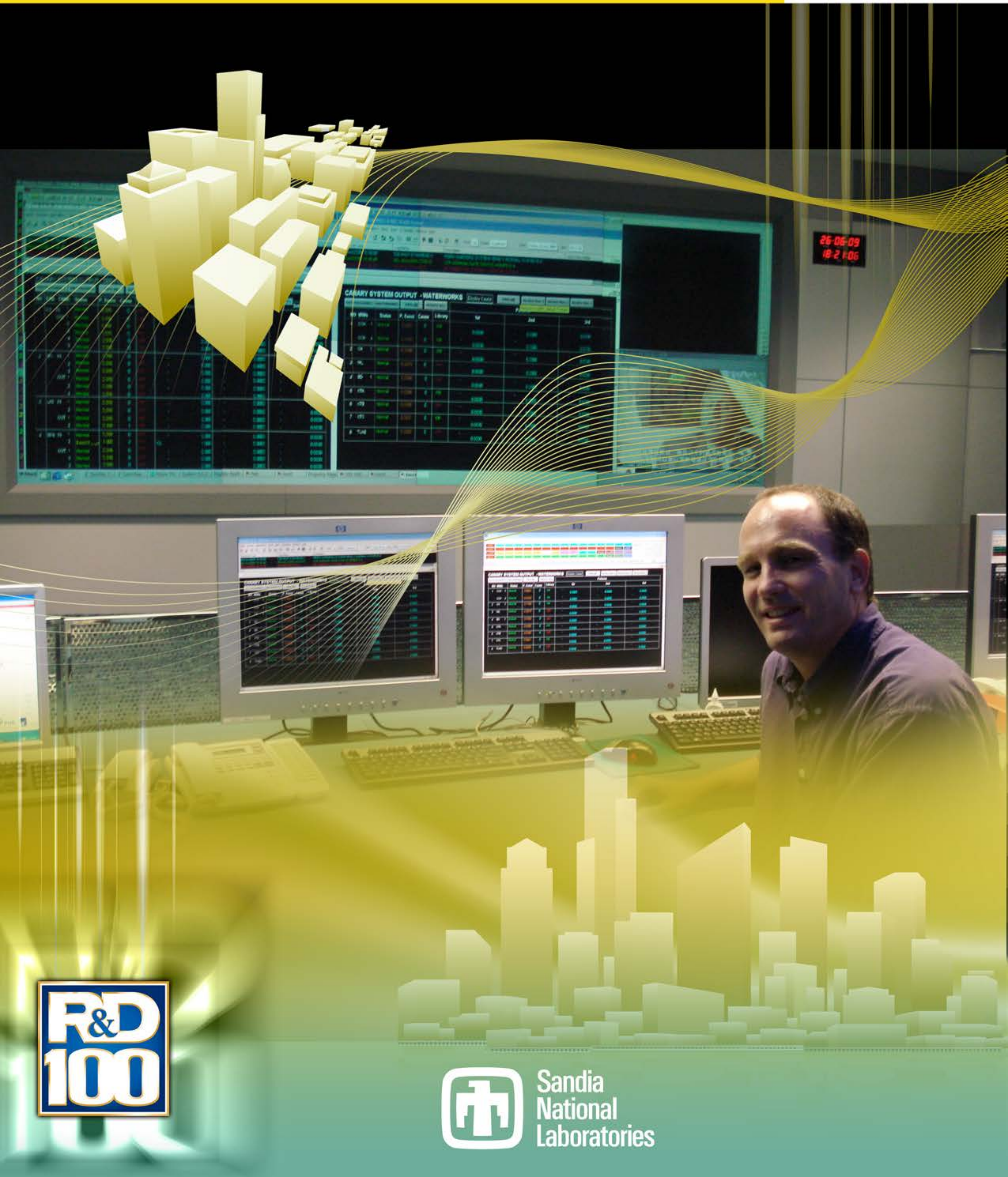
CANARY: Event Detection Software

2010



R&D 100

Sandia
National
Laboratories

Submitting Organization

Sandia National Laboratories
PO Box 5800, MS 0185
Albuquerque, NM 87185-0185
Sean A. McKenna, PhD
Distinguished Member of the Technical Staff
P.O. Box 5800
MS 0751
Albuquerque, NM 87185-0185
Phone (505) 844-2450
Fax (505) 844-7354
samcken@sandia.gov

Contact Person

Glenn D. Kubiak
Director, Biological and Materials Sciences Center
Sandia National Laboratories
PO Box 969
MS 9405
Livermore, CA 94551-0969
USA
Phone (925) 294-3375
Fax (925) 294-3403
kubiak@sandia.gov

Joint Entry with

U. S. Environmental Protection Agency
26 W Martin Luther King Dr. (NG 16)
Cincinnati, Ohio 45268
USA
Dr. Regan Murray
Phone (513) 569-7031
Fax (513) 487-2559
Murray.Regan@epa.gov

Product Name

CANARY: Event Detection Software

Brief Description

CANARY provides continuous monitoring of water quality from networked sensors for automated event detection, which enables improved security and operations within water distribution systems worldwide.

Product First Marketed or Available for Order

CANARY was first made publicly available in May of 2009.

Inventors or Principal Developers

Sean A. McKenna, PhD
Distinguished Member of the Technical Staff
Sandia National Laboratories, National Security Applications
PO Box 5800, MS 0751
Albuquerque, NM 87185-0751
USA
Phone (505) 844-2450
Fax (505) 844-0735
samcken@sandia.gov

David Hart
Member of the Technical Staff
Sandia National Laboratories, National Security Applications
PO Box 5800, MS 0751
Albuquerque, NM 87185-0751
USA
Phone (505) 844-4674
Fax (505) 844-7354
dbhart@sandia.gov

Katherine Klise
Senior Member of the Technical Staff
Sandia National Laboratories, National Security Applications
PO Box 5800, MS 0751
Albuquerque, NM 87185-0751
USA
Phone (505) 284-4456
Fax (505) 844-7354
kaklise@sandia.gov

Eric Vugrin, PhD
Principal Member of the Technical Staff
Sandia National Laboratories, Infrastructure and Economic Systems
PO Box 5800, MS 1138
Albuquerque, NM 87185-1138
USA
Phone (505) 284-8494
Fax (505) 284-3850
edvugri@sandia.gov

Mark W. Koch, PhD
Principal Member of the Technical Staff
Sandia National Laboratories, Sensor Exploitation Applications Dept.
PO Box 5800, MS 1163
Albuquerque, NM 87185-1163
USA
Phone (505) 844-4731
Fax (505) 844-4157
mwkoch@sandia.gov

Shawn Martin, PhD
Senior Member Technical Staff
Sandia National Laboratories, Computer Science and Informatics
PO Box 5800, MS 1316
Albuquerque, NM 87185-1316
USA
Phone (505) 284-3601
Fax (505) 845-7442
smartin@sandia.gov

William E. Hart, PhD
Distinguished Member of the Technical Staff
Sandia National Laboratories, Computer Software Rsrch/Dvlpmnt
PO Box 5800, MS 1318
Albuquerque, NM 87185-1318
USA
Phone (505) 844-2217
Fax (505) 845-7442
wehart@sandia.gov

Regan Murray, PhD
Water Infrastructure Protection, Mathematical Statistician
U.S. Environmental Protection Agency
26 W. Martin Luther King Dr. (NG 16)
Cincinnati, OH 45268
USA
Phone (513) 569-7031
Fax (513) 487-2559
Murray.Regan@epa.gov

Terranna Haxton, PhD
Water Infrastructure Protection, Environmental Engineer
U.S. Environmental Protection Agency
26 W. Martin Luther King Dr. (NG 16)
Cincinnati, OH 45268
USA
Phone (513) 569-7810
Fax (513) 487-2559
Haxton.Terra@epa.gov

John Hall
Water Infrastructure Protection, Physical Scientist
U.S. Environmental Protection Agency
26 W. Martin Luther King Dr. (NG 16)
Cincinnati, OH 45268
USA
Phone (513) 487-2814
Fax (513) 487-2559
Hall.John@epa.gov

Katie Umberg
Water Infrastructure, Physical Scientist
U.S. Environmental Protection Agency
26 W. Martin Luther King Dr. (MS 140)
Cincinnati, OH 45268
USA
Phone (513) 569-7925
Fax (513) 569-7191
Umberg.Katie@epa.gov

Product Price

CANARY is distributed as open-source software (https://software.sandia.gov/trac/canary). The public domain licensing allows for third-party development of proprietary commercial software systems that incorporate CANARY. Executable versions of CANARY can be obtained from the US EPA (see: http://www.epa.gov/nhsrc/water/teva.html)

Patents or Patents Pending

No.

Product's Primary Function

Rapid and accurate detection of contamination incidents in drinking water is critical for notifying consumers of threats and risks to public health and for making remediation and recovery decisions. Sandia National Laboratories (SNL) and the United States Environmental Protection Agency (EPA) developed the CANARY event detection software to enable online contaminant event detection for time-critical decision making in both routine and emergency water quality assessments. As a free software tool, CANARY is available to drinking water utilities of all sizes worldwide striving to provide the best quality water to their customers.

CANARY is a software package that performs on-line, multivariate, event detection from networked sensor data. Employing statistical forecasting and classification algorithms, CANARY continuously analyzes time series signals for anomalous conditions. By combining standard statistical methods in an innovative framework, noisy data is filtered to accurately identify anomalous events while minimizing false positive detections. Through testing on data from partner water utilities and a two-year pilot study, CANARY has been shown to be effective and its performance documented. The software is compatible with any sensor technology or information technology platform, and can be easily modified for specific applications. In contrast to proprietary systems, CANARY provides the end-user with transparency in the algorithms and their parameterization, which is important for utility-specific customization. In addition to achieving homeland security goals, CANARY can be used to enhance day-to-day water quality management. Development of CANARY has focused on providing enhanced monitoring of water quality within distribution networks; however, its capabilities are general and applications of CANARY to other online event detection applications are being pursued.

> “Through testing on data from partner water utilities… CANARY has been shown to be effective and its performance documented.”

Through an open-source licensing approach, CANARY allows all utilities access to state-of-the-art event detection capabilities that can leverage their existing investments in water quality sensors.

The Need for Water Security

Continuous, reliable delivery of safe drinking water to customers is essential to the viability of large metropolitan areas, and the distribution networks used to deliver water are a critical component of municipal infrastructure systems. The scale, diversity, and complexity of these networks render them susceptible to accidental and intentional contamination events. The potentially high public health and economic consequences of such events have focused recent research on strategies to make both the physical and cyber components of water distribution systems robust against contamination. The concept of a contamination warning system (CWS) has been proposed as an integrated tool that employs *in situ* sensors, supervisory control and data acquisition (SCADA) systems, and water quality event detection systems (EDS) to continuously monitor network conditions and warn operations personnel of potential contamination events.

Issued in December, 2003, and January, 2004, Homeland Security Presidential Directives 7 and 9 (HSPD-7, HSPD-9) establish a national policy for Federal departments and agencies to identify and prioritize critical infrastructure for protection against terrorist attacks, including a mandate to "…develop robust, comprehensive, and fully coordinated surveillance and monitoring systems for…water quality that provides early detection and awareness of disease, pest, or poisonous agents."[1]

> "Continuous, reliable delivery of safe drinking water to customers is essential to the viability of large metropolitan areas, and the distribution networks used to deliver water are a critical component of municipal infrastructure systems."

---

[1] *President George W. Bush, HSPD-9, January, 30th, 2004*

A major challenge for water security is the ability to rapidly and reliably detect the presence of contaminants in drinking water distribution systems. To date, large investments in contaminant-specific sensors utilizing micro and nano technologies have not yet demonstrated the engineering reliability necessary for continuous 24/7 monitoring of water in the ambient conditions of municipal distribution systems. In contrast, the installation of commercial, off-the-shelf water quality sensors (e.g., pH, residual chlorine, specific conductivity) within distribution networks has expanded. Controlled testing of chemical and biological contaminants injected into pipe loops at EPA's Test and Evaluation Facility demonstrated that for all contaminants tested, at least one water quality sensor responded to the introduction of the contaminant. These results demonstrated that a suite of commercially available off-the-shelf water quality sensors could provide broad-based indication of contamination events in a water distribution system. Event detection in real-world situations presents challenges including reliable recognition of signals above noisy backgrounds, effective integration of changes in the hydraulic operations that impact water quality, and flexibility in connecting to existing SCADA systems with a wide variety of sensor hardware and database software. CANARY has been developed to meet these challenges.

CANARY directly addresses this national mandate by providing advanced techniques for the continuous monitoring of water quality within municipal distribution networks and real-time notification of adverse changes in water quality.

Event Detection

The incoming signals analyzed by CANARY are noisy measurements of changing environmental conditions. CANARY uses adaptive filtering to process these noisy signals and detect significant changes within them. These changes are indicative of water quality degradation within the distribution network due to intentional or accidental contamination events. CANARY employs several novel algorithms to incorporate information on operational changes within the utility into the event detection process and to recognize recurring signal patterns indicative of changes in background conditions. A series of embedded graphical editors facilitate user-based selection and parameterization of the algorithms, as well as the creation and editing of pattern libraries and the generation of graphics illustrating event detection results. CANARY leverages existing investments by connecting to a utility's SCADA database either directly or through third-party software, providing alerts to the system operator when significant water quality changes are detected.

In the basic mode of operation, CANARY uses a four-step process to examine recent water quality data and identify significant deviations from those data:

1) Estimation: For each time series of data from a single sensor, CANARY adaptively predicts the expected water quality value for the next time step. CANARY looks backwards within a user-defined moving-window of previous time steps and uses the data in this window to estimate the value of the next time step. The data are first normalized to remove the units of measurement so that different signals with different units of measurement can be easily combined.
Two estimation approaches are available within CANARY: linear filtering and multivariate nearest neighbor.

A) 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

> The fundamental issue here is to reliably detect rare events in a noisy environment.

of previous values in the prediction of the next value no matter how far in the past that value has occurred. These weights are calculated automatically within CANARY and are updated at each time step to dynamically adapt the prediction to recent changes in the water quality data. The weighted average of the set of previous values serves as the prediction of the water quality value at the next time step.

B) Multivariate Nearest Neighbor: This estimation approach also uses the normalized water quality values within the predefined set of previous time steps. The set of values at each time step across *n* different water quality sensors can be considered as a point in *n*-dimensional space. At each new time step, a new point in *n*-dimensional space is created, and its “nearest neighbor,” or the closest point in the set of previous values, serves as the predicted value for this time step.

2) Residual Calculation and Fusion: As the observation at the current time step becomes available through the SCADA system, it is normalized for comparison to the predicted value and a residual (predicted - measured) value is calculated. This process is repeated for each water quality signal at the monitoring station. In the linear filtering 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, there is only one

residual distance no matter how many different sensors are used because the residual is measured as a single distance within the multi-dimensional space.

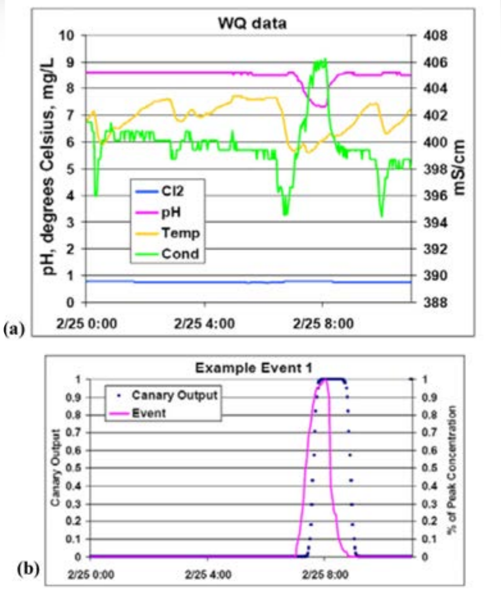
3) Residual Classification: The maximum residual value across all of the different water quality sensors for the current time step is compared to a user-defined threshold value, also defined in units of standard deviations. Residual values exceeding the threshold are classified as "outliers" and are excluded from the history window used to predict future water quality values. Other approaches to fusing residuals have been examined, including summing and averaging residuals, but results show that retaining the maximum residual for each time step provides the best overall results.

4) Probability Calculation: A Binomial Event Discriminator (BED) was developed for CANARY to create a time-integrated probability of an event [*P(event)*]. The BED models the occurrence of outliers as a binomial process and defines *P(event)* as a function of three user-defined parameters: 1) the number of outliers within the BED integration window; 2) the length of the BED integration window; and 3) the probability of an outlier occurring at any given time step under an assumption of background water quality conditions. User specification of these parameters allows for maximum flexibility in the definition of an event and sensitivity of the event detection process. Integrating results over multiple time steps creates a lag time between the true onset of the event and the time at which

CANARY detects an event. However, testing at partnering water utilities has shown that the disadvantage of increased time to detection is significantly outweighed by the advantage of decreased false alarms that result from using BED to integrate evidence for an event over multiple time steps.

The figure at the top right shows the inputs to CANARY and compares CANARY output over the course of a water quality contamination event. Note the effectiveness of the BED algorithm in keeping the probability of an event equal to zero prior to the actual event despite the noise in the input signals. This ability to keep *P(event)* near zero is a key element of the false positive reduction within CANARY. The response of CANARY to the event is delayed by the parameters of the BED that integrate responses for several time steps before increasing the probability of an event above zero. This lag is specified by the user and provides a tradeoff between false positive alarms and time-to-event detection.

CANARY improves on current industry-wide standard approaches to event detection that employ "set points" – fixed threshold values outside of which water quality is deemed anomalous and an alarm is sounded. Set points cannot provide the detection sensitivity of the CANARY algorithms, which focus on identifying relative changes that do not exceed the set point values. However, set points do provide value in alerting the system operator to persistent and gradual changes in water quality. Therefore, CANARY incorporates user-specified set points into event detection by increasing *P(event)* as the observed water quality moves towards a set point value. Results from the set point algorithm can be combined with other algorithms through a consensus approach to provide event detection for either absolute or relative changes in water quality.



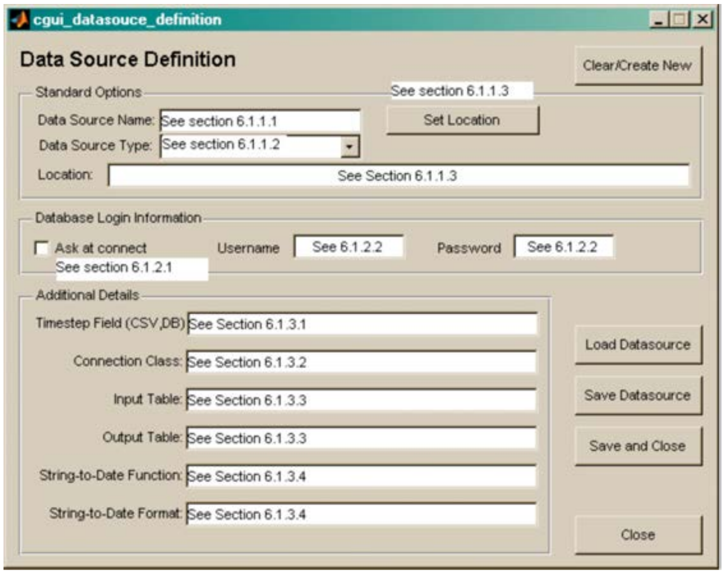
Water quality signal points (a) and CANARY response with BED activated (blue square in b) compared to the normalized concentration of the contaminant (magenta line in b). Note that CANARY maintains zero probability of an event outside of the actual event duration despite the noisy signals. See text for additional details.

The algorithms within CANARY are designed and implemented to allow for fast data input and output as well as rapid analysis. This design approach makes CANARY both efficient and essentially unlimited in terms of both the number of monitoring locations analyzed simultaneously and the number of water quality and hydraulics signals available at every monitoring station. Using a single processor desktop computer, a current application of CANARY at one of the largest utilities in the world easily and simultaneously analyzes data from 70 monitoring stations with four to six signals reporting on a five-minute sampling interval.

CANARY detects anomalous events using the four-step approach described above, and this approach has proven to be robust in deployments at several operating utilities in the past nine months. However, other applications may exist where additional algorithms are required or new algorithms need to be tested. CANARY is designed to be extensible such that additional algorithms written in Java can be incorporated directly into the event detection process. CANARY provides an application programming interface (API) to connect additional algorithms.

Parameters are entered into CANARY through a configuration file. The configuration file is written in extensible markup language (XML) and can be edited using any text editor. To improve user-friendliness, CANARY has a graphically driven configuration file editor. This editor uses a series of dialog windows to lead the user through the logical sequence of steps for connecting CANARY to the data source, defining the water quality signals to be used in event detection, choosing algorithms, and selecting algorithm parameters. The figures below show some examples of the screens used in the configuration file editor.

Examples of the CANARY graphical user interface showing the ability to define data sources (below), create a pattern library from historical water quality data (top right), and then visualize and edit the pattern library (bottom right).



cgui_datasouce_definition

Data Source Definition

Standard Options

Data Source Name: See section 6.1.1.1

Data Source Type: See section 6.1.1.2

See section 6.1.1.3

Location: See Section 6.1.1.3

Database Login Information

☐ Ask at connect

See section 6.1.2.1

Username: See 6.1.2.2

Password: See 6.1.2.2

Additional Details

Timestep Field (CSV,DB): See Section 6.1.3.1

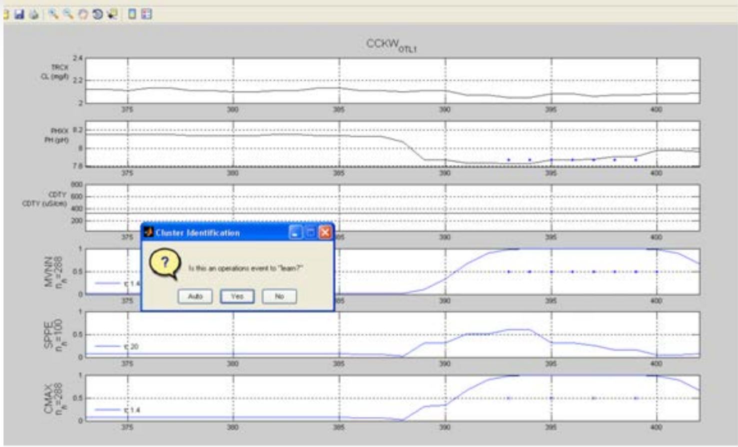
Connection Class: See Section 6.1.3.2

Input Table: See Section 6.1.3.3

Output Table: See Section 6.1.3.3

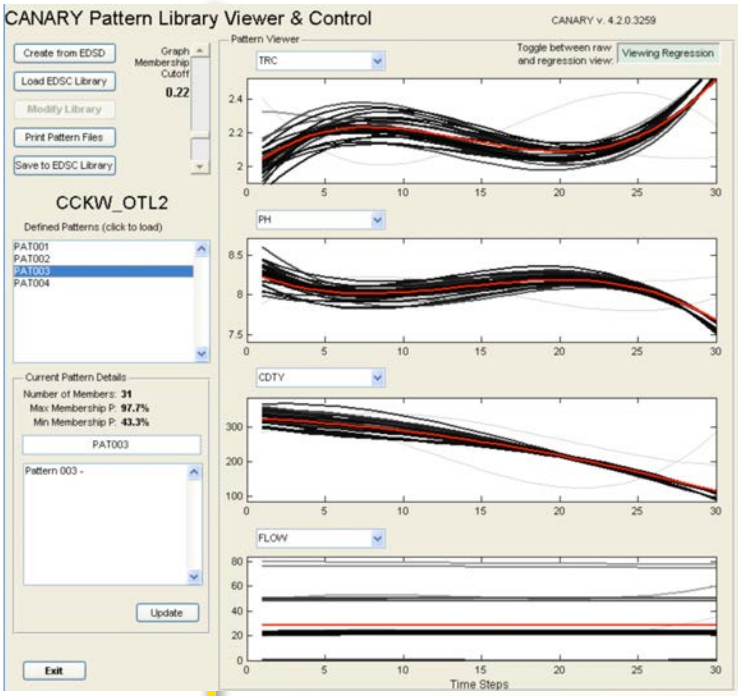
String-to-Date Function: See Section 6.1.3.4

String-to-Date Format: See Section 6.1.3.4



Cluster Identification

Is this an operations event to "learn?"



CANARY Pattern Library Viewer & Control

CANARY v. 4.2.0.3259

Graph Membership Cutoff 0.22

CCKW_OTL2

Defined Patterns (click to load)

PAT001
PAT002
PAT003
PAT004

Current Pattern Details

Number of Members: 31

Max Membership P: 97.7%

Min Membership P: 43.3%

PAT003

Pattern 003 -

Toggle between raw and regression view: Viewing Regression

Key Innovations: Limiting False Alarms and Distributed Detection

Event detection from water quality data, as well as in other security-focused monitoring applications, is a prototypical case of searching for high-consequence events that have a low probability of occurrence. A perfect EDS would have the sensitivity to detect all events (no false negatives) as well as the specificity to alarm only on water quality changes that are due to true events (no false positives). A challenge to achieving this goal of a perfect EDS in water distribution systems are the significant water quality changes caused by hydraulic operations of the network (e.g., valves opening and closing, pumps starting and stopping, changes in flows). In addition to the BED described previously, two innovative approaches to reducing false positives associated with hydraulic operational changes are available within CANARY. These approaches – composite signals and trajectory clustering - can integrate operational data directly into the water quality event detection process.

Composite Signals

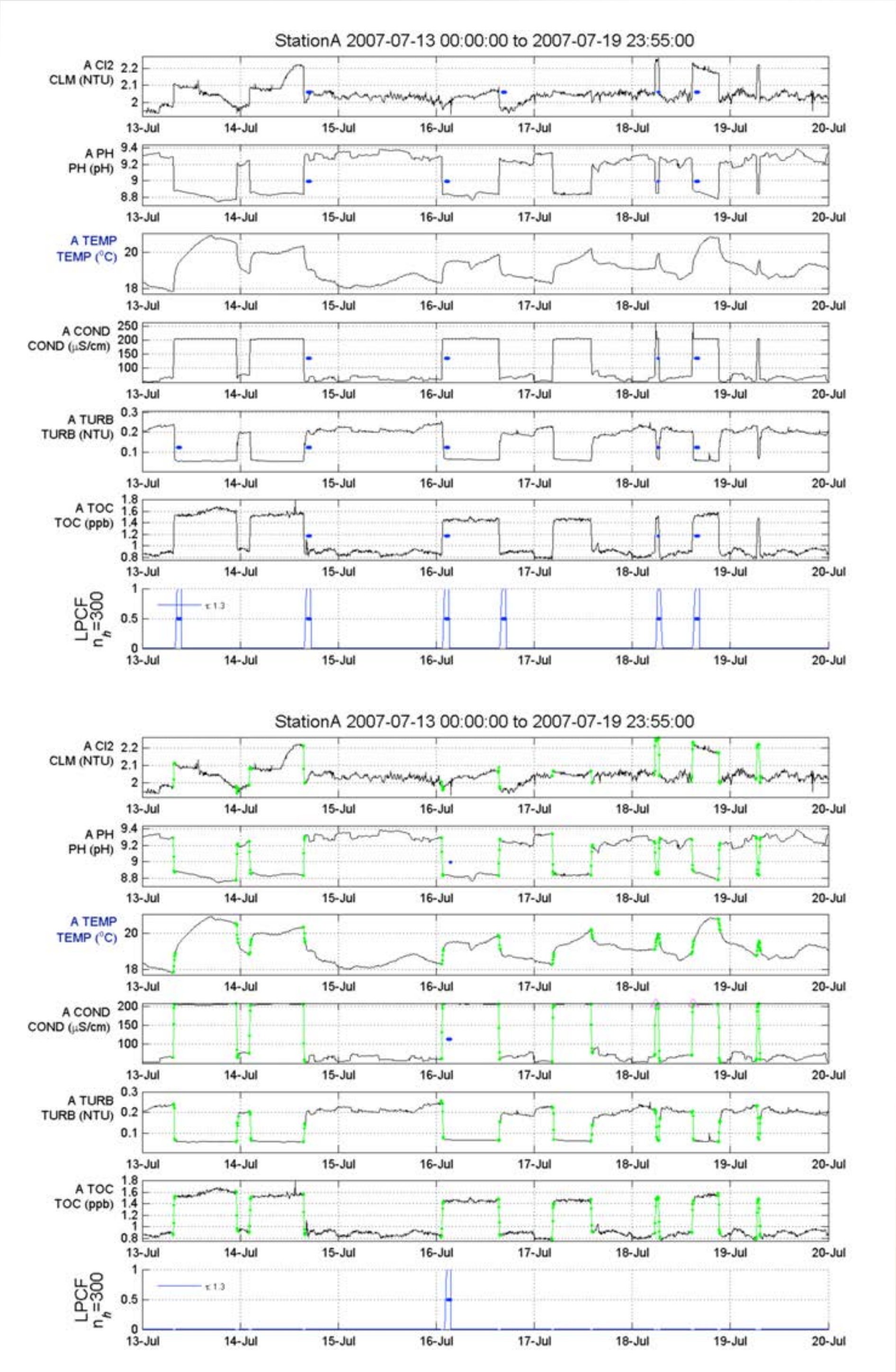
Water quality changes can be a direct and immediate result of nearby changes in network operations. In these situations, the relationship between the operational change and the water quality change is known or can be deduced by the utility operators. Often, multiple operational signals (e.g., at least one of three pumps changing status) need to be combined and reduced into a concise measure that informs CANARY of an operational change that impacts water quality at that location.

CANARY provides a simple scripting language that enables utility personnel to utilize their knowledge of the system operations to easily customize the event detection algorithms to recognize specific operational changes. This scripting language operates using reverse Polish notation similar to that found in programmable calculators, allowing the user to perform algebraic operations on any scalar data values. Through this capability, CANARY provides the user a range of options to define composite signals that are algebraic combinations of existing signals. These combinations can include differences, ratios, sums, or log transforms, etc., of one or more existing signals at the current or previous time steps. In addition, new signals can be created from differences between values obtained at different time steps from the same signal. The flexibility of the scripting language within CANARY makes it feasible to use an essentially infinite number of combinations of operational and water quality data to customize event detection at each water quality monitoring location within a utility. The approach is general and provides any utility the capability to integrate specific operational knowledge into event detection through custom-built scripts.

> “This degree of site-specific customization available through composite signals allows CANARY to meet a national security need without pushing a “one-size-fits-all” solution onto water utilities.”

This degree of site-specific customization available through composite signals allows CANARY to meet a national security need without pushing a “one-size-fits-all” solution onto water utilities.

The figure on page 18 shows event detection results with and without using composite signals for a set of water quality data that are influenced by nearby hydraulic operations. In the lower image set, scripting was used to create a composite signal that identified when a flow value recorded in at least one of three nearby pipes had exceeded a threshold level of five gallons per minute within the previous ten minutes. This new composite signal was used to cue the event detection algorithms of times of water quality change (green signals in lower image set) and temporarily decrease the event detection sensitivity during those periods.



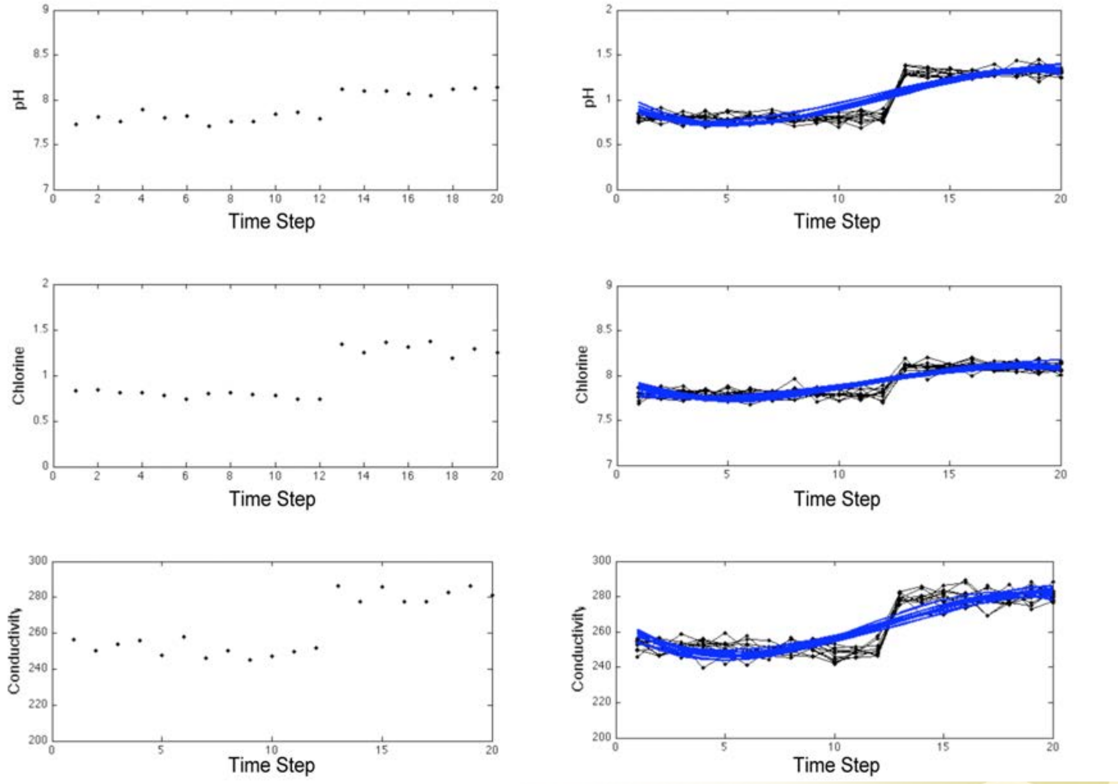
Six water quality signals (black lines) and the probability of an event (blue dots and lines) are calculated by CANARY. Detected water quality events are show as blue dots within the plot of the signal that caused the event. One week of data and results are shown. In the top image set, no accounting for operational information is made and there are six distinct periods of events. In the lower image set, operational data are incorporated as discussed in the text and time periods identified as being influenced by changes in nearby flow rates are colored green. The event detection sensitivity within CANARY is reduced during these periods of operational influence resulting in a single event identification (lower image set).

Trajectory Clustering

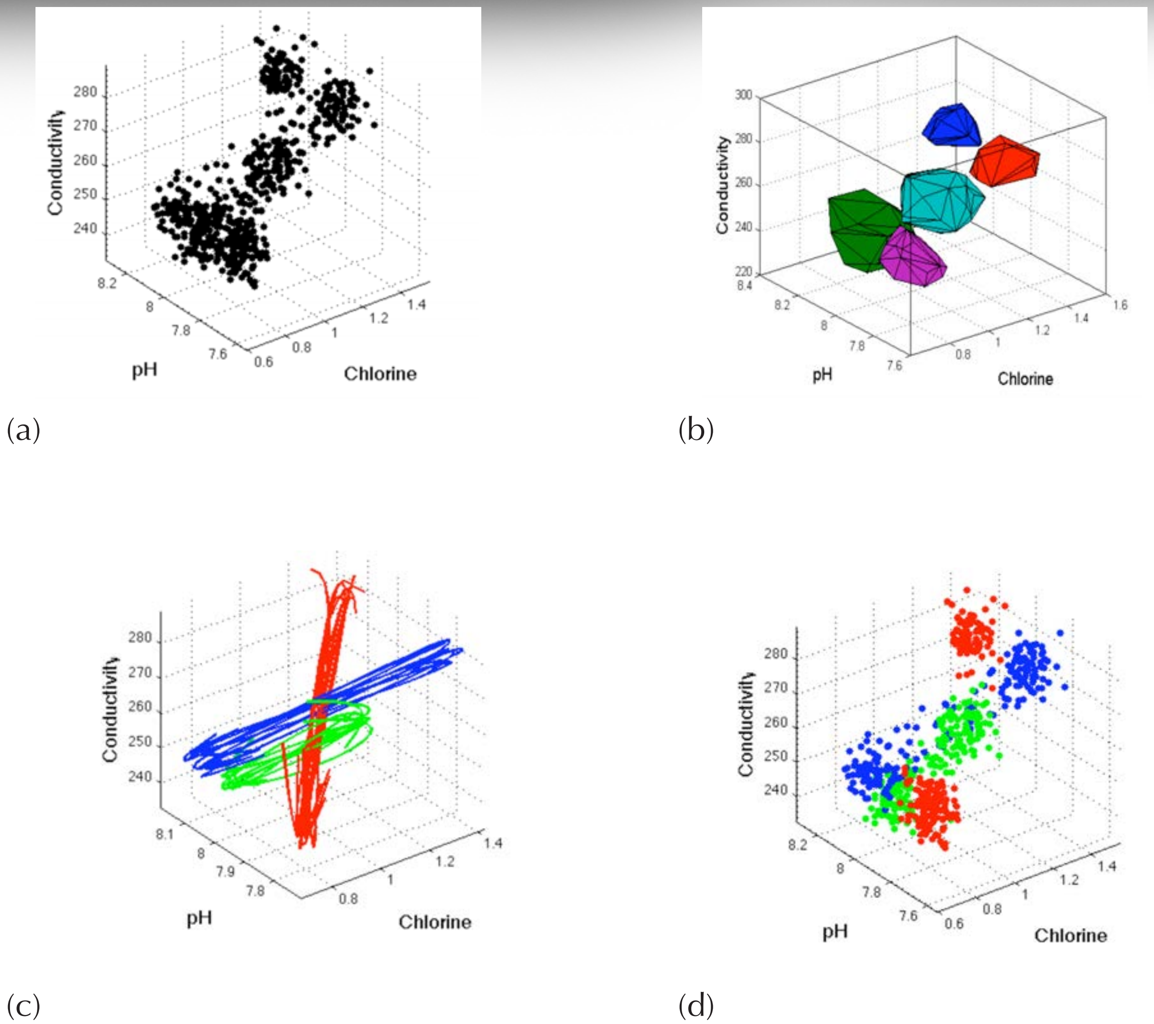
Integration of operational data into water quality event detection is more complicated for water quality monitoring stations far from the source of an operational change. The lag time between operational and water quality changes, as well as the character of the water quality change, are often variable due to changing flows within the distribution network. CANARY implements a recent development in data mining research, trajectory clustering, along with fuzzy clustering algorithms to create an efficient mechanism for classifying recurring multivariate water quality changes within a pattern library. Additionally, CANARY accesses this pattern library in real time to evaluate any potential water quality event against previously seen changes, thereby integrating pattern matching into robust event detection.

Typical clustering approaches treat the multivariate measurement vector at each time step as an independent feature, resulting in loss of information on the sequential nature of the water quality signal values across time. However, time series event detection depends on understanding the relationship of any measurement with those directly preceding it. Therefore, CANARY does not cluster the actual data, but a representation of the pattern created by sequential data prior to and including the event. CANARY implements low-order polynomial regression models to define each potential event as a relatively smooth trajectory through multi-dimensional water quality space as a function of time. Multivariate clustering is then employed on the coefficients of the regression functions, not on the actual data, to classify the trajectories into distinct clusters.

This trajectory clustering pattern matching approach is implemented in a two-step process within CANARY. First, historical data from a monitoring station are used in an offline analysis to develop a multivariate pattern library. The clustering of the regression coefficients is conducted using a fuzzy C-means clustering algorithm to define a degree of membership for each water quality event within each cluster. In the second step, online analysis compares any potential water quality event to the existing pattern library. If the current water quality pattern matches an existing pattern within a user-specified tolerance, the current water quality pattern is accepted as background water quality and the event is added to the appropriate existing pattern. Otherwise, the current pattern is considered unknown and an event is signaled. A novel aspect of this approach is that operational signals (e.g., flow rate, temperature) can be directly included in the multivariate pattern definitions. The following sequence of three figures demonstrates the concepts of trajectory clustering and results of an application of pattern matching in CANARY.



A change in water quality is detected by CANARY (left images). This change is defined by increases in all three water quality signals, and could be a contamination event or a variation caused by changes in the operational state of the distribution network. Examination of historical data shows that this is a recurring pattern, most likely due to changes in network operations, and the similar patterns in the historical data can be approximated by a third order polynomial (right images, blue lines).



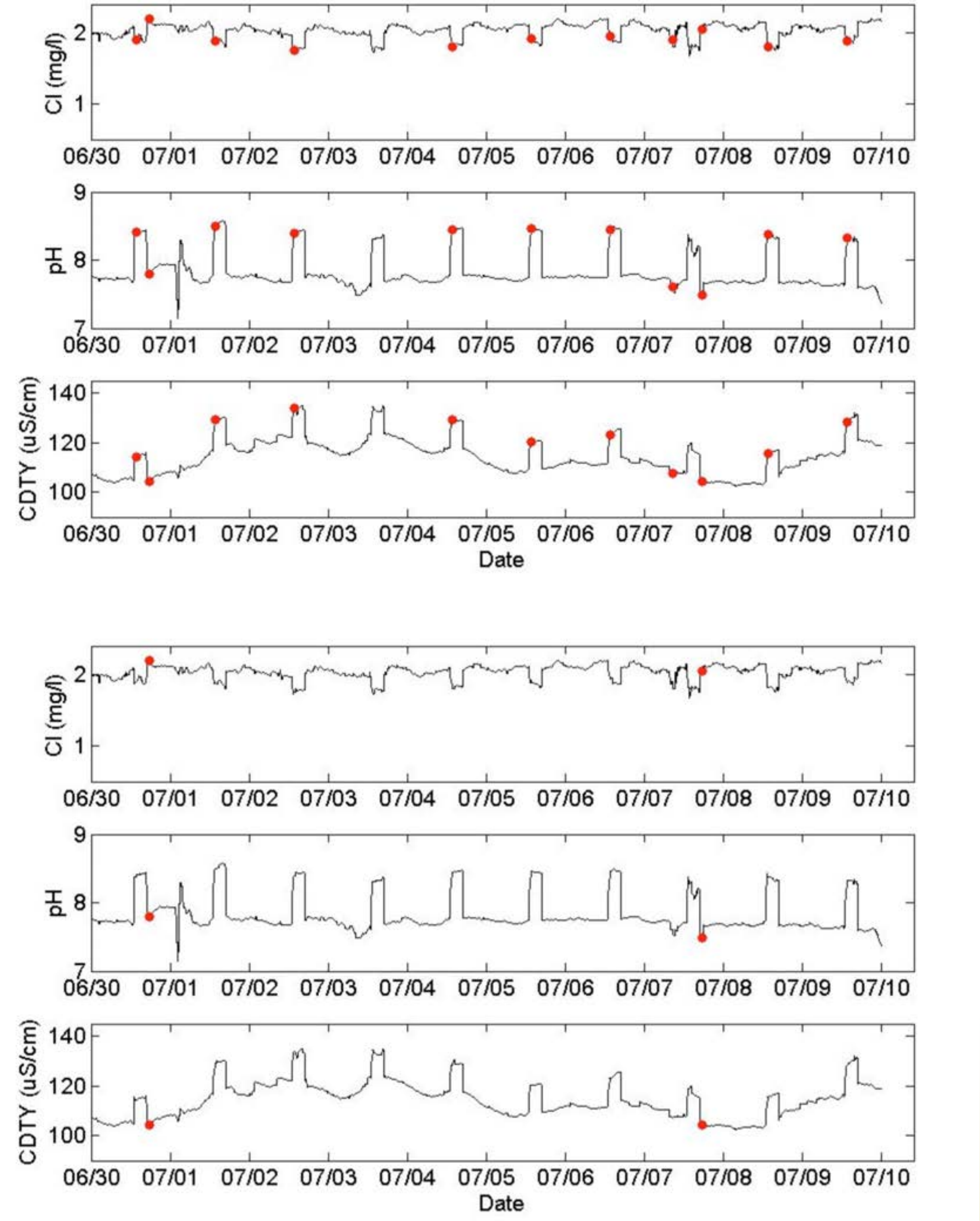
(a)

(b)

(c)

(d)

The historical data defining a pattern as shown in previous figure are combined with data taken from two other recurring patterns and plotted in multidimensional space (a). Traditional approaches to multivariate clustering do not utilize information on the sequential nature of the data and define five separate groupings for these data (b). Trajectory clustering retains the sequential nature of the data throughout the change in water quality by applying multivariate clustering to the regression coefficients, not to the data. The three groups of change patterns identified by trajectory clustering are seen clearly in (c). The correct classification of the raw data is readily apparent from the trajectory clustering (d).



Recurring daily changes in water quality: decreases in chlorine (Cl) coupled with increases in pH and conductivity (CDTY) are identified as events (red dots) in the upper figure. Creation of a pattern library from historical water quality data captures these daily changes and allows for real-time pattern matching eliminating the false alarms as shown in the bottom figure. Ten days of data are shown. Over the entire 72-day data set, the pattern matching capability reduced the number of false alarms by 79 percent.

> “...the pattern matching capability reduced the number of false alarms by 79 percent.”

Distributed Event Detection

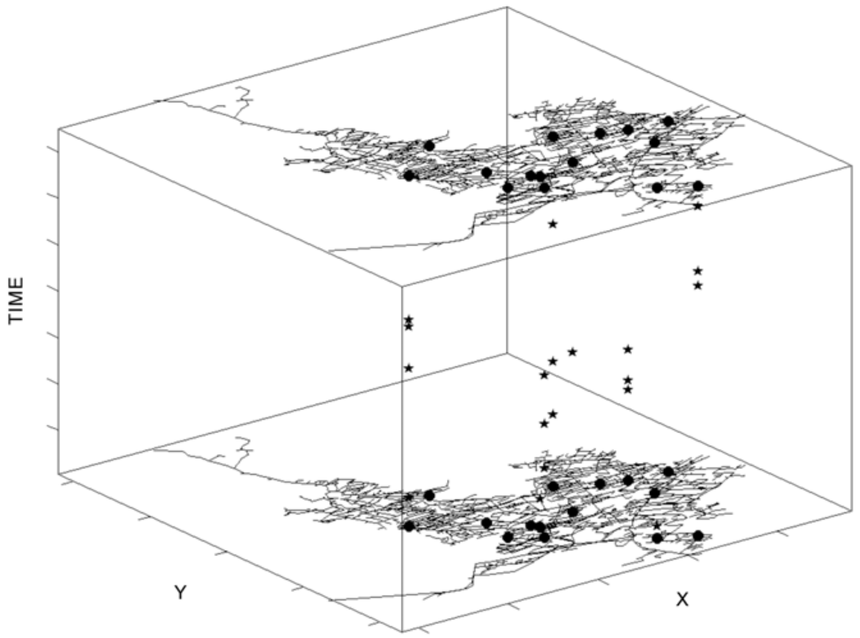
CANARY has been developed to operate independently and simultaneously on data from multiple locations within a distribution network and this aspect of CANARY is similar to other existing EDS tools. An obvious outstanding need is to be able to connect event detection results at multiple locations within a network into an integrated "network-wide" event detection.

Recasting this National Academy of Sciences research priority to utilize an array of *event detection results*, rather than using the sensors directly, Sandia researchers developed and demonstrated an efficient approach to integrating event detection results from multiple independent instances of CANARY running within a utility. This approach treats events detected by CANARY as a spatial-temporal point process and uses scan tests to identify significant clusters of events, i.e., zones in space and time with event detections that are significantly above that expected from a background false alarm rate. The figure on page 24 shows a space-time "cube" with observations of alarm locations in space and time over a 24-hour period for an example network. The centralized processing on a single computer used by CANARY enables rapid integration of event detection results from across the network.

> "Improved event detection architecture could possibly reduce the number of false positives. In this approach, a water system would install an array of sensors linked in a way that only triggers an alarm when a statistically significant number of sensors detect abnormal levels. This should reduce or eliminate the false positives caused by independent sensor malfunctions..."

NAS (National Academy of Sciences), 2007, Improving the Nation's Water Security: Opportunities for Research, Water Science and Technology Board (WSTB), National Academies Press, 170 pp. (available at http://www.nap.edu/catalog/11872.html)

The distributed detection approach considers the topology of the distribution network directly in defining the degree of connection between all water quality monitoring stations and in determining the likelihood that two or more CANARY alarms could be caused by the same contamination event. In an example calculation with an individual false alarm rate of one-per-day for each monitoring station, the spatial-temporal fusion approach was able to reduce the number of network-wide false alarms per month by *three orders of magnitude or more* depending on the size of the contaminant plume and the number of monitoring stations.



The layout of the Metropolis distribution network in geographical space (X and Y axes) with time (24 hours) shown on the vertical axis and locations of event detections (stars) in space and time. The Metropolis network contains 3323 pipe junctions and, for this example, monitoring stations have been randomly placed at 25 junctions.

CANARY Application, Impact, and Dual-Use Benefits

Since being released in May 2009, CANARY has been installed at several water utilities across the US and is running on the national distribution system in Singapore. Several of the largest water utilities in the US are installing CANARY as part of EPA's Water Security Initiative. Additionally, multiple software vendors, both US and foreign, are interested in extending their existing products to integrate CANARY capability.

> "Since being released in May 2009, CANARY has been installed at several water utilities across the U.S. and is running on the national distribution system in Singapore."

The public health and economic impacts of a water contamination event are significant and real.
As an example, the 1993 *Cryptosporidium* incident in Milwaukee, Wisconsin was the largest known outbreak of waterborne disease in US history.[2] More than 400,000 people were infected, which resulted in over 4,400 hospitalizations and at least 69 deaths. The total cost of this outbreak-associated illness was $96.2 million. With CANARY in place, such an event would be detected earlier, resulting in fewer illnesses and deaths. Murray, et al. (2009) estimate that CWS's developed for water utilities participating in the EPA's Threat Ensemble Vulnerability Assessment (TEVA) Research Program could reduce expected fatalities by 48% and associated economic consequences by over $19 billion. This analysis assumes that water quality sensor data can be analyzed with CANARY to reliably detect contaminants.

In addition to water security concerns, water utilities are interested in the dual-use benefits of online event detection to improve management of their distribution networks. The growing number of installed online water quality sensors and their connection to SCADA systems has significantly expanded the amount of water quality data to the point where system operators and network analysts are "drowning in data."

---

[2] *MacKenzie et al., 1994. A massive outbreak in Milwaukee of cryptosporidium infection transmitted through the public water supply. New England J. Medicine 331 161–167.*

Online tools, such as CANARY, that can be customized to the specific water quality and operational data characteristics at a utility are needed to aid the operators in better managing their systems. Experience shows that as the online analysis capabilities of CANARY are explored, utility operators are able to find new applications for these capabilities to improve system management.

> “Today, the event detection software CANARY is running online in PUB’s water supply control centre. This represents a quantum leap in PUB’s business practices...
>
> PUB is planning to place CANARY in water treatment plants for improving process control, a move which is radically different from CANARY’s traditional function as a water security tool.”

Mr. Harry Seah,
Director Technology and Water Quality Office,
PUB, (see letter in Appendices)

> “Even for the most experienced chemists or operators, it is an overwhelming and time-consuming task to monitor and interpret the enormous amount of sensor output, much less correlate it with thousands of pieces of hydraulic and operational data.”

American Water Works Association Research Foundation Request for Proposals (4182, March 2008)

Product's Competitors

To date, the competing solutions to event detection for water security have been developed by commercial water quality sensor vendors who have focused on connecting their proprietary event detection capabilities exclusively to their sensor hardware. Event detection systems that only operate with sensors from a single manufacturer are incompatible with operations at most utilities where prior investments in water quality sensors from a variety of manufacturers must work together simultaneously to monitor water quality. CANARY solves this problem by working with sensor hardware from any manufacturer as long as sensor performance characteristics (i.e., precision, accuracy) are known. CANARY provides the end-user with transparency in the algorithms and their parameterization that "black box" proprietary systems do not. This transparency is critical for utility-specific and location-specific customization of an event detection system and development of associated response strategies.

Currently, there are no other publicly available event detection systems. Two commercial water security event detection systems that are closest in functionality to CANARY are *Event Monitor*, manufactured by Hach, and *con::stat,* manufactured by S::can. These two products have recently been developed to be more open to integration with sensor hardware from other manufacturers.

Comparison Matrix

	Hach, Guardian Blue	**S::can con::stat**	**CANARY**
Cost: Event Detection Software (10 stations)	$92,500	$60,000	$0.00
Cost: Required Computing Hardware (10 stations)	$0.00 (included)	$0.00 (included)	$3500
Cost: Total	$92,500	$60,000	$3500
Algorithm Transparency	Proprietary	Proprietary	Fully Transparent
Direct Integration of Operational Data into Event Detection	No	No	Yes
Centralized processing on a single computing platform	No	No	Yes
Ability to work with sensors from multiple vendors	Custom Request	Yes	Yes

The following considerations were made in the construction of this comparison table.

- The costs are based on installation of an EDS applied to 10 water quality monitoring stations. It is assumed that the monitoring stations exist, are fully operational, and connected to a central SCADA system. Costs of the sensor hardware are not considered here and are assumed to be the same for all three event detection solutions.
- The cost estimates for the two commercial systems are based on generally available information from the vendors.
- Both commercial systems include integrated computing hardware for each monitoring station and this hardware is included in the cost estimates. CANARY only requires a single computer.
- The cost estimate for CANARY is based on purchase of a single desktop computer having moderate performance and memory capabilities with a standard MS Windows installation. Previous experience has shown this level of computing power is easily capable of processing all data produced by 10 monitoring stations.
- CANARY processes all signals at a central location and therefore requires a single computer connected to the existing SCADA system. If the number of monitoring stations were to double to 20, the total costs (third row) of the commercial systems would also double, while the cost of CANARY would remain the same.

How Product Improves on Competition

CANARY has been designed from the start to fully leverage the existing investments in water quality sensors and SCADA systems made by utilities. Commercial event detection systems have been developed by water quality sensor vendors under the design philosophy of creating additional markets for sensor hardware. Several advantages result from the design philosophy taken in development of CANARY, including:

- CANARY is not tied to sensors from a single vendor. This protects a utilities' investment in sensor hardware and, in most cases, enables immediate start up of event detection on the existing data streams at a utility.
- CANARY utilizes the existing utility SCADA system to enable analysis of data from all monitoring stations at a single central location. Centralized processing saves costs by only requiring a single computer for event detection no matter how many monitoring stations are connected to CANARY.
- Centralized processing enables execution of follow-on analyses that integrate event detection results from across the network. These follow-on analyses include distributed event detection (discussed above), fusion of event detection results with other data streams from outside the utility (e.g., public health data), and development of strategies to respond to a contamination event.

The other guiding principle in the development of CANARY was to make the event detection approach completely open to the users. An end-goal of this open-source model is to increase the collective knowledge and experience of the water quality event detection community by allowing transparency in the algorithms and their operations as well as the ability to modify and extend CANARY functionality. Competing technologies have taken a strongly proprietary, "black box" approach to event detection. CANARY meets the goal of creating an open resource in multiple ways:

- The software is open source and can be modified and extended by anyone to meet their specific needs.

- The open source license fully provides for linking of CANARY with proprietary software packages allowing software vendors to integrate functionality of CANARY with commercial products.
- The parameters controlling the event detection algorithms are well documented in the CANARY User's Manual and all of them can be modified by the end-user.
- CANARY is extensible in that new algorithms can be added to CANARY through external implementation in Java.

Extensive interaction with utilities has led the CANARY team to a deeper understanding of how operational changes can impact water quality and the importance of integrating operational signals into event detection. This realization has led to development of two approaches for integrating operational signals available through the existing SCADA systems within CANARY: composite signals and pattern matching through trajectory clustering. Competing technologies are focused solely on water quality signals and do not access additional data from the SCADA systems and therefore cannot capitalize on the additional information contained in operational data.

> “Extensive interaction with utilities has led the CANARY team to a deeper understanding of how operational changes can impact water quality and the importance of integrating operational signals into event detection.”

Product's Prinicpal Applications

The principal application for CANARY is real-time, online detection of anomalous water quality events, ranging from accidental introduction of poor quality water to intentional injection of chemical, biological, or radiological agents within municipal water distribution systems. This water security application is closely tied to improved management of the distribution network. The dual-use benefits of CANARY to assist operators in making sense of increasing amounts of online data and to provide better understanding of operational factors that alter water quality have made significant impacts within utilities using CANARY. These impacts include improved efficiency of utility operations as well as increased appreciation for the knowledge that can be extracted from data being collected within these networks.

Other Applications

CANARY is written to be generally applicable to online event detection from multivariate time series data collected in noisy environments. Other applications being investigated include:

- Computer network traffic logs: Detecting periods of anomalous behavior from online monitoring of measured parameters regarding the type and volume of internet traffic through a particular network node has many similarities with online water quality event detection. Internet traffic exhibits trends and periodicity as a function of the time of day and day of the week, much like water quality values. Identification of anomalous periods of traffic may provide early warning of denial-of-service attacks or other periods of concern.

- Geophysical log analysis: Geophysical logs provide a multivariate description of the subsurface. Here the time dimension is replaced by the depth or distance along the borehole. In particular, real-time event detection provides advantage in measurement while drilling (MWD) situations. The goals of this analysis are early detection of overpressured zones prior to drilling into them and repeatable detection of hydrocarbon-bearing pay zones.

- Satellite telemetry: Application of CANARY to analysis of satellite telemetry data is currently being assessed. Satellite sensor platforms also work in noisy environments and there is a need for real-time segregation of signals of interest from background noise.

CANARY: Event Detection Software 2010

Summary

CANARY provides water utilities around the globe free access to state-of-the art online event detection capabilities in an operational software package. It is specifically designed to identify significant changes in signals within noisy data while reducing false alarms. For water quality monitoring, CANARY analyzes data from commecially available off-the-shelf water quality sensors and has been demonstrated to identify events resulting from injection of less than 1.0 parts per million of a chemical contaminant into a pipe. In an operating water distribution system with an adequate water quality monitoring network, contamination events caused by sewer cross-connections, breaks in pipes that introduce material into the water system and injection of chemical and or biological contaminants are several of the types of events CANARY is designed to detect.

CANARY plays a significant role in protection of critical infrastructure and the public health and economic functions that rely on that infrastructure. CANARY's capabilities are critical to the effective deployment of contamination warning systems (CWSs) using water quality sensors. The widespread deployment of CWSs can significantly reduce the risks associated with catastrophic contamination incidents.

CANARY provides a user-friendly and fully customizable event detection capability that has been developed with extensive feedback from the end-user utility operators. CANARY brings to bear a novel blend of technologies from different fields on the problem of online event detection. To leverage existing investments in water quality monitoring, CANARY is able to connect with common existing data formats and SCADA databases as well as work with water quality sensors from a wide variety of manufacturers. CANARY provides complete transparency in algorithms and parameterization as well as in the integration of operational data allowing the end-user to fully customize the settings to meet conditions throughout the distribution network.

> “...CANARY analyzes data from commercially available off-the-shelf water quality sensors and has been demonstrated to identify events resulting from injection of less than 1.0 parts per million of a chemical contaminant into a pipe.”

AFFIRMATION

By uploading this form to R&D Magazine's website you affirm that all information submitted as a part of, or supplemental to, this entry is a fair and accurate representation of this product.

Sean A. McKenna

Appendices

Appendix Item A:

Letters of Support:

PUB

Mr. Harry Seah, Director, Technology and Water Quality Office

(PUB is the Singaporean National Water Management Authority and was honored for its exemplary management of water resources and application of novel technologies in 2007 with the Stockholm Industry Water Award. This award recognizes outstanding contributions made by businesses and industries to improve the world water situation: (see: http://www.siwi.org/stockholmindustrywateraward)

Metropolitan Water District (MWD), Southern California

Mr. Eric Crofts, Interim Chemistry Unit Manager and Water Quality EDS Project Manager

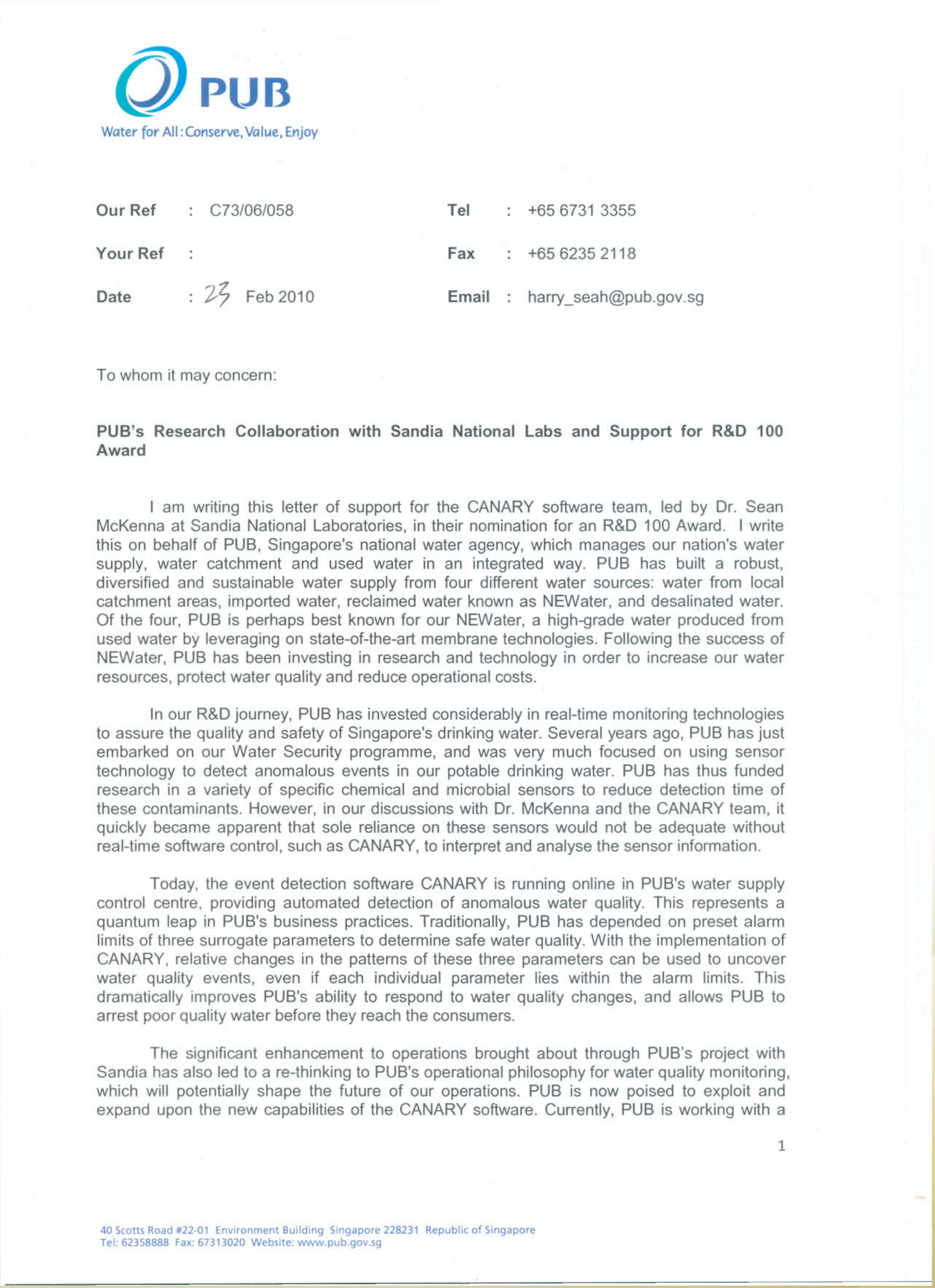
(MWD is the largest water distribution utility in the US with over 18 million customers)

Philadelphia Water Department (PWD)

Mr. Thomas Taggart, Environmental Engineer, Scientific and Regulatory Affairs

(PWD has provided integrated water management for Philadelphia for nearly 200 years)

Appendix Item B: Research articles documenting technological advances and testing results made in the development of CANARY



PUB
Water for All: Conserve, Value, Enjoy

Our Ref : C73/06/058
Your Ref :
Date : 23 Feb 2010

Tel : +65 6731 3355
Fax : +65 6235 2118
Email : harry_seah@pub.gov.sg

To whom it may concern:

PUB's Research Collaboration with Sandia National Labs and Support for R&D 100 Award

I am writing this letter of support for the CANARY software team, led by Dr. Sean McKenna at Sandia National Laboratories, in their nomination for an R&D 100 Award. I write this on behalf of PUB, Singapore's national water agency, which manages our nation's water supply, water catchment and used water in an integrated way. PUB has built a robust, diversified and sustainable water supply from four different water sources: water from local catchment areas, imported water, reclaimed water known as NEWater, and desalinated water. Of the four, PUB is perhaps best known for our NEWater, a high-grade water produced from used water by leveraging on state-of-the-art membrane technologies. Following the success of NEWater, PUB has been investing in research and technology in order to increase our water resources, protect water quality and reduce operational costs.

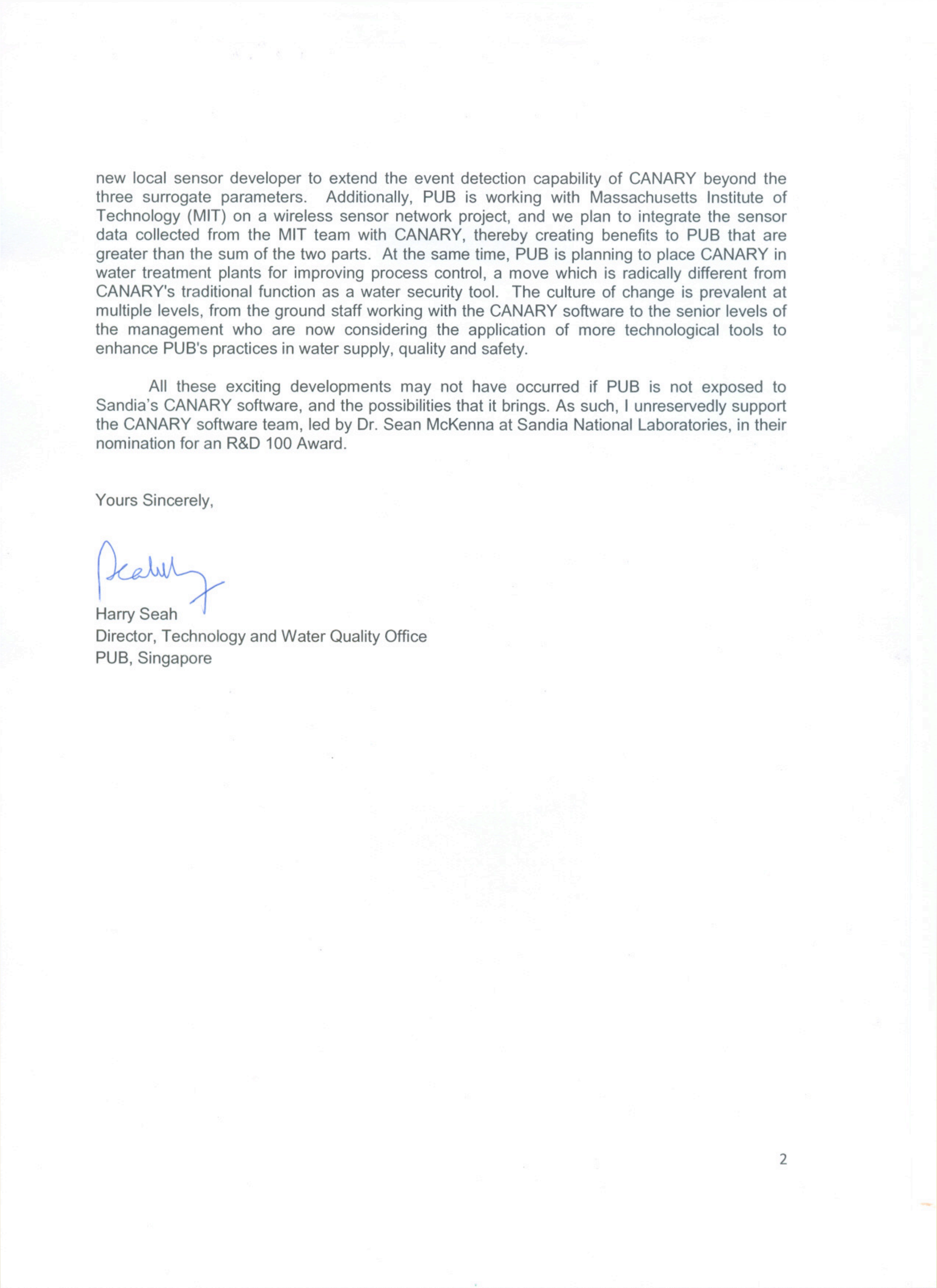
In our R&D journey, PUB has invested considerably in real-time monitoring technologies to assure the quality and safety of Singapore's drinking water. Several years ago, PUB has just embarked on our Water Security programme, and was very much focused on using sensor technology to detect anomalous events in our potable drinking water. PUB has thus funded research in a variety of specific chemical and microbial sensors to reduce detection time of these contaminants. However, in our discussions with Dr. McKenna and the CANARY team, it quickly became apparent that sole reliance on these sensors would not be adequate without real-time software control, such as CANARY, to interpret and analyse the sensor information.

Today, the event detection software CANARY is running online in PUB's water supply control centre, providing automated detection of anomalous water quality. This represents a quantum leap in PUB's business practices. Traditionally, PUB has depended on preset alarm limits of three surrogate parameters to determine safe water quality. With the implementation of CANARY, relative changes in the patterns of these three parameters can be used to uncover water quality events, even if each individual parameter lies within the alarm limits. This dramatically improves PUB's ability to respond to water quality changes, and allows PUB to arrest poor quality water before they reach the consumers.

The significant enhancement to operations brought about through PUB's project with Sandia has also led to a re-thinking to PUB's operational philosophy for water quality monitoring, which will potentially shape the future of our operations. PUB is now poised to exploit and expand upon the new capabilities of the CANARY software. Currently, PUB is working with a

1

40 Scotts Road #22-01 Environment Building Singapore 228231 Republic of Singapore
Tel: 62358888 Fax: 67313020 Website: www.pub.gov.sg

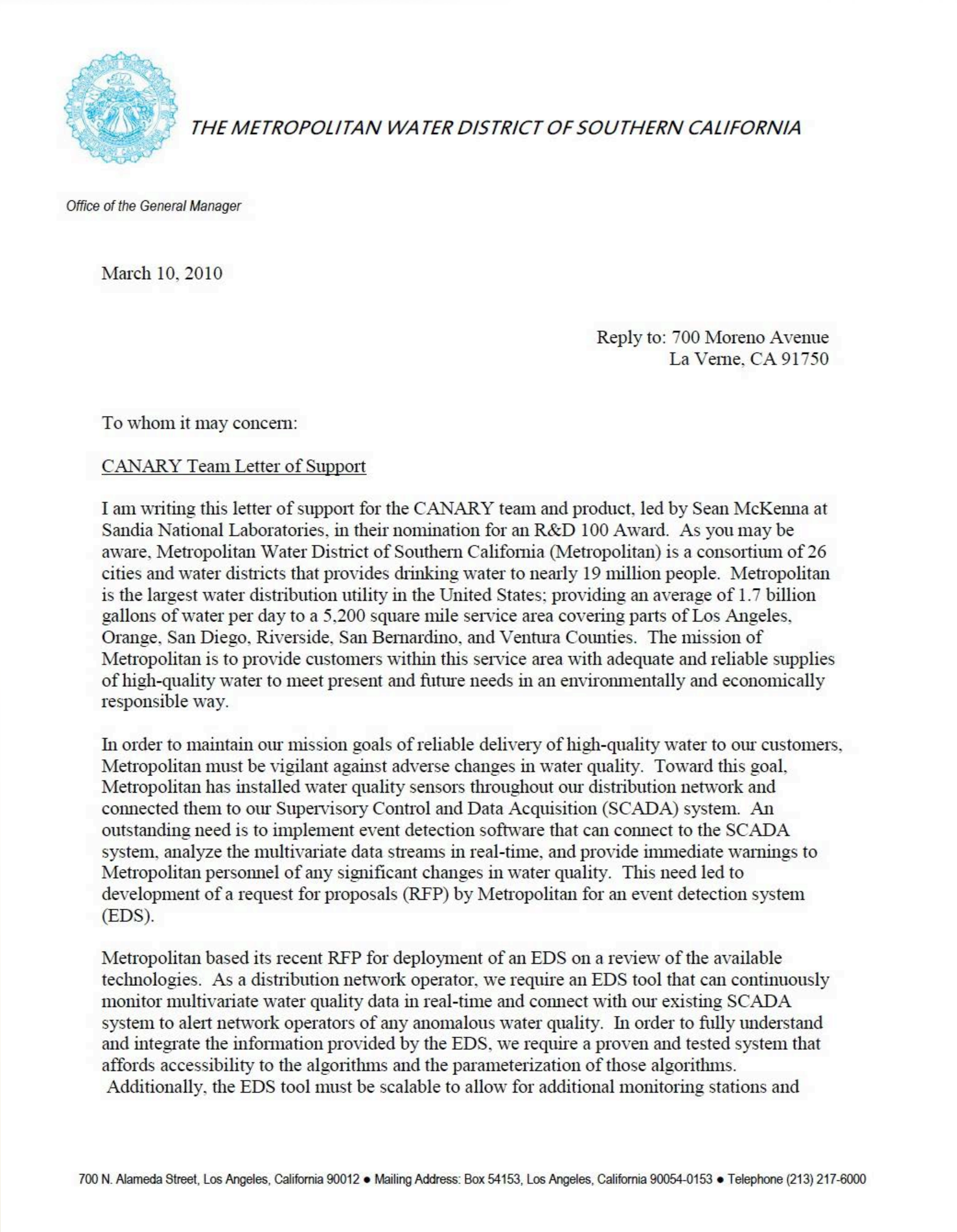


new local sensor developer to extend the event detection capability of CANARY beyond the three surrogate parameters. Additionally, PUB is working with Massachusetts Institute of Technology (MIT) on a wireless sensor network project, and we plan to integrate the sensor data collected from the MIT team with CANARY, thereby creating benefits to PUB that are greater than the sum of the two parts. At the same time, PUB is planning to place CANARY in water treatment plants for improving process control, a move which is radically different from CANARY's traditional function as a water security tool. The culture of change is prevalent at multiple levels, from the ground staff working with the CANARY software to the senior levels of the management who are now considering the application of more technological tools to enhance PUB's practices in water supply, quality and safety.

All these exciting developments may not have occurred if PUB is not exposed to Sandia's CANARY software, and the possibilities that it brings. As such, I unreservedly support the CANARY software team, led by Dr. Sean McKenna at Sandia National Laboratories, in their nomination for an R&D 100 Award.

Yours Sincerely,

Harry Seah
Director, Technology and Water Quality Office
PUB, Singapore



THE METROPOLITAN WATER DISTRICT OF SOUTHERN CALIFORNIA

Office of the General Manager

March 10, 2010

Reply to: 700 Moreno Avenue
La Verne, CA 91750

To whom it may concern:

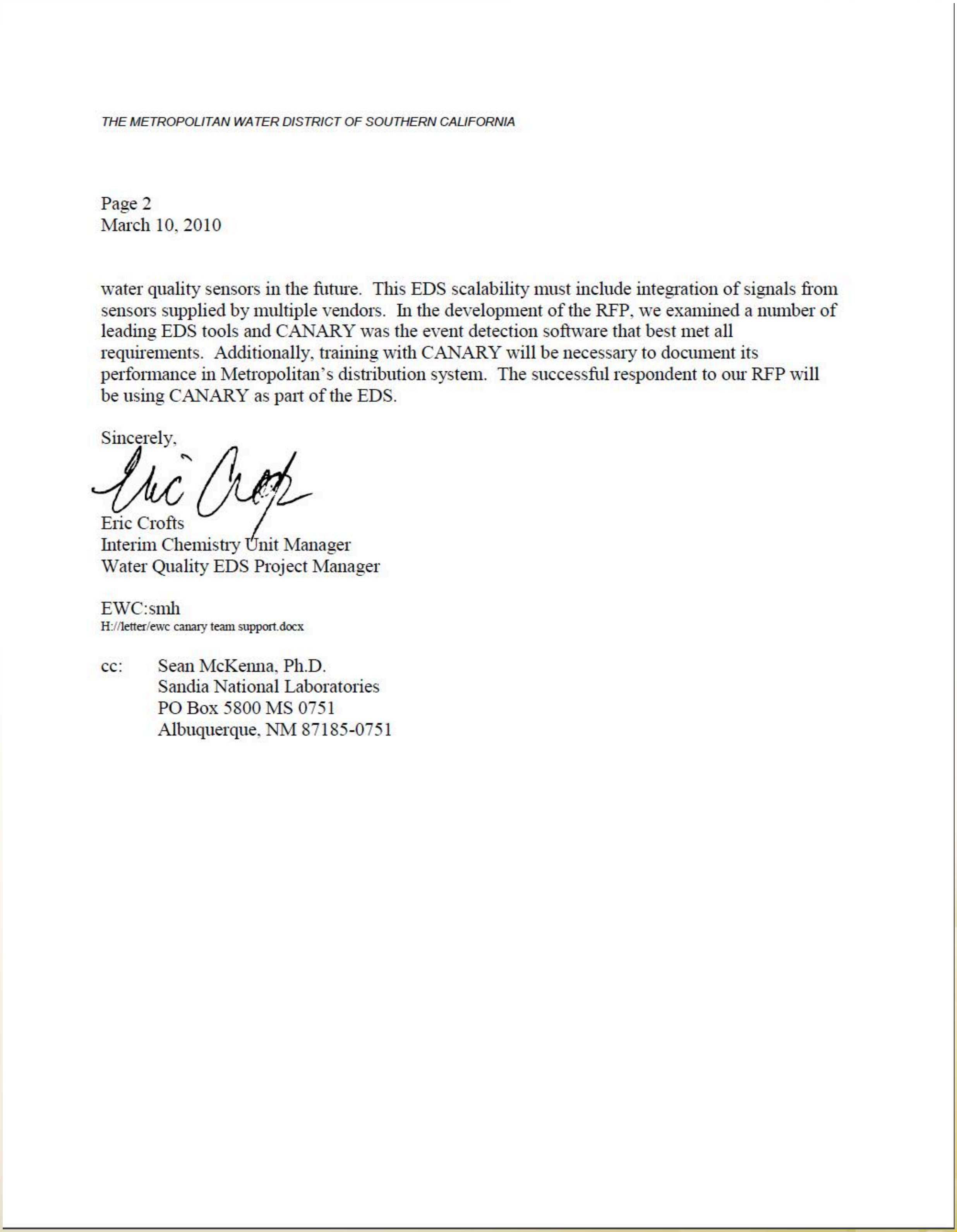
CANARY Team Letter of Support

I am writing this letter of support for the CANARY team and product, led by Sean McKenna at Sandia National Laboratories, in their nomination for an R&D 100 Award. As you may be aware, Metropolitan Water District of Southern California (Metropolitan) is a consortium of 26 cities and water districts that provides drinking water to nearly 19 million people. Metropolitan is the largest water distribution utility in the United States; providing an average of 1.7 billion gallons of water per day to a 5,200 square mile service area covering parts of Los Angeles, Orange, San Diego, Riverside, San Bernardino, and Ventura Counties. The mission of Metropolitan is to provide customers within this service area with adequate and reliable supplies of high-quality water to meet present and future needs in an environmentally and economically responsible way.

In order to maintain our mission goals of reliable delivery of high-quality water to our customers, Metropolitan must be vigilant against adverse changes in water quality. Toward this goal, Metropolitan has installed water quality sensors throughout our distribution network and connected them to our Supervisory Control and Data Acquisition (SCADA) system. An outstanding need is to implement event detection software that can connect to the SCADA system, analyze the multivariate data streams in real-time, and provide immediate warnings to Metropolitan personnel of any significant changes in water quality. This need led to development of a request for proposals (RFP) by Metropolitan for an event detection system (EDS).

Metropolitan based its recent RFP for deployment of an EDS on a review of the available technologies. As a distribution network operator, we require an EDS tool that can continuously monitor multivariate water quality data in real-time and connect with our existing SCADA system to alert network operators of any anomalous water quality. In order to fully understand and integrate the information provided by the EDS, we require a proven and tested system that affords accessibility to the algorithms and the parameterization of those algorithms. Additionally, the EDS tool must be scalable to allow for additional monitoring stations and

700 N. Alameda Street, Los Angeles, California 90012 ● Mailing Address: Box 54153, Los Angeles, California 90054-0153 ● Telephone (213) 217-6000



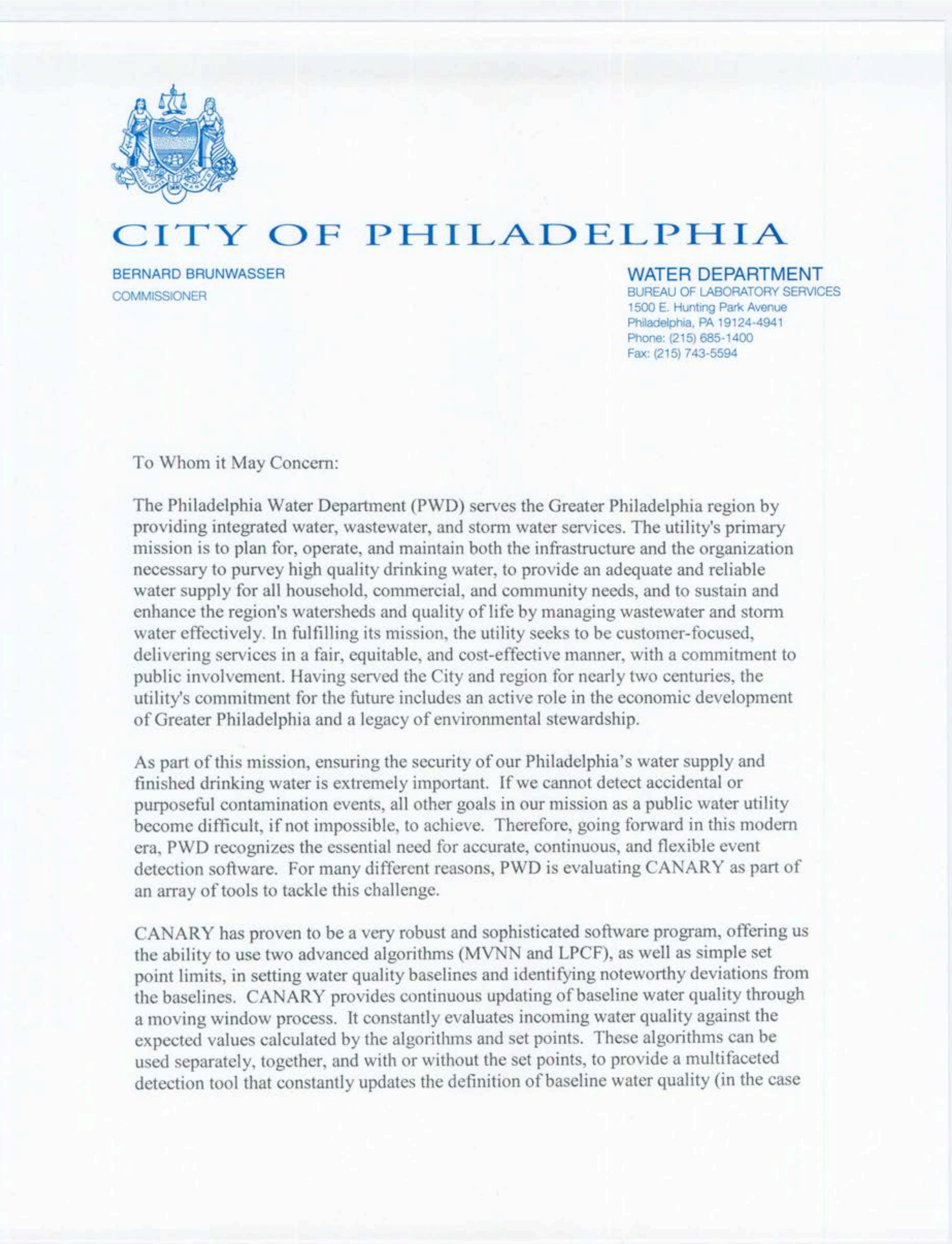
water quality sensors in the future. This EDS scalability must include integration of signals from sensors supplied by multiple vendors. In the development of the RFP, we examined a number of leading EDS tools and CANARY was the event detection software that best met all requirements. Additionally, training with CANARY will be necessary to document its performance in Metropolitan's distribution system. The successful respondent to our RFP will be using CANARY as part of the EDS.

Sincerely,

Eric Crofts
Interim Chemistry Unit Manager
Water Quality EDS Project Manager

EWC:smh
H://letter/ewc canary team support.docx

cc: Sean McKenna, Ph.D.
Sandia National Laboratories
PO Box 5800 MS 0751
Albuquerque, NM 87185-0751



CITY OF PHILADELPHIA

BERNARD BRUNWASSER
COMMISSIONER

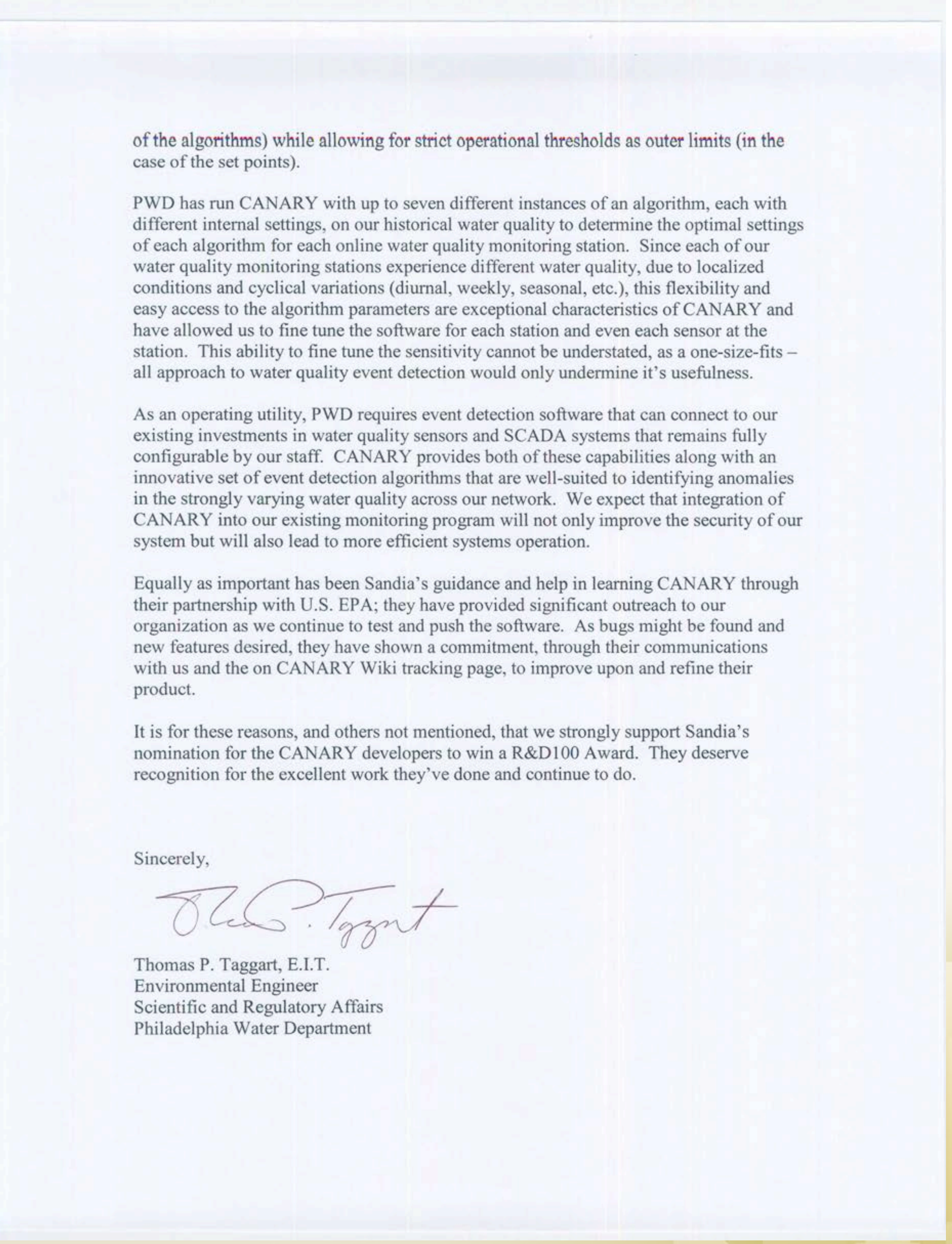
WATER DEPARTMENT
BUREAU OF LABORATORY SERVICES
1500 E. Hunting Park Avenue
Philadelphia, PA 19124-4941
Phone: (215) 685-1400
Fax: (215) 743-5594

To Whom it May Concern:

The Philadelphia Water Department (PWD) serves the Greater Philadelphia region by providing integrated water, wastewater, and storm water services. The utility's primary mission is to plan for, operate, and maintain both the infrastructure and the organization necessary to purvey high quality drinking water, to provide an adequate and reliable water supply for all household, commercial, and community needs, and to sustain and enhance the region's watersheds and quality of life by managing wastewater and storm water effectively. In fulfilling its mission, the utility seeks to be customer-focused, delivering services in a fair, equitable, and cost-effective manner, with a commitment to public involvement. Having served the City and region for nearly two centuries, the utility's commitment for the future includes an active role in the economic development of Greater Philadelphia and a legacy of environmental stewardship.

As part of this mission, ensuring the security of our Philadelphia's water supply and finished drinking water is extremely important. If we cannot detect accidental or purposeful contamination events, all other goals in our mission as a public water utility become difficult, if not impossible, to achieve. Therefore, going forward in this modern era, PWD recognizes the essential need for accurate, continuous, and flexible event detection software. For many different reasons, PWD is evaluating CANARY as part of an array of tools to tackle this challenge.

CANARY has proven to be a very robust and sophisticated software program, offering us the ability to use two advanced algorithms (MVNN and LPCF), as well as simple set point limits, in setting water quality baselines and identifying noteworthy deviations from the baselines. CANARY provides continuous updating of baseline water quality through a moving window process. It constantly evaluates incoming water quality against the expected values calculated by the algorithms and set points. These algorithms can be used separately, together, and with or without the set points, to provide a multifaceted detection tool that constantly updates the definition of baseline water quality (in the case



of the algorithms) while allowing for strict operational thresholds as outer limits (in the case of the set points).

PWD has run CANARY with up to seven different instances of an algorithm, each with different internal settings, on our historical water quality to determine the optimal settings of each algorithm for each online water quality monitoring station. Since each of our water quality monitoring stations experience different water quality, due to localized conditions and cyclical variations (diurnal, weekly, seasonal, etc.), this flexibility and easy access to the algorithm parameters are exceptional characteristics of CANARY and have allowed us to fine tune the software for each station and even each sensor at the station. This ability to fine tune the sensitivity cannot be understated, as a one-size-fits – all approach to water quality event detection would only undermine it's usefulness.

As an operating utility, PWD requires event detection software that can connect to our existing investments in water quality sensors and SCADA systems that remains fully configurable by our staff. CANARY provides both of these capabilities along with an innovative set of event detection algorithms that are well-suited to identifying anomalies in the strongly varying water quality across our network. We expect that integration of CANARY into our existing monitoring program will not only improve the security of our system but will also lead to more efficient systems operation.

Equally as important has been Sandia's guidance and help in learning CANARY through their partnership with U.S. EPA; they have provided significant outreach to our organization as we continue to test and push the software. As bugs might be found and new features desired, they have shown a commitment, through their communications with us and the on CANARY Wiki tracking page, to improve upon and refine their product.

It is for these reasons, and others not mentioned, that we strongly support Sandia's nomination for the CANARY developers to win a R&D100 Award. They deserve recognition for the excellent work they've done and continue to do.

Sincerely,

Thomas P. Taggart, E.I.T.
Environmental Engineer
Scientific and Regulatory Affairs
Philadelphia Water Department

Appendix B: Publications resulting from development of CANARY

Koch, M.W. and S.A. McKenna, (in review), Distributed Sensor Fusion in Water Quality Event Detection, submitted to: *ASCE Journal of Water Resources Planning and Management*, November, 2009

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

Hart, D.B. and S.A. McKenna, 2009, CANARY User's Manual, Version 4.2, U.S. Environmental Protection Agency, Office of Research and Development, National Homeland Security Research Center, EPA 600/R-08/040A, 51 pp.

McKenna, S.A., M. Wilson and K.A. Klise, 2008, Detecting Changes in Water Quality Data, *American Water Works Association Journal*, Vol. 100, No. 1, pp. 74-85.

McKenna, S.A. and D.B. Hart, 2008, On-Line Identification of Adverse Water Quality Events from Monitoring of Surrogate Data: CANARY Software, In Proceedings of: Singapore International Water Week, June 23rd-27th, Singapore

Koch, M.W., and S.A. McKenna, 2008, Distributed Network Fusion for Water Quality, in Proceedings of ASCE World Environmental and Water Resources Congress, Honolulu, Hawaii, May 13-16.

McKenna, S.A., D.B. Hart, K.A. Klise, V.A. Cruz and M.P. Wilson, 2007, Event Detection from Water Quality Time Series, in proceedings of: ASCE World Environmental and Water Resources Congress, Tampa, FL, May, 2007, May 15-19th.

McKenna, S.A., K.A. Klise and M.P. Wilson, 2006, Testing Water Quality Change Detection Algorithms, in Proceedings of the 8th Annual Water Distribution System Analysis Symposium, Cincinnati, OH, August 27-30, 2006.

Klise, K.A. and S.A. McKenna, 2006, Multivariate Applications for Detecting Anomalous Water Quality, in Proceedings of the 8th Annual Water Distribution System Analysis Symposium, Cincinnati, OH, August 27-30, 2006.

McKenna, S.A., D.B. Hart and L. Yarrington, 2006, Impact of sensor detection limits on protecting water distribution systems from contamination events, *Journal of Water Resources Planning and Management*, Special Issue on Drinking Water Distribution Systems Security, 132 (4), pp. 305-309.

Klise, K.A. and S.A. McKenna, 2006, Water quality change detection: multivariate algorithms, in Proceedings of SPIE (International Society for Optical Engineering), Defense and Security Symposium 2006, April 18-20, Orlando. Florida, 9pp.

Koch, M.W. and McKenna, S.A., (in review), Distributed Sensor Fusion in Water Quality Event Detection, submitted to: *ASCE Journal of Water Resources Planning and Management*, November, 2009

Distributed Sensor Fusion in Water Quality Event Detection

Mark W. Koch[1] , Sean A. McKenna[2]

Abstract

To protect drinking water systems, a contamination warning system can use in-line sensors to indicate possible accidental and deliberate contamination. Currently, reporting of an incident occurs when data from a single station detects an anomaly. This paper proposes an approach for combining data from multiple stations to reduce false background alarms. By considering the location and time of individual detections as points resulting from a random space-time point process, Kulldorff's scan test can find statistically significant clusters of detections. Using EPANET to simulate contaminant plumes of varying sizes moving through a water network with varying amounts of sensing nodes, it is shown that the scan test can detect significant clusters of events. Also, these significant clusters can reduce the false alarms resulting from background noise and the clusters can help indicate the time and source location of the contaminant. Fusion of monitoring station results within a moderately sized network show false alarm errors are reduced by three orders of magnitude using the scan test.

---

[1] Sandia National Laboratories, P.O. Box 5800, MS 1163, Sensor Exploitation Applications Department, Albuquerque, NM; mwkoch@sandia.gov

[2] Sandia National Laboratories, P.O. Box 5800, MS 0751, National Security Applications Department, Albuquerque, NM; samcken@sandia.gov

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

World Environmental and Water Resources Congress 2009: Great Rivers © 2009 ASCE 590

Trajectory Clustering Approach for Reducing Water Quality Event False Alarms

Eric Vugrin[1], Sean A. McKenna[1], David Hart[1]

[1]Sandia National Laboratories, Energy, Resources and Systems Analysis Center,
PO Box 5800 MS 1138, Albuquerque, NM 87185-1138,
{edvugri,samcken,dbhart}@sandia.gov

ABSTRACT

Event Detection Systems (EDS) performance is hindered by false alarms that cause unnecessary resource expenditure by the utility and undermine confidence in the EDS operation. Changes in water quality due to operational changes in the utility hydraulics can cause a significant number of false alarms. These changes may occur daily and each instance produces similar changes in the multivariate water quality pattern. Recognizing that patterns of water quality change must be identified, we adapt trajectory clustering as a means of classifying these multivariate patterns. We develop a general approach for dealing with changes in utility operations that impact water quality. This approach uses historical data water quality data from the utility to identify recurring patterns and retains those patterns in a library that can be accessed during online operation. We have implemented this pattern matching capability within CANARY and describe several example applications that demonstrate a decrease in false alarms.

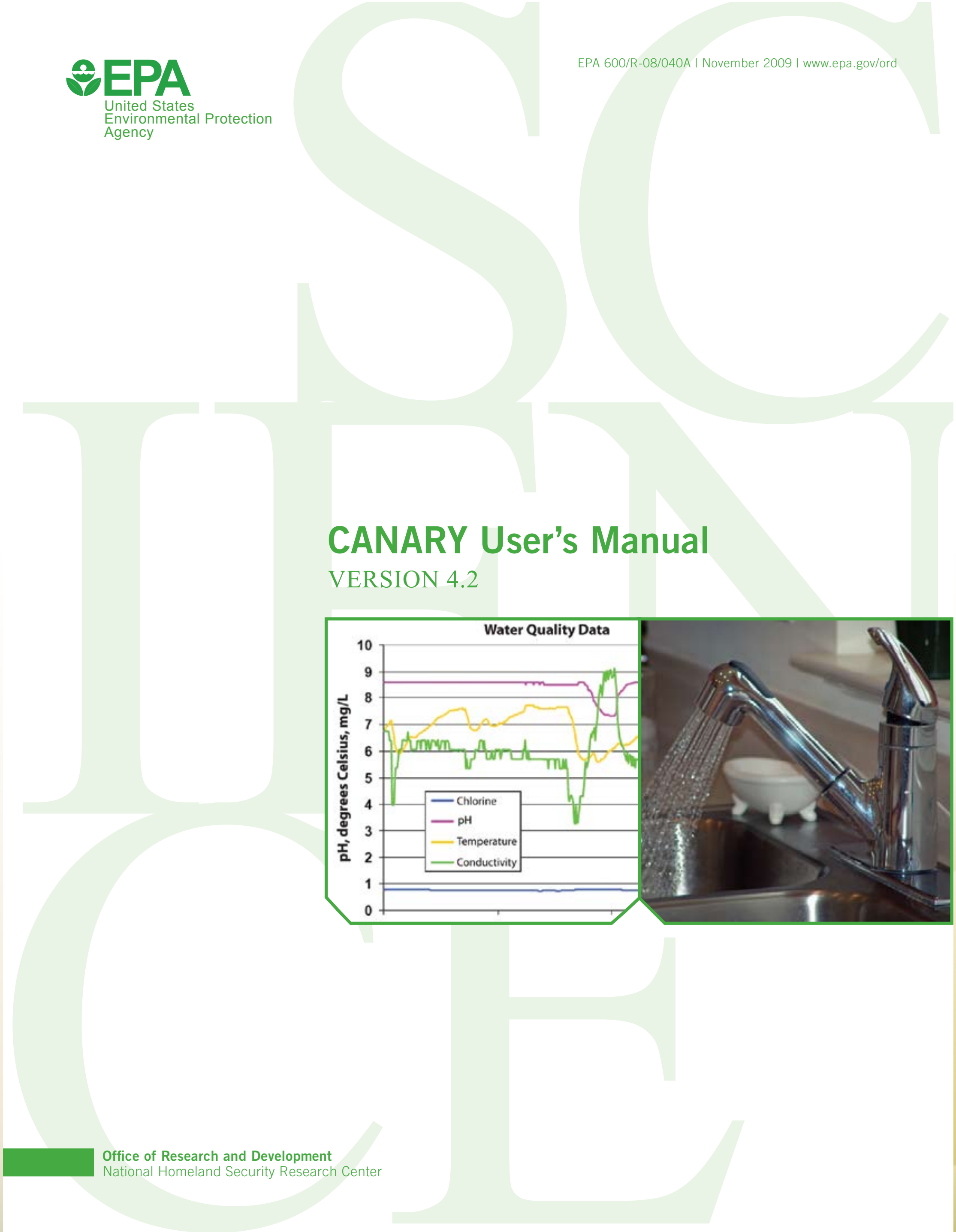
INTRODUCTION

Event detection software (EDS) tools are an integral part of contaminant warning systems being developed and deployed at water utilities. EDS constitute an integral part of contaminant warning systems (Hasan et al., 2004) and can also provide utilities with enhanced monitoring and management of daily operations. Here we focus on the CANARY EDS developed at Sandia National Laboratories (Hart et al., 2007).

A key component of the effectiveness of these EDS tools in actual operating distribution networks is their ability to reduce false positive (false alarm) detections. EDS parameters can be tuned to be more or less sensitive to changes in water quality with increasing sensitivity resulting in fewer missed detections (false negatives), but also increased false alarms. The tradeoff between false negatives and false alarms can be defined through a receiver operating characteristic curve built on historical data with simulated events and an acceptable EDS sensitivity can be defined from the receiver operating characteristic curve (McKenna et al., 2008).

Deployment of EDS tools in distribution networks has shown that changes in water quality at a monitoring location that are caused by changes in the hydraulic operations of the utility are often responsible for false positive alarms. As an example, examination of deployment of the EDS CANARY at the Greater Cincinnati Water Works (GCWW) over six month period has shown a number of instances where

Hart, D.B. and S.A. McKenna, 2009, CANARY User's Manual, Version 4.2, U.S. Environmental Protection Agency, Office of Research and Development, National Homeland Security Research Center, EPA 600/R-08/040A, 51 pp.



EPA
United States
Environmental Protection
Agency

EPA 600/R-08/040A | November 2009 | www.epa.gov/ord

CANARY User's Manual
VERSION 4.2

Office of Research and Development
National Homeland Security Research Center

McKenna, S.A., M. Wilson and K.A. Klise, 2008, Detecting Changes in Water Quality Data, American Water Works Association Journal, Vol. 100, No. 1, pp. 74-85.

Timely deployment of contaminant warning systems requires on-line sensors and advancement of data analysis and decision support systems to accurately detect water quality changes. As a demonstration of event detection in water quality data, three water quality change-detection algorithms were developed and used to detect changes in water quality observed at four locations within a distribution system. Each data set was "spiked" with simulated anomalous water quality values of 1 h duration and 10 levels of spike strength. The receiver operating characteristic (ROC) curve is proposed as an objective means of assessing and comparing water quality change-detection algorithm results that is also able to provide the tradeoff between missed detections and false alarms on a single plot. The area under the ROC curve provides a measure of the sensitivity and specificity of each algorithm for each spike strength and location. Resulting areas under the ROC curve range from 0.46 for spike strengths of 1.0 (background) to 0.98 for strengths of 3.5 standard deviations from the mean, where an ROC curve area of 1.0 indicates perfect detection.

Detecting Changes in Water Quality Data

BY SEAN A. MCKENNA,
MARK WILSON,
AND KATHERINE A. KLISE

Protection of drinking water systems from accidental and malevolent contamination events has taken on increased importance in recent years because of heightened security concerns (Roberson & Morley, 2005). A number of data sources, including customer complaints and public health monitoring, can be used to identify contamination in water distribution systems (Roberson & Morley, 2005). However, the primary data source for detecting contamination events is a contaminant warning system (CWS) composed of on-line sensors, connected supervisory control and data acquisition (SCADA) systems, data analysis algorithms, and decision-support tools (Hasan et al, 2004; Grayman et al, 2001). An important component of any CWS is the ability to rapidly and accurately detect a significant change in water quality as it occurs within a distribution system. In the absence of contaminant-specific on-line sensors, detection of a significant change must depend on the signals generated by existing on-line water quality sensors and the proper analysis of these signals. At its essence, change detection is a way to separate anomalous conditions from normal background operating conditions. Detection of a change does not necessarily imply that the cause of that change has also been identified.

To date, two general approaches to developing and testing event detection using water quality signals have been examined. First, laboratory and test-

McKenna, S.A. and D.B. Hart, 2008, On-Line Identification of Adverse Water Quality Events from Monitoring of Surrogate Data: CANARY Software, In Proceedings of: Singapore International Water Week, June 23rd-27th, Singapore

On-Line Identification of Adverse Water Quality Events from Monitoring of Surrogate Data: CANARY Software

Sean A. McKenna, David Hart***

*Sandia National Laboratories, P.O. Box 5800, MS-0751, Albuquerque, NM 87185-0751, U.S.A., (E-mail: samcken@sandia.gov)
**Sandia National Laboratories, P.O. Box 5800, MS-0751, Albuquerque, NM 87185-0751, U.S.A., (E-mail: dbhart@sandia.gov)

Abstract:
Real-time monitoring of water quality parameters in drinking water distribution networks is becoming commonplace and this monitoring provides information critical for the efficient and safe operation of these networks. Water quality monitors deployed to date are focused on measuring the general water quality within the network (e.g., residual chlorine levels, pH, electrical conductivity, etc.). Sensors that can provide reliable, real-time, in-situ monitoring of specific contaminants and pathogens are not yet deployed. Therefore, it is essential to use the currently available information as effectively as possible to identify significant changes in water quality that can be a sign of adverse conditions in the network.

Here we demonstrate an approach to automatic on-line detection of anomalous water quality events from surrogate monitoring data. The three steps in this approach are: 1) State estimation where statistical models applied to previous measurements of multiple water quality variables are used to predict the next measured value for each water quality variable; 2) Integration of the residuals between the predicted and measured values across all water quality variables as the measurements become available; and 3) Analysis of the residuals across consecutive time steps against an expected failure rate to determine the probability of an adverse water quality event.

This three-step process for on-line water quality event detection has been implemented in the CANARY software. Examples of off-line analysis of water quality data using CANARY allow for determination of false positive rates for different parameter settings in background water quality where no adverse water quality events have occurred. Additionally, water quality data containing simulated contamination events allows for the quantification of false positive and false negative results. The impact of different event detection operating parameters on the false positive rate and the number of missed detections is examined for two locations in a water distribution system.

Keywords Water quality, Event detection, Surrogate parameters, SCADA system

INTRODUCTION

Continuous, reliable delivery of safe drinking water to customers is an essential component of the viability of large metropolitan areas. The distribution networks used to deliver water represent a critical component of municipal infrastructure systems. Accidental or intentional contamination events that could degrade water quality within water distribution systems have focused recent discussion on various means of hardening both the physical and cyber components of these systems against contamination events. The concept of a contaminant warning system (CWS) has been proposed as an integrated tool that employs in-situ sensors, supervisory control and data acquisition (SCADA) systems, and water quality event detection systems (EDS) to continuously monitor network conditions and warn operations personnel of any potential contamination events (see Hasan, et al., 2004; Grayman, et al., 2001).

The sensor component of a CWS can be comprised of various water quality sensing platforms including contaminant-specific sensors that make use of recent developments in "*chem-lab on a microchip*" technologies, or more commonly, of existing water quality sensors (e.g., pH, Cl,

Koch, M.W., and S.A. McKenna, 2008, Distributed Network Fusion for Water Quality, in Proceedings of ASCE World Environmental and Water Resources Congress, Honolulu, Hawaii, May 13-16.

Distributed Network Fusion for Water Quality[+]

Mark W. Koch[1] , Sean A. McKenna[2]

[1] Sandia National Laboratories[*], P.O. Box 5800, MS 1163, Sensor Exploitation Applications Department, Albuquerque, NM; mwkoch@sandia.gov

[2] Sandia National Laboratories, P.O. Box 5800, MS 0735, Geohydrology Department, Albuquerque, NM; samcken@sandia.gov

Abstract

To protect drinking water systems, a contamination warning system can use in-line sensors to detect accidental and deliberate contamination. Currently, detection of an incident occurs when data from a single station detects an anomaly. This paper considers the possibility of combining data from multiple locations to reduce false alarms and help determine the contaminant's injection source and time. If we consider the location and time of individual detections as points resulting from a random space-time point process, we can use Kulldorff's scan test to find statistically significant clusters of detections. Using EPANET, we simulate a contaminant moving through a water network and detect significant clusters of events. We show these significant clusters can distinguish true events from random false alarms and the clusters help identify the time and source of the contaminant. Fusion results show reduced errors with only 25% more sensors needed over a nonfusion approach.

1. Introduction

To maintain the safety and security of drinking water, water utilities need innovative technologies to detect deliberate or accidental contamination in water distribution systems. One approach uses water quality sensors in the water distribution system and measures attributes of the water such as free chlorine, total organic carbon, pH, temperature, and electrical conductivity. While these measurements do not necessarily measure contaminant levels directly, a sudden change in their readings can indicate contamination or an abnormal operation of the water distribution system. One approach uses change detection algorithms to compare the current measurements with models of the background. We call each location with sensors and algorithms a *sensing-node* and a change in water quality detected by the algorithms an *event*.

In conjunction with the National Homeland Security Research Center we are extending research from detection at a single sensing-node to detection at multiple nodes distributed throughout the water distribution network. Here, we want to use the topology of the water distribution network and sensor fusion to combine multiple

---

[+] This work was funded by the U.S. EPA National Homeland Security Research Center (NHSRC)

[*] 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

McKenna, S.A., D.B. Hart, K.A. Klise, V.A. Cruz and M.P. Wilson, 2007, Event Detection from Water Quality Time Series, in proceedings of: ASCE World Environmental and Water Resources Congress, Tampa, FL, May, 2007, May 15-19th.

World Environmental and Water Resources Congress 2007: Restoring Our Natural Habitat

Event Detection from Water Quality Time Series

Sean A. McKenna[1], David Hart[1], Katherine Klise[1], Victoria Cruz[1] and Mark Wilson[2]

[1]Geohydrology Department, Sandia National Laboratories, PO Box 5800 MS 0735, Albuquerque, NM 87185-0735, ph: 505 844 2450; fax: 505 844 7354; em: samcken@sandia.gov

[2]GRAM Incorporated, 8500 Menaul Blvd., NE, Suite B-335, Albuquerque, New Mexico, 87112, em: mpwilso@graminc.com

Abstract

Detection of anomalous events in water distribution systems is of interest for both daily operations focused on delivery of high quality water as well as for identification of accidental or intentional contamination events. In lieu of network-wide deployment of in-situ contaminant-specific sensors, data streams resulting from in-situ monitoring of ambient water quality are employed as input to event detection algorithms to identify periods of anomalous water quality. The basis of these approaches is prediction of the future water quality values (state estimation) and then comparison of the prediction errors, the differences between predicted and measured water quality signals, to identify outliers in an on-line framework. These algorithms generally rely on a stationary time series and large, sudden changes within the time series make outlier detection difficult. Here we propose an approach to improving the identification of events, defined as a cluster of outliers, that will also identify changes in the baseline water quality. This approach is called the binomial event discriminator (BED) and it uses a failure model based on the binomial distribution to determine the probability of an event existing based on *r* outliers occurring within *n* time steps. If the consecutive number of outliers exceeds an upper limit, a change in the baseline water quality is declared. The BED is applied to observed water quality collected at a location within a utility distribution system. The BED is able to reduce the number of false positive event identifications by several orders of magnitude compared to not using the BED. The BED is also identifies two locations as baseline water quality changes.

Introduction

Water distribution systems represent a critical component of municipal infrastructure systems. Accidental or malevolent contamination events within water distribution systems have focused recent discussion on various means of hardening both the physical and cyber components of these systems against these types of events. Additionally, the rapid identification of contamination events using networks of in-situ water quality sensors and subsequent analysis algorithms has also received attention (Hasan, et al., 2004; Grayman, et al., 2001). The various components of a system designed to detect a contamination event are known as a contaminant warning system (CWS). The CWS must have the capability to rapidly and accurately detect a

McKenna, S.A., K.A. Klise and M.P. Wilson, 2006, Testing Water Quality Change Detection Algorithms, in Proceedings of the 8th Annual Water Distribution System Analysis Symposium, Cincinnati, OH, August 27-30, 2006.

8th Annual Water Distribution Systems Analysis Symposium, Cincinnati, Ohio, USA, August 27-30, 2006

TESTING WATER QUALITY CHANGE DETECTION ALGORITHMS

Sean A. McKenna[1], Katherine A. Klise[1] and Mark P. Wilson[2]
[1]Sandia National Laboratories
Albuquerque, New Mexico
samcken@sandia.gov, kaklise@sandia.gov

[2]GRAM Incorporated
Albuquerque, New Mexico
mpwilso@graminc.com

Abstract

Rapid detection of anomalous operating conditions within a water distribution network is desirable for the protection of the network against both accidental and malevolent contamination events. In the absence of a suite of in-situ, real-time sensors that can accurately identify a wide range of contaminants, we focus on detecting changes in water quality through analysis of existing data streams from in-situ water quality sensors. Three different change detection algorithms are tested: time series increments, linear filter and multivariate distance. Each of these three algorithms uses previous observations of the water quality to predict future water quality values. Large deviations between the predicted or previously measured values and observed values at future times indicate a change in the expected water quality. The definition of what constitutes a large deviation is quantified by a threshold value applied to the observed differences.

Both simulated time series of water quality as well as measured chlorine residual values from two different locations within a distribution network are used as the background water quality values. The simulated time series are created specifically to challenge the change detection algorithms with bimodally distributed water quality values having a square wave and sin wave time series, with and without correlated noise. Additionally, a simulated time series resembling observed water quality time series is created with different levels of variability. The algorithms are tested in two different ways. First, background water quality without any anomalous events are used to test the ability of each algorithm to identify the water quality value at the next time step. Summary statistics on the prediction errors as well as the number of false positive detections quantify the ability of each algorithm to predict the background water quality. The performance of the algorithms with respect to limiting false positives is also compared against a simpler "set point" approach to detecting water quality changes. The second mode of testing employs events in the form of square waves superimposed on top of modeled/measured background water quality data. Three different event strengths are examined and the event detection capabilities of each algorithm are evaluated through the use of receiver operating characteristic (ROC) curves. The area under the ROC curve provides a quantitative basis of comparison across the three algorithms. Results show that the multivariate algorithm produces the lowest prediction errors for all cases of background water quality. A comparison of the number of false positives reported from the change detection algorithms and a set point approach highlights the efficiency of the change detection algorithms. Across all three algorithms, most prediction errors are within one standard deviation of the mean water quality. The event detection results show that the best performing algorithm varies across different background water quality models and simulated event strength.

Keywords

Contaminant warning system, Water quality, Change detection, ROC curves

Klise, K.A. and S.A. McKenna, 2006, Multivariate Applications for Detecting Anomalous Water Quality, in Proceedings of the 8th Annual Water Distribution System Analysis Symposium, Cincinnati, OH, August 27-30, 2006.

MULTIVARIATE APPLICATIONS FOR DETECTING ANOMALOUS WATER QUALITY

Katherine A. Klise and Sean A. McKenna
Geohydrology Department, Sandia National Laboratories
Albuquerque, New Mexico
kaklise@sandia.gov, samcken@sandia.gov

Abstract

The ability to detect deliberate or accidental contamination of a water distribution system is of real concern to the safety and security of our nation's drinking water. To address these concerns, increased attention has been placed on sophisticated monitoring of water distribution systems and the use of robust statistical analysis. Using existing data from in-situ water quality sensors, this paper explores the ability to detect anomalies in water quality using multivariate techniques. The algorithm developed in this study uses a multivariate distance measure between the current water quality measurement and the closest observation in multivariate space within a moving window of previous observations. To discriminate between normal and anomalous water quality, the distance measure is compared to a constant threshold. To test the algorithm, we utilize both simulated anomalous events and laboratory based events that correspond to real contaminants. These events are superimposed onto in-situ water quality recorded at four different locations within a single utility network. Measured water quality parameters include free chlorine, pH, temperature and electrical conductivity. Robust discrimination methods have a high probability of detecting anomalies with a low false alarm rate. Here, receiver operating characteristic (ROC) curves are used to test the ability of the multivariate classification algorithm to detect anomalous water quality while keeping false alarms low. This analysis explores the false alarm rate associated with detecting a range of anomalous water quality observations.

Keywords
Contaminant warning system, Homeland security, Drinking water, Change detection, ROC curves

1. Introduction

In light of growing concern over the safety and security of our nation's drinking water, increased attention has been focused on advanced monitoring of water distribution systems. The ability to detect anomalous changes in water quality has application to both maintaining normal operations of water distribution systems as well as detecting anomalous events as a result of accidental or malicious contamination within the system. With increased interest in contaminant warning systems (Grayman et al., 2001), efforts must focus on protecting public health while minimizing public concern (Hasan et al., 2004). Developing accurate contaminant warning systems is a critical step to correctly initiate a response if anomalous water quality is detected.

Under an interagency agreement between the U.S. EPA National Homeland Security Research Center and Sandia National Laboratories, this project aims to develop automated monitoring techniques that will classify real time water quality data to sound an alarm when anomalous water types are detected. Using all available water quality sensor data, the algorithms calculate the multivariate distance between the current water quality measurement and the closest observation in multivariate space within a moving window of previous measurements. Expected water quality is defined as the closest observation in multivariate space within a window of previous measurements. ROC curves define the ability to discriminate between normal and anomalous water quality using a variety of thresholds and anomalous events. Simulated events are first added to the data to track the detection ability over a variety of anomalous water quality deviations superimposed onto the measured data. Simulated events cover a range of event magnitude, duration, and a range of rapid to gradual changes from measured data. Next, laboratory-based events are superimposed onto the data. These laboratory events track changes to free

McKenna, S.A., D.B. Hart and L. Yarrington, 2006, Impact of sensor detection limits on protecting water distribution systems from contamination events, *Journal of Water Resources Planning and Management*, Special Issue on Drinking Water Distribution Systems Security, 132 (4), pp. 305-309.

Impact of Sensor Detection Limits on Protecting Water Distribution Systems from Contamination Events

Sean A. McKenna[1]; David B. Hart[2]; and Lane Yarrington[3]

Abstract: Real-time water quality sensors are becoming more commonplace in water distribution systems. However, field deployable, contaminant-specific sensors are still in the development stage. As development proceeds, the necessary operating parameters of these sensors must be determined to protect consumers from accidental and malevolent contamination events. This objective can be quantified in several different ways including minimization of: the time necessary to detect a contamination event, the population exposed to contaminated water, the extent of the contamination within the network, and others. We examine the ability of a sensor set to meet these objectives as a function of both the detection limit of the sensors and the number of sensors in the network. A moderately sized distribution network is used as an example and different sized sets of randomly placed sensors are considered. For each combination of a certain number of sensors and a detection limit, the mean values of the different objectives across multiple random sensor placements are calculated. The tradeoff between the necessary detection limit in a sensor and the number of sensors is evaluated. Results show that for the example problem examined here, a sensor detection limit of 0.01 of the average source concentration is adequate for maximum protection. Detection of events is dependent on the detection limit of the sensors, but for those events that are detected, the values of the performance measures are not a function of the sensor detection limit. The results of replacing a single sensor in a network with a sensor having a much lower detection limit show that while this replacement can improve results, the majority of the additional events detected had performance measures of relatively low consequence.

DOI: XXXX

CE Database subject headings: Hydraulic networks; Sensors; Hydraulic models; Water distribution systems; Water pollution.

Introduction

The concept of a contaminant warning system (CWS) that can be deployed within a water distribution network to give rapid and accurate indication of conditions within the distribution system that may be adverse to human health has recently gained interest (ILSI 1999; Grayman et al. 2001). The majority of effort in CWS research has concentrated on the development of contaminant sensor technologies; however, as pointed out by Hasan et al. (2004), an effective CWS must also include analysis of sensor data towards making decisions that protect public health while minimizing community concern.

Considerable CWS research has been done to determine the optimal locations for sensors within a distribution network to provide for the greatest protection of human health [see Ostfeld and Salomons (2004); Uber et al. (2004); Watson et al. (2004); Berry et al. (2006)]. The majority of these sensor location optimization studies have focused on algorithm development and, to avoid other complications, have typically assumed a "perfect sensor." This assumption states that any amount of contamination reaching the location of the sensor will cause the sensor to correctly indicate a positive contamination event, or in the case of quantitative measurements, to accurately measure the contaminant concentration at that location. Under the perfect sensor assumption, issues associated with sensors deployed in actual distribution networks such as: detection limits, accuracy and precision of the sensor, false positive and false negative readings, the ability of the sensor to integrate a contaminant level within a volume of water, and the robustness of the sensor readings against calibration drift and changes in the distribution system operating conditions are not considered.

This work begins to examine the effects of varying sensor performance on the ability of a CWS to mitigate adverse consequences from a contamination event within a water distribution system. This goal is accomplished by injecting a contaminant at a single node in a distribution network and placing sensors at nodes to detect this contaminant. Sensor locations are varied randomly and different numbers of sensors are examined. In each case, the perfect sensor assumption is relaxed by varying the detection limit relative to the average injection concentration.

Problem Formulation

For a contaminant injected into the system at a network node (junction), the average source concentration, C_s^*, over the period of the contaminant source injection, T_s, is

---

[1] Geohydrology Dept., Sandia National Laboratories, P.O. Box 5800, MS 0735, Albuquerque, NM 87185-0735 (corresponding author). E-mail: samcken@sandia.gov

[2] Geohydrology Dept., Sandia National Laboratories, P.O. Box 5800, MS 0735, Albuquerque, NM 87185-0735. E-mail: dbhart@sandia.gov

[3] Geohydrology Dept., Sandia National Laboratories, P.O. Box 5800, MS 0735, Albuquerque, NM 87185-0735. E-mail: lyarrin@sandia.gov

Note. Discussion open until December 1, 2006. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this technical note was submitted for review and possible publication on August 23, 2005; approved on December 30, 2005. This technical note is part of the ***Journal of Water Resources Planning and Management***, Vol. 132, No. 4, July 1, 2006. ©ASCE, ISSN 0733-9496/2006/4-1–XXXX/$25.00.

Klise, K.A. and S.A. McKenna, 2006, Water quality change detection: multivariate algorithms, in Proceedings of SPIE (International Society for Optical Engineering), Defense and Security Symposium 2006, April 18-20, Orlando. Florida, 9pp.

Water quality change detection: multivariate algorithms

Katherine A. Klise[*] and Sean A. McKenna
Geohydrology Department, Sandia National Laboratories, PO Box 5800, MS 0735,
Albuquerque, NM 87185

ABSTRACT

In light of growing concern over the safety and security of our nation's drinking water, increased attention has been focused on advanced monitoring of water distribution systems. The key to these advanced monitoring systems lies in the combination of real time data and robust statistical analysis. Currently available data streams from sensors provide near real time information on water quality. Combining these data streams with change detection algorithms, this project aims to develop automated monitoring techniques that will classify real time data and denote anomalous water types. Here, water quality data in 1 hour increments over 3000 hours at 4 locations are used to test multivariate algorithms to detect anomalous water quality events. The algorithms use all available water quality sensors to measure deviation from expected water quality. Simulated anomalous water quality events are added to the measured data to test three approaches to measure this deviation. These approaches include multivariate distance measures to 1) the previous observation, 2) the closest observation in multivariate space, and 3) the closest cluster of previous water quality observations. Clusters are established using kmeans classification. Each approach uses a moving window of previous water quality measurements to classify the current measurement as normal or anomalous. Receiver Operating Characteristic (ROC) curves test the ability of each approach to discriminate between normal and anomalous water quality using a variety of thresholds and simulated anomalous events. These analyses result in a better understanding of the deviation from normal water quality that is necessary to sound an alarm.

1. INTRODUCTION

The ability to detect anomalous changes in water quality has application to both maintaining normal operations of water distribution systems as well as detecting anomalous events as a result of accidental or malicious contamination within the system. To address the concerns of the safety and security of our nation's drinking water, enhanced monitoring of water distribution systems is vital. Contaminant warning systems are critical to correctly initiate a response if anomalous water quality is detected. Contaminant warning systems are gaining interest[1], and efforts must focus on protecting public health while minimizing public concern[2]. To do this, water quality monitoring needs to integrate real time data and robust statistical analysis.

Under an interagency agreement between the U.S. EPA National Homeland Security Research Center and Sandia National Laboratories, this project aims to develop automated monitoring techniques that will classify real time data to sound an alarm when anomalous water types are detected. Using all available water quality sensor data, the algorithms calculate the multivariate distance between the current water quality measurement and expected water quality within a moving window of previous measurements. Three approaches to defining the distance to expected water quality are tested. These approaches include multivariate distance measures to 1) the previous observation, 2) the closest observation in multivariate space, and 3) the closest cluster of previous water quality observations. ROC curves define the ability of each approach to discriminate between normal and anomalous water quality using a variety of thresholds and simulated anomalous events. Water quality data in 1 hour increments collected from sensors at 4 locations within a single utility are used for the study. Simulated events are added to the data to test the ability of each approach to find a variety of anomalous water quality deviations within the measured data.

The purpose of this study is to develop a change detection algorithm and assess its ability to discriminate normal water quality and anomalous water quality while tracking the probability of detection and the false alarm rate. This study details the specific aspects of multivariate algorithms, and uses measured and simulated data to test the reliability of such methods. The analysis results in a better understanding of the magnitude and type of deviation from normal water quality that is needed to alert a utility of uncertain water quality.

2. WATER QUALITY DATA

Evaluating change detection algorithms requires a set of known anomalous events against which the algorithm is tested. Measured water quality data streams obtained from in-line utility sensors represent normal operating conditions.

Sandia National Laboratories is a multi-program laboratory operated by
Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation,
for the U.S. Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000. SAND2010-2228P. Designed by the Sandia Creative
Group. (505) 284-3181. SP•144947•04/10

National Nuclear Security Administration