DISSERTATION

STOCHASTIC ANALYSIS OF FLOW AND SALT TRANSPORT MODELING IN IRRIGATION-DRAINAGE SYSTEMS

Submitted by

Ayman H. Alzraiee

Department of Civil and Environmental Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2012

Doctoral Committee:

Advisor: Luis A. Garcia
Co-Advisor: Timothy K. Gate

Domenico Bau
Greg Butters

ABSTRACT

STOCHASTIC ANALYSIS OF FLOW AND SALT TRANSPORT MODELING IN IRRIGATION-DRAINAGE SYSTEMS

Sustainability of crop production in the Lower Arkansas River Basin in Colorado is seriously threatened by the continuous degradation of irrigated lands by the dual impact of soil salinization and waterlogging problems. Integration of improved irrigation practices, upgrades to the irrigation systems, and subsurface drainage are essential components of any plan to stop the deterioration of irrigated lands. Numerical simulations of irrigation and drainage systems are necessary to justify the consequent management actions. Despite the uncertainty of their predictions, numerical models are still indispensable decision support tools to investigate the feasibility of irrigation and drainage systems management plans. However, the uncertainties in input parameters to these models create a risk of misleading numerical results. That is beside the fact that the numerical models themselves are conceptual simplifications of the complex reality.

The overarching objective of this dissertation is to *investigate the impact of parameters uncertainty on the response of simulated irrigation-drainage systems*. In the first part of the research, a Global Sensitivity Analysis (GSA) is conducted using a one-dimensional variably saturated problem to prioritize parameters according to their

importance with respect to predefined performance indices. A number of GSA methods are employed for this purpose, and their comparative performances are investigated. Results show that only five parameters out of 18 parameters are responsible for around 73% of crop yield uncertainty.

The second part introduces a method to reduce the computational requirements of Monte Carlo Simulations. Numerical simulation of variably saturated three-dimensional fields is typically a computationally intensive process, let alone Monte Carlo Simulations of such problems. In order to reduce the number of model evaluations while producing acceptable estimates of the output statistical properties, Cluster Analysis (CA) is used to group the input parameter realizations, e.g. hydraulic conductivity. The potentials of this approach are investigated using different: 1) clustering schemes; 2) clustering configurations, and 3) subsampling schemes. . Results show that response of 400 realizations ensemble can be efficiently approximated using selected 50 realizations.

The third part of the research investigates the impact of input parameter uncertainty on the response of irrigation-drainage systems, particularly on crop yield and root zone hydrosalinity. The three-dimensional soil parameters, i.e. hydraulic conductivity, porosity, the pore size distribution (van Genuchten β) parameter, the inverse of the air entry pressure (van Genuchten α) parameter, the residual moisture content parameter, and dispersivity; are treated as spatial random processes. A sequential multivariate Monte Carlo simulation approach is implemented to produce correlated input parameter realizations. Other uncertain parameters that are considered in the study are irrigation application variability, irrigation water salinity, irrigation uniformity,

preferential flow fraction, drain conductance coefficient, and crop yield model parameters. Results show that as the crop sensitivity to salinity increases, the crop yield standard deviation increases.

The fourth part of the research investigates an approach for optimal sampling of multivariate spatial parameters in order to reduce their uncertainty. The Ensemble Kalman Filter is used as instrumentation to integrate the sampling of the hydraulic conductivity and the water level for a two-dimensional steady state problem. The possibility of combining designs for efficient prediction and for efficient geostatistical parameter estimation is also investigated. Moreover, the effect of relative prices of sampled parameters is also investigated. A multi-objective genetic algorithm is employed to solve the formulated integer optimization problem. Results reveal that the multi-objective genetic algorithm constitutes a convenient framework to integrate designs that are efficient for prediction and for geostatistical parameter estimation.

ACKNOWLEDGEMENTS

I would like to greatly thank my Advisor, Dr. Luis Garcia, for his sincerity, encouragement, and continuous support during all stages of this research. I would like also to thank my co-advisor Dr. Timothy Gates, and my committee members Dr. Domenico Bau, and Dr. Greg Butters, for their valuable insights and comments to me throughout my study at Colorado State University.

This work was partially funded by a project from the United States Bureau of Reclamation, and the author is grateful for the help provided by Mr. Roger Burnett, an Agricultural Engineer, with the United States Bureau of Reclamation

Last but not least, I am deeply grateful to my parents, Hadba and Hajjaj, whose continuous prayers to the almighty God, enlightened my way throughout my life. Especially, I would like to give my special thanks to my wife, Ayah, whose patience, support, and inestimable proofreading enabled me to complete this work.

Dedication

This work is dedicated to my mother Hadba, my father Hajjaj, my wife Ayah, my daughter Hanan, and all my six brothers and three sisters.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 General

The ever-increasing world population causes a continuous increase in demand for food and fiber. The disturbance of food production processes could have detrimental consequences for local and global social, economic, and political stability. The world relies heavily on agriculture to secure these needs. Irrigated land nowadays constitutes 20% of the world's cultivated land and produces up to 40% of the food and fiber that humans need (Hoffman, 2007). Therefore, irrigation is the largest consumer of water on earth, accounting for 80% of fresh water diverted for human use (Hoffman, 2007). The vital importance of the irrigation sector is not limited to its direct effects on the crop production industry, but extends to affect the social fabric and the income of communities, especially in arid and semi-arid regions.

Recently, the demand for energy has increased and the cultivation of crops for biofuel production is gaining greater attention. Several sources (for example, The Washington Post, 2011) attribute the food crisis, that affected the world in 2008, to the increased demand on biofuels, which uses crops that otherwise would have been used for human consumption.

Several issues have challenged the use of irrigation water in crop production. For instance, the irrigation sector in many parts of the world faces a real competition

from other water users such as municipal, industrial, and recreational demands. Although the fresh water resources are limited (in terms of quantity, quality and accessibility), the demand continues to increase at a pace that exceeds the supply. The improvement in the living standards in many places around the world produces an un-proportional increase in water demand. Namely, a two-fold increase in population since 1900 induced a six-fold increase in water use (World Water Council, 2011).

Another dimension of challenges facing irrigation is the environmental consequences of current irrigation practices. In particular, diverting water for irrigation use has resulted in negative consequences in water quality, soil erosion, groundwater levels, and stream flow quantity and quality. The changes induced by irrigation activities in hydrological systems are adversely affecting aquatic and riparian ecosystems. As a result, strict environmental regulations aimed at protecting endangered species add more burden to the crop production process.

The general management framework of the irrigation industry can be summarized by the following points (Hoffman et al. 2007):

- Secure the required water supply in terms of the quantity.
- Conserve the soil/water quality by managing the salinity of the soil and water.
- Minimize soil erosion resulting from some types of irrigation activities.
- Maximize crop productivity for the amount and quality of the available water.

- Achieve the above goals within economical costs and in a sustainable manner. Sustainability should be achieved environmentally and economically.

1.2 Salinity and Waterlogging

The deterioration of crop production due to salinity buildup in the root zone is a global problem. The interactions of water with geological salt deposits results in dissolution of salts in water. Using such water in irrigation adds non-native salts to the land where it is applied. Soil evaporation and root extraction of fresh water concentrate the salinity in the root zone. Accordingly, the repetition of this process over a long period results in the emergence of large-scale salinization problems. In order to reduce root zone salinity, extra water is usually applied to leach salinity out from the root zone. Beside the fact that this practice might increase stresses on the limited available water budget, it also results in the development of high saline groundwater tables, especially in fields with low drainage capacity. Thus, the dual impact of soil salinization and waterlogging (shallow water tables) causes a significant reduction in the productivity of crops. The high salinity increases the pressure that crops need to exert in order to extract fresh water, while shallow water table results in reduction of air circulation which cause decay to the roots.

1.3 Site Description

The Lower Arkansas River basin in Colorado is part (Figure 1.1) of the Arkansas River Basin, the sixth longest river in the US. Irrigation activities were

introduced to the region in the 1870's (Miles 1977). More than 1,000 miles of irrigation cannels were constructed to divert water for irrigation purposes (Gates et al. 2006). Irrigation activities are mainly based on open-ditch furrow irrigation, in addition to a limited number of center pivot sprinklers and a few drip systems (Burkhalter et al. 2005). The major crops that are produced in the valley are alfalfa, corn, melons, onions, beans, and wheat (Farm Service Agency 1999-2000). The soil type in the region is alluvial deposits that consist of a silty loam clay layer in the upper surface and loam to sandy loam substrata (USDA, 1971a; USDA, 1971b).

Since the introduction of irrigation to the region, the salinity of the water has been increasing due to the accumulation of salts in the hydrologic system. Extensive data collection, as part of Water Management in the Lower Arkansas Valley project (Gates et al. 2006; Burkhalter et al. 2005), show that the average groundwater table depth in a section of the valley upstream of John Martin Reservoir ranges between 1.21m to 1.65m, and a minimum depth of 0.08m. The average groundwater salinity ranges between 2.6 dS/m to 4.62 dS/m at different times in the season. The overall average extract soil salinity for the same region is around 4.1 dS/m (Morway et al.2011); however, the maximum soil salinity reaches over 20 dS/m. The over irrigation of crops to leach the root zone salinity has resulted in field scale water table increases in several areas of the valley. This increase became a regional problem after construction of John Martin Reservoir in 1948 and Pueblo Reservoir in 1975. The reservoirs were blamed for reducing sedimentation of fine soils in the canals, which reduce bed lining. Therefore, seepage from canals has been increasing, resulting in a shallower water table.

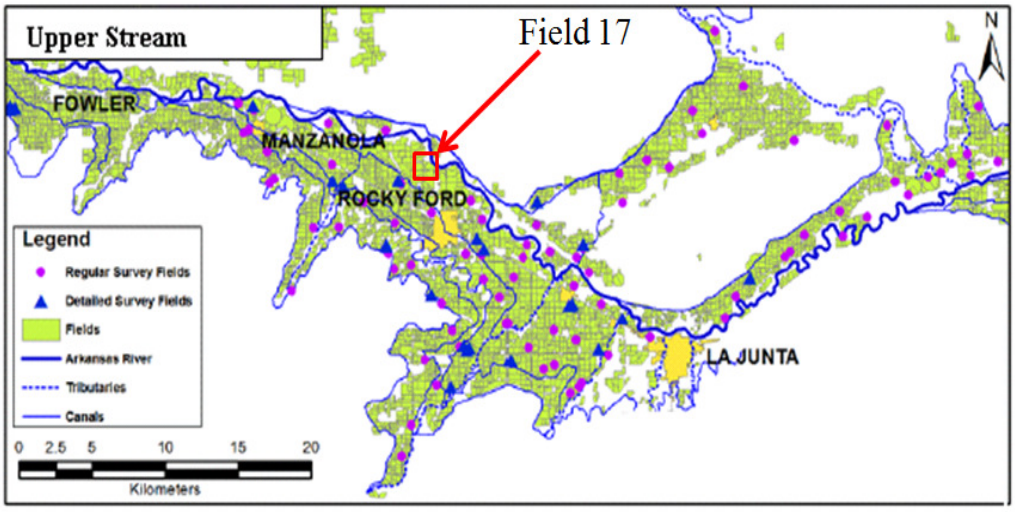

Figure 1.1: General Arial Map of the Lower Arkansas Basin (Modified from Morway et al. 2011)

Figure 1.2: Aerial Photo for Study Site, Field 17, Rocky Ford, CO

1.4 Research Organization

Countering the dual impact of salinity and waterlogging requires a wide scale multidisciplinary intervention. As part of Water Management in the Lower Arkansas Valley Project, an extensive data collection effort was launch to characterize the regional hydrological system and to aid in developing feasible management plans. However, the ability to develop a set of management plans is always hampered by the uncertainty in our knowledge of the characteristics of the hydrologic system. This research is an attempt to cast light on how to approach this problem on a field scale. Regional scale is not included in this study due the computational difficulties associated with large scales.

The research herein is organized into four parts that tackle the uncertainty issue from different angles. These parts are as follow:

- The first part explores and applies a Global Sensitivity Analysis (GSA) for a one-dimensional variably saturated flow and transport problem. The sensitivities of the input parameters to crop yield and root zone hydrosalinity were investigated. Commensurate with the theme of this research, i.e. prediction uncertainty, four GSA techniques are utilized to approach the same problem. The comparative performances of the GSA techniques are studied.
- The second part provides an approach that alleviates the computational burden required by Monte Carlo Simulations. This burden is particularly intensive and time consuming in simulating three-dimensional variably

saturated problems. The approach uses cluster analysis to stratify the ensemble of realizations. Numbers of realizations are selected from each stratum to represent the entire ensemble.

- The third part of the study investigates the effect of input uncertainly on crop yield and root zone hydrosalinity for a three-dimensional field scale problem. Multivariate Monte Carlo simulation of the soil properties is implemented to generate correlated realizations of the input parameters. The statistical properties of crop yield for two crops, i.e. alfalfa and corn, are obtained.
- The fourth part applies a methodology that reduces the parameters uncertainty in the data collection stage. The ensemble Kalman Filter is utilized to provide design criteria that are optimized via a Multi-objective Genetic Algorithm technique. Different design schemes that are efficient for prediction, covariance parameter estimation, and cost are investigated.

This research is intended to improve the understanding of the role that different sources of uncertainty play in soil salinity and waterlogging prediction.

1.5 References

Burkhalter, J.P., Gates, T.K. & others, 2005. Agroecological impacts from salinization and waterlogging in an irrigated river valley. *Journal of Irrigation and Drainage Engineering*, 131, 197.

Gates, T.K., Garcia, L.A. & Labadie, J.W., 2006. Toward Optimal Water Management in Colorado's Lower Arkansas River Valley: Monitoring and Modeling to Enhance Agriculture and Environment. *Colorado Water Resources Research Institute Completion Report*, (205).

Hoffman, G.J. et al., 2007. *Design And Operation Of Farm Irrigation Systems* 2nd ed., American Society of Agricultural & Biological.

Miles, D.L., 1977. Salinity in the Arkansas Valley of Colorado."Interagency Agreement Report EPA-IAG-D4-0544. *Environmental Protection Agency, Denver, Colorado*.

Morway, Eric D., and Timothy K. Gates. 2011. "Regional Assessment of Soil Water Salinity across an Intensively Irrigated River Valley." Journal of Irrigation and Drainage Engineering. doi:10.1061/(ASCE)IR.1943-4774.0000411. http://link.aip.org/link/doi/10.1061/(ASCE)IR.1943-4774.0000411.

The Washington Post. 2011. Tim Searchinger - *How biofuels contribute to the food crisis*. February 11. http://www.washingtonpost.com/wp-dyn/content/article/2011/02/10/AR2011021006323.html.

U.S. Dept. of Agriculture (USDA). 1972a. Soil survey of Otero County, Colorado, USDA, SCS, La Junta, Colo.

U.S. Dept. of Agriculture (USDA). 1972b. Soil survey of Bent County, Colorado, USDA, SCS, La Junta, Colo.

World Water Council, 2011, *Water crisis*, Available online 28 September, 2011 http://www.worldwatercouncil.org/index

2 GLOBAL SENSITIVITY ANALYSIS OF VARIABLY SATURATED FLOW AND TRANSPORT PARAMETERS AND ITS IMPLICATION FOR CROP YIELD AND ROOT ZONE HYDROSALINITY

2.1 General

Modeling of crop yield and root zone hydrosalinity usually requires large number of parameters that are often expensive to obtain and can be associated measurement errors. Consequently, identifying the most relevant parameters, and their contributions to the uncertainty of the output, might be used as a basis to focus research resources in an efficient manner. Global Sensitivity Analysis (GSA) is a powerful tool that can be employed to achieve this goal. However, some GSA methods perform better than others, which introduces the risk of mistakenly prioritizing a secondary parameter while neglecting a primary one. This paper evaluates the usage of four GSA methods to rank the importance of input parameters with respect to five performance indices, which summarize the output of a flow and transport model, and its implication for the root zone hydrosalinity and crop yield. Results show that 73% of crop yield variance is controlled by only five of eighteen parameters that were considered in this study. Moreover, it was found that the van Genuchten pore size parameter is very important to the relative crop yield prediction and the water availability index. In addition, it was

found that the variance decomposing method, the screening method, and the Monte Carlo Filtering method are generally consistent in their performance; whereas the partial correlation coefficient method is significantly different.

2.2 Introduction

Since the advent of physically based numerical models in hydrology in the early 1960's, their ability to simulate reality has been limited. This has been mainly due to the lack of the field data required to justify a representative conceptual model of the system under study, and to the limited information available about the controlling parameters. Furthermore, the amount and accuracy of the parameters' field data are always constrained by budget and regulatory pressures. Therefore, hydrologists usually find themselves faced with ill-posed hydrological problems where the size of the available information is not enough to produce a unique prediction (Beven and Binley 1992).

In spite of this chronic problem, numerical models are still vital tools in most research efforts, as well as in most regulatory settings. A careful use of numerical models should be based on a good understanding of the model's input-output dynamics. Such a relationship can be efficiently revealed by sensitivity analysis investigations. Classically, the One At a Time (OAT) local derivative-based sensitivity analysis was used in groundwater modeling as a diagnostic tool of the models (Anderson and Woessner 1992), in calibration of groundwater models (Hill and Tiedeman 2007), and in optimization of groundwater systems. For instance, local sensitivity of a certain parameter is simply obtained by calculating the derivative of the output with respect to

one input parameter at a specified base point. Such measure does not mirror the prior knowledge of input factors and is not efficient when dealing with complex nonlinear models. Moreover, sometimes it is essential to establish the *comparative importance* of input parameters with respect to a predefined output index.

Global Sensitivity Analysis (GSA) (Saltelli et al. 2008b) provides an attractive alternative to the local derivative-based sensitivity approach, in which the input-output relationship can be established in light of the prior uncertainty of the input factors. The analysis is *'global'* in the sense that it covers the whole uncertainty space of the input parameters, and apportions the uncertainty in the output indices to the uncertainty in the input factors. A number of approaches have recently emerged to compute the global sensitivity measures without computing the derivatives, for example, the sampling-based method (Helton et al. 2006), the Bayesian method (Oakley and O'Hagan 2004), and variance decomposing methods (Sobol 2001; Homma and Saltelli 1996; Saltelli et al. 2008) . Another valuable method is the screening method (Morris 1991; Campolongo et al. 2007), which can be seen as a global sensitivity method despite its derivative-based root. Using this method, the mean and standard deviation of the derivatives of the output with respect to a sample of the input parameters are used as sensitivity measures.

Typically, crop prediction models are established by simulating the complex plant-soil-climate system. These models are usually highly parameterized; and the estimation of parameters is an expensive process and is highly prone to errors (Varella et al. 2010). Uncovering the relative importance of the different factors allows the focusing of research resources on factors that make a major contribution to the output

variance. Number of studies applied the global sensitivity analysis to groundwater problems. Specifically, Pan et al. (2011) studied the sensitivity of unsaturated flow and contaminant transport parameters. Mishra et al. (2009) reviewed the application of three GSA techniques to groundwater problems and their practical implications. Recently, several studies have been conducted to study the sensitivity of crop models, for example (Ruget et al. 2002; Jongschaap 2007; Makowski et al. 2006; Pathak et al. 2007; Varella et al. 2010).

In this study, the relative importance of the input parameters, which control flow and transport in a variably saturated soil, and their impacts on crop yield and hydrosalinity of the root zone are investigated. Eighteen different input factors were considered as input parameters to the Colorado State University Irrigation and Drainage (CSUID) Model, a three-dimensional finite difference model (Alzraiee and Garcia 2009). In this paper, the Global Sensitivity Analysis (GSA) is structured in a manner that fulfills the following major objectives:

- Determine the relative importance of input parameters with respect to different root zone hydrological processes, and with respect to different GSA goals.
- Highlight the differences among global sensitivity methods by comparing their performances.

In order to achieve these goals, four GSA experiments were conducted using variance decomposing method, Morris method, partial correlation method, and Monte Carlo filtering method. These methods are used to calculate the sensitivities measures for five performance indices. The indices are formulated to reflect a specified

hydrological process. These indices are Relative Crop Yield (RCY), Water Availability Index (WAI), Water Excess Index (WEI), Root Zone Salinity Index (SI), and Deep Percolation Index (DPI).

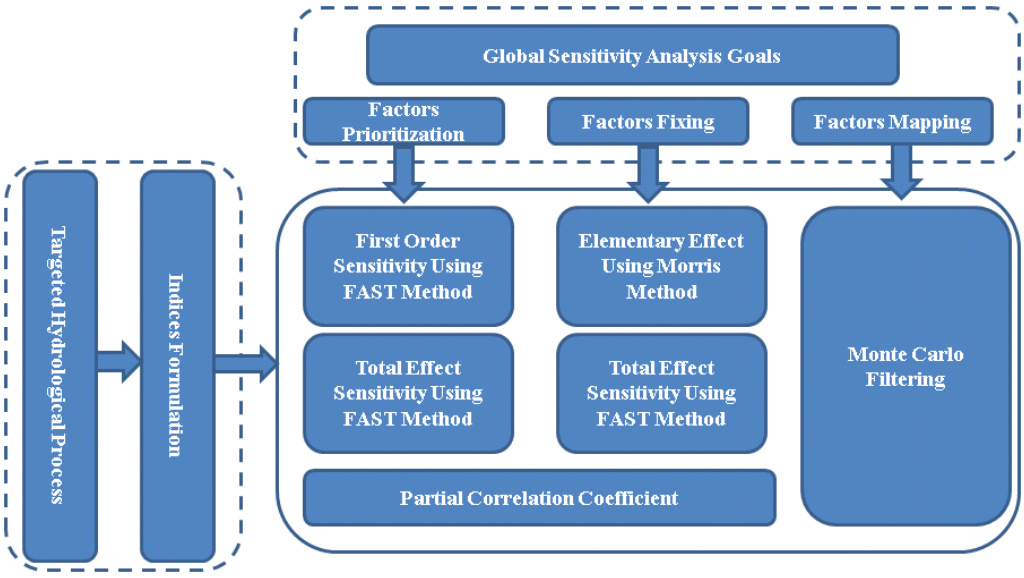

Figure 2.1: Illustration of the General GSA Framework

2.3 Methodology

The general approach adopted herein is illustrated in Figure 2.1. As shown in the figure, prior to the application of GSA, it is required to identify, firstly, the targeted hydrological process of concern to the modeler, e.g. crop yield prediction, root zone salinity prediction, among others; and, secondly, the objectives of GSA study, e.g. factor prioritization, factor fixing, among others. These two decisions are discussed in the following sections.

2.3.1 Targeted Hydrological Process

Modeling variably saturated flow and transport of subsurface systems is usually motivated by different research and regulatory objectives. To illustrate, models can be used to predict the crop productivity for fields (Feddes et al. 1976; Xevi et al. 1996), to estimate the deep percolation of pesticides and fertilization, to assess reclamation of saline soils, or to evaluate different irrigation designs. This array of modeling motivations might be used to formulate the indices that reflect the performance of targeted hydrological process.

Normally flow and transport numerical models generate a large amount of outputs, i.e. at every numerical cell and at each time step. These outputs can be interpreted differently by modelers according to their objectives. A summary of the output is essential to facilitate the sensitivity analysis. That is, a scalar-valued index should be formulated to accurately measure the performance of each of the modeling objectives. As an example, the spatio-temporal mean of moisture content in the root zone might be used as an index of water deficit stress. The four targeted hydrological processes of concern to this paper are:

- Predicting the relative crop yield,
- Predicting the possibility of water deficit stress throughout the growth season,
- Predicting the possibility of water excess stress (waterlogging) ,
- Predicting the root zone salinity.

- Predicting the vertical flux entering or leaving the root zone.

2.3.2 Global Sensitivity Analysis Goals

This step answers the question about the objectives of the GSA study. A major component of the GSA framework (Figure 2.1) is to define precisely the objectives of the sensitivity analysis in advance of the analysis. Saltelli et al. (2008b) suggested a general framework, or setting as they called it, to define the objectives of the sensitivity analysis and, thus, determine the most suitable sensitivity method. In this study, the framework proposed by Saltelli (2008) is used, which can be summarized as follows:

- The *Factor Prioritization* setting is a GSA framework in which the objective of the analysis is to identify factors that when fixed to their true values result in the greatest reduction in variance of the output. For instance, to guide data collection, the Factor Prioritization setting could be the basis upon which the important factors are determined and then intensively and accurately sampled. The extended Fourier Amplitude Sensitivity Test (FAST) is used herein to rank different parameters according to their First Order Sensitivities (FOS).
- The *Factor Fixing* setting identifies input factors which if fixed at any value in their range, do not produce significant change in the output. This setting can be used to reduce the complexity of models by choosing a small number of factors without impacting the output (Parsimony Principle). The Screening Method (Morris 1991) and the total effect coefficients resulting from the extended Fourier Amplitude Sensitivity Test (FAST) are obtained

to achieve this objective. Using the screening method is computationally more efficient, since it requires a smaller number of model evolutions.

- The *Factor Mapping* (FM) setting is used in cases when the decision maker is interested in a certain region in the index's CDF, e.g. rare events, extreme contamination, among others. The Monte Carlo Filtering method is employed for this purpose. The method is capable of identifying regions in the input space that produce a certain region in the output space.

2.4 Model Description

In this section, the theory of flow and transport model used herein is introduced. The numerical prediction of crop productivity depends on several factors; among them are the hydrosalinity conditions of the root zone. In this study, it is assumed that the agronomic conditions are excellent, and the only limiting factors are the soil hydrosalinity conditions. The spatio-temporal status of the water content and salinity are mathematically described using the continuity partial differential equation of flow; i.e. modified Richard's equation (Eq. 2.1) and the dispersion-advection partial differential equation (Eq. 2.2), respectively.

$$\frac{\partial}{\partial x_i}\left(K_i(\psi)\frac{\partial h}{\partial x_i}\right) + Q_s = \left(\frac{\theta}{\theta_s}S_s + C(\psi)\right)\frac{\partial h}{\partial t} \tag{2.1}$$

$$\frac{\partial}{\partial x_i}\left(\theta D_{ij}\frac{\partial C}{\partial x_j}\right) - \frac{\partial}{\partial x_i}(\theta v_i C) + Q_s C_s = \frac{\partial(\theta C)}{\partial t} \tag{2.2}$$

Where $K_i(\psi)$ is the hydraulic conductivity [L/T], ψ is the capillary head [L], h is the total head [L] ($\psi = h - z$), Q_s is the sink or source term per unit volume [T^{-1}], θ is the

moisture content [L^3/L^3], θ_s is the soil porosity [L^3/L^3], S_s is the specific storage [L^{-1}], $C(\psi)$ is the specific capacity [L^{-1}], x is a space vector [L] and $i = 1,2,3$ represents three-dimensional space, t is time [T]. D_{ij} is the hydrodynamic dispersion [L/T^2], C is the salinity concentration [M/L^3], and v_i is the seepage velocity [L/T].

Solving equations 2.1 and 2.2 requires the knowledge of the constitutive relationship between moisture content and capillary head which are modeled via the van Genuchten (1980) model in equation (2.3).

$$\boldsymbol{\theta(\psi) = \theta_r + \frac{\theta_s - \theta_r}{(1+(\alpha|\psi|)^{\beta})^{1-\frac{1}{\beta}}}} \tag{2.3}$$

Where θ_r is the residual moisture content [L^3/L^3], α is a fitting parameter related to the inverse of the air entry suction, $\alpha > 0$ [L^{-1}], β is a measure of the pore size distribution, $\beta > 1$.

The sinks/sources term Q_s is the summation of the irrigation and root extraction rates (Eq. 2.4), where Q_i is the irrigation rate [L^3T^{-1}], which is a model input parameter (positive value), Q_r is the root uptake rate [L^3T^{-1}] (negative value); and ΔV is the cell volume.

$$Q_s = \frac{(Q_r + Q_i)}{\Delta V} \tag{2.4}$$

While the irrigation application rate is an input parameter, the root extraction rate is internally computed according to equations 2.5 to 2. 8. The overall sink term that accounts for root density and geometry, water matric and osmotic pressure and root

growth stage are summarized in equation 2.5 (Hopmans and Bristow 2002; Feddes et al. 1976).

$$\boldsymbol{Q_r(z,t) = \lambda(z,t).\frac{ET(t)}{\Delta A}.\alpha(\psi,\psi_o).C_o}\ , \tag{2.5}$$

where $ET(t)$ is the reference evaporation [L/T], C_o is the crop growth coefficient at time t, $\Delta \boldsymbol{A}$ is area [L^2]; and $\lambda(z,t)$ is the root density equation that describes the density and the geometry of the root network with respect to the depth and is calculated using the S function as described in equation (2.6).

$$\boldsymbol{\lambda(z,t) = \frac{-1.6z}{D(t)^2} + \frac{1.8}{D(t)}}\ , \tag{2.6}$$

where z is the depth at which the root density is calculated [L]; and $D(t)$ is the root depth at current time [L]. The temporal root growth can be approximated using the Hanks and Hill (1980) equation (2.7).

$$\boldsymbol{D(t) = \frac{D_{max}}{\left(1+\exp\left(a-b\frac{t}{t'}\right)\right)}}\ , \tag{2.7}$$

where $D(t)$ is the root depth at time t, D_{max} is the maximum root depth, t' is the end of the third stage of the crop's growth, and a and b are empirical coefficients.

Van Genuchten (1987) pioneered describing the sink term as a function of the water content and extended it to incorporate the osmotic head. In this paper, the Cardon and Letey (1992) equation was modified, which is a slight modification of the Feddes et al. (1976) equation, to account for root uptake reduction due to waterlogging. Equation 2.8 is the final equation that accounts for water deficit stress, salinity stress and water excess stress (waterlogging).

$$\alpha(\psi,\psi_o)=\begin{cases}\dfrac{1}{1+\left(\dfrac{\psi}{\psi_{50}}+\dfrac{\psi_o}{\psi_{o50}}\right)^{P}} & \psi(z,t)<\psi_s \qquad (2.8.a)\\ \dfrac{\left(\dfrac{\psi}{\psi_s}\right)}{1+\left(\dfrac{\psi}{\psi_{50}}+\dfrac{\psi_o}{\psi_{o50}}\right)^{P}} & 0\geq\psi_s>\psi(z,t) \qquad (2.8.b)\end{cases}$$

where p is a parameter close to 3, ψ_{50} is the capillary head at which the root uptake is reduced by 50% and $\psi_o = 0$ [L]; $\boldsymbol{\psi_{o50}}$ is the osmotic head at which root uptake is reduced by 50% and $\psi = 0$[L]; $\psi(z,t)$ [L] is the capillary head [L]; $\psi_o(z,t)$ is the osmotic head [L]; ψ_s is the head threshold after which oxygen deficiency starts to occur[L]. It is recognized that the water excess stress (near saturation cases) does not affect the root uptake instantaneously (Harbaugh et al. 2000), but could take the crop a few days (for example, 2 days) to affect the root uptake. As a result, Equation 2.8.b will not be active until the matric head is equal or above ψ_s for a period of two days.

The total actual evapotranspiration ET_a is approximated by integrating the temporal extraction rate over the growing season and over the root zone depth (Equation. 2.9).

$$\boldsymbol{ET_a=\int_0^T\int_0^{D(t)}Q_r(z,t)\,dzdt} \qquad (2.9)$$

where $Q_r(z,t)$ is the temporal root extraction [L^3/T] at a vertical depth z per unit soil volume, T is the growing season [T], D is the root depth [L] at time t.

Finally, the relative crop yield is approximated using equation (2.10) which is based on the assumption of a linear relationship between relative evapotranspiration and relative crop yield (Doorenbos et al. 1986)

.

$$\boldsymbol{RCY = \frac{Y_a}{Y_m} = 1 - k_y\left(1 - \frac{ET_a}{ET_m}\right)} \quad (2.10)$$

Where Y_a is the actual dry matter yield [M], Y_m is the maximum harvested dry matter yield [M], k_y is the yield response factor, ET_a is the total (seasonal) actual evapotranpiration [L], and ET_m is the reference evapotranpiration which can be obtained from climatic data [L].

2.5 Formulation of Indices

The CSUID model, described in section 3, calculates the temporal and spatial variability of the moisture content and salt concentration at each numerical cell. These two variables are the major factors affecting the relative crop yield calculation as shown in equations 2.5 to 2.10.

One of the objectives of the decision maker might be to maximize crop production while reducing long-term environmental risks, e.g. reduction of deep percolation of pesticides and fertilizers. Another concern for decision makers is to maintain the sustainability of the crop production processes, i.e. preventing root zone salinization. In order to put these goals in quantitative measures, the following indices are set up:

1- Relative Crop Yield Index (RCY): This index is the numerical value of the simulated relative crop yield as shown in equation (2.10). This prediction is the resultant of temporal and spatial variability of moisture and salinity along the vertical dimensions of the root zone and throughout the growing season of the crop.

2- Water Availability Index (WAI): The availability of moisture in the root zone is a major factor that is used to evaluate the water deficit conditions. Since the moisture content in the root zone is spatially and temporally variable, the mathematical integration of the Readily Available Water (RAW) over the root zone depth and over the growing season is used as an index (Equation 2.11). By definition, RAW is the moisture available to the plant to extract. Equation (2.12) is widely used to calculate it.

$$WAI = \frac{1}{T.D}\int_{t=0}^{t=T}\int_{z=0}^{z=D} RAW(t,z)\,dz\,dt \tag{2.11}$$

$$RAW(t,z) = \theta(t,z) - \theta_w(z) - 0.5(\theta_{FC}(z) - \theta_w(z)) \tag{2.12}$$

where $\theta(t,z)$ is the moisture content, $\theta_w(z)$ is the water content at the wilting point (or water content at capillary head of -15,300cm (Meyer et al. 1997)), and $\theta_{FC}(z)$ is the moisture content at field capacity (or water content at capillary head of -340cm).

3- Water Excess Index (WEI): This index measures the waterlogging condition of the soil profile. In other words, it calculates the total time that the moisture content of the root zone is close to saturation as a percentage of the total simulation time. The correlation between this index and crop yield can be used to estimate the contribution of excess water stress in crop yield reduction.

$$I_E(t) = \begin{cases} 1 & \theta(t) > 0.95\theta_s \\ 0 & else \end{cases} \tag{2.13}$$

$$WEI = \frac{\sum_{t=0}^{t=T} I_E(t).\Delta t}{T} \tag{2.14}$$

4- Root Zone Salinity Index (SI): This index measures the average salinity status of the root zone during the simulation period (Eq. 2.15).

$$SI = \frac{1}{T.D}\int_{t=0}^{t=T}\int_{z=0}^{z=D} C(t,z)\, dz\, dt \qquad (2.15)$$

5- Deep Percolation Index (DPI): Knowing the deep percolation fraction is of great importance for several reasons. The index is the cumulative temporal deep percolation. Note that the index can be positive when the net vertical flux is out of the root zone or negative when the net vertical flux is into the root zone. This index can be used to measure the contribution of groundwater to subirrigation and to upflux of salt. In addition, the design of subsurface drainage system requires the knowledge of the deep percolation fraction, where large deep percolation requires smaller drain spacing. Moreover, the transport of pesticides and fertilizers as well as salts carried by deep percolation into the groundwater might be an environmental concern.

Beside the previously mentioned indices, it is required to identify parameters responsible for a particular region in the index's CDF (Factor Mapping). These regions are usually of particular importance to decision makers. In this paper, the following thresholds in the indices' CDFs are defined:

- The relative crop yield $\leq 40\%$, used as a definition of low relative crop yield RCY.
- The highest 30% region of the root zone salinity index (SI) CDF, used as a definition of extreme salinization.
- The lowest 20% region of the CDF of Water Availability Index, used to define the dry root zone conditions.

2.6 Global Sensitivity Analysis Review

Four types of sensitivity analysis methods were utilized in this research, and all of them are sampling-based methods. The sampled input factors were assumed statistically uncorrelated. The general Global Sensitivity Analysis procedures for each index can be outlined as follows:

- Define the sensitivity analysis objectives or settings; e.g. Factor Prioritization, Factor Fixing or Factor Mapping.
- Choose the proper sensitivity analysis method given the GSA objective (Figure 2.1).
- Define a probability distribution function (PDF) for each input parameter, which reflects the degree in uncertainty in the parameter value.
- Choose a sampling scheme (e.g. Monte Carlo, Latin Hyper Cubic sampling, etc.), and generate a sample for each of the input factors.
- Evaluate the model using the generated samples.
- Use the produced input-output set to calculate the sensitivity coefficients using one of the GSA methods.

The following sections, introduce a brief description of the theoretical background of each of the GSA methods used in this study.

2.6.1 Variance Decomposing Method

The generic model (Eq. 2.16), describe the relation between the output (***Y)*** and number ***n*** of uncertain factors x_i, where $(i = 1, \dots, n)$. No assumption is made regarding the complexity of the model, namely, whether the function is linear, monotonic, or additive.

$$Y = f(x_1, x_2, x_3 \dots, x_n) \tag{2.16}$$

The variance decomposing method suggests the decomposing of the variance of the output as a set of terms of increasing dimensionality. Sobol (1993) provided a straight forward Monte Carlo-based implementation of this method. In other words, a Monte Carlo sampling scheme can be set to draw realizations from the joint distribution $f(x_1, x_2, .. x_n)$, and consequently, each of the realizations is evaluated in the model (Eq. 2.16) to obtain the response Y. Next, the variance of the output Y is decomposed into 2^n terms according to equation (2.17).

$$V(Y) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + \cdots + V_{1,2,..n} \tag{2.17}$$

Where $V(Y)$ is the output variance, $V_i = V[E(Y \mid x_i)]$ is the first order (one-way) effect of x_i, and $V_{ij} = V[E(Y \mid x_i, x_j)] - V_i - V_j$ is the second order (two-way) effect. By dividing both sides of the equation by $V(Y)$, a normalized sensitivity measure can be obtained (Eq. 2.18).

$$\sum_i S_i + \sum_i \sum_j S_{ij} + \ldots + S_{1,2,\ldots,n} = 1 \tag{2.18}$$

Where S_i is the first order sensitivity measure of parameter x_i and S_{ij} is the second order sensitivity measure of interacting parameters x_i and x_j.

Computing all of the sensitivity measures could be computationally prohibitive. For example, calculating Sobol's indices requires $N(2^n + 1)$ model evaluations, where N is the sample size used to estimate one individual effect.

As an alternative, the Fourier Amplitude Sensitivity Test (FAST) (Cukier et al. 1973; Saltelli and Bolado 1998) provides a computationally affordable approach to calculate the total effect of each of the parameters and the first order effect. For instance, the total effect of any factor x_i is the summation of all sensitivity measures in equation (2.18) that corresponds to $\boldsymbol{i}$ and can be computed as

$$STi = 1 - \frac{V[E(Y \mid X_{\sim i})]}{V(y)} \tag{2.19}$$

where $X_{\sim i}$ means all factors except i. The extended Fourier Amplitude Sensitivity Test (FAST) within the SIMLAB package (Giglioli and Saltelli 2000) was used to calculate the first and the total effect of each factor.

2.6.2 Elementary Effect Method (EEM)

Morris (1991) proposed a One At a Time (OAT) sensitivity analysis method. The method is known for its effective identification of a few important factors in models that have many factors. For the model in equation (2.16), the elementary effect

(EE) of any parameter x_i is computed using approximations of the derivative of the response function at a base point.

$$EE_i = \frac{[f(x_1,....x_i + \Delta,...x_n) - f(x_1,....x_i,...x_n)]}{\Delta} \tag{2.20}$$

The EE_i values are computed using $p^{n-1}[p - \Delta(p-1)]$ elementary effects, where p is a preselected number of levels that divide the parameter space, and $\Delta = p/(2(p-1))$. The mean μ_i of the elementary effects is a measure of the importance of the factor, while the standard deviation σ_i determines whether the response function is nonlinear in factor *i* and/or if the factor is interacting with other factors. Campolongo et al. (2007) suggested replacing the mean μ_i with μ_i^*, the mean of the absolute values of EE_i.

2.6.3 Monte Carlo Filtering

The Monte Carlo Filtering Method is used within the context of the Factor Mapping (FM) framework. Sometimes it is important to recognize the input factors that produce a targeted region in the output space. In other words, the realizations of the input factors are categorized 'filtered' as behavioral and non-behavioral depending on whether the realization produces an output value within the targeted region or not. The discrepancy between the cumulative distribution function CDF of the behavioral $F(X_i|B)$ and non-behavioral $F(X_i|\bar{B})$ realizations is used to accept or reject the hypothesis regarding the importance of the factor. To illustrate, if the two CDFs are significantly different, then this implies that the factor plays a major role in producing

the targeted region in the output. The statistical test used is the Smirnov two-sample test (Eq. 2.21).

$$D(X_i) = \max \ |F(X_i|B) - F(X_i|\bar{B})| \tag{2.21}$$

2.6.4 Partial Correlation Coefficient (PCC)

This coefficient is a regression-based sensitivity measure (Helton 1993). Efficient usage of this measure requires a linear model; however, for nonlinear models, a rank transformation might be an effective linearizing technique. Conceptually, a linear response surface is fitted between the input and output, and then a sensitivity analysis is performed on this fitted model. The PCC can be computed using the following equation:

$$r_{x_jY} = \frac{\sum_{i=1}^{m}(x_i - \bar{x_j})(Y_i - \bar{Y})}{\left[\sum_{i=1}^{m}(x_{ij} - \bar{x_j})^2\right]^{\frac{1}{2}}\left[\sum_{i=1}^{m}(Y_i - \bar{Y})^2\right]^{\frac{1}{2}}} \tag{2.22}$$

2.7 Method Application

Four numerical experiments are conducted using the four GSA methods outlined in section 5. Each of these methods was used to calculate the sensitivity of input parameters with respect to the five indices described in (section 2.5).

2.7.1 Model Settings

Because of the nonlinearity of the Richard's flow equation, small cell sizes and small time steps are required to aid in the convergence of the solution. This, obviously, makes the simulation computationally extensive and a time-consuming process. On the other hand, the implementation of the GSA requires a large number of model runs. Therefore, it was found to be extremely difficult to conduct the GSA study on the three-dimensional space for this example. Instead, the numerical model was simplified to a one-dimensional vertical column (Figure 2.2).

The number of vertical layers was set to 30 with a thickness of 10 cm each. The growth period of a hypothetical crop was presumed to be 30 days. The model evaluates the time steps internally, and they range between a maximum time step of 0.01 days to a minimum time step of 0.00001 days.

A root zone depth of 0.5 m was assumed. The initial growth stage was taken as 10 days, the development stage as 10 days, the middle stage as 5 days, and the late stage as 5 days. The crop coefficients were presumed constant and equal to unity for the initial, middle and late crop coefficients in order to simplify the evaluation of the results. The initial moisture content was chosen to represent the equilibrium moisture content.

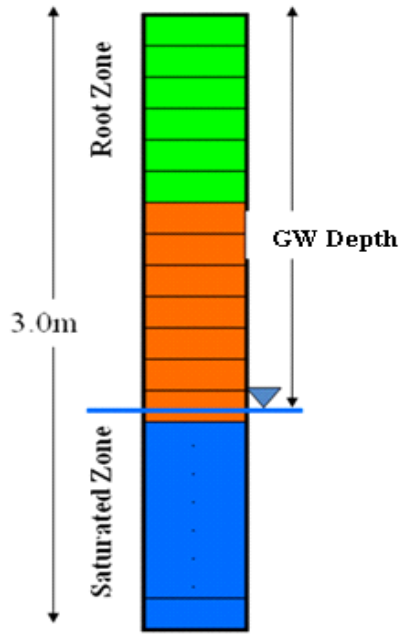

Figure 2.2: Schematic Illustration of the One-dimensional Soil Profile

2.7.2 Input Factors

Eighteen input factors were considered for the GSA study, and a prior statistical distribution was assigned to each of the factors, as shown in Table 2.1. The soils properties were obtained from the soil database provided by Schaap et al. (2001) and from field measurements obtained from Field 17, near Rocky Ford, Colorado. The distributions were truncated to prevent the sampling of computationally or physically unacceptable large or small values. As an example, the normal distribution of the Van Genuchten pore size parameter was truncated at the right side by 30% to prevent sampling of values less than or equal to 1.

The irrigation application depth and salinity statistical properties were chosen to be consistent with irrigation conditions in Field 17. A record of the total volume of water diverted to Field 17 was used to set up the statistical properties of the irrigation depth while the nearby canal water salinity data was used to define the statistical properties of the applied water salinity.

To randomize the pattern of irrigation practices, the value of the frequency of irrigation events was randomly sampled from a discrete distribution, 3 to 8 events. After the sampling of the number of irrigation events, the irrigation date was determined so that the irrigation events were evenly distributed over the season. Next, the total applied water depth was divided by the number of irrigation events to determine the application depth per event.

The initial groundwater depth was presumed to range between 0.5 m to 2.5 m, which was the range of the groundwater depth measured in Field 17. The water table can be seen as a controllable parameter, rather than a system parameter, where it can be controlled by the subsurface drainage system. The initial salinity of the water phase was determined by field measurements.

The statistical properties of the reference evapotranspiration statistics were obtained by processing the climatic data for the Colorado Meteorological Agricultural Network at Rocky Ford for the period May 15th to June 15th, 2010 and was modeled using a normal distribution with a mean evapotranspiration value of 7.2 mm and a standard deviation of 1.3 mm.

The crop yield model parameters were obtained by calibrating the model to the dry biomass of alfalfa from Field 17 as well as from published literature such as Veenhof et al. (1994); Cardon et al. (1992); and Shalhevet et al. (1986).

2.7.3 Sensitivity's Coefficient Computation

The SIMLAB sensitivity and uncertainty package (Giglioli et al. 2000a) interfaced with MATLAB was used to automate the analysis. The CSUID model used the randomly generated input factors (Table 2.1) to calculate the temporal water content, capillary heads, salt concentrations and sink terms at each time step and at all numerical nodes. Consequently, the performance indices are calculated using equations 2.10 to 2.15.

The SIMLAB package internally determines the required number of samples. The Extended FAST variance decomposing method, for example, requires 1,170 model evaluations or 65 runs per parameter for a one dimensional example. On the other hand, the random samples used for the screening method are obtained by the optimal sampling scheme suggested by Morris (1991) and implemented in the SIMLAB package (Giglioli and Saltelli 2000a). Thus, the input space for each input factor was divided into 8 levels; and 30 trajectories were used. The eight levels correspond to the 6.25^{th}, 18.75^{th}, 31.25^{th}, 43.75^{th}, 56.25^{th}, 68.75^{th}, 81.25^{th} and 93.75^{th} quantiles of the input factor CDF. The number of trajectories ($\boldsymbol{r}$) is the number of successive points starting from a random initial vector of input factors, where two successive elements differ only at one component (Giglioli and Saltelli 2000a). Accordingly using $\boldsymbol{k}$ input

factors (***k=18*** in this study), the total model evaluations required should be ***r(k+1),*** which results in 570 model runs.

To reduce the computational requirement, the combined input-output sets for both the Extended FAST and Morris methods were used in the regression-based sensitivity, as well as in the Monte Carlo filtering.

2.8 Results and Discussion

The results of the four experiments are presented in this section. The sensitivity coefficients for each of the five indices (Section 2.5) with respect to the 18 input factors (Table 2.1) were obtained using four sensitivity techniques. The results of analysis are presented for each performance index separately. This is in order to obtain the rank of factors for each performance index, and also to compare the performance of each GSA techniques at the same index.

Table 2.1 Statistical Distributions of Input Factors

Input Factor	Symbol	Prior PDF	PDF Parameters	Right Truncation	Left Truncation
Hydraulic Conductivity, log[cm/day]	K	LogN	$\mu = -0.20, \sigma = 0.80$	5%	95%
Porosity	θs	Normal	$\mu = 0.41, \sigma = 0.073$	5%	95%
Residual Moisture Content	θr	Normal	$\mu = 0.11, \sigma = 0.068$	5%	95%
VG Air Entry Parameter [1/cm]	α	Normal	$\mu = 0.048, \sigma = 0.015$	5%	95%
VG Pore Size Parameter	β	Normal	$\mu = 1.7, \sigma = 1.05$	30%	95%
Dispersivity [cm]	γ	Normal	$\mu = 0.54, \sigma = 0.28$	5%	95%
Root Growth Parameter 1	a	Uniform	L = 1, U=10	–	–
Root Growth Parameter 2	b	Uniform	L = 1, U=10	–	–
Crop Yield Model suction Parameter [cm]	ψ_{50}	Uniform	L = -3000, U=-800	–	–
Crop Yield Model Salinity Parameter [cm]	ψo_{50}	Uniform	L = -7000, U=-1000	–	–
Crop Yield Model Exponent Parameter	p	Uniform	L = 2, U= 4	–	–
Crop Model Waterlogging Parameter [cm]	ψs	Uniform	L = -30, U = -1	–	–
Irrigation Depth [cm]	Id	Normal	$\mu = 5, \sigma = 1$	1%	99%
Salinity of Irrigation Water [mg/L]	Is	Normal	$\mu = 456, \sigma = 109$	1%	99%
Number of Irrigation Events	If	Discreet Uniform	L = 3, U = 8	–	–
Initial Water Table Level [m]	wt	Uniform	L=0.5, U= 2.5	–	–
Initial Profile Salinity[mg/L]	si	Normal	$\mu = 1200, \sigma = 300$	2%	98%
Reference Evapotranspiration [mm]	ET	Normal	$\mu = 7.2, \sigma = 1.3$	1%	99%

2.8.1 Sensitivity of the Relative Crop Yield

The first order sensitivity coefficients were determined and ranked using the Extended FAST variance decomposing method (Figure 2.4). The results show that the key parameters controlling the RCY are $\beta, ET, wt, and\ \theta_s$. The value of the first order sensitivity analysis resides in its usage within the context of the factor prioritization framework, where the objective of the decision maker is to determine factors that need further inspection.

Although the first order sensitivity provides us with the main influence of the parameters, it is limited in determining the overall importance of each parameter. The Total Effect, estimated using the Extended FAST method, is an economical method to reveal the overall importance of different factors. Results show (Figure 2.3) that only five factors are responsible for 73% of the relative crop yield variance. These factors, in order of importance are: $\beta, ET, wt, \theta_s, and\ b$. Interestingly, none of the crop yield model parameters is among them. Of particular importance for the crop yield numerical prediction is the van Genuchten pore-size distribution factor (β). The results of the higher order effects of the parameters (Table 2.2) show that only 22.9% of β's total effect comes from the interaction with other parameters while the remaining 78.1% comes from the first order sensitivity. Similar results were noticed for the ET and wt factors. In contrast, the majority of the influence of the θ_s and b factors occurred as a product of the interaction with other parameters; their high order sensitivity measures are 66.7% and 86.5%, respectively.

While the previous first order and higher order sensitivity results are valuable in the context of the factor prioritization setting, it is important for numerical modelers to exclude factors that are none influential (Factor Prioritization setting) from any uncertainty analysis (the principle of parsimony). The screening method is ideal for such goals, given its relatively low computational demand. The average elementary effect (μ^*) and it's standard deviation σ are presented as bar charts in Figure (2.5). Although $\beta\ and\ ET$ preserve their high ranking, it can be seen that the irrigation depth (Id), van Genuchten air entry α and hydraulic conductivity K factors look more important in the ranked list. An interesting result is that the $\boldsymbol{b}$ factor occupies the 14th position in the Morris method while it occupies the 5th position in the extended FAST method.

The PCC sensitivities (Table 2.3) show significantly different ranking results except for β which maintained its ranking as the most important parameter. The root growth parameters and crop yield model parameters ($a, b, \psi_{50}, and\ \psi_s$) are ranked high. The PCC measures are more effective for linear models, which is not the case for the nonlinear Richard's flow equation. Rank transformation was suggested by Iman and Conover (1979) to reduce the effect of the nonlinearity, however; due to the independence of input factors, the Partial Rank Correlation Coefficient (PRCC) and the PCC are the same.

To identify the input factors responsible for low crop yield, the relative crop yields of less than 40% were mapped to the input factor space. The input factors were categorized as behavioral or non-behavioral based on this condition. The Simonov two-sample test was used to test whether the behavioral or non-behavioral CDFs were

identical. Table 4 shows the Simonov *D*-statistics as computed with Equation 2.21. The factors $\beta, Id, ET, \psi_{o50}\ and\ wt$ contribute highly to the low RCY. It is worth noticing that the factor ψ_{o50} (osmotic head at which the root extraction was reduced by 50%) plays a significant role in the low RCY while it has a small importance in general.

A general notion about the RCY index is that the van Genuchten pore-size parameter (β) is, by far, the most influential factor in the numerical simulation of crop yield. Actually, this is not a surprise because β is the exponent in the van Genuchten equation (2.3).

2.8.2 Sensitivity of the Water Availability Index (WAI)

Estimating the available water in the root zone provides a numerical basis for verifying the proposed irrigation design, specifically the application depth and event frequency. Seven parameters were found to account for 78% of the WAI index variance. The water table factor occupies the top of the list (22% of the WAI variance); this shows the importance of subirrigation in providing moisture to the crop regardless of irrigation design efficiency. The hydraulic conductivity and β are second and third, respectively, on the list and their major impact is through retaining moisture in the soil profile and enhancing the upflux of water into the root zone. The water table level and the β affect the WAI index mostly through their first order sensitivity (FO = 83.35% and 84.13% of the total effect, respectively), whereas the hydraulic conductivity effect take place primarily from interaction with other parameters (FO = 26.9% of the total effect).

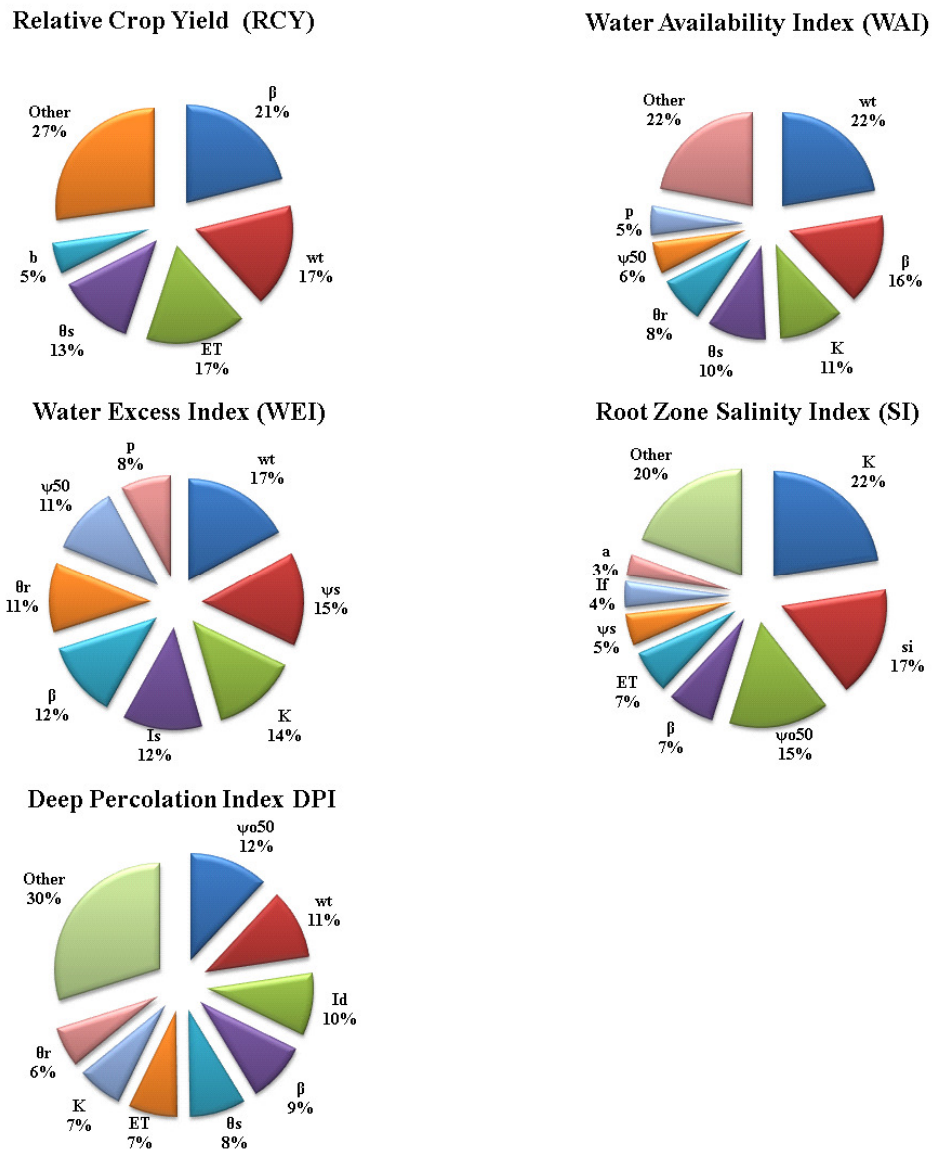

Figure 2. 3: Total Effect Sensitivities Using Extended FAST Method

Table 2.2 Higher Order Effect as a Percentage of the Total Effect Using the Extended FAST Method

Input Factor	Higher Order Effect (%) of the Total Effect				
	RCY	WAI	WEI	SI	DPI
K	91.23	73.05	87.06	85.96	88.32
θs	66.72	51.12	0.00	55.65	37.86
θr	81.77	15.71	87.22	85.51	76.28
α	60.66	79.91	0.00	92.52	95.22
β	22.88	16.65	87.19	86.02	74.70
γ	97.49	95.88	0.00	95.95	87.32
a	94.71	86.62	0.00	95.65	93.46
b	86.46	84.43	0.00	99.05	96.61
ψ50	99.07	92.56	85.89	91.39	90.31
ψo50	80.58	95.73	0.00	85.39	84.35
p	83.25	93.72	87.93	91.99	86.36
ψs	92.76	97.51	87.06	96.90	94.97
Id	84.36	92.47	0.00	63.98	70.54
Is	89.14	95.18	86.79	97.18	87.67
If	83.52	92.88	0.00	70.39	90.86
wt	34.73	15.87	79.58	90.39	83.04
si	75.25	94.10	0.00	14.34	74.95
ET	30.29	87.43	0.00	74.41	63.65

The Morris coefficients μ^* ranks β and θ_r at the top of the list, while the water table is in third place. The irrigation depth is more important according to the Morris Coefficients than the FAST method, but its importance is dependent on the interactions

with other parameters. For example, high irrigation depths at low irrigation frequency and for highly permeable soil should have negligible influence on the WAI index especially in shallow groundwater tables.

The partial correlation coefficient measures of parameters are orderly ranked as $\beta, b, a, si\ and\ K$ which is significantly different from the FAST and Morris methods. However, the sign of the partial correlation coefficient provides some insight about how the factors affect the output. For example, higher values of β produce lower WAI.

The Monte Carlo filtering is employed to map the factors controlling the dry conditions. Dry conditions are defined as the lowest 20% WAI values. The factors $\theta_r\ and\ wt$ were found to contribute significantly to the dry conditions. A possible explanation is that the upflux from the shallow water table keeps the moisture content high in the root zone, while θ_r resembles the ability of the soil to retain moisture. Moreover, the effect of weather conditions on dry conditions is evident by the importance of the evapotranspiration factor.

2.8.3 Water Excess Index (WEI)

The WEI is a quantitative predictor of waterlogging. According to the extended FAST method, 100% of the WEI variance results from only eight factors (Figure 2.2). The water table, Ψ_s and hydraulic conductivity were identified as the most influential factors. This result is reasonable because the higher initial water table and the high hydraulic conductivity values control the capacity of the subsurface system to drain the water out of the root zone. Morris average elementary effect (EE) gives more weight to

the residual moisture content and the irrigation frequency, however their effect is dependent on the interactions with other factors as indicated by the variance of the EE's. The limitation of the PCC coefficient for nonlinear problems is obvious here because the water table occupies the 14th place which contradicts common sense that the water table should play a significant role in waterlogging problems.

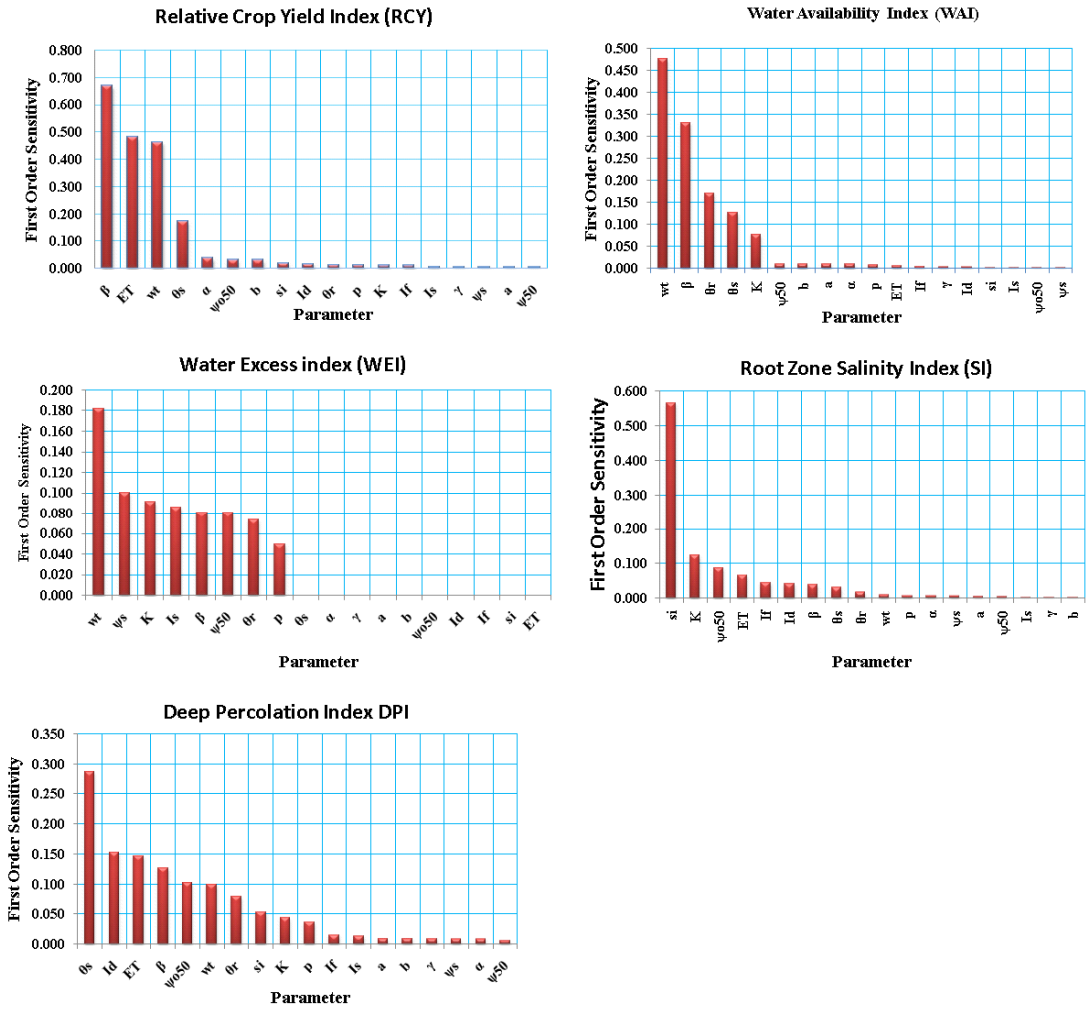

Figure 2. 4: The First Order Sensitivities Using Extended FAST Method

2.8.4 Root Zone Salinity Index (SI)

The pie chart in Figure 2.3 shows the total sensitivity of input factors with respect to SI. Eight factors were found to be responsible for 80% of the SI variance. The hydraulic conductivity by itself contributes 22% of the index variance. A possible explanation is that the conductivity of the soil controls the leaching efficiency of the roots and the saline groundwater up flux rate. Moreover, 86% of the hydraulic conductivity contribution is generated through interaction with other parameters.

Unsurprisingly, the initial root zone salinity is second in importance. It was expected that the water table depth should be one of the top controlling parameters; however, this is only true if the groundwater salinity is also high. This notion is supported by the higher order sensitivity percentage of the water table factor in Table 2.2, in which it has a value that amounts to 90.4% of the total effect.

The crop yield model parameter (ψ_{050}) is in third place (15% of the total effect). Since this term determines the salinity concentration (as osmotic head), at which the root extraction drops by 50%; the root extraction rate should be very sensitive to the root zone salinity especially at high ψ_{050} values (or small absolute values). For very small ψ_{050} values, the root extraction continues even at high root zone salinity values, and thus the driving force of the root zone salinization process would continue. The Partial Correlation Coefficient (PCC) ranks ψ_{050} as the most important parameter with a positive sign, which supports the previously mentioned explanation.

For irrigation designers, it is seen that irrigation frequency (**If**) is responsible for only 4% of the SI variance and the FOS measure for irrigation efficiency and irrigation depth are the fifth and sixth positions respectively. On the other hand, the Morris importance factor (μ^*) ranks the application depth as second in importance. Again, it is important to recall that the Morris sensitivity results should be understood in the context of the Factor Fixing framework. In other words, the method is efficient in determining the important factors but not the relative importance of each of them.

Filtering input factors that contribute to high root zone salinity are shown in Table 2.4. The Semirnov statistic ranks the initial salinity, irrigation frequency, and the residual moisture content as the major factors controlling extreme salinization of the root zone. The irrigation water salinity is in the sixth position.

The scatter plot (Figure 2.6) shows the Relative Crop Yield and Root Zone Salinity. The figure does not show a strong correlation between the two indexes. This may be due to the higher order interactions of the crop salt tolerance factor ψ_{050} with other factors.

2.8.5 Deep Percolation Index (DPI)

The results show that eight input factors control 70% of the deep percolation variance as shown in Figure 2.2. The Total Effect sensitivities of factors $\psi_{050}, wt\ and\ Id$ are ranked as first, second and third, respectively. It is worth noticing that all input factors, except for θ_s, have higher order interactions of more than 63%, which reveals the complex dynamics of deep percolation. Take for example the

hydraulic conductivity, which at higher values might increase the deep percolation; however, the index definition does not distinguish between up flux and down flux since they are physically one process. As a result, the high ***K*** value might enhance up flux and down flux equally, which makes the ***K*** effect highly interactive with other factors. The high EE standard deviation and high order effect in Table 2.2 supports this result.

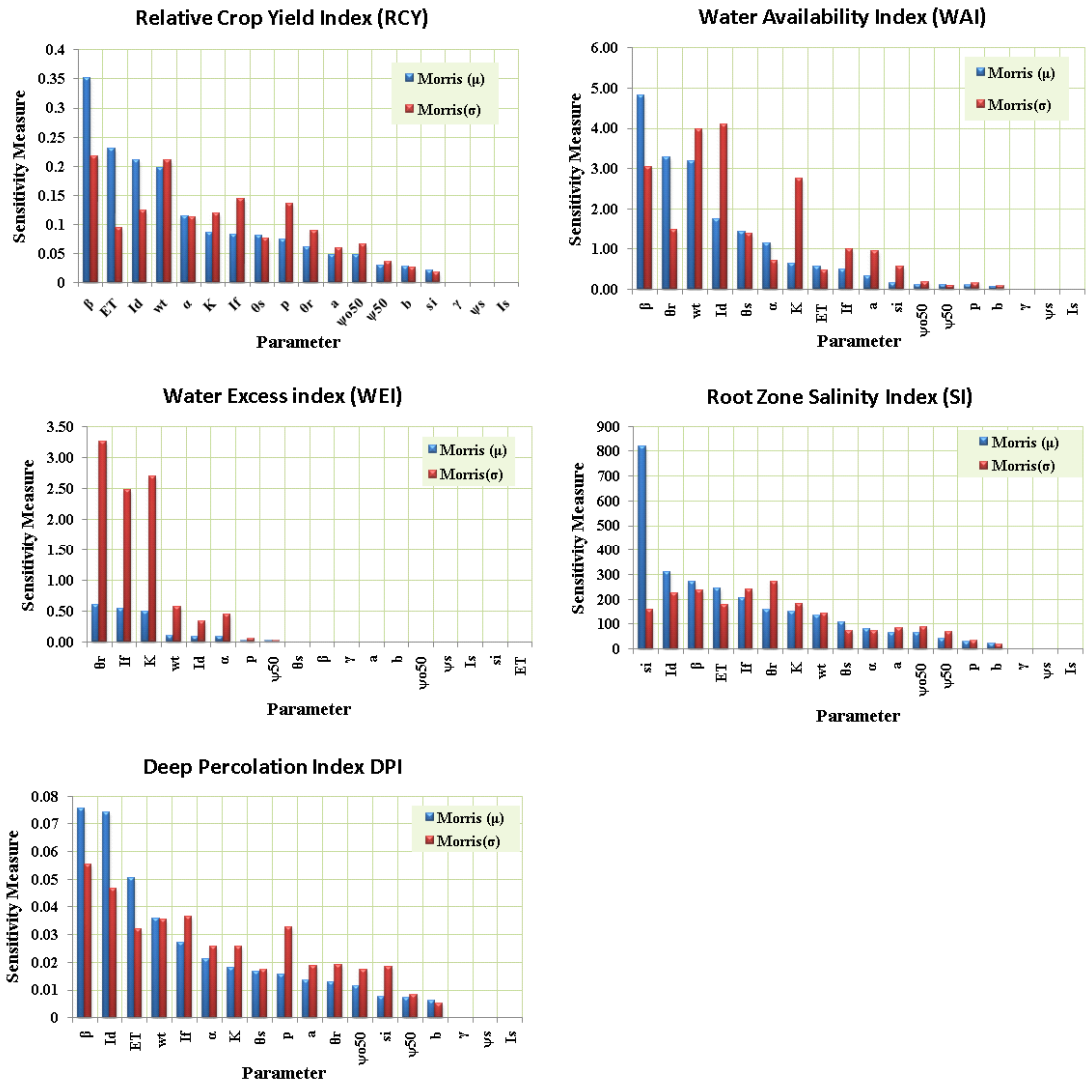

Figure 2. 5: The Average and the Standard Deviation of the Elementry Effect Using Screening Method

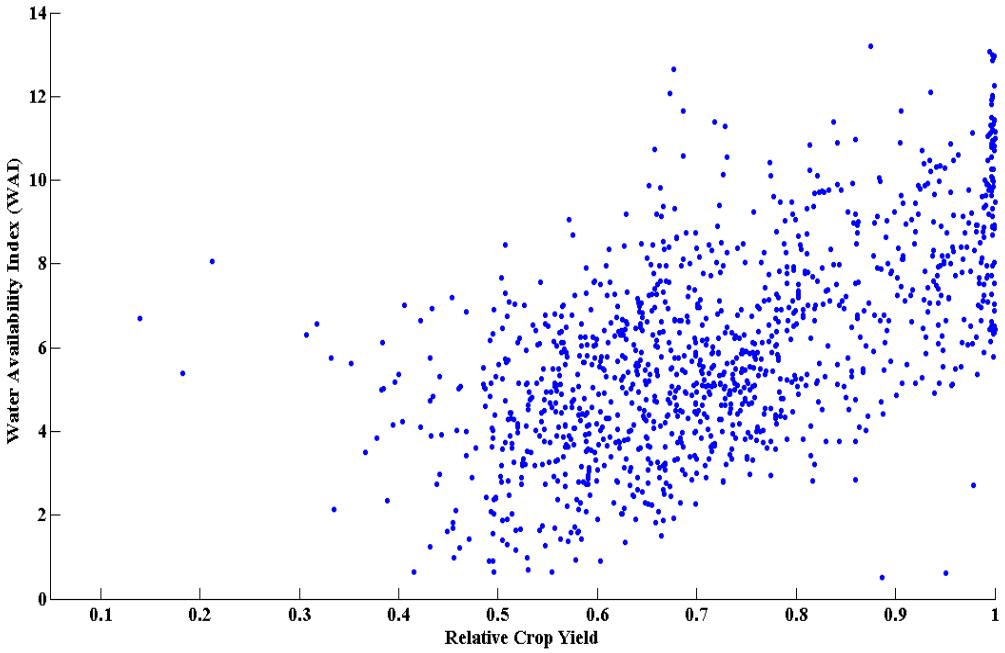

Figure 2. 6: Water Avialability Index (WAI) Vs. Relative Crop Yield

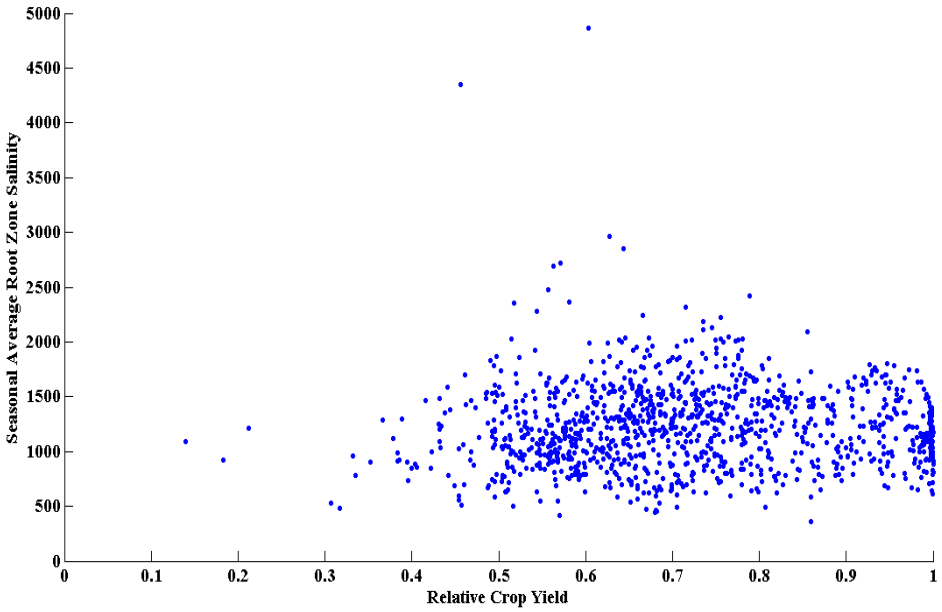

Figure 2. 7: Root Zone Average Salinity (SI) Vs. Relative Crop Yield

Table 2.3 Ranking the Importance of Input Factors Using Partial Correlation Coefficients (PCC), where R is the rank.

Input Factor	RCY		WAI		WEI		SI		DPI	
	PCC	R	PCC	R	PCC	R	PCC	R	PCC	R
K	0.016	8	0.038	5	0.041	3	0.014	9	0.004	16
θs	0.001	18	0.020	7	0.018	8	0.004	15	0.022	7
θr	0.021	6	0.006	13	0.011	10	0.026	7	0.020	8
α	0.005	15	0.024	6	0.000	18	0.001	17	0.014	10
β	0.045	1	0.057	1	0.064	1	0.011	11	0.019	9
γ	0.018	7	0.001	17	0.001	17	0.021	8	0.041	3
a	0.041	2	0.043	3	0.025	6	0.013	10	0.054	1
b	0.031	3	0.044	2	0.011	11	0.027	6	0.044	2
ψ50	0.026	4	0.005	15	0.038	4	0.008	14	0.032	6
ψo50	0.001	17	0.020	8	0.023	7	0.066	1	0.032	5
p	0.008	14	0.006	14	0.009	13	0.008	13	0.013	11
ψs	0.021	5	0.009	11	0.046	2	0.036	3	-0.007	13
Id	-0.009	13	0.012	10	0.007	15	0.045	2	0.005	15
Is	0.009	12	0.004	16	0.010	12	0.003	16	0.012	12
If	0.011	11	0.016	9	0.030	5	0.031	5	0.000	18
wt	0.013	10	0.001	18	0.008	14	0.001	18	0.036	4
si	0.014	9	0.040	4	0.004	16	0.033	4	0.002	17
ET	0.004	16	0.007	12	0.014	9	0.009	12	0.005	14

Table2. 4: Monte Carlo Filtering of Low Crop Yield, Saline Root Zone and Dry Conditions

Input Factor	Low Crop Yield		Saline Root Zone		Dry Conditions	
	D-stat	Rank	D-stat	Rank	D-stat	Rank
K	0.132	14	0.186	10	0.127	8
θs	0.188	9	0.171	11	0.195	5
θr	0.122	15	0.341	3	0.646	1
α	0.146	12	0.085	18	0.106	10
β	0.478	1	0.236	9	0.315	3
γ	0.220	8	0.258	8	0.123	9
a	0.166	11	0.106	16	0.158	7
b	0.228	7	0.152	13	0.085	15
ψ50	0.108	17	0.106	17	0.099	12
ψo50	0.301	4	0.260	7	0.030	18
p	0.121	16	0.130	15	0.186	6
ψs	0.142	13	0.329	4	0.075	17
Id	0.398	2	0.157	12	0.081	16
Is	0.083	18	0.315	6	0.089	14
If	0.251	6	0.395	2	0.103	11
wt	0.291	5	0.139	14	0.408	2
si	0.171	10	0.595	1	0.094	13
ET	0.328	3	0.329	5	0.226	4

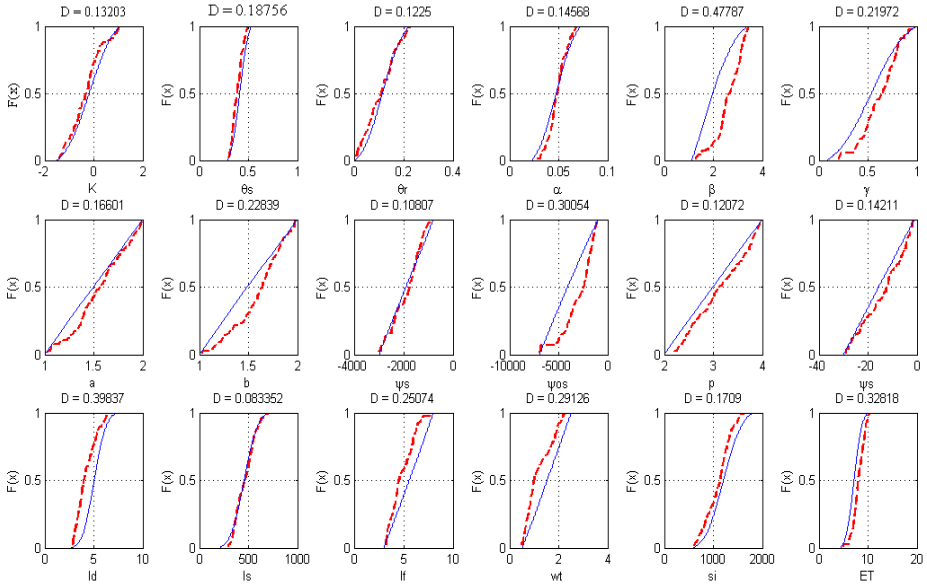

Figure 2. 8: Behavioral Vs. Non-behavioral CDF's of Input Factors for RCY < 40%

2.9 Conclusion

The strength of the Global Sensitivity Analysis resides in its ability to improve the overall understanding of the input-output relationship by exposing the role each factor plays by itself and through interactions with other factors. Such understanding can be employed within the factor prioritization framework to determine factors that need further attention during either the sampling or modeling stages. In addition, the Global Sensitivity Analysis could be employed within the Factor Fixing framework to reduce the complexity of models by fixing non-influential factors. Lastly, the Global Sensitivity Analysis might be used within the Factor Mapping setting to determine the input factors responsible for a certain output cumulative distribution function region. In this paper, four Global Sensitivity Analysis methods were used to obtain the relative importance of 18 uncertain input factors with respect to five output indices. These

indices reflect different research and regulation interests. A one-dimensional variably saturated flow and transport case was used to simulate these indices.

Results show that a large portion of the output variance can be attributed to a few parameters. For example, 73% of the relative crop yield variance is contributed by only five factors. Similar results were obtained for other indices. Of particular importance for most of the indices is the pore-size van Genuchten parameter. Besides its top ranking, its major influence on Relative Crop Yield and Water Availability Index occurs through the main effect of the parameter; i.e. not through interactions with other parameters.

The Monte Carlo Filtering (MCF) for extreme salinity and dry conditions reveals the importance of the residual moisture content. This observation was not possible using other Global Sensitivity Analysis methods. Roughly speaking, the MCF, Extended FAST, and Morris methods produce similar results while the regression-based PCC measure is significantly different. The severe nonlinearity of the Richard's flow equation might be the source of the deficiency of the PCC measure; however the sign of the PCC ranking reveals the sign of the correlation between the input and the output parameters.

2.10 Recommendations for Future Investigations

Several assumptions were made throughout this study; e.g. the absence of correlation between input factors. As a result, this assumption might introduce two

types of errors. Unfortunately, the available data were not enough to favor neither the correlation nor an independence decision.

The expensive computational demand required to simulate the variably saturated problem push toward simplifying the problem from a three dimensional field to a one-dimensional soil column. Overcoming this obstacle might be achieved by parallelizing the models evaluations on a large cluster of computers.

Another assumption that was made in this paper is the spatial homogeneity of the soil properties. This assumption can be relaxed by considering each soil property at each layer as an individual parameter. Certainly, this would increase the number of input factors and consequently increase the computational requirements of the GSA.

Acknowledgements

This work was partially funded by a project from the United States Bureau of Reclamation and the authors are grateful for the help provided by Mr. Roger Burnett an Agricultural Engineer, with the United States Bureau of Reclamation, Denver Technical Services Center.

2.11 References

Alzraiee, A., Garcia, L. and Burnett, R. (2009) "Modeling Spatial and Temporal Variability in Irrigation and Drainage Systems: Improvements to the Colorado State University Irrigation and Drainage Model (CSUID)", Presented and

published in the proceedings of the USCID Conference on Irrigation and Drainage for Food, Energy and the Environmental, November 3-6, Salt Lake City, Utah.

Anderson, Mary P., and William W. Woessner. 1992. Applied Groundwater Modeling. 1st ed. Academic Press, January 15.

Beven, Keith, and Andrew Binley. 1992. "The future of distributed models: Model calibration and uncertainty prediction." Hydrological Processes 6 (3): 279-298.

Campolongo, Francesca, Jessica Cariboni, and Andrea Saltelli. 2007. "An effective screening design for sensitivity analysis of large models." Environmental Modelling & Software 22 (10) (October): 1509-1518. doi:doi: DOI: 10.1016/j.envsoft.2006.10.004.

Cardon, G. E, and J. Letey. 1992. "Plant Water Uptake Terms Evaluated for Soil Water and Solute Movement Models." Soil Sci. Soc. Am. J. 56 (6): 1876-1880.

Cukier, R. I., C. M. Fortuin, K. E. Shuler, A. G. Petschek, and J. H. Schaibly. 1973. "Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory." The Journal of Chemical Physics 59 (8): 3873-3878.

Doorenbos, J., J. Plusje, AH Kassam, V. Branscheid, and CLM Bentvelsen. 1986. Yield response to water. Vol. 3.

Feddes, Reinder A., Piotr Kowalik, Krystina Kolinska-Malinka, and Henryk Zaradny. 1976. "Simulation of field water uptake by plants using a soil water dependent root extraction function." Journal of Hydrology 31 (1-2) (September): 13-26. doi:doi: DOI: 10.1016/0022-1694(76)90017-2.

Van Genuchten, M. T, and US Salinity Laboratory. 1987. A numerical model for water and solute movement in and below the root zone. United States Department of Agriculture Agricultural Research Service US Salinity Laboratory.

van Genuchten, M. Th. 1980. "A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils1." Soil Science Society of America Journal 44 (5): 892. doi:10.2136/sssaj1980.03615995004400050002x.

Giglioli, N., and A. Saltelli. 2000. "SimLab 1.1, Software for Sensitivity and Uncertainty Analysis, tool for sound modelling." Arxiv preprint cs/0011031.

Hanks, R. J, and R. W Hill. 1980. Modeling crop responses to irrigation in relation to soils, climate and salinity. 6. International Irrigation Information Center.

Helton, J.C., J.D. Johnson, C.J. Sallaberry, and C.B. Storlie. 2006. “Survey of sampling-based methods for uncertainty and sensitivity analysis.” Reliability Engineering & System Safety 91 (10-11): 1175-1209. doi:doi: DOI: 10.1016/j.ress.2005.11.017.

Helton, Jon C. 1993. “Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal.” Reliability Engineering & System Safety 42 (2-3): 327-367. doi:doi: 10.1016/0951-8320(93)90097-I.

Hill, Mary C., and Claire R. Tiedeman. 2007. Effective Groundwater Model Calibration: With Analysis of Data, Sensitivities, Predictions, and Uncertainty. Wiley-Interscience, January 22.

Homma, Toshimitsu, and Andrea Saltelli. 1996. “Importance measures in global sensitivity analysis of nonlinear models.” Reliability Engineering & System Safety 52 (1) (April): 1-17. doi:doi: DOI: 10.1016/0951-8320(96)00002-6.

Hopmans, J. W, and K. L Bristow. 2002. “Current capabilities and future needs of root water and nutrient uptake modeling.” Advances in Agronomy 77: 103–183.

Iman, Ronald L., and W. J. Conover. 1979. “The Use of the Rank Transform in Regression.” Technometrics 21 (4) (November 1): 499-509.

Jongschaap, Raymond E.E. 2007. “Sensitivity of a crop growth simulation model to variation in LAI and canopy nitrogen used for run-time calibration.” Ecological Modelling 200 (1-2) (January 10): 89-98. doi:doi: DOI: 10.1016/j.ecolmodel.2006.07.015.

Makowski, C. Naud, M.-H. Jeuffroy, A. Barbottin and H. Monod. 2006. “Global sensitivity analysis for calculating the contribution of genetic parameters to the variance of crop model prediction.” Reliability Engineering & System Safety 91 (10-11): 1142-1147. doi:doi: DOI: 10.1016/j.ress.2005.11.015.

Meyer, PD and Rockhold, ML and Gee, GW. 1997. Uncertainty analyses of infiltration and subsurface flow and transport for SDMP sites. US Nuclear Regulatory Commission Report NUREG/CR--6565.

Mishra, Srikanta, Neil Deeds, and Greg Ruskauff. 2009. “Global Sensitivity Analysis Techniques for Probabilistic Ground Water Modeling.” Ground Water 47 (5): 727-744.

Molz, Fred J. 1981. “Models of water transport in the soil-plant system: A review.” Water Resour. Res. 17 (5): 1245-1260.

Morris, Max D. 1991. “Factorial Sampling Plans for Preliminary Computational Experiments.” Technometrics 33 (2) (May): 161. doi:10.2307/1269043.

Oakley, Jeremy E., and Anthony O’Hagan. 2004. “Probabilistic sensitivity analysis of complex models: a Bayesian approach.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66 (3): 751-769. doi:10.1111/j.1467-9868.2004.05304.x.

Pan, Feng, Jianting Zhu, Ming Ye, Yakov A. Pachepsky, and Yu-Shu Wu. 2011. “Sensitivity analysis of unsaturated flow and contaminant transport with correlated parameters.” Journal of Hydrology 397 (3-4) (February 3): 238-249. doi:doi: DOI: 10.1016/j.jhydrol.2010.11.045.

Pathak, T., C. Fraisse, J. Jones, C. Messina, and G. Hoogenboom. 2007. “Use of global sensitivity analysis for CROPGRO cotton model development.” Transactions of the ASABE 50 (6): 2295–2302.

Ruget, Françoise, Nadine Brisson, Richard Delécolle, and Robert Faivre. 2002. “Sensitivity analysis of a crop simulation model, STICS, in order to choose the main parameters to be estimated.” Agronomie 22 (2) (March): 133-158. doi:10.1051/agro:2002009.

Saltelli, A., K. Chan, and E. M. Scott. 2008. Sensitivity Analysis. Wiley, December 30.

Saltelli, A., Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. 2008. Global Sensitivity Analysis: The Primer. Wiley-Interscience, March 11.

Saltelli, Andrea, and Ricardo Bolado. 1998. “An alternative way to compute Fourier amplitude sensitivity test (FAST).” Comput. Stat. Data Anal. 26 (4): 445-460.

Schaap, M. G, F. J Leij, and M. T van Genuchten. 2001. “Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions.” Journal of Hydrology 251 (3-4): 163–176.

Shalhevet, J., A. Vinten, and A. Meiri. 1986. “Irrigation interval as a factor in sweet corn response to salinity.” Agronomy journal (USA).

Sobol, I. M. 1993. “Sensitivity analysis for non-linear mathematical models.” Mathematical Modelling and Computational Experiment 1 (1): 407–414.

Sobol', I. M. 2001. "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates." Mathematics and Computers in Simulation 55 (1-3) (February 15): 271-280. doi:doi: DOI: 10.1016/S0378-4754(00)00270-6.

Varella, Hubert, Martine Guérif, and Samuel Buis. 2010. "Global sensitivity analysis measures the quality of parameter estimation: The case of soil parameters and a crop model." Environmental Modelling & Software 25 (3) (March): 310-319. doi:doi: DOI: 10.1016/j.envsoft.2009.09.012.

Veenhof, D. W., and R. A. McBride. 1994. "A preliminary performance evaluation of a soil water balance model (SWATRE) on corn producing croplands in the RM of Haldimand-Norfolk." Soil compaction susceptibility and compaction risk assessment for corn production. Centre for Land and Biological Resources Research AAFC, Ottawa: 112–142.

U.S. Dept. of Agriculture (USDA). 1972a. Soil survey of Otero county, Colorado, USDA, SCS, La Junta, Colo.

U.S. Dept. of Agriculture (USDA). 1972b. Soil survey of Bent county, Colorado, USDA, SCS, La Junta, Colo.

Xevi, E., J. Gilley, and J. Feyen. 1996. "Comparative study of two crop yield simulation models." Agricultural Water Management 30 (2) (April): 155-173. doi:doi: DOI: 10.1016/0378-3774(95)01218-4.

3 USING CLUSTER ANALYSIS OF HYDRAULIC CONDUCTIVITY REALIZATIONS TO REDUCE COMPUTATIONAL TIME FOR MONTE CARLO SIMULATIONS

3.1 General

Despite the conceptual simplicity of the Monte Carlo Simulation methods in assessing the uncertainty in hydrogeological systems, their use is limited by the expensive computational requirements in terms of the large number of realizations required to be processed. Cluster Analysis is applied in this paper to reduce the number of realizations to be processed by flow simulators while efficiently approximating the flow response statistics. Different clustering techniques are used to partition the realizations ensemble into a few clusters that are significantly different from each other and have maximum intra cluster similarity. The clustering step is achieved by using different similarity metrics. Then a subsample of the realizations is collected to represent the uncertainty in the whole ensemble. Two methods for collecting the subsample are investigated; the stratified sampling and the centriod based sampling. The performance of different clustering and sampling techniques is tested by evaluating the mismatch between the statistics of the ensemble response, the reference response, and the statistics of the subsample response which are estimated from the clusters.

Results show that 25% of the realizations in the ensemble could be sufficient to estimate the uncertainty in the flow responses using a suitable clustering method and suitable similarity measures.

3.2 Introduction

The application of numerical methods to solve the continuity flow and transport equations requires reasonable knowledge of the hydraulic properties of the porous media. Typically, our knowledge of the subsurface porous properties is developed by using scarce and expensive field or laboratory measurements. These measurements - limited in number and accuracy - would usually not be sufficient to provide accurate insight into the hydrogeological system. Moreover, the spatial variability of soil properties, which are essential to understanding the flow and transport processes, require large number of measurements to be accurately reconstructed. This is, of course, beside the errors in the measurements themselves.

A more realistic approach is to consider the soil properties in a probabilistic framework where a soil property, for example the hydraulic conductivity field, is described by a random function. This certainly does not imply that the hydraulic conductivity field is random, but that our knowledge is incomplete. The concept of a spatial random field was introduced by Matern (1960) and Matheron (1962) to analyze the uncertainty in geological formations and for use in the mining industry and it has been widely adopted (Freeze 1975; Tang and Pinder 1977; Dagan 1982; Yeh 1992; Gotovac et al. 2009) to quantify uncertainty in hydrogeological systems. The

probabilistic representation of soil properties is incorporated with the physical flow equation to produce the statistical properties of the hydraulic responses of the system.

Since the mid 1970's the uncertainty in the spatial properties of aquifers has been studied by researchers using either "analytical stochastic" approaches or the "Monte Carlo simulation" approach. The analytical approaches (Bakr et al. 1978; Tang and Pinder 1977) are usually based on several restrictive assumptions that are not easy to relax. For example, a small variance of the log conductivity, unbounded domains, steady-state flow, and uniform-in-the-average flow are necessary assumptions for analytical solutions; and the results are limited to the first two moments (Rubin 2003). On the other hand, the Monte Carlo (MC) method (Freeze 1975; Smith and Hebbert 1979) is known for its conceptual simplicity and its generality (Rubin 2003). Despite these advantages, large computational efforts are required in most practical problems.

To apply the MC simulations, large numbers of realizations are normally generated and processed in flow or/and transport simulators to approximate the uncertainty in the response variables. Such methodology could be unpractical in cases where fine grids are used to model the spatial properties or when the unsaturated-saturated flow numerical models are used to simulate a field scale problem. Specifically, solving the nonlinear unsaturated flow equation requires small time steps and small spatial discretizations which make the simulation time considerably longer.

Several techniques have been used to reduce the number of realizations to be analyzed while producing a useful prediction of the uncertainty. A comprehensive review of these techniques can be found in Deutsch (2002). These techniques are more

popular in the oil industry than in the hydrogeology field. Previous attempts to approach the problem have mainly focused on ranking the realizations based on the response of flow or transport models. The concept of connectivity of the porous media was suggested by Deutsch (2002) as a way to rank realizations. Connectivity of the porous media, defined as the sets of net geological numerical cells that are connected in a three-dimensional space (Deutsch 2002), was used in two ways, the static connectivity and the dynamic connectivity. In the static connectivity, the ratio of the volume of geological objects; defined at a certain permeability threshold, to the total reservoir volume, could be used as a rank measure. On the other hand, the dynamic methods use the *lengths or the times* that a particle needs to move between two defined points (for example, injection and production wells), as the index to rank the realizations. Another approach is to use flow simulators to rank the realizations based on the flow responses after upscaling the fine realizations to coarse ones (Kupfersberger and Deutsch 1999). Gómez-Hernández and Carrera (1994) used a linear approximation of the groundwater flow equation, instead of the flow models themselves, to approximate the rank of the realizations.

In this paper, Cluster Analysis (CA) (Anderberg 1973; Everitt, et al. 2009) is implemented as a novel technique to reduce the number of realizations to be processed in a flow simulator while still covering the uncertainty space. The underlying assumption behind using clustering to group the realizations is that similar realizations have similar responses and there is no need to run all the realizations in the ensemble; instead, a subsample, a specified number of realizations, are collected from each cluster

and the probability (frequency) of the of the flow response is approximated based on the size of the cluster from which the realizations were sampled.

The overall procedure was implemented in two major steps: (i) the realizations were grouped into numbers of clusters, and (ii) a number of realizations were sampled from each cluster. For the first step, the hierarchal clustering (deterministic clustering) and the K-means clustering (iterative data partition) methods are evaluated using different similarity criteria to group the realizations (Anderberg 1973; Everitt, et al. 2009). In the second step, two sampling schemes are evaluated, stratified sampling and centriod based sampling. A synthetic unconfined aquifer system is simulated to verify the feasibility of these methods. A large number of realizations (400 realizations) of the hydraulic conductivities are generated using a geostatistical simulation method. All of these realizations are processed in the flow simulator to produce the reference cumulative distribution function (CDF) of the response variable (i.e. the hydraulic heads). Then, the ability of the clustering methods to produce a good estimation of the reference CDFs, the reference means and variances using a subsample of realizations, is investigated.

3.3 Cluster Analysis (CA)

Clustering is a method by which the data can be grouped into clusters based on certain similarity measures. The data could be scalar or multi-dimensional. Different types of clustering can be found in the literature. In this paper two methods are utilized to cluster an ensemble of realizations, namely the hierarchical method and the K-means method.

3.3.1 Hierarchical Clustering

The hierarchical clustering algorithm(Anderberg 1973; Abonyi et al. 2007; Everitt, et al. 2009) agglomerates *similar* individual realizations (leaves level) into small clusters and subsequently, based on the proximity of the clusters, larger clusters are produced by merging smaller clusters using a certain *linkage criteria*. The cluster scheme can be represented in a tree structure called a dendrogram. Based on the number of clusters or based on the required precision of the clustering, the tree could be cut at a certain level. Different types of similarity measures and linkage criteria are used in the literature. For example, the *Euclidian distance* between two realizations can be used as an indicator of similarity; thus small distance implies similar realizations. The linkage criteria is used to merge clusters; for example the *single linkage* criterion merges clusters based on the shortest distance between two clusters while the furthest distance linkage criteria uses the farthest distance between two clusters. For more details about different distance measures and linkage criteria see Anderberg (1973); Everitt et al. (2009); and Jones (1997). The distance measures and linkage criteria available in the MATLAB's Statistical Toolbox (Jones 1997) are used in this paper. Once the clustering is achieved, the dissimilarity of the produced clusters can be evaluated using a *cophenetic correlation coefficient* which produces values less than one. Cophenetic correlation coefficients closer to one mean higher dissimilarity between clusters. The hierarchical clustering has a deterministic clustering output and once a realization is set in a cluster it will stay in it.

3.3.2 K-means method

The K-means method (Abonyi et al. 2007; Everitt, et al. 2009) is an iterative data partitioning method in which an objective function is minimized. The number of clusters is determined in advance and a random number of initial points (realizations) equal to the number of clusters are chosen. Each of the realizations is associated with one of the initial points (realizations) based on its proximity. Next, new centriods are calculated for each cluster and using these new centriods the association process is repeated. The iterations proceed until the intra-cluster variance is minimized. The K-means method produces different clustering each time they are implemented. However, due to the simplicity and speed of the K-means method, it can be repeated several times and the clustering scheme that achieves the minimum intra cluster variance is used.

3.4 Methodology

The uncertainty in parameters of a given geological formation can be represented statistically using the joint distribution function $\boldsymbol{f_k(u_1, u_2, \ldots u_n)}$ for the hydraulic parameter $\boldsymbol{k}$ and at $\boldsymbol{u_1, u_2, \ldots u_n}$ spatial positions. This joint distribution function demands a large amount of data to be established which is usually expensive to obtain. A practical approach to characterizing the random field is through the use of the first two moments of $\boldsymbol{f_k(u_1, u_2, \ldots u_n)}$, the mean function $\boldsymbol{m(u)}$ and the spatial covariance function $\boldsymbol{C_{kk}(u_1, u_2)}$. The experimental covariance should be modeled using specified functions to ensure that the calculated variance is positive "*positive definiteness condition*". For stationary random fields, the covariance function can be reduced to

$\boldsymbol{C_{kk}(h)}$ where $\boldsymbol{h = u_1 - u_2}$. The final modeled covariance function takes the form $\boldsymbol{C_{kk}(h, \sigma^2, I_x, I_y, I_z)}$ where σ^2 is the variance at the sill of the equivalent variogram and $\boldsymbol{I_x, I_y, I_z}$ are the integral scale in the $\boldsymbol{x, y}$ and $\boldsymbol{z}$ directions.

In sequential Gaussian Monte Carlo simulations, the mean function and the spatial covariance function are used to generate N_r number of equal probable realizations $\boldsymbol{l_i}$: $\{\boldsymbol{k(u), u \in A}\}$ for the spatial domain $\boldsymbol{A}$ and $\boldsymbol{i} = \{\boldsymbol{1}, \ldots \boldsymbol{N_r}\}$. The spatial position in a three-dimensional numerical grid with three coordinates $\boldsymbol{k(x, y, z)}$ can be rearrange using Equation 4.1 into a one dimensional column vector $\boldsymbol{l}$ with dimension $[\boldsymbol{n}, \boldsymbol{1}]$ where $\boldsymbol{n}$ is the number of nodes.

$$\boldsymbol{u = (iz - 1) * nx * ny + (iy - 1) * nx + ix} \quad \textbf{(3.1)}$$

In Equation (4.1) $\boldsymbol{iz, iy,}$ and $\boldsymbol{ix}$ represent the coordinate index of the cell node in the x, y and z directions respectively; and $\boldsymbol{nx}$ and $\boldsymbol{ny}$ are the number of rows and columns of the grid.

Large numbers of realizations usually need to be generated to cover the uncertainty space. Consider the ensemble of realizations $\boldsymbol{Q_k = [l_1, l_2, .. l_{N_r}]}$ where $\boldsymbol{N_r}$ is the number of realizations. In Monte Carlo simulations each of the realizations is processed deterministically in a flow simulator to produce the response:

$$\boldsymbol{h_i = M(l_i)} \quad \textbf{(3.2)}$$

The function $\boldsymbol{M(l_i)}$ represents the flow (or transport) simulator. The hydraulic heads for all cells are summarized in the response vector $\boldsymbol{h_i}$. All of the vectors $\boldsymbol{h_i}$ can be concatenated to form the response matrix $\boldsymbol{H = [h_1, h_2, .. h_{N_r}]}$. Each row $\boldsymbol{H_{i:}}$ in the

response matrix represents all of the possible heads at cell $\boldsymbol{i}$, and thus the expectations and the variances of the response variable at all cells can be computed using Equations (3.3) and (3.4):

$$\widehat{\boldsymbol{h}_{\boldsymbol{i}}} = \boldsymbol{E}[\boldsymbol{h}_{\boldsymbol{i}}] = \frac{\mathbf{1}}{\boldsymbol{N}_{\boldsymbol{r}}} \sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{N}_{\boldsymbol{r}}} \boldsymbol{H}_{\boldsymbol{ij}} \tag{3.3}$$

$$\boldsymbol{VAR}[\boldsymbol{h}_{\boldsymbol{i}}] = \frac{\mathbf{1}}{\boldsymbol{N}_{\boldsymbol{r}}-\mathbf{1}} \sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{N}_{\boldsymbol{r}}} \left(\boldsymbol{H}_{\boldsymbol{ij}} - \widehat{\boldsymbol{h}_{\boldsymbol{i}}}\right).\left(\boldsymbol{H}_{\boldsymbol{ij}} - \widehat{\boldsymbol{h}_{\boldsymbol{i}}}\right)^{\boldsymbol{T}} \tag{3.4}$$

A complete description of the uncertainty in the response variable can also be developed in the form of the Cumulative Distribution Function (CDF) $\boldsymbol{F_i(h)}$**.** This estimation will be referred to as the reference CDF of the response variable $\boldsymbol{h}$ at node $\boldsymbol{i}$.

To reduce the computational loads required in calculating the response variables in Equation (3.2), the realizations in the ensemble $\boldsymbol{Q_k}$ can be clustered into a few groups where realizations in each cluster $\boldsymbol{G_i}: \{\boldsymbol{i} = \mathbf{1}, \mathbf{2} \ldots \boldsymbol{N_c}\}$ are similar according to certain similarity measures $\boldsymbol{d}(\boldsymbol{r_1}, \boldsymbol{r_2})$ and $\boldsymbol{N_c}$ is the number of clusters. The clusters have different sizes (number of realizations) and each has a size $\boldsymbol{R_i}$. Instead of processing all the realizations in the flow simulator, a sample of the realizations can be used. There are two possibilities to achieve the sampling step, the *stratified sampling* and the *centriod based sampling*.

3.4.1 Stratified Sampling

A total number of samples $(\boldsymbol{N_s} \ll \boldsymbol{N_r})$ are selected and processed in the simulator. The number of samples to be *randomly* selected (Rubinstein and Kroese 2007; Gilbert 1987) from each cluster should be calculated according to Equation (3.5).

$$S_i = \frac{R_i}{N_r} N_s \tag{3.5}$$

The estimated first two moments should be calculated according to the following equations:

$$\hat{h}^* = \sum_{j=1}^{Nc} \sum_{i=1}^{S_j} \frac{R_j}{N_r S_j} h_{ij} \tag{3.6}$$

$$VAR^*(h) = \frac{1}{(N_s - 1)} \sum_{j=1}^{Nc} \sum_{i=1}^{S_j} (h_{ij} - \hat{h}^*).(h_{ij} - \hat{h}^*)^T \tag{3.7}$$

The estimated Cumulated Distribution Function $F_i^*(h)$ at any cell can be determined by using the $\frac{R_j}{N_r . S_j}$ as the probability of the head.

3.4.2 Centriod based sampling

In stratified sampling, realizations are randomly sampled from each cluster. This may raise the question about the consistency of the computed $F_i^*(h)$ distributions each time the sampling is implemented. By *consistency* it is meant that each time the stratified sampling is implemented the difference $|F_i^*(h) - F_i(h)|$ is always less than a specified tolerance error. In other words, the estimated Cumulative Distribution Function does not significantly change if the sampling is repeated. Unfortunately this might not be the case especially when the cluster sizes are significantly different.

An alternative sampling strategy is the centriod based sampling, in which the ensemble of realizations is clustered into N_s clusters (the same number of realizations to be processed in the flow simulator). Only one realization is selected from each cluster. This realization should represent the average of the realizations in the cluster.

The average of the realizations in a cluster can be thought of as the centriod of the cluster. For any cluster $\boldsymbol{G_m}$, where $\boldsymbol{m}$ is the cluster index, the centriod can be calculated using Equation (3.8).

$$\boldsymbol{C_{G_m}} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{R_m}} \boldsymbol{l_{ij}} \tag{3.8}$$

Since the centriod of the cluster might not carry the same statistical properties as other realizations generated from the sequential Gaussian simulations and it does not belong to the ensemble, the closest realization to the centriod of each cluster is selected. The Euclidian distance $\boldsymbol{d(C_m, l_{im})}$ between realization i and the centriod of cluster m could be used to define the proximity to the centriod according to Equation (3.9), where n is the number of cells in the numerical domain.

$$\boldsymbol{d(C_m, l_{im})} = \sqrt{\sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \big(\boldsymbol{C_m(j) - l_{im}(j)} \big)^2} \tag{3.9}$$

The selected realization is simulated and the response variables are associated with a probability value ($\frac{\boldsymbol{R_j}}{\boldsymbol{N_r}}$) proportional to the size of the cluster.

For the purpose of verifying the method, two mismatch measures are utilized. Equation (3.10) provides the mismatch measure between the reference and the estimated CDF with equal weights given to errors in the CDF values, while the mismatch measure calculated using Equation (3.11), which was introduced by Kupfersberger and Deutsch (1999), gives higher weights to the tails of the CDF.

$$e_i' = \int_{hi_{min}}^{hi_{max}} |F_i^*(h) - F_i(h)| \; \boldsymbol{dh} \tag{3.10}$$

$$e_i = \int_{hi_{min}}^{hi_{max}} \left| \frac{1 - \frac{F_i^*(h)}{F_i(h)}}{1 - F_i(h)} \right| dh \qquad (3.11)$$

3.5 Experimental Example

A synthetic aquifer system will be used as an example. The aquifer is a two dimensional unconfined aquifer that extends 3,000m x 3,000m in the horizontal plane. A uniform cell size of 100m x 100m is used which yields a total of 900 cells. The statistical parameters of the hydraulic conductivity random field with a mean of 10m/day and a coefficient of variation of 1.5 are chosen to reflect a heterogeneous aquifer. The spatial variability is modeled using a spherical variogram that has an isotropic horizontal correlation scale of 500m (1/6 of the model domain). The sequential Gaussian simulation is used to generate 400 equal probable realizations. MODFLOW (Harbaugh et al. 2000) is used to simulate the steady state flow equation. The domain is bounded by two constant head boundary conditions. Constant head boundary conditions of 50m and 45m are used at the upstream and downstream ends respectively. The side boundary conditions are taken to be no flow boundary conditions. All of the realizations generated by the geostatistical simulator are processed to produce the reference cumulative distribution function, the reference mean, and the reference standard deviation for the hydraulic heads at each node in the numerical grid.

Two clustering methods are used to cluster the realization ensembles, the hierarchical clustering and the K-means clustering methods. Different combinations of

distance measures and linkage criteria are used in the hierarchical clustering. The cophenetic correlation coefficient for each combination of the linkage criterion and the distance measure is calculated to verify the dissimilarity of the produced clusters. The stratified sampling scheme and the centriod based sampling scheme are performed for each clustering set.

Each stratified sampling was repeated 50 times (for this example) to investigate the consistency of the hydraulic head statistics. The mismatches between the estimated and the reference CDF and the errors in the first two moments are used to verify the performance of each clustering setting. The hierarchical clustering with stratified sampling can be summarized in the following steps:

1. Choose the distance measure and linkage criteria,
2. Determine the number of clusters. (For this example, the analysis was started using ten clusters.)
3. Determine the total number of realizations to be run in the flow simulator, 50 realizations were chosen for this analysis,
4. The number of realizations to be sampled N_s (subsample size) from each cluster is calculated according to Equation (3.5).
5. The sampled realizations are processed in the flow simulator to calculate the heads at each node.
6. The CDF of the estimated heads, means and standard deviations are calculated. The CDF mismatch and errors in the first two moments are calculated.

7. Optionally, steps 4, 5 and 6 are repeated for a user specified number of times (50 times in this example) to investigate the consistency of the estimated hydraulic head statistics.

To perform the hierarchical clustering with centriod based sampling the following steps are taken:

a. The same step 1 as above;

b. Determine the number of realizations, and cluster the ensemble into N_s clusters;

c. One realization is chosen from each cluster. This realization should be the closest to the centriod of the cluster according to Equations (3.8) and (3.9);

d. The same as steps 5 and 6 above.

To use the K-means method, the same procedures can be followed. However, since the K-means method produces different clustering results, the clustering procedures are repeated 50 times in this example and the clusters that produce the minimum intra cluster variances are used.

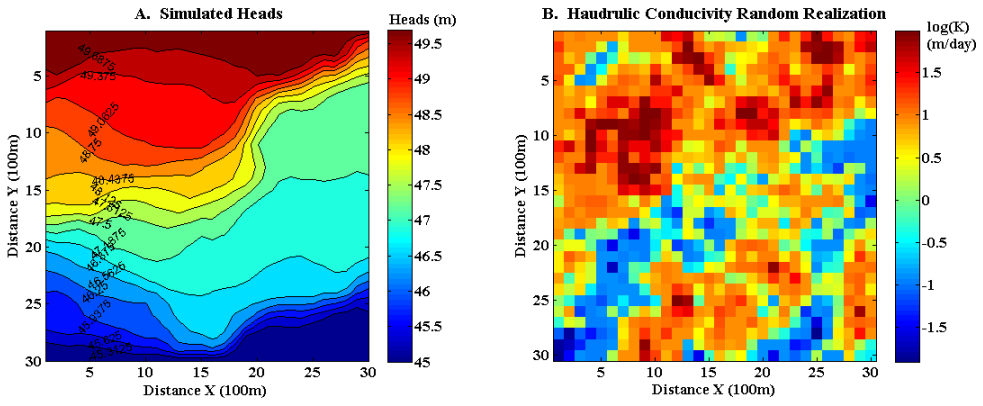

Figure 3. 1: Hydraulic Conductivity Realization and the Simulated Heads

3.6 Results

Upon clustering the realizations in the ensemble, the first question that might arise is whether the clustering scheme naturally divides the ensemble into dissimilar clusters with minimum dissimilarity among realizations within the same cluster. Producing dissimilar clusters, or significantly different clusters, ensures that the major differences among realizations are represented in the resulting clusters. In other words, the clusters produced are a coarse mirror of the fine uncertainty spectrum in the whole ensemble. Minimum dissimilarity among realizations within the same cluster enables us to sample a small number of realizations as a representative sample of the whole cluster. Table 3.1 summarizes the cophenetic correlation coefficients for each of the distance and linkage combinations used in this paper. Those distance-linkages with a cophenetic correlation close to one produce dissimilar clusters. For example, the *Standardized Euclidian* distance combined with the *Centriod linkage* has the highest cophenetic correlation coefficients (around 0.92). Generally, the Euclidian based metrics, such as *Euclidian*, *Standardize Euclidian*, *Cityblock* and *Minkowski* with the *Average, Centriod,* and *Single* linkages, produce high cophenetic correlations, around 0.9.

It's straight forward to say that dividing the ensemble into a large number of clusters increases the chance of producing dissimilar clusters with minimum intra cluster dissimilarity. This, by necessity, poses the question about the number of clusters required to produce a good estimate of the response statistics. For the time being, the subsample size is taken to be 50 realizations (12.5% of the realizations ensemble). Later, the impact of the number of realizations in the subsample on the errors in the estimated statistics will be quantitatively studied. So the initial focus is on the impact

of different clustering methods, different similarity measures, and different sampling schemes; on the accuracy of the estimated statistics of the response variable (i.e. the head).

Table 3. 1: Cophenetic Correlation Coefficient for the Hierarchical Sampling

Distance	Euclide-an	Standarized Euclidean	Cityblock	Minko-wski	Cosine	Corre-lation	Spea-rman	Cheb-ychev
Average	0.90	0.91	N/A	N/A	N/A	N/A	N/A	N/A
Centroid	0.91	0.92	0.90	0.91	0.28	0.28	0.27	0.68
Complete	0.68	0.76	0.49	0.68	0.31	0.31	0.22	0.35
Single	0.90	0.91	0.88	0.90	0.22	0.25	0.25	0.56
Ward	0.46	0.48	N/A	N/A	N/A	N/A	N/A	N/A
Weighted	0.70	0.60	0.74	0.70	0.35	0.40	0.28	0.53

N/A – Not Applicable

For each node in the numerical grid, the mismatch measures (Equations 3.10, 3.11) and errors in the means and standard deviations are computed as shown in step (6) of the hierarchical clustering with stratified sampling procedure explained above. However, in order to determine the overall performance of any clustering scheme, the mean of the squared errors of all nodes in the grid is determined. More specifically, the means of the squared errors $\boldsymbol{mse_e}, \boldsymbol{mse_{e'}}, \boldsymbol{mse_m}$, and $\boldsymbol{mse_s}$ of the estimated $\boldsymbol{e}$, $\boldsymbol{e'}$, mean, and the standard deviation respectively, are calculated to compare between different linkage-distance criteria combinations. These single number statistics are used as an indicator of the performance of the methodology.

The results of using the hierarchical clustering are summarized in Table 3.2 for both stratified sampling and centriod based sampling. Six linkage methods and 8 distance metrics, which are available in the Matlab statistical toolbox (Jones 1997), were used.

For the centriod based sampling, it can be noticed that the minimum equal weight mismatch measure ($\boldsymbol{mse_{e'}}$) occurred at the *Complete-Spearman* linkage-distance combination, while the minimum tail weighted mismatch measure ($\boldsymbol{mse_e}$) was achieved at the *Single-Spearman* criteria. In some practical cases, the analyst main concern is on the first two moments, namely the mean and the standard deviation. The mean squared errors of the estimated mean and standard deviation ($mse_m, and\ mse_s$) show that the minimum error achieved in the estimated mean is 0.003 m^2 (0.05 m) at the *Average-Spearman*; and for the estimated standard deviation is 0.004 m^2 at the *Complete-Cosine* criteria. In general, it can be seen that the errors in the estimated means range between 5.4 cm and 53 cm, while the errors in the estimated standard deviation range between 6 cm to 37 cm.

For the stratified sampling, the averages of the calculated $mse_e, mse_{e'}, mse_m, and\ mse_s$ of the 50 repeated samplings are, also, reported in Table 3.2. As mentioned previously, the logic behind repeating the stratified sampling is to understand the consistency of the estimated statistics. The coefficient of variation (CV) of the equal weight mismatch (e') can be seen as a measure of the consistency in the estimated parameters. For example, smaller CV values reveal more consistence sampling results. From Table 3.2, the *Averaged-Squared Euclidian* criterion produces the lowest CV value.

Table 3.3 summarizes the results of using K-means clustering. The Four distance metrics available in the statistical Matlab toolbox (Jones 1997) were used. For the centriod based sampling, the equal mismatch measures ($mse_{e'}$) range between 0.021 and 0.038 and the lowest error occurred at the *cosine* metric. The errors in the means

(mse_m) are between 0.002 m^2 to 0.020 m^2 which are equivalent to absolute residuals (not squared) of the errors of 4.7 cm and 13.7 cm. The errors in the estimated standard deviations (mse_s) range between 0.004 m^2 - 0.025 m^2 (equivalent absolute residuals are 6.3 cm -15.8 cm). The stratified sampling average errors ($mse_{e'}$) are roughly the same, which are around 0.033. The average errors in the estimated means and estimated standard deviations are around 0.004 and 0.008 respectively.

Tables 3.2 and 3.3 provide different statistical measures to evaluate the overall performance of the different clustering methods, the error in the mean and standard deviation of different realization reduction methods were plotted (Figure 3.3 and 3.4) at each cell in the numerical grid. The spatial distribution in the mean and standard deviation errors are shown in Figure 3.3 for the hierarchical method at the *Complete-Correlation* criteria and for the K-means method using the *Correlation* distance. It can be seen that errors in the means range between 0.1 m to -0.23 m for the K-means method (Figure 3.4) and between 0.11 m to -0.20 m for the hierarchical method (Figure 3.3). It is worthy of mention that the errors displayed in Figures 3.3 and 3.4 appear to be spatially correlated.

Table 3. 2: Mismatch Errors for Hierarchical Clustering

Linkage Type	Distance	Centriod sampling				Stratified Sampling				
		mse_e	$mse_{e'}$	mse_m	mse_s	mse_e	$mse_{e'}$	mse_m	mse_s	CV(e')
Average	Euclidean	0.262	0.044	0.169	0.100	0.425	0.026	0.023	0.011	0.090
	Seuclidean	0.249	0.044	0.177	0.122	0.424	0.025	0.024	0.011	0.085
	Cityblock	0.261	0.045	0.172	0.111	0.405	0.025	0.024	0.011	0.085
	Minkowski	0.262	0.044	0.169	0.100	0.447	0.025	0.021	0.011	0.091
	Cosine	0.345	0.026	0.005	0.009	0.583	0.027	0.016	0.007	0.120
	Correlation	0.451	0.027	0.008	0.012	0.678	0.028	0.019	0.007	0.141
	Spearman	0.293	0.024	0.003	0.007	0.525	0.028	0.018	0.006	0.136
	Chebychev	0.222	0.036	0.094	0.050	0.478	0.028	0.015	0.010	0.134
Centroid	Euclidean	0.238	0.045	0.179	0.123	0.427	0.027	0.025	0.011	0.134
	Seuclidean	0.242	0.044	0.181	0.126	0.388	0.027	0.026	0.011	0.135
Complete	Euclidean	0.422	0.036	0.064	0.016	0.539	0.026	0.014	0.007	0.126
	Seuclidean	0.309	0.038	0.065	0.025	0.498	0.026	0.011	0.007	0.124
	Cityblock	0.442	0.041	0.035	0.018	0.561	0.026	0.015	0.007	0.123
	Minkowski	0.422	0.036	0.064	0.016	0.513	0.027	0.015	0.007	0.122
	Cosine	0.333	0.022	0.009	0.004	0.479	0.027	0.020	0.008	0.126
	Correlation	0.270	0.023	0.012	0.006	0.562	0.027	0.023	0.009	0.138
	Spearman	0.226	0.019	0.004	0.007	0.577	0.027	0.019	0.008	0.143
	Chebychev	0.245	0.031	0.011	0.016	0.533	0.028	0.014	0.007	0.143
Single	Euclidean	0.247	0.045	0.177	0.116	0.421	0.027	0.023	0.011	0.139
	Seuclidean	0.242	0.044	0.181	0.126	0.466	0.027	0.023	0.011	0.139
	Cityblock	0.253	0.044	0.179	0.118	0.411	0.027	0.023	0.011	0.138
	Minkowski	0.247	0.045	0.177	0.116	0.440	0.027	0.024	0.012	0.137
	Cosine	0.218	0.050	0.179	0.129	0.472	0.027	0.022	0.011	0.137
	Correlation	0.940	0.066	0.279	0.097	0.444	0.027	0.027	0.011	0.136
	Spearman	0.181	0.044	0.186	0.143	0.445	0.027	0.024	0.012	0.136
	Chebychev	0.224	0.045	0.176	0.124	0.480	0.027	0.021	0.012	0.135
Ward	Euclidean	0.367	0.028	0.013	0.008	0.616	0.027	0.022	0.007	0.139
	Seuclidean	0.345	0.029	0.008	0.008	0.518	0.027	0.014	0.007	0.140
Weighted	Euclidean	1.107	0.045	0.107	0.020	0.494	0.027	0.014	0.008	0.151
	Seuclidean	0.285	0.041	0.089	0.053	0.430	0.027	0.023	0.012	0.151
	Cityblock	0.362	0.039	0.070	0.024	0.453	0.027	0.021	0.010	0.150
	Minkowski	1.107	0.045	0.107	0.020	0.445	0.027	0.015	0.008	0.150
	Cosine	0.359	0.025	0.006	0.009	0.568	0.027	0.017	0.008	0.150
	Correlation	0.410	0.023	0.007	0.007	0.698	0.028	0.021	0.007	0.151
	Spearman	0.272	0.023	0.006	0.010	0.528	0.028	0.018	0.008	0.151
	Chebychev	0.238	0.037	0.066	0.033	0.421	0.028	0.014	0.008	0.150

Table 3. 3: Mismatch Errors for K-means Clustering

Distance	Centriod Sampling				Stratified Sampling				
	mse_e	$mse_{e'}$	mse_m	mse_s	mse_e	$mse_{e'}$	mse_m	mse_s	$CV(e')$
SqEuclidean	0.324	0.029	0.008	0.013	0.565	0.033	0.021	0.007	0.116
Cityblock	1.015	0.038	0.020	0.025	0.581	0.032	0.020	0.008	0.111
Cosine	0.312	0.021	0.002	0.004	0.541	0.033	0.025	0.010	0.114
Correlation	0.443	0.022	0.011	0.005	0.543	0.033	0.027	0.009	0.109

The impact of the distance criteria on the spatial errors are plotted for the hierarchical (Figure 3.5) and the K-means method (Figure 3.6). Figure 3.6 shows the equal weight mismatch (**e'**) for the K-mean method. The mismatch errors tend to be highest in the middle of the model domain and gradually decrease in the proximity of the boundary conditions. A possible explanation is that near the boundary conditions the hydraulic head variability range is narrower compared to that in the middle of the model domain.

To visually evaluate the mismatch between the reference CDF and the estimated CDF, two points were chosen to make the comparison; the first point is point A (200 m, 1600 m) which is adjacent to the upper stream head boundary condition, and the second point is point B (1600 m, 1600 m) which is in the middle of the domain. The estimated CDFs obtained from the hierarchical clustering and K-means method were plotted. The reference CDF at point A is negatively skewed because of its proximity to the boundary condition while point B is approximately symmetric. As shown by Figures 3.7 and 3.8 there appears to be good correlation between the estimated and the reference CDFs for both points.

Previous results were limited to a subsample consisting of 50 realizations (12.5% of the ensemble size). To evaluate the impact of the subsample size on the accuracy of the estimation statistics, the procedures were repeated at different sample sizes. Initially, the subsample size consists of 10 realizations and the size was increased gradually by 10; at each subsample size the mismatch error was evaluated and plotted as shown in Figures 3.9 and 3.10 for hierarchical and K-means methods respectively. Hierarchical clustering using *Cosine* and *Correlation* metrics produce mismatch errors of 0.06 using 10 realizations and dropped to 0.024 using 50 realizations. The *Spearman* metric performs better than the *Cosine* and *Correlation* metrics for sample sizes less than 150 realizations, after which the three metrics tend to have the same performance. The K-means method's performance using *Cosine*, *Correlation* and C*ityblock* metrics are plotted in Figure 3.10. The *Cosine* and *Correlation* mismatch errors at 10 realizations are 0.047 and 0.038 respectively, both of which are less than the hierarchical method. The *Cityblock* metric is not as efficient as the other metrics for subsample sizes less than 320 realizations.

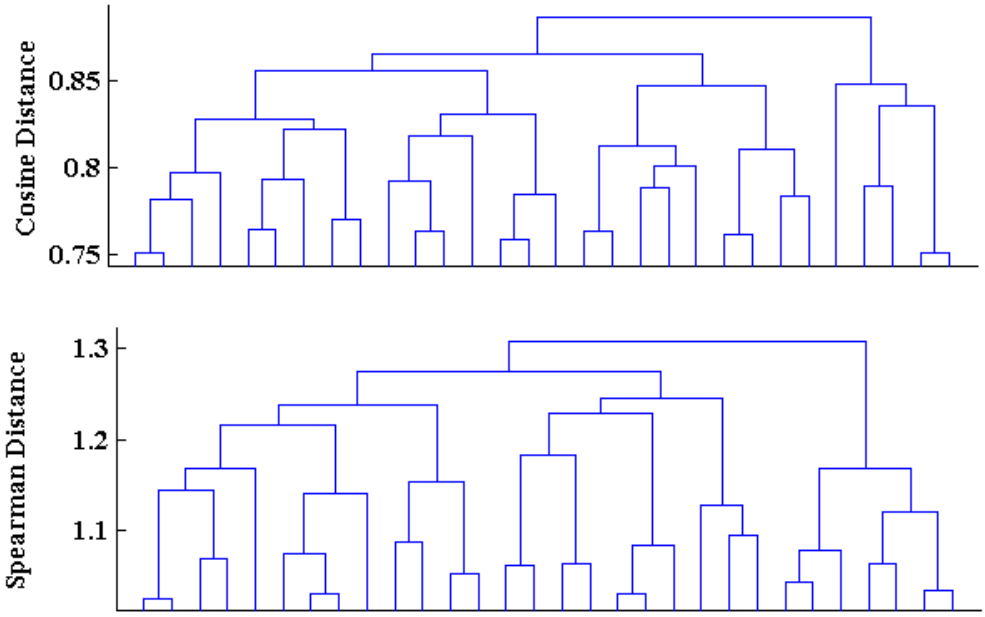

Figure 3. 2: Clusters tree for Hierarchal Clustering Using Furthest Linkage for Cosine and Spearman Distances

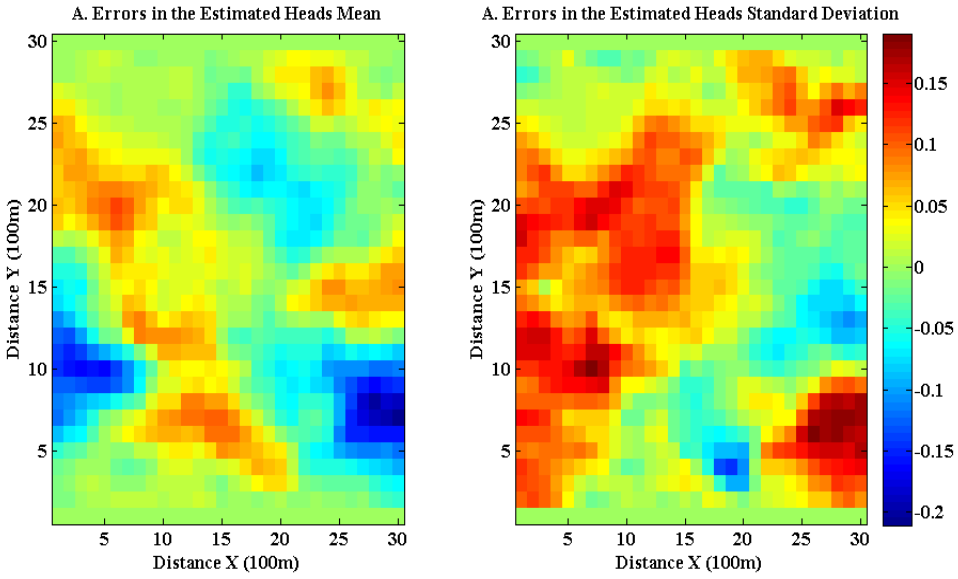

Figure 3. 3: Errors in Estimated Means and Standard Deviations Using the Hierarchical Method

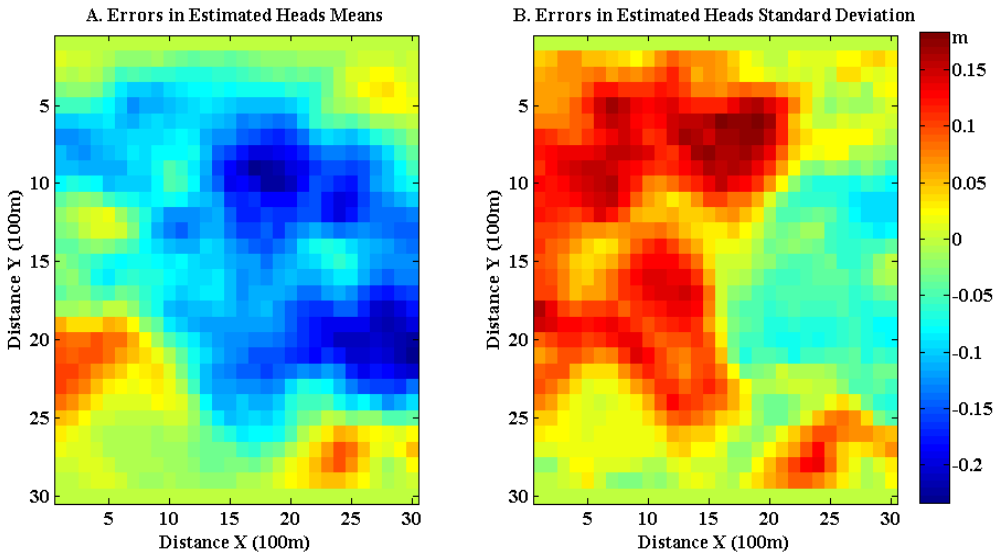

Figure 3. 4: Errors in Estimated Means and Standard Deviations Using the K-means Method

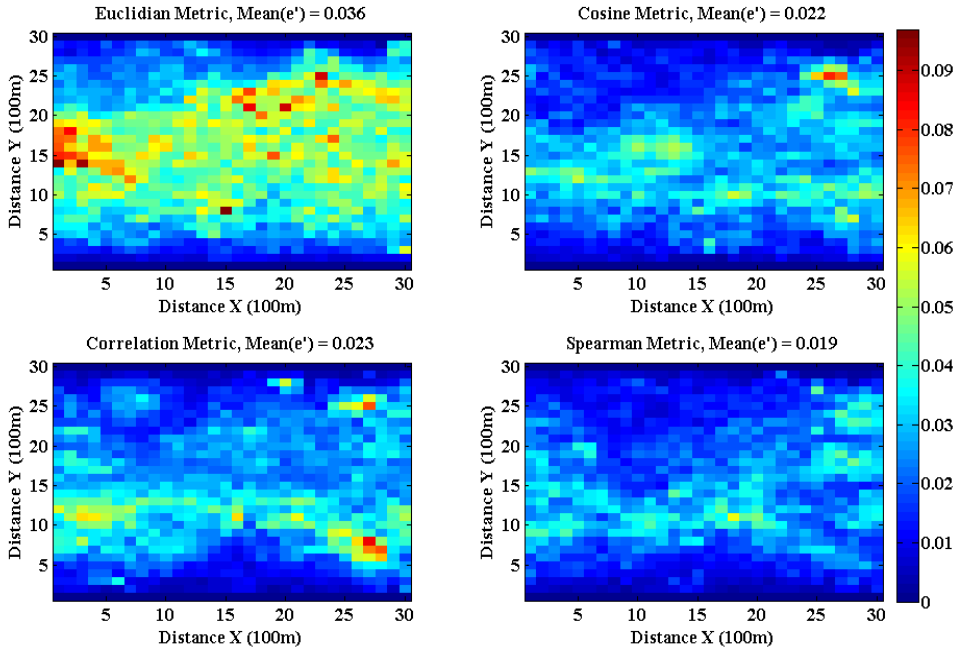

Figure 3. 5: Mismatch Measures (e') Using Different Distance Metrics in the Hierarchical Method

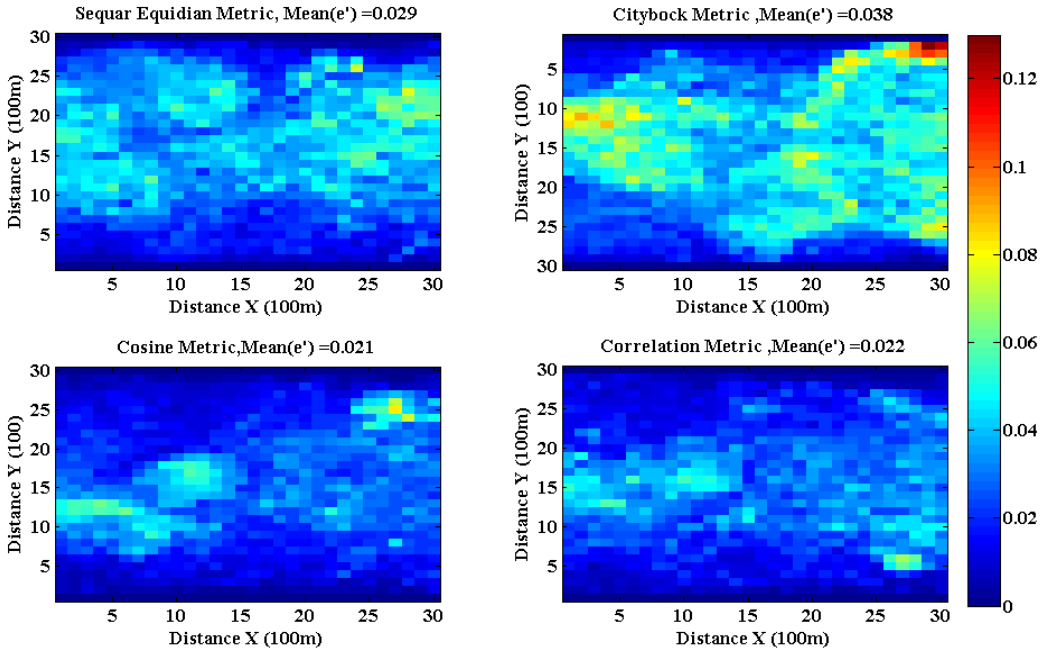

Figure 3. 6: Mismatch Measures (e') Using Different Distance Metrics in the K-means Method

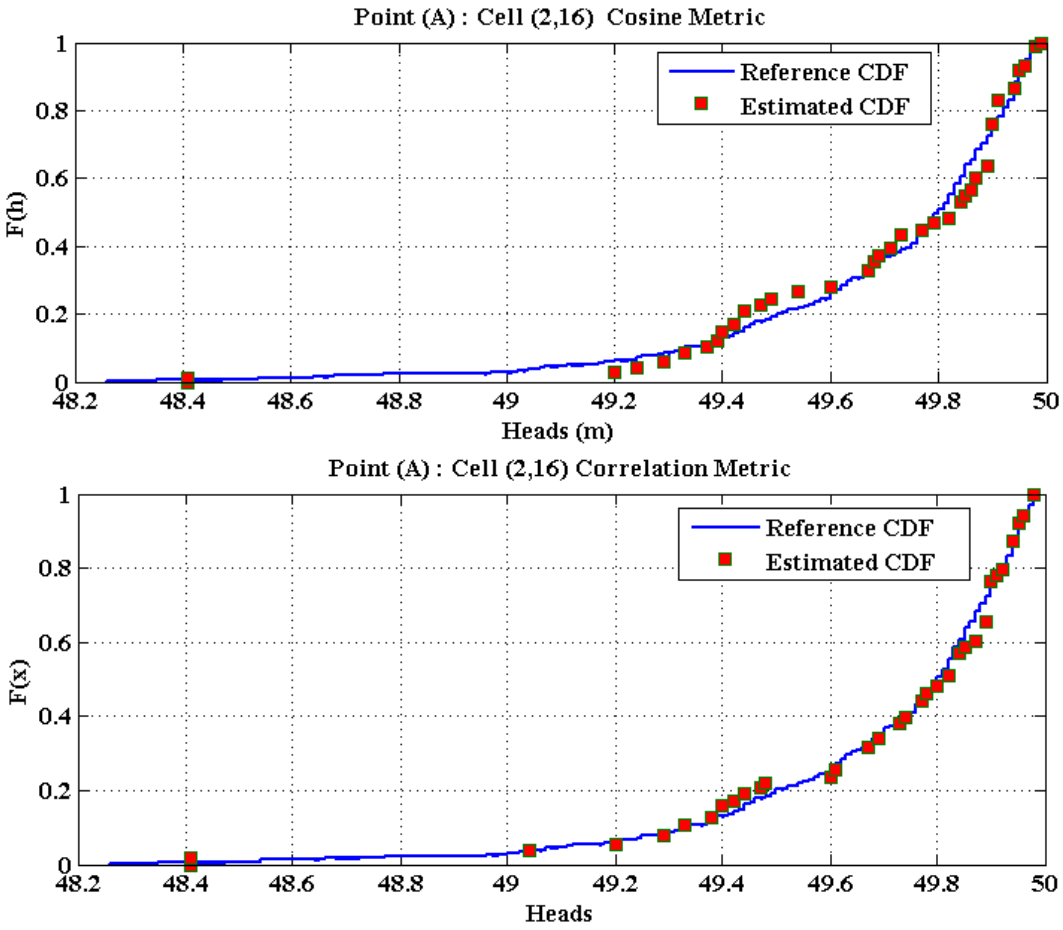

Figure 3. 7: Estimated and Reference CDF Using K-means Clustering at Point (A)

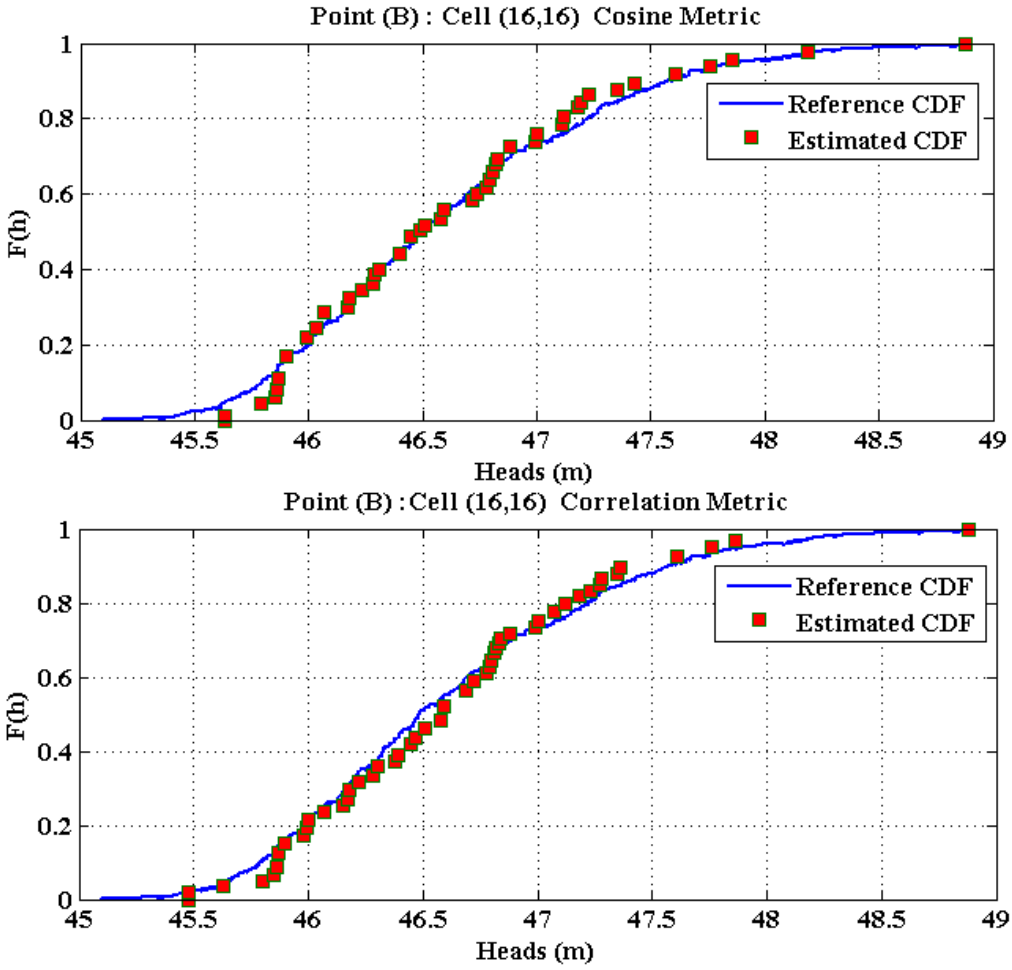

Figure 3. 8: Estimated and Reference CDF Using K-means Clustering at Point (B)

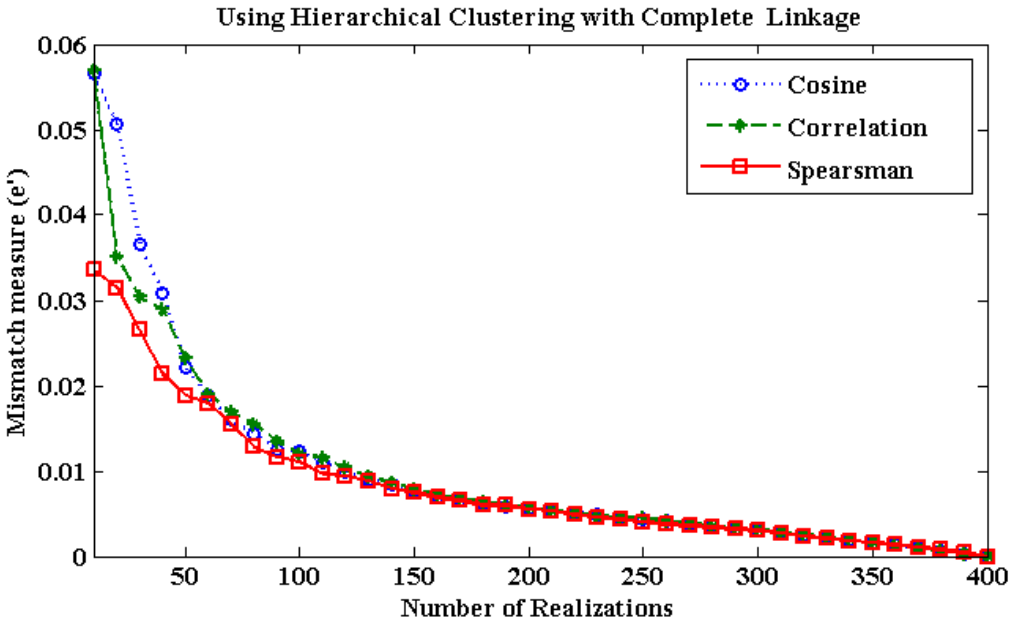

Figure 3. 9: Impact of Subsample Size on the Mismatch Measure (e') Using Hierarchical Clustering

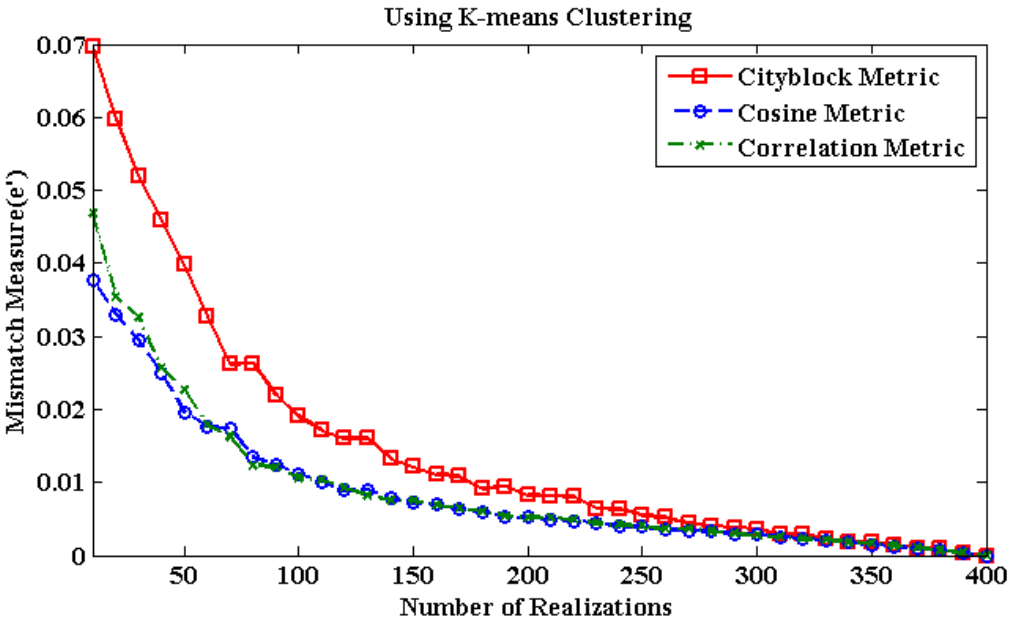

Figure 3. 10: Impact of Subsample Size on the Mismatch Measure (e') Using K-means Clustering

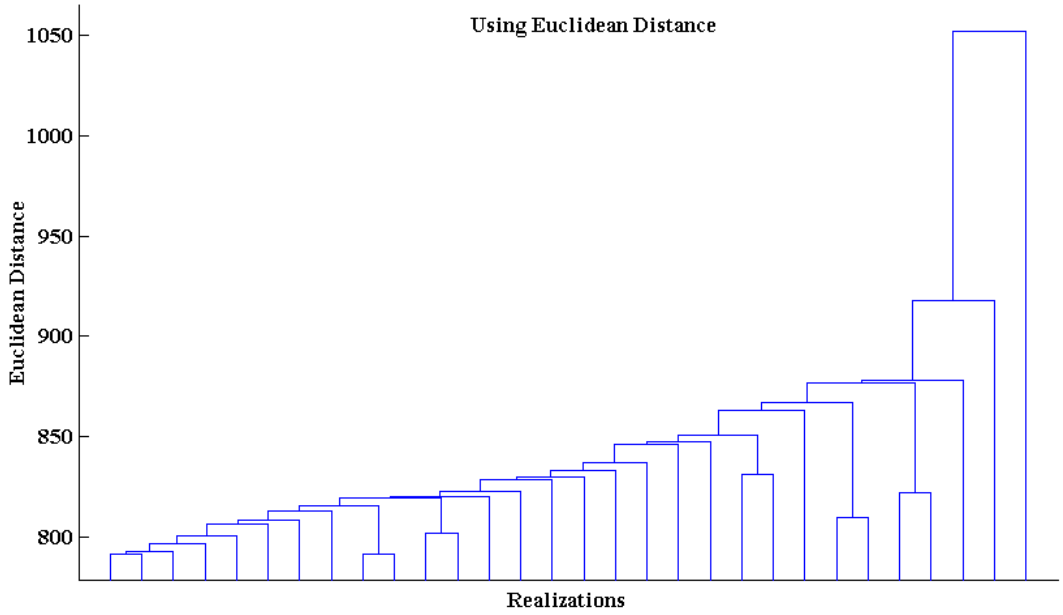

Figure 3. 11: Cluster Tree Shows the Unbalance in the Cluster Size

3.7 Discussion

As discussed in the methodology section, the implementation of cluster analysis to reduce the computational effort involved in MC simulation requires the following four main decisions: 1) choosing the clustering method, namely the hierarchical versus the K-means methods; 2) choosing the similarity measure and the linkage criteria in the case of hierarchical clustering; 3) selecting which sampling method to use, namely the stratified sampling versus the centriod based sampling; and 4) determining the size of the subsample. Each of these decisions affects the estimation accuracy as was presented in the results section.

3.7.1 Choosing a Clustering Method

The comparison in this paper is limited to the hierarchical and K-means method. The advantage of the hierarchical method is that it produces unique and deterministic clusters for a specified distance-linkage combination. This seems to be attractive, however the final structure of the clusters is confined to the merging of clusters at the leaves level (individual realizations). Once a realization is assigned to a cluster it will stay in it. Generally speaking, the intra variance of the realizations in each cluster should be at a minimum to guarantee similarity between realizations in the same cluster. Unfortunately this requirement is not part of the hierarchical clustering algorithm. On the other hand, the K-means method is an iterative clustering algorithm that seeks minimum intra cluster variances. The clusters produced using the K-means method are different each time the realizations are partitioned. This disadvantage can be avoided by repeating the clustering process several times and selecting the cluster scheme that produces the minimum intra cluster variances. Fortunately, repeating the K-means clustering is not computationally intensive.

In terms of comparison between the two methods, the K-means method performs slightly better than the hierarchal method. For example the errors ($\boldsymbol{mse_{e\prime}}$) associated with the cosine metric at 10, 50 and 100 realizations are 0.038, 0.021 and 0.012, respectively for the K-means method and 0.06, 0.023, 0.011 respectively for the hierarchical method. The same error ($\boldsymbol{mse_{e\prime}}$) behavior (K-means method performs better at low subsample sizes) can be seen for the *Correlation* metric. It's clear that the performance of the K-means method at small subsample sizes is better than the

hierarchical method and this advantage diminishes as the subsample size increases. A possible explanation for this is that the K-means algorithm seeks a minimum intra cluster variance and that for this work clustering was repeated 50 times and the clustering scheme that achieves the minimum intra cluster variance was selected. Since the objective of the analysis is to reduce the subsample size, the optimized K-means method performed better in this example.

3.7.2 Choosing a Distance-linkage Criteria

Thinking of each realization as a point in hyperspace is important to understanding the differences between different distance metrics. Two points, or realizations, that are separated by a small Euclidian distance have similar responses, if it is assumed that the flow equation is continuous. Results show that the *Euclidian* distance measure and the *Standard Euclidian* distance measure usually produce dissimilar clusters (see Table 3.2). However, by analyzing the cluster trees for them (Figure 3.11) it can be seen that *Euclidian* based distances usually produce one large cluster (cloud of points around the centriod) that contains most of the realizations and other clusters that contain fewer realizations. This unbalance in cluster sizes result in a poor estimation of a reference CDF. If stratified sampling is used, most of the samples will be taken from the large cluster, according to Equation 3.5, leaving the extreme realizations under represented.

On the other hand, the *Cosine*, *Correlation* and *Spearman's* distances are quite similar in that they provide a measure of dispersion of the realizations. The *Correlation* and *Spearman's* distances provide measures of dispersion around the centriod of the

points while the *Cosine* provides the dispersion of the points around the origin point. The magnitude of dispersions around the centriod of the points is usually larger than the magnitude of dispersions around the origin point. It is worth noting that *Spearman* distances perform much better than the *Cosine* and the *Correlation* distances which are quite similar. The *Cityblock* and *Minkowski* metrics are similar in that they provide the summation of the absolute difference between two vectors. These metrics usually have the same performance as the Euclidian distance metrics, namely they produce unbalance clustering trees.

Choosing suitable linkage criteria has a major impact on the clusters that are produced. By comparing the $\boldsymbol{mse}_{e'}$ errors in Table 3.2 for different linkage methods but at the same distance metric, it can be seen that the *Complete* linkage method generated the smallest errors. The *Average* linkage method performs second best and the *Single* linkage method generated poor results in general.

3.7.3 Choosing a Sampling Scheme

The major disadvantage of stratified sampling is that it produces different response statistics every time it is implemented. In this paper stratified sampling is repeated multiple times to investigate the variance in the estimated statistics. In a practical problem, the luxury of repeating the stratified sampling and choosing the minimum variance distance-linkage combination is not an option. This is because it requires large amounts of CPU resources and it will render using cluster analysis, to reduce the effort associated with the numerical simulations, impractical. On the contrary, the centriod based sampling provides a practical and consistent sampling

scheme for three reasons: 1) sampling will be carried out only one time and the results are unique, 2) the number of required clusters is larger compared to the stratified sampling, and 3) the realizations chosen are the mean realizations and they can be deemed as legitimately representing the whole cluster in contrast to the stratified sampling where sampling is achieved randomly. The centriod based sampling is similar to the composite sampling; the only difference is that the closest sample to the centriod is collected instead of using the centriod realization itself (the composite sample). The mismatch errors ($\boldsymbol{mse_{e'}}$) in using the centriod based sampling are usually larger than those of the stratified sampling (Tables 3.2 and 3.3). However, the reported values for the stratified sampling's ($\boldsymbol{mse_{e'}}$) mismatch errors in Table 3.2 and Table 3.3 are the average of 50 sampling repetitions. Despite repeating the stratified sampling 50 times, the centriod based sampling still produces the same ($\boldsymbol{mse_{e'}}$) errors especially for the *Complete* linkage criteria and for the *Cosine*, *Correlation*, and *Spearman* distance metrics.

3.7.4 Subsample Size

In practical applications of the cluster analysis to reduce CPU time to approximate the flow response uncertainty, it is required to determine the size of the subsample in advance. It is clear from Figures 3.9 and 3.10 that the increase in the subsample size reduces $\boldsymbol{mse_{e'}}$ errors significantly in the first 100 realizations (25% of the ensemble size) after which the decrease rate in the error is smaller. For example, in Figure 3.9 and at the *Cosine* metric, the mismatch error ($\boldsymbol{mse_{e'}}$) dropped from around 0.057 to 0.023 when the subsample size increased from 10 realizations to 50

realizations. Suggesting a subsample size that is optimum for all problems is not an easy task, however it can be seen that a subsample size that is 25% of the ensemble size could be a good estimate.

3.8 Conclusion and summary

The computational effort required by the MC simulation method to account for uncertainty in heterogeneous aquifers is the main drawback of this powerful and simple method. In this paper, a methodology to reduce CPU time by reducing the number of realizations to be processed has been outlined. Cluster analysis (CA) has been widely used to cluster scalar field data and multi-dimensional field data. In this study, CA is employed to cluster the large output of the geostatistical simulators. To efficiently apply this method, the clusters generated should be significantly different from each other, while producing minimum intra cluster variance.

Two clustering methods have been utilized, namely the hierarchical and the K-means methods. Within each method, numbers of similarity metrics have been tested. The hypothesis behind this paper is that similar realizations produce similar responses, and there is no necessity to process all realizations in the ensemble. Strictly speaking, this assumption is valid if similarity measures are accurately partitioning the ensemble and that the governing flow equation is continuous in the clustered parameter (for example, the hydraulic conductivity).

The next step following clustering the ensemble is to collect a subsample of realizations that represents the whole ensemble. The stratified sampling and the

centriod sampling were investigated. Results show that in general the centriod based sampling is equivalent to repeating the stratified sampling several times at *Complete* linkage with *Cosine*, *Correlation* and *Spearman* metrics. This result is promising in the sense of the practicality and consistency of the centriod based sampling.

Results also show that different clustering methods using different similarity metrics have different abilities to reproduce the ensemble response. In general, we found out that the dispersion metrics, such as the *Cosine*, *Correlation* and *Spearman* metrics are more adequate to cluster the realizations. Results also show that sample sizes of more than 25% of the ensemble size can achieve a practical approximation for the ensemble statistical responses.

3.9 References

Abonyi, János, and Balázs Feil. 2007. *Cluster analysis for data mining and system identification*. Springer, August 17.

Anderberg, Michael R. 1973. *Cluster Analysis for Applications*. Academic Pr, December.

Bakr, Adel A., Lynn W. Gelhar, Allan L. Gutjahr, and John R. MacMillan. 1978. "Stochastic Analysis of Spatial Variability in Subsurface Flows 1. Comparison of One- and Three-Dimensional Flows." *Water Resources Research* 14 (2): 263-271. doi:10.1029/WR014i002p00263.

Dagan, G. 1982. "Stochastic modeling of groundwater flow by unconditional and conditional probabilities: 1. Conditional simulation and the direct problem." *Water Resources Research* 18 (4): 813. doi:10.1029/WR018i004p00813.

Deutsch, Clayton V. 2002. *Geostatistical Reservoir Modeling*. 1st ed. Oxford University Press, USA, April 4.

Everitt, Brian S., Sabine Landau, and Morven Leese. 2009. *Cluster Analysis*. 4th ed. Wiley, January 20.

Freeze, R. A. 1975. A stochastic-conceptual analysis of one-dimensional groundwater flow in nonuniform homogeneous media, Water Resour. Res., 11(5), 725–741.

Gilbert, Richard O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. Wiley, February 1.

Gómez-Hernández, J. Jaime, and Jesús Carrera. 1994. "Using Linear Approximations to Rank Realizations in Groundwater Modeling: Application to Worst Case Selection." *Water Resources Research* 30 (7): 2065. doi:10.1029/94WR00322.

Gotovac, Hrvoje, Vladimir Cvetkovic, and Roko Andricevic. 2009. "Flow and travel time statistics in highly heterogeneous porous media." *Water Resources Research* 45 (7) (July). doi:10.1029/2008WR007168. http://www.agu.org/pubs/crossref/2009/2008WR007168.shtml.

Harbaugh, A. W, E. R Banta, M. C Hill, and M. G McDonald. 2000. *MODFLOW-2000, The U. S. Geological Survey Modular Ground-Water Model-User Guide to Modularization Concepts and the Ground-Water Flow Process*. United States Geological Survey.

Jones, B. 1997. "MATLAB, Statistics Toolbox: User's Guide." *The Math Works Inc., Version* 5.

Kupfersberger, H., and C. V. Deutsch. 1999. "Ranking stochastic realizations for improved aquifer response uncertainty assessment." *Journal of Hydrology* 223 (1-2) (September 22): 54-65. doi:doi: DOI: 10.1016/S0022-1694(99)00113-4.

Matern, B. 1986. *Spatial Variation*. 2nd ed. Springer, December 12.

Matheron, G. 1962. "Traité de géostatistique appliquée."

Rubin, Yoram. 2003. *Applied Stochastic Hydrogeology*. Oxford University Press, USA, March 27.

Rubinstein, Reuven Y., and Dirk P. Kroese. 2007. *Simulation and the Monte Carlo Method*. 2nd ed. Wiley-Interscience, December 19.

Smith, R. E., and R. H. B. Hebbert. 1979. "A Monte Carlo analysis of the hydrologic effects of spatial variability of infiltration." *Water Resources Research* 15 (2).

Tang, D. H., and G. F. Pinder. 1977. "Simulation of groundwater flow and mass transport under uncertainty." *Advances in Water Resources* 1 (1) (September): 25-30. DOI: 10.1016/0309-1708(77)90005-7.

Yeh, T.-C.J., 1992. "Stochastic modelling of groundwater flow and solute transport in aquifers." *Hydrological Processes* 6 (4): 369-395.

4 MULTIVARIATE STOCHASTIC ANALYSIS OF FLOW AND SALINITY TRANSPORT IN A SUBSURFACE DRAINED FIELD

4.1 General

Spatial heterogeneity of soil, flow and transport properties and uncertainty in root-uptake model parameters make numerical prediction of crop yield prone to high degree of uncertainty. This paper discusses a method for accounting for the uncertainty of correlated regionalized soil parameters using multivariate Monte Carlo simulation. Sequential indicator simulation is used to generate three-dimensional correlated realizations for hydraulic conductivity, porosity, Van Genuchten parameters, and dispersivity. Other semi empirical parameters that control crop water uptake, irrigation efficiency, and subsurface drainage conductance were randomized. The generated ensembles for each of the soil parameters were processed in the variably saturated flow and transport model (CSUID) to obtain the spatial statistical moments of the relative crop yield for two crops, i.e. alfalfa and corn. Moreover, the spatial variability of statistical moments of root zone salinity and vertical soil flux were obtained. Furthermore, the statistical properties of hydrographs of drainage outflow and salinity were explored. Results show that parameter uncertainty significantly affects the predicted spatial variability of hydrosalinity responses, which consequently affects the

in-field relative crop yield variability. A maximum standard deviation of 30% for the predicted corn crop yield was observed and of 26% for alfalfa yield.

4.2 Introduction

As the demand on food, fiber, and biofuel increases (Schnepf, 2010), the burden on agricultural crop production to fulfill these demands is increasing as well. Several factors make such a burden restrictive to the long-term sustainability of crop production. For example, despite the increase in crop production induced by modern irrigation practices, the continuous salinization of root zone salinity in several countries around the world poses a major problem for the sustainability of their production process.

Leaching salts out of the root zone is commonly achieved by increasing the volume of irrigation water; however, this practice places pressure on the tight fresh water budget, particularly in arid and semi-arid regions. Furthermore, excessive application of irrigation might cause a high groundwater table to develop, especially in regions where a shallow impermeable layer exists or where the storage capacity of the vadose zone is limited. Additionally, seepage losses from nearby irrigation canals, reservoirs, and return flow to streams make the salinity and waterlogging problem not only a field scale problem, but also a regional or basin scale problem.

The combined effect of shallow groundwater table with high soil salinity is widely recognized as a plague that affects crop production worldwide. According to several resources (Tanji 1990; Postel 1989; Umali and Umali-Deininger 1993;

Ghassemi, et al. 1995; and Wichelns 1999), lands affected by excessive salinity comprise up to 28% of all irrigated land in the US, 23% in China and 21% in Pakistan. Ghassemi et al. (1995) estimated annual worldwide economic losses to be around 11 billion dollars. This number is expected to be higher today.

Devising efficient irrigation and drainage management plans is necessary to counteract the current deterioration in land productivity. Consequently, the performance of any proposed management plan should be evaluated based on, among other factors; the expected improvement in the crop yield, the extent of potential environmental risks that might result from drain effluent; and the long-term performance of field productivity (i.e. the sustainability of the cultivation activities and the prevention of salt accumulation in the root zone). However, such evaluations could not be achieved at a reasonable cost without the use of numerical models, which approximate the complex interactions between soil, water, plants and the atmosphere, on one side, and the proposed management plan on the other side.

Despite of the importance of numerical models in the decision making process, modelers still debate the validity of their predictions (Konikow et al. 1992; Oreskes et al. 1994). For example, Konikow et al. (1992) argued that calibrating the model's parameters to match historical data of the system's response is usually not enough to guarantee the model's validity since the number of unknowns is usually larger than the number of parameters to be estimated. That is to say, it seems that one unique and trusted prediction of numerical models is far from being attainable, and it is more reasonable to deal with the issue of prediction in a probabilistic framework. In this framework, model inputs are described by using probability distribution functions to

reflect parameter uncertainty. Then, the uncertainty in the input parameters is propagated analytically or via Monte Carlo simulations to obtain the response's statistical properties.

This study is part of a wider effort to investigate the problem of salinization and waterlogging in the Lower Arkansas River Basin in Colorado (Gates, et al. 2006). The Arkansas River is more affected by salinity than any other river in the US (Miles 1977; Tanji 1990). An extensive data sampling effort has been carried out from 1999 to 2009 to characterize the spatial and temporal extent of the soil salinization and waterlogging problem. Burkhalter, et al. (2005) developed a regional numerical model to investigate several management scenarios. Houk et al. (2004) estimated the direct average forgone profit to be around $4.3 million/year in Otero County ($68/acre per year); additionally, the indirect and induced costs associated with waterlogging and soil salinization in Otero County are estimated to have increased by approximately 20%.

In this study, the uncertainty aspect of the waterlogging and salinization problem is tackled on a field scale. Generally speaking, understanding the role that spatial variability plays in crop yield prediction is vital to guide future data-collection efforts and to avoid risks that arise from incomplete knowledge of the controlling parameters. A Multivariate Monte Carlo Analysis is implemented herein to include a wide array of independent and dependent input parameters that control crop yield and subsurface drainage. The controlling parameters are found to be either spatially random correlated soil properties such as hydraulic conductivity, Van Genuchten parameters, dispersivity, porosity, and irrigation uniformity; or semi empirical parameters that control root growth, water uptake and drainage outflow simulation.

A number of researchers have studied the spatial variability of crop yield. For example, Warrick et al. (1983) studied the impact of soil heterogeneity, represented as field capacity, wilting point, and irrigation uniformity, on crop yield. A linear response function between crop yield and soil and uniformity variability was assumed. Bresler et al. (1988) studied the impact of uncertainty of soil parameters and uptake model parameters for a one-dimensional model on yield uncertainty. The parameters are assumed statistically independent and uniform in the vertical direction. Rubin et al. (1993) used a stochastic analytic perturbation method to study the impact of soil spatial variability on water uptake by plants using a one dimensional steady state unsaturated flow case. Muralidharan et al. (2009) showed that the variability of spatial infiltration increases the applied irrigation water and deep percolation flows by very substantial amounts compared to uniform infiltration. Recently, Montazar (2010) studied the impact of irrigation spatial uniformity on the yield of alfalfa hay.

Using numerical models to analyze the uncertainty associated with subsurface drainage systems is the subject of several studies. For example, Haan et al. (2003) studied the impact of input parameter uncertainty on the drain outflow and relative crop yield using DRAINMOD. Monte Carlo simulations and first order approximation methods were used to determine the most sensitive uncertain parameters for a simplified layered subsurface system. Wang et al. (2006) employed the generalized likelihood uncertainty estimation (GLUE) procedures to evaluate the uncertainty in DRAINMOD predictions of the subsurface drain flow. To the best of the authors knowledge, simulation of the performance of subsurface drainage systems in a fully

three dimensional heterogeneous aquifer system is still absent from the published literature.

This study presents a comprehensive approach to dealing with crop production uncertainty in a subsurface drained field scale problem. More specifically, this paper tackles the following points: first, investigates the temporal and spatial variability of root zone water content, salinity, root extraction rates and groundwater table depths as a response to uncertainty in input parameters; second, studies the spatial variability of crop yield statistical moments for two crops, alfalfa and corn; and third, evaluates the performance of subsurface drainage in a three-dimensional random soil domain.

The paper is organized as follows: In section 4.3, the theoretical background of the variably saturated flow and transport model is outlined. Then, the methodology of the multivariate stochastic analysis is illustrated in section 4.4. The baseline conditions and site description are presented in section 4.5. Next, the statistical PDF's of the input parameters are developed in section 4.6. Subsequently, the numerical implementation of the method is shown in section 4.7, followed by the results and the discussion.

4.3 Theoretical Framework

A large number of parameters such as soil fertility, local pests, soil chemistry, crop type, climate conditions and cultivation practices affects crop yield. In this study, it is assumed that the agronomic conditions are excellent and the only limiting factors are the soil hydrosalinity conditions. Crop yield is modeled using the following

relationship, which is based on the assumption of a linear relationship between the relative evapotranspiration and relative crop yield:

$$\frac{Y_a}{Y_o} = 1 - k_y\left(1 - \frac{ET_a}{ET_o}\right) \tag{4.1}$$

Where Y_a is the actual dry matter yield [M], Y_o is the maximum harvested dry matter yield [M], k_y is the yield response factor, ET_a is the total (seasonal) actual evapotranspiration, and ET_o is the maximum seasonal evapotranpiration (The reference ET), which is obtained from climatic data.

According to Equation 4.1, the calculation of the relative crop yield is equivalent to the calculation of actual *ET*. The total actual evapotranspiration ET_a is approximated by double integrating the temporal root extraction rate $\boldsymbol{Q_r(z,t)}$ over the growing season and over the root zone depth (Equation 4.2).

$$\boldsymbol{ET_a = \int_0^T \int_0^{D(t)} Q_r(z,t)\, dzdt}\,, \tag{4.2}$$

where $Q_r(z,t)$ is the temporal root extraction [L^3/T] at a vertical depth z, T is the overall season length [T], D is the root depth [L]at time t.

It is appropriate in this study to adopt macroscopic modeling of the root uptake, in which the uptake rate is represented as a nonlinear sink term in the flow and transport equations. The nonlinearity of calculating the root uptake $Q_r(z,t)$ stems from its dependency on the capillary head and the osmotic head induced by salinity. For overviews of root uptake models, readers are referred to Molz (1981) and Hopmans and

Bristow (2002). The overall sink term that accounts for root density, root geometry, capillary head, and osmotic head can be modeled via Equation (4.3).

$$\boldsymbol{Q_r(z,t) = \lambda(z,t).\frac{ET(t)}{\Delta A}.\alpha(\psi,\psi_o).K_c} \tag{4.3}$$

Where $ET(t)\ is$ the temporal reference evapotranspiration rate [L/T], K_c is the crop growth coefficient at time t, $\boldsymbol{\Delta A}$ is area [L^2], and $\lambda(z,t)$ is the root density term that describes the density and the geometry of the root network with respect to the depth; and can be modeled using the S function (Equation 4.4).

$$\boldsymbol{\lambda(z,t) = \frac{-1.6z}{D(t)^2} + \frac{1.8}{D(t)}} \tag{4.4}$$

Where z is the depth at which the root density is calculated [L] and $D(t)$ is the root depth at time t [L]. The temporal root growth can be approximated using the following equation from Hanks and Hill (1980)

$$\boldsymbol{D(t) = \frac{D_{max}}{\left(1+\exp\left(a-b\frac{t}{t'}\right)\right)}} \tag{4.5}$$

where $D(t)$ is the root depth at time t, D_{max} is the maximum root depth, t' is the end of the crop's third stage of growth, and a, b are empirical coefficients.

Feddes et al. (1976) pioneered describing the sink term as a function of water content; and Van Genuchten (1987) extended it to incorporate osmotic head. This model does not take into account yield reduction due to the long saturation of the root zone. To overcome this obstacle, we modify the Cardon and Letey (1992a) equation, which is a slight modification of Van Genuchten (1987), to account for waterlogging. That is, as $\boldsymbol{\psi}$ in Equation 4.6 increases, or become more saturated; the term $\boldsymbol{\alpha(\psi,\psi_o)}$

is linearly reduced by multiplying it by$\left(\frac{\psi}{\psi_s} < 1\right)$. The final equation that accounts for water deficit stress, salinity stress and water excess stress (waterlogging) is

$$\alpha(\psi, \psi_o) = \begin{cases} \dfrac{1}{1 + \left(\dfrac{\psi}{\psi_{50}} + \dfrac{\psi_o}{\psi_{o50}}\right)^P} & \psi(z,t) < \psi_s \qquad (4.6.a) \\ \dfrac{\left(\dfrac{\psi}{\psi_s}\right)}{1 + \left(\dfrac{\psi}{\psi_{50}} + \dfrac{\psi_o}{\psi_{o50}}\right)^P} & 0 \geq \psi(z,t) > \psi_s \qquad (4.6.b) \end{cases}$$

Where p is an empirical parameter close to 3, ψ_{50} is the capillary head at which root uptake is reduced by 50% when $\psi_o = 0$ [L], ψ_o is the osmotic head, at which root uptake is reduced by 50% when $\psi_{50} = 0$[L], $\psi(z,t)$ [L]is the capillary head [L] and $\psi_o(z,t)$ is the osmotic head [L], ψ_s is the head threshold after which oxygen deficiency starts to occur[L]. It is recognized that stress due to water excess (near saturation conditions) does not influence root uptake instantaneously (Feddes et al. 1976), but could take a few days (for example, 2 days) to affect the root uptake. As a result, Equation 4.6.b will not be active until the capillary head is equal to or greater than ψ_s for a period of two days.

The evaluation of Equation 4.6 requires the calculation of the capillary head and the salt concentration. Thus, the continuity equation for flow and transport of water and salts in a variably saturated aquifer are mathematically modeled using two partial differential equations for flow and transport (Equations 4.7 and 4.8 respectively).

$$\frac{\partial}{\partial x_i}\left(K_i(\psi)\frac{\partial h}{\partial x_i}\right) + Q_s = \left(\frac{\theta}{\theta_s}S_s + C(\psi)\right)\frac{\partial h}{\partial t} \qquad (4.7)$$

$$\frac{\partial}{\partial x_i}\left(\boldsymbol{\theta D_{ij}}\frac{\partial C}{\partial x_j}\right) - \frac{\partial}{\partial x_i}(\boldsymbol{\theta v_i C}) + \boldsymbol{Q_s C_s} = \frac{\partial(\boldsymbol{\theta C})}{\partial t} \tag{4.8}$$

Where $K_i(\psi)$ is the hydraulic conductivity [L/T], ψ is the capillary head [L], h is the total head [L] ($\psi = h - z$), Q_s is the sink or source term per unit volume [T^{-1}], θ is the moisture content [L^3/L^3], θ_s is the soil porosity [L^3/L^3], S_s is the specific storage [L^{-1}], $C(\psi)$ is the specific capacity [L^{-1}], x is a space vector [L] and $i = 1,2,3$ represents three-dimensional space, t is time [T]. D_{ij} is the hydrodynamic dispersion [L/T^2], C is the salinity concentration [M/L^3], v_i is the seepage velocity [L/T].

The sink term Q_s is the net sinks/sources term [T^{-1}]. Q_i is the irrigation rate [L^3T^{-1}] which is a model input parameter, Q_d is the drain outflow, Q_r is the root uptake [L^3T^{-1}] calculated from equation (4.3). Source terms have a positive sign whereas the sink terms have negative sign.

$$Q_s = \frac{(Q_r + Q_i + Q_d)}{\Delta V} \tag{4.9}$$

It can be noted that Q_r is function of the head and concentration. The Q_d sink term depends on the head at the drain pipe and is calculated using Equation 4.10 (Harbaugh et al. 2000)

$$Q_d = C_d(h - Z_d) \tag{4.10}$$

Where Q_d is the drain outflow [L^3/T], C_d is the conductance [L^2/T], h is the hydraulic head [L] at the drain pipe and Z_d is the drain elevation [L].

Solving Equations 4.7 and 4.8 requires knowledge of the constitutive relationship between moisture content and capillary head, which is represented by the van Genuchten (1980) model,

$$\boldsymbol{\theta(\psi) = \theta_r + \frac{\theta_s - \theta_r}{(1+(\alpha|\psi|)^{\beta})^{1-\frac{1}{\beta}}}} \tag{4.11}$$

Where θ_r is the residual moisture content [L^3/L^3], α is a fitting parameter related to the inverse of the air entry suction, $\alpha > 0$ [L^{-1}], β is a measure of the pore size distribution, $\beta > 1$.

4.4 The Stochastic Analysis

The uncertainty in the predictions made by the set of deterministic Equations 4.1 to 4.11 in the preceding section is assumed to stem mainly from uncertainty of the parameters. Other sources of uncertainty, such as geostatitical parameters uncertainty and conceptual uncertainty (simplification to mathematical models), are not considered in this study. The input parameters are grouped into four categories. The first is the three-dimensional soil properties, which include the hydraulic conductivity K; the porosity; the Van Genuchten Model's parameters θ_r, α, β; soil specific storativity S and dispersivity α_z. The second group is the two-dimensional parameters such as irrigation weight w_i (irrigation uniformity) and preferential flow fraction p_i (i.e. the fraction of irrigation water that reaches the water table instantaneously). The third group comprises the irrigation system parameters such as the amounts of diverted water, its salinity, and the fraction of infiltrating water. The fourth group includes the semi-empirical parameters that control yield, water uptake and drainage flow. This group includes

parameters such as the yield response factor K_y, uptake model parameters ψ_{50}, ψ_{o50}, ψ_s, P; root growth rate parameters a, b; crop growth parameter K_c, and drain conductance coefficient C_d.

The following section illustrates the methodology of the Multivariate Monte Carlo Analysis of the correlated soil properties within the first group.

4.4.1 Multivariate Simulation of the Soil Properties

Generally speaking, consider N regionalized and correlated soil properties. Also, assume that each soil property has a number of field measurements $D_1, D_2, \ldots D_N$ respectively. The correlated *N* soil properties are presumed to be normally distributed, or that they could be transformed to be normal. Given this setting, the objective of the geostatistical simulation is to generate *N dependent* realizations for each soil property.

The Sequential Indicator Simulations (SIS) method (Deutsch et al. 1997) is employed herein due to its flexibility in incorporating hard and soft information about simulated parameters. In particular, the SIS method aims to calculate a least-squares estimate of the conditional cumulative distribution function $\mathrm{F}(\mathrm{Z_{i,k}})$ at pre specified cutoffs $\mathrm{Z_{i,k}}$, where $\boldsymbol{i}$ is the soil property index and $\boldsymbol{k}$ is the cutoff index. The cutoffs are a set of $\left(Z_i, F(Z_i)\right)$ pairs that can be used to approximate the CDF of Z_i. Usually 4 to 10 cutoff values are sufficient to obtain a good approximation of the CDF (Deutsch 2002).

In order to facilitate the multivariate SIS, we used the same number of cutoffs for each of the soil variables ($\mathrm{Z_1}, \mathrm{Z_2}, \ldots \mathrm{Z_N}$). Moreover, the cutoff values are chosen in a

consistent fashion for each variable; that is, the k[th] cutoff is calculated for any variable Z_i using Equation 4.12.

$$Z_{i,k} = Z_{o,k}\sigma_{Z_i} + \mu_{Z_i} \tag{4.12}$$

Where $Z_{o,k}$ is an arbitrary cutoff value in the standard normal distribution ($\mu_o = 0, \sigma_0 = 1$), σ_{Z_i} and μ_{Z_i} are the standard deviation and the mean of the variable Z_i. That is to say, Equation 4.12 calculates cutoff values for each of the soil variables at the same standardized CDF cutoff $Z_{o,k}$.

Without loss of generality, the sequential simulation of the variables is initiated at parameter Z_1. The CDF value $F(Z_{1,k}(x))$ at the k[th] cutoff and at x spatial position must be either zero or one at locations where Z_1 field measurement are available (Equation 4.13).

$$F(Z_{1,k}(x)) = \begin{cases} 1 & Z_1(x) < Z_{1,K} \\ 0 & \text{otherwise} \end{cases} \tag{4.13}$$

However, for any other location x' where no field measurements for Z_1 are available, $F(Z_{1,k}(x'))$ can be estimated using Equation 4.14 according to Deutsch et al. (1997), which is the best linear unbiased estimate of the $F(Z_{1,k}(x'))$ or *Indicator Kriging* (IK).

$$F\left(Z_{1,k}(x')\right) = \sum_{x=1}^{n} \lambda_x . F\left(Z_{1,k}(x)\right) + [1 - \sum_{x=1}^{n} \lambda_x] F(Z_{1,k}) \tag{4.14}$$

where $F(Z_{1,k})$ is the global CDF based on the data D_1, n is the number of field measurements, and λ_x is the simple kriging weight that can be calculated from a set of linear equations as in Equation 4.15.

$$\sum_{y=1}^{n} \lambda_y \, C_I^{Z_{1,k}}(y-x) = C_I^{Z_{1,k}}(x'-x), \quad x = 1,..n \tag{4.15}$$

The term $C_I^{Z_{1,k}}(x'-x)$ is the indicator covariance of variable Z_1 at the k cutoff.

It is possible to incorporate the field measurement of the other variables, namely, $Z_2,..Z_N$, as soft data to estimate $F\left(Z_{1,k}(x')\right)$ from Equation 4.14. To illustrate, hence Z_1 is presumed normally distributed and correlated with $Z_2,..Z_N$, therefore the conditional CDF of the variable Z_1, given the data $Z_2,..Z_N$, is normally distributed and can be exhaustively described by its conditional mean and conditional variance as expressed in Equations 4.16 and 4.17.

$$\boldsymbol{m(Z_1(x)|D_2,...D_N) = m(Z_1(x)|D_2,..D_{N-1}) + \rho_{1,N-1}\left(\frac{\sigma(Z_1(x)|D_2,..D_N)}{\sigma(Z_{N-1})}\right)(Z_{N-1} - m(Z_{N-1}))} \tag{4.16}$$

$$\boldsymbol{\sigma(Z_1(x)|D_2,...D_N) = \sigma(Z_1(x)|D_2,...D_{N-1})\left(1 - \rho_{1,N-1}\right)^{0.5}} \tag{4.17}$$

Accordingly, knowing the conditional mean and standard deviation is all that is required to determine the $F\left(Z_{1,k}(x)|D_2,...D_N\right)$, which can be used in Equation 4.14. In the same fashion, the CDF values at all other cutoffs are estimated resulting in an approximate estimate of the $F\left(Z_1(x')\right)$. Thereupon, it is straightforward to sample a

random value for the variable Z_1 at location x' from the approximated CDF. Next, the simulation is repeated at all other cell nodes to produce a three dimensional realization of the variable Z_1. Recall that every new simulated Z_1 should be conditioned on previously simulated values by treating simulated values as hard data. The resulting realization can be expressed mathematically using Equation 4.18 which says that l_{Z_1} is a realization of the variable Z_1 conditioned on the data $D_1, .. D_N$.

$$l_{Z_1}: \{(Z_1(x) | D_1, .. D_N) , x \in A\} \tag{4.18}$$

Given the simulated soil property realization l_{Z_1}, it is required to continue the simulation to the next soil property Z_2. The same steps can be followed, however; this time, the simulation of Z_2 is conditioned on the field measurement $(D_1, .. D_N)$ as well as the previously simulated Z_1. To generalize the procedures, any variable Z_i can be simulated by conditioning the simulation on all previously simulated variables $Z_1, \ldots Z_{i-1}$ and field measurements $D_i, .. D_N$ as shown in Expression 4.19.

$$l_{Z_i}: \{(Z_i(x) | (Z_1, .. Z_{i-1}, D_i, .. D_N) , x \in A\} \tag{4.19}$$

The order of the simulation of variables, that is adopted here, is to start simulating the parameter that has the largest number of field measurements; in this field, it is the hydraulic conductivity, followed by the pore scale parameter, which is the parameter that is highly correlated with the conductivity and so on.

4.4.2 Covariance Inference

The solution of Equation 4.15 requires the knowledge of the covariance model at the k^{th} cutoff and requires that the transformed variables using Equation 4.13 are stationary at the k^{th} cutoff. The assumption that the variables are statistically homogenous is usually difficult to verify; however, it may still be a convenient working assumption .Under the stationary assumption, the variogram and covariance models are equivalent, i.e. $\gamma(h) = C(0) - C(h)$. Denoting the number of parameters as $\boldsymbol{N}$, the total number of covariance models required is N^2. For example, in our case, if five soil properties are intended to be simulated, then the number of variograms required is 25 variograms. Obviously, the inference of variograms for all the variables requires a large number of field measurements that are usually not available.

However, it is possible, assuming the validity of the intrinsic coregionalization model (Wackernagel 2003), to make use of the abundant data for a certain variable to infer the variograms for other variables. For example, the abundant data for hydraulic conductivity can be used to infer the variogram for Van Genuchten parameters. Intrinsic coregionalization models can be valid if the multivariate correlation structure of a set of variables has the same spatial correlation scale. This model is based on a Markov screening hypothesis, in which a co-located primary data screens the influence of distant measurements on the secondary variable (Journel 1999). As a result, the covariance model for any variable Z_i can be obtained from Equation 4.20.

$$C_{Z_i}(h) = \sigma_{Z_i}\rho(h) \tag{4.20}$$

where $\rho(h)$ is the correlogram that is inferred from another variable Z_j that has abundant field measurements.

$$\rho(h) = \frac{C_{Z_j}(h)}{{\sigma_{Z_j}}^2} \tag{4.21}$$

Therefore, Equation 4.21 reveals that all variables have the same spatial correlation scale, but the sill value is scaled by the variance of the variable.

4.5 Site Description

The site of this study (Figure 4.1), known as Field 17, is located in the neighborhood of the town of Rocky Ford, Colorado. The groundwater flow direction is from the southwest to the northeast, and the water table depths range between 0.6 m in the southwest to 2.4 m in the northeast. The groundwater salinity in field 17 (Figure 4.8) was found in 2009 to range between 1.33 – 2.49 dS/m. The average groundwater salinity is 2 ds/m with a standard deviation of 0.44 dS/m. According to FAO report 48 (Rhoades et al. 1992), this level of salinity is classified as slight to moderate salinity. The soil type in the region is alluvial deposits that consist of a silty loam clay layer in the upper surface and loam to sandy loam substrata (USDA, 1971a; USDA, 1971b). The extracted soil salinity EC_e of the root zone ranges between 2.8-4.3 dS/m (right of Figure 4.7). The water table and groundwater salinity were monitored using 31 observation wells. Part of the field was chosen for the numerical simulation. This part has a shallow water table and relatively high soil salinity. Additionally, a subsurface set of drainage pipes is intended to be installed to help alleviate the waterlogging and salinity problems.

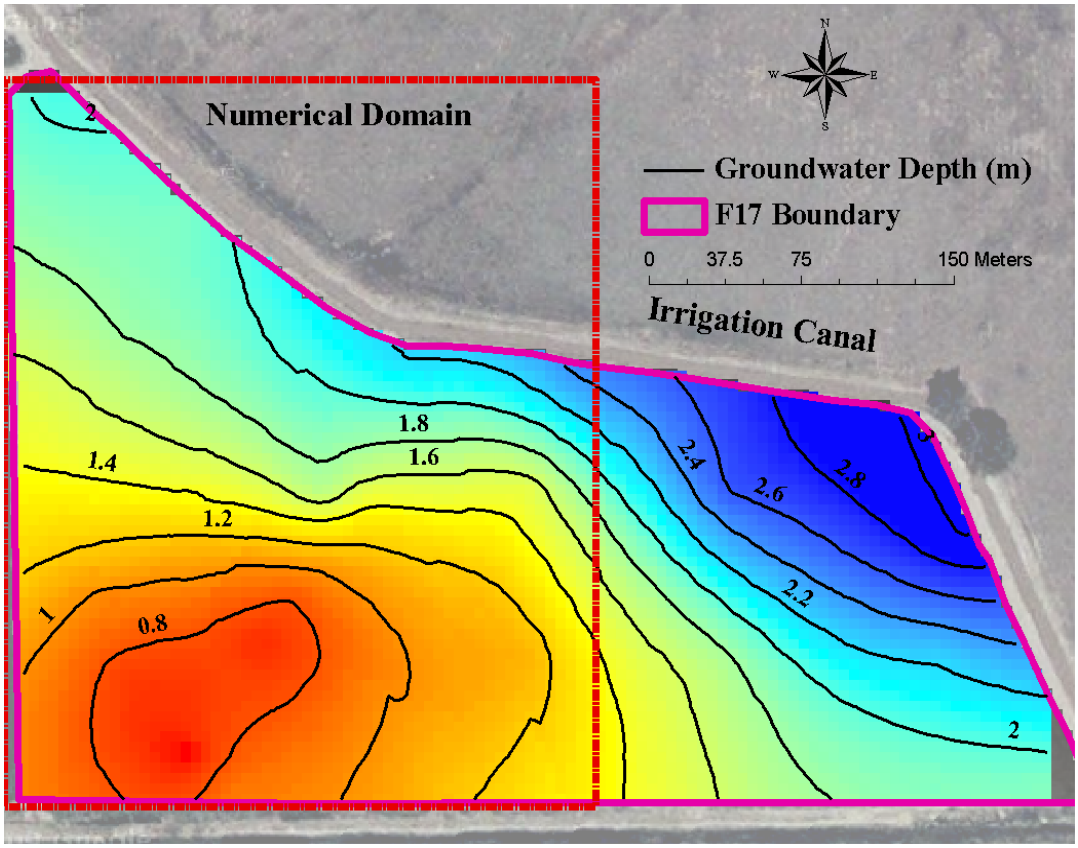

Figure 4. 1: Field 17 Site Map, Groundwater Depth and the Numerical Domain

4.6 Statistical Distribution of Parameters

The statistical properties of the input variables were inferred using field data, soil databases for other sites, and published literature. The following section discusses the statistical properties for each of the soil properties.

4.6.1 Three-dimensional Soil Parameters

All the three-dimensional soil parameters are randomized except for the specific storativity, which has a small impact especially in shallow unconfined aquifers such as the field under study. The multivariate Sequential Indicator Simulation, outlined in

section 4.4, is employed to generate equally probable three-dimensional realizations for (K, θ_s, θ_r, α, β), whereas dispersivity is estimated based on a regression model as will be shown later in section 4.6.2. Data from a Cone Penetration Test (CPT) are used to estimate the hydraulic conductivity vertical profiles at 15 positions in Field 17 with vertical depths ranging between 6m to 20m. The resulting measurements are averaged at 10 cm vertical intervals resulting in 665 hydraulic conductivity estimates. A number of core samples (37 cores) were used to obtain the Van Genuchten parameters and porosity. Only eight of these core samples have hydraulic conductivity values. Obviously, these eight samples are not enough to obtain the correlation between the hydraulic conductivity and the Van Genuchten parameters. Therefore, a soil database (ROSSETA database) for the five parameters (K, θ_s, θ_r, α, β) (Schaap et al. 2001) which include 650 records are used for two purposes, first, to estimate a better correlation coefficients among the five parameters and, second to select a suitable normal transformation scheme as required by the multivariate Monte Carlo simulation in section 4.4.

The soil parameters were transformed using three transformations and the one that produced the lowest Chi-squared Goodness of Fit Test statistic was selected (Table 4.1). The Johnson transformations family (Johnson et al. 1995) was used for this purpose. The transformations are the Lognormal (LN), the Log Ratio (SB) (Equation 4.22) and Hyperbolic Arcsine (SU) (Equation 4.23).

$$Y = ln\left(\frac{x-A}{B-x}\right) \tag{4.22}$$

$$Y = -sinh^{-1}(x) = ln(x + (1 + x^2)^{0.5}) \tag{4.23}$$

where A is the minimum value of x and B is the maximum value.

The statistical properties of the transformed variables are illustrated in Table 4.2. The transformed field measurements are simulated using the multivariate SIS using five cutoffs (Table 4.3 and Figure 4.2). The transformed hydraulic conductivity field measurements are used to calculate the experimental indicator horizontal and vertical variograms at each of the five cutoffs. Thus, ten experimental variograms were obtained, five for the horizontal variograms and another five for the vertical variograms; subsequently, the spherical variogram functions were fitted. A summary of the variogram fittings is shown in Table 4.4. Recall that according to the intrinsic coregionalization model adopted in section 4.4.2 the correlation scales of the hydraulic conductivity indicator variograms are the same for θ_s, θ_r, α and β. The Geostatistical Library (GSLIB) (Deutsch et al. 1997) Sequential Indicator Simulation method is used successively to simulate the parameters in the following order: $K, \beta, \theta_s, \theta_r,$ and α.

Table 4. 1 : Choosing the Best Normal Transformation Scheme

Parameter	Mean	Variance	Chistat
θr	0.11	0.002	97.36
θs	0.41	0.01	175.27
α	0.05	0.006	65.95
β	2.08	1.12	320.02
log(θr)	-1.02	0.07	11.33
log(θs)	-0.39	0.01	95.82
log(α)	-1.34	0.02	33.12
log(β)	0.27	0.04	482.12
SB(θr)	-0.47	0.29	11.69
SB(θs)	-0.35	0.2	74.35
SB(α)	-0.36	0.15	107.75
SB(β)	-0.79	0.67	55.27
SU(θr)	-1.65	0.35	11.33
SU(θs)	-0.22	0.03	95.82
SU(α)	-2.4	0.11	33.12
SU(β)	1.31	0.2	482.12

Table 4. 2 : Statistical Properties of Transformed Data

				Correlation Coefficients				
	Mean	**STD**	**Transfor-mation**	**Log(K)**	***θs***	***θr***	***α***	***β***
Log(K)	-0.574	1.788	Log10	1.00	-0.19	-0.42	-0.26	0.77
θs	0.451	0.047	SB	-0.19	1.00	0.59	0.60	-0.36
θr	0.160	0.050	Log10	-0.42	0.59	1.00	0.55	-0.29
α	0.046	0.030	Log10	-0.26	0.60	0.55	1.00	-0.47
β	1.529	0.329	SB	0.77	-0.36	-0.29	-0.47	1.00

Table 4. 3 : Cutoff Values of the Transformed Parameters

Soil Property Cutoff	θr	Θs	α	β	K
C1	-1.00	-2.09	-1.64	-2.39	-3.01
C2	-0.93	-1.35	-1.51	-1.39	-2.00
C3	-0.80	0.11	-1.26	0.58	0.00
C4	-0.73	0.84	-1.13	1.57	0.99
C5	-0.67	1.57	-1.00	2.55	1.99

Table 4. 4 : Horizontal and Vertical Indicator Variogram Parameters

Normalized Cutoff Value (Z)	-1.37	0.8	0.32	0.88	1.44
Vertical Correlation Length (m)	1.1	1.3	1.4	1.5	1.5
Horizontal Correlation Length (m)	126	117	153	72	75
Sill Value	0.2	0.23	0.3	0.11	0.03

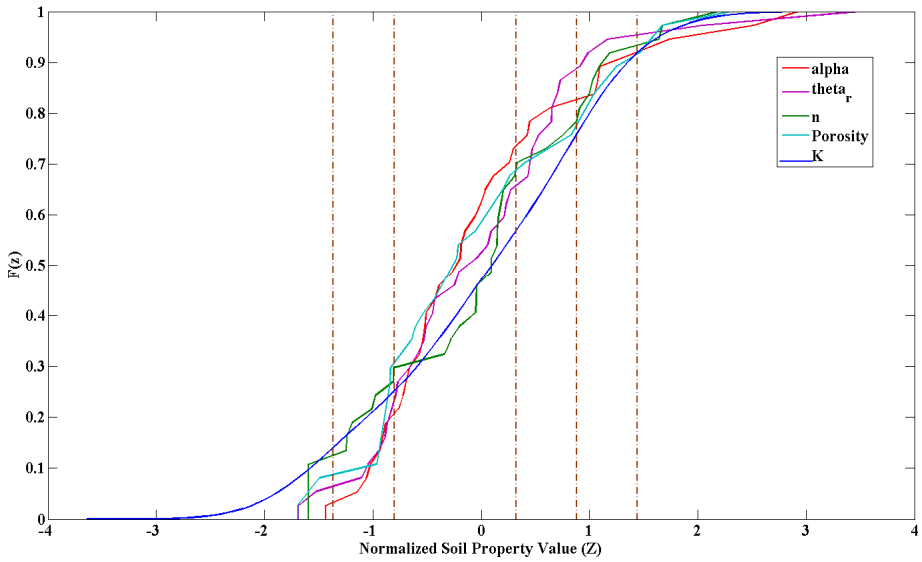

Figure 4. 2: Normalized CDF of the Soil Properties and the Indictor Cutoffs in the Distribution

4.6.2 Hydrodynamic Dispersivity (α_x)

It is expensive and time consuming to obtain site-specific dispersivity values (for example, tracer tests); on the other hand, laboratory column tests usually reflect scales that are much smaller than site scales. Other methods are correlation methods (Xu et al. 1997) that use correlations between dispersivity and other easier to obtain soil properties. In this research, a simple correlation method from Xu et al. (1997) was chosen. It was found that the correlation between α_x and the reciprocal of porosity is 0.84 and the following model can describe this relationship with an r^2 of 0.74,

$$\alpha = -25.47 + 12.40\left(\frac{1}{\theta_s}\right) \tag{4.24}$$

Unfortunately, the limitation of this regression model has the inability to deal with porosity values greater than (0.486) due to the resulting negative dispersivity. To circumvent this problem, the previous model was refitted to a polynomial function that has its root at $\theta_s = 0.668$, which is a very rare event. Both models have almost the same performance for $\theta_s < 0.486$.

$$\alpha_z = -1268.9\theta_s^3 + 1950.6\theta_s^2 - 990.64\theta_s + 169.72 \tag{4.25}$$

Where α_z is the dispersivity value in (mm). Equation 4.25 enables us to calculate the vertical dispersivity at each cell using the porosity field.

4.6.3 Irrigation spatial depth and Uniformity

The uniformity of irrigation depth is significantly impacted by the irrigation system used. For instance, the sprinkler system usually has high uniformity coefficients; however, its Christiansen's Coefficient (CU) is highly sensitive to the wind speed and direction. Other factors that affect the CU for sprinklers are the layout and spraying hydraulics. In surface irrigation systems, the topography, bed geometry, vegetation density and soil properties affect the uniformity of the irrigation. In this study, a sprinkler system is assumed to be used. The CU values of less than 84% are considered low according to Bliesner et al. (2001). Assuming a unity irrigation depth (called *'irrigation weight'*), it is reasonable to statistically model this property using a normal distribution $N\sim(\mu = 1, \sigma_w^2)$. According to Montazar (2010), the normal distribution is a good model for the case of sprinkler systems, but might not be proper for surface or drip irrigation systems due to the role of land topography affecting surface irrigation and due to the design and the hydraulics of drip irrigation systems.

The actual irrigation depth can be calculated by multiplying the average irrigation depth by the irrigation weight. The question is how to select a value of σ_w^2 based on the uniformity coefficient. First, it is needed to randomly select a Christiansen's Coefficient CU from the noninformative uniform distribution $U\sim(a = 95\%, b = 85\%)$, which is within the typical sprinklers coefficients. Next using the sampled CU, the irrigation depths variance can be computed from:

$$\sigma_w^2 = \frac{\pi}{2}\left(1 - \frac{CU}{100}\right)^2 \tag{4.26}$$

The previous equation is derived (Appendix B) by substituting the mean absolute deviation, in the CU equation (Hoffman et al. 2007), by the irrigation depth variance.

4.6.4 Irrigation Efficiency and Salinity

The total diverted water depth (D_T) to Field 17 was found to range between 56 cm-68 cm per meter square per summer season; noninformative uniform distribution is used to model its uncertainty. The fraction of infiltration water is highly impacted by the irrigation scheme used. Assuming a sprinkler irrigation system, the PDF of the infiltration fraction (I_f) is taken as $N(\mu = 85\%, \sigma = 7\%)$. This distribution was obtained by analyzing data for a number of sprinkler-irrigated fields in the Arkansas Basin (Figure 4.3). The sampling sequence starts by randomly sampling a value of the total diverted water from $(a = 56cm, 68cm)$; then the infiltration fraction is sampled from $N(\mu = 85\%, \sigma = 7\%)$. Consequently, the total amount of water infiltrated (I_i) can be directly calculated, i.e.$I_i = \frac{D_T . I_f}{100}$, and then divided on seven irrigation events.

The salt concentration, as Total Dissolved Solids (TDS), was measured for the irrigation canal near Field 17 in the summer of 2005 and found to be approximately normal $N(\mu = 0.76, \sigma = 0.18)$ (mmoh/cm) (Figure 4.4).

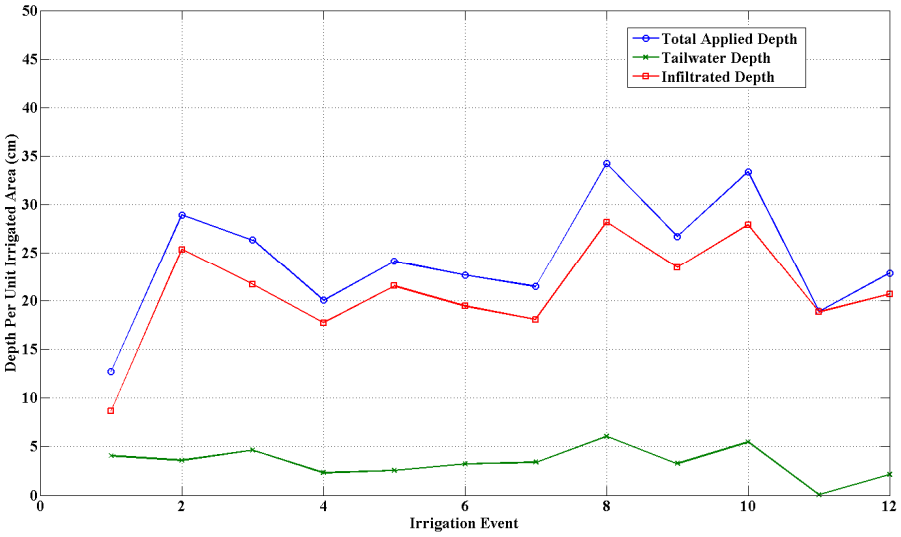

Figure 4. 3: Total Applied Water Depths, Tailwater Depths and Infiltration Depth for a Sprinkler Irrigation System

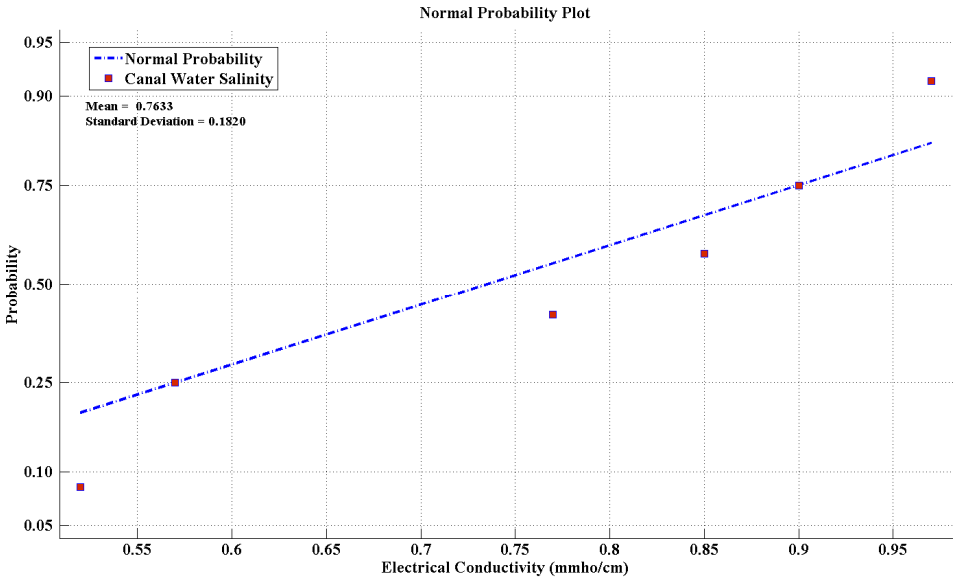

Figure 4. 4: Statistical Properties of Irrigation Canal Water Salinity

4.6.5 Preferential flow

The flow and transport in porous media can be either rapid macro pores flow and transport (direct drainage) or matrix slow flow (general drainage) (Steenhuis et al. 1994). The spatial distribution of shrinking cracks or bio-holes is complex and unpredictable. A simplified method is adopted to quantify the fraction of irrigation or rainfall water that rapidly reaches the water table. To the extent of the authors' knowledge, there are no published data that quantify the fraction of surface water that rapidly reaches the groundwater table. Alternatively, it is assumed that the bypass flow fraction is a spatially random, uniformly distributed, and spatially independent regionalized variable. The uniform distribution of the bypass fraction adopted is $\mathrm{U}\sim(\mathrm{a}=0,\mathrm{b}=10\%)$. This distribution reflects a high degree of uncertainty regarding bypass flow.

4.6.6 Root Uptake Model Parameters

Equation 4.6 in Section 4.3 describes the root uptake model where four semi empirical parameters are required to predict the root extraction ($\psi_s, \psi_{50}, \psi_{o50}, p$). The parameter ψ_s represents the capillary head value at which the yield will decrease due to oxygen deficiency. Veenhof and McBride (1994) suggested values for ψ_s to be between -1cm to -30cm. In line with these findings, a noninformative uniform distribution is used $\mathrm{U}\sim(\mathrm{a}=-1,\mathrm{b}=-30)$cm. Cardon and Letey (1992) used a value for the salinity tolerance parameter of $\psi_{o50} = -4{,}300$ cm and then they estimated

water deficit parameter ψ_{50} to be within the range of $-2{,}500\ to\ -6{,}500$ cm. Shalhevet et al. (1986) used a $\psi_{o50} = -6{,}400cm$ for alfalfa.

In this paper, dry biomass data for alfalfa that were collected from Field 17 and used to calibrate the parameters using Equation 4.6. Using $\psi_{o50} = -6{,}400cm$ as in Vinten and Meiri (1986) the values of ψ_{50} are estimated at different capillary heads. The value of ψ_{50} was plotted versus the summed square errors of alfalfa yield estimation at different capillary heads (Figure 4.5). From Figure 4.5 it can be seen that a reasonable distribution of ψ_{50} is U~(a $= -800$, $b\ -3{,}000$) cm. The value for the corn maize (Zea Mays) is taken as $\psi_{o50} = -4{,}100cm$. The parameter $\boldsymbol{p}$ is taken as a deterministic value equal to 3 (Ayers and Westcot 1994).

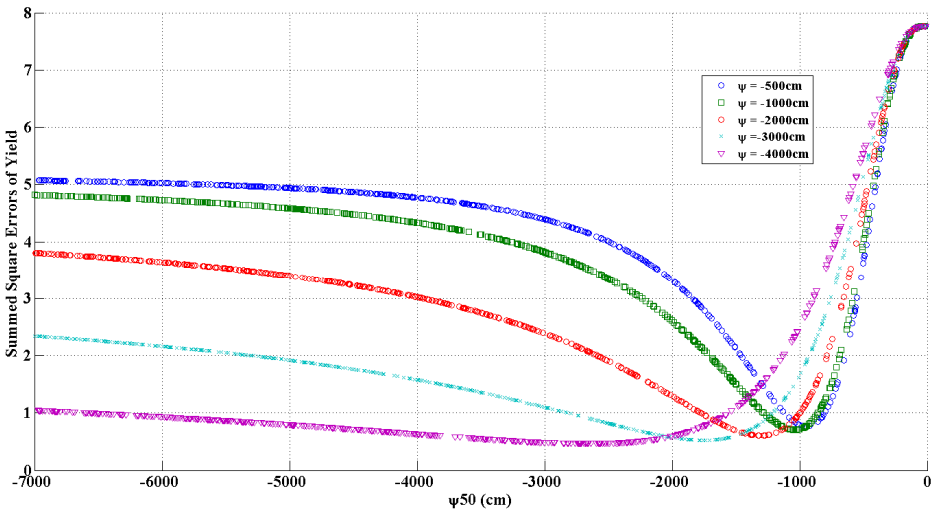

Figure 4. 5: Estimating ψ_{50} Values Using Dry Alfalfa Biomass Data

4.6.7 Drain Conductance

The drain outflow in Equation 4.10 is a head dependent flow. The conductance [L^2/T] quantifies the resistance that water flow experiences to enter the drain. Specifically, the value of conductance reflects entrance resistance and gravel envelop resistance. Usually the values of conductance are determined during model calibration procedures. Deverel et al. (1991) used field data of head at the drain and outflows to calculate conductance values based on Equation 4.10. The values obtained were within the range of 0.27 m^2/day/m to 0.44m^2/day/m for different type of soil, gravel and drain pipe materials.. Conductance values estimated via calibration by Goswami and Kalita (2009) are within the range of 0.15 and 0.58 m^2/day/m. In this study and in accordance with published conductance values, a wide uniform distribution is used $U\sim(a = 0.01, b = 2)$ m^2/day/m.

4.7 Numerical Simulation

The Equations 4.1 to 4.11 are solved using the CSUID model (Alzraiee et al. 2009). This model is a three-dimensional variably saturated flow and transport model in a heterogonous porous media. The resulting nonlinear finite difference equations of flow and transport are solved using the precondition conjugate gradient method (Harbaugh et al. 2000). The horizontal cell sizes used is 10m x 10m and the vertical cell size is 0.25m. The number of cells in the horizontal plane is 30 for the east-west direction and 38 cells in the north-south direction (Figure 4.6). Twenty layers, each 25cm, were used. The general boundary conditions are used to describe the boundary of

the field. The upstream and downstream boundaries are each divided into 3 sections to represent the head variability along the boundary. The salinity of the lateral flux induced by the general boundary condition was chosen to be consistent with groundwater salinity measurement at these boundaries.

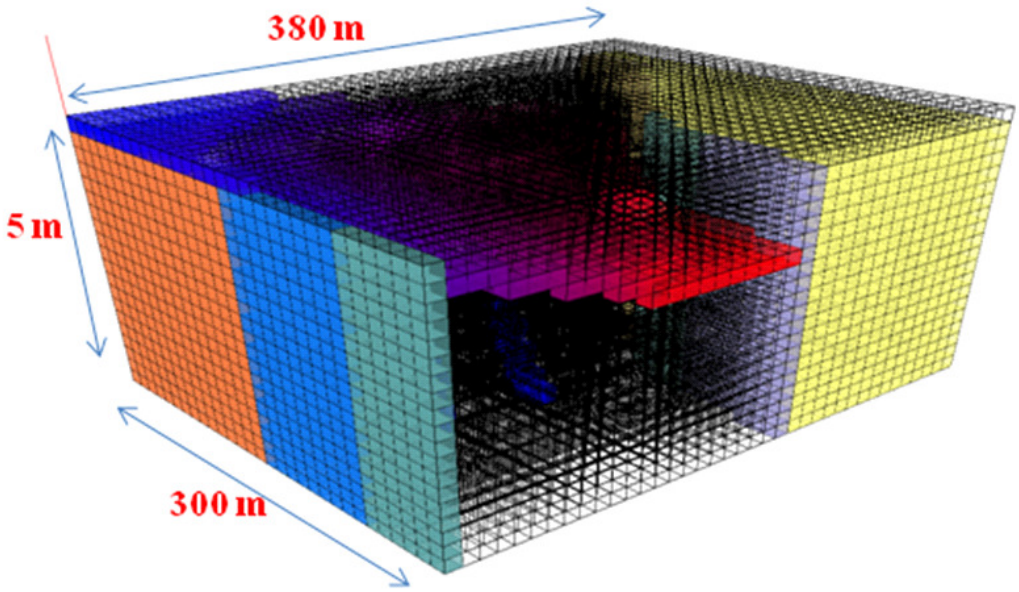

Figure 4. 6: Dimensions and Spatial Discretization of the Numerical Domain

The simulation periods are chosen to be 60 days for alfalfa first cycle and 125 days for corn . A root zone depth of 1.5 m is used for alfalfa crop and 1.3 m for the corn. Table 4.5 shows alfalfa and corn growth properties. The reference evapotranspiration was obtained from the weather data for Rocky Ford between May 1, 2010 and July 15, 2010 (Figure 4.7).

Table 4. 5 : Crops growth properties (Hoffman 2007)

Crop Type	Growth Period (days)	Kc ini	Kc mid	Kc end	Depth of Root Zone (m)	%50 Yield Reduction Osmosis Pressure (cm)
Alfalfa	**60**	**0.4**	**0.95**	**0.9**	**1.5**	**-6500**
Corn	**125**	**0.3**	**1.15**	**0.4**	**1.3**	**-4100**

The initial water table is treated as a deterministic surface and kriged using 31 observation wells. The initial salt concentration is obtained by kriging the salinity measurements, and it is assumed that the vertical salinity profile is uniform due to the lack of information about the salinity stratification.

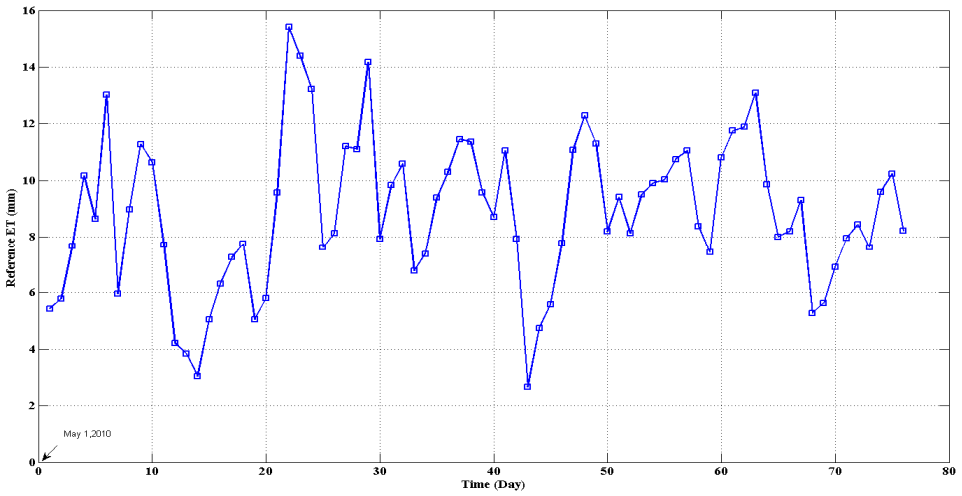

Figure 4. 7: Dialy Reference Evapotranspiration in Rocky Ford, CO (May 1, 2010 – June 15 ,2010)

4.8 Results and Discussion

The hydrological responses of the aquifer systems, which include water table level, root zone salinity, relative crop yield, and deep percolation, are obviously spatial and temporal variables. Due to the large amount of numerical outputs, i.e. at each time step and at each numerical node, the discussion of the results, herein, is limited only to the output at the end of the modeled crop season. In addition, time series for water content salinity, root extraction and vertical flux are presented. At first, the baseline condition (before drain installation) of the field is outlined; therefore, the impact of installing a subsurface drainage system to help alleviate the problems of waterlogging, root zone salinity, deep percolation and relative crop yield is investigated.

4.8.1 Pre-Drainage Water Table Conditions

Waterlogging spatial delineation requires the spatial interpolation of groundwater table depths, computed as the difference between the ground surface elevations and the groundwater table elevations. Figure 4.8 shows that groundwater table depth, prior to drain installation, ranges between 0.66m to 2.43m. In this paper, the field is classified as (i) waterlogged area of groundwater table depth < 1.0m, (ii) partially waterlogged area with groundwater table depth between 1m and 2m, and (iii) waterlogging free area with groundwater depth ≥ 2m. The first category of waterlogged areas constitutes 33% of the field, and partially waterlogged areas constitute 58% of the field.

The existing conditions (no drainage system) is simulated in CSUID to predict the relative crop yield for alfalfa crop. Results presented in figure 4.10 show that the

relative crop yield is between 20% and 62%. The lowest yield occurs in region with high salinity and shallow groundwater table, which demonstrate the negative impact of waterlogging and high soil salinity on crop yield.

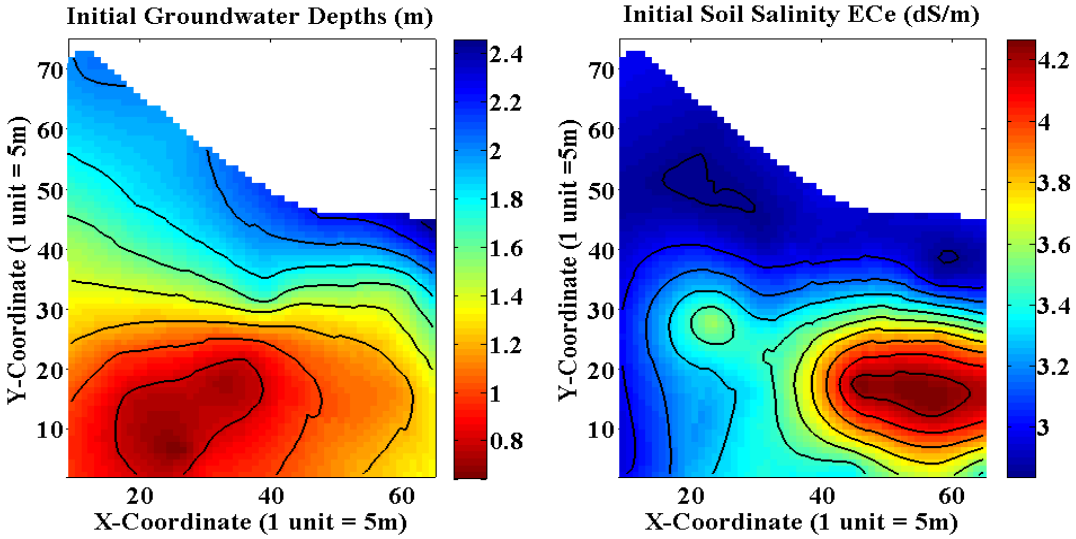

Figure 4. 8: Initial Groundwater Table Depths and Ground Surface Soil Salinity

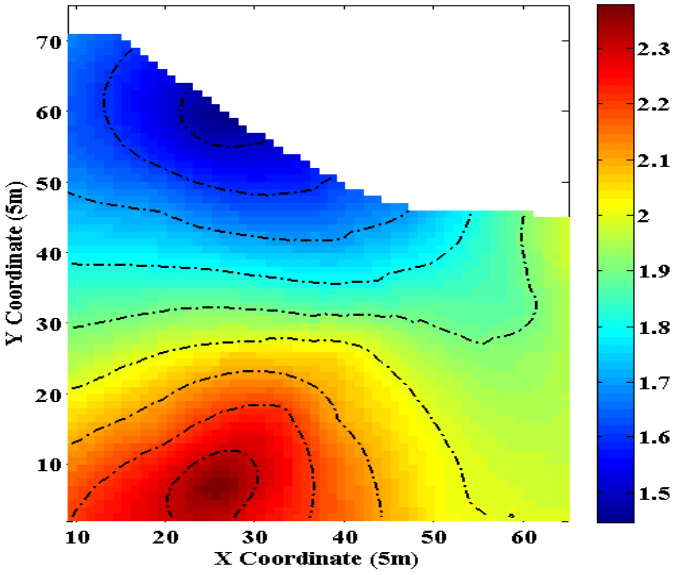

Figure 4. 9: Initial Groundwater Salinity ECw (dS/m)

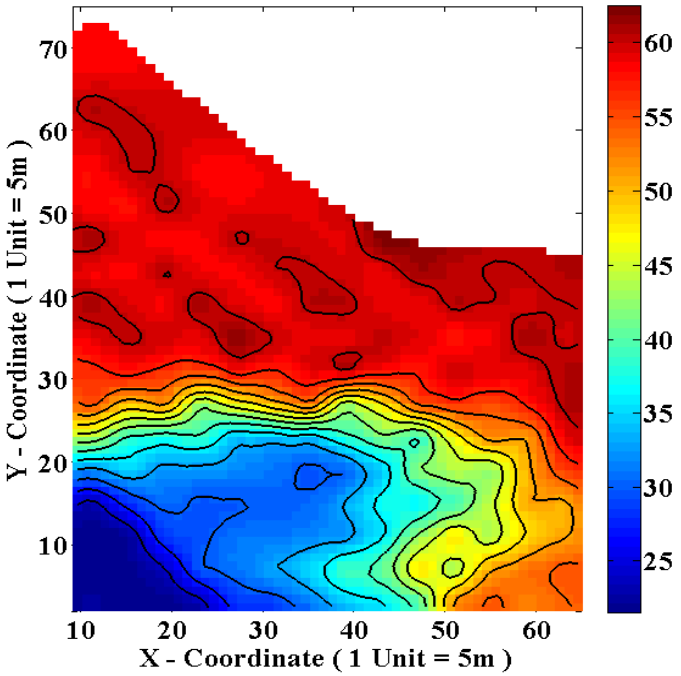

Figure 4. 10: Simulated Relative Crop Yield before Drain Installation

4.8.2 Expected Post-Drainage Groundwater Table Depths

The spatial distribution of the first and the second statistical moments of the water depth at the end the growing season are shown in Figures 4.11. The average water table depth range between 1.05 m and 2.37 m, and the standard deviation of the of water table depths ranges between 0.018 m in the field downstream and 0.37 m in the field upstream(south boundary of the field). The overall average groundwater depth was 1.42 m before installing the drains and is expected to be 1.79 m after installing the drains. This means that the water depth before and after installing the drains is reduced and the waterlogging problem in the southern west part of the field has disappeared; in particular, groundwater table depth is greater than 1 m in the entire field.

Coefficient of variation is in the range of 0.01 to 0.21, which reflects a narrow range of variability. This is due, in part, to the impact of the drains, which fix the groundwater table elevation approximately at the drains elevation. The highest groundwater table depth variability occurs at the upstream end of the middle drain where the general boundary condition and the drain meet. Working as an interceptor of the lateral groundwater flow, the drainage system reduced the water table in downstream (south of the field) region but has a limited impact on the upstream region (north of the field).

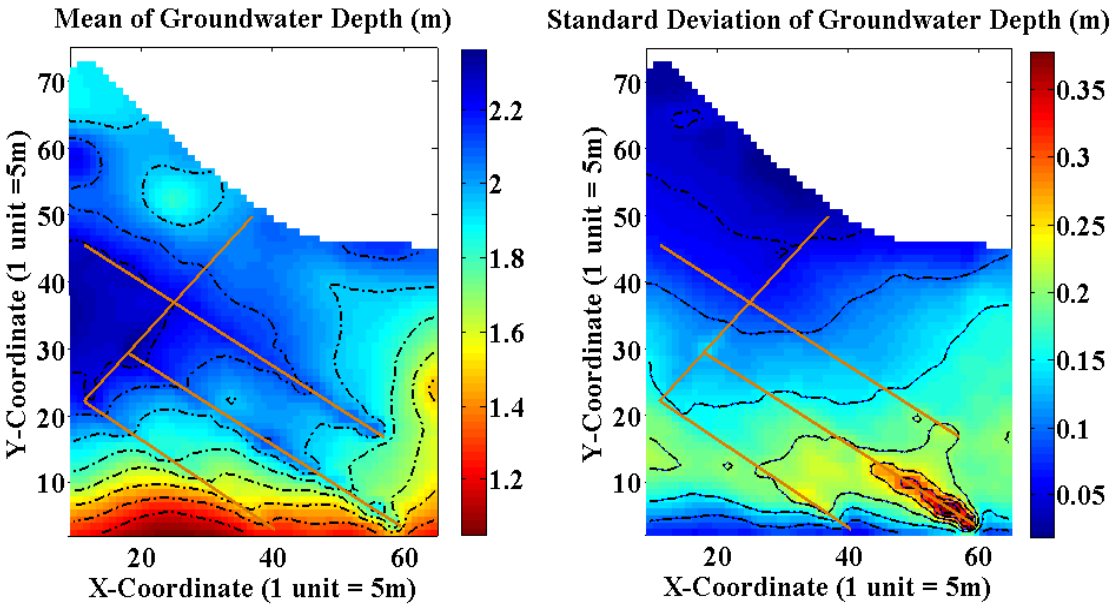

Figure 4. 11: Mean and Standard Deviation of Groundwater Table Depth at the End of the Season

4.8.3 The Expected Root Zone Salinity

The simulated average root zone soil water salinity (ECe) (Figure 4.13) is found to range between 1.05 dS/m and 2.3 dS/m. The initial salinity (ECe) before the

installation of the drain was within the range of 2.8-4.3 dS/m. The highest salinity occurs mainly in the upstream of the drain pipes (south of the field) where the groundwater table is relatively high (shallow). This might be explained by the limited leaching capacity of shallow water table, which also contribute to the upflux that mobilizes salt from the groundwater to the root zone.

The overall simulated average $ECe \approx 1.6$ dS/m whereas the overall initial soil salinity was $ECe \approx 3.5$ dS/m. This demonstrates the drainage systems ability to remove the salts from the soil water profile in most of the field, particularly downstream of the drain.

The soil water salinity (ECe) spatial standard deviation (right of Figure 4.12) ranges between 0.18 and 0.45 dS/m, and the coefficient of variations (CV) ranges between 0.12 and 0.31, which are higher than the variability of the water table depth. The highest salt concentration standard deviation occurs in the upstream section of the field and coincides with the region of the highest variability of the vertical cumulative flow as shown in Figure 4.15. This might reflects the role of upflux in the soil salinization of fields underlain by shallow groundwater table.

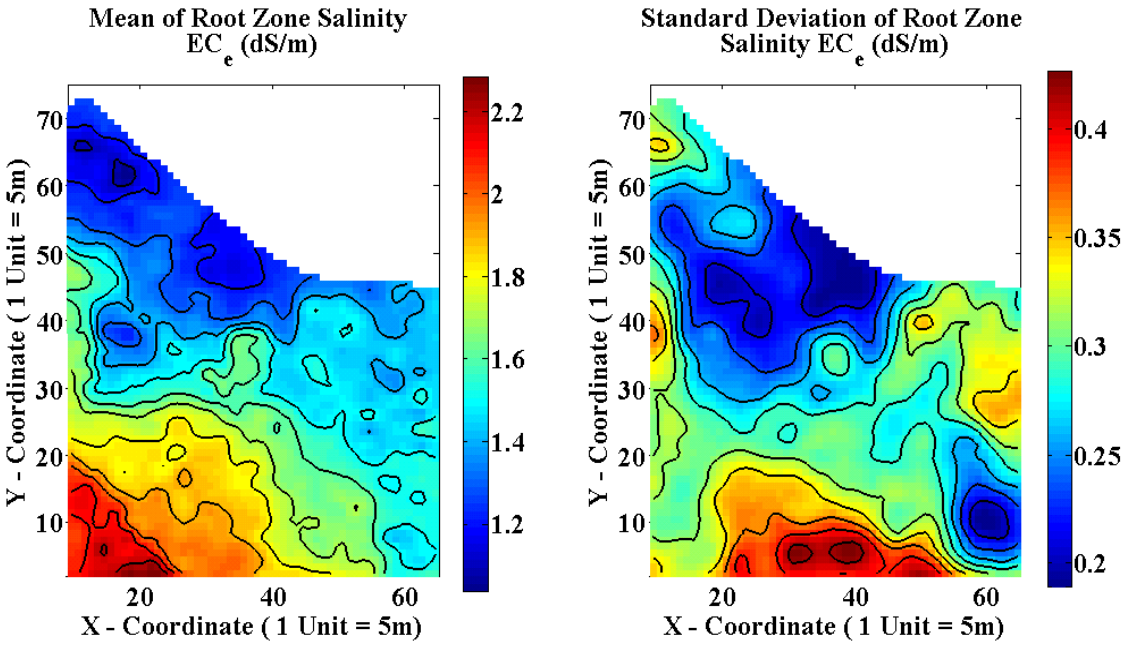

Figure 4. 12: Mean and Standard Deviation of the Root Zone Salinity at the End of the Season

4.8.4 Expected Relative Crop Yield

The simulated relative crop yield is the result of the interaction between plant and hydrosalinity conditions of the root zone as summarized by Equation 4.6. Figures 4.13 and 4.14 show the simulated ensemble mean and standard deviation of the relative crop yield for the alfalfa and corn over the field, respectively. The mean alfalfa relative yield are found to between 91 % and 67% and for the corn between 89% and 64%, which are similar in value and spatial distribution. This observation might seem to contradict the fact that alfalfa and corn have different salinity tolerance (the salinity tolerance parameter for alfalfa is $\psi_{o50} = -6{,}400cm$, and for the corn (Zea Mays) is $\psi_{o50} = -4{,}100cm$). However, this disagreement can be explained by noting that the average soil water salinity in the field is around 3 dS/m, which corresponds to osmotic

pressure of around -1,070 cm, calculated using the equation $\Phi(cm) = 356.89EC$. Obviously, this osmotic pressure is significantly smaller than the 50% reduction threshold for both alfalfa and corn; in other words, the salinity condition in the field after drainage installation is not, *on the average*, the major factor affecting the crop yield.

As the concentration of salts in the root zone deviates from the average, its impact on the crop becomes apparent. This observation can be seen by the different yield standard deviation for alfalfa and corn. The spatial standard deviations (right of Figures 4.13 and Figures 4.14) are found to be between 26% and 11% for alfalfa, and 30% and 11% for corn (Zea Mays), which reflects slight higher corn sensitivity to salinity.

It is worth noting that the maximum relative crops yield for both alfalfa and corn coincide in south of the field which has cumulative deep percolations 0.25m. This indicates the contribution of subirrigation that provides water to crops despite the high salinity concentration in this part which is ECe = 2.2 dS/m (less than the 50% yield reduction of corn and alfalfa).

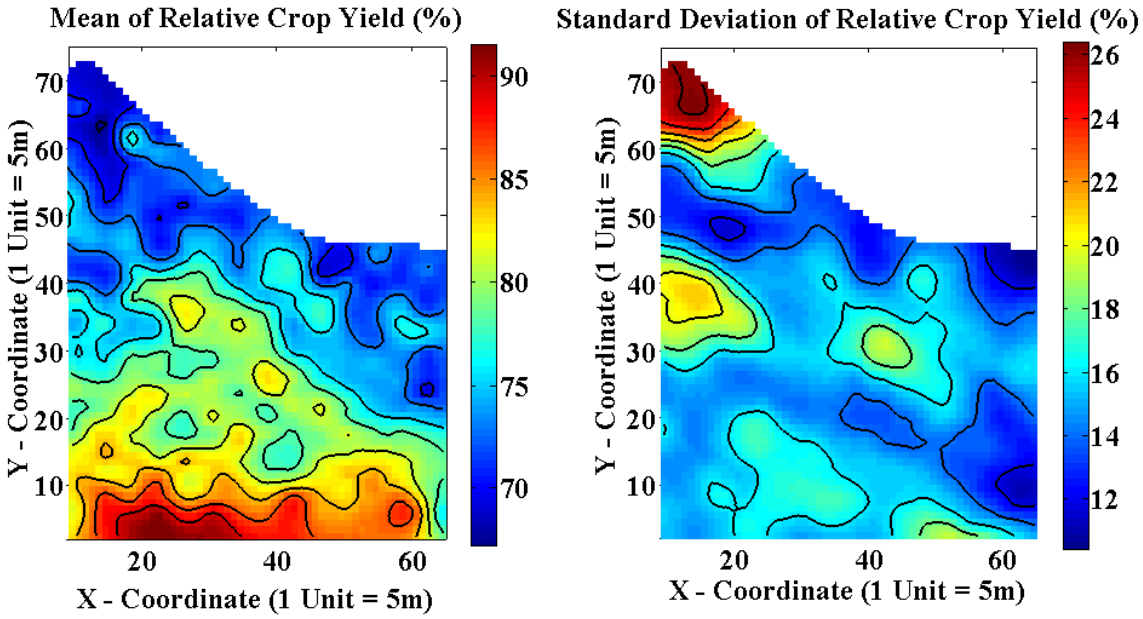

Figure 4. 13: Spatial Expectation and Spatial Standard Deviation of the Relative Crop Yield of Alfalfa

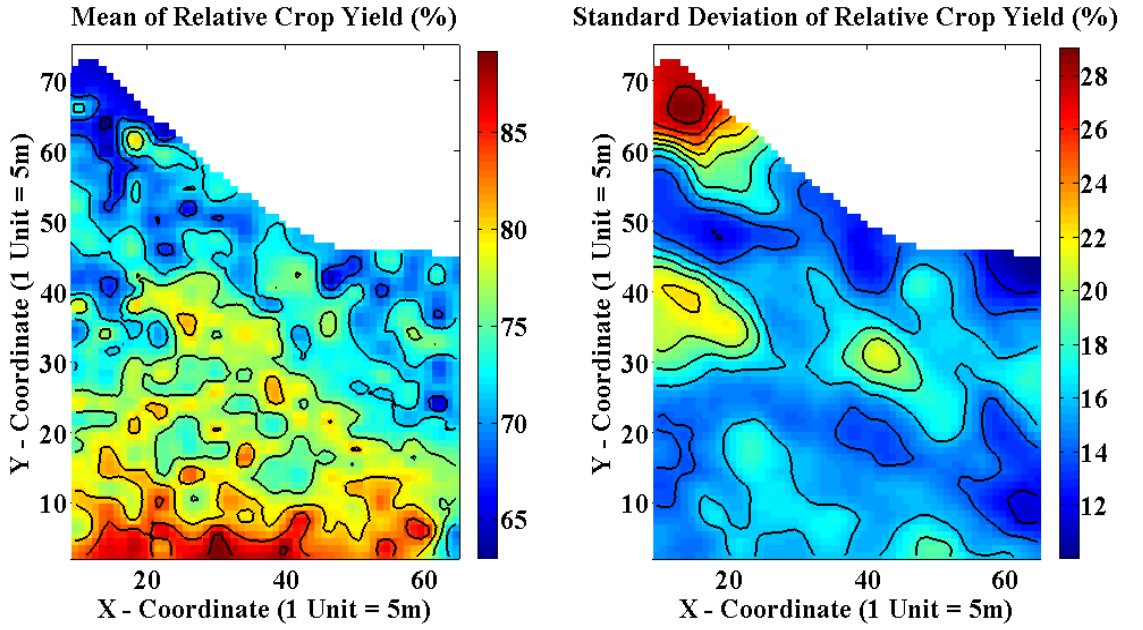

Figure 4. 14: Mean and Standard Deviation of the Relative Crop Yield for Corn

4.8.5 Expected Vertical Flux

Reclamation of fields with high root zone salinity or maintaining root zone salinity within a tolerable range for the crop requires quantifying of the vertical flux. It is important to note that we define the vertical flux as the portion of irrigation water that leaves (or enters) the root zone and that does not necessarily reach the groundwater table. Other environmental implication for the vertical flux calculation is quantifying the loading of pesticides or/and fertilizers to the groundwater.

The spatial expectation and the standard deviation of the vertical flux (defined as the cumulative vertical flux) are shown in Figure 4.15. Net vertical flux mean values klrange between -0.11m and 0.28m with low absolute values in the southern part of the field. The negative sign indicates a downward flow while the positive sign indicates an upward flow. The shallow water table in the southern part seems to be the factor behind the low downward cumulative vertical flux value. The standard deviation values range between 0.065 m and 0.11 m. It is noticed that the standard deviation is very small at region underlain by the drains.

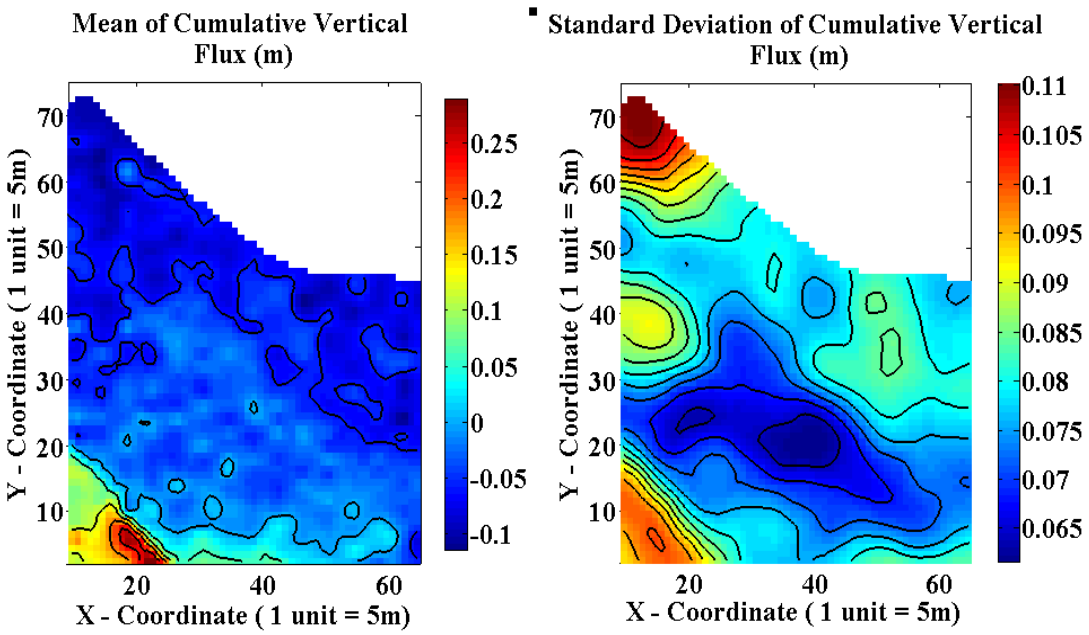

Figure 4. 15: Spatial Expectation and Spatial Standard Deviation of the Simulated Cumulative Vertical Flux

4.8.6 Expected Drain's Flow and Salinity Hydrograph

Investigating the drainage effluent, in terms of quantity and quality, is important to determine the feasibility of drainage water reuse and to assess the environmental risks (for example, disposal of salt loads to streams) associated with the drain installation. The uncertainties in drain conductance, soil hydraulic properties are certainly impact the predicted hydrographs. In order to illustrate the statistical properties of the hydrographs, the flow rate mean $\boldsymbol{\mu}$ and $\boldsymbol{\mu} \mp \boldsymbol{\sigma}$, where $\boldsymbol{\sigma}$ is the standard deviation of flow rate, are plotted in Figures 4.15. The same results are plotted for the salinity of the effluent. The drainage flow rate mean fluctuates around 120m^3/day while the mean of the drainage effluent salinity fluctuates around 2 dS/m.

It is worth noting that the average salinity of the effluent is slightly above the salinity of the lateral flow from the southern boundary condition, which emphasizes the belief that the major source for salinity load in Field 17 is from the lateral saline flow.

The change in flow rate due to irrigation events can be observed as local spikes in the flow that last for around 1 day after which the flow approaches the same rate of the lateral flow. The same observation can be seen for the effluent salinity that experiences a drop in its value due to the application of irrigation water that has salt concentration less than the groundwater.

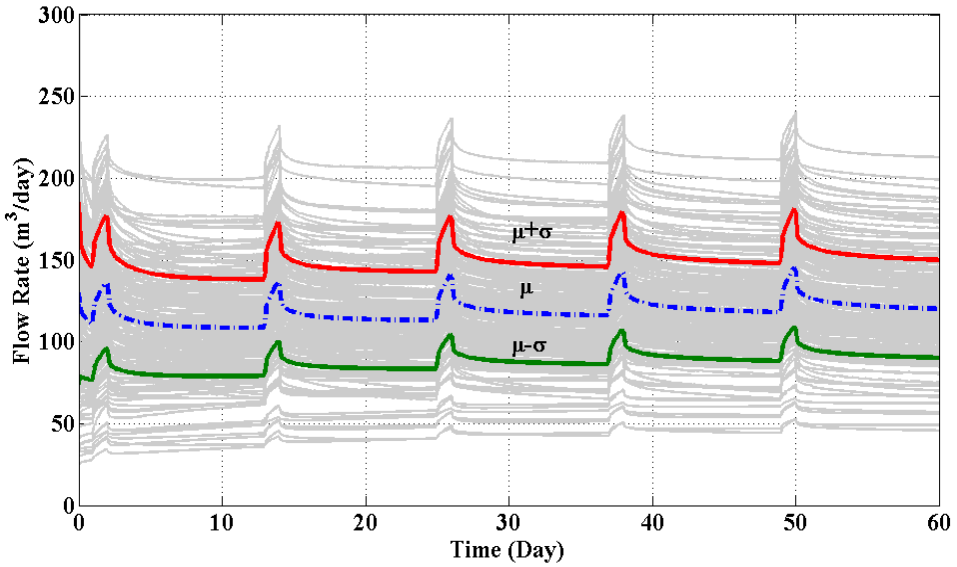

Figure 4. 16: Drain Outflow Hydrographs Include the Mean μ and $\mu \mp \sigma$, where σ is the Standard Deviation

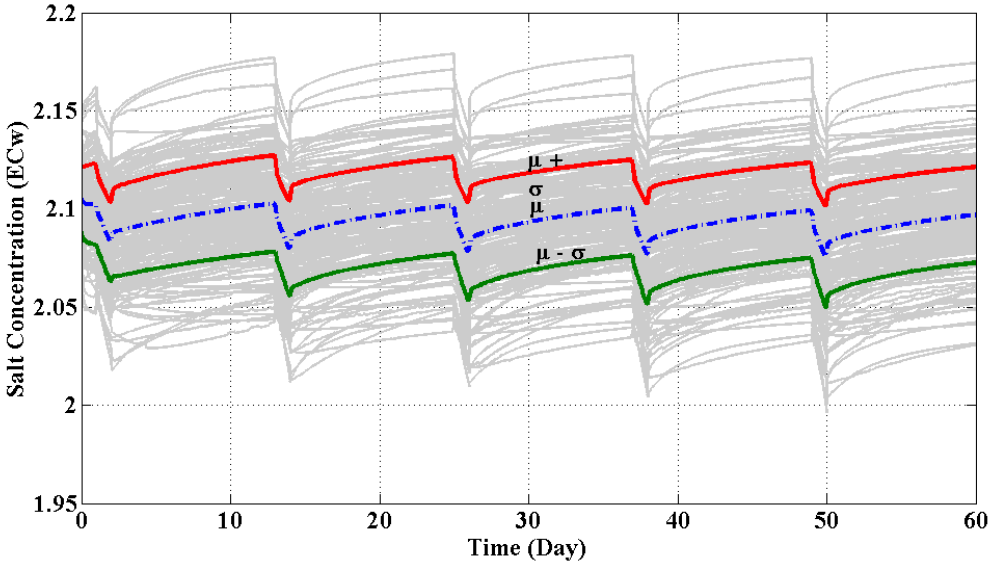

Figure 4. 17: Drain's Effluent Salinity Hydrographs Include the Mean μ and μ ∓ σ, where σ is the Standard Deviation

4.8.7 Temporal Variability of Root Zone Hydrosalinity

The statistical properties of the root zone hydrosalinity shown in section 4.8.1 to 4.8.5 illustrate their spatial variability; however, the root zone hydrosalinity varies across time as well. It is difficult to visualize the change across time for the two-dimensional root zone without resorting to animation. Alternatively, the simulated hydrograph of root zone variables are plotted for a point that is located middle of the field at the coordinate (150m, 150m). Figure 4.18 shows the simulated hydrographs of the root zone moisture content. The effect of the biweekly irrigation events can be easily noticed. The median (50% percentile) of the water content fluctuate between a maximum water content of 0.37 and a minimum of 0.17.

The driving force of the moisture content change, from the model perspective, is the root extraction of fresh water, which also drives the temporal change of salt concentration in water phase (Figure 4.19). The potential of salt build-up can be clearly noticed from the 95% percentile salinity hydrograph and to a lesser degree from the 50% percentile salinity hydrograph. On the other hand, the 5% percentile salinity hydrograph experience a slight decrease in the trend of salinity with time. Modifying the irrigation schedule and the application rate might alleviate the salt build-up demonstrated by the 95% and 50% percentile.

The temporal root extraction rate is shown in Figure 4.20. The impact of the daily reference evapotranspiration can be noticed from the high fluctuation of the uptake rate; though, the impact of irrigation events can be seen as a spike in the uptake rate.

An interesting aspect of the root zone hydrology is the time variability of the vertical flux (Figure 4.21), which experiences a change not only in magnitude of the flow, but also its direction. Directly following the irrigation event, the downward flow is dominant. Then, the vertical flow reverses to the upward direction in a response to the high root extraction following an irrigation event. An up-flux rate of 8mm/day is highly probable in periods between subsequent irrigation events. The variability of the up-flux rate is a reflection to the variability of the evapotranspiration. One can notice that immediately before an irrigation event, the up-flux rate reached its lowest value; this might be due to the relatively dry condition that the root zone experienced, which results in a low conductivity of the porous media.

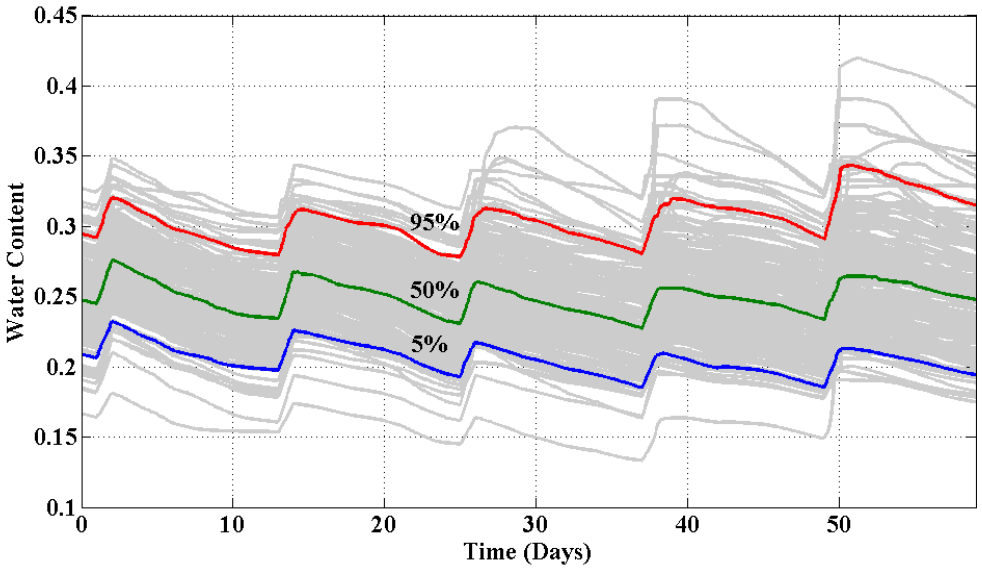

Figure 4. 18: The Temporal Variability of the Root Zone Average Water Content

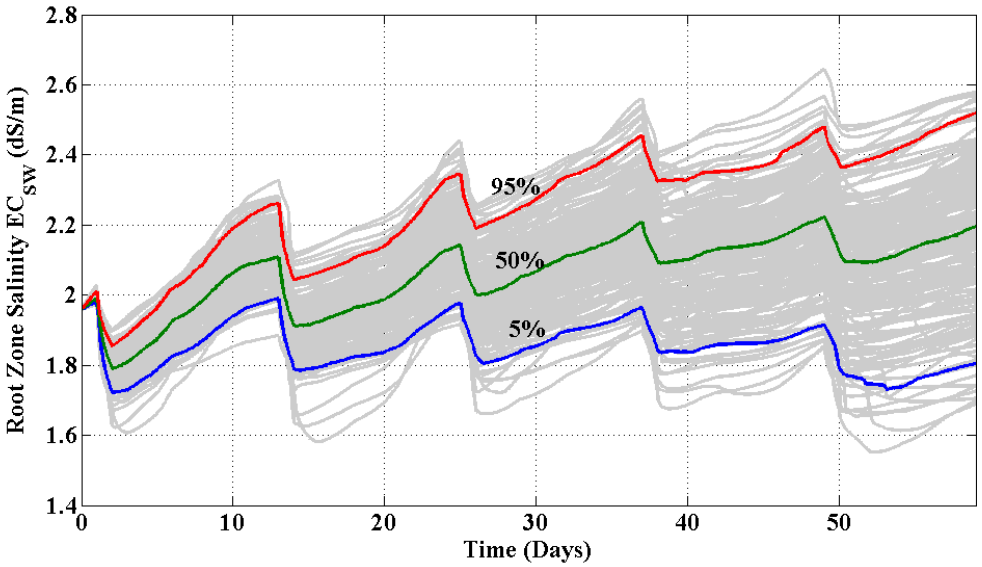

Figure 4. 19: The Temporal Variability of the Root Zone Average Salinity

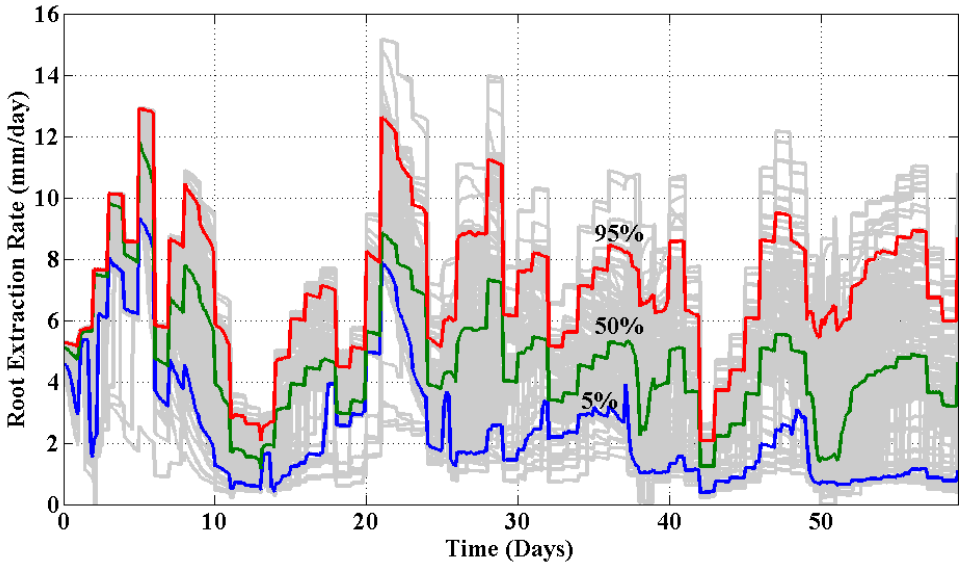

Figure 4. 20: The Temporal Variability of the Root Extraction Rates

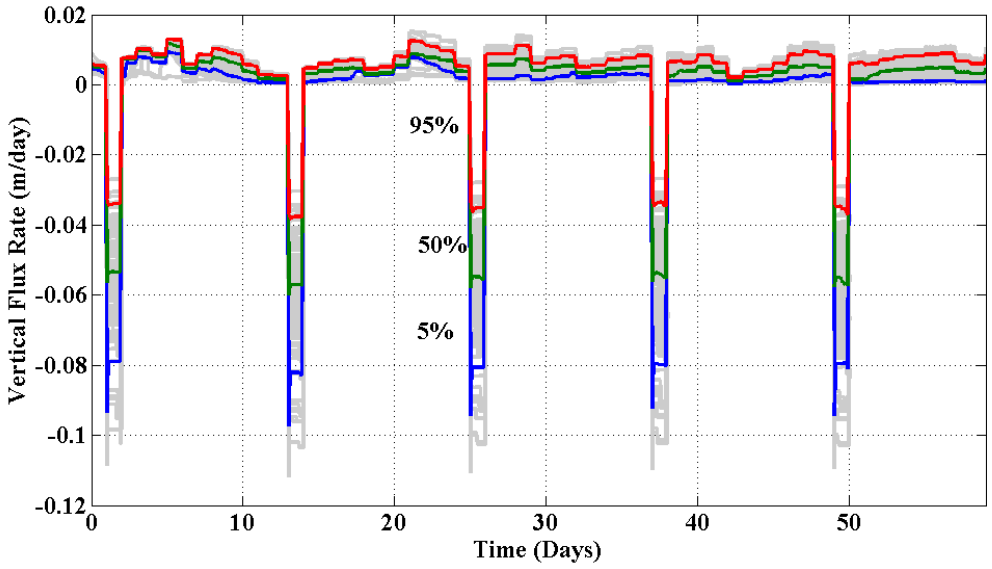

Figure 4. 21: The Temporal Variability of the Vertical Water Flux

4.9 Conclusion and Summary

The prediction of the relative crop yield under irrigation and drainage conditions using numerical models requires the knowledge of a large number of input parameters that are usually sparse and contaminated with measurement errors. This paper attempts to quantify the impact of parameter uncertainty on the prediction of crop yield and the hydrosalinity conditions of the root zone. The input parameters are categorized into four groups, 1) three-dimensional soil properties, 2) two dimensional parameters such as irrigation uniformity and preferential flow fraction, 3) irrigation application parameters such as diverted volumes, irrigation efficiency, and irrigation water salinity, and finally 4) semi empirical scalar parameters that control drainage flow, root water uptake, and crop yield. The three-dimensional parameters are randomized using the

multivariate Sequential Indicator Simulation of the correlated soil properties. Other parameters are assumed independent and sampled from their respective PDFs, which were inferred from field data and published data. The statistical moments of model's responses are evaluated spatially, such as groundwater depths, relative crop yield, root zone salinity, and deep percolation.

As interceptor of the lateral flow, the subsurface drainage system controls the groundwater table efficiently in the downstream; while the upstream area is only slightly affected by the drain. The reduction in the groundwater table in the field's downstream area improves the leaching capacity of salts, while upstream salinity increased due to poor leaching. Results show that the installation of the subsurface drainage is necessary to intercept the saline lateral flow from the southwestern direction and, consequently, improves the field crop yield. The crop yield standard deviations for corn are almost the same of alfalfa.

It is vital to note that the simulated crop yield and hydrosalinity conditions presented in this paper are responses of a specific assumed irrigation design, a sprinkler irrigation system that applies irrigation water regularly. Undoubtedly, these assumptions are not applicable to the entire basin. Although expanding the study to account for the entire Lower Arkansas River basin sounds appealing to decision makers, it is faced by major obstacles. For instance, simulating the variably saturated flow and transport on a regional scale could be computationally prohibitive. A simplified conceptualization of the unsaturated zone, e.g. the unsaturated zone package in MODFLOW (UZF1 package, Niswonger et al. (2006)), might be a pragmatic

response of modelers, though the conceptual model uncertainty introduced by this simplification still needs to be quantified.

4.10 References

Alzraiee, A., Garcia, L. and Burnett, R. (2009) "Modeling Spatial and Temporal Variability in Irrigation and Drainage Systems: Improvements to the Colorado State University Irrigation and Drainage Model (CSUID)", Presented and published in the proceedings of the USCID Conference on Irrigation and Drainage for Food, Energy and the Environmental, November 3-6, Salt Lake City, Utah.

Ayers, RS, and DW Westcot. 1994. "Water Quality for Agriculture: FAO Irrigation and Drainage Paper 29 Rev. 1." FAO. Rome.

Bliesner, Ron D., and Jack Keller. 2001. Sprinkle and Trickle Irrigation. The Blackburn Press, March 1.

Bresler, Eshel, and Gedeon Dagan. 1988. "Variability of yield of an irrigated crop and its causes: 2. Input data and illustration of results." Water Resources Research 24 (3): 389. doi:10.1029/WR024i003p00389.

Burkhalter, J. P, T. K Gates, and others. 2005. "Agroecological impacts from salinization and waterlogging in an irrigated river valley." Journal of Irrigation and Drainage Engineering 131: 197.

Cardon, G. E, and J. Letey. 1992. "Plant Water Uptake Terms Evaluated for Soil Water and Solute Movement Models." Soil Sci. Soc. Am. J. 56 (6): 1876-1880.

Deutsch, Clayton V. 2002. Geostatistical Reservoir Modeling. 1st ed. Oxford University Press, USA, April 4.

Deutsch, Clayton V., and André G. Journel. 1997. GSLIB. Oxford University Press, January 1.

Deverel, S., and J. Fio (1991), Groundwater Flow and Solute Movement to Drain Laterals, Western San Joaquin Valley, California 1. Geochemical Assessment, Water Resour. Res., 27(9), 2233-2246.

Feddes, Reinder A., Piotr Kowalik, Krystina Kolinska-Malinka, and Henryk Zaradny. 1976. "Simulation of field water uptake by plants using a soil water dependent root extraction function." Journal of Hydrology 31 (1-2) (September): 13-26. doi:doi: DOI: 10.1016/0022-1694(76)90017-2.

Gates, T. K., L. A. Garcia, and J. W. Labadie. 2006. "Toward Optimal Water Management in Colorado's Lower Arkansas River Valley: Monitoring and Modeling to Enhance Agriculture and Environment." Colorado Water Resources Research Institute Completion Report (205).

Van Genuchten, M. T, and US Salinity Laboratory. 1987. A numerical model for water and solute movement in and below the root zone. United States Department of Agriculture Agricultural Research Service US Salinity Laboratory.

van Genuchten, M. Th. 1980. "A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils1." Soil Science Society of America Journal 44 (5): 892. doi:10.2136/sssaj1980.03615995004400050002x.

Ghassemi, F., A. J. Jakeman, and H. A. Nix. 1995. Salinisation of Land and Water Resources: Human causes, extent, management and case studies. CABI, March 2.

Goswami, Debashish, and Prasanta K. Kalita. 2009. "Simulation of base-flow and tile-flow for storm events in a subsurface drained watershed." Biosystems Engineering 102 (2) (February): 227-235. doi:doi: DOI: 10.1016/j.biosystemseng.2008.11.004.

Haan, P. K., and R. W. Skaggs. 2003. "Effect of parameter uncertainty on DRAINMOD predictions: I. Hydrology and yield." Transactions of the ASAE 46 (4): 1061–1067.

Hanks, R. J, and R. W Hill. 1980. Modeling crop responses to irrigation in relation to soils, climate and salinity. 6. International Irrigation Information Center.

Harbaugh, A. W, E. R Banta, M. C Hill, and M. G McDonald. 2000. MODFLOW-2000, The U. S. Geological Survey Modular Ground-Water Model-User Guide to Modularization Concepts and the Ground-Water Flow Process. United States Geological Survey.

Hoffman, Glenn J., Robert G. Evans, Marvin Eli Jensen, Derrel L. Martin, and Ronald L. Elliott. 2007. Design And Operation Of Farm Irrigation Systems. 2nd ed. American Society of Agricultural & Biological, October 30.

Hopmans, J. W, and K. L Bristow. 2002. "Current capabilities and future needs of root water and nutrient uptake modeling." Advances in Agronomy 77: 103–183.

Houk, Eric, W. Marshall Frasier, and Eric Schuck. 2004. The Regional Effects Of Waterlogging And Soil Salinization On A Rural County In The Arkansas River Basin Of Colorado. Western Agricultural Economics Association. http://ideas.repec.org/p/ags/waeaho/36229.html.

Johnson, Norman L., Samuel Kotz, and N. Balakrishnan. 1995. Continuous Univariate Distributions, Vol. 2. 2nd ed. Wiley-Interscience, May 8.

Journel, A. 1999. "Markov Models for Cross-Covariances." Mathematical Geology 31 (8) (November 1): 955-964.

Konikow, Leonard F., and John D. Bredehoeft. 1992. "Ground-water models cannot be validated." Advances in Water Resources 15 (1): 75-83. doi:doi: DOI: 10.1016/0309-1708(92)90033-X.

Miles, D. L. 1977. “Salinity in the Arkansas Valley of Colorado.“Interagency Agreement Report EPA-IAG-D4-0544.” Environmental Protection Agency, Denver, Colorado.

Molz, Fred J. 1981. “Models of water transport in the soil-plant system: A review.” Water Resour. Res. 17 (5): 1245-1260.

Montazar, Aliasghar. 2010. “Predicting alfalfa hay production as related to water distribution functions.” Irrigation and Drainage 59 (2): 189-202.

Muralidharan, Daya, and Keith C. Knapp. 2009. “Spatial dynamics of water management in irrigated agriculture.” Water Resources Research 45 (5) (May). doi:10.1029/2007WR006756. http://www.agu.org/pubs/crossref/2009/2007WR006756.shtml.

Niswonger, R.G., D.E. Prudic, *and* R.S. Regan. 2006. Documentation of the Unsaturated-Zone Flow (UZF1) Package for modeling unsaturated flow between the land surface and the water table with MODFLOW-2005. *USGS Techniques and Methods 6-A19. Reston, Virginia: USGS.*

Oreskes, Naomi, Kristin Shrader-Frechette, and Kenneth Belitz. 1994. “Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences.” Science 263 (5147). New Series (February 4): 641-646.

Postel, Sandra. 1989. Water for agriculture : facing the limits. Washington D.C.: Worldwatch Institute.

Rhoades, J. D., A. Kandiah, and A. M. Mashali. 1992. “The use of saline waters for crop production.” FAO Irrigation and Drainage Paper (FAO) 48.

Rubin, Yoram, and Dani Or. 1993. “Stochastic modeling of unsaturated flow in heterogeneous soils with water uptake by plant roots: The Parallel Columns Model.” Water Resources Research 29 (3): 619. doi:10.1029/92WR02292.

Schaap, M. G, F. J Leij, and M. T van Genuchten. 2001. “Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions.” Journal of Hydrology 251 (3-4): 163–176.

Schnepf, R. 2010. “Agriculture-Based Biofuels: Overview and Emerging Issues.”

Shalhevet, J., A. Vinten, and A. Meiri. 1986. “Irrigation interval as a factor in sweet corn response to salinity.” Agronomy journal (USA).

Steenhuis, T. S, J. Boll, G. Shalit, J. S Selker, and I. A Merwin. 1994. “A simple equation for predicting preferential flow solute concentrations.” J. Environ. Qual 23 (5): 1058–1064.

Tanji, Kenneth K. 1990. Agricultural Salinity Assessment and Management. Amer Society of Civil Engineers, August.

Umali, Dina L., and Dina Umali-Deininger. 1993. Irrigation-induced salinity: a growing problem for development and the environment. World Bank Publications, November.

U.S. Dept. of Agriculture (USDA). 1972a. Soil survey of Otero county, Colorado, USDA, SCS, La Junta, Colo.

U.S. Dept. of Agriculture (USDA). 1972b. Soil survey of Bent county, Colorado, USDA, SCS, La Junta, Colo.

Veenhof, D. W., and R. A. McBride. 1994. "A preliminary performance evaluation of a soil water balance model (SWATRE) on corn producing croplands in the RM of Haldimand-Norfolk." Soil compaction susceptibility and compaction risk assessment for corn production. Centre for Land and Biological Resources Research AAFC, Ottawa: 112–142.

Wackernagel, Hans. 2003. Multivariate Geostatistics. 3rd ed. Springer, April 10.

Wang, X., J. R. Frankenberger, and E. J. Kladivko. 2006. "Uncertainties in DRAINMOD predictions of subsurface drain flow for an Indiana silt loam using the GLUE methodology." Hydrological Processes 20 (14): 3069-3084.

Warrick, A. W., and W. R. Gardner. 1983. "Crop yield as affected by spatial variations of soil and irrigation." Water Resources Research 19 (1): 181. doi:10.1029/WR019i001p00181.

Wichelns, Dennis. 1999. "An economic model of waterlogging and salinization in arid regions." Ecological Economics 30 (3) (September): 475-491. doi:doi: DOI: 10.1016/S0921-8009(99)00033-6.

Xu, Moujin, and Yoram Eckstein. 1997. "Statistical Analysis of the Relationships Between Dispersivity and Other Physical Properties of Porous Media." Hydrogeology Journal 5 (4) (April 1): 4-20.

Appendix A: Convertion Equations

- $EC_{sw} = \frac{\rho_b}{\rho_s}\left(\frac{n}{1-n}\right)\frac{1}{\theta}EC_e$
- Conversion from salinity concentration to osmotic head $\Phi(cm) = 356.89EC$
- Conversion from salinity concentration to osmotic head $\Phi(cm) = -339.0 * (C/640)^{1.05}$

Appendix B: Relationship between Christiansen's Uniformity Coefficient

The uniformity of a system can be defined for a sprinkler using the Christiansen's Uniformity Coefficient CU using the following equation:

$$CU = 100\left[1 - \frac{\sum |x_i - x_m|}{\sum x_i}\right]$$

where the variables $\{x_1, \dots, x_n\}$ are measured depth of water in equally spaced cans on a grid

The average absolute deviation is defined as

$$D = \frac{1}{n}\sum_{i=1}^{i=n} |x_i - x_m|$$

The relationship between standard deviation and average absolute deviation can described as follows:

$$\frac{1}{n}\sum_{i=1}^{i=n} |x_i - x_m| = \sigma\sqrt{\frac{2}{\pi}}$$

$$CU = 100\left[1 - \frac{n\sigma\sqrt{\frac{2}{\pi}}}{\sum x_i}\right]$$

Assume that the average irrigation depth is unity, then

$$\sum x_i = n$$

Then the standard deviation as a function of CU is given by the following equation:

$$\sigma^2 = \frac{\pi}{2}(1 - CU/100)^2$$

5 MULTI-OBJECTIVES AQUIFER SAMPLING USING ENSEMBLE KALMAN FILTER FOR OPTIMAL SPATIAL PREDICTIONS AND COVARIANCE-PARAMETERS ESTIMATION

5.1 General

Effective sampling of groundwater systems is an essential effort toward gaining insight into the system's behavior. Data collection is usually motivated by different objectives; for example, minimizing prediction errors at unsampled locations; estimating the spatial covariance parameters; using minimum cost of installing and operating the monitoring network. Essentially, sampling of a groundwater system is not limited to a single parameter but multiple parameters (multivariate sampling problem) that fully characterize the system. This paper employs the Ensemble Kalman Filter as a flexible tool to incorporate the sampling of different system parameters (e.g. hydraulic conductivity) and system variables (e.g. hydraulic head). The approach is investigated by applying it to a two-dimensional steady state groundwater problem. The formulated objective function is a multi-objective integer optimization where the decision variables include the number of hydraulic conductivity measurements, the number head measurements, and their spatial locations. The optimization problem is searched using

the multi-objective Genetic Algorithm. Several design scenarios are investigated; and the implications of different sampling cost are also studied. Results show that the Ensemble Kalman Filter is a very flexible tool for tackling multivariate sampling problems. Moreover, a tradeoff design for minimizing prediction errors and spatial covariance parameters estimation can be approached as a multi-objective optimization problem.

5.2 Introduction

The unknown spatial variability of the hydrogeological controlling parameters, e.g. hydraulic conductivity, contributes significantly to the uncertainty of the flow and transport model predictions. The geostatistical method (Matheron 1962, Isaaks and Srivastava 1990, Diggle and Ribeiro 2007) has been widely used to model spatial variability. Within the geostatistical framework, the field measurements are typically analyzed to infer a spatial statistical model as well as its geostatistical parameters, i.e. spatial correlation length, variance, nugget effect, and trend surface. This inferred model is supposed to serve the ultimate objective of geostatistics analysis, which is to provide unbiased predictions at unsampled locations (Kriging). Obviously, the accuracy of the predictions is substantially affected by the relevancy of the inferred structural parameters.

Bridging the gap between the unknown reality of groundwater systems and our status of knowledge can solely be achieved by collecting more data. In particular, groundwater monitoring networks are supposed to provide insight into the behavior of groundwater systems. A quantitative characterization of the behavior of the

groundwater systems might be determined by identifying the spatial and temporal variability of the *system's variables*, e.g. piezometeric levels, groundwater velocities and water chemistry, etc.; and the *system's parameters*, e.g. the hydraulic conductivity, storativity, porosity, etc.

Unfortunately, the amount of data collected is usually constrained by logistical and budgetary considerations. This is beside the fact that groundwater systems are distributed parameter systems (DPS) that cannot be determined uniquely using a finite set of measurements. Several studies explore the optimal design scheme for distributed systems (Ucinski 2004). The kriging-based design is one of the widely used methods to optimize spatial data collection. Specifically, the optimal sampling design of a univariate spatial variable might be achieved by minimizing the prediction variance (Kriging Variance), which is a function of the spatial locations of the data (See for example (Yfantis, Flatman, and Behar 1987); (Cressie, Gotway, and Grondona 1990). The optimization criterion in these works is either the maximum kriging variance or the average kriging variance.

Although the term *optimal designs* usually refers to efficient prediction design, another design objective might be to provide optimal estimates of the random field's model; more specifically, the spatial covariance function's parameters. (Bogaert and Russo 1999) provided a methodology for variogram parameters estimation based on the generalized least square approach and under the hypothesis that the random field is Gaussian second-order stationary. Other works on design for variogram estimation include (Müller et al. 1997, Zhang et al. 2010). Zhu et al. (2005) introduced a design that is optimal for the maximum likelihood estimation of the covariance parameters.

The geostatistical design (Diggle and Ribeiro 2007) incorporate the uncertainty of the geostatistical parameters in design using a Bayesian weighted average of predictive distribution as a design criterion.

Combining a design that is efficient for prediction as well as for covariance parameter estimation in one design scheme seems to be natural and promising. Zimmerman (2006) proposed a design that is optimal for both prediction and covariance-parameters estimation using the Empirical Best Linear Unbiased Estimation (Empirical Kriging). Zimmerman (2006) noticed that design for prediction and for covariance-parameters estimation is antithetical. In other words, while optimal prediction requires a wide and uniform spread of measurements, the optimal design for covariance-parameters estimation requires spatial clustering of the measurements.

The increased awareness of environmental issues in 1970's, crowned by the passing of the Clean Water Act 1972, focused more attention on optimal data collection, particularly water quality data. In the field of groundwater, several studies have been conducted with the goal of optimally designing groundwater-monitoring systems (Loaiciga et al. 1992). As an example, Carrera et al. (1984) used a kriging-based approach to optimally sample fluoride concentrations. Hsu et al. (1989) proposed an experimental design for parameter identification of groundwater system to identify the number and the locations of pumping and observation wells.

The inability of kriging-based methods to provide a design that monitors the physical process across space and time increased interest in the Kalman Filter (Kalman 1960) . In groundwater hydrology, the filter was used as a framework to determine the

spatiotemporal distribution of the sampling of groundwater systems (Van Geer 1987, Wu 2004). Andricevic (1990) employed a branch and bound technique and Kalman Filtering to obtain optimal solution for a discrete set of samples. Herrera et al. (2005) used the Kalman filter coupled with a stochastic transport model to provide a minimum cost design. Zhang et al. (2005) combined the genetic algorithm with a static Kalman filter and a stochastic groundwater flow and contaminant transport model to obtain a least cost design for groundwater quality monitoring. Recently, Kollat et al. (2011) employed the bias-aware ensemble Kalman filtering to improve the long-term monitoring while accounting for model errors.

In this research, the optimal design theory is applied to obtain the optimal number and locations of hydraulic conductivity and hydraulic head. This can be seen as a multivariate optimal sampling problem. Fortunately, the Ensemble Kalman Filter (Evensen 2009) provides a natural instrumentation for the multivariate sampling design algorithm. The advantage of this instrumentation is that the designer can include the measurement errors associated with a given measurement method. Moreover, cross-covariance of conductivity and head can be approximated using ensembles produced from numerical simulations, that is to say no field data is required to establish the cross-covariance of conductivity and head.

Another objective of this paper is to seek a local optimal design for prediction and conductivity covariance parameters estimation as a multi-objective optimization problem. This differs from Zimmerman (2006) in that the Empirical Kriging approximate criterion is not used; instead both objectives are competing equally to produce a tradeoff design. This approach enables decision makers to choose a design

from a set of optimal designs that have different efficiencies for prediction and covariance parameter estimation. The classical tradeoff between performance and cost is also investigated at different relative pricing of measurements. The multi-objective genetic algorithm is utilized to solve the optimization problem.

This approach is applied to augment the existing monitoring network in Field 17 in Rocky Ford, Colorado. The existing network consists of a number of observation wells and Cone Penetration Test (CPT) measurements of hydraulic conductivity.

The paper is organized as follows: a brief review of deign paradigms are outlined in section 5.3; section 5.4 outlines the methodology used and the setting of the multi-objective GA optimizations. Section 5.5 illustrates the setup of the computational experiments and the different scenarios investigated. Results are shown and discussed in section 5.6. Final notes and conclusion are presented in section 5.7.

5.3 Network Design Paradigms

A suitable design framework depends on the particular circumstances of the problem to be tackled. There is a rich published literature about optimal experimental designs. A comprehensive treatment of spatial data collection appears in Müller (2007) and Ucinski (2004). In terms of design methodology (Zidek et al. 2010), the design could be geometry-based, probability-based, or model-base design. In the geometry-based design, the sample locations can be chosen to optimize a specified geometrical pattern; one the other hand, the probability-based method is based on sampling of an assumed PDF of the underplaying process. The model-based design optimizes a

specified statistical criterion; this method is widely used in environmental monitoring network design.

Another perspective on categorizing experimental designs is based on the status of prior knowledge about the field (Diggle et al. 2007). The design can be *retrospective design* if an existing monitoring network is required to be augmented by deleting some sensors or adding new ones; or *prospective design* if the design does not consider any prior knowledge about the field. Sometimes the concern of the monitoring network is not limited to one variable, i.e. *univariate optimal design*, but requires sampling several spatial variables, i.e. *multivariate optimal design* (Li 2009). In terms of the design objective, the design goal could be to optimize prediction, estimate the random process geostatistical parameters, determine the trend of spatial process, and determine the extreme values for spatial process, among others.

5.4 Methodology

The general continuity Equation (5.1) is typically used to describe the flow of fluids in porous media. This equation establishes the relationship between aquifer soil parameters, e.g. hydraulic conductivity and storativity, and the response variables, e.g. head field and flow rate.

$$\boldsymbol{\nabla.(K.\nabla h) + Q = S\frac{\partial h}{\partial t}} \qquad \textbf{(5.1)}$$

Where K is the hydraulic conductivity [L/T] , h the hydraulic head [L], Q is a sink or source term per unit volume of the porous media [1/T], and S the storativity term [1/L].

The numerical solution of Equation (5.1) requires the knowledge of the hydraulic conductivity field within a discretized domain of field under study D, as well as the boundary conditions and initial conditions of the state variable. In other words, the random field $S(x)$, where $S(x) = \log \mathrm{K}(x)$, should be defined at every numerical node, the vector $x = (x_1, \dots, x_n)^T, x_i \in D$, where n is the number of active cells in the domain D.

It is convenient to decompose $S(x)$ into a deterministic component (trend model) and a stochastic component such that the term ε_s is stationary and has a Gaussian distribution $\varepsilon_s \sim N(0, \Sigma(\theta))$.

$$\boldsymbol{S = X\beta + \varepsilon_s,} \tag{5.2}$$

In equation 5.2, $[X\beta]$ is the vector form of the trend surface. To simplify the analysis further, it is presumed that $[X\beta] = \mu$; as a result, the spatial random process is modeled as $S(x) \sim N(\mu, \Sigma(\theta))$; where μ is the stationary average of $\log \mathrm{K}(x)$ and $\Sigma(\theta)$ is the spatial covariance matrix. Several functional forms of the spatial covariance have been used; herein the spherical covariance function is adopted to model the spatial correlation. Accordingly, the $(i, j)th$ element of the covariance matrix might be computed using Equation 5.3 (assuming a zero nugget effect).

$$\boldsymbol{\sigma_{ij}(\theta)} = \begin{cases} \boldsymbol{c - c\left(1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3\right)} & \boldsymbol{h \le a} \\ \boldsymbol{0} & \boldsymbol{h \ge a} \end{cases} \tag{5.3}$$

Where $\theta = (c, a)$ is the structural parameters vector, c is the stationary variance, a is the correlation length, and h is the Euclidian distance between two points x_i and x_j.

The assumption of the zero nugget effect is compensated for by allowing specified measurement errors in the Kalman Filter in section 5.4.1, instead of dealing with it as unknown parameters.

In the multivariate optimal design setting, the intent is to sample n_K hydraulic conductivity measurements and n_h head measurements, as well as their spatial locations, in order to optimize a specified statistical inference about the multivariate random field of the conductivity and head. In the next section, the Ensemble Kalman Filter is utilized as an instrument to calculate cross-covariance of the $\log(K)$ and h, which will be used later in calculating the design criteria.

5.4.1 Ensemble Kalman Filter

The Kalman Filter was initially proposed by (Kalman 1960) and was widely used as a data assimilation technique. The underlying hypothesis for the use of the filter, in the estimation of linear dynamic systems, is that the systems noises are multivariate Gaussian processes. Consider the state of an evolving in time model (5.4)

$$\boldsymbol{x_{k+1} = f(P; x_k; q_k; w_k),} \tag{5.4}$$

In the context of groundwater modeling, this transition function may resemble the flow in the porous media (Equation 5.1), where x_{k+1} is the forecasted hydraulic head vector (The state), P is hydraulic conductivity parameter vector, q_k, and w_k represent deterministic and stochastic stressing terms. Note that k represents the time index. The predictions made by the transition function (5.4) are usually in discrepancy with field

measurements. Conveniently, the relation between model predictions and field measurement is described in the following equation:

$$\boldsymbol{z_{k+1} = H_{k+1}.x_{k+1} + v_{k+1}} \quad \textbf{(5.5)}$$

In equation (5.5), $z_{k+1} \in R^{m\times 1}$ is the field measurements vector, $H_{k+1} \in R^{m\times n}$ is the incident matrix of binary constants, which maps the state vector space to the measurements vector space. The term m represents the number of measurements while n is the number of active numerical cells in the simulated domain. The vector v_{k+1} $\in R^{m\times 1}$is a measurement noise vector. The process noise w_k and the measurement noise v_k are assumed Gaussian with the following PDFs:

$$\boldsymbol{p(w_k) \sim N(0, Q)} \quad \textbf{(5.6)}$$

$$\boldsymbol{p(v_k) \sim N(0, R)} \quad \textbf{(5.7)}$$

With the ensemble Kalman Filter approach (Evensen 2009), data assimilation is a two-stage process. In the first stage (the forecast stage), an ensemble of the system's parameters *P* is used as an approximation to the probability distribution. Each realization in the parameter ensemble is processed to obtain the state vector. At the end of the forecast stage, the parameters ensemble and state are augmented as follows:

$$\boldsymbol{X_{k+1/k}} = \begin{bmatrix} \boldsymbol{x_{11}} & \cdots & \boldsymbol{x_{1n_r}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{x_{n1}} & \cdots & \boldsymbol{x_{n.n_r}} \\ \boldsymbol{P_{11}} & \cdots & \boldsymbol{P_{1n_r}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{P_{n1}} & \cdots & \boldsymbol{P_{n.n_r}} \end{bmatrix} \quad \textbf{(5.8)}$$

The term n_r in equation (5.8) is the number of realization; and n is the number of active numerical cells. The expectation of the augmented state ensemble and forecast error covariance are calculated as follows:

$$\widehat{\boldsymbol{X}}_{\boldsymbol{k+1/k}} = \boldsymbol{E}[\boldsymbol{X}_{\boldsymbol{k+1/k}}] \tag{5.9}$$

$$\boldsymbol{C}_{\boldsymbol{k+1/k}} = \boldsymbol{E}\left[\left(\boldsymbol{X}_{\boldsymbol{k+1/k}} - \widehat{\boldsymbol{X}}_{\boldsymbol{k+1/k}}\right).\left(\boldsymbol{X}_{\boldsymbol{k+1/k}} - \widehat{\boldsymbol{X}}_{\boldsymbol{k+1/k}}\right)^{\boldsymbol{T}}\right] \tag{5.10}$$

In the second stage (update stage), the expected state $\hat{X}_{k+1/k} \in R^{2n\times 1}$ can be further improved by assimilated newly collected data, following a Bayesian least-square estimate:

$$\widehat{\boldsymbol{X}}_{\boldsymbol{k+1/k+1}} = \widehat{\boldsymbol{X}}_{\boldsymbol{k+1/k}} + \boldsymbol{K}_{\boldsymbol{k+1}}(\boldsymbol{z}_{\boldsymbol{k+1}} - \boldsymbol{H}_{\boldsymbol{k+1}}.\widehat{\boldsymbol{X}}_{\boldsymbol{k+1/k}}) \tag{5.11}$$

$$\boldsymbol{K}_{\boldsymbol{k+1}} = \boldsymbol{C}_{\boldsymbol{k+1/k}}.\boldsymbol{H}_{\boldsymbol{k+1}}^{\boldsymbol{T}}.\left(\boldsymbol{H}_{\boldsymbol{k+1}}.\boldsymbol{C}_{\boldsymbol{k+1/k}}.\boldsymbol{H}_{\boldsymbol{k+1}}^{\boldsymbol{T}} + \boldsymbol{R}_{\boldsymbol{k+1}}\right)^{-\boldsymbol{1}} \tag{5.12}$$

Where $K_{k+1} \in R^{2n\times m}$is a gain matrix, $C_{k+1/k} \in R^{2n\times 2n}$ is the forecasted covariance matrix, and $R_{k+1} \in R^{m\times m}$ is the measurement errors covariance matrix. Note that in this case of augmented matrix $H_{k+1} \in R^{m\times 2n}$. The covariance of update error is calculated as:

$$\boldsymbol{C}_{\boldsymbol{k+1/k+1}} = [\boldsymbol{I} - \boldsymbol{K}_{\boldsymbol{k+1}}.\boldsymbol{H}_{\boldsymbol{k+1}}].\boldsymbol{C}_{\boldsymbol{k+1/k}}.[\boldsymbol{I} - \boldsymbol{K}_{\boldsymbol{k+1}}.\boldsymbol{H}_{\boldsymbol{k+1}}]^{\boldsymbol{T}} + \boldsymbol{K}_{\boldsymbol{k+1}}.\boldsymbol{R}_{\boldsymbol{k+1}}.\boldsymbol{K}_{\boldsymbol{k+1}}^{\boldsymbol{T}} \tag{5.13}$$

The updated covariance matrix (Equation 5.13) will be used in the next section to calculate the design criterion. The calculation of the gain matrix in equation 5.12 requires the inversion of the term$(H_{k+1}.C_{k+1/k}.H_{k+1}^T + R_{k+1})$, which could be singular

matrix. This singularity results from the inclusion of measurements (e.g. conditioning the realizations to field measurements) that have zero or very small measurement error, which results in a row and a column that is close to zero. In order to circumvent this obstacle, the known matrix inversion was substituted with the Moore–Penrose pseudoinvers (Moore 1920, Penrose et al. 2008).

5.4.2 Monitoring Network Design

Equation (5.13) represents the updated covariance matrix of the estimation error. Since $C_{k+1/k+1}$ is independent of the measurement value z_{k+1}, it is possible to design a monitoring network (spatial locations) and measurements frequency (time schedule) that minimizes the estimation error based on a specified statistical criterion.

Several alphabetical optimality criteria were traditionally used (Steinberg and Hunter 1984); for example, the A-optimality minimizes the trace of the inverse information matrix, the D-optimality maximizes the determinant of the information matrix, and E-optimality maximizes the minimum eigenvalue of the information matrix. Herein, the A-optimality is used:

$$\boldsymbol{argmin}_{M \subset D}\ \boldsymbol{f}_1(\boldsymbol{M}) = \boldsymbol{tr}\left(\boldsymbol{C}_{k+1/k+1}\right) \qquad \textbf{(5.14)}$$

Since the updated covariance matrix (5.13) is the augmented covariance matrix for state variable and parameter vectors (for example conductivity and head), then the optimal design seeks the optimal number of hydraulic conductivity and head measurements as well as their optimal spatial locations. This design is efficient for spatial prediction. However, the drawback of this design is the implicit assumption that

the spatial covariance parameters (θ) which, were used to generate number of realizations for the hydraulic conductivity, are known.

From the several studies that attempted to provide designs for variogram parameter estimation, Zhu and Stein (2005) is of particular importance to this study. Zhu and Stein (2005) suggested using the maximum likelihood method to approximate the covariance of θ .This method suggests that under certain regularity conditions, the maximum likelihood is asymptotically efficient and the asymptotic covariance matrix of the estimators is given by the inverse of the Fisher information matrix. The ijth element of information matrix $I(\theta)$ is given by:

$$\boldsymbol{I_{i,j}(\theta) = 0.5\, tr\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_i}\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_j}\right)} \quad \textbf{(5.15)}$$

The D-optimal design for minimizing the determinant of the inverse of the information matrix can be given by:

$$\boldsymbol{argmin_{M\subset D}\, f_2 = -\log\, \det I\,(\theta, M)} \quad \textbf{(5.16)}$$

Since the design criterion depends on the parameter vector θ, the maximum likelihood estimation of the parameters $\hat{\theta}$ is used; and the design obtained is considered, consequently, as locally optimum.

It is intuitive to conclude that more field measurements results in a better prediction; however the budget resources are usually limited. Therefore, the measurements cost can enter the tradeoff through a third objective function,

$$\boldsymbol{argmin_{M\subset D}\, f_3 = c_k.n_k + c_h.n_h} \quad \textbf{(5.17)}$$

Objective functions in equations 5.14, 5.15 and 5.17 can be combined to find the optimal M design variables that minimize the multi-objective problem

$$\boldsymbol{argmin}_{\boldsymbol{M}\subset\boldsymbol{D}}\, \boldsymbol{F}(\boldsymbol{M}) = \begin{bmatrix} \boldsymbol{f_1} \\ \boldsymbol{f_2} \\ \boldsymbol{f_3} \end{bmatrix} \tag{5.18}$$

Where the decision variables vector $M = \{x_{1k}, .., x_{nk}, x_{1h}, .., x_{nh}\}$, such that x_{ik} is the spatial location of the i^{th} conductivity measurement, x_{ih} spatial location of the i^{th} head measurement, and nk and nh are the number of conductivity measurements and head measurements respectively.

5.4.3 Multi Objective Genetic Algorithm

Optimizing the design objective function (Equation 5.18) is an NP hard problem where it is usually not possible to find exact solution in a reasonable time. Some researchers used the Simulating Annealing Algorithm (Zhu et al. 2005, Zimmerman 2006). Note that the optimizing problem herein is an integer-optimization problem where the solution space consists of spatial indices and the number of K and h measurements, which are also integers. The Genetic Algorithm (GA) is a promising heuristic search algorithm for optimization of such problems, as well as the multi objective problems that are too complex to be solved using deterministic techniques such as gradient-based methods.

A multi objective optimization approach seeks to find optimal trade-offs to obtain solutions that are optimal in some sense or acceptable to a decision maker (Coello et al. 2002). Normally, there is no unique solution to this problem, but rather a set of

solutions called the Pareto optimal set. The Pareto Optimality concept is essential to approach the problem. In non-formal language, the Pareto optimal solution is one wherein improvement in one objective function results in degradation in another. In this paper, the nondominated sorting genetic algorithm (NSGA-II) (Deb et al. 2002) is applied to obtain a set of trade-offs. The method is known for its computational efficiency, and it uses elitism and a crowded comparison operator to produce diverse solutions.

5.4.4 Genetic Algorithm Setting

The vector of decision variables M is coded using a chromosome (Figure 5.1) in which the first two genes carry the number of conductivity measurements n_k and the number of head measurements n_h respectively. The next n_k genes carry the spatial location indices of conductivity measurements followed by n_h genes for the spatial location indices of head measurements. The initial population is generated such that measurement locations are within the active cells domain and that spatial location indices of measurements are integers. The initial population creation was so that the chromosome does not replicate the same genes; and the locations of existing measurements excluded.

The selection process is based on Roulette wheel selection where each parent is proportionally represented in the wheel according to the rank of its fitness value. The crossover is designed to occur at a single point such that the crossover occurs at any location as long as the total length of the chromosome is constant. A crossover rate of 0.8 was used. The mutation rate was chosen to be 0.11 and to produce only locations

within the active domain. Elite selection was allowed so that the best three solutions will survive in each evolution step in order to guarantee not losing the best solutions.

Figure 5. 1: The Scheme for Decision Variables Vector (Chromosome)

5.5 Computational Experiments

The methodology outlined in section 5.3 is applied to two examples. The first example is a one-dimensional synthetic groundwater steady-state flow problem. This example serves as simple illustrative problem. Moreover, the obtained design can be easily compared to exact theoretical designs. The second example is a two-dimensional steady-state groundwater flow problem, which is a conceptual simplification of the groundwater flow regime in Field 17 in Rocky Ford, Colorado.

5.5.1 One-Dimensional Synthetic Case

The purpose of this example is to examine the validity of the GA coupled with the Ensemble Kalman Filter approach to obtain optimal sampling design for a one-dimensional groundwater flow problem. The system is 1500m in length with a constant upper stream head of 50m. The flow path is divided into 150 cells. The hydraulic heads (Figure 5.2) at each cell were computed using equation (5.19)

$$\boldsymbol{h_i = h_o + q \sum_{j=0}^{j=i} \frac{\nabla x_j}{K_i}} \quad \textbf{(5.19)}$$

The term $h_o = 50m$ is the upstream constant head, q is the specific flow [L/T] and is assumed to equal 4.3×10^{-4} m/day. The unconditional hydraulic conductivity realizations were generated using Gaussian Sequential Simulation (Deutsch and Journel 1997), where $\log(K) \sim N(0,1)$ and the correlation scale is 300m (Figure 5.2). To be consistent with the theoretical requirements of the Kalman Filter, the head field should be transformed to normality before using the filter. Consequently, the conductivity ensemble and head ensemble are used to calculate the forecasted cross-covariance matrix (Left of Figure 5.3). It can be seen that the diagonal of the conductivity covariance matrix is almost uniform since the simulation is unconditional. On the other side, the covariance diagonal of the head is zero at the upper stream end, and increases as it approaches the downstream; this is because of the constant head boundary condition imposed upstream and the high variability of the head downstream.

The GA optimization is implemented to minimize the trace of the cross-covariance. The simulation converged within a relatively short time (50 generations). The resulting design was found to have almost regular spatial intervals in the hydraulic conductivity field. This result is consistent with the theoretical work by Papageorgiou et al. (1998) for finite correlated samples. The updated cross-covariance matrix shows the relatively regular pattern of the design in the hydraulic conductivity field.

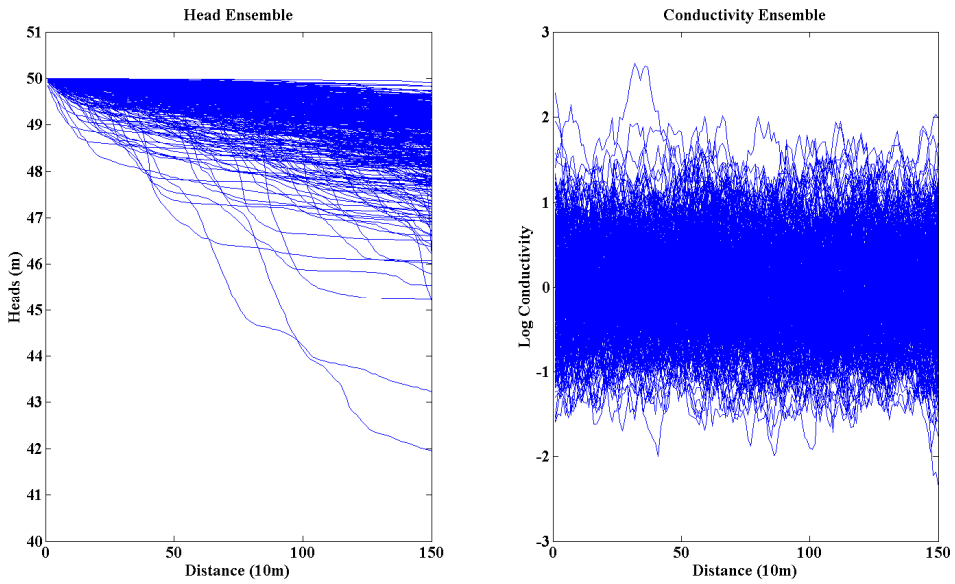

Figure 5. 2: Hydraulic Conductivity and Head Realization for One-dimensional Groundwater Flow

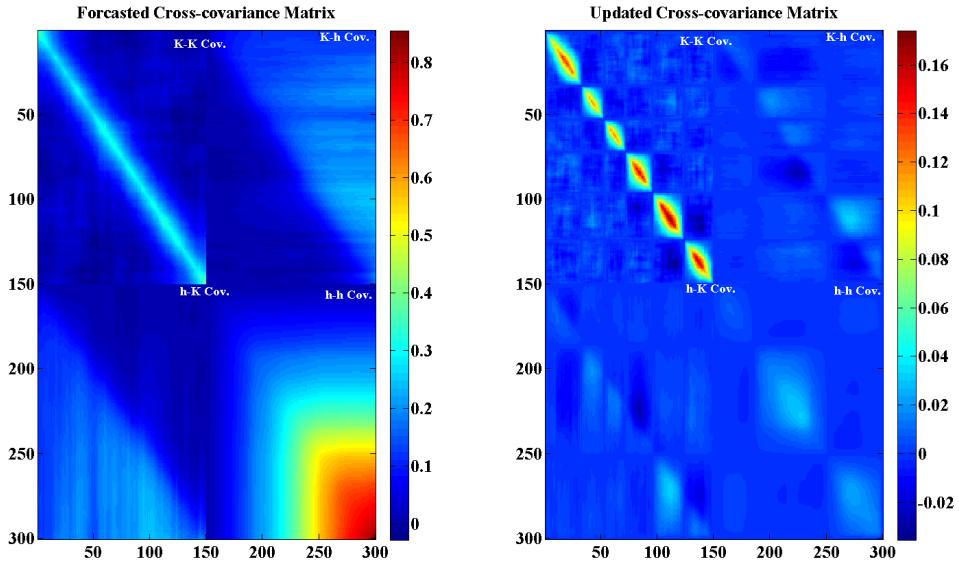

Figure 5. 3: The Forecasted and the Updated Cross Covariance for One-dimensional Flow Problem

5.5.2 Two-Dimensional Field Application

This section outlines the redesign of the monitoring network in Field 17 in Rocky Ford, Colorado. The east-west dimension of the field is 600m and the north-south dimension is 420m. The Cone Penetration Test (CPT) was used to obtain the hydraulic conductivity vertical profile at 15 locations. In addition, 31 observation wells were used to monitor the groundwater table (Figure 5.4). MODFLOW (Harbaugh et al. 2000) is used to simulate the groundwater flow regime in the field. The model domain is discretized into 60 columns and 42 rows (10m by 10m). The upstream boundary condition is assumed to be constant head and was divided into 5 sections to capture the variability of the head at the boundary condition. The same thing is assumed for the downstream constant head boundary condition. The in-field yearly averaged evapotranspiration map was obtained using the ReSET model (Remote sensing of evapotranspiration, Elhaddad et al. 2011).

The vertically averaged conductivity measurements are used to generate 500 conditional realizations, where each is processed in the flow model to obtain the head field. Prior to the calculation of the forecasted cross-covariance matrix, the conductivity realizations and the head realizations were transformed to normal distribution $N\sim(\mu = 0, \sigma = 1)$ using a normal score transformation. The spatial random field is assumed isotropic with horizontal correlation length of around 95m.

The simulation study is divided into four experiments; the objective of the first is to obtain optimal design that is efficient for prediction only. The second seeks optimal design that is efficient for covariance parameter estimation (CPE) and prediction. The

third experiment seeks cost effective design for prediction, and the fourth experiment combines prediction, CPE and cost objectives in one general formulation.

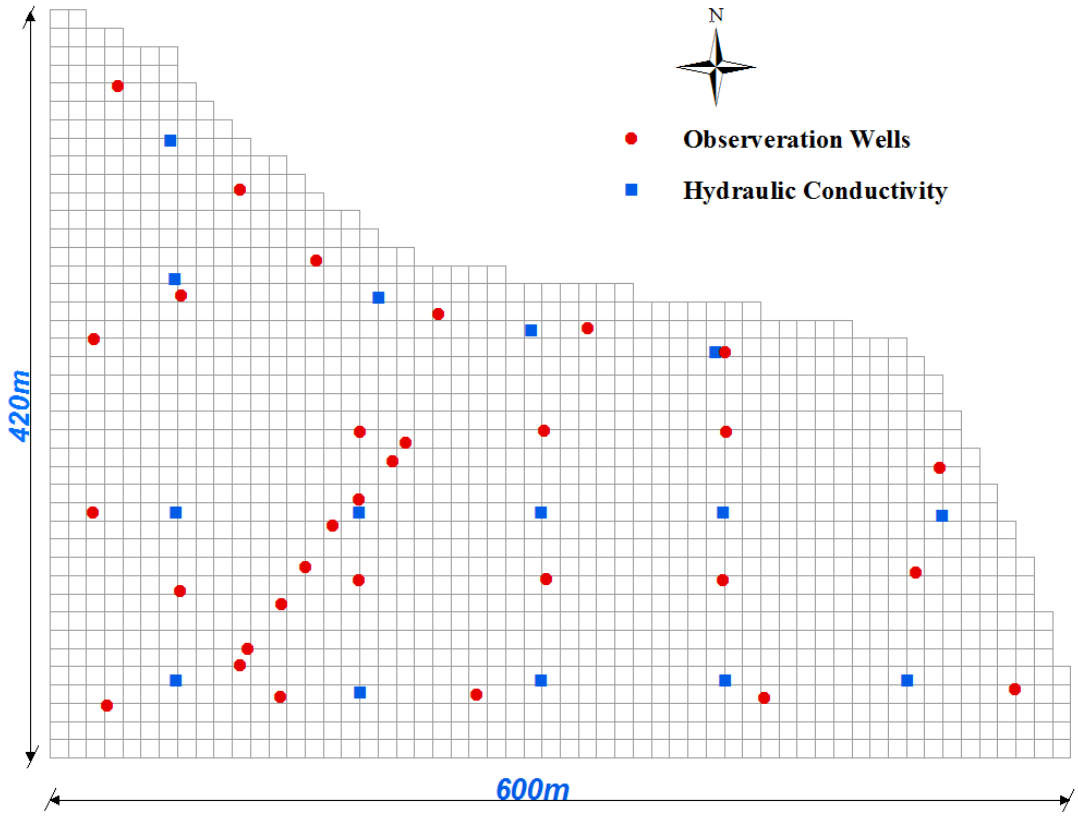

Figure 5. 4: Existing CPT Conductivity Measurements and Observation Wells

5.5.2.1 Design for Prediction

In this experiment, ten new measurements (Conductivities and Heads) are intended to be optimized. The optimal design is investigated using the following scenarios:

- **Single Objective Optimization (Scenario A-1)**: This scenario seeks the optimal design for minimum prediction error using equation (5.14) as the objective function. Since the single objective genetic algorithm does not guarantee a global optimal solution, and in order to ensure a better scan of the solution space, the optimization procedures are repeated eight times, and the one that achieves the minimum fitness value is taken as the best design. The initial population size is chosen to be 32 possible solutions, which were left to evolve for 500 generations. In the event that the best solution remains unchanged continuously for 150 consecutive generations the optimization is stopped.
- **Multi Objective Optimization (Scenario A-2)**: In the previous scenario, the objective function was the trace of the cross-covariance matrix. In this scenario, the trace of conductivity covariance matrix and the trace head covariance matrix are treated as two separate objectives. Scenarios A-1 can be seen as aggregation of the two objectives; this might result in a design that favors one objective over the other. Within the multi objective optimization, each of the objectives will be equally contributing to the tradeoff.

5.5.2.2 Design for Prediction and Covariance Parameter Estimation (Scenario B-1)

This optimization experiment seeks an optimal tradeoff design for minimum prediction error (equation 5.14) and minimum structural parameters estimation error (equation 5.16). The design for the covariance parameters estimation applied in this paper is a local optimal design, in which an estimate of the covariance parameters is

used. This is to distinguish it from the global optimal design where no estimates of the parameters are available. Zhu and Stein (2005) applied a minimax approach to provide a global optimal design, where they initially obtain the covariance parameters that maximize equation 5.16, and then the same equation is minimized to obtain the best design. This approach is computationally intensive as it requires searching the covariance parameter space as well as the decision variables space. Alternatively, the designs presented in this work were obtained at correlation scale $a \in \{\mathbf{50}\boldsymbol{m}, \mathbf{100}\boldsymbol{m}, \mathbf{200}\boldsymbol{m}\}$**.** In all of these designs the variance is assumed the same. The multi-objective GA optimization is employed to determine a total number of 20 new measurements (conductivities and heads) and their spatial locations. The population size is assumed 30 and the maximum number of generations is 500.

5.5.2.3 Design for Prediction and Cost (Scenario C-1)

A practical concern for sampling designs is how the cost of each variable's sampling affects the tradeoff. In this paper, it is assumed that the cost of conductivity measurement is 1 monetary unit and the cost of the head measurement is a ratio (α) of the conductivity cost. So equation 17 is replaced by the following equation:

$$\boldsymbol{argmin}_{\boldsymbol{M}\subset\boldsymbol{D}}\, \boldsymbol{f_3} = \boldsymbol{n_k} + \boldsymbol{\alpha}.\boldsymbol{n_h} \qquad \textbf{(5.20)}$$

where $= \frac{\boldsymbol{c_h}}{\boldsymbol{c_k}}$. The cost function could be complex if the measurement cost is also a function of its location. Herein, it is assumed that the cost does not change with location; and the design is repeated at ratios $\boldsymbol{\alpha} \in \{\mathbf{0.1}, \mathbf{0.5}, \mathbf{0.9}\}$. The multi-objective GA optimization is employed to determine a total number of 10 new measurements

(conductivities and heads) and their spatial locations. The population size and the maximum number of generations are as in scenario B-1.

5.5.2.4 Design for Prediction, Covariance Parameter Estimation and Cost (Scenario D-1)

In this experiment, the three objective functions (Equation 5.18) are combined. The Covariance parameter estimation objective function is at local correlation scale equal to 100m; and the relative cost objective is evaluated at price ratio $\alpha = 0.1$. Note that the Pareto front in this case is a three-dimensional surface.

5.6 Results and Discussion

The computations were executed on two four-core computers where the MATLAB Parallel toolbox was used to parallelize the computational tasks. The Genetic Algorithmic function in MATLAB (Chipperfield et al. 1994) was modified to fit the needs of integer programming required to solve the design optimization. The results of the numerical experiments and the obtained designs for each of the scenarios outlined in Section 5.4.2, are illustrated and discussed in the flowing subsections.

5.6.1 Optimal Prediction Designs (Scenario A-1)

Figure 5.5 shows the evolution of the eight GA simulations. The lowest fitness values are found ranging between 535.98 and 529.48. The corresponding designs for the eight simulations are shown in Figure 5.6. The best design is shown separately in

Figure 5.7. One might observe that the eight designs are roughly similar; this might indicate a consistent result of the GA simulations. An obvious observation is that the proposed hydraulic conductivity measurements are larger in number than the proposed head in the eight designs. The reason for that might be that the number of the existing water level wells is double the number of conductivity measurements. Another possible explanation is that the head field is bounded from downstream and upstream by constant heads; this might restrict the variability of the head comparing to the conductivity.

Another observation about the resulting designs is that the measurements are concentrated in the southern part of the field. There are 5 conductivities measurements along the 530m southern border, whereas the 700m northern border also has 5 measurements. No measurements for conductivity are taken on western boundary.

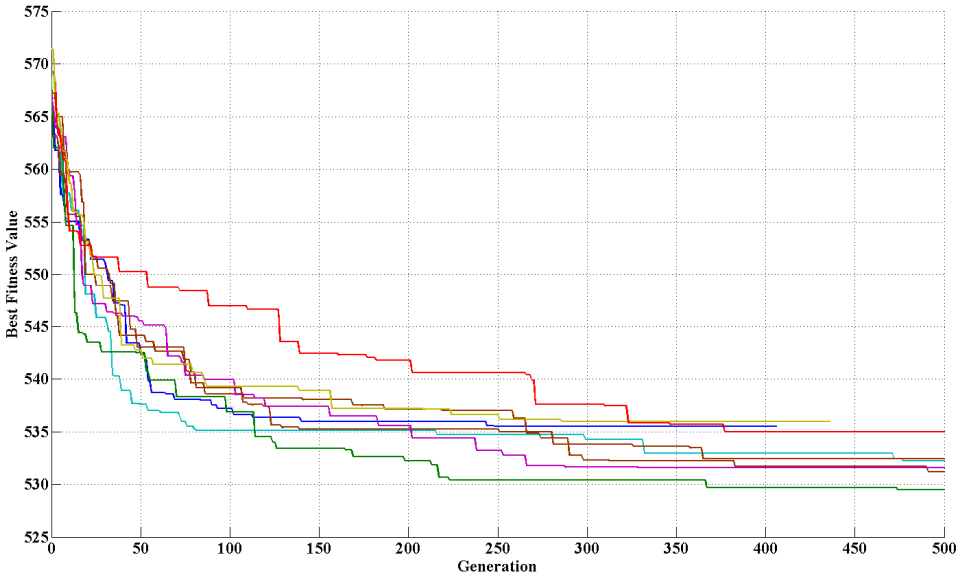

Figure 5. 5: Multi Evaluations of GA Optimization Problem and Their Best Fitness Value Evolutions

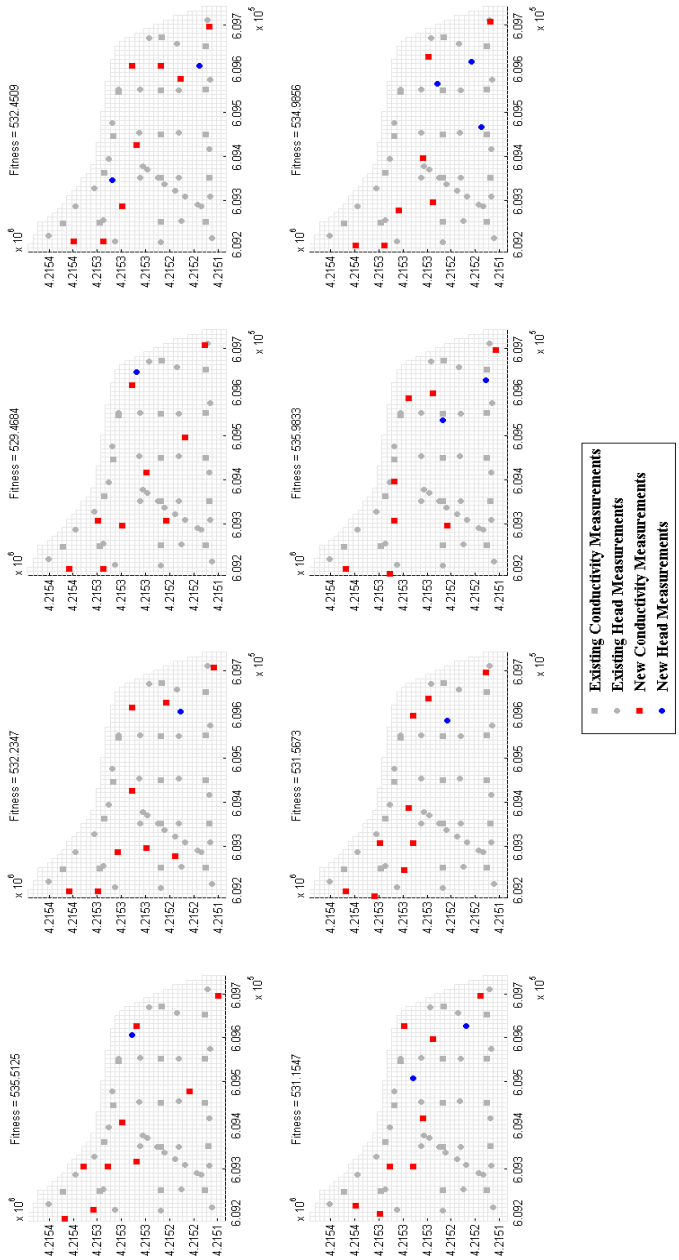

Figure 5. 6: Design Results for Multiple Evaluations of GA Optimization

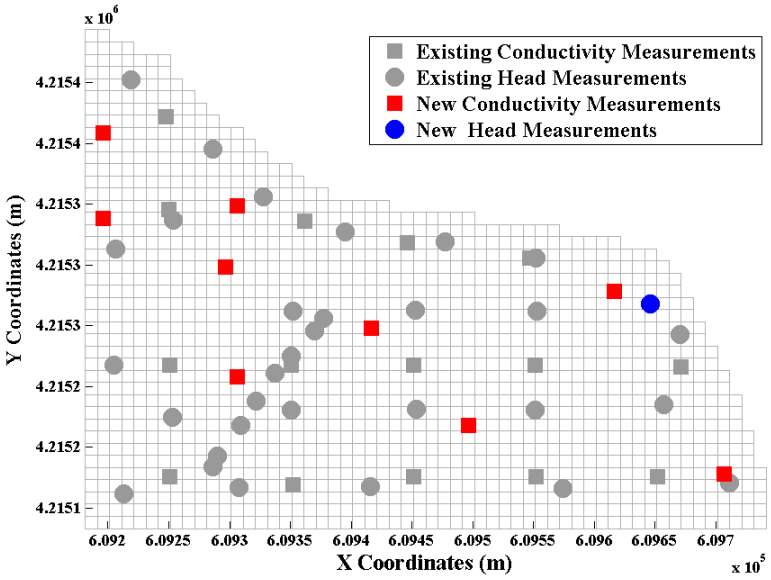

Figure 5. 7: The Best-fitness Design that Minimizes Prediction Errors

5.6.2 Optimal Prediction Designs (Scenario A-2)

The design criterion in Scenario A-1 is based on the trace of the cross-covariance matrix, which can be seen as an aggregation of two objective functions, i.e. the traces of the conductivity covariance matrix and the head covariance matrix. Aggregation of the two traces might mask contribution of one of the variables to the final design. In this scenario each of these traces is presented as a separate objective function. Figure 5.8 shows the objective function evaluations for each objective function as well as the Pareto optimal front. The multi-objective GA typically produces a set of optimal

designs that have different preferences for each objective function. Take, for example, design point A, which represents an optimal design for the conductivity, while the design point C produces an optimal design for the heads; point B, on the other hanf, is the midpoint design that places equal emphasis on both functions. The three designs are plotted in Figures 5.9, 5.10 and 5.11; it can be noticed that the number of conductivity measurements for design A is dominant while the number of head measurements in design C is dominant. Design B has an equal number of conductivity and head measurements.

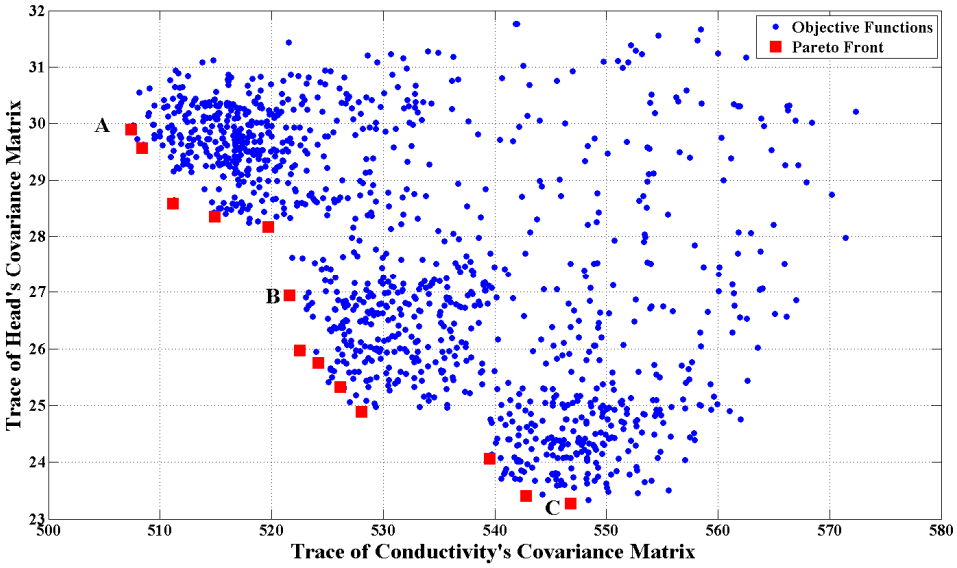

Figure 5. 8: Pareto Optimal Set for the Tradeoff of Conductivity and Head Predictions

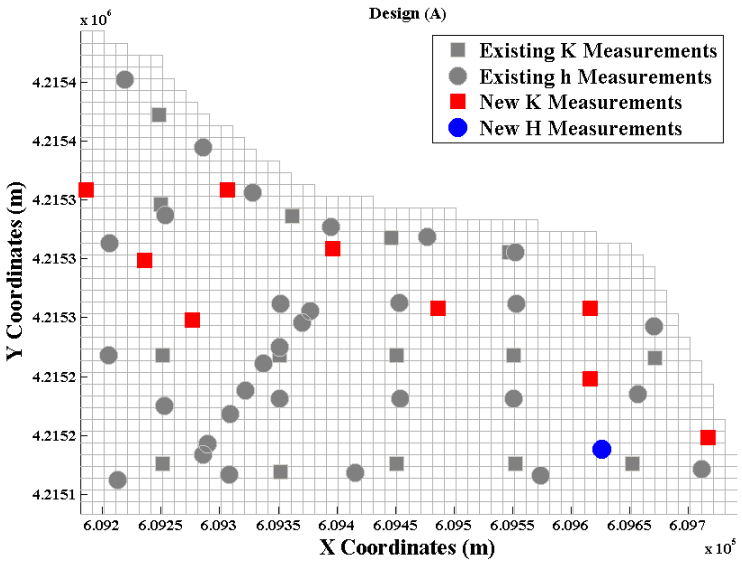

Figure 5. 9: Resulting Design for Point A in Figure 5.8

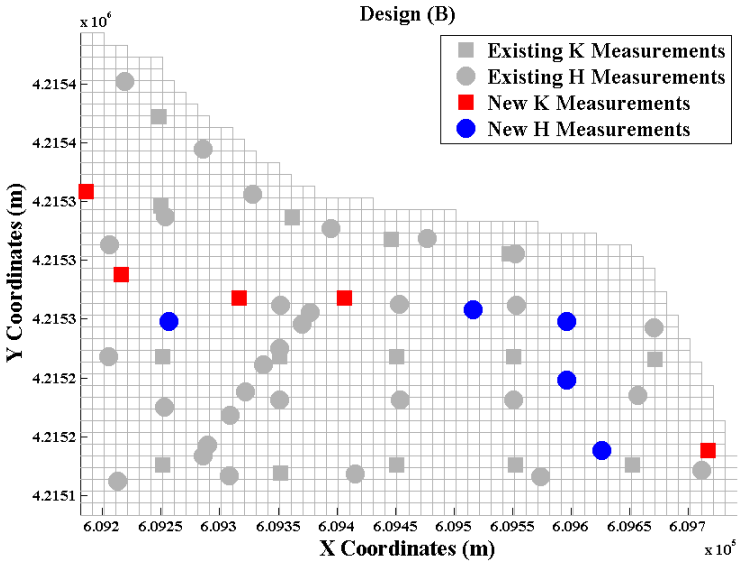

Figure 5. 10: Resulting Design for Point B in Figure 5.8

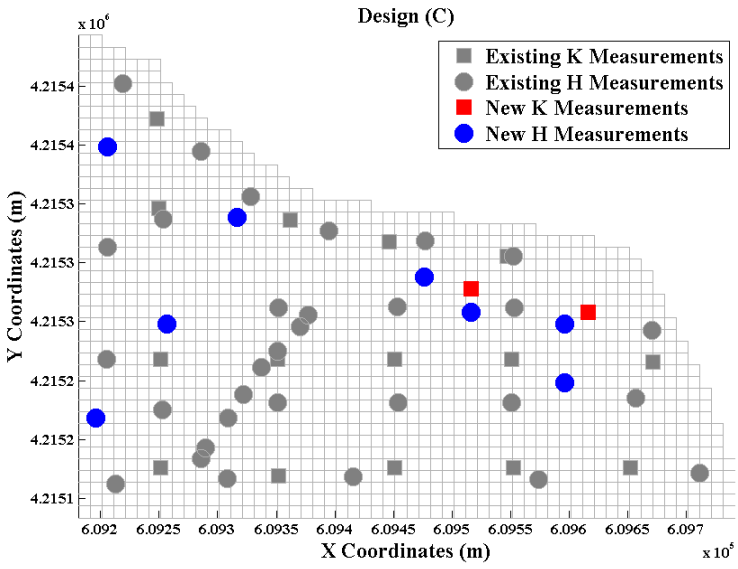

Figure 5. 11: Resulting Design for Point C in Figure 5.8

5.6.3 Optimal Prediction and CPE design (Scenario B-1)

The results of integrating the prediction and CPE objective functions are shown in Figures 5.12 to 5.16. Figure 5.12 shows the optimal Pareto set. It can be seen that as prediction function decreases the CPE objective is deteriorating; this is consistent with the observations of Zimmerman (2006). Note that the designs provided here are at local estimate of the correlation scale. In order to investigate different possibilities for the design, the design is repeated at correlation scales of the conductivity field, $a \in \{50m, 100m, 200m\}$. Figure 5.13 shows the Pareto front for each correlation scale. It is possible to notice that at minimum prediction error, the three curves get closer to each other and the differences become bigger at high prediction error (or at smaller CPE error). This might be attributed to the fact that as the objective function of CPE

increases, i.e. the design deteriorates, and the design converges to the prediction only design (Scenario A-1). Another observation is that as the correlation scale decreases, the prediction function general deteriorates. This is because small correlation scale requires small spatial sampling intervals to be detected, in contrast to the wide sampling spread required for optimal prediction. On the other hand, the large correlation scale makes the design widely spread, which is required for good prediction sampling.

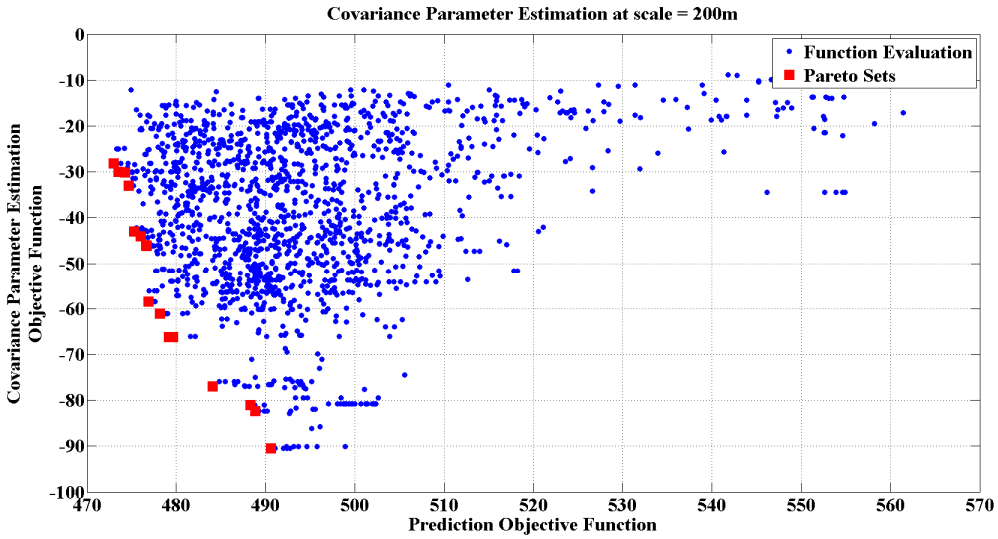

Figure 5. 12: Pareto Front Optimal Set for the Combined Prediction and CPE at Local Scale of 200m

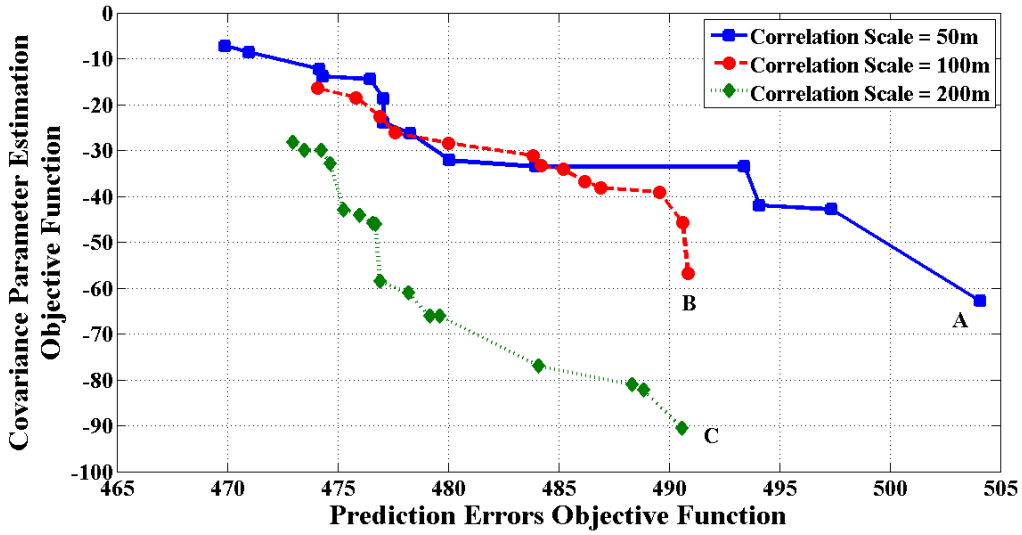

Figure 5. 13: Pareto Front Optimal Set for the Combined Prediction and CPE at Local Correlation Scale of 50m, 100m, and 200m

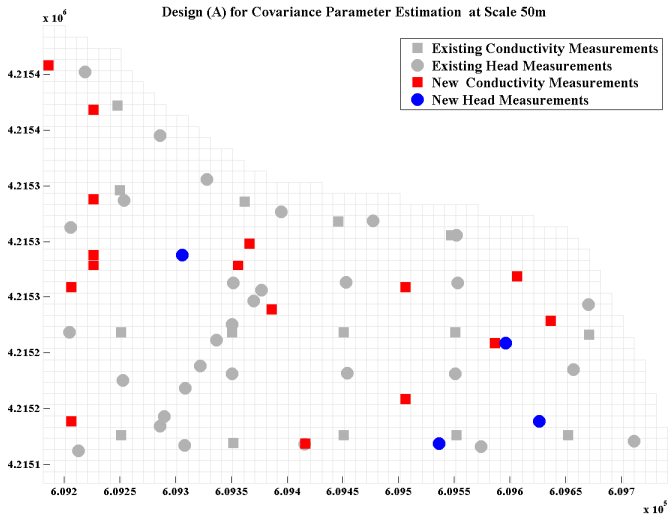

Figure 5. 14: Design at Point (A) in Figure 13 at Correlation Scale = 50m

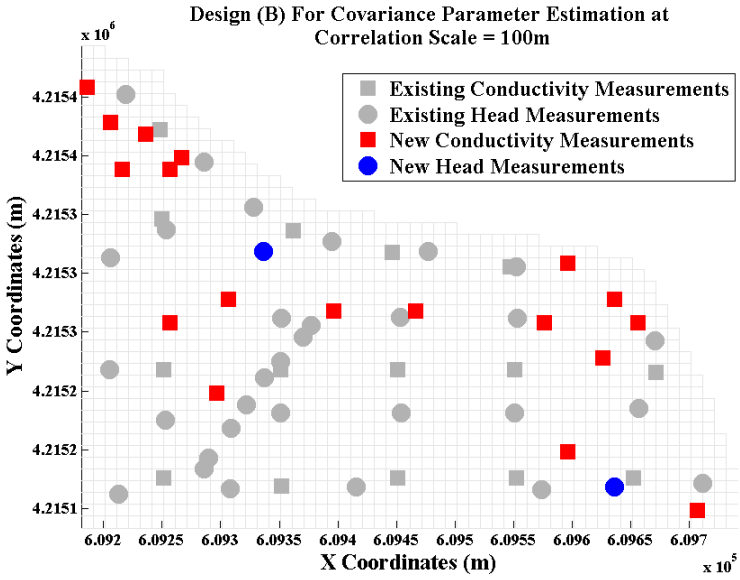

Figure 5. 15: Design at Point (B) in Figure 13 at Correlation Scale = 100m

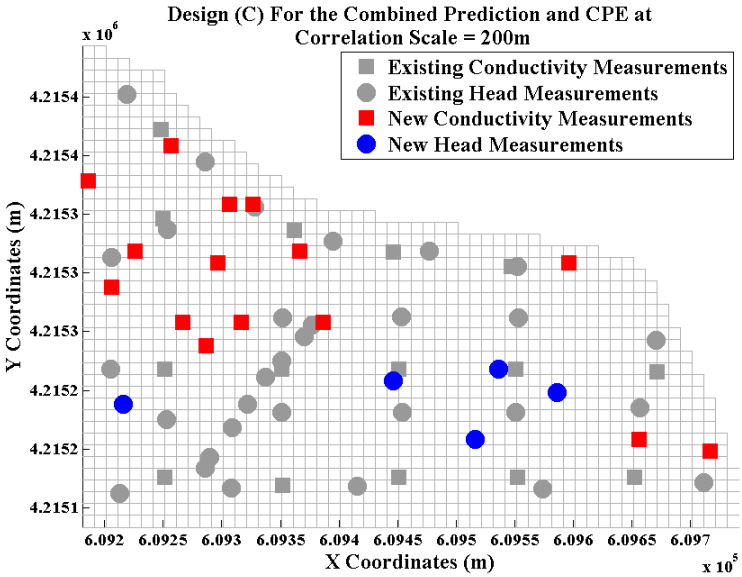

Figure 5. 16: Design at Point (C) in Figure 13 at Correlation Scale = 200m

5.6.4 Optimal Prediction and Relative Cost (Scenario C-1)

In scenario A-1, it was noticed that the obtained design suggests sampling more conductivity than heads measurements; however, the relative cost might produce a different preferred design than in scenario A-1. That is, if the cost of conductivity measurements is high compared to that of head measurements, then an economical design will prefer sampling more heads. Obviously, this case does not minimalize prediction errors. This notion is explored quantitatively by optimizing the multi-objective problem that consists of equation 5.14 and equation 5.20. The Pareto front is plotted in Figure 5.17 for the three cost ratios. It can be seen that as the ratio between head measurement cost and conductivity measurement cost approach unity, the Pareto front tends to be flat. In other words, the cost objective function will be neutral and the solution will be equivalent to a single objective function as in scenario A-1. It is worth noting that in practical problems, the cost might appear as a constraint rather than a separate objective function; as the proposed design cost must not exceed a predetermined budget.

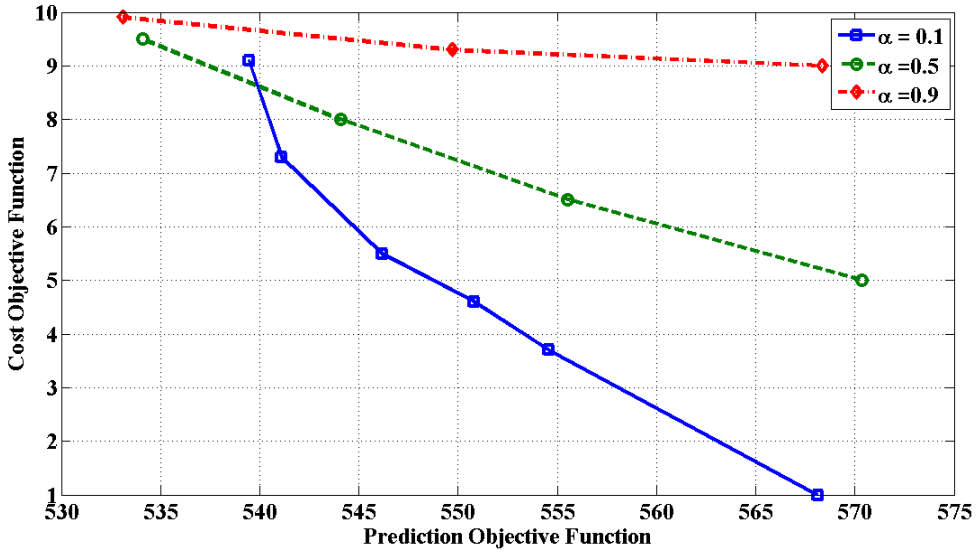

Figure 5. 17: Pareto Front Optimal Set for the Combined Prediction and Cost Objective Functions

5.6.5 Optimal Design for Prediction, CPE, and Relative Cost (Scenario D-1)

The objective of this experiment is to explore the designs that result from the interaction of the three objective functions. Figure 5.18-A shows the resulting Pareto front surface. The Side view (5.18-C) in the figure, shows that the relationship between optimal cost and optimal prediction can be clearly seen; however, no distinct relationship is apparent for the CPE objective and the cost as shown in Figure 5.18-D. The relationship between optimum prediction and optimum CPE, such as was shown in figure 5.12, is distorted after including the cost objective function (Figure 5.18-B).

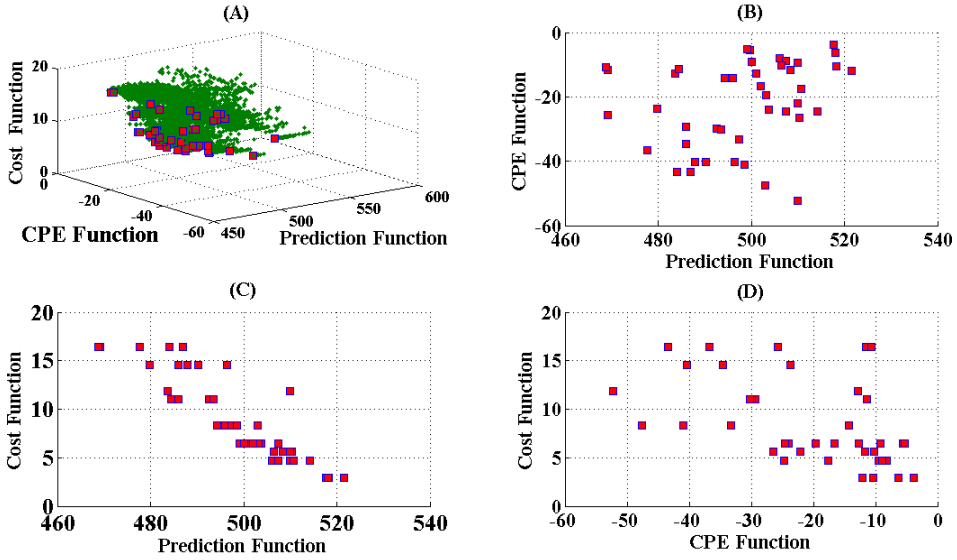

Figure 5. 18: Pareto Front Optimal Set for the Combined Prediction, Covariance Parameter Estimation and Cost Objective Functions in Figure (A) and the Corresponding Side Views in B, C and D

5.7 Conclusions

This paper investigated the use of the Ensemble Kalman Filter (EKF) to optimally sample multiple spatial variables that represent groundwater systems. These variables can be system parameters (e.g. conductivity, storativity, porosity, etc.) and state variables (head field, pollutant concentrations, velocities, etc.).

The efficiency of EKF stems from the ease of incorporating the physical relationships among the system variables in order to calculate the cross-covariance of systems variables. Moreover, the measurement's error of each variable can be included

in the design process. The possibility of obtaining the spatial and the temporal frequency of new samples further increases the efficiency.

The major drawback of the EKF approach is that the covariance matrix doubles for each new variable considered in the design. In cases of transient sampling schemes, the cross-covariance matrix size is the number of cells in the numerical domain multiplied by the number of variables, which change with time, multiplied by the number of time steps. The optimization procedures require multiple evaluations of equation 5.12, which includes the inversion of the cross-covariance matrix.

Integrating different design objectives, e.g. minimizing prediction errors and optimal design for CPE, was achieved through usage of multi-objective GA. The advantage of multi-objective optimization is that a range of good designs (optimal Pareto set) are obtained; each of these designs has different performance with respect to the objective functions. This might give decision makers more flexibility in preferring a specific design given a specific case.

5.8 References

Andricevic, R. 1990. "Cost-effective network design for groundwater flow monitoring." Stochastic Hydrology and Hydraulics 4 (1) (March): 27-41. doi:10.1007/BF01547730.

Bogaert, Patrick, and David Russo. 1999. "Optimal spatial sampling design for the estimation of the variogram based on a least squares approach." Water Resour. Res. 35 (4): 1275-1289.

Carrera, Jesus, Eduardo Usunoff, and Ferenc Szidarovszky. 1984. "A method for optimal observation network design for groundwater management." Journal of Hydrology 73 (1-2) (July 25): 147-163. doi:doi: 10.1016/0022-1694(84)90037-4.

Chipperfield, A., P. Fleming, H. Pohlheim, and C. Fonseca. 1994. "Genetic algorithm toolbox for use with MATLAB."

Coello, Carlos A. Coello, David A. Van Veldhuizen, and Gary B. Lamont. 2002. Evolutionary Algorithms for Solving Multi-Objective Problems. 1st ed. Springer, June 30.

Cressie, Noel, Carol A. Gotway, and Martine O. Grondona. 1990. "Spatial prediction from networks." Chemometrics and Intelligent Laboratory Systems 7 (3) (February): 251-271. doi:doi: 10.1016/0169-7439(90)80115-M.

Deb, K., A. Pratap, S. Agarwal, and T. Meyarivan. 2002. "A fast and elitist multiobjective genetic algorithm: NSGA-II." Evolutionary Computation, IEEE Transactions on 6 (2): 182-197.

Deutsch, Clayton V., and André G. Journel. 1997. GSLIB. Oxford University Press, January 1.

Diggle, P.J., and Paulo Justiniano Ribeiro. 2007. Model-based Geostatistics. 1st ed. Springer, March 12.

Elhaddad, Aymn, and Luis A. Garcia. 2011. "ReSET-Raster: Surface Energy Balance Model for Calculating Evapotranspiration Using a Raster Approach." Journal of Irrigation and Drainage Engineering 137: 203. doi:10.1061/(ASCE)IR.1943-4774.0000282.

Evensen, Geir. 2009. Data Assimilation: The Ensemble Kalman Filter. Springer, September 1.

Van Geer, F. C. 1987. "Applications of Kalman filtering in the analysis and design of groundwater monitoring networks." Delft University of Technology, Delft, The Netherlands.

Harbaugh, A. W, E. R Banta, M. C Hill, and M. G McDonald. 2000. MODFLOW-2000, The U. S. Geological Survey Modular Ground-Water Model-User Guide to Modularization Concepts and the Ground-Water Flow Process. United States Geological Survey.

Herrera, Graciela S., and George F. Pinder. 2005. "Space-time optimization of groundwater quality sampling networks." Water Resour. Res. 41 (12) (December 1): W12407.

Hsu, Nien-Sheng, and William W-G. Yeh. 1989. "Optimum experimental design for parameter identification in groundwater hydrology." Water Resour. Res. 25 (5): 1025-1040.

Isaaks, Edward H., and R. Mohan Srivastava. 1990. An Introduction to Applied Geostatistics. Oxford University Press, USA, January 11.

Kalman, RE. 1960. "A New Approach to Linear Filtering and Prediction Problems." Transactions of the ASME – Journal of Basic Engineering (82 (Series D)): 35-45.

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 (2) (February 18): W02529.

Li, Jie. 2009. "Spatial multivariate design in the plane and on stream networks." Theses and Dissertations (January 1). http://ir.uiowa.edu/etd/395.

Loaiciga, Hugo A., Randall J. Charbeneau, Lorne G. Everett, Graham E. Fogg, Benjamin F. Hobbs, and Shahrokh Rouhani. 1992. "Review of Ground-Water Quality Monitoring Network Design." Journal of Hydraulic Engineering 118 (1): 11. doi:10.1061/(ASCE)0733-9429(1992)118:1(11).

M\üller, W., and D. L Zimmerman. 1997. "Optimal design for variogram estimation."

Matheron, G. 1962. "Traité de géostatistique appliquée."

Moore, E.H. 1920. "On the reciprocal of the general algebraic matrix." Bull. Amer. Math. Soc 26: 394–395.

Müller, Werner G. 2007. Collecting spatial data: optimum design of experiments for random fields. Springer.

Papageorgiou, Ioulia, and K. X. Karakostas. 1998. "On Optimal Sampling Designs for Autocorrelated Finite Populations." Biometrika 85 (2) (June 1): 482-486.

Penrose, R., and J. A. Todd. 2008. "A generalized inverse for matrices." Mathematical Proceedings of the Cambridge Philosophical Society 51 (October 24): 406. doi:10.1017/S0305004100030401.

Steinberg, David M., and William G. Hunter. 1984. “Experimental Design: Review and Comment.” Technometrics 26 (2) (May 1): 71-97.

Ucinski, Dariusz. 2004. Optimal Measurement Methods for Distributed Parameter System Identification. 1st ed. CRC Press, August 27.

Wu, Y. 2004. “Optimal design of a groundwater monitoring network in Daqing, China.” Environmental Geology 45 (4) (February): 527-535. doi:10.1007/s00254-003-0907-x.

Yfantis, Evangelos, George Flatman, and Joseph Behar. 1987. “Efficiency of kriging estimation for square, triangular, and hexagonal grids.” Mathematical Geology 19 (3) (April 1): 183-205.

Zhang, Yingqi, George F. Pinder, and Graciela S. Herrera. 2005. “Least cost design of groundwater quality monitoring networks.” Water Resour. Res. 41 (8): W08412.

Zhenghao Zhang, Husheng Li, and Changxing Pei. 2010. Optimum experimental design for estimating spatial variogram in cognitive radio networks. In 2010 44th Annual Conference on Information Sciences and Systems (CISS), 1-6. Princeton, NJ, USA, March. doi:10.1109/CISS.2010.5464833. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5464833.

Zhu, Z, and M Stein. 2005. “Spatial sampling design for parameter estimation of the covariance function.” Journal of Statistical Planning and Inference 134 (2) (October 1): 583-603.

Zidek, James, and Dale Zimmerman. 2010. Monitoring Network Design. In Handbook of Spatial Statistics, ed. Alan Gelfand, Peter Diggle, Montserrat Fuentes, and Peter Guttorp, 20103158:131-148. CRC Press, March 19. http://www.crcnetbase.com/doi/abs/10.1201/9781420072884-c10.

Zimmerman, Dale L. 2006. “Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction.” Environmetrics 17 (6): 635-652.

6 CONCLUSIONS AND RECOMMENDATIONS

Specific conclusions about each chapter are described in chapters 2, 3, 4 and 5. This chapter outlines some remarks and thoughts that were not explicitly mentioned, this is in addition to some recommendation for future work.

6.1 Expansion from Local Scale to Regional Scale models

The study herein was conducted on a relatively small field, which can be seen as representative of conditions in the whole Lower Arkansas Basin in Colorado. However, the expansion of the study to the regional scale is not a linear transition; and some major concerns should be tackled beforehand.

In regional models, the uncertainty of input parameters will not be only limited to errors in the absolute value of the measurements, but also include the scale effect of parameters. It is widely recognized that soil properties generally depend on the support volume of the measurement experiment. For example, for the same soil type, it was found that dispersivity changes by orders of magnitude as the scale of measurement change. This poses a challenge for modelers in terms of interpreting the available measurements. Moreover, large numerical cell sizes are usually needed to reduce the computations required by regional models; and upscaling measurements to represent the whole numerical cell should be addressed.

The conceptual uncertainty resulting from wrongly adopting an optimistic simplification of reality might have a detrimental effect on the accuracy of predictions made by models. Usually such errors produce systematically biased predictions. Developing several conceptual models for the same system is one of the options that modelers should consider.

6.2 Future of Numerical Modeling

Advancements in numerical computation capabilities in terms of hardware, e.g. multi-core PCs and cloud computing, and in terms of software, e.g. parallel programming; open the door wide for a new era of numerical simulations in the field of hydrology. An example of such efforts is the ParFlow project by Colorado School of Mines (Kollet et al. 2006); however, it is too early to talk about practical employment of these models.

In light of these advancements, integrating multi-disciplinary models, such as groundwater models, surface water models, atmospheric models, among others, seems to be a possible task in the near future.

6.3 Future of Data Collection

Use of the new computationally efficient models is of no value if the resolution of the available data does not match the high capacity of the models. As a result, a wide scale and comprehensive characterization of watershed (surface and subsurface systems) parameters is a necessity that is far from being achieved, especially for subsurface parameters. For example, the classical way of measuring hydraulic

conductivity is through a pumping test or slug test, which yields a local measurement of the wide scale process; and it is expensive to conduct sufficient numbers of them. The hope is in a new geophysical technology that might thoroughly image the subsurface system in a short time and for a reasonable price. Assimilation and inversion of seismic data, geoelectric methods, and subsurface electromagnetic methods might be a promising topic of the research in the effort to revolutionize aquifer characterization in the future.

6.4 Decision Making under Uncertainty

Usually, the uncertainty analysis of a system prediction is reported to decision makers in the form of a statistical distribution of the response, which might be wide and non-informative. This undoubtedly makes the decision making process a challenge. A different way of seeing the decision process can be illustrated as follows: assuming that the decision making process can be conceptualize as a game theory problem, and assuming that the modeler and the decision maker are two separate players, then it is the goal of the modeler to provide the decision maker with a wide range predictions of the system response to avoid any future blame. As an extreme example, if the goal is to determine a contamination concentration at a point, then it is very safe for the modeler to report a range of 0 to infinity, which of course will be of no help to the decision maker. While narrowing this range from the modeler perspective means collecting more data about the field, the decision maker can see this as unacceptable increase on a limited budget. In reality, the interplay between the technical and the political aspects of the problem could be extremely complex, which, in turn, makes the decision making

a challenging process. A decision support system that acknowledges the diverse interests of players, and that encourages players to adopt a positive cooperation scenario of the game and share the risk, cab be an important component of decision-making process under uncertainty.

6.5 Other Options for Drainage System Design

The subsurface drainage system is redesigned herein using the design-simulation approach. Different drainpipes layout (figures 6.2 to 6.6) and drain depths are investigated, and for each the average groundwater depth and the drainage effluent rate are reported (Table 6.1). Changing the number of drains and their layout did not result in significant change in the average groundwater depth because the most important part of the drain is the one that first intercept the lateral flow (adjacent to the southern boundary).

In reality, however, other factors make the design options limited to few ones. For example, in field 17 case the elevation of manhole outlet must be the lowest point in the system. This of course limits our ability to change the drain depth. The previous design (Figure 6.7) shows that slope of 0.2% (20cm in 100m) are the smallest that could be achieved. Small slopes are typically difficult to achieve in field due the flexibility of the drainpipe.

Deeper drains usually results in larger outflow rates due to high groundwater head on top of the drainpipes. This might raise some environmental concerns, especially about the disposal of the large volumes of saline groundwater. The major

function of the drainage system in Field 17 is to intercept the lateral saline groundwater flow. As a result, installing drainage pipes in the southern part of the field might the best option to intercept the flow and ensure low water table in the remaining of the field.

Table 6. 1 : Average groundwater depth, standard deviation, drainage outflow rate, and depth of drain for each design option

Design	Average GW Depth(m)	Standard Deviation of GW Depth (m)	Drainage outflow Flow Rate (m/day)	Depth of Drain Pipe (m)
A	1.98	0.313	605.6	3
B	1.95	0.317	620.5	3
C	1.94	0.317	595.3	3
D	1.86	0.321	426.8	2.5
E	1.52	0.218	216.2	2

Depth of drainage Pipe are measured from Ground surface level in the southwest Corner

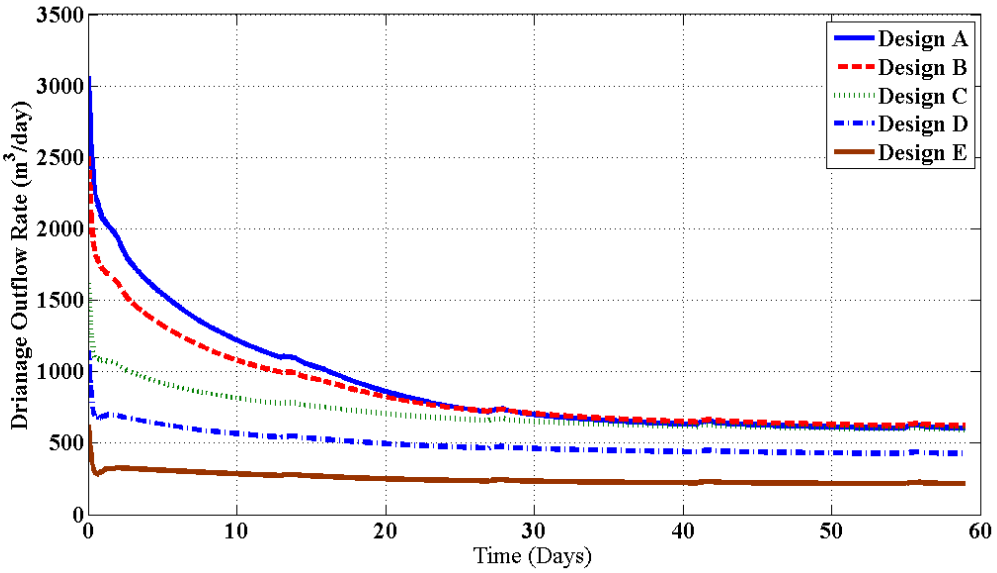

Figure 6. 1 : Drainage Outflow for Different Design Options

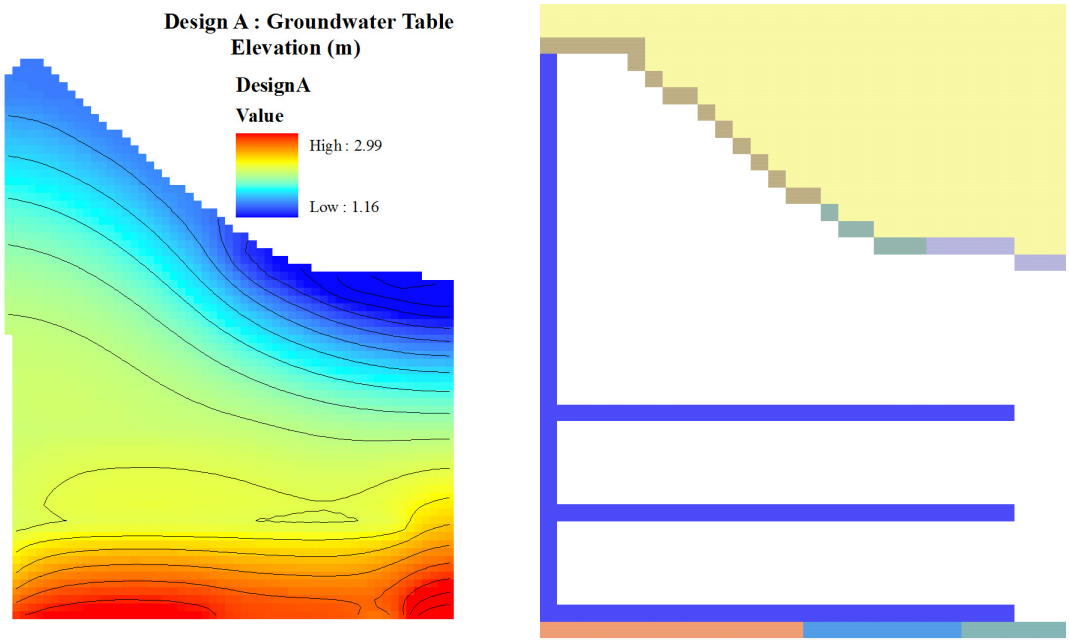

Figure 6. 2 : Layout of Design (A) and the resulting groundwater elevations

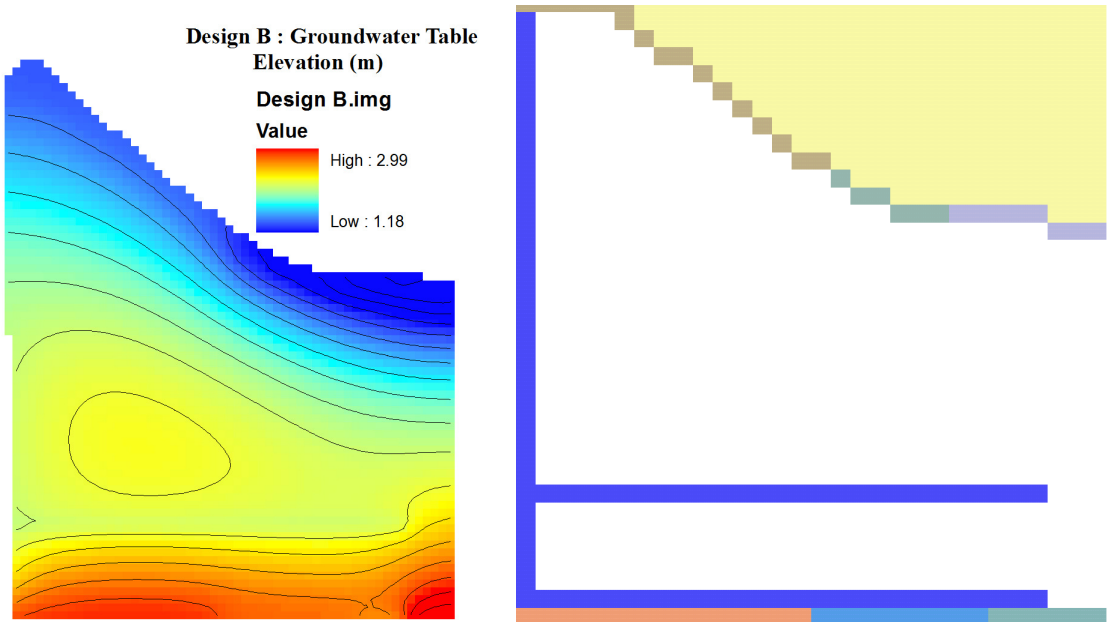

Figure 6. 3 : Layout of Design (B) and the resulting groundwater elevations

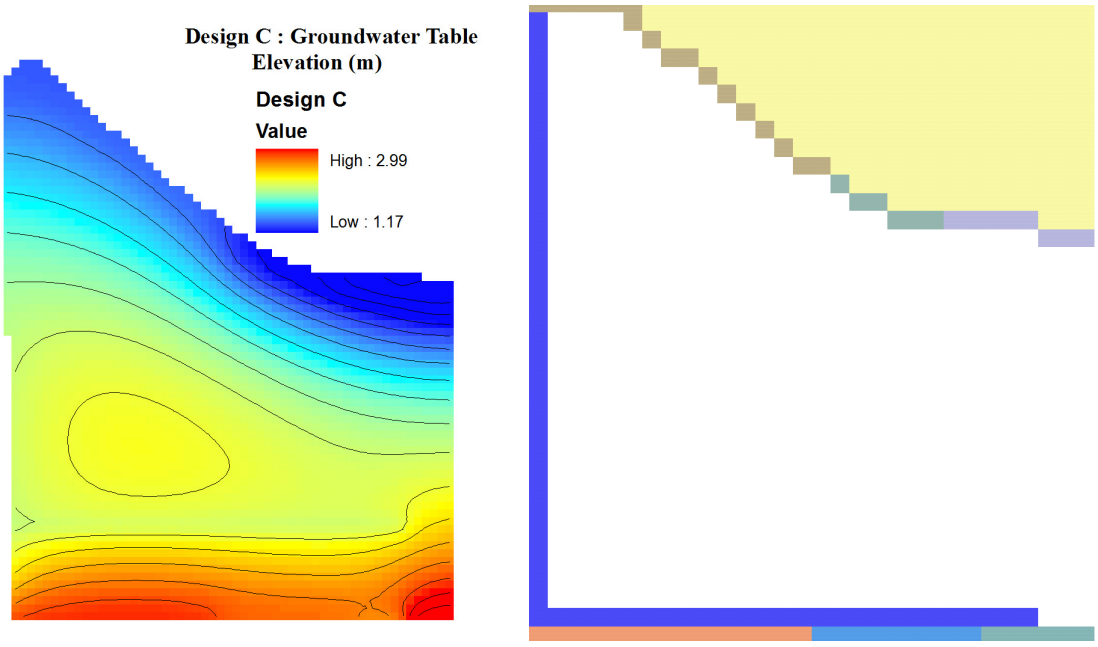

Figure 6. 4 : Layout of Design (C) and the resulting groundwater elevations

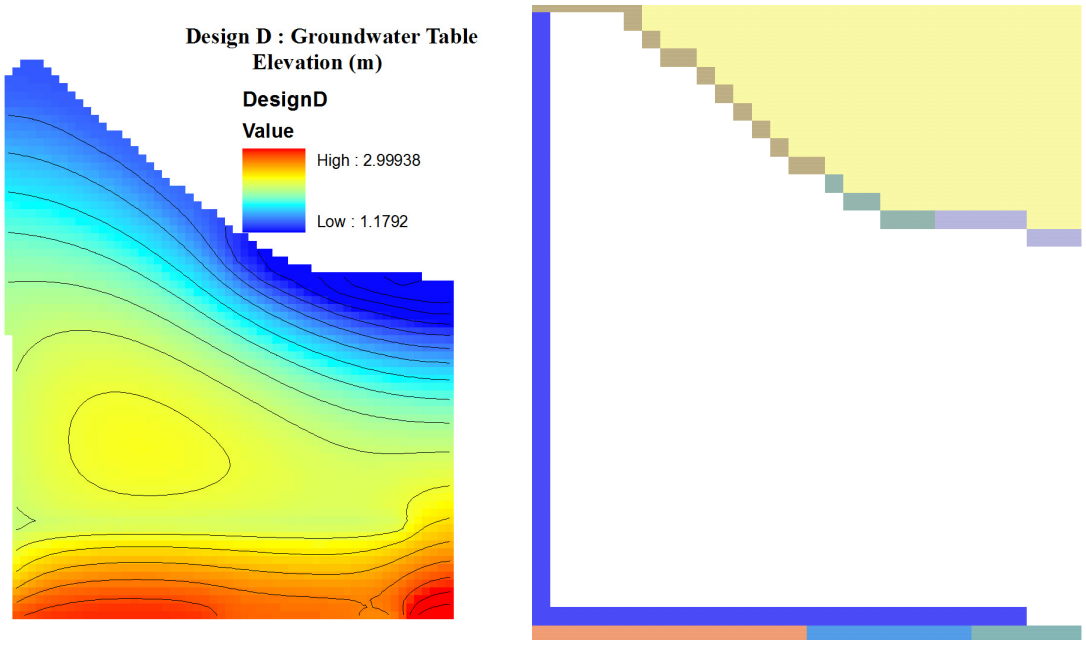

Figure 6. 5 : Layout of Design (D) and the resulting groundwater elevations

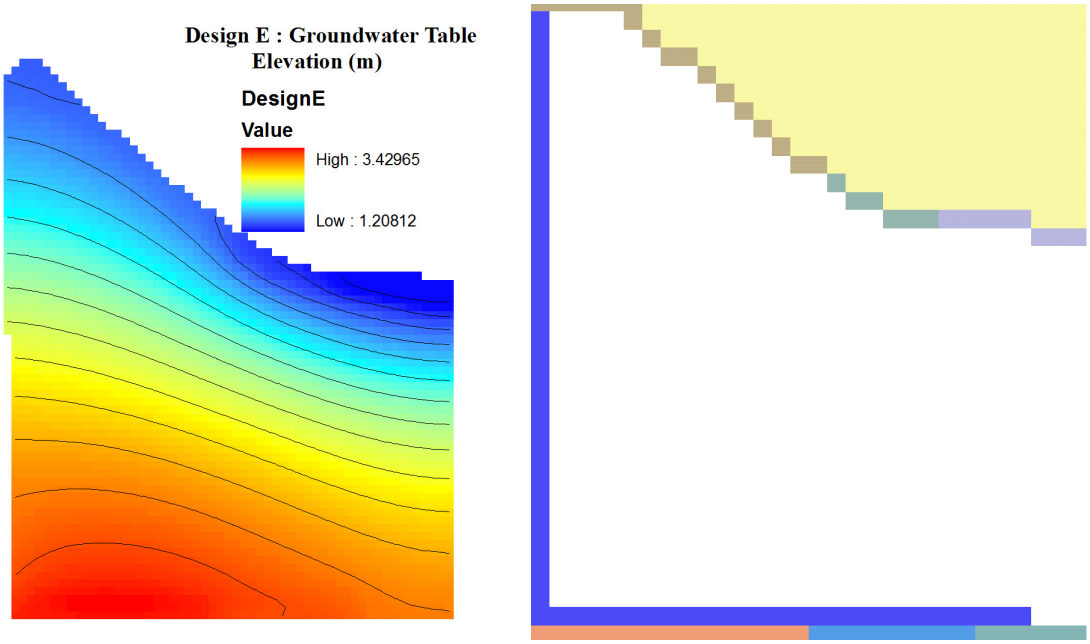

Figure 6. 6 : Layout of Design (E) and the resulting groundwater elevations

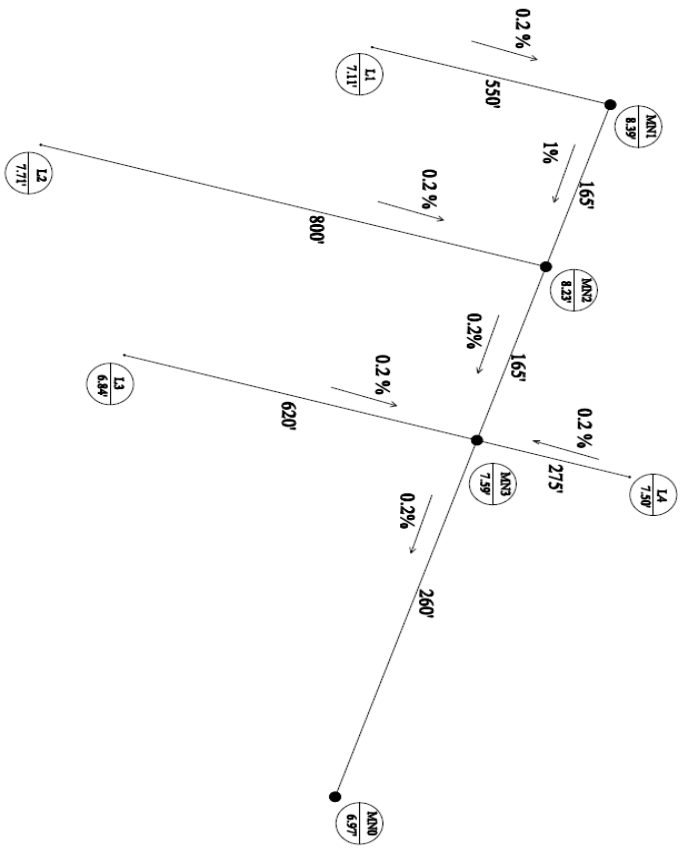

Figure 6. 7 : Actual proposed design drainage system showing pipe slopes

6.6 References

S. J. Kollet and R. M. Maxwell (2006), Integrated surface-groundwater flow modeling: A free-surface overland flow boundary in condition a parallel groundwater flow model, Advances in Water Resources, (29)7, 945-958

INDEX

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y

Z