An Integrated Framework for Response Actions for a Drinking Water Distribution Security Network

Problem and Research Objectives

The use of non-specific water quality sensors have provided the foundation for developing contamination warning systems (CWS) for drinking water distribution systems to protect public health from (un)intentional intrusion events. These non-specific water quality sensors (e.g., chlorine, pH, conductivity, etc.) can be linked with data driven event detection algorithms to determine when anomalous water quality conditions occur. The ability to detect water quality events has led to the development of forensic tools, such as contaminant source identification, and response strategies to mitigate the impact on the population. However, there has been little activity associated with translating the CWS signals and contaminant source information into predicting the future transport of a contamination event to inform response activities. Thus, there is a critical need to develop an integrative framework that can propagate the impacts of contaminant source uncertainty through distribution system transport to provide more information for utilizing response tools. The objective of this project, which is the next step towards developing an integrated real-time security application, is to develop a framework that will forecast contaminant spread throughout the distribution system based on the current estimated state of potential contaminant sources, which will then be utilized to inform confirmatory sampling locations to improve our estimates of contaminant spread and source location.

Methodology

Our central hypothesis is that contaminant source identification algorithms can be utilized to assess the potential contaminant spread, and provide needed information to select confirmatory sampling locations to improve our estimates of contaminant source identification and spatial distribution of an event. The rationale for developing this framework is to better utilize the information generated via CWS for assessing the overall impact throughout the distribution system and provide more appropriate response actions. We will test our central hypothesis and associated objectives by pursuing the following: 1) implement a forecasting algorithm to propagate probabilistic information associated with potential contaminant source locations to assess contaminant spread; 2) develop an approach for identifying confirmatory sampling locations that seeks to maximize new information associated with a contamination event; and 3) create an output format conducive for visualization purposes.

Forecasting Algorithm. For a given sensor network design, the forecasting algorithm will rely on the probabilistic contaminant source identification (PCSI) algorithm of Yang and Boccelli (2013) and the EPANET distribution system modeling software (Rossman, 2000) – to identify and characterize the probability of a specific location as a potential contaminant source. As sensors report positive or negative alarms (i.e., indicating an event has occurred or safe conditions exist, respectively), the PCSI algorithm utilizes the backtracking algorithm (Shang et al., 2002) to determine the upstream location-time pairs that are hydraulically connected to the observed sensor signal. Then, a Bayesian updating procedure – a Beta-Binomial conjugate pair – is used to update the probability that the location-time pair was the source (a positive alarm increases the probability; a negative alarm decreases the probability). For large networks, the PCSI algorithm will identify multiple potential source locations. The backtracking algorithm, in conjunction with hydraulic information, will be used to efficiently forecast the short-term (e.g., up to 6 hours) spread of contaminant from the individual sources represented as a "conservative tracer." For each of the downstream nodes, the flow-weighted probabilities from the

potential source locations will be assumed to characterize the probability of a contaminant being at the downstream location.

Identifying Confirmatory Sampling Locations. The following sections first introduce the metric used to quantify the information associated with sampling locations, and then how the best sampling location is determined.

Entropy as Information. Within the area of Information Theory, "entropy" is provided as a metric of information – more specifically, entropy represents the average information contained within a specific distribution (Reza, 1961). For example, if we are provided a discrete distribution of n classes, where each class has an associated probability p_i, $\sum p_i = 1$, the entropy of that distribution is defined as

$$H(P) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

where P represents the overall distribution and $H(P)$ the expected information, or entropy, contained within that distribution. For a uniform distribution (i.e., p_i is the same for all classes), $H(P)$ has maximum entropy defined as $\log_2(n)$. When we have perfect information regarding one class (i.e., $p_i = 1$, $p_j = 0$ when $i \neq j$) then $H(P)$ has a minimum entropy of 0. Thus, decreasing entropy represents increased information associated with the discrete distribution.

Confirmatory Sampling Selection. The intent behind confirmatory sampling is to increase the information associated with the potential source or forecasted locations. The initial entropy estimates for the potential source and contaminant spread locations can be estimated using the probabilities developed from the PCSI algorithm. To determine the amount of information gained (or lost) by confirmatory sampling, we need to perform the following steps for each potential sampling location: i) identify the probability that the sampling location could result in a contaminant observation; ii) assume that the resulting confirmatory sample returned a positive or negative alarm and separately update the PCSI results; iii) for each set of updated source probabilities, update the probabilities of the forecasted contaminant spread; iv) calculate the updated entropies for the source and forecasted locations assuming both positive and negative alarms; and v) use the probability value associated with observing the contamination event at the sampling location (from step i) to calculate the expected entropies for the source and forecasted locations based on the updated entropies from step iv. The differences between the initial and updated entropies represent the expected amount of information gained (or lost) by performing confirmatory sampling at that individual location. Confirmatory samples with the largest information differences indicate the locations that would provide the most new information.

The benefit to this approach is that an increase in information can occur by either selecting locations that would reinforce higher probability source locations, or, vice-versa, that would reinforce lower probability source locations. In either case, the overall information would be increased by generating a distribution such that the probability of the most likely candidate location(s) also increased. Computationally, this approach is also attractive because updating the probabilities is relatively straightforward since all of the necessary hydraulic and water quality simulations have been performed leaving only the algebraic calculations to update the probabilities. The relative ease of computation will allow an enumeration approach to be utilized to identify confirmatory sampling locations rather than using a formalized optimization algorithm (e.g., a mixed-integer non-linear programming approach) that would likely be more computationally intensive.

Principal Findings and Significance

In order to test and evaluate the forecasting and sampling algorithm, two example networks were utilized: 1) a small test network (the Net 3 example) included with EPANET (Rossman, 2000), and 2) a large network most recently utilized in the "Battle of the Water Sensor Networks" (Ostfeld et al, 2008).

Small Network. Figure 1 presents the small test network model as well as the placement of five water quality sensor locations (blue symbols). A 1-hour contaminant injection was simulated at node 10 starting at the 3rd hour of the simulation. The first detection of the contaminant occurs at the 5th hour at the water quality sensors located at Node 193.

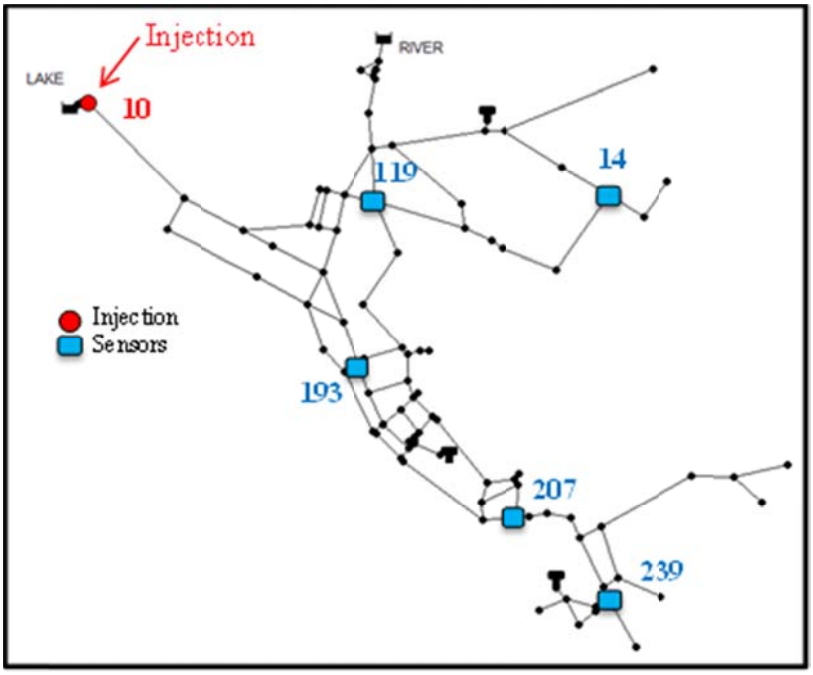

Figure 1. Small example test network for evaluating the forecasting and sampling algorithm.

Once the contaminant reaches a water quality sensor, the detection of the contaminant along with the hydraulics can be utilized with the PCSI algorithm to determine the probabilities of the upstream nodes. Figure 2 presents the upstream locations that have probabilities of greater than [red] or less than [green] a 50% probability of being a source, as determined by the PCSI algorithm.

Figure 3 presents classification results associated with data collected up until hour 6. Prior to hour 6, the data represents the classification results from the PCSI algorithm (solid lines). After hour 6, the data represents the forecasted classification performance (dashed lines). The performance of the PCSI algorithm decreases until, at the current time (hour 6), both the percent correct and incorrect identifications for the PCSI reach zero because there are no hydraulically connected nodes at the current time. With respect to the forecasted data, the increased percentage of correct identification occurs because the results are dependent upon hydraulic connections with previously connected nodes, not current connectivity with the sensors. The performance of the spread forecasting algorithm decreases as the forecasting horizon increases, which is to be expected.

Figure 4 [top] shows the forecasted change in entropy for all of the 97 possible sampling locations; a line plot (instead of a scatter plot) is used to more clearly show the results [the Node Index is an internal EPANET variable; the Node ID shown in the graph is the identifier of the actual location]. The peaks within each box are associated with the Node ID value provided, and represent the larger changes in entropy (remember, decreasing entropy suggests more information). For Nodes 1 (the biggest change), 40, and 179, these locations are either the tank or associated with the pipe to/from the tank. Node 237 is potentially downstream of contamination spread. In addition to Node 237, Nodes 206, 208, 209, 211, and 213 all have similar drops in entropy as these are all hydraulically "close" to Node 237.

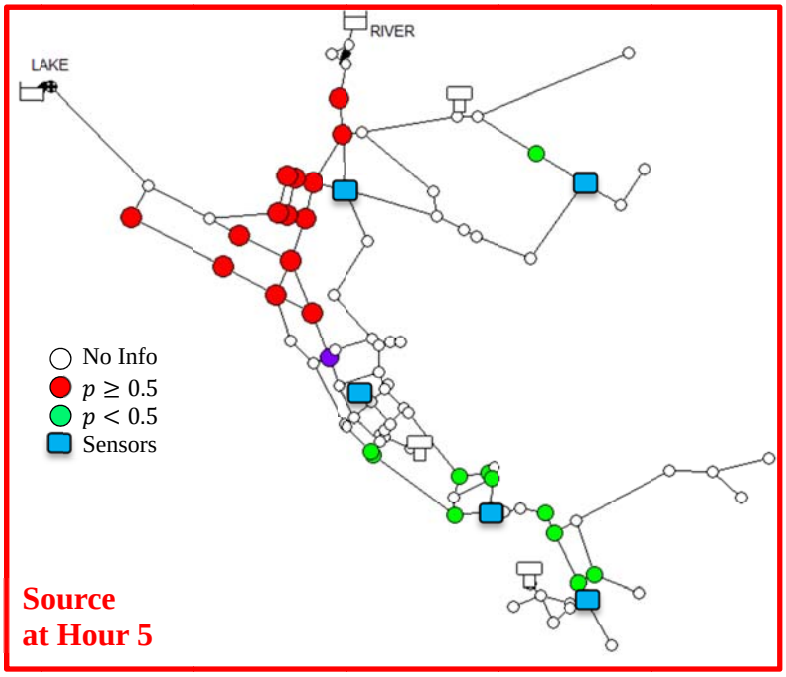

Figure 2. Estimated source probabilities at hour 5 of the simulation; red symbols represent locations with probabilities of being a source greater than 50%, green symbols represent locations with probabilities of being a source greater than 50%.

Figure 5 [top] shows the resulting change in entropy for all of the 97 possible sampling locations *after* the simulation is moved ahead one hour; a line plot is used to more clearly show the results [the Node Index is an internal EPANET variable; the Node ID shown in the graph is the identifier of the actual location]. These results provide information associated with the "true" information of the system at the next hour. The peaks within each box are associated with the Node ID value provided, and represent the larger changes in entropy (remember, decreasing entropy suggests more information). The nodes near the tank (1, 40, 179 and 271) would have provided the greatest decrease in information. The approach correctly identified Node 1 as the best sampling location and, with the exception of Node 271, also identified Nodes 40 and 179 as sampling locations with more significant decreases in entropy (Figure 4). Additionally, Nodes 199, 201 and 202, which are just downstream of Tank 1, also appear to be significant during the next hour. For the forecasted sampling, Nodes 206, 208, 209, 211, 213, and 237 were all shown to have similar impacts on entropy (Figure 3). While Nodes 211, 213 and 237 show a slightly less drop in "actual" entropy relative to the forecasted entropy, these locations were still identified. However, Nodes 206, 208 and 209 did not appear to have as significant an impact on entropy

as forecasted, but Node 241 appears to have been a good place to sample that was not one of the higher locations identified in the forecasting portion of the algorithm.

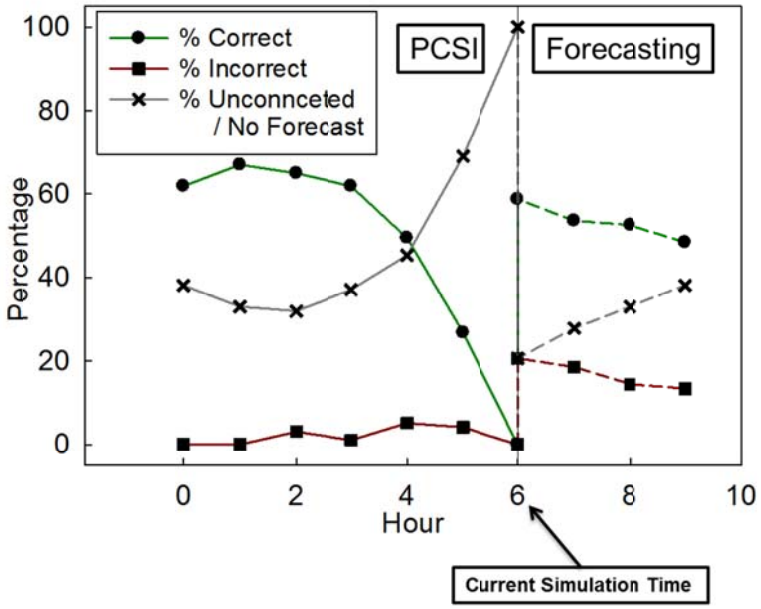

Figure 3. Spread forecasting summary for Net 3. PCSI results (solid lines) up to hour 6 are shown along with forecasting results (dashed lines) for hours 6 through 9; sensor data up to hour 6 was used for this forecasting.

Figure 6 shows the change in the PCSI results when sampling at nodes 1 or 2 along with the baseline case (no sampling). The shaded region shows the range of changes for sampling at all other locations. Sampling at nodes 1 and 2 produced the greatest increases in percent correct identification and greater decreases in percent unconnected. The percent incorrect classifications were not impacted by sampling at nodes 1 and 2. With respect to the forecasting, there was little improvement in forecasting accuracy via confirmatory sampling.

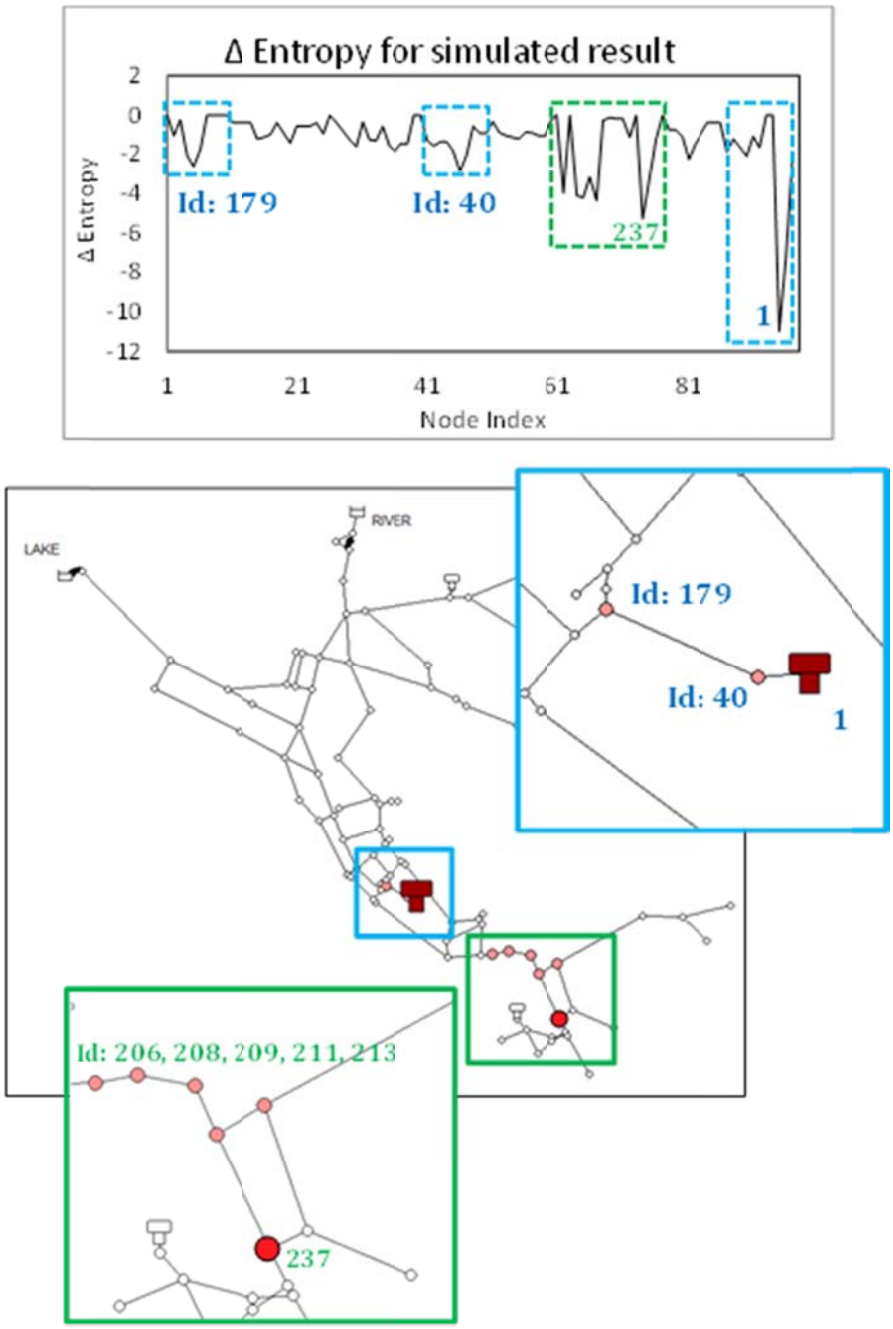

Figure 4. Plots of the forecasted change in entropy for each potential sampling location [top], as well as the spatial location of the sampling nodes that result in the greatest decrease in entropy (i.e., largest increase in information).

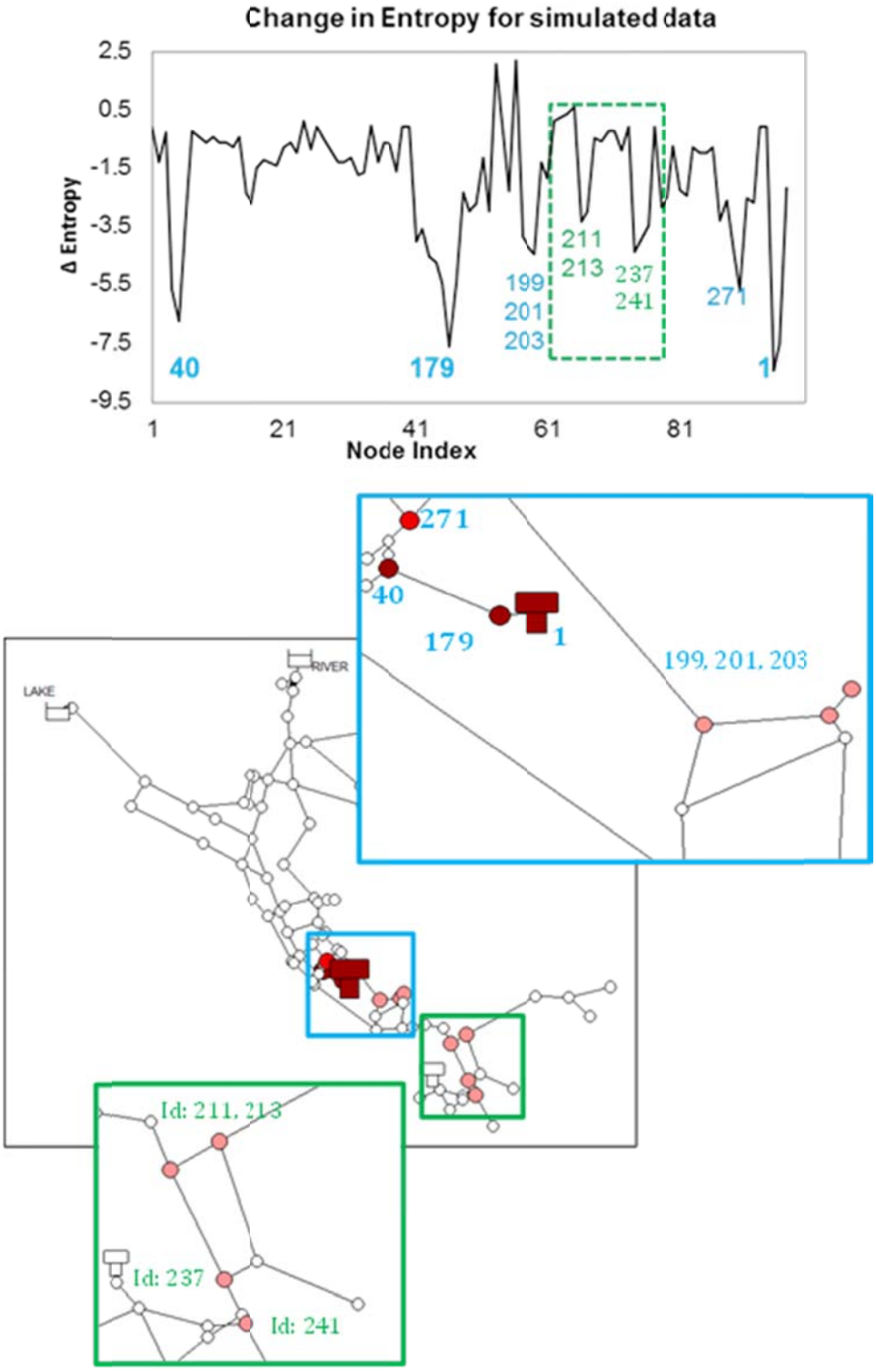

Figure 5. Plots of the forecasted change in entropy for each potential sampling location [top], as well as the spatial location of the sampling nodes that result in the greatest decrease in entropy (i.e., largest increase in information).

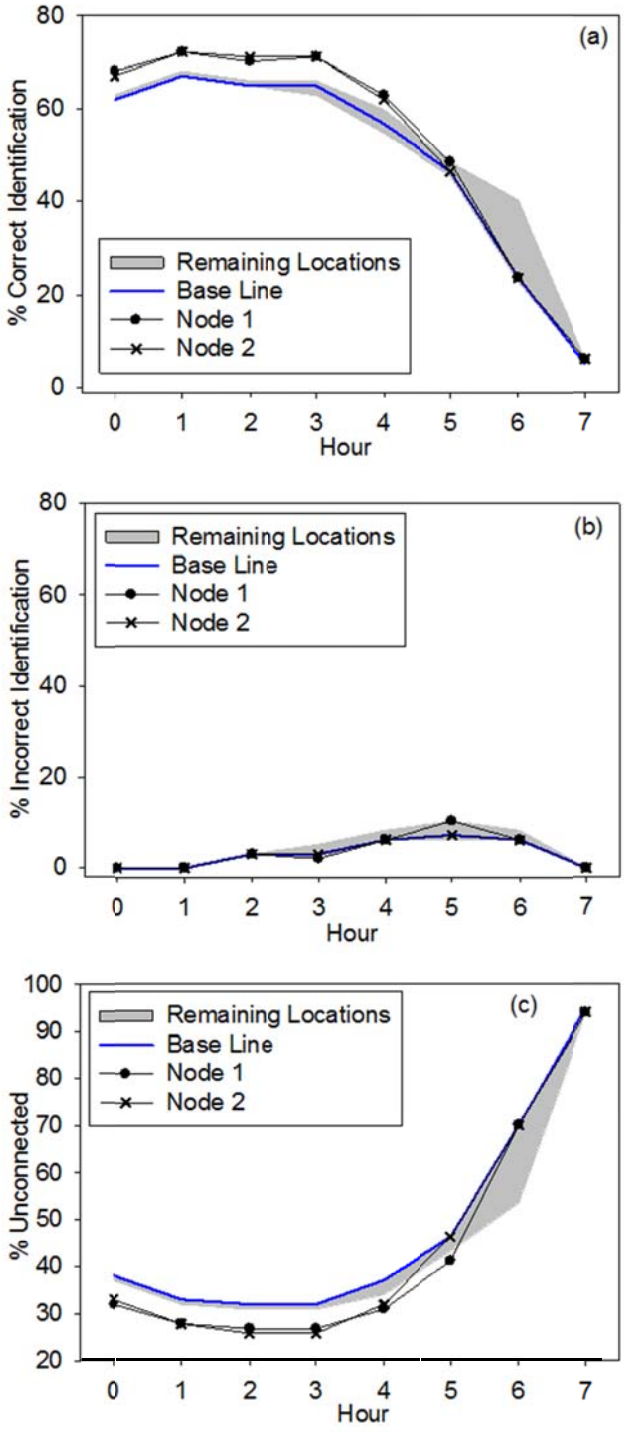

Figure 6. Changes in: (a) % correct identification; (**b**) % incorrect identification; and (c) % unconnected, in the PCSI results for Net 3 from a single grab sample during hour 7.

Large Network. Figure 7 presents the large network model as well as the placement of five water quality sensor locations (blue symbols); additional studies were performed with 10, 20 and 50 sensors. A 1-hour contaminant injection starting at the 3rd hour was simulated at node 5416 (shown by the star), which was close to the source and a high-impact node. The first detection of the contaminant occurs at the 11th hour.

Figure 7. Layout of the large test network with the injection location indicated by a star and the five sensor locations indicated by the rectangles.

Figure 8 shows the accuracy results from the PCSI algorithm and spread forecasting. These results are similar to those shown for the small network except that the percentage of correctly identified nodes and unconnected nodes are significantly decreased and increased, respectively, relative to the small network. This is due to placing the same number of sensors within a much larger network. Increasing the number of sensors improved both metrics (not shown).

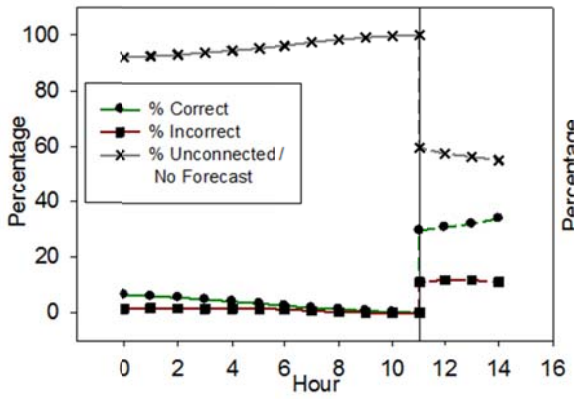

Figure 8. PCSI and spread forecasting results for the large network including the perfect correct, incorrect, and unconnected/no forecast for the 5 sensor case.

When using the forecasted results to identify the sampling locations, the sampling location that resulted in the greatest decrease in network entropy did not always result in the greatest decrease in actual network entropy once the simulations were continued. However, the locations associated with the best performance using the expected and actual network entropies were always shown to be in the top 1% of possible nodes. Figure 9 shows the top 1% of estimated confirmatory sampling locations. The circle represents the sampling location that provides the most information using the forecasted data; the triangle represents the best sampling location using perfect information (i.e., the actual hydraulic conditions, not forecasted). As can be seen in the inset, while the two possible sampling locations are not located at the same spatial locale, the sampling locations are in a similar spatial region. This observed spatial similarity occurred for all of the results if the two locations were not identical.

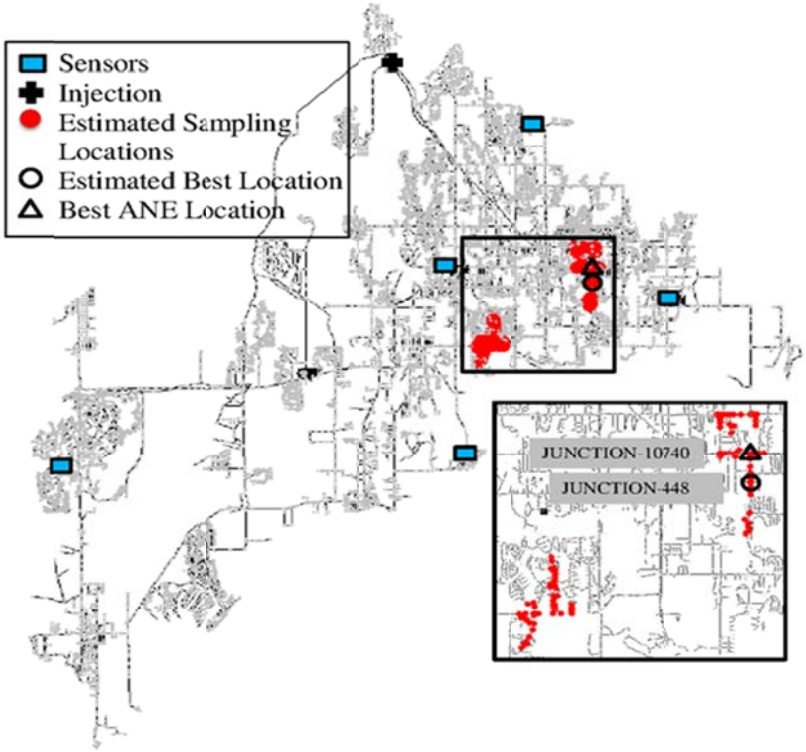

Figure 9. Spatial locations of the top 1% of the estimated sampling locations for the 5 sensor scenario in the large network; the circle represents the best sampling location based upon forecasted information and the triangle represents the best sampling location under perfect information.

Figure 10 illustrates the changes in the percentage of correct and incorrect identification and percentage of unconnected nodes in the large network for the sampling location based upon the forecasted data (JUNCTION-448), perfect information (JUNCTION-10740), and all other possible nodes (grey shaded area). While the sampling locations based on the forecasted and perfect information were not located at the exact same spatial locations, both locations provided the greatest improvement in the PCSI results relative to the other locations. In general, the confirmatory sampling algorithm identified locations that would increase the percentage of correctly identified nodes, and reduce percentage of incorrectly identified and unconnected nodes.

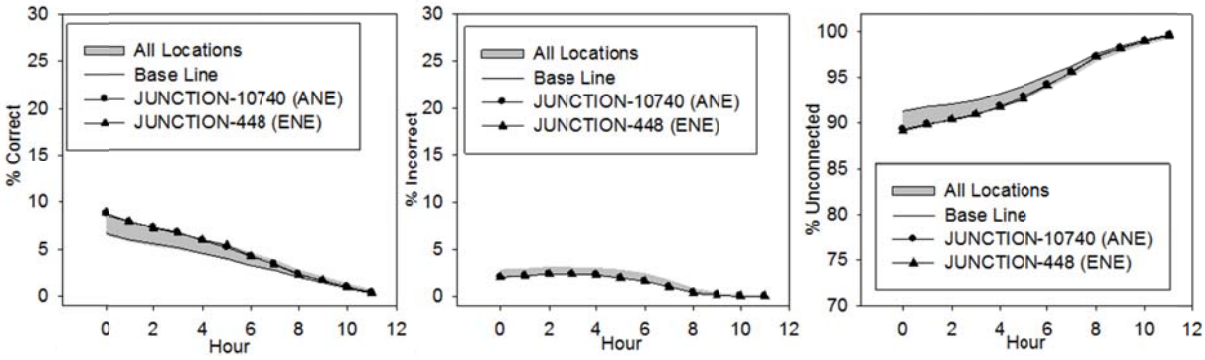

Figure 10. Changes in % correct identification (left), % incorrect identification (middle) and % unconnected (right), in the PCSI results for the large network with 5 sensors.

Figure 11 shows the future classifications without confirmatory sampling and with the confirmatory sampling locations based upon the forecasted and perfect information. For all three categories, the confirmatory samples improved our forecasted estimates. Interestingly enough, the confirmatory sampling location based on forecasted information had a greater impact on the performance of the three categories than if we had perfect information, even though the overall entropy was less. When investigating this result, the confirmatory sampling locations based upon the forecasted data tended to "uncover" more of the unconnected network whereas the sampling locations based upon the perfect information sought to identify and strengthen existing hydraulic pathways. Thus, while the latter improved the information along a given hydraulic path, not as much new information was generated

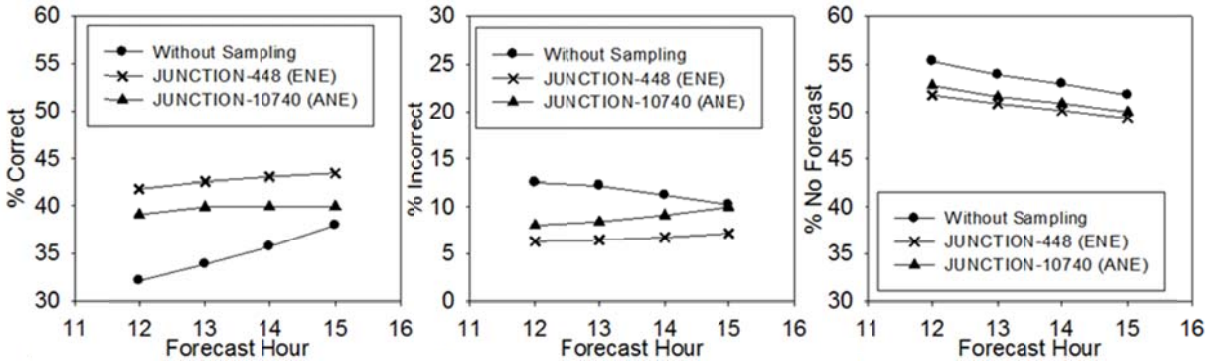

Figure 11. Change in % correct identification (left), % incorrect identification (middle) and % unconnected (right), in the forecasted spread results for the large network with 5 sensors.

about the remainder of the network.

Significance. The contribution of this research is significant as this study is effectively the first research that bridges the gap between the information generated from a contaminant warning system (via sensor response and contaminant source identification) and the initial phases of response (via confirmatory sampling). The forecasting and confirmatory sampling is particularly important for larger networks that cannot adequately cover the spatial scale of large networks with costly sensor systems.

Long-term, these results will provide the foundation for developing more robust response activities when attempting to mitigate the impact of an intrusion event.

Publication Citations

Ostfeld, A., Uber, J. G., Salomons, E., Berry, J. W., Hart, W. E., Phillips, C. A., Watson, J.-P., Dorini, G., Jonkergouw, P., Kapelan, Z., di Pierro, F., Khu, S.-T., Savic, D., Eliades, D., Polycarpou, M., Ghimire, S. R., Barkdoll, B. D., Gueli, R., Huang, J. J., McBean, E. A., James, W., Krause, A., Leskovic, J. Isovitsch, S., Xu, J., Guestrin, C., VanBriesen, J., Small, M. Fischbeck, P., Preis, A., Propato, M. Pillier, O., Trachtman, G. B., Wu, Z. Y., and Walski, T. (2008) "The Battle of the Water Sensor Networks (BWSN): A design challenge for engineers and algorithms." *Journal of Water Resources Planning and Management*, 134(6), 556-568.

Reza, F. M. (1961) *An Introduction to Information Theory*. McGraw-Hill, New York.

Rossman, L. A. (2000) *EPANET2 User's Manual*. Cincinnati, OH: Risk Reduction Engineering Laboratory, U.S. Environmental Protection Agency.

Shang, F., Uber, J. G., and Polycarpou, M. M. (2002) "Particle Backtracking Algorithm for Water Distribution System Analysis." *Journal of Environmental Engineering*, 128(5), 441-450.

Yang, X. and Boccelli, D. L. (2013) "A Bayesian Approach for Real-Time Probabilistic Contaminant Source Identification." *Journal of Water Resources Planning and Management*, (under review).