United States
Environmental Protection
Agency

EPA/600/R-010/036 | May 2010 | www.epa.gov/ord

Water Quality Event Detection Systems for Drinking Water Contamination Warning Systems

Development, Testing, and Application of CANARY

Office of Research and Development
National Homeland Security Research Center

EPA/600/R-10/036 | May 2010 | www.epa.gov/ord

Water Quality Event Detection Systems for Drinking Water Contamination Warning Systems

Development, Testing, and Application of CANARY

Regan Murray and Terra Haxton
National Homeland Security Research Center
Cincinnati, OH 45268

Sean A. McKenna, David B. Hart, Katherine Klise, Mark Koch, Eric D. Vugrin, Shawn Martin, Mark Wilson, Victoria Cruz, Laura Cutler
Sandia National Laboratories
Albuquerque, NM 87185

NATIONAL HOMELAND SECURITY RESEARCH CENTER
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
CINCINNATI, OH 45268

Disclaimer

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.

This work was performed under Interagency agreement DW89921928 with Sandia National Laboratories. 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.

The CANARY software described in this manual is subject to copyright. It is free software that can be redistributed and/or modified under the terms of the GNU Lesser General Public License as published by the Free Software Foundation and to the terms of other third-party software licenses. Specifications of these terms are included with the CANARY software distribution.

The authors and the U.S. Environmental Protection Agency are not responsible and assume no liability whatsoever for any results or any use made of the results obtained from this software, nor for any damages or litigation that result from the use of this software for any purpose.

Foreword

Since 1970, EPA has been working toward a cleaner and healthier environment for the American people. To help meet this mandate, EPA's research program provides data and technical support for solving environmental problems today and for building the scientific base necessary to manage our ecological resources wisely, understand how pollutants affect our health, and prevent or reduce environmental risks in the future.

Following the events of September 11, 2001, EPA's mission was expanded to address critical needs related to homeland security. Presidential Directives identified EPA as the primary federal agency responsible for the country's water supplies and for decontamination following a chemical, biological, and/or radiological (CBR) attack. To provide scientific and technical support in meeting this expanded mission, EPA's National Homeland Security Research Center (NHSRC) was established. NHSRC is focused on conducting research and delivering products that improve the capability of the Agency to carry out its homeland security responsibilities.

As a part of this mission, NHSRC conducts research and provides technical assistance to support America's drinking water utilities so they can improve their security preparedness, response, and recovery. Over the last several years, NHSRC has been developing new methods to help design, implement, and evaluate drinking water contamination warning systems. These new systems integrate a variety of monitoring technologies to rapidly detect contamination. One important question for contamination warning system implementation is how to analyze water quality sensor data to determine if a contamination event has occurred in a water distribution network. Since water quality data are often noisy, it is difficult to visually determine if contaminants are present in the network. **This publication summarizes a large body of research addressing event detection issues and provides critical information for water utilities considering how to analyze water quality data to detect contamination in their own distribution networks.**

NHSRC works with many partners to meet its responsibilities. This research was conducted in collaboration with EPA's Office of Water, across the federal government working with the U.S. Department of Energy's Sandia National Laboratories, and with the American Water Works Association and their member utilities.

This publication provides a comprehensive resource on event detection methods and case studies, and is intended for a broad audience of water utility staff, policy makers, and researchers. NHSRC has made this publication available to assist the water community in improving its security and optimizing the quality of our nation's drinking water. This research moves EPA one step closer to achieving its homeland security goals and its overall mission of protecting human health and the environment.

Cynthia Sonich-Mullin, Acting Director
National Homeland Security Research Center

Acknowledgments

The National Homeland Security Research Center would like to acknowledge the following organizations and individuals for their support in the development of this report, whether in contributing to the research, providing a peer review, sharing data, or pilot testing the software.

American Water Works Association
Kevin Morley

EPA Office of Research and Development
John Hall
Robert Janke

EPA Office of Water
Steve Allgeier
Mike Henrie
Katie Umberg

Sandia National Laboratories
William E. Hart

Drinking water utility partners
Yeongho Lee, Greater Cincinnati Water Works (pilot utility)
Chit Pin Teo, PUB, Singapore (pilot utility)
Cheng Sheun Lee, PUB, Singapore (pilot utility)
Zia Bukhari, New Jersey American Water
Dan Quintanar, Tucson Water

University of Cincinnati
Ernesto Arandia
Sam Hatchett

Shaw Group Inc.
Srinivas Panguluri

Table of Contents

List of Acronyms and Abbreviations

ACM	Association for Computing Machinery
AR	Autoregressive
ARMA	Autoregressive and Moving Average
ARIMA	Autoregressive Integrated Moving Average
ASCE	American Society of Civil Engineers
ASME	American Society of Mechanical Engineers
AWWA	American Water Works Association
AwwaRF	American Water Works Association Research Foundation (name changed to Water Research Foundation)
BED	Binomial Event Discriminator
CBRS	Case-Based Reasoning System
CDF	Cumulative Distribution Function
CDTY	Conductivity
Cl	Chlorine
CSV	Comma Separated Value
CUSUM	Cumulative Sum
CWS	Contamination Warning System
DHS	U.S. Department of Homeland Security
EDS	Event Detection System
EKF	Extended Kalman Filter
EnKF	Ensemble Kalman Filter
EPA	U.S. Environmental Protection Agency
FA	False Alarm
FAQ	Frequently Asked Questions
FCM	Fuzzy C-Means
FFT	Fast Fourier Transform
GCWW	Greater Cincinnati Water Works
ISSRE	International Symposium on Software Reliability Engineering
KDD	Knowledge Discovery and Data
KF	Kalman Filter
LD	Levinson-Durbin
LPCF	Linear Prediction-Correction Filter
µS/cm	microsiemens per centimeter
MA	Moving Average
MD	Missed Detection
mg/L	milligrams per Liter
MSET	Multivariate State Estimation Technique
MVNN	Multivariate Nearest Neighbor
NHSRC	National Homeland Security Research Center
NIST	National Institute of Standards and Testing
NTU	Nephelometric Turbidity Unit
ORP	Oxidation Reduction Potential
PBM	Pakhira, Bandyopadhyay, and Maulik (authors of PBM–index)
ppm	parts per million

List of Acronyms and Abbreviations

PUB	Singapore Public Utility Board
ROC	Receiver Operating Characteristic
SCADA	Supervisory Control and Data Acquisition
T&E	U.S. EPA Testing and Evaluation facility
TCMD	Thousand Cubic Meters per Day
TEVA	Threat Ensemble Vulnerability Assessment
TEVA-SPOT	Threat Ensemble Vulnerability Assessment Sensor Placement Optimization Tool
TOC	Total Organic Carbon
TRC	Total Residual Chlorine
U.S.	United States of America
WQ	Water Quality
WS	Water Security
XML	Extensible Markup Language

1. Background and Purpose

Protecting our nation's critical infrastructure from terrorist attacks has become a federal and local priority over the last several years. Under Homeland Security Presidential Directive 7, the United States Environmental Protection Agency (EPA) is the lead federal agency for protecting the water infrastructure in the United States. In this capacity, EPA has worked with public and private water utilities, federal, state and local agencies, and the public health community to develop assistance and research programs to improve the safety and security of drinking water systems. Water associations, community water systems, academia, private industry, and others have focused attention and research on developing new methods, policies, and procedures to secure drinking water and wastewater systems.

The Public Health Security and Bioterrorism Preparedness and Response Act of 2002 (Bioterrorism Act of 2002) required drinking water systems serving more than 3,300 people to conduct vulnerability assessments and prepare or update emergency response plans that address a range of potential terrorist threats. In 2006, a report on the fourteen features of an active and effective security program informed the water community about the most important organizational, operational, infrastructure, and external features of resilient and secure systems (U.S. EPA 2006). Many representatives of the water sector joined together to prepare a sector-specific plan that coordinates activities across organizations (U.S. DHS et al. 2007). These activities have reduced water sector vulnerabilities through increasing awareness, hardening of critical assets, improved physical security, and more comprehensive response plans.

Recently, water security research efforts have focused on the advancement of methods for mitigating contamination threats to drinking water systems. A promising approach for the mitigation of both accidental and intentional contamination is a *contamination warning system* (CWS), a system to deploy and operate online sensors, other surveillance systems, rapid communication technologies, and data analysis methods to provide an early indication of contamination. CWSs with multiple approaches to monitoring – such as water quality sensors located throughout the distribution system, public health surveillance systems, and customer complaint monitoring programs – are theoretically capable of detecting a wide range of contaminants in water systems. However, CWSs are expensive to purchase, install, and maintain. To make CWSs a viable option, there is a clear need to minimize the investment required by individual drinking water systems.

The purpose of this report is to provide documentation on strategies and tools needed to assist in the application of an *event detection system* (EDS) as part of a CWS. EDSs are required to analyze the large volume of data from online water quality monitors, to differentiate normal water quality patterns from anomalous conditions, and to alert the operator to these situations. This report focuses on the event detection methodologies that have been developed by EPA's Threat Ensemble Vulnerability Assessment (TEVA) Research Team, which is composed of researchers from EPA, Sandia National Laboratories, the University of Cincinnati, and Argonne National Laboratory. This team has developed several water tools including TEVA-SPOT – Threat Ensemble Vulnerability Assessment Sensor Placement Optimization Tool – (Berry et al. 2008; U.S. EPA 2009b) and CANARY (Hart et al. 2007; Hart et al. 2009). This report focuses on the research and development activities that led to the CANARY software.

Chapters 1-3 of this report are intended for a broad audience composed of water utility staff, policy makers, and researchers. These chapters describe the challenges in developing an EDS, the CANARY software and discuss frequently asked questions about how to implement the software at a field site. The remaining chapters of this report are intended for researchers and others who want to understand the methods implemented within CANARY in greater detail. These chapters present the algorithms underlying the methodology and some recent improvements and new features, some of which are still in the research stage and not yet implemented in CANARY. **Appendix A** presents a more detailed review of the event detection literature.

Drinking Water Contamination Warning Systems

Research on methods to mitigate the impacts of contamination incidents have converged over the last several years on the concept of a CWS. CWSs have been proposed as a promising approach for the early detection and management of contamination incidents in drinking water distribution systems (ASCE 2004; AWWA 2005; U.S. EPA 2005a). Through the Office of Water's Water Security (WS) initiative (formerly WaterSentinel), EPA is piloting CWSs at a series of drinking water utilities.

An effective response to a water contamination incident is based on minimizing the time between the detection of a contamination incident and the implementation of effective response actions that mitigate further consequences. Implementation of a robust CWS can achieve this by providing earlier indications of potential contamination incidents than would be possible in the absence of a CWS. A CWS is a proactive approach that uses advanced monitoring technologies and enhanced surveillance activities to collect, integrate, analyze, and communicate information that provides a timely warning of potential contamination incidents.

The WS initiative promotes a comprehensive CWS that is theoretically capable of detecting a wide range of contaminants, covering a large spatial area of the distribution system, and providing early detection in time to mitigate impacts (U.S. EPA 2005b). Components of the WS initiative include:

- **Online water quality monitoring.** Continuous online monitors for water quality parameters, such as free chlorine, total organic carbon, pH, conductivity, and turbidity, help to establish expected baselines for these parameters in a given distribution system. An EDS, such as CANARY (Hart et al. 2009), can be used to analyze data from these monitors in real-time to detect anomalous changes from the baseline and provide an indication of potential contamination. Other monitoring technologies can be used as well, such as contaminant specific monitors, although the goal is to detect a wide range of possible contaminants.
- **Consumer complaint surveillance.** Consumer complaints regarding unusual taste, odor, or appearance of the water are often reported to water utilities, which track the reports as well as steps taken by the utility to address these water quality problems. The WS Initiative is developing a process to automate the compilation and tracking of information provided by consumers. Unusual trends that might be indicative of a contamination incident can be rapidly identified using this approach.
- **Public health surveillance.** Syndromic surveillance conducted by the public health sector, including information such as sales of over-the-counter medication, reports from emergency medical service logs, calls from 911 centers, and calls into poison control hotlines, could serve as a warning of a potential drinking water contamination incident. Information from these sources can be integrated into a CWS by developing a reliable and automated link between the public health sector and drinking water utilities.
- **Enhanced security monitoring.** Security breaches, witness accounts, and notifications by perpetrators, news media, or law enforcement can be monitored and documented through enhanced security practices. This component has the potential to detect a tampering event in progress, potentially preventing the introduction of a harmful contaminant into the drinking water system.
- **Routine sampling and analysis.** Water samples can be collected at a predetermined frequency and analyzed to establish a baseline of contaminants of concern. These samples will provide a baseline for comparison during the response to detection of a contamination incident. In addition, this component requires continual testing of the laboratory staff and procedures so that everyone is ready to respond to an actual incident.

A CWS is not merely a collection of monitors and equipment placed throughout a water system to provide an intrusion or contamination alert. Fundamentally, it is an exercise in information acquisition and management. Different information streams must be captured, managed, analyzed, and interpreted in time to recognize potential contamination incidents and mitigate the impacts. Each of these information streams can independently provide some value in terms of timely initial detection. However, when these streams are integrated and used to evaluate a potential contamination incident, the credibility of the incident can be established more quickly and reliably than if any of the information streams were used alone. While the primary purpose of a CWS is to detect contamination incidents, implementation of a CWS is expected to result in dual-use benefits for network operations that will help to ensure its sustainability within a utility.

Although many utilities are currently implementing some monitoring and surveillance activities, these activities are either lacking critical components or have not been integrated in a manner sufficient to meet the primary objectives of a CWS – timely detection of a contamination incident. For example, although many utilities currently track consumer complaint calls, a CWS requires a robust spatial-temporal analysis system that, when integrated with data from public health surveillance, online water quality monitoring, and enhanced security monitoring, will provide specific, reliable, and timely information for decision makers to establish credibility and respond in an effective manner. Beyond each individual component of the CWS, coordination between the utility, the public health agency, local officials, law enforcement, and emergency responders, among others, is needed to develop an effective consequence management plan that ensures appropriate actions will occur in response to detection by the different components. An advanced and integrated laboratory infrastructure to support baseline monitoring as well as analysis of samples collected in response to initial detections is critical to timely response. In the absence of a reliable and sustainable CWS, a utility's ability to respond to contamination incidents in a timely and appropriate manner is limited.

Online Monitoring and Event Detection Systems

The online monitoring component of a CWS is composed of multiple sensor stations that collect data continuously and transmit it to a central database in a control room, most commonly a *Supervisory Control and Data Acquisition* (SCADA) database (see **Figure 1-1**). In the rest of the report, "SCADA" is used to represent any type of data storage system. Various types of sensors, which can be categorized as direct or surrogate, have been considered as part of a CWS. Direct sensors detect specific contaminants whereas surrogate sensors indirectly detect the presence of one or more contaminants through changes in water quality values. For example, pH, chlorine, electrical conductivity, oxygen-reduction potential, and total organic carbon can be considered as surrogate sensors for multiple contaminants. These typical water quality parameters tend to vary significantly in water distribution systems due to normal changes in the operations of tanks, pumps, and valves, and daily and seasonal changes in the source and finished water quality, as well as fluctuations in demands. Therefore, event detection systems are needed to distinguish between periods of normal and anomalous water quality variability from measures made with surrogate sensors.

A critical premise for online monitoring of surrogate parameters is that recognizable variations in water quality signals will occur in the presence of certain contaminants of concern. A number of recent experiments conducted in laboratory and pipe test loop systems have explored this assumption and concluded that many contaminants cause surrogate parameters to diverge significantly away from background levels (Byer et al. 2005; Cook et al. 2005; Hall et al. 2007). In particular, Hall et al. (2007) tested the response of a number of commercially available water quality sensors in the presence of nine different contaminants introduced to a pipe loop at different concentrations and found that at least one of the surrogate parameters changed in response to the presence of every contaminant.

Figure 1-2 shows data from another laboratory study by Hall examining the response of a chlorine sensor to the introduction of 15 different contaminants in a pipe-loop. These data were collected 24.3 m (80 feet) downstream of the contaminant injection point at EPA's Test and Evaluation Facility in Cincinnati, Ohio (Hall et al. 2009). The black line shows the response of the chlorine sensor to two separate injections of each of the 15 contaminants. For more information about the experiments, see Hall et al. (2009).

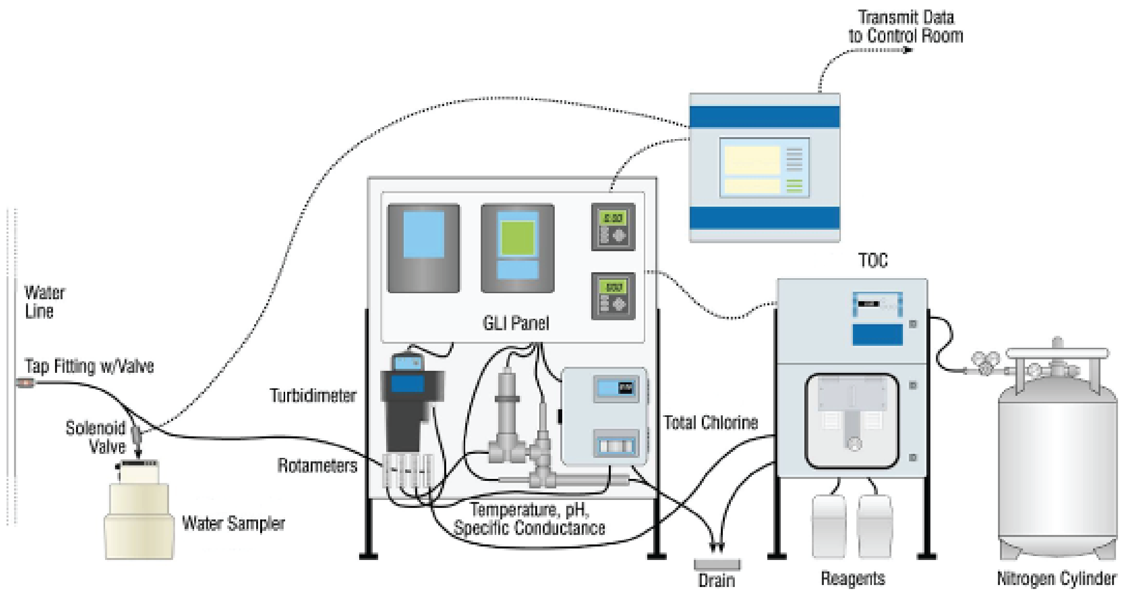


Figure 1-1. Typical sensor station configuration for the first Water Security Initiative pilot city.

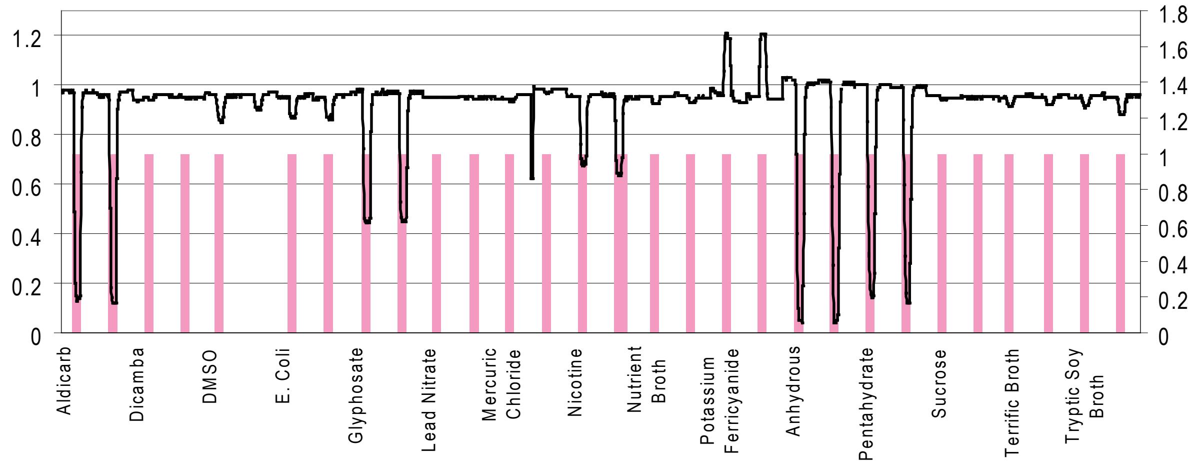


Figure 1-2. Response of a free chlorine monitor (black line) to the introduction of different contaminants into the pipe loop. The Y-axis on the left defines the free chlorine level in mg/L. The pink bars indicate the time period over which the contaminant is affecting water quality at the sensor station. The blue lines separate the results for each contaminant.

Figure 1-2 shows that for 10 of the 15 contaminants, there is a significant deviation from the baseline chlorine levels as measured by the sensor. For five of these ten contaminants that cause changes in free chlorine, the decrease in the free chlorine concentration is 80% or more from the background levels. These rapid and significant shifts away from the background concentration of the free chlorine values can be considered as potential indicators of the presence of contamination.

While **Figure 1-2** demonstrates the response of a single free chlorine sensor to various introduced contaminants, the goal of an EDS is not to simply analyze the response of a single sensor to the introduction of the contaminants, but to analyze the collective response of all sensors at the monitoring station. Other sensors (pH, dissolved oxygen, total organic carbon, and specific conductivity) were also examined in the same laboratory study. These results indicated that free chlorine and total organic carbon (TOC) are the most responsive surrogate parameters to the widest range of contaminants (Hall et al. 2007).

These laboratory studies suggest that water quality parameters (surrogates) will change rapidly and significantly in the presence of many contaminants of concern. However, real-world conditions in distribution systems involve much more complex background variations in water quality parameters than found in the laboratory (see Chapter 2). The purpose of an EDS is to automatically and rapidly distinguish between changes caused by the presence of contaminants and changes caused by normal background variability.

Typically, an EDS reads in SCADA data (e.g., water quality signals and operations data), performs an analysis in near real-time, and then returns the calculated probability of a water quality event occurring at the current time step. A water quality event is defined as a time period over which water with anomalous characteristics is detected. The working definition of “anomalous” can be set by the user by selecting configuration parameters that govern the sensitivity and operation of the EDS. The values of these configuration parameters might vary from one utility to the next and could even vary across monitoring stations within a single utility.

Increasing installation of online water quality sensors in distribution networks and their connection to SCADA systems has significantly increased the amount of water quality data available to system operators and network analysts. As an example, a modest online monitoring system consisting of ten monitoring stations with five water quality parameters monitored at a 5 minute sampling interval will provide 14,400 water quality records per day, or 5.26 million records per year. The possibility of massive amounts of real-time data overwhelming the operators and analysts is real, and automated approaches to making sense of these data are needed. Investment in automated approaches such as EDS will allow a utility to detect and characterize changes in the water quality as well as mine the historical data for recurring patterns and trends. Information derived from these data can then be used to more effectively operate the distribution network.

Summary of EDS Literature Review

Event detection from data collected in series over a length of time is a research topic in a large number of fields, including tsunami detection, traffic accidents analysis, mechanical component failure, system fault detection, data mining, and network intrusion detection among others. Based on reviewing developments in these other fields for their relationship and applicability to the EDS problem in water distribution networks, two main categories of event detection can be identified: *offline and online*. Offline, or batch mode, analysis is done on previously collected, or historical, data sets. Online, or real-time, analysis is done in real-time on data that are input to the EDS tool as soon as they become available.

A number of offline approaches to analyzing data collected over time can be classified as *change point* detection (see Ge et al. 2000b; Lai 1995; Raftery 1994; West et al. 1997). Change points are defined as the point in time where an abrupt change in the nature of a signal occurs. For example, the time at which there is a change in the source of water supplying a monitoring station can be a change point for the water quality at that station. A number of approaches to the change point detection problem exist, but the essential element is to examine data from opposite sides of a proposed change point to determine if those two data sets are significantly different. If they are, the point that best divides the two data sets is a change point. For offline analyses, the full data set has already been recorded and is available for analysis. In the online scenario, only the data recorded up to the present time are available, and the goal is to identify the change point as close to the time at which it occurs as possible. The constraint of making a determination as near to real-time as possible generally precludes the use of offline change point detection methods for water quality applications. The goal of an efficient water quality EDS is to develop an online approach to water quality event detection that can warn system operators in real-time about unexpected water quality conditions. Achievement of this goal generally requires use of online event detection algorithms.

Online EDS tools generally consist of a two-stage approach to event detection. The first stage predicts a future water quality value. This prediction is most often based on recently observed, historical water quality values. A wide variety of prediction tools are available, including neural networks, support vector machines, and calibrated water quality models. Our focus to date has been on traditional time series and multivariate statistical approaches (e.g., Box et al. 1976; Bras et al. 1993). Different statistical models applied to the previously observed data can provide predictions of future water quality values. The process of making the prediction is referred to as *state estimation*. In the second stage of event detection, the prediction of the expected water quality value is compared to the observed water quality value as it becomes available. The difference between the prediction and the observation is termed the *residual* and classification of the residual is used to determine if the water quality at that time step is either expected or anomalous. If the residual is relatively small, the predicted and observed water quality values are similar and the water quality is as expected or representative of the background water quality. If the residual is relatively large, the observed water quality value is quite different from what was predicted, and this indicates an anomalous observation. This second stage is called *residual classification*.

To date, the majority of event detection methods for drinking water distribution networks involve monitoring of surrogate parameters. Observations of changes in common water quality measures such as free chlorine, pH and specific conductivity serve as surrogates for more specific monitoring of individual contaminants. A number of studies have demonstrated how different surrogate monitors react to the introduction of contaminants. Current approaches to event detection in drinking water distribution systems are described in: Byer et al. 2005; Cook et al. 2006; Jarrett et al. 2006; Kroll et al. 2006; McKenna et al. 2008; Yang et al. 2009. Other issues of event detection that are important to water quality monitoring include approaches to event detection that simultaneously incorporate information from more than one monitoring station and techniques that can be used to quantitatively evaluate the performance of event detection algorithms.

An in-depth review of existing literature in fields of change point and event detection is included in **Appendix A.** These topics are of interest in a number of technical fields with the majority of recent research driven by event and intrusion detection in computer science. The amount of published literature in the broad area of anomaly detection from time series data is vast. The majority of the water quality event detection publications are quite recent, reflecting a growing interest in water security. **Appendix A** also includes a glossary of event detection terms. Terms contained in the glossary are italicized on first use throughout this document.

CANARY Software Tool

The CANARY EDS software has been developed at Sandia National Laboratories in collaboration with EPA's National Homeland Security Research Center (NHSRC). Additional functionality for reducing false alarms has been added to CANARY through engagement with the Singapore Public Utility Board (PUB). CANARY was written using the MATLAB® (MathWorks 2008) programming language and is distributed as both the MATLAB® source and as an executable program under an open source license. CANARY can be connected to a utility SCADA database directly or through a third party software connection. All water quality signals contained in the SCADA database can be used as input to CANARY. In addition to water quality data, these signals can also include hydraulic data such as tank levels, flow rates and valve settings as well as sensor hardware alarms and calibration alarms.

CANARY provides a platform within which different event detection algorithms can be developed and tested. These algorithms process the water quality data at each time step to identify periods of anomalous water quality. The end result of processing the water quality data at each time step is an indication of the probability of a water quality event existing at that time step. This probability is calculated with respect to the recent water quality values. Recent additions to CANARY allow for multivariate water quality pattern recognition to reduce false alarms in the presence of changes in water quality signals created by utility operations.

CANARY is intended as a research tool to help water utilities and others in the water community better understand normal background fluctuations in water quality and to begin to identify anomalies that are potentially indicative of contamination incidents. To be used as part of a CWS, the utility must integrate CANARY with a well-tested consequence management plan in order to respond effectively and in a timely manner to potential contamination threats.

Report Overview

This report is divided into several distinct chapters that do not need to be read in order. The first three chapters, including this Background section, are intended for a broad audience composed of water utility staff, policy makers, and researchers. The next four chapters are intended for researchers and others who want to understand the EDS methods in greater detail. The final chapter outlines outstanding challenges and research needs.

The report is organized as follows:

- Chapter 2 presents a series of frequently asked questions (FAQs) that provide the EDS user with some general understanding of the event detection approach utilized. These FAQs also explore some of the fundamental assumptions and constraints on event detection.
- Chapter 3 provides an introduction and background to the CANARY water quality event detection software.
- Chapter 4 provides a technical overview of the event detection algorithms and their use within CANARY.
- Chapter 5 summarizes results of testing and sensitivity analysis of the CANARY algorithms.
- Chapter 6 describes a pattern matching technique to reduce false positive alarms.
- Chapter 7 provides the basis for a distributed approach to event detection, fusing data from multiple monitoring stations.

2. A Discussion on Event Detection

This chapter attempts to address some commonly asked questions about the need for an event detection system (EDS), the role of an EDS in routine monitoring, and the ability of an EDS to perform well amidst great variability in water quality data.

Why Not Just Use Set Points (Thresholds)?

Currently, many water utilities utilize set points in order to set alarms on online water quality monitors. "Trigger levels" or "set points" are a simple way to identify when water quality parameters are outside of an expected range of values. For example, if the free chlorine levels drop below 0.6 mg/L or rise above 1.5 mg/L, it is outside the expected range as defined by a utility operator, and further investigation into the source of the anomalous water quality is warranted. Traditionally, water utilities use set points to identify changes in water quality parameters that are undesirable no matter what the cause. For example, free chlorine levels near zero are a problem that needs to be communicated immediately to an operator.

Set points provide alarms when the actual value of the water quality signal goes above or below the set point value. EDS tools are designed to identify water quality values that are significantly different from the background values whether or not they exceed the set point limits. The event detection algorithms in CANARY continuously adapt to changing water quality values and look for significant deviations from that changing background. As an example, the blue line in **Figure 2-1** shows water quality data from a particular monitoring station that routinely varies over time due to changes in the utility operations. The blue line represents normal behavior in the system, and reasonable set points of 0.5 and 3.5 mg/L are used. The magenta line in **Figure 2-1** shows an example of a water quality event that would not be detected by set points; however, it nearly triples the value of the background water quality for those time steps. This significant change in water quality would be detected by an EDS. In other words, some contamination incidents might not cause water quality parameters to move outside of set point boundaries, but still cause significant changes in water quality and will be detected by an EDS.

This example points out a key tenet of effective event detection: the changes of interest are not just changes in the absolute values of the water quality but *changes relative to the water quality values expected to occur at the specific location and time.* EDS algorithms can improve upon the use of fixed threshold values in many situations.

The use of multiple sensors at a monitoring station also complicates the simple set point approach to event detection. As demonstrated schematically in **Figure 2-2,** a change in water quality occurs that affects all three water quality sensors (e.g., pH, chlorine, and total organic carbon), yet only one of those sensors (Signal 1) registers a change that exceeds a set point threshold.

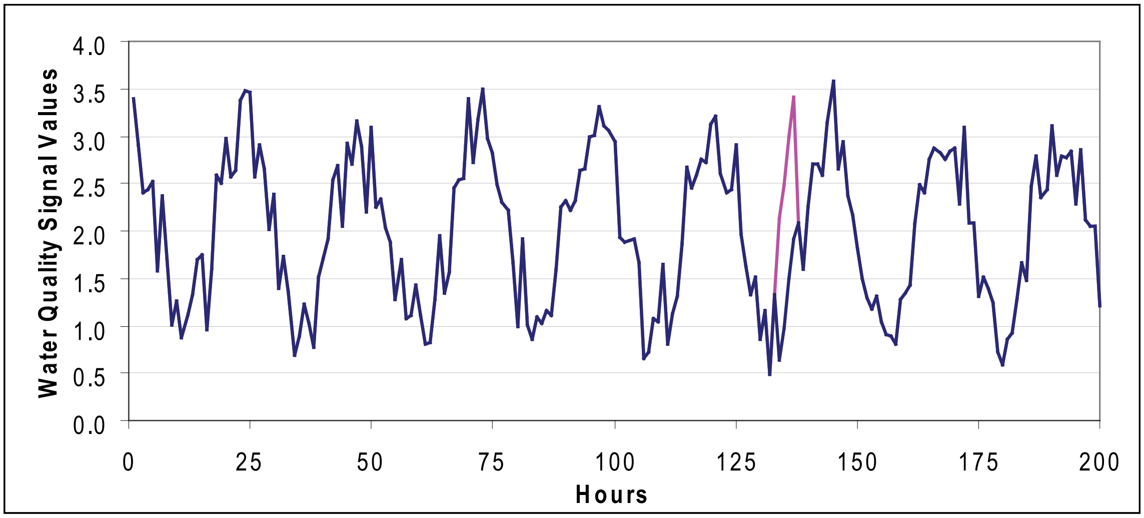


Figure 2-1. Background water quality signal (blue) with superimposed water quality event (magenta).

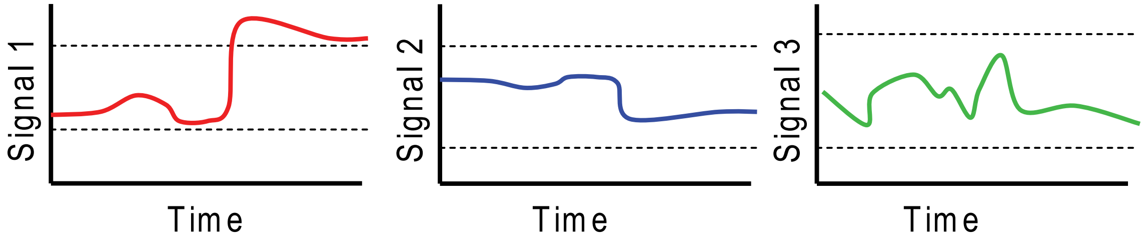


Figure 2-2. Schematic diagram of changes in three different water quality signals over time. The dashed lines represent set points for each signal.

The water quality response demonstrated in **Figure 2-2** would only be recognized by a set point in Signal 1. The strong relative changes in Signals 2 and 3 would go unrecognized and would not provide any information to better identify the cause of the water quality change. In contrast, an EDS tool examines relative changes in *all three* signals in order to detect a water quality event. In general, EDS tools can identify relative changes in water quality while simultaneously incorporating set point limits into the EDS algorithms.

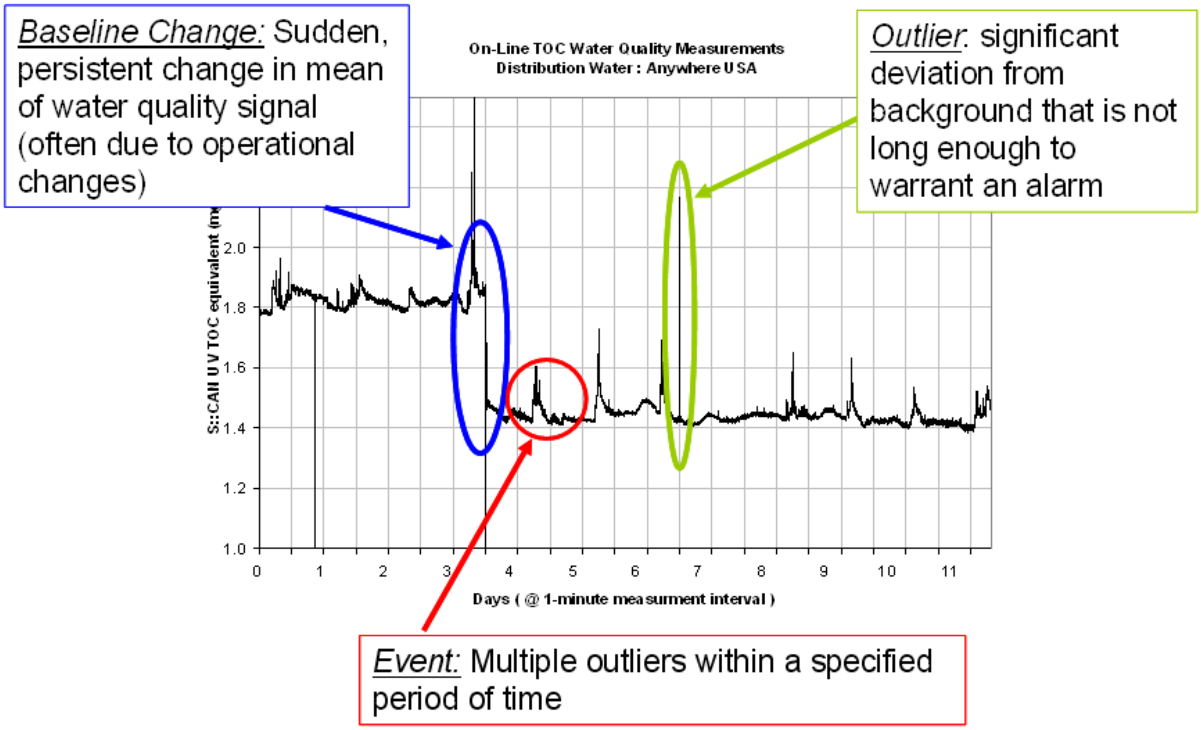
What Can One Expect to Find With an EDS?

Experience in examining water quality data from a number of utilities and multiple monitoring stations at each utility has shown that a variety of different types of anomalous signals do occur in utility water quality datasets. The root cause of these different types of anomalous signals is not always clear and is often specific to the utility and the monitoring station. Example causes include one or more of the following: changes in hydraulic operations at the utility, recalibration of sensors, missing or spurious data within the Supervisory Control And Data Acquisition (SCADA) system, unexpected failure in utility infrastructure such as a break in a main or failure of a booster station, water quality sensor malfunction, or other relatively common acts that can cause a change in water quality.

These commonly seen water quality changes are classified by the CANARY EDS into several different groups and examples of these groups are shown in **Figure 2-3.** The terms used for each group are:

- *Baseline Change:* A sudden and persistent change in the mean value of a water quality signal over several hours or more of data. Baseline changes are often due to changes in utility operations that cause water with different quality characteristics to flow through the monitoring station. For example, a baseline change could be caused by a valve or a pump being turned on or off that delivers water with a different background water quality or the draining of a nearby tank with water of a different age than was previously at the monitoring station.
- *Outlier:* An unexpected value in water quality at a single time step. Water quality values are estimated ahead of the measurement and if there is a significant difference between the estimated and observed water quality values, that time step is considered as an outlier. Nearly every water quality data set examined by the authors has outliers similar to those shown in **Figure 2-3.** These are single time steps where the water quality value suddenly rises or falls and then returns back to the expected value again in the next time step. In general, these types of outliers are thought to be due to noise in the SCADA system and should not be the cause of an alarm from an EDS tool. CANARY allows the user to determine how many outliers are needed prior to declaring a water quality event.
- *Event:* Measured water quality values that are significantly different from the expected water quality values for at least a specified minimum number of time steps. This definition does not define the cause of the water quality event. The number of time steps and the level of difference from the expected water quality needed to declare an "event" will often vary from one monitoring station to another within a utility and these parameters are adjustable by the user within the CANARY software.

In general, an outlier is a single time step with water quality that is significantly different than expected. Both events and baseline changes are groups, or clusters, of outliers that occur within a specified time frame. The major difference between a baseline change and an event is the length of time over which the outliers occur. The early stages of a baseline change are no different than an event, and the water quality analyst will often need to look at other information, most notably network operations information, to discriminate a baseline change from an event.



Baseline Change: Sudden, persistent change in mean of water quality signal (often due to operational changes)

Outlier: significant deviation from background that is not long enough to warrant an alarm

Event: Multiple outliers within a specified period of time

Figure 2-3. Examples of different types of changes in a water quality signal. Only a single example of each type of change is highlighted in the figure.

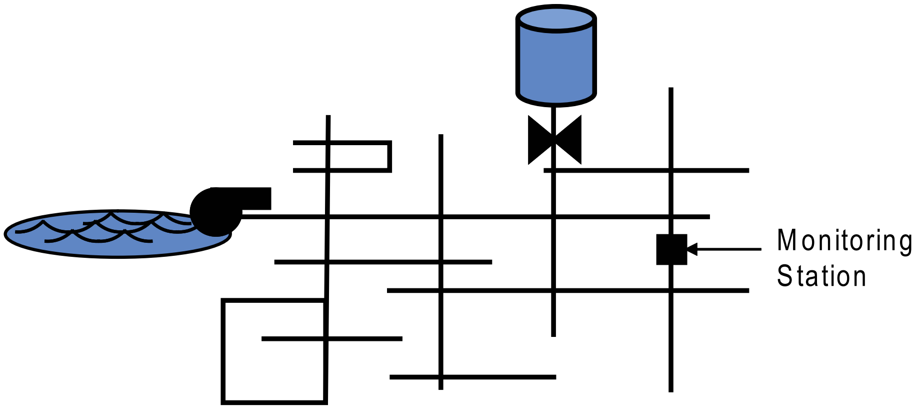


Figure 2-4. Schematic network containing two sources of water: reservoir and tank.

What About Variation in Water Quality and Regular Changes Due to Operations?

Online continuous monitoring at many utilities shows that there are significant variations in water quality that can be linked to operation of the utility. The underlying mechanisms can be demonstrated using a schematic network in **Figure 2-4.** Here a single water quality monitoring station can receive water from two different sources: the tank and the reservoir. The amount of each type of water at the monitoring station for any given time is a function of the valve setting at the tank and the pump operations at the reservoir. Both of these variables are known and recorded by utilities, but the uncertain, random quantity and spatial distribution of water demands between the two water sources and the monitoring station make it nearly impossible to predict the exact mix of the two waters reaching the monitoring station at any time. This situation is further complicated by the network having more than one pathway from each water source to the monitoring station. For the same reasons that it is difficult to predict the ratio of the two water sources arriving at the monitoring station, it is also difficult to track the effects of a change in water quality operations from the point where it occurs to the monitoring station. A closing of the tank valve would mean that the entire network will now be supplied solely by the reservoir, yet the time needed for the remaining tank water in the network to be removed through demand and the time it takes for the reservoir water to comprise 100% of the water at the monitoring station remain unknown, again due to uncertain demands.

Figure 2-5 shows roughly one week of water quality data observed at one monitoring station within a large municipal distribution network. Free chlorine (Cl) and pH data were collected at a sample interval of 2 minutes and 5000 measurements are shown for a total of one week. No water quality events are known to have occurred during this period and thus all data is assumed to be indicative of background water quality conditions. The chlorine data (green line) show relatively long (duration of approximately 200 time step, or 400 minutes) increases of roughly 0.25 mg/L spread throughout the observed data with seven of these increases occurring during this week. These increases in chlorine are matched by changes in pH (black line). At the beginning of the week, the pH decreases when Cl increases but by the end of the week, the increases in Cl are coupled with increases of 0.1 in pH. Some shorter-term changes of approximately 50 time steps (100 minutes) in length where Cl increases and pH decreases are also present. These shorter-term changes only occur in the first half of the data shown. **Figure 2-5** demonstrates the magnitude of the changes that can be caused by utility operations, in this case pumps turning on and off, as well as the complexity of these changes.

EDS algorithms can be trained to expect variations such as those shown in **Figure 2-5.** Once the EDS has learned that these variations are to be expected on a regular basis, the EDS will no longer alarm during these regular periods of change (see Chapter 6 for a more detailed discussion). In addition, judicious selection of parameters within the EDS software can be used to reduce false positive event detections in these circumstances.

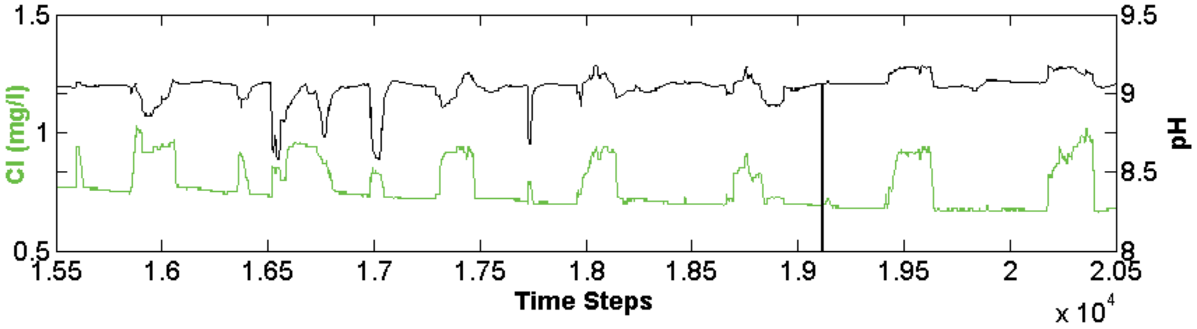


Figure 2-5. Example of water quality data changes caused by operations changes.

3. Canary Software: The Basics

An introduction to the CANARY software, how it operates, and some key parameters are discussed in this chapter. Additional technical details on all of these aspects are contained within the later chapters of this report, and can be found in the CANARY User's Manual (Hart et al. 2009).

What Is CANARY?

The CANARY software reads water quality monitoring data in real-time from a utility's Supervisory Control and Data Acquisition (SCADA) system. Water quality event detection algorithms within CANARY automatically identify significant deviations from expected water quality values at each time step for which the sensors provide data.

CANARY has been developed at Sandia National Laboratories in collaboration with and with funding from EPA's National Homeland Security Research Center (NHSRC). CANARY is distributed as open-source software and was first released publicly at the American Water Works Association (AWWA) Water Security Congress in early April of 2008. CANARY implements statistical algorithms for estimating the expected value of the water quality, state estimation, and subsequent classification of the residuals between measured and predicted water quality values. More detail and examples of using these algorithms are provided in Chapters 4, 5 and 6 of this document. CANARY is designed to be extensible, allowing researchers to modify existing algorithms or incorporate new algorithms easily. CANARY has been deployed at Greater Cincinnati Water Works (GCWW), the first EPA Water Security (WS) Initiative pilot utility, for over a year. In addition, several other U.S. water utilities are planning to deploy CANARY in the near future. Sandia National Laboratories is currently also working with Singapore national water utility, PUB to deploy CANARY. For more information about obtaining CANARY see: http://www.epa.gov/nhsrc/water/teva.html.

CANARY reads in time series data to identify anomalous water quality events. CANARY can read data from any sensor manufacturer for any type of water measurements and any number of sensors. Typical applications have included five to seven sensors, including some combination of free chlorine, pH, specific conductivity, total organic carbon (TOC), oxidation reduction potential (ORP), temperature, and turbidity.

Water quality monitoring data are transmitted through a SCADA system to a central database. CANARY is capable of linking to this database directly or through third-party interface software. CANARY can then gather and read the data in real-time. Experience to date shows that most utilities use a sampling interval of somewhere between 2 to 15 minutes as dictated by the particular requirements and SCADA system of each utility. CANARY processes these data in real-time and outputs the probability of a water quality event occurring at that monitoring station. This probability value can be transmitted back to the SCADA system for storage and display on the network operator's console.

CANARY is operated through use of an Extensible Markup Language (XML) formatted text file known as a configuration file. All data inputs and event detection algorithm parameters are defined in this configuration file. Sensor and SCADA alarm flags are defined in this file such that data received during a sensor hardware failure or during a manual calibration of the sensor are automatically recognized as not being quality data and are ignored in the event detection algorithms. To assist users, a configuration file editor with a graphical user interface is distributed as part of CANARY and is detailed in the CANARY User's Manual (Hart et al. 2009).

An example of event detection by CANARY is provided in **Figure 3-1.** In this example, four water quality signals from a real water system that have been modified to simulate an event are used as input to CANARY: free chlorine, pH, specific conductivity and temperature as shown in **Figure 3-1a** (Allgeier et al. 2008). The chlorine signal is exceedingly stable during this time period while the temperature and conductivity signals are variable. **Figure 3-1b** shows the probability of an event as predicted by CANARY (blue squares) and the relative concentration of the contaminant as it moves past the monitoring station (magenta line). Examination of **Figure 3-1b** shows that CANARY detects the contamination event approximately 15 time steps (30 minutes) after the contamination arrives at the monitoring station. The probability of an event occurring as calculated by CANARY rapidly increases to a value of 1.0 over approximately 15 time steps and stays at 1.0 until the event passes the monitoring station. The lag time between the arrival of the contamination and the increase in the calculated probability of an event is determined by the user-defined parameter settings in CANARY. Integrating results over greater numbers of time steps prior to increasing the probability of event generally results in a fewer number of false positive detections, but at the expense of increasing the delay between arrival of an event and declaration of that event.

Figure 3-1 also demonstrates a key ability of CANARY in that it continues to calculate a probability of event equal to zero prior to the time of the true event even though the water quality signals display considerable variation during this period. This ability is essential for reducing false event detections when presented with noisy data typical of most water quality monitoring networks.

CANARY provides the probability of an event at each time step for each monitoring station. These probabilities

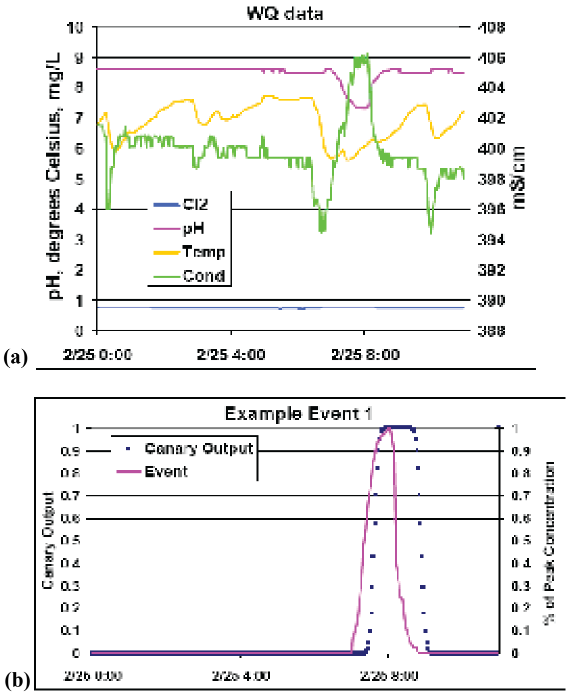


Figure 3-1. Example event detection using CANARY; (a) shows the water quality data and (b) shows the relative concentration of the contaminant and the CANARY output, which is the probability of an event.

are calculated independently for each monitoring station. The number of monitoring stations that can be analyzed simultaneously in real-time is also theoretically unlimited. At the pilot utility, CANARY is running on a single processor desktop computer reading in real-time data across 15 to 20 monitoring stations. Each monitoring station includes four or five water quality signals sampling on a 2 minute interval. Experience shows that CANARY could handle many additional sensors and still achieve real-time performance.

CANARY provides a stable software platform on which to extend the event detection algorithms to incorporate utility operations data. All of the necessary connections to real-time SCADA systems, the ability to incorporate sensor alarms, the text file user interface, and the experiences gained through previous deployment of CANARY will facilitate efficient extension and testing of the algorithms within CANARY to incorporate network operations data.

How Does CANARY Work?

CANARY works by reading in real-time (online) or historical data (offline), using a set of algorithms to analyze the data, and returning the probability of an anomalous water quality event. All of the algorithms within CANARY are based on the premise that past water quality observations can be used to accurately predict future water quality values under normal conditions. Additionally, CANARY recognizes that not all of the past data contains useful information and therefore less emphasis is placed on some previous information through filtering. Event detection within CANARY can be conceptualized as examination of a signal (e.g., pH) to identify its component parts. For an observed water quality signal (*S*) coming from a sensor, the signal is composed of the background water quality (*B*), any deviation from that background (*D*) due to an anomalous event, such as a contamination process upstream of the sensor, and noise (*N*) inherent in the water quality monitoring system.

For the most part, there are no anomalous events, and therefore, the deviations away from the background values are zero and the observed signal is simply the value of the background water quality along with the noise inherent in the measurement. The algorithms in CANARY are designed to continuously update and learn the characteristics of this background water quality signal and then account for it when presented with a new water quality observation. Each component of the observed signal (*B, D,* and *N*) could be further dissected into various sub-components. For example, noise is due to sensor imprecision, drift in the instrument, and transmission errors in the SCADA system, but that level of detail is beyond the scope of this discussion (see Einfeld et al. 2008). This simple model of a water quality signal is sufficient to understand how CANARY works.

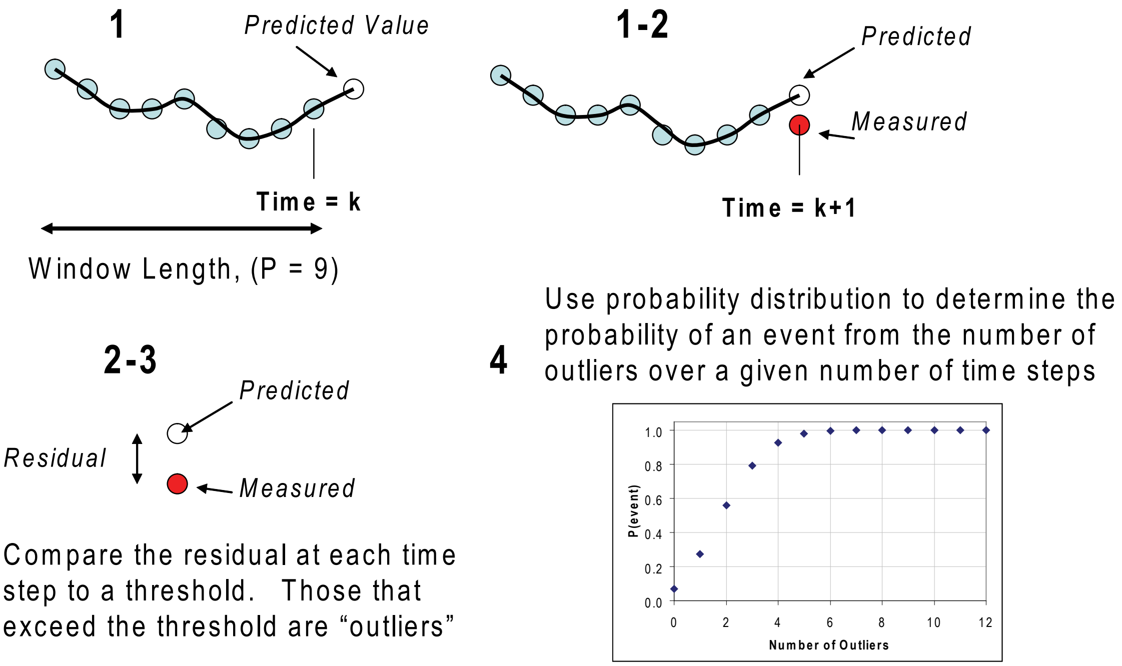
Four steps are involved in the event detection algorithms deployed in CANARY: 1) Estimation of the future water quality values; 2) Comparison of the estimated values against observed values as they become available and calculation of the “residual” as the difference between estimated and observed values; 3) Integration or “fusion” of the residuals across all water quality sensors at the monitoring station; and 4) Calculation of the probability of a water quality event occurring at each measurement time for each monitoring station from the residual data using a binomial event discriminator (BED). These steps are shown schematically in **Figure 3-2** and each step is defined in more detail below.

Step 1: Estimation

For each time series of data provided by a single sensor, CANARY looks across a pre-defined set of previous time steps and uses the data in this window to predict the value of the next time step. The data values within the window are first normalized to have a mean of zero and standard deviation of 1.0. This normalization removes the units of measurement so that the different signals with potentially very different units of measurement can be easily combined later. Two approaches to estimation are available within CANARY: linear filtering and multivariate nearest neighbor.

Linear filtering: At each time step, an optimal set of weights is determined to apply to each of the previously measured standardized observations for each water quality signal. The weights are calculated using an auto-covariance function computed independently for each signal. This calculation allows the assigned weights to reflect the importance of the previous value in the prediction of the next value no matter how far in the past that value has occurred. For example, in many systems where tanks are filled at night, then drained during the day, the free chlorine value observed 24 or even 48 hours ago often has greater bearing on the prediction of the current free chlorine value than does an observation from only 4 or 6 hours ago. These weights are calculated automatically within CANARY and are updated at each time step. The weighted average of the predefined set of previous values then serves as the prediction of the water quality value at the next time step. If the background water quality is perfectly understood and its characteristics do not change over time (no seasonal effects) and the noise in the system was zero, the linear filter algorithm would be able to perfectly predict the water quality value at each new time step. In such an ideal situation, the background signal would be completely accounted for, or filtered out of the observed signal, and any deviations from the background would be readily apparent.

Nearest neighbor lookup: The second approach to estimation also uses the water quality values at the predefined set of previous time steps. Grouping together water quality values from n different sensors at one monitoring station (e.g., if measuring for free chlorine, pH, and specific conductivity measurement, then $n = 3$), the set of values at each time step can be considered as a point in n-dimensional space. All of the data from previous time steps can be mapped as points in this space, and then the distance from any point to another can be calculated. At each new time step, a new point in n-dimensional space is created, and its "nearest neighbor," or the closest point in the previous set serves as the predicted value for this time step.



Compare the residual at each time step to a threshold. Those that exceed the threshold are "outliers"

Use probability distribution to determine the probability of an event from the number of outliers over a given number of time steps

Figure 3-2. Steps in the CANARY event detection process, including 1) estimation, 2) comparison, 3) residual classification, and 4) probability calculation.

Step 2: Comparison

As the observation at the current time step becomes available through the SCADA system, it is normalized using the same process as the background data. This new normalized observation is then compared to the predicted value and a residual (difference between the predicted and measured) value is calculated. The units of this residual are the number of standard deviations that the estimated value is away from the observed value. This process is repeated for each water quality sensor at the monitoring station. The end result of this comparison step is a residual value for each water quality sensor. For the linear filter approach, the residual values are in common units of standard deviations away from each respective estimated water quality signal value. In the multivariate nearest neighbor approach, the residual values are also in units of standard deviations. Because the residual is measured within the *n*-dimensional space, there is only one residual distance no matter how many different sensors are used.

Step 3: Residual Classification

The maximum residual value across all different water quality sensors at the monitoring station for the current time step is compared to a user defined threshold value, which is also in units of standard deviations. If this maximum residual value exceeds the threshold, the water quality at that time step is classified as an "outlier" and is excluded from the values used to predict water quality at the next time step. Other approaches to combining residuals have been examined including summing and averaging residuals, but it has been found that retaining the maximum residual for each time step provides the best overall results.

Step 4: Probability Calculation

Finally, the number of water quality outliers (observations that are significantly different from the estimated or expected) is analyzed to provide the probability of a water quality event existing at the current time step and in a given monitoring station. Experience with monitoring water quality data in several distribution systems has shown that outliers will occur with some frequency due to spreading of different types of source waters (e.g., ground water and surface water) throughout the network at different times, sensor hardware performance issues, as well as issues within the SCADA system.

Given that some number of outliers is to be expected, CANARY counts the number of outliers that occur in a user-specified time frame. For example, some outliers only last for a few time steps and could be due to SCADA communication failures or sensor malfunction. The number of outliers occurring within that time frame, along with knowledge of the likelihood of an outlier occurring under normal background water quality conditions, provides the necessary information for CANARY to calculate the probability of this number of outliers occurring under normal background conditions. A mathematical function called the binomial distribution is used to calculate a time-integrated probability of a water quality event (*P(event)*). The probability of an event (*P(event)*) is 1.0 minus the probability of background conditions prevailing (*P(backgrd)*).

As an example: if 12 time steps in a row are examined, and the chance of an outlier in any single one of those 12 time steps was 20%, then the chance of observing 8 outliers out of the 12 time steps examined would be extremely low. The probability that those observations are indicative of background conditions is extremely low. The binomial distribution can be used to predict the probability that these observations were caused by background water quality conditions, *P(backgrd)*, which can be quantified: *P(backgrd)* = 0.0005 (this calculation was done with Microsoft® Excel® software using the binomdist function). For this example, the probability of a water quality event *P(event)* causing these 8 outliers is 0.9995. In other words, it is so rare to observe 8 outliers in 12 time steps that the probability that this is a water quality event is nearly 100%. The graph in **Figure 3-2** demonstrates how *P(event)* changes as the number of outliers observed within 12 time steps increases from zero to 12 under the assumptions of the binomial distribution as discussed above.

The Binomial Event Discriminator (BED) within CANARY provides a means of combining the results of multiple successive time steps into a time-integrated probability of an event (*P(event)*). The BED employs the properties of the binomial distribution as described above to define *P(event)* as a function of the number of outliers within the BED integration window, the length of the BED integration window, and the probability of an outlier occurring at any given time step under an assumption of background water quality conditions (see McKenna et al. 2007 for more details). User specification of these parameters allows for maximum flexibility in the definition of an event and sensitivity of the event detection process. The drawback of integrating results over multiple time steps in this way is that there will be an additional lag time between the true onset of the event and the time at which CANARY detects an event (which depends on the time it takes the BED to increase the *P(event)* value). However, testing at partnering water utilities has shown that this disadvantage is significantly outweighed by the advantage of decreased false alarms that result from the BED integrating evidence for an event over multiple time steps.

How Accurate Are CANARY Predictions?

Historical water quality data from water utilities can be used to measure the accuracy of CANARY predictions. However, these data sets generally are not known to contain true water quality events of interest (i.e., contamination and cross-connections), as these types of events are generally quite rare in utility operations. Using such historical data, two performance measures can be derived: the accuracy of the water quality estimations made by CANARY and the number of false positive alarms. Both of these performance measures are explained below.

At each time step, CANARY uses previous data to estimate the value of each water quality signal at the next time step and then calculates the difference between the estimated and measured values. An example of the estimated water quality values compared to the actual measured water quality values is shown in **Figure 3-3.** The blue line is the historical data and the pink line is CANARY's estimate of water quality based on the data from previous time steps. **Figure 3-3** shows the estimated and measured values in the original units of the water quality signal (e.g., ppm). The results in **Figure 3-3** show that the algorithms within CANARY are able to estimate the next water quality value with high accuracy. The estimated signal value tracks the observed signal value closely through an increase of the chlorine values of 0.35 ppm (30%) over a three hour period. The estimated values also track the relatively quick 0.3 ppm decrease in chlorine at about 17:00. The largest difference between the estimated and observed values is seen at the end of this decrease and is approximately 0.07 ppm.

Within CANARY, the observed values are normalized so that the mean observation is zero and the standard deviation of the observations is one. Examples of the observations and estimates of the same data in this normalized scale are shown in **Figure 3-4.** For this example, the mean and standard deviation are calculated over a moving window that is 1440 time steps in length, which is much longer than the 180 time steps shown in the figure. The example in **Figure 3-4** shows chlorine values below the moving average (negative values) at both ends of the figure and a period of time where the chlorine values rise to more than 3 standard deviations above the moving average.

The pattern of observed and estimated values is identical to that shown in **Figure 3-3;** however, the values in **Figure 3-4** are now in units of standard deviation and they can be compared directly to any other water quality signal that has also been normalized. **Figure 3-4** also shows the residual values (grey line) calculated as: *residual = observed-estimated*. The residuals are also in units of standard deviation and provide the number of standard deviations between the observed and estimated values. The absolute value of each residual is compared against a threshold defined by the user. A typical threshold value is 1.0 standard deviation. In **Figure 3-4,** all residuals are less than this typical threshold value, with the largest residual (-0.83) occurring as the chlorine level drops near a time of 17:00. Therefore, none of the residuals in **Figure 3-4** exceed the threshold and none of the data are considered to be outliers.

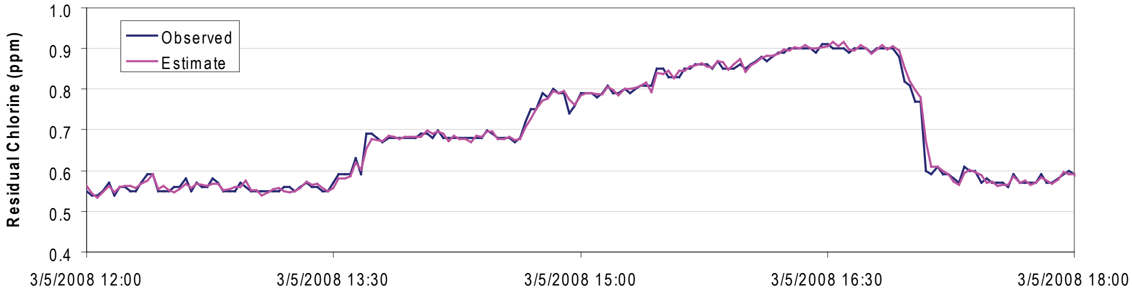


Figure 3-3. Observed (blue line) and estimated (pink line) chlorine values for a 6 hour period at a utility monitoring station. The sampling interval is 2 minutes and 6 hours (180 time steps) of data are shown.

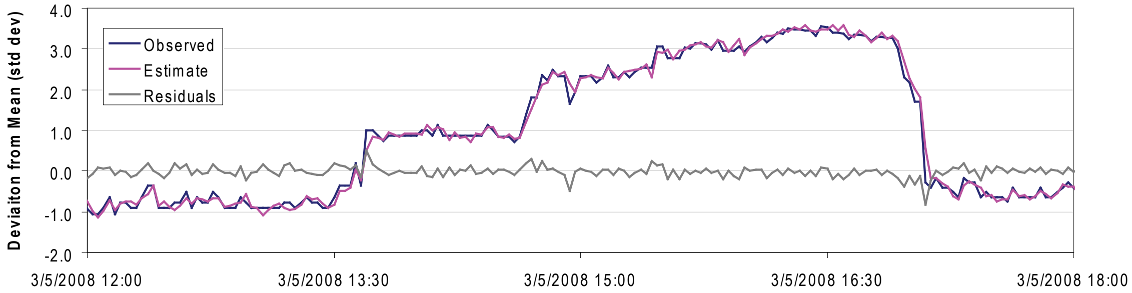


Figure 3-4. Observed and estimated chlorine values in standard normal space for a 6 hour period. The residuals between the observed and estimated data in standard normal space are also shown.

The residual values can be used to define the first performance measure that calculates how well the current parameter settings in CANARY are able to predict the water quality values. Here the average absolute error was calculated as the average of the absolute values of all residuals. For the data shown in **Figure 3-3** and **Figure 3-4,** the average estimation error is 0.094 standard deviations. Given that the fluctuations of the background water quality have a standard deviation of 1.0, this result indicates that CANARY is able to predict water quality with an average error of less than 0.1 the standard deviation of the background variation. This result reflects the ability of CANARY to not only track major trends in the water quality, but also predict the minor fluctuations in water quality around those trends. Both aspects of these prediction results are seen in **Figure 3-4.**

Figure 3-5 helps to explain the second performance measure: the false positive rate. If an event occurs, then the EDS tool can make the correct decision and detect the event (green box) or make a incorrect decision and not detect the event, so this is called the false negative (red box). Similarly, if no event occurs, then the EDS tool can make a correct decision by not detecting the event (green box) or make an incorrect decision by detecting an event and this is called the false positive (blue box).

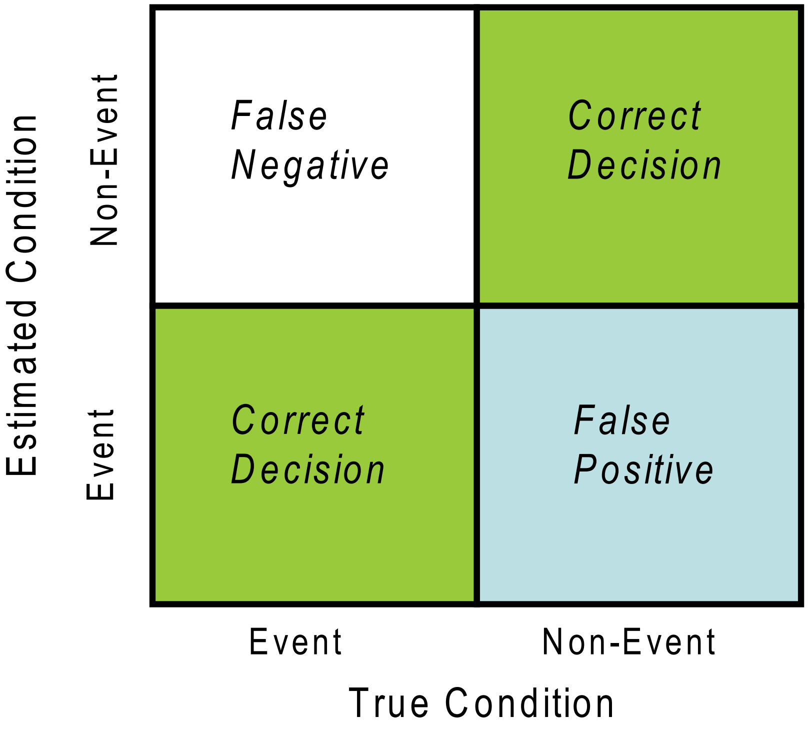
Typically, false positives are a nuisance for the utility operator, since the utility must then take steps to investigate whether it is a true event or not. The false negative results are more serious, since, in this case, a true event is not detected and the utility will not be aware of this situation. Using historical data with no known events, only the EDS results in the right column of **Figure 3-5** can be assessed.

To evaluate the performance of an EDS with respect to false negatives, it is necessary to have a water quality data set that contains actual events. Some historical data sets might contain a few water quality events resulting from main breaks or other unexpected factors that change water quality, but typically there are not enough of these events available to fully evaluate EDS performance. A solution to this lack of event data is to add simulated water quality events to the existing background water quality data. For example, the deviations in water quality due to contaminant injection, recorded by Hall et al. (2007), can be superimposed onto an existing historical water quality data set to provide events for testing.

Rather than count the number of individual time steps that are classified as false positives, the number of distinct clusters, or groups, of false positives could be used as a performance measure. Calculation of the mean prediction error and the number of clusters of false positives within a data set are the measures used to evaluate the performance of CANARY. As shown in the next FAQ, these performance measures can be used to determine some of the parameters needed for running CANARY on a data set.

How Are the CANARY Algorithm Parameters Set?

The algorithms used in CANARY rely on a number of user-defined parameters (see Chapter 6 of the CANARY User's Manual, Hart et al. 2009). The selection of these parameters will influence CANARY's performance as measured by false positive and false negative rates. Experimentation using historical water quality data allows for determining



		True Condition	
		Event	Non-Event
Estimated Condition	Non-Event	*False Negative*	*Correct Decision*
	Event	*Correct Decision*	*False Positive*

Figure 3-5. Possible event classification results when compared to the true condition.

the optimal parameter set for a desired false alarm (false positive) rate. This approach is demonstrated here and more detailed examples are given in Chapter 5 of this report.

For simplicity, only two of the parameters within CANARY are considered and adjusted here. These two parameters are: 1) the length of the time window used to predict the next water quality value, *P*; and 2) the value of the threshold (*thresh*) against which the largest residual is compared to determine whether or not it is an outlier in the residual classification process. In general, the larger the number of previous measurements included in the prediction of the future values, the more accurate that prediction will be. However, if *P* becomes too large, the computational load could increase to the point of making real-time estimation impractical. Also, the lower the value of *thresh*, the more sensitive CANARY will be to outliers in the data and the number of outliers identified and therefore the number of events identified will increase. Each identified outlier is then kept out of the history window used to estimate values of water quality at future time steps. If too many previous values are excluded from the estimations of future water quality, the accuracy of the predictions will degrade.

As an example of how to set these two parameters within CANARY, two water quality data sets are used to demonstrate how changing these two parameters affects the ability of CANARY to estimate the next water quality values and the number of false positive event declarations. As discussed above, in most practical settings, access to historical data containing a large number of actual water quality events is not possible. Therefore, the impact of various parameter settings on the accuracy and precision of the predicted water quality values as well as the number of false positive alarms are used to evaluate the parameter choices. This same approach would be used with historical water quality data in most practical applications of CANARY.

The data sets used here are from a large metropolitan water distribution network in the U.S. Four water quality parameters are examined: free chlorine (Cl), pH, total organic carbon (TOC), and specific conductivity (CDTY). For each monitoring station, a data set consisting of 28,082 time steps with a 2 minute sampling interval (approximately 39 days), collected between March 1st and April 9th is employed (**Figure 3-6** and **Figure 3-7**). Location A represents an environment with a relatively quiet background water quality, and Location B represents a monitoring station with significant changes in water quality due to daily changes in utility operations. Data sets from both monitoring stations have periods where the TOC values are missing. CANARY will keep running as long as there is one data signal available. It will calculate probabilities based on the available other signals, and as soon as TOC data is available it will include it in the analysis of probabilities.

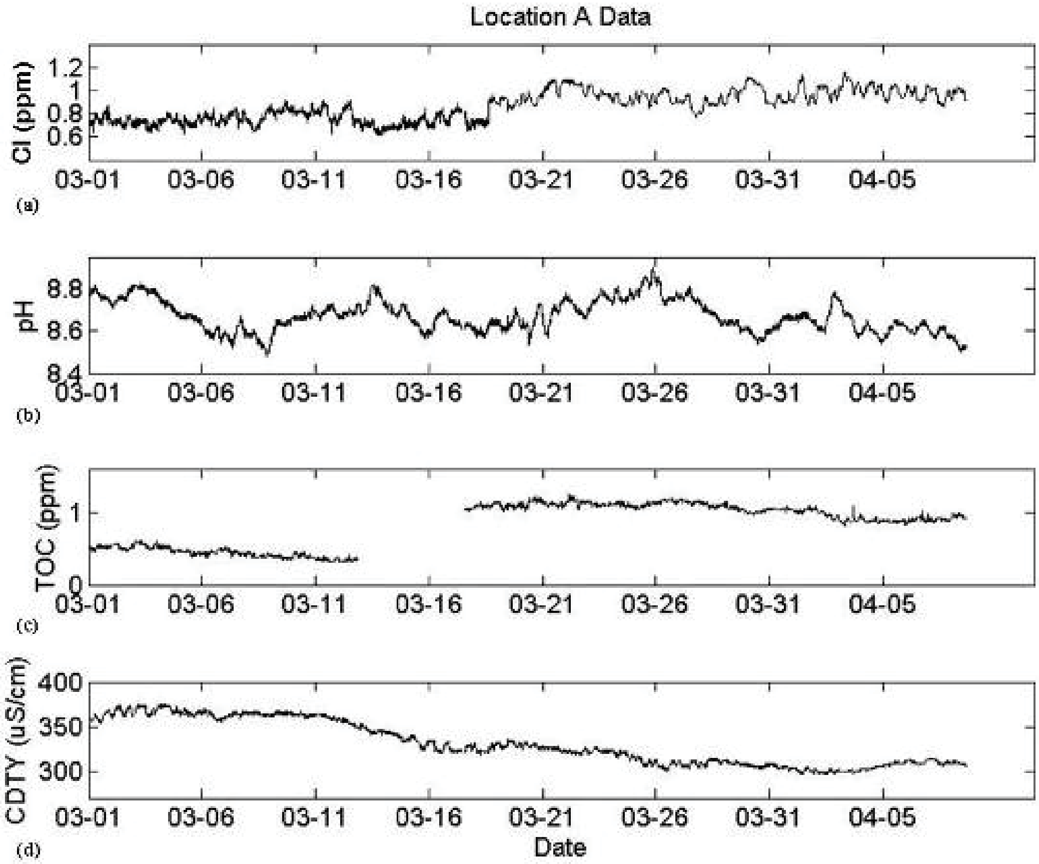


Figure 3-6. Water quality data from four water quality sensors as recorded at Location A. The water quality signals are (a) free chlorine (Cl), (b) pH, (c) total organic carbon (TOC), and (d) specific conductivity (CDTY).

Five levels of the time window are examined and five levels of the threshold value are also considered for 25 combinations of *P* and *thresh*. These parameter values are shown in **Table 3-1.** The parameters that are used in the binomial distribution to define the probability of an event are fixed here such that out of 25 consecutive time steps, 18 of them must be outliers to signal an event. If an event is identified, the next 125 time steps must be considered outliers as well before a baseline change in the water quality is declared. The parameters used in the calculation of the probability of an event and then declaring an event or a baseline change are held constant throughout this exercise.

Table 3-1. Parameter settings examined in example calculations.

Window Length, *P*, Time steps (days)	Threshold, "thresh" (std. dev)
360 (0.5)	0.6
720 (1)	0.8
1080 (1.5)	1.0
1440 (2)	1.2
1800 (2.5)	1.4

The results of the CANARY analysis on these data sets are presented in **Figure 3-8** and **Figure 3-9** for Locations A and B, respectively. Each of these figures shows the mean estimation residual (difference between the actual and predicted values) and the number of event clusters for each of the 25 parameter combinations. An event cluster is a continuous sequence of time steps which CANARY determines to be part of an event that is bounded on either end by periods of normal background water quality. Even though the available data are quite noisy, event clusters must be considered as false positives since no known adverse water quality events are present in these data. Several general trends occur in both figures. As the window length is increased, the mean prediction residual and the number of event clusters decreased for both monitoring stations. Also, for both stations, the changes in mean estimation error and number of event clusters are not a strong function of the threshold. The threshold could impact mean prediction error because time steps at which there are false alarms (residual is above threshold) are not used to predict future water quality values.

For Location A, the change in the mean prediction error is relatively small because both the window length and the threshold change. The maximum mean estimation error is 0.27 and the maximum number of clusters of false positives is 54 and these clusters contain 673 time steps, or 2.4% of all time steps examined. Both of these occur when a window length of 360 time steps (12 hours) is used. At Location A, all of the mean estimation errors are considerably less

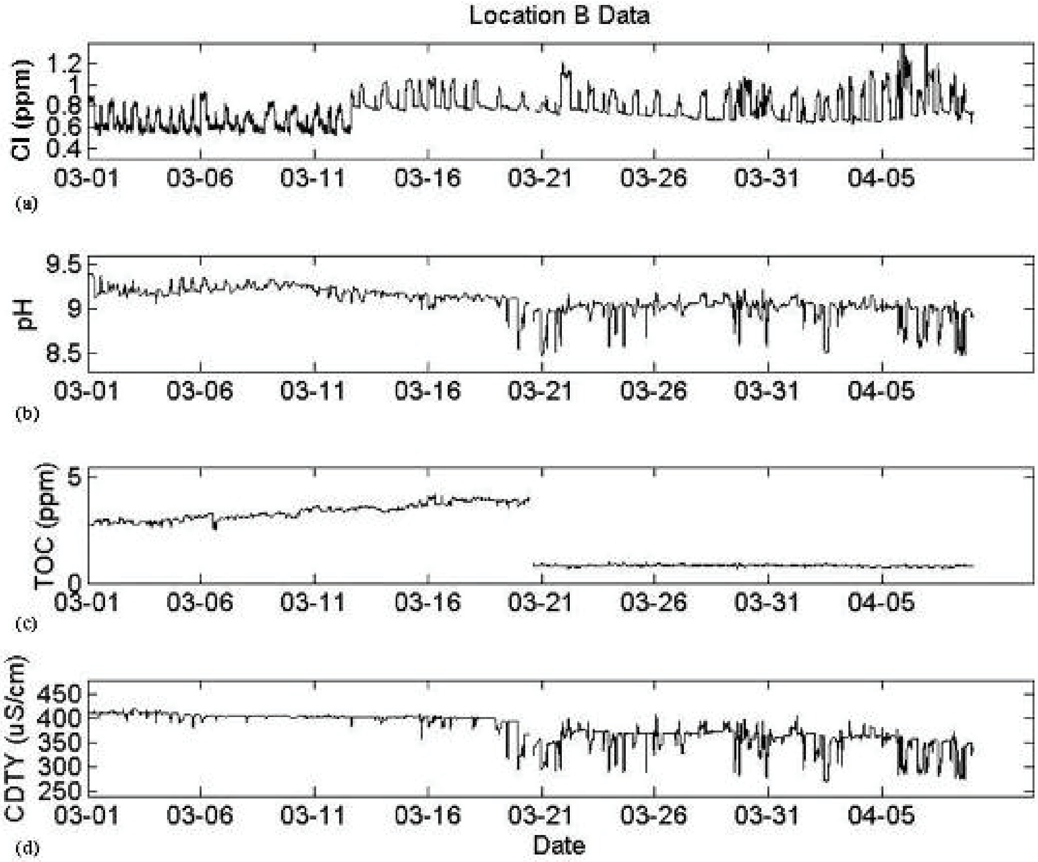
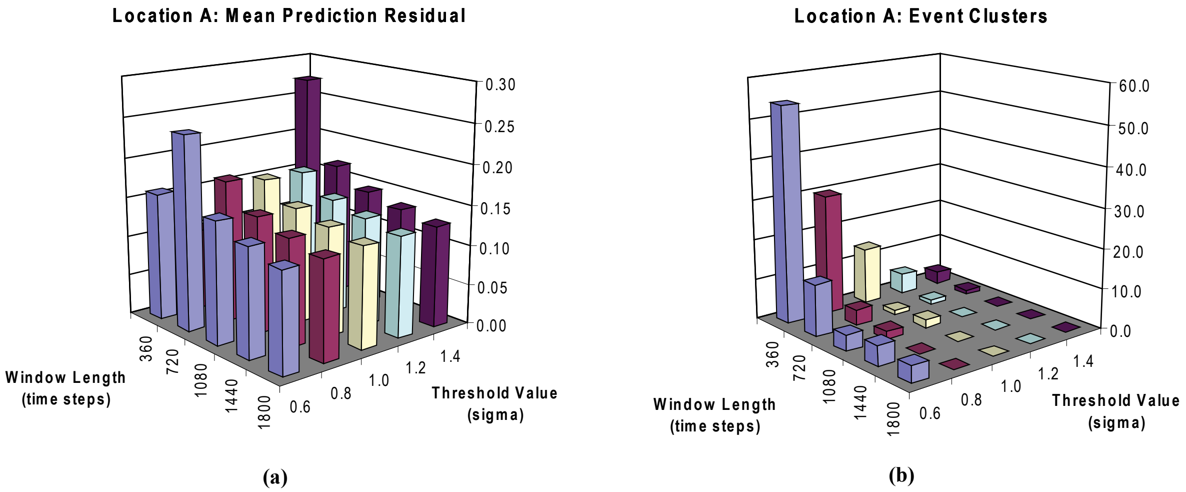


Figure 3-7. Water quality data from four water quality sensors as recorded at Location B. The water quality signals are (a) free chlorine (Cl), (b) pH, (c) total organic carbon (TOC), and (d) specific conductivity (CDTY).

than the thresholds used to identify outliers (0.6 through 1.4). Therefore, the threshold value has little effect on the mean error of the estimated values (**Figure 3-8a).** For window lengths of 1080 time steps (1.5 days) or more, mean estimation errors are less than 0.15 standard deviations for all thresholds. This accuracy in water quality value estimation provides for relatively sensitive event detection at this monitoring station.

Figure 3-8b shows that for Location A, a window length of 360 time steps (one-half of a day) is not adequate to reduce false positive results to an acceptable level. Additionally, a threshold value of 0.6 for any window length is too low to reduce the number of false positives. Window lengths of 1440 and 1800 time steps (2 and 2.5 days) result in zero false positives for any threshold value greater than 0.6. These results show that a number of different combinations of window length and threshold will result in zero false positive results at Location A for the time period examined.

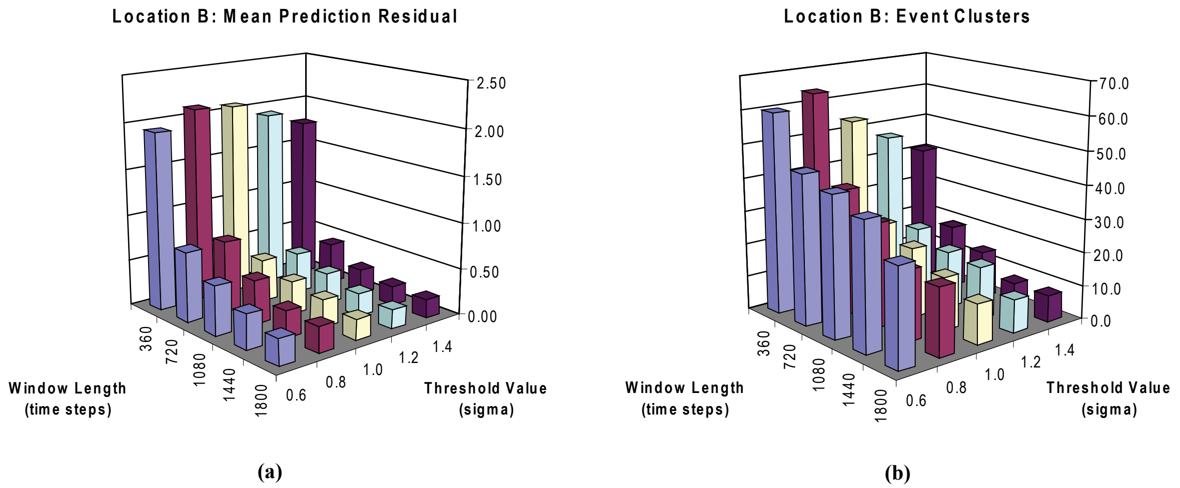
Results at Location B are considerably different than Location A due to the higher variability of the background water quality data at Location B. Note the scale used for the vertical axes in **Figure 3-9** and in **Figure 3-8.** The maximum average estimation error is 2.1 standard deviations. The maximum number of event clusters is 64, which contain 1913 time steps, or 6.8% of all time steps examined. Both the maximum average estimation error and the maximum number of event clusters occur when a window length of 360 time steps (12 hours) is used and therefore, this length is considered to be too short for use at Location B. At Location B it is not possible to reduce the number of false positive clusters to zero with these parameter settings. The minimum number of false positive clusters is 8 and this occurs when using a threshold value of 1.4 and a window length of 1440 or 1800 time steps. The mean estimation errors are 0.21 and 0.20 standard deviations respectively for these parameter combinations.



(a)

(b)

Figure 3-8. (a) Mean prediction error and (b) number of event clusters as a function of window size and threshold for Location A.



(a)

(b)

Figure 3-9. (a) Mean prediction error and (b) number of event clusters as a function of window size and threshold for Location B.

The approach to parameter selection and examination of the mean estimation errors and the number of event clusters demonstrated here provides the basics of how parameters can be estimated using historical water quality data. Currently, this approach to parameter estimation is done with a different run of CANARY for each parameter set. Future versions of CANARY will allow for automatic parameter estimation using historical water quality data using an approach similar to that described here, but with minimal user intervention. The results shown here are consistent with results obtained at other monitoring stations and in other distribution networks in that at least one day of data must be contained in the history window to minimize the mean prediction error.

What Are the Input/Output Options for CANARY?

Data can be input to CANARY through one of several formats:

- CSV files: SCADA data can be written to CSV (.csv) files that can be read directly by CANARY. Transfer of data using CSV formatted files is the most common path for using CANARY in offline mode to analyze historical data sets.
- Direct database connections: CANARY can connect directly to the SCADA database or through a third-party database connection tool such as EDDIES (U.S. EPA 2009a).
- XML files.
- Other software specific formats (for more information, see Chapter 4 of Hart et al. 2009).

No inherent limits exist within CANARY to the maximum number of water quality signals at any one monitoring station or to the number of unique monitoring stations that can be analyzed simultaneously. However, as the size of the data streams increases, the processing time also increases.

Outputs from CANARY can also be obtained in several formats:

- CSV files: CSV files that will be generated that include the following for each timestep: raw data output (for verification and capturing online values), estimation residuals, estimated probabilities of events, and event codes.
- HDF-5 formatted binary files.
- Database files: Connected SCADA databases will receive an event code, a probability of event, and an optional message with the name of the parameter(s) that could be responsible for causing the probability of an event to increase.
- Software specific files.

For more information about these output files, see Chapter 5 of Hart et al. (2009).

How Can CANARY Connect to a SCADA System?

Most utilities collect, transmit, and store water quality data using a SCADA system. Typical elements are the connections to the sensors through dedicated phone or Ethernet lines and, in some cases, through radio communications, the central receiving station for this information, and a database for storage of all water quality records.

Several third-party packages are designed to serve as middleware between a SCADA system and CANARY. One of these is the EDDIES system developed by EPA's Office of Water (U.S. EPA 2009a). This system acts as an intermediary between CANARY and SCADA, so that CANARY does not interact with the SCADA system directly. The information needed to connect CANARY to EDDIES is the same information needed to connect EDDIES to the database system. These third-party software interfaces are desirable when the utility maintains the SCADA database with a secure network and does not want to open that network to direct connections to EDS tools such as CANARY.

To directly connect to a database system, the user must have the proper Java™ JDBC™ Connector for the database software. This is typically available for no cost online from the database vendor. The Internet address for the database, as well as log-in credentials, will be needed to connect with the SCADA system database. Once the database is connected, the user will need to know:

- the database table where monitor values are stored
- the format of the table (are SCADA tags stored in a particular field or are the field names the same as the SCADA tags?)
- the names of the SCADA tags that are to be monitored at a given site

Due to timing issues, retrieving real-time data can be difficult. It is necessary to make sure that all data associated with a given time step are available to be read by CANARY. If CANARY is accessing the database at the same time the data are being written to the database from the SCADA system, there can be problems with the data transfer. To resolve this problem, it is advisable to set the clock slightly slow (a minute at most) compared to the database server, or ensure that CANARY is selecting data on, for example, the even minutes, while the database is being updated on odd minutes. Avoiding this data transfer problem is one benefit of using a middleware package like EDDIES, where the package handles the timing and timing messages rather than the system clock.

4. Event Detection Algorithms

Several algorithms for predicting water quality values have been developed and implemented within CANARY. Detailed descriptions of these algorithms are provided in this section. The residual time series resulting from the application of these algorithms are classified using a threshold comparison approach that takes into account the relative variability in the background water quality signal. The outcomes of the residual classification over multiple consecutive time steps are combined to provide a probability of an event at each time step using a binomial event discriminator (BED).

State Estimation Models

Three different state estimation models are implemented in the prediction algorithms and described in this chapter: time series increments, a linear filter, and a multivariate nearest neighbor algorithm. The linear filter and multivariate nearest neighbor algorithms have proven to be the most effective and, beyond the brief introduction below, the time series increments algorithm is not discussed further in this document.

Time Series Increments

The time series increments model is an implicit estimation model where the prediction of the value of a water quality parameter at the next time step, $\hat{z}(t+1)$, is simply the value measured at the previous time step, $z(t)$:

$$\hat{z}(t+1)=z(t) \qquad (4\text{-}1)$$

where t is time. Time series increments depend on only the single previous measurement and, thus, fit the definition of a Markov model (see Taylor et al. 1998). The time series increments, $\delta(t)$, are defined as the difference between the estimated and measured water quality parameter value at the next time step:

$$\delta(t+1)=\hat{z}(t+1)-z(t+1)=z(t)-z(t+1) \qquad (4\text{-}2)$$

These differences are calculated on standardized data (mean = 0 and standard deviation = 1) within a moving window of P previous measurements so that the units of the δ's are also in units of standard deviation. The value of δ, then, is the number of standard deviations the estimated value is away from the predicted value.

Linear Filter

The linear prediction-correction filter (LPCF) model uses a linear predictor to estimate the current value of a time series based on a weighted sum of past values. In its most general form, this approach is also known as an autoregressive (AR) model (Bras et al. 1993). The most common representation of the AR model is:

$$\hat{z}(t+1)=a_1z(t)+a_2z(t-1)+...+a_Pz(t-P+1)+\delta(t+1) \qquad (4\text{-}3)$$

or more compactly:

$$\hat{z}(t+1)=\sum_{i=1}^{P}a_iz(t-i+1)+\delta(t+1), \qquad (4\text{-}4)$$

where a_i are the estimation coefficients, P is the order of the estimation filter polynomial (number of previous measurements), and $\delta(t+1)$ is the estimation error. The error, or residual, generated by this estimate is

$$\delta(t+1)=z(t+1)-\hat{z}(t+1) \qquad (4\text{-}5)$$

where a mean-zero Gaussian distribution defines δ. Several methods are available to estimate the values of the parameters a. In CANARY, the autocorrelation method of AR modeling is used for such estimation. This formulation is set up as a linear system:

$$\boldsymbol{Za}\approx\boldsymbol{b} \qquad (4\text{-}6)$$

where $\boldsymbol{Z}$ is a function of time. Expansion of this equation results in:

$$\boldsymbol{Z}=\begin{bmatrix} z(t) & 0 & \cdots & 0 \\ z(t-1) & z(t) & \ddots & \vdots \\ \vdots & z(t-1) & \ddots & 0 \\ z(t-P+1) & \vdots & \ddots & z(t) \\ 0 & z(t-P+1) & \ddots & z(t-1) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & z(t-P+1) \end{bmatrix}, \quad \boldsymbol{a}=\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} \quad \boldsymbol{b}=\begin{bmatrix} z(t+1) \\ z(t) \\ \vdots \\ z(t-P+2) \end{bmatrix}$$

Here for online operation, all entries in the linear system are updated at every time step and use only the most recent P observations such that $\boldsymbol{Z}$ has dimension of P. Updating at every time step allows the coefficients, $\boldsymbol{a}$**,** to adapt to the changing water quality values contained in the moving window of previous values. To the extent possible within the AR model, non-stationarity and periodicity in the water quality data are captured by calculation of the appropriate coefficients at each time step. Note that this system of equations is solved separately for each water quality variable at each time step.

The solution that minimizes the estimation error through linear least squares is generally solved as:

$$\boldsymbol{a}=(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{b} \qquad (4\text{-}7)$$

where $\boldsymbol{Z}^{\mathrm{T}}$ is the transpose of $\boldsymbol{Z}$**.** The parameter estimation method exploits the fact that there is a direct correspondence between the parameters a and the correlation function of the water quality signals. Consequently, the Yule-Walker equations might be used to estimate the parameters by inverting such correspondence. Thus, the correlation coefficients, ρ, calculated from the P previous measurements provide a solution for the coefficients in $\boldsymbol{a}$

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{P-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{P-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{P-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{P-1} & \rho_{P-2} & \rho_{P-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \rho_P \end{bmatrix} \qquad (4\text{-}8)$$

The subscripts in **Equation 4-8** indicate the size of the lag spacing, in time steps, for which each correlation coefficient is calculated. Here the correlation coefficients are calculated in the frequency domain using an inverse and forward Fast Fourier Transform (FFT) on the previous P measurements of z. Since **Equation 4-8** is a Toeplitz matrix, the use of Levinson-Durbin (LD) recursion provides the most efficient solution for $\boldsymbol{a}$. Once $\boldsymbol{a}$ has been determined, it is inserted back into **Equation 4-4** and the current value of the signal is estimated.

Examination of the autocorrelation structure for several water quality time series provides an intuitive feel for how the values of the coefficients will vary for different water quality signals and for different monitoring stations. **Figure 4-1** shows two example plots of the correlation coefficient as a function of the number of time steps between sample values. A total of 2900 data values were examined with a 5 minute sampling interval (288 time steps per day). For parameter 1, the autocorrelation decreases linearly with increasing time between the samples. For parameter 2, periodicity in the water quality signal causes the correlation coefficient to decrease and increase with varying time lags between the sample data. For example, in **Figure 4-1b,** values separated by one day (288 time steps) are more strongly correlated than values separated by only 50-60 time steps.

Multivariate Nearest Neighbor

Another approach to state estimation that uses all water quality signals at each time step simultaneously to define the background state of the water quality is the multivariate nearest neighbor (MVNN) approach (see Klise et al. 2006a; Klise et al. 2006b).

For each time step, all J water quality signals are combined into a vector:

$$[z^{j=1}(t), z^{j=2}(t), z^{j=3}(t), \ldots, z^{j=J}(t)] = \mathbf{z}^J(t) \qquad (4\text{-}9)$$

The vector defines a point in the J-dimensional space at time t. If multivariate clustering is used to define K clusters, or classes, of water quality, the mean coordinate of the kth cluster in the J-dimensional space calculated over the previous P time steps is denoted by $\bar{\mathbf{z}}_k^J(t-P,t)$. **Figure 4-2** shows a schematic example of this calculation in J=3 dimensional space. The data in **Figure 4-2a** have been classified into five water quality classes and the extent of these classes are shown in **Figure 4-2b.** The distance between a new data point, red star in **Figure 4-2b,** and the centroid of each existing cluster is calculated as a Euclidean measure.

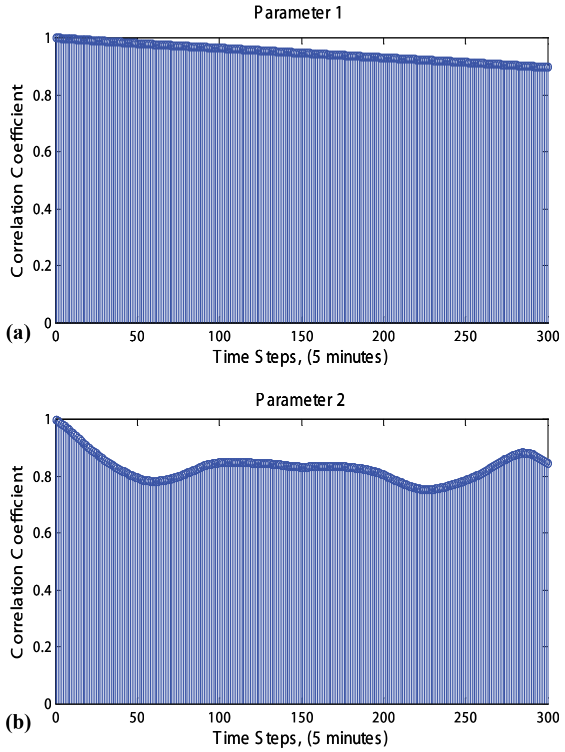


Figure 4-1. Example autocorrelation functions calculated for two different water quality parameters; (a) Parameter 1 and (b) Parameter 2.

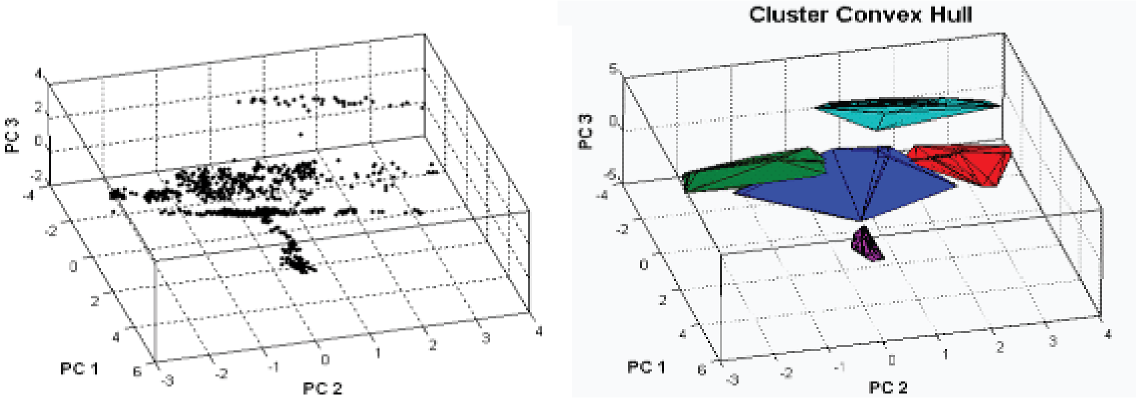


Figure 4-2. Example of data classification in three-dimensional space. The normalized data in (a) are classified into five clusters (b). A new data vector to be compared to the existing cluster centroids is shown by the red star in (b).

The MVNN approach does not provide an estimate of the water quality at a future time step. Instead, MVNN provides a measure of similarity of that sampled water quality with the *P* previously measured samples contained in the history window. The distance between the new water quality sample, $z^J(t+1)$, and the closest of the previous *P* water quality samples is measured as the Euclidean distance between the samples within the *J*-dimensional space. The minimum distance between the points is retained as the distance, Δ, which is compared to the threshold:

$$\Delta = Min_{i=1\ldots P}\left|\sqrt[J]{\sum_{j=1}^{J} z^j(t+1) - z^j(P-i+1)}\right| \qquad (4\text{-}10)$$

The Δ value can be calculated between the current water quality value and the mean locations of *K* previously defined clusters, or it can be calculated for every previous sample separately. Work by Klise and McKenna (2006b) demonstrated that as *K* was allowed to approach *P*, where each time step defined an individual cluster, event detection results improved. Contrary to the linear filter approach described above, the distance calculated with the MVNN is not a function of any individual water quality signal, but is a combined measure of the distance using all signals simultaneously.

Residual Classification

The residual (**Equation 4-5**) at each time step must be classified as either being consistent with background water quality or as an outlier. Comparison to a threshold that is a function of the standard deviation of the signal within the moving window allows for adaptive residual classification that is relative to the variation in the background water quality. Here the threshold is defined as the number of standard deviations away from the expected water quality value below which the estimated and measured water quality values are considered to be consistent with each other. The relative acceptable deviation from expected water quality is fixed to a specified number of standard deviations, although the absolute value of the threshold in raw units (e.g., ppm, NTU) can vary as background water quality values vary. As an example, a threshold value of 1.0 standard deviations applied to the chlorine signal might correspond to a threshold of 0.04 ppm during periods of relatively stable chlorine concentrations, or it might increase to 0.25 ppm throughout times of larger background water quality variation.

State estimation techniques that fuse all signals (e.g., the MVNN technique or approaches using cross-correlations between signals) generally result in a single estimate of the residual that is combined from all input signals. In the linear filter approach, independent state estimation results in a residual for each signal that must be combined, or "fused" in some way, to identify an outlier at that time step. The linear filter approach allows for determination of the specific sensor responsible for the outlier, whereas the sensor fusion within the MVNN approach does not. Residual classification using the maximum residual across all sensors makes it easy to record the sensor that is responsible for the outlier at each time step and this is the approach currently implemented in CANARY, e.g.,

$$\max_{j=1\ldots J}\left|\partial_j(t+1)\right| \qquad (4\text{-}11)$$

Binomial Event Discriminator (BED)

Previous application of outlier detection algorithms has been focused on classification of the water quality measurement vector (e.g., pH, chlorine, and total organic carbon) at every time step as either an event or background. One result of this approach was the large number of false alarms that are tied to significant, but very short-term changes in the water quality, including significant decreases and increases in consecutive time steps most likely due to noise in the SCADA system. The BED was developed to integrate events over multiple consecutive time steps before declaring the sequence of time steps to be a true event, background water quality, or a change in the baseline of the background water quality (**Figure 2-3**). The BED works on the results of any event detection algorithm that produces a binary result (success/failure) for every time step. The BED provides an additional filtering of the data after the LPCF or MVNN algorithms and decreases the impact of any one time step that provides unexpected data.

The result of any outlier detection algorithm is conceptualized to define any time step with an outlier as a "failure" and any other with a residual consistent with background quality as a "success." The binomial probability distribution gives the probability that r "failures" occur in n trials, when the expected probability of any one trial failing is p. The corresponding probability that any one trial will succeed is $q = 1\text{-}p$.

The probability that the water quality observed in the n trials is indicative of background water quality conditions is *P(backgrd) = b(r;n,p)* and is given by **Equation 4-12.** The complementary probability of an anomalous water quality event occurring within n trials is *P(event)* = 1.0 – *b(r;n,p).*

$$b(r;n,p) = \frac{n!}{r!(n-r)!} p^r (1-p)^{(n-r)} = \frac{n!}{r!(n-r)!} p^r q^{(n-r)} \qquad (4\text{-}12)$$

In online analysis, the concern is that the number of failures within a specified time period increases towards the positive tail of the binomial distribution of failures. This relatively large number of failures would indicate a rather unlikely occurrence of anomalies coming from the background. To more efficiently identify such sequences of events, the cumulative distribution function (cdf) of the binomial distribution is used:

$$P(r \le z_c) = \sum_{i=1}^{n(r \le z_c)} b(r;n,p) \qquad (4\text{-}13)$$

where z_c is the probability threshold value. Using the cdf function ensures that the probability of an event is increasing as the number of failures increases.

The binomial probability distribution describes the outcome of a Bernoulli process which must have the following properties (Walpole et al. 1989):

1) n repeated trials in the experiment.
2) Each trial can only have one of two outcomes: success or failure.
3) The probability of failure, p, remains constant from one trial to the next.
4) Repeated trials are independent of one another.

Each time step for which water quality data are available is considered a trial. A user-defined window length within CANARY, *bed-window-TS*, defines the number of repeated trials (the n time steps) that are input to the BED

The outlier detection algorithms (time series increments, linear filter, or MVNN) are designed to produce a sequence of binary flags (0/1 indicating whether the data from the time step is an outlier or not) as output, which fits the requirement of a Bernoulli process having only success or failure outcomes.

The third requirement for a Bernoulli process is that the probability of failure, p, remains constant from one trial to the next. The use of a threshold that is relative to the current, or recent, variation of the water quality signals maintains a constant failure rate independent of the variation in the water quality. This approach also allows for a much more efficient detection algorithm than can be obtained using a constant threshold such as the "set point" approach often employed (see comparison in McKenna et al. 2006b).

The fourth property of the Bernoulli process that must be met for application of the BED is the most restrictive: repeated trials are independent of one another. This requirement is equivalent to stating that the values of the estimation error, ε, are uncorrelated in time. Autocorrelation of the estimation errors can occur when the estimation algorithm tends to create a smoothed version of the observed water quality such that over or under prediction of the measured water quality could occur in sequences of consecutive time steps. The simulation of varying amounts of autocorrelation in error series has been examined and caused deviations from the expected binomial behavior.

Operation of the estimation algorithms under ideal conditions would result in ε being uncorrelated Gaussian noise and, for any z_c, the expected proportion of outliers could be determined from properties of the Gaussian distribution. This proportion corresponds to the probability of any single trial resulting in a failure, p, and could be used directly in the definition of the binomial parameters. However, experience has shown that serial correlation in the errors and other factors do not allow for this theoretical interpretation. In addition, experience at multiple testing stations has shown that by keeping $p = 0.50$ and altering both the size of the binomial window (*bed-window-TS*) and the probability threshold that must be exceeded to declare an event (*event-threshold-P*), a wide range of event detection sensitivity can be achieved.

Within CANARY, a second window length, *event-timeout-TS*, is defined as the number of consecutive time steps beyond the BED window, *bed-window-TS*, in which every time step must contain an outlier in order to identify a baseline change. The length of *event-timeout-TS* is not directly tied to the binomial probability distribution, but is set by the water quality analyst.

Testing Strategies

The central problem in fully evaluating EDS tools is that very few data sets exist where a contamination event that changed water quality is known to be recorded by sensors in the distribution network. Although it might be possible to find a few cases at different utilities where this has occurred, a database of such events that were large enough to develop a statistically significant number of results does not exist. An alternative approach to EDS testing is to simulate water quality events on top of previously recorded water quality data at a utility. This type of simulation can be done with varying levels of sophistication: from simply adding a square wave that increases/decreases the measured water quality values by a set amount for a prescribed time period; to defining and solving the chemical reactions between an injected contaminant and any substances in the bulk water or on the pipe-walls.

An intermediate approach that superimposed the measured responses of sensors to actual contaminants injected into a pipe loop onto previously recorded water quality data was utilized. A Microsoft® Excel® software-based tool has been developed that uses internal Microsoft® Excel® functions and Visual Basic® programming to provide a convenient means of superimposing any sensor response onto water quality data observed within a water distribution network. The water quality sensor responses to injected contaminants as recorded by EPA in the National Homeland Security Research Center (NHSRC) Test and Evaluation (T&E) Facility are included in the event simulator program. Hall et al. (2007) provide additional details on the experimental program that recorded such responses. The event simulator program considers the sensor response to an injected contaminant to be a template for the change in that water quality parameter and then places this response on top of observed water quality as input by the user. The response on top of the water quality is repeated multiple times as requested by the user. For best results, the user can input measured sensor responses as recorded from experiments using water and pipe materials that are specific to the utility of interest.

Five steps are involved in the event simulation process within the Microsoft® Excel® software-based tool:

1) "Load Data." The baseline water quality (WQ) data needs to be pasted into the columns of the *InputData* worksheet, in the order indicated by the column titles. Note that the order of the water quality variable columns must be kept the same as in the existing worksheet template. No blanks can exist in any column with data in it, as this will cause the calculation to stop at that point. If missing values need to be represented, use either "NaN" or "#VALUE!" instead of blanks. If the monitoring station does not have water quality data for all of the water quality sensors, those missing columns need to be filled with the missing value indicators.

2) "Create Pattern." This step defines the portion of the water quality pattern and the spacing between events to be superimposed onto the recorded water quality data. This portion must be done manually with 'Pattern Label' values of zero indicating no event, and pattern values other than zero indicating an event. Pattern values indicate the deviation from background water quality that results from the introduction of contaminant. Event time steps range from 1 (start of event) to 39 (end of event), where 39 is the maximum number of time steps in any experiment (Hall et al. 2009). The event might be lengthened or shortened. The water quality value at step 20, the midpoint of the event, generally defines the maximum deflection of the sensor away from the background values. Repeating the value at step 20 will lengthen the event, which kept the shape of the pattern at the beginning and end of the event consistent with those measured. A default pattern that is 20 time steps long, followed by 80 zeros denoting background water quality after the event, is provided in the *Pattern* worksheet.

3) "Select Contaminant." The event simulator program comes with a database of sensor responses as determined at EPA's T&E facility. Instrument responses are currently available for 15 different contaminants. Each contaminant was run with three different injection strengths and, for each, two replicates were completed. Therefore, a total of 90 (15 x 3 x 2) different experiments are available from which a pattern can be chosen. Some contaminants create a response only in one sensor, while others affect multiple sensors. In addition to the contaminant responses from the experiments, two additional synthetic patterns were added to the event simulation tool: the square wave and sawtooth (triangular) wave for both chlorine (Cl) and pH signals. Each synthetic wave applies to only a single sensor, and three different levels of maximum deflection can be chosen. Users of the event simulator are encouraged to add events from other experimental efforts to the event simulator database. **Figure 4-3** shows example deflections in four sensors from the introduction of a 2.2 ppm, 20 minute pulse of aldicarb (pesticide), 24.4 meters upstream of the monitoring station.

4) "Create Events." This is the calculation step where the selected contaminant response is superimposed onto the measured water quality data. The results are automatically pasted into the *OutputData* worksheet. These results contain a column of time step indices from 1 to *n*, where *n* is the total number of time steps, an event indicator column, where each time step is either a "0" for background or a "-1" for an event, and then one column for each of the modified water quality data.

5) "Save Data." The last step is to save the created data set to an external file. For example, these data can be saved to a CSV file that can then be modified by adding the correct header names and used as input to CANARY.

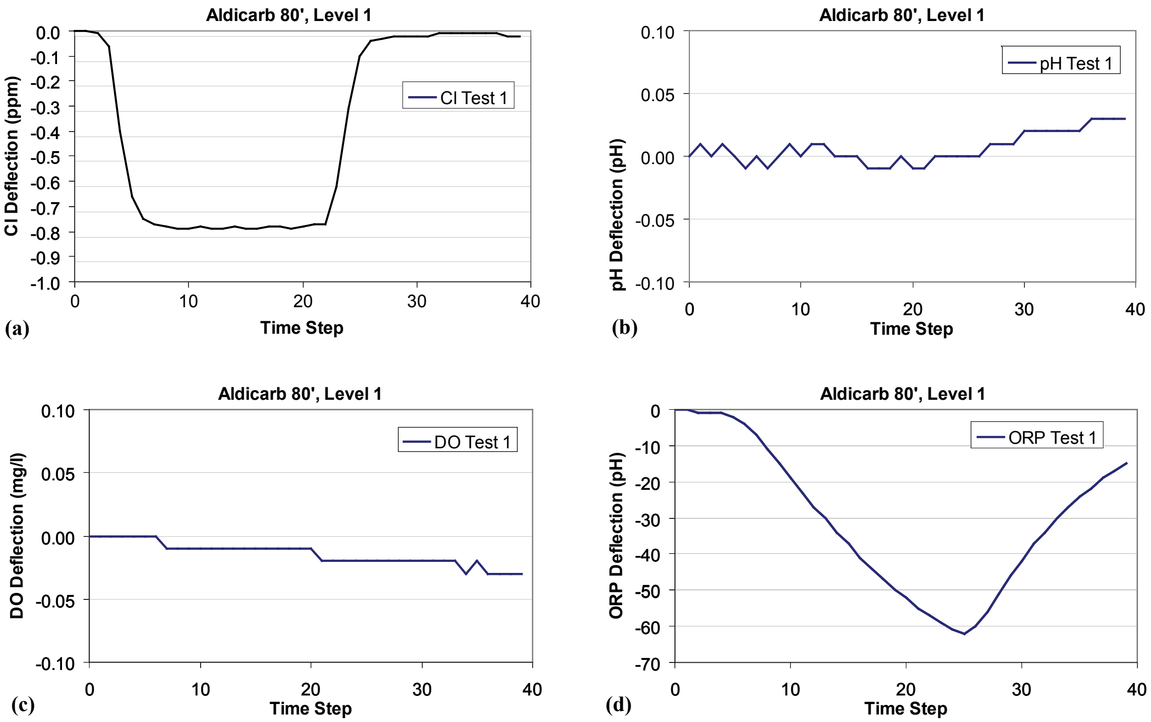


Figure 4-3. Example deflections of four water quality sensors from the introduction of a 2.2 ppm pulse of aldicarb, 24.4 meters upstream of the sensor station. The sensors are (a) free chlorine (Cl), (b) pH, (c) dissolved oxygen (DO), and (d) oxidation reduction potential (ORP).

As an example of this process, the Cl squarewave with a deflection of -0.15 ppm is added to an existing water quality data set. The data prior to adding the events are shown in **Figure 4-4** and the data set containing the events is shown in **Figure 4-5.** In this example, only the Cl data are affected: the pH and specific conductivity data are not changed. For this data set, the sampling interval is 5 minutes and each event is 24 time steps (2 hours) long. The events are spaced 1000 time steps apart, which means that the start of each event is 83 hours and 20 minutes after the start of the previous event. Ten events are visible in **Figure 4-5.**

The major advantage of the event simulator is that it is possible to superimpose the sensor responses of actual contamination events as recorded in a pipe loop onto water quality data collected within an operating water distribution network. These responses are most likely as close as possible to the actual responses of the sensors had that contaminant been injected into the distribution network. Each sensor response is discretized into 40 time steps and the shape and length of any sensor response can be varied by including or not including particular time steps of the response pattern. To make longer patterns representative of a longer injection period, central portions of the pattern can be repeated for as many time steps as necessary to extend the pattern to the desired length.

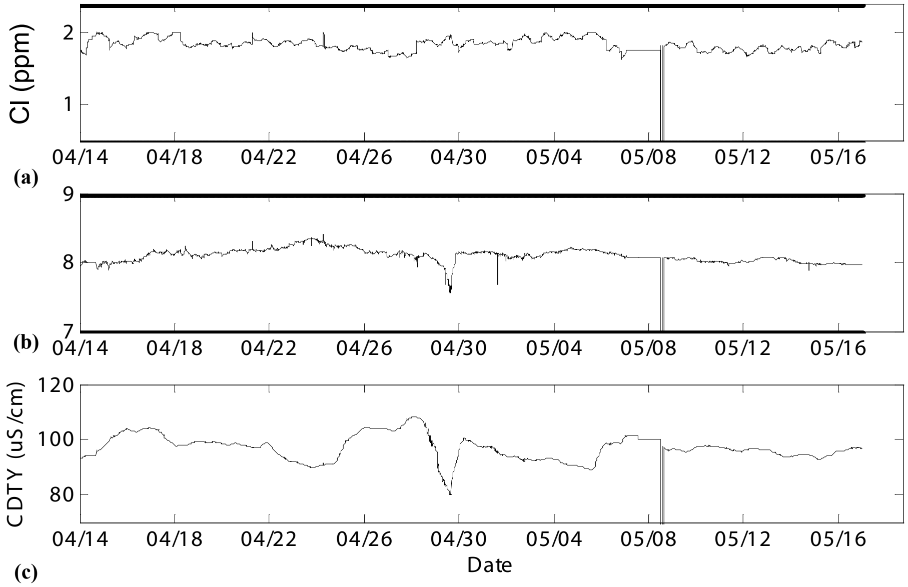


Figure 4-4. Example water quality data prior to adding events. The water quality data are (a) free chlorine (Cl), (b) pH, and (c) specific conductivity (CDTY).

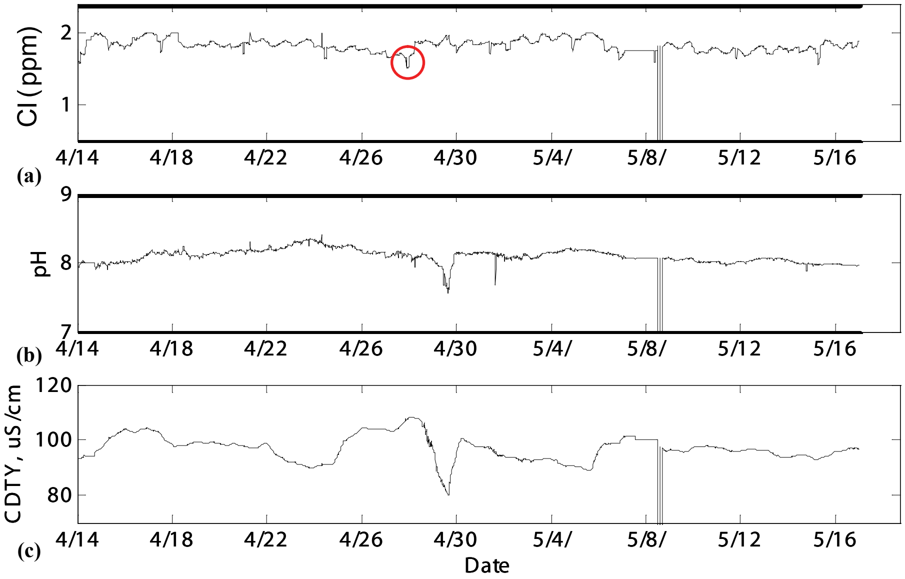


Figure 4-5. Example water quality data with square wave deflections in the Cl signal added. The water quality data are (a) free chlorine (Cl), (b) pH, and (c) specific conductivity (CDTY). Red circle indicates one of the added events.

The main assumption in using the event simulator is that the responses of the sensors as recorded at EPA's T&E facility would also be representative of the responses of the sensors when operating in the utility from which the water quality data were recorded. Background water quality at the utility might be different than that at EPA's T&E facility, and this could cause the actual sensor deflections at the utility to be different from those recorded. Ideally, this assumption would be validated for each application at a different utility, but that would require injection of these contaminants into the network, which is not feasible. A way to address this problem might be to repeat the experiments performed by EPA at any utility using local water. Then, the sensor responses recorded in those experiments might be added to the event simulator database.

A secondary assumption is that the sensor response patterns measured at EPA's T&E facility can be mapped onto the water quality time steps measured at the utility of interest. The experiments done at the T&E facility used a 1 minute sampling interval, whereas most utilities have a longer sampling interval, typically 5 minutes. It is possible to map only every fifth time step from the sensor response patterns onto the measured water quality data, but this will degrade the resolution of the water quality pattern. Another option is to assume that the time steps can be mapped one to one. For the typical longer sampling intervals at utilities, this results in slower changes in the sensor responses that should be consistent with a slower injection of the contaminant.

5. Canary Testing and Sensitivity Analysis

The performance of CANARY is particularly sensitive to two parameters: the window size and the threshold. This chapter provides a step-by-step example of how to select the values of these two parameters.

In what follows, both the linear prediction-correction filter (LPCF) and multivariate nearest neighbor (MVNN) prediction algorithms are used on data collected at three monitoring stations within an operating U.S. water distribution system. The stations were chosen because of their different water quality characteristics. Location A has a relatively stable background signal; Location B has similar characteristics to Location A, but with additional periodic variations; and Location C is strongly influenced by utility's operational changes.

The following water quality event detection issues are investigated:

1) Determination of appropriate event detection parameters from background data only (training).
2) Simulation of events with different contaminant concentrations for testing the detection capabilities of CANARY's algorithms.
3) Application of algorithms with parameters identified in training step for detection of events added to the background water quality data (testing).
4) Detailed examination of the events (false alarms and actual events) identified by the CANARY algorithms.
5) Evaluation of different parameterizations and the effects on event detection and baseline change identification (sensitivity analysis).

Available Data Sets

As mentioned above, the three monitoring stations are labeled “A,” “B,” and “C” and were selected to provide three distinctly different sets of water quality data for training and testing. For each monitoring station, there are 31 days (22,320 time steps at 2 minute intervals) of training data from July 8th through August 7th. Each station has four water quality signals: chlorine (Cl), pH, conductivity (CDTY), and total organic carbon (TOC). These training data are shown in **Figure 5-1** through **Figure 5-3.**

The stability of the background water quality of Location A is noticed in the signals shown in **Figure 5-1.** The signals vary only gradually throughout the training data period with the exception of a sharp change in pH on July 11th and a sharp change in TOC on July 29th. Location B is another example of a monitoring station with a relatively stable background water quality (**Figure 5-2**)**.** The signals of Location B also exhibit more regular daily periodicity relative to Location A (note that the water quality axis are different in **Figures 5-2** and **5-3** because the dynamics at each monitoring station are so different). As shown in **Figure 5-3** the water quality at Location C is strongly impacted by utility operations. The strong daily periodicity in the chlorine, pH, and conductivity signals are caused by water from different sources passing the monitoring station each day.

The training data are used to identify the parameter settings in the event detection algorithms. These algorithms and parameters are then applied to a second set of testing data. Water quality events of varying strengths are added to these testing data sets to evaluate the event detection algorithms. It is assumed throughout these steps that the characteristics of the background water quality do not change between the training and testing data sets.

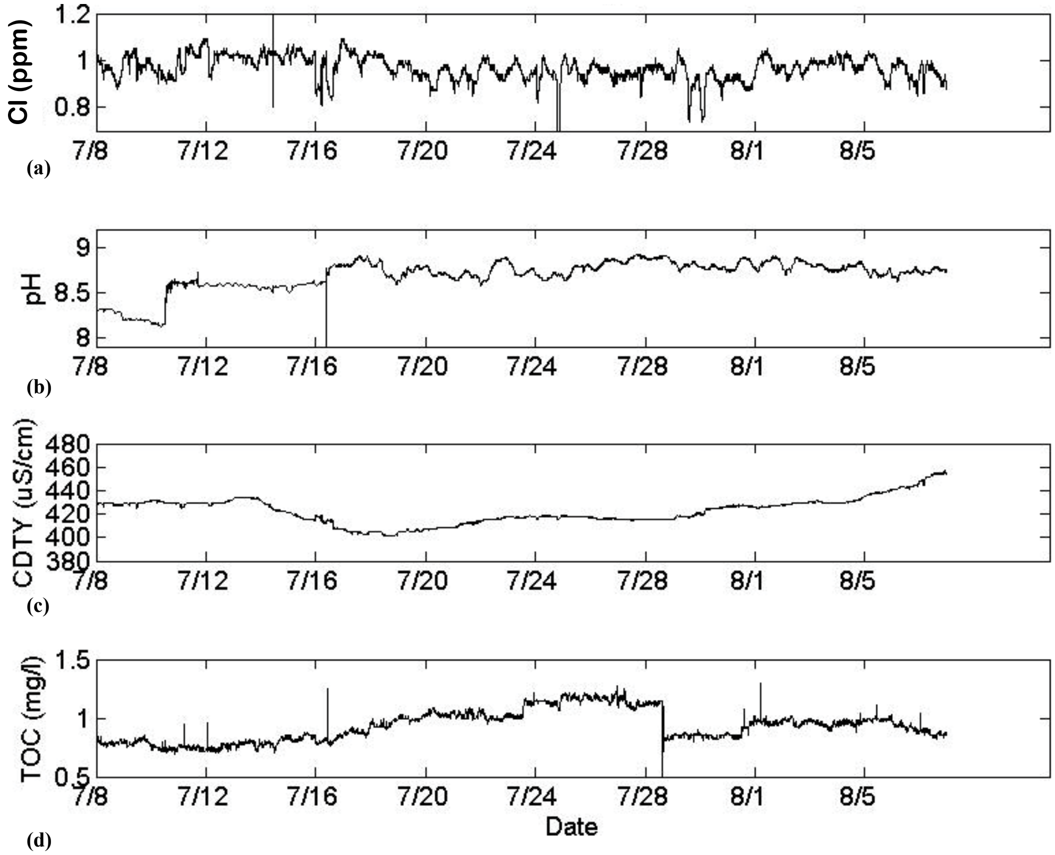


Figure 5-1. Training data for Location A. The four water quality signals used are (a) chlorine (Cl), (b) pH, (c) conductivity (CDTY), and (d) total organic carbon (TOC).

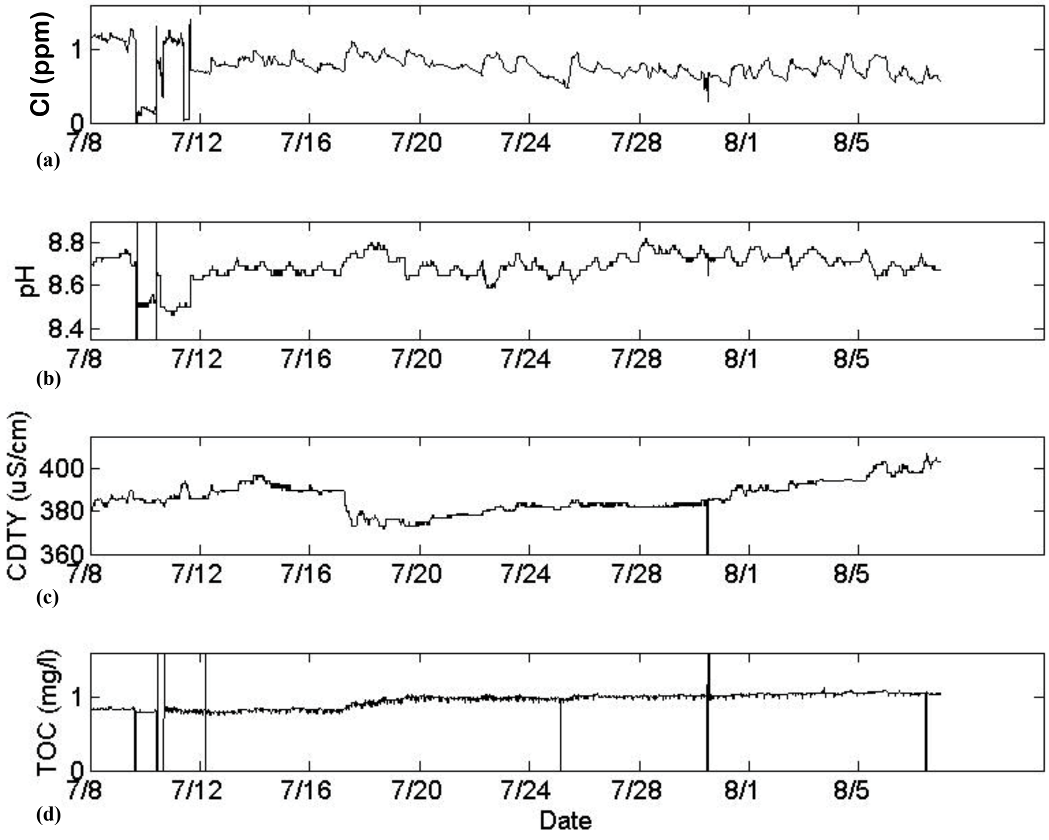


Figure 5-2. Training data for Location B. The four water quality signals used are (a) chlorine (Cl), (b) pH, (c) conductivity (CDTY), and (d) total organic carbon (TOC).

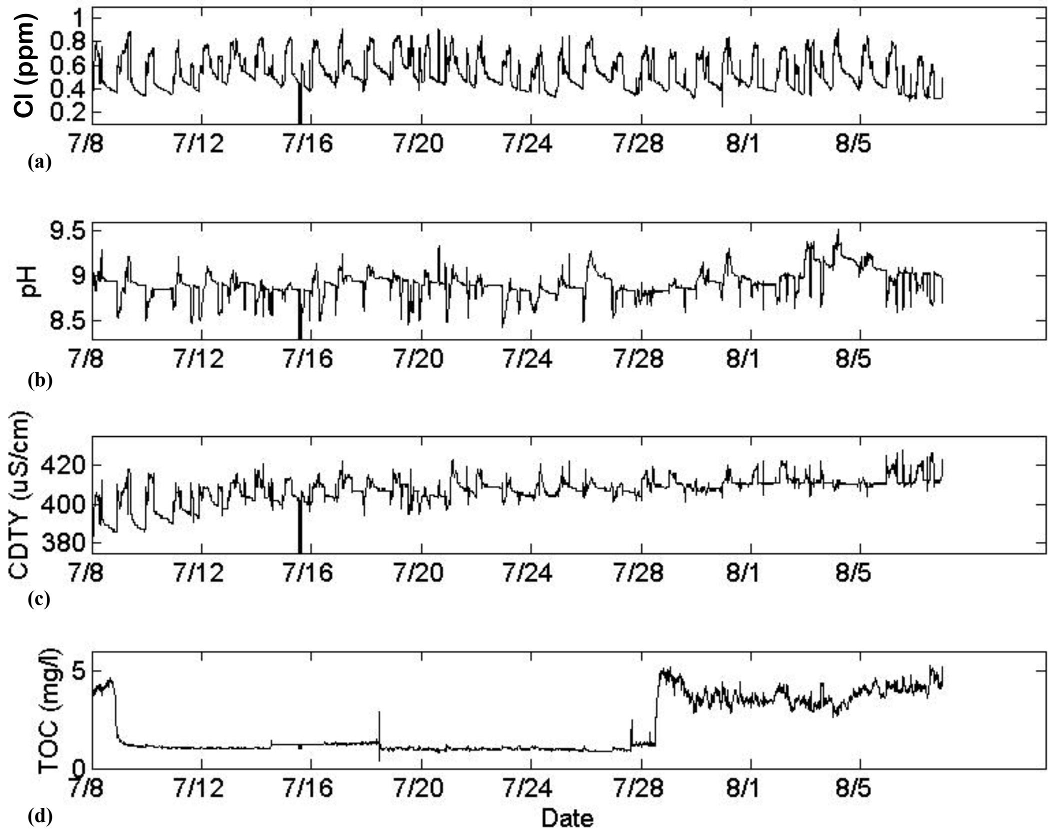


Figure 5-3. Training data for Location C. The four water quality signals used are (a) chlorine (Cl), (b) pH, (c) conductivity (CDTY), and (d) total organic carbon (TOC).

Analysis Step 1: Window Size and Prediction Errors

The first step in setting CANARY's algorithm parameters at a specific monitoring station is to determine the value of the window size. This moving window defines the number of previous time steps that are used to predict the water quality value at the next time step (LPCF algorithm) or to compare against the water quality value at the next time step (MVNN algorithm). The values in the window are normalized (mean of zero and standard deviation of one) prior to any analysis within CANARY. The best window size is determined by using the LPCF and MVNN algorithms on a training data set to predict each future water quality value. The quality of the predictions is defined by the average absolute value of the residual between the observed and predicted water quality values and the standard deviations of these residuals. These two performance measures are calculated as a function of the window size. For this case study, ten different window sizes ranging from 180 time steps (6 hours) to 1800 time steps (2.5 days) are examined. The results of these calculations are shown in **Figures 5-4** and **5-5.** The parameters controlling the integration of results across time steps using the BED algorithm are held constant across all testing runs. These parameters are set such that 14 outliers within 18 consecutive time steps (18 trials) are necessary before an event can be identified. These parameters were determined through testing on historical data.

For both algorithms, all stations and water quality signals show a general decrease in the performance measures with increasing window size. The exceptions to this observation are the standard deviation of the TOC and CDTY signals for Location B. These results are attributed to the variation in these signals at early times in the training data sets. All results show that the variation in the accuracy of the predictions across the different signals at one monitoring station is of the same order of magnitude as the variation in accuracy across all three monitoring stations. This result is remarkable given the strong increases in the variation of the water quality signals from Location A to B to C and demonstrates how the prediction algorithms in CANARY are able to adapt to different water quality characteristics at different monitoring stations.

Lower values of the average absolute residual and the standard deviation of the residuals indicate increased accuracy and precision, respectively, in the predictions of future water quality values. An obvious choice for the window size would be the one that produces the lowest values of the performance measures. In this case, the largest window size (1800 time steps) performs best across all stations, water quality signals, and algorithms (as shown in **Figure 5-4** and **Figure 5-5**). Statistical testing showed that the changes in the performance measures from one window size to the next were significant at all window sizes, indicating that even larger window sizes would continue to reduce these performance measures. The drawback of increased window sizes is the longer computational time needed to update the parameters and predict the future water quality at each time step. Experience with other monitoring stations and other water utilities have shown that window sizes between one and two days are enough to provide reasonably accurate and useful predictions of future water quality values.

The results for the LPCF algorithm shown in **Figure 5-4** indicate that a window size of at least 1200 time steps is needed to reduce the standard deviation of the residuals to near their final minimum value. Therefore, for the LPCF, a window size of 1440 (2 days) is selected. The results for the MVNN algorithm in **Figure 5-5** generally show the same shape, but have lower values than those for the corresponding LPCF calculations. Based on the similar shapes of the curves, a window size of 1440 time steps is also used for the MVNN algorithm.

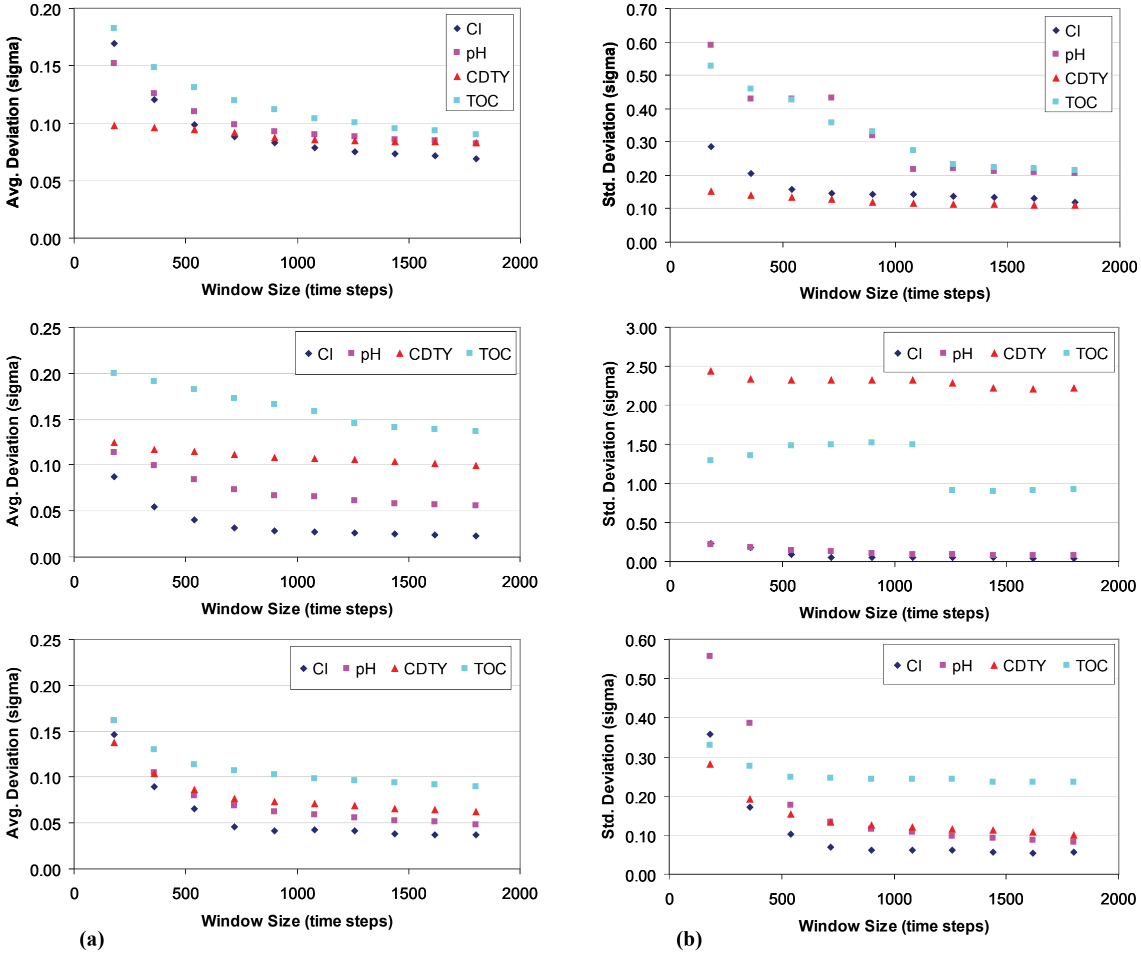


Figure 5-4. (a) Average deviation and (b) standard deviation of the prediction errors as a function of the window size for Locations A, B, and C from top to bottom, respectively. These results are from the LPCF algorithm.

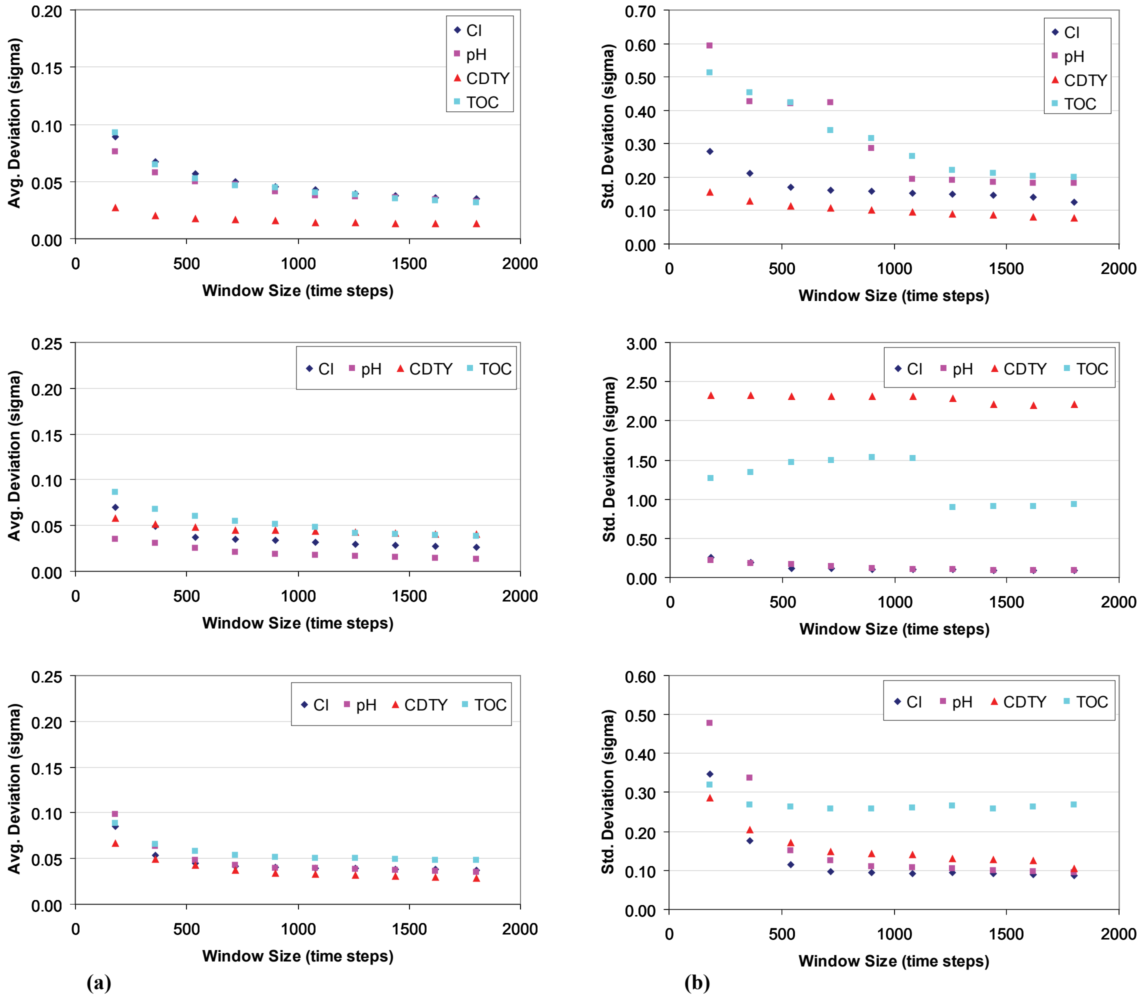


Figure 5-5. (a) Average deviation and (b) standard deviation of the prediction errors as a function of the window size for Locations A, B, and C from top to bottom, respectively. These results are from the MVNN algorithm. The scale of the Y-axes are held constant with those in Figure 5-4.

Analysis Step 2: Threshold Value and False Alarms

In addition to setting the window size, the event detection algorithms also require a threshold value to classify residuals as being indicative of either background or outlier water quality. The authors have determined that a useful rule of thumb for setting the minimum practical threshold is typically given by:

$$thresh_{\text{min}} = \overline{\varepsilon} + 2\sigma_{\varepsilon} \qquad (5\text{-}1)$$

where $\overline{\varepsilon}$ and σ_{ε} are the maximum values of the mean and standard deviation of the residuals across all signals analyzed.

The LPCF algorithm, with a window size of 1440 time steps, produces a mean deviation of approximately 0.10 and a maximum standard deviation of 0.20 to 0.25 across the four signals analyzed. The exception to this observation is Location B, which has higher standard deviation values as discussed previously. Based on the rule of thumb and the results across the multiple window sizes, the minimum threshold tested here is 0.60. A total of six threshold values are tested with increments of 0.10, such that the maximum threshold is 1.10. The MVNN algorithm results show generally lower mean deviations of approximately 0.05 and slightly lower standard deviations of less than 0.20 as compared to the LPCF results. For consistency, the series of threshold values evaluated are held constant for both the MVNN results and the LPCF results.

Both the LPCF and MVNN algorithms are used on the training data set for the range of thresholds from 0.60 to 1.10. The EDS results were examined qualitatively to determine the best threshold value. Threshold values that resulted in event declaration on obvious significant changes in water quality and minimized events and outliers throughout the rest of the data set were chosen. The thresholds selected for use on the testing data are shown in **Table 5-1.** Even though there are no known water quality events in the training data sets, alarms from CANARY are expected. These alarms are due to significant changes in the background water quality that occur at most monitoring stations. Examples are in the training data (**Figure 5-1** through **Figure 5-3**): the sharp drop in TOC on July 28th (7/28 in **Figure 5-1**) or July 29th at Location A, the drop in conductivity at Location B on July 17th, and the drop in TOC on July 9th followed by the increase in TOC on July 29th at Location C.

The results of running CANARY on the training data sets using the final selected parameters are shown for each station and each algorithm in **Table 5-2.** Four different measures are used to summarize these results: the total number of events identified by CANARY (i.e., the number of alarms produced); the proportion of all time steps that are identified as events; the average event length; and the average probability of an event *P(event)*, for those time steps classified as background (non-event) water quality. The BED parameters used here are the same as in Step 1 and limit the maximum length of an event to 45 time steps.

The results in **Table 5-2** show that for Locations A and B, 1.2 to 1.7% of the time steps are classified as events. At Location C, the estimated events make up 2.0 to 2.3% of all time steps. For a given monitoring station, the results from the two different algorithms are approximately the same. The average probability of an event outside of the areas classified as events ranges from 0.016 to 0.029 with the highest value occurring at Location C. These values are well below the probability threshold of 0.995 and indicate that outside of the events identified, the chances of a false alarm are very low.

Analysis Step 3: Simulation of Water Quality Events

A separate set of testing data is available for each monitoring station from August 8th through September 18th (29,606 time steps, or approximately 41 days). Simulated water quality events are added to these testing data sets. Here the contaminant simulator spreadsheet was not used. Instead, simulated events were designed to represent changes in water quality that would be observed from the introduction of a small amount of a contaminant into the distribution network. The simulated events change the background water quality by adding a deviation to that background:

$$Z_E\,(t) = Z_0\,(t) + E_{ind}\,(t) \cdot e \cdot E_{max} \cdot \sigma_z \qquad (5\text{-}2)$$

where $Z_E\ (t)$ is the event modified water quality value at time t, $Z_0\ (t)$ is the original background water quality at the same time step, E_{ind} is an event indicator equal to zero at all time steps outside of an event or between zero and one during an event, e defines a decrease (-1.0) or increase (1.0) in the water quality signal in response to the contamination event, and Emax is a coefficient applied to σ_z, the standard deviation of the water quality sensor data. An E_{ind} value of 1.0 indicates that the contaminant concentration is at full strength and the maximum change in the water quality sensors is occurring. Values of E_{ind} less than 1.0 indicate time steps within an event at which the contaminant concentration is less than full strength, such as at each end of the event where the effects of dispersion in the pipe have created transitional concentrations of the contaminant between zero and the maximum concentration. The maximum deviation of $Z_E\ (t)$) from $Z_0\ (t)$ is plus or minus the quantity (E_{max})(σ_z).

The initial shape of the simulated contaminant pulse is a square wave. Inclusion of the E_{ind} term in the event simulation allows for the shape of the leading and trailing edges of the contaminant pulse to be modified to represent varying amounts of smoothing that occur due to dispersion and diffusion of the pulse within the pipe network. As an example, **Figure 5-6** shows the values of E_{ind}, fraction of the event strength, as a function of the time step within the contaminant pulse. Both ends of the original square wave of the injected pulse (**Figure 5-6**) have been smoothed. The example in **Figure 5-6** has four time steps on each end of the pulse where the concentration is intermediate between the background (0.0) and the maximum strength of the event (1.0). The shape of the transition from background to maximum strength is modeled using a Gaussian cumulative distribution function and the total event length in is 34 time steps.

Table 5-1. Event detection parameters used in the analyses.

Monitoring Station	Window	threshold
Location A, LPCF	1440	0.90
Location A, MVNN	1440	1.10
Location B, LPCF	1440	1.00
Location B, MVNN	1440	1.10
Location C, LPCF	1440	1.10
Location C, MVNN	1440	1.10

Table 5-2. CANARY results on training data prior to addition of events.

Monitoring Station	Number of Events	Proportion of Time Steps within Events	Average Event Length (time steps)	Average P(event) Outside of Events
Location A, LPCF	7	0.014	39.7	0.016
Location A, MVNN	7	0.017	45.0	0.016
Location B, LPCF	9	0.014	31.9	0.021
Location B, MVNN	8	0.012	30.4	0.016
Location C, LPCF	14	0.020	29.4	0.021
Location C, MVNN	17	0.023	27.5	0.029

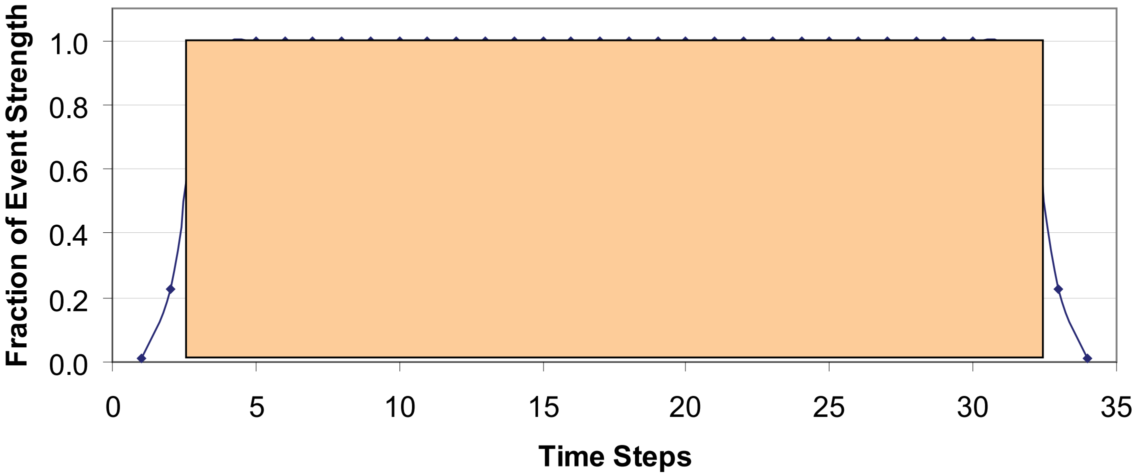


Figure 5-6. Use of the event indicator (E_{ind}) to define the shape of the event. The dotted line represents the shape of the contaminant pulse as witnessed at the monitoring station. The shaded box represents the initial square pulse of the simulated contaminant.

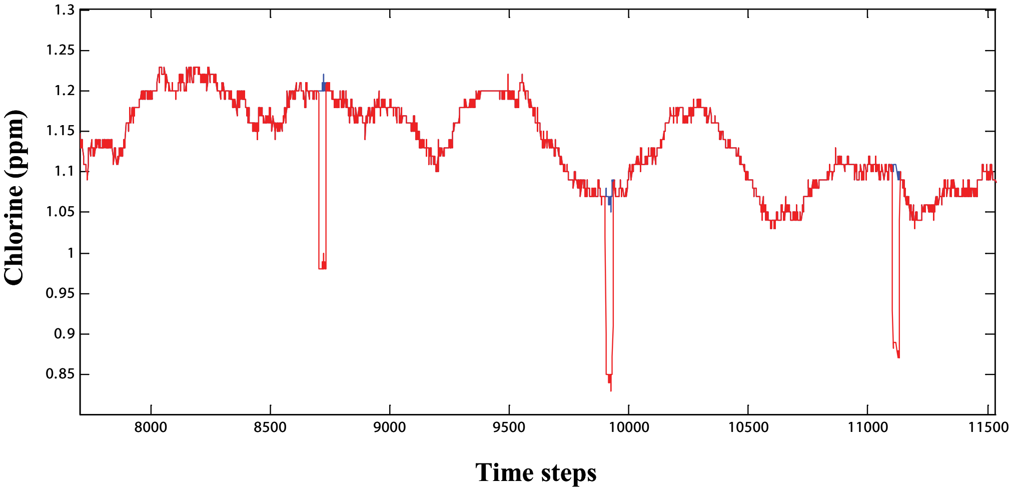


Figure 5-7. Example of the simulated response of the free chlorine sensor to the introduction of three contamination events. The blue lines indicate the original background sensor signal. The time between events is 40 hours.

Figure 5-7 provides an example of the simulated change in the response of the free chlorine sensor due to the injection of a contaminant. The maximum deviation of the sensor response from the background reading in this example is 1.5 times the standard deviation of the signal value (E_{max}= 1.5). This parameterization results in events which decrease the free chlorine concentrations by approximately 0.22 mg/L. The shape of the events is as defined in **Figure 5-6.** The spacing between events is 1200 time steps (40 hours).

Testing data sets were created by adding simulated events to experimental data. Hall et al. (2007; Table 3) showed that many contaminants decreased free chlorine and/or increased total organic carbon. For the majority of the contaminants tested, changes in pH and specific conductance were minimal.

For all testing data sets, the shape of the event is shown in **Figure 5-6** with the characteristics that were described above.

The effect of an event is to decrease the value measured by the Cl sensor and increase the value measured by the TOC sensor. The first event begins at time step 1501 and the subsequent events begin at intervals of 1200 time steps (40 hours) from time step 1501. Twenty-four events are added to each testing data set.

As mentioned above, the size of the maximum deviation away from the background water quality signal for Cl and TOC is defined as Emax times the standard deviation of the observed water quality. The standard deviations of the Cl and TOC data for the three training data sets are given in **Table 5-3.** The corresponding maximum deviation in the background signal for each monitoring station and each E_{max} are given in **Table 5-4** through **Table 5-6.**

Table 5-3. Standard deviation of the Cl and TOC signals for the three monitoring stations.

Monitoring Station	Cl (mg/L)	TOC (mg/L)
Location A	0.1469	0.1635
Location B	0.1818	0.0724
Location C	0.1775	1.8776

Table 5-4. Maximum signal deviation for each event at Location A.

Max Event Strength (E_{max})	Max Cl Deviation (mg/L)	Max TOC Deviation (mg/L)
0.50	0.073	0.082
0.75	0.110	0.123
1.00	0.147	0.164
1.25	0.184	0.204
1.50	0.220	0.245
1.75	0.257	0.286
2.00	0.294	0.327
2.25	0.331	0.368
2.50	0.367	0.409
2.75	0.404	0.450
3.00	0.441	0.491

Table 5-5. Maximum signal deviation for each event strength at Location B.

Max Event Strength (E_{max})	Max Cl Deviation (mg/L)	Max TOC Deviation (mg/L)
0.50	0.091	0.036
0.75	0.136	0.054
1.00	0.182	0.072
1.25	0.227	0.091
1.50	0.273	0.109
1.75	0.318	0.127
2.00	0.364	0.145
2.25	0.409	0.163
2.50	0.455	0.181
2.75	0.500	0.199
3.00	0.545	0.217

Table 5-6. Maximum signal deviation for each event strength at Location C.

Max Event Strength (E_{max})	Max Cl Deviation (mg/L)	Max TOC Deviation (mg/L)
0.50	0.089	0.939
0.75	0.133	1.408
1.00	0.178	1.878
1.25	0.222	2.347
1.50	0.266	2.816
1.75	0.311	3.286
2.00	0.355	3.755
2.25	0.399	4.225
2.50	0.444	4.694
2.75	0.488	5.163
3.00	0.533	5.633

Across all three monitoring stations, the decreases in free Cl range from less than 0.1 mg/L to near 0.5 mg/L. The simultaneous increases in the TOC are less uniform due to the larger variation in the TOC standard deviation values across the three stations. The TOC increases range from less than 0.1 mg/L at Locations A and B to greater than 5.0 mg/L at Location C. The much larger changes in TOC during the events at Location C are due to the high standard deviation of the TOC signal at that monitoring station.

In addition to the calculations done with the simulated event sizes shown in **Table 5-4** through **Table 5-6,** the three original data sets (unmodified) are also analyzed with CANARY. Analysis of the un-modified testing data sets corresponds to $E_{max} = 0.0$ and these results provide the baseline event detection results. **Figure 5-8** through **Figure 5-13** show the results of analyzing the unmodified testing data sets from all three sites using the LPCF and MVNN algorithms and as well as the time steps where CANARY identified water quality events for each signal (red dots in the figures).

A noticeable attribute of the three data sets is the loss of nearly 24 hours of data between September 14th and 15th (9/14 and 9/15 in **Figure 5-8** through **Figure 5-13**), which is easily noticeable in the figures below. During the periods of data loss, CANARY waits for the data to be available again and then continues to process the new data using the previous data left in the window prior to the data loss. Any significant change in the values of the signals from one side of the data loss to the other will cause an event. This scenario causes both algorithms to sound an alarm at all three stations at the end of the data loss.

In addition to periods with missing data, CANARY ignores those in which the sensor is offline. Such periods are identified by CANARY through sensor hardware alarms, which are not shown in the figures below. In particular, Location A and Location B have TOC sensor hardware alarms indicating that the TOC sensor is offline. At Location A, this occurs from September 13th to 14th (9/13 to 9/14 in **Figure 5-8** and **Figure 5-9**) (about 3.3% of the data). At Location B, the TOC sensor hardware alarm is on for part of September 10th (9/10 in **Figure 5-10** and **Figure 5-11**) and then from late on September 10th through the end of the data set (greater than 18% of the testing data). No specific TOC sensor hardware alarms occur in the data for Location C, although the pattern of missing data is a bit more complex and there is a period of sensor recalibration earlier in the data set, during the morning of August 13th (8/13 in **Figure 5-12** and **Figure 5-13**). Throughout periods of TOC sensor hardware alarms at Locations A and B, CANARY will continue processing and using the other three water quality signals to detect events. Because the simulated events only alter the Cl and TOC signals, during the periods of TOC sensor hardware alarms, CANARY will only be able to detect events on the basis of the changes in the Cl signal. The impacts of sensor hardware alarms on the CANARY results will be strongest at Location B.

In addition to the loss of data and the TOC sensor hardware alarms, there also appear to be some issues with the pH and CDTY signals at Locations A and B. These signals seem not to change at all beginning on September 9th or 10th (9/9 or 9/10 in **Figure 5-8** through **Figure 5-11**) up until the loss of data. These signals have no alarms during this period, but this behavior is unusual in water quality monitoring data.

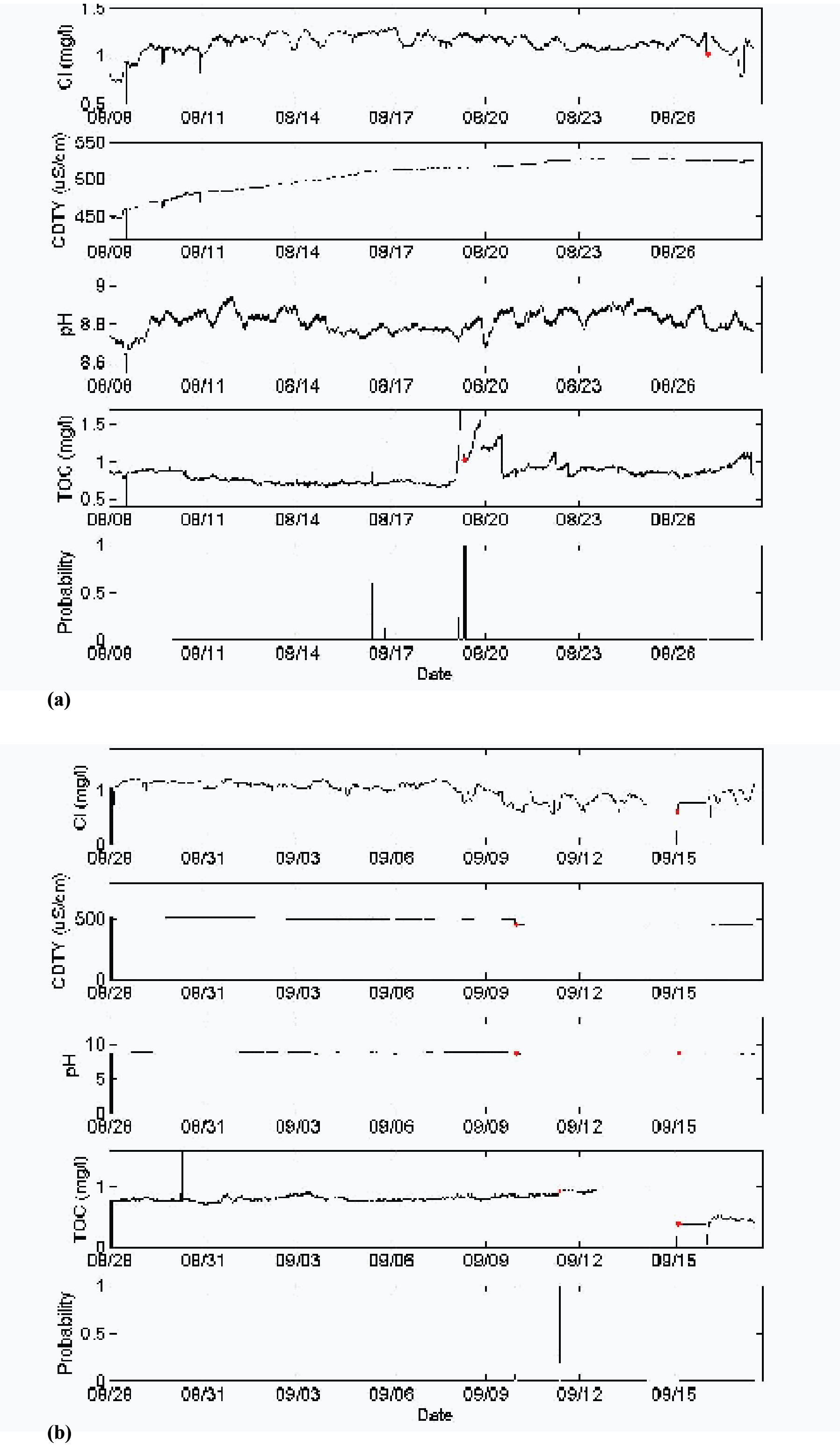
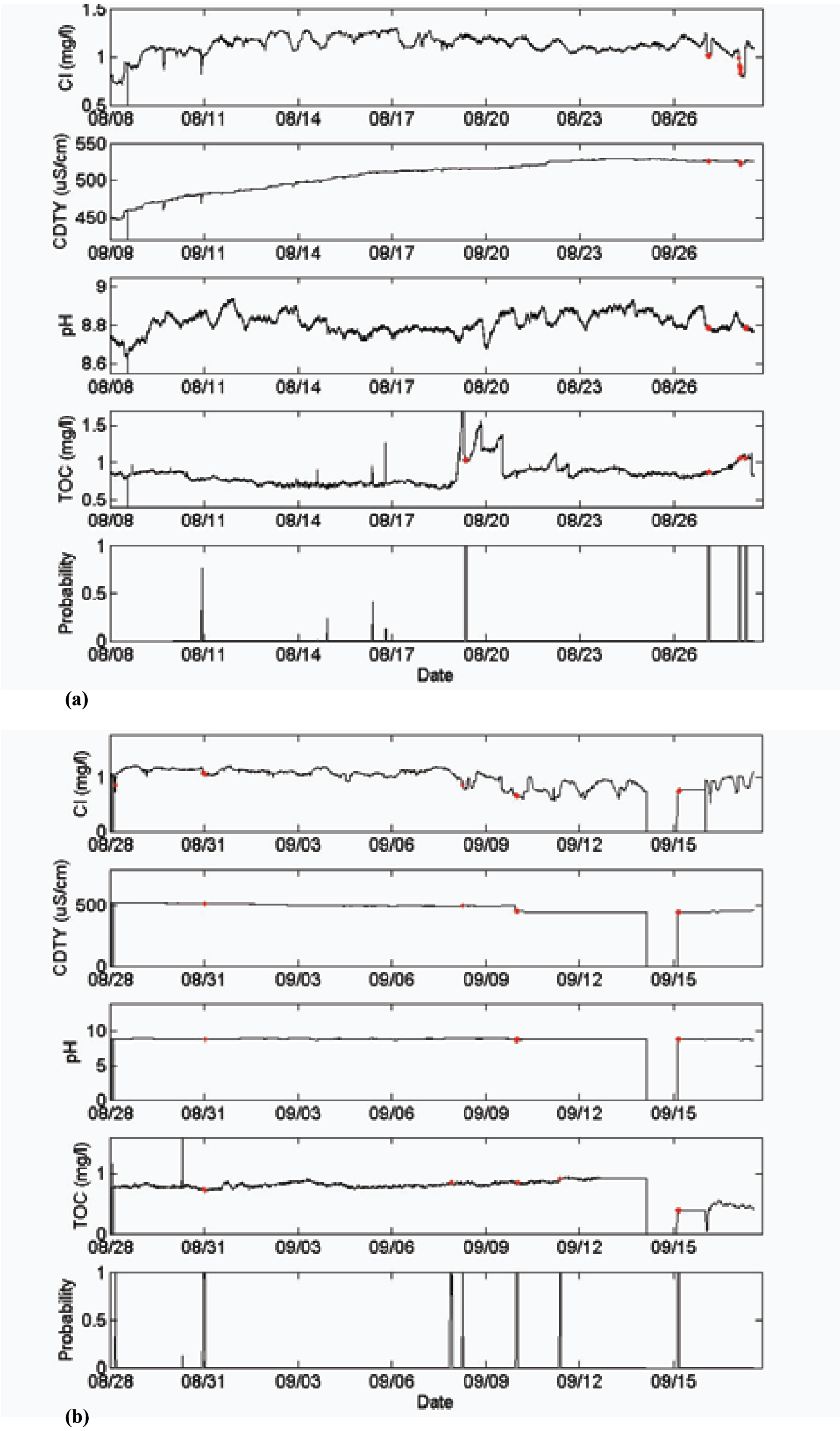


Figure 5-8. LPCF event detection results for the Location A testing data - no events added. The two images show the (a) first and (b) second halves of the data set.



(a)

(b)

Figure 5-9. MVNN event detection results for the Location A testing data - no events added. The two images show the (a) first and (b) second halves of the data set.

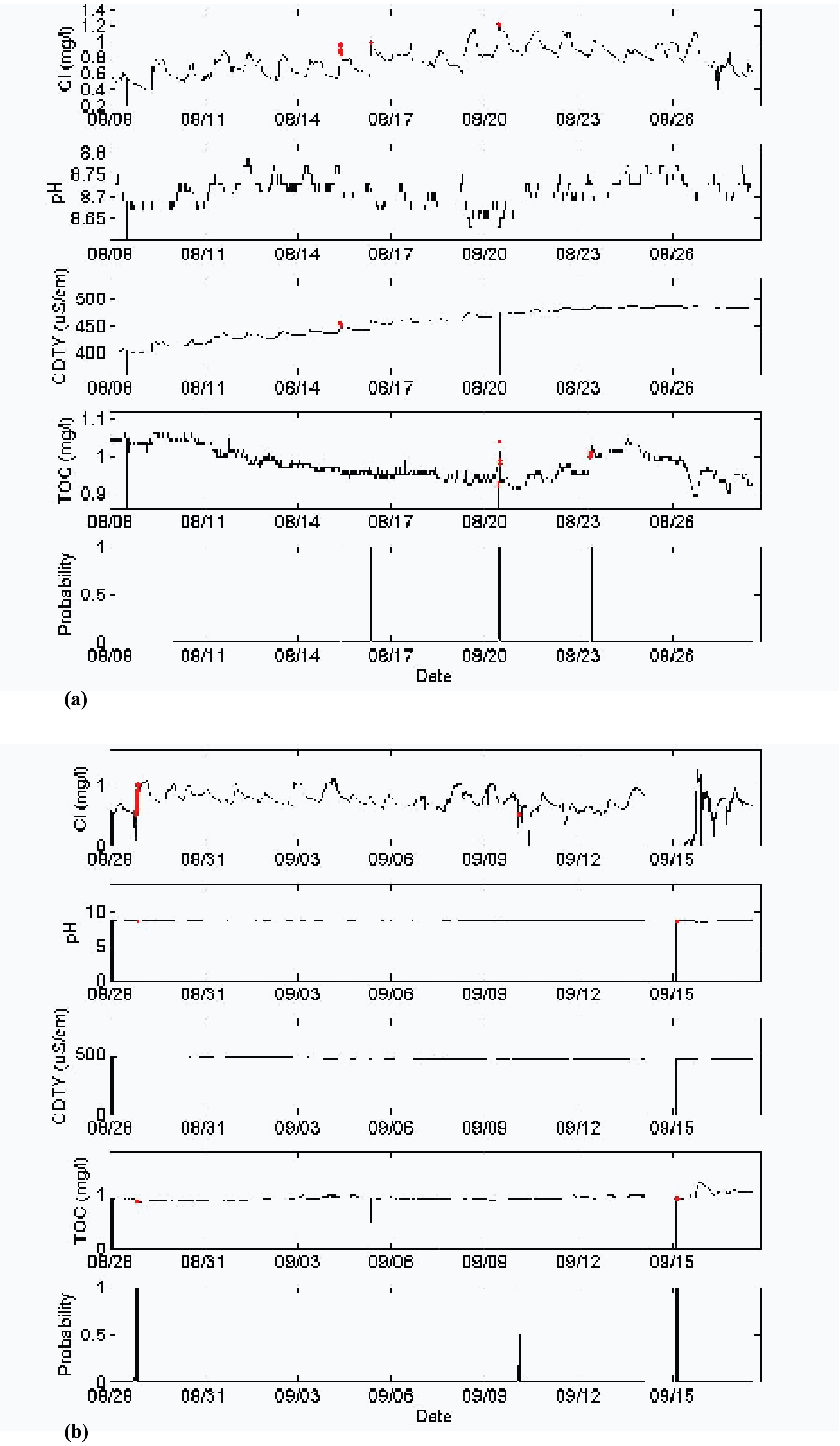
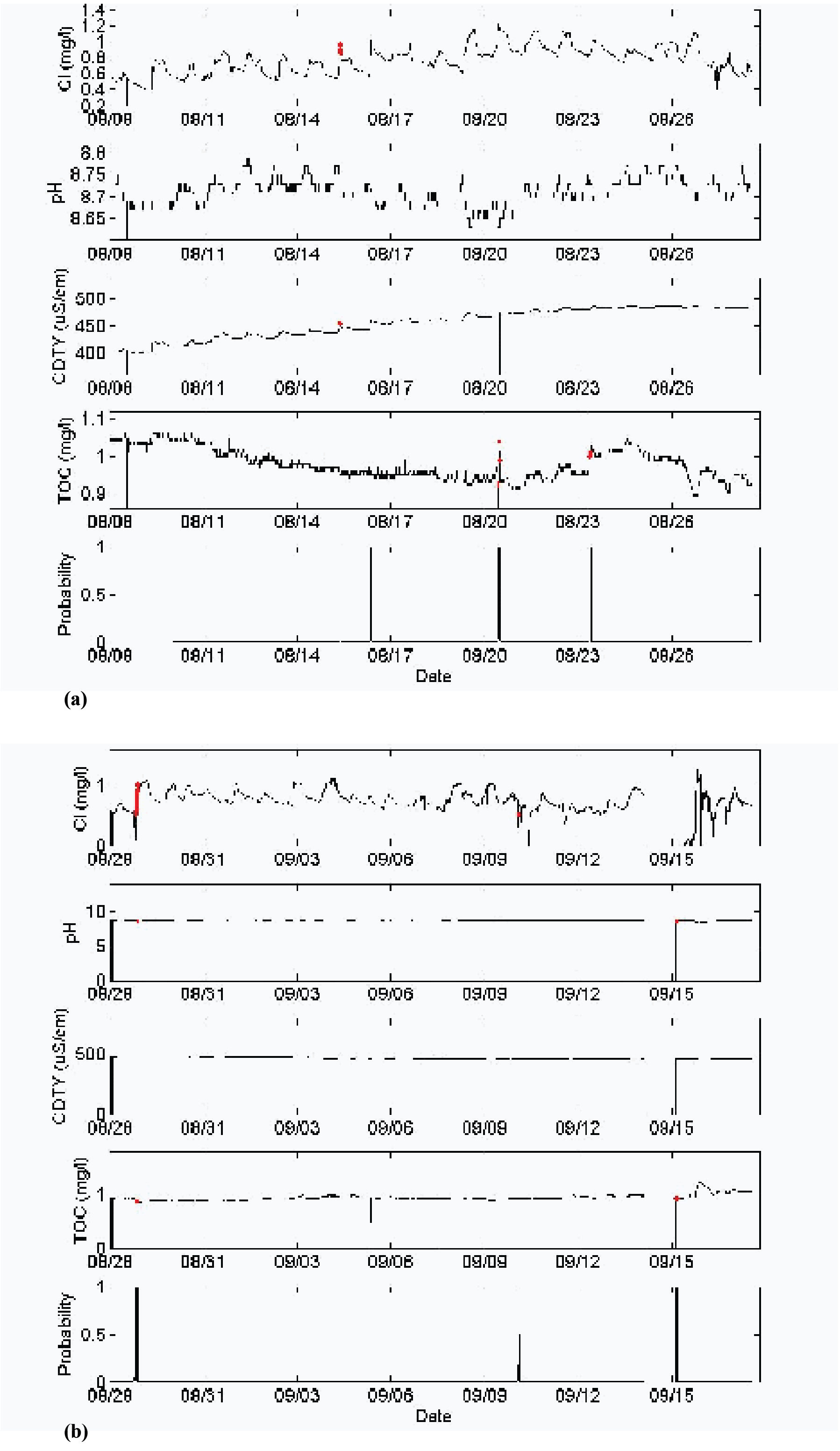


Figure 5-10. LPCF event detection results for the Location B testing data - no events added. The two images show the (a) first and (b) second halves of the data set.



(a)

(b)

Figure 5-11. MVNN event detection results for the Location B testing data - no events added. The two images show the (a) first and (b) second halves of the data set.

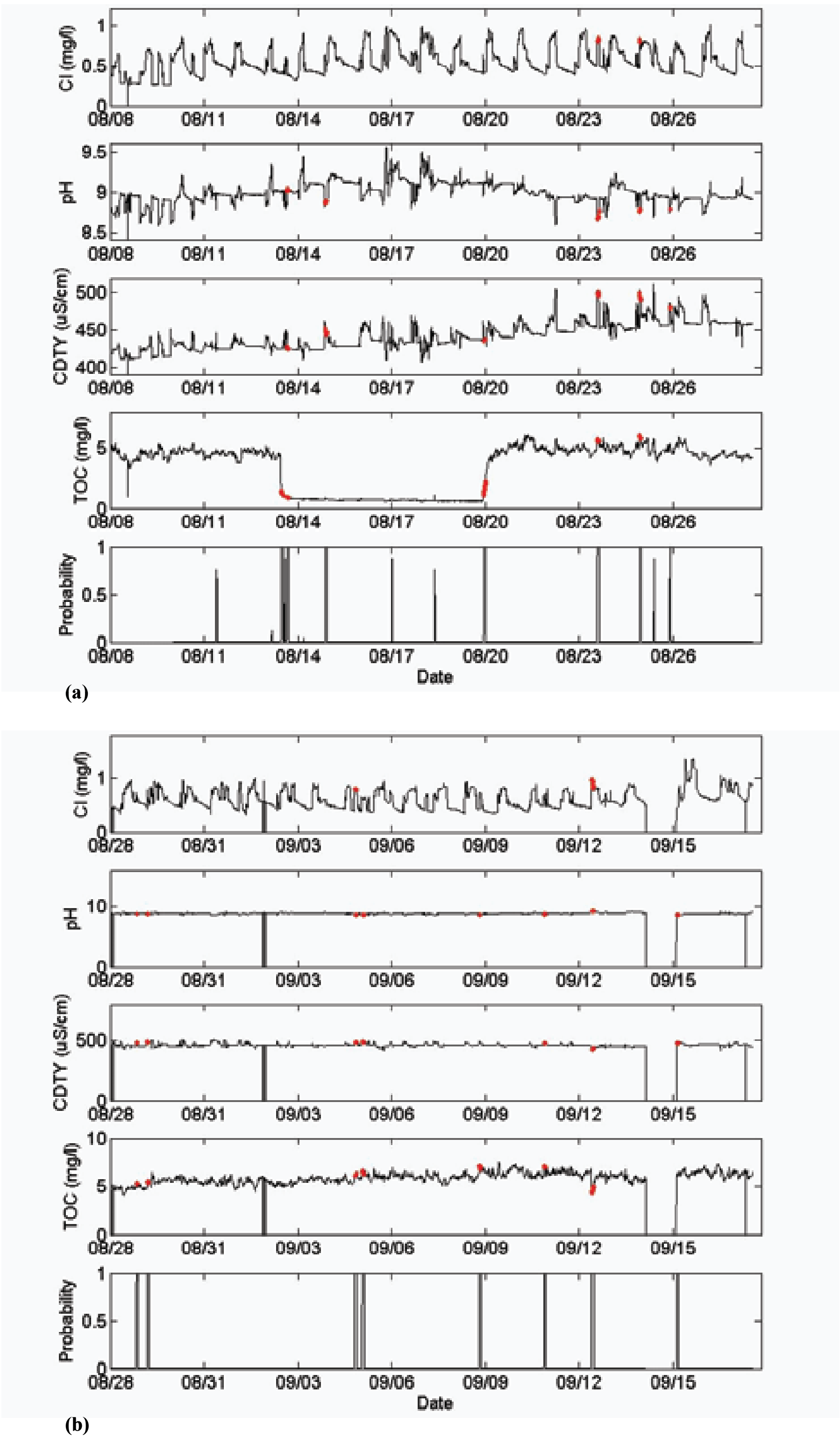
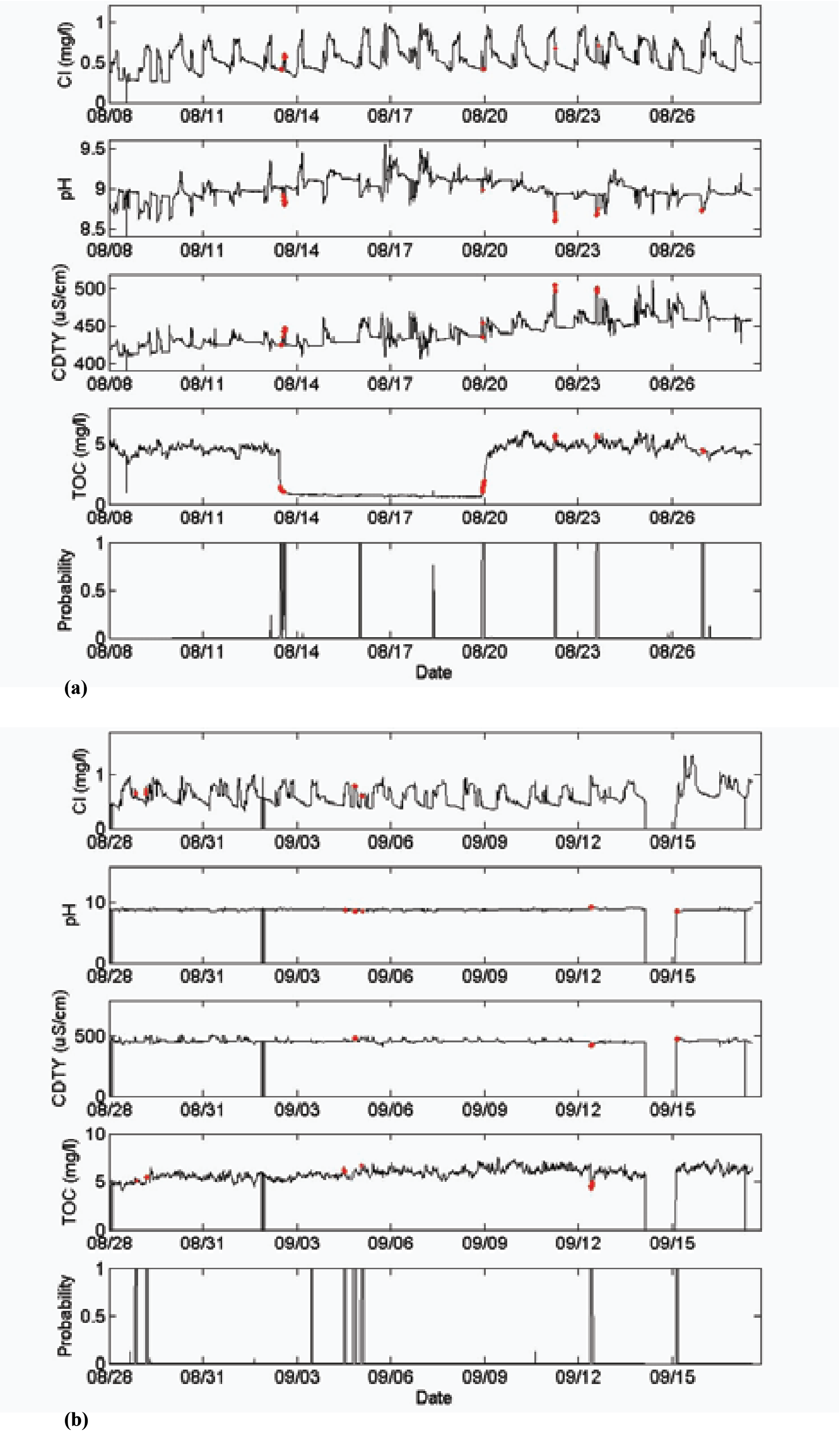


Figure 5-12. LPCF event detection results for the Location C testing data - no events added. The two images show the (a) first and (b) second halves of the data set.



(a)

(b)

Figure 5-13. MVNN event detection results for the Location C testing data - no events added. The two images show the (a) first and (b) second halves of the data set.

Table 5-7. CANARY results on testing data prior to addition of events.

Monitoring Station	Number of Events	Proportion of Time Steps within Events	Average Event Length (time steps)	Average *P(event)* Outside of Events
Location A, LPCF	5	0.033	37.8	0.009
Location A, MVNN	12	0.040	29.9	0.018
Location B, LPCF	7	0.035	32.9	0.011
Location B, MVNN	16	0.049	37.5	0.025
Location C, LPCF	15	0.048	37.3	0.025
Location C, MVNN	16	0.049	35.9	0.024

The results of running CANARY on the testing data sets with no events added are summarized in **Table 5-7.** The performance measures in **Table 5-7** are the same used on the training data set as shown in **Table 5-2.** Across all monitoring stations and algorithms, approximately 3 to 5% of the time steps are classified as events by CANARY. These results are indicative of the fairly sensitive parameter settings and at least double the proportion of the time steps classified as events in the training data. Such increase might indicate a change in the nature of the water quality signals between the training and testing data.

Analysis Step 4: Event Detection Results

Several different measures are employed to evaluate the performance of the event detection algorithms on the testing data sets. The known times of the simulated events are considered to be the "true" events, whereas the times identified by CANARY are called the "estimated" events. The performance measures are:

1) The area under the receiver operating characteristic (ROC) curve;
2) The proportion of true events for which there is at least one CANARY detection;
3) The proportion of the total time of the true events that overlap with the estimated events;
4) The average delay in the time of detection from the beginning of the true event;
5) The average length of the estimated events compared to the same measures for the true events.

The ROC curve has been widely used in evaluating decisions made in medical and engineering applications, including evaluating water quality event detection algorithms (McKenna et al. 2008). The ROC curve defines the tradeoff between missed detections (*MD*) and false positive decisions in a single curve. The two axes of the ROC curve are defined by the false alarm rate (*FAR*):

$$FAR = \frac{FP}{P + TN} \qquad (5\text{-}3)$$

and the probability of detection (*PD*):

$$PD = 1 - MD = \frac{TP}{TP + FN} \qquad (5\text{-}4)$$

where *TP* and *TN* are the true positives and true negatives, respectively, as defined by the extent of the simulated event, and *FP* and *FN* are the false positives and the false negatives (blue and red boxes in **Figure 3-5**). An *FP* occurs when CANARY estimates an event when no true event has occurred at the same time. An *FN* occurs when a true event remains undetected by CANARY. A *TP* occurs when CANARY estimates an event and a true event occurred at the same time. A *TN* occurs when CANARY does not estimate an event and there is no true event at that time.

The ROC curve provides a single plot that demonstrates the tradeoff between *FAR* and *PD* across all ranges of the probability of an event, *P(event)*. Typically, as the sensitivity of the algorithm is increased, the level of *PD* increases, but this also results in increased *FP's*. The area under the ROC curve varies from 0.5, indicating the decision results are only as good as those created by random guesses, to one, which indicates perfect decision making – the case of *PD* = 1 and *FAR* = 0. The ROC curve area is used here as a performance measure.

The results of the ROC curve calculations are summarized by the area under the ROC curve for each monitoring station and algorithm (**Figure 5-14a**). The areas under the ROC curves increase from approximately 0.5 to 0.7, at an event strength of 0.5, to above 0.8, at event strengths greater than 1.5. Beyond the event strength of 1.5, the ROC curve areas are nearly constant. Some level of variation exists in the ROC curve areas across the three stations and the two different algorithms. This variation is greatest at the smallest event strength (0.5), where the MVNN algorithm provides significantly better performance than the LPCF algorithm at Locations A and B. At event strengths greater than 1.5, the highest ROC curve areas (0.875) occur at Location C when using the LPCF algorithm and the lowest areas (0.8) occur at Location B when using the MVNN algorithm. The reason for Location C having the highest ROC curve areas is most likely due to the large absolute event strengths for TOC at that monitoring station (**Table 5-6**). These are due to the large standard deviation of the TOC data at Location C (**Table 5-3**).

The ROC curve analysis is accomplished by evaluating the decision result at each individual time step. This approach can be misleading as the parameters in CANARY are set to identify water quality events composed of groups of consecutive time steps where the water quality is anomalous. In particular, the settings of the BED algorithm used here require that at least 14 time steps be classified as outliers before an event can be identified. This intentional delay in the event identification works to reduce the number of false positive alarms, but also creates a large number of time steps where the true event is already occurring prior to CANARY identifying it. These time steps are considered as missed detections (false negatives). This delay in the event identification and the associated — relatively large — number of time steps considered to be false negatives leads to a characteristic shape in the ROC curve.

Figure 5-15 provides an example ROC curve as simulated using data for Location A. The most noticeable feature is the strong change in slope of the curve at the *PD* value (Y-axis) of approximately 0.69. This change in slope is caused by the delay in detection. No matter what threshold is applied to the probability of event values from CANARY, the first 14 time steps, at least, of every true event cannot be detected due to the delay built into the BED algorithm. Therefore, the ROC curve cannot rise any higher along the Y-axis. This delay mechanism limits the ability of CANARY to increase the probability of detection. The impact of changing the BED parameters on the ROC curve areas and the delay in the time to detection is evaluated further in the sensitivity analysis section of this chapter.

In addition to the ROC curve analysis, another evaluation approach is to consider each water quality event as an individual entity and determine the proportion of these events during which CANARY displays an alarm for at least one time step. This approach considers the resolution of the event to be the entire duration of the event and, therefore, is a less precise measure of the detection capabilities of CANARY. However, from a practical perspective, the bottom line for event detection is whether or not the events are detected at all, and this evaluation answers that question. For all signal strengths evaluated, the proportion of events that contain at least one time step of an alarm are evaluated and shown in **Figure 5-14b.**

When the event strength is 1.5 or larger, the proportion of detected events is greater than 0.85 for all monitoring stations and all algorithms. For the majority of the monitoring stations and event strengths, the LPCF algorithm performs better than the MVNN algorithm by detecting one or two more of the 24 true events, a 0.04 or 0.08 increase in the proportion detected, for each event strength. The best performance occurs at Location B where the LPCF algorithm is able to detect 23 of the 24 events (Proportion Detected = 0.96) for event strengths of 1.5 and greater.

Results of the other performance measures: proportion of overlap, average delay, and average event length, are all consistent with the results discussed above showing that CANARY is able to identify the majority of events when the strength is 1.5 or larger. To summarize these results, CANARY displays an alarm, on average, for about 40% of the time steps associated with each event. This corresponds to alarms for approximately 14 of the 34 time steps in each of the true events. For Location C, the average overlap proportion increases to approximately 50% (i.e., 17 out of 34 time steps) at the higher signal strengths. The average delay between the start of the true event and the first alarm from CANARY is 16 to 17 time steps depending on the monitoring station and the algorithm. This delay is consistent with the settings of the BED parameters that require a delay of 14 time steps before alarming. Additionally, several more time steps of delay are needed to account for CANARY not recognizing the first two or three time steps of each event that have transitional concentrations between the background and the full strength. The average event lengths identified by CANARY are 26 to 27 time steps compared to the 34 time steps of the true events. This result shows that not only is there a delay in the CANARY detections of 16 to 17 time steps, but that the CANARY detections continue beyond the end of the true events by approximately 10 time steps.

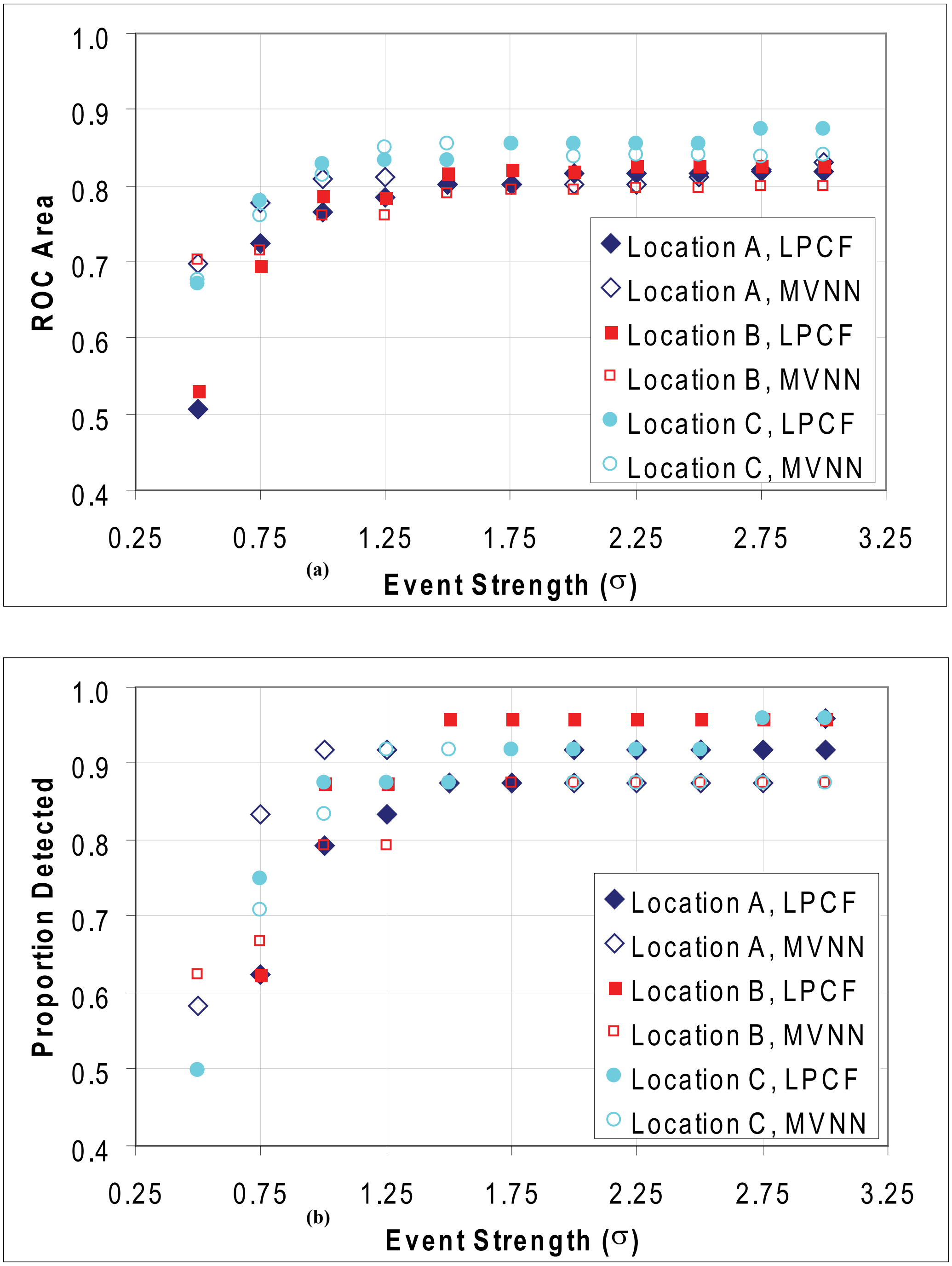


Figure 5-14. (a) Areas under the ROC curves and (b) proportions of true events with at least one detection as a function of the event strength in terms of standard deviation (σ).

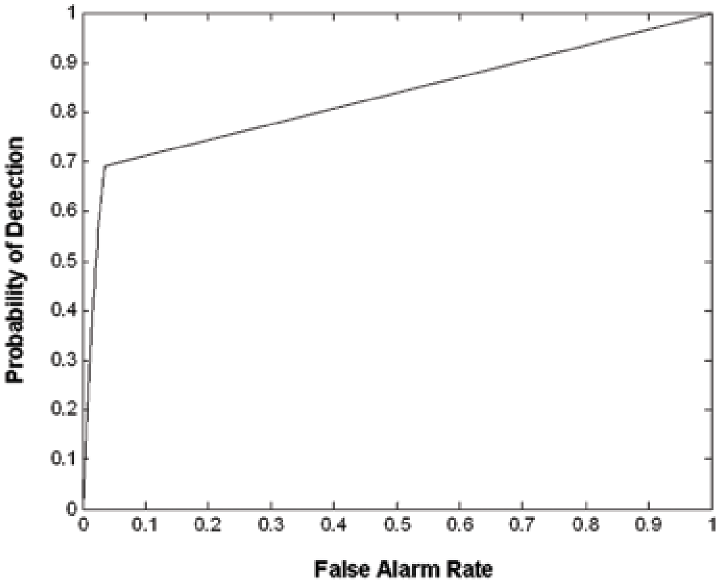


Figure 5-15. Example ROC curve showing the characteristic shape caused by a large delay in the detection of the true events. This calculation was done on the Location A testing data using the MVNN algorithm with simulated events of strength = 3.

Analysis Step 5: Sensitivity Analysis

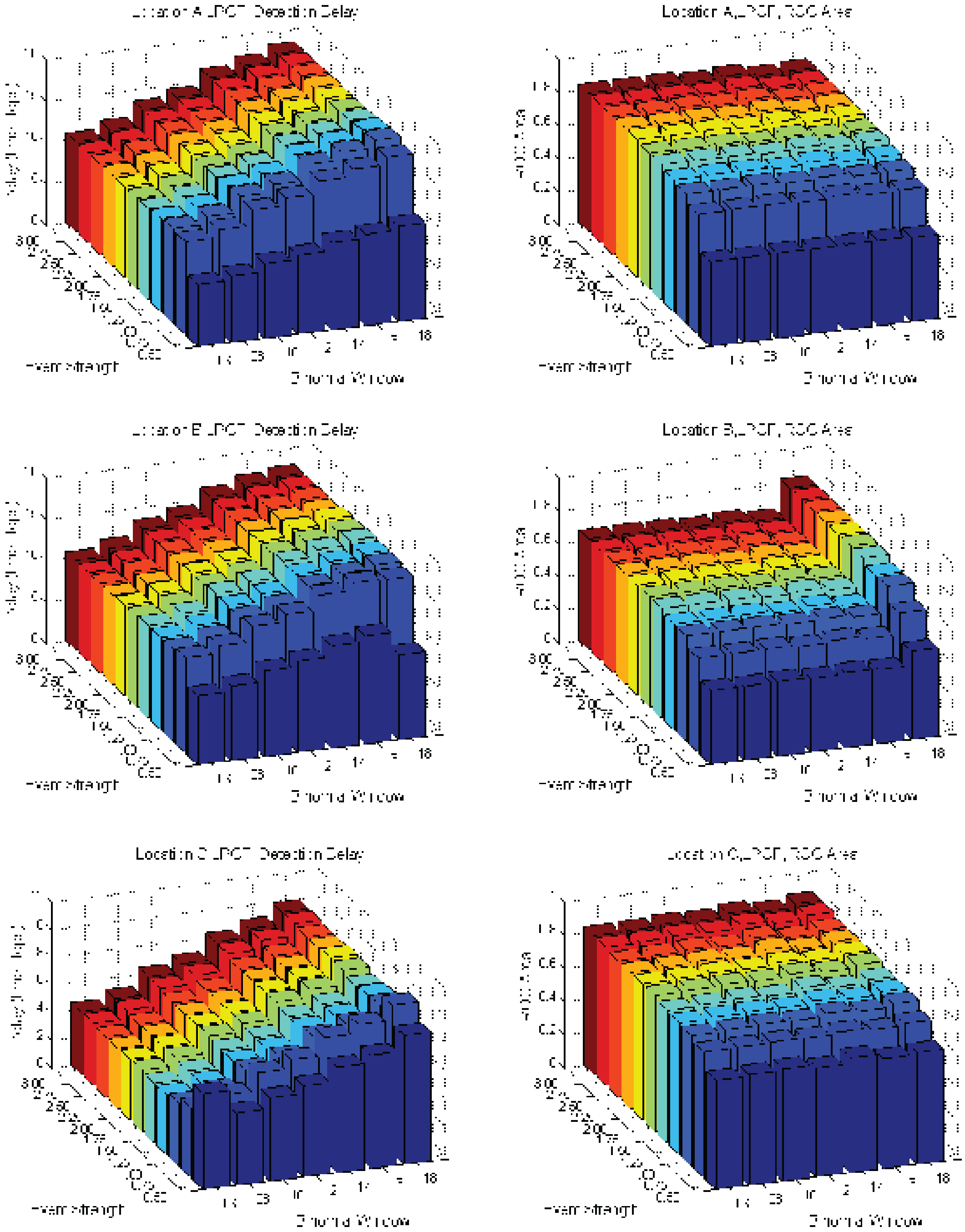
The BED was proposed as a means of gathering evidence of an anomalous period of water quality across several consecutive time steps (McKenna et al. 2007), but at this time the performance of the BED still has not been rigorously evaluated. The key parameters of the BED algorithm are the number of trials in each binomial probability calculation and the probability threshold compared to the probability of an event, *P(event)*. In CANARY, such parameters are defined as *bed-window-TS* (shown as the "binomial window" in the figures below) and *event-threshold-P,* respectively. The *bed-window-TS* parameter is evaluated to determine its impact on the previously described EDS performance measures and also on the delay in detection of an event. Of particular focus is the question of whether or not changes in the parameterization of the BED can reduce the average delay time between the onset of an event and the detection of that event, while simultaneously increasing, or at least maintaining, the area under the ROC curve.

Calculation of *P(event)* through the binomial model and comparison to the probability threshold of 0.995 specifies that 14 outliers within 18 time steps are necessary in order to declare a water quality event. Here the same results are used to examine how the performance measures are affected by a decrease in the number of outliers needed for an event declaration. Changes to the number of trials in the binomial experiment are made to decrease the number of outliers necessary for identification of a water quality event. Thus, the value of *bed-window-TS* is decreased from 18 to 6 in steps of two.

Figures 5-16 and **5-17** show the results of changing the value of *bed-window-TS*. **Figures 5-16a** and **5-17a** show the detection delay as a function of the value of *bed-window-TS* and the event strength. Figures **5-16b** and **5-17b** show the area under the ROC curve as a function of the same two parameters. **Figure 5-16** shows results for all three monitoring stations obtained using the LPCF algorithm, and **Figure 5-17** shows the same results for the MVNN algorithm. Several observations are clear from these figures:

- Decreasing the number of trials used in the BED decreases the detection delay in a near linear manner for all event strengths above 0.5 standard deviations. This behavior is expected, given that a larger number of trials increase the delay prior to being able to detect an event.
- The ROC curve area is not strongly dependent on the value of bed-window-TS. For most cases, decreasing the detection delay does not significantly change the area under the ROC curve. A strong exception to this observation occurs at Location B using the LPCF algorithm, since a bed-window-TS value of 18 results in a jump in the ROC curve area relative to smaller values of bed-window-TS. This jump is due to CANARY identifying 23 of the 24 true events when bed-window-TS is set to 18 and only identifying 20 of the 24 true events when the bed-window-TS drops to 16 or less.
- The values of the ROC curve areas remain relatively stable as the detection delays decrease. Therefore, while such a decrease produces a decrease in the number of false negatives, it also produces an increase in the number of false positives. This relationship is further explored below.
- The detection delay results at Location C are significantly shorter than those at the other two monitoring stations. The large standard deviation of TOC data at Location C makes the absolute values of the simulated events quite large relative to background values, and these events are, therefore,

detected faster than events of similar strengths at the other two monitoring stations. The resulting detection delays of approximately four time steps indicate that CANARY is capable of identifying events at Location C during the transitional period from background to full contaminant strength.



(a)

(b)

Figure 5-16. (a) Detection delay and (b) ROC curve area results for the LPCF algorithm at all three monitoring stations.

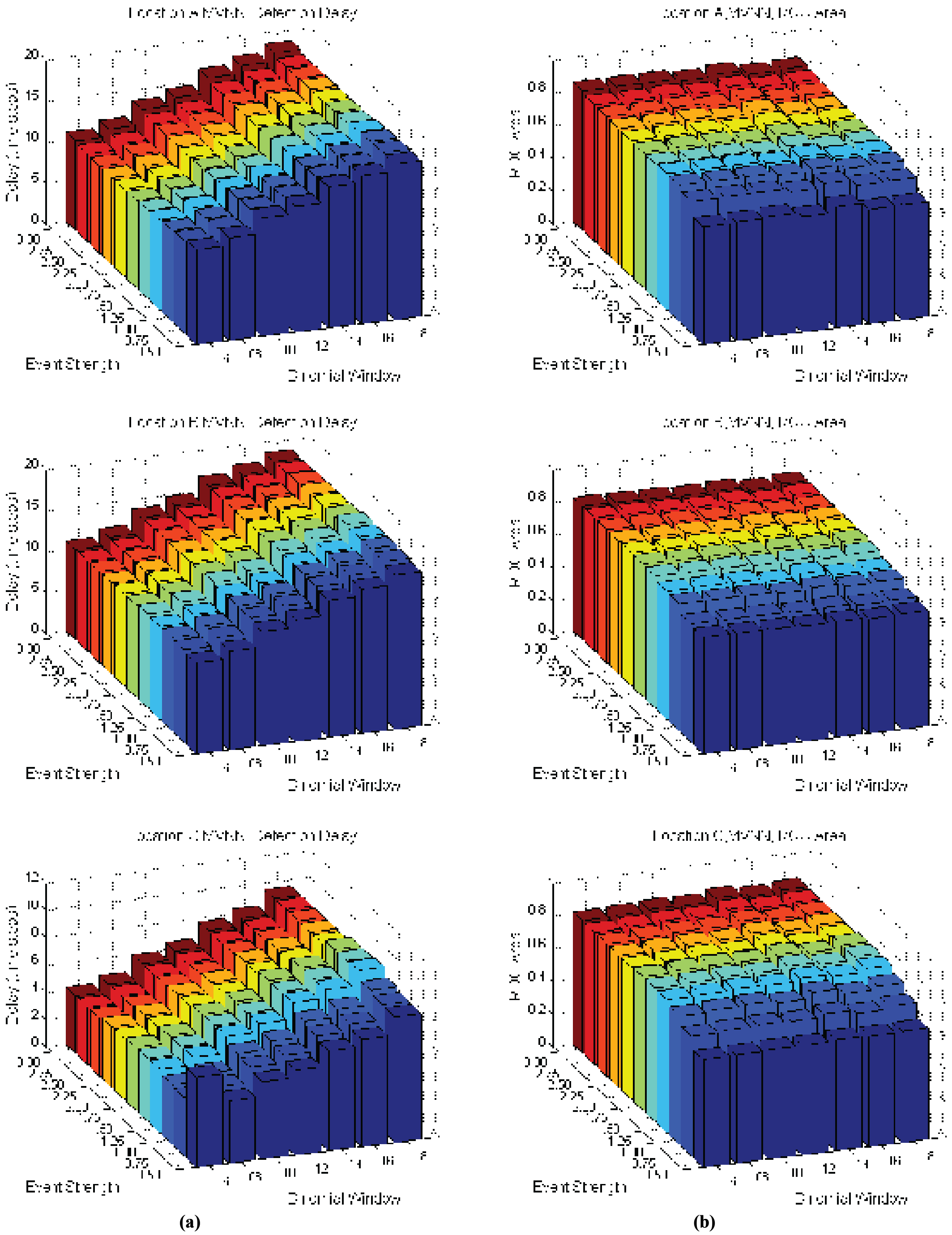


Figure 5-17. (a) Detection delay and (b) ROC curve area results for the MVNN algorithm at all three monitoring stations.

Discussion

The event detection results presented show that the differences between the LPCF and MVNN algorithms are minimal. In theory, the MVNN algorithm should require a larger threshold to get the same results as the LPCF algorithm, based on the mechanism for calculating the threshold.

An example with two signals provides a simple basis for comparison: Cl and TOC. The normalized distance (residual) between the predicted and observed water quality for each signal is one. The LPCF algorithm will retain the maximum residual for comparison to the threshold. The MVNN algorithm will calculate the Euclidean distance between the current observation and the closest previous observation. If the predicted water quality value turns out to be one standard deviation for both the TOC and Cl values, the Euclidean distance will be the square-root of two or 1.41. For a threshold value between 1 and 1.4, only the LPCF algorithm will identify this time step as an outlier. The residual calculation differences also lead to a broader distribution

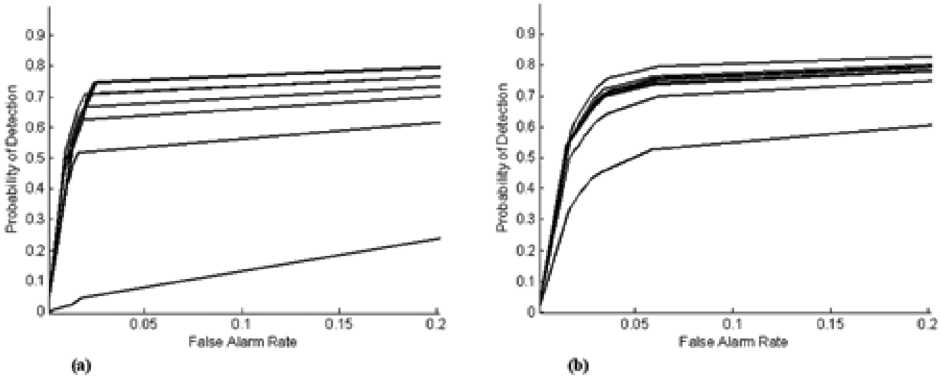


Figure 5-18. Comparison of example ROC curves for 11 event strengths at Location A with a *bed-window-TS* value of 8; (a) from the LPCF algorithm and (b) from the MVNN algorithm.

of residual values from the MVNN algorithm, which is a combination of signals, than from the LPCF algorithm, which only selects a single maximum value at each step. These differences influence the shape of the ROC curves.

Examination of the actual ROC curves shows that the maximum *PD* is reached with a very low false alarm rate, less than 5%. In general, the LPCF algorithm results in lower false alarm rates and a sharper break in slope than the MVNN algorithm (**Figure 5-18).** Note the expanded scale on the x-axis in **Figure 5-18.** The relative sharpness of the break in the slope is due to the differences in the LPCF and MVNN algorithms discussed above. The wider distribution of residual values created by the MVNN algorithm relative to the LPCF algorithm leads to the smoothed change in slope displayed in **Figure 5-18b.**

A series of simple calculations using the properties of the simulated true events can provide additional understanding of the values in the ROC curves. For every 1200 time steps, 34 are the true event and 1166 are background. Outside of the results at Location C, the minimum delay times calculated were near 11 time steps (**Figures 5-16** and **5-17**). If the estimated events have a delay of 11 time steps and there are no extra time steps estimated as events at the end of the event- no false positives - then: *TP* = (34-11) = 23, *FN* = 11, and *PD* = 23/(23+11) = 0.68. This value is near that of the break in slope for many of the ROC curves shown in **Figure 5-18**. A hypothetical decrease in the delay to 8 time steps, increases the *PD* value to 26/(26+8) = 0.76. A drawback of considering every time step as an independent result when using the ROC curve as an evaluation tool is that the calculation is dependent on the length of the event. If the detection delay remains constant at 8 time steps and the length of the simulated events is simply made twice as long, 68 time steps, the resulting *PD* value would be 60/(60+8) = 0.88, a 16% improvement over the case of the shorter events.

Examination of the results calculated here show that false positives occur most commonly by overestimation of the length of the event. If a delay of 11 time steps is needed to identify each event and, at the end of each event, CANARY continues to estimate an event for ten time steps beyond the end of the true event, the *FAR* calculation is: *FP* = 10, *TN* = (1166-10) = 1156, and *FAR* = 10/(10 + 1156) = 0.0086. Again, this value is close to the break in slope in **Figure 5-18**.

This sensitivity analysis demonstrates that changes in the EDS parameterization can be completed to decrease detection times at the expense of a higher number of false positives. Operational reasons might exist to bias this tradeoff towards faster detection times and higher false positives, for example during a period of heightened security around a community event. The increased false alarm rate would most likely prohibit operating in this mode for extended periods of time.

Summary and Conclusions

Three monitoring stations were selected to provide data for testing of CANARY EDS. The water quality signals in these three data sets demonstrated varying levels of periodicity due to network operations. The three data sets were split into training and testing sets. The training data sets were used to determine the appropriate window and threshold parameters for CANARY using two different algorithms: LPCF and MVNN. Within CANARY, a window length of two days, 1440 time steps, and threshold values of one standard deviation were used.

An event simulation approach was developed to create changes in water quality assuming the introduction of small amounts of potential contaminants. The basic shape of the contaminant pulse is a square wave and can be controlled to represent more or less smoothing of the leading and trailing edges of the pulse.

Several issues complicated testing of the event detection capabilities provided by CANARY. Some change in the nature of the water quality data appeared to exist between the training and testing data sets. This change roughly doubled the number of false positive events at all three monitoring stations between the training and testing data sets using the same parameters when no events were added to either data set. This change also challenged the underlying assumption that the training and testing data have the same statistical characteristics. The TOC testing data for Location C had a unique bimodal distribution where approximately one week of TOC data had values below 1.0 mg/L, athough the remainder of the data set was near 5.0 mg/L. This bimodal distribution caused the TOC data set to have a large standard deviation and, therefore, large absolute values for the TOC events relative to the other signals and other monitoring stations.

Across all monitoring stations examined here, CANARY was able to detect more than 90% of the simulated events for event strengths greater than 1.5 standard deviations of the background water quality (a change of approximately 0.25 mg/L in the Cl and TOC signals). Event detections remained at 80% or greater for event strengths between 1.0 and 1.5 standard deviations (a change of 0.15 to 0.20 mg/L in Cl and TOC). These results are remarkable when considering that the daily changes in the background Cl were as much as 0.5 mg/L at Locations B and C and daily changes in TOC were of a similar magnitude at Location C.

The delay between the onset of an actual event and the declaration of that event by CANARY is controlled by the BED algorithm. The BED algorithm can be considered a post-processor of the outliers determined by the LPCF or MVNN algorithms. The initial probability threshold and *bed-window-TS* values used in testing resulted in a minimum of 14 outliers before an event could be declared. Given the smoothed leading edge of the contamination events, this requirement generally meant that at least 18 outliers, or 36 minutes, were needed prior to declaring an event. Changes in the BED parameter, *bed-window-TS*, reduced the average delay to as little as six time steps (12 minutes) while keeping the area under the ROC curve the same. This result means that the reduction of false positives created by the decreased delay to detection is offset by an increase in false positives at other points in the data set. These results provide some guidelines on how to handle CANARY’s settings based on the tradeoff between sensitivity in detection and generation of false positives.

6.
Water Quality Pattern Matching

Deployment of EDS tools has shown that false positive alarms are often caused by routine changes in water quality due to the hydraulic operations of a utility. As an example, during a six-month period when two EDS tools were deployed at the Greater Cincinnati Water Works (GCWW), both tools frequently produced false alarms due to opening or closing of valves, draining of tanks, and changes in the status of pumps within the distribution system (Allgeier et al. 2008).

False positives caused by these types of regular operational events can be minimized in one of two ways: 1) locate monitoring stations in areas with stable water quality characteristics (far from the influence of tanks and pumps); or 2) incorporate algorithmic approaches to reliably recognize changes in water quality due to hydraulic operations. The rest of this chapter focuses on the second option, since the purpose of this report is to discuss event detection and not sensor placement. For more information on sensor placement, refer to Murray et al. 2010.

One algorithmic approach for reducing false alarms caused by operational actions is to include more than water quality information as input to the EDS. The main difficulty with this approach is that changes in hydraulic operations often occur far away from monitoring stations. Therefore, it is difficult to predict which monitoring stations will be impacted, the lag time between the operational action and the resulting water quality change, and the exact strength of the change. Extensive knowledge and experience of the utility operations can be used to try to predict the relationship between operational and water quality changes, but this can be difficult, if not impossible, to quantify due to the almost limitless possible combinations of system configurations and operations. In addition, this approach is utility specific and not easily generalized across utilities. Because of these significant difficulties, a new approach is proposed that adapts a method called trajectory clustering in order to identify regular patterns in water quality changes.

Trajectory Clustering

Previous applications of multivariate clustering algorithms (see Chapter 4) cluster the actual water quality data values to fit within a finite number of water quality classes (e.g., Klise et al. 2006a; Klise et al. 2006b). However, for the purpose of event detection, one is most interested in knowing when the data changes from one classification (e.g., background) to another (e.g., event). Therefore, the focus is on classification of water quality changes into background or "other" categories.

One promising technique for identifying patterns within time series data is trajectory clustering (Gaffney 2004). In this technique, time series of data (i.e., trajectories) are clustered rather than considered to be discrete data points. Typical applications of multivariate clustering focus on classification of discrete points or vectors of data measured on different features and are not inherently designed to integrate multiple measurements made in series along a curve or trajectory (Xu et al. 2009). As an example, focused applications of trajectory clustering have been developed to help classify historical storm tracks (Camargo et al. 2007; Gaffney et al. 2007). Storm track data are considered the prototypical data sets for trajectory clustering because they provide a series of latitude and longitude pairs that define the center of the storm at discrete time intervals.

Previous developments in trajectory clustering are expanded to increase the dimensionality of the trajectory clustering framework to examine the water quality within n-dimensional parameter space. Additionally, the clustering is placed into an online framework that constantly updates the current water quality pattern with new data and compares it to a previously defined pattern library. The proposed approach for reducing EDS false alarms consists of two key steps: 1) creation of a cluster library of events from historical data and 2) comparison of current water quality signals against entries within the library. The following sections detail those processes and provide several example calculations.

Creation and Clustering of Water Quality Template Libraries

The analysis of historical water quality data to create water quality template libraries is a multi-step process. The key steps in this process are:

1. Identification of water quality events in historical data;
2. Creation of a template library;
3. Water quality change clustering;
4. Calculation of cluster statistics.

The clustering approach developed here provides a concise summary of common water quality patterns against which any new observed water quality pattern can be quickly compared.

Identification of Water Quality Events

The first step in creating water quality event template libraries is to identify the routine events associated with operational actions. To do so, the user first runs CANARY with established configuration parameters on a set of historical data to generate an event probability, $P_C(t)$, for each time step. The data set should be long enough to capture the pattern of interest many times: a conservative recommendation would be 30 times. CANARY compares the event probability, $P_C(t)$ (the same as *P(event)* from previous chapters), with the user-defined threshold probability, P_{thresh}. For the purposes of creating the template library, an event is defined as a continuous interval of time steps during which

Table 6-1. An illustrative example to describe how events are identified: two events, colored red, begin at time steps 3 and 7.

P_{thresh}	0.5											
$P_C(t)$	0	0.2	**0.6**	**0.8**	0.1	0.2	**0.6**	**0.8**	**1**	**1**	**1**	0
$P_C(t) > P_{thresh}$?	N	N	**Y**	**Y**	N	N	**Y**	**Y**	**Y**	**Y**	**Y**	N
Initiation of event?	N	N	**Y**	**N**	N	N	**Y**	**N**	**N**	**N**	**N**	N
Time Step	1	2	**3**	**4**	5	6	**7**	**8**	**9**	**10**	**11**	12

the event probability exceeds the threshold probability ($P_C(t) > P_{thresh}$) and is identified with the first time step in this interval. **Table 6-1** contains hypothetical data to illustrate this process.

Creation of the Template Libraries

For each event identified by CANARY in the dataset, CANARY removes any missing or bad data. The pattern matching capability of CANARY then fits a series of low order regression models on the remaining data[1] using the MATLAB® (MathWorks 2008) function `polyfit`. For each water quality signal, time is considered the independent variable in the regression model. For a particular signal, a regression model is determined for the data points that immediately precede the initiation of an event.

The orders of the regression models and the numbers of data points to which the models are fit must be specified. The orders and number of data points are constant across events. The regression coefficients for an event are stored in a matrix that is termed the template library. That is, the template library is an N_E x O_{Total} matrix, where N_E is the total number of events identified in the historical data and O_{Total} is the sum, over all of the water quality signals, of the orders of polynomial regression plus the number of signals considered (since a n^{th} order polynomial has n+1 coefficients). **Figure 6-1** contains a flow chart of how the template library is created.

For example, the authors typically use free chlorine, pH, and conductivity signals as input to CANARY. The authors' empirical trials have determined that third- to fifth-order regression models typically work well when fitting 2290 time steps prior to the detection of an event. If each signal is fit with a third-order polynomial, then the first four entries of a row in the template library row contain regression coefficients for free chlorine data, the fifth through eighth entries are regression coefficients for pH, and the last four row entries are regression coefficients for conductivity data. Thus, the template library would have twelve columns. Thus, the template library would have 12 columns for each event. These events are then clustered into water quality patterns.

[1] For some signals, "2"s are removed from the regression data since Supervisory Control and Data Acquisition (SCADA) systems sometimes report powers of 2 for signals (e.g., pH) when there are SCADA errors.

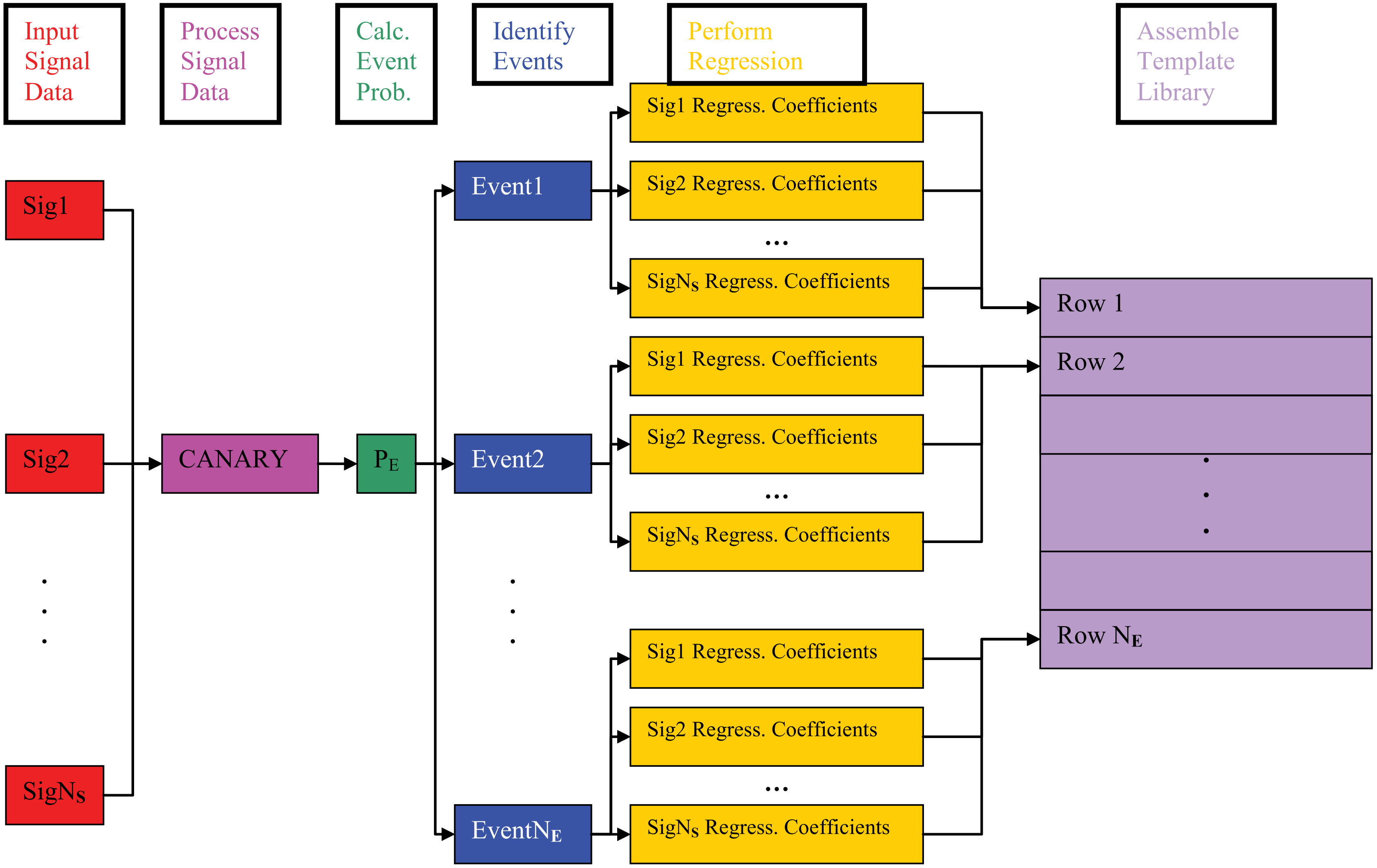


Figure 6-1. Flow diagram for creating the water quality template library.

Water Quality Change Clustering

Following the creation of the template library, the water quality change events within the library are clustered. A trajectory clustering algorithm has been implemented in CANARY, because it is the pattern of water quality change, not the actual water quality values during the change, which must be identified. The algorithm simultaneously clusters the regression coefficients for all signals rather, than the actual data values corresponding to the events.

CANARY uses the fuzzy c-means (FCM) algorithm to cluster the regression coefficients. The FCM algorithm is an iterative clustering algorithm developed by Dunn (1973) and further refined by Bezdek (1981). It is a "soft" clustering algorithm that permits events (indicated by a small number of regression coefficients) to belong to multiple clusters and, thereby differs from a "hard" clustering technique, like the k-means algorithm (Hartigan et al. 1978), which assign events to a single cluster. For each event, the FCM algorithm calculates the degree to which each event belongs to each cluster.

The basis of the FCM algorithm is the minimization of the following objective function:

$$J = \sum_{i=1}^{N_E}\sum_{j=1}^{N_C} u_{ij}^{m}\left\|x_i - c_j\right\|^2, \qquad 1 \le m < \infty, \tag{6-1}$$

where

- N_E is the number of events being clustered;
- N_C is the number of clusters;
- $\boldsymbol{x}_i$ is the event that is being clustered;
- $\boldsymbol{c}_j$ is the cluster center for the jth cluster;
- u_{ij} is the degree of membership of x_i to cluster j. Note $0 \le u_{ij} \le 1$, and $\sum_{j=1}^{N_C} u_{ij} = 1$;
- $\|\ \|$ is a norm for measuring the distance of events from cluster centers; and
- m is a "fuzziness" parameter that can be adjusted to affect cluster membership. This parameter must be assigned a value greater than one, and larger values lead to more overlap of the clusters.

The FCM algorithm is an iterative algorithm, and it is composed of the following steps:

1. Initialize the cluster membership matrix U^0, i.e., the matrix that contains u_{ij}.
2. At the kth step, calculate the cluster centers c^k_j using the cluster membership matrix U^K in the following equation

$$c_j^k = \frac{\sum_{i=1}^{N_E}(u_{ij}^k)^m x_i}{\sum_{i=1}^{N_E}(u_{ij}^k)^m}. \tag{6-2}$$

3. Update the cluster membership matrix U^K with the following equation

$$u_{ij}^{k+1} = \frac{1}{\sum_{p=1}^{N_C}\left[\frac{\left\|x_i - c_j^k\right\|}{\left\|xi - c_p^k\right\|}\right]^{(2/m-1)}}. \tag{6-3}$$

4. Repeat steps 2 through 5 until $\left\|U^K - U^{K+1}\right\|_U < \varepsilon$ or $k > N_{term}$. The term ε is a positive constant used to establish convergence criteria for the FCM algorithm, and N_{term} is a positive integer that establishes additional termination criteria. The notation $\|\ \|_U$ is used to represent a matrix norm.

Since Dunn (1973) assigned m a value of 2 in the first presentation of the FCM, this convention is often still followed. The CANARY implementation of the FCM algorithm also follows this convention. Values for the other FCM parameter values are fixed within CANARY, and sensitivity analyses were conducted to determine the other parameter values. **Table 6-2** lists parameter values that are assigned in CANARY's implementation of the FCM algorithm.

Table 6-2. Fuzzy c-means clustering algorithm parameters in CANARY.

Parameter	m	ε	N_{term}	$\|\ \|_U$
Value	2	0.1	100	$\|\ \|_U$ for matrices

Several considerations had to be made when implementing the FCM algorithm in CANARY. The distance norm that was implemented is defined as follows:

$$\|v\| = \sqrt{\sum_{i=1}^{l}\left(\frac{v_i}{\sigma_i}\right)^2} \tag{6-4}$$

where

- l is the length of the vector;
- v_i denotes the ith element of the vector v; and
- σ_i denotes the standard deviation of all of the events' ith regression coefficients that are being clustered.

Often, specific conductivity values are one to two orders of magnitude larger than the other water quality signals, and if the standard Euclidian distance is used to define the norm in the FCM algorithm, the clustering algorithm would more heavily weight the patterns in conductivity signals than patterns in the other signals. (The clustering methodology was tested on data in which free chlorine values typically ranged between 1-3 mg/L, pH values from 7 to 9, and conductivity values from 90 to 120 and 170 to 200 µS/cm.) To avoid this problem, **Equation 6-4** is used to equally weight the regression coefficients for all the signals that are clustered.

The FCM algorithm also requires an "initial guess" for the degree of cluster memberships (U^0 in Step 1 of the algorithm). It is common practice to assign random values to this matrix, but the efficiency of the algorithm can be sensitive to the initial guess. Thus, a different approach for assigning initial cluster membership values was implemented. The template library was initially clustered using MATLAB®'s (MathWorks 2008) hierarchical clustering function **`clusterdata`**. Hierarchical clustering is a "hard" clustering technique in which events are assigned to a single cluster. If an event was assigned to a particular cluster using the hierarchical clustering approach, the initial cluster membership degree for that event to the cluster was assigned a value of δ, and the degrees of membership for that event to all the other clusters were assigned a value equal to

$$(1-\delta) \big/ (N_C - 1)\,. \tag{6-5}$$

That is,

$$u_{ij} = \begin{cases} \delta\,, x_i \in cluster \quad j \\ \left(\dfrac{1-\delta}{N_C - 1}\right), x_i \notin cluster \quad j\,. \end{cases} \tag{6-6}$$

The parameter δ is assigned a value of 0.8 in CANARY's FCM algorithm. This value was determined through trial and error.

Finally, the FCM algorithm requires that the analyst determine the number of clusters *a priori*. This can be difficult if the data are difficult to visualize or a large number of events are being clustered. At best, relying on the analyst's judgment is a subjective process. Thus, CANARY uses the clustering index developed by Pakhira, Bandyopadhyay, and Maulik (PBM) (Pakhira et al. 2004) to determine the optimal number of clusters. This index, termed the PBM-index, is defined as follows:

$$PBM(N_C) = \left(\frac{1}{N_C} \times \frac{E_1}{E_{N_c}} \times D_{N_c}\right) \tag{6-7}$$

where E represents the cluster membership weighted norm of the distance between the data within an event and the center of an existing cluster. The E_1/E_{NC} term in the PBM-index is the sum of all intra-cluster distances for the full data set if there was only a single "super cluster" over the sum of all distances for the multi-cluster system (Pakhira, et al., 2004). All that is needed for the PBM calculation is the value of E for the case of a single cluster, E_1, and the sum of E over all N_C clusters, E_{NC}.

$$E_1 = \sum_{j=1}^{N_E} u_{1j} \left\| x_j - c_1 \right\| \tag{6-8}$$

$$E_{N_c} = \sum_{i=1}^{N_C} \sum_{j=1}^{N_E} u_{ij} \left\| x_j - c_i \right\| \tag{6-9}$$

$$D_{N_c} = \max \left\| c_i - c_j \right\| \tag{6-10}$$

Note that c_1 is used to define the single "super cluster" for calculation of E_1 and otherwise x_i, c_j, u_{ij}, and $\|\;\|$ are defined in the same manner as they were in the FCM algorithm.

Pakhira et al. (2004) assert that the positive integer that maximizes the PBM-index is optimal in the sense that it minimizes the number of clusters while increasing compactness and separation between clusters. Hence, CANARY assigns the parameter representing the number of clusters in the FCM algorithm to the integer value between 2 and 10 (inclusive) that maximizes the PBM-index. The upper bound on the number of clusters is arbitrarily set to 10 since most examples that have been analyzed optimize the PBM-index with three to six clusters.

Calculating Cluster Statistics

In order to perform real-time comparison of water quality events with an existing template library, it is necessary to calculate cluster statistics. The events in the clusters are assumed to be normally distributed and the following equations are used to calculate the entries in the cluster means and covariance matrices, μ_j and COV_j, respectively:

$$\mu_j = \frac{\sum_{i=1}^{N_E} (u_{ij}) x_i}{\sum_{i=1}^{N_E} u_{ij}} \tag{6-12}$$

$$COV_j = \frac{\sum_{i=1}^{N_E} (u_{ij})(x_i - m_j)(x_i - m_j)^T}{\sum_{i=1}^{N_E} u_{ij}} \tag{6-13}$$

The subscript j denotes the cluster number.

Comparison of Incoming Data With the Template Library

Creation and clustering of the template library is performed in an offline mode using historical data. Once the library is established, when CANARY is run in online mode, it monitors incoming data and compares it to patterns already contained in the template library. If a change in water quality is significantly different from the template library, then CANARY will alarm. This section describes the process that CANARY uses to compare the real-time signals with the template library.

In its online mode, data read by CANARY from SCADA will be processed to compute event probabilities. If such probabilities exceed the user-defined probability threshold, the software will perform polynomial regression fits to the same signals and number of time steps considered in the template library. This regression step must use all of the same parameters that were used to create the template library. The following calculations compare current regression coefficients for each cluster to the mean and covariance of events in the pattern library:

$$p_j = 1 - \left[\chi^2_{DOF}\right]^{-1} (x_{RT} - \mu_j)^T COV_j^{-1} (x_{RT} - \mu_j) \quad \text{(6-14)}$$

where x_{RT} denotes the regression coefficients for the new event and $[X^2_{DOF}]^{-1}$ denotes the inverse cumulative distribution function (CDF) for the chi-squared distribution with degrees of freedom equal to the total number of regression coefficients. Under the assumption that the clusters follow multivariate normal distributions, the term p_j denotes the percentile of each cluster's distribution to which x_{RT} corresponds. If the new event does not fall within a certain percentile of any cluster, then CANARY will detect an event. If any p_j is less than the tolerance level, no new clusters are added to the template library. Rather, the regression coefficients corresponding to the new event are added to the library, and the FCM algorithm is re-run with on the entire supplemented library. Means and covariance matrices are then recalculated for each cluster.

Example Calculations

Several example calculations are provided to demonstrate application of the pattern matching capability within CANARY. The first example (Location A) employs a simulated pattern of daily changes in the water quality signals. The simulated values provide a repeatable pattern with precisely known characteristics and beginning and ending times. Simulated water quality events are also added to the observed data to evaluate event detection with recurring water quality patterns. The second example (Location B) considers data collected from the output of a water works within an operating network and allows for examination of the effects of plant production on influencing water quality patterns.

In both examples, the pattern library is constructed by analyzing approximately four months of data. The goal of the pattern library construction is to identify a large number of events that are potential members of a pattern. To achieve this goal, parameters for the event detection algorithm are set to be more sensitive than for a typical analysis. The length of the history window is set to be one-half of the typical window length and the residual threshold is set to be slightly lower than typically used in online event detection. For each event, a decision is then made to include it within the pattern clustering or not.

Example 1: Simulated Patterns

The Location A data set used for this example has three water quality parameters: free chlorine (Cl), pH, and specific conductivity (CDTY). The background water quality is relatively stable and strong recurring patterns were not observed. Simulated changes in the observed water quality values are added such that the Cl values decrease by approximately 0.25 mg/L and the pH and CDTY values increase by 0.5 and approximately 12 μS/cm, respectively. This modification occurs for a four-hour period each day between 1:00 AM and 5:00 AM, after which the water quality values return to the actual measured values. The images of **Figure 6-2** show the data with the simulated patterns. The daily, four-hour long changes in the water quality signals are obvious in the 10 days of data shown.

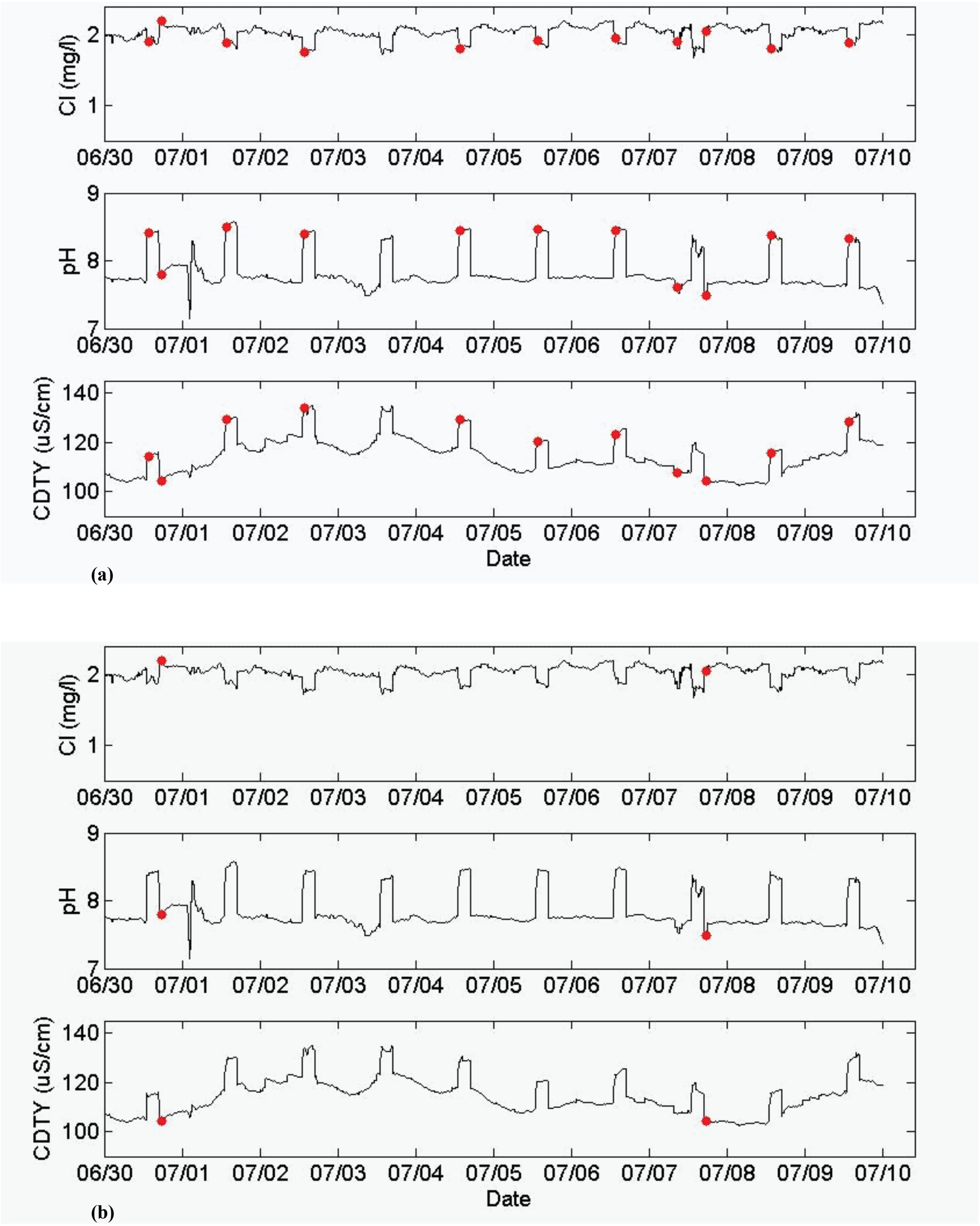


Figure 6-2. Water quality data and example event detection results for Location A with simulated patterns; (a) without the pattern matching and (b) with the pattern matching activated. Red dots indicate water quality events identified by CANARY.

A training data set consisting of 105 days of the modified data from March 10th to June 20th was examined to construct the pattern library. The linear prediction-correction filter (LPCF) algorithm with a window of 144 time steps (12 hours) and a threshold of 0.9 standard deviations was used. During this period, 121 events were identified by CANARY and these events were placed into three clusters. The polynomial fits to the events (grey lines), the mean polynomial fit for each cluster (black lines), and each water quality signal are shown in **Figure 6-3.** Water quality pattern 3 (**Figure 6-3,** right column) has the largest number of events (89) and corresponds to the simulated pattern of decreased Cl and increased pH and CDTY. The two other patterns account for 32 of the 121 total patterns identified. Note that it is not just the value but also the shape of the event that determines its pattern membership.

The testing data set consists of 72 days of data from June 20th to August 30th. These data were analyzed using the same parameters in the LPCF algorithm both with and without the pattern matching activated. Without the pattern matching, there were a total of 65 false alarms over the 72 day period. When the pattern matching was activated, the same data set produced only 14 alarms, a 79% reduction in the number of false alarms. Example results over a 10-day period are shown in **Figure 6-2** for the case (a) without the pattern matching and (b) with the pattern matching activated.

The ability of the pattern matching approach, both to reduce false positives and still identify water quality events, is evaluated by simulating water quality events on top of the water quality data containing the simulated pattern changes. An example of these components is shown in **Figure 6-4.** The simulated events decrease Cl and pH while increasing CDTY. The perturbations of the water quality signals during the events are of a similar order as the changes in the water quality patterns. It is easiest to distinguish the events from the water quality changes by looking at the pH signal in **Figure 6-4** where the water quality changes increase pH and the simulated events decrease pH. Each simulated event lasts for 1 hour (12 time steps) and a new event begins every 300 time steps so that the events take place at different times of the day throughout the testing data set. The events can occur during the simulated change in water quality due to the pattern or they can occur during times of unmodified water quality.

The testing data are analyzed again with and without the pattern matching activated. The same patterns identified in the training data (**Figure 6-3**) are used in the pattern matching. Without pattern matching, there are 104 false alarms and 95% of the true events are detected. With pattern matching, there are 39 false alarms and 92% of the true events are detected. These results show that although use of the pattern matching approach can reduce the number of false

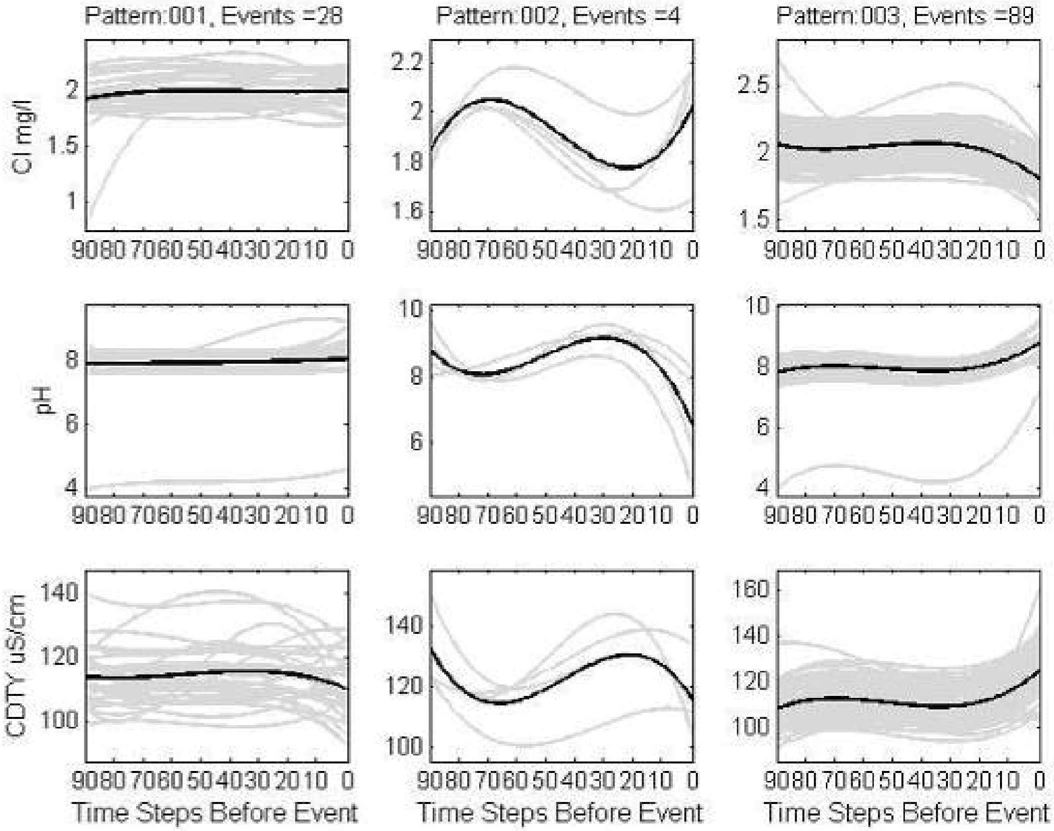
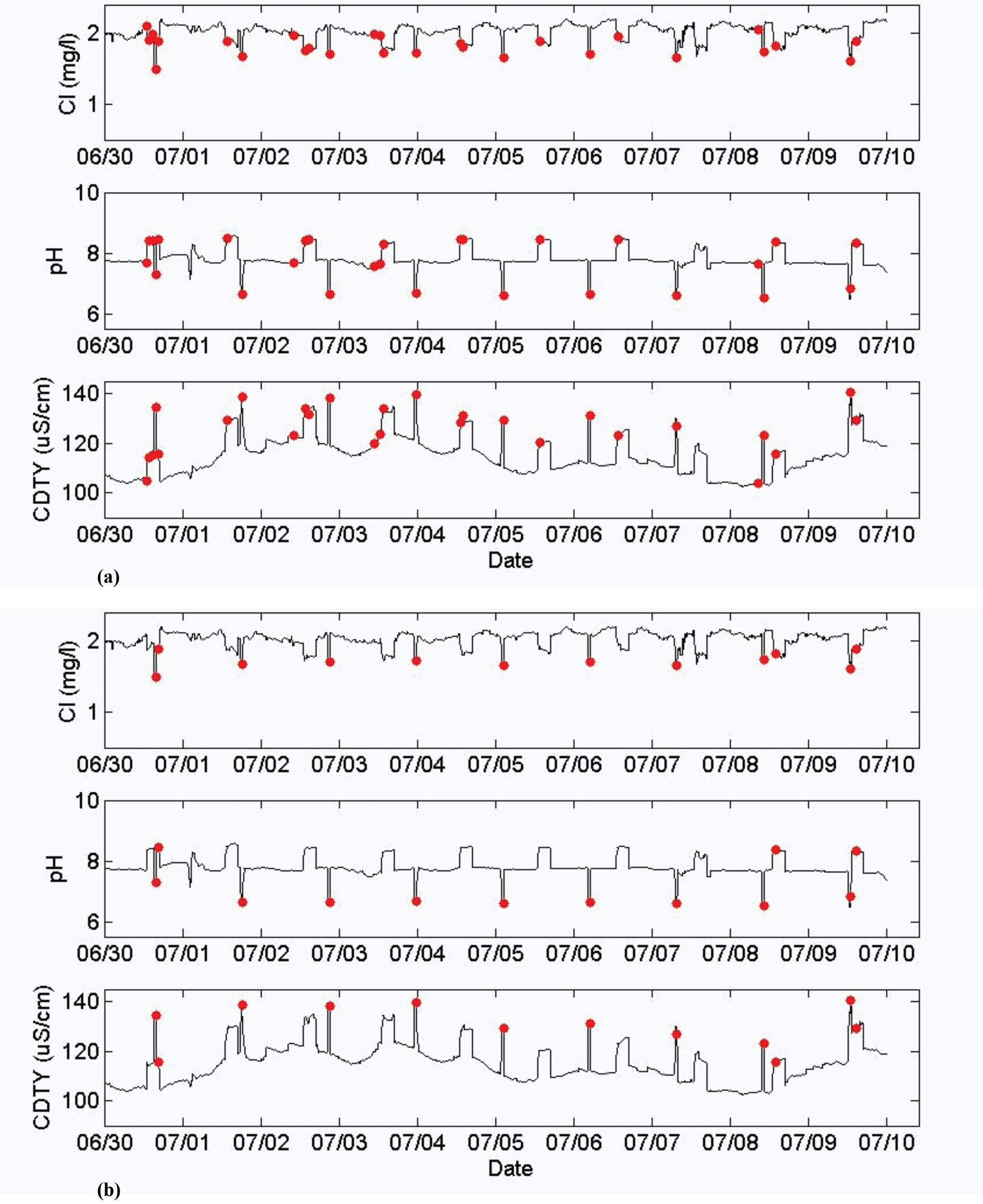


Figure 6-3. Water quality patterns identified in the training data for Location A. Each grey line is the polynomial fit to the 90 time steps prior to a water quality event. The black lines are the mean polynomial fit for each water quality signal and each pattern.

alarms by 62.5%, likewise it can decrease detection of true events by 3%. Comparison of event detection results with and without the pattern matching are shown in **Figure 6-4.**

Note the water quality events are simulated as repeatable patterns, but they are not incorporated into the pattern library. Therefore, CANARY correctly detects them as real events.



(a)

(b)

Figure 6-4. Water quality data and example event detection results for Location A with simulated patterns and simulated events; (a) without the pattern matching and (b) with the pattern matching activated. Red dots indicate water quality alarms, which correspond to the simulated events detected by CANARY.

Example 2: Treatment Plant Output

Data from Location B include the Cl, pH, and CDTY water quality signals, as well as the total output (flow) of the treatment plant in thousand cubic meters per day (TCMD). As in Example 1 (Simulated Patterns), the treatment plant output data were recorded with a 5 minute sample interval and are broken into two separate data sets for training and testing. The training data set is 121 days long (34,849 time steps) from January 1st through April 30th. The observed data in the training set are used directly for construction of the pattern library without any simulated patterns. The pattern library is constructed using two different approaches: 1) the three water quality signals and the total output (flow) are used for construction of the library; and 2) only the three water quality signals are used in construction of the pattern library.

The patterns constructed with the three water quality signals and the total flow values are shown in **Figure 6-5.** A total of 155 water quality events are identified in the training data and classified into two distinct patterns. **Figure 6-5** shows that the main differences between the patterns are in the Cl and flow values, while the pH and CDTY values are relatively flat for both patterns. In the case of Pattern 1, events are triggered by simultaneous decreases in Cl and flow, while in the case of Pattern 2, the onset of events occurs when there are slight increases in Cl and flow. Note that the changes in the water quality signals within the patterns are subtle with average changes in Cl on the order of 0.1 mg/L and changes in pH of approximately 0.1 to 0.2. These subtle changes are considered significant at this monitoring station, which is located at the outlet of the treatment plant.

A second pattern library is constructed from the training data using only the three water quality signals (Cl, TOC, and CDTY). The same 155 water quality events are identified, which is not surprising given that CANARY does not use operations data in event detection, only in the definition of the patterns after event detection. The number of resulting patterns (i.e., the number of clusters) increases from two, when flow was included, to eight, derived solely on the water quality data (**Figure 6-6**). Flow dominates the selection of patterns in **Figure 6-5:** one corresponds to flow decreasing (pattern 1) and one to flow increasing (pattern 2). When flow is removed from the analysis, the number of patterns increases, due to the more subtle water quality patterns not being overwhelmed in the clustering process by the more consistent flow data.

The testing data set is 61 days long and covers the period May 1st through June 30th (17,568 time steps). For this example, only the reduction in false positives (between the case in which no pattern library is used and the case that uses two pattern libraries described above) is examined; additional events are not added to the observed data. Single time steps are counted as individual false positives even though the majority of the events occurred in clusters of consecutive time steps. Without the pattern library engaged, CANARY identified 321 time steps as being anomalous compared to the background. With the pattern matching engaged, using either pattern library, CANARY identified 96 time steps as being anomalous. The pattern matching enabled a 70% reduction in the number of alarms. Results of an example 10 day period are shown in **Figure 6-7** for the case of (a) no pattern matching, (b) pattern matching using the pattern library created with flow data, and (c) without flow data.

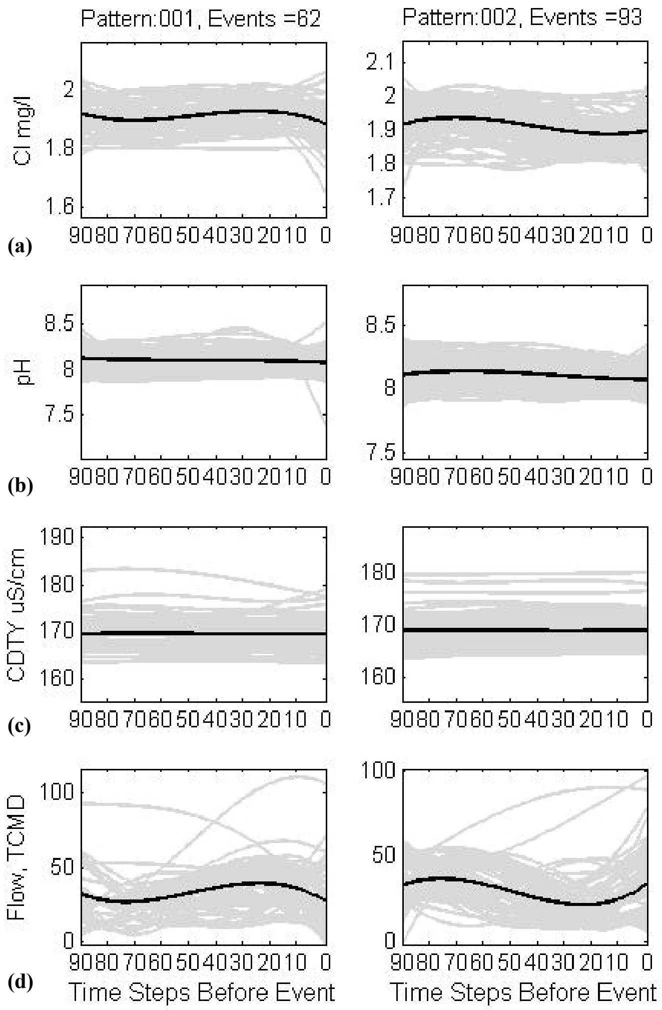


Figure 6-5. Water quality patterns identified in the training data for Location B when flow is included in the pattern definition; (a) Cl results in mg/L, (b) pH results, (c) CDTY results in μS/cm, and (d) total flow in TCMD. Each grey line is the polynomial fit to the water quality or flow data for 90 time steps prior to a water quality event. The black lines are the mean polynomial fit for each water quality signal and each pattern.

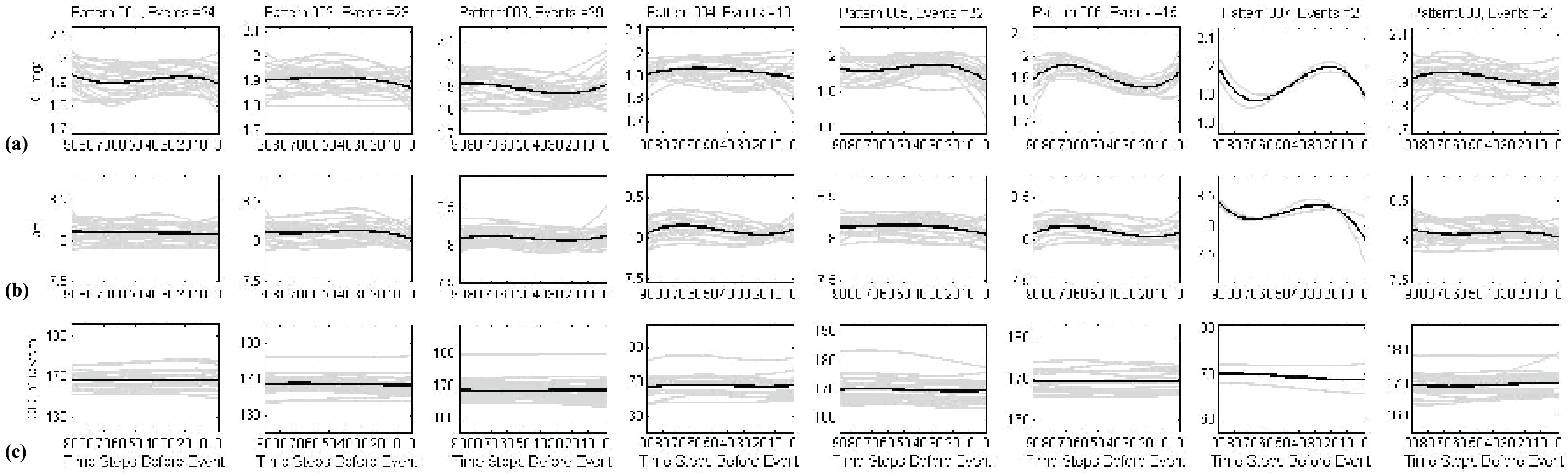


Figure 6-6. Water quality patterns identified in the training data for Location B, when flow is not included in the pattern construction; (a) Cl results in mg/L, (b) pH results, and (c) CDTY results in µS/cm. Each grey line is the polynomial fit to the water quality for the 90 time steps prior to a water quality event. The black lines are the mean polynomial fit for each water quality signal and each pattern.

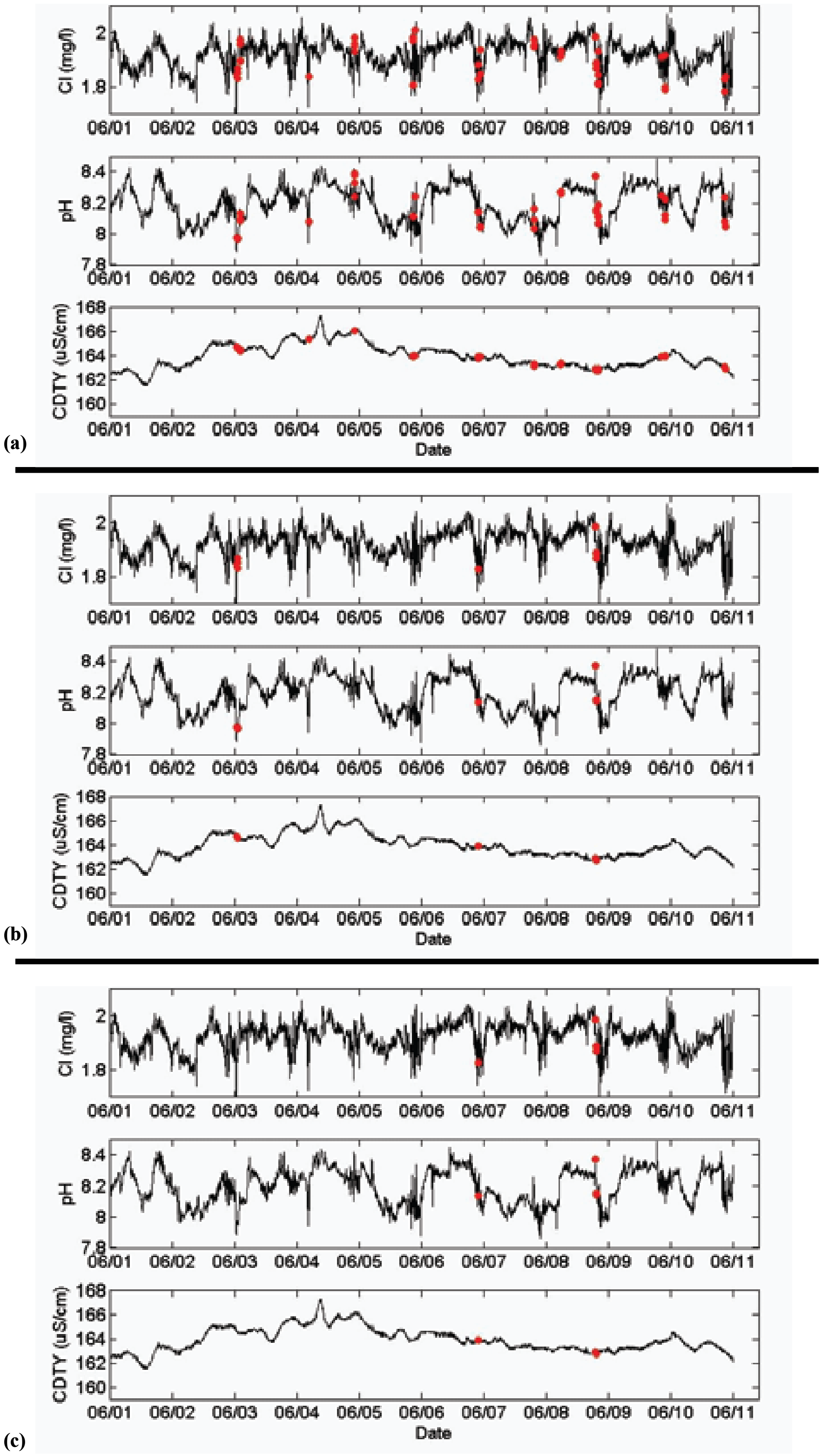


Figure 6-7. Event detection results on 10 days of Location B testing data with (a) no pattern library, (b) the pattern library constructed with flow data, and (c) the pattern library constructed with only water quality data. Events are noted by red dots.

In general, the two different pattern libraries created for Location B identified the water quality patterns and reduced the false positives by the same amount. Some differences are seen in **Figure 6-7b,** which shows an event near June 3rd that was not classified as a pattern using the library created with the flow data. The event was classified as a pattern when the library created without flow data was used (**Figure 6-7c**). The best approaches for integrating hydraulic data, such as flow, in combination with water quality data in this trajectory clustering method still need to be refined and additional experimentation is necessary.

Conclusions

Water quality patterns are not repeated exactly in distribution networks, but often at slightly different times of day and with slightly different expressions each time. These recurring patterns can create false alarms. Online application of trajectory clustering and use of a pattern library are proposed here as a means of recognizing water quality patterns and reducing false alarms. Application of this pattern recognition approach to two example problems demonstrates the ability to reduce false positive detections and still identify water quality events.

In the first example, three distinct patterns were identified that correspond to recognizable water quality changes in historical data. Using test data, CANARY was able to reduce the number of false positive detections from 65 to 14 (79%, compared to not using pattern matching); the reduction in false positive detections was 65% even if simulated water quality events were added on top of the data set. Detection of the simulated events for this data set was 92%.

The second example application demonstrates a case in which small variations in Cl of less than 0.2 mg/L or variations in pH of less than 0.3, over the course of several time steps, can trigger a water quality event. Pattern matching allows for the CANARY event detection algorithms to be set to this level of sensitivity while maintaining a reasonable number of false alarms. The false alarm reduction in this example was 70% compared to not using pattern matching and is consistent with reductions observed in the first example. Monitoring with this level of sensitivity might not be routinely necessary but can be implemented during times of heightened security. Additionally, as demonstrated here, monitoring at this level of sensitivity at the outlet of a treatment plant allows the rapid identification of subtle variations in treatment plant output. This capability opens the possibility of using CANARY output as feedback for improved treatment plant control and optimization during routine operations.

This work demonstrates an adaptation of trajectory clustering for online event monitoring and real-time pattern matching. This approach to pattern matching was specifically developed to be an extremely general approach to integrating multivariate water quality signals and other information into online identification of recurring patterns. In addition to water quality signals, any operations data signal such as tank levels, flow rates, pressure data, and valve settings can be used in the definition of the water quality patterns. These additional signals could even include a "time of day" signal to improve recognition of patterns that occur at similar times each day. Integration of these other data streams allows for tighter coupling between changes in network operations and resultant changes in water quality within the event detection framework Integration of these additional data streams is generally applicable and does not rely on creation of complicated rule sets for use of these additional data streams.

Additional testing and application of the trajectory clustering approach to pattern matching will identify improvements that can be made. Two current ideas for improvement are sequential comparison of water quality data to the pattern library and more efficient pattern library construction. The current implementation of pattern matching compares only the current water quality to patterns in the library at the time of event detection. Future extensions of the trajectory clustering approach will allow for comparison of the current water quality to the pattern library at multiple times both prior to and after the declaration of an event. This sequential comparison to patterns in the library will allow for increasing evidence for or against a pattern match to be developed through time. Construction of the pattern library currently requires the water quality analyst to examine several months of data and make a decision on every identified event during that period as to whether or not it should be included in the library. This approach is necessary for each change in the event detection parameters. A more efficient way of doing this would be to retain only the days containing events in which the analyst is interested in classifying rather than examining several months of data.

7.
Distributed Network Fusion for Water Quality

If CANARY is running at multiple monitoring stations in a water distribution system, how can the separate outputs be interpreted at the same time, so as to improve event detection? This chapter presents an approach for fusing CANARY outputs from multiple monitoring stations. It should be noted that this is a current area of research and has not yet been implemented in CANARY.

The previous chapters of this document have focused on the detection of anomalous water quality at a single monitoring station that has one or more sensors, assuming that results obtained at one monitoring station are independent of results from other monitoring stations. This independent analysis approach is designed for situations where the size of the plume resulting from any contaminant injection is small relative to the spatial density of the sensors. In such cases, it is expected that any given contaminant plume would only intercept a single monitoring station.

The concept of distributed detection, however, is based on the assumption that a typical contaminant plume will be detected by more than one monitoring station, and that the outputs of multiple monitoring stations can be combined to provide a network-wide indication of a water quality event. A way to combine the results from multiple monitoring stations together is needed in order to achieve distributed detection capability.

To the authors' knowledge, no study has been completed on the typical number of monitoring stations necessary to detect a contamination incident. To some extent, this is not surprising because the number of sensors, the design of the sensor network, the location and duration of the contaminant injection, and the layout and demand patterns of the underlying distribution networks are all variable and make comparison of results difficult.

The goal of this work is to use the topology of the water distribution network and a sensor fusion approach to combine multiple detected events together. Each monitoring station with sensors and associated event detection system (EDS) is referred to as a sensing-node, and a change in water quality detected by the EDS is called an event. The approach serves both to reduce the number of false alarms and missed detection errors, as well as to determine the injection location of the contaminant accurately.

In detection, two types of errors are: false alarm (FA) errors and missed detection (MD) errors. FA errors occur when the change detection algorithm generates an alarm on a nonexistent event and MD errors occur when the change detection algorithm misses an actual event. The two errors are linked, so that decreasing one of the errors increases the other. If the FA errors are too numerous, then operations personnel will begin to mistrust the system, but if there are too many MDs, then the system becomes ineffective. The number of FA errors increase when more sensing-nodes, which determine the injection location and extent of the contaminant plume, are added. Suppose a single sensing-node has one FA per day, then a network of 100 sensing-nodes would have four FA errors per hour! This, of course, would be an unacceptably high rate in a working system.

Figure 7-1 shows a space-time cube for a water distribution system called Anycity with 100 nodes randomly selected as containing sensors. The width and depth dimensions of the cube show the spatial dimension with the water distribution network shown at the top and bottom of the cube. The time dimension is along the height of the cube with time increasing from bottom to top. The circles represent detections from a change detection algorithm operating on a simulated contamination event. The open circles show randomly generated FAs, assuming a sensor at every junction and a single sensing-node FA rate of once per day. The entire cube represents sensor activity over 25 hours. For one FA per day at each monitoring station, an average of 104 FAs is expected within the cube. Using EPANET (Rossman 2000), a tracer injected into the network is simulated. This tracer represents the contaminant and the solid circles show the detections of this tracer. In actuality, it is not known if the detections were real or FAs (solid or open circles). Using sensor fusion, the idea is to separate the real detections from the FAs and reduce the errors of the entire system.

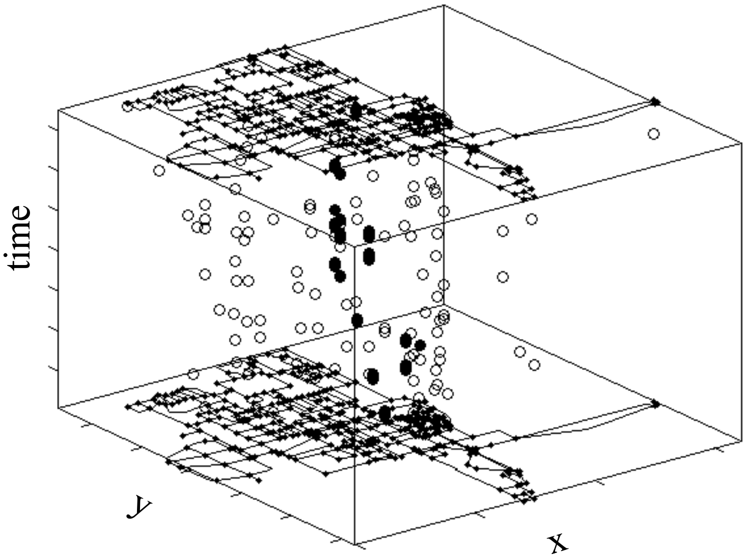


Figure 7-1. Space-time cube of Anycity with simulated sensors at every junction. The water distribution network is shown in the space dimensions (width and depth) and time is the height dimension from bottom to top. The circles represent detected events with open circles shown as false alarms and filled circles as correct detections.

The location and time of an event is considered as a point resulting from a random space-time point process (Kulldorff 1997). Clusters in space and time that are significantly different from background indicate a set of true detections, whereas a set of purely random events indicate false detections. Kulldorff's scan statistic (Kulldorff 1997) is used to fuse the detections of the sensing-nodes in the distribution network (**Figure 7-1**) through the identification of statistically significant clusters of events in space and time. The location and size of the significant clusters indicates the location and the extent of the contamination. Scan tests count events within sliding windows over a multidimensional area A and use the counts contained within A to determine if there is a cluster of significant events. For a water distribution network, the space within which the event detection occurs is not continuous, but is defined by the topology of the distribution network. The proximity of any two sensing nodes is defined by the velocity of the pipe flows between them. The distribution network's topology and a rough measure of flow velocity are used to define the space dimension.

To test the distributed detection algorithms, EPANET is used to simulate a city's water distribution system. By combining EPANET's simulation of the transport of a tracer and the performance models of the event detection algorithms, it is shown how multiple sensing-nodes can improve the event detection performance relative to the use of a single sensing-node. How the distributed detection system's performance changes with the number of sensing-nodes, and how well the scan test can determine the injection location and time of the contamination, is also demonstrated.

Related Work

Recent research on using water quality measurements to identify periods of anomalous water quality has focused on data obtained at a single monitoring station. Various algorithms have been applied to these data sets to extract anomalous signals from the often noisy water quality background (e.g., Cook et al. 2006; Jarrett et al. 2006; Kroll et al. 2006). Research at Sandia National Laboratories has involved development and testing of multiple, robust multivariate statistical algorithms (Klise et al. 2006a; Klise et al. 2006b; McKenna et al. 2007; McKenna et al. 2006b), which are embedded in the CANARY software (Hart et al. 2007; Hart et al. 2009). The algorithms provide a means of automatically detecting changes in water quality sensor measurements by comparing the current measurements to their predicted values based on their previous history. Essentially, the algorithms create a current measurement vector from all the available sensors. This measurement vector is compared to a prediction vector based on previous sensor data (see Chapters 3 and 4).

The concepts of distributed detection, where sensor responses from multiple monitoring stations across a network are fused to provide a "network-wide" detection capability, have not been fully applied to water distribution networks. Initial work towards integrating responses from more than one monitoring station has recently been reported (O'Halloran et al. 2006; Yang et al. 2008). In both of these studies, the authors use water quality sensors at two monitoring stations to improve the water quality signal. One of the monitoring stations acts as a reference that allows for adaptive compensation at the second station to account for variable time delays between the two sensors as well as calibration errors and background noise in the second (downstream) sensor.

Approach and Implementation

As a hypothetical base case calculation, randomly placed sensor nodes in a network are considered. Each node is assumed to have perfect detection capability and work independently, which is the non-fusion approach. The hypergeometric distribution provides the probability of detection under these conditions. If *m* sensor nodes are randomly placed in a network of *M* junctions, then the probability of having at least *x* detections in a contaminant plume that has a size of *X* junctions is

$$P(x \mid m) = \sum_{i=x}^{X} \binom{X}{i}\binom{M-m}{X-i} \Big/ \binom{M}{m} \qquad (7\text{-}1)$$

where $\begin{pmatrix} M \\ m \end{pmatrix} = \dfrac{M!}{m!(M-n)!}$ represents the number of combinations of *M* things taken *m* at a time. The non-fusion approach uses the Anycity network (M = 396) and assumes a contaminant plume size of 20 nodes. In order to have a 0.99 probability that one sensor is in the contaminant plume, 80 sensors are needed (**Figure 7-2**). The sensors are assumed to be perfect and placed randomly throughout the network.

The Y-axis of **Figure 7-2** is the complement of the fraction of missed detections that is often used as a performance measure in sensor optimization studies (e.g., Watson et al. 2004). The underlying assumptions of Equation 7-1 are that the sensors have perfect behavior (i.e., no FAs and no MDs), as is often assumed in sensor network design studies and that the sensors are randomly placed, which is in contrast to most network design studies that use some form of optimization to place the sensors to minimize public health or economic impacts.

In reality, sensors are not perfect, and for large numbers of imperfect sensors, there is the potential for a large increase in FAs, as discussed above. To remedy this problem, a modest increase in the number of sensing-nodes and a distributed fusion approach to combine results at multiple nodes to reduce the level of FA errors is proposed.

To combine detections from multiple sensing-nodes, Kulldorff's scan statistic (Kulldorff 1997) is used. Kulldorff calls this set of sliding windows used to examine the multidimensional area, A, *zones*. Significance is determined by comparing the count value to a null distribution of counts that would be obtained from a random point process.

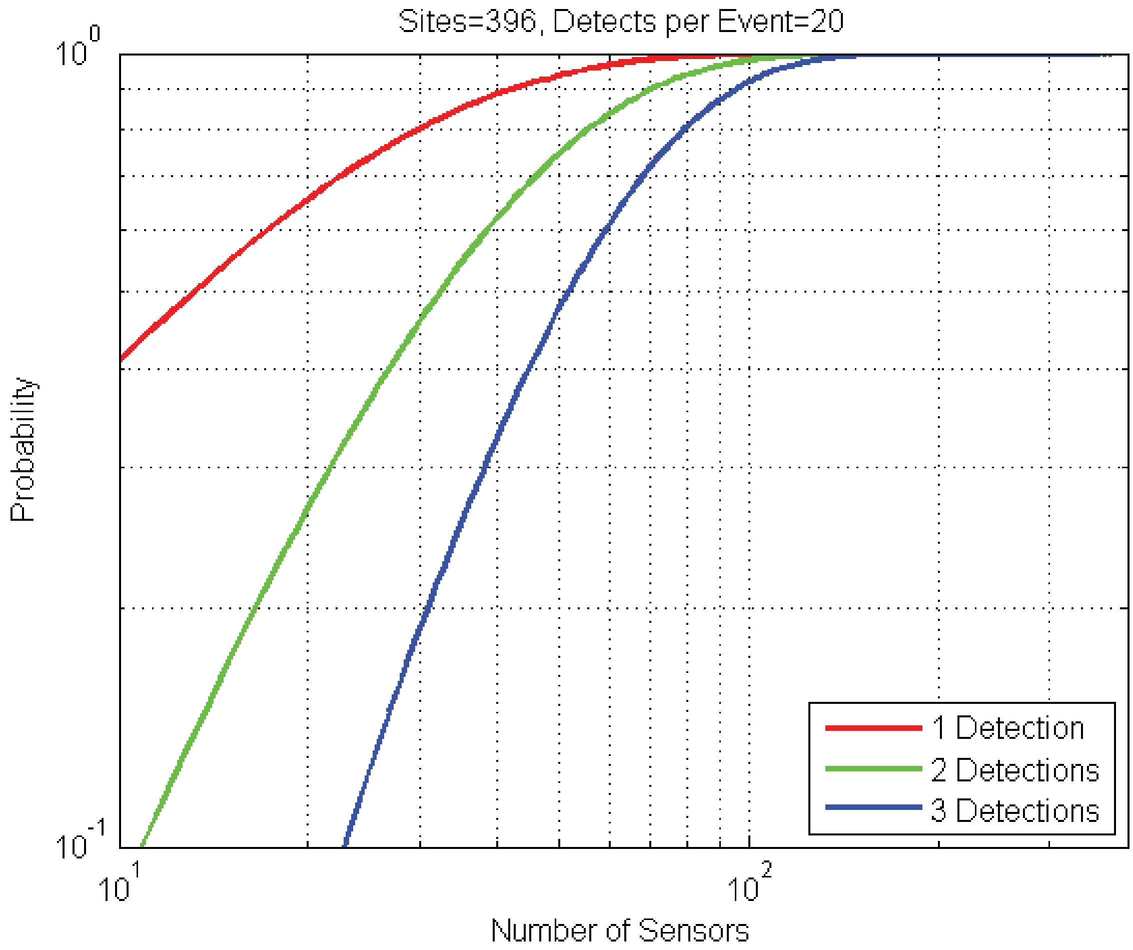


Figure 7-2. Probability of one, two, or three sensors detecting a plume with a size of 20 junctions within a 396 junction network as a function of the total number of sensors deployed.

One possible approach would test each possible cluster for significance. Each test would not necessarily be independent, since the clusters are overlapping and events in one cluster could appear in another cluster. The multiple-testing problem is also a concern. If enough tests are made, then it is more feasible to disprove the null hypothesis H_0 (no significant clusters) and, thus, increase the FA error. Adjustments can be made to solve the multiple-testing problem, but these become too conservative, especially for dependent tests.

Kulldorff avoids the multiple and dependent testing problem by clearly defining the alternative hypothesis H_1. Kulldorff's H_1 constrains the search to at least one significant cluster in an area, A, of space-time. This in turn defines the null distribution H_0 and, thus, the threshold for decision of a significant cluster. This threshold can be used as a conservative estimate for multiple clusters and/or over smaller areas, but never areas larger than A.

Although computationally intensive for estimating the null distribution, Kulldorff's approach can handle multiple dimensions and overlapping zones of different sizes and shapes and directly determine the locations of the clusters. By using a likelihood ratio and a clearly defined alternative hypothesis, it avoids the multiple and dependent testing problem. It is also a unique test, making it unnecessary to perform a separate test for each cluster size and location. The binomial version of the test (Kulldorff et al. 1995) is used, and it is assumed that each sensing-node produces a yes/no or 1/0 decision on the presence/absence of an event.

Kulldorff's scan test is conditioned on the knowledge of the total number of events C. The geographic area and time interval of interest is needed to define A, as well as how the region is covered with the set of all zones Z. Kulldorff's test has two hypotheses:

1. Null hypothesis H_0: For all the zones, the probability of an event inside the zone, p, is the same as that outside the zone, q, i.e., $p=q$.
2. Alternative hypothesis H_1: At least one zone has the probability of an event inside the zone being greater than the probability outside, i.e., $\exists z \in Z \mid p > q$.

The likelihood function $L(z,p,q)$ for the scan test is:

$$L(z,p,q) = p^{c_z}(1-p)^{n_z-c_z} q^{C-c_z}(1-q)^{(N-n_z)-(C-c_z)} \quad (7\text{-}2)$$

and represents the likelihood that the number of events inside zone z is c_z and the number of events outside zone z is $C - c_z$. N represents the total number of possible events in A and nz represents the number of possible events in z. Using **Equation 7-2,** the likelihood ratio becomes:

$$\frac{L(z)}{L_0} = \frac{\sup\limits_{z \in Z, p>q} L(z,p,q)}{\sup\limits_{p=q} L(z,p,q)} \quad (7\text{-}3)$$

where L_0 indicates the value of the likelihood function under the null hypothesis and sup denotes the *supremum* of the set, or the minimum value in the set that is greater than or equal to every number in the set. Thus, the scan test uses the largest likelihood ratio to combine results from multiple zones. The scan test statistic λ is:

$$\lambda = \frac{\max\limits_{z} L(z)}{L_0} \quad (7\text{-}4)$$

In general, the distribution of λ has no simple analytical form. To determine the distribution of λ for the null hypothesis, Kulldorff suggests using Monte Carlo randomization. Since the test is conditioned on the number of cases C, random examples using $p = C / N$ (sensing-node FA error estimate) can be generated and the scan test for each can be computed. As long as the number and performance of the sensing-nodes stays the same, then estimation of the null distribution can be accomplished offline and prior to application.

Implementation

To fuse information from multiple sensing-nodes in a water distribution system, Kulldorff's scan test is used. **Figure 7-3a** shows a hypothetical time series for three sensing-nodes, which assumes sensing-nodes B and C are downstream from sensing-node A and that the sensing-node produces a 1/0 decision for event and no event, respectively. In the figure, a dot with a stem represents a 1 and just a dot represents a 0. For this problem, there are two dimensions: space and time. The network of pipes and junctions represents the space dimension. The space dimension is represented as a series of travel time connections between sensing-nodes. EPANET simulations are used to find the median value of the absolute value (i.e., no direction) of the flow velocity over a 24-hour period for each pipe in the network, and then these median values are used to compute an estimate for the travel time between any two nodes. The direction of the flow is not recorded or used, only the median time delay between nodes. The median of the absolute values of the simulated travel time over a 24-hour period is an acknowledged rough estimate of the actual travel time between nodes, which could be significantly shorter than 24 hours and also dependent on the time of day. In any practical application, the actual travel times will always be uncertain and the median value over a 2 hour period is used to represent an estimated travel time.

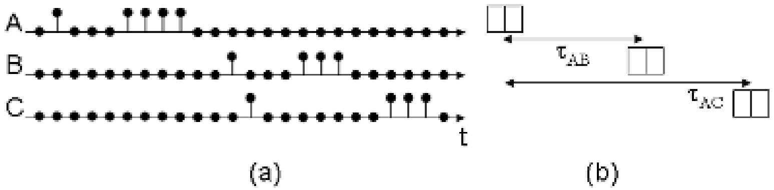


Figure 7-3. (a) Example hypothetical time series for three sensing-nodes A, B, and C. A dot with a stem represents a 1 or a detection and a dot with no stem represents a zero or no detection. (b) A 3 x 2 space-time template for a space-time cluster centered at node A.

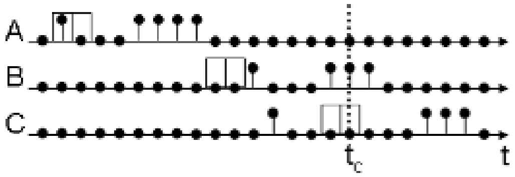


Figure 7-4. The 3 x 2 template aligned with the time series in space and time at current time t_c.

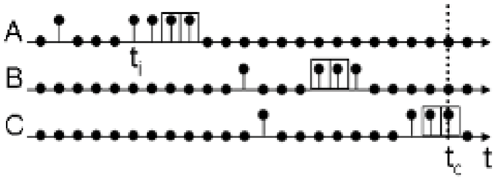


Figure 7-5. The 3 x 2 template aligned with the time series in space and time at a new time t_c.

To search for clusters of detections, the zone sizes in space and time as $s \times \tau$ are specified. Here s represents the size of the zone in space and s represents the size of the zone in time. The space size s represents the number of sensing-nodes closest to and including the center of the zone (in units of travel times). For example, **Figure 7-3b** shows a template for a 3 x 2 space-time zone centered at node A. The number of adjacent boxes represents the zone size in time ($\tau = 2$) and the number of sets represents the zone size in space ($s = 3$). The horizontal distance τ_{AB} represents the estimated travel time between nodes A and B and τ_{AC} represents the estimated travel time between nodes A and C. This zone template searches for clusters that have a contaminant injection at node A.

Figure 7-4 shows the 3 x 2 template aligned with the time series in space and time for the current time t_c. The leading edge of the template is aligned with the current time for sensing-node C. The total number of detections in this template is one as seen at node A. This becomes c_z in **Equation 7-2.** Even though an event was detected at node A, there is no correlating evidence at the other nodes, so the scan test would not find the single detection at node A to be a significant cluster for this space and time.

Figure 7-5 shows the template advancing to a new time in which six detections occur. If a significant cluster is assumed to be detected and the contaminant was introduced at time step ti, then the detection delay is given by $t_{\mathrm{c}} - t_i$. The scan test zones are defined such that when a significant cluster is identified, the central node, node A in **Figure 7-5**, is the source node for that cluster. The distributed detection approach provides both event detection and injection location identification.

Since the actual size of the space-time cluster is unknown, multiple zone templates of different sizes need to be tested. The contaminant injection location is also unknown, so zone templates that assume an injection at the other sensing locations within the template need to be tested. At each point in space and time these zones are combined by taking the one that produces the largest scan test score (**Equation 7-4**). Note, for $\tau > 1$, counts in the zone template at one time could be used for counts for a zone template at neighboring times. Because of this overlapping in time and the different zone sizes, the random variables representing the counts are not independent. This dependence makes analytical determination of the null hypothesis difficult.

Evaluation and Results

To evaluate this distributed fusion approach, an event simulator called DetectNet was built using MATLAB® (MathWorks 2008) and the EPANET developer's toolkit (Rossman 1999). The objective of DetectNet is to simulate sensing-node detections from an algorithm like CANARY (Hart et al. 2007) in a water distribution system.

The heart of DetectNet is EPANET. EPANET takes a description of a water distribution system, including stochastic demands and a chemical tracer, and determines the concentration of the tracer throughout the network at different time steps. This tracer serves as a proxy for a contaminant introduced into the water distribution system. The parameters specified for the tracer are initial concentration, start time, and length of the tracer injection. DetectNet takes the tracer simulation results and produces a set of detections based on the performance characteristics of a suite of sensors and the associated event detection algorithm (e.g., CANARY). The performance characteristics of the imperfect sensors and EDS algorithms are combined and are based on FA and MD errors. A pseudo random number generator is used to add extra detections to the "ground truth" tracer simulation based on the FA error and remove detections based on MD errors. EPANET is also used to extract the network constraints for the sensor fusion algorithm. These constraints are the connectivity of the network and median travel times between junctions.

Figure 7-1 shows the Anycity network used for the simulation. The network has 396 junctions, 534 pipes, 2 tanks, 4 valves, and no pumps. The simulation runs for 24 hours with 1 minute time steps. For randomly selected junctions with zero base demand, a 30 minute tracer injection with a concentration of 50 mg/L is simulated. The tracer's concentration decreases as it moves through the network and mixes with non-tracer water from other parts of the network. Assuming there is a sensor at a junction, if the concentration at that junction is greater than 5 mg/L (detection limit), then a possible true detection by the sensing-node is allowed, otherwise only false detections are allowed. Simulation runs that gave an average plume size, at concentrations above the detection limit, equivalent to a portion of the network that would contain 20 nodes are selected.

For sensing-node performance, a 10 minute sample interval, a FA rate of 1/144 (once per day), and a 0.01 MD error are assumed. This FA error was selected as a plausible worst case performance that demonstrated the fusion algorithm's abilities to identify a contaminant in background clutter. The FA error does not necessarily reflect current or projected sensor node performance. Kulldorff's scan test is applied with varying numbers of sensor nodes whose locations were randomly selected. Sensors at 396 (all junctions), 200, 150, 100, 50, and 20 sensing-nodes are investigated. Using Equation 7-1 and 20 sensing-nodes, very good results are not expected, since there is less than a 20% chance that at least two nodes will randomly be placed in the contaminant plume for the contaminant injection characteristics used here. For each set of sensing-nodes, 100 different days of background clutter data using the FA error rate are generated and the scan statistics for each time step and cluster location are computed. The exact location of the sensing-nodes does not change the null distribution, since the space dimension is based on the s closest nodes. For all sensor configurations, all combinations of clusters sizes of 1, 3, 6, and 12 in space combined with 1, 3, and 6 in time are used, except for the single cell case, $s \times \tau = 1$ x 1. These cluster sizes were chosen based on the size of the network and estimated median travel times between nodes.

Using EPANET, the introduction of a contaminant at five different junctions is simulated. For each separate injection location, 100 one day simulations with different randomly selected sensing locations and different background FAs are generated. This gives a total of 500 different simulations used to evaluate the performance of the approach. **Figure 7-6** shows a portion of the Anycity network overlaid with sensing-node detections and significant clusters. A sensor at every junction is indicated by the black dots. Circles represent all the detections up to and including the time shown in lower left corner. Circles filled with white are true detections, and circles not filled (shows junction and links) represent FAs. The relatively high FA rate results in every sensor node experiencing at least one FA by the end of the simulation (see **Figure 7-6b**). In actuality, the truth of the detections is unknown, but this labeling makes it easy to see how the distributed detection is performing. The contamination is introduced at 12:00 AM.

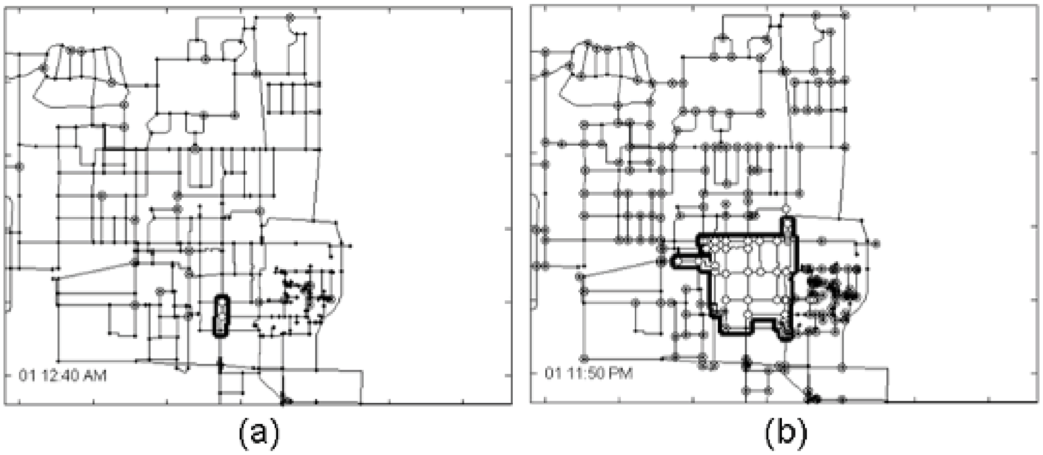


Figure 7-6. Example results for scan test with 396 sensors. Circles represent detections. Circles filled with white represent true detections and circles with no fill (show junctions and links) are false detections. The solid black line indicates the extent of the significant cluster at this time step. (a) First significant cluster detected. (b) Intersection of all significant scan clusters after 24 hours.

In **Figure 7-6a**, the heavy black line shows the first significant cluster detected by the scan test at 12:40 AM. Thus, it took 40 minutes after the introduction of the contaminant to detect a significant cluster. **Figure 7-6b** shows the intersection of all the significant clusters for that day. The scan test accurately reflects the extent of the contamination plume, though it does include some FAs near the bottom. This is because any knowledge of the flow direction in any one link is not used. Note the actual contaminant did not spread significantly in the southerly direction.

Figure 7-7 shows the results for 100 sensing-nodes or 25% coverage for a scenario where two contaminant injection locations were used. **Figure 7-7a** shows that it takes 4.5 hours to detect the plumes from both contaminant injections. **Figure 7-7b** shows the results after 24 hours.

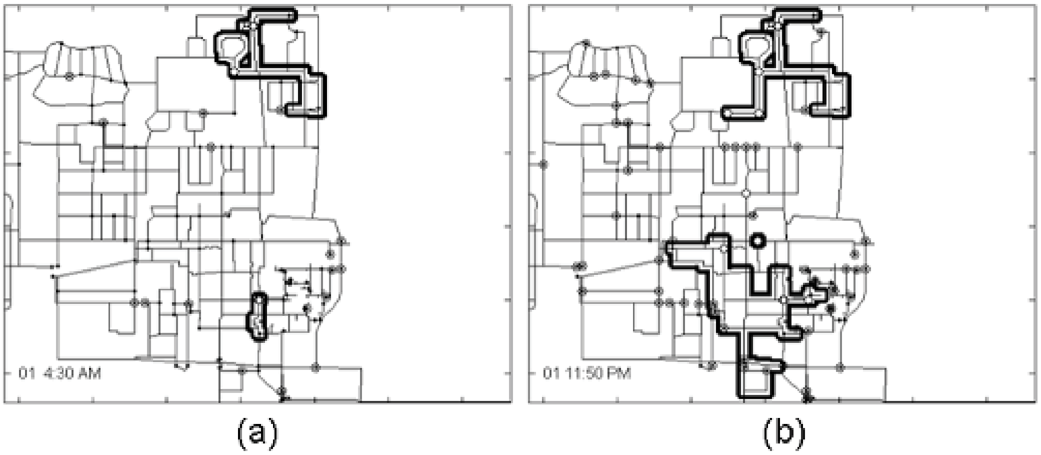


Figure 7-7. Scan test results for 100 sensing-nodes and two contaminant injections. (a) Results when both injections are first detected. (b) Results after 24 hours.

Figure 7-8 shows the operating characteristics for different numbers of sensors. A correct detection is identified if the scan test finds a significant cluster intersecting the contaminant plume. A false detection is identified if the scan test finds a significant cluster during a day of background clutter. The scan test has excellent performance until the number of sensors drops to 50 or below. At this point, the chances that at least two sensors will be within the contaminant plume start to drop rapidly. For more than 50 sensors, the results are excellent considering the high numbers of individual sensing-node FAs. At 100 sensors, there are very low errors. Recall that 80 sensors are needed for the non-fusion single-detection approach, assuming perfect detection, to achieve 99% detection of plumes with a size of at least 20 junctions. Thus, given imperfect sensors with the FA and MD rates specified here, the distributed fusion approach requires a 25% increase in numbers of sensing-nodes relative to having perfect sensors in the non-fusion approach.

Figure 7-9 shows histograms for the time to detection (hours) and distance to detection (number of links) for the 500 simulated cases. These statistics are based on the time and location of the first significant cluster to be detected (assuming there is a detection). The threshold for distributed detection is selected to achieve one FA per 100 days. As expected, the time to detection and distance to detection increases as the number of sensors decrease. If there is a sensor at every node, then the contaminant is detected within one hour and the correct injection location is identified more than 50% of the time. For 100 sensors, those results increase to 3.5 hours and the injection location is estimated with a maximum error of two network links (pipes) away from the true injection location.

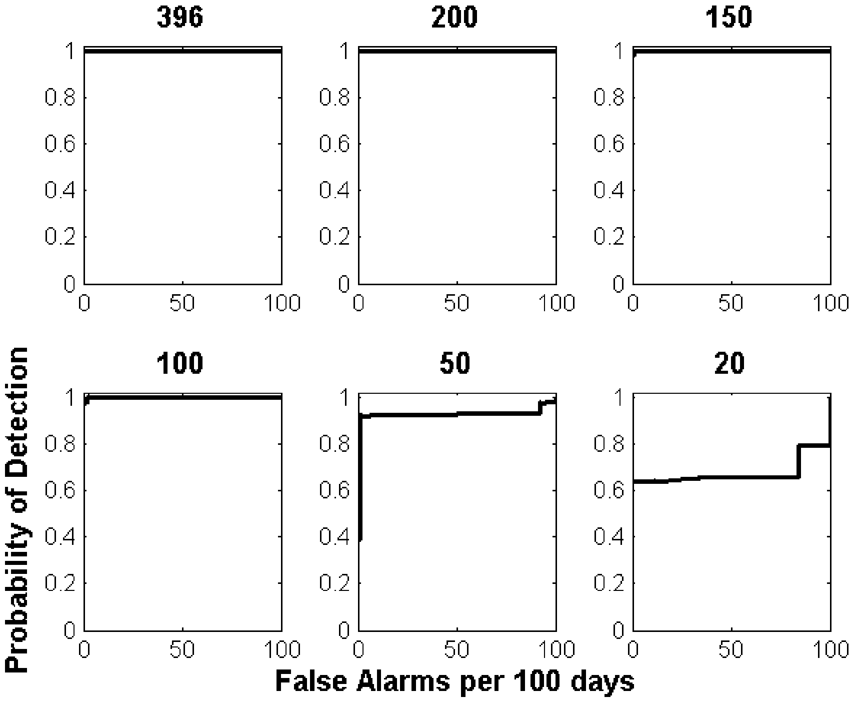


Figure 7-8. Operating characteristics for varying numbers of sensors.

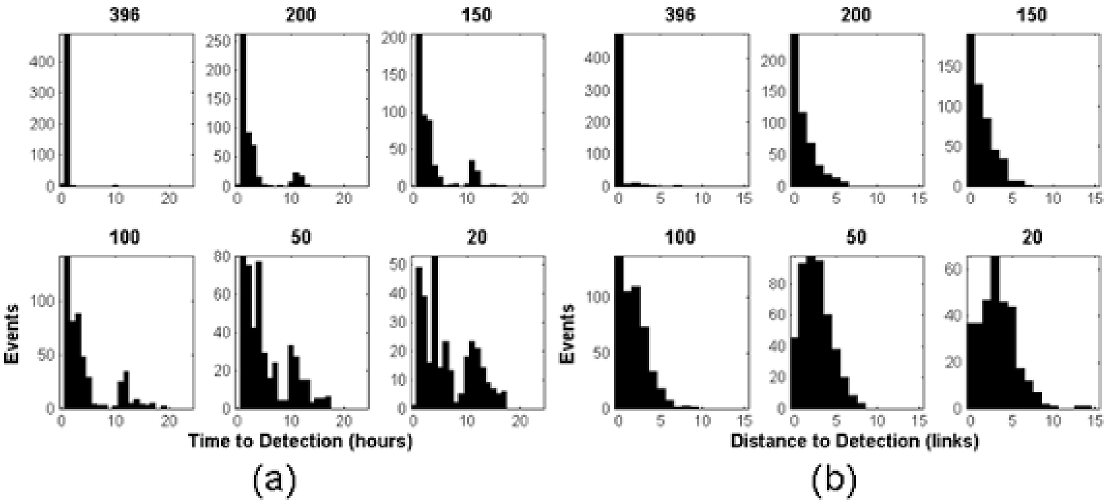


Figure 7-9. Detection statistics. (a) The time to detection (hours) for different numbers of sensors. (b) The distance to detection (links) in terms of network links. Both are based on the center of the first cluster to be detected.

Note that the time to detection histograms become bimodal as the number of sensors decrease. The authors hypothesize that as the number of sensors decreases, it is more likely that the first sensor to encounter the contaminant will be on the edge of a plume than the center, since the plume expands as time increases. Sensors in the plume center eventually detect the contaminant, but the delay to detection increases.

Conclusions and Future Work

This work has presented a new approach to integrate independent event detection results into a consistent network-wide event detection strategy. This approach is designed to use the binary output (event/non-event) of an EDS such as CANARY, along with basic information on the network connectivity to identify events that impact multiple monitoring stations within the network. Kulldorff's scan test has been applied to the problem of detecting contamination using multiple sensors in a water distribution network. Kulldorff's test identifies significant clusters in space and time and can distinguish between clusters of true events from random background alarms. As the number of sensors in the water distribution network increases, the chance of a FA increases too. This makes it difficult to separate false detections from true detections. The approach developed here is general enough to handle improvements in change detection algorithms, such as potential contaminant identification, and real-time estimation of flow rates and directions from the network model. In the example network studied, a 25% increase in sensing-nodes from the non-fusion single-detection approach allows the perfect sensor assumption to be dropped and the distributed fusion results in very low error rates.

Currently, Monte Carlo simulation is used to estimate the null distribution. Re-estimation of this distribution is required if the number of sensors changes or the sensor performance characteristics change. Another approach would use a Bayesian scan test that would make more assumptions about the characteristics of the null distributions. The Bayesian approach would not require the time-consuming Monte Carlo techniques to estimate the null distribution. It is noted that, although time consuming, the current Monte Carlo calculation of the null distributions is done offline using the assumed FA rate prior to the detection data becoming available. This makes the distributed detection approach developed here capable of functioning in a real-time mode.

Tracking the detections through the network and improving how to determine the extent of contamination is important for knowing how to respond to an event. Tracking involves determination of which clusters are associated at different time steps and which belong to different contaminant plumes.

The distributed detection approach provides both event detection and injection location identification. In this present approach, the detections are projected back in time to determine the best estimate of the injection location and time. To improve determinations of the extent of contamination, the detections could also be projected forward in time. This would give more support to the detections at the edge of the plume and might guide the placement of portable sampling units to further identify and characterize the contamination event.

Appendix A
Literature Review and Glossary

This literature review covers methods for analyzing offline and online data using what are called change point or event detection schemes. These topics have generated a considerable number of research publications in a diverse set of fields over the past 20 to 30 years. Only key developments and publications containing illustrative examples of those developments are included in this review. Specific applications of these methods to water quality monitoring are explored in more depth.

The publications reviewed are considered from the perspective of a two stage approach to event detection. The first stage provides a prediction of a future water quality value. This prediction is most often based on previous water quality values and the process of making the prediction is referred to as *state estimation*. In the second stage of event detection, the prediction of the expected water quality value is compared to the observed water quality value as it becomes available. The difference between the prediction and the observation is termed the *residual*, and the residual is used to classify the water quality at that time step as either being expected (or representative of the background water quality) or anomalous. This second stage of *residual classification* determines whether or not the observed water quality is significantly different from the expected water quality. These two stages to event detection are further defined below and example publications for each step are examined. Finally, a glossary of terms is added to this literature review.

General Approaches

Water quality data collected from online sensors in water distribution systems can be thought of as a time series – a sequence of data values collected over time. Similar to many time series, water quality data streams are typically recorded at a constant sampling rate (e.g., every 10 minutes). The general problem of detecting anomalous behavior in time series data is a subject of research in a number of disparate fields, including tsunami detection, traffic accidents analysis, mechanical component failure, system fault detection, data mining, and network intrusion detection among others. This work can be categorized into two distinct approaches described here as *"online"* and *"offline."* The online approaches receive data in discrete time steps ranging from milliseconds to minutes and provide a determination of the presence or absence of an anomalous reading immediately after the receipt of each new data point. Offline approaches generally require all data to have been collected, and a retrospective survey of the data is then completed to identify change points within the data set. The use of online and offline to describe event detection algorithms is not necessarily the same meaning as when these terms are used to describe *Supervisory Control and Data Acquisition (SCADA)* connections. *Event detection system (EDS)* tools can be run in online mode where they are connected to a SCADA system and receiving data and providing analyses in real-time. These same tools can be also be run in offline mode where historical data are analyzed as if they were coming to the EDS in real-time.

These approaches are considered below, in greater detail, in the section on event detection. Offline approaches are used to analyze previously collected, or historical, data and are often employed to identify the correct set of EDS parameters for use in subsequent online studies at the same monitoring station. An *event* is defined as a series of time steps in which the water quality is significantly different than would be expected based on background water quality states. The number of time steps containing anomalous water quality values required in order to be called an event can vary and depends on the specific application and the goals of the event detection study.

Offline Change Point Detection

In contrast to the online approaches, a number of offline or "after the fact" approaches to analyzing time series data have also been developed. A significant amount of literature in this area exists and it is reviewed here only briefly. The majority of the offline approaches to identifying anomalous behavior are based on the detection of *change points*. Change points are defined as abrupt changes in the nature of a signal as generally indicated by statistical measures of that signal. As an example, the time at which there is a change in the source water supplying a monitoring station can be a change point for the water quality at that station. A sudden and significant shift in the average of a water quality value that is due to a change in network operations is called a *baseline change*.

Multiple approaches to change point detection are described in the literature. A change point is generally determined by fitting a statistical model to the data and identifying the time when parameterization of that model changes significantly. This is typically done offline and is often described as "retrospective segmentation" (Adams et al. 2007) in which the change points define the ends of the various segments (e.g., Tsihrintzis et al. 1995). Additionally, two separate models can be used to fit the data before and after the change point. Application of regression models to continuous data and Poisson models to discrete count data are common (See Raftery 1994 for a review). Conceptualizing the process generating the signal as a Markov process is another natural approach to change point detection. For these cases, the change points define the switch between model states and Markov models can be applied (e.g., Ge et al. 2000a). Some approaches require that the number of change points in the time series be known, or defined by the user, prior to the analysis. Other techniques are less restrictive and will determine the necessary number of change points to fit the data to some specified tolerance.

In particular, change point detection has been an active area of research with diverse applications: for example, the annual rate of coal mining accidents (e.g., West et al. 1997); highway traffic patterns (Ihler et al. 2006); and semi-conductor manufacturing process control (Ge et al. 2000b). Some recent work has been done in the area of merging change point detection approaches with those of online event detection (Takeuchi et al. 2006).

The general approach to offline change point detection is to examine data from opposite sides of a proposed change point to determine if those two data sets are significantly different from each other. If they are, the point that separates the two data sets is a change point. For offline analyses, the full data set would already be recorded and is available for analysis. In the online world, only the data recorded up to the present time are available, and the goal is to identify the change point as close to the time at which it occurs as possible. This constraint of making a determination as near to real-time as possible limits the available number of measurements that occur after the change point to as low a number as possible. The goal of an efficient water quality EDS is to develop an online approach to water quality event detection that can warn analysts and system operators in real-time of unexpected water quality conditions. To meet this goal, the offline approaches discussed above are not viable.

Online Detection

A number of online (real-time) approaches to identifying anomalous observations in time series data have been developed for use in a variety of fields. Some of the fundamental tools and the basic approaches to online event detection are covered below, with example citations provided. The literature covering online event detection is vast. This review covers only a fraction of techniques that have been published, emphasizing fundamental techniques that have been incorporated into water quality EDS tools.

Control Charts

Perhaps the oldest approaches to online event detection are the Shewhart charts and Cumulative Sum (CUSUM) charts developed in the 1920s. These approaches were originally developed for quality control in manufacturing and industrial processes and are now used in a number of other applications as well.

The CUSUM chart shows the cumulative sum of differences between the measured values and the average value. These differences are calculated by subtracting the average from each value. Increases in the cumulative sum value indicate a time of values that are continuously above the average. The resulting CUSUM chart will have an upward slope during such a period of relatively high values. The opposite results hold for periods of relatively lower values.

Shewhart charts calculate a chosen statistic of the observed data (e.g., mean) using a moving window through time. Control limits are calculated for the value of the statistic and any values of the statistic that deviate beyond the confidence bounds are identified as *outliers*. The control limits may be calculated using the expected variation in the range of data values or using more common statistical tools based on an assumption of Gaussian variation in the calculated statistic.

The choice between applying a CUSUM versus a Shewhart chart depends on the nature of the process being monitored. In general, CUSUM charts are thought to be better at detecting small, yet sustained changes in the mean value of a process (NIST/SEMATECH 2008), whereas Shewhart charts are often better suited to incorporating knowledge of the operating conditions held by the analyst.

Applications of both CUSUM and Shewhart charts typically employ standard statistical approaches to determine the range of control or the confidence limits for the process. A limitation of these tools is that they generally rely on assumptions of stationary independent and identically distributed variables and typically invoke the Gaussian distribution to define these variables. Adaptations to these charts have been made to accommodate time series data with autocorrelation, non-Gaussian distributions, non-stationarity in the data, and various ways of calculating the control limits (see Lai 1995; Zhang 1997).

Water quality time series are inherently non-stationary. Both daily (diurnal) cycles and seasonal patterns are the cause of these non-stationarities. Additionally, short term (daily to weekly) and longer term (multiple week) trends in water quality data are caused by varying levels of control of water treatment, source water changes, and hydraulic operations within the utility along with drift in the water quality sensors. It might be possible to employ techniques that have been developed for stationary time series, such as Shewhart and CUSUM charts, but first it would be necessary to remove (detrend) the non-stationary aspects of the observed data. Therefore, a robust means of modeling the background variation in water quality is a necessary step in being able to separate that background from the water quality events. This modeling of the background variation is generally referred to as "state estimation" and is the first step in modeling of non-stationary time series.

Two Step Approach to Event Detection

A common model for the online detection of changes in time series data incorporates two components that work in concert: 1) a state estimation model and 2) a residual classification algorithm. The state estimation model uses previous observations of one or more time series measurements to estimate future values of the process. These estimates could also be made by a physical process model (e.g., chemical reactions and solute transport within the pipe network). The residual classification algorithm then uses the differences between the predicted state and the observed state to determine whether or not the observed state represents an anomalous condition. This basic approach has been employed for detection of anomalous conditions in time series data in various fields including tsunami detection (Gower et al.

2006), component degradation in nuclear power plants (Yuan et al. 2005) and aging in computer software (Vaidyanathan et al. 2003).

State Estimation

Modeling of background water quality falls into the general time series modeling category of state estimation. The goal is to provide an accurate estimate of the unknown state and do this iteratively so that the state estimate is updated at every time step. The state estimate is most often quantified by parameter estimates in a statistical model of the time series. For each parameter in the model of the state, the best estimate of that parameter is provided and, depending on the complexity of the state estimation approach, a measure of uncertainty about the estimated parameter might also be determined.

A number of modern statistical and mathematical advances facilitate time series forecasting and have been applied to state estimation, including neural networks (e.g., Boznar et al. 1993), support vector machines (e.g., Müller et al. 1999), and wavelets (e.g., Lueck et al. 2000). However, this literature review focuses on traditional techniques derived in the fields of signal processing and time series analysis. These approaches have proven effective in the development and application of water quality event detection tools. Additionally, drinking water quality time series are significantly less complex than time series data obtained from natural systems (e.g., Phoon et al. 2002; Yu et al. 2004), and can be modeled under assumptions of linear processes.

Traditional approaches to time series analysis provide data driven models based on the theory of time series analysis as defined by Box and Jenkins (1976). These approaches include the popular autoregressive (AR) and moving average (MA) models as well as the various hybrids of these approaches (ARMA) and autoregressive integrated moving average (ARIMA). These models use observed data to estimate the parameters of the models and then use these estimated parameters to predict the expected data values at future observation times. These models are designed to provide a measure of uncertainty on the resulting predicted value of the time series. In essence, the time series models can be thought of as a filtering process where the noise in the underlying physical processes and in the measurements is filtered out to leave the best estimate of the water quality. Linear filters as used in signal processing can be built from AR and MA models. A thorough treatment of these approaches is given in Bras and Rodriguez-Iturbe (1993). These models are heavily used in signal processing, surface water hydrology, and econometrics applications and have also been adapted to estimate spatially correlated properties in 2 and 3 dimensions (see Goovaerts 1997; Journel et al. 1978). Traditional application of these models considers the estimated parameters as point estimates with no uncertainty, although Bayesian approaches to time series modeling can incorporate parameter uncertainty into these models.

The many variations of Kalman filters represent the next level of complexity in state estimation. Kalman filters are currently popular for data assimilation where observed data can be used to iteratively update parameters of physical process models, as well as estimate future observations of the process. Kalman filters incorporate uncertainty in the underlying model and/or its parameters along with uncertainty in the observed data in predictions of future values of the time series. Original development of the Kalman filter (Kalman 1960) was focused on state estimation for linear systems with assumed Gaussian errors, model and observation, and covariance structures. The extended and ensemble Kalman filters (EKF and EnKF, respectively) were motivated by both the need to solve more highly non-linear problems and the inadequacy of the KF for solving these problems (Evensen 1992, 1994). In particular, the EnKF replaces the analytical calculation of covariances for both the model error and observational error and the assumptions necessary for those calculations with a numerical approximation where the covariance terms are calculated across a stochastic ensemble of model states and resulting model predictions (see Evensen 2003; Moradkhani et al. 2005b).

A known disadvantage of the EnKF approach is that it has been developed for models with non-linear relationships between inputs and outputs, but it relies on a linear updating process. Additionally, the probabilistic approach to uncertainty estimation, for all variants of the Kalman filter is only valid up to second order (i.e., the output of any KF is a mean estimate and a variance defining uncertainty about that estimate, but no shape to the uncertainty distribution is provided). The second-order basis of the uncertainty estimation for the predictions essentially limits the validity of the KF approach to distributions that are at least symmetric, if not moderately Gaussian. For state estimation problems where uncertainty estimates that take into account higher-order moments of the predictive distribution are needed, particle filtering techniques (see Arulampalam et al. 2002; Gordon et al. 1993; Moradkhani et al. 2005a) provide the next level of uncertainty quantification along with additional complexity in applications.

The time series models and the Kalman filter approaches to state estimation employ some optimal weighting of previous measurements to predict the future state of the water quality. Another decidedly simpler approach to state estimation is to use just a single previous water quality measurement as the state estimate. Two approaches to using a single previous measurement as the state estimate known as time series increments and multivariate nearest neighbor are discussed further below. The time series increments approach uses the single most recent observation as the state estimate. This approach is equivalent to the Markov model, which is often referred to as the Thomas-Fiering model in surface water hydrology, where it has been applied to modeling stream flows (Bras et al. 1993). The multivariate nearest neighbor approach (Klise et al. 2006a; Klise et al. 2006b) uses the measurement within a window of recent measurements that is closest to the current observation as measured within the multivariate space defined by the observed water quality signals.

The field of signal processing provides another means of state estimation that uses cross-correlations between different signals as well as the autocorrelations within each signal to estimate the future water quality values. For example, the best estimate of the next pH value might be a weighted combination of the 10 previous pH values, the chlorine (Cl) value from 24 hours ago, and the temperature value from 12 hours ago. This approach to state estimation is a common tool in the signal processing field and has been incorporated into event detection schemes (e.g., Zavaljevskl et al. 2000).

State estimation approaches that exploit cross-correlations between signals are well-suited to situations where sensor and data transmission reliability are not an issue (e.g., engine monitoring), but when a sensor fails, the entire state estimation model fails. In environmental monitoring situations, sensor and/or data transmission failure can be common and these cross-correlation based approaches might not be best suited to these situations.

Residual Classification

Residual classification is the process of classifying each deviation between the observed and predicted water quality values (state) as either being part of the background or being a significant deviation from the background. Small residuals (deviations) can be considered as arising from incomplete parameterization of the state estimation model and measurement error. Large deviations are deemed to be significant departures from the expected background water quality and therefore are indicative of a critical change in the system (an outlier). The simplest approach to residual classification is to apply a single threshold value to the residuals, and those that exceed the threshold are considered outliers. Two issues complicate this simple approach: 1) A single constant threshold value might not apply equally well to all times in the data set, so an adaptive thresholding approach could make more sense; and 2) The thresholding approach needs to take into account the fact that state estimation and residual calculation might be done separately for each water quality signal and therefore a multivariate approach to residual classification is necessary.

The threshold used in residual classification can often be made more efficient by adapting the size of the threshold to the size, or variability, of the residuals. One approach is to make the threshold a multiplier of the standard deviation of the signal such that the threshold adapts to the variability of the signal. Normalization of the signal values to a fixed variance within a moving window allows for a threshold that is a constant multiplier of the variance, but scales relative to the un-normalized signal variance. This approach is used in the CANARY software (see Hart et al. 2007; McKenna et al. 2007). Breitgand et al. (2005) demonstrate a logistic regression based algorithm for setting adaptive thresholds in the context of computer performance monitoring.

State estimation techniques that use cross-correlations between signals generally result in a single estimate of the state that is integrated over all input signals. For these approaches a single residual is calculated at each time step. Independent state estimation for each signal results in a residual for each signal and these must be combined, or fused in some way to identify an outlier at that time step. A simple approach is to make a single classification for each time step using some combination of the residual values from all sensors operating at that time step. Equivalent results are obtained by using the average or the sum of the residuals. Classification using the maximum residual across all sensors also makes it easy to record the sensor that is responsible for the outlier at each time step. More complicated approaches to decision fusion are examined by Dasarathy (1991).

Independent state estimation followed by residual fusion allows for the number of sensors providing information at each time step to change over time. The structure of the event detection approach does not have to change to accommodate the loss or addition of sensors. This flexibility is in contrast to state estimation tools that employ cross-correlation between signals where a change in the number of sensors requires reconstruction of the model.

Takeuchi and Yamanishi (2006) integrate their deviation scores (essentially residuals) over time by calculating a moving average value. A threshold is then applied to the moving average value to detect outliers and change points. The sensitivity of this algorithm to short-lived events is controlled by the length of the moving average applied to the residuals. The Multivariate State Estimation Technique (MSET) uses a sequential probability ratio test to examine how well the distribution of residuals fits a predefined Gaussian distribution (Zavaljevskl et al. 2000). Several different hypothesis tests are examined in the MSET approach to look for mean residuals that are above or below the expected value (mean = zero) as well as variances that deviate from the expected variance value. Applications of the MSET approach to problems of computer reliability have shown it to be especially adept at detecting early stages of component degradation (Vaidyanathan et al. 2003). McKenna et al. (2007) demonstrated the binomial event discriminator for mapping outliers to events in a water quality event detection application and this approach is discussed further below.

Water Quality Event Detection

Development of event detection tools for water security has been an area of recent interest. A number of published approaches to this problem are reviewed below. These papers demonstrate the response of surrogate parameters to various contaminants and the approaches developed to detect events. The majority of these approaches work with sensor data from a single monitoring station; however, several of them have been designed to integrate sensor information from more than one station.

Byer and Carlson (2005) examined the response of surrogate monitors to the introduction of various contaminants in both laboratory beakers and in bench scale tests using water from a local utility. Their results clearly indicate the response of several surrogate parameters to the introductions of a range of contaminants at various concentrations. These results are also used by Cook et al. (2006) in testing an event detection system. More recently, Hall et al. (2007) tested the response of a number of commercially available water quality sensors in the presence of nine different contaminants introduced to a pipe test loop at different concentrations and found that at least one of the surrogate parameters responded to the presence of every contaminant.

Byer and Carlson (2005) conducted event detection through the relatively simple approach of comparing the measurement at any time to a predefined mean baseline level and defining anomalous values as those that exceed +/- 3 standard deviations from the mean of the baseline values. The baseline values were considered to be stationary and calculated by either using all of the available data, approximately 16,000 observations, or using the 100 observations immediately prior to arrival of the contaminant at the sensor. This approach represents a relatively simple example of state estimation where a large number of previous measurements are used to represent a stationary estimate of the water quality state.

Cook et al. (2006) outlined the development of a case-based reasoning system (CBRS) for the identification of multivariate data patterns that represent acceptable changes in water quality. The CBRS acts as a classifier to identify the current state of the system. Patterns that cannot be classified into existing groups are considered outliers. The work by Cook et al. (2006) highlights the need for accurate and reliable sensing of the water quality data; sophisticated software cannot make up for low quality input data.

Jarrett et al. (2006) focused their analysis of water quality data on the control exerted by the time of day and the day of the week on the expected water quality value. In the systems they examined, operation of the distribution network was responsible for a significant portion of the water quality variation, and those operations followed a reasonably predictable behavior. However, the temporal patterns controlling the water quality tended to change over time and therefore it was difficult to accurately predict water quality based solely on the time of day. Jarrett et al. (2006) proposed that a control chart approach applied to the first differences (increments) of the water quality data might prove useful in event detection if the center line and widths of the control region were both allowed to vary temporally (a non-stationary control chart approach).

Kroll and King (2006) provided a rough outline of a proprietary EDS that includes both a baseline (state) estimation component and a multivariate classification component. The classification step used the deviations in the measured signals from the baseline along with a library of previously recorded deviations to classify the cause of the event as being either a particular contaminant or a change in water quality caused by a change in operations at the utility. Patterns that did not match any of the library patterns were declared “unknown” and can be added to the library by the operator. The ability of the algorithm to “learn” through time was used to lower the number of false positives upon deployment.

Klise and McKenna (2006b) examined the utility of multivariate classification schemes for event detection. The state estimation approach in this work was to define every time step of the baseline water quality as belonging to one of a finite number of clusters within the multivariate space defined by the vector of surrogate parameter measurements, or by a lower dimensional representation of that space as defined by principal components. Results showed that increasing the number of clusters, which defined the baseline to the point where every recent time step was considered to be a separate cluster, improved results over smaller numbers of clusters. The result of this work was the development of the multivariate nearest neighbor (MVNN) algorithm in which the distance from any new measured water quality vector to the nearest previously measured vector in the multidimensional space is recorded. If that distance exceeds a specified threshold, the new data point is considered to be an outlier. Klise and McKenna (2006a) further tested the MVNN algorithm using event data from EPA’s Test and Evaluation (T&E) Facility (Hall et al. 2007) that were superimposed on water quality data collected at a U.S. utility and found that event detection results with MVNN are sensitive to the contaminant type and the background water quality variability at each monitoring station.

McKenna et al. (2006b) compared three approaches to state estimation for each water quality time series: time series increments (the previous measured value is the predictor of the next value), linear filters, and the MVNN approach. For the increment and linear filter approaches, the residuals between the predicted and measured water quality values were fused across all water quality signals, and this final fused residual was compared to a threshold to define whether or not it was an event. The multivariate distances in the MVNN algorithm were compared directly to the same threshold as they already represent a measure of prediction accuracy that takes into account all water quality signals. Testing was completed on simulated time series and actual measured water quality time series data. Simulated events were added to all data sets. Results showed that the MVNN algorithm was able to best predict the water quality background in all cases, but that the ability to predict the background did not necessarily translate into the best detection capabilities as measured by the false positive and false negative rates.

Prior work in event detection algorithms by Klise and McKenna (2006a; 2006b) and McKenna et al. (2006b) evaluated every time step against a threshold. Those time steps with residuals from the baseline that exceeded the threshold were classified as events (a single outlier equals an event). This approach led to a large number of false positives as water quality within utility systems can be quite noisy and additional noise is added to the water quality data as it is transmitted through the SCADA system. McKenna et al. (2007) introduced the binomial event discriminator (BED) as a means of aggregating results over multiple time steps to determine whether or not an event was occurring. Each individual time step is now considered to be part of the background or an outlier and the number of outliers (failures) within a given number of time steps (trials) as inputs to the binomial distribution defines the probability that a water quality event is occurring. Addition of the BED to the water quality prediction algorithms allowed for order of magnitude reductions in FAs for the data sets tested.

The papers discussed above relied on various statistical models to estimate the state by tracking the baseline water quality conditions so that a comparison between the expected baseline value and the observed values can be made. Another approach to determination of the baseline conditions would be to employ a model that directly simulates all water quality parameter values through the physical and chemical processes occurring in the distribution network between the treatment plant and the monitoring station (Shang et al. 2008). This model would have to be continuously updated with real-time information and could provide the predicted water quality values for each future time step in the same manner that the linear filter and multivariate nearest neighbor algorithms are currently used in CANARY.

Monitoring at Multiple Stations

The majority of the research in this field is focused on analysis of data from each monitoring station independently. However, several publications have shown additional benefit that can be gained from combining water quality data from more than one monitoring station. O'Halloran et al. (2006) developed a water parcel tracking approach that matched the "fingerprint" of water quality recorded at two different monitoring stations along the same flowpath in the network. They were able to use this automated technique to determine the transit time of the water between the two monitoring stations, although an assumption of steady flow was required. Yang et al. (2007) defined a technique for improving event detection and reducing false positives by combining water quality monitoring data from two monitoring stations along the same pipe. Data from two stations in series allows for transport modeling to be applied to the water quality between the two stations, where data from the first station essentially provides the initial conditions for the transport solution. This transport modeling provides improved state estimation at the downstream monitoring station. A more general approach to integrating data from two or more monitoring stations, which might not be in direct hydraulic connection, has recently been tested, with promising results (Koch et al. 2008).

Evaluating Event Detection Algorithms

An issue of considerable importance in water quality monitoring is the appropriate evaluation of an event detection algorithm. As a rule, utilities do an extremely good job of supplying high quality water to their customers without fail, and water quality events — even those due to routine causes such as main breaks, faults in a primary treatment system, or failures of a chlorine booster station — are rare. Documented accounts of malevolent contamination of a utility are even rarer, and this routine delivery of high quality water makes it nearly impossible to completely test event detection systems in real-world situations. Testing of EDS with experimental data obtained in laboratory settings (e.g., Byer et al. 2005; Kroll et al. 2006) or in specially designed pipe loops (e.g., Hall et al. 2007; Yang et al. 2007) provides direct measurement of sensor responses to controlled contamination events, but typically the variation in background water quality for these tests is considerably less than that experienced within operating distribution networks. Therefore, the most direct means of evaluating event detection systems is to simulate the response of water quality monitoring sensors to the introduction of a contaminant. Simulation based evaluation of EDS tools has been done by a number of authors using varying levels of sophistication in the contaminant simulation approach (e.g., Allgeier et al. 2008; McKenna et al. 2006a; McKenna et al. 2008; Shang et al. 2008; Uber et al. 2007; Umberg et al. 2008).

Reports of early work in event detection from water quality data focused on the number of known events that were detected. However, as pointed out by Rizak and Hrudey (2006), when monitoring for events that are expected to occur with a very low probability, dealing with false positive events will consume the largest amount of the monitoring organization's resources. McKenna et al. (2008) employed the receiver operating characteristic (ROC) curve approach from the signal processing and medical diagnostics fields to evaluate water quality event detection systems. ROC curves demonstrate the tradeoff between the rate of false events and the probability of detecting true events on a single graph. Typical ROC curve shapes quantify the increase in false positive events as the sensitivity of the algorithm is increased to improve the probability of detecting true events.

Glossary

Baseline Change
A baseline change is a significant change in a statistical parameter, generally the mean, of the observed data. The point at which a baseline change occurs is a change point. The change in behavior of the observed signal from one side of the change point to the other is referred to as the baseline change. Baseline changes are common in some water distribution systems, due to changes in mixing of source waters within the network at different time of the day.

Change Point
Change points are defined as the point in time or space where an abrupt change in behavior or mode of operation is observed. Change points are generally identified by comparing data collected on both sides of the proposed change point. If that comparison shows the data on either side of the proposed change point to be significantly different, then that proposed change point is confirmed as a change point. Change points are most accurately identified using offline techniques for data analysis where adequate amounts of data have already been collected on both sides of any proposed change point.

Contaminant Warning System (CWS)
The CWS is composed of all the hardware and software components that are necessary for monitoring the water distribution system for contamination events. These components include the sensors, the Supervisory Control and Data Acquisition (SCADA) system to collect and transmit the data from the sensors to a central facility, the database to hold sensor information, and the event detection system (EDS) that processes the data to provide some indication of the occurrence of an event. Other components of a CWS consist of monitoring non-water quality data streams (from public health and physical security) and the decision support approaches and responses applied to these monitoring systems.

Event
An activity or behavior that is unusual relative to normal modes of operation. Abnormal activity or operation relative to the background or ambient modes of operation can be classified as an event. An event is a sustained period of such abnormal activity that is of a longer duration than an outlier, but of shorter duration than a baseline change.

Event Detection System (EDS)
The software system that contains the data handling, algorithms, and input/output functions necessary to identify events from water quality time series. Typically, the inputs to an EDS are the water quality data streams from the sensors as stored in a SCADA database. The output of an EDS is an indication of the state of the water quality. This indication can be a binary signal such as “alarm/no-alarm” or it can be a continuous indication of the water quality state such as the probability of an event occurring at every time step. The EDS itself is a component of the more comprehensive CWS.

Offline
Analysis or operations that are completed without connection to any real-time source of data or any means of implementing real-time control are considered offline. Analysis of previously collected “historical” data is done in an offline manner.

Online
Analysis or operations that are completed with connection to a real-time source of data or that have some means of implementing real-time control are considered online. Online analysis typically implies that data are received in a periodic manner and the analysis is completed prior to the receipt of additional data. Analysis of real-time data as provided through connection to a SCADA system is completed in an online manner.

Outlier
An outlier is defined here to be a single time step with behavior that is considered anomalous relative to the background or expected behavior for that time step. A large enough number of outliers within a prescribed time interval could constitute an event.

Residual
The difference between the predicted and observed water quality values at a single time step. The size of the residual is classified as being indicative of background water quality conditions or of anomalous conditions representing water quality events.

Supervisory Control and Data Acquisition (SCADA) system
The hardware and software components that transmit water quality and operations data from in-situ sensors throughout the network to a central facililty. The SCADA system also includes the database hardware and software to store collected data.

Appendix B.
Quality Assurance

EPA's quality systems cover the collection, evaluation, and use of environmental data by and for the Agency, and the design, construction, and operation of environmental technology by the Agency. The purpose of EPA's quality systems is to support scientific data integrity, reduce or justify resource expenditures, properly evaluate of internal and external activities, support reliable and defensible decisions by the Agency, and reduce burden on partnering organizations.

All research presented in this report that was performed by the authors was completed under approved EPA and DOE quality practices adapted from the *Advanced Simulation and Computing (ASC) Software Quality Plan* and *EPA Guidance for Quality Assurance Project Plans*. The *ASC Software Quality Plan* was generated to conform with the SNL corporate and DOE QC-1 revision 9 standards.

The quality assurance (QA) practices followed under this research included:

- Project Management
- Computational Modeling and Algorithm Development
- Software Engineering
- Data Generation and Acquisition
- Model and Software Verification
- Training

Project management is the systematic approach for balancing the project work to be done, resources required, methods used, procedures to be followed, schedules to be met, and the way that the project is organized. The project management QA practices included: performing a risk-based assessment to determine level of formality and applicable practices; identifying stakeholders and other requirements sources; gathering and managing stakeholders' expectations and requirements; deriving, negotiating, managing, and tracking requirements; identifying and analyzing project risk events; defining, monitoring, and implementing the risk response; creating and managing the project plan; and tracking project performance versus project plan and implementing needed corrective actions.

Modeling and algorithm development are often closely related activities; modeling is the process of mathematically formulating a problem, while algorithm development is the process of finding a method to solve the problem computationally. These activities can be distinguished from software engineering efforts, which are more specifically focused on ensuring that software generated has high quality itself. The modeling and algorithm development QA practices included: documenting designs for models and algorithms; conducting peer reviews of modeling assumptions and algorithmic formulations; documenting preliminary software implementation; documenting sources of uncertainty in modeling and algorithmic methods; and completing peer-review of modeling and algorithmic outputs.

Software engineering is a systematic approach to the specification, design, development, test, operation, support, and retirement of software. The modeling and algorithm development QA practices included: communicating and reviewing software design; creating required software and product documentation; identifying and tracking third party software products and follow applicable agreements; identifying, accepting ownership, and managing assimilation of other software products; performing version control of identified software product artifacts; recording and tracking issues associated with the software product; ensuring backup and disaster recovery of software product artifacts; planning and generating the release package; and certifying that the software product (code and its related artifacts) was ready for release and distribution.

Input data for model development and application efforts are typically collected outside of the modeling effort or generated by other models or processing software. These data need to be properly assessed to verify that a model characterized by these data would yield predictions with an acceptable level of uncertainty. The data generation and acquisition QA practices included: documenting objectives and methods of model calibration activities; documenting sources of input data used for calibration; identifying requirements for non-direct data and data acquisition; developing processes for managing data; and documenting hardware and software used to process data.

The purpose of software verification is to ensure (1) that specifications are adequate with respect to intended use and (2) that specifications are accurately, correctly, and completely implemented. Software verification also attempts to ensure product characteristics necessary for safe and proper use are addressed. Software verification occurs throughout the entire product lifecycle. The software verification QA practices included: developing and maintaining a software verification plan; conducting tests to demonstrate that acceptance criteria are met and to ensure that previously tested capabilities continue to perform as expected; and conducting independent technical reviews to evaluate adequacy with respect to requirements.

The goal of training practices is to enhance the skills and motivation of a staff that is already highly trained and educated in the areas of mathematical modeling, scientific software development, algorithms, and/or computer science. The purpose of training is to develop the skills and knowledge of individuals and teams so they can fulfill their process and technical roles and responsibilities. The training QA practices included: determining project team training needed to fulfill assigned roles and responsibilities; and tracking training undertaken by project team.

References

Adams, R. P., and MacKay, D. J. C. (2007). *Bayesian online change point detection,* <http://arxiv.org/abs/0710.3742v1>.

Allgeier, S. C., and Umberg, K. (2008). "Systematic evaluation of contaminant detection through water quality monitoring." *Proc., AWWA Water Security Congress,* AWWA, Denver, CO.

Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (2002). "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking." *IEEE Transactions on Signal Processing,* 50(2), 174–188.

ASCE. (2004). *Interim voluntary guidelines for designing an online contaminant monitoring system*, American Society of Civil Engineers, Reston, VA.

AWWA. (2005). *Contamination warning systems for water: an approach for providing actionable information to decision-makers,* American Water Works Association, Denver, CO.

Berry, J. W., Boman, E., Riesen, L. A., Hart, W. E., Phillips, C. A., and Watson, J.-P. (2008). *User's manual: TEVA-SPOT toolkit 2.2,* EPA/600/R-08/041, U.S. Environmental Protection Agency, Office of Research and Development, National Homeland Security Research Center, Cincinnati, OH.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms,* Kluwer Academic Publishers, Norwell, MA.

Box, G. E. P., and Jenkins, G. M. (1976). *Time series analysis: forecasting and control,* Holden-Day series in time series analysis, Holden-Day, San Francisco, CA.

Boznar, M., Lesjak, M., and Mlakar, P. (1993). "A neural-network-based method for short-term predictions of ambient SO_2 concentrations in highly polluted industrial-areas of complex terrain." *Atmospheric Environment Part B-Urban Atmosphere,* 27(2), 221-230.

Bras, R. L., and Rodriguez-Iturbe, I. (1993). *Random functions and hydrology,* Dover Publications, Mineola, NY.

Breitgand, D., Henis, E., and Shehory, O. (2005). "Automated and adaptive threshold setting: enabling technology for autonomy and self-management." *Proc., ICAC 2005: Second International Conference on Autonomic Computing,* 204–215.

Byer, D., and Carlson, K. H. (2005). "Real-time detection of intentional chemical contamination in the distribution system." *Journal American Water Works Association,* 97(7), 130–133.

Camargo, S. J., Robertson, A. W., Gaffney, S. J., Smyth, P., and Ghil, M. (2007). "Cluster analysis of typhoon tracks. Part I: General properties." *Journal of Climate,* 20(14), 3635–3653.

Cook, J., Roehl, E., Daamen, R., Carlson, K., and Byer, D. (2005). "Decision support system for water distribution system monitoring for homeland security." *Proc., AWWA Water Security Congress,* AWWA, Denver, CO.

Cook, J. B., Byrne, J. F., Daamen, R. C., and Roehl Jr., E. A. (2006). "Distribution system monitoring research at Charleston Water System." *Proc., 8th Annual Water Distribution Systems Analysis Symposium,* ASCE, Reston, VA.

Dasarathy, B. V. (1991). "Decision fusion strategies in multi-sensor environments." *IEEE Transactions on Systems Man and Cybernetics,* 21(5), 1140–1154.

Dunn, J. C. (1973). "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters." *Cybernetics and Systems: An International Journal,* 3(3), 32–57.

Einfeld, W., McKenna, S. A., and Wilson, M. P. (2008). *A simulation tool to assess contaminant warning system sensor performance characteristics,* AwwaRF Report 91219, American Water Works Association Research Foundation (AwwaRF), Denver, CO.

Evensen, G. (1992). “Using the extended Kalman filter with a multilayer quasi-geostrophic ocean model.” *Journal of Geophysical Research-Oceans,* 97(C11), 17905–17924.

Evensen, G. (1994). “Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics.” *Journal of Geophysical Research-Ocean*s, 99(C5), 10143–10162.

Evensen, G. (2003). “The ensemble Kalman filter: theoretical formulation and practical implementation.” *Ocean Dynamics,* 53(4), 343–367.

Gaffney, S. J. (2004). “Probabilistic curve-aligned clustering and prediction with regression mixture models.” Dissertation, University of California, Irvine, Irvine, CA.

Gaffney, S. J., Robertson, A. W., Smyth, P., Camargo, S. J., and Ghil, M. (2007). “Probabilistic clustering of extratropical cyclones using regression mixture models.” *Climate Dynamics,* 29(4), 423–440.

Ge, X., and Smyth, P. (2000a). *Deformable Markov model templates for time series pattern matching,* Technical Report UCI-ICS 00-10, University of California, Irvine, Irvine, CA. <http://www.datalab.uci.edu/papers/trseghmm.pdf>.

Ge, X., and Smyth, P. (2000b). *Segmental semi-Markov models for change point detection with applications to semiconductor manufacturing,* Technical Report UCI-ICS 00-08, University of California, Irvine, Irvine, CA. <http://www.datalab.uci.edu/papers/trchange.pdf>.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation,* Applied Geostatistics Series, Oxford University Press, Inc., New York, NY.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). “Novel approach to non-linear and non-Gaussian Bayesian state estimation.” *Proceedings of the Institute of Electrical Engineers (IEE), Part F, 140,* 107–113.

Gower, J., and González, F. (2006). “U.S. warning system detected the Sumatra Tsunami.” *EOS, Transactions of the American Geophysical Union,* 87(10), 105–108.

Hall, J., Zaffiro, A. D., Marx, R. B., Kefauver, P. C., Krishnan, E. R., and Herrmann, J. G. (2007). “Online water quality parameters as indicators of distribution system contamination.” *Journal American Water Works Association,* 99(1), 66–77.

Hall, J. S., Szabo, J. G., Panguluri, S., and Meiners, G. (2009). *Distribution system water quality monitoring: sensor technology evaluation methodology and results, a guide for sensor manufacturers and water utilities,* EPA/600/R-09/076, U. S. Environmental Protection Agency, Office of Research and Development, National Homeland Security Research Center, Cincinnati, OH.

Hart, D., McKenna, S. A., Klise, K., Cruz, V., and Wilson, M. (2007). “CANARY: a water quality event detection algorithm development tool.” *Proc., World Environmental and Water Resources Congress,* ASCE, Reston, VA.

Hart, D. B., and McKenna, S. A. (2009). *CANARY user’s manual, version 4.1,* EPA/600/R-08/040A, U.S. Environmental Protection Agency, Office of Research and Development, National Homeland Security Research Center, Cincinnati, OH.

Hartigan, J. A., and Wong, M. A. (1978). “Algorithm AS 136: a k-means clustering algorithm.” *Applied Statistics,* 28, 100–108.

Ihler, A., Hutchins, J., and Smyth, P. (2006). “Adaptive event detection with time-varying Poisson processes.” *Proc., 12th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-06),* ACM (Association for Computing Machinery), New York, NY, 207–216.

Jarrett, R., Robinson, G., and O’Halloran, R. (2006). “Online monitoring of water distribution systems: data processing and anomaly detection.” *Proc., 8th Annual Water Distribution Systems Analysis Symposium,* ASCE, Reston, VA.

Journel, A. G., and Huijbregts, C. J. (1978). *Mining geostatistics,* Academic Press, San Diego, CA.

Kalman, R. E. (1960). “A new approach to linear filtering and prediction problems.” *Transactions of the ASME-Journal of Basic Engineering,* 82(Series D), 35–45.

Klise, K. A., and McKenna, S. A. (2006a). “Multivariate applications for detecting anomalous water quality.” *Proc., 8th Annual Water Distribution Systems Analysis Symposium,* ASCE, Reston, VA.

Klise, K. A., and McKenna, S. A. (2006b). “Water quality change detection: multivariate algorithms.” *Proc., SPIE (International Society for Optical Engineering), Defense and Security Symposium 2006.*

Koch, M. W., and McKenna, S. A. (2008). “Distributed network fusion for water quality.” *Proc., World Environmental & Water Resources Congress,* ASCE, Reston, VA.

Kroll, D., and King, K. (2006). “Laboratory and flow loop validation and testing of the operational effectiveness of an online security platform for the water distribution system.” *Proc., 8th Annual Water Distribution Systems Analysis Symposium,* ASCE, Reston, VA.

Kulldorff, M. (1997). “A spatial scan statistic.” *Communications in Statistics-Theory and Methods,* 26(6), 1481–1496.

Kulldorff, M., and Nagarwalla, N. (1995). “Spatial disease clusters-detection and inference.” *Statistics in Medicine,* 14(8), 799–810.

Lai, T. L. (1995). “Sequential change point detection in quality-control and dynamical-systems.” *Journal of the Royal Statistical Society Series B-Methodological,* 57(4), 613–658.

Lueck, R. G., Driscoll, F. R., and Nahon, M. (2000). "A wavelet for predicting the time-domain response of vertically tethered systems.” *Ocean Engineering,* 27(12), 1441–1453.

MathWorks. (2008). MATLAB, 2008b, The Mathworks, Natick, MA. <http://www.mathworks.com/products/matlab/>.

McKenna, S. A., Hart, D., Klise, K., Cruz, V., and Wilson, M. (2007). “Event detection from water quality time series.” *Proc., World Environmental and Water Resources Congress,* ASCE, Reston, VA.

McKenna, S. A., Hart, D. B., and Yarrington, L. (2006a). “Impact of sensor detection limits on protecting water distribution systems from contamination events.” *Journal of Water Resources Planning and Management,* 132(4), 305–309.

McKenna, S. A., Klise, K. A., and Wilson, M. P. (2006b). “Testing water quality change detection algorithms.” P*roc., 8th Annual Water Distribution Systems Analysis Symposium,* ASCE, Reston, VA.

McKenna, S. A., Wilson, M., and Klise, K. A. (2008). “Detecting changes in water quality data.” *Journal American Water Works Association,* 100(1), 74–85.

Moradkhani, H., Hsu, K. L., Gupta, H., and Sorooshian, S. (2005a). “Uncertainty assessment of hydrologic model states and parameters: sequential data assimilation using the particle filter.” *Water Resources Research,* 41(5), W05012.

Moradkhani, H., Sorooshian, S., Gupta, H. V., and Houser, P. R. (2005b). “Dual state-parameter estimation of hydrological models using ensemble Kalman filter.” *Advances in Water Resources,* 28(2), 135–147.

Müller, K.-R., Smola, A. J., Rätsch, G., Schökpf, B., Kohlmorgen, J., and Vapnik, V. (1999). “Using support vector machines for time series prediction.” *Advances in kernel methods: support vector learning,* MIT Press, Cambridge, MA, 243–253.

Murray, R., Haxton, T., Janke, R., Hart, W. E., Berry, J. W., and Phillips, C. A. (2010). *Sensor network design for drinking water contamination warning systems: a compendium of research results and case studies using the TEVA-SPOT software,* U.S. Environmental Protection Agency, Office of Research and Development, National Homeland Security Research Center, Cincinnati, OH.

NIST/SEMATECH. (2008). *Univariate and multivariate control charts, e-handbook of statistical methods: engineering statistics handbook,* National Institute of Standards and Testing (NIST)/SEMATECH, <http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc3.htm>.

O'Halloran, R., Yang, S., Tulloh, A., Koltun, P., and Toifl, M. (2006). "Sensor-based water parcel tracking." *Proc., 8th Annual Water Distribution Systems Analysis Symposium,* ASCE, Reston, VA.

Pakhira, M. K., Bandyopadhyay, S., and Maulik, U. (2004). "Validity index for crisp and fuzzy clusters." *Pattern Recognition,* 37(3), 487–501.

Phoon, K. K., Islam, M. N., Liaw, C. Y., and Liong, S. Y. (2002). "Practical inverse approach for forecasting nonlinear hydrological time series." *Journal of Hydrologic Engineering,* 7(2), 116–128.

Public Health Security and Bioterrorism Preparedness and Response Act of 2002. PL 107-188. <http://www.fda.gov/EmergencyPreparedness/Counterterrorism/BioterrorismAct/default.htm>.

Raftery, A. E. (1994). "Change point and change curve modeling in stochastic processes and spatial statistics." *Journal of Applied Statistical Science,* 1(4), 403–424.

Rizak, S. N., and Hrudey, S. E. (2006). "Misinterpretation of drinking water quality monitoring data with implications for risk management." *Environmental Science & Technology,* 40(17), 5244–5250.

Rossman, L. A. (1999). "The EPANET programmer's toolkit for analysis of water distribution systems." *Proc., 26th Annual Water Resources Planning and Management Conference,* ASCE, Reston, VA.

Rossman, L. A. (2000). *EPANET 2: users manual,* EPA/600/R-00/057, U.S. Environmental Protection Agency, Office of Research and Development, National Risk Management Research Laboratory, Cincinnati, OH. <http://www.epa.gov/nrmrl/wswrd/dw/epanet/EN2manual.pdf>.

Shang, F., Uber, J., Murray, R., and Janke, R. (2008). "Model-based real-time detection of contamination events." *Proc., Water Distribution Systems Analysis 2008,* ASCE, Reston, VA.

Takeuchi, J., and Yamanishi, K. (2006). "A unifying framework for detecting outliers and change points from time series." *IEEE Transactions on Knowledge and Data Engineering,* 18(4), 482–492.

Taylor, H. M., and Karlin, S. (1998). *An introduction to stochastic modeling,* Third Edition, Academic Press, San Diego, CA.

Tsihrintzis, G. A., and Nikias, C. L. (1995). "Robust change point detection and segmentation in data streams." *Proc., Milcom 95 - Conference Record, Vols 1-3,* 125–129.

U.S. DHS and U.S. EPA. (2007). *Water: critical infrastructure and key resources sector-specific plan as input to the national infrastructure protection plan,* U.S. Department of Homeland Security and U.S. Environmental Protection Agency, Washington, D.C. <http://www.dhs.gov/xlibrary/assets/Water_SSP_5_21_07.pdf >.

U.S. EPA. (2005a). *Technologies and techniques for early warning systems to monitor and evaluate drinking water quality: a state-of-the-art review,* EPA/600/R-05/156, U.S. Environmental Protection Agency, Office of Research and Development, National Homeland Security Research Center, Cincinnati, OH.

U.S. EPA. (2005b). *WaterSentinel system architecture,* EPA/817/D-05/003, U.S. Environmental Protection Agency, Office of Water, Office of Ground Water and Drinking Water, Washington, D.C.

U.S. EPA. (2006). *Active and effective water security programs: a summary report of the National Drinking Water Advisory Council recommendations on water quality,* EPA/817/K-06/001, U.S. Environmental Protection Agency, Office of Water, Washington, D.C.

U.S. EPA. (2009a). *EDDIES-RT 4.0 user's guide,* U.S. Environmental Protection Agency, Office of Water, Office of Ground Water and Drinking Water, Washington, D.C.

U.S. EPA. (2009b). *Tutorial threat ensemble vulnerability analysis – sensor placement optimization tool (TEVA-SPOT) graphical user interface, Version 2.2.0 Beta,* EPA/600/R-08/147, U.S. Environmental Protection Agency, Office of Research and Development, National Homeland Security Research Center, Cincinnati, OH.

Uber, J. G., Murray, R., Magnuson, M., and Umberg, K. (2007). “Evaluating real-time event detection algorithms using synthetic data.” *Proc., World Environmental and Water Resources Congress,* ASCE, Reston, VA.

Umberg, K., and Allgeier, S. (2008). “Evaluation of water quality event detection systems deployed at the first water security initiative pilot utility.” *Proc., AWWA Water Security Congress,* AWWA, Denver, CO.

Vaidyanathan, K., and Gross, K. (2003). “MSET performance optimization for detection of software aging.” *Proc., 14th IEEE International Symposium on Software Reliability Engineering (ISSRE’03), Fast Abstract ISSRE.*

Walpole, R. E., and Myers, R. H. (1989). *Probability and statistics for engineers and scientists,* Fourth Edition, MacMillan Publishing Company, New York, NY.

Watson, J.-P., Greenberg, H. J., and Hart, W. E. (2004). “A multiple-objective analysis of sensor placement optimization in water networks.” *Proc., The 2004 World Water and Environmental Resources Congress,* ASCE, Reston, VA.

West, R. W., and Ogden, R. T. (1997). “Continuous-time estimation of a change point in a Poisson process.” *Journal of Statistical Computation and Simulation,* 56(4), 293–302.

Xu, R., and Wunsch, D. C. (2009). *Clustering*, Institute of Electrical and Electronics Engineers Press, Hoboken, NJ.

Yang, Y. J., Goodrich, J. A., Clark, R. M., and Li, S. Y. (2008). “Modeling and testing of reactive contaminant transport in drinking water pipes: chlorine response and implications for online contaminant detection.” *Water Research,* 42(6–7), 1397–1412.

Yang, Y. J., Haught, R. C., and Goodrich, J. A. (2009). “Real-time contaminant detection and classification in a drinking water pipe using conventional water quality sensors: techniques and experimental results.” J*ournal of Environmental Management,* 90(8), 2494–2506.

Yang, Y. J., Haught, R. C., Hall, J., Szabo, J., Clark, R. M., and Meiners, G. (2007). “Adaptive water sensor signal processing: experimental results and implications for online contaminant warning systems.” *Proc., World Environmental and Water Resources Congress,* ASCE, Reston, VA.

Yu, X. Y., Liong, S. Y., and Babovic, V. (2004). “EC-SVM approach for real-time hydrologic forecasting.” *Journal of Hydroinformatics,* 6(3), 209–223.

Yuan, C., Neubauer, C., Cataltepe, Z., and Brummel, H. G. (2005). “Support vector methods and use of hidden variables for power plant monitoring.” *Proc., 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vols 1-5,* 693–696.

Zavaljevskl, N., and Gross, K. (2000). “Sensor fault detection in nuclear power plants using multivariate state estimation technique and support vector machines.” *Proc., Third International Conference of the Yugoslav Nuclear Society (YUNSC 2000).*

Zhang, N. F. (1997). “Detection capability of residual control chart for stationary process data.” *Journal of Applied Statistics,* 24(4), 475–492.

EPA
United States
Environmental Protection
Agency

Office of Research and Development
National Homeland Security Research Center
Cincinnati, OH 45268
Official Business
Penalty for Private Use
$300