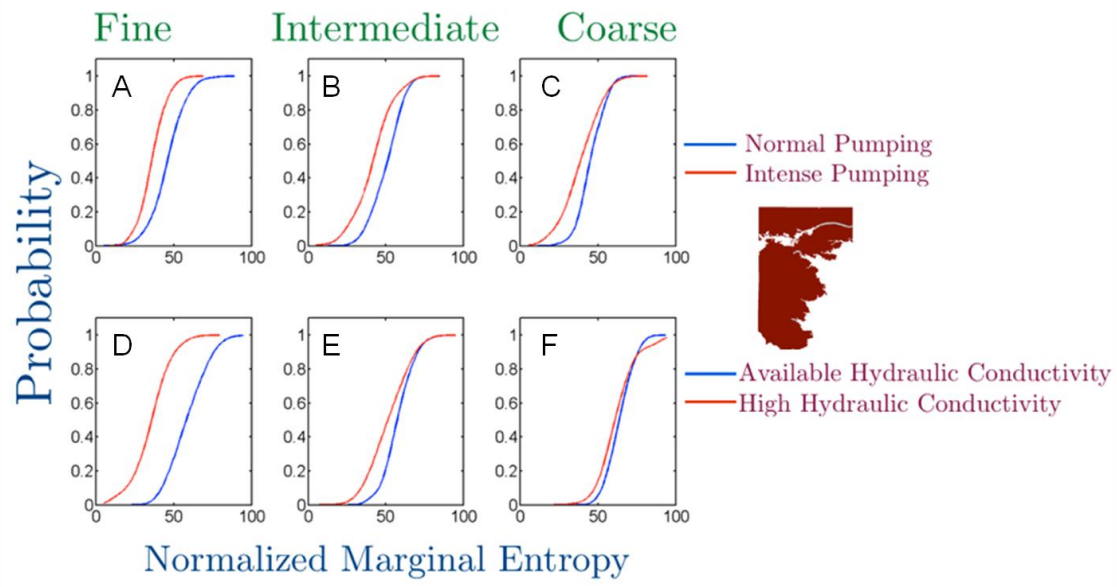

Figure 4.10: A numerical study using Visual MODFLOW in the Ogallala Aquifer was conducted to verify the effect of different factors on spatial variability of nitrate-N at multiple scales. Probability distribution functions of the normalized marginal entropy values (%) of nitrate-N in the Ogallala Aquifer are plotted for pumping and hydraulic conductivity scenarios. The results demonstrate that the spatial variability of nitrate-N is controlled by the effect of pumping at the fine grid (A) as compared to intermediate (B) or coarse (C) grids; hydraulic conductivity also plays a key role at the small grid (D) as compared to intermediate (E) or coarse (F) grids. The higher NME values indicate lower entropy values.

MT3DMS v5.1 was used for numerical modeling with perturbations in factors such as pumping, hydraulic conductivity, presence and absence of a river, and effects of layering. MODFLOW—2000 (Visual MODFLOW) was run prior to MT3DMS.

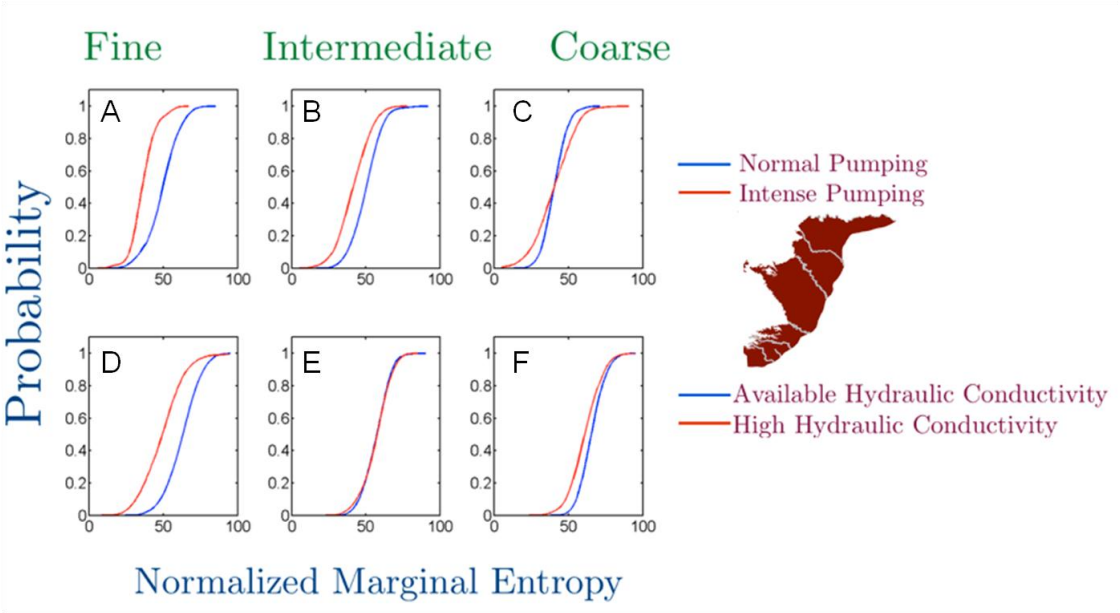

Figure 4.11: A numerical study using Visual MODFLOW in the Trinity Aquifer was conducted to verify the effect of different factors on spatial variability of nitrate-N at multiple scales. Probability distribution functions of the normalized marginal entropy values (%) of nitrate-N in the Trinity Aquifer are plotted for pumping and hydraulic conductivity scenarios. The results demonstrate that the spatial variability of nitrate-N is controlled by the effect of pumping at the fine grid (A) as compared to intermediate (B) or coarse (C) grids; hydraulic conductivity also plays a key role at the small grid (D) as compared to intermediate (E) or coarse (F) grids. The higher NME values indicate lower entropy values.

Cumulative Probability distribution functions (CDFs) of NME values in the wells of the Ogallala and Trinity aquifers were compared for perturbed factors at the three spatial scales.The center of mass (mid- point of the CDF) of a CDF shows where NME values are positioned and spread around. Therefore, if the center of mass of a CDF is located at 40 as compared to another CDF whose center of mass is located at 60, the CDF with NME value of 40 will have more variability (see section 4.3.2). The CDFs also show the percentage of NME values greater than a particular number.

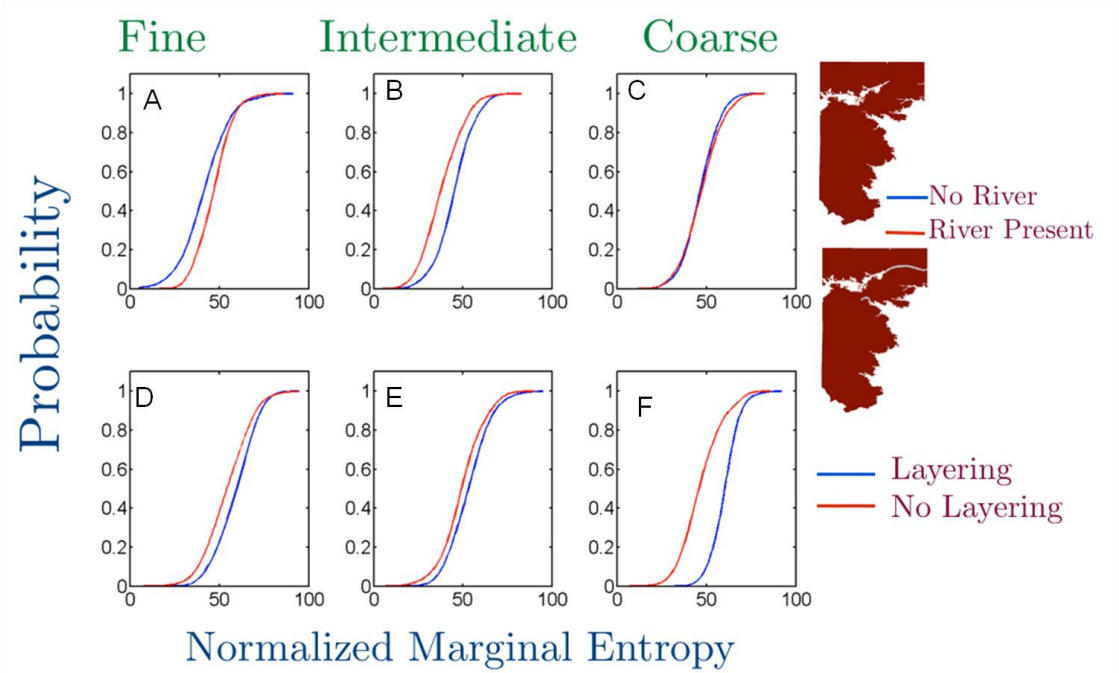

Figure 4.12: A numerical study using Visual MODFLOW was conducted in the Ogallala Aquifer to verify the effect of different factors on spatial variability of nitrate-N at multiple scales. Probability distribution functions of the normalized marginal entropy values (%) of nitrate-N in the Ogallala Aquifer are plotted for river and layering scenarios. The results demonstrate that the spatial variability of nitrate-N is controlled by the presence of a river at the intermediate grid (B) as compared to fine (A) or coarse (C) grids; geology (layering vs. no layering) plays a key role at the coarse grid (F) as compared to fine (D) or intermediate (E) grids. The higher NME values indicate lower entropy values.

Figures 4.12 (A, B, and C) and 4.13 (A, B, and C) show the CDFs of NME values calculated by numerical simulations mimicking the effect of rivers on the spatial variability of nitrate-N at three spatial scales in the Ogallala and Trinity aquifers, respectively.

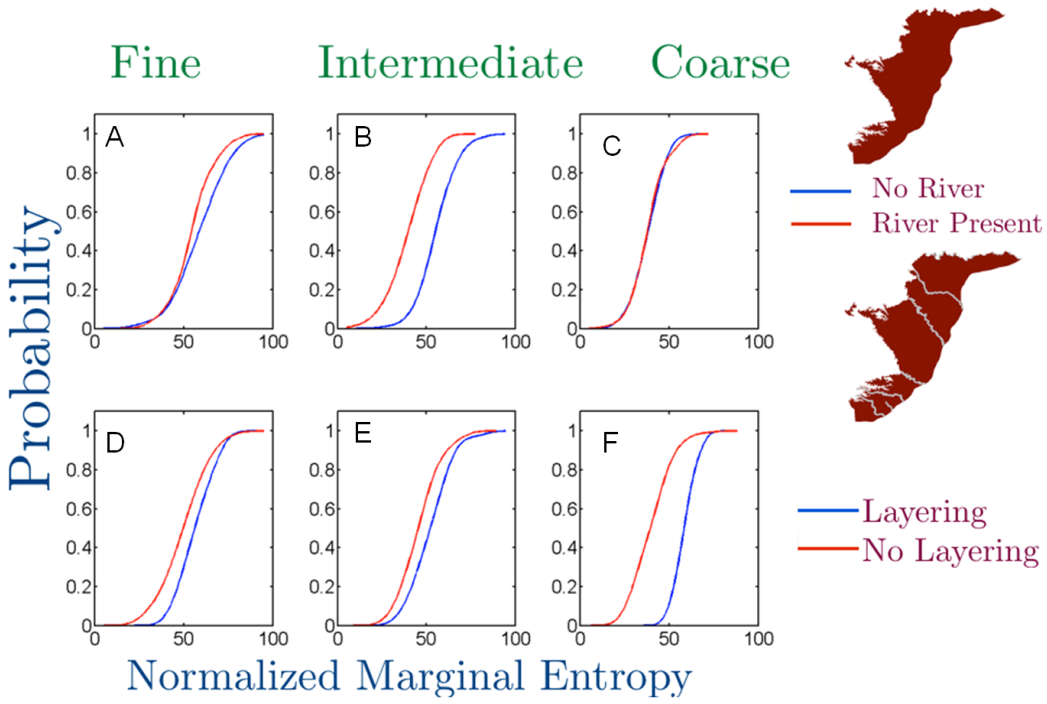

Figure 4.13: A numerical study using Visual MODFLOW was conducted in the Trinity Aquifer to verify the effect of different factors on spatial variability of nitrate-N at multiple scales. Probability distribution functions of the normalized marginal entropy values (%) of nitrate-N in the Trinity Aquifer are plotted for river and layering scenarios. The results demonstrate that the spatial variability of nitrate-N is controlled by the presence of a river at the intermediate grid (B) as compared to fine (A) or coarse (C) grids; geology (layering vs. no layering) plays a key role at the coarse grid (F) as compared to fine (D) or intermediate (F) grids. The higher NME values indicate lower entropy values.

Presence of rivers exhibits lower NME values at the intermediate scale, whereas the NME of nitrate-N has not changed significantly between fine and coarse scales. At the large scale, geology becomes important for the spread of nitrate in the aquifer. Point sources of nitrate-N can contribute to GW nitrate at large scale through the unsaturated zone. If there is a confining stratum, then nitrate-N will spread only through mixing in GW, which is slow at the large scale. The numerical study corroborates the dominance of geological framework and layering at the coarse scale for the Ogallala (Figure 4.12 D, E, and F) and Trinity aquifers (Figure 4.13 D, E, and F).

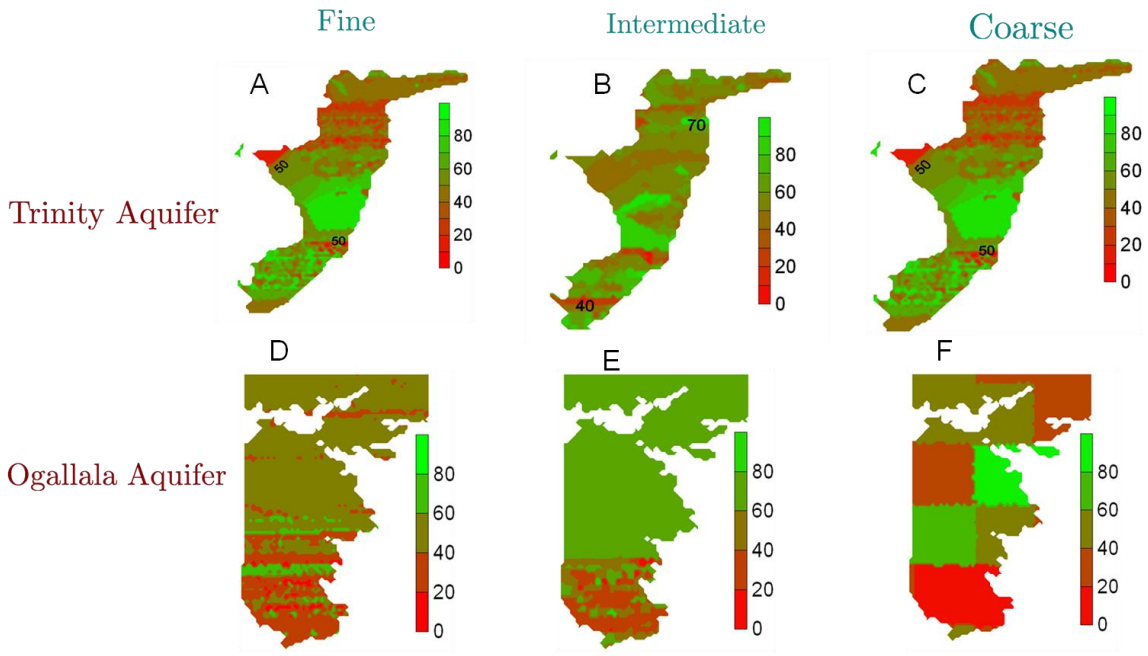

Figure 4.14: Map shows the contour plots of entropy values of Cl^- in the Trinity and Ogallala aquifers at three spatial scales: fine grid, intermediate grid, and coarse grid, respectively.

4.5.2 *Spatial variability of Cl^- and mutual information sharing between nitrate-N and Cl^-*

Figure 4.14 outlines the spatial variability of Cl^- (Trinity and Ogallala aquifers), which is in accordance with the spatial variability of nitrate-N at the fine scale but not over the intermediate and coarse scales. Cl^- is naturally present in GW, particularly in deep bedrock aquifers, and it may also originate from septic tank leachate or fertilizers. [*http://walrus.wr.usgs.gov/infobank/programs/html/factsheets/pdfs/2004_3120.pdf*].

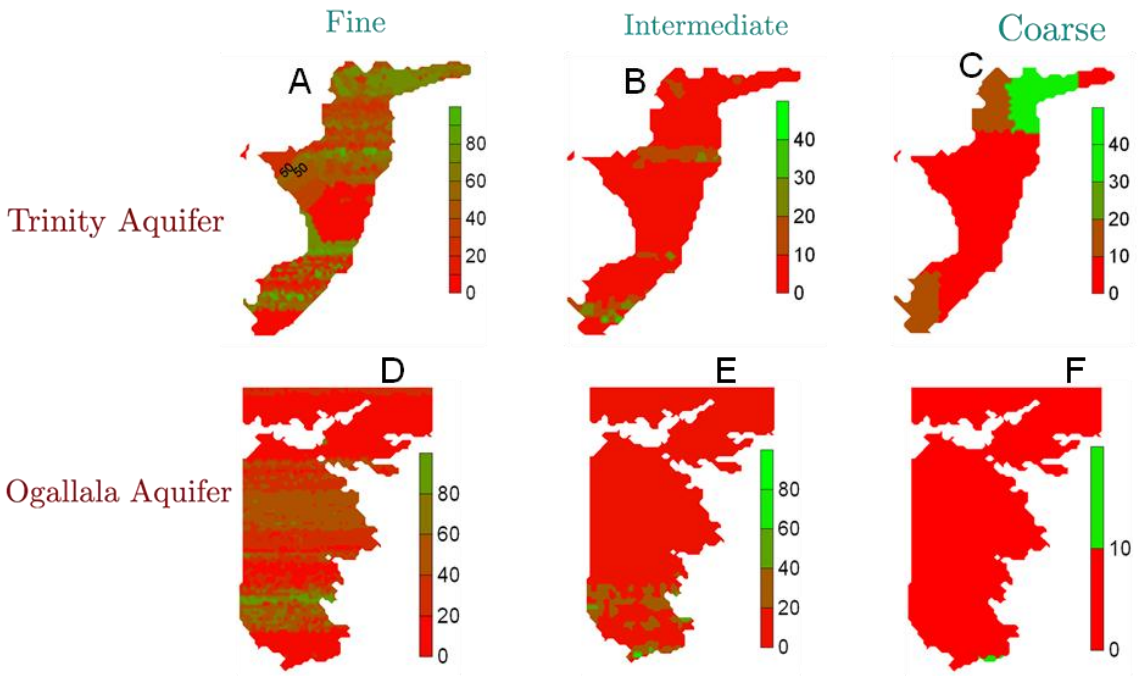

Figure 4.15: Map presents the Normalized Mutual Information between nitrate-N and Cl^- in the Trinity and Ogallala aquifers at three spatial scales: fine (A, D), intermediate (B, E), and coarse (C, F) grids, respectively. The green color indicates the codependency of the two variables (nitrate-N and Cl^-).

If Cl^- has any anthropogenic sources, it$^-$ is likely to manifest spatial variability at all scales. In section 4.5.1, we have shown that pumping and hydraulic conductivity are the key controls for variability of nitrate-N at the local scale.

Mutual information (NMI) between nitrate-N and Cl^- are presented in Figure 4.15. It is worth mentioning that the mutual variability (covariance of nitrate-N and Cl^-) is different from the mutual information of nitrate-N and Cl^-. The difference is that mutual variability does not state anything about the common source or origin of nitrate-N and Cl^- whereas the mutual information indicates the codependency of the two variables. As shown in Figure 4.15, the flow of information decreases from the small scale (NMI >30) to the coarse scale (NMI<20) in both the aquifers. In other words, nitrate-N and Cl^- have a better codependency at the small scale. The mutual information in Figure 4.15 exhibits a rich pattern of information flow in the Trinity Aquifer where hot spots of nitrate contamination are present such that regions with NMI>30 of Figure 4.15 match to hot spot regions (wells having nitrate-N>10 mg/L) of Figure 4.8. On the outcrop regions of the Trinity Aquifer, nitrate-N and Cl^- have a maximum of relative mutual information of 50%, 15%, and 5% at the fine, intermediate, and coarse scales, respectively. This suggests that approximately 50%, 15%, and 5% of the total Shannon entropy of one variable (say nitrate-N) is explained by the other variable (Cl^-). In the Trinity Aquifer at the coarse scale, there are high NMI values in the northern and southern parts. In the Ogallala Aquifer, nitrate-N and Cl^- have high relative mutual information in the western part of the aquifer equaling to 40%, 20%, and 5% at the fine, intermediate, and coarse scales, respectively. These results suggest that nitrate-N and Cl^-

may be partly originating from leachate and that soil type dominates the shared information at the small scale. Other than soil type, irrigation with GW is a crucial factor for the flow of information between nitrate-N and Cl^-. Unlike the Ogallala Aquifer, the Trinity Aquifer has less intense GW-based irrigation, and there are limited recharge areas (aquifer outcrop areas (Figure 4.1).

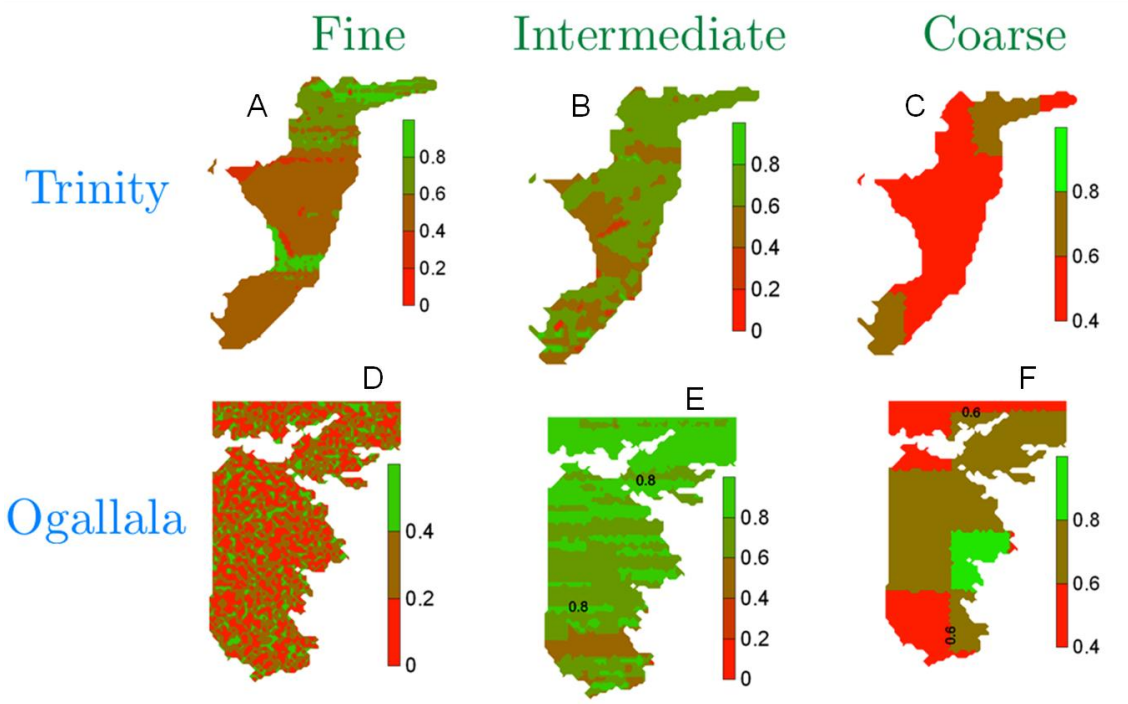

Figure 4.16: The Hurst exponent (H) of nitrate-N in the Trinity Aquifer at the small (A), intermediate (B), and coarse (C) scales, as well as in the Ogallala Aquifer at the small (D), intermediate (E), and coarse (F) scales. The Hurst exponent varies from 0 to 1, 1 being the most persistent, 0.5 being the random changes in the trend, and 0 being the most anti-persistent (or mean reversion) behavior.

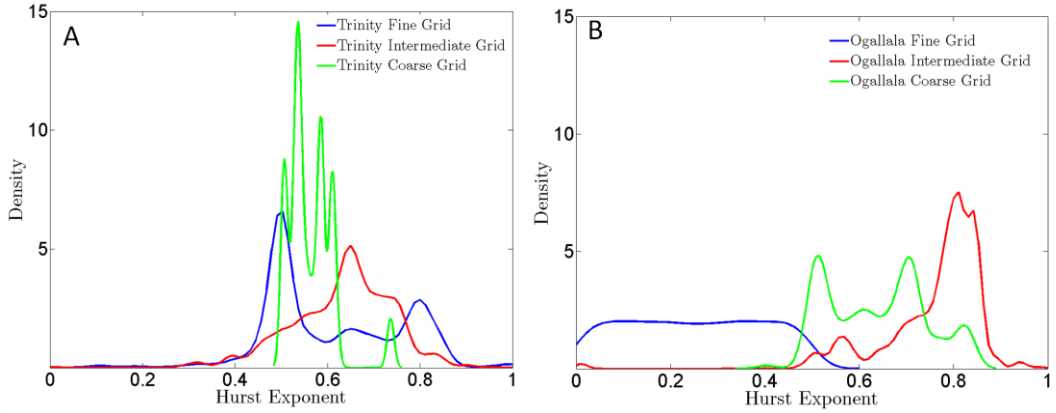

Figure 4.17: Probability distribution (PDF) of the Hurst exponent of nitrate-N in the (A) Trinity and (B) Ogallala aquifers across different spatial scales (fine, intermediate, and coarse).

4.5.3 Trend and persistence of variability of nitrate-N

The Hurst exponents (**H**) of nitrate-N in both the aquifers at the fine, intermediate, and coarse scales are presented in Figure 4.16. Although there are small/fine scale variations of **H** in both the aquifers, trends are more persistent for both aquifers at the intermediate scale. In section 4.5.1, we have shown that there is maximum variability of nitrate-N at the intermediate scale. The Hurst exponent signifies that the variability of nitrate-N continues to exist at the intermediate scale.

There are regions where more variation (anti-persistence) is observed in both the aquifers at the coarse scale. Figure 4.17 shows the probability distribution (PDF) of **H** at the fine, intermediate, and coarse scales in the Trinity and Ogallala aquifers. The PDF illustrates that more than 50% grids at the fine scale, 70% at the intermediate scale, and 10% at the coarse scale show persistence in the Trinity Aquifer. On the other hand, the PDF shows that more than 80% grids at the small scale, 90% at the intermediate scale,

and 70% at the coarse scale show persistence in the Ogallala Aquifer. The constantly changing behavior of nitrate-N at the small scale in the Ogallala Aquifer, as shown in Figure 4.16 (in contrast to the Trinity Aquifer), may be explained by the presence of prevalent local flow systems. A local flow system is typically present in sandy and/ or unconfined aquifers, and such aquifers respond quickly to increased GW recharge. At the coarse scale in both the aquifers, PDFs show multimodality. In the Trinity Aquifer, these modes are located around 0.5, which means rapidly changing behavior of the variability of nitrate-N. However, in the Ogallala Aquifer, these modes are located around 0.5, 0.6, and 0.7, which means there are regions signifying the persistence of nitrate-N. The Ogallala Aquifer is intensive agricultural land, so the use of fertilizer may be a reason for the persistence.

4.5.4 Normalized risk index of nitrate-N

The Normalized Risk Index (NRI) measures the distribution of hot spots (nitrate-N > 10 mg/L). The higher the NRI, the lesser are the number of hot spots in the aquifer at a particular spatial scale. In Table 4.3, the NRI values are given for both the aquifers at different scales. Results indicate that number of hot spots of nitrate contamination in GW decrease with the increase in the spatial scale (from fine to coarse grids). This finding may suggest that there are concentrated hot spots in both the aquifers where nitrate-N contamination is localized.

Table 4.3: Normalized risk index (%) in the Trinity and Ogallala aquifers. The NRI values show that the likelihood of hot spots is higher for larger NRI.

Scale	Trinity Aquifer	Ogallala Aquifer
Fine	32.2	24.7
Intermediate	35.8	38.7
Coarse	60.9	59.1

4.5.5 Temporal variability of nitrate-N

For understanding the spatial variability of nitrate-N over different decades, inter-decadal variation of nitrate-N was analyzed (Figures 4.18 and 4.19). Decadal mean (μ), standard deviation (SD), percent samples having nitrate-N > 10 mg/L, and normalized marginal entropy (NME) were plotted for different scales (fine, intermediate, and coarse). In the Trinity Aquifer, overall decadal mean nitrate-N has declined at all scales. Standard deviation has also decreased with time implying lesser uncertainty (Figure 4.18). Furthermore, NME also suggests a slight decline in the temporal variability of nitrate-N at the coarse scale (Figure 4.19). However, percent samples having nitrate-N > 10 mg/L are almost constant over the last 7 decades (1940-2010). Based on these findings, it seems that nitrate-N levels in GW have stabilized in the Trinity Aquifer.

In the Ogallala Aquifer, overall decadal mean nitrate-N has increased at all scales from 1940 to 1970, which may be because of higher use of fertilizers to improve agricultural productivity since 1940s. After 1970, overall mean nitrate-N has decreased at fine and coarse scales (Figure 4.18). Standard deviation has increased in the Ogallala Aquifer at all scales (Figure 4.18). However, percent samples having nitrate-N > 10

mg/L decreased significantly over the last 7 decades. This decrease is suggested to be because of more efficient water management methods employed in irrigation practices of the region, particularly after the early 1980s. Hence, the irrigation return flow, which contributes significant amounts of recharge to the Ogallala Aquifer, has declined through time.

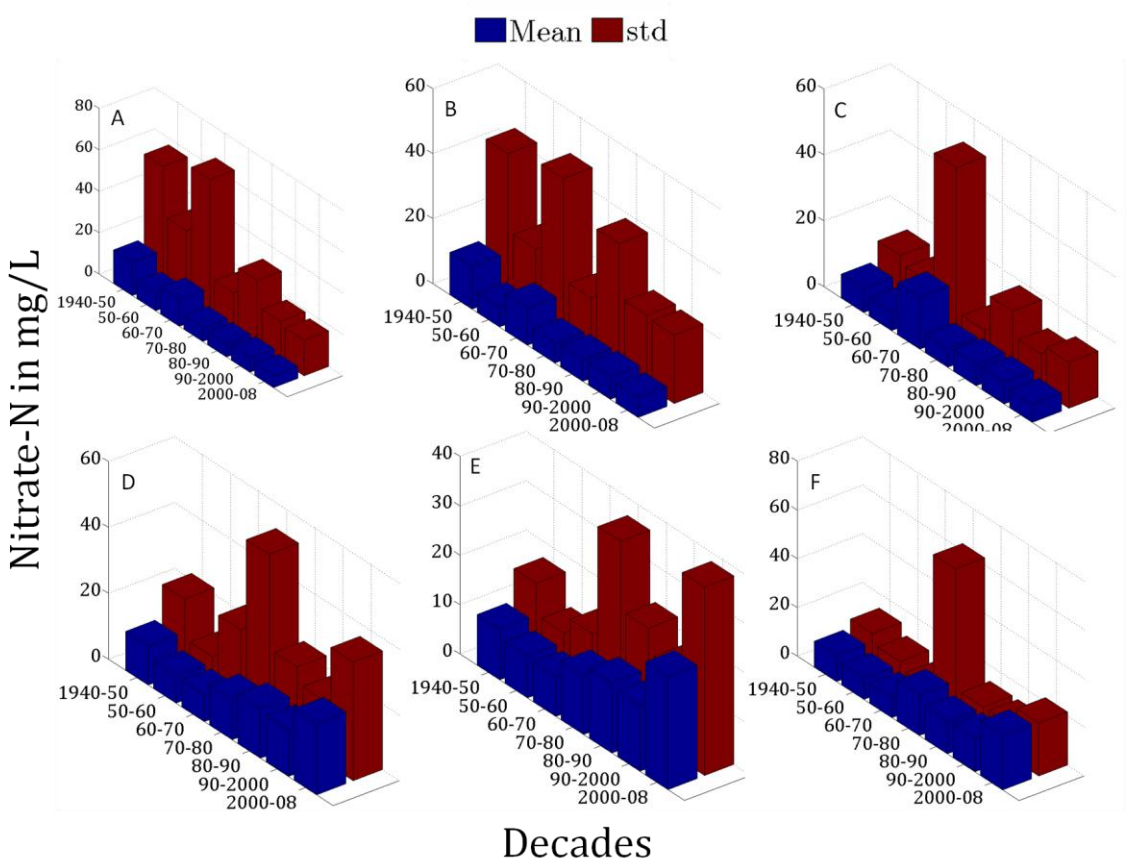

Figure 4.18: Decadal analysis (mean and standard deviation) of nitrate-N in the Trinity Aquifer across (A) fine, (B) intermediate, (C) coarse scales, and in the Ogallala Aquifer across (D) fine, (E) intermediate, and (F) coarse scales.

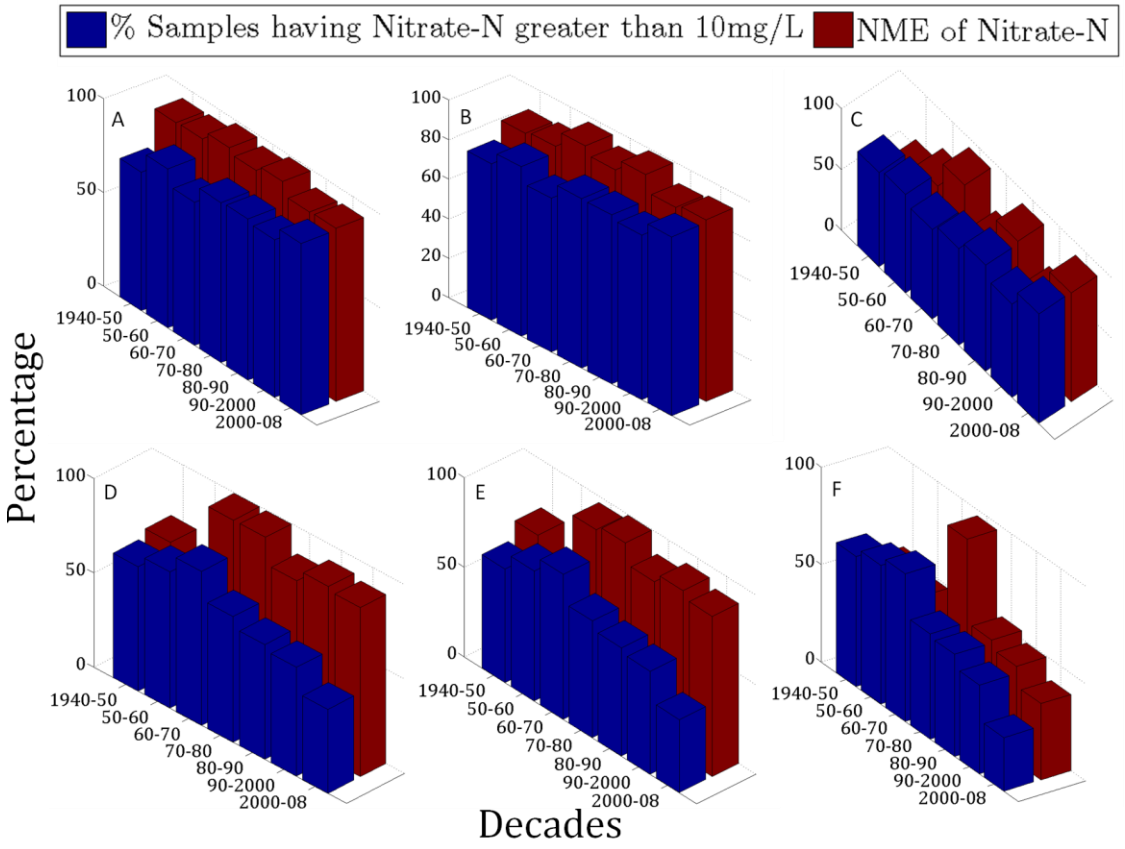

Figure 4.19: Decadal analysis (% samples having nitrate-N more than 10 mg/ L and NME) of nitrate-N in the Trinity Aquifer across (A) fine, (B) intermediate, (C) coarse scales, and in the Ogallala Aquifer across (D) fine, (E) intermediate, and (F) coarse scales.

4.6 Summary and conclusions

The entropy based approach provides a physical interpretation of spatial and temporal variability of nitrate-N in the two Texas aquifers. The primary conclusion of this study is that multiple controlling factors exist and dominate the variability of nitrate-N at different spatial scales (fine, intermediate, and coarse) in the Trinity and Ogallala aquifers of Texas. The variability of nitrate-N is controlled by pumping, and hydraulic conductivity at the fine scale. The intermediate scale variability of nitrate-N results from the complex interactions between rivers and the aquifer. The coarse scale variability of nitrate-N is caused by lithology and geology control. The maximum variability of nitrate-N occurs at the intermediate scale. The number of hot spots of nitrate-N contamination in GW decreases with increase in the spatial scale (from fine to coarse grids).

The sharing of information between (nitrate-N and Cl^-) is an effective way to identify process interactions in complex subsurface systems. At the fine scale, nitrate-N and Cl^- have similar sources (most likely fertilizers). Soil type and GW-based irrigation are important factors, which regulate the transport of nitrate-N and Cl^- to GW.

The trends of nitrate-N variability show long term persistence at the intermediate scale. In the Ogallala Aquifer, there is more persistence of nitrate-N, which is attributed to more fertilizer use.

In the Trinity Aquifer, overall mean nitrate-N (at the fine scale 50 mg/L to 6 mg/L; at the intermediate scale 40 mg/L to 15 mg/L; at the coarse scale 15 mg/L to 10 mg/L) has declined with not much change in the entropy (at the fine scale 88 to 90; at

the intermediate scale 88 to 90; at the coarse scale 55 to 89) over the decades between 1940 and 2008. Moreover, percent samples having nitrate-N > 10 mg/L did not change much across time (at the fine scale 60% to 70%; at the intermediate scale 70% to 75%; at the coarse scale 60% to 65%). It can be inferred from these results that overall consumption of fertilizers (or nitrate-N emanating from other sources as well) has gone down, but there are hot spots in the Trinity Aquifer. These hot spots are also responsible for the slight increase in the entropy values. In the Ogallala Aquifer, at all scales, overall mean nitrate-N has increased till 1970 because of enhanced use of fertilizer since 1940 (at the fine scale 20 mg/L to 50 mg/L; at the intermediate scale 15 mg/L to 35 mg/L; at the coarse scale 15 mg/L to 60 mg/L). However, percent samples having nitrate-N > 10 mg/L has significantly decreased (at the fine scale 60% to 32%; at the intermediate scale 60% to 38%; at the coarse scale 60% to 27%) over the last seven decades due to the use of more efficient methods of irrigation.

CHAPTER V

ADDRESSING UNCERTAINTY IN NITRATE TRANSPORT IN GROUNDWATER USING THE ENSEMBLE KALMAN FILTER

5.1 Synopsis

A fundamental problem in the analysis of nitrate-N contamination of complex GW systems is the fusion of data and modeling. We present the fusion of the ensemble Kalman filter (EnKF) with the GW flow model MODFLOW and the solute transport model MT3DMS. The EnKF is a sequential data assimilation approach, which is applied to quantify and reduce the uncertainty of flow and nitrate transport in GW. We conducted numerical simulation experiments from Jan 1996 to Dec 2005 with MODFLOW and MT3DMS models for variably saturated GW flow in a synthetic 2D Aquifer, and then tested the EnKF algorithm in the Ogallala Aquifer. The EnKF was used to update model parameters such as hydraulic conductivity, aquifer recharge, and first order decay coefficient of nitrate-N. Once these parameters are optimized, then updated parameters are used to predict hydraulic heads and nitrate concentration in aquifers. Results indicate that the EnKF method notably improves the estimation of nitrate-N concentrations as compared to hydraulic heads, and therefore, uncertainty is not propagated from MODFLOW to MT3DMS model. Hydraulic conductivity is found to be the most important parameter in improving estimations of hydraulic heads and nitrate-N concentrations using the EnKF, followed by recharge, and then by the decay coefficient. Results suggest that nitrate-N concentrations are more sensitive to

intermediate recharge values as compared to high or low recharge events. Furthermore, the optimization of parameters using the EnKF approach demonstrated the asymptotic behavior of hydraulic conductivity and first order decay coefficient.

5.2 Introduction

Nitrate is one of the most widespread contaminants in groundwater (GW) and known to pose a serious health risk. Despite its critical importance, fate and transport of nitrate-N in GW shows significant uncertainty as a result of poor spatial sampling and heterogeneity. Other factors contributing towards uncertainty are interaction among multiple geophysical factors such as source availability (land use), thickness and composition of the vadose zone [*Williams et al.*, 1998], types of aquifers (confined or unconfined), aquifer heterogeneity (geological and alluvial) [*Spalding* and *Exner*, 1993], and precipitation characteristics [*Koren et al.*, 1999]. Many uncertainty quantification frameworks exist in the environmental modeling literature [for example, *Carrera* and *Neuman*, 1986 a; b; c; *Beven* and *Binley*, 1992; *Woodbury* and *Ulrych*, 2000; *Thiemann et al.*, 2001; *Vrugt et al.*, 2005; *Vrugt* and *Robinson*, 2007].

A fundamental problem in the analysis of nitrate-N contamination of complex GW systems is the fusion of data and modeling. Data on hydrologic properties controlling flow and transport are needed to address contamination in aquifer systems. Accurate predictions of fluid flow and contaminant transport are dependent on the values assigned to input parameters and the environmental forcing required for the simulation, which are often not known with certainty. In order to reduce uncertainty due to system parameters (e.g., hydraulic conductivity, reaction rate constants, etc.), numerous

parameter estimation methodologies have been proposed in the hydrologic literature [*Ahmed* and *Demarsily*, 1987; *Dagan*, 1979; *Dagan* and *Lessoff*, 2007; *Indelman* and *Dagan*, 1993; *Indelman et al.*, 1998; *Neuman*, 1980; 2003; *Neuman* and *Yakowitz*, 1979; *Neuman et al.*, 1980; *Rubin et al.*, 1990].

In the last decade, ensemble-based forecasting methods established on data assimilation approaches have become increasingly popular for quantification and reduction of state uncertainty. Methods of probabilistic prediction and data assimilation (DA) have been extensively explored in the atmospheric and ocean sciences [*Anderson*, 2007b; *Daley*, 1992; 1997]. Data assimilation and uncertainty analysis have also been employed for flux estimation [*Ng et al.*, 2009], identification of diffuse recharge mechanisms [*Ng et al.*, 2009]), surface energy balance assessment [*Sun et al.*, 2011], and other hydrological applications (e.g., soil moisture [*Flores et al.*, 2010; *Margulis et al.*, 2002; *Sun et al.*, 2011]). In particular, the ensemble Kalman filter (EnKF) has been successfully employed in the estimation of hydrologic system parameters [*Aksoy et al.*, 2006a; b; *Bailey* and *Bau*, 2010; *Ebtehaj et al.*, 2010]. DA schemes have also been used in contaminant transport studies to estimate the distribution of solute concentration in aquifer systems [*Chang* and *Latif*, 2010; *Zou* and *Parr*, 1995], and to design adaptive strategies for sampling in space and time [*Kollat et al.*, 2011]. DA approaches have been employed to update solute transport parameters such as dispersivity values [*G S Liu et al.*, 2008; *Y Q Liu* and *Gupta*, 2007], sorption rates [*Vugrin et al.*, 2007], and first order chemical reaction rates [*Bailey and Bau*, 2011]. None, however, have addressed the temporal evolution of parameters (e.g., hydraulic conductivity, first order chemical

reaction rate) by assimilation of flow and concentration data. DA methods provide the flexibility to investigate possible temporal evolution of the model parameters. These adaptive estimation methods become more appropriate when the forecast lead-time is short in comparison to the response time of the aquifer system [*Kitanidis* and *Bras*, 1980; *Moradkhani et al.*, 2005], which is a typical scenario in the real world problems. Therefore, this study investigates the temporal progression of parameters like hydraulic conductivity, first order chemical reaction rate, and recharge in aquifer systems.

The rationale behind selecting these parameters is that these three parameters are of critical importance for studying the fate and transport of nitrate-N in aquifers. For example, *Tebes-Stevens et al.* [2001] found that hydraulic conductivity is the most sensitive parameter for reactive transport models. In addition, previous studies have established that denitrification is an important mechanism for nitrate-N removal from aquifers containing reactive reducing substrates such as organic carbon [*Bohlke et al.*, 2002; *Kolle et al.*, 1983; *Komor and Anderson*, 1993; *Mariotti et al.*, 1988; *Vogel et al.*, 1981]. It has been observed that microbially mediated denitrification is a dominant process and can be represented as a first-order decay reaction [*Frind et al.*, 1990]. Therefore, first order chemical decay rate is an important factor in studying nitrate-N transport in GW. Several studies have also reported that recharge conditions become crucial for nitrate-N transport especially at large spatial scales (e.g., aquifer scale) [*Strebel et al.*, 1989; *Bohlke*, 2002].

In DA approaches, the parameters are estimated by establishing an initial (prior) state of the aquifer system by generating an ensemble of parameter fields and

subsequently, establishing flow and concentration fields using GW models (MODFLOW and MT3DMS). Then, a posterior ensemble of parameter fields are determined that best match with the measurement from the actual aquifer system. In this study, the integration of the ensemble Kalman filter (EnKF) with the numerical GW flow model MODFLOW [*McDonald and Harbaugh*, 2003] and the solute transport model MT3DMS [*Wang and Zheng*, 2000] to condition aquifer parameters is develpoed. The observed data (heads and nitrate-N concentration) are sparse in the Ogallala Aquifer; therefore, a synthetic aquifer is constructed to test the accuracy of the EnKF algorithm. Subsequently, the EnKF algorithm is applied with the available data in the Ogallala Aquifer. The specific objectives of this study are to improve forward groundwater modeling capabilities through parameter estimation (hydraulic conductivity, reaction rate constant, and recharge) and investigate the temporal evolution of these flow and transport parameters in the EnKF framework.

5.3 Methodology

In order to reduce the uncertainty associated with nitrate-N processes in GW, a data assimilation framework is applied to a synthetic 2D aquifer and the Ogallala Aquifer system. An Ensemble Kalman Filter (EnKF) approach is used to sequentially update relevant flow and transport parameters, such as hydraulic conductivity, recharge, and decay rate coefficient, by assimilating the hydraulic head and nitrate-N concentration data. Subsequently, updated parameters (once they are optimized with reference to the system response) are used to predict hydraulic heads and nitrate-N concentrations in the synthetic and Ogallala Aquifers. In this study, we numerically

solve the flow equation with MODFLOW-2000 and the solute transport equation with MT3DMS, and integrate these modeling results with measured data.

5.3.1 Data assimilation theory and parameter estimation

Data assimilation is an operation whereby measurement data are integrated into model simulation results to provide an updated estimate of the system state. The improvement in the corrected system state depends upon the uncertainty associated with both the simulation results and the measurement data. The data assimilation algorithm used in this study is the Ensemble Kalman Filter (EnKF) [*Evensen*, 1994], which is an extension of the Kalman Filter method [*Kalman*, 1960]. The Ensemble Kalman Filter (EnKF) is a Monte Carlo implementation in the Bayesian framework. In this framework, the probability density function (pdf) of the state of the modeled system (the prior) and the data (likelihood) are combined using the Bayes theorem to obtain the posterior pdf. This Bayesian update is combined with advancing the model in time, and incorporating new data as they become available.

Data assimilation is generally applied to update the state vector in real-time applications. However, in GW hydrology, an important part of the estimation error is associated with incorrect parameter values as a result of the variability of the GW system in both space and time domains. Various studies have used the EnKF approach to jointly estimate states and parameters [*Franssen and Kinzelbach*, 2008; *Naevdal et al.*, 2005]. However, *Wen and Chen* [2007] showed that the joint updating of states and parameters can introduce inconsistencies in the analysis, especially in a heterogeneous system. There are different methods to address this issue such as the rerun option [*Wen and*

Chen, 2007], a dual approach [*Moradkhani et al.*, 2005], etc. We used the dual approach, in which, the Ensemble Kalman filter is first applied to update the parameters. Then, the updated parameters are again applied to the Ensemble Kalman Filter to obtain the final update of the states. Another problem associated with the EnKF approach is filter inbreeding [*Franssen and Kinzelbach*, 2008]. Filter inbreeding problems occur due to a deteriorating quality of the analyzed covariance matrix during the real-time assimilation of observations in the EnKF. In such cases, the ensemble variance is increasingly underestimated over time. Adaptive covariance inflation [*Anderson*, 2007a; *X G Wang et al.*, 2007], dual filter [*Houtekamer and Mitchell*, 1998], and analysis of prediction errors [*Moradkhani et al.*, 2005] are a few methods to deal with the filter inbreeding problem. Another method used to address this problem is the variance inflation approach in which both the ensemble and the covariance matrix are multiplied with a certain constant factor (fixed covariance inflation), which is obtained heuristically. For simplicity, we used this fixed variance inflation approach and the covariance inflation factor was chosen heuristically to give a filtering solution that does not diverge from the observations while keeping the prior covariances small.

5.3.2 The Kalman filter

A brief description of the Kalman filter methodology and its Bayesian framework is provided here, and further details can be obtained elsewhere [*Kalman*, 1960]. The Kalman filter methodology follows the sequential forecast-update (predictor-corrector) cycle. The Kalman Filter assumes that all pdfs are Gaussian:

$$p(x) \propto e^{\left(-\frac{1}{2}(x-\mu)^T Q^{-1}(x-\mu)\right)} \tag{5.1}$$

where **X** denotes the n-dimensional state vector of the model, and it is assumed that it has Gaussian probability distribution with mean **μ** and covariance **Q**, and p(x) is called the prior.

It is assumed that the error distribution of the data is known as well. The data (d) is assumed to have a Gaussian pdf with covariance **R** and mean $\boldsymbol{H_{mean}}$. **H** is the observation matrix. Therefore, the likelihood of the data p($d|x$) is calculated as:

$$p(d|x) \propto e^{(-\frac{1}{2}(d-H_{mean})^T R^{-1}(d-H_{mean}))} \tag{5.2}$$

where RE measures the density of hot spots in an aquifer.

The prior of the state and the likelihood are combined to yield the posterior pdf of the state **X** conditioned on the value of d. As new data become available, the posterior is calculated using the Bayesian framework as:

$$p(x|d) \propto p(x|d)p(x) \tag{5.3}$$

The posterior state is denoted as $\hat{x}$ and is given by:

$$p(\hat{x}) = e^{(-\frac{1}{2}(\hat{x}-\hat{\mu})^T \hat{Q}^{-1}(\hat{x}-\hat{\mu})^T} \tag{5.4}$$

The posterior mean $\hat{\mu}$ and covariance $\hat{Q}$ are calculated as follows:

$$\hat{\mu} = \mu + K(d - H\mu) \tag{5.5}$$

$$\hat{Q} = (1 - KH)Q \tag{5.6}$$

where **K** is the Kalman gain and is given by:

$$K = QH^T(HQH^T + R)^{-1} \tag{5.7}$$

However, maintaining the covariance matrix Q is not feasible computationally for high-dimensional systems [*Evensen*, 1994]. For this reason, EnKFs were developed.

5.3.3 The ensemble Kalman filter (EnKF)

The EnKF is an extension of the Kalman filter method by using Monte Carlo approximation of the Kalman filter for application to high-dimensional systems. The EnKF uses an ensemble of model realizations to define the error statistics of the predicted system state [*Evensen*, 1994]. The pdf of the state vector **X** is represented by an ensemble

$$X = [x_1, x_{2,} \cdots, x_N] = [x_i] \tag{5.8}$$

where **X** is an $n \times N$ matrix, whose (n) columns are the ensemble members, and it is called the prior ensemble. Ideally, ensemble members would form a sample from the prior distribution**.**

If m is numbers of ensemble and N is the number of observations, then data **D** is given as an $m \times N$ matrix:

$$D = [d_1, d_{2,} \cdots, d_N] = [d_i] \tag{5.9}$$

$$d_i = d + \epsilon_i \tag{5.10}$$

$$\epsilon_i = (0, R) \tag{5.11}$$

The basic form of the algorithm follows a prediction-correction cycle, with corrections made to the system state whenever measurement data are available for assimilation. The prediction step involves forecasting an ensemble of model states $\mathbf{X_{k+1}}$ at forward time k+1 based on the solution to the GW model **M**, system parameters **P**, initial conditions **I**, forcing terms **q**, and boundary conditions **B.C.**, thereby generating the predicted state:

$$X_{k+1}^{p} = M(X_K; P; q; I; B.C.) \tag{5.12}$$

In GW modeling applications, each realization of the ensemble is run forward in time using a different set of system parameters (**P**). This creates an ensemble of model states in which model results at a given location are spread over a range of values signifying the uncertainty in the system prediction. At time $k + 1$, a set of measurements (e.g., hydraulic head and solute concentration) is collected, perturbed to account for measurement error, and assimilated into the system state X_{k+1}^{p} to produce a corrected system state X_{k+1}^{c} using the following correction equation:

$$X_{k+1}^{c} = X_{k+1}^{p} + K_{k+1}(d_{k+1} - BX_{k+1}^{p}) \tag{5.13}$$

Here, d_{k+1} contains the perturbed measurements, and **B** contains binary constants (0 or 1) that map model results at measurement locations to actual measurements, creating a residual at measurement locations between the predicted and actual value. The matrix **K** is termed the Kalman Gain matrix, and has the following structure:

$$K = C^{f}B^{T}(BC^{f}B^{T} + R)^{-1} \tag{5.14}$$

where C^{f} is the forecast error covariance matrix associated with the model forecast X_{k+1}^{p} and **R** is the measurement error covariance matrix associated with the perturbed measurements **d**. The formulation of **K** performs the dual role of (1) spreading information from measurement locations to regions between these locations, and thereby allowing the measurement information to correct predicted values throughout the model domain, and (2) weighting the correction terms according to model and measurement error. As **R** approaches zero, signifying low error in the measurement data, the influence of **K** increases and the residual is weighted more heavily. The model forecast values thus approach the measurement values. In contrast, as C^{f} approaches 0, signifying relative

agreement among the model realizations, the influence of **K** decreases, and the residual is weighted less heavily. The model forecast values thus receive little to no correction from the measurement data.

5.3.4 Parameter update using the EnKF

The EnKF is used to update model parameters such as hydraulic conductivity, aquifer recharge, and first order decay coefficient of nitrate-N to quantify uncertainty in hydraulic head and nitrate-N concentrations in two aquifer systems. All three parameters are considered to be random. Estimation of the parameters is performed using the EnKF in the following manner. The hydraulic conductivity [*Mallants et al.*, 1997] and the first order rate coefficient [*Frind et al.*, 1990] are known to vary log normally. An ensemble of random log spatial parameter fields, with each field encompassing the model domain, is generated using a sequential Gaussian simulation algorithm. This algorithm works using a specified mean $\boldsymbol{\mu}$, standard deviation $\boldsymbol{\sigma}$, and correlation length l as described in the Stanford Geostatistical Earth Modeling Software (S-GEMS) [*Deutsch and Journel*, 1992]. The recharge rate is assigned from the available data from the Texas Water Development Board (TWDB). All three parameters are inversely estimated by matching the model-simulated hydraulic heads and nitrate-N concentrations to the reference fields of the synthetic and Ogallala Aquifers.

It is seen that the performance of the EnKF analysis is enhanced by the ensemble size [*Evensen*, 2009]. In most published applications of the EnKF, a typical ensemble size is around 100 members. Therefore, an ensemble of 100 realizations is used in this study. These parameter fields (100 realizations) are used in a numerical contaminant

transport model to generate an ensemble of nitrate-N concentration (**C**) fields, with a value of **C** calculated for each model grid cell. These **C** ensembles and predicted parameters populate the predicted system state matrix X_{k+1}^{p}, and the correlation between parameters and nitrate-N concentrations **C** is established in the matrix C^{f}. Then, measurement data from selected grid cells of the reference **C** field are chosen and perturbed with a specified coefficient of variation to mimic error in the measurement data, and the matrix **d** is populated with these perturbed values. The matrix **d** is then used in the EnKF update routine to correct the predicted parameters and **C** ensembles. The strength of correlation, between parameters and **C** obtained by the MT3DMS model, is the criterion to correct the parameters values.

The update stage of the EnKF approach consists of populating the forecast model state matrix with system response variables and the ensemble of **K** fields (hydraulic conductivity or recharge or decay coefficient), obtaining measurements from the true state, and providing an updated estimate of the model states given the collected measurements. The update of the parameters using the EnKF approach is tested using the root mean square error (RMSE) criterion:

$$RMSE(X,t) = \sqrt{E((\hat{\theta} - \theta)^2)} \tag{5.15}$$

where E is the expectation of the difference of the estimated nitrate-N concentration against the reference field**.**

5.3.5 *Application of the modeling framework*

A synthetic 2D transient state GW flow problem is used to test the EnKF update routine to condition hydraulic conductivity, recharge, and decay coefficient fields using nitrate-N concentration data. The synthetic aquifer is constructed because the data availability (heads and nitrate-N concentrations) in the Ogallala Aquifer is limited. Therefore, the synthetic aquifer is used to examine the EnKF methodology. Subsequently, the EnKF algorithm is applied to the Ogallala Aquifer. The ENKF update algorithm using GW flow simulations is shown in Figure 5.1. The forecast stage comprises of initializing an ensemble of GW flow simulations with a separate ensemble of hydraulic conductivity, recharge, and decay coefficient fields, and initial conditions. In the following sections, details of the synthetic 2D transient flow problem and the Ogallala Aquifer system are described.

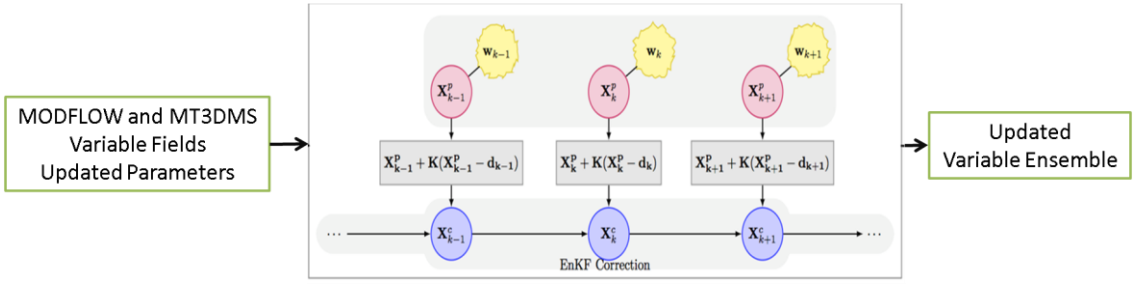

Figure 5.1: The data assimilation algorithm is presented using the ensemble Kalman filter in a GW and nitrate-N transport modeling framework. The forecast stage consists of generating the initial ensemble of K (recharge, hydraulic conductivity, or decay coefficient individually) fields and the resulting simulated concentration fields. X is the state, K is the Kalman gain, w is the noise, and *d* is the binary matrix.

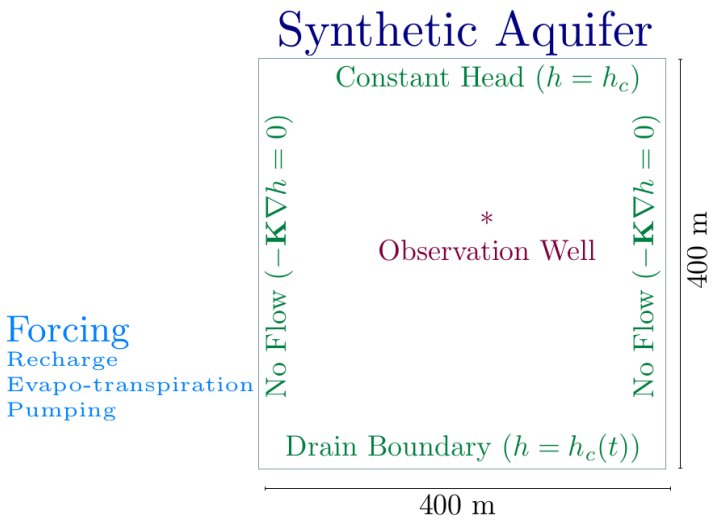

Figure 5.2: Conceptual model domain showing observation well and boundary conditions (B.C.) for the transient GW flow problem.

5.3.6 Synthetic 2D transient state GW flow model

The 2D transient flow problem consists of an areal aquifer of 4000 m (easting) by 4000 m (northing) as shown in Figure 5.2. Nitrate-N was simulated at 50 m x 50 m grids. The hydrological inputs consist of rainfall, evapo-transpiration, and surface runoff, and were applied as top boundary conditions. These inputs determine the recharge rate to the aquifer. Recharge updates account for uncertainty in the hydrological inputs. Irrigation, drainage, and pumping from wells were assigned throughout the modeling domain. No flow boundary was placed on the north and south ends of the aquifer, while all other ends were assigned drain boundaries, with an average saturated thickness of 30 m.

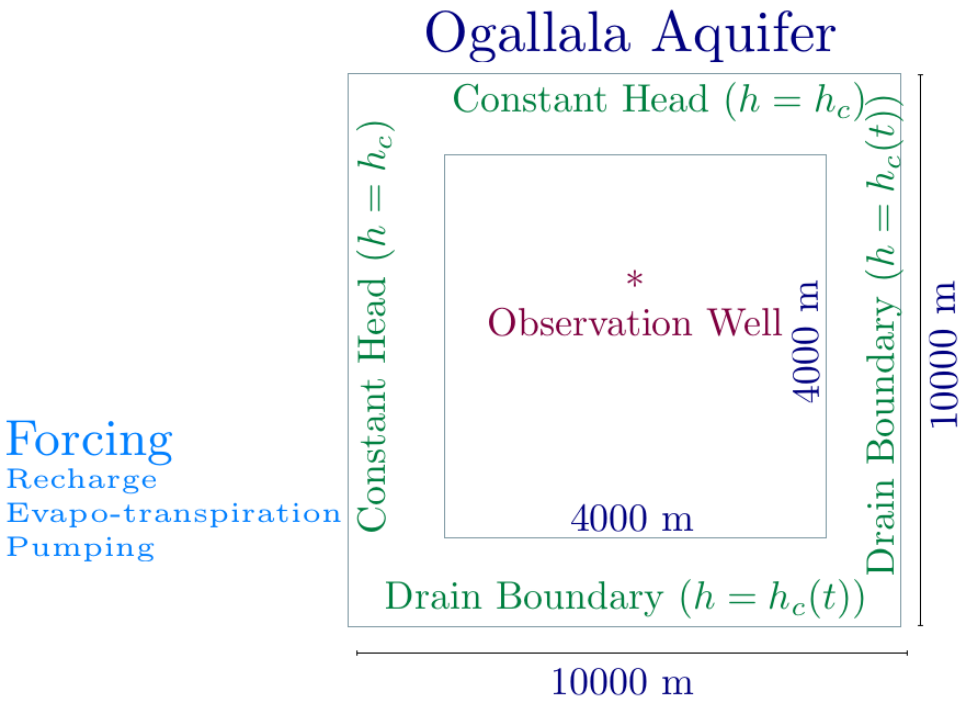

Figure 5.3: Plan view of the Ogallala Aquifer shows the observation well, forcing, and boundary conditions (B.C.). A larger domain (10,000 m by 10,000 m) was selected to import boundary condition in the problem domain to reduce the edge effect.

5.3.7 The Ogallala Aquifer GW flow model

A basic setup of a single layer model was adopted for the Ogallala Aquifer. The model is 4000 m (easting) by 4000 m (northing). Nitrate-N was simulated at 50 m x 50 m grids. The model domain for the aquifer is shown in Figure 5.3. For temporal discretization, each year was divided into 12 stress periods wherein each stress period corresponds to one month during which all inputs are constant. The top of the model domain was assigned a recharge boundary, which incorporates pumping and return flow from irrigation. A no-flow boundary was assigned as the bottom boundary condition for the Ogallala Aquifer. The outer limits of the model domain were defined by physical and

hydraulic boundaries. Therefore, the eastern, western, northern, and southern boundaries of the model domain were defined by drain conditions. The pseudo modeling domain was initially taken as 10000 m (easting) by 10000 m (northing). After model spin off (MODFLOW and MT3DMS initial runs from 1990 to 1996), the hydraulic heads obtained in the 4000 m by 4000 m modeling domains were employed as initial conditions.

Aquifer properties and model parameters for the synthetic and Ogallala aquifers are provided in Tables 5.1 and 5.2. All simulations were carried out for a 10 year period (1996-2005) for both the aquifers.

Table 5.1: Hydrogeologic and other properties of the Ogallala Aquifer used for modeling the Ogallala Aquifer and the synthetic flow problem [*Blandford et al.*, 2003*; Long et al.*, 2003; *Dutton et al.*, 2004]*.

Property	Ogallala Aquifer
	Parameter range
Hydraulic conductivity in longitudinal direction K_x (m/sec)	7×10^{-7}
Hydraulic conductivity in lateral direction K_y (m/sec)	7×10^{-8}
Hydraulic conductivity in vertical direction K_z (m/sec)	7×10^{-8}
Transmissivity (m^2/sec)	1×10^{-5}
Specific storage S_s (1/m)	2.6×10^{-3}
Specific Yield S_y (-)	0.05
Effective Porosity (-)	0.2
Total Porosity (-)	0.3

* Hydraulic conductivity and recharge values were used to initialize the respective fields, later updated by the EnKF algorithm.

Table 5.2: Parameters for nitrate-N transport [obtained from *Gelhar*, 1992; Mehta et al., 2000; Burton, 2007; *Bronson et al*., 2009]. First order reaction rate values were used to initialize the decay coefficient field, and were later updated by the EnKF algorithm.

Model properties	Values
Longitudinal Dispersivity (m)	10
Horizontal/Longitudinal Dispersivity (-)	0.1
Vertical/Longitudinal Dispersivity (-)	0.01
Diffusion Coefficient (m^2/sec)	5.7×10^{-10}
First Order Reaction Rate for Dissolved Phase (1/sec)	3.5×10^{-6}
First Order Reaction Rate for Sorbed Phase (1/sec)	0

5.3.8 Numerical study

For completeness, a brief description of the numerical model is given below. Numerical experiments (Figures 5.2 and 5.3) were conducted using MODFLOW [*McDonald and Harbaugh*, 1984]. The governing equations for GW flow and nitrate-N transport are given in 4.20 and 4.21, respectively.

5.4 Results and discussion

The data assimilation (DA) algorithm using the EnKF filter in a GW and nitrate-N transport modeling framework was developed to estimate hydraulic heads and nitrate-N concentrations in aquifers by sequentially updating parameters, and the model state. In this section, first we describe the applicability and validity of the EnKF algorithm in a synthetic 2D aquifer. Subsequently we will discuss the applicability and performance of the EnKF algorithm in real field conditions (Ogallala Aquifer), when only limited observations are available. We also describe how hydraulic conductivity, first order decay coefficient, and recharge parameters change over time, when the EnKF updates these parameters. As described in Figure 5.1, the MODFLOW and MT3DMS models

were used to assimilate hydraulic heads and nitrate-N concentrations Figures 5.2 and 5.3 show the conceptual framework of the synthetic 2D and the Ogallala aquifers, respectively.

5.4.1 Parameter optimization using the EnKF algorithm

Figure 5.4 demonstrates the change in the fractional normalized entropy (FNME) values of hydraulic conductivity, first order decay coefficient, and recharge in aquifers during EnKF assimilations. Normalized marginal entropy (NME) is described in section 4.3.2. FNME is the NME divided by 100. Figure 5.4 reveals that hydraulic conductivity manifests asymptotic behavior in both aquifers, after fifty EnKF assimilations. First order decay coefficient does not change like other two parameters; however, decay coefficient shows more uncertainty in its estimation in the Ogallala Aquifers as compared to the synthetic aquifer. FNME of recharge values describes the importance of the stress periods (recharge and its frequency) in real field conditions. For a controlled system, like the synthetic aquifer, recharge approaches asymptotic values with respect to the reference field after fifty EnKF assimilations; whereas recharge does not display asymptotic behavior in the Ogallala Aquifer.

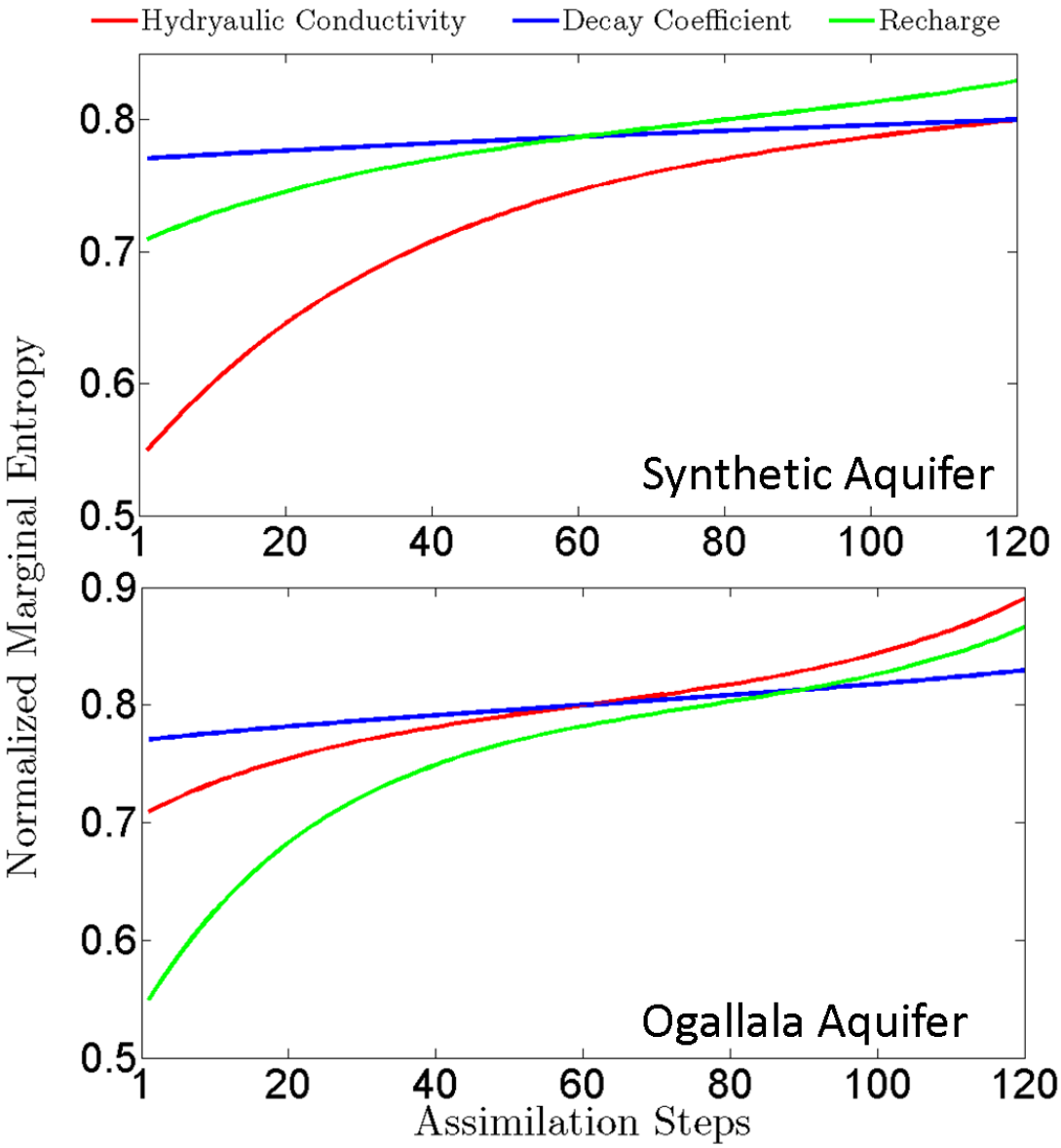

Figure 5.4: The temporal evolution of the log hydraulic conductivity, first order decay coefficient, and recharge fields for the synthetic and Ogalalla aquifers shows the asymptotic behavior.

Figures 5.5 and 5.6 show the temporal evolution of the log hydraulic conductivity fields in the synthetic and Ogallala Aquifers. The log hydraulic conductivity fields shown in the Figures 5.5 and 5.6 are the realizations of log saturated hydraulic conductivity fields when 2, 10, 50 and 100 measurements of nitrate-N concentrations are assimilated in the synthetic and Ogallala Aquifers.

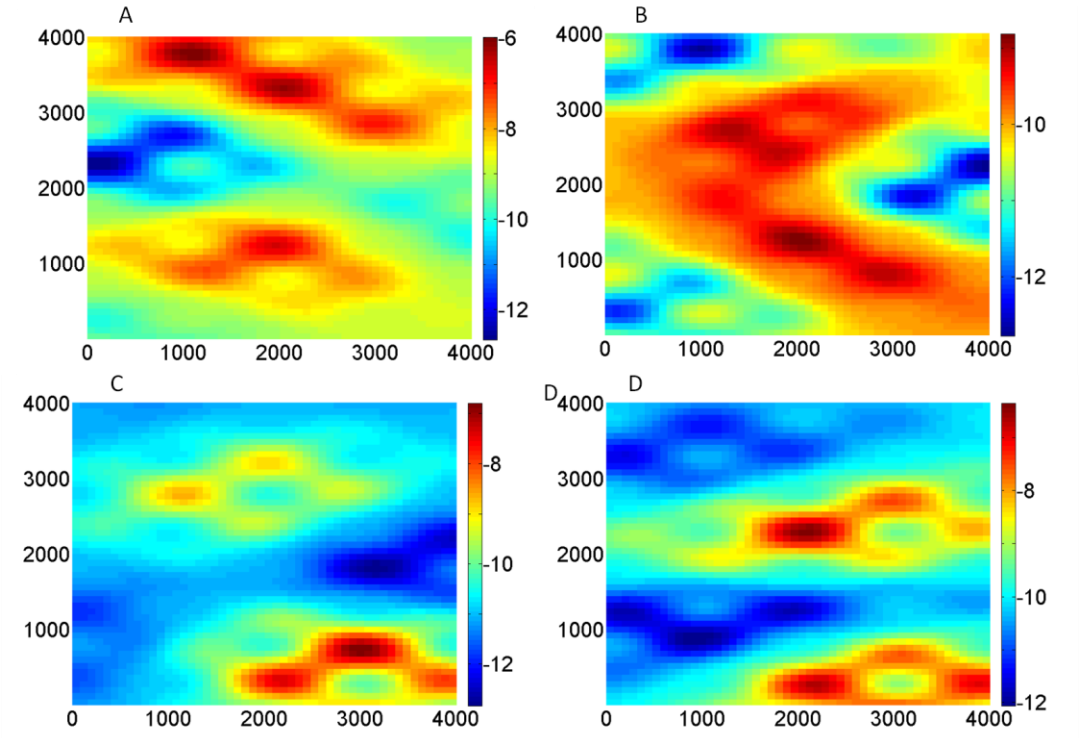

Figure 5.5: The temporal evolution of the log hydraulic conductivity fields (m/sec) when (A) 2, (B) 10, (C) 50, and (D) 100 measurements of nitrate-N concentrations are assimilated in the synthetic Aquifer.

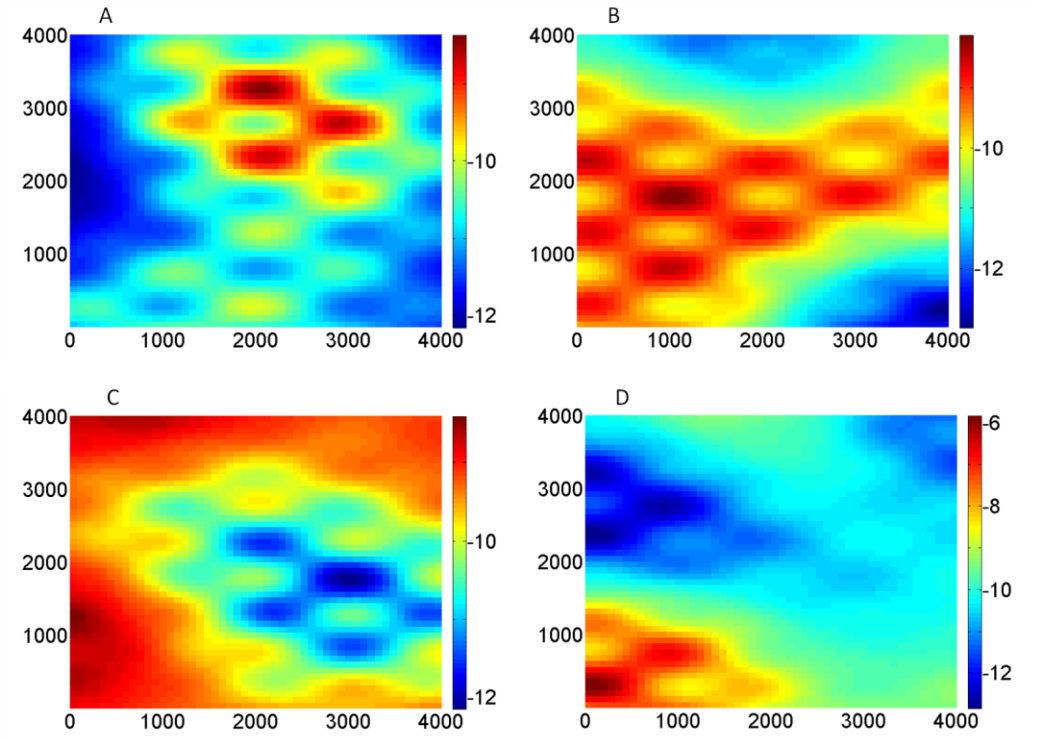

Figure 5.6: The temporal evolution of the log hydraulic conductivity fields (m/sec) when (A) 2, (B) 10, (C) 50, and (D) 100 measurements of nitrate-N concentrations are assimilated in the Ogallala Aquifer.

It is evident that hydraulic conductivity fields change over time. This change is more visible when 2 or 10 nitrate-N measurements are assimilated in the synthetic Aquifer (Figures 5.5 A and 5.5 B). This evolution of temporal hydraulic conductivity fields can be attributed to uncertainty in the hydraulic conductivity parameter. The hydraulic conductivity further changes when 50 nitrate-N measurements are assimilated, but there is no significant change from the assimilation of 50 nitrate-N measurements to 100 nitrate-N measurements. This reflects that hydraulic conductivity keeps evolving, and after 50 assimilations its estimation is closer to the true hydraulic conductivity field of the system. On the other hand, the hydraulic conductivity fields keep evolving with

the assimilation of 2, 10, 50 and 100 concentrations of nitrate-N measurements in the Ogallala Aquifer. Therefore, more uncertainty is associated with the hydraulic conductivity fields of the Ogallala Aquifer as compared to the synthetic Aquifer. This is because the available hydraulic conductivity fields in the Ogallala Aquifer are at the coarse scale (km scale) and the reference fields (hydraulic heads and nitrate-N measurements) are not available at each time step.

Figure 5.7 shows the probability distributions (PDFs) of the log hydraulic conductivity fields in the synthetic and Ogallala Aquifers. The PDFs show the temporal evolution of the mean and spread of the hydraulic conductivity. In the synthetic Aquifer, the mean of hydraulic conductivity decreases slightly and the spread of the PDF increases with the increase in the number of measurements assimilated. In the Ogallala Aquifer, the mean of hydraulic conductivity increases from the assimilation of 2 to 10 measurements, and remains constant beyond that. However, the spread of the PDF decreases with the increase in the number of measurements assimilated. Therefore, the temporal evolution of hydraulic conductivity demonstrates that conditioning of the parameters is a function of the number of assimilated measurements.

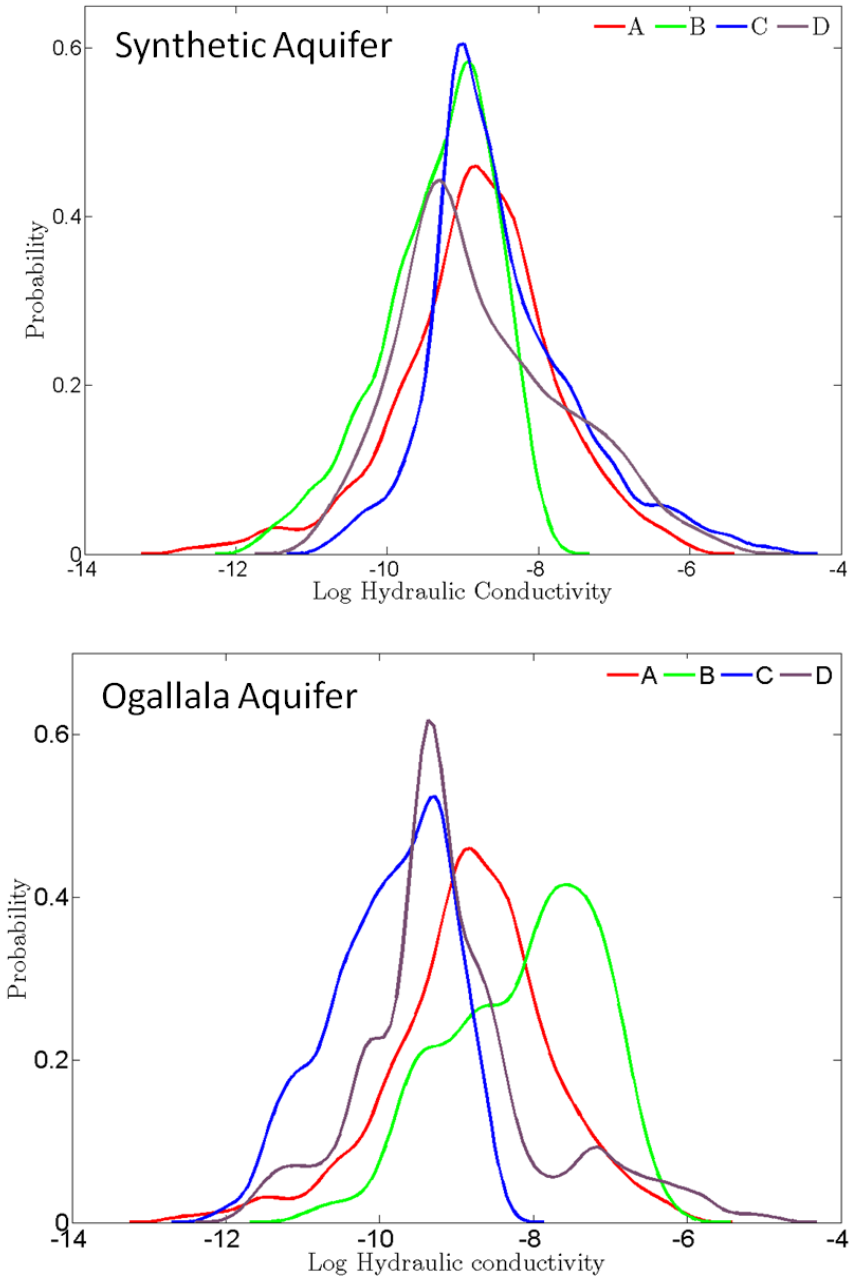

Figure 5.7: The temporal evolution of the log hydraulic conductivity fields (m/sec) when (A) 2, (B) 10, (C) 50, and (D) 100 measurements of nitrate-N concentrations are assimilated in the Ogallala Aquifer.

Figures 5.8 and 5.9 show the conditioning of the recharge values in the synthetic and Ogallala Aquifers when 100 measurements of nitrate-N concentrations are assimilated in the synthetic and Ogallala Aquifers. Recharge values (as input to the flow-transport model) are moderately conditioned for uncertainty, which may result from multiple factors such as runoff generation, focused recharge through playa, return flow from irrigation, etc.

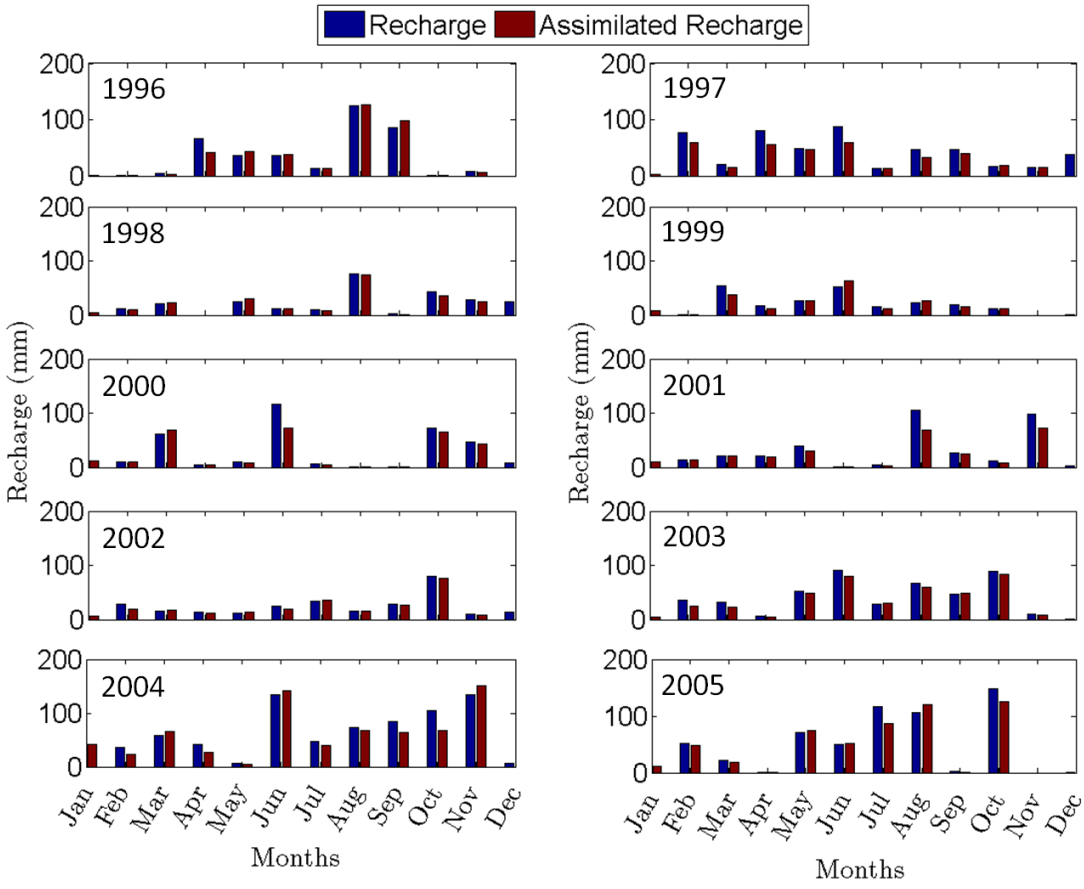

Figure 5.8: The reference recharge and assimilated recharge values for each month (from 1996 to 2005) in the synthetic Aquifer when 100 measurements of nitrate-N concentrations are assimilated.

It is evident from Figure 5.8 that recharge values are conditioned mostly for intermediate values (less than 70 mm and more than 30 mm) in the synthetic Aquifer. Possible reasoning for this is the importance of intermediate recharge values for flow and transport processes. When there are very high or low recharge events, it becomes less a sensitive parameter. Similar results are obtained for the conditioning of recharge values in the Ogallala Aquifer.

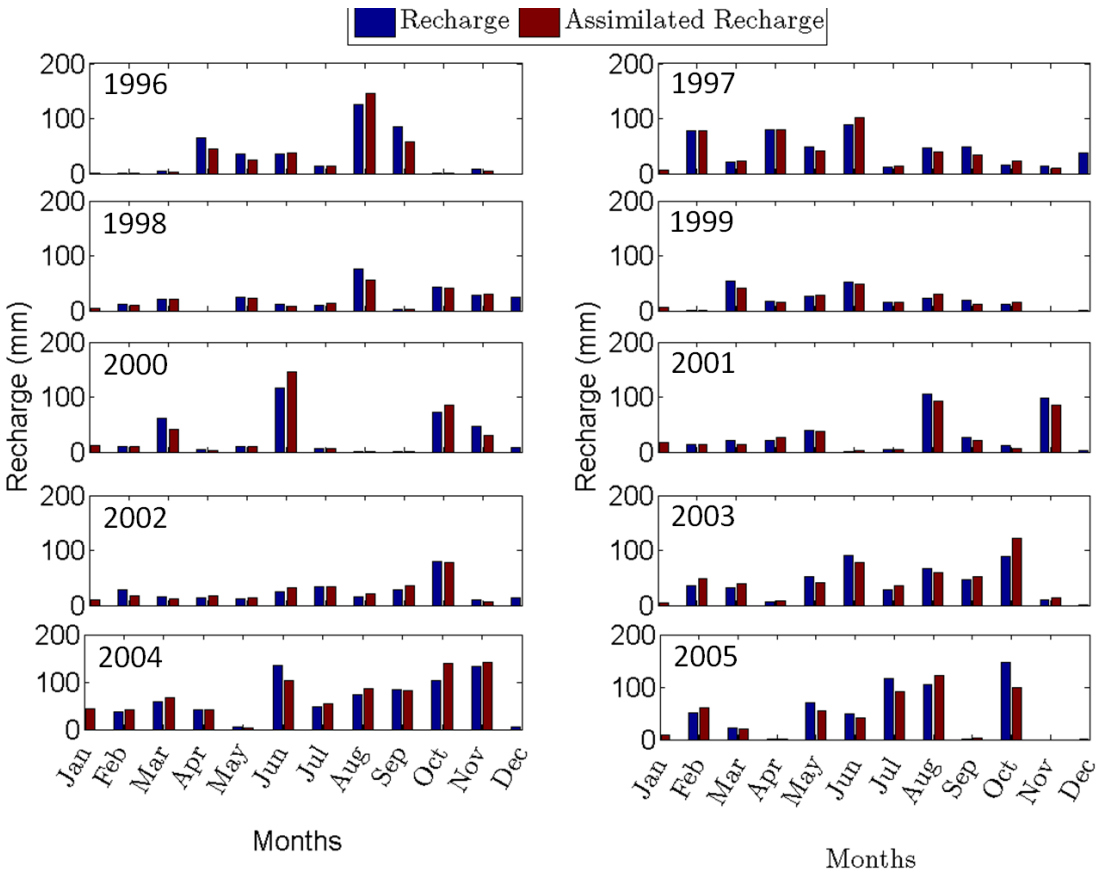

Figure 5.9: The reference recharge and assimilated recharge values for each month (from 1996 to 2005) in the Ogallala Aquifer when 100 measurements of nitrate-N concentrations are assimilated.

5.4.2 DA algorithm in the synthetic 2D Aquifer

Figure 5.10 shows the estimated hydraulic head and nitrate-N concentrations in the synthetic 2D Aquifer after the hydraulic head observations from the reference hydraulic head and nitrate-N concentration fields are assimilated using the EnKF approach. In this step, the simulation analysis involves solving the flow (4.20) and transport (4.21) equations by updating first order decay coefficient, recharge, and hydraulic conductivity independently. Figure 5.10A displays the ensembles of hydraulic heads and Figure 5.10D presents the ensembles of nitrate-N concentrations of the 100 realizations after the first order decay coefficient was updated in time. Figure 5.10B shows the ensembles of hydraulic heads and Figure 5.10E presents the ensembles of nitrate-N concentrations of the 100 realizations after the recharge was updated in time. Similarly, Figure 5.10C displays the ensembles of hydraulic heads and Figure 5.10F presents the ensembles of nitrate-N concentrations of the 100 realizations after the hydraulic conductivity was updated.

It is visible from Figure 5.10 that updates of the parameters are successful in capturing the variability of hydraulic heads and nitrate-N concentrations in the synthetic 2D Aquifer. Regardless of some local mismatch, the general trend is consistent between the simulated heads and nitrate-N concentrations and the reference field. Figures 5.10 (A, B, and C) show that hydraulic conductivity is the most important parameter in improving the estimation of hydraulic head, followed by recharge. Since recharge is not continuous in nature, hydraulic heads are not influenced by recharge as much as they are

influenced by the hydraulic conductivity parameter. The first order decay coefficient does not play any role in improving the estimation of hydraulic heads (Figure 5.10A).

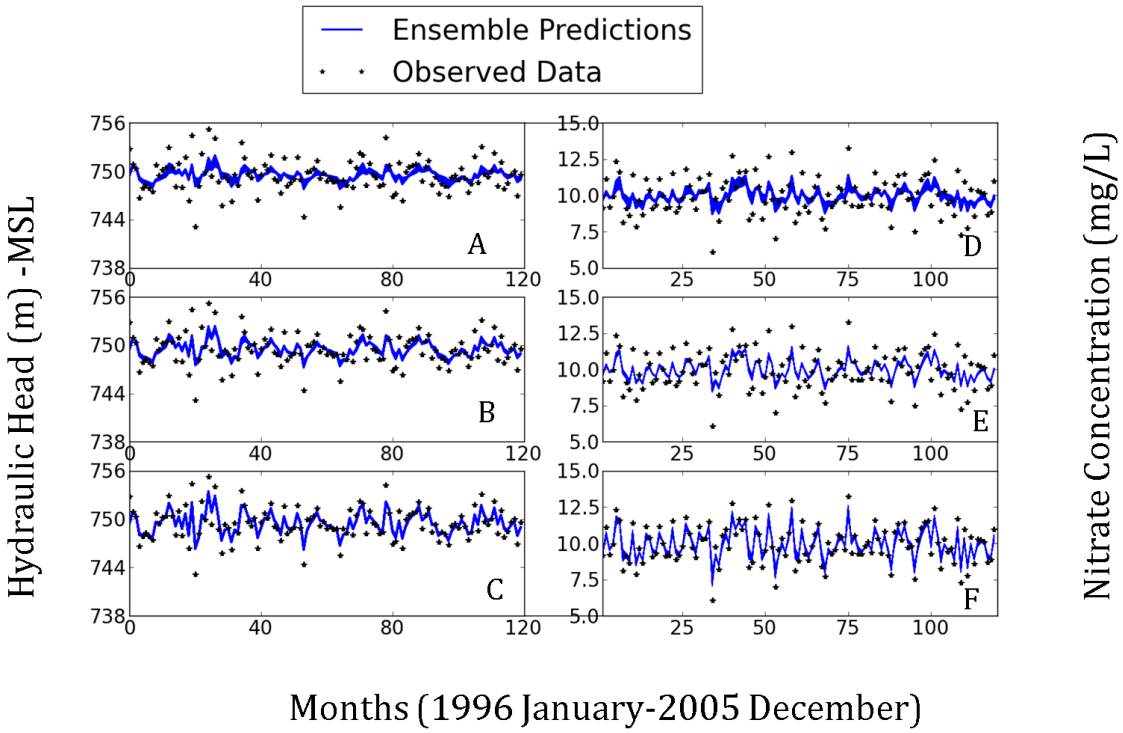

Figure 5.10: Estimation of hydraulic heads and nitrate-N concentrations in the synthetic 2D problem using the EnKF algorithm with 100 ensembles. Estimation of (A) the ensemble hydraulic heads by updating the first order decay coefficient; (B) the ensemble hydraulic heads by updating the recharge; (C) the ensemble hydraulic heads by updating the hydraulic conductivity; (D) the ensemble nitrate-N concentrations by updating the first order decay coefficient; (E) the ensemble nitrate-N concentrations by updating the recharge; and (F) the ensemble nitrate-N concentrations by updating the hydraulic conductivity.

It is evident from Figures 5.10 (D, E, and F) that hydraulic conductivity is again the most important parameter in improving the estimation of nitrate-N concentrations, followed by recharge, and then by the first order decay coefficient. This is expected because advection plays an important role in nitrate-N transport so the impact of hydraulic conductivity is more than the other two parameters. The EnKF method assimilates nitrate-N concentrations after the hydraulic heads are assimilated and therefore, nitrate-N concentrations also incorporate improved hydraulic heads.

The vertical spread of these ensembles in Figure 5.10 represents the uncertainty in the prediction. The uncertainty band is not wide, however as we move forward from the first month to the next month, the EnKF improves the predicted state through the Bayesian update. It is also apparent that the EnKF reduces the mismatch between predictions and observations at different measurement locations and thereby produces a narrow uncertainty band.

It is seen in Figure 5.10 that the uncertainty band is getting reduced from the Bayesian updates of all three parameters (first order decay coefficient, recharge, and hydraulic conductivity) in estimating hydraulic heads and nitrate-N concentrations. This uncertainty is further quantified by plotting the histograms of the root mean square error (RMSE) in the section 5.3.4. The width of the uncertainty bands further substantiate that the hydraulic conductivity parameter improves the ensemble predictions the most, followed by recharge, and then by the first order decay coefficient in the estimation of nitrate-N concentrations.

5.4.3 DA algorithm in the synthetic 2D Aquifer

Figure 5.11 shows the estimated hydraulic head and nitrate-N concentrations in the Ogallala Aquifer after the hydraulic head observations from the reference hydraulic head and nitrate-N concentration fields are assimilated using the EnKF approach. The difference in assimilating hydraulic heads and nitrate-N concentrations in real field conditions (Ogallala Aquifer) and in the synthetic Aquifer is the availability of data (reference fields). The EnKF approach adjusts the unknown model parameters based on the observed data with time. The hydraulic heads and nitrate-N concentrations were estimated in the Ogallala Aquifer from January 1996 to December 2005, a total of 120 months. In this span of time, 14 observations of hydraulic heads and nitrate-N concentrations were available. These observations were used as the reference field for the EnKF algorithm.

It is apparent from Figure 5.11 that updates of the parameters resulted in capturing the variability of hydraulic heads and nitrate-N concentrations. Figure 5.11 shows that hydraulic conductivity is again the most important parameter in improving the estimation of hydraulic head and nitrate-N, followed by recharge, and then by the decay coefficient. However, the first order decay coefficient improves nitrate-N estimations more in the case of the Ogallala Aquifer as compared to the synthetic 2D Aquifer.

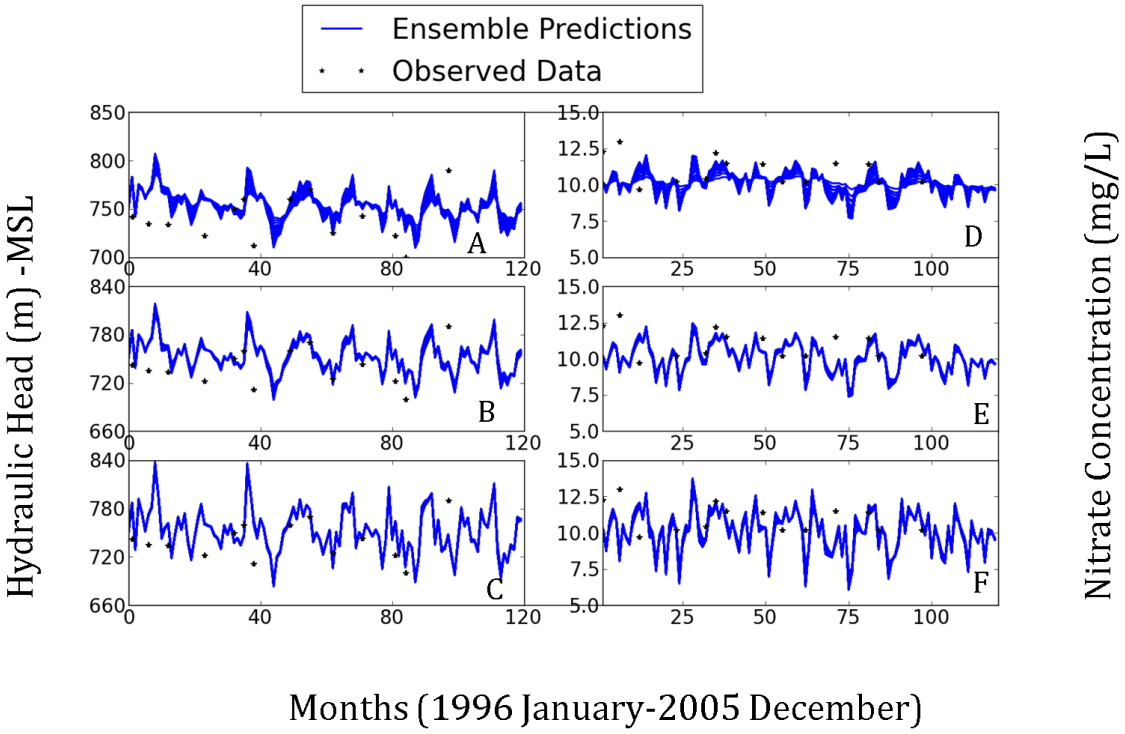

Figure 5.11: Estimation of hydraulic heads and nitrate-N concentrations in the Ogallala Aquifer using the EnKF algorithm with 100 ensembles. Estimation of (A) the ensemble hydraulic heads by updating the first order decay coefficient; (B) the ensemble hydraulic heads by updating the recharge; (C) the ensemble hydraulic heads by updating the hydraulic conductivity; (D) the ensemble nitrate-N concentrations by updating the first order decay coefficient; (E) the ensemble nitrate-N concentrations by updating the recharge; and (F) the ensemble nitrate-N concentrations by updating the hydraulic conductivity.

The decay coefficient plays a key role in real field situations because it represents important processes that govern the fate of nitrate-N in GW. To understand the importance of the first order decay coefficient, we present here a brief historical background of nitrate-N contamination in the Ogallala Aquifer. Figures 4.18 and 4.19 show mean nitrate-N concentrations and percentage samples having nitrate-N exceeding 10 mg/L for different decades in the Ogallala Aquifer at various spatial scales. Nitrate-N concentrations start to rise around 1970-80 in accordance with the increasing load at

surface level (fertilizers). Nitrogen applications in the 1960s and 1970s ranged from 250 to 300 kg ha^{-1}, whereas recent surveys suggest most applications have declined to 150 to 180 kg ha^{-1} [*Spalding et al.*, 2001]. This suggests almost 50% reduction in the application rate. The response of nitrate-N concentrations in GW to the higher load (in 1960-70) at the land surface is thus noticeably fast. However, the concentrations in the GW keep decreasing beyond 1990 even though the nitrate-N concentrations in the recharge water decrease around 1985. This implies that the nitrate-N application at the land surface arrives in the groundwater at least 10 years later, and there are other processes (such as denitrification, nitrogen cycling) that govern the fate of nitrate-N.

The uncertainty in the estimation of hydraulic heads and nitrate-N concentrations is larger in the Ogallala Aquifer than the uncertainty in the synthetic Aquifer. This is because the reference field data in the Ogallala Aquifer was scarce and has more variation as compared to the reference field data of the synthetic Aquifer. The width of the uncertainty bands in Figure 5.11 also corroborate that the hydraulic conductivity parameter improves the ensemble predictions of hydraulic head and nitrate-N concentrations the most, followed by recharge, and then the first order decay coefficient.

5.4.4 Root mean square error in the EnKF predictions

Figures 5.12 and 5.13 show the root mean square error in the estimation of hydraulic head and nitrate-N concentrations in the synthetic and Ogallala Aquifers while updating the first order decay coefficient, recharge, and hydraulic conductivity parameters. In the synthetic Aquifer, the RMSE values while estimating the hydraulic

heads range from 1.25 to 1.60, 0.80 to 1.20, and 0.30 to 0.70 for the Bayesian updates of the first order decay coefficient, recharge, and hydraulic conductivity, respectively.

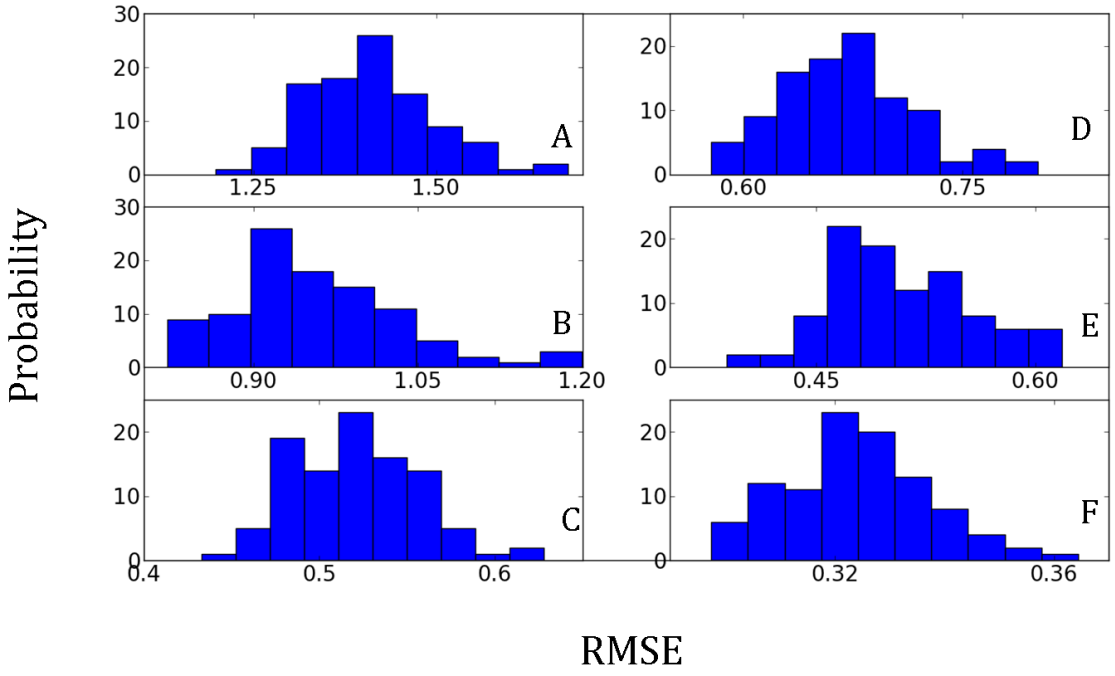

Figure 5.12: Root mean square error (RMSE) in estimating ensemble hydraulic heads in the synthetic 2D problem by updating the (A) first order decay coefficient; (B) recharge; (C) hydraulic conductivity; and the RMSE in estimating ensemble nitrate-N concentrations by updating the (D) first order decay coefficient; (E) recharge; and (F) hydraulic conductivity.

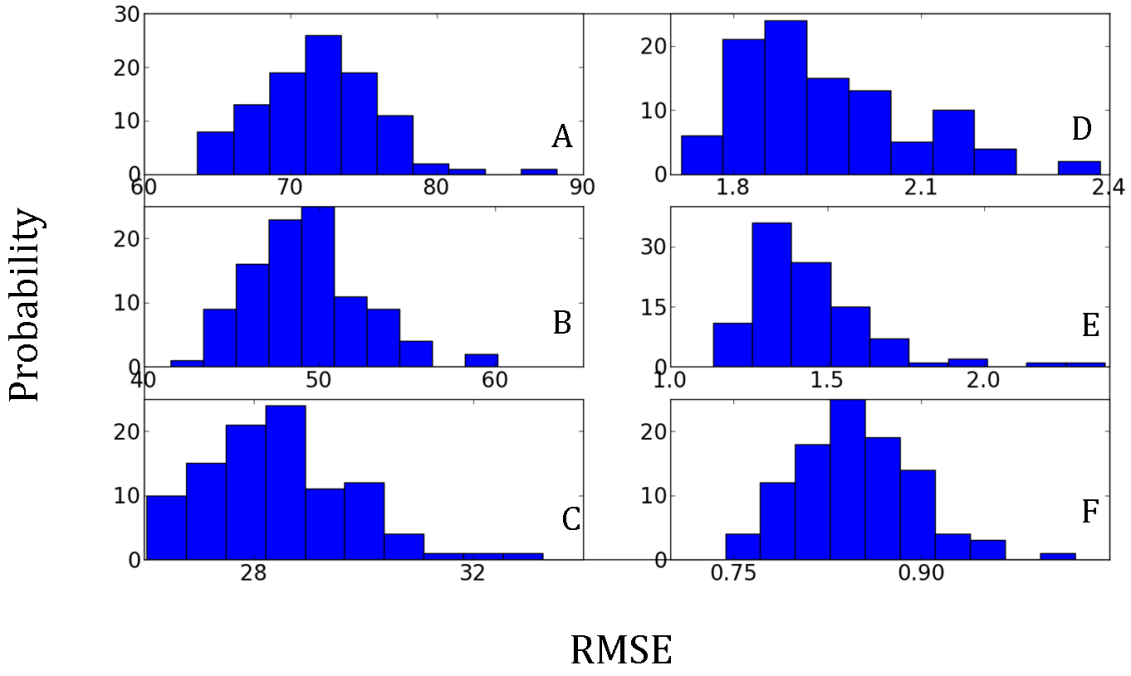

Figure 5.13: Root mean square error (RMSE) in estimating ensemble hydraulic heads in the Ogallala Aquifer by updating the (A) first order decay coefficient; (B) recharge; (C) hydraulic conductivity; and the RMSE in estimating ensemble nitrate-N concentrations by updating the (D) first order decay coefficient; (E) recharge; and (F) hydraulic conductivity.

On the other hand, the RMSE values while estimating the nitrate-N concentrations range from 0.60 to 0.80, 0.30 to 1.20, and 0.20 to 0.40 for the Bayesian update of the first order decay coefficient, recharge, and hydraulic conductivity, respectively. In the Ogallala aquifer, the RMSE values while estimating the hydraulic heads range from 60 to 90, 45 to 60, and 0 to 35 for the EnKF updates of the first order decay coefficient, recharge, and hydraulic conductivity, respectively. In contrast, the RMSE values while estimating the nitrate-N concentrations range from 1.2 to 2.4, 1.2 to 2.0, and 0.75 to 1.0 for the Bayesian update of the first order decay coefficient, recharge, and hydraulic conductivity, respectively. It is evident that the synthetic Aquifer has

lesser RMSE in all simulations. This shows the importance of the reference field and the strength of the EnKF approach when data are available. It is also visible from Figures 5.12 and 5.13 that the RMSE values are smaller in the estimation of nitrate-N concentrations as compared to the hydraulic heads.

5.5 Summary and conclusions

The EnKF, a data assimilation algorithm that conditions uncertain parameters with measurement data within a Bayesian framework, was applied to investigate the flow and nitrate transport processes at the Ogallala Aquifer. MODFLOW and MT3DMS models were used for estimating hydraulic heads and nitrate-N concentrations. Hydraulic conductivity, recharge, and first order decay coefficient parameters associated with flow and transport models were updated using EnKF, and evaluated for a synthetic groundwater and the Ogallala Aquifer. This study highlights several key findings. The EnKF improved nitrate-N estimates more than hydraulic heads. It was also found that the EnKF reduces the mismatch between observations and predictions over time for both the synthetic and Ogallala Aquifer systems. Hydraulic conductivity was found to be the most important parameter in estimating hydraulic heads and nitrate-N concentrations in aquifers, followed by recharge, and then by the decay coefficient.

The EnKF can be a potential alternative approach for inverse estimation of parameters. The optimization of parameters using the EnKF approach demonstrated the asymptotic behavior (hydraulic conductivity and first order decay coefficient). The temporal conditioning of recharge parameter showed that nitrate-N concentrations are more sensitive to intermediate recharge values.

CHAPTER VI

A HYBRID DETERMINISTIC-STOCHASTIC MODELING FRAMEWORK FOR CONTAMINANT TRANSPORT IN THE SUBSURFACE

6.1 Synopsis

Microbes are major contaminants of surface water while nitrate is the most ubiquitous contaminant in groundwater. Hydrological interactions between surface water bodies and groundwater are of fundamental concern to the migration of contaminants (microbes and nitrate) [*McMahon et al.*, 1995; *Bethune et al.*, 1996]. The vadose zone acts as an interface between surface water and groundwater. Once these contaminants enter the subsurface environment, they are subjected to a variety of coupled hydrological, geochemical, and biological processes. There is significant uncertainty associated with geochemical and microbiological processes due to a lack of easily available data and heterogeneity in terms of redox potential of the subsurface systems. Since most hydrologic analyses focus exclusively on the optimization of model parameters and ignore inadequate model structure (structural uncertainty), we present a conceptual framework that incorporates different model structures for complex biogeochemical processes. In the proposed hybrid conceptual modeling framework physical processes (e.g., advection, dispersion) are modeled as deterministic partial differential equations, while chemical (e.g., nitrification, denitrification) and biological processes (e.g., population growth) are modeled as stochastic processes (e.g., chemotaxis). In this hybrid deterministic-staocastic modeling scheme, we focus here on

capturing the influence of stochastic chemical and biological processes under deterministic hydrological feedbacks on nitrate and *E. coli* transport (as two separate case studies) in a 1-D soil column and understanding their dynamics under perturbed conditions. We demonstrated that the hybrid-stochastic-deterministic model improves the predictive capabilities. Results from the nitrate transport study indicate that there is higher uncertainty in predicting ammonium concentrations in the soil column as compared to nitrate and nitrite concentrations when bio-chemical processes are modeled stochastically. Perturbations to soil temperature and pH cause variability in the vertical distribution of nitrate and ammonia. Results from the *E. coli* transport study suggest that the *E. coli* moves faster in the column with a slower rate of deposition. *E. coli* concentrations in the soil profile reach the breakthrough faster if both the growth and die-off rates are higher, whereas perturbations to chemotactic motility of *E. coli* suggest a delay in the breakthrough.

6.2 Introduction

In surface water, microbes have been identified as major contaminants of water resources in the US [USEPA, 2006]. In groundwater, nitrate from both natural and anthropogenic sources, has been identified as the most widespread contaminant [*Nolan et al.*, 2002]. Surface water and groundwater interact directly at the recharge interfaces or indirectly through the vadose zone. As a result, much of the contamination of groundwater occurs through the vadose zone [*Feyen et al.*, 1998]. Therefore, the importance of vadose zone in cycling and containing contaminants is substantial. Interest in the vadose zone function has remarkably increased in recent years due to the growing

concern of the subsurface being adversely affected by agricultural, industrial, and municipal activities.

Once contaminants are released into the subsurface environment, they are subjected to a variety of coupled physical, chemical, and biological processes, such as advection, dispersion, ion-exchange, sorption-desorption, biotransformation, biodegradation [*Ginn et al.*, 2002]. A variety of analytical and numerical models, encompassing such processes, are available to predict water and/or solute transport in the vadose zone. These predictive models provide the framework for integration in terms of processes and their linkages under various scenarios of subsurface conditions for assessing and optimizing the fate and transport of contaminants. Some of the earlier one- and multi-dimensional analytical transport models include AT123D [*Yeh*, 1981], CXTFIT [*Parker and van Genuchten*, 1984], and 3DADE [*Leij et al.*, 2000]. However, other models developed afterwards can cope with much more complex system-dependent boundary conditions evaluating mass and energy balances. Such models include DAISY [*Hansen et al.*, 1990], TOUGH2 [*Pruess*, 1991], SHAW [*Flerchinger et al.*, 1996], SWAP [*van Dam et al.*, 1997], HYDRUS [*Šimůnek et al.*, 2006], UNSAT-H [*Fayer*, 2000], COUP [*Jansson and Karlberg*, 2001], MIN3P [*Mayer et al.*, 2001], and STOMP [*Lenhard et al.*, 1995; *White et al.*, 1995]. The widely-used models for predicting fate and transport of contaminants in the subsurface are typically based on the Richards' Equation for variably saturated flow and the Fickian-based Advection-Dispersion Equation (ADE) for solute transport [*Šimůnek et al.*, 2006]. These models are used to address subsurface fate and transport of nitrate and *E. coli* in the vadose zone.

Despite advances in subsurface modeling, there is growing evidence that many subsurface systems are not accurately predictable using these models. This situation can occur when subsurface systems involve one of the following, over simplified parameterization of complex processes, inadequate model structure (conceptualization), or processes triggered by microorganisms. For instance, relatively complex processes, such as adsorption and cation exchange, are often accounted for by means of empirical linear or nonlinear adsorption isotherms [*Goncalves et al.*, 2006]. This leads to uncertainty in the output because of simplified parameters. An example illustrating inadequate conceptualization is the existence of more than fifty models representing the denitrification process [*Heinen*, 2006]. Thus, it is always challenging to decide which denitrification model is best suited for a given scenario. Microorganisms also pose challenges in system understanding and prediction, as they are adaptive life forms existing in an environment difficult to observe and measure.

There are various approaches to improve predictions of fate and transport of contaminants in the subsurface, such as parameter estimation methodologies based on optimization and data assimilation algorithms [*Ahmed and Demarsily*, 1987; *Dagan*, 1979; *Dagan and Lessoff*, 2007; *Indelman and Dagan*, 1993; *Kitanidis and Bras*, 1980; *Neuman*, 1980; *Neuman and Yakowitz*, 1979; *Rubin et al.*, 1990], selection of the best models for a given scenario, and using these approaches jointly to render optimum predictions under uncertainty (Bayesian Model Averaging) [*Gaganis and Smith*, 2001; *Neuman and Wierenga*, 2003; *Ye et al.*, 2004], as well as stochastic modeling [*Freeze*, 1975; *Gedeon*, 1986; *Gelhar and Axness*, 1983; *Gelhar et al.*, 1979; *Neuman et al.*,

1987]. Yet, most hydrologic analyses focus exclusively on the optimization of model parameters and ignore inadequate model structure (structural uncertainty). This often leads to overconfidence in the predictions of the model, which are rarely substantiated with the sparsely available hydrologic data [*Neuman and Wierenga*, 2003].

Therefore, we present a conceptual framework that incorporates different model structures for complex subsurface processes associated with nitrate and *E. coli* transport. However, we must account for the advantages and limitations of conducting stochastic simulations. In this regard, *Dagan* [1987; 1988] formulated a linear model of stochastic transport, which united the work of [*Dagan*, 1982; 1984; *Gelhar and Axness*, 1983; *Neuman et al.*, 1987]. In this formulation, all nonlinear terms are neglected, such as those arising from the deviation of solute particles from their mean trajectory. Furthermore, the simulated water content and solute concentration distributions are quantified in terms of space averages and two-point autocorrelation functions, and the time evolution of the solute plume is quantified in terms of its first two normalized spatial moments [*Russo*, 1991]. Finally, the output concentration is depicted as the ensemble mean concentration. Hence, such stochastic models for contaminant transport in the vadose zone exist and can describe the general trend of the solute plume behavior in time but fail to predict the actual instantaneous spread of a single plume [*Russo*, 1991], which signifies the limitations of stochastic models. Apart from solute transport, uncertainty in physical processes has been extensively explored in the hydrologic literature of [*Dagan*, 1982; 1984; *Gelhar and Axness*, 1983; *Neuman et al.*, 1987]. Many microbial processes can also be described using either deterministic partial-differential

equations or stochastic simulation algorithms [*Erban and Chapman*, 2007]. However, the stochastic simulation algorithms provide a more comprehensive and precise picture of the microbial processes [*Erban and Chapman*, 2007].

In this study, we present a conceptual model with physical processes represented in a deterministic framework, and chemical and biological processes as stochastic processes. This hybrid deterministic-stochastic conceptual framework is tested on a 1-D soil column where the emphasis is on capturing the influence of deterministic hydrological as well as stochastic bio-chemical processes separately on (1) nitrate and (2) *E. coli* transport. The dynamics of nitrate transport and *E. coli* deposition and transport are further explored under perturbed systems (i.e., soil temperature and soil pH for nitrate transport; *E. coli* growth-death and chemotaxis for *E. coli* deposition and transport).

6.3 Methodology

In this study, we present a novel approach to understand the interactions between biogeochemical and hydrological processes on nitrate and *E. coli* transport and their individual behavior in the subsurface environment under perturbed conditions. First, we will discuss the physical, chemical, and biological processes governing the fate and transport of nitrate and *E. coli* in subsurface systems, and their mathematical formulation. Then, we will present the application of this modeling framework to a 1D soil column study.

6.3.1 *Factors affecting nitrogen transport through soils and their mathematical formulations*

Many aspects of the soil nitrogen cycle have been described elsewhere [*Del Grosso et al.*, 2000; *Parton et al.*, 1996; *Shaffer*, 2002; *Tanji et al.*, 1979]. Here, we briefly describe nitrogen cycle components relevant to nitrate transport in the vadose zone (Figure 6.1A).

Advective-dispersive transport in the vadose zone is the major mechanism for chemical transport. Richards' equation is used for describing variably-saturated flow:

$$\frac{\partial \theta}{\partial t} = \vec{\nabla}.K(h)\nabla h - S \tag{6.1}$$

where t is time [T], θ is the water content [L^3L^{-3}], h is the pressure head [L], K is the unsaturated hydraulic conductivity [LT^{-1}], S is a sink term. $\vec{\nabla}$ is the *gradient* operator, and ∇ is the divergence *operator*.The general advection dispersion equation (ADE) is used for modeling solute transport:

$$\frac{\partial C}{\partial t} = D_L \frac{\partial^2 C}{\partial x^2} - v_x \frac{\partial C}{\partial x} - \frac{B_d}{\theta}\frac{\partial C_s}{\partial t} + \left(\frac{\partial C}{\partial t}\right)_{rxn} \tag{6.2}$$

where c is the concentration of nitrate in the liquid phase, D_L is the longitudinal dispersion coefficient, v_x is the pore water velocity, B_d is the bulk density, C_s is the amount of solute sorbed (which is 0 for nitrate), the subscript rxn describes the chemical transformation.

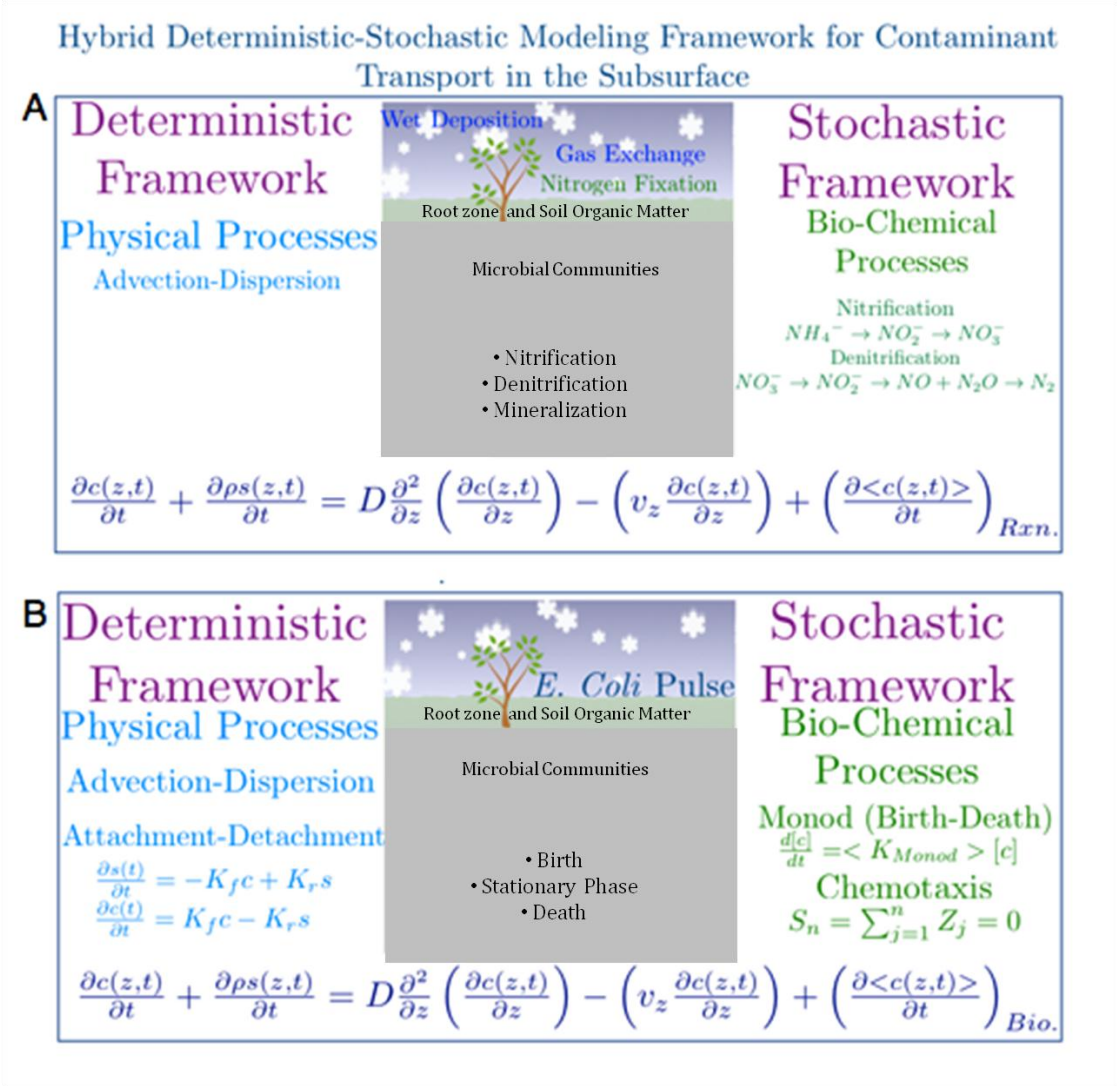

Figure 6.1: A conceptual framework showing some key processes for (A) nitrate transport and (B) *E. coli* transport and deposition behavior in the subsurface.

In addition to the advective-dispersive processes, the chemical form of N is also critical. N exists in many chemical forms in the soil. For example, nitrate (NO_3^-) is highly soluble and very mobile in water. On the other hand, ammonium (NH_4^+) and organic forms of N (Organic Nitrogen, ON) are sorbed by the soil and not readily transported by flowing water. N exists in many chemical forms in the soil. Moreover,

chemical transformations of N are a common occurrence, mainly mediated by a broad gamut of microorganisms that extensively inhabit near-surface soils. These microorganisms can potentially transform N via more than one pathway, and under various conditions of temperature, pH, water content, substrate, as well as electron acceptor and inhibitor concentrations in the subsurface environment [*Wrage et al.*, 2001; *Shrestha et al.*, 2002].

Mineralization is the microbial conversion of organic N to mineral N in the crop root zone (Equation 6.3). The mineralization process is represented by the first order kinetics equation, which can be expressed as [*Hutson and Wagenet*, 1991]:

$$\mathrm{ON} \rightarrow \mathrm{NH_4^+} \qquad \text{(Weeks to years)} \tag{6.3}$$

$$\frac{\mathrm{d[ON]}}{\mathrm{dt}} = -\mathrm{k_{mi}[ON]} \tag{6.4}$$

where *ON* represents the concentration of organic nitrogen in the soil, and k_{mi} is first-order rate constant for mineralization.

NH_4^+ produced in the root zone is oxidized to nitrate in the presence of oxygen (Equation 6.5), known as nitrification. This process is called nitrification. The rate of nitrification is expressed as [*Parton et al.*, 1996]:

$$\mathrm{NH_4^+} \rightarrow \mathrm{NO_2^-} \rightarrow \mathrm{NO_3^-} \qquad \text{(Days)} \tag{6.5}$$

$$\mathrm{R_N} = \frac{\mathrm{d[NO_3^-]}}{\mathrm{dt}} = \mathrm{Net_{min}} \times \mathrm{K_1} + \mathrm{K_{max}} \times \mathrm{NH_4} \times \mathrm{F(T)} \times \mathrm{F}(\boldsymbol{\theta}) \times \mathrm{F(pH)} \tag{6.6}$$

where R_N is the rate of nitrification [$ML^{-3}T^{-1}$], Net is the daily net N mineralization from the organic nitrogen, K_1 is the fraction of net nitrogen that is assumed to be nitrified each day, NH_4^+ is the soil ammonium concentration [ML^{-3}], K_{max} is the maximum fraction of NH_4^+ nitrified, F(T) is the effect of soil temperature on nitrification, F(θ) is the effect of

soil water content and soil texture on nitrification, and F(pH) is the effect of soil pH on nitrification.

There are several processes with which nitrate can be removed from the subsurface environment. Denitrification is the biological reduction of nitrate to nitrogen (Equation 6.7). The rate of denitrification is expressed as [Del Grosso et al., 2000]:

$$NO_3^- \rightarrow NO_2 \rightarrow N_2O \uparrow \rightarrow N_2 \uparrow \qquad \text{(Days)} \qquad (6.7)$$

$$R_D = \min[F_d(NO_3), F_d(CO_2)]F_d(\boldsymbol{\theta}) \qquad (6.8)$$

where R_D is the rate of denitrification [$ML^{-3}T^{-1}$], and $F_d(CO_2)$ **a**nd $F_d(NO_3)$ are general functions relating N to soil respiration and nitrate levels. This is modified by $F_d(\boldsymbol{\theta})$, a dimensionless multiplier based on water content status [*Del Grosso et al.*, 2000].

In this study, we consider only nitrification processes (for validation). For completeness, we have provided a summary of the key processes that govern nitrate transport in the subsurface. Nitrification and denitrification are potentially affected by CO_2, temperature, and N concentration through a wide variety of complex, interacting mechanisms. Some of factors have a direct impact (e.g., N addition increases substrate availability for both processes), and others have an indirect influence. For instance, nitrification is aerobic and denitrification is anaerobic. Therefore, O_2 concentrations play a key role in controlling these processes [*Barnard et al.*, 2005].

Mineralization and immobilization are N-transformation processes that occur simultaneously in the soil. Both NH_4^+ and nitrate can be immobilized, i.e. taken up by the soil microorganisms (Equation 6.9). This immobilization is a very rapid process and is also considered to follow first order kinetics equation [*Murphy et al.*, 2003]:

$$NH_4^+ \text{ or } NO_3^- \rightarrow ON \tag{6.9}$$

$$\frac{d[ON]}{dt} = -k_{ON}[ON] \tag{6.10}$$

NH_4^+ produced in the root zone can also be removed via volatilization (Equation 6.11). Volatilization is a first-order process in the surface layer and is given by [*Hutson and Wagenet*, 1991]:

$$NH_4^+ \text{ in alkaline form } \rightarrow NH_3 \uparrow \tag{6.11}$$

$$\frac{d[NH_4^+]}{dt} = -k_{vol}[NH_4^+] \tag{6.12}$$

where NH_4^+ represents the concentration of ammonium salt available in the soil for volatilization, and k_{vol} is the first-order rate constant.

6.3.2 Factors affecting E. coli transport through soils and their mathematical formulations

Many physical, chemical, and biological processes that govern microbial transport in the subsurface have been described elsewhere [*Bradford et al.*, 2006; *Ginn et al.*, 2002; *Murphy and Ginn*, 2000]. Here, we present a brief summary of processes that are relevant for *E. coli* transport in the vadose zone (Figure 6.1B). The physico-chemical processes include advection, diffusion, dispersion, exclusion, straining, and physical filtration; and they are affected by the structure and properties of the porous media. The biological processes include growth/decay, chemotaxis, predation, physiological adaptation (survival), and adhesion or active detachment [*Murphy and Ginn*, 2000].

E. coli transport in the vadose zone is strongly guided by the physico-chemical framework. Microbes undergo convective transport as particulates or dissolved species

moving with the pore water whose velocity is governed by the hydraulic pressure gradient, porosity, and permeability distribution of the porous media:

$$\frac{\partial C_m}{\partial t} = D_L \frac{\partial^2 C_m}{\partial x^2} - v_x \frac{\partial C_m}{\partial x} - \frac{B_d}{\theta} \frac{\partial C_{ms}}{\partial t} + \left(\frac{\partial C_m}{\partial t}\right)_{bio;} \quad (6.13)$$

where C_m is the concentration of microbes in the liquid phase, t is time, D_L is the longitudinal dispersion coefficient, v_x is the pore water velocity, B_d is the bulk density, θ is the water content, C_{ms} is the amount of solute sorbed (which is 0 for nitrate), and subscript *bio* describes the biological transformation.

Physical partitioning of microbes between aqueous (Equation 6.14) and attached (Equation 6.15) phases can be represented by different first- or second-order models. We consider only first order models in this study. For example, a first-order model can be represented as [*Murphy and Ginn*, 2000]:

$$\frac{\partial C}{\partial t} = -K_f C + K_r C_s \quad (6.14)$$

$$\frac{\partial C_s}{\partial t} = K_f C - K_r C_s \quad (6.15)$$

where K_f and K_r are the rates for the attachment of mobile and immobile microbes, respectively.

Biological processes (e.g., growth and decay) are critical for *E. coli* transport in the subsurface and are generally linked to spatial and temporal variations in nutrient flux. These can be described using the Monod reaction kinetics. The Monod equation is given as:

$$\mu = \mu_{max} \frac{S}{K_{1/2} + S} \quad (6.16)$$

where μ is the specific growth rate of the microorganisms, μ_{max} is the maximum specific growth rate of the microorganisms, S is the concentration of the limiting substrate for growth, and $K_{1/2}$ is the "half-velocity constant"—the value of S when μ /μ_{max} =0.5.

E. coli's search for food and escape from harm can be adequately represented by a biased random walk [*Berg*, 2004]. *E. coli* propel themselves with the help of flagella, and this process is called chemotaxis. The presence of a substrate gradient (food) gives rise to the biased random walk. A random walk is defined by the equal probability of *E. coli* movement in all possible directions. A biased random walk will have different probabilities of *E. coli* movement in different directions depending upon the susbtrate gradient. A simple random walk is defined as follows:

$$S_n = \sum_{j=1}^{n} Z_j = 0 \tag{6.17}$$

where Z_j are independent random variables and their expectation is 0.

6.3.3 Stochastic formulation

In this hybrid approach for addressing nitrate and *E. coli* transport in the subsurface, physical processes (e.g., advection, dispersion) are modeled in the deterministic framework, and chemical (e.g., nitrification, denitrification) and biological processes (e.g., population growth, chemotaxis) are modeled as stochastic partial differential equations. For nitrate transport, relevant bio-chemical processes, such as nitrification, mineralization, immobilization, and denitrification, follow first order kinetics. For *E. coli* transport, biological processes, such as population growth, die-off,

also follow first order kinetics. Therefore, we will provide a mathematical formulation of the first order stochastic differential equation.

The first order stochastic differential equation is of the form:

$$\frac{dX}{dt} = b(t, X_t) + \sigma(t, X_t).W_t \tag{6.18}$$

where b and σ are given functions, t is time, X is the process (e.g., nitrification), and W is the noise. It is assumed that the noise is independent and stationary with zero expectation. Rewriting this equation in a discrete form:

$$X_{K+1} - X_K = b(t_K, X_K)\Delta t_K + \sigma(t_K, X_K)W_K\Delta t_K \tag{6.19}$$

This can be rewritten as [*Knight*, 1981]:

$$X_K = X_0 + \sum_{j=0}^{k-1} b\left(t_j, X_j\right)\Delta t_j + \sum_{j=0}^{k-1} \sigma\left(t_j, X_j\right) B_j \tag{6.20}$$

where B is a Brownian motion approximation of the noise term. It is possible to prove that the limit of the right hand side of Equation 6.20 exists, and therefore this equation can be rewritten in an integral form:

$$X_t = X_0 + \int_0^t b(s, X_s)ds + \int_0^t \sigma(s, X_s)dB_s \tag{6.21}$$

Other processes that do not follow the first order kinetics are biological processes such as chemotaxis. Chemotaxis can be modeled as a biased random walk [*Berg*, 2004] as described in Equation 6.17.

The Richard's equation and advection-dispersion equations (Equations 6.1, 6.2, and 6.13) are solved using the HYDRUS model [*Simunek et al*., 2006]. After each time step the concentrations (nitrate or *E. coli*) are updated for chemical and biological processes. This is done for each ensemble at every time step. In most applications of the stochastic framework, a typical ensemble size is around 100 members [*Evensen*, 2009].

Therefore, an ensemble of 100 members is generated in this study using parameter distributions. Uniform distributions are employed to describe all stochastic parameters (biological and chemical) because little prior information is available for these parameters in the literature. The range for each parameter is chosen to cover all realistic values. The generated ensemble of parameter fields (e.g., nitrification rate, growth rate for *E. coli* transport) is then used in the conceptual transport model to generate 100 realizations of ammonium, nitrite, and nitrate concentrations for nitrate transport; and *E. coli* concentrations for *E. coli* transport. The predicted concentrations were then compared with the data using the root mean square error (RMSE) criteria:

$$RMSE = \sqrt{E((\hat{\theta} - \theta)^2)} \tag{6.22}$$

where $\boldsymbol{E}$ is the expectation of the difference of the predicted concentrations ($\hat{\theta}$) against the reference concentrations (θ) (analytical solutions explained in section 6.3.4).

6.3.4 Model testing

A synthetic one-dimensional steady state column is used to verify the accuracy of the conceptual framework for nitrate transport. Numerical results are compared with an analytical solution for advective dispersive transport of nitrate [*van Genuchten*, 1985]. Subsequently, we compare the analytical solution with the deterministic numerical solution, where parameters are estimated using the concentrations (ammonium, nitrite, and nitrate) from the analytical solution. The analytical solution for solute transport assumes a homogeneous, isotropic porous medium for the one-dimensional steady state column, and only considers nitrification process from NH_4^+. It is assumed that there are

aerobic conditions in the column for nitrification process to occur. Ammonium is applied at the start of the experiment, and the experiment is simulated for 200 days. The soil column is 200 cm long. For the numerical study, a spatial discretization of 1 cm is adopted. An initial time step of 1 day and minimum and maximum time steps of 0.001 and 100 day are employed. The results are described with respect to observation nodes located at 20, 40, 80, and 120 cm from the top of the soil column (Table 6.1). The conceptual model is governed by the advection dispersion equation for solute transport (Equations 6.1 and 6.2), and the nitrification process is described using first order kinetics (Equations 6.5 and 6.6). Initial input parameters for nitrate transport are provided in Table 6.2.

Similarly, one-dimensional column experiments are conducted to investigate the transport and deposition behavior of *E. coli* involving attachment-detachment mechanism. The soil column is 200 cm long. The experiment is simulated for 700 days as *E. coli* transport occurs at a slower rate than nitrate. It is assumed that there is enough soil organic matter for *E. coli* survival, and there is no inhibition for *E. coli* growth. For the numerical study, the observation nodes as well as the spatial and temporal discretization are the same as applied to the nitrate transport experiment. The conceptual model for *E. coli* transport is governed by the attachment-detachment mechanism (Equations 6.14 and 6.15) other than advection and dispersion, and growth is described using the Monod model (Equation 6.16). Initial input parameters for *E. coli* transport are provided in Table 6.3.

Table 6.1: Description of the one-dimensional soil column.

Feature	Depth (from top in cm)
Column Depth	200
Observation Node 1	20
Observation Node 2	40
Observation Node 3	80
Observation Node 4	120
Boundary Conditions	Top: $q = -K\frac{\partial h}{\partial z}=q_0$ $\frac{\partial c}{\partial t} = 0$
	Bottom: if h=0; $q = -K\frac{\partial h}{\partial z}$ Else $q = 0$

Table 6.2: Initial input parameters for nitrate transport [*van Genuchten*, 1985].

Parameter	Value
Average pore water velocity (cm/hour)	1.0
Dispersion coefficient (cm^2/hour)	0.18
First order degradation coefficient for NH_4^+ (-/sec)	4.6×10^{-8} to 6.9×10^{-8}
First order degradation coefficient for NO_2^- (-/sec)	9.2×10^{-6} to 1.4×10^{-8}
First order degradation coefficient for NO_3^- (-/sec)	0.0

Table 6.3: Initial input parameters for *E. coli* transport [*Bradford et al.*, 2006].

Parameter	Value
Average pore water velocity (m/sec)	1.15×10^{-7}
Dispersion coefficient (m^2/sec)	2.1×10^{-10}
Attachment coefficient (-/sec)	6.1×10^{-5} to 9.1×10^{-5}
Detachment coefficient (-/sec)	4.1×10^{-7} to 6.3×10^{-7}

6.4 Results and discussion

The results presented here provide insight into the different physical, chemical, and biological processes that are significant in the modeling of nitrate and *E. coli* transport in the subsurface. First, we test the conceptual framework of nitrate and *E. coli* transport at the column scale independently. Then, we perturb some system properties to develop a better understanding of nitrate and *E. coli* transport in the subsurface. The significance of this modeling framework is that it provides a novel way of integrating physical and bio-chemical processes, and enables us to investigate the importance of individual processes and their mutual interactions.

For nitrate transport, the one-dimensional column involves the application of a solute pulse (ammonium from fertilizers) resulting in the chemical transformation of ammonium to nitrite and then to nitrate. Nitrate transport is also analyzed under a perturbed system by changing soil temperature and pH. Advection dispersion (Equations 6.1 and 6.2) are modeled as deterministic processes, and nitrification (Equations 6.5 and 6.6) as a stochastic differential equation in the conceptual framework. The reason for modeling nitrification in a stochastic framework is that nitrification is a microbial mediated process and not possible to parameterize in a column for an extended period of time.

Similarly, for *E. coli* transport, the one-dimensional column involves the application of a solute pulse (*E. coli* pulse from waste water) resulting in *E. coli* deposition. *E. coli* transport is also analyzed under a perturbed system by changing the growth and die-off rates of *E. coli* and imposing chemotaxis in *E. coli*. Advection

dispersion (Equations 6.1 and 6.13) are modeled as deterministic processes, and attachment-detachment mechanism (microbial processes) (Equations 6.14 and 6.15) and growth (Equation 6.16) as stochastic differential equations in the conceptual framework.

6.4.1 Nitrate transport in one-dimensional soil column

Chemical transformation of ammonium in soils takes place primarily through nitrification. Nitrification is an aerobic process carried out primarily by autotrophic bacteria [*Anderson et al.*, 1993; *Parton et al.*, 1996; *Schimel et al.*, 1989]. In the nitrification process, part of ammonium first converts into nitrite, which is a daughter product of ammonia. Then, part of nitrite converts into nitrate, which is a daughter product of nitrite. The net result of the complete pathway is oxidation of ammonium to nitrate.

In the column study, ammonium solute pulse was applied for the entire duration of the experiment. Figures 6.2 and 6.3 demonstrate ammonium, nitrite, and nitrate concentration at different observation nodes (1-4) corresponding to depths 20, 40, 80, and 120 cm (Table 6.1). However, Figure 6.2 presents a comparison of the analytical solution of ammonium, nitrite, and nitrate at different observation depths with the deterministic numerical solution.

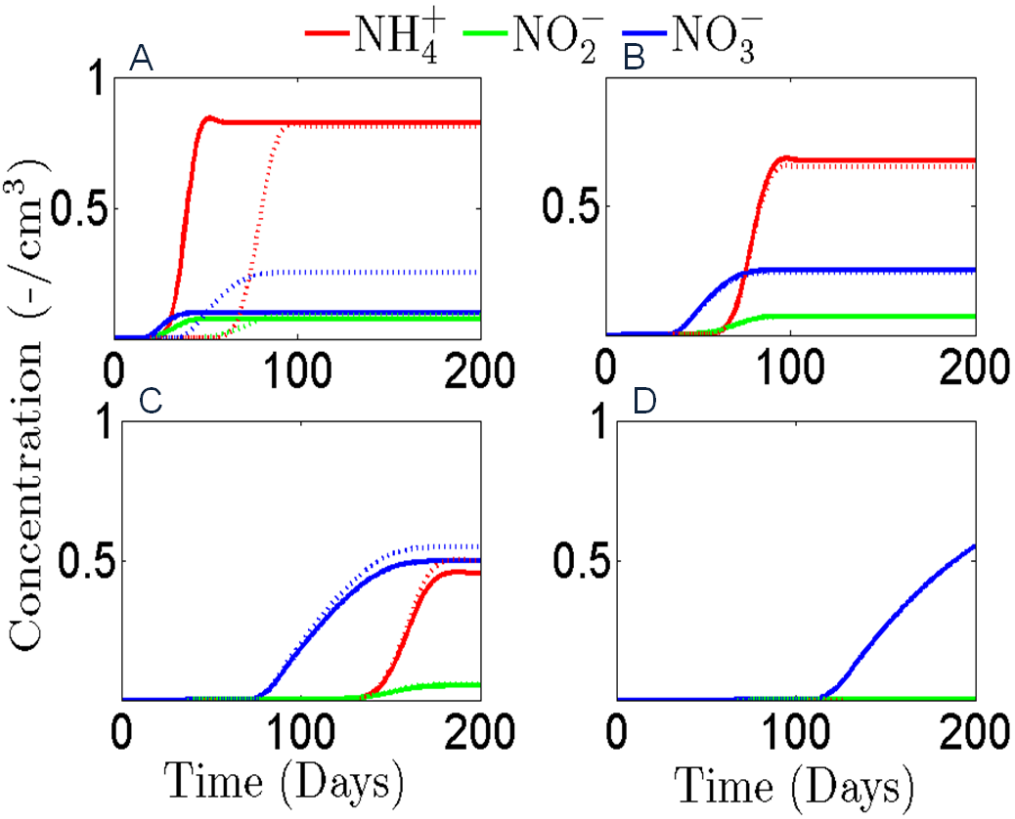

Figure 6.2: Nitrate transport with three-species nitrification chain is shown at (A) observation node 1, (B) observation node 2, (C) observation node 3, and (D) observation node 4. Y-axis represents the breakthrough concentration (C/C_0). Solid lines represent the analytical solution, and dotted lines represent the deterministic numerical solution.

Figure 6.3 presents a comparison of the analytical solution with the hybrid-stochastic-deterministic solution. In Figure 6.3, 100 realizations of ammonium, nitrite, and nitrate are plotted, and the vertical spread of these realizations represents the uncertainty in the prediction of these concentrations.

It is clearly visible from Figures 6.2 and 6.3 that the hybrid-stochastic-deterministic solution is able to reproduce the analytical solution, whereas the

deterministic numerical solution shows some mismatch at observation node 1. In other words, the proposed methodology offers an improvement.

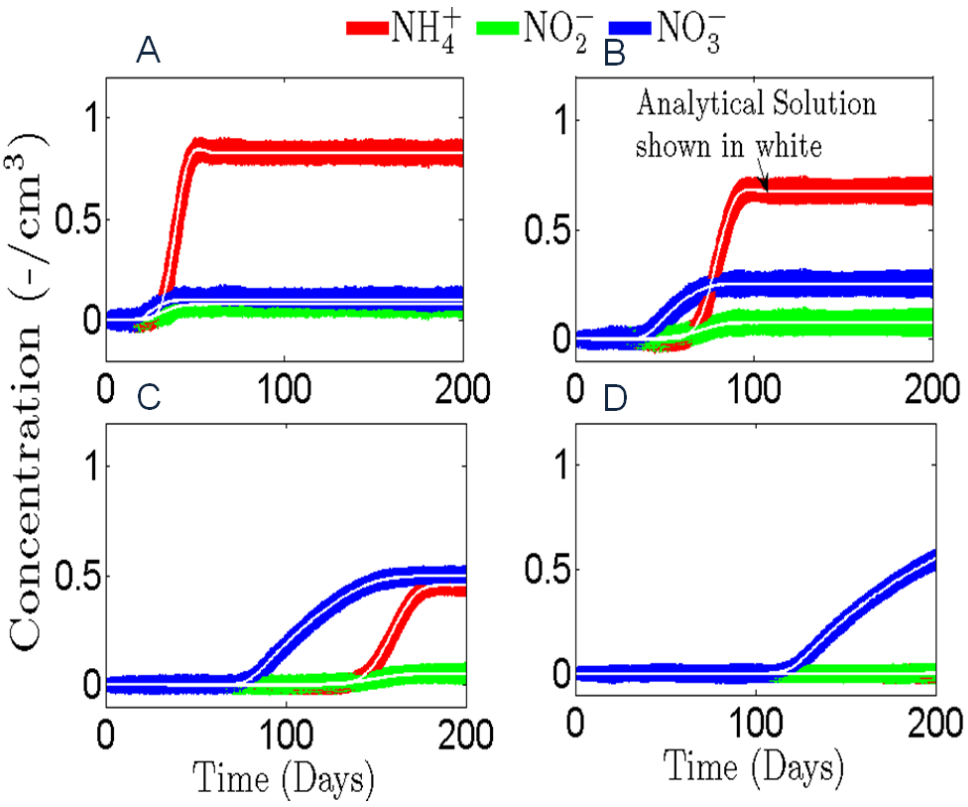

Figure 6.3: Nitrate transport with three-species nitrification chain is shown at (A) observation node 1, (B) observation node 2, (C) observation node 3, and (D) observation node 4. Y-axis represents the breakthrough concentration (C/C_0). Solid white lines represent the analytical solution, and other lines represent the hybrid deterministic-stochastic numerical solution.

It is evident from Figure 6.3A that it takes almost 20 days for NH_4^+ to reach the breakthrough at observation node 1. The chemical transformation of ammonia to nitrite occurs after a small time interval and therefore, nitrite is discernible at this observation node after 1 day. The transformation to nitrate is almost immediate. Observation node 2 (Figure 6.3B) shows a similar behavior as well. However, observation nodes 3 and 4 behave differently. Ammonium reaches observation node 3 after 140 days, but nitrate reaches observation node 3 after 80 days. This is because transport of NO_3^- is faster as compared to NH_4^+. Ammonium also gets adsorbed in soil and therefore, it is not observed at node 4 (Figure 6.3D) in the simulated time frame. It is clear from Figure 6.3 that the predictions of ammonium, nitrite, and nitrate by the conceptual framework are in agreement with the analytical solution (white solid line). Moreover, the conceptual model provides a broader view of the integrated system behavior as it simulates bio-chemical processes in a stochastic framework.

This uncertainty is further quantified by plotting the probability distributions (PDFs) of the root mean square error (RMSE). Figure 6.4 depicts RMSE distributions of predicted ammonium, nitrite, and nitrate concentrations at observation node 4 against the analytical solution.

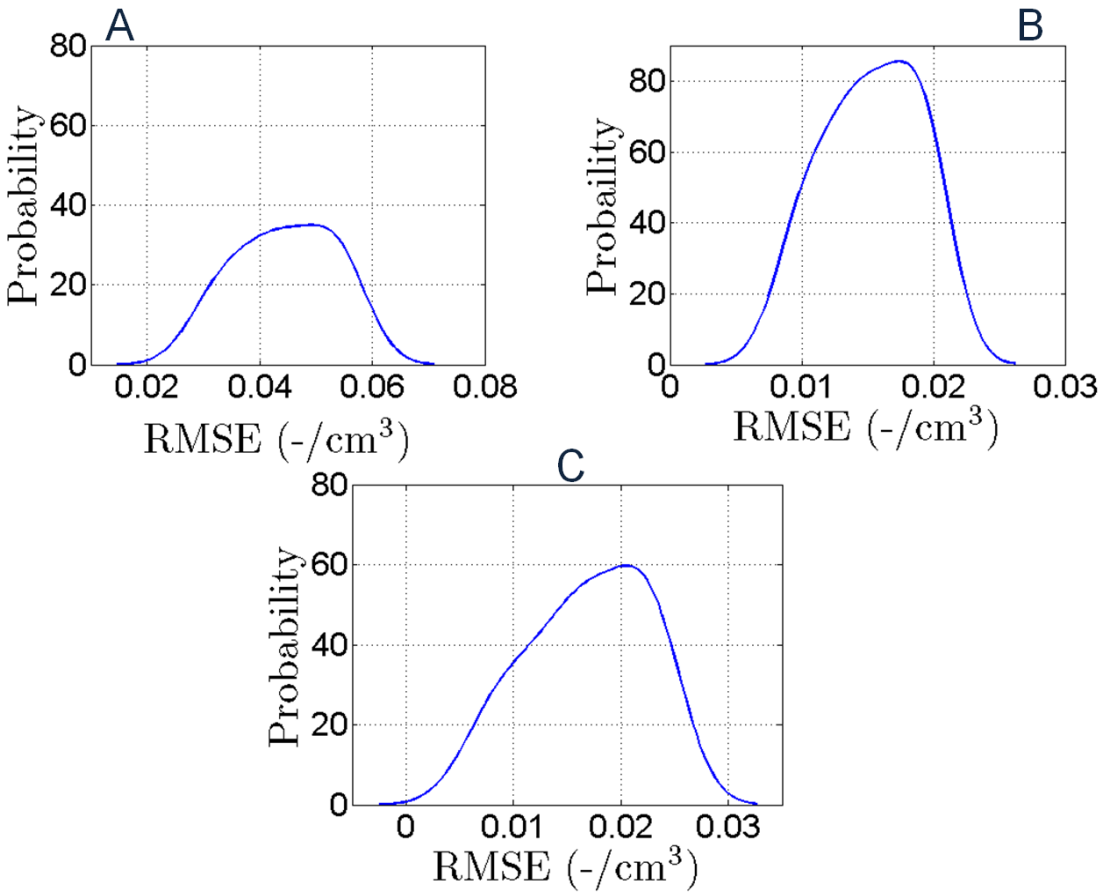

Figure 6.4: The probability distribution functions (PDFs) of the root mean square error (RMSE) of predicted (A) ammonium, (B) nitrite, and (C) nitrate concentrations at observation node 4 against the analytical solutions. The PDFs demonstrate the uncertainty in the predictions.

It is evident from Figure 6.4 that the maximum spread of the PDF is for ammonium, followed by nitrite, and then, by nitrate. Therefore, the maximum uncertainty is associated with predictions of ammonium concentrations in the subsurface, followed by nitrite and nitrate, respectively. Apart from advective dispersive transport, ammonium concentrations in the soil are affected by adsorption as well as oxidation to nitrite (nitrification).

Ammonium concentrations are influenced by multiple processes and therefore, this may be the reason for its higher RMSE values (0.01 to 0.07) as compared to nitrite (0.005 to 0.02) and nitrate (0 to 0.03). Nitrite shows lesser uncertainty than nitrate. This may be because the concentration of nitrite is smaller and nitrite is short lived in the system.

The factors which influence nitrification are pH, water content, temperature, and the existing ammonium concentration in the soil profile as given in Equation 6.6. Nitrification follows first order kinetics and is therefore dependent on the soil ammonium concentration. Among the factors, we perturbed soil temperature and pH to understand their effect on nitrate transport.

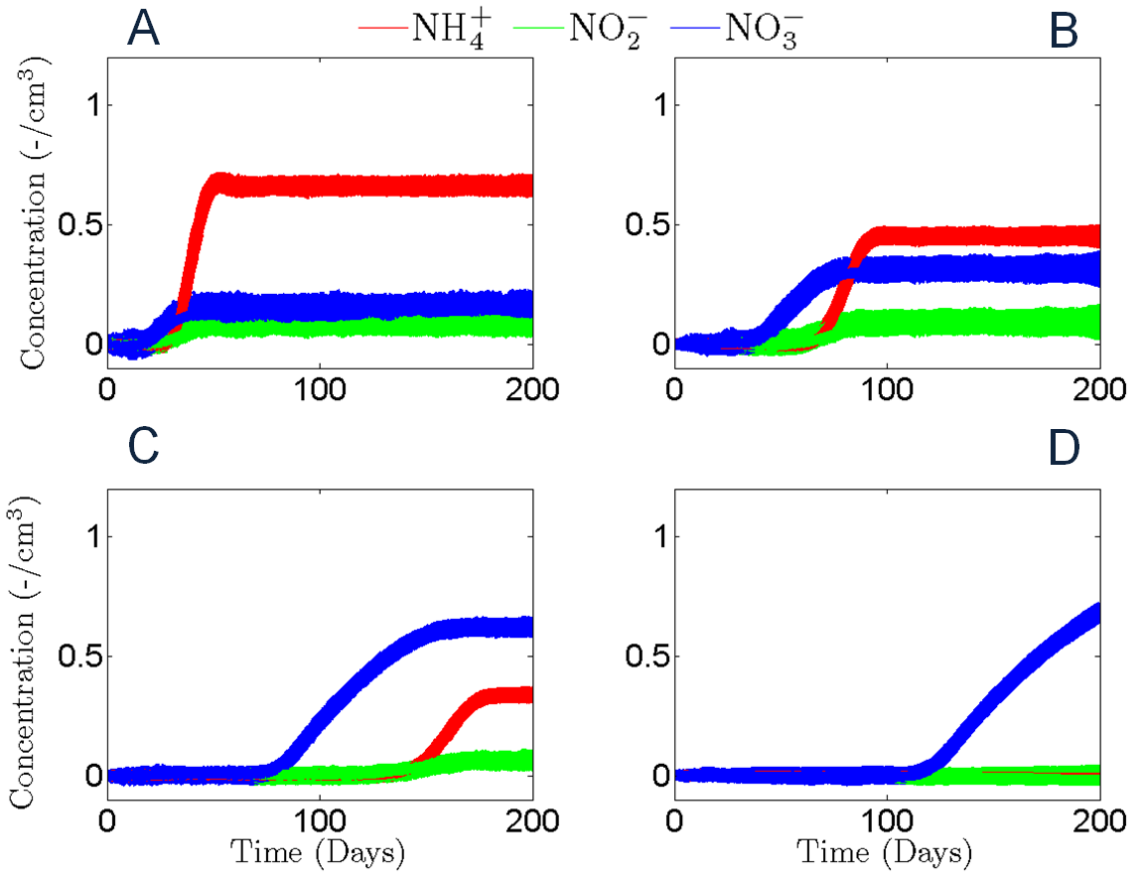

Figure 6.5: Nitrate transport with perturbations to soil temperature at (A) observation node 1, (B) observation node 2, (C) observation node 3, and (D) observation node 4. Y-axis represents the breakthrough concentration (C/C_0).

6.4.2 Nitrate transport with perturbations to soil temperature

Figure 6.5 demonstrates one-dimensional nitrate transport in the subsurface with perturbed soil temperature. In soils, temperature profile varies quickly (5 ºC within 20 cm of depth and then it varies gradually) [*Raju et al.*, 1995], and may have a significant influence on nitrate transport. There is evidence in the literature that nitrification rates increase with increasing soil temperature. This increase is exponential [*Parton et al.*, 1996]. Therefore, nitrification rates vary along the soil profile.

For this study, soil temperature was varied with depth (starting at 30 ºC at the surface, 5 ºC variations till 20 cm depth, and 0.5 ºC variations from 20 cm till the length of the column) and follow temperature profiles described in *Parton et al.* [1996]. As expected, nitrification rate changes along the soil profile; it decreases with depth quickly (up to 20 cm i.e., observation node 1) and then decreases gradually. It is evident from Figure 6.5A that observation node 1 has less ammonium concentrations as compared to the unperturbed system (Figure 6.3A). This is because more ammonium gets nitrified at the top due to a higher nitrification rate. The ammonium concentration is also lower at observation nodes 2 and 3 but nitrite and nitrate concentrations are slightly higher as compared to the unperturbed system. The reason for higher nitrate concentration at observation node 2 is the higher nitrification rate associated with the soil temperature profile. However, the higher concentrations of nitrate at deeper observation nodes (such as node 3) are associated with high mobility of NO_3^- in water. These results suggest that soil temperature variations cause nitrification rates to vary along the soil profile and consecutively, nitrate concentrations arrive earlier at greater depths, and ammonium concentrations are smaller along the soil profile.

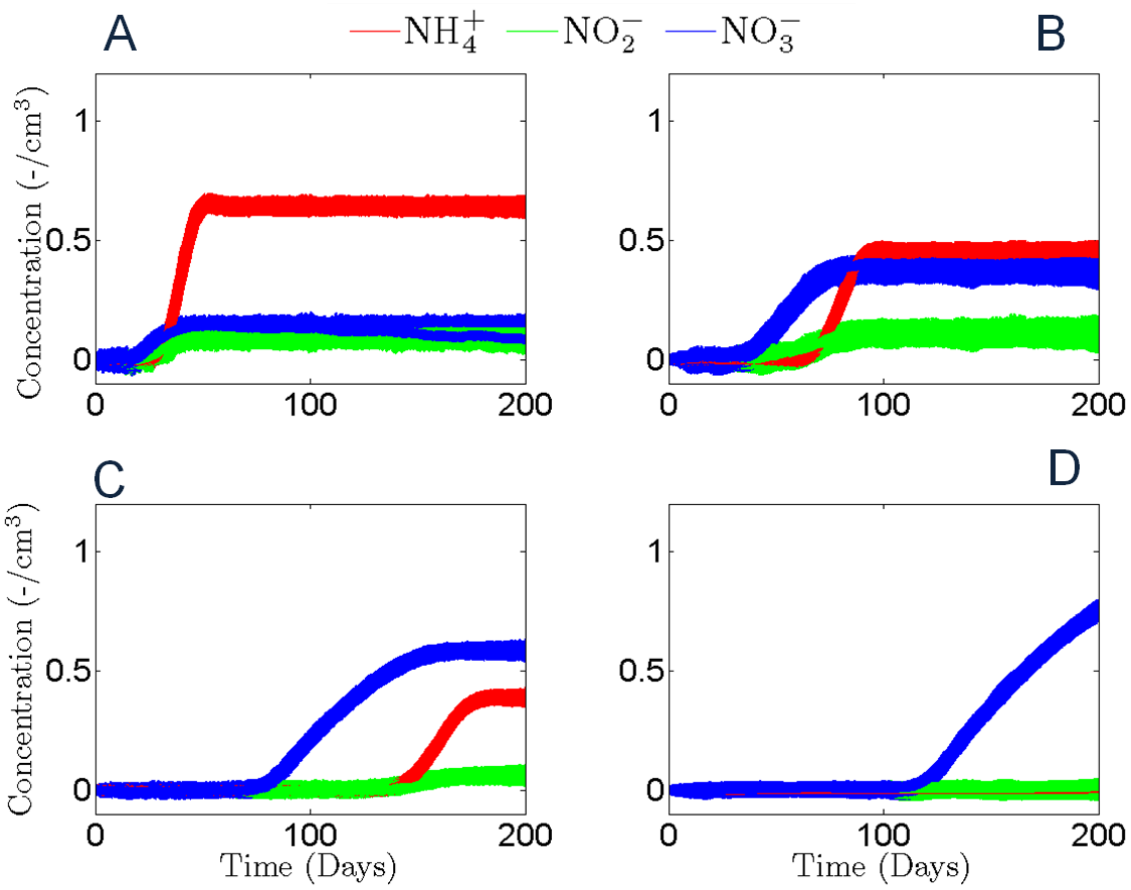

Figure 6.6: Nitrate transport with perturbations to soil pH at (A) observation node 1, (B) observation node 2, (C) observation node 3, and (D) observation node 4. Y-axis represents the breakthrough concentration (C/C_0).

6.4.3 Nitrate transport with perturbations to soil pH

Figure 6.6 demonstrates one-dimensional nitrate transport in the subsurface with perturbed soil pH. The majority of soils are known to be alkaline with pH ranging from 7.0 to 7.8 and above. Soil pH vertical profile remains in a range of pH from 6 to 7, and it may have a significant influence on nitrate transport. There is evidence in the literature that nitrification rates increase with increasing soil pH (say from 6 to 7). This increase in nitrification rate can be described as inverse tangent function (arc tan) of soil pH [*Parton*

et al., 1996]. Nitrification rate is the highest under alkaline soil pH conditions [*Pang et al.*, 1975].

For this study, soil pH was slightly depressed (5 to 7) and follows the trend of temperature profiles (as described above). Further details are available in *Parton et al.* [1996]. As expected, nitrification rate is slightly lower along the soil profile as compared to the unperturbed system. It is clear from Figure 6.6D that observation node 4 has some discernible amount of ammonium as compared to the unperturbed system (Figure 6.3D). This is because less ammonium gets nitrified at these depths (around 120 cm) due to a lower nitrification rate and therefore, it is transported to deeper depths in the soil column. The ammonium concentration is lower at observation nodes 1-3, but nitrite and nitrate concentrations are slightly higher as compared to the unperturbed system. Nitrate concentrations are significantly higher at observation node 2. These results suggest that nitrification rates are higher at the top of the soil column (20-80 cm) with lower ammonia concentrations and higher nitrate concentrations. Therefore, it is evident that soil pH variations cause nitrification rates to vary along the soil profile and consecutively, ammonium concentrations reach deeper in the column.

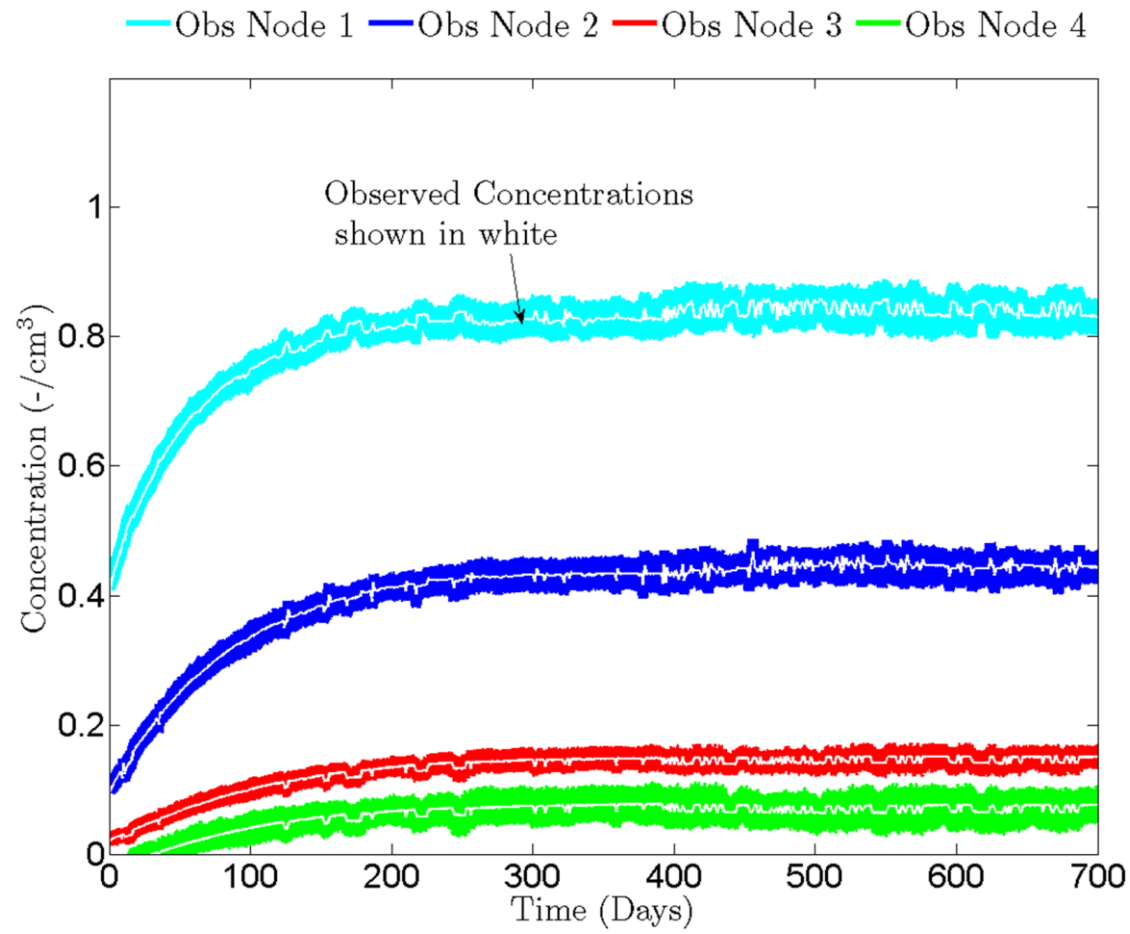

Figure 6.7: Transport and deposition behavior of *E. coli* with attachment-detachment model is shown at observation nodes 1, 2, 3 and 4. Y-axis represents the breakthrough concentration (C/C_0).

6.4.4 E. coli *transport in one-dimensional soil column*

Similarly, we present our findings of the *E. coli* transport and response to perturbations in the one-dimensional soil column in the following sections. We investigated *E. coli* transport using the conceptual framework that incorporates the various physical, chemical, and biological processes. *E. coli* transport takes place primarily through advection. We used an attachment-detachment model to describe *E. coli* movement in the subsurface because *E. coli* can reside in the porous media as monodispersed and aggregated species [*Šimůnek et al.*, 2006].

In the column study, wastewater pulse was applied for the entire duration of the experiment. A typical concentration of *E. coli* in raw municipal waste is of the order of 10^6 CFU/100 mL [*Pang et al.*, 2004]. Figure 6.7 describes transport and deposition behavior of *E. coli* in the soil column at different observation nodes (1-4) as shown in Table 6.1. It is evident from Figure 6.7 that the breakthrough in the deposition of *E. coli* occurs almost at the same time in all observation nodes. However, *E. coli* concentrations are in decreasing order from observation node 1 to observation node 4.

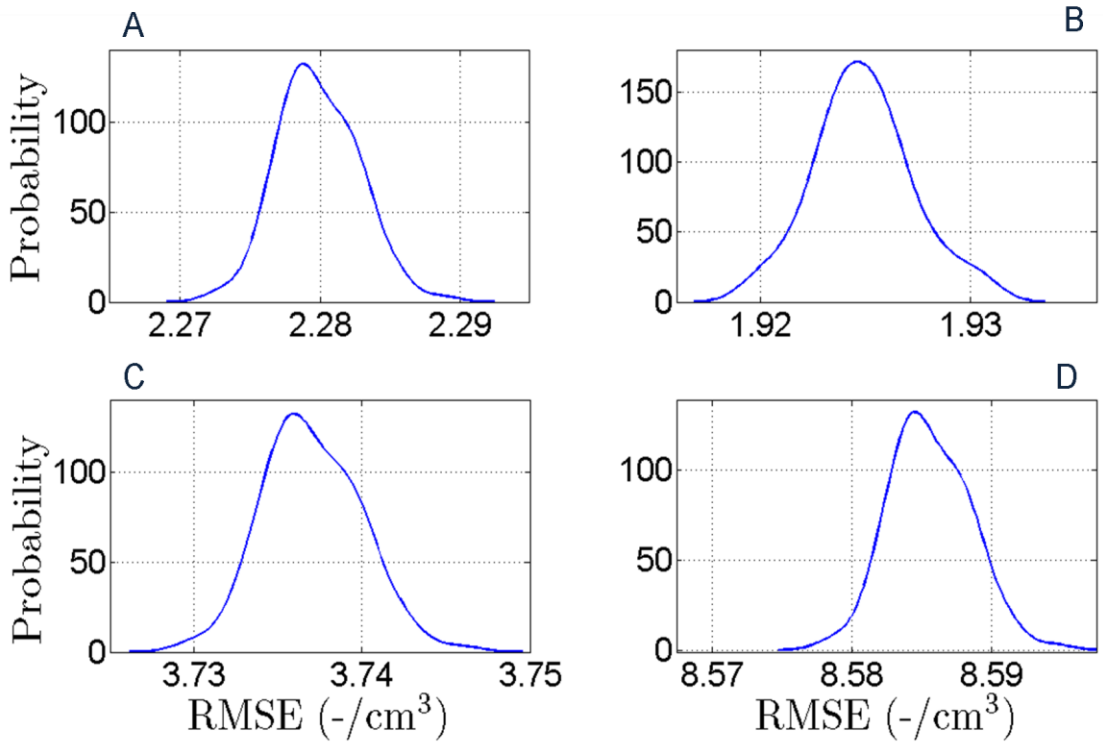

Figure 6.8: The probability distribution functions (PDFs) of the root mean square error (RMSE) of predicted *E. coli* concentrations at A) observation node 1, (B) observation node 2, (C) observation node 3, and (D) observation node 4 against the numerical solution. The PDFs demonstrate the uncertainty in the predictions.

The presence of *E. coli* at all nodes suggests that *E. coli* moves faster in the column due to advective transport while deposition occurs at a slower rate. It is also evident from Figure 6.7 that the majority of *E. coli* deposition occurs near the surface (observation node 1). Drawing a comparison with nitrate transport (Figure 6.3), *E. coli* concentrations are much lower at deeper depths. The reason is that *E. coli* is attached to the soil matrix whereas nitrate is not adsorbed in the soil matrix. Figure 6.7 further indicates that predictions of *E. coli* deposition by the conceptual framework are in agreement with the numerical solution (white solid line). In order to explore the

uncertainty in the estimation of *E. coli* concentrations, 100 realizations of *E. coli* depositions are plotted in Figure 6.7.

This uncertainty is further quantified by plotting the probability distributions (PDFs) of the root mean square error (RMSE). Figure 6.8 demonstrates the root mean square error (RMSE) distributions of the predicted *E. coli* concentrations (at observation nodes 1-4) against the numerical solution. It is evident from Figure 6.8 that the range of RMSEs is approximately similar for all four observation nodes. This suggests that the uncertainty does not change with depth in *E. coli* transport. However, the conceptual modeling framework is able to predict near surface *E. coli* deposition with lesser RMSE values than the deeper *E. coli* deposition. For example, observation nodes 1 and 2 have RMSE values ranging from 2.27 to 2.29 and 1.92 to 1.93, respectively; observation nodes 3 and 4 have RMSE values ranging from 3.73 to 3.75 and 8.57 to 8.59, respectively. These results suggest that the complexity in the behavior of *E. coli* in the subsurface increases with depth.

The factors that influence *E. coli* deposition and transport are growth, die-off, and random motility [*Berg*, 2004], as described in Equations 6.16 and 6.17. Therefore, we perturbed growth and die-off rates as well as random motility to investigate their effect on *E. coli* deposition and transport.

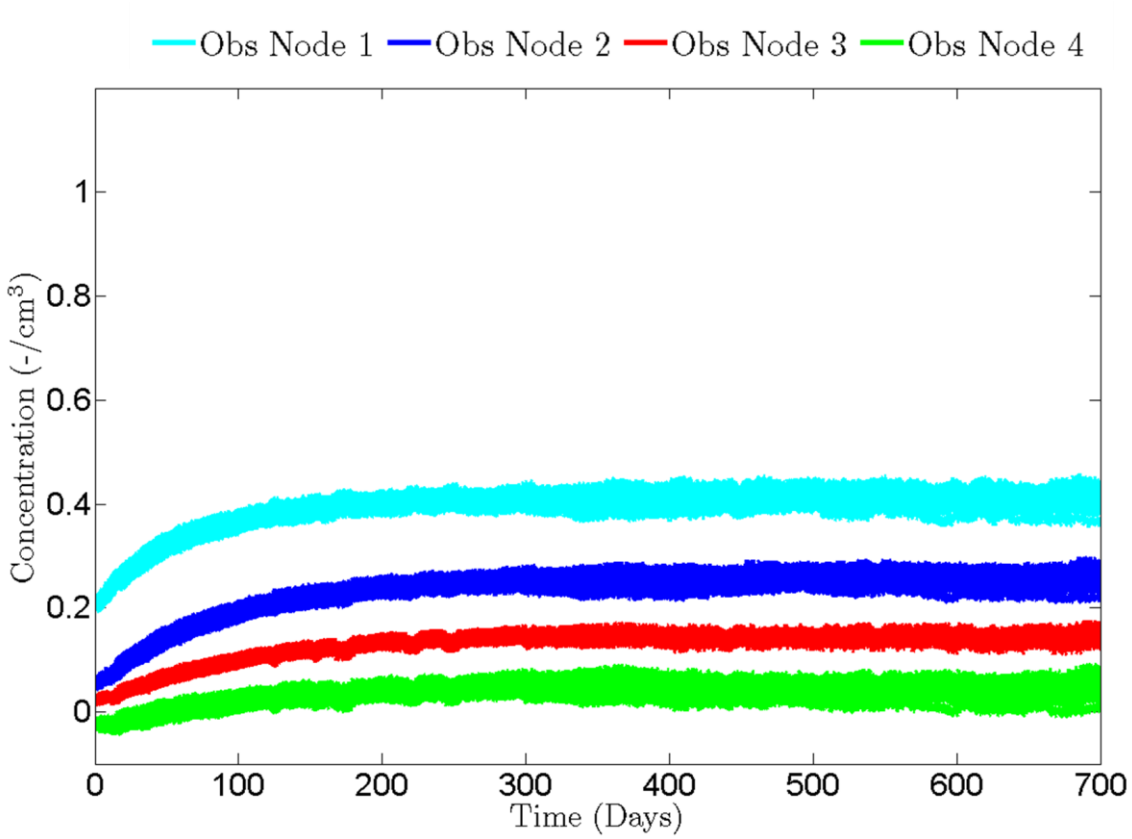

Figure 6.9: Transport and deposition behavior of *E. coli* with perturbations to growth and die-off rates is shown at observation nodes 1, 2, 3 and 4. Y-axis represents the breakthrough concentration (C/C_0).

6.4.5 E. coli *transport with perturbations to growth and die-off rates*

Figure 6.9 demonstrates one-dimensional *E. coli* transport in the subsurface with perturbed growth and die-off rates. *E. coli* population has four growth stages: lag phase, exponential growth, stationary phase, and logarithmic decay phase [*Berg*, 2004]. If any system has enough supply of substrate, *E. coli* population will manifest itself in the stationary phase.

For this study, we increased growth and die-off rates by 20% to investigate the deposition behavior of *E. coli* in the subsurface. It is evident from Figure 6.9 that

observation nodes (1-3) reach breakthrough earlier as compared to the unperturbed system (Figure 6.7). In addition, there is no discernible deposition of *E. coli* at observation node 4. This is because *E. coli* reaches the stationary phase faster and therefore, the breakthrough is also quicker. Moreover, the life span of *E. coli* is shorter with increased growth and die-off rates, which prohibits *E. coli* to reach deeper in the soil profile.

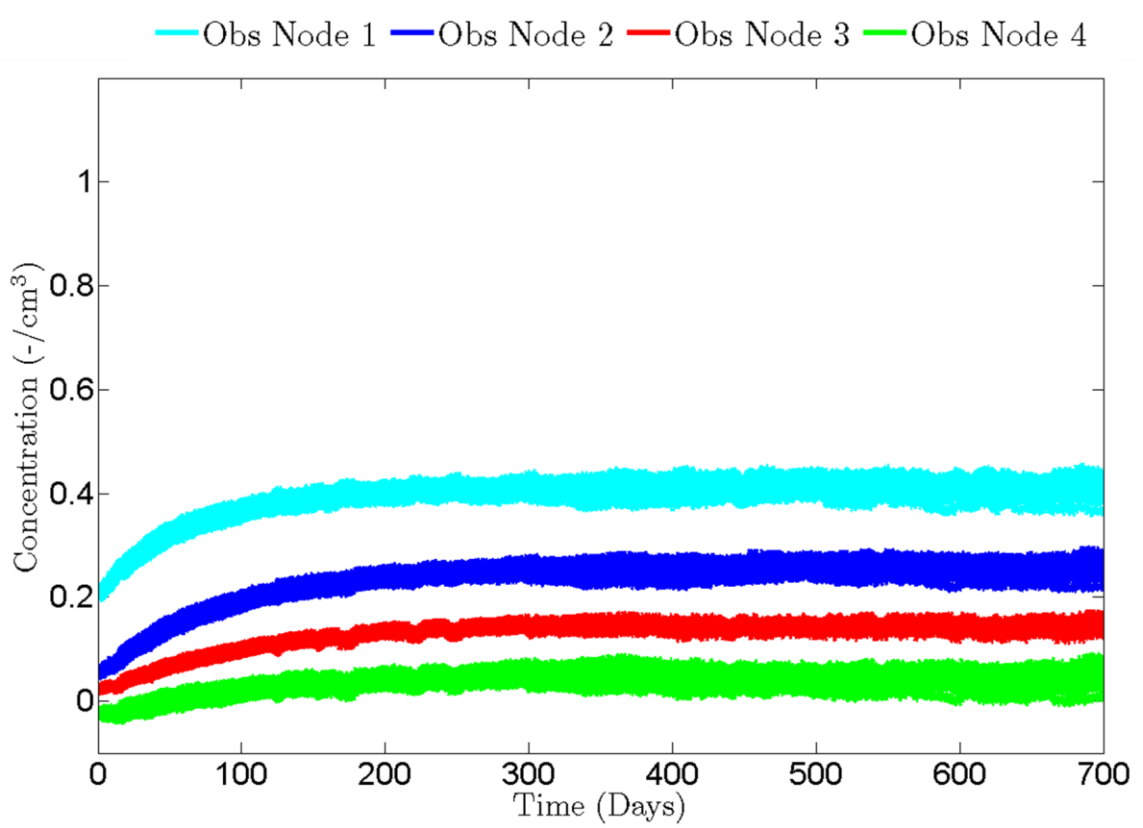

Figure 6.10: Transport and deposition behavior of *E. coli* with perturbations to chemotaxis is shown at observation nodes 1, 2, 3 and 4. Y-axis represents the breakthrough concentration (C/C_0).

6.4.6 E. coli *transport with perturbations to chemotaxis*

Figure 6.9 demonstrates one-dimensional *E. coli* transport in the subsurface with perturbations to chemotaxis. Chemotaxis is the directed movement of *E. coli* towards or away from a chemical gradient. The presence of hot spots, such as food supply gradient, adds bias to the random walk [*Macnab and Koshland*, 1972].

In this study, we introduced chemotaxis in the behavior of *E. coli* that was compared with the unperturbed system (no chemotaxis). As suggested earlier, the motility of *E. coli* in the subsurface can be modeled as a random walk [*Berg*, 2004]. In this situation, *E. coli* will have equal probabilities to go from one point to another in space. In the one-dimensional column, there are two directions (up and down) for the movement of *E. coli*. In our conceptual model, we introduced a biased random walk by allowing *E. coli* to move 70% times in the downward direction and 30% times in the upward direction. The reason for introducing this bias is that as carbon sources (substrates) are exhausted at the top of the soil column, *E. coli* will have a tendency to go down in search of food.

Figure 6.10 indicates that observation nodes (2, 3, and 1) reach breakthrough after a lag as compared to both the unperturbed and the growth and die-off perturbed systems (Figures 6.7 and 6.8). There is higher deposition of *E. coli* at observation nodes 1 and 2 as compared to the unperturbed system (Figure 6.7). However, there is very small deposition of *E. coli* at observation nodes 3 and 4. These results suggest that *E. coli* deposition rates are strongly influenced by chemotaxis. Therefore, an increase in

near surface *E. coli* deposition is observed, and consequently, *E. coli* concentrations are not observed at greater depths in the soil column.

6.5 Summary and conclusions

A novel conceptual framework, which provides an entirely new way of evaluating the importance of individual processes, was developed for modeling the fate and transport of contaminants in the subsurface. This modeling framework is particularly useful when biological and chemical processes with significant uncertainty are involved.

We simulated nitrate transport involving various physical, chemical, and biological processes in a one-dimensional soil column and investigated the impact of perturbations to soil temperature (along the soil profile) and pH on nitrate transport. We showed that the hybrid-stochastic-deterministic model improves the predictive capabilities. The results indicate that there is higher uncertainty associated with predicting ammonium concentrations, followed by nitrite, and then, by nitrate in the subsurface. Variability in soil temperature causes nitrification rates to vary along the soil profile. This causes nitrate to move faster in the subsurface, and decreases ammonium concentrations along the soil profile. Variability in soil pH results in higher nitrate concentrations and lower ammonia concentrations at the top of the soil profile. Thus, perturbations to soil pH also cause nitrification rates to vary along the soil profile, higher rates at the top and lower at the bottom, and consecutively, ammonium concentrations reach deeper in the column.

We also simulated *E. coli* transport involving various physical, chemical, and biological processes in a one-dimensional soil column and investigated the impact of

imposing chemotaxis in *E. coli* and perturbations to growth and die-off rates. The results suggest that the majority of *E. coli* deposition occurs near the surface of the soil profile, and it is associated with the smaller RMSE values. *E. coli* concentrations in soil profile reach the breakthrough faster if both the growth and die-off rates are higher. Furthermore, chemotactic motility of *E. coli* delays the breakthrough of *E. coli* concentrations in the soil profile.

CHAPTER VII

GENERAL CONCLUSIONS

Freshwater resource is a limited supply and is under threat because of various contaminants. The most ubiquitous contaminants of water resource include: microbes (in surface water) and nitrate (in groundwater). For the sustainable management, numerical models of flow and transport are frequently used. Despite advances in numerical modeling, threats from the ubiquitous contaminants are not accurately identified due to structural uncertainty, parameter uncertainty, and scale disparity (observation and process scales).

This dissertation is specifically focused on understanding the linkages between bio-chemical and hydrologic processes, exploring the spatio-temporal variability, quantifying uncertainty, and developing models for both *E. coli* and nitrate transport to better characterize the threats to water resources. Chapters II and III highlight surface water (SW) contamination from microbes. *E. coli* is an indicator organism. For that reason, *E. coli* was used to explore the physical, chemical and biological linkages as precursors to SW contamination from microbes. Chapters IV and V provide a description of spatio-temporal variability of nitrate in groundwater (GW) and a quantification of parameter uncertainty for nitrate transport in GW. Since, surface water and GW interact directly or indirectly through the vadose zone, chapter VI advances our understanding of the subsurface modeling for *E. coli* and nitrate

transport, independently. The major conclusions from this dissertation are summarized below.

Chapter II provides a Bayesian Neural Network (BNN) model for *E. coli* prediction in SW (streams). Results demonstrate six key variables, from a selection of physical, chemical, and biological factors, which influence *E. coli* loads in surface streams. Physical factors include temperature and DO; chemical factors include phosphate and ammonia; and biological factors include SS and chlorophyll.

Results from chapter III demonstrate the impact of flow regime (shallow GW table and unsaturated flow in the vadose zone) and seasonal variability (climatic) on *E. coli* contamination to SW through the subsurface route. The results show slightly increased mobility of *E. coli* in GW during winter time as compared to summer.

Chapter IV depicts that multiple controlling factors exist and dominate the variability of nitrate at different spatial scales (fine, intermediate, and coarse). The entropy based approach provides a physical interpretation of spatial and temporal variability of nitrate in the Trinity and Ogallala aquifers. Results demonstrate that the variability of nitrate is controlled by— pumping and hydraulic conductivity at the fine scale; complex interactions between rivers and the aquifer at the intermediate scale; and lithology and geology at the coarse scale. The trends of nitrate variability manifested long term persistence at the intermediate scale.

Chapter V establishes the Ensemble Kalman Filter (EnKF) as a potential alternative approach for inverse estimation of parameters. The EnKF—integrated with the flow and transport models: MODFLOW and MT3DMS—improves estimation of

nitrate in GW. Parameter conditioning by the EnKF demonstrates that parameter update is a function of the number of assimilated measurements.

In chapter VI, a novel conceptual framework is developed for modeling the fate and transport of contaminants (*E. coli* and nitrate, independently) in the subsurface. This modeling framework is particularly valuable when bio-chemical processes with significant uncertainty are involved. The modeling framework allows us to investigate the impact of perturbations to soil temperature (along the soil profile) and pH on nitrate transport. Results indicate that variability in soil temperature causes nitrification rates to vary along the soil profile. This causes nitrate to move quickly, and ammonium to move slowly along the soil profile. Variability in soil pH results in higher nitrate concentrations and lower ammonia concentrations near the surface.

E. coli transport is explored in the subsurface for perturbed factors such as chemotaxis and growth and die-off rates. The results corroborate the findings of chapter III that the majority of *E. coli* deposition occurs near the surface. *E. coli* concentrations in the soil profile reach the breakthrough (peak concentration) faster if both the growth and die-off rates are higher. Furthermore, chemotactic motility of *E. coli* delays the breakthrough of *E. coli* deposition in the soil profile.

This research develops tools and techniques to adequately address the issues in contaminant hydrology such as, structural uncertainty, parameter uncertainty, and scale disparity, alongside advancing process understanding with regards to fate and transport of contaminants in water resources. Results from this dissertation are valuable in identifying threats to water resources which are critical for decision

makers and environmental managers to design targeted monitoring programs and support real-time decision-making.

REFERENCES

Ahmed, S., and G. Demarsily (1987), Comparison of Geostatistical Methods for Estimating Transmissivity Using Data on Transmissivity and Specific Capacity, *Water Resour. Res.*, *23*(9), 1717-1737.

Aksoy, A., F. Q. Zhang, and J. W. Nielsen-Gammon (2006a), Ensemble-based simultaneous state and parameter estimation with MM5, *Geophys Res. Lett.*, *33*(12), doi:10.1029/2006GL026186.

Aksoy, A., F. Q. Zhang, and J. W. Nielsen-Gammon (2006b), Ensemble-based simultaneous state and parameter estimation in a two-dimensional sea breeze model, *Mon. Weather Rev.*, *134*, 2951-2970.

Alley, W. M., R. W. Healy, J. W. LaBaugh, and T. E. Reilly (2002), Hydrology - Flow and storage in groundwater systems, *Science*, *296*(5575), 1985-1990.

Almasri, M. N. and J. J. Kaluarachchi (2007), Modeling nitrate contamination of groundwater in agricultural watersheds, *J. Hydrol.*, *343*(3-4), 211-229.

Amorocho, J. and B. Espildor (1973), Entropy in Assessment of Uncertainty in Hydrologic Systems and Models, *Water Resour. Res.*, *9*(6), 1511-1522.

Anderson, I. C., M. Poth, J. Homstead, and D. Burdige (1993), A Comparison of NO and N_2O Production by the Autotrophic Nitrifier Nitrosomonas-Europaea and the Heterotrophic Nitrifier Alcaligenes-Faecalis, *Appl. Environ. Microb.*, *59*(11), 3525-3533.

Anderson, J. L. (2007a), An adaptive covariance inflation error correction algorithm for ensemble filters, *Tellus A*, *59*(2), 210-224.

Anderson, J. L. (2007b), Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter, *Physica D*, *230*(1-2), 99-111.

Andrieu, C., N. de Freitas, and A. Doucet (2001), Robust full Bayesian learning for radial basis networks, *Neural Computation*, *13*, 2359-2407.

Arnold, J. G. and N. Fohrer (2005), SWAT2000: current capabilities and research opportunities in applied watershed modelling, *Hydrol. Process*, *19*(3), 563-572.

Ashworth, J. B. (1983), Ground-water availability of the Lower Cretaceous formations in the Hill Country of south-central Texas, *Tex. Dept. Water Resour. Rep. 273*, Austin, Texas, 65 pp.

Ashworth, J. B. and J. Hopkins (1995), Aquifers of Texas, *Tex. Water Dev. Board Rep. 345*, Austin, Texas, 69 pp.

Babbar-Sebens M. and R. Karthikeyan (2009), Consideration of Sample Size For Estimating Contaminant Load Reductions Using Load Duration Curves, *J. Hydrol.*, *372*(1–4), 118–123.

Baez-Cazull, S.E., McGuire, J.T., Cozzarelli, I.M., and Voytek, M.A. (2008), Determination of dominant biogeochemical processes in a contaminated aquifer-wetland system using multivariate statistical analyses, *J. Environ. Qual., 37*(1), 30- 46.

Bailey, R., and D. Bau (2010), Ensemble smoother assimilation of hydraulic head and return flow data to estimate hydraulic conductivity distribution, *Water Resour. Res.*, *46*, W12543, doi:10.1029/2010WR009147.

Bailey, R., and D. Bau (2011), Estimating spatially-variable first-order rate constants in groundwater reactive transport systems, *J. Contam. Hydrol.*, *122*(1-4), 104-121.

Bak, P., M. Paczuski, and M. Shubik (1997), Price variations in a stock market with many agents, *Physica A*, *246*(3-4), 430-453.

Barnard, R., P. W. Leadley, and B. A. Hungate (2005), Global change, nitrification, and denitrification: A review, *Global Biogeochem. Cy.*, *19*(1), GB1007, doi:10.1029/2004GB002282.

Bates, B.C. and Campbell, E.P. (2001), A Markov Chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall–runoff modeling. *Water Resour. Res.*, *37*(4), 937–947.

Bengraïne, K. and T. F. Marhaba (2003), Using principal component analysis to monitor spatial and temporal changes in water quality, *J. Hazard. Mater.*, *100*, 179–195.

Benham, B. L., et al. (2006) Modeling bacteria fate and transport in watersheds to support TMDLs, *Trans. ASABE, 49*, 987–1002.

Berg, H. C. (2004), *E. coli in Motion*, Springer Verlag, New York, New York.

Bethune, D. N., R. N. Farvolden, M. C. Ryan, and A. L. Guzman (1996), Industrial contamination of a municipal water-supply lake by induced reversal of groundwater flow, Managua, Nicaragua, *Ground Water*, *34*(4), 699-708.

Beven, K., and A. Binley (1992), The Future of Distributed Models - Model Calibration and Uncertainty Prediction, *Hydrol. Process. 6*(3), 279-298.

Blandford, T. N., D. J. Blazer, K. C. Calhoun, A. R. Dutton, T. Naing, R. C. Reedy, and B. R. Scanlon (2003), Groundwater Availability of the Southern Ogallala Aquifer in Texas and New Mexico: Numerical Simulations Through 2050, *Tex. Water Dev. Board.*, Austin, Texas, 158 pp., (available at http://www.twdb.state.tx.us).

Bohlke, J. K. (2002), Groundwater recharge and agricultural contamination, *Hydrogeol. J.*, 10(1), 153-179.

Bohlke, J. K., R. Wanty, M. Tuttle, G. Delin, and M. Landon (2002), Denitrification in the recharge area and discharge area of a transient agricultural nitrate plume in a glacial outwash sand aquifer, Minnesota, *Water Resour. Res.*, *38*(7), 1105, doi:10.1029/2001WR000663.

BRA and ECI (2010), Lake Granbury watershed protection plan, *Brazos River Authority and Espey Consultant Inc.*, pp. 10-18.

Bradford, S. A., J. *Šimůnek*, and S. L. Walker (2006), Transport and straining of *E-coli O157 : H7* in saturated porous media, *Water Resour. Res.*, *42*(12), W12S12, doi:10.1029/2005WR004805.

Brender, J., J. Olive, M. Felkner, L. Saurez, K. Hendricks, and W. Marckwardt (2004), Intake of nitrates and nitrites and birth defects in offspring, *Epidemiology*, *15*(4), S184-S184.

Bronson, K. F., A. Malapati, J. D. Booker, B. R. Scanlon, W. H. Hudnall, and A. M. Schubert (2009), Residual soil nitrate in irrigated Southern High Plains cotton fields and Ogallala groundwater nitrate, *J. Soil Water Conserv.*, *64*(2), 98-104.

Burton, M. (2007), Application of a nitrate fate and transport model to the Abbotsford-Sumas Aquifer, Whatcom County, Washington, M.S. thesis, Western Washington University, Bellingham, Washington.

Carle, S.F., B.K. Esser, and J.E. Moran, (2006), High-resolution simulation of basin-scale nitrate transport considering aquifer system heterogeneity, *Geosphere*, *4*, 195-209.

Carrera, J. and S. P. Neuman (1986), Estimation of Aquifer Parameters under Transient and Steady-State Conditions: 1. Maximum-Likelihood Method Incorporating Prior Information, *Water Resour. Res.*, *22*(2)a, 199-210.

Carrera, J. and S. P. Neuman (1986), Estimation of Aquifer Parameters under Transient and Steady-State Conditions: 3. Application to Synthetic and Field Data, *Water Resour. Res.*, *22*(2), 228-242.

Chang, S. Y. and S. M. I. Latif (2010), Extended Kalman Filtering to Improve the Accuracy of a Subsurface Contaminant Transport Model, *J. Environ. Eng.-ASCE*, *136*(5), 466-474.

Chapman, T. G. (1986), Entropy as a Measure of Hydrologic Data Uncertainty and Model Performance, *J. Hydrol.*, *85*(1-2), 111-126.

Cilek, E. C. and B. Z. Yilmazer (2003), Effects of hydrodynamic parameters on entrainment and flotation performance, *Miner. Eng.*, *16*, 745-756.

Ciocoiu, I. B. (2002), Hybrid feed forward neural networks for solving classification problems, *Neural Process Lett.*, *16*, 81-91.

Cohn, T. A. (2005), Estimating contaminant loads in rivers: An application of adjusted maximum likelihood to type 1 censored data, *Water Resour. Res.*, *41*, W07003, doi:10.1029/2004WR003833.

Dagan, G. (1979), Models of Groundwater Flow in Statistically Homogeneous Porous Formations, *Water Resour. Res.*, *15*(1), 47-63.

Dagan, G. (1982), Analysis of Flow through Heterogeneous Random Aquifers .2. Unsteady-Flow in Confined Formations, *Water Resour. Res.*, *18*(5), 1571-1585.

Dagan, G. (1984), Solute Transport in Heterogeneous Porous Formations, *J. Fluid Mech.*, *145*, 151-177.

Dagan, G. (1987), Theory of Solute Transport by Groundwater, *Annu. Rev. Fluid Mech.*, *19*, 183-215.

Dagan, G. (1988), Time-Dependent Macrodispersion for Solute Transport in Anisotropic Heterogeneous Aquifers, *Water Resour. Res.*, *24*(9), 1491-1500.

Dagan, G. and S. C. Lessoff (2007), Transmissivity upscaling in numerical models of steady aquifer flow: Conditional statistics, *Water Resour. Res.*, *43*(10), W10412, doi:10.1029/2007WR006028.

Daley, R. (1992), Estimating Model-Error Covariances for Application to Atmospheric Data Assimilation, *Mon. Weather Rev.*, *120*(8), 1735-1746.

Daley, R. (1997), Atmospheric data assimilation, *J. Meteorol. Soc. Jpn.*, *75*, 319-329.

Del Grosso, S. J., W. J. Parton, A. R. Mosier, D. S. Ojima, A. E. Kulmala, and S. Phongpan (2000), General model for N2O and N-2 gas emissions from soils due to dentrification, *Global Biogeochem. Cy.*, *14*(4), 1045-1060.

Dembo, A., T. M. Cover, and J. A. Thomas (1991), Information Theoretic Inequalities, *IEEE Trans. Inf. Theory*, *37*(6), 1501-1518.

Dennehy, K. F., D. W. Litke, and P. B. McMahon (2002), The High Plains Aquifer, USA: groundwater development and sustainability, in *Sustainable Groundwater Development*, edited by K. Hiscock, M. O. Rivett, and R. M. Davison, Geol. Soc., Special Publication 193, London.

Deutsch, C. V. and A. G. Journel (1992), *GSLIB: Geostatistical Software Library and User's Guide*, Oxford University Press, New York, New York.

Domagalski, J. L., S. Ator, R. Coupe, K. McCarthy, D. Lampe, M. Sandstrom, and N. Baker (2008), Comparative Study of Transport Processes of Nitrogen, Phosphorus, and Herbicides to Streams in Five Agricultural Basins, USA, *J. Environ. Qual.*, *37*, 1158-1169, doi:10.2134/jeq2007.0408.

Dorner, S. M., W. B. Anderson, R. M. Slawson, N. Kouwen, and P. M. Huck (2006), Hydrologic modeling of pathogen fate and transport, *Environ. Sci. Technol.*, *40*(15), 4746-4753.

Dorner, S. M., W. B. Anderson, R.M. Slawson, N. Kouwen, and P. M. Huck (2006), Hydrologic modeling of pathogen fate and transport, *Environ. Sci. Technol.*, *40*, 4746-4753.

Dutton A., S. W. Tinker, John A., and Katherine G. (2004), Adjustment of Parameters to Improve the Calibration of the Og-n Model of the Ogallala Aquifer, Panhandle Water Planning Area, *Tex. Water Dev. Board*, Austin, Texas, (available at http://www.twdb.state.tx.us).

Ebtehaj, M., H. Moradkhani, and H. V. Gupta (2010), Improving robustness of hydrologic parameter estimation by the use of moving block bootstrap resampling, *Water Resour. Res.*, *46*, W07515, doi: 10.1029/2009WR007981.

Eltahir, E. A. B. (1998), A soil moisture rainfall feedback mechanism 1. Theory and observations, *Water Resour. Res.*, *34*(4), 765-776.

Entekhabi, D., I. Rodriguez-Iturbe, and F. Castelli (1996), Mutual interaction of soil moisture state and atmospheric processes, *J. Hydrol.*, *184*(1-2), 3-17.

Erban, R., and S. J. Chapman (2007), Reactive boundary conditions for stochastic simulations of reaction-diffusion processes, *Phys. Biol.*, *4*(1), 16-28.

Evensen, G. (1994), Sequential Data Assimilation with a Nonlinear Quasi-Geostrophic Model Using Monte-Carlo Methods to Forecast Error Statistics, *J. Geophys. Res.*, *99*(C5), 10143-10162.

Evensen, G. (2009), Ensemble methods, in *Data Assimilation: the ensemble Kalman filter*, pp. 119-137, Springer, Berlin.

Fayer, M. J. (2000), UNSAT-H Version 3.0: Unsaturated Soil Water and Heat Flow Model, Theory, User Manual, and Examples, *Rep. 13249*, Battelle Pacific Northwest Laboratory, Hanford, Washington.

Federle, T. W., D. C. Dobbins, J. R. Thorntonmanning, and D. D. Jones (1986), Microbial Biomass, Activity, and Community Structure in Subsurface Soils, *Ground Water*, *24*(3), 365-374.

Feyen, J., D. Jacques, A. Timmerman, and J. Vanderborght (1998), Modelling water flow and solute transport in heterogeneous soils: A review of recent approaches, *J Agr. Eng. Res.*, *70*(3), 231-256.

Fincher L. M., C. D. Parker, and C. P. Chauret (2009), Occurrence and Antibiotic Resistance of *Escherichia coli O157:H7* in a Watershed in North-Central Indiana, *J. Environ. Qual.*, *38*, 997–1004, doi:10.2134/jeq2008.0077.

Flerchinger, G. N., C. L. Hanson, and J. R. Wight (1996), Modeling evapotranspiration and surface energy budgets across a watershed, *Water Resour. Res.*, *32*(8), 2539-2548.

Flint, K. P. (1987), The Long-Term Survival of *Escherichia-Coli* in River Water, J. *Appl. Bacteriol.*, *63*, 261-270.

Flores, A. N., D. Entekhabi, and R. L. Bras (2010), Reproducibility of soil moisture ensembles when representing soil parameter uncertainty using a Latin Hypercube-based approach with correlation control, *Water Resour. Res*, *46*, W04506, doi:10.1029/2009WR008155.

Franssen, H. J. H., and W. Kinzelbach (2008), Real-time groundwater flow modeling with the Ensemble Kalman Filter: Joint estimation of states and parameters and the filter inbreeding problem, *Water Resour. Res.*, *44*(9), W04506, doi:10.1029/2009WR008155.

Fraser, A. M. (1989), Information and Entropy in Strange Attractors, *IEEE Trans. Inf. Theory*, *35*(2), 245-262.

Freedman, D. and P. Diaconis (1981), On the Histogram as a Density Estimator: L2 Theory, *Z. Wahrsch. verw. Gebiete*, *57*(4), 453-476.

Freeze, R. A. (1975), Stochastic-Conceptual Analysis of One-Dimensional Groundwater Flow in Nonuniform Homogeneous Media, *Water Resour. Res.*, *11*(5), 725-741.

Frind, E. O., W. H. M. Duynisveld, O. Strebel, and J. Boettcher (1990), Modeling of Multicomponent Transport with Microbial Transformation in Groundwater - the Fuhrberg Case, *Water Resour. Res.*, *26*(8), 1707-1719.

Gaganis, P. and L. Smith (2001), A Bayesian approach to the quantification of the effect of model error on the predictions of groundwater models, *Water Resour. Res.*, *37*(9), 2309-2322.

Gagliardi, J. V. and J. S. Karns (2000), Leaching of *Escherichia Coli* O157 : H7 in diverse soils under various agricultural management practices, *Appl. Environ. Microb.*, *66*(9), 877-883.

Gedeon, D. (1986), Mean-Parameter Modeling of Oscillating Flow, *J. Heat Trans-T. Asme*, *108*(3), 513-518.

Gelhar, L. W., A. L. Gutjahr, and R. L. Naff (1979), Stochastic-Analysis of Macrodispersion in a Stratified Aquifer, *Water Resour. Res.*, *15*(6), 1387-1397.

Gelhar, L. W. and C. L. Axness (1983), 3-Dimensional Stochastic-Analysis of Macrodispersion in Aquifers, *Water Resour. Res.*, *19*(1), 161-180.

Gelhar, L. W., C. Welty, and K. R. Rehfeldt (1992), A Critical-Review of Data on Field-Scale Dispersion in Aquifers, *Water Resour. Res.*, *28*(7), 1955-1974.

Gelhar, L. W., C. Welty, and K. Rehfeldt (1992), A critical review of data on field-scale dispersion in aquifers, *Water Resour. Res.*, *28* (7), 1955–1974.

Gelman, A., B. J. Carlin, H. S. Stern, and D. B. Rubin (1995), *Bayesian Data Analysis*, Chapman and Hall, London.

Geman, S., E. Bienenstock, and R. Doursat (1992), Neural Networks and the Bias Variance Dilemma, *Neural Computation*, *4*, 1-58.

Gerba, C. P., J. L. Melnick, and C. W. (1975), Fate of Wastewater Bacteria and Viruses in Soil, *J. Irr. Drain. Div-ASCE*, *101*(3), 157-174.

Ginn, T. R., B. D. Wood, K. E. Nelson, T. D. Scheibe, E. M. Murphy, and T. P. Clement (2002), Processes in microbial transport in the natural subsurface, *Adv. Water Resour.*, *25*(8-12), 1017-1042.

Goldberger, A. L., L. A. N. Amaral, J. M. Hausdorff, P. Ch. Ivanov, C. K. Peng, and H. E. Stanley (2002), Fractal dynamics in physiology: Alterations with disease and aging, *Proc. Natl. Acad. Sci. U.S.A.*, *99*, 2466-2472.

Goncalves, M. C., J. *Šimůnek*, T. B. Ramos, J. C. Martins, M. J. Neves, and F. P. Pires (2006), Multicomponent solute transport in soil lysimeters irrigated with waters of different quality, *Water Resour. Res.*, *42*(8), W08401, doi:10.1029/2005WR004802.

Gormly, J. R. and R. F. Spalding (1979), Sources and Concentrations of Nitrate-Nitrogen in Ground-Water of the Central Platte Region, Nebraska, *Ground Water*, *17*(3), 291-301, W08401, doi:10.1029/2005WR004802.

Green, C. T., J. K. Böhlke, B. A. Bekins, and S. P. Phillips (2010), Mixing effects on apparent reaction rates and isotope fractionation during denitrification in a heterogeneous aquifer, *Water Resour. Res.*, *46*, W08525, doi: 10.1029/2009WR008903.

Hansen, S., H. E. Jensen, N. E. Nielsen, and H. Svendsen (1990), DAISY: Soil Plant Atmosphere System Model, *NPO Report No. A 10*, The National Agency for Environmental Protection, Copenhagen, 272 pp.

Harmel, R. D., K. W. King, C. W. Richardson, and J. R. Williams (2003), Long-term precipitation analyses for the central Texas Blackland Prairie, *T. ASAE*, *46*(5), 1381-1388.

Harter, T., H. Davis, M. C. Mathews, and R. D. Meyer (2001), Shallow groundwater quality on dairy farms with irrigated forage crops, *J. Contam. Hydrol.*, *55*, 287-315.

Haykin, S. (1996), Neural networks expand SP's horizons, *IEEE Signal Process Mag.*, *13*, 24-49.

Haznedaroglu, B. Z., C. H. Bolster, and S. L. Walker (2008), The role of starvation on *Escherichia Coli* adhesion and transport in saturated porous media, *Water Res*, *42*(6-7), 1547-1554.

Heinen, M. (2006), Simplified denitrification models: Overview and properties, *Geoderma*, *133*(3-4), 444-463.

Helton, J. C. and W. L. Oberkampf (2004), Alternative representations of epistemic uncertainty, Reliab.Eng. *Syst. Saf.*, *85*, 1-10.

Hipsey, M. R., J. P. Antenucci, and J. D. Brookes (2008), A generic, process-based model of microbial pollution in aquatic systems, *Water Resour. Res.*, *44*, W07408, doi:10.1029/2007WR006395.

Holmes, C. C. and B. K. Mallick (1998), Bayesian radial basis functions of variable dimension, *Neural Computation*, *10*, 1217-1233.

Houtekamer, P. L., and H. L. Mitchell (1998), Data assimilation using an ensemble Kalman filter technique, *Mon. Weather Rev.*, *126*(3), 796-811.

http://walrus.wr.usgs.gov/infobank/programs/html/factsheets/pdfs/2004_3120.pdf

http://www.gbra.org/CRP/Sites; accessed on May, 12, 2009

http://en.wikipedia.org/wiki/Hurst; accessed on July, 08, 2009

Hudak, P. F. (2000), Regional trends in nitrate content of Texas groundwater, *J. Hydrol.*, *228*(1-2), 37-47.

Hurst, H. E. (1951), Long-Term Storage Capacity of Reservoirs, *Trans. Am. Soc. Civ. Eng.*, *116*, 770-799.

Hutson, J. L. and R. J. Wagenet (1991), Simulating Nitrogen Dynamics in Soils Using a Deterministic Model, *Soil Use Manage*, *7*(2), 74-78.

Indelman, P. and G. Dagan (1993), Upscaling of Conductivity of Heterogeneous Formations - General-Approach and Application to Isotropic Media, *Transport Porous Med.*, *12*(2), 161-183.

Indelman, P., I. Touber-Yasur, B. Yaron, and G. Dagan (1998), Stochastic analysis of water flow and pesticides transport in a field experiment, *J. Contam. Hydrol.*, *32*(1-2), 77-97.

Jamieson, R. C., R. J. Gordon, S. C. Tattrie, and G. W. Stratton (2003), Sources and persistence of fecal coliform bacteria in a rural watershed, *Water Qual. Res. J. Can.*, *38*(1), 33-47.

Jansson, P. E., and L. Karlberg (2001), *Coupled Heat and Mass Transfer Model for Soil-Plant-Atmosphere Systems*, Royal Institute of Technology, Department of Civil and Environmental Engineering, Stockholm.

Jiang, G., M. J. Noonan, G. D. Buchan, and N. Smith (2007), Transport of Escherichia coli through variably saturated sand columns and modeling approaches, *J. Contam. Hydrol.*, *93*(1-4), 2-20.

Jin, G., A. J. Englande, and A. Liu (2003), A preliminary study on coastal water quality monitoring and modeling, *J. Environ. Sci. Health*, *A38*, 493– 509.

Jolliffe, I.T. (2002), *Principal component analysis*, Vol. 2, Springer-Verlag, New York.

Jonge, De, V. (1997), High remaining productivity in the dutch western wadden sea despite decreasing nutrient inputs from riverine sources, *Marine Pollution Bulletin*, *34*(6), 427-436.

Kalman, R. E. (1960), A New Approach to Linear Filtering and Prediction Problems, *J. Basic Eng-T ASME, 82*, 35-45.

Keeney, D. (1986), Sources of Nitrate to Groundwater, in *Crit. Rev. Environ. Control*, vol. 16(3), edited by C. P. Staub, pp. 257-304, CRC Press, Boca Raton, FL.

Khan, S., A. R. Ganguly, S. Bandyopadhyay, S. Saigal, D. J. Erickson, III, V. Protopopescu, and G. Ostrouchov (2006), Nonlinear statistics reveals stronger ties between ENSO and the tropical hydrological cycle, *Geophys. Res. Lett.*, *33*, L24402, doi:10.1029/2006GL027941.

Kitanidis, P. K. and R. L. Bras (1980), Adaptive Filtering through Detection of Isolated Transient Errors in Rainfall-Runoff Models, *Water Resour. Res.*, *16*(4), 740-748.

Knight, F. B. (1981), *Essentials of Brownian Motion and Diffusion*, American Mathematical Society, Providence, Rhode Island.

Kollat, J. B., P. M. Reed, and R. M. Maxwell (2011), Many-objective groundwater monitoring network design using bias-aware ensemble Kalman filtering, evolutionary optimization, and visual analytics, *Water Resour. Res.*, *47*, W02529, doi:10.1029/2010WR009194.

Kolle, W., P. Werner, O. Strebel, and J. Bottcher (1983), Denitrification by Pyrite in a Reducing Aquifer, *Vom Wasser*, *61*, 125-147.

Komor, S. C., and H. W. Anderson (1993), Nitrogen Isotopes as Indicators of Nitrate Sources in Minnesota Sand-Plain Aquifers, *Ground Water*, *31*(2), 260-270.

Koren, V. I., B. D. Finnerty, J. C. Schaake, M. B. Smith, D. J. Seo, and Q. Y. Duan (1999), Scale dependencies of hydrologic models to spatial variability of precipitation, *J. Hydrol.*, *217*(3-4), 285-302.

Kozubowski, T. J., and S. T. Rachev (1999), Univariate geometric stable laws, *J. Comp. Anal. Appl.*, *1*(2), 177-217.

Lanouette, R., J. Thibault, and J. L. Valade (1999), Process modeling with neural networks using small experimental data sets, *Comput. Chem. Eng.*, *23*, 1167-1176.

Leij, F. J., E. Priesack, and M. G. Schaap (2000), Solute transport modeled with Green's functions with application to persistent solute sources, *J. Contam. Hydrol.*, *41*(1-2), 155-173.

Lenhard, R. J., M. Oostrom, and M. D. White (1995), Modeling Fluid-Flow and Transport in Variably Saturated Porous-Media with the Stomp Simulator .2. Verification and Validation Exercises, *Adv. Water Resour.*, *18*(6), 365-373.

Lessard, E. J. and J. M. Sieburth (1983), Survival of Natural Sewage Populations of Enteric Bacteria in Diffusion and Batch Chambers in the Marine-Environment, *Appl. Environ. Microbiol.*, *45*, 950-959.

Levin, S. A. (1998), Ecosystems and the biosphere as complex adaptive systems, *Ecosystems*, *1*(5), 431-436.

Liu, Y. Q. and H. V. Gupta (2007), Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, *Water Resour. Res.*, *43*(7), W07401, doi:10.1029/2006WR005756.

League, K. and G. Gander (1990), R-5 Revisited 1. Spatial Variability of Infiltration on a Small Rangeland Catchment, *Water Resour. Res.*, *26*(5), 957-971.

Long, A. J., L. D. Putnam, and J. M. Carter (2003), Simulated Ground-Water Flow in the Ogallala and Arikaree Aquifers, Rosebud Indian Reservation Area, South Dakota, *U.S. Geol. Surv. Water Resour. Invest. Rep. 03-4043*, Denver, Colorado (available at http://pubs.usgs.gov/wri/wri034043/wri034043_files/wri034043.pdf).

Lowe, K. S. and R. L. Siegrist (2008), Controlled field experiment for performance evaluation of septic tank effluent treatment during soil infiltration, *J. Environ Eng-ASCE*, *134*(2), 93-101.

Lundberg, J. O. and M. Govoni (2004), Inorganic nitrate is a possible source for systemic generation of nitric oxide, *Free Radic. Biol. Med.*, *37*(3), 395-400.

Mace, R. E., A. H. Chowdhury, R. Anaya, and S. C. Way (2000), Groundwater availability of the Trinity Aquifer, Hill Country Area, Texas—Numerical simulations through 2050, *Tex. Water Dev. Board Rep.,* Austin, Texas, *(available at http://www.twdb.state.tx.us).*

Macnab, R. M. and D. E. Koshland (1972), Gradient-Sensing Mechanism in Bacterial Chemotaxis, *Proc. Nat. Acad. Sci. USA*, *69*(9), 2509–2512.

Maeda, K., Y. Imae,J. I. Shioi, and F. Oosawa (1976), Effect of temperature on motility and chemotaxis of *Escherichia Coli*, *J. Bacteriol.*, *127*, 1039-1046.

Mallants, D., B. P. Mohanty, A. Vervoort, and J. Feyen (1997), Spatial analysis of saturated hydraulic conductivity in a soil with macropores, *Soil Technol.*, *10*(2), 115-131.

Margulis, S. A., D. McLaughlin, D. Entekhabi, and S. Dunne (2002), Land data assimilation and estimation of soil moisture using measurements from the Southern Great Plains 1997 Field Experiment, *Water Resour. Res.*, *38*(12), 1299, doi:10.1029/2001WR001114.

Mariethoz, G., P. Renard, and R. Froidevaux (2009), Integrating collocated auxiliary parameters in geostatistical simulations using joint probability distributions and probability aggregation, *Water Resour. Res.*, *45*, W08421, doi:10.1029/2008WR007408.

Mariotti, A., A. Landreau, and B. Simon (1988), N-15 Isotope Biogeochemistry and Natural Denitrification Process in Groundwater - Application to the Chalk Aquifer of Northern France, *Geochim Cosmochim Ac.*, *52*(7), 1869-1878.

Maritan, A., F. Colaiori, A. Flammini, M. Cieplak, and J. R. Banavar (1996), Universality classes of optimal channel networks, *Science*, *272*(5264), 984-986.

Marquardt, D. W. (1963), An Algorithm for Least-Squares Estimation of Nonlinear Parameters, *J. Soc. Ind. Appl. Math.*, *11*(2), 431-441.

Mayer, K. U., D. W. Blowes, and E. O. Frind (2001), Reactive transport modeling of an in situ reactive barrier for the treatment of hexavalent chromium and trichloroethylene in groundwater, *Water Resour. Res.*, *37*(12), 3091-3103.

Mays, D. C., B. A. Faybishenko, and S. Finsterle (2002), Information entropy to measure temporal and spatial complexity of unsaturated flow in heterogeneous media, *Water Resour. Res.*, *38*(12), 1313, doi:10.1029/ 2001WR001185.

McCarthy, D. T., A. Deletic, V. G. Mitchell, and C. Diaper (2008), Uncertainties in storm water *E. Coli* levels, *Water Res.*, *42*(6-7), 1812-1824.

McCorquodale, J. A., I. Georgiou, S. Carnelos, and A. J. Englande (2004), Modeling coliforms in storm water plumes, *J. Environ. Eng. Sci.*, *3*, 419– 431.

McDonald, M. G. and A. W. Harbaugh (1984), A modular three-dimensional finite-difference groundwater flow model, *U.S. Geological Survey Open-File Report 83*, 875 pp.

McDonald, M. G. and A. W. Harbaugh (2003), The history of MODFLOW, *Ground Water*, *41*(2), 280-283.

McGuire, V. L. (2003), Water-level changes in the High Plains aquifer, predevelopment to 2001, *U.S. Geological Survey Fact Sheet 078-03*, 4 pp.

McKergow, L. A. and R. J. Davies-Colley (2009), Stormflow dynamics and loads of *Escherichia Coli* in a large mixed land use catchment, *Hydrol. Process*, *24*(3), 276-289, doi: 10.1002/hyp.7480.

McMahon, P. B., J. A. Tindall, J. A. Collins, K. J. Lull, and J. R. Nuttle (1995), Hydrologic and geochemical effects on oxygen uptake in bottom sediments of an effluent-dominated river, *Water Resour. Res.*, *31*(10), 2561-2569.

McMahon, P. B., B. W. Bruce, M. F. Becker, L. M. Pope, and K. F. Dennehy (2000), Occurrence of nitrous oxide in the central High Plains aquifer, 1999, *Environ. Sci. Technol.*, *34*(23), 4873-4877.

McMahon, P. B., J. K. Bohlke, L. J. Kauffman, K. L. Kipp, M. K. Landon, C. A. Crandall, K. R. Burow, and C. J. Brown (2008), Source and transport controls on the movement of nitrate to public supply wells in selected principal aquifers of the United States, *Water Resour. Res.*, *44*(4), W04401, doi: 10.1029/2007WR006252.

McMahon, P. B., K. F. Dennehy, B. W. Bruce, J. K. Bohlke, R. L. Michel, J. J. Gurdak, and D. B. Hurlbut (2006), Storage and transit time of chemicals in thick unsaturated zones under rangeland and irrigated cropland, High Plains, United States, *Water Resour. Res.*, *42*(3), W03413, doi: 10.1029/2005WR004417.

Mead, P. S., and P. M. Griffin (1998), *Escherichia coli* O157:H7, *Lancet*, *352*, 1207-1212.

Medema, G. J. and J. F. Schijven (2001), Modelling the sewage discharge and dispersion of Cryptosporidium and Giardia in surface water, *Water Res.*, *35*, 4307– 4316.

Mehta, S., A. E. Fryara, R. M. Bradyb, and R. H. Morinc (2000), Modeling regional salinization of the Ogallala aquifer, Southern High Plains, TX, USA, *J. Hydrol.*,*238*(1-2), 44-64.

Mishra, A. K., O. Mehmet, and V. P. Singh (2009), An entropy-based investigation into the variability of precipitation, *J. Hydrol.*, *370*(1-4), 139-154.

Mogheir, Y. and V. P. Singh (2002), Application of Information Theory to Groundwater Quality Monitoring Networks, *Water Resources Management*, *16*, 37–49.

Money, E. S., G. P. Carter, and M. L. Serre (2009), Modern Space/Time Geostatistics Using River Distances: Data Integration of Turbidity and *E. coli* Measurements to Assess Fecal Contamination Along the Raritan River in New Jersey. *Environ. Sci. Technol.*, *43*, 3736-3742. doi: 10.1021/es803236j.

Moore, A. W. and M.S. Lee (1994), Efficient algorithms for minimizing cross validation error, in *Proceedings of the Eleventh International Conference on Machine Learning*, edited by W. Cohen and H. Hirsh, pp. 190–198, San Francisco, California.

Moradkhani, H., S. Sorooshian, H. V. Gupta, and P. R. Houser (2005), Dual state-parameter estimation of hydrological models using ensemble Kalman filter, *Adv. Water Resour.*, *28*(2), 135-147.

Murphy, D. V., S. Recous, E. A. Stockdale, I. R. P. Fillery, L. S. Jensen, D. J. Hatch, and K. W. T. Goulding (2003), Gross nitrogen fluxes in soil: Theory, measurement and application of N-15 pool dilution techniques, *Adv. Agron.*, *79*, 69-118.

Murphy, E. M. and T. R. Ginn (2000), Modeling microbial processes in porous media, *Hydrogeol J.*, *8*(1), 142-158.

Naevdal, G., L. M. Johnsen, S. I. Aanonsen, and E. H. Vefring (2005), Reservoir monitoring and continuous model updating using ensemble Kalman filter, *Spe J.*, *10*(1), 66-74.

Neuman, S. P. (1980), A Statistical Approach to the Inverse Problem of Aquifer Hydrology: 3. Improved Solution Method and Added Perspective, *Water Resour. Res.*, *16*(2), 331-346.

Neuman, S. P. (2003), Accounting for conceptual model uncertainty via maximum likelihood Bayesian model averaging, in *Calibration and Reliability in Groundwater Modelling: A Few Steps Closer to Reality*, pp. 303-313, IAHS Publication 195, Wallingford.

Neuman, S. P. and P. J. Wierenga (2003), A comprehensive strategy of hydrogeologic modeling and uncertainty analysis for nuclear facilities and sites, *Rep. NUREG/CR-6805*, U.S. Nucl. Regul. Comm., Washington, D. C.

Neuman, S. P. and S. Yakowitz (1979), Statistical Approach to the Inverse Problem of Aquifer Hydrology .1. Theory, *Water Resour. Res.*, *15*(4), 845-860.

Neuman, S. P., C. L. Winter, and C. M. Newman (1987), Stochastic-Theory of Field-Scale Fickian Dispersion in Anisotropic Porous-Media, *Water Resour. Res.*, *23*(3), 453-466.

Neuman, S. P., G. E. Fogg, and E. A. Jacobson (1980), A Statistical Approach to the Inverse Problem of Aquifer Hydrology .2. Case-Study, *Water Resour. Res.*, *16*(1), 33-58.

Nevers B. M. and R. L. Whitman (2005), Nowcast modeling of *Escherichia Coli* concentrations at multiple urban beaches of southern Lake Michigan, *Water Res.*, *39*(20), 5250–5260, doi.org/10.1016/j.watres.

Ng, G. H. C., D. McLaughlin, D. Entekhabi, and B. Scanlon (2009), Using data assimilation to identify diffuse recharge mechanisms from chemical and physical data in the unsaturated zone, *Water Resour. Res.*, *45*, W09409, doi:10.1029/2009WR007831.

Niswonger, R. and D. Prudic (2009), Comment on "Evaluating Interactions between Groundwater and Vadose Zone Using the HYDRUS-Based Flow Package for MODFLOW" by Navin Kumar C. Twarakavi, Jirka Simunek, and Sophia Seo, *Vadose Zone J.*, *8*(3), 820-821.

Noguchi K., H. Nakajima, and R. Aono (1997) Effects of oxygen and nitrate on growth of *Escherichia Coli* and *Pseudomonas aeruginosa* in the presence of organic solvents, *Extremophiles*, *1*, 193–198.

Nolan, B. T., K. J. Hitt, and B. C. Ruddy (2002), Probability of nitrate contamination of recently recharged groundwaters in the conterminous United States, *Environ Sci Technol*, *36*(10), 2138-2145.

Nolte, B. (2010), Nitrogen from Fertilizers, *ChemMatters*, *4*, 5-7.

Olden, J.D. and Poff NL (2003), Redundancy and the choice of hydrologic indices for characterizing streamflow regimes, *River Res. Appl.*, *19*, 101–121.

Or, D., B. F. Smets, J. M. Wraith, A. Dechesne, and S. P. Friedman (2007), Physical constraints affecting bacterial habitats and activity in unsaturated porous media - a review, *Adv. Water Resour.*, *30*(6-7), 1505-1527.

Pachepsky, A., M. Sadeghi, K. M. Brannan, M. L. Soupir, and M. J. Habersack (2006), Modeling bacteria fate and transport in watershed models to support TMDLs, *Trans. ASABE*, *49*(4):987-1002.

Pang, L. P., M. Close, M. Goltz, L. Sinton, H. Davies, C. Hall, and G. Stanton (2004), Estimation of septic tank setback distances based on transport of *E-coli* and F-RNA phages, *Environ. Int*., *29*(7), 907-921.

Pang, P. C., C. M. Cho, and R. A. Hedlin (1975), Effects of pH and Nitrifier Population on Nitrification of Band-Applied and Homogeneously Mixed Urea Nitrogen in Soils, *Can. J. Soil Sci.*, *55*(1), 15-21.

Parker, J. C., and M. T. van Genuchten (1984), Determining Transport Parameters from Laboratory and Field Tracer Experiments, *Bulletin 84-3*, Va Agricultural Experiment Station, Blacksburg, Virginia.

Parton, W. J., A. R. Mosier, D. S. Ojima, D. W. Valentine, D. S. Schimel, K. Weier, and A. E. Kulmala (1996), Generalized model for N_2 and N_2O production from nitrification and denitrification, *Global Biogeochem. Cy.*, *10*(3), 401-412.

Personne, J. C., F. Poty, L. Vaute, and C. Drogue (1998), Survival, transport and dissemination of *Escherichia Coli* and enterococci in a fissured environment. Study of a flood in a karstic aquifer, *J. Appl. Microbiol.*, *84*(3), 431-438.

Powelson, D. K. and A. L. Mills (2001), Transport of *Escherichia Coli* in sand columns with constant and changing water contents, *J. Environ. Qual.*, *30*(1), 238

Pruess, K. (1991), Tough2 – A General-Purpose Numerical Simulator for Multiphase Fluid and Heat Flow, *Rep. LBL-29400,* Lawrence Berkeley Laboratory, Berkeley, California.

Raju, S., A. Chanzy, J. P. Wigneron, J. C. Calvet, Y. Kerr, and L. Laguerre (1995), Soil-Moisture and Temperature Profile Effects on Microwave Emission at Low-Frequencies, *Remote Sens. Environ.*, *54*(2), 85-97.

Reckhow, K. H. (1999), Water quality prediction and probability network models, *Can. J. Fish. Aquat. Sci.*, *56*, 1150-1158.

Refsgaard, J. C., M. Thorsen, J. B. Jensen, S. Kleeschulte, and S. Hansen (1999), Large scale modelling of groundwater contamination from nitrate leaching, *J. Hydrol.*, *221*(3-4), 117-140.

Robakis N., Y. Cenatiempo, L. Meza-Basso, N. Brot, and H. Weissbach (1983), A coupled DNA-directed in vitro system to study gene expression based on di- and tripeptide formation, *Methods Enzymol.*, *101*, 690–706.

Robert, C. P. and G. Casella (1999), *Monte Carlo Statistical Methods*, Springer, New York, New York.

Rubin, Y., J. P. L. Ferreira, J. D. Rodrigues, and G. Dagan (1990), Estimation of the Hydraulic Parameters of the Rio-Maior Aquifer in Portugal by Using Stochastic Inverse Modeling, *J. Hydrol.*, *118*(1-4), 257-279.

Ruddell, B. L. and P. Kumar (2009), Ecohydrologic process networks: 1. Identification, *Water Resour. Res.*, *45*, W03419, doi:10.1029/2008WR007279.

Runkel, R., C. G. Crawford, and T. A. Cohn (2004), Load estimator (LOADEST): A Fortran program for estimating constituent loads in streams and rivers, *U.S. Geol. Surv. Techniques Methods*, Book 4, Chap. A5.

Russo, D. (1991), Stochastic-Analysis of Simulated Vadose Zone Solute Transport in a Vertical Cross-Section of Heterogeneous Soil during Nonsteady Water-Flow, *Water Resour. Res.*, *27*(3), 267-283.

Ryder, P. D. (1996), Ground Water Atlas of the United States: Oklahoma, Texas, *U.S. Geol. Surv. Contents of HA 730-E*, Reston, Virginia (Available at http://pubs.usgs.gov/ha/ha730/ch_e/index.html).

Ryker, S. J. (2001), Mapping arsenic in ground water—A real need, but a hard problem, *Geotimes*, *46*(11), 34–36.

Sakalauskienė, G. (2003), The Hurst phenomenon in hydrology, *Environ. Res. Engng. Manage.*, *3*(25), 16–20.

Santos, J. F., I. P. Calvo, and M. M. Portela (2010), Spatial and temporal variability of droughts in Portugal, *Water Resour. Res.*, *46*, W03503, doi:10.1029/2009WR008071.

Scandura, J. E. and M. D. Sobsey (1997), Viral and bacterial contamination of groundwater from on-site sewage treatment systems, *Water Sci. Technol.*, *35*(11-12), 141-146.

Scanlon B. R., A. Dutton, and M. Sophocleous (2003), Groundwater recharge in Texas, *Tex. Water Dev. Board Rep.*, Austin, Texas, 80 pp.

Scanlon, B., R. Reedy, and K. Kier (2003), Evaluation of Nitrate Contamination in Major Porous Media Aquifers in Texas, *Technical Rep.,* Bureau of Economic Geology, University of Texas at Austin, Austin, Texas.

Schimel, J. P., L. E. Jackson, and M. K. Firestone (1989), Spatial and Temporal Effects on Plant Microbial Competition for Inorganic Nitrogen in a California Annual Grassland, *Soil Biol. Biochem.*, *21*(8), 1059-1066.

Scott, D. W. (1979), Optimal and Data-Based Histograms, *Biometrika*, *66*(3), 605-610.

Shaffer, M. J. (2002), Nitrogen modeling for soil management, *J. Soil Water Conserv.*, *57*(6), 417-425.

Shannon, C. E. (1948), A Mathematical Theory of Communication, *Bell System Technical Journal*, *27*(3), 379-423.

Sharma, A., K. C. Luk, I. Cordery, and U. Lall (2000), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 2 - Predictor identification of quarterly rainfall using ocean-atmosphere information, *J. Hydrol.*, *239*(1-4), 240-248.

Showers, W. J., B. Genna, T. McDade, R. Bolich, and J. C. Fountain (2008), Nitrate contamination in groundwater on an urbanized dairy farm, *Environ. Sci. Technol.*, *42*(13), 4683-4688.

Shrestha, N. K., S. Hadano, T. Kamachi, and I. Okura (2002), Dinitrogen production from ammonia by Nitrosomonas europaea, *Appl. Catal. a-Gen.*, *237*(1-2), 33-39.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Sims, C. A. and T. Zha (1998), Bayesian methods for dynamic multivariate models, *Int. Econ. Rev.*, *39*, 949-968.

Šimůnek, J., C. M. He, L. P. Pang, and S. A. Bradford (2006), Colloid-facilitated solute transport in variably saturated porous media: Numerical model and experimental verification, *Vadose Zone J.*, *5*(3), 1035-1047.

Šimůnek, J., M. T. v. Genuchten, and M. Šejna (2006), The HYDRUS software package for simulating two- and three-dimensional movement of water, heat, and multiple solutes in variably-saturated media, Version 1.0, PC Progress, Prague, Czech Republic, pp. 241.

Sjogren, R. E. and M. J. Gibson (1981), Bacterial Survival in a Dilute Environment, *Appl. Environ. Microbial.*, *41*, 1331-1336.

Skalak, D. B. (1994), Prototype and feature selection by sampling and random mutation hill climbing algorithms. in *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 293-301, New Brunswick, New Jersey.

Smith, L. and F. W. Schwartz (1980), Mass transport: 1. A stochastic analysis of macroscopic dispersion, *Water Resour. Res.*, *16*(2), 303–313, doi:10.1029/WR016i002p00303.

Smith, L. and F. W. Schwartz (1981), Mass transport: 2. Analysis of uncertainty in prediction, *Water Resour. Res.*, *17*(2), 351–369, doi:10.1029/WR017i002p00351.

Smith, L., and F. W. Schwartz (1981b), Mass-Transport: 3. Role of Hydraulic Conductivity Data in Prediction, *Water Resour. Res.*, *17*(5), 1463-1479.

Smith, M. S., G. W. Thomas, R. E. White, and D. Ritonga (1985), Transport of *Escherichia-Coli* through Intact and Disturbed Soil Columns, *J. Environ. Qual.*, *14*(1), 87-91.

Spalding, R. F. and M. E. Exner, (1993), Occurrence of Nitrate in Groundwater—A Review, *J. Environ. Qual., 22*, 392-402.

Spalding, R. F., D. G. Watts, J. S. Schepers, M. E. Burbach, M. E. Exner, R. J. Poreda, and G. E. Martin (2001), Controlling nitrate leaching in irrigated agriculture, *J. Environ. Qual.*, *30*(4), 1184-1194.

Steets, B. M. and P. A. Holden (2003), A mechanistic model of runoffassociated fecal coliform fate and transport through a coastal lagoon, *Water Res.*, *37*, 589– 608.

Strebel, O., W. H. M. Duynisveld, and J. Bottcher (1989), Nitrate pollution of groundwater in Western Europe, *Agr. Ecosyst. Environ.*, *26*, 189-214.

Strehl, A. and J. Ghosh (2002), A Knowledge Reuse Framework for Combining Multiple Partitions, *J. Machine Learning Res.*, *3*, 583-617.

Sturges, H. A. (1926), The choice of a class interval Case I Computations involving a single, *J. Am. Stat. Assoc.*, *21*, 65-66.

Sun, J. A., G. D. Salvucci, D. Entekhabi, and L. Farhadi (2011), Parameter estimation of coupled water and energy balance models based on stationary constraints of surface states, *Water Resour. Res.*, *47*, W02512, doi:10.1029/2010WR009293.

Tanji, K. K., F. E. Broadbent, M. Mehran, and M. Fried (1979), Extended Version of a Conceptual-Model for Evaluating Annual Nitrogen Leaching Losses from Croplands, *J. Environ. Qual.*, *8*(1), 114-120.

Teague, A., R. Karthikeyan, M. Babbar-Sebens, R. Srinivasan, and R. A. Persyn (2009), Spatially Explicit Load Enrichment Calculation Tool to Identify Potential *E. coli* Sources in Watersheds, *Trans. ASABE*, *52*(4): 1109-1120.

Thiemann, M., M. Trosset, H. Gupta, and S. Sorooshian (2001), Bayesian recursive parameter estimation for hydrologic models, *Water Resour. Res.*, *37*(10), 2521-2535.

Tian, Y. Q., P. Gonga, J. D. Radkeb, and J. Scarborough (2002), Spatial and temporal modeling of microbial contaminants on grazing farmland, *J. Environ. Qual.*, *31*, 860–869.

Toth, J. (1963), A theoretical analysis of groundwater flow in small drainage basins, *J. Geophys. Res.*, *68*(16), 4795-4812.

Townley, L. R. and M. G. Trefry (2000), Surface water-groundwater interaction near shallow circular lakes: Flow geometry in three dimensions, *Water Resour. Res.*, *36*(4), 935–948.

Unc, A. and M. J. Goss (2004), Transport of bacteria from manure and protection of water resources, *Appl. Soil. Ecol.*, *25*(1), 1-18.

USEPA (1997), National water quality inventory, report to Congress, *U.S. Environmental Protection Agency.*

USEPA (2002), Onsite Wastewater Treatment Systems Manual, *EPA/625/R-00/008*.

USEPA (2005), (available at http://www.epa.gov/owm/septic/pubs/onsite_handbook.pdf).

USEPA (2006), http: //www.epa.gov/volunteer/stream/vms50.html., accessed on 26 June, 2009.

USEPA (2006), http: //www.epa.gov/volunteer/stream/vms50.html., accessed on 26 June, 2009.

van Dam, J. C., J. Huygen, J. G. Wesseling, R. A. Feddes, P. Kabat, P. E. V. van Valsum, P. Groenendijk, and C. A. van Diepen (1997), Theory of SWAP, Version 2.0. Simulation of Water Flow, Solute Transport and Plant Growth in the Soil-WaterAtmosphere-Plant Environment, *Rep. 71*, Department of Water Resources, Wageningen Agricultural University, Wageningen.

Van der Steen, P., A. Brenner, Y. Shabtai, and G. Oron (2000), Improved fecal coliform decay in integrated duckweed and algal ponds, *Water Sci. Technol.*, 363-370.

van Genuchten, M. T. (1985), Convective-Dispersive Transport of Solutes Involved in Sequential 1st-Order Decay Reactions, *Comput Geosci*, *11*(2), 129-147.

Vastano, J. A. and H. L. Swinney (1988), Information Transport in Spatiotemporal Systems, *Phys. Rev. Lett.*, *60*(18), 1773-1776.

Vega, M., R. Pardo, E. Brrrado, and L. Deban (1998), Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis, *Water Resour. Res.*, *32*(12), 3581-3592.

Vidon, P., L. P. Tedesco, J. Wilson, M. A. Campbell, L. R. Casey, and M. Gray (2008), Direct and Indirect Hydrological Controls on Concentration and Loading in Midwestern Streams, *J. Environ. Qual.*, *37*(5), 1761-8, doi: 10.2134/jeq2007.0311.

Vogel, J. C., A. S. Talma, and T. H. E. Heaton (1981), Gaseous Nitrogen as Evidence for Denitrification in Groundwater, *J. Hydrol.*, *50*(1-3), 191-200.

Vrugt, J. A. and B. A. Robinson (2007), Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resour. Res.*, *43*(1), W01411, doi:10.1029/2005WR004838.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resour. Res.*, *41*(1), W01017, doi:10.1029/2004WR003059.

Vugrin, E. D., S. A. McKenna, and K. W. Vugrin (2007), Markov Models and the Ensemble Kalman Filter for Estimation of Sorption Rates, *SAND2007-5975*, Sandia National Laboratories, Albuquerque, New Mexico.

Walker, F. R. and J. R. Stedinger (1999), Fate and transport model of Cryptosporidium, *J. Environ. Eng.*, *125*, 325– 333.

Wang, P. P. and C. M. Zheng (2000), Contaminant transport modelling under random sources, *IAHS-Aish*, *265*, 317-323.

Wang, X. G., T. A. Hamill, J. S. Whitaker, and C. H. Bishop (2007), A comparison of hybrid ensemble transform Kalman filter-optimum interpolation and ensemble square root filter analysis schemes, *Mon. Weather Rev.*, *135*(3), 1055-1076.

Wen, X. H. and W. H. Chen (2007), Some practical issues on real-time reservoir model updating using ensemble Kalman filter, *Spe. J.*, *12*(2), 156-166.

Western, A. W. and R. B. Grayson (1998), The Tarrawarra data set: Soil moisture patterns, soil characteristics, and hydrological flux measurements, *Water Resour. Res.*, *34*(10), 2765-2768.

WHI (2002), Visual PEST-ASP Model-Independent Parameter Estimation, *Waterloo Hydrogeologic Inc.*, 2.1-2.10.

White, M. D., M. Oostrom, and R. J. Lenhard (1995), Modeling Fluid-Flow and Transport in Variably Saturated Porous-Media with the Stomp Simulator .1. Nonvolatile 3-Phase Model Description, *Adv. Water Resour.*, *18*(6), 353-364.

Whitman, R. L., M. B. Nevers, G. C. Korinek, and M. N. Byappanahalli (2004), Solar and temporal effects on *Escherichia Coli* concentration at a lake Michigan swimming beach, *Appl. Environ. Microbiol.*, *70*, 4276-4285.

Wilkinson, J., A. Jenkins, M. Wyer, and D. Kay (1995), Modelling faecal coliform dynamics in streams and rivers, *Water Res.*, *29*, 847– 855.

Williams, A. E., J. A. Johnson, L. J. Lund, and Z. J. Kabala (1998), Spatial and temporal variations in nitrate contamination of a rural aquifer, California, *J. Environ. Qual.*, *27*(5), 1147-1157.

Winkel, L., M. Berg, M. Amini, S. J. Hug, and C. A. Johnson (2008), Predicting groundwater arsenic contamination in Southeast Asia from surface parameters, *Nat. Geosci.*, *1*(8), 536–542, doi:10.1038/ngeo254.

Woodbury, A. D. and T. J. Ulrych (2000), A full-Bayesian approach to the groundwater inverse problem for steady state flow, *Water Resour. Res.*, *36*(8), 2081-2093.

Wrage, N., G. L. Velthof, M. L. van Beusichem, and O. Oenema (2001), Role of nitrifier denitrification in the production of nitrous oxide, *Soil Biol. Biochem.*, *33*(12-13), 1723-1732.

Ye, M., S. P. Neuman, and P. D. Meyer (2004), Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resour. Res.*, *40*(5), W05113, doi:10.1029/2003WR002557.

Yeh, G. T. (1981), Analytical Transient One-, Two-, and ThreeDimensional Simulation of Waste Transport in the Aquifer System, *ORNL-5602*, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Zheng, C., and P. P. Wang (1999), MT3DMS: Documentation and user's guide, *Contract Rep. SERDP-99-1*, U.S. Army Eng, R&D Center, Vicksburg, Mississipi, 220pp.

Zou, S. and A. Parr (1995), Optimal Estimation of 2-Dimensional Contaminant Transport, *Ground Water*, *33*(2), 319-325.