SAND2010-6049C

Effective Integration of Operations Data with Water Quality Data

Sean A. McKenna[1]*, David Hart*[1]*, Terra Haxton*[2]*, Regan Murray*[2]

[1] *National Security Applications Department, Sandia National Laboratories, Albuquerque, NM*
[2] *National Homeland Security Research Center, US Environmental Protection Agency, Cincinnati, OH*

Water quality signals are processed by event detection systems (EDS) to determine the probability of a water quality event occurring at each sample time. Inherent noise in sensor data and rapid changes in water quality due to changes in the hydraulic operations of the distribution network can cause false alarms from EDS's. Here we examine the practical problem of utilizing information on the operations of the network within water quality event detection and present two different approaches for integrating operational and water quality data.

We consider two situations, proximal and distal, that define the relationship between the locations of a water quality monitoring station and the operational changes that impact the water quality. In the proximal case, the monitoring station is located near the source of the operational changes and those changes have a direct and nearly immediate impact on water quality. In the distal case, the monitoring station and the operational changes are located in different parts of the network and the lag time between the operational and water quality changes are variable and unknown.

For the proximal case, we construct aggregated signals as composites of the raw operational signals. Compositing operations are defined as algebraic combinations and transformations of one or more raw signals into a new signal. For example, water quality changes are observed to result when at least 2 of 3 pumps at an adjacent location start or stop running. A composite "pump change" signal is created to indicate when at least two pumps have changed status. This new signal serves as the input to a second composite signal that looks for a change in the "pump change" signal during any of the previous five time steps. A change within that time window is used to decrease the sensitivity of the EDS. This approach is equivalent to creating a hardware calibration alarm based on the pump operations.

In the distal case, historical water quality data and operational data are used to develop a multivariate pattern library of frequently observed changes. As an example, simultaneous changes in water quality signals are accompanied by changes in flow rate and water temperature at the monitoring station. All of these signals define multivariate patterns of change in a pattern library. During real-time operations, any new potential event is compared against the library. Current data that match an existing pattern within a tolerance are not considered as water quality events. The CANARY software was used to test and demonstrate these techniques for integrating operational data to improve event detection sensitivity and decrease false alarms.

Introduction

Potential contamination of water distribution networks by contamination events is concern for network operators and the customers they service. The infrastructure of these networks was designed primarily for customer service and firefighting, retrofitting them to be more robust against contamination events is difficult and costly. One solution to making them more robust to contamination events is enhanced monitoring. Deployment of an EDS within a distribution network can provide around-the-clock monitoring of water quality data signals with near-realtime notification of the presence of anomalous water quality (see Murray et al., 2010a).

An potential issue with EDS deployment are larger than desired numbers of false positive event notifications. Too many false positives can decrease confidence in the EDS by the water quality analysts. False positive water quality alarms can often be caused by changes in the hydraulic operations of the distribution network. As an example, deployment of EDS tools at the Greater Cincinnati Water Works (GCWW) over a six month period has shown a number of instances where false alarms were produced due to opening or closing of valves, draining of tanks and changes in the status of pumps within the distribution system (Allgeier and Umberg, 2008). Data processing algorithms are needed to recognize water quality changes that are due to operational changes and improve the water quality event detection.

Here we examine two approaches to integrating operational data into water quality event detection: composite signals and trajectory clustering. A brief description of both approaches is provided and example calculations are documented. The event detection examples and results shown here are done using the CANARY software (Hart and McKenna, 2010; Murray et al., 2010b). CANARY has been developed as an open-source platform for water quality event detection and contains a number of algorithms for detecting anomalous periods of water quality. These algorithms include adaptive filtering of multivariate water quality signals, data fusion techniques, a binomial failure model to aggregate information over time steps and multivariate pattern recognition (Hart and McKenna, 2010; Murray et al., 2010b). CANARY connects to existing SCADA systems and provides the user with full customization of the sensitivity of the event detection algorithms for each monitoring station within the network.

Composite Signals

Operational data from a utility often provide indirect information on water quality changes. For a given utility, a rule-set could be developed that uses recorded changes in operational signals to inform the analyst of upcoming changes in water quality. An alternative to utility specific rule-sets is a more general approach to incorporating operational signals into water quality event detection. Here we employ *composite signals* to integrate both operational and water quality signals into enhanced event detection.

Composite signals within CANARY are defined as a set of algebraic operations that can be completed on individual water quality signals or on combinations of water quality signals. These operations create new signals that are sent, along with the original water quality signals to CANARY for analysis. As examples, a moving average or a set of differences between single time steps can easily be created from a water quality signal. Additionally, ratios of two or more signals at the current time step or at different previous time steps can be created. Additional details and applications of composite signals to water quality event detection are available in Hart et al., (2010).

Here, we are interested in using composite signals to transform changes in operational signals into calibration alarms. This is a simple approach use a pattern or a change in operations to temporarily suspend event detection capability. The change in operational data is used as a cue to notify the event detection algorithms that the system is undergoing a temporary calibration and the incoming water quality signals are not representative of the background water quality. This approach works best for the case of the water quality monitoring station being proximal to the source of the operational change.

Trajectory Clustering

Trajectory clustering considers the multivariate pattern of water quality through time. Historical data are used to construct a library of these patterns. Typically, the patterns are defined by the measured water quality signals within a prescribed number of time steps prior to a water quality event. The approach to pattern library construction is designed to provide a concise summary of common multivariate water quality patterns against which any new water quality pattern can quickly be compared. The steps in this approach are to: 1) Identify water quality events in historical data; 2) Summarize the change in water quality that defines those events by fitting each signal within the event data with a low-order polynomial; 3) Classify the regression models into a small number of patterns through multivariate cluster analysis applied to the coefficients of the polynomials regression models; and 4) Calculate statistics on the resulting patterns.

During online event detection, the statistical description of the resulting patterns is used to determine the proximity of any new water quality event to one of the existing patterns. If the water quality event in question is close enough to a known pattern, it is considered part of the background water quality and no alarm is sounded. The current water quality pattern is then added to the appropriate cluster within the library.

Patterns can be constructed from historical water quality data or they can be constructed from a combination of water quality and operational data. Adding operational data can reduce the variability in the patterns and provide a link between changes in water quality and causal operational changes

Example Calculations

Three example calculations are used to demonstrate the effective integration of operational data into water quality event detection. The first two examples cover the case of the water quality monitoring station being proximal to the cause of the operational changes and utilize composite signals to 1) utilize information on changes in pumping rates to suppress false positive alarms during those periods; and 2) integrate supplemental water quality data into event detection. Both of these example calculations also use composite signals to suppress water quality alarms after a period of sensor calibration, although these results are not called out specifically here. The final example calculation examines the case of a water quality monitoring station being distal from the operational change. For this example, a pattern library is constructed from historical data and then trajectory clustering is used to match current water quality to those patterns. Additional details on trajectory clustering and its application to water quality data can be found in Vugrin et al. (2008). TRAJECTORY CLUSTERING EXAMPLE

Post-Calibration Alarm Suppression

A common issue in water quality monitoring is the creation of alarms following the calibration of a monitoring station. Post-calibration water quality is often significantly different than water quality prior to the calibration due to changes in the sensor settings. EDS tools will often see this change as an event and provide an alarm. A composite signal can be created to suppress alarms for a time period following a calibration event. In the two proximal examples shown below, EDS alarms are suppressed for 15 time steps (30 minutes) following any sensor calibration.

Water Quality Changes Caused by Operational Changes

A common cause for water quality change at a monitoring station within a water distribution networks is a change in pumping such that water from a different source, or water of a different age, is now moving past the monitoring station. In this example analysis, a pump that connects a reservoir to a main as well as two other pumps in the network impact water quality at the monitoring station (Figure 1). The locations of the pumps and details on their connections to the main containing the water quality sensor are known by the utility, but were not available for this analysis. For this monitoring station, five water quality signals are monitored: residual chlorine (Cl), pH, specific conductance (COND), turbidity (TURB) and total organic carbon (TOC). Additionally, the water temperature is recorded but is considered and operational parameter here.

The data in Figure 1 show a relationship between changes in the pump status and changes in the water quality. However, it is not clear if all pumps impact water quality equally or if the timing between the change in pump status and the change in water quality is the same for all pumps. A composite signal was developed to look back in time over the previous five time steps of the pumping rate and retain the maximum absolute change. This same composite signal was applied to each of the three pumps. Another composite signal was written to combine these results and retain the maximum change (absolute value) over all three pumps during the past five time steps. A final composite signal compares the maximum change value against a threshold, here 5.0 gpm, and suspends the event detection if the change exceeds that threshold. The maximum time for which the event detection will be suspended is set by the length of how far back changes in the pump are recorded – 5 time steps (10 minutes) in this example.

Implementation of these composite signals results in short periods of suspended event detection that are initiated by significant changes in operational data. Comparison of results with and without the composite signals activated is shown in Figure 2. For the week of data shown in Figure 2, there are three events (see blue dots on signals responsible for causing the events). When the information on changes in pumping is integrated through the composite signals, two of the events are seen to be associated with changes in pumping and are ruled out as potential water quality events and only one event remains. Over the entire data set of 105 days, 0.85 percent of all time steps are considered water quality events without the use of composite signals and these occur in 20 different groups. When the composite signals are used to identify water quality changes due to changes in pumping, the percent of time steps considered as events is reduced to 0.50 and these occur in 10 groups. For this example, integration of changes in pumping reduced the proportion of false positive time steps by approximately 40 percent. Certainly, utility-specific knowledge could be used to further improve these results by refining the numbers of time steps and the thresholds used within the composite signals.

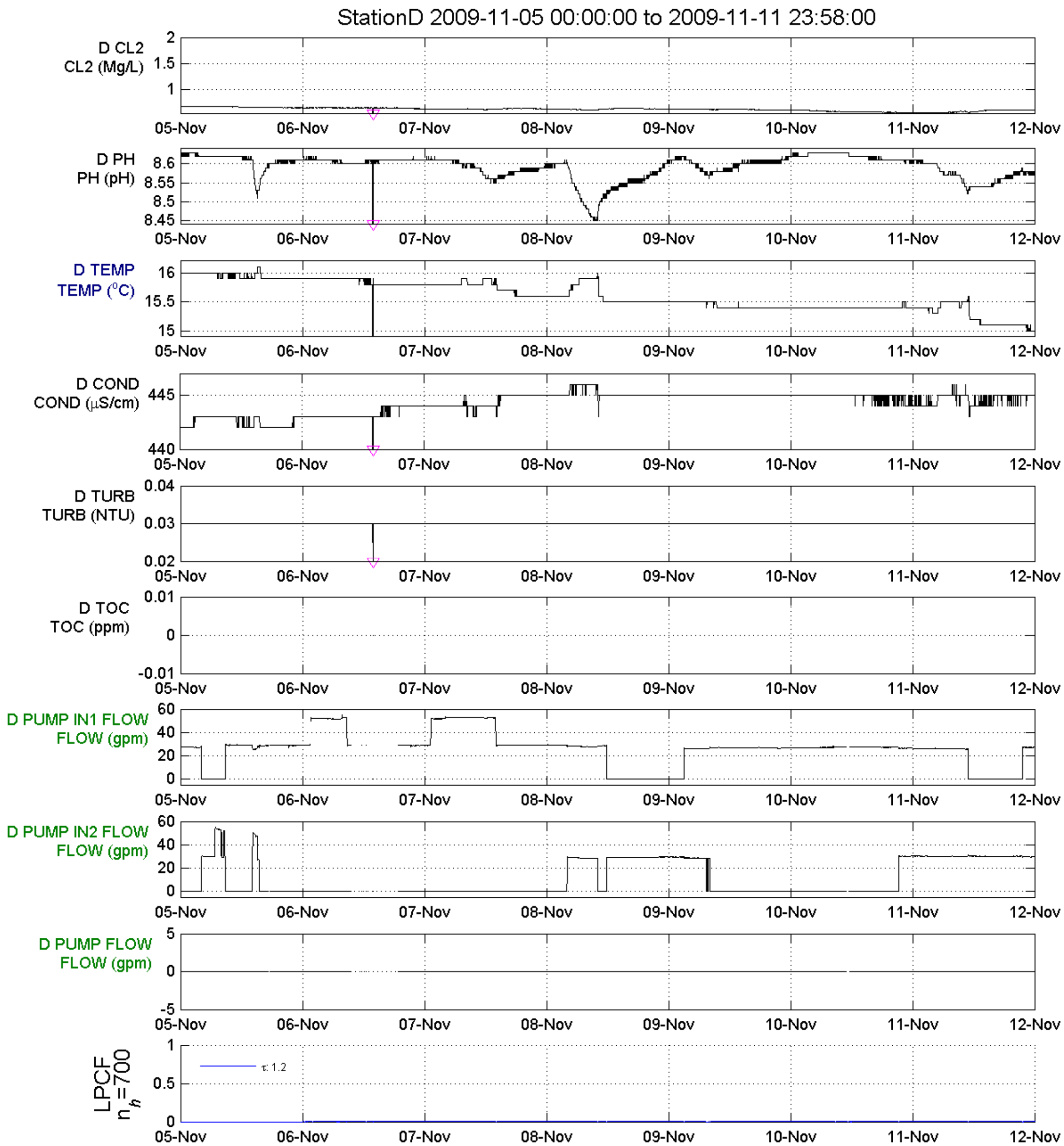

Figure 1. Relationships between pumping rates (three graphs with green labels on Y axis) and water quality data.

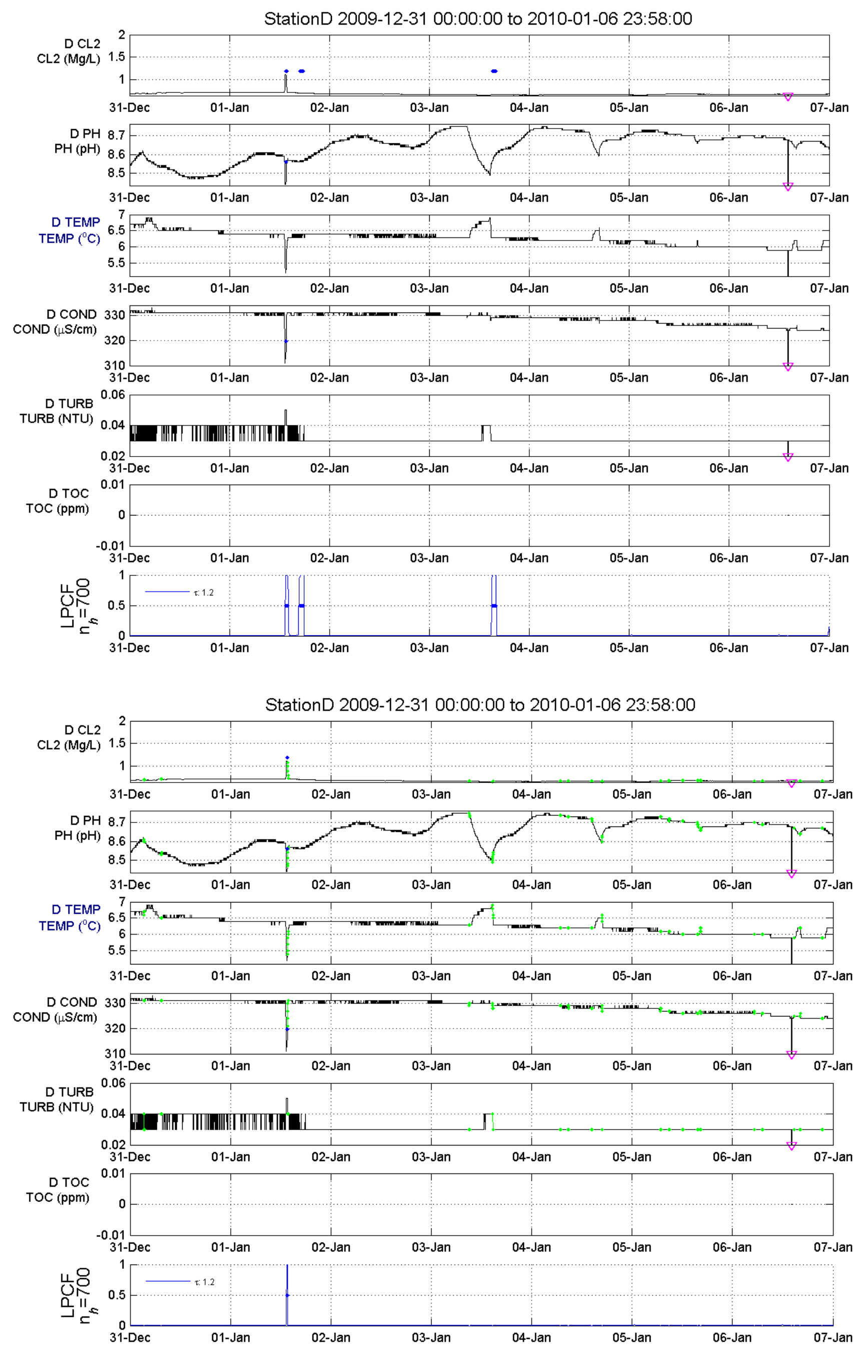

Figure 2. Comparison of water quality data and CANARY results without the use of composite signals (top) and with the addition of composite signals (bottom). Time periods where event detection is suspended are shown by the green lines in the bottom graphs. The probability of an event as calculated by CANARY is shown by the blue lines in the bottom graph of each set.

Supplemental Water Quality Data
Primary water quality monitoring stations are often supplemented by nearby secondary water quality monitoring stations. A common example is supplemental monitoring of a single water quality signal, e.g., residual chlorine, at a nearby tank outlet or pump station. While supplemental monitoring stations may only employ one water quality signal, this can be enough information to reliably cue the event detection system to significant changes in water quality at a downstream monitoring station.

In this example, a nearby tank outlet that feeds into the network upgradient of the monitoring station has a residual chlorine monitor. The data from this monitor are used to construct a composite signal that is then used as input to the event detection running on the monitoring station. Here, the composite signal identifies the largest (absolute value) change in the tank outlet chlorine value between the current time step and either of the two previous time steps. This maximum change value is then compared against a threshold of 0.06 mg/l. Changes larger than this threshold are considered indicative of an impending significant change in water quality at the down-gradient monitoring station and the monitoring station is considered to be in "calibration mode" for that time step. Because the maximum change is calculated between the current time step and either of the two previous time steps, the typical result is that a change in the tank outlet chlorine value will cause the monitoring station to remain in calibration mode for two consecutive time steps (four minutes for this monitoring station).

Results comparing the use of the composite signals to the case where they are not employed for a single day of analysis are shown in Figure 3. The upper set of graphs is for the case of no composite signals and the lower set is for the case of the composite signals being activated. In the lower set of graphs, the signals are colored green during periods of water quality event suppression. Three water quality signals are analyzed: residual chlorine (CL2X), pH (PHXX) and specific conductance (COND). Additionally, the water temperature is recorded, but is not used in the water quality event detection. The upstream Cl monitor values are also shown in Figure 3. The bottom graph in each set (blue line) shows the probability of a water quality event as calculated by CANARY. For the day of data shown in Figure 3, there are three event periods identified by CANARY without the composite signals. When the composite signals are activated, all of these events are identified as being due to changes in the water entering the system from the reservoir as signaled by the changes in the chlorine residual values at the upstream monitoring station.

A total of 41 days of data were analyzed for this monitoring station. The base case analysis here with no composite signals results in 2.87 percent of all time steps classified as water quality events. Addition of the composite signals to utilize the upstream Cl monitor and to allow a smooth transition after calibration events reduces the amount of time steps classified as events to 0.47 percent. This result is close to an 85% reduction in the number of false positives for this monitoring station. Similar to the first example case, it may be possible to improve these results with additional tuning of the composite signals

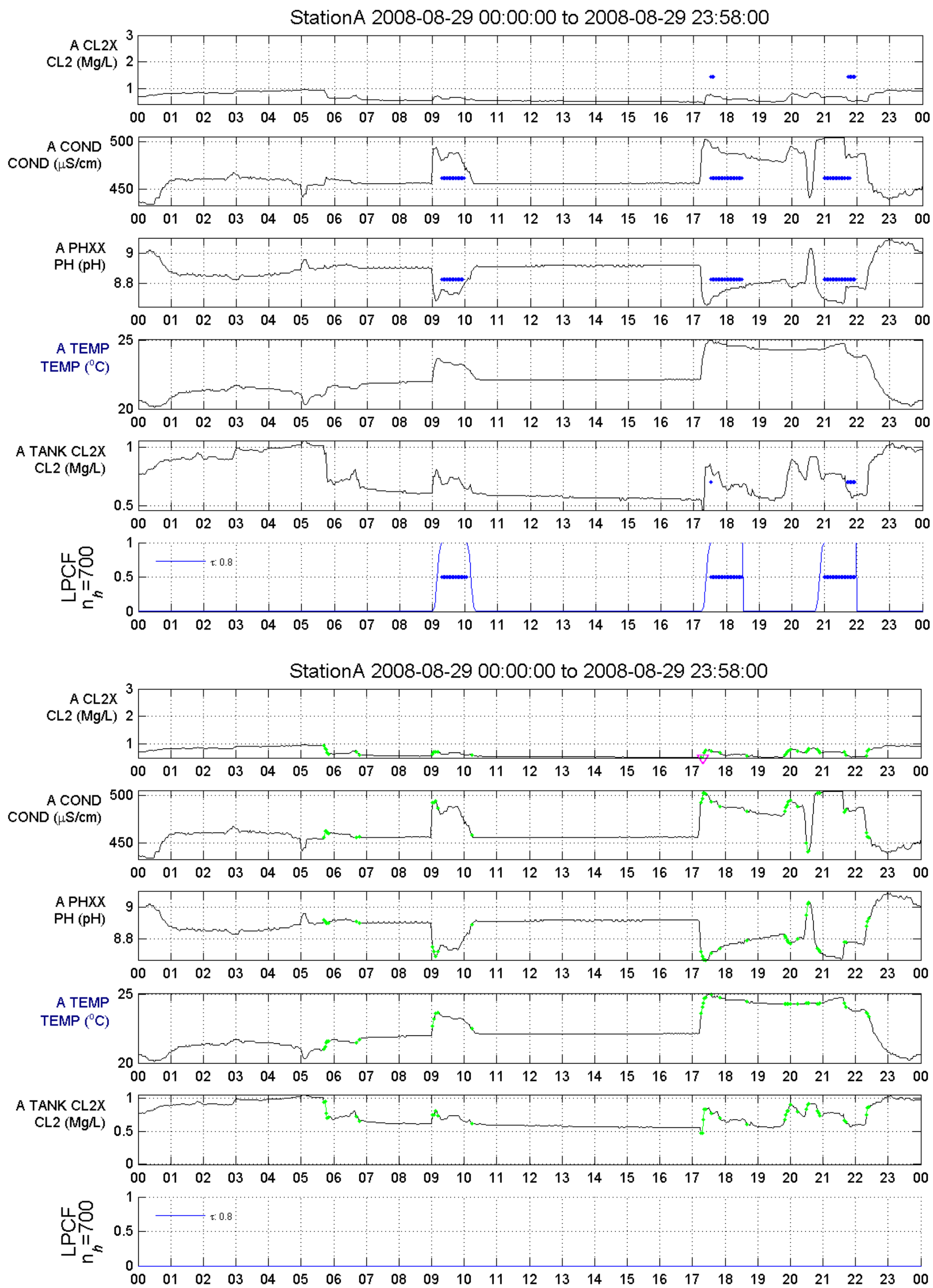

Figure 3. Example 2 comparison of water quality data and CANARY results without the use of composite signals (top) and with the addition of composite signals (bottom). Time periods where event detection is suspended are shown by the green lines in the bottom graphs. The probability of an event as calculated by CANARY is shown by the blue lines in the bottom graph of each set.

Building Water Quality Pattern Libraries

In the case of a monitoring station that is distal from the source of operational changes, the timing and nature of the relationship between the operational change and the water quality change may be complex and time varying. In these cases, construction of a library of observed water quality changes that are considered part of the background water quality variability can be completed for later use in realtime monitoring. As an example of library construction, 46 days of water quality data were collected and analyzed. The data are sampled every 2 minutes and there are three water quality signals: chlorine residual (CL2), pH and specific conductance (COND). Additionally, the water temperature (TEMP) was also recorded at this station.

Multivariate patterns were constructed by running CANARY on the historical data and capturing the 20 time steps of data in each signal prior to and including the time step at which an event was identified. Each of these segments of data is then fit with a third-order polynomial and the coefficients of the polynomials are clustered using a fuzzy-C-means clustering algorithm. The analysis was done on the three water quality parameters. In a second run, temperature was included in the clustering. Temperature is treated as an operational signal here and does not contribute to the event detection, but it is used as a separate signal in the pattern construction. The attributes of the multivariate patterns identified in the data with and without consideration of temperature as a signal are compared in Table 1. In Table 1, the colors denote the general change in each signal over the 20 time steps captured in the pattern: Red is a decrease, yellow is no change and green is an increase.

Table 1. Comparison of patterns detected in historical data with and without inclusion of temperature.

	N_{event}	**CL2**	**pH**	**COND**	**TEMP**
Pattern 1	24				
Pattern 2	171				
Pattern 3	34				
Pattern 4	15				
Pattern 5	1				
Pattern 1	37				
Pattern 2	165				
Pattern 3	17				
Pattern 4	25				
Pattern 5	1				

The addition of temperature to the pattern definition has minimal effect on the resulting patterns with the exception of Pattern 3. Pattern 3 decreases in size (number of events) from 34 to 17 and demonstrates an increase in chlorine residual values when temperature is added relative to decreasing values in the original pattern.

Summary
Two approaches for integrating operational data into water quality event detection have been described here. The composite signal approach provides a flexible platform for defining signals that are created from operations on one or more input data streams. Examples here show how changes in pumping rates and changes in upstream water quality can be used to reduce false positive event detections. Additionally, composite signals are used to suppress water quality event alarms after a sensor calibration. Results in these examples show that integration of operations data through composite signals can reduce false positive events by 40 and 85 percent in examples 1 and 2, respectively.

The conference presentation will use the pattern libraries constructed here in event detection. The results with the addition of the operational data, temperature, will be compared to the results where temperature is not included in the pattern library construction.

A difference in the two approaches used here is that the composite signal approach provides a calibration signal that is used to suppress alarms when a significant change in water quality occurs. While the results above show the proportion of time the station is put into calibration mode to be less than 1.5% and typically less than 0.1%, the EDS is not operating during these time periods. The trajectory clustering approach does not require the monitoring station to be taken offline and therefore provides uninterrupted event detection.

Acknowledgements
Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. This work was performed under Interagency agreement DW89921928 with the US Environmental Protection Agency.

The information in this document has been funded wholly or in part by the U.S. Environmental Protection Agency (EPA). It has been subjected to the Agency's peer and administrative review and has been approved for publication as an EPA document. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

References
Allgeier, S.; Umberg, K. 2008, Operational Experience with Water Quality Anomaly Detection During the Cincinnati Water Security Initiative Pilot (presentation) AWWA WQTC, Cincinnati, OH, November Cincinnati, OH

Hart, D.B. and S.A. McKenna, 2010, CANARY User's Manual, Version 4.2, U.S. Environmental Protection Agency, Office of Research and Development, National Homeland Security Research Center, EPA 600/R-08/040A, 54 pp.

Hart, D.B., S.A. McKenna. R. Murray and T. Haxton, 2010, Combining Water Quality and Operational Data for Improved Event Detection, Water Distribution Systems Analysis (WDSA) conference 2010, Tucson, AZ, Sept 12-15.

Murray, R. T. Haxton, D. Hart, S. McKenna, K. Umberg, J. Hall, Y. Lee, M. Tyree and D. Hartman, 2010a, Case Study Application of the CANARY Event Detection Software, presentation at AWWA Water Quality Technology Conference, Savannah, Georgia, November 14 - 17.

Murray, R., T. Haxton, S.A. McKenna, D.B. Hart, K. Klise, M. Koch, S. Martin, M. Wilson, V. Cruz, L. Cutler, 2010b, Water Quality Event Detection Systems for Drinking Water Contamination Warning Systems, Development, Testing, and Application of CANARY, EPA/600/R-010/036U.S. Environmental Protection Agency, National Homeland Security Research Center, Cincinnati, Ohio, 91 pp.

Vugrin, E., S.A. McKenna and D. Hart, 2009, Trajectory Clustering Approach for Reducing Water Quality Event False Alarms, in proceedings of ASCE Annual World Environmental and Water Resources Congress, Kansas City, Missouri, May 17-21