USING NICHED CO-EVOLUTION STRATEGIES TO ADDRESS NON-UNIQUENESS IN CHARACTERIZING SOURCES OF CONTAMINATION IN A WATER DISTRIBUTION SYSTEM

A Thesis

by

KRISTEN LEIGH DRAKE

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

MASTER OF SCIENCE

August 2011

Major Subject: Civil Engineering

Using Niched Co-Evolution Strategies to Address Non-Uniqueness in Characterizing Sources of Contamination in a Water Distribution System

USING NICHED CO-EVOLUTION STRATEGIES TO ADDRESS NON-UNIQUENESS IN CHARACTERIZING SOURCES OF CONTAMINATION IN A WATER DISTRIBUTION SYSTEM

A Thesis

by

KRISTEN LEIGH DRAKE

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

MASTER OF SCIENCE

Approved by:

Co-Chairs of Committee,	Emily Zechman
	James Kelly Brumbelow
Committee Member,	Justin Yates
Head of Department,	John Niedzwecki

August 2011

Major Subject: Civil Engineering

ABSTRACT

Using Niched Co-Evolution Strategies to Address Non-Uniqueness in Characterizing Sources of Contamination in a Water Distribution System. (August 2011)

Kristen Leigh Drake, B.S., Texas A&M University

Co-Chairs of Advisory Committee: Dr. Emily Zechman
Dr. James Kelly Brumbelow

Threat management of water distribution systems is essential for protecting consumers. In a contamination event, different strategies may be implemented to protect public health, including flushing the system through opening hydrants or isolating the contaminant by manipulating valves. To select the most effective options for responding to a contamination threat, the location and loading profile of the source of the contaminant should be considered. These characteristics can be identified by utilizing water quality data from sensors that have been strategically placed in a water distribution system. A simulation-optimization approach is described here to solve the inverse problem of source characterization, by coupling an evolutionary computation-based search with a water distribution system model. The solution of this problem may reveal, however, that a set of non-unique sources exists, where sources with significantly different locations and loading patterns produce similar concentration profiles at sensors. The problem of non-uniqueness should be addressed to prevent the misidentification of a contaminant source and improve response planning. This paper aims to address the problem of non-uniqueness through the use of Niched Co-Evolution Strategies (NCES).

NCES is an evolutionary algorithm designed to identify a specified number of alternative solutions that are maximally different in their decision vectors, which are source characteristics for the water distribution problem. NCES is applied to determine the extent of non-uniqueness in source characterization for a virtual city, Mesopolis, with a population of approximately 150,000 residents. Results indicate that NCES successfully identifies non-uniqueness in source characterization and provides alternative sources of contamination. The solutions found by NCES assist in making decisions about response actions. Once alternative sources are identified, each source can be modeled to determine where the vulnerable areas of the system are, indicating the areas where response actions should be implemented.

ACKNOWLEDGEMENTS

I would like to thank my committee co-chairs, Dr. Emily Zechman and Dr. Kelly Brumbelow, and committee member, Dr. Justin Yates, for their guidance and support throughout the course of this research.

I would also like to thank my friends and colleagues in the water resources engineering division and the department faculty and staff for making my time at Texas A&M University a great experience. I also want to extend my gratitude to the Department of Civil Engineering at North Carolina State University, which allowed me to utilize the Neptune computer cluster for my research.

Finally, I would like to thank my family for their encouragement and support and my husband for his patience and love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Water Distribution Systems (WDS) are comprised of several components: pipes, storage tanks, pump stations, water treatment plants, and pipes. Due to the wide range of access points, WDS are considered to be vulnerable to accidental outbreaks of bacteria and intentional injection of harmful contaminants. A contaminant can enter the system through any one of the components if precautions have not been taken to ensure their security. Since WDS provide communities with clean drinking water, it is essential to protect the public should a contamination event occur. During a contamination event, decision makers, such as city or utility managers, should select effective mitigation strategies to protect public health. Several measures can be taken to minimize consequences if a contaminant is introduced to the system: open fire hydrants to release water; close or open valves to isolate the contaminant; increase chlorine concentration in water; and notify consumers to stop or limit usage of water. Knowledge about the location and timing of the source of contamination can provide the necessary insight to select the most effective decision for implementing response actions. The source characterization can be performed using data from sensor networks placed in the WDS, which provide information about the contaminant as it moves through the system, including the observed contamination profile data.

Data observed from a sensor network can be used to solve for the initial characteristics of a contamination event by posing source identification as an inverse

---

This thesis follows the style of *Journal of Water Resources Planning and Management.*

problem. The inverse problem can be solved by coupling a simulation model with an optimization method; however, due to the ill-posed nature of inverse problems, a set of solutions may exist that match the observed data, while demonstrating significantly different source characteristics. This issue, known as non-uniqueness, may occur due to lack of sufficient data or the presence of error in the data. If not properly addressed, non-uniqueness in a source identification problem may lead to faulty identification of the location of the source, and response actions that are based on misidentifications may fail to protect public health. While optimization methodologies have been developed to solve the source identification problem, the issue of non-uniqueness has been addressed to only a limited extent. The goal of this research is to develop a method that addresses non-uniqueness in source identification for water distribution contamination events through refinement and application of an evolutionary computation-based method, Niched Co-Evolution Strategies (NCES).

2. PROBLEM

2.1 Water Distribution System Security

Water systems are classified as critical infrastructure by the Department of Homeland Security (DHS). Critical infrastructure is a sector that is vulnerable to attacks that can lead to a large amount of illnesses and casualties and/or disruptions in critical services to the public. Because water systems are considered to be critical, there should be extensive planning and preparation on behalf of the government or managing entities to protect the public. The Public Health Security and Bioterrorism Preparedness and Response Act of 2002 (Bioterrorism Act of 2002) outlines the requirements for preparing for water related terrorism. Through this act, the United States Environment Protection Agency (EPA) is required "to conduct assessments of their vulnerabilities to terrorist attack or other intentional acts and to defend against adversarial actions that might substantially disrupt the ability of a system to provide a safe and reliable supply of drinking water" for water systems with 3,300 or more customers (Public Health Security 2002). There are two main types of water security events: accidental and intentional. Accidental events include biological/parasitic outbreaks, chemical spills, and natural disasters. In 1993, the city of Milwaukee experienced a cryptosporidium outbreak in the water distribution system after heavy rains. The contaminant caused over 400,000 people to become ill and 104 deaths (Mac Kenzie et al. 2004). Intentional attacks can be

in the form of biological/chemical contamination, damage to physical infrastructure, and computer system attack.

Throughout history, various forms of biological and chemical attacks have been witnessed during times of war, such as poisoning water wells. In Rome in 2002, a group of terrorists planned to introduce cyanide to the city's WDS during a festival. Fortunately, the attack was foiled by security personnel before anyone was harmed. Damage to physical infrastructure, through, for example, bombing a water treatment plant, pump station, or water storage facility, would greatly impair a city's ability to deliver safe drinking water for consumption and fire-fighting capabilities. An attack on a computer system that controls the daily operations of a water utility is also considered to be a terrorist attack that could harm the public. It is commonly known that the terrorist group Al Qaeda considers water as an option to cause terrorism (Kroll 2006). Anthrax and cholera are among many bacteria, viruses, and bio toxins that survive in water and have the potential to harm consumers (Clark and Deininger 2000).

2.2 Source Identification

Source identification is a problem that utilizes data, often from a sensor network, to solve for the source of contamination once an event has occurred. Sensor data consists of the concentration profile (concentration over time) of a contaminant. For the source identification problem, the solution is comprised of the location of the source, the

start time of contamination, and the contaminant loading profile. The optimization problem is defined in Equation (1) as:

$$\underset{\{M,n,t_I\}}{Minimize}\ E = \sum_{i}^{N}\sum_{t}^{T}\left(c_i^{pred}(t) - c_i^{obs}(t)\right)^2 \quad (1)$$

where E is the error, $c_i^{pred}(t)$ is the predicted contaminant concentration, and $c_i^{obs}(t)$ is the observed contaminant concentration. The time step is t and i is the node index of a water quality sensor. The specified number of time steps is *T* and the number of sensors in the network is *N*. The difference between the predicted and observed contamination profile is calculated for all sensors in the system at each simulated time step. The decision variables for this problem are the contaminant loading profile, *M*; the contaminant node location, *n*; and the start time of contamination, t_I.

3. SOLUTION APPROACH

3.1 Evolutionary Computation for Source Identification

The problem of source identification can be approached as an inverse problem, where the output of a model is used to identify input parameters. One approach to solve an inverse problem is to use a simulation-optimization approach (Fig. 1).

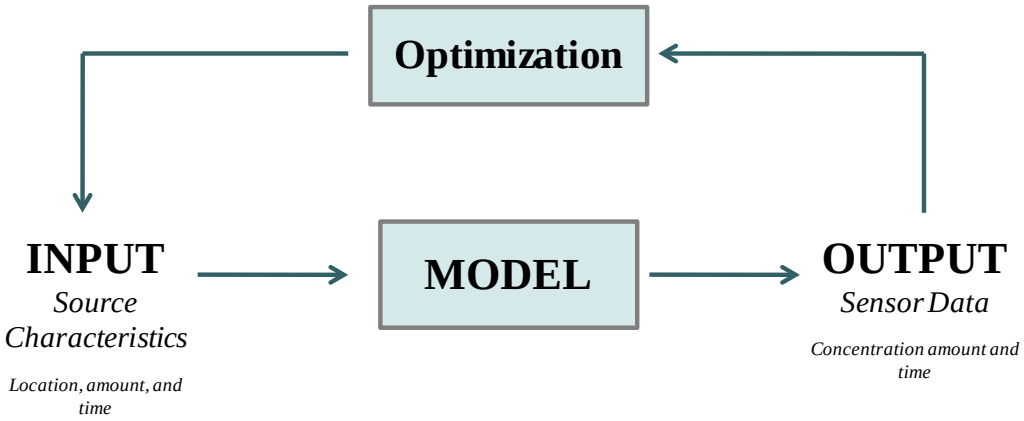

Figure 1. Source Identification as an Inverse Problem.

The problem of source identification has been addressed through both linear and non-linear programming, regression trees, logistic regression, and Evolutionary Computation (EC) methods. Van Bloemen Waanders et al. (2003) approach the source identification problem using non-linear programming and gradient based methods. They use a convection-diffusion approach to the source inversion problem and solve the problem using several different gradient based methods. Laird et al. (2005) approach the

inverse problem using an origin tracking method through non-linear programming. The goal of this methodology is to decrease the complexity of the problem, while continuing to identify realistic sources of contamination. Guan et al. (2006) describes an algorithm that utilizes simulation-optimization and a reduced gradient method (RGM) to solve the source identification problem. Their use of RGM aims to reduce the computation time of the simulation-optimization process. Guan et al. (2006) also examined the effect of error in sensor data on the algorithms ability to correctly identify a source. They found that the algorithm could still correctly identify a source, but it was not able to accurately define the release history. Preis and Ostfeld (2006) utilized tree based methods to solve the source identification problem. Model trees are calibrated by using EPANET and linear trees are constructed to solve for source characteristics. Linear programming is used on both trees to solve the inverse problem. Liu et al. (2008) presents a new method to reduce the search space in solving the source identification problem. Logistic regression is used to assign a probability of a given node being the source to each node in the system. A local search is then performed around nodes with a high probability of being a source.

The research presented in this thesis uses evolutionary computation (EC) within a simulation-optimization framework to solve the source identification problem. EC is an optimization approach based on the theory of natural selection, where a population of individuals represents potential solutions to a problem and converges to nearly global optima over repeated iteration of genetic operators, including selection and mutation (Rechenberg 1973, Schwefel 1981, and Schwefel 1995). Genes (the decision variables)

make up the individual (the solution). Evolution Strategies (ES) and Genetic Algorithms (GA) are common forms of EC, where ES uses mutation as its primary operator and GA uses crossover and mutation as operators. Preis and Ostfeld (2007) describe the use of a GA in the source identification simulation-optimization problem. The use of a GA allows for more exploration of possible solutions as opposed to non-EC methods. GAs are also better equipped to solve large network problems more efficiently.

Zechman and Ranjithan (2009) describe an ES approach to characterize sources of contamination during a WDS event. For the source identification problem, ES is the optimizer applied to minimize the difference between the simulated contaminant and the observed contaminant concentration by adjusting decision variables, which represent the location, loading profile, and timing of contaminant release. For a small WDS, a significant amount of non-uniqueness was revealed in the identification of centrally-located sources.

3.2 Modeling-to-Generate Alternatives

One approach to address non-uniqueness that is inherent in the source identification problem is the Modeling-to-Generate Alternatives (MGA) modeling methodology. MGA was first developed as a method to assist human decision making through mathematical programming (Brill 1979). Alternatives generation is beneficial to a problem containing non-uniqueness because other possible solutions are identified; therefore addressing the problem by providing multiple solutions. Alternative solutions

should have similar objective values and be maximally different in their decisions. An original optimization problem can be represented in Equations (2) and (3) as:

$$Maximize\ Z = f(X) \tag{2}$$

$$Subject\ to\ g_i(X) \geq b_i\ \ i = 1, \ldots, M \tag{3}$$

where Z is the objective function and X is a vector representing the decision variables. Equation (3) represents a set of constraints on the problem. Optimization of Equation (1) yields Z^* as the best solution and X^* as the corresponding decisions. A set of alternatives is generated using the following model, represented as Equations (4), (5), and (6):

$$Maximize\ D = d(X, X^*) \tag{4}$$

$$Subject\ to\ f(X) \geq T(Z^*) \tag{5}$$

$$Subject\ to\ g_i(X) \geq b_i\ \ i = 1, \ldots, M \tag{6}$$

where D represents the difference between decision vectors X and X^*. An allowable relaxation in the objective Z^* is represented by the target T, which allows for exploration of different decisions. Though this relaxation encourages inferior solutions, nearly optimal solutions may be considered as viable options in decision making. For the problem of non-uniqueness, MGA will identify other good, but different solutions to a problem containing non-uniqueness. The amount of difference among the alternatives

can indicate the level of non-uniqueness in the problem. For example, a large amount of difference between alternatives may indicate a high degree of non-uniqueness; while a small difference indicates that the problem may be unique.

3.3 Evolutionary Computation for Generating Alternatives

One implementation of MGA was performed using genetic algorithms to generate alternatives and was presented by Harrell and Ranjithan (2003) for a detention pond design problem. Rather than specifying a difference function to find alternatives, the problem objectives and constraints were adjusted for different scenarios to find alternative solutions. Allowing components of the problem to be flexible, such as land use, different solutions were derived with lower costs. Incorporating different approaches to a problem can lead to the identification of many different, but good, solutions.

Niching algorithms have also been used for generating alternatives. Traditional niching methods encourage solutions to be found with similar objective values but different decisions compared with an optimal solution. Three popular niching methods include clearing, crowding, and sharing. In these traditional niching methods, several parameters must be defined to guide the search (Mahfoud 1995; Singh and Deb 2006). To avoid setting parameters, other methods can be adopted to generate alternatives. A Genetic Algorithm for Modeling-to-Generate Alternatives (GAMGA) was explored by

Loughlin et al. (2001). GAMGA used a genetic algorithm with the maximal difference function to generate alternatives.

Zechman and Ranjithan (2004) presented the Evolutionary Algorithm to Generate Alternatives (EAGA) to explore complex engineering problems. EAGA uses multiple subpopulations to solve for good solutions with maximally different decisions, where each subpopulation converges to one alternative solution. EAGA operators primarily include binary tournament selection, mutation, and crossover, which are executed separately within each subpopulation. Zechman and Ranjithan (2007) applied EAGA for a water resources problem to generate alternative designs.

3.4 Niched Co-Evolution Strategies

Zechman et al. (2006) extended EAGA to an ES-based implementation, Niched Co-Evolution Strategies (NCES), and applied the algorithm for source identification in groundwater pollution. NCES utilizes ES optimization and the MGA modeling approach to generate alternatives. NCES uses a set of subpopulations, where the first subpopulation searches for the best solution to the original optimization problem, and secondary subpopulations search for maximally different alternatives. The relaxation used for deriving alternative solutions is based on the fitness of the best solution in the first subpopulation. Secondary subpopulations are guided to different areas of the decision space by a selection mechanism that encourages solutions based on a difference

function and satisfaction of the objective function target. NCES uses the same evolutionary operators for all subpopulations in the search for alternatives.

3.4.1 Difference Calculation

The difference function is based on the location of each subpopulation and how close one subpopulation is to all subpopulations. The difference function for each individual is calculated as the minimum Euclidean distance to the centroids of all other subpopulations. The difference function ($D^{k,p}$) represents the centroid calculation and is defined in Equations (7) and (8) as:

$$D^{k,p} = Min\left\{\frac{\sum_{j=1}^{K} d\left(X^{k,p}, X^{j,q}\right)}{K}; q = 1, \dots, P, q \neq p\right\} \quad (7)$$

$$Euclidian\ Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (8)$$

where $d\left(X^{k,p}, X^{j,q}\right)$ is the Euclidian distance between the centroids of two subpopulations $X^{k,p}$ and $X^{j,q}$, K is the number of individuals in a subpopulation, and P is the number of subpopulations (Zechman et al. 2006). The Euclidean distance in Equation (8) is defined as the distance between two points, (x_1,y_1) and (x_2,y_2).

3.4.2 Search Operators

NCES utilizes two main search operators: mutation and selection. These operators search the decision space for different solutions that meet the constraints of the optimization problem. The mutation operator makes changes in the decisions variables based on probability and produces λ new individuals in each subpopulation. The mutation is performed on the genes by randomly sampling from a normal distribution based on the current values of the genes, where the mean is the current value. The standard deviation is determined for all subpopulations and is set as the mutation parameter. For NCES, the mutation operator is adaptive. In adaptive mutation, the standard deviation is determined while the search is occurring and is adjusted using a separate normal distribution to mutate the mutation parameter.

The main selection operators are ranking selection and Elitist Graduated Over-Selection (EGOS) (Fernandez and Evett 1997). Ranking involves sorting the solutions from best to worst based on their fitness values. The ranking process in NCES is different depending on the subpopulation. In the first subpopulation, selection is based solely on fitness values. Individuals are ranked from best fitness to worst fitness and the best μ individuals are selected to survive for the next generation. In subsequent populations, selection depends on feasibility. Feasible individuals meet the target value set by the first subpopulation. The feasible individuals are first ranked from maximally different to least different. Infeasible individuals are ranked from most different to least

different and are ranked below all feasible individuals. The best μ individuals survive to the next generation (Schwefel 1995).

EGOS is used to increase the chance that highly fit solutions are selected in each subpopulation. The step size (upper quantile) is set by the user at the beginning of the search. After individuals are ranked, a set of individuals are placed into a pool of candidates. The size of the pool is specified as the upper quantile. Each individual in this pool has an equal probability to be selected, and one individual is randomly selected to survive to the next generation. The solution that is selected is placed back into the pool, and the size of the pool is increased by adding the next individual from the ranked list. Individuals are selected from the increasing pool of candidates until the population for next generation has been developed. This increases the selection pressure because highly ranked individuals have more opportunities to be selected than poorly ranked individuals, and duplicates of highly fit solutions are likely to be copied into the next generation.

3.4.3 Algorithmic Steps

The algorithmic steps as defined by Zechman et al. (2006) are listed below:

Step 1. Initialize a population with P subpopulations, each of size μ, where P is the number of alternatives the algorithm is searching for and μ is the number of individuals in each subpopulation. Each subpopulation is represented by the index SP_p $(p=1,\ldots,P)$.

SP_1, the first subpopulation, searches for the best solution with respect to the objective function. $SP_{p\neq1}$, the subsequent subpopulations, search for alternative solutions.

Step 2. Apply adaptive mutation to all subpopulations, yielding λ new individuals in each subpopulation.

Step 3. Evaluate the fitness of each individual in $\mu + \lambda$ in the first subpopulation and select the best individual with respect to fitness using Equation (2). The fitness of the best individual is relaxed by the target T for the generation of alternatives in the subsequent subpopulations using Equation (5).

Step 4. In the subsequent subpopulations ($SP_{p\neq1}$), evaluate the fitness of each individual in $\mu + \lambda$. The fitness of each individual is designated as feasible if it meets the target constraint in Equation (5). Individuals not meeting the target constraint are designated as infeasible.

Step 5. Calculate the difference $D^{k,p}$ for all individuals in the subsequent populations ($SP_{p\neq1}$) using Equation (7).

Step 6. Apply the ranking and EGOS selection to all subpopulations.

Step 7. Check termination criteria (e.g. number of generations). If met, stop. Otherwise go to Step 2 for the next generation.

4. CASE STUDY: MESOPOLIS

4.1 Virtual City of Mesopolis

The virtual city of Mesopolis was designed by Brumbelow et al. (2007) to model realistic events in WDS without compromising the water security of actual cities. Mesopolis is a city of approximately 150,000 residents and includes a naval base, university, urban and suburban housing throughout the system, an industrial area, an airport, and commercial areas. Sources of contamination were chosen based on a vulnerability analysis by Zechman et al. (2011). An original source is required to acquire the sensor data needed to use NCES for the identification of sources. Fig. 2 shows the locations of the sources of contamination tested for NCES. A conservative contaminant was placed in the system at hour seven with a load of 60 mg/min for three hours, as shown in the loading profile in Fig. 2. The total simulation time for these contamination events was 72 hours; however, only the first 24 hours of data is shown (remaining data provides no important information). All simulations were performed using EPANET, a modeling package provided by the EPA to simulation events in a WDS (Rossman 2000).

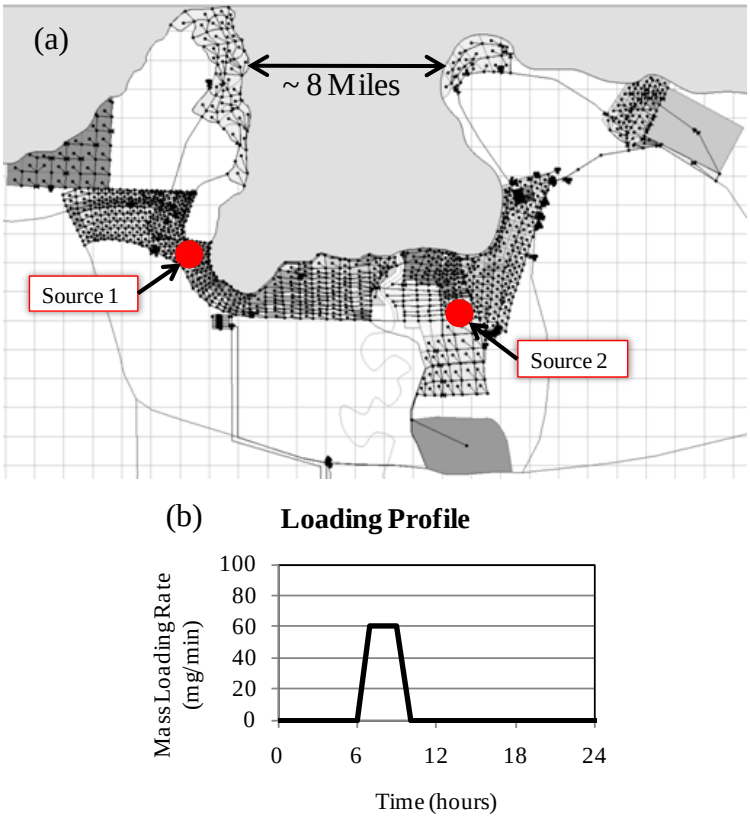

Figure 2. (a) Source Locations in Mesopolis with (b) Contaminant Loading Profile.

4.2 Sensor Network Design

Several different water quality sensor networks were designed for Mesopolis. The number of sensors varied from three to ten and sensors were strategically placed in the system. For the three sensor network (Fig. 3a), the sensors were placed along water mains, tanks, and pump stations. These high flow areas were selected because of their ability to provide large amounts of water to populated areas of Mesopolis. Other sensors were added to the initial network to provide additional coverage of vulnerable areas. Fig. 3b shows the five sensor network designed for Mesopolis.

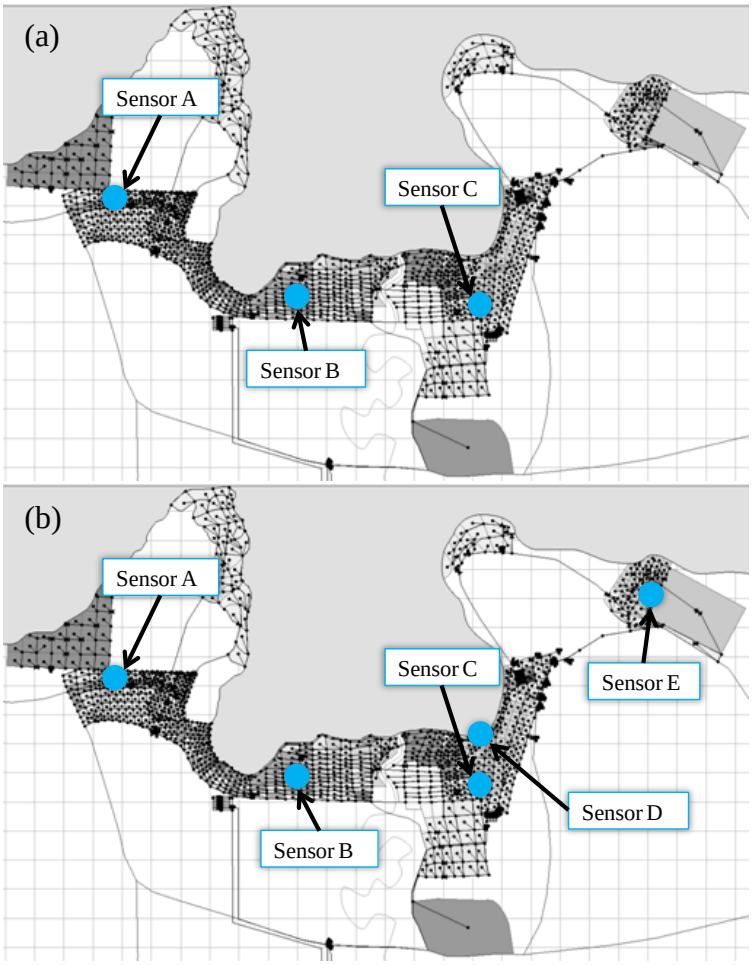

Figure 3. Sensor Networks in Mesopolis. (a) Three Sensors and (b) Five Sensors.

4.3 Results and Discussion

NCES was tested for three source and sensor network ensembles: source 1 and sensor network ABC, source 1 and sensor network ABCDE, and source 2 and sensor network ABC. The performance of the algorithm and the results of each ensemble tested will be discussed in the following sections.

4.3.1 Algorithm Performance

The algorithmic parameters used when testing NCES for Mesopolis are shown in Table 1. These values were selected after many trials using different values. The population size is fairly large due to the large amount of potential solutions in Mesopolis. Terminal nodes within the WDS were not considered; only intermediate nodes were considered in the search for sources of contamination. For Mesopolis, NCES searched for three alternatives and therefore used three subpopulations. Fig. 4 shows three graphs representing the convergence of the objective function (error value) while NCES was operating. The convergence shows that the first alternative (the optimal solution) experienced small changes as it converged to the best solution, which is to be expected by using the adaptive mutation. The subsequent alternatives also followed typical ES convergence. The movements of the average objective, shown in red in Fig 4., exist due to changes in feasibility cause strong mutations. Once the majority of the subpopulation becomes feasible (they meet the target value), the mutation operator changes the individuals to maximize the distance between the other subpopulations. The convergence graphs show that NCES is co-evolving by making small changes based on the performance of each of the alternatives.

Table 1. NCES Parameters for Mesopolis

Algorithmic Parameter	Value
Number of Subpopulations	3
Number of Generations	300
Population Size μ	400
Mutated Individuals λ	400
Step Size (Upper Quantile)	80
Target T	1.5

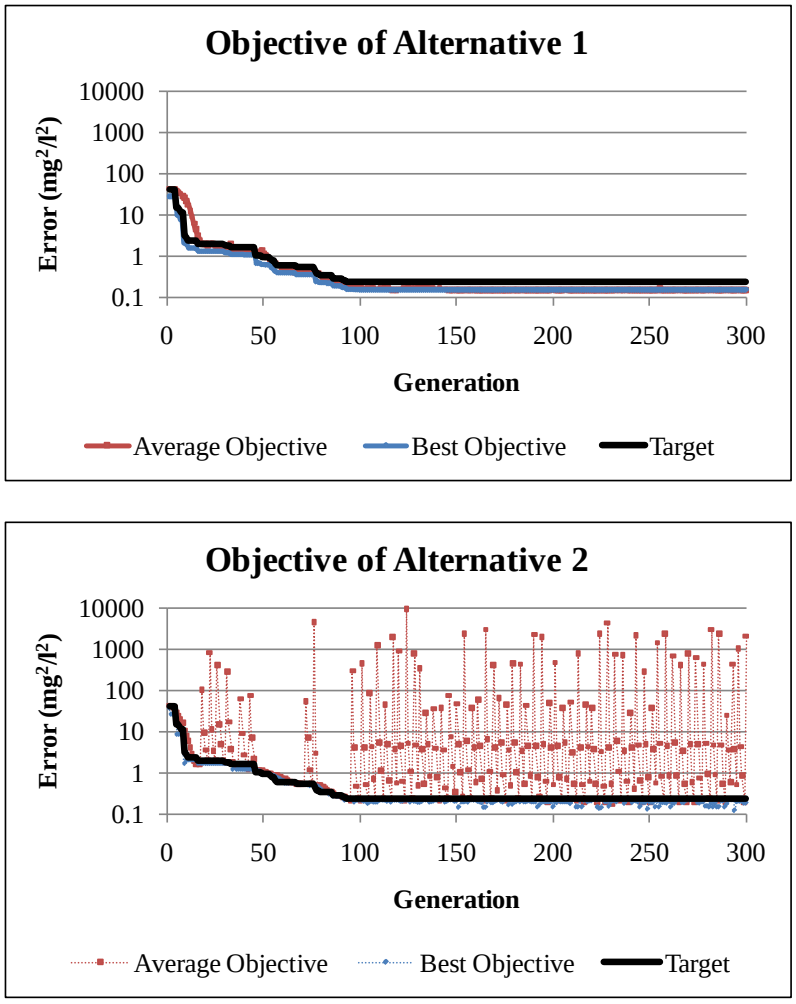

Figure 4. Objective Function Convergence for Source 2 and Sensor Network ABC.

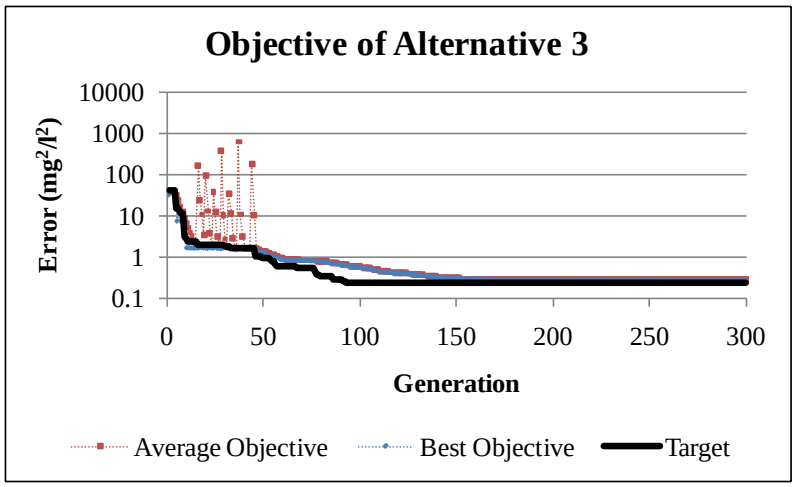

Figure 4 continued.

NCES uses a Java framework coupled with the WDS modeling package, EPANET. The parameters shown in Table 1 and the computation time of approximately 83 days on an average desktop computer reflect the need to have a large number of individuals generated to search Mesopolis. To reduce the computation time, NCES was executed on a computer cluster containing eleven nodes with two 2.2 GHz processors, four GB RAM, and 80 GB HD per node (Mahinthakumar et al. 2006). Using parallelized versions of the Java framework and EPANET, the computation time for one optimization trial using the settings in Table 1 was reduced to approximately seven hours.

Table 2 shows a summary of the results found by NCES for three different source and sensor network combinations in Mesopolis. Each scenario was executed for 20 trials. The algorithm found results that yielded low error values for all combinations. Good alternatives are identified as solutions with non-zero readings at the sensor that observes non-zero concentration values. Solutions are labeled “good” though the

concentration profiles do not exactly match the observations; good solutions can be further ranked based on the value of the error function. The error value must be evaluated while looking at the different between the concentration profiles to determine if the error value is good or not. In general, good sets of solutions matched concentration profiles relatively well. In comparing the scenarios with Source 1 and sensor networks ABC and ABCDE, there is little difference in the error value and the number of trials that identified good alternatives. Even though two sensors were added to network ABC, these new sensors did not provide any new information and therefore did not improve the search for good alternatives. In comparing Source 1 and Source 2 with the same sensor network ABC, there is a difference in the error value and the number of trials that identified good alternatives. Source 2 error values were lower on average and more good alternatives were identified. This is most likely due to Source 2's proximity to Sensor C, the only sensor that received non-zero concentration values for Source 2. The distance among the alternatives, however, is smaller than the distance among the alternatives found for Source 1. The following sub-sections further investigate the results of the three scenarios.

Table 2. Summary of NCES Results as Demonstrated for Mesopolis

Ensembles/Parameters	**Source 1 and Sensor Network ABC**	**Source 1 and Sensor Network ABCDE**	**Source 2 and Sensor Network ABC**
Average Error (mg^2/L^2)	1.2	1.5	0.5
Number of Trials that Identified Good Alternatives	9	9	13
Average Distance Between Alternatives in Best Trial (feet)	5,175	13,478	4,471

4.3.2 Source 1 and Sensor Network ABC

Source 1 was placed on the western side of Mesopolis along a high flow water main. Nine trials, of the 20 tested, identified three good alternatives. A good alternative is defined as a non-zero reading at a minimum of one of the sensors in the network. This means that the alternative sources identified did contain a loading profile that yielded sensor data. The average error for all 20 trials was 1.2 mg^2/L^2. Figs. 5 and 6 show the location and loading profile, with ensuing sensor data, for each alternative in the best trial of the 20 trials. Sensor A is the only sensor that received data and is therefore the only concentration profile shown; sensors B and C had zero concentration values. The average Euclidian distance between the alternatives is 5,175 feet. The error values are 0.0, 1.7, and 0.8 mg^2/L^2 for alternatives 1, 2, and 3, respectively.

For this representative solution, the first alternative correctly identified Source 1 as the true source of contamination. Two other alternatives were successfully generated with similar error values and different locations. The loading profiles in Fig. 6 show minimal difference in the start time and duration and a large difference in the amount of contaminant entering the system. The concentration profiles at Sensor A show the observed sensor data in black marks and the predicted sensor data in solid, colored lines. The sensor data matches exactly for the first alternative, while the sensor data closely matches for the subsequent alternatives. The differences in the decision variables (location and amount of loading) that yield similar objective values indicate that good alternatives were generated for this ensemble, showing non-uniqueness in Mesopolis.

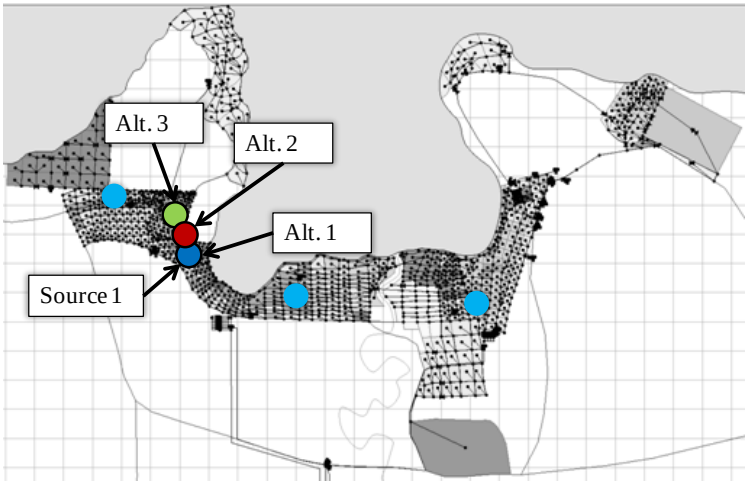

Figure 5. Location of Alternatives Found for Source 1 and Sensor Network ABC.

4.3.3 Source 1 and Sensor Network ABCDE

Similar to the first ensemble, NCES was executed for 20 trials and nine trials identified three good alternatives. The average error was 1.5 mg^2/L^2 and the only sensor that received data was Sensor A. Therefore, the addition of Sensors D and E did not provide any new information to assist in minimizing the error in the concentration profiles. The average Euclidian distance between the alternatives is 13,478 feet. Figs. 7 and 8 show the location, loading profile, and sensor data for each alternative in the best trial of the 20 trials. The error values for the best trial shown are 0.3, 0.4, and 0.4 mg^2/L^2 for alternatives 1, 2, and 3, respectively.

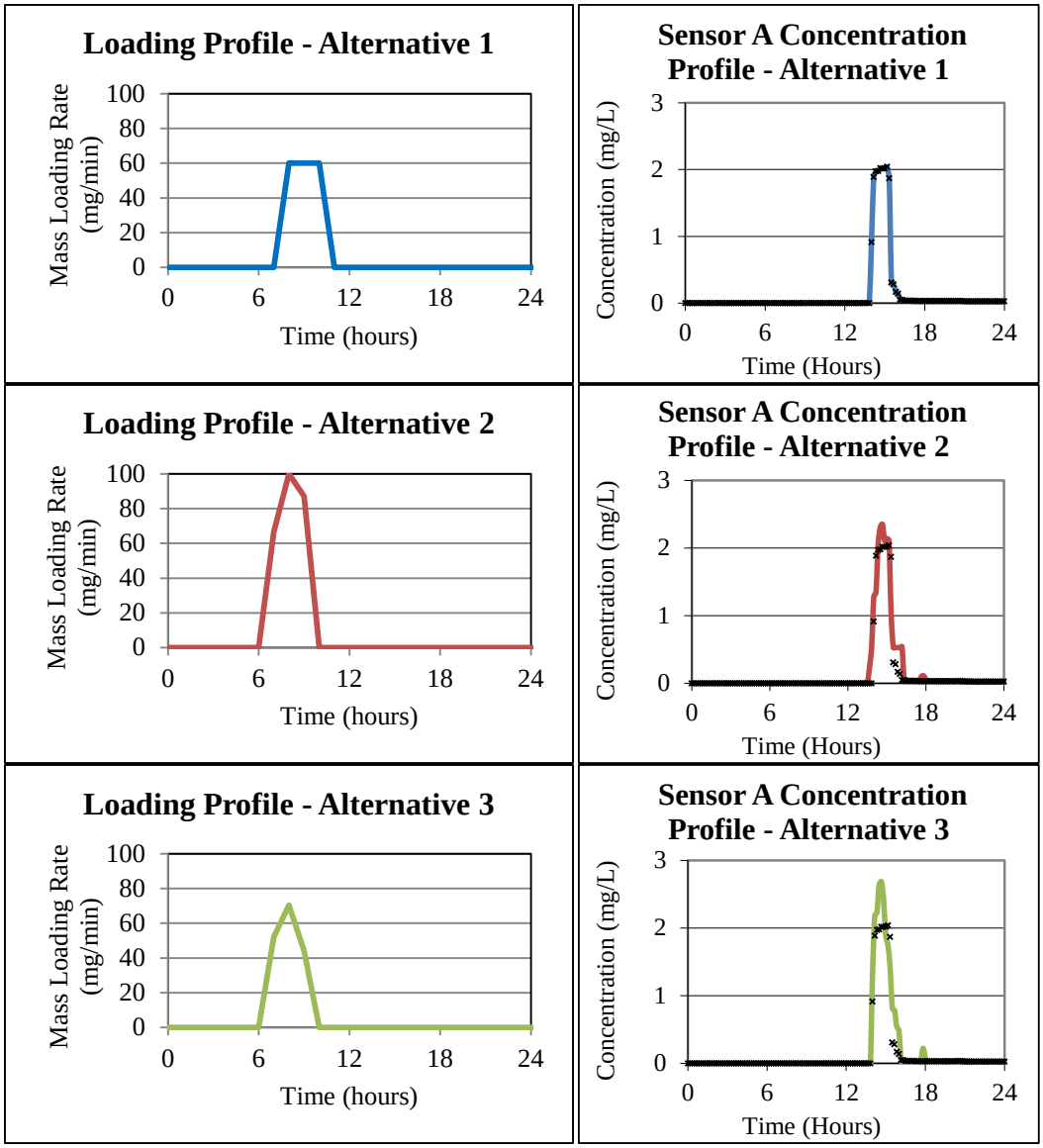

Figure 6. Loading Profiles and Sensor Concentration Profiles for Alternatives Found for Source 1 and Sensor Network ABC.

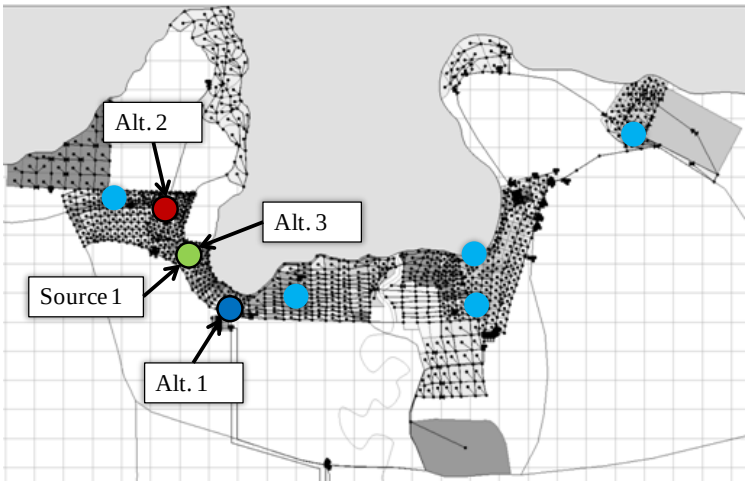

Figure 7. Location of Alternatives Found for Source 1 and Sensor Network ABCDE.

The third alternative identified the true source of contamination, Source 1, while the first and second alternatives identified different sources in their location and loading profile. The loading profiles vary among the three alternatives, with the third alternative being different from the loading profile of the true source. This difference contributes to the error value, due to a variation in the sensor data. Identical to the first ensemble, the concentration profiles at Sensor A show the observed sensor data in black marks and the predicted sensor data in solid, colored lines. The sensor data closely matches the observed sensor data for all alternatives generated, yielding low error values. Again similar to the first ensemble, good alternatives were generated as indicated by the large difference in the decision variables with low error values, suggesting non-uniqueness was present in the system.

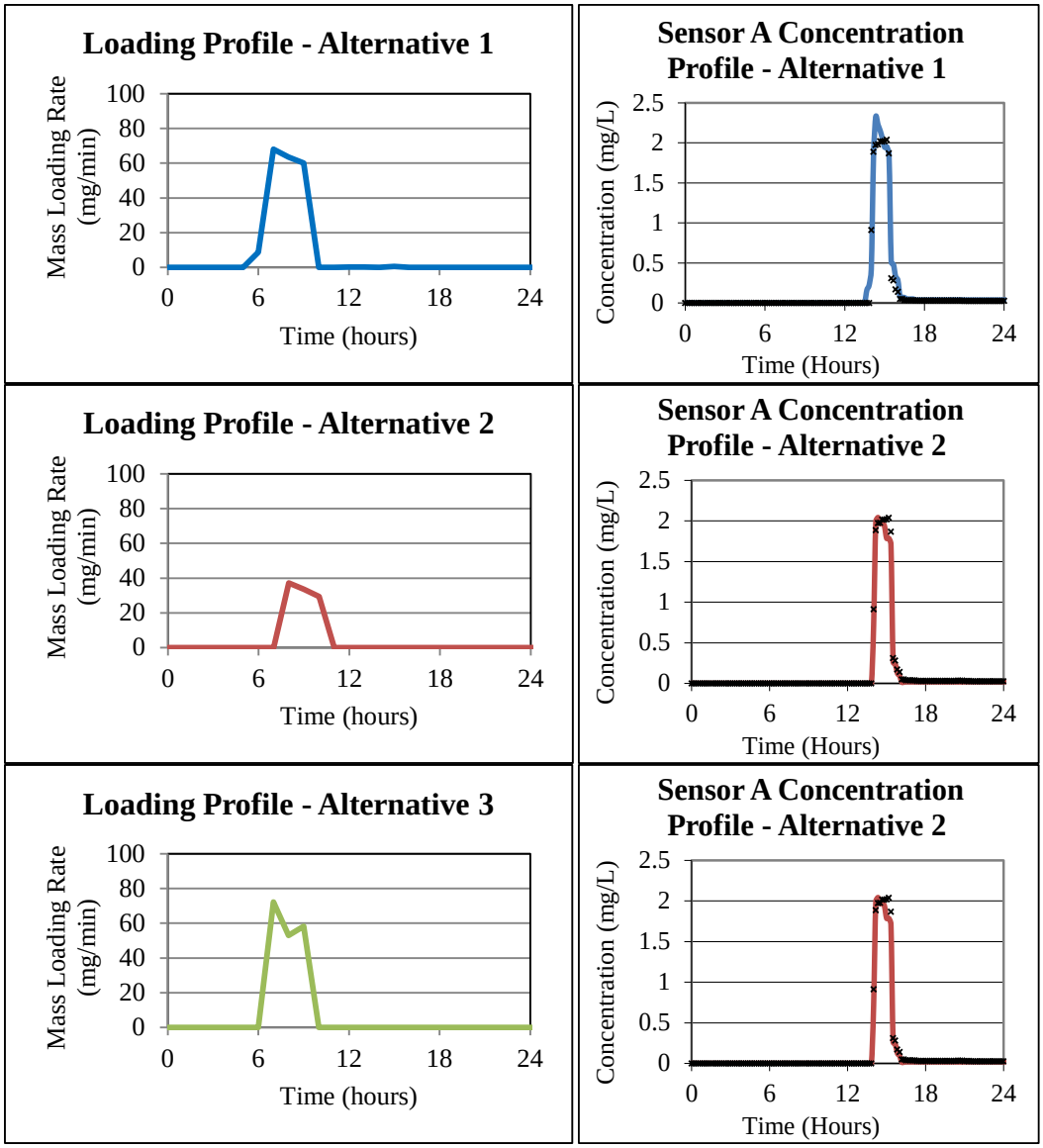

Figure 8. Loading Profiles and Sensor Concentration Profiles for Alternatives Found for Source 1 and Sensor Network ABCDE.

4.3.4 Source 2 and Sensor Network ABC

The third combination utilized the smaller three sensor network ABC for a source on the eastern side of Mesopolis, Source 2, and was tested for 20 trials. Thirteen

trials identified three good alternatives, where "good" means the alternative source yielded non-zero readings at a sensor. The average error was 0.5 mg^2/L^2 and the only sensor that received data was Sensor C. Sensors A and B did not receive sensor data. The average Euclidian distance between the alternatives is 4,471 feet. Figs. 9 and 10 show the location, loading profile, and sensor data for each alternative in the best trial of the 20 trials. For the best trial the error values are 0.0, 1.2, and 0.5 mg^2/L^2 for alternatives 1, 2, and 3, respectively.

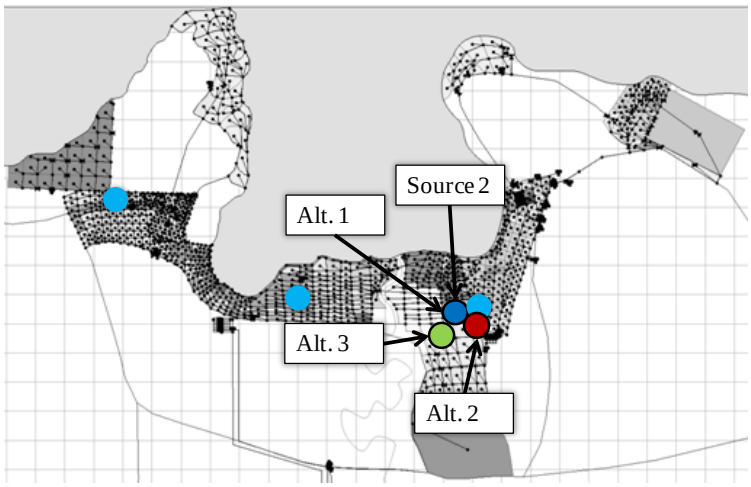

Figure 9. Location of Alternatives Found for Source 2 and Sensor Network ABC.

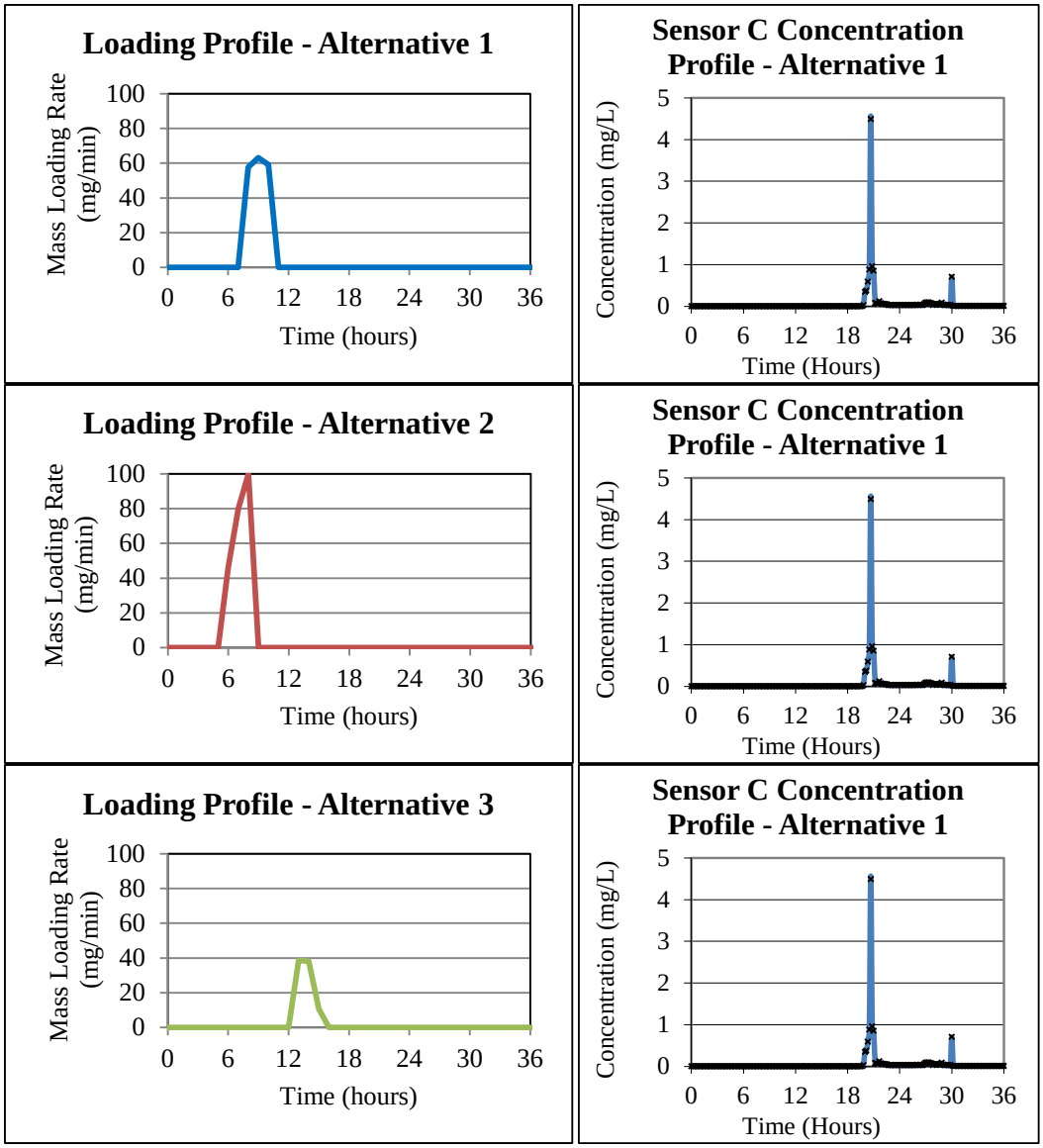

Figure 10. Loading Profiles and Sensor Concentration Profiles for Alternatives Found for Source 2 and Sensor Network ABC.

The first alternative correctly identified the true source, Source 2. Alternatives two and three identified two different sources with strong variability in their loading profiles. This difference among the loading profiles is to be expected due the varying location of alternatives two and three. The graphs for Sensor C show the observed

sensor data in black marks and the predicted sensor data in solid, colored lines. There was no difference between the predicted and observed sensor data for alternative one and there was minimal difference for the subsequent alternatives. The large difference among decision variables with similar objective values indicates that non-uniqueness is present in the system due to a good generation of alternatives.

5. SUMMARY AND CONCLUSIONS

Niched Co-Evolution Strategies is a method that combines the optimization of Evolution Strategies with a modeling approach to generate alternatives to address the problem of non-uniqueness in source identification. NCES was executed for three source and sensor network ensembles in the virtual city of Mesopolis. All three ensembles indicated that non-uniqueness was present in the system by identifying alternatives that were different in their loading profiles but produced sensor data that closely, and in some cases exactly, matched the observed sensor data. One metric that can be used to determine the amount of non-uniqueness in the system is the distance calculation. In general, a greater distance between alternatives indicates that there is strong non-uniqueness present in the system. This conclusion is supported by the results shown in the previous sectiom. From a management and WDS security standpoint, the presence of non-uniqueness suggests that more data should be collected to confidently identify the true source of contamination. If more data cannot be acquired in a timely manner, NCES provides alternative source locations where response actions should be implemented to protect consumers. NCES is a helpful tool for emergency managers and others responsible for WDS security. Aside from addressing the problem of non-uniqueness in source identification, NCES can be coupled with other modeling and optimization approaches to address a wide range of problems including adaptive source identification, sensor placement, and source identification using diverse sources of information.

While NCES can be extended to different algorithms for future studies, the algorithm can be improved to make the results more accurate and reduce the computational time. Since computer technology is constantly updated, improvements for NCES may come in the form of algorithmic and hydraulic system changes. Using a regular centroid calculation, as opposed to the weighted centroid calculation mentioned previously, may improve the identification of sources that are geographically distant. Sensor placement algorithms can be executed for Mesopolis to improve the value of the data that is collected. One hydraulic system adjustment that could be made to improve NCES's ability to perform in a more practical situation is to include more realistic sensors that better reflect current technology. For this research, it was assumed that all sensors were perfect and provided accurate data. Realistically, sensors occasionally provide false readings or incomplete readings. There is research in the area of using imperfect sensor data for source identification that could be incorporated for future work with NCES (Preis and Ostfeld 2008).

Since Mesopolis is such a large system with many possible nodes, it could be of benefit to use a method to decrease the search space for NCES. The search space cannot be reduced too much, though, because the goal of NCES is to search different areas of the decision space to find alternative solutions. Liu et al. (2008) explores the use of a local search to decrease the search space for the source identification problem, and this type of algorithm can be integrated as an initial step in NCES.

NCES follows a long line of tools developed to aid in the decision making process. Liebman (1976) defined the role of optimization in public sector decision

making as a tool to assist in the decision making process, rather than a single solution identifier. Optimization and model methods should provide insight to a problem as opposed to delivering an answer. Most problems in the public sector are classified as "wicked" and are difficult to solve using traditional methods. Brill's development of MGA (1979) was a significant step in the direction of providing insight to a problem instead of simply solving the problem. NCES is the next step in coupling the modeling approach of MGA with optimization methods to provide information about a problem. For example, in the source identification problem, NCES indicates if non-uniqueness is present in a problem and the amount of non-uniqueness in the problem. Though NCES provides solutions in the form of alternatives, these alternatives are not the final answer. They merely show that there are several possible areas of contamination and suggest that more data is needed to identify the true source or a different respond strategy is needed to protect the public from any of the possible sources.

At this time, NCES has been applied to the technical model of the hydraulic system, while ignoring the dynamic interactions of consumers and utility operators that could change the propagation of the contaminant plume. Another area of future work for NCES is incorporating the social aspects of a contamination event with the technical system through a socio-technical system analysis approach. Research is already underway in this field (Zechman 2011). Using NCES to generate alternatives instead of finding one outcome can provide an immense amount of insight for socio-technical problems.

REFERENCES

Brill, E. D. Jr. (1979). “The use of optimization models in public-sector planning.” *Water Resour. Res.*, 15(4), 750-756.

Brumbelow, K., Torres, J., Guikema, S., Bristow, E., Kanta, L. (2007). “Virtual cities for water distribution and infrastructure system research.” *Proc. World Environmental and Water Resources Congress*, ASCE, Tampa, FL.

Clark, R.M. and Deininger, R.A. (2000). “Protecting the nation’s critical infrastructure: the vulnerability of U.S. water supply systems.” *J. of Contingencies and Crisis Manage.*, 8(2), 73-80.

Fernandez, T., and Evett, M. (1997). “The impact of training period size on the evolution of financial trading systems.” *Proc. Genetic Programming Second Annual Conference*, AAAI, Stanford, CA.

Guan, J., Aral, M., Morris, L. M., and Grayman, W. M. (2006). “Identification of contaminant source in water distribution systems using simulation-optimization method: Case study.” *J. of Water Resour. Plann. and Manage.*, 132(4), 252-262.

Harrell, L. J. and Ranjithan, S. (2003). "Integrated detention pond design and land use planning for watershed management." *J. of Water Resour. Plann. and Manage.*, 129(2), 98-106.

Kroll, D. (2006). *Securing Our Water Supply: Protecting a Vulnerable Resource*. PennWell Corporation, Tulsa, OK.

Laird, C. L., Biegler, L. T., van Bloemen Waanders, B. G., and Bartlett, R. A. (2005). “Contamination source characterization for water networks.” *J. of Water Resour. Plann. and Manage.*, 131(2), 125-134.

Liebman, J. C. (1976). “Some simple-minded observations on the role of optimization in public systems decision-making.” *Interfaces*, 6(4), 102-108.

Liu, L., Brill, E. D. Jr., Mahinthakumar, G., and Ranjithan, S. (2008). “Contaminant source characterization using logistic regression and local search methods.” *Proc. World Environmental and Water Resources Congress*, ASCE, Honolulu, HI.

Loughlin, D. H., Ranjithan, S. R., Brill, E. D. Jr., and Baugh, J. W. Jr. (2001). “Genetic algorithm approaches for addressing unmodeled objectives in optimization problems.” *Engineering Optimization*, 33(5), 549-569.

Mac Kenzie, W. R., Hoxie, N. J., Proctor, M. E., Gradus, M. S., Blair, K. A., Peterson, D. E., Kazmierczak, J. J., et al. (1994). "A massive outbreak in Milwaukee of *Cryptosporidium* infection transmitted through the public water supply." *The New England J. of Medicine*, 331(3), 161-167.

Mahfoud, S. W. (1995). “Niching methods for genetic algorithms.” Ph.D. Dissertation, University of Illinois, Urbana- Champaign.

Mahinthakumar, K., von Laszewski, G., Ranjithan, S., Brill Jr., E. D., Uber, J., Harrison, K., Sreepathi, S., and Zechman, E. M. (2006). “An adaptive cyberinfrastructure for threat management in urban water distribution systems.” *Proc., International Conference on Computation Science*, Reading, UK.

Preis, A., and Ostfeld, A. (2006). “Contamination source identification in water systems: A hybrid model trees-linear programming scheme.” *J. of Water Resour. Plann. and Manage.*, 132(4), 263-273.

Preis, A., and Ostfeld, A. (2007) “A contamination source identification model for water distribution system security.” *Engineering Optimization*, 39(8), 941-951.

Preis, A. and Ostfeld, A. (2008). “Genetic algorithm for contaminant source characterization using imperfect sensors.” *Civil Engineering and Environmental Systems*, 25(1), 29-39.

Public Health Security and Bioterrorism Preparedness and Response Act of 2002. 4 USC. Sec. 401. 2002. *Requirements of the Public Health Security and Bioterrorism Preparedness and Response Act of 2002 (Bioterrorism Act)*. United States Environmental Protection Agency. Web. 30 Apr. 2011. <http://water.epa.gov/infrastructure/watersecurity/lawsregs/bioterrorismact.cfm>.

Rechenberg, I. (1973). *Evolutionsstrategie: Optimierung technischer systeme und prinzipien derbiologischen evolution.* Frommann-Holzboog, Stuttgart.

Rossman, L. A. (2000). *EPANET 2 Users Manual*, National Risk Management Research Laboratory, U.S. EPA, Cincinnati.

Schwefel, H. P. (1981). *Numerical Optimization of Computer Models*. Wiley, Chichester, UK.

Schwefel, H. P. (1995). *Evolution and Optimum Seeking*. Wiley & Sons, New York.

Singh, G. and Deb, K. (2006) “Comparison of multi-modal optimization algorithms based on evolutionary algorithms.” *Proc., Genetic and Evolutionary Computation Conference,* Seattle, WA.

Van Bloemen Waanders, B.G., Bartless, R.A., Bigler, L.T., and Laird, C.D. (2003). “Nonlinear programming strategies for source detection of municipal water networks.” *Proc. World Environmental and Water Resources Congress*, ASCE, Philadelphia, PA..

Zechman, E. M. and Ranjithan, S. (2004). “An evolutionary algorithm to generate alternatives (EAGA) for engineering optimization problems.” *Engineering Optimization*, 36(5), 539-553.

Zechman, E. M., Liu, L., and Ranjithan, S. (2006). “Niched co-evolution strategies (NCES) to address non-uniqueness in engineering design.” *Proc., Genetic and Evolutionary Computation Conference,* Seattle, WA.

Zechman, E. M. and Ranjithan, S. (2007). “Generating alternatives using evolutionary algorithms for water resources and environmental management problems.” *J. of Water Resour. Plann. and Manage.*, 133(2), 156-165.

Zechman, E. M. and S. Ranjithan. (2009). “Evolutionary computation-based methods for characterizing contaminant sources in a water distribution system.” *J. of Water Resour. Plann. and Manage.*, 135(5), 334-343.

Zechman, E.M., K. Brumbelow, M. Lindell, J. Mumpower, A. Rasekh, and M. Shafiee. (2011). “Agent-based modeling for planning emergency response to contamination emergencies in water utilities.” *NSF CMMI Engineering Research and Innovation Conference*, Atlanta, GA.

Zechman, E. M. (2011). “Agent-based modeling to simulate contamination events and evaluate threat management strategies in water distribution systems.” *Risk Analysis*, 31(5), 758-772.

VITA

Kristen Leigh Drake received her Bachelor of Science degree in civil engineering from Texas A&M University in May 2009. She remained in the civil engineering program at Texas A&M University and received her Master of Science degree in August 2011. Her research interests include water distribution systems, optimization, and systems engineering. She plans to pursue her Doctor of Philosophy in civil engineering at North Carolina State University starting August 2011. Ms. Drake may be reached at kristenldrake@hotmail.com. Her mailing address is c/o Dr. Kelly Brumbelow, WERC 205L, 3136 TAMU, College Station, TX 77843-3136.