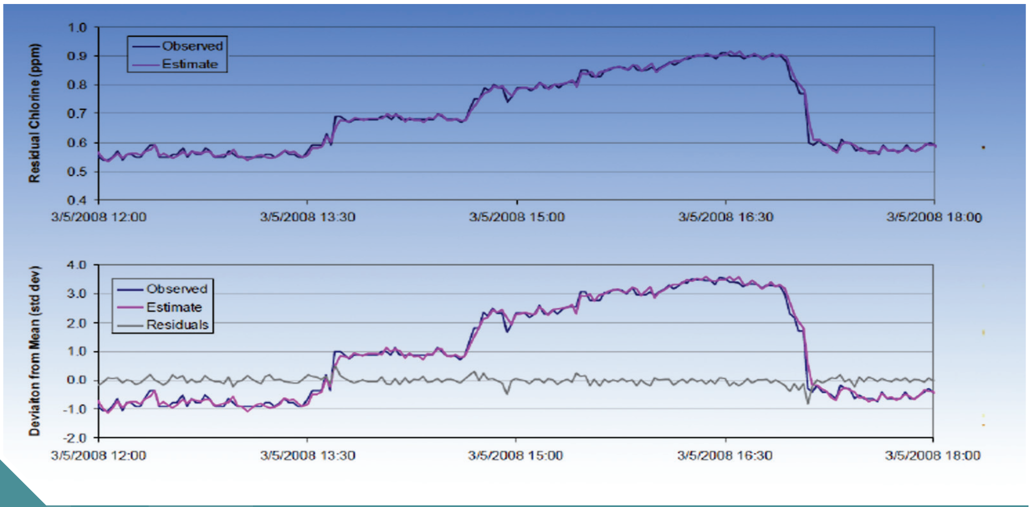
United States
Environmental Protection
Agency

EPA 600/R-14/254 | September 2014 | www.epa.gov/research

# Configuring Online Monitoring Event Detection Systems



**Office of Research and Development**
National Homeland Security Research Center

United States
Environmental Protection Agency

EPA 600/R-14/254 | September 2014 | www.epa.gov/research

# Configuring Online Monitoring Event Detection Systems

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

CINCINNATI, OHIO 45268

## Executive Summary

The CANARY event detection software was developed to enhance the detection of contamination in drinking water systems. Working in conjunction with a network of water quality sensors placed strategically throughout a water distribution system, CANARY increases the likelihood and speed of detection by interpreting sensor data in real time, identifying anomalies, and alerting the operator when a contaminant might be present. CANARY has been adopted by several water utilities around the world to help continuously monitor their water quality. One barrier to more widespread use of CANARY has been the lack of guidance on how to configure the software. This report presents a logical process for configuring CANARY, as well as a set of rule-of-thumb configuration parameters that can be used by water utilities as they begin implementing CANARY.

In order to develop reliable rule-of-thumb configuration parameters, this report analyzes sensor data from two real world water systems. Eight months of data from five water quality sensor stations is studied, four historical datasets from the Singapore Public Utility Board and one from Greater Cincinnati Water Works (monitored at EPA's Testing and Evaluation Facility), with data from the latter station containing laboratory-controlled contamination events. Multiple configuration parameter combinations are evaluated in order to demonstrate the effect of each parameter on CANARY's performance, and to determine the most useful rule-of-thumb parameters. Results using sensor data from Greater Cincinnati Water Works demonstrate the ability to configure CANARY to detect 100% of true contamination events, while reducing the false alarm rate to below one alarm per week. Results based on data from the Singapore Public Utility Board demonstrated that the alarm rate can be reduced to below one alarm per day.

This report focuses on four configuration parameters: the *history window*, the *binomial event discriminator* (*BED) window*, the *outlier threshold* and the *event threshold*. The *history window* is the number of historical data points used to calculate the baseline variability of a water quality signal. Water quality signals are time series of data produced by sensors measuring, for example, free chlorine, electrical conductivity, or total organic carbon. The *outlier threshold* is the number of standard deviations away from the mean a data point must be in order to be declared an outlier. The *BED window* is the number of historical data points examined to look for the onset of water quality events, and is typically a subset of the *history window*. The *event threshold* is the value of probability that must be exceeded in order for a group of outliers to be declared as a water quality event.

These parameters are important because they determine CANARY's performance in terms of detection sensitivity (i.e., the proportion of events detected) and specificity (i.e., proportion of non-events that are correctly identified). As the analysis shows, changing the values of the configuration parameters resulted in general trends across all five stations. Increasing either the *BED window* or the *outlier threshold* parameters reduced false alarm rates; however, these parameter changes also decreased the number of true events detected. *History window* parameter values that correspond to 1.5 or 2 days generally reduced the number of alarms, while lower values increased alarms and higher values did not change results significantly. The number and type of signals also impacts results: removing certain water quality signals from analyses, specifically turbidity, resulted in fewer alarms in all stations. Overall, alarm rates were most sensitive to the *outlier threshold* parameter.

Based on the analysis conducted in this report, the configuration parameter values presented in Table 1 are recommended as a starting point for most water utilities. The recommendations vary depending on the frequency at which water quality sensor data is recorded; for online monitoring, recording data at least every 5 minutes is recommended. This frequency is labelled the *data interval* in the CANARY software. A 2-day *history window* is recommended as it captures a baseline behavior that encompasses most day-to-

day activity. An *outlier threshold* of 1.15 treats 75% of sensor data variability as normal behavior, which allows some noise to be accepted as normal. A *BED window* of 15 (for a 2-minute *data interval*) and *event threshold* of 0.90 is recommended; this combination means that 2/3 of the data points over the previous 30 minutes must be outliers in order for CANARY to alert that a possible event has occurred. This helps to ensure that a single outlier does not result in a CANARY detection and reduces the number of false alarms. These parameters can capture events as short as 20 minutes, or those that onset over a 30 minute period, and can be detected within 20 minutes after the initial deviation from normal behavior is picked up by sensors. These parameters should be adjusted in tandem if the detection of shorter events is desired. Similarly, a *BED window* of 12 and *event threshold* of 0.90 is recommended for a 5-minute *data interval*, which means that 2/3 of the data points over the previous 60 minutes must be outliers for CANARY to detect an event.

**Table 1: Recommended Configuration Parameter Values**

| **Parameter** | **Recommended Values** | |
|---|---|---|
| *Data interval* | 2 minutes | 5 minutes |
| *History window* | 1440 data points | 576 data points |
| *Outlier threshold* | 1.15 | 1.15 |
| *BED window* | 15 data points | 12 data points |
| *Event threshold* | 0.90 | 0.90 |

*BED, binomial event discriminator*

The goal of this report is to help users develop a more intuitive understanding of the role of these important parameters, and to show how these recommended parameter values were derived from a study of two water systems. The starting configuration shown above performs well for most water utilities. However, a simplified approach to optimizing parameter values shows how to improve this rule-of-thumb configuration, or optimize the configuration, at a specific location within a water utility's distribution system. This helps to fine-tune a parameter set for specific sensor locations, or enable the utility to detect longer or shorter water quality events.

In most cases, the configuration parameter values do not need to be reconfigured for each new season. CANARY automatically recalculates the baseline variability of water quality parameters at each timestep using data from the *history window*; thus, it already incorporates changes over time. If there is an abrupt change in water quality due to a change in seasons, CANARY might produce an alarm, but should quickly stop producing an alarm as it adjusts to the new seasonal values. Sensor station locations play a large role in the number of false alarms produced by CANARY: sensors near tanks or pumps that witness large variability produce more false alarms than locations further from such facilities. Utilizing additional features within CANARY, such as the set point algorithms or the *precision* and *valid range* configuration parameters, can also help to reduce false positives and increase the detection rate. Finally, using highly reliable sensors, and properly calibrating and maintaining sensors reduces false alarm rates.

## Acknowledgements

The National Homeland Security Research Center would like to acknowledge the following organizations and individuals for their support in the development and/or technical review of this report, or for sharing data:

EPA Office of Research and Development – National Homeland Security Research Center (NHSRC)
Jonathan Burkhardt (ORISE Fellow)
John Hall
Terra Haxton
Regan Murray
Ruben Kertesz (ORISE Fellow)

EPA Office of Water – Water Security Division
Katie Umberg

Greater Cincinnati Water Works

Sandia National Laboratories
David Hart
Katherine Klise

Singapore Public Utility Board

Questions concerning this document or its application should be addressed to:

Regan Murray
USEPA/NHSRC (NG 16)
26 W Martin Luther King Drive
Cincinnati OH 45268
(513) 569-7031
Murray.Regan@epa.gov

## Disclaimer

The U.S. Environmental Protection Agency (EPA) through its Office of Research and Development funded and collaborated in the research described here under Inter-Agency Agreements with Oak Ridge National Laboratory (Oak Ridge Institute for Science and Education [ORISE] Program) and with the Department of Energy's Sandia National Laboratories. This document has been subjected to the Agency's review and has been approved for publication. Note that approval does not signify that the contents necessarily reflect the views of EPA. EPA does not endorse the purchase or sale of any commercial products or services.

## Contents

## List of Tables

## List of Figures

## List of Acronyms and Abbreviations

| | |
|---|---|
| BED | binomial event discriminator |
| EDS | event detection system |
| EPA | U.S. Environmental Protection Agency |
| GCWW | Greater Cincinnati Water Works |
| LPCF | linear prediction coefficient filter |
| µS/cm | microsiemens per centimeter |
| mg/L | milligram per liter |
| MVNN | multivariate nearest neighbor |
| NHSRC | National Homeland Security Research Center |
| NTU | nephelometric turbidity unit |
| ORISE | Oak Ridge Institute for Science and Education |
| ORP | oxidation reduction potential |
| ppm | parts per million |
| PUB | Singapore Public Utility Board |
| SCADA | supervisory control and data acquisition |
| T&E | EPA Testing and Evaluation Facility |
| TOC | total organic carbon |
| U.S. | United States |
| WDS | Water distribution system |

# 1.0 Introduction

The CANARY event detection software was developed to enhance the detection of contamination in drinking water systems. Working in conjunction with a network of water quality sensors placed strategically throughout a water distribution system, CANARY increases the likelihood and speed of detection by interpreting sensor data in real time, identifying anomalies, and alerting the operator when a contaminant might be present. CANARY has been adopted by several water utilities around the world to help continuously monitor their water quality. One barrier to more widespread use of CANARY has been the lack of guidance on how to configure the software. This report presents a logical process for configuring CANARY, as well as a set of rule-of-thumb configuration parameters that can be used by water utilities as they begin implementing CANARY.

## 1.1 Background

The security of the water infrastructure of the United States (U.S.) gained increased awareness after the events of September 11, 2001. The U.S. Environmental Protection Agency (EPA), as the lead federal agency for water security, helped develop tools, procedures and documentation to support water utilities and other agencies in protecting the water supply. A primary focus of this effort has been to develop and to demonstrate components of drinking water contamination warning systems–monitoring and surveillance systems that can detect contamination in time to allow for mitigation of human health and economic consequences (U.S. EPA, 2008).

As part of the contamination warning system development process, EPA has worked closely with the Greater Cincinnati Water Works (GCWW) and other water utilities to test the ability of continuous sensors to detect a wide range of contaminants (Pickard et al., 2011; Allgeier et al., 2011a; Hall et al., 2007; Szabo et al., 2008; Hall and Szabo, 2010; Hall et al., 2009 and U.S. EPA, 2012a). This work has shown that commercially available water quality sensors, such as for electrical conductivity, free chlorine, and total organic carbon can be used to indirectly detect the presence of contaminants in water (Hall et al., 2007; Szabo et al., 2008; Hall and Szabo, 2010; Hall et al., 2009 and U.S. EPA, 2012a). As water quality sensors detect changes in water quality but not the presence of specific contaminants, an automated data analysis tool, or event detection system (EDS), is needed to determine when such water quality changes indicate a potential contamination event.

In 2003, EPA and Sandia National Laboratories began to work in partnership with the American Water Works Association and member utilities to investigate the ability of EDSs to automate the indirect detection of contamination using water quality sensors (Morley et al., 2007; Murray et al., 2010). EDSs read in sensor data in real time and use statistical and data mining techniques to identify anomalous patterns indicative of contamination events. The result of this research was the CANARY Event Detection Software, a freely available EDS (Murray et al., 2010, and Hart and McKenna, 2012). CANARY was developed using water quality data from several U.S. water utilities and was also tested for its ability to detect real contamination events by using data from laboratory-controlled experiments involving more than 20 chemical and biological contaminants. In addition, the software has been piloted over several years in real time at multiple water utilities, including GCWW and the Singapore Public Utility Board.

CANARY was designed to work with data from any type (or brand) of sensor by interfacing with databases or data files rather than with the sensor hardware. CANARY can communicate directly with supervisory control and data acquisition (SCADA) databases for easy integration into any water utility system. This provides flexibility to the water utilities or other end users by allowing any type of filed data to be used for analysis. In addition to the built-in event detection algorithms, CANARY can also be expanded with user-developed algorithms (Hart and McKenna, 2012). Algorithms within CANARY function as more than fixed set-point alarms; they can also detect anomalous behavior within the range of normal operating values. CANARY uses a binomial event discriminator (BED), a moving time period over which signal data is examined to look for the onset of events, to require that multiple timesteps be anomalous before establishing that an event has occurred or is occurring (Hart and McKenna, 2012).

CANARY was recently used in the "Water Quality Event Detection Challenge" to compare its ability to detect simulated events against other available EDSs (U.S. EPA, 2013a). CANARY performed well overall and was able to detect 70% of the simulated events in data from six separate monitoring stations at several water utilities, each consisting of multiple water quality sensors; in two stations, CANARY was able to detect 86% or more of the simulated events (U.S. EPA, 2013a). CANARY and one other EDS attempted to analyze data from all six stations. Both detected 70% of known events; however, CANARY produced fewer total invalid alarms for all six stations (U.S. EPA, 2013a). This highlights CANARY's ability to perform well across a wide variety of water quality data from multiple water utilities. See Appendix C for more information.

Today, the software is available for free from EPA's website and can be used by any water utility to detect anomalous water quality events (U.S. EPA, 2012c). Although its intended use is for water security, and it has been tested primarily on water quality data, CANARY can be used to identify events in any time-based data stream (see for example, Kertesz et al., 2014).

## 1.2 CANARY Configuration

In order to use CANARY at a specific water quality monitoring station, the software needs to be configured. Configuration parameters are set within a configuration, or input, file that defines the features of the data being analyzed and sets the parameters for the event detection algorithms. A complete account of all the elements that are required in a configuration file can be found within the CANARY User's Manual (U.S. EPA, 2012a).

The configuration file specifies details about how CANARY is run (real time vs. batch mode), where to find the sensor data (in a database or spreadsheet), what sensor data to include (the time period and the *data interval* of the data to be analyzed), which water quality sensor signals to include (chlorine, pH, etc.) and where to find them in the file or database, which algorithms to use for event detection, and additional details for each monitoring station (see CANARY User's Manual for further details (U.S. EPA, 2012a)). Much of this information is straightforward to complete in the input file; however, the choice of algorithms and the associated parameters can be a more difficult task. These parameters are important because they determine the performance of the EDS in terms of detection sensitivity and specificity. The detection sensitivity is the proportion of events detected, or the true positive rate. The specificity is the proportion of non-events that are correctly identified, or the true negative rate, which is complementary to the false positive rate.

The selection of the algorithm configuration parameters is often referred to as training or optimizing the EDS and it requires analysis of historical data as well as user judgment. CANARY does not include a machine learning capability in which the configuration parameters are automatically updated and improved by the software over time; instead, users must specify values in the input file. Four configuration parameters listed below are particularly important for users to select carefully. (Note that, for clarity, configuration parameter names appear in italics throughout this document.) Another configuration parameter, *data interval*, is important for understanding the real meaning of several of these parameters. The *data interval* parameter sets how frequently CANARY expects new signal data (e.g., every two minutes). This value is usually the frequency at which a SCADA system or other database records sensor data.

While discussed in more detail in the next section, the four critical configuration parameters are:

- *history window* – the number of historical data points used to calculate the baseline variability of a water quality signal.
- *outlier threshold* – the number of standard deviations away from the mean baseline variability of the water quality signal data must be in order to be declared an outlier.
- *BED window* – the number of data points over which signal data is examined for the onset of events.
- *event threshold* – the value of probability that must be exceeded in order for a group of outliers to be considered an event.

Of the parameters listed above, *history window* and *BED window* are reasonably easy to understand in terms of real world timeframes. If these two parameters are multiplied by the *data interval*, they represent windows of historical data that move over time as CANARY analyzes data. The statistics involved in defining the *event threshold* and *outlier threshold* are more difficult to understand intuitively. A systematic approach to selecting these parameter values has been used in the past (Murray et al., 2010; Rosen and Bartrand, 2013) in which several values of each parameter is evaluated against a set of historical data. Since most historical data does not contain true contamination events, the optimal parameter values for each sensor station, then, are often selected to be the ones that minimize false positive rates; this approach can have the unintended effect of reducing detection rates of true events. True contamination events can be uncommon, but water events that are the result of other disruptions within a water distribution system are quite common (Hagar, 2013; also see Appendix B); however, unless the timing of these events is known, it is difficult to use such data to determine true detection rates. Without any guidance on good starting points, a user might begin with tens or hundreds of different values for each parameter, resulting in thousands of CANARY runs, and utilizing weeks of computer time. This approach can be daunting to new users and might not be the best use of water utility staff time or computer resources.

Fortunately, application of CANARY to the pilot cities has resulted in many lessons learned about the configuration process that can provide some practical starting points for new CANARY users. For example, a *history window* of about two days captures normal day-to-day signal behavior. A shorter *history window* likely increases the false alarm rate, and although increasing the *history window* might, in some cases, result in fewer false alarms, changing the other three parameters has a greater impact. The goal of this report is to help users develop a more intuitive understanding of the role of these important parameters as well as to develop rule-of-thumb guidance for the selection of these four parameter values. This is accomplished by analyzing data from both pilot cities and comparing performance results for multiple parameter values.

## 1.3 Overview of Report

The rest of this report is organized as follows: in section 2, an overview of the configuration parameters is presented to provide a more intuitive understanding; in section 3, the data and methods used in this report are presented; in section 4, the results of the data analysis and testing are reported; in section 5, a discussion compares the results for the two pilot cities and selects the recommended rule-of-thumb parameter values; in section 6, a simplified optimization protocol is presented and applied to a subset of the data.  Finally, an example CANARY output is provided in Appendix A, a discussion of what constitutes an event is in Appendix B and an application of the simplified optimization approach to the Water Quality Event Detection Challenge data is provided in Appendix C.  Appendix D contains full tabulated results for analyses presented in section 4.

## 2.0 An Overview of the Configuration Parameters

From a user's perspective, the configuration process begins by collecting information about the water quality time series data to be analyzed, the types of sensor signals and their variability at each location, and the type and duration of events that the user would like to be able to detect. These factors can be translated into the statistical configuration parameter values: the *history window*, the *BED window*, the *outlier threshold* and the *event threshold.*

### 2.1 The Four Parameters

When performing event detection with the most commonly used algorithm, linear prediction coefficient filter (LPCF), CANARY uses historical water quality signal data to predict the signal value for the next timestep. This predicted signal value is compared to the actual reported signal value, and the difference between the two is calculated and called the residual. Figure 1 shows the observed and predicted residual chlorine concentrations at a specific sensor location over a 6-hour period in the top plot. In the bottom plot, the concentrations have been normalized to their mean value over a given *history window*, and the residual, or the difference between the two, is also shown.

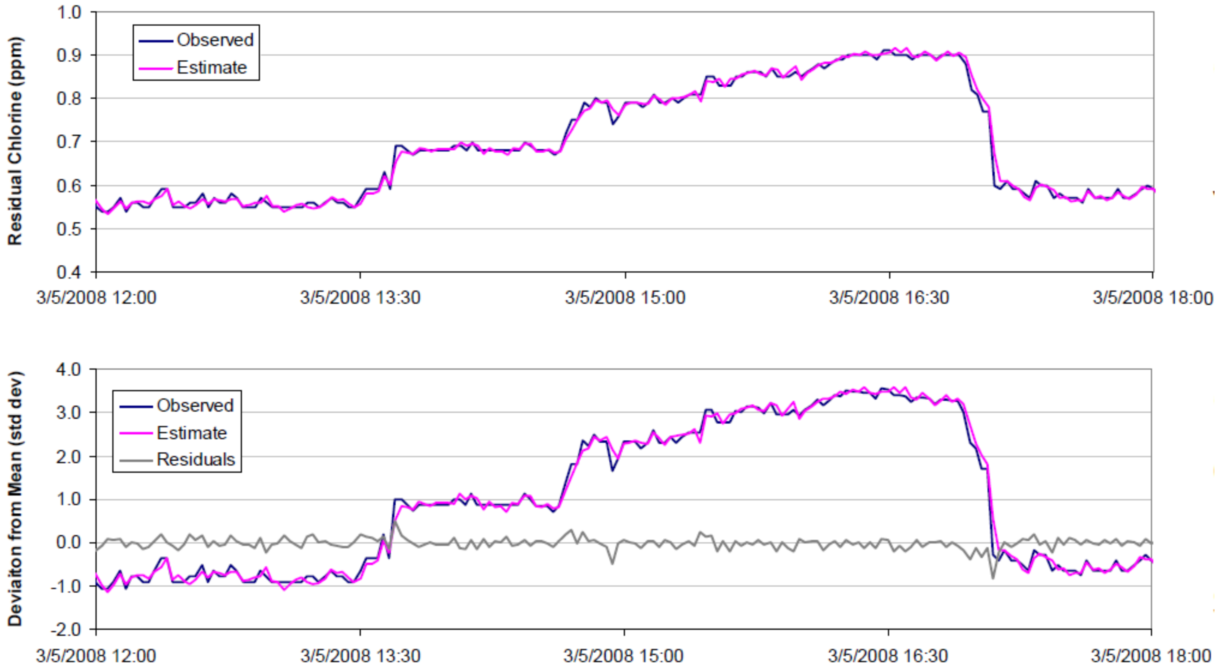


**Figure 1: Observed (blue) and predicted (pink) chlorine concentrations over time, and normalized observed, predicted and residual (grey) concentrations over time.**

Configuration parameters define how CANARY interprets a calculated residual. The *history window* and the *outlier threshold* determine when the residual is large enough to label the signal value an outlier, or an anomalous data point. CANARY uses the binomial event discriminator (BED) and an associated probability distribution function to determine if enough outliers have occurred to declare an event. The two key parameters that describe the probability distribution within CANARY are the *BED window* parameter and the *event threshold*. The parameter *history window* also affects how CANARY determines when an event is occurring. The parameter *data interval* is relevant to all of the time dependent parameters.

For a user, the first parameter to define is the *data interval* of the data being analyzed. In most cases, this is defined by the fixed frequency (in minutes) at which the SCADA system polls the sensors at a given station. For example, if a SCADA system records data from chlorine, pH, and TOC sensors every 2 minutes, the *data interval* is set to 2. This *data interval* can be used to understand the real timeframes associated with the *history window* and *BED window* parameters. Each of those CANARY parameter values is measured in terms of the number of data points; when multiplied by the *data interval* value, their values in real time can be understood. For example, if the *data interval* were 2 minutes and the *history window* were 1440 data points, then the *history window* would be equivalent to two days (2880 minutes). If the *data interval* were 5 minutes and the *history window* were 1440 data points, then the *history window* would be equivalent to five days (7200 minutes).

Previous work has shown that EDSs perform better when sensor data is available at shorter *data intervals* – specifically, five minutes or shorter (U.S. EPA, 2013a; Allgeier et al., 2011b). Additionally, a shorter *data interval* relates to more overall data, which in turn provides more flexibility in parameter selection. For example, with a *data interval* of 1-hour, if an event were to last for 90 minutes, only one or two data points might be outliers. However, with a 2-minute *data interval*, multiple (from 2-45) outliers can be grouped using the *BED window* to decide whether they are significant enough to indicate an event. In this way, longer *data intervals* require that a smaller group of outliers have a higher statistical significance in order to detect contamination events that occur over a short timeframe.

> ***history window***: This parameter is the number of historical data points used to determine the normal or baseline variability of a water quality signal. The *history window* establishes a moving time frame over which the mean and standard deviation of the signal data are calculated. This parameter is multiplied by *data interval* to establish the corresponding real world length of time. For example, a *history window* of 1080 data points is equivalent to 1.5 days when using a *data interval* of 2 minutes (i.e., 1080 data points × 2 minutes = 2160 minutes or 36 hours or 1.5 days).

The *history window* helps CANARY define the normal, baseline behavior for a signal. As stated above, the *history window* is a moving time period over which the mean and standard deviation of the signal data are calculated. Most water utility locations experience consistent day-to-day variability and benefit from a *history window* of 1.5 to 2 days (see the results in section 4). By selecting a *history window* larger than the typical diurnal variability of the system, CANARY is better able to predict the baseline behavior of the signal and therefore, able to distinguish normal behavior from anomalous behavior. As the *history window* moves in time, the baseline is recalculated, allowing CANARY to automatically adapt to changing conditions over time. Figure 2 shows a two day snapshot of an example sensor signal and how the *BED window* (green filled box) and *history window* (blue outline) change as CANARY analyzes the signal; the upper figure represents time $t_1$, and the lower figure shows how these windows have shifted for the analysis of time $t_2$. The red dashed line shows the baseline mean value recalculated over the *history window* at each time and the red dotted lines show the standard deviation.

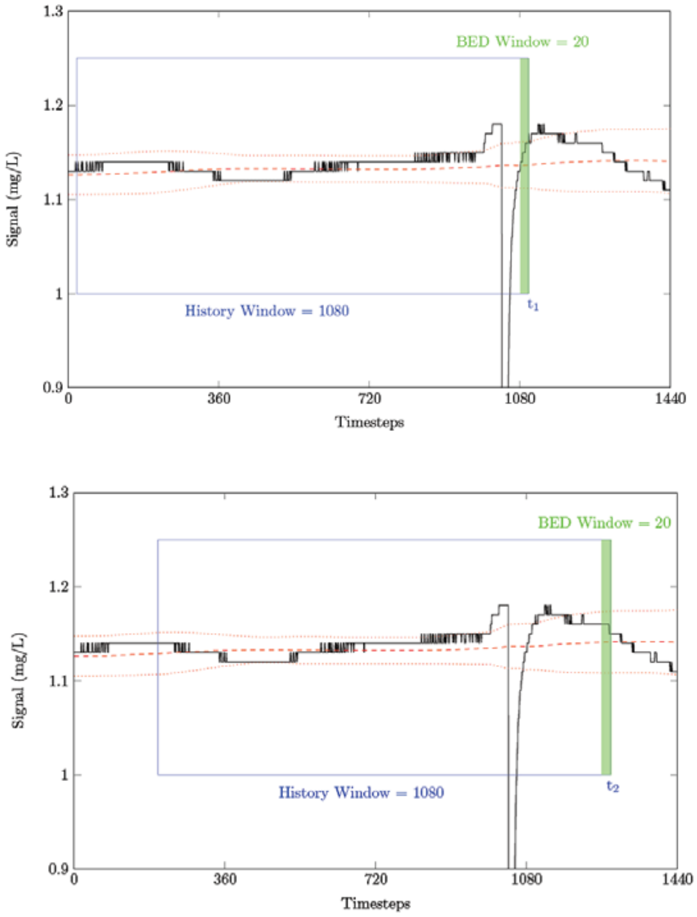


**Figure 2: Water quality signal over time showing the *history window* (blue box) and the *BED window* (green box) at time $t_1$ and at time $t_2$. The black line shows the water quality signal values over time, the red dashed line is the mean value and the red dotted lines show the standard deviation from the mean calculated using data in the moving *history window*.**

***outlier threshold***: This parameter is the number of standard deviations away from the mean a data point must be to be labeled an outlier. This parameter is multiplied by the standard deviation of a signal (as calculated within the moving *history window)* to produce the acceptable maximum normal deviation of a signal. The statistics of a signal within the *history window* are calculated and normalized. The difference between the recorded (current value) and predicted value of a signal is normalized, producing a residual. The data point is considered to be an outlier if the calculated residual exceeds the *outlier threshold* times the standard deviation.

The *outlier threshold* parameter defines when CANARY treats a data point as an outlier, and thus, a possible indicator of an event. It defines how large the difference between the data point and the predicted value (i.e., the residual) must be in order for the signal to be considered an outlier. This parameter is a multiplier of a signal's standard deviation, and so an *outlier threshold* value of 1.0 means that a signal value greater than one standard deviation from the mean signal value is labelled an outlier. Assuming the normal bell-shaped probability curve, the range of plus and minus one standard deviation ($\pm 1\sigma$) contains 68.3% of the baseline data. Using the *outlier threshold* parameter, the user is defining what CANARY considers acceptable, or good, variations within the signal. Anything that CANARY treats as an outlier, or anomalous, can ultimately contribute to an alarm. Lower values of the *outlier threshold* result in more outliers, because less of the data is treated as good relative to the predicted behavior. Higher values are less sensitive; if it is too high, almost all data might be considered good and real events might not be detected. CANARY normalizes residuals relative to a signal's value such that only one *outlier threshold* value is required for all signals. It should also be noted that this normalization occurs over the moving *history window* so that the definition of good changes over time. In the second plot of Figure 1, the normalized residual values are shown; in this case, an *outlier threshold* of one would not result in any detected events as the residual values never cross ± 1.

***BED window***: This parameter is the number of historical data points examined to look for the onset of an event. This parameter establishes how many data points CANARY uses to determine whether or not an event is occurring. CANARY calculates the probability that an event is occurring based on the number of outliers present in the *BED window*. For example, a *BED window* of 1 indicates that CANARY will decide if a single outlier is significant, whereas a *BED window* of 10 indicates that CANARY will determine the significance of multiple outliers within that timeframe before producing an alarm.

The *outlier threshold* determines when a data point is declared an outlier, but in most cases, it is undesirable for CANARY to produce an alarm every time a single outlier is detected. While CANARY uses the *outlier threshold* parameter to determine when a data point is significantly different from a predicted, or normal, value, the *event threshold* and *BED window* parameter values work together to tell CANARY how many outliers are necessary in order to trigger an alarm. The *BED window* parameter defines the size of the historical data window that CANARY uses when grouping outliers. A *BED window* value greater than one ensures that at least two outliers are needed before CANARY detects an event; this helps to greatly reduce false positives caused by noisy or bad data. The *event threshold* value defines the value of probability that must be exceeded in order for a group of outliers to be considered an event, and thereby, trigger an alarm. CANARY's statistical algorithms calculate the probability that an event is occurring at each time step; if this value exceeds the *event threshold*, CANARY alerts that a potential event has been detected.

***event threshold***: This parameter defines the value of probability that must be exceeded in order for a group of outliers to be considered an event and to trigger an alarm by CANARY. This value determines the number of outliers that must be located within a *BED window* to trigger an alarm.

In practical terms, these two parameters represent: (1) the time period in which to look for the onset of an event and (2) the number of outliers that need to occur within that time period to indicate an event. The combination of the *BED window* and *event threshold* parameters also affects the delay between the start of an event and the alarm produced by CANARY, which might also be a consideration for water utilities. The *BED window* parameter does not have to be selected to capture the entirety of an event, only the length of time necessary to signify that a water quality change has begun. For detection, CANARY is concerned with the significance of the change, not the total length of time that a signal remains abnormal.

To illustrate the relationship between *BED window* and *event threshold*, assume that the *data interval* is 2 minutes, the *history window* is set to 1080 (1.5 days) and the *BED window* to 15 (30-minutes). If one would like to ensure that 10 of the 15 timesteps in this window are anomalous before declaring an event, Table 2 shows the corresponding *event threshold* that is required. The table contains *BED window* parameter values corresponding to a real world window of 30 minutes, for *data intervals* of 1-minute, 2-minutes and 5-minutes. Assuming that 2/3 (67%), or more, of the data points within a window must be outliers in order to trigger an alarm, the number of outliers in the *BED window* must be 20/30, 10/15 and 4/6, respectively. It is equally valid to say that 20 minutes out of the 30-minute window must be anomalous in order to trigger an alarm. Binomial distribution theory is used to calculate the *event threshold* range required to satisfy this scenario (Table 2, also see equations 1 and 2, and the CANARY User's Manual for further details (U.S. EPA, 2012a)).

**Table 2: Examples of How *data interval*, *BED window* and *event threshold* Parameters Interact**

| ***data interval*** | ***BED window* value (≅ 30 min)** | **# of outliers =2/3 of *BED window*** | **Resulting range of *event threshold* values** |
|---|---|---|---|
| 1 minute | 30 | 20 | 0.951–0.9785 |
| 2 minutes | 15 | 10 | 0.85–0.94 |
| 5 minutes | 6 | 4 | 0.657–0.89 |

*BED, binomial event discriminator; BED window value* = (30 min)/(*data interval*)

The following conceptual process for calculating *BED window* and *event threshold* can be adopted. The *BED window* parameter value can be calculated based on the timeframe over which an extended period of outliers is likely to relate to an event. For most systems, this is likely to correspond to a timeframe of 30 minutes to 2 hours. This timeframe is divided by the *data interval* to establish the *BED window* value. The number of required outliers ($N_{RO}$) can be chosen as a number less than the *BED window*, or can be thought of as the percentage of the *BED window* that must be outliers in order to trigger an alarm ($N_{RO}$ = (BED)(% outliers)). Then the *event threshold* value can be calculated based on these two values (*BED* and $N_{RO}$) as follows:

$$\text{Minimum Event Threshold} = \sum_{i=0}^{N_{RO}-1} \left(\frac{\text{BED!}}{i!\,(\text{BED}-i)!}\right)\left(\frac{1}{2}\right)^{BED} \qquad (1)$$

$$In\ Excel = BINOMDIST(N_{RO}-1, BED, 0.5, TRUE)$$

$$\text{Maximum Event Threshold} = \sum_{i=0}^{N_{RO}} \left(\frac{\text{BED!}}{i!\,(\text{BED}-i)!}\right)\left(\frac{1}{2}\right)^{BED} \qquad (2)$$

$$In\ Excel = BINOMDIST(N_{RO}, BED, 0.5, TRUE)$$

As shown in the equations, this calculation can be easily done in Microsoft Excel® (Redmond, WA) which has a built in binomial distribution function. As the *BED window* value increases, the range of *event threshold* values corresponding to the number of required outliers decreases, as can be seen in Table 2. For most situations, it is sufficient to calculate the minimum *event threshold* and round this value up to three significant digits. Additionally, increasing the *event threshold* increases the likelihood that a series of outliers are contiguous and reduce the likelihood of noise triggering an alarm.

The *BED window* and *event threshold* also affect the timeliness of detection and the length of events detected. The average detection delay is tied to the required number of data points that must be outliers in order to trigger an alarm. For example, in Table 2, if the *data interval* is 2 minutes, the *BED window* of 15 timesteps equals 30 minutes, and the *event threshold* is 0.9, then 10 data points are required to be outliers, and an alarm is not triggered until at least 20 minutes after CANARY detects the first outlier. If an event lasts less than 20 minutes, it would result in fewer than 10 required outliers, and would not be detected with these configuration settings. The shortest detection delay corresponds to the scenario when all outliers are consecutive between the first detected deviation and the alarm, however, if there are gaps in the outliers then there is a longer delay to detection. The delay before alarm may exceed the *BED window* value when the initial deviation from normal behavior is accepted as normal, rather than being treated as an outlier.

If the length of an event is less than or equal to the *BED window*, and the required number of outliers is satisfied, CANARY would only produce an alarm for a single timestep after the event occurs. If, however, the event is longer than the *BED window*, CANARY would continue to produce an alarm as long as the event probability is larger than the *event threshold* up to the *event timeout* value. The *event timeout* parameter is specified by the user and is the number of consecutive data points after an event is found before the alarm is silenced automatically. In this way, it is better to have a shorter *BED window* and allow for a prolonged alarm, than a large *BED window* that might only detect events at one timestep with a long delay.

## 2.2 Additional Considerations

Once the user has selected these important configuration parameters, CANARY can be run on historical or real time data. The user might find, however, that the selected configuration parameters do not result in adequate sensitivity and specificity (i.e., they result in high false positive rates) of the EDS at one or more sensor stations. An EDS that is being used to make real time decisions about water treatment for maintaining a finished water standard (i.e., a control system) might have different requirements when compared to an EDS that is looking for changes that could indicate large scale contamination. The parameter values can be adjusted to reduce false alarm rates while ensuring adequate detection sensitivity.

Although CANARY can be adjusted to minimize the effect of noisy data or missing data, using highly reliable sensors, and properly calibrating and maintaining sensors, helps reduce the need to optimize CANARY parameters (Allgeier et al., 2011a; Pickard et al., 2011). If maintaining water quality within a fixed set-point range is desirable (for system control, or for maintaining quality within a regulatory standard), algorithms such as the set-point proximity algorithms can be used in addition to the statistical algorithms to provide better event detection for a given system. CANARY offers two proximity algorithms that incorporate both probability analysis and set-points when producing alarms; they use either the Set-Point Proximity algorithm using Beta distribution (SPPB), or using Exponential distribution (SPPE) to calculate how the probability of an event increases as a fixed set-point is approached (Hart and McKenna, 2012).

Thought should also be given to where sensors are located within a system. Sensors located near pumps, tanks, chemical injections, mixing points or water sources might have much higher signal variability relative to sensors located further from these facilities in the distribution system. A water quality change that results in a sharp signal response at a pumping station results in a much broader, or muted, sensor response in the middle of a distribution system. This is the result of diffusion and mixing of water within a pipe. This broadening of sensor response causes two things: (1) it lengthens the duration of the event and (2) it decreases the maximum signal response at an individual timestep. If the water quality change is a baseline shift, rather than a pulse, the timeframe over which the change occurs is lengthened, but the water quality eventually reaches its new baseline throughout the system.

Sensors in the middle of a water distribution system might demonstrate less variability. They might contain well-mixed water with fewer sharp changes in water quality readings; a sensor station in this type of location might provide good alarm behavior with a shorter *BED window* (8-12). For sensors near facilities, the key is to determine how much a signal should change under normal conditions, and for how long. The *BED window* should be set to be long enough to exceed the normal duration of a change in a signal. Additionally, choosing an *outlier threshold* value from 1.0 to 1.5 ensures that normal signal variability is treated by CANARY as normal, and that only significant deviations from normal behavior are considered outliers.

Changes in alarm behavior associated with seasonal changes were not explored in detail. Although changing seasons could affect alarm behavior, this effect is expected to be small for most systems. CANARY relies on the moving *history window* to establish baseline behavior for signals. This means that CANARY is constantly recalculating the predicted behavior, and other statistics, associated with a signal. Shifting of a baseline or an increase in the standard deviation of a signal is automatically adjusted as new data points are added to the *history window* and old data points are removed. While an increase in alarms at the start or end of a season might occur, the rate of alarms during a season is likely to be unaffected because new baseline values are being used. If a seasonal change does cause an increase in signal variability, a higher *outlier threshold* value than those discussed within this report might produce more favorable alarm rates. In general, an *outlier threshold* value should be selected with periods of the highest normal variability in mind. Simulating events within highly variable periods of signal data can ensure that a parameter set is still able to detect real events. In practice, a 1.5 to 2 day *history window* helps limit the effect of day-to-day variability and seasonal changes; possibly only causing a slight increase in alarms for a few days around seasonal changes.

In addition to the statistical parameters discussed above, two other parameters should be set for each sensor during the initial setup process: *precision* and *valid range*. These parameters can be found by examining each sensor's technical documentation. The *valid range* corresponds to the range of values that can be reported by a sensor; for example, a pH sensor might have a *valid range* from 0 to 14. The *valid range* parameter is not related to a set-point range; it only provides CANARY a frame of reference for whether a sensor is accurate if it reports a given value. The *precision* is related to a sensor's ability to report at a specific increment; in other words, if a sensor can only report with a precision of 1, then CANARY does not treat a change in that signal of 1 unit in the same way as for a sensor that can report in increments of 0.01. These values would require only a single set up, with changes to these values only being necessary if new hardware were installed. CANARY can operate successfully without setting either of these values; however, setting these parameters lessens alarms caused by invalid data.

## 3.0 Data and Methods

In order to develop rule-of-thumb guidance for the selection of these four parameter values, historical water quality data from two pilot cities were analyzed and performance results compared for multiple parameter values. In this section, the data and methods used for this analysis are presented.

### 3.1 Station Information

The CANARY EDS software was used to analyze sensor data from two different drinking water systems; one station from the GCWW distribution system (measured at EPA's Testing and Evaluation (T&E) facility) and four datasets from the Singapore Public Utility Board (PUB) water distribution system.

The T&E facility is a unique research facility that provides researchers an opportunity to perform real time sensor deployment in a municipal water system and to test sensor response to injected contaminants within a controlled water distribution system setting. This provides an opportunity to test CANARY's ability to detect real contamination events. The research nature of the T&E facility produces an abundance of available sensor data.

GCWW treats water from the Ohio River, and maintains a chlorine disinfectant residual in the distribution system. During the test period, GCWW tap water entered the T&E facility and was stored in a 750 gallon polyethylene tank that was refreshed several times per day. The travel time of the tap water between the GCWW treatment facility and T&E was approximately 12 hours. The tap water was gravity fed into a 1,200 feet (ft) long pipe loop made up of 3-inch diameter glass-lined ductile iron pipe, with a total capacity of 440 gallons. The flow rate through the system varied from 3 to 22 gallons per minute depending on test conditions. The contaminant injection port was located immediately after the storage tank, and two water quality sensor stations were located at 80 ft and 1180 ft downstream of the injection port. After flowing through the entire loop, the water was discharged to the public sewer. Note that the experiments were designed so that the effluent did not exceed contaminant-specific discharge limits allowed. For more information about the design and operation of the pipe loop as well as the implementation of these experiments, see Hall et al., 2009.

PUB is Singapore's national water agency. PUB manages water collection, treatment and reclamation as part of their effort to supply drinking water to the population of Singapore. PUB supplies approximately 5 million residents 300 million gallons per day from nine treatment facilities and 17 reservoirs. Fifty percent of Singapore's water is imported. A large number of monitoring stations are located within Singapore's water distribution system. Signal data from four of these monitoring stations was analyzed within this report.

Table 3 presents the ranges and averages of water quality sensor signal data during the 8-month test period for each of the five stations used in this study. The data shows that water quality parameters for treated drinking water can vary considerably even within a single water utility. CANARY analysis was performed on two groups of sensors for each system as described further in the methods section of this report. Sensor signals listed in bold were included in both analyses. Averages include all data in the tested date range; no attempt was made to remove bad data, or real events.

**Table 3: Ranges and Averages for Sensor Signals from T&E and PUB Datasets**

| Facility | Sensor† | Units | Data Range | Average | Std. Dev. |
|---|---|---|---|---|---|
| T&E | **Chlorine** | ppm | -0.45 – 2.00* | 1.13 | ± 0.11 |
| | **pH** | – | 5.03 – 10.01 | 8.35 | ± 0.18 |
| | Temperature | °C | 0.00 – 31.44 | 16.69 | ± 4.78 |
| | **Spec. Conductivity** | µS/cm | 0 – 526 | 338 | ± 32 |
| | Chlorine | ppm | -1.07 – 5.25* | 1.65 | ± 0.27 |
| | pH | – | 0.00 – 9.98 | 8.92 | ± 0.44 |
| | ORP | mV | -192 – 763 | 135 | ± 399 |
| | Turbidity | NTU | -3.3 – 774.4* | 502.1 | ± 333.1 |
| | **UVA** | $m^{-1}$ | -0.7589 – 4.0037* | 1.1601 | ± 0.4848 |
| PUB1 | **Chlorine** | ppm | 0.000 – 2.356 | 1.920 | ± 0.062 |
| | **pH** | – | 2.00 – 8.86 | 8.146 | ± 0.140 |
| | **Spec. Conductivity** | µS/cm | 159.829 – 187.912 | 166.855 | ± 4.001 |
| | Turbidity | MNTU | 0.040 – 2.995 | 0.096 | ± 0.051 |
| PUB2 | **Chlorine** | ppm | 0.000 – 5.000 | 2.118 | ± 0.220 |
| | **pH** | – | 2.499 – 10.766 | 8.170 | ± 0.341 |
| | **Spec. Conductivity** | µS/cm | 0.977 – 300.000 | 87.015 | ± 7.959 |
| | Turbidity | MNTU | 0.071 – 5.585 | 0.130 | ± 0.096 |
| PUB3 | **Chlorine** | ppm | 0.000 – 5.000 | 2.192 | ± 0.229 |
| | **pH** | – | 4.15 – 9.55 | 8.05 | ± 0.11 |
| | **Spec. Conductivity** | µS/cm | 0.000 – 445.000 | 312.993 | ± 39.371 |
| | Turbidity | MNTU | 0.00 – 1.99 | 0.08 | ± 0.08 |
| PUB4 | **Chlorine** | ppm | 0.000 – 5.000 | 2.151 | ± 0.235 |
| | **pH** | – | 6.67 – 8.94 | 8.06 | ± 0.10 |
| | **Spec. Conductivity** | µS/cm | 0 – 753 | 315 | ± 41 |
| | Turbidity | MNTU | 0.01 – 1.99 | 0.07 | ± 0.11 |

NTU, nephelometric turbidity units; MNTU, milli NTU; ORP, oxygen reduction potential; PUB, Singapore Public Utility Board; Spec., specific; T&E, U.S. EPA Testing and Evaluation Facility

*Negative values for these sensor signals occurred during the test period, however, they were outside the valid range for each sensor. No attempt was made to correct these values within the data file.

† Analysis was performed on two groups of sensors for each facility. Sensor signals listed in bold were included in both analyses.

Table 4 contains information regarding the completeness of the data available from each station for the tested timeframes. In particular, some data points were missing from each location, perhaps caused by sensor malfunctions or data transmission errors. CANARY, however, is designed to manage missing data.

**Table 4: Data Completeness of T&E and PUB Datasets**

| Data Location | Actual Timesteps | Theoretical Timesteps | Percent Complete (%) |
|---|---|---|---|
| T&E | 171292 | 171359 | 99.96 |
| PUB1 | 70247 | 70270 | 99.97 |
| PUB2 | 68782 | 70270 | 97.88 |
| PUB3 | 67851 | 70270 | 96.56 |
| PUB4 | 62616 | 70270 | 89.11 |

PUB, Singapore Public Utility Board; T&E, U.S. EPA Testing and Evaluation Facility

Table 5 contains the information about the 14 contaminant testing events at the T&E facility that were performed during the analyzed timeframe of this study. Contaminants were injected for either 2 minutes or 20 minutes. Three contaminants and the dechlorinating agent sodium thiosulfulate were injected for a total of 14 testing events as listed. See Hall et al., 2009 for more information about these tests.

**Table 5: Contaminant and Injection Times for Contaminant Tests at the T&E**

| ID | Contaminant | Concentration | Time of Injection | |
|---|---|---|---|---|
| 1 | Sodium thiosulfate | * | 11/14/2011 | 10:30 AM |
| 2 | *Escherichia coli* | 1.0×$10^3$ CFU/mL | 11/15/2011 | 10:30 AM |
| 3 | *E. coli* | 1.0×$10^3$ CFU/mL | 11/15/2011 | 12:00 PM |
| 4 | *E. coli* | 1.0×$10^4$ CFU/mL | 11/15/2011 | 1:30 PM |
| 5 | *E. coli* | 1.0×$10^4$ CFU/mL | 11/15/2011 | 3:00 PM |
| 6 | KCN | 10 ppm | 11/28/2011 | 11:50 AM |
| 7 | KCN | 20 ppm | 11/28/2011 | 1:40 PM |
| 8 | Atrazine | 10 mg/L | 11/28/2011 | 2:40 PM |
| 9 | Atrazine | 10 mg/L | 01/26/2012 | 11:40 AM |
| 10 | Atrazine | 10 mg/L | 01/26/2012 | 3:00 PM |
| 11 | Atrazine | 1 mg/L | 02/01/2012 | 1:00 PM |
| 12 | Atrazine | 1 mg/L | 02/01/2012 | 2:30 PM |
| 13 | Atrazine | 0.1 mg/L | 02/02/2012 | 2:00 PM |
| 14 | Atrazine | 0.1 mg/L | 02/02/2012 | 3:30 PM |

CFU, colony forming units; T&E, U.S. EPA Testing and Evaluation Facility; *Concentration not recorded.

## 3.2 Methods

CANARY version 4.3.2 (Hart and McKenna, 2012) was used for the analysis of the T&E and PUB data. Results presented in Sections 4 and 5 utilized CANARY's LPCF algorithm. This algorithm was chosen because it is the most commonly used and it applies a predictive algorithm to each signal individually when performing its analysis and thus provides a high likelihood of detecting contamination events that might only change one water quality signal.

In order to explore the importance of the configuration parameters, several different combinations of parameter values were used as input to CANARY and the results were compared. A summary of tested parameters is outlined in Table 6. An 8-month timeframe was chosen for each station. These date ranges were selected because they contained data for over 89% of the theoretically available data for the station PUB4, and over 96.6% of available data for the other datasets (see Table 4). Data might not have been present for several reasons: power loss within the system, communication loss with SCADA system and sensor errors or failures.

Water quality data from T&E and PUB datasets had *data intervals* of 2-minutes and 5-minutes, respectively. Factoring in the system's *data interval*, the tested *history windows* vary from 12 hours to one week, spanning the values of 1.5 and 2 days that were recommended in the CANARY User's Manual (Hart and McKenna, 2012). One hundred ninety-two parameter combinations were tested using T&E data. One hundred eight parameter combinations were tested for each PUB dataset.

The *outlier threshold* was adjusted around one standard deviation (± 1 σ) in increments of 0.15 (producing values of 0.85, 1.0 and 1.15) for all datasets. A value of 1.4 was also included with T&E to see if a higher value would prevent CANARY from detecting true contamination events. Increasing the *outlier threshold* increases the amount of variation that is considered normal. A normal bell-shaped distribution and an *outlier threshold* of 0.85, 1.0, 1.15 and 1.4 correspond to 60.5%, 68.3%, 75% and

83.8% of variations being accepted as being normal, respectively. The range of *BED window* values were selected in order to detect signal changes (i.e., events) that occurred for approximately 30 to 60 minutes. Real timeframes of 30, 40 and 60 minutes were used; for T&E, these *BED window* values were 15, 20 and 30 and for PUB, values of 6, 8 and 12 were used. For T&E, a *BED window* of 8 was also included, which is equivalent to 16 minutes. The equivalent value of 3 for the PUB datasets 5-minute *data interval* was not used as it would result in less controllability of alarm behavior because only 2 or 3 anomalous data points would be required to trigger an alarm. The *event threshold* is fixed at 0.9 in order to limit the total number of permutations that were tested, and because it cannot be varied independently of the *BED window* as discussed in section 2.1. The significance of the *event threshold* is discussed in conjunction with the *BED window* in sections 4.1.2 and 4.2.2.

**Table 6: Testing Information for the T&E and PUB Datasets**

| Configuration Variables and Parameters | T&E data sets | PUB data sets |
|---|---|---|
| Data Range | 11/01/2011 – 06/26/2012 | 01/01/2008 – 08/31/2008 |
| Days Analyzed | 238 | 244 |
| Number of Sensor Signals tested | 4 & 9 | 3 & 4 |
| *data interval* (min) | 2 | 5 |
| *history windows* (number of data points) tested | 360, 720, 1080, 1440, 2160 & 5040 | 144, 288, 432, 576, 864 & 2016 |
| *BED windows* (number of data points) tested | 8, 15, 20 & 30 | 6, 8 & 12 |
| *outlier thresholds* (probability) tested | 0.85, 1.0, 1.15 & 1.4 | 0.85, 1.0 & 1.15 |

*BED, binomial event discriminator;* PUB, Singapore Public Utility Board; T&E, U.S. EPA Testing and Evaluation Facility

CANARY analyses were conducted using signal data from four or nine sensor signals for T&E, and either three or four sensor signals for PUB datasets (see Table 3). For some of the analysis, the total number of sensor signals was reduced to 4 for the T&E data and 3 for PUB datasets. This removed the turbidity signal from all datasets; previous work showed that the signal to noise ratio for the turbidity signal was not sufficient to be useful in contaminant event detection (Hall et al., 2009). The four additional sensor signals that were removed from the T&E analysis were temperature, oxidation reduction potential (ORP), one pH and one chlorine. The pH and chlorine signals that were removed from the T&E analysis had higher variability than the pH and chlorine signals that were retained (bolded in Table 3). For the reduced signal testing scenario, there was a chlorine, pH and specific conductivity sensor for all datasets. In addition to these signals, a UVA (ultraviolet) sensor signal was present in the T&E signal data, which responds to many of the same contaminants as a total organic carbon (TOC) sensor (U.S. EPA, 2012a). Chlorine and TOC sensors have been shown to respond to the widest variety of contaminants (U.S. EPA, 2012a), and pH and specific conductivity have also been shown to respond to contaminants (Hall et al., 2009).

In general, the results are reported as the total number of alarms calculated by CANARY and the false alarm rate per day. The number of false alarms per day is reported in order to provide a metric for translating these results to other systems. For the contamination events at T&E, the number of true detections, and the detection delay is also reported. For the purposes of this report, an alarm is produced when CANARY determines that a water quality event has occurred. Alarms can be divided into four categories: a true positive (CANARY alarms and an event occurred), a false positive (CANARY alarms even though no event occurred), a true negative (CANARY did not trigger an alarm and no event occurred) and a false negative (CANARY did not trigger an alarm but an event did occur). In this report, a true positive detection is defined to be any alarm associated with a known testing event that occurred at the T&E facility. All alarms at PUB and any remaining alarms at T&E are reported as false alarms; however, these alarms might have a valid root cause that is not known by the authors. Appendix B: *What Constitutes an Event?* discusses other types of water quality or operational changes that could be considered by water utilities to be true events.

For T&E, detected events, or true positives, were determined based on the injection times listed in Table 5. The alarm times reported by CANARY were compared to the known contaminant injection times to determine if each alarm corresponds to a real testing event. Alarms that correspond to a real testing event were reported as a detected event. The delay to detection was reported and discussed in relation to the various parameters being tested. For T&E, any alarm that did not correspond to a real testing event was considered a false alarm. All alarms for PUB analyses were treated as false alarms for the purpose of this report. The rationale for this treatment is discussed further in the Section 4.0.

These analyses were performed using CANARY version 4.3.2, running on Windows 7, updated to the version of the day available on August 6, 2013. The specific version information is CANARY 4.3.2 (build b580:r3777, 2013-05-31 00:28:18). This version of CANARY relies on the MATLAB® (MathWorks, Natick, MA) Compiler Runtime version R2008b and only runs on a single processor. Analyses discussed in this report were performed using CANARY's batch mode on historical data contained in comma separated value files (CSV). Two computer systems – designated Computer 1 and 2 – were used to perform all analyses.

Runtime is not an issue when running CANARY in real-time; in that case, the software runs quickly and typically waits for the SCADA system to send real data. However, runtime results are presented here in order to demonstrate how long it might take to run multiple configuration options in order to optimize the configuration. Each analysis configuration file, and consequently each run, consisted of only a single algorithm with a signal set of configuration parameters. The performance improvement related to using multiple algorithms is discussed in section 5.3. Runtime statistics for these analyses are reported in terms of total runtime, runtime per day of historical data and runtime per timestep for two different computer systems. Four designations are used to discuss runtime statistics:

- Computer 1: Intel Pentium Dual E2180 (@2.00 GHz) processor with 4 GB of RAM. A single CANARY analysis running at a time.
- Computer 1a: *Same hardware as above*. Two CANARY analyses running simultaneously (one per core).
- Computer 1b: *Same hardware as above*. A single CANARY analysis running on a machine running other programs, at full CPU utilization.
- Computer 2: Dual Processor Intel Xeon E5430 (@2.66 GHz) with 4 GB of RAM running – eight total cores. This machine was able to run four to six analyses simultaneously, without appreciable loss of performance.

# 4.0 Results

This section is divided into results and discussion for T&E and PUB individually. Throughout this section, the term "alarm" refers to the total number of alarms reported by CANARY. No attempt was made to merge alarm events that were reported close together. Alarms that occurred shortly after another alarm were not common, and were only present in a few parameter combinations that had higher alarm rates in PUB3. Additionally, true contamination detections at T&E were only considered to be correct if an alarm was triggered after the start of an event, not as part of a previous alarm. The term "false alarm" is used to describe all alarms reported from analyzing PUB datasets and any alarm at T&E that cannot be attributed to a contamination testing event. These alarms might have a root cause; however, no information was available relating to these causes.

In order to highlight trends, graphical information is presented as an average over multiple parameter values unless specifically mentioned otherwise. Total or false alarm values and true detections are tabulated at the beginning sections 4.1.2 and 4.2.2 for T&E and PUB data, respectively. Graphs show the average of two or more values that satisfy the listed parameters. For example, a graph that shows alarm behavior with *BED window* and *outlier threshold* categories averages all values that satisfy each group; that is, each reported value is the average of both *history window* values that have the same *BED window* and *outlier threshold*. This was done because, while the *history window* does affect alarm behavior, this effect was generally much smaller than changes associated with the other two parameters.

## 4.1 T&E

Within this section, three main performance metrics are discussed: detected events, false alarms and delay time to detect real events. The effect of *history window* is investigated. The effect of *BED window* and *outlier threshold* are discussed together. Finally, the effect of the number of signals is investigated.

### 4.1.1 *History Window*

The CANARY User's Manual (Hart and McKenna, 2012) suggests using a *history window* of approximately 1.5 to 2 days – for this system with a 2-minute *data interval*, this corresponds to *history window* values of 1080 to 1440. As shown in Table 6, most of the analysis results focus on these two values of *history window*; however, this subsection investigates why those values were recommended by considering a larger range of *history window* values, from ½ day (360) to 7 days (5040). The analysis in this subsection uses data from four sensors (not the full set of nine sensors).

Figure 3 shows the number of false alarms produced by CANARY over the eight month time period for the five *history window* values included in this testing: ½ day (360), 1 day (720), 1.5 days (1080), 2 days (1440) and 3 days (2160), and for four *BED window* values: 8, 15, 20 and 30. Note that these results are averaged over all of the *outlier threshold* values given in Table 6, and the *event threshold* is fixed at 0.9. The minimum number of false alarms occurs when using a *history window* of 1080; however, the results are nearly equivalent for a *history window* of 1440. For this dataset, *history window* values of 360 and 720 resulted in an increased number of false alarms, and a *history window* value of 2160 resulted in the same or slightly larger number of false alarms relative to a *history window* value of 1440.

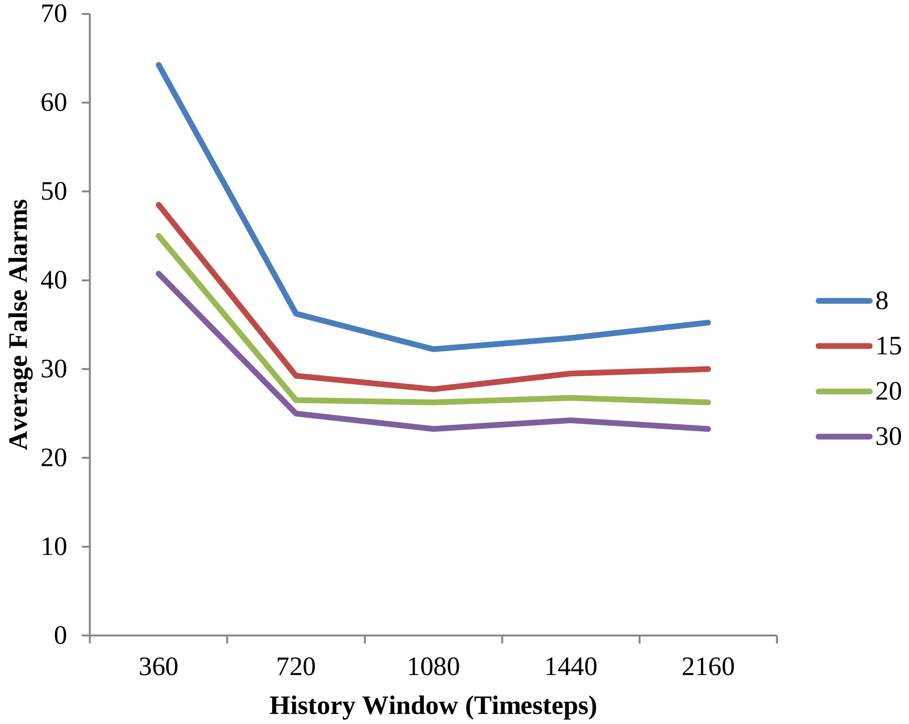


**Figure 3: Average false alarms per *history window* from CANARY using four signals from U.S. EPA Testing and Evaluation (T&E) Facility data from 11/01/2011 to 06/26/2012. Blue indicates a *binomial event discriminator* (*BED) window* of 8, red is 15, green is 20 and purple is 30.**

Figure 4 shows the number of false alarms for a range of *history window* values when using a single *BED window* of 15, an *outlier threshold* of 1.4 and an *event threshold* of 0.9. This analysis includes a *history window* of 7 days (5040), which shows a continued increase in false alarms beyond the three days included in Figure 3. This figure shows a clear minimum at 1440, although the difference is still small relative to 1080. Increasing the *history window* from 2 days to 7 days translates to a difference of approximately 10 alarms in a 238 day analyzed window, or 0.042 alarms per day, which is unlikely to be noticed in real use.

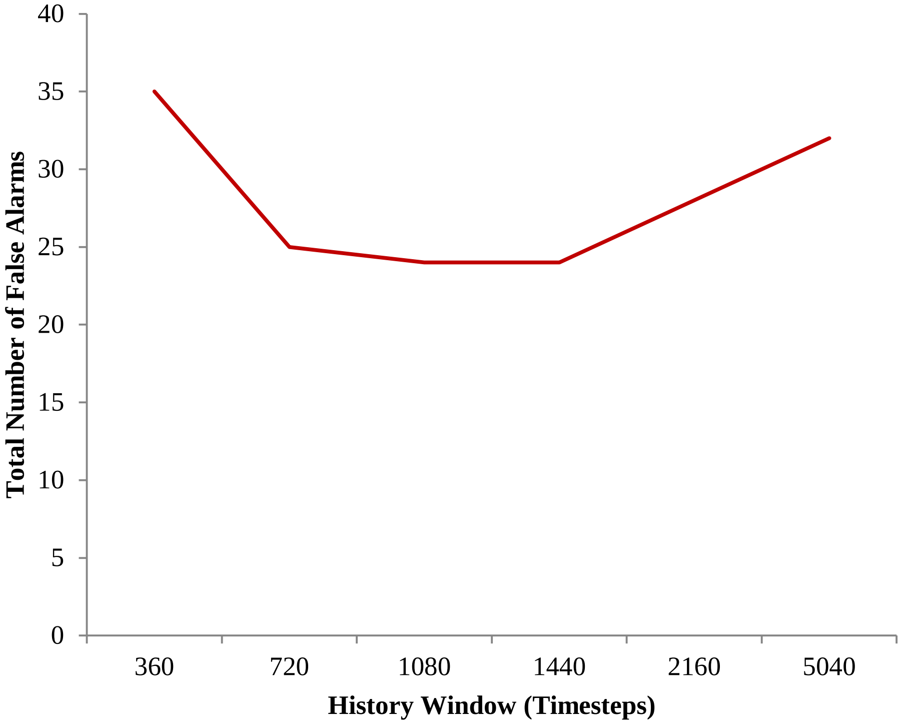


**Figure 4: False alarms per *history window* from CANARY using four signals from U.S. EPA Testing and Evaluation Facility data from 11/01/2011 to 06/26/2012. Line indicates false alarm behavior when using a *binomial event discriminator* (*BED) window* of 15, *outlier threshold* of 1.4 and *event threshold* of 0.9.**

Figure 5 shows CANARY's runtime for each of the 8-month analyses reported in Figure 4. A more thorough discussion of runtime is presented in section 5.3, however it is clear from Figure 5 that there is a large penalty for using a *history window* of 7 days for this system. The analysis using a *history window* of 5040 took approximately 18 hours to analyze 8 months of signal data, compared to less than 5 hours for a *history window* of 1440. This increase in runtime is caused by more data being processed to predict future behavior. The ability to run real time CANARY event detection using longer *history windows* should not be affected; however, the runtime might be an issue if a large set of parameter combinations were run in batch in order to optimize configuration parameters.

Further results based on T&E data focus on analyses that used *history windows* of 1080 and 1440. These values of *history window* minimize false positives and produce approximately the same number of alarms when other parameters are the same. Limiting the discussion to these values highlights general trends when examining the effect of other parameters. Moreover, changing the *history window* parameter within this range does not impact CANARY's ability to detect true events. Full alarm results can be found in Appendix D: *Full Alarm Data*.

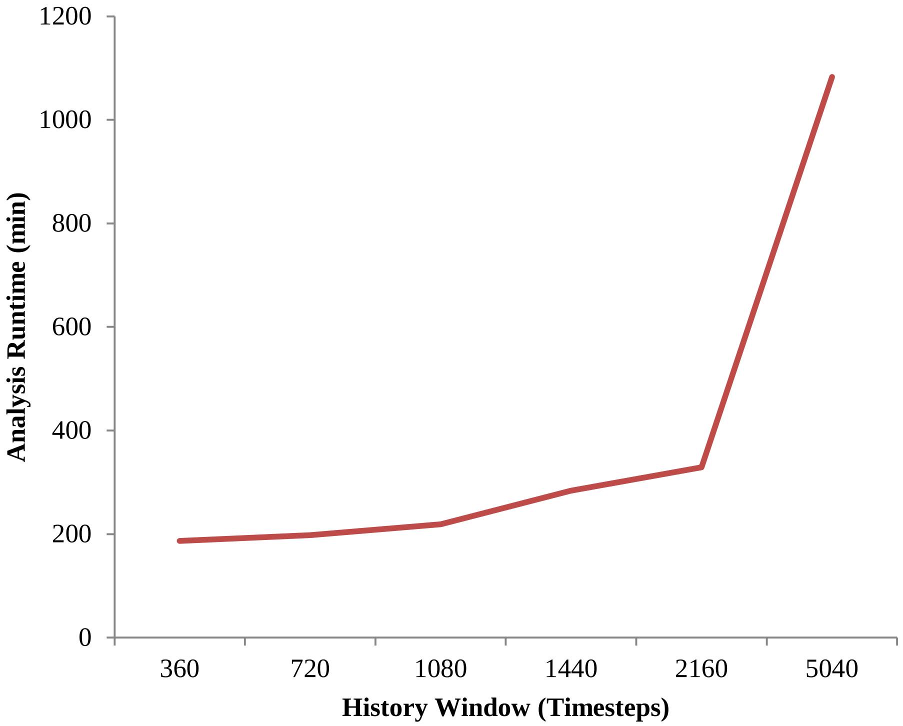


**Figure 5: Analysis runtime per *history window* for CANARY using four signals from U.S. EPA Testing and Evaluation Facility data from 11/01/2011 to 06/26/2012 using linear prediction coefficient filter (LPCF). Line indicates runtime behavior when using Computer 2, a *binomial event discriminator* (*BED) window* of 15, *outlier threshold* of 1.4 and *event threshold* of 0.9.**

### 4.1.2 *BED Window* and *Outlier Threshold* Parameters

Table 7 contains the total number of alarms reported by CANARY for the analyzed timeframe – reported for each combination of *BED window*, number of signals, *outlier threshold* and *history window*. Each parameter's impact on the total number of alarms is discussed individually for each parameter. Values in Table 7 that are in parentheses indicate the number of true event detections based on events listed in Table 5. The number of false alarms for each parameter combination is shown in Table 8.

**Table 7: Summary of Total Alarms Reported by CANARY on T&E Data from 11/01/2011 to 06/26/2012 Using the LPCF Algorithm.**

| ***BED window*** | **Number of Signals** | ***outlier threshold*** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 0.85 | | 1.0 | | 1.15 | | 1.4 | |
| | | ***history window*** | | | | | | | |
| | | 1080 | 1440 | 1080 | 1440 | 1080 | 1440 | 1080 | 1440 |
| | | **Total Alarms (True Detections)** | | | | | | | |
| 8 | 4 | 58 (14) | 62 (14) | 50 (14) | 49 (14) | 41 (14) | 43 (14) | 36 (14) | 36 (14) |
| | 9 | 115 (14) | 107 (14) | 84 (14) | 77 (14) | 68 (14) | 64 (14) | 53 (14) | 48 (14) |
| 15 | 4 | 49 (14) | 53 (14) | 46 (14) | 46 (14) | 38 (14) | 41 (14) | 34 (14) | 34 (14) |
| | 9 | 88 (14) | 84 (14) | 71 (14) | 66 (14) | 55 (14) | 51 (14) | 47 (14) | 42 (14) |
| 20 | 4 | 43 (10) | 45 (10) | 40 (10) | 39 (10) | 33 (10) | 34 (10) | 29 (10) | 29 (10) |
| | 9 | 76 (10) | 77 (10) | 60 (10) | 58 (10) | 49 (10) | 44 (10) | 39 (10) | 36 (10) |
| 30 | 4 | 34 (4) | 36 (4) | 29 (4) | 31 (4) | 24 (3) | 25 (4) | 20 (3) | 21(3) |
| | 9 | 65 (8) | 65 (8) | 53 (8) | 51 (8) | 39 (7) | 37 (8) | 32 (7) | 30 (7) |

*BED, binomial event discriminator;* LPCF, linear prediction coefficient filter; T&E, U.S. EPA Testing and Evaluation Facility

These tables show clear trends in how each parameter affects the results. As the *outlier threshold* increases, the number of false alarms decreases. For these parameter combinations, the *outlier threshold* does not impact the number of true detections. As the *history window* increases, the number of false alarms changes little, and in most cases, the *history window* does not impact the number of true detections. However, as shown in the previous subsection, *history window* can impact the results if the values are much smaller or larger than considered here. As the *BED window* increases, the number of true detections and the number of false alarms decrease. Finally, as the number of signals decreases from 9 to 4, the number of false alarms decreases and the number of true detections decreases, but only for large *BED windows*. These results are discussed in greater detail below.

**Table 8: Summary of False Alarms Reported by CANARY on T&E Data from 11/01/2011 to 06/26/2012 using the LPCF Algorithm**

| ***BED window*** | **Number of Signals** | ***outlier threshold*** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 0.85 | | 1.0 | | 1.15 | | 1.4 | |
| | | ***history window*** | | | | | | | |
| | | 1080 | 1440 | 1080 | 1440 | 1080 | 1440 | 1080 | 1440 |
| 8 | 4 | 44 | 48 | 36 | 35 | 27 | 29 | 22 | 22 |
| | 9 | 101 | 93 | 70 | 63 | 54 | 50 | 39 | 34 |
| 15 | 4 | 35 | 39 | 32 | 32 | 24 | 27 | 20 | 20 |
| | 9 | 74 | 70 | 57 | 52 | 41 | 37 | 33 | 28 |
| 20 | 4 | 33 | 35 | 30 | 29 | 23 | 24 | 19 | 19 |
| | 9 | 66 | 67 | 50 | 48 | 39 | 34 | 29 | 26 |
| 30 | 4 | 30 | 32 | 25 | 27 | 21 | 21 | 17 | 18 |
| | 9 | 57 | 57 | 45 | 43 | 32 | 29 | 25 | 23 |

*BED, binomial event discriminator;* LPCF, linear prediction coefficient filter; U.S. EPA Testing and Evaluation Facility

Here the results for the *BED window* and the *outlier threshold* are presented together as their trends are similar. Table 8 shows that the number of false alarms decreases as the *BED window* and the *outlier threshold* increase, for all parameter combinations studied. In this subsection, the effect of the *BED window* and the *outlier threshold o*n false alarm rates, true detection rates and detection delays is further investigated. In addition, the number of signals is also considered.

Four *BED window* parameters were used with the T&E data: 8, 15, 20 and 30. Table 9 summarizes the real timeframes analyzed and reports how the *event threshold* value of 0.9 relates to each *BED window* used. For example, a *BED window* of 8 is counting the number of outliers that have occurred within a 16-minute timeframe (window) for the *data interval* of 2-minutes used at T&E. With an *event threshold* of 0.9, 6 out of 8 timesteps must be considered to be outliers by CANARY to trigger an alarm. This means that 12 minutes out of a 16-minute timeframe must be considered an outlier in order for CANARY to trigger an alarm.

**Table 9: *BED Window* Parameters and Number of Required Outliers for T&E Data with an *Event Threshold* of 0.9 with a 2-minute *Data Interval***

| ***BED* window** | ***BED* Timeframe (min)** | **Required outliers in BED** | **Required duration of outliers (min)** |
|---|---|---|---|
| 8 | 16 | 6 | 12 |
| 15 | 30 | 10 | 20 |
| 20 | 40 | 13 | 26 |
| 30 | 60 | 19 | 38 |

*BED, binomial event discriminator*; min, minutes; U.S. EPA Testing and Evaluation Facility

Figure 6 shows the average number of false alarms per day for the studied parameter values. *Outlier threshold* values are shown as 0.85 in blue, 1.0 in red, 1.15 in green and 1.4 in purple. Darker shades indicate scenarios that use nine sensor signals and lighter shades indicate four sensors were used. Each graphed value is an average of the total false alarms over both *history window* values for a fixed *BED window/outlier threshold/# signals* combination. Figure 6 shows that increasing either *BED window* or *outlier threshold* results in a reduction in false alarms. Examination of the data also reveals that the alarm rates are more sensitive to – i.e. have a larger change because of – the changes to *outlier threshold* relative to the *BED window* parameter. In general, using only four signals rather than all nine reduced the number of false alarms significantly – by 40%. The effect of the number of signals on alarm behavior is discussed further below.

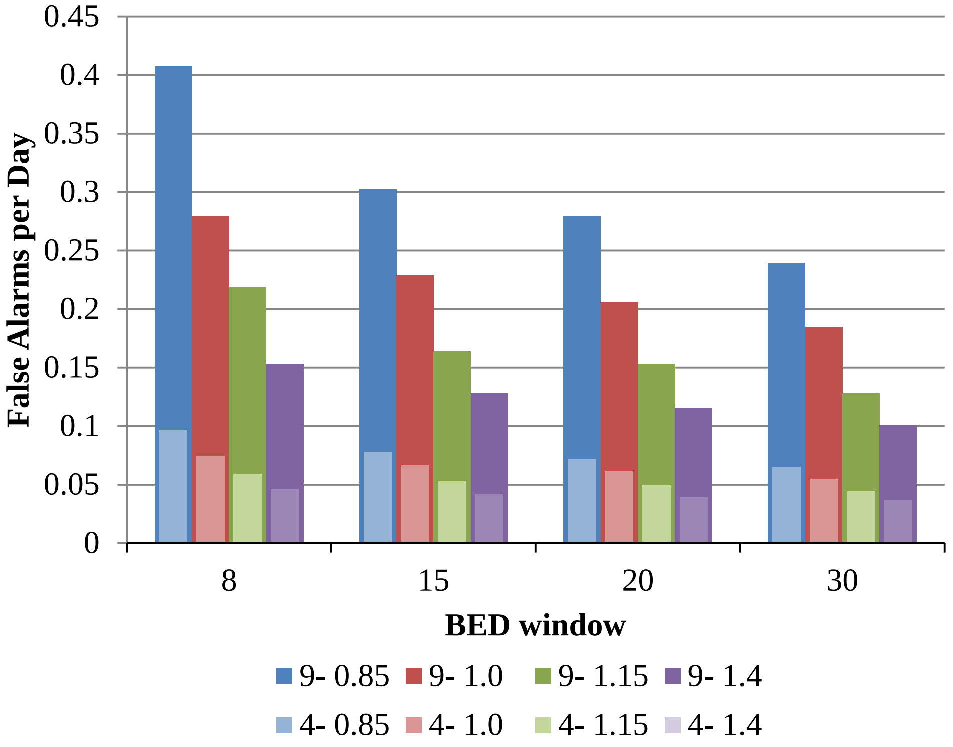


**Figure 6: False alarms per day from CANARY using U.S. EPA Testing and Evaluation Facility data from 11/01/2011 to 06/26/2012. Darker shades indicate scenarios that use nine sensor signals and lighter shades (overlaid) indicate four sensors used. *Outlier threshold* values are shown as 0.85 in blue, 1.0 in red, 1.15 in green and 1.4 in purple. The values are the average of alarm rates with different *history window* values.**

The maximum individual false alarm rate was 0.42 false alarms per day, when using nine signals, *BED window*=8, *outlier threshold*=0.85 and *history window*=1080. The majority of the tested scenarios resulted in an alarm every three to ten days (0.1 – 0.33 false alarms per day). When only analyzing four signals, the alarm rates do not exceed 0.1 false alarms per day (see lighter colors in Figure 6).

Figure 7 contains the average number of true contamination events (as listed in Table 5) detected by CANARY for each combination of *BED window* and *outlier threshold* values. A total of 14 testing events occurred at T&E during the analyzed timeframe. Of those testing events, 14 (100%) were detected with all combination of parameters that used *BED windows* of 8 or 15. Using a *BED window* of 20 resulted in only 10 (71%) events being detected. Contaminant injections ID#12 – ID#15 were not detected with *BED windows* of 20 or 30. A *BED window* of 30 resulted in as few as three events being detected, but averaged five or six events detected. Below a *BED window* of 20, there was no difference in true event detection between analyses that used four or nine signals; therefore, the number of signals is not shown in the figure. The four events that were not detected using a *BED window* of 20 did trigger an increase in the probability of an event; however, this probability did not exceed the *event threshold* of 0.9. These events correspond to the four Atrazine injections having concentrations of either 0.1 or 1.0 ppm (see Table 5).

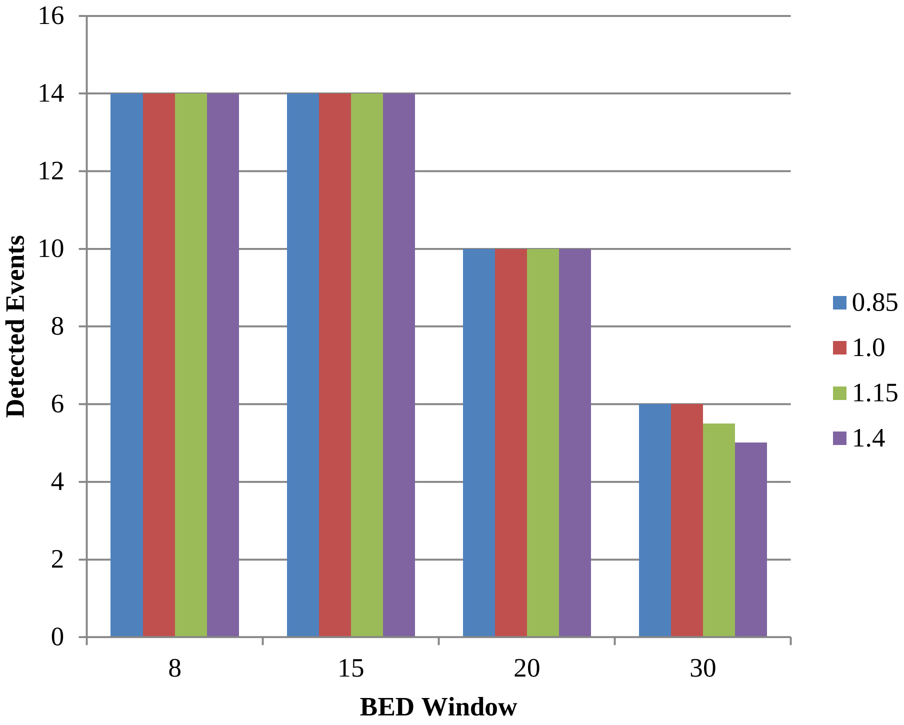


**Figure 7: Number of true detections of contamination events using CANARY on U.S. EPA Testing and Evaluation Facility data from 11/01/2011 to 06/26/2012 using linear prediction coefficient filter (LPCF).** ***Outlier threshold*** **values are shown as 0.85 in blue, 1.0 in red, 1.15 in green and 1.4 in purple. The values are the average of alarm rates with different number of signal and** ***history window*** **values.**

It should be noted that the reduction in the true detection rate for longer *BED windows* is linked to the specific experimental testing conditions at T&E, and in particular, the contaminant injection length. Contaminant injections were limited to twenty minutes, which translates to detectable changes in sensor signals only slightly longer than twenty minutes. Table 9 shows that, for the parameter combinations used for this testing, the signal changes would have to be at least 26 or 38 minutes for *BED windows* of 20 and 30, respectively, in order to be detected by CANARY. Therefore, for these parameter combinations, CANARY would not be able to detect these short events. This points to the importance of understanding the characteristics of the events that a utility would like to use CANARY to detect; using the *BED window, event threshold,* and the *outlier threshold*, the software can be configured to detect short or long events. In general, events with both long and short durations can be detected with smaller values of the *BED window* parameter, whereas, only longer events can be detected using larger *BED window* values.

Comparing results from Figure 6 to Figure 7, it is clear that there is a tradeoff between increased detection sensitivity with lower *BED window* values and decreased false positive rates with higher *BED window* values. The false alarm rate can be reduced to 0.084 false alarms per day (20 false alarms in 238 days, light purple), using a *BED window* of 15, an *outlier threshold* of 1.4 and four signals, while maintaining the ability to detect 14 out of 14 true contamination events. Increasing the *BED window* to 30 only reduced the number of false alarms to 17 or 18, but failed to detect 11 events that were detected when the *BED window* was 15.

When detecting real contamination events, another important factor is the detection time. In all 14 detected events, there was a delay between the time the contaminant was injected and the time that CANARY detected the event. The delays are due to several factors: the flow time from the injection point to the sensors, time for the sensors to measure and report water quality values, and the time required for CANARY to witness enough outliers to determine an event has occurred and calculate results. The *BED window* and the *event threshold* determine the number of outliers needed before CANARY identifies an event, and therefore affect the delay time.

Figure 8 shows how this delay is related to the *BED window*; the *event threshold* is fixed at 0.9. To minimize the number of figures, the delays were averaged over results for *history windows* of 1.5 and 2 days, and four and nine signals. In all cases, delay times were approximately five minutes longer than the required duration of outlier values prior to CANARY detection as listed in Table 9 . The values listed in Table 9 represent the minimum delay, from the first detected outlier, associated with an *event threshold* of 0.9 and the listed *BED window* values. The additional 5-minute delay is attributed to the other factors listed above. Contaminants were injected into a pipe with water flowing with a linear velocity of 1 ft/s approximately 90 feet from the junction where water is diverted into the sensor stations (U.S. EPA, 2012a), corresponding to a delay of 90 seconds. Further delay can be attributed to the instrumentation delays, and delays caused by the flow from the main pipe to the sensors. Delay time was not significantly affected by changes to the *outlier threshold* or *history window* parameters. This highlights how CANARY's detection time is linked to the combination of *BED window* and *event threshold* values (see Table 9). The additional 5-minute delay is specific to the testing setup found at the T&E facility and cannot to be translated to other applications.

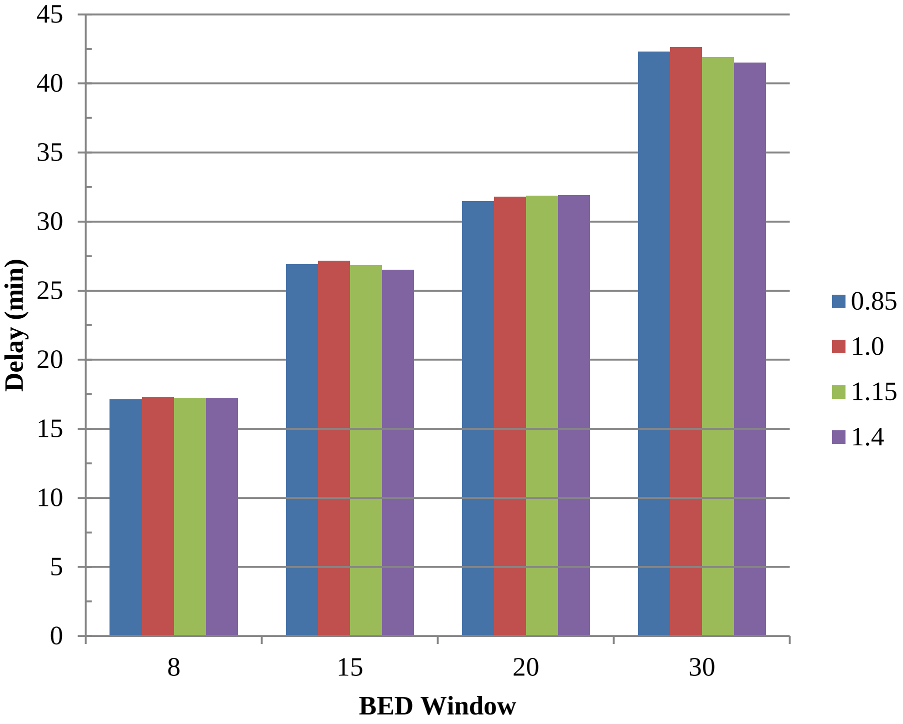


**Figure 8: Delay from contaminant injection to the event alarm for U.S. EPA Testing and Evaluation Facility data using the linear prediction coefficient filter (LPCF) algorithm. *Outlier threshold* values are shown as 0.85 in blue, 1.0 in red, 1.15 in green and 1.4 in purple. The values are the average of alarm rates with different number of signal and *history window* values.**

### 4.1.3 Number of Input Signals

The T&E facility evaluates many different sensors under real world conditions. This leads to an abundance of sensor data, and some overlaps in sensor signals due to comparisons between technologies or manufacturers. For example, two residual chlorine sensors and two pH sensors provided the analyzed data (see Table 3). Additional sensors were also in use at T&E but were excluded because they were not operational for the majority of the studied timeframe. Sensors that were included in the four signals analyses (pH, chlorine, specific conductivity and UVA as a surrogate for TOC) were selected because those signals have been previously shown to provide good detection rates for a wide variety of contaminants (Hall et al., 2010; Hall et al., 2009; U.S. EPA, 2012a); and to remove duplicate pH and chlorine sensors (those chosen had lower variability during the analyzed timeframe). In practice, most utilities might have from one to ten sensor signals available at a given sensor station; however, not all of the data might be useful for event detection.

Results in Figure 6 were broken down by the number of input signals. Dark colors within Figure 6 represent analyses with nine signals and their lighter counterparts represent those analyses with only four signals. Relative to nine input signals, the analyses that used only four input signals had fewer false alarms in all tested scenarios (see Figure 6).

With the exception of when the *BED window* was set to 30, there was no difference between the number of actual contamination events detected by either the four or nine sensor scenarios (see Table 7). All combinations of parameters that contained *BED window* values of 8, 15 or 20 were able to detect the same number of real contamination events for the four or nine input signal cases. In the case of the *BED window* parameter of 30, the runs with nine input signals were able to detect four more real contamination events relative to when using four inputs signals (i.e., three or four events detected for four inputs, and seven or eight events detected for nine signals). Examination of the alarm reports shows that the ORP sensor, which was not included when only using four sensors, had the longest deviation for the events that were only detected in the nine sensor case. The chlorine sensor also had a strong response to those injection events; however, that signal change alone did not increase the probability enough to exceed the *event threshold* – in other words, the signal change for that sensor was not long enough to trigger an alarm.

For practical reasons, most utilities install the fewest number of sensors necessary to capture water quality data that is relevant for each station. For the majority of CANARY parameter combinations, the scenario with four sensor signals was able to detect the same number of contamination events as the corresponding nine signal scenario, while reducing the number of false alarms. CANARY is able to cope with signal variability; however, it performs best with signals from good-quality and well-maintained sensors. Previous reports outline the effectiveness of various sensors for true contamination event detection (U.S. EPA, 2012a; Hall et al., 2007; Hall and Szabo, 2010; Szabo et al., 2008). The sensors that were found to respond to a large number of contaminants were specific conductivity, free chlorine, chloride, and ORP; TOC sensors had the best response to organic containing compounds (Hall et al., 2007; U.S. EPA, 2012a).

Based on the data in Table 7 and Table 8, and the above discussion of each parameter, the parameter combination that results in the fewest false alarms while still being able to detect 14 of the 14 true contamination events is: a *history window* of 1440 (2 days), a *BED window* of 15 (30 minutes), an *outlier threshold* of 1.4 and an *event threshold* of 0.9.

## 4.2 Singapore – PUB

Historical data from four stations within the Singapore Public Utility Board (PUB) water distribution system was analyzed for dates ranging from January 1, 2008 to August 31, 2008. These datasets are designated PUB1, PUB2, PUB3 and PUB4, to differentiate their originating location, and contain data from four water quality signals and two operational signals: water quality – free chlorine, pH, conductivity and turbidity, and operational – total output and pressure (see Table 3). No real contamination events were believed to have occurred during the analyzed date range at PUB datasets, so all alarms for these stations are classified as false alarms. CANARY was in development at the time the data was collected, and no actual CANARY alarms or problems at any of those sites should be inferred.

### 4.2.1 *History Window*

The recommended values of 1.5 to 2 days for the *history window* correspond to 432 and 576, respectively, for the 5-minute *data interval* used at PUB datasets. Figure 9 contains the average alarm per day results for the four PUB datasets investigated. These values are the average of all combinations with the same *history window*. The tested *history window* values range from half a day, 144, to a week, 2016. The turbidity signal was omitted from this analysis, leaving three signals. Tested *BED window* and *outlier threshold* values are listed in Table 6. Trends in Figure 9 are consistent across all four datasets. Increasing the *history window* leads to fewer alarms per day.

These results are consistent with previously reported conclusions that increasing the *history window* results in fewer alarms (Murray et al., 2010); as shown in Figure 10. As the *history window* increases, more data points are used to determine the baseline normal variability in the sensor signals. As more of the data is considered normal, CANARY produces alarms less frequently. These results are slightly different than those reported in section 4.1.1, where the number of alarms reaches a steady state and does not continue to decrease (see Figure 3). This is probably due to the low variability in the data at T&E compared to the PUB data. In T&E and PUB systems, *history windows* of 1.5 to 2 days provide good alarm behavior. Alarm behavior can be refined using the *BED window* and *outlier threshold* parameters, as discussed below. Full alarm results can be found in Appendix D: *Full Alarm Data*.

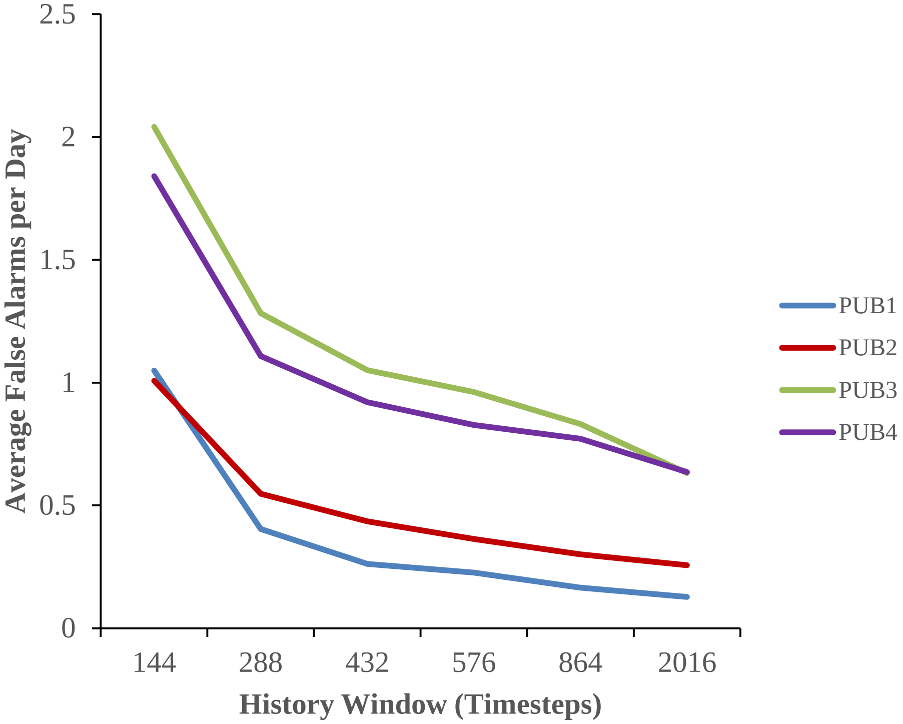


**Figure 9: Average false alarms per day when changing *history window* for Singapore Public Utility Board (PUB) dataset data. Blue corresponds to data from PUB1, red is PUB2, green is PUB3 and purple is PUB4. The values are the average of all combinations with the same *history window*.**

### 4.2.2 *BED Window* and *Outlier Threshold* Parameters

Results in the previous section showed a reduction in alarms per day with increasing *history windows*. This section focuses on alarm behavior when using *history windows* of 432 and 576, 1.5 and 2 days, respectively. Although there was a reduction in average alarm behavior when increasing the *history window*, there was not a significant decrease in alarms relative to the recommended 1.5 to 2 day range.

Table 10 contains a summary of the number of alarms reported by CANARY for the tested parameter combinations using the four water quality sensor signals. Each dataset is independent, so no direct comparisons can be made between datasets; however, general trends can be seen in all four datasets.

**Table 10: Summary of Total Alarms Reported by CANARY on Historical PUB Data from 01/01/2008 to 08/31/2008 Using the LPCF Algorithm and Four Sensor Signals**

| Dataset | *BED window* | *outlier threshold* | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0.85 | | 1.0 | | 1.15 | |
| | | *history window* | | | | | |
| | | 432 | 576 | 432 | 576 | 432 | 576 |
| PUB1 | 6 | 240 | 196 | 136 | 131 | 106 | 88 |
| | 8 | 198 | 167 | 114 | 110 | 88 | 73 |
| | 12 | 166 | 125 | 88 | 83 | 58 | 50 |
| PUB2 | 6 | 227 | 186 | 156 | 128 | 114 | 100 |
| | 8 | 214 | 168 | 146 | 123 | 107 | 93 |
| | 12 | 187 | 146 | 130 | 100 | 93 | 80 |
| PUB3 | 6 | 561 | 518 | 407 | 361 | 304 | 262 |
| | 8 | 510 | 470 | 375 | 330 | 283 | 248 |
| | 12 | 460 | 429 | 324 | 286 | 253 | 211 |
| PUB4 | 6 | 517 | 443 | 378 | 324 | 293 | 261 |
| | 8 | 469 | 405 | 343 | 301 | 265 | 229 |
| | 12 | 419 | 351 | 301 | 266 | 231 | 206 |

*BED, binomial event discriminator*; LPCF, linear prediction coefficient filter

The table shows clear trends in how each parameter affects the results. As the *outlier threshold* increases, the number of false alarms decreases. As the *history window* increases, the number of false alarms decreases. As the *BED window* increases, the number of false alarms decreases. These results are discussed in greater detail below.

Three *BED window* parameter values were used for PUB datasets in this study: 6, 8 and 12 timesteps. Table 11 summarizes the real timeframes analyzed and reports how the *event threshold* value of 0.9 relates to each *BED window* used. For example, a *BED window* of six is counting the number of outliers that have occurred within a 30-minute timeframe (window) for the 5-minute *data interval* used at PUB datasets. With an *event threshold* of 0.9, five out of six timesteps must be considered to be outliers by CANARY to trigger an alarm. This means that 25 minutes out of a 30 minute timeframe must be considered an outlier in order for CANARY to trigger an alarm.

**Table 11: *BED Window* Parameters and Number of Required Outliers for PUB Station Data with an *Event Threshold* of 0.9 with a 5-minute *Data Interval***

| ***BED* *window*** | ***BED* Timeframe (min)** | **Required outliers in *BED*** | **Duration of outliers (min)** |
|---|---|---|---|
| 6 | 30 | 5 | 25 |
| 8 | 40 | 6 | 30 |
| 12 | 60 | 8 | 40 |

*BED, binomial event discriminator; min, minutes*

Figure 10 contains the average number of alarms per day (averaged over both *history window* results) for each combination of *BED window* and *outlier threshold* for all PUB datasets. Blue corresponds to an *outlier threshold* value of 0.85, red is 1.0 and green is 1.15. The alarm rates for datasets PUB1 and PUB2 ranged from approximately one alarm per day to one alarm per five days (0.2–1 alarms per day). Alarm rates for the PUB3 and PUB4 datasets ranges from approximately one to two alarms per day. Increasing either the *BED window* or *outlier threshold* is seen to decrease the alarm rate in all datasets. Alarm rates decrease more due to an increase of the *outlier threshold* value when compared to increasing the *BED window* value.

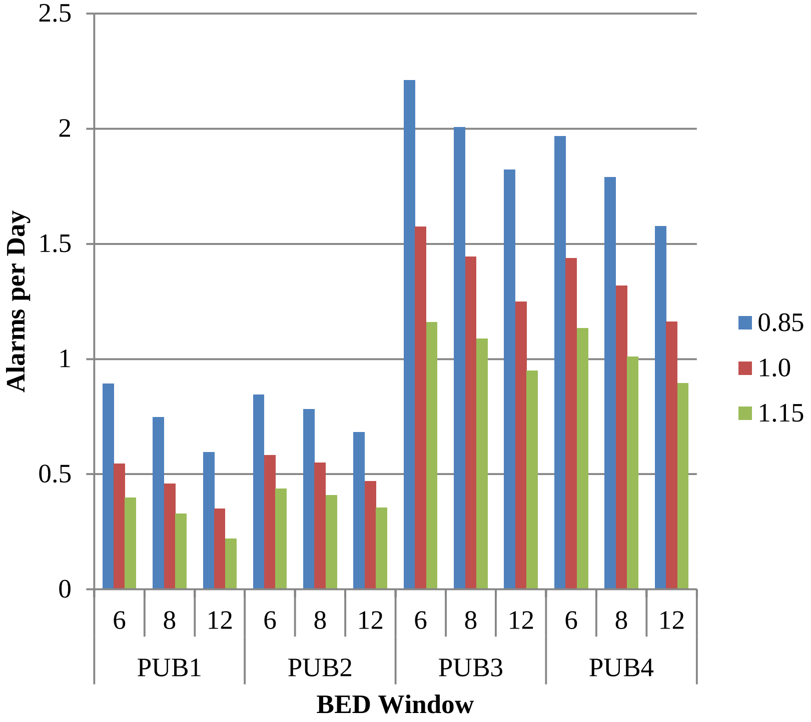


**Figure 10: Alarms per day for Singapore Public Utility Board (PUB) datasets using the linear prediction coefficient filter (LPCF) algorithm. Blue corresponds to an *outlier threshold* value of 0.85, red is 1.0 and green is 1.15. The values are the average of alarm rates with different *history window* values.**

### 4.2.3 Number of Signals

Previous testing involving PUB datasets suggested that the turbidity signal alone triggered a large number of false alarms. Removing the turbidity signal data resulted in an average alarm reduction of 28% for all four datasets (see Table 12). For datasets PUB3 and PUB4, that had a higher initial alarm rate, this corresponds to an average of 85 fewer alarms or one fewer alarm per three day period. The maximum reduction in total alarms was 124 fewer alarms for PUB3 when a *BED window* of 6, *history window* of 576 and *outlier threshold* of 0.85 were used. For these datasets, the turbidity signal contributes to approximately one-quarter of all alarms and could be an unreliable indicator of real events.

**Table 12: Summary of Total Alarms Reported by CANARY on Historical PUB Data from 01/01/2008 to 08/31/2008 using the LPCF Algorithm Without the Turbidity Sensor Signal**

| **Station** | ***BED window*** | ***outlier threshold*** | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0.85 | | 1.0 | | 1.15 | |
| | | ***history window*** | | | | | |
| | | 432 | 576 | 432 | 576 | 432 | 576 |
| PUB1 | 6 | 173 | 149 | 86 | 75 | 50 | 34 |
| | 8 | 145 | 125 | 65 | 65 | 37 | 24 |
| | 12 | 119 | 100 | 52 | 48 | 21 | 16 |
| PUB2 | 6 | 163 | 135 | 119 | 97 | 95 | 80 |
| | 8 | 153 | 125 | 112 | 91 | 88 | 76 |
| | 12 | 136 | 112 | 104 | 85 | 86 | 68 |
| PUB3 | 6 | 438 | 394 | 299 | 290 | 221 | 205 |
| | 8 | 407 | 357 | 284 | 221 | 213 | 195 |
| | 12 | 355 | 325 | 251 | 240 | 189 | 166 |
| PUB4 | 6 | 399 | 349 | 270 | 245 | 191 | 191 |
| | 8 | 362 | 307 | 245 | 227 | 178 | 173 |
| | 12 | 323 | 274 | 219 | 208 | 155 | 160 |

*BED, binomial event discriminator;* LPCF, linear prediction coefficient filter; PUB, Singapore Public Utility Board

Removing the turbidity signal resulted in a reduction in total alarms for all PUB datasets. Turbidity was also removed from the analysis of T&E data when using four sensors; however, no attempt was made to determine its specific effect on alarm behavior in that system.

Based on the data in Table 10 and Table 12, and the above discussion of each parameter, the parameter combination that results in the fewest total alarms in the PUB data are as follows: a *history window* of 576 (2 days), a *BED window* of 12 (60 minutes), an *outlier threshold* of 1.15 and an *event threshold* of 0.9.

# 5.0 Discussion

This section is divided into three parts: a comparison of results from both systems, a discussion about CANARY's runtime and the selection of recommended parameter values.

## 5.1 Comparison

To develop a fair comparison between the two data sets, the number of false alarms from T&E is compared to the number of total alarms of the PUB systems. No contamination events were believed to have occurred during the eight-month period analyzed at PUB, so all PUB alarms were thought to be false. No attempt was made to establish whether false alarms from any dataset had a valid root cause related to operational changes, or equipment malfunctioning or failure.

No comparison of the effect that *data interval* has on true contamination event detection can be made since there are no contamination events in the PUB data. Previous work suggests that a five-minute interval is likely the maximum *data interval* that is able to reliably detect true contamination events (U.S. EPA, 2013a; Allgeier et al., 2011b).

In both T&E and PUB datasets, alarm rates were more sensitive to changes in the *outlier threshold* when compared to the *BED window* and the *history window*. Increasing either *outlier threshold* or *BED window* resulted in fewer alarms. This trend in alarm reduction is the result of two different behaviors. When the *outlier threshold* is increased more signal variability is accepted as being normal and thus, not used in the event detection analysis. Increasing the *BED window* increases the number of timesteps that are used to define when an event is occurring; in other words, increasing the *BED window* results in more outliers being necessary to trigger an alarm. Increasing the *history window* also resulted in fewer alarms in the majority of cases, although the effect was smaller. Increasing the *history window* lengthens the timeframe over which the baseline behavior of a signal is established; this results in more normal signal variability being factored into the predictive algorithm, helping to reduce the number of alarms.

The primary difference between these two sets of data is that PUB sensors are located near source water or treatment facilities whereas and T&E sensors are located within the distribution system away from the treatment facility. Sharp quality changes that occur at a water processing or pumping station show a broader signal response further into the distribution system due to diffusion and mixing. Visual inspection of the PUB data reveals significant variability due to operational changes. Despite that variability, CANARY's alarm rates can be reduced to manageable values in all four datasets. This same variability was not present the T&E facility. Since no attempt was made to introduce fictitious events in the PUB data, it is not possible to determine from this analysis how the tested parameters would respond to contamination events; however, even the maximum alarm rate reported here should be acceptable by most utilities.

Sensor data analyzed here ranged from fairly stable, for T&E, to moderately variable, for PUB3 and PUB4 (see Table 3); however, results show that, for all datasets tested, it is possible to reduce alarm rates to below one alarm per day. For T&E, the alarm rate was reduced to approximately 0.1 alarms per day while maintaining the ability to detect 14 out of 14 real testing events.

## 5.2 Recommended Rule-of-Thumb Parameters

Based on the analysis presented in section 4.0 Results, the following parameter values are recommended (Table 13) as a good starting configuration, or rule-of-thumb parameters, for CANARY for a 2-minute or a 5-minute *data interval:*

**Table 13: Recommended Rule-of-Thumb Values for Parameters**

| Parameter | Recommended Values | |
|---|---|---|
| *Data interval* (minutes) | 2 | 5 |
| *History window* (number of data points) | 1440 | 576 |
| *Outlier threshold* (number of standard deviations) | 1.15 | 1.15 |
| *BED window* (number of data points) | 15 | 12 |
| *Event threshold* (probability) | 0.90 | 0.90 |

*BED, binomial event discriminator*

The rationale behind these parameters is as follows:

1. A two-day *history window* captures a baseline behavior that encompasses most day-to-day activity. This was the optimal setting for all five sensor datasets included in this report.
2. An *outlier threshold* of 1.15 treats 75% of signal variability as normal behavior, which allows some noise to be accepted as normal. A larger value resulted in fewer false positives, but also increased detection time and reduced the number of true events detected. This is the optimal setting for the four PUB datasets, and performed close to optimal for the T&E station.
3. For a *data interval* of 2 minutes, the *BED window* of 15 and *event threshold* of 0.9 combine to require 10 out of 15 timesteps to be outliers in order to produce an alarm. This requires that 2/3 of a *BED window* contains outliers before producing an alarm. This combination can capture events as short as 20 minutes, and can be detected as early as 24 minutes after the initial deviation from normal behavior. These parameters should be adjusted in tandem if the detection of shorter events is desired. Note that larger *BED window* values resulted in fewer false positives, but also increased detection time and reduced the number of true events detected. This was the optimal setting for the T&E station.
4. For a *data interval* of 5 minutes, the *BED window* of 12 and *event threshold* of 0.9 combine to require 8 out of 12 timesteps to be outliers in order to produce an alarm. This combination can capture events as short as 40 minutes. This was the optimal setting for the PUB datasets.

From the analysis reported in section 4, the rule-of-thumb parameters resulted in a false positive rate of 0.17/day for the T&E station and an alarm rate of 0.067/day for PUB1, 0.28/day for PUB2, 0.69/day for PUB3, and 0.67/day for PUB4 (see raw data in Table 7 with 4 signals for T&E, and Table 12 with 3 signals for PUB datasets). For the T&E station, these parameters detected 14 out of 14 true events.

### 5.2.1 Validation of Parameter Selection

In this subsection, the selection of the recommended rule-of-thumb parameter values is validated by using them to analyze a different data set. The original T&E data set used in the initial analysis contained water quality data collected from 11/01/2011 – 06/26/2012. A second data set is used here to validate the parameter settings; this data set extends these dates to two years, from 01/01/2011 – 12/31/2012.

A summary of the parameter set is as follows:

- Date range: 01/01/2011 – 12/31/2012
- Algorithm: LPCF
- *outlier threshold*: 1.15
- *history window*: 1440 (2 days)
- *BED window:* 15
- *event threshold:* 0.9
- *event timeout* 30
- *signals* chlorine, pH, conductivity, UVA

Table 14 summarizes the contamination testing events that occurred during the entire two year window (which includes the 14 events previously studied); a total of 74 contamination experiments were conducted. Using the parameters listed above to analyze the data with CANARY resulted in 130 total alarms, including 51 true detections (out of 74 possible contamination events for a 69% detection rate) and 79 false alarms (or 0.11 false alarms per day). (Appendix A provides a graphic that shows the water quality signal data for one day and the CANARY output in more detail.)

**Table 14: Summary of Testing Events and Detection by CANARY for Data from T&E between 1/1/2011 and 12/31/2012**

| ID | Time | Injection | Concentration | Note | DET | ID | Time | Injection | Concentration | Note | DET |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 03/15/2011<br>08:30 AM | KHP | 10 ppm | | DET | 20 | 04/21/2011<br>12:30 PM | Ethylene Glycol | 1 ppm | NVC | - |
| 2 | 03/15/2011<br>10:15 AM | KHP | 10 ppm | | DET | 21 | 04/21/2011<br>02:00 PM | Ethylene Glycol | 10 ppm | NVC | - |
| 3 | 03/15/2011<br>01:45 PM | KHP | 10 ppm | | DET | 22 | 04/21/2011<br>03:31 PM | Ethylene Glycol | 10 ppm | NVC | - |
| 4 | 04/12/2011<br>11:15 AM | *Escherichia coli* | $1.26\times10^4$ CFU/mL | | DET | 23 | 04/26/2011<br>02:00 PM | Coolant | 1 ppm | | - |
| 5 | 04/12/2011<br>12:40 PM | *E.coli* | $1.26\times10^4$ CFU/mL | | DET | 24 | 04/26/2011<br>03:30 PM | Coolant | 1 ppm | | - |
| 6 | 04/12/2011<br>02:15 PM | *E.coli* | $1.26\times10^4$ CFU/mL | | DET | 25 | 04/27/2011<br>10:00 AM | Coolant | 10 ppm | | DET |
| 7 | 04/14/2011<br>02:30 PM | Bleach | 4 ppm | | DET | 26 | 04/27/2011<br>11:30 AM | Coolant | 10 ppm | | DET |
| 8 | 04/15/2011<br>11:00 AM | Bleach | 4 ppm | | DET | 27 | 04/28/2011<br>10:00 AM | Deicer | 1 ppm | NVC | - |
| 9 | 04/15/2011<br>02:00 PM | Bleach | 4 ppm | | DET | 28 | 04/28/2011<br>11:30 AM | Deicer | 1 ppm | NVC | - |
| 10 | 04/19/2011<br>11:00 AM | *E.coli* | $5.74\times10^2$ CFU/mL | | DET | 29 | 04/28/2011<br>01:00 PM | Deicer | 10 ppm | | - |
| 11 | 04/19/2011<br>12:30 PM | *E.coli* | $5.74\times10^2$ CFU/mL | | DET | 30 | 04/28/2011<br>02:30 PM | Deicer | 10 ppm | | - |
| 12 | 04/19/2011<br>02:00 PM | *E.coli* | $5.74\times10^3$ CFU/mL | | DET | 31 | 05/02/2011<br>09:30 AM | Dispersant | 1 ppm | | DET |
| 13 | 04/19/2011<br>03:45 PM | *E.coli* | $5.74\times10^3$ CFU/mL | | DET | 32 | 05/02/2011<br>11:00 AM | Dispersant | 1 ppm | | DET |
| 14 | 04/20/2011<br>10:15 AM | Pepsin | 10,000 ppm | | DET | 33 | 05/02/2011<br>12:30 PM | Dispersant | 10 ppm | | DET |
| 15 | 04/20/2011<br>12:00 PM | Pepsin | 10,000 ppm | | DET | 34 | 05/02/2011<br>02:00 PM | Dispersant | 10 ppm | | DET |
| 16 | 04/20/2011<br>01:30 PM | Pepsin | 1,000 ppm | | DET | 35 | 05/03/2011<br>11:00 AM | Diesel Fuel | 1 ppm | | - |

| ID | Time | Injection | Concentration | Note | DET | ID | Time | Injection | Concentration | Note | DET |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 17 | 04/20/2011 03:00 PM | Pepsin | 1,000 ppm | | DET | 36 | 05/03/2011 12:30 PM | Diesel Fuel | 1 ppm | | - |
| 18 | 04/20/2011 04:30 PM | Sodium Thiosulfate | – | | DET | 37 | 05/03/2011 02:00 PM | Diesel Fuel | 5 ppm | | - |
| 19 | 04/21/2011 10:15 AM | Ethylene Glycol | 1 ppm | NVC | - | 38 | 05/03/2011 03:30 PM | Diesel Fuel | 5 ppm | | - |
| 39 | 08/15/2011 11:55 AM | *E.coli* | $1.2\times10^4$ CFU/mL | | DET | 57 | 11/15/2011 03:00 PM | *E.coli* | $1.0\times10^4$ CFU/mL | | DET |
| 40 | 08/15/2011 02:40 PM | *E.coli* | $1.2\times10^4$ CFU/mL | | DET | 58 | 11/28/2011 11:50 AM | KCN | 10 ppm | | DET |
| 41 | 08/15/2011 03:30 PM | *E.coli* | $1.2\times10^3$ CFU/mL | | DET | 59 | 11/28/2011 01:40 PM | KCN | 20 ppm | | DET |
| 42 | 08/18/2011 04:15 PM | *E.coli* | $1.15\times10^4$ CFU/mL | | DET | 60 | 11/28/2011 02:40 PM | Atrazine | 10 mg/L | | Merged |
| 43 | 08/19/2011 10:00 AM | Sodium Thiosulfate | – | | DET | 61 | 01/26/2012 11:40 AM | Atrazine | 10 mg/L | | DET |
| 44 | 08/19/2011 11:01 AM | *E.coli* | $1.15\times10^4$ CFU/mL | | DET | 62 | 01/26/2012 03:00 PM | Atrazine | 10 mg/L | | DET |
| 45 | 08/23/2011 11:00 PM | Sodium Thiosulfate | – | | DET | 63 | 02/01/2012 01:00 PM | Atrazine | 1 mg/L | | DET |
| 46 | 08/24/2011 03:45 PM | *E.coli* | $1.15\times10^5$ CFU/mL | | DET | 64 | 02/01/2012 02:30 PM | Atrazine | 1 mg/L | | DET |
| 47 | 08/25/2011 02:00 PM | *E.coli* | $1.15\times10^5$ CFU/mL | | DET | 65 | 02/02/2012 02:00 PM | Atrazine | 0.1 mg/L | | DET |
| 48 | 09/01/2011 09:20 AM | *B. subtilis* Spheres | Unknown | | DET | 66 | 02/02/2012 03:30 PM | Atrazine | 0.1 mg/L | | DET |
| 49 | 10/26/2011 10:00 AM | Sodium Basagran | 1 ppm | | DET | 67 | 09/20/2012 12:00 PM | *E.coli* | $1.0\times10^5$ CFU/mL | No Cl2 | - |
| 50 | 10/26/2011 12:30 PM | Sodium Basagran | 1 ppm | | DET | 68 | 09/24/2012 02:30 PM | *E.coli* | $1.0\times10^5$ CFU/mL | No Cl2 | - |
| 51 | 10/26/2011 02:00 PM | Sodium Basagran | 10 ppm | | DET | 69 | 10/16/2012 09:00 AM | *E.coli* | $1.09\times10^5$ CFU/mL | No data | - |
| 52 | 10/26/2011 03:30 PM | Sodium Basagran | 10 ppm | | DET | 70 | 10/16/2012 11:15 AM | *E.coli* | 1.09×105 CFU/mL | No data | - |
| 53 | 11/14/2011 10:30 AM | Sodium Thiosulfate | – | | DET | 71 | 10/16/2012 01:15 PM | *E.coli* | $1.09\times10^5$ CFU/mL | No data | - |
| 54 | 11/15/2011 10:30 AM | *E.coli* | $1.0\times10^3$ CFU/mL | | DET | 72 | 10/23/2012 01:20 PM | *E.coli* | $6.85\times10^4$ CFU/mL | No data | - |
| 55 | 11/15/2011 12:00 PM | *E.coli* | $1.0\times10^3$ CFU/mL | | DET | 73 | 10/24/2012 12:35 PM | *E.coli* | $1.37\times10^5$ CFU/mL | No Cl2 | - |
| 56 | 11/15/2011 01:30 PM | *E.coli* | $1.0\times10^4$ CFU/mL | | DET | 74 | 10/24/2012 02:35 PM | *E.coli* | $1.37\times10^5$ CFU/mL | No Cl2 | - |

NVC, No Visible Change; DET, Detected by CANARY; U.S. EPA Testing and Evaluation Facility. Merged = Event timeout was longer than the delay between injections, consequently CANARY was still producing an alarm from the original event when the second event had begun. KHP = Potassium Hydrogen Phthalate

A closer examination of the injection events shows that of the 74 total testing events, only 66 could have been captured by CANARY. Eight of the injections occurred during periods of missing sensor data, so no true or false detections can be established for those events. Of the 66 remaining events, 51 (77%) were captured by CANARY with these configuration settings. Two events (ID#59 and 60) were merged since they were injected only one hour apart, and the length of the *BED window* and the *event timeout* parameters caused these alarms to be merged into one. Two coolant injections (ID#23 & 24) were not detected; both had a concentration of 1 ppm and there was a small (~0.01 ppm), but visible, signal change to the UVA signal; however, this change would likely go unnoticed when plotting the data if the range was set from 0 to 2. In the case of the two undetected deicer events (ID#27 & 28), the tested concentration was low, which produced no visible change in any sensor signal.  Other deicer injections (ID#29 & 30) caused visible changes in the UVA sensor data, suggesting that the first two deicer injections were below the detection limit of the tested sensors.  Only the UVA signal responded significantly to the deicer injections, but this was not sufficient to trigger an alarm using the specified configuration parameters.  The sensors used in this analysis did not respond to capture ethylene glycol; however, previous work showed that TOC sensors were able to detect the presence of ethylene glycol (U.S. EPA, 2012a).  Injection ID#60 is listed as “merged,” because CANARY was still in an alarm state due to the previous injection.  Diesel fuel injections did lead to a change in the UVA signal, but this did not trigger an alarm with this parameter set.  Eight of the undetected events show discernable changes in UVA and, given the right configuration parameters, CANARY would be able to detect these events.

It would be possible to optimize CANARY’s performance further for this specific data set (in particular, to detect the eight events that changed the UVA signal); however, the purpose here is to demonstrate that the parameters selected for a subset of data also work quite well for a much larger set of data.

## 5.3 Runtime Analysis

The rule-of-thumb parameters developed in the last section provide a useful starting point for most utilities.  Some, however, would prefer to do a more in-depth analysis to select parameter values.  Some papers in the literature might have given the impression that such an analysis would require enormous computational capabilities and would take a significant amount of time (Murray et al., 2010; Rosen and Bartrand, 2013).  To show how long such an analysis might take, this section summarizes the runtimes associated with performing the analyses described in this report.  The runtimes for each analysis correspond to using CANARY’s batch mode.  This section reports the runtime per single analysis.  These values can be used to approximate the time it takes to perform multiple parameter testing.  Throughout this section, the word “instance” represents a single CANARY analysis; multiple analyses, or instances, can be run simultaneously on multi-core computers.  This results in a reduction in the total parameter testing time.

Figure 11 shows the average total runtimes for the all water quality stations broken down by the computer used to conduct the analysis (see section 3.2 Methods for a definition of the computer capabilities).  The longest runtimes were for systems with nine input streams on Computer 1 running two analyses at a time (1a).  Under those conditions, the eight month window took an average of approximately nine hours to analyze. Computer 2 was able to analyze the equivalent system in an average of approximately five hours. Figure 11 shows that increasing the number of parameters in the analysis increases the total runtime. Figure 11 also reveals an increased runtime associated with increasing the *history window* value.

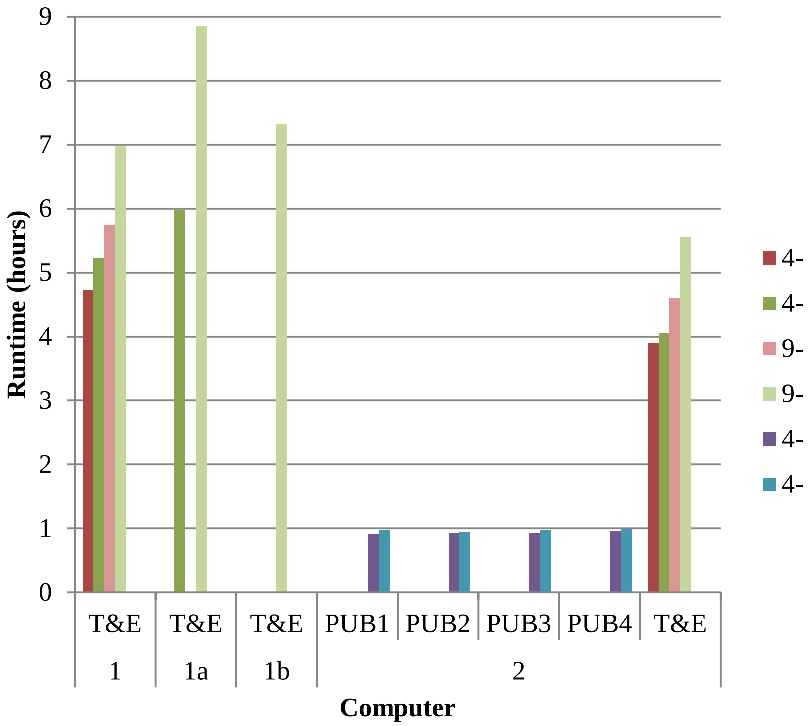


**Figure 11: Average total analysis runtime for stations using the linear prediction coefficient filter (LPCF) algorithm. For U.S. EPA Testing and Evaluation Facility, red indicates four signals and a *history window* of 1080 (i.e., 4- 1080), green indicates 4- 1440, pink indicates 9- 1080 and light green indicates 9- 1440. For Singapore Public Utility Board (PUB) datasets, purple indicates a *history window* of 432 and blue indicates 576.**

The average runtime using one algorithm for all PUB datasets was approximately one hour. PUB analyses required an average of 0.23 minutes per analyzed day (60 minutes/244 days) on Computer 2, whereas, analyses on T&E data on Computer 2 for four signals required an average of 1.0 minutes per analyzed day. The higher total runtime for T&E analyses can be attributed to the larger number of data points when using a two-minute *data interval*; there are two and a half times more data points when using a *data interval* of two-minutes relative to a five-minute interval. This suggests that analyzing one day of two-minute data would be expected to take two and a half times longer than five-minute data. Actual runtimes are closer to four times longer for similar four signal data streams. This suggests that *history window* values also play a role in runtimes.

The parameter combination testing outlined above utilized 18 to 20 combinations for each dataset. This provided a good picture of how CANARY performed for each system, and resulted in the rule-of-thumb parameter set. Users might want to conduct analyses on their own data to select configuration parameters. Runtimes discussed in this document were run with only a single algorithm per input file. CANARY is capable of accepting multiple algorithms per input file and reporting results for each of those algorithms in a single output file. This ability can be used to test multiple algorithms at a time, and might perform these tests in less time than if the tests were run individually. See the CANARY User's Manual (Hart and McKenna, 2012) for more information.

Based on the runtime results presented here for an eight-core machine (Computer 2), which is capable of running six instances of CANARY at once, the following timetables can be established to run 50 parameter combinations on eight months of data that use four input signals:

- **5-minute *data interval*** – Using four parameter combinations (four algorithms) per instance (this maintains a runtime of approximately one hour per algorithm) and six concurrent instances of CANARY produce results for 24 parameter combinations in four hours. This means that 50 combinations can be tested in approximately eight hours, much less than a single day. Fifty combinations should be sufficient to refine the input parameters in order to produce manageable alarm behavior.
  - A standard dual-core computer can run one or two instances of CANARY at a time. The same four algorithm per input file example (described above) would require approximately one to two days to analyze 50 combinations. This could be accomplished over a weekend.
- **2-minute *data interval*** – Using four parameter combinations (four algorithms) per instance and six concurrent instances of CANARY will produce alarm results for 24 combinations in approximately one day. Fifty combinations could be tested in two days.
  - Using a dual-core computer would limit the analysis to three algorithms per instance and two instances concurrently to produce six combinations per day. This would require approximately nine to ten days to analyze 50 combinations on a single dual-core computer. A single quad-core computer could run three concurrent instances, cutting the required time to approximately six days.

## 6.0 Optimization

The above discussion lays out a rationale for establishing rule-of-thumb parameters for any system. This rationale should help CANARY users reduce the number of unwanted alarms caused by noisy signals, without needing to run thousands of test runs to fully optimize their system. This section discusses a simplified optimization process for users who want to optimize their CANARY analysis. In addition, an example of the optimization process is given for the T&E data.

### 6.1 A Simple Optimization Protocol

If the rule-of-thumb parameters listed in section 5.2 produce too many false alarms, the following protocol can be used to understand how varying each parameter impacts alarm behavior and ultimately how to improve the configuration settings for a given sensor station. This protocol outlines a test for 48 parameter combinations that can be used to understand alarm behavior at a sensor station.

- ***BED window*** – A good starting point should correspond to 15 to 60 minutes of real time. Values should be incremented in steps of approximately 5–15 minutes of real time. It is worth considering what the shortest event duration of concern might be for a given application. The entire event does not have to be within a *BED window* in order to trigger an alarm; in fact, CANARY has a shorter alarm delay with shorter *BED windows*. Increasing *event threshold* and *outlier threshold* can reduce the number of alarms for a given *BED window,* while maintaining the benefits of having a shorter *BED window* (short delay). For 2-minute *data interval*, testing values between 8 and 20 will provide a good indication of how alarm behavior will change over this range. This range responds to events as short as approximately 15 minutes (for 8), or on the high end (20) have a delay of around 40 minutes from the first signal change.
- ***Event threshold*** – As previously discussed (see section 2.1), this parameter value works with the *BED window* to determine the number of outliers required to trigger an alarm. The number of required outliers is a subset of the *BED window*; in other words, if the *BED window* is six, then the number of required outliers can be from one to six. As such, the *BED window* and *event threshold* are intricately linked and a user should not vary the *event threshold* independently. In this example, running 100 variations of this parameter does not make sense because it only results in six distinct behaviors corresponding to the one to six required outliers; at most, six values of this parameter should be investigated. Instead of varying this parameter independently, *event threshold* should be calculated using the equation for minimum *event threshold* (see equation 1 in section 2.1). It is likely that most users only want alarms to occur when more than 50% of the data points in a *BED window* are considered outliers; for example, a user would only consider testing *event threshold* values relating to 4, 5 or 6 outliers are in a *BED window* of 6. For larger *BED windows*, the increment between the number of required outliers could be expanded to two or three in order to reduce the number of testing steps (e.g., 10, 12, or 15 out of a *BED window* of 15 may provide a good understanding of behavior).
    - It is worth mentioning again that *BED window* and *event threshold* work together. Increasing the percentage of required outliers within a *BED window* results in alarms that are trigged by contiguous sets of outliers. For example, if five out of six timesteps must be outliers to trigger an alarm (*event threshold*=0.9), they are likely to all be contiguous; however, if only 5 out of 10 timesteps must be outliers (*event threshold*=0.38), then an alarm could be triggered by noisy data. In most applications, a shorter *BED window* coupled

with a higher *event threshold* likely alarms only for events caused by contiguous outliers, while maintaining a short alarm delay.

- ***Outlier threshold*** – Table 15 shows the percentage of points that fall within ± ***x*** standard deviations of the mean on a normal probability curve, where ***x*** is the *outlier threshold*. These values help show how changing the *outlier threshold* changes the likelihood that a new signal value is considered normal. CANARY treats a point that fell within the range equal to ± ***x*** standard deviations as being part of the normal variability. A good starting point should be from 1 to 1.5. Incrementing this value by 0.01 only changes the amount of variation that is accepted as normal by an average of 0.27% within the *outlier threshold* range of 1 to 2. This small change would likely not lead to significantly different alarm behavior between most tested increments. For this reason, *outlier threshold* values should be incremented in steps of 0.1, 0.15 or 0.25. The Δ % column in Table 15 indicates the effective increase in the percent of data that is accepted as normal for each 0.25 step increase in *outlier threshold* (e.g., 16.4% more data that fits in a Gaussian distribution is accepted as normal when increasing *outlier threshold* from 0.5 to 0.75).

**Table 15: Effect of *Outlier Threshold* - Percentage of Data Surrounding the Mean that is Accepted as Normal Based on the *Outlier Threshold* Value (x)**

| ***outlier threshold*** | **% Treated as Normal** | **Δ %** |
|---|---|---|
| 0.5 | 38.3% | |
| 0.75 | 54.7% | 16.4% |
| 1 | 68.3% | 13.6% |
| 1.25 | 78.9% | 10.6% |
| 1.5 | 86.6% | 7.8% |
| 1.75 | 92.0% | 5.3% |
| 2 | 95.4% | 3.5% |
| 2.5 | 98.8% | 3.3% |
| 3 | 99.7% | 1.0% |

- ***History window*** – Users should begin with a *history window* equivalent to 1.5 or 2 days (based on their *data interval).* Adjustments to this parameter should be made in increments that are equivalent to one-quarter of a day or more. This value can be fixed for much of the optimization, with variations tested after other parameters have been optimized.

Using the approach described above, the optimization process should require analysis of fewer than 200 individual parameter configurations – in most cases, a logical approach should reduce this process to 100 or fewer tested variations. The optimization process can be simplified by remembering that slightly different combinations of parameters yield essentially the same alarm response; for example, requiring five outliers out of a *BED window* of seven might produce the same results as five outliers out of six. For this reason, initial testing can be done with larger tested increments to establish how parameters impact a system. Once large increments have been explored, smaller changes can be made to completely optimize the parameters.

Combinations of the following parameters provide a good initial understanding of how parameters might perform on a data set, values are listed for a station with a 2-minute *data interval*:

- *History window*: 1440 (2 days)
- *Outlier threshold*: 0.85, 1.0, 1.15, 1.3
- *BED window*: 8, 10, 12, 15
- *Event threshold*: Calculated using equation 1. These should correspond to N-0, N-1 and N-2 for each *BED window* (e.g., 6, 7 and 8 required outliers for *BED window* of 8 – 0.8555, 0.9649 and 0.9961, respectively).

This corresponds to 48 parameter combinations. After this initial set of tests, trends in behavior should be apparent, and an optimum set of parameters could be selected. If further refinement is still desired, more parameters can be tested based on these results. Further tests might only incrementally change one of the four parameters, focusing on the parameter that impacts that alarm behavior most. The set of parameters that results in the lowest number of alarms might not be the best set of parameters. If detecting short events, or events after only a short delay, is the primary goal of optimization, then further exploration of parameters using, for example, a *BED window* of 8 or 10, could provide the best match , even if a *BED window* of 15 results in fewer alarms.

To aid in this testing process, CANARY can be configured to run using multiple sets of parameters within a single configuration file. Each parameter set is defined as a different "algorithm" even though they might all use the same EDS algorithm (e.g., LPCF, see USEPA, 2013b for an example). In this way, it is possible to group optimization test steps and thereby, reduce the total number of CANARY runs and the computational runtime. Multi-algorithm configuration files have longer total runtimes, but can have better performance on a per test basis. Analyses for PUB datasets that included four to six algorithms had runtimes that were at least as fast or faster per algorithm than the corresponding single algorithm analysis. This merging of testing allows a user to set up a prolonged run that incorporates multiple tests without having to restart CANARY for each individual test or having to program a script to automate the starting of multiple runs. In this way, multiple analyses could be run overnight, or over a weekend, without much effort.

The goal of optimization is to minimize false alarms, not all alarms. This distinction is crucial to the optimization process. It is easy to eliminate all, or at least most, alarms; however, the resulting analyses would not yield good real-event detection. For example, as demonstrated in Section 4.1, increasing the *BED window* reduced false alarms, but it also reduced the number of true events detected and increased the detection delay. It is important to keep the shortest event duration that might be significant in mind as parameter selection, or testing, is being done. Low *BED window* values are recommended as they produce short delay-to-alarm times while maintaining the ability to detect short or long events. However the *BED window* should be long enough to contain several data points so that CANARY does not produce an alarm on a single outlier. Adjusting the *outlier threshold* and *event threshold* values help reduce false alarm rates.

The *Water Quality Event Detection System Challenge* (U.S. EPA, 2013a) utilized simulated events to test the true detection rate for a variety of EDSs. This test provides some guidance on how to superimpose simulated events on top of real signal data. In addition to simulated or real contamination events, it is important to consider events that could be linked to water quality, operations or treatment that an EDS might be able to detect (e.g., pipe break, chemical overfeed, or nitrification) (Hagar et al., 2013). Appendix B: *What Constitutes an Event?* discusses why an EDS might produce an alarm, which could help guide the parameter testing process by establishing when a disturbance is meaningful to a utility.

Users who chose to perform controlled testing in their system (with tracers or other materials) or artificially create events within their data should remember that the type of response selected biases CANARY's alarm behavior. A variety of concentrations and event durations should be tested to best understand how CANARY will behave for a variety of simulated events.

## 6.2 Optimization Example Using T&E Data

This section presents an example of the simplified optimization process using the 8-month 4-signal T&E data. As reported above, the 4-signal analyses were able to detect the same number of contamination events as their 9-signal counterparts for the majority of parameter sets. These signals were also felt to better represent the types of signals that would be present in a typical utility installation.

### 6.2.1 Methods

During this optimization process both the LPCF and multivariate nearest neighbor (MVNN) algorithms were used with each set of parameters. This is simply to demonstrate the several algorithms available within CANARY; however, LPCF is the most commonly used. A total of 48 parameter combinations were tested; each configuration had two algorithms, LPCF and MVNN, which produced a total of 96 combinations. The parameters tested are as follows:

| | |
|---|---|
| - Algorithms: | MVNN, LPCF |
| - *History window*: | 1080 |
| - *Outlier threshold*: | 0.85, 1.0, 1.15, 1.4 |
| - *BED window*: | 6, 8, 12, 15 |
| - *Event threshold*: | BED-1, BED-2, BED-3 |

Table 16 contains the specific *event threshold* that relates to the BED-# value. Values in parentheses in Table 16 are the number of required outliers in a given *BED window*. *BED window* values larger than 15 were not tested because initial testing (in section 4.1) revealed that above 15 the number of true detections began to decline. Only a *history window* of 1080 was selected for this example to simplify the discussion, and previous testing (see section 4.1) revealed that 1080 and 1440 performed similarly.

**Table 16: *Event Threshold* Values Used in Optimization Testing on T&E Data**

| ***event threshold*** | ***BED window*** | | | |
|---|---|---|---|---|
| | **6** | **8** | **12** | **15** |
| BED-1 | 0.8907 (5) | 0.9649 (7) | 0.9969 (11) | 0.9996 (14) |
| BED-2 | 0.6563 (4) | 0.8555 (6) | 0.9808 (10) | 0.9964 (13) |
| BED-3 | 0.3438 (3) | 0.6368 (5) | 0.9271 (9) | 0.9825 (12) |

*BED, binomial event discriminator*; U.S. EPA Testing and Evaluation Facility

### 6.2.2 Optimization Results & Discussion

Table 17 contains the number of false alarms and true detections reported by CANARY for the period of 11/01/2011 to 06/26/2012 for all parameter and algorithm combinations studied; the total number of alarms is the sum of these two values. In general, the number of false alarms produced by both LPCF and MVNN algorithms are similar. The exception is when using an *outlier threshold* of 1.0 and the MVNN algorithm. This dramatically increased the number of false alarms reported by CANARY. Examination of the alarm information showed that many of the extra alarms reported by CANARY were related to the conductivity signal. It was expected that using an *outlier threshold* of 1.0 would result in an alarm total between 0.85 and 1.15 false alarm totals, as was seen when using the LPCF algorithm. It is unclear why this *outlier threshold* value would increase the number of false alarms in this way.

**Table 17: False Alarms Reported by CANARY on T&E Data from 11/01/2011 to 06/26/2012**

| ***BED window*** | ***event threshold*** | ***outlier threshold*** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 0.85 | | 1.0 | | 1.15 | | 1.4 | |
| | | **Algorithm** | | | | | | | |
| | | LPCF | MVNN | LPCF | MVNN | LPCF | MVNN | LPCF | MVNN |
| | | **False Alarms (True Contaminant Detections)** | | | | | | | |
| 6 | 0.3438 | 118 (14) | 110 (12) | 59 (14) | 1138 (10) | 37 (14) | 50 (12) | 26 (14) | 40 (10) |
| | 0.6563 | 68 (14) | 63 (12) | 42 (14) | 475 (10) | 30 (14) | 38 (10) | 25 (14) | 31 (10) |
| | 0.8907 | 47 (14) | 50 (12) | 38 (14) | 330 (10) | 29 (14) | 36 (10) | 24 (14) | 30 (10) |
| 8 | 0.6368 | 56 (14) | 55 (12) | 39 (14) | 430 (10) | 29 (14) | 37 (10) | 24 (14) | 31 (10) |
| | 0.8555 | 44 (14) | 48 (12) | 36 (14) | 325 (10) | 27 (14) | 33 (10) | 22 (14) | 26 (10) |
| | 0.9649 | 38 (14) | 41 (11) | 33 (14) | 289 (11) | 25 (14) | 30 (10) | 22 (14) | 24 (10) |
| 12 | 0.9271 | 36 (14) | 39 (10) | 32 (14) | 283 (9) | 24 (14) | 28 (10) | 20 (14) | 23 (10) |
| | 0.9808 | 34 (14) | 38 (10) | 32 (14) | 270 (10) | 23 (14) | 26 (10) | 20 (14) | 23 (10) |
| | 0.9969 | 33 (12) | 38 (8) | 31 (12) | 265 (8) | 22 (12) | 26 (8) | 19 (12) | 23 (8) |
| 15 | 0.9825 | 32 (10) | 37 (7) | 28 (10) | 270 (6) | 23 (10) | 26 (6) | 19 (10) | 23 (6) |
| | 0.9964 | 32 (10) | 36 (6) | 28 (10) | 263 (5) | 22 (10) | 26 (6) | 19 (10) | 23 (6) |
| | 0.9996 | 32 (9) | 36 (6) | 28 (8) | 254 (4) | 22 (7) | 26 (5) | 19 (6) | 23 (4) |

*BED, binomial event discriminator;* LPCF, linear prediction coefficient filter; MVNN, multivariate nearest neighbor; U.S. EPA Testing and Evaluation Facility

Figure 12 shows the false alarms per day reported by CANARY on T&E data from 11/01/2011 to 06/26/2012. The minimum number of false alarms occurred when using an *outlier threshold* value of 1.4 (purple bars), a *BED window* of 12 or 14, and the LPCF algorithm; in this case, the false alarm rate was reduced to 0.08 false alarms per day or 19 false alarms over the 8-month studied period.

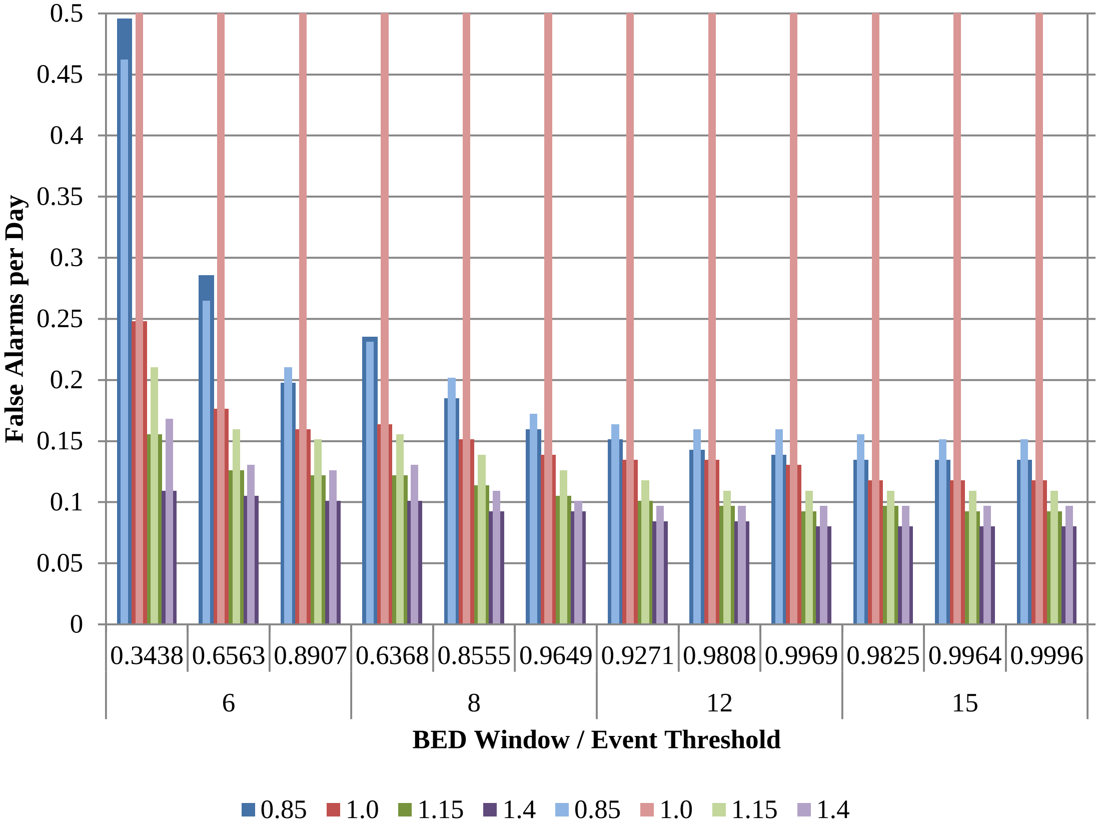


**Figure 12: False alarms per day reported by CANARY on U.S. EPA Testing and Evaluation Facility data from 11/01/2011 to 06/26/2012. Blue represents an *outlier threshold* value of 0.85, red is 1.0, green is 1.15 and purple is 1.4. Dark colors indicate results from the linear prediction coefficient filter (LPCF) algorithm and lighter shades (overlaid) indicate results from the multivariate nearest neighbor (MVNN) algorithm.**

Figure 13 shows the number of true detections by CANARY on T&E data from 11/01/2011 to 06/26/2012 for parameter and algorithm combinations. This shows that the LPCF algorithm is able to detect 14 out of 14 contamination events for all parameter combinations up to a *BED window* of 12 and an *event threshold* of 0.9808. The corresponding false alarm rate for this combination is 0.084 false alarms per day (or 20 false alarms in the studied period). Note that this same low alarm rate was reached during the initial testing (section 4.1) when using a *BED window* of 30 and *event threshold* of 0.9; however, the true detection total dropped to only 3 detected events (see Table 7). This highlights that the false alarm rate can be reduced with a short *BED window* when the *event threshold* value is high enough to require more contiguous outliers to trigger an alarm.

The reduction in true detections beyond a *BED window* of 12 and an *event threshold* of 0.9808 can be attributed to the testing conditions at T&E. As noted in the previous discussion concerning T&E results (see section 4.1), the average contamination event duration was 20 minutes, or 10 timesteps. The combination of the *BED window* of 12 and *event threshold* of 0.9808 corresponds to 10 out of 12 timesteps must be an outlier to trigger an alarm. Any combination of *BED window* and *event threshold* that require more than 10 data points to be outliers results in lower detection rates.

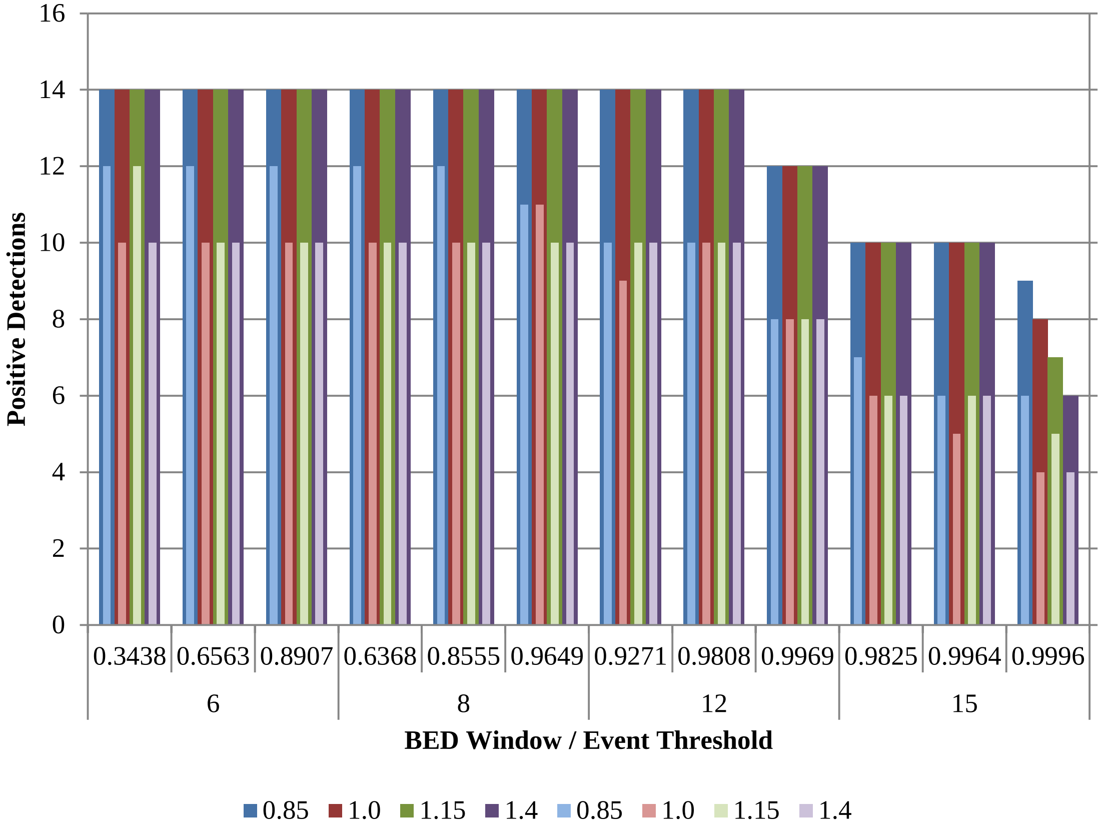


**Figure 13: True detections reported by CANARY on U.S. EPA Testing and Evaluation Facility data from 11/01/2011 to 06/26/2012. Blue represents an *outlier threshold* value of 0.85, red is 1.0, green is 1.15 and purple is 1.4. Dark colors indicate results from the linear prediction coefficient filter (LPCF) algorithm and lighter shades (overlaid) indicate results from the multivariate nearest neighbor (MVNN) algorithm.**

Results from the MVNN algorithm (Figure 12, Figure 13 and Table 17) contain similar trends to those produced by the LPCF algorithm. In general, increasing the *outlier threshold* parameter decreased the number of alarms produced. One obvious difference relates to all analyses that use an *outlier threshold* of 1.0 (light red in Figure 12). The total number of alarms when using MVNN and an *outlier threshold* of 1.0 resulted in a 10-fold increase in alarms, relative to the LPCF counterpart. The alarm values when using the MVNN algorithm and an *outlier threshold* of 1.0 were expected to fall between the 0.85 and 1.15 results, as was seen when using LPCF; no explanation for this behavior is available at this time.

The number of true detections reported by CANARY using MVNN was lower with all tested parameter combinations relative to LPCF. The maximum number of true detections by the MVNN algorithm was 12 out of 14 contaminations events; two fewer than when the LPCF algorithm was used. Using the MVNN algorithm, the false alarm rate can be reduced to 0.2 false alarms per day while maintaining the ability to capture 12 out of 14 contamination events. Additionally, no combination that included an *outlier threshold* of 1.0 exceeded a true detection capability of 11 out of 14 contamination events, despite the high number of false alarms present.

It is important to note that these results do not show that the LPCF algorithm is better than the MVNN algorithm for all systems; only that the LPCF algorithm performed better for the combinations of parameters tested for the T&E system. Previous testing also showed that the percentage of true detections differed between these algorithms (Murray et al., 2010). Although the LPCF algorithm was able to detect a higher percentage of events for two of the three locations used in previous testing, the MVNN algorithm was able to detect a higher percentage for the third location (Murray et al., 2010). The MVNN and LPCF algorithms were able to detect 80% or more of the short contamination events present in the T&E data for the eight month period that was analyzed.

This parameter optimization case study considered 48 combinations. The total runtime for this analysis was 52 hours. These results suggest that the optimal configuration would use the LPCF algorithm with a *history window* of 1.5 days, a *BED window* of 12, an *outlier threshold* of 1.4, and an *event threshold* of 0.9271-0.9808. This combination results in 14 out of 14 events detected and 20 false positives over the 8-month period (0.084 false alarms per day or about one alarm every 12 days).

Table 18 shows how these results compare to the rule-of-thumb parameter settings. The number of false alarms is reduced in half by this optimization procedure; however, the rate was already low at only one alarm about every 6 days (0.17/day).

**Table 18: Rule-of-Thumb vs. Optimized Parameter Performance**

| **Parameters** | **Rule-of-thumb parameter values** | **Optimized parameter values** |
|---|---|---|
| *History window* | 1440 | 1080 |
| *BED window* | 15 | 12 |
| *Outlier threshold* | 1.15 | 1.4 |
| *Event threshold* | 0.9 | 0.9808 |
| **Output metrics** | | |
| Number of false alarms (rate) | 41 (0.17/day) | 20 (0.084/day) |
| Number of events detected | 14 out of 14 | 14 out of 14 |

*BED, binomial event discriminator*

## 7.0 Conclusion

CANARY is a powerful and customizable EDS software, capable of differentiating real water quality events from background variability. It requires a configuration file to get started, and some of the statistical parameters can be difficult for new users to determine; in particular, the *BED window*, *outlier threshold, event threshold*, and the *history window*.

In section 2, an in-depth discussion of these four parameters was presented in order to provide a more intuitive understanding of each parameter. The *history window* is a moving time period over which historical data is used to calculate the baseline variability of a water quality signal. *The outlier threshold* is the number of standard deviations away from the mean the signal data must be in order to be declared an outlier and potentially indicate an event. *The BED window* is a moving time period over which signal data is examined to look for the onset of events, and helps to reduce false positives by eliminating alarms from single data outliers. *The event threshold* is a probability that, if exceeded by CANARY's event probability, indicates an event has occurred. The *BED window* and *event threshold* are linked and should not be considered independent of one another.

In sections 3 and 4, an analysis of data from five real-world sensor datasets is conducted in order to investigate their effect on CANARY's alarm behavior. Increasing the *BED window*, *outlier threshold* or *event threshold* parameters was shown to decrease the number of alarms generated by CANARY in all five datasets. Values of the *history window* parameter from 1.5 to 2 days generally minimized alarm rates. The number and type of signals also impacts results: removing certain water quality signals from analyses, specifically turbidity, resulted in fewer alarms in all datasets. Overall, alarm rates were most sensitive to the *outlier threshold* parameter.

A set of rule-of-thumb configuration parameters was developed and are recommended as a starting point for new users of CANARY, see

Table 13 in section 5.2. Using these parameters, CANARY was successfully able to detect 14 out of 14 real contamination events that occurred during the analyzed timeframe for the T&E facility while the number of false alarms was 41 (0.17 alarms per day). For PUB1 and PUB2, alarm rates were reduced below 0.5 alarms per day. For PUB3 and PUB4, alarm rates were reduced below one alarm per day.

A simple optimization procedure is outlined in section 6 to aid in testing more parameter combinations if the rule-of-thumb parameters produce too many alarms. This procedure is tested on the T&E data, and an improved configuration is generated that reduces the number of false alarms to 20 (or 0.083 alarms per day) while maintaining the ability to detect all test contaminant injections.

## References

Allgeier, S. C., Haas, A. J., Pickard, B. C. (2011a). “Optimizing Alert Occurrence in the Cincinnati Contamination Warning System.” *Journal American Water Works Association*, 103(10), p. 55–66.

Allgeier, S., USEPA-OGWDW, Korbe, C., Salomons, E., Edthofer, F., McKenna, S., Zach-Maor, A. (2011b). “The Impact of Polling Frequency on Water Quality Event Detection.” *Proceedings of AWWA Water Quality Technology Conference*, Phoenix, AZ. p. 1497.

Hagar, J., Murray, R., Haxton, T., Hall, J., McKenna, S. (2013) “The Potential for Using the CANARY Event Detection Software to Enhance Security and Improve Water Quality.” *Proceedings, Environmental & Water Resources Institute*, May 19-23, 2013, Cincinnati, OH.

Hall, J. Szabo, J. (2010). “On-line Water Quality Monitoring in Drinking Water Distribution Systems: A Summary Report of USEPA Research and Best Practices.” *Journal American Water Works Association*, 102 (8), p. 20-22.

Hall, J. S., Szabo, J. G., Panguluri, S., Meiners, G. (2009). *Distribution System Water Quality Monitoring: Sensor Technology Evaluation Methodology and Results – A Guide for Sensor Manufacturers and Water Utilities.* Washington, DC: U.S. Environmental Protection Agency. EPA/600/R-09/076. [http://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=212368 Accessed October 30, 2013]

Hall, J., Zaffiro, A. D., Marx, R. B., Kefauver, P. C., Krishnan, E. R., Radha, E., Haught, R. C., Herrmann, J. G., (2007). “On-line Water Quality Parameters as Indicators of Distribution System Contamination.” *Journal American Water Works Association*, 99 (1), p. 66-77.

Hart, D. B., McKenna, S.A. (2012). *CANARY User’s Manual: Version 4.3.2*. Washington, DC: U.S. Environmental Protection Agency. EPA/600/R-08/040B. [http://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=253555 Accessed October 1, 2013]

Kertesz, R., Burkhardt, J., Panguluri, S. (2014). “Real-time analysis of moisture and flow data to describe wet weather response in a permeable pavement parking lot.” *Proceedings of the World Environmental and Water Resources Congress*, Portland OR, June 1-5, 2014.

Morley, K., Janke, R., Murray, R., Fox, K. (2007). “Drinking Water Contamination Warning Systems: Water Utilities Driving Water Security Research.” *Journal American Water Works Association*, 99(6), p. 40-46.

Murray, R., Haxton, T., McKenna, S.A., Hart, D. B., Klise, K., Koch, M., Vugrin, E. D., Martin, S., Wilson, M., Cruz, V., Cutler, L. (2010). *Water Quality Event Detection Systems for Drinking Water Contamination Warning Systems: Development, Testing, and Application of CANARY*. Cincinnati, OH: U.S. Environmental Protection Agency. EPA/600/R-10/036. [http://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=221394 Accessed October 1, 2013]

Pickard, B. C., Haas, A. J., Allgeier, S. C. (2011). “Optimizing Operational Reliability of the Cincinnati Contamination Warning System.” *Journal American Water Works Association*, 103(1), p. 60–68.

Rosen, J., Bartrand, T. (2013). ”Using Online Water Quality Data to Detect Events in a Distribution System.” Journal of the American Water Works Association 105(7), p. 22-26, July 2013.

Szabo, J. G., Hall, J. S., Meiners, G. (2008). "Sensor Response to Contamination in Chloraminated Drinking Water." *Journal American Water Works Association*, 100 (4), p. 33-40.

U.S. EPA (2008). *Interim Guidance on Developing an Operational Strategy for Contamination Warning Systems.* September 2008. Washington DC: U.S. Environmental Protection Agency, Office of Water. EPA 817-R-08-002. [http://water.epa.gov/infrastructure/watersecurity/upload/2008_10_24_watersecurity_pubs_guide_interim_operational_strategy_wsi.pdf Accessed August 26, 2014]

U.S. EPA (2012a). *Detection of Contamination in Drinking Water Using Fluorescence and Light Adsorption Based Online Sensor*. Cincinnati, OH: U.S. Environmental Protection Agency. EPA/600/R-12/672. [http://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=246568 Accessed October 31, 2013]

U.S. EPA (2012b). *CANARY Quick Start Guide*, Washington, DC: U.S. Environmental Protection Agency. EPA/600/R-12/010.

U.S. EPA (2013a). *Water Quality Event Detection System Challenge: Methodology and Findings*. Washington, DC: U.S. Environmental Protection Agency. EPA/817/R-13/002. [http://water.epa.gov/infrastructure/watersecurity/lawsregs/upload/epa817r13002.pdf Accessed October 1, 2013]

U.S. EPA (2013b). *CANARY Training Tutorials*. Washington, DC: U.S. Environmental Protection Agency. EPA/600/R-13/201. [http://cfpub.epa.gov/si/si_public_file_download.cfm?p_download_id=516747]

## Appendix A: One-Day Detail for Contamination Detection for T&E

This graphic shows the water quality signal data and CANARY output on one day in 2012 in order to demonstrate CANARY detection of an event in more detail.

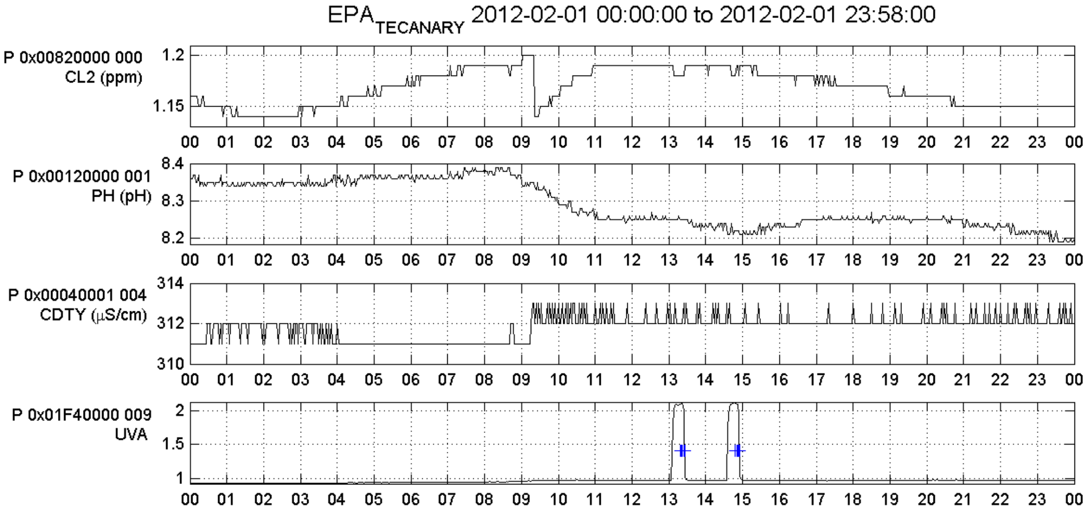


**Figure 14: CANARY output for 02/01/2012 with four signals and two linear prediction coefficient filter (LPCF) algorithms from U.S. EPA Testing and Evaluation Facility (T&E) data.**

Figure 14 shows water quality data collected at T&E on 2/1/2012, including chlorine, pH, conductivity, and UVA signals. On that day, two experiments with the chemical Atrazine were conducted, one at 1:00 PM and the second at 2:30 PM. These events caused visible changes in the UVA signal and slight decreases in the chlorine signal. Two events are detected (as shown by the blue dots) caused by rapid changes in the UVA data.

## Appendix B: What Constitutes an Event?

In order to be consistent with other CANARY documentations and studies, this report uses the term “false alarms” to describe alarms that do not correspond to known contamination events. The EPA report for the event detection system (EDS) Challenge uses the term “invalid” to describe alarms that did not have a known origin in the challenge data (U.S. EPA, 2013). The report’s corresponding discussion highlights that a false or invalid alarm might have a valid root cause that was not known by the challenge designers.

A utility might wish to reduce the total number of alarms, and the associated follow-up investigations; however, it is important to establish why an EDS might produce an alarm in order to gain the most value out of an EDS. Hagar et al. (2013) discuss the likelihood of CANARY, or another EDS, detecting six types of common water distribution system disruptions or issues that might be of interest to a utility (summarized in Table 19).

The value of any EDS is found in its ability to detect changes in sensor signals. Some EDSs focus on detecting when a signal exceeds a fixed bound (set-point), while others attempt to predict signal behavior to produce better results. A signal change that is large enough to trigger an alarm from an EDS may have an underlying cause that has significance to a utility, even if it is not associated with a malicious event.

In general, the reason an EDS might produce an alarm fall into five categories:

1. material additions from an outside source (contamination event, cross-connection)
2. component failure (pipe breaks, sensor failure or loss-of-calibration)
3. operational changes within the system (variability of finished water quality)
4. variability of source-water quality
5. signal noise

Of those listed reasons, only signal noise (5) generates truly false alarms. It is possible to dramatically reduce, or eliminate, alarms caused by a noisy signal; this can be done within CANARY, which does not require any pretreatment of the signal data. This leaves four reasons that an EDS would produce an alarm.

For the purpose of this report, real events were classified as being those related to material additions from an unknown source (1). At the T&E facility, contamination injections are performed in order to test the response of an EDS or sensor to contaminants and their ability to detect contamination events. From a water security perspective, these material additions would be events that should be detected.

Utilities might also wish to detect events associated with component failures (2). Pipe breaks, specifically, could result in customer complaints and disruption in the water supply network. Early detection of pipe breaks will result in a prompt response and limit disruptions to the system. Additionally, an EDS is only as good as the sensor data that it is analyzing, so alarms that might be indicative of sensor problems could be equally desirable.

The detection of the two remaining reasons an EDS might produce an alarm, operational changes and variability of source water, might or might not be desired by a utility. Operational changes and water quality variability can trigger alarms because they will cause a change in the sensor signal at a station. These types of alarms could be considered invalid because they do not relate to a security concern. On the one hand, setting CANARY to detect such events would provide confidence that CANARY is working, and would detect rarer events (such as 1 and 2). On the other hand, an operational change or variability in source or finished water quality that does not exceed set-points is probably anticipated and therefore the operators do not need to be alerted to such a change.

Some utilities might wish to monitor the water quality of source water. In some communities, the source water is the drinking water, so this is also their finished water quality; whereas, for other utilities, knowledge about the source water quality might aid in the treatment process. Water quality at a source could change because of a contamination event, which would be considered a valid alarm; however, source water quality could change due to rainfall or other natural events might be considered an invalid event. Monitoring source water quality might produce some valid alarms; however, source water quality could produce a higher number of invalid alarms. A contaminant found in source water might be removed as part of the treatment process. If the contaminant is removed during the treatment process, then the treatment has worked; if it was not, then any signal change associated with that contaminant should occur in the finished water's signal as well as a source-water's signal.

While changes in source water quality could indicate valid events, many utilities might consider operational changes to be invalid alarms. These changes include water treatment chemical additions or switching, or mixing, of water sources. They would be considered invalid alarms when they do not exceed set-points for a water quality parameter. These alarms might provide useful information to a utility. A chemical overfeed will trigger an alarm because it results in changes in water quality signals. These could possibly be caused by a control system or valve problem, and a utility would wish to correct this because it will reduce their chemical usage. Mixing two finished water sources would likely also be considered an invalid alarm; however, if the water quality of these sources is different an EDS can produce an alarm. Alarms related to mixing events could also be valuable, in that they verify that a system change occurred. For example, if additional water is pumped into the network from a storage tank during peak demand periods, an alarm might occur at the beginning of this process change; this alarm would validate that a control system change had occurred in the system. Control systems might include built in system integrity checks; however, this type of alarm would provide a secondary verification of a system change. Additionally, a decrease in the quality of water in a storage tank could indicate a problem in that tank.

An EDS is designed to produce an alarm whenever a signal change meets certain criteria. For set-point EDSs, the criteria would be whether a signal exceeds a set-point range. For algorithms within CANARY that utilize signal prediction, the criteria are whether a signal deviates from a predicted behavior and for how long. If a signal deviates from a normal behavior for a sufficiently long time, then CANARY will produce an alarm. Many normal operational changes will be captured as normal behavior; however, some changes might trigger an alarm.

The preceding discussion outlined why an EDS might produce an alarm, but is not meant to suggest that normal operations will always trigger alarms; however, there could be some unrealized value in invalid alarms. Alarms associated with noisy signals can be reduced or eliminated within CANARY by altering configuration parameters. Valid alarms would be related to the detection of contamination events and system disruptions (i.e., pipe breaks). Utilities could choose to consider alarms related to source water quality changes or operational changes as valid or invalid depending on how they hope to use their EDS.

It is possible to reduce invalid or false alarms reported by CANARY while maintaining a high true detection rate. A parameter combination that eliminates all events triggered by normal operational variability will likely drastically reduce or eliminate CANARY's ability to detect true contamination events. Furthermore, these invalid alarms might provide valuable information to utilities.

**Table 19: Summary of Sensor Responses and Likelihood of Detection for Common Issues in Water Distribution Systems**

| Incident Type | Free Chlorine | Total Chlorine | pH | Turbidity | Conductivity | ORP[a] | TOC[b] | Temperature | DO[c] | Ammonia | Pressure | Total dissolved solids | Time Scale | EDS Detection Potential § |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pipe Break | – | † | | + | | | | | | | – | | Hours | High |
| Corrosion | – | | – | | | | | | + | | | + | Weeks to years | Low |
| Cross-connection | – | | | + | | | | ± | | | – | | Hours | Medium |
| Nitrification | | – | – | | | | | + | – | + | | | Weeks | Medium |
| Decay in Water Quality | – | | | | | – | – | + | | | | | Days to months | Medium |
| Pressure Transients | | | | | | | | | | | – | | Seconds to minutes | Low* |
| Caustic Overfeed | + | + | – | | + | | | | | | | | Hours | High |
| Disinfectant Overfeed | + | + | + | + | | | | | | | | | Hours | High |

[a] Oxidation Reduction Potential, [b] Total Organic Carbon, [c] Dissolved Oxygen

† Shading = not likely a viable indicator of that type of incident

§ The environmental detection system (EDS) detection potential of each type of event is predicted based on the strength of signal responses, proximity of an event to a sensor station, length of a typical event and the responsiveness of the EDS. (Information consolidated from Hagar et al., (2013) "The Potential for Using the CANARY Event Detection Software to Enhance Security and Improve Water Quality." *Proceedings, Environmental & Water Resources Institute*, Cincinnati, OH.)

* Pressure transients are listed as having a low potential for detection because the duration is so short. These could potentially be detected; however, it is likely that false alarm rates would increase in order to capture these transient events. If desired, an analysis of only a pressure signal might provide utilities the ability to detect these pressure events at a higher likelihood.

## REFERENCES

Hagar, J., Murray, R., Haxton, T., Hall, J., McKenna, S. (2013) "The Potential for Using the CANARY Event Detection Software to Enhance Security and Improve Water Quality." *Proceedings, Environmental & Water Resources Institute*, May 19-23, 2013, Cincinnati, OH.

U.S. EPA (2013). *Water Quality Event Detection System Challenge: Methodology and Findings*, Washington, DC: U.S. Environmental Protection Agency. EPA/817/R-13/002. [http://water.epa.gov/infrastructure/watersecurity/lawsregs/upload/epa817r13002.pdf Accessed October 1, 2013]

# Appendix C: Event Detection Software Challenge

In addition to the analyses described in this report, data from the *Water Quality Event Detection Software Challenge* (Challenge) (U.S. EPA, 2013) was also analyzed using the same methods described in the optimization section. The Challenge data included simulated events (in comparison to the real contamination experiments conducted at T&E); these events simulated sensor responses for six different contaminants, at four different times, with two different concentrations and two different durations for a total of 96 simulated events per station. Refer to the original Challenge report for a full discussion on how the Challenge was conducted, including how events were simulated (U.S. EPA, 2013).

During the EDS Challenge, water quality sensor signal data for six stations was given to each EDS team; this data did not contain any simulated events and was assumed to be free of any type of water quality events. The teams provided a set of configuration parameters back to the designers of the Challenge to be used in the actual testing, based on this training data. The teams involved were not aware of the specific nature of the simulated events.

As the Challenge provided a rich set of simulated events with which true detection rates can be measured, this appendix applies the rule-of-thumb parameters to the Challenge data in order to provide more evidence that these parameters can be used to detect events. The results in this section should not be compared directly to the Challenge results as the same rigorous testing approach was not used here. In particular, this appendix focuses on true detection rates, but does not analyze all the data needed to produce comparable false alarm rates.

The rule-of-thumb parameters used in this analysis were selected based on the process described in this report (and its results) without any consideration of the Challenge data. In addition to the rule-of-thumb parameter testing, an optimization process was undertaken to find out how many true detections would be found for each of the six stations. This analysis focused on two questions: (1) how many true detections were possible, and (2) how many parameters were able to maintain a high level of true detections. These results are presented here as further testing for the configuration parameter selection process described above, and the optimization process described in Section 6.0, for a dataset that includes simulated events.

### C.1 Methods

The results discussed in this appendix focus on analyses performed on previously published Challenge data (U.S. EPA, 2013). Challenge data was analyzed by CANARY using data from six stations (designated A, B, D, E, F and G).

The following parameters were selected for the initial assessment of how CANARY performs with Challenge data:

| | |
|---|---|
| Name: | *Initial Parameters* |
| Algorithm: | LPCF |
| *BED window*: | 10 |
| *Event threshold*: | 0.99 |
| *Outlier threshold*: | 1.4 |
| *History window*: | 1.5 days (calculated based on the *data interval* present in each dataset) |

These initial parameters were chosen based on a slight variation of the rule-of-thumb approach to parameter selection (as described below). These parameters were tested on all stations. The stations had different *data intervals* (2, 5, or 20 minutes) and thus the *history window* was calculated to be 1.5 days based on the *data interval* found in each system. The combination of a *BED window* of 10 and an *event threshold* of 0.99 results in 9 out of 10 timesteps being outliers in order to trigger an alarm. The *BED window* of 10 was a compromise value that would work for all of the *data intervals* utilized in the six stations. An *outlier threshold* of 1.4 was used because two of the six stations had high variability in water quality (as reported in U.S. EPA, 2013a). (Note that for real world application of CANARY, configurations should be specific to each station. For this analysis, some of the parameters were kept constant across stations for simplicity.)

The above parameters performed reasonably well for all stations; however, the true detection rate from one station was well below the detection rate from another station. This preliminary result prompted further testing on the Challenge data. Analyses using the initial parameters present both a true and false alarm value; however, for the second part of this testing only true alarm rates are discussed. This second set of testing was geared towards determining the maximum number of detectable events. Only the best parameter set from each station will be used to analyze the minimum false alarm rate that maintains the maximum number of detected events.

A script was used to test a series of parameters. Ninety-six combinations of parameters and algorithms were tested, as follows:

- Name: *Parameter Testing*
- Algorithm: LPCF, MVNN
- *History window*: 1.5 days (calculated based on the *data interval* present in each dataset)
- *BED window*: 6, 8, 10, 12
- *Event threshold*: BED-1, BED-2, BED-3
- *Outlier threshold*: 1.0, 1.15, 1.3, 1.5

The *event threshold* value was calculated for each *BED window* to satisfy the three values listed. For example, *event threshold* values were calculated for 3/6, 4/6 and 5 out of 6 required outliers. The resulting *event threshold* values for this example were 0.3438, 0.6563 and 0.8907, respectively.

Both LPCF and MVNN algorithms were tested in order to determine if there was a difference in true detection rates.

In order to be consistent with Challenge results, an alarm was considered to be a true detection if it occurred within the timeframe of a simulated event. Multiple alarms within a simulated event were counted as a singular detection, ensuring that a maximum of 96 true detections could occur for each station.

### C.2 Results

The results reported in this Appendix are divided into two sections. Section C.2.1 presents the results using the LPCF algorithm and an initial set of parameters derived from the rule-of-thumb approach of parameter selection. Section C.2.2 presents results from an optimization process, focused on determining the maximum number of detectable events for each station. These analyses used both the LPCF and MVNN algorithms.

#### C.2.1 Initial Parameters

Results from CANARY analyses using the LPCF algorithm and the initial parameters are summarized in Table 20. The percent detected is also presented graphically in Figure 15. The initial parameter set was able to detect a higher percentage of simulated events in data from four out of six stations (B, D, E & G), relative to the CANARY results reported in the Challenge (U.S. EPA, 2013). The results for Station A data was approximately the same, with the initial parameters detecting only 3% fewer events. The initial parameters detected 24% fewer events for Station F data.

**Table 20: True Detection and Invalid Alarm Rate Comparison between Rule-of-Thumb Parameters and Challenge Results**

| | | Initial Parameters | | | Event Detection System Challenge* | | |
|---|---|---|---|---|---|---|---|
| **Location** | ***Data interval*** | **Detection** | **(%)** | **Invalid Alarms** | **Detection** | **(%)** | **Invalid Alarms** |
| **Station A** | 5 | 67 | 69.8 | 99 | 70 | 72.9 | 38 |
| **Station B** | 20 | 70 | 72.9 | 126 | 37 | 38.5 | 54 |
| **Station D** | 2 | 89 | 92.7 | 151 | 62 | 64.6 | 96 |
| **Station E** | 10 | 82 | 85.4 | 159 | 71 | 74.0 | 23 |
| **Station F** | 2 | 60 | 62.5 | 112 | 83 | 86.5 | 1146 |
| **Station G** | 2 | 91 | 94.8 | 106 | 83 | 86.5 | 90 |
| | **Total** | 459 | 79.7 | 753 | 406 | 70.0 | 1447 |

*Source: U.S. EPA, 2013, EPA/817/R-13/002

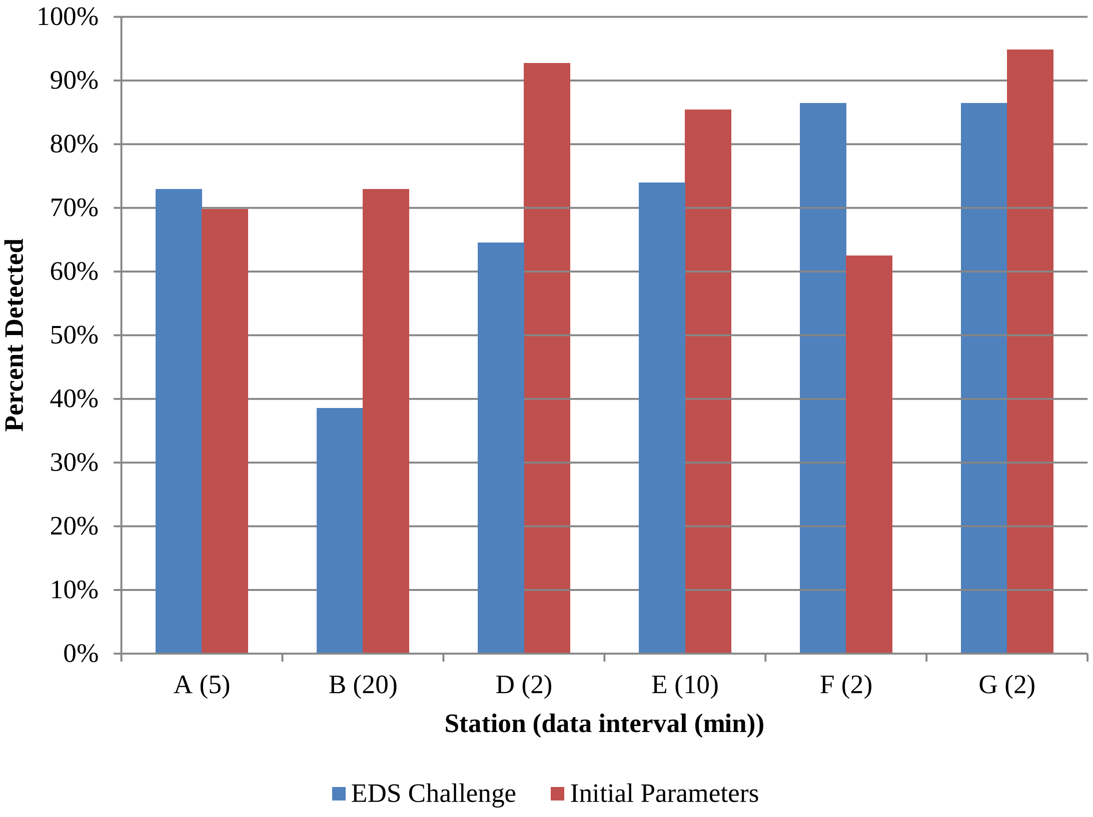


**Figure 15: Percentage of detected events reported by CANARY for Challenge dataset. Blue indicates results reported in the Environmental Detection System Challenge (U.S. EPA, 2013, EPA/817/R-13/002) and red indicates results for the initial parameter set (listed above).**

Table 20 also contains the number of invalid alarms for each station for the initial parameters and as reported by the Challenge. With the exception of Station F, parameters used in the Challenge resulted in fewer invalid alarms. Averaging all six stations, these initial parameters produce an invalid alarm 0.488

times per day (approximately once every two days), compared to 0.239 invalid alarms per day when Station F is excluded (or 0.792 invalid alarms per day when Station F is included).

The initial parameters performed well for the six stations studied here. An average detection rate of approximately 80% for all six stations, with a variety of *data intervals*, confirms that these parameters are a good starting point for further optimization. The invalid alarm rates are higher than those reported for the Challenge results; however, the average invalid alarm rate was less than one alarm every two days.

### C.2.2. Parameter Testing

The initial parameter tests revealed that the initially selected parameters performed well when using a variety of *data intervals* (ranging from 2 to 20 minutes). Those parameters were able to detect more than 60% of the simulated events at each individual station, and over 90% for Stations D and G. The percentage of detected events using Station F data was lower than expected; however, this station proved challenging for all EDSs in the Challenge. This prompted further testing – focused on determining how many events could be detected for each station.

In what follows, the simplified optimization process discussed within this report is utilized. The *history window* for all tests was 1.5 days – calculated for each station's *data interval*. Forty-eight parameter combinations were tested with either the LPCF or the MVNN algorithm, as described above.

This section focuses on the maximum number of events that could be detected for each station. The invalid alarm rates are presented for the parameter combination that achieved the maximum event detection for each algorithm for each station. Figure 16 contains a graphical representation of the average and maximum percentage of true detections for each station, divided by algorithm (CANARY Challenge results [green bars] are included for comparison). The averages are calculated over all variations of parameters tested; in comparison, the maximum values are for a single set of parameter values that achieved the maximum number of simulated events detected. Averages are included to highlight how much latitude in parameter selection is available for each station, based on parameters tested; that is, if the average is near the maximum percent detected, then CANARY is likely to maintain a high true positive rate while providing some ability to reduce invalid alarms. For example, Stations D and G maintain a high true detection rate throughout the range of parameters tested, which will allow for more flexibility in selecting parameters that reduce invalid alarms.

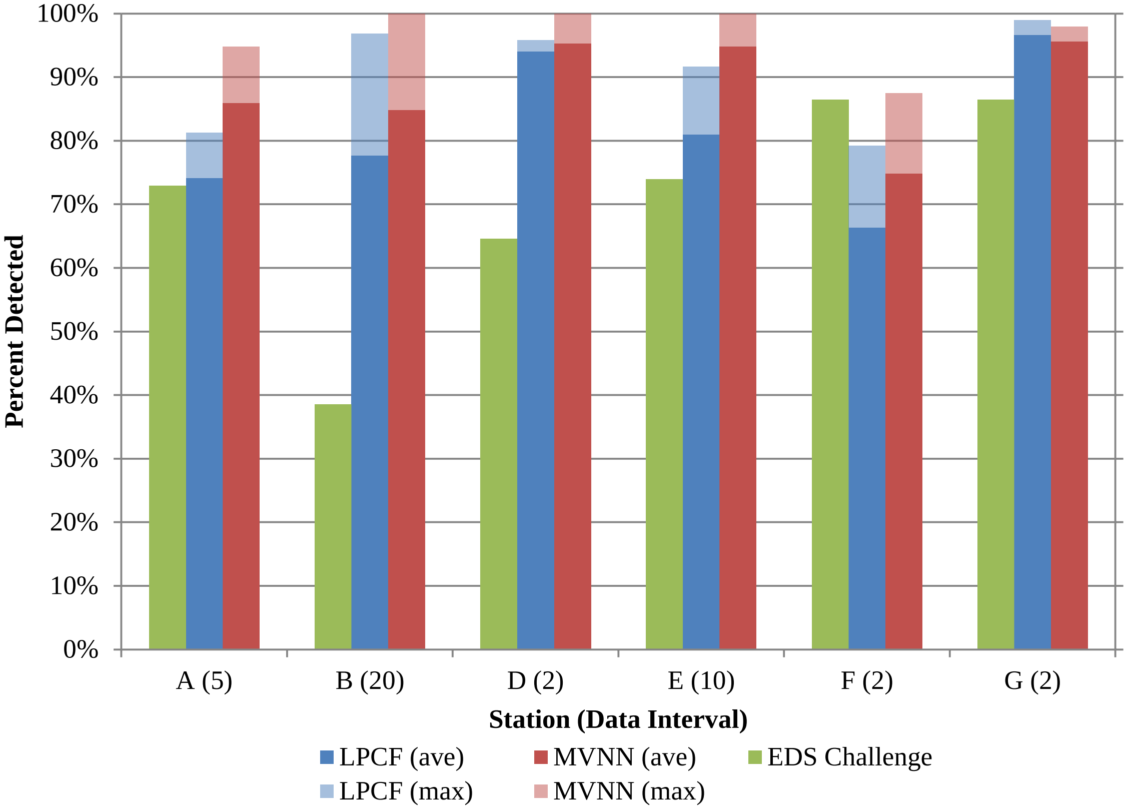


**Figure 16: Summary of average and maximum percent detected reported by CANARY for the Challenge dataset. Green indicates results reported in the Challenge (U.S. EPA, 2013, EPA/817/R-13/002), blue indicates the linear prediction coefficient filter (LPCF) algorithm and red indicates the MVNN algorithm. Dark shades are average (over all tested parameters) and light shades are the maximum percentage of detected events.**

**Table 21: Maximum Percentage of Detected Events by CANARY for the LPCF and MVNN Algorithms for the Parameter Testing Combinations**

| Location | Maximum Percent Detected LPCF | Maximum Percent Detected MVNN | Water Quality Variability* |
|---|---|---|---|
| **Station A** | 81.25 | 94.79 | Medium |
| **Station B** | 96.88 | 100.0 | Low |
| **Station D** | 95.83 | 100.0 | Medium |
| **Station E** | 91.67 | 100.0 | Low |
| **Station F** | 79.17 | 87.50 | High |
| **Station G** | 98.96 | 97.92 | High |

*Source: reported in Challenge (U.S. EPA, 2013, EPA/817/R-13/002)
LPCF, linear prediction coefficient filter; MVNN, multivariate nearest neighbor

Table 21 contains the maximum percentage of true detections of simulated events reported by CANARY for each station for both algorithms using the Parameter Testing combinations. These results show that at least one combination of parameters was able to achieve this level of true detections. Also contained in this table are the water quality variability that was reported for each station in the Challenge report (U.S. EPA, 2013); this is only intended to provide a general idea of how variable the water quality was during the analyzed period. CANARY was able to detect one-hundred percent of the simulated events in three of the stations using the MVNN algorithm and at least one combination of parameters tested (Stations B, D & E). At least one parameter combination resulted in a true detection rate of over 87.5% for all stations using the MVNN algorithm. For the LPCF algorithm, the best true detection rate for any combination was 98.96%; with CANARY able to detect over 90% of simulated events for four out of six stations. The MVNN algorithm had a higher maximum detection rate within the range of parameters tested for five out of six stations – the only exception being Station G.

Figure 17 – Figure 22 contain graphical representations of the percentage of simulated events detected for each parameter combination tested for Station A – Station G. The *outlier thresholds* of 1.0, 1.15, 1.3 and 1.5 are represented by blue, red, green and purple, respectively. Dark shades indicate the use of the LPCF algorithm and lighter shades represent when the MVNN algorithm was used. Results from the MVNN algorithm were overlaid on the results from the LPCF algorithm.

The results of these analyses are applicable only to the stations that were tested and during the period that was tested. While universal predictions cannot be made based on these results, general trends can help in the parameter selection process.

In general, event detection rates should decrease with increasing *outlier threshold* values. As the *outlier threshold* is increased, more variation is included in the baseline behavior and therefore it is less likely to trigger an alarm. This type of behavior can be seen for stations that utilize a *data interval* of five minutes or less (Stations A, D, F and G shown in Figure 17, Figure 19, Figure 21 and Figure 22). Stations B and E (20- and 10-minute *data intervals*, respectively) exhibit much less consistent behavior throughout the range of tested parameters – specifically using the LPCF algorithm (see Figure 18 and Figure 20). Two factors might play a role in this inconsistency: the method for simulating events; or, the loss of detail associated with increasing the *data interval*.

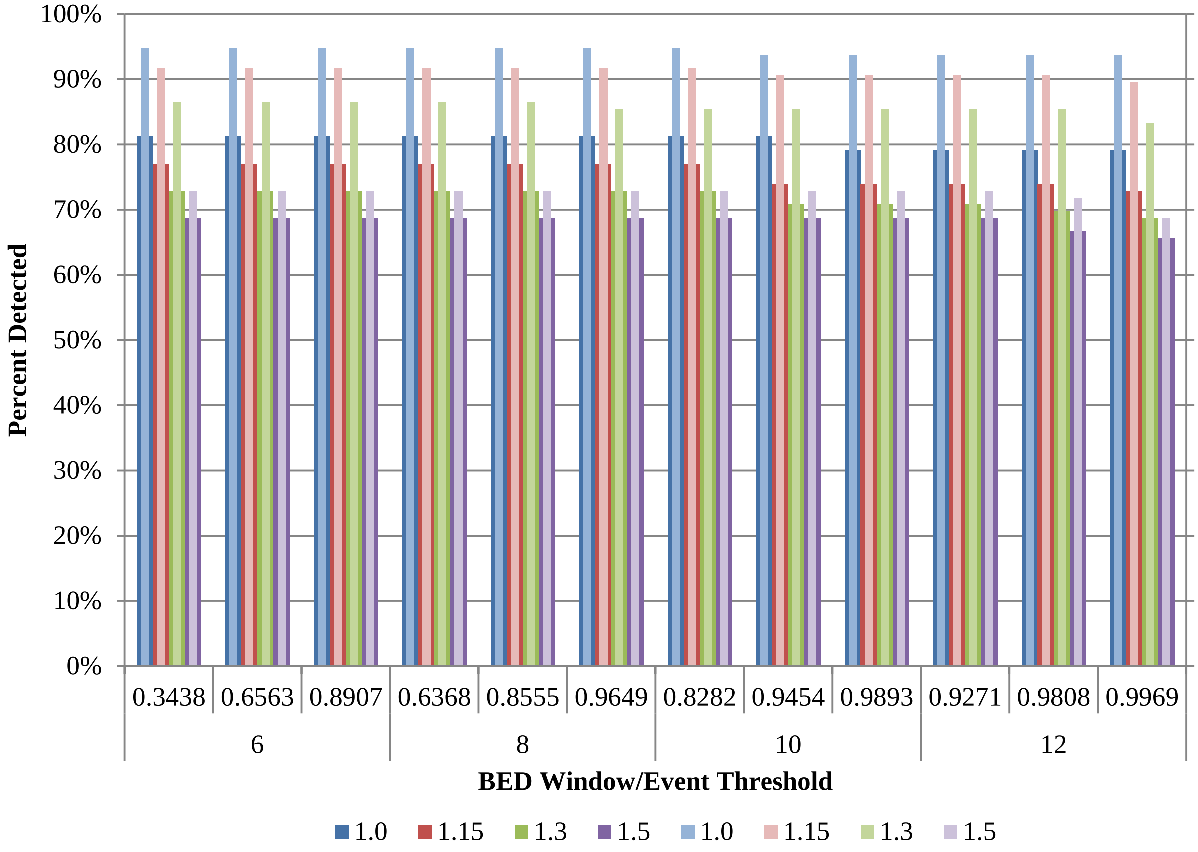


**Figure 17: Percent of simulated events detected by CANARY for Station A (5-minute *data interval*). *Outlier threshold* values of 1.0, 1.15, 1.3 and 1.5 are indicated by blue, red, green and purple. Dark shades indicate linear prediction coefficient filter (LPCF) and light shades indicate MVNN (overlaid).**

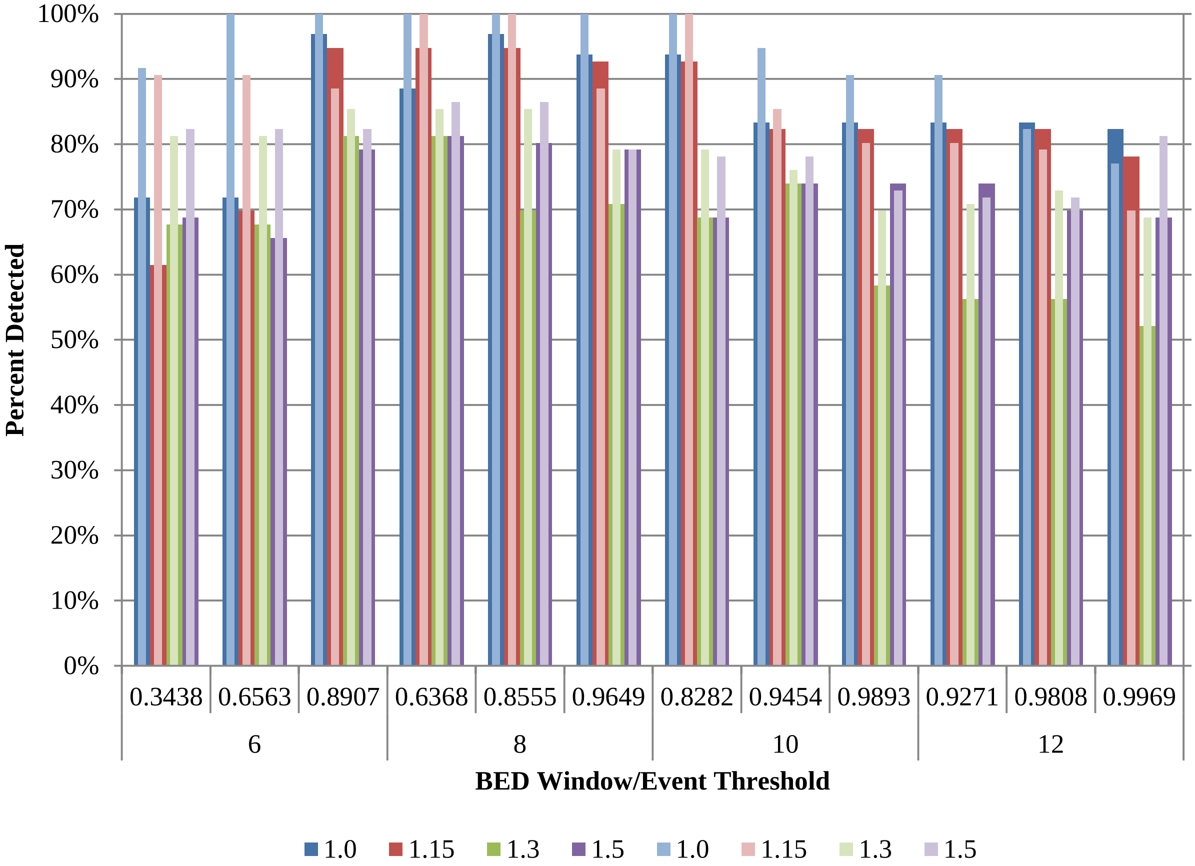


**Figure 18: Percent of simulated events detected by CANARY for Station B (20-minute *data interval*). *Outlier threshold* values of 1.0, 1.15, 1.3 and 1.5 are indicated by blue, red, green and purple. Dark shades indicate linear prediction coefficient filter (LPCF) and light shades indicate MVNN (overlaid).**

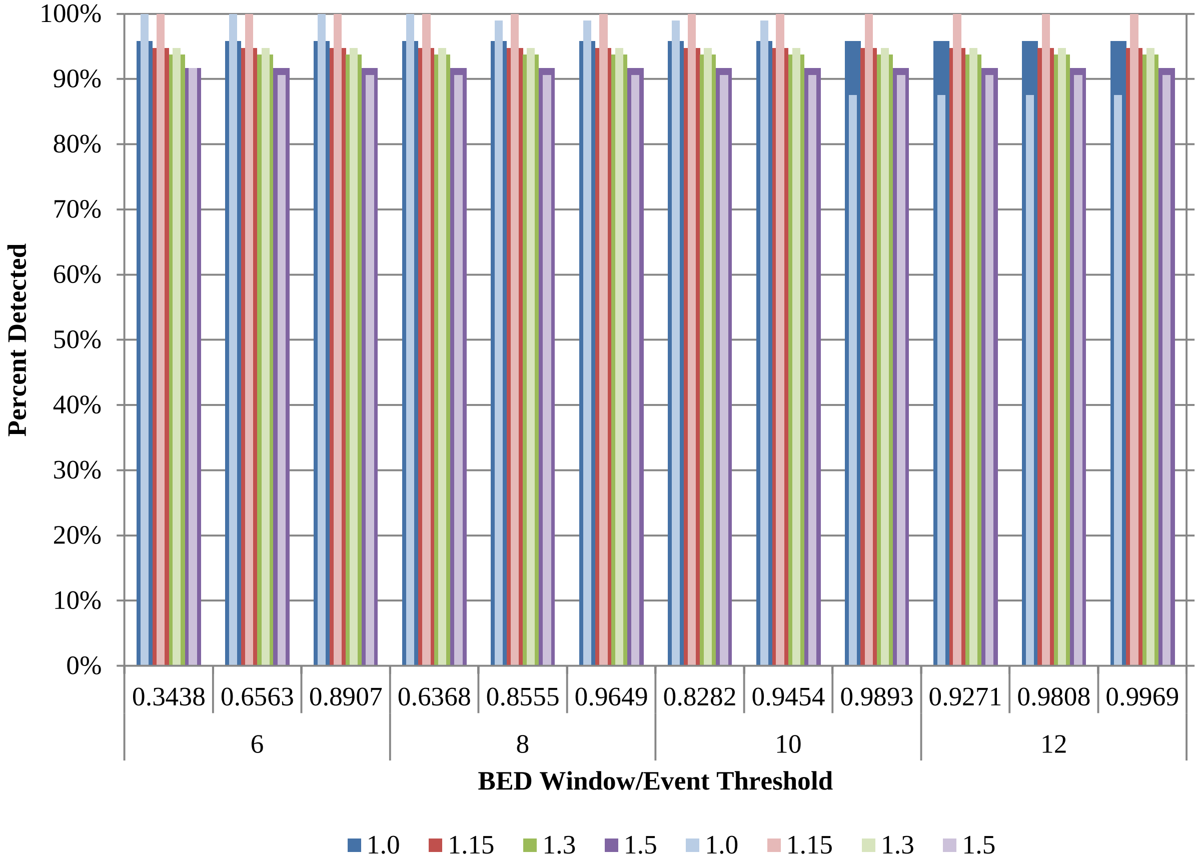


**Figure 19: Percent of simulated events detected by CANARY for Station D (2-minute *data interval*). *Outlier threshold* values of 1.0, 1.15, 1.3 and 1.5 are indicated by blue, red, green and purple. Dark shades indicate linear prediction coefficient filter (LPCF) and light shades indicate MVNN (overlaid).**

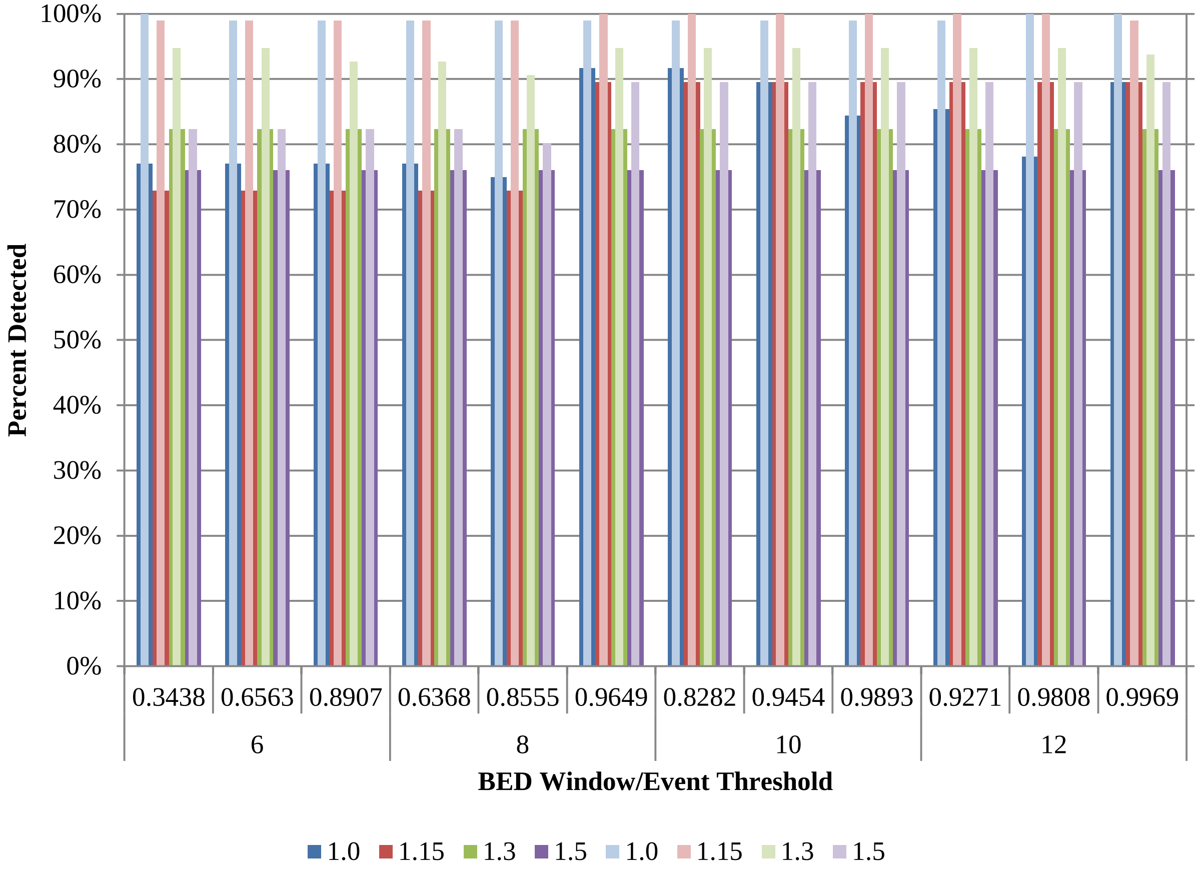


**Figure 20: Percent of simulated events detected by CANARY for Station E (10-minute *data interval*). *Outlier threshold* values of 1.0, 1.15, 1.3 and 1.5 are indicated by blue, red, green and purple. Dark shades indicate linear prediction coefficient filter (LPCF) and light shades indicate MVNN (overlaid).**

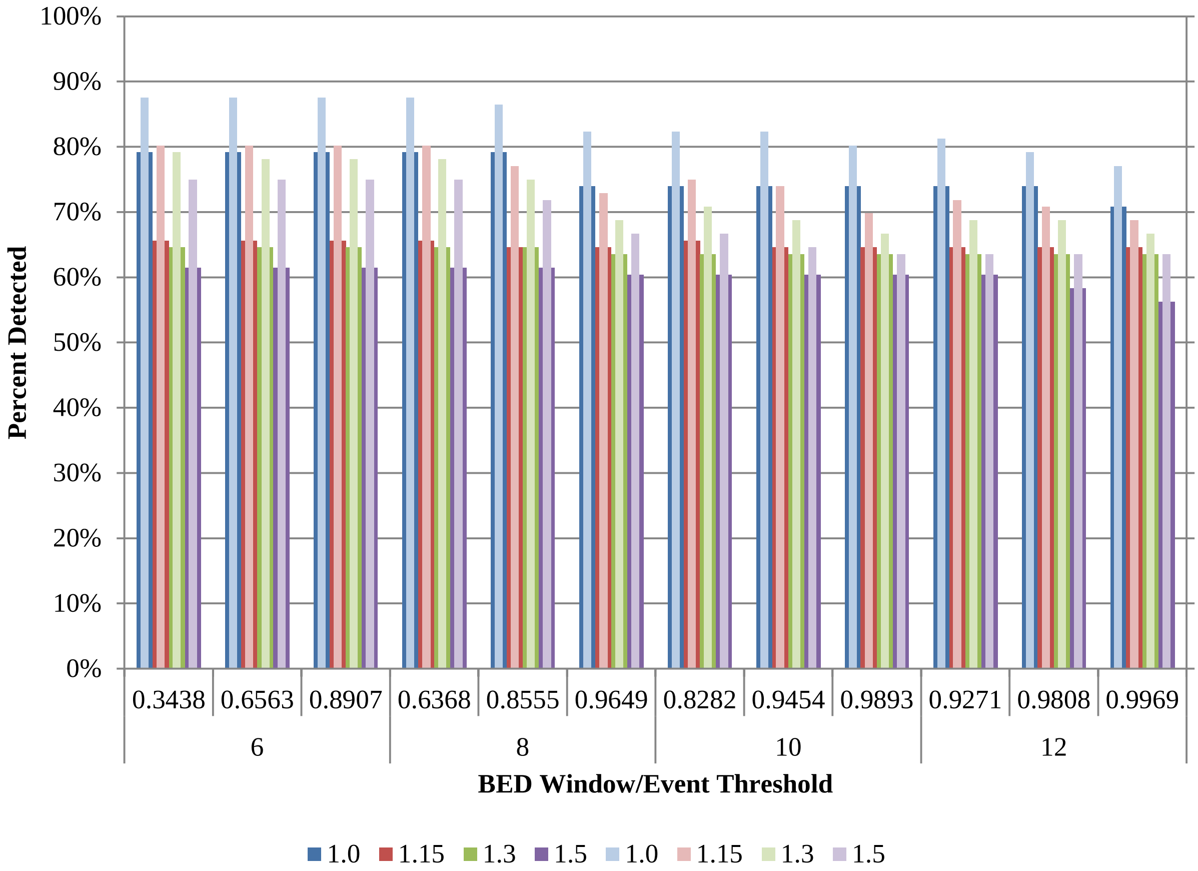


**Figure 21: Percent of simulated events detected by CANARY for Station F (2-minute *data interval*). *Outlier threshold* values of 1.0, 1.15, 1.3 and 1.5 are indicated by blue, red, green and purple. Dark shades indicate linear prediction coefficient filter (LPCF) and light shades indicate MVNN (overlaid).**

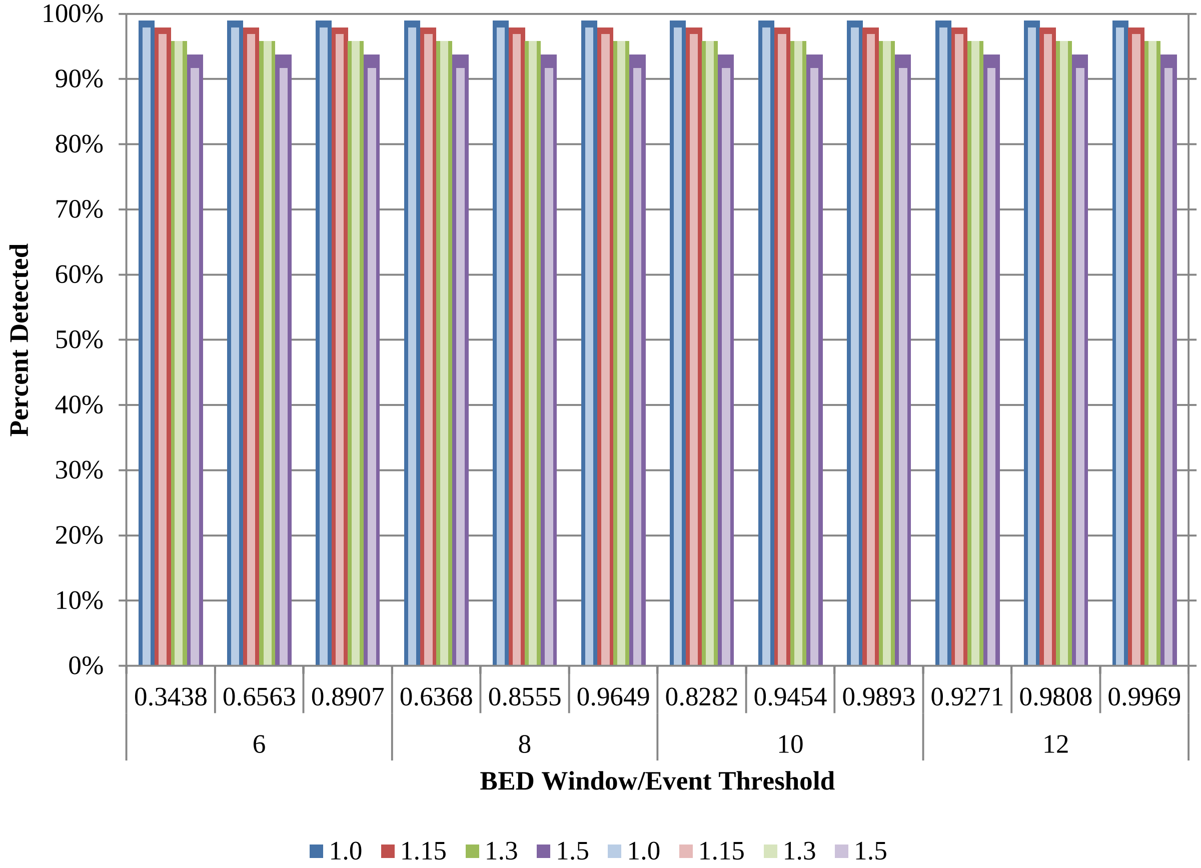


**Figure 22: Percent of simulated events detected by CANARY for Station G (2-*minute data interval*). *Outlier threshold* values of 1.0, 1.15, 1.3 and 1.5 are indicated by blue, red, green and purple. Dark shades indicate linear prediction coefficient filter (LPCF) and light shades indicate MVNN (overlaid).**

The effect of *data interval* and the effect of how the events were simulated have different impacts on this discussion. Previous work showed that increasing the *data interval* reduced the detection rate of CANARY and other EDSs (Allgeier et al., 2011). When the duration of an event is shorter than the *data interval*, water quality sensor signals might not show any changes in values; they might miss the event entirely. In addition, if the event duration is not much larger than the *data interval*, the number of corresponding outliers detected will be small; therefore, events are only detected for appropriate values of *BED window* and *event threshold*. In both cases, a smaller *data interval* is preferred to maximize the probability of detecting a high percentage of events. While events with long durations could be captured using a variety of *data intervals*, the probability of detecting shorter events is reduced for longer *data intervals*.

However, the Challenge results did not show this expected trend of decreasing true detection rates with increasing *data intervals*. This is because in the Challenge, event durations were not constant but were a function of *data interval*. Previous testing (Allgeier et al., 2011) compared the detectability of an event with a given duration with various *data intervals*; whereas, the Challenge simulated events that were shortened or lengthened depending on the *data interval*. Simulated events lasted one or two hours when using a *data interval* of two minutes and 9 or 19 hours with a *data interval* of 20 minutes. It is expected that the detectability of 1- to 2-hour events when using a *data interval* of 20 minutes would be decreased relative to the results presented here.

This approach to testing in the Challenge makes these results not comparable to the previous study by Allgeier et al., and make it difficult compare across stations. However, some general trends can be observed. Results from Station G (Figure 22) are very consistent throughout all 48 tested combinations and between both algorithms tested – all 48 parameter combinations resulted in a greater than 90% detection rate for both algorithms. Station D (Figure 19) results also show a greater than 90% detection rate for the majority of tested parameter sets. This consistent true detection rate throughout the tested range of parameters provides greater flexibility when selecting parameters that will minimize false alarms. In other words, the user does not have to try too many configuration settings before finding one that performs very well, and the initial parameter settings might perform adequately. In contrast, Station B (Figure 18) only has 9 out of 48 sets of parameters that maintain a greater than 90% detection rate using the LPCF algorithm and 15 sets when using the MVNN algorithm. For Stations D and G, it is likely that there is a tested parameter set that would maintain a high percentage of true detections while minimizing the invalid alarms. For Station B, there is a clear tradeoff in order to reduce false alarms, the true event detection rate must also be reduced.

Stations D and G show a wide latitude in parameter values that result in a high detection rate; this range is expected to be greater for systems with short *data intervals*. Of the Challenge Stations, D and G show the highest latitude (both with 2-minute *data intervals)*, followed by A and F (5 and 2-minute *data intervals*), and Stations B and E (20 and 10-minute *data intervals* respectively) have the lowest latitudes of those tested. Available latitude in parameter selection should not be confused with maximum detectability of events. High latitudes suggests that a user can reduce false alarms while not appreciably reducing the true positive rate; it does not always follow that the true positive rate is high, just that it is less sensitive to changes in parameter values.

Within the set of Testing Parameters, no combination was found that could detect 100% of events at three of the six stations. At least one combination of parameters was able to detect all simulated events using the MVNN algorithm for three out of the six stations. Ninety-one out of 96 simulated events or more were detected in two other stations using the MVNN algorithm. Three out of six stations detected at least 91 simulated events using the LPCF algorithm and at least one combination of parameters tested. At least one of the combinations of parameters tested was able to detect at least 76 of the simulated events for all stations.

The results from the Challenge and T&E datasets highlight the power of CANARY to detect true events when using either algorithm. In contrast to the results of CANARY testing with data from T&E (see sections 4.1 and 6.0), for five of the six stations studied, the MVNN algorithm performed better than the LPCF algorithm. The MVNN algorithm's ability to detect more events might have been related to the method of simulating the events in the Challenge's stations; however, these results show that both algorithms can produce a high percentage of true positives.

This testing did not include a full investigation of invalid alarms for each parameter combination; however, it is clear that there is a significant amount of latitude in parameter selection that would enable a utility to reduce invalid alarms while maintaining a high degree of true event detections. Table 22 contains the parameters used for each algorithm with each station that were expected to produce the fewest invalid alarms while maintaining the highest true detection percentage. Table 23 contains the maximum true detection percentage and number of invalid alarms reported for the parameters listed in Table 22.

**Table 22: Summary of CANARY Parameters that Maintained the Maximum True Detection Rate**

| Location | Algorithm | *BED Window* | *Event Threshold* | *Outlier Threshold* |
|---|---|---|---|---|
| **Station A** | LPCF | 10 | 0.9454 | 1.0 |
| | MVNN | 10 | 0.8282 | 1.0 |
| **Station B** | LPCF | 8 | 0.8555 | 1.0 |
| | MVNN | 10 | 0.8282 | 1.15 |
| **Station D** | LPCF | 12 | 0.9969 | 1.0 |
| | MVNN | 12 | 0.9969 | 1.15 |
| **Station E** | LPCF | 10 | 0.8282 | 1.0 |
| | MVNN | 12 | 0.9969 | 1.0 |
| **Station F** | LPCF | 8 | 0.8555 | 1.0 |
| | MVNN | 8 | 0.6368 | 1.0 |
| **Station G** | LPCF | 12 | 0.9969 | 1.0 |
| | MVNN | 12 | 0.9969 | 1.0 |

*BED, binomial event discriminator*; LPCF, Linear Prediction Coefficient Filter; MVNN, multivariate nearest neighbor

**Table 23: Summary of Invalid Alarms Reported by CANARY for Parameter Sets that Resulted in the Maximum True Detection Rates**

| Location | LPCF<br>% Detected | LPCF<br>Invalid Alarms | MVNN<br>% Detected | MVNN<br>Invalid Alarms |
|---|---|---|---|---|
| **Station A** | 81.25 | 135 | 94.79 | 166 |
| **Station B** | 96.88 | 221 | 100.0 | 427 |
| **Station D** | 95.83 | 224 | 100.0 | 282 |
| **Station E** | 91.67 | 256 | 100.0 | 928 |
| **Station F** | 79.17 | 303 | 87.50 | 331 |
| **Station G** | 98.96 | 147 | 97.92 | 385 |

LPCF, Linear Prediction Coefficient Filter; MVNN, multivariate nearest neighbor

Parameters in Table 22 were selected to minimize invalid alarms while maximizing true detections. An alternative approach would be to ensure that a set of parameters produced at least a minimum level of true event detection (e.g., 90%), or that a minimum percentage of events with longer durations are detected. Invalid alarm rates (see Table 23) correspond to below two alarms per day for all stations except when the MVNN algorithm was used at Station E. The MVNN algorithm was able to detect 100% of the simulated events; however, the invalid alarm rate was nearly four per day, which is likely unacceptable for most applications. This case highlights the direct tradeoff between true event detection and invalid alarm rates; that is, it was possible to detect all of the simulated events, but in this case it resulted in a higher than acceptable invalid alarm rate. At other stations, the invalid alarm rate ranged from 1.62 alarms per day down to 0.57 alarms per day; 6 of the 12 combinations resulted in an alarm rate of less than one invalid alarm per day. Invalid alarm rates could be further reduced when selecting parameters to maintain a minimum acceptable detection rate rather than maximizing detection.

For the Challenge data, the true detection rate could also be affected by the nature of the simulated events. Figure 23 contains two simulated events of the same contaminant with the same start time but with differing durations. The event with the shorter duration (dashed line) reaches its maximum concentration – that is, the largest deviation from baseline – much sooner than in the longer event (solid line). Both examples clearly have deviations from baselines; however, it is possible that the difference in the shape of signal response will affect detectability – specifically when simulating lower concentration events. Both algorithms were better able to detect the shorter duration simulated events in five out of the six stations. Some parameter combinations resulted in equal detection of short and long duration simulated events; however, in five out of the six stations, when a difference in detection percentage was observed, a higher percentage of these cases favored detection of the shorter event. Station B (with a 20-minute *data interval*) was the only station where the longer simulated events were more readily detected, with results from the LPCF algorithm having approximately the same behavior for both length events and only slightly favoring longer events.

Despite having different simulated sensor response profiles, the majority (or in some cases, all) of the events were detected with at least some of the parameter combinations tested. This suggests that CANARY is able to detect events with differing sensor response profiles and those with different durations. This difference in simulated response profiles is introduced as a reason why parameter latitude is reduced for some stations. In addition to testing two different event durations, two concentrations were also tested for each simulated contamination event. For shorter events, detection of events was approximately equivalent for the majority of tested concentrations. For longer simulated events, those with a lower concentration were detected with a lower frequency than those with a higher concentration. This effect was not unexpected, as a small signal change over a long timeframe might not likely to be considered significant with many combinations of parameters. Detection of events with a minimal change might require a combination of parameters that is very sensitive – that is, those that will result in a large number of false positives.

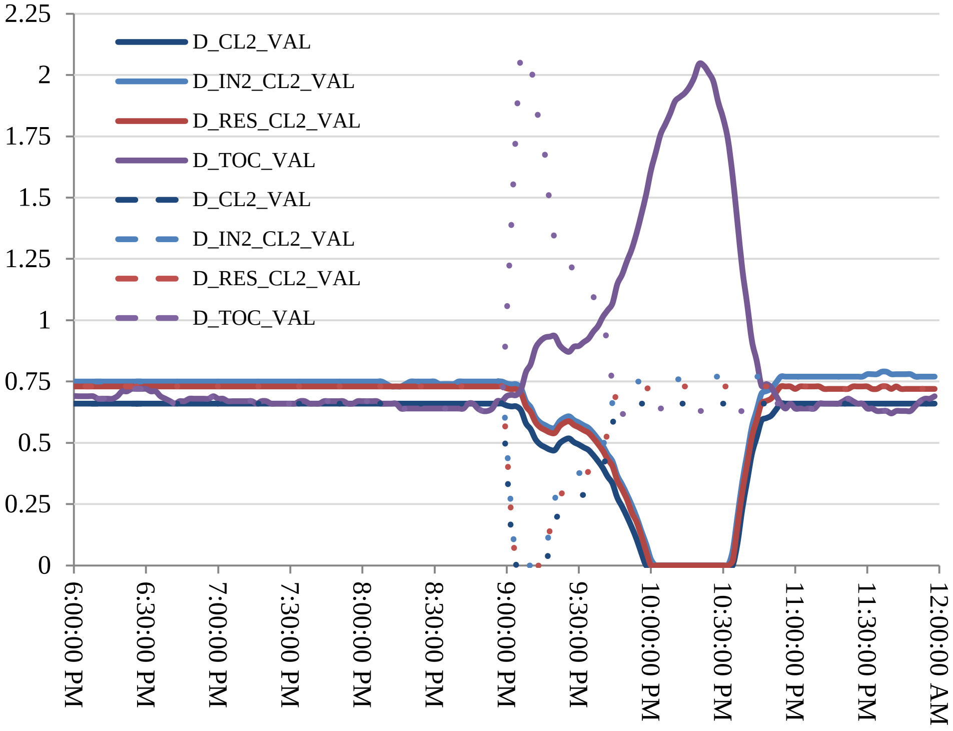


**Figure 23: An example comparison of simulated events from a Challenge dataset. Solid lines represent a simulated long event and dashed lines represent a simulated short event.**

In simulating future events, it might be appropriate to test not only different concentrations and durations, but also sensor response shape to gain a better understanding of how CANARY will behave with a wide variety of simulated events. Sensor response profiles could differ between distribution systems due to differences in pipe size, flow rates, mixing behavior and water chemistry; so, events must be simulated based on knowledge of the system and how contaminants affect sensors. Sensors monitoring large water mains might respond more slowly than a smaller pipe in a testing facility (e.g., the ones found at T&E), because contaminants must become mixed into the water in order to register a water quality signal change.

For comparison, Figure 24 contains adjusted signal responses for a contamination event performed at the T&E facility. In order to better visual signal changes, Figure 24 normalizes the signal data; such that 100% equals the maximum signal change (positive or negative) and 0% represents the normal signal behavior before the event. In general, signal responses to contamination events performed at the T&E facility tended to be abrupt (i.e., have a sharp response) for chlorine (blue and green) and UVA (cyan) sensors and resemble a normal distribution for conductivity (red) or turbidity (not shown for the plotted example due to lack of response to this contaminant) sensors. Changes in the ORP signal (purple) occurred more slowly (i.e., the ORP reached a maximum deviation after other signals had returned to normal, and it did not return to normal until approximately 1.5 hours after the injection).

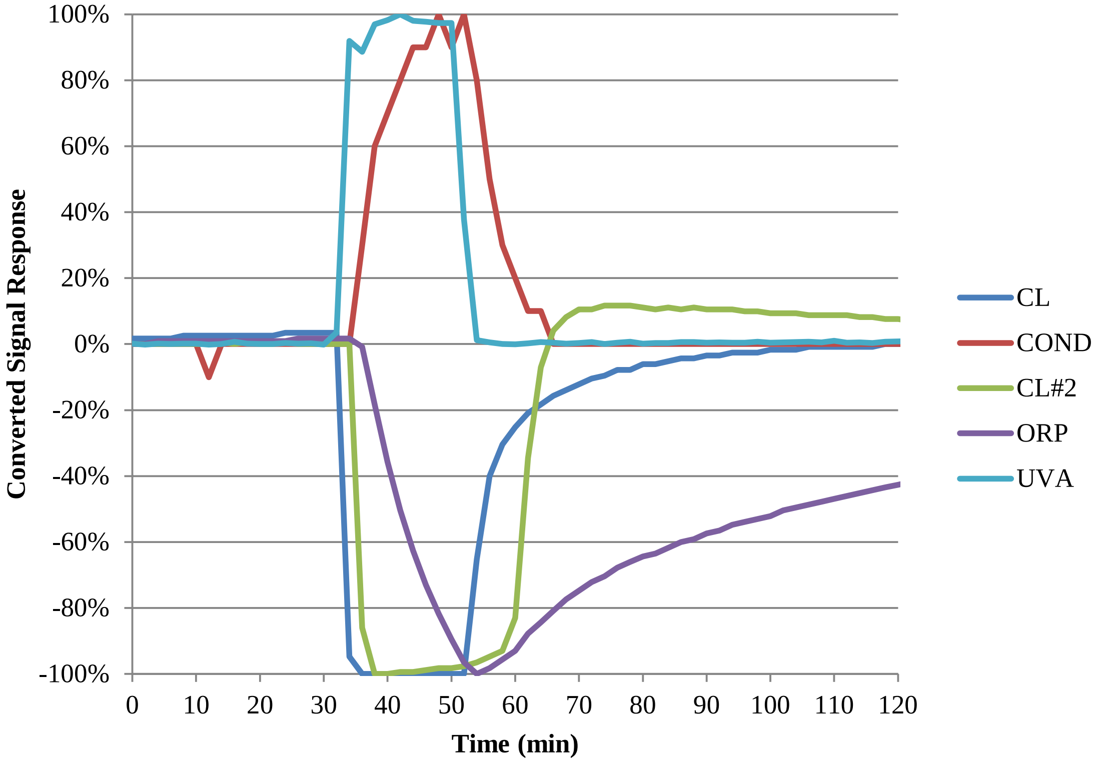


**Figure 24: An example of an adjusted signal responses for example contamination event at the U.S. EPA Testing and Evaluation Facility (T&E) facility.**

In addition, the results presented for the Challenge stations were obtained without incorporating alarms, calibration or operational signals in the CANARY analyses described in this section. This was done because not all stations provided equivalent data related to calibration, alarms or operational signals; so, omitting these signals equalized the analysis process. The use of these signals might have reduced the invalid alarm rates below the levels reported in this section.

### C.3 Conclusions

The results of retesting the Challenge data provided further information into how CANARY performs, and additional evidence that the recommended rule-of-thumb parameters are capable of detecting real or simulated events. At least one combination of the tested parameters was able to detect over 90% of the simulated events in four stations using the LPCF algorithm and five stations using the MVNN algorithm. At least one combination of the tested parameters was able to detect over 75% of the simulated events in all stations using either algorithm.

These results highlight that CANARY is able to maintain a high percentage of true positive detections using water quality data from a variety of stations. Stations with *data intervals* less than five minutes had a wider latitude in parameter selections, which will provide more choices in parameters in order to minimize false alarms.

## REFERENCES

Allgeier, S., USEPA-OGWDW, Korbe, C., Salomons, E., Edthofer, F., McKenna, S., Zach-Maor, A. (2011). “The Impact of Polling Frequency on Water Quality Event Detection”, *Proceedings of AWWA Water Quality Technology Conference*, Phoenix, AZ. p. 1497.

U.S. EPA, 2013. *Water Quality Event Detection System Challenge: Methodology and Findings*, Washington, DC: U.S. Environmental Protection Agency, EPA/817/R-13/002. [http://water.epa.gov/infrastructure/watersecurity/lawsregs/upload/epa817r13002.pdf Accessed October 1, 2013]

# Appendix D: Full Alarm Data for Testing of all Stations

This appendix provides complete results from section 4 for all values of *history window*, *binomial event discriminator (BED) window*, and *outlier threshold*. These tables expand on Table 7 and Table 10 from section 4.

**Table 24: Total Alarm and True Detections Reported by CANARY for 11/01/2011 to 6/27/2012 for T&E Data**

| ***outlier threshold*** | ***BED window*** | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **8** | | | | | **15** | | | | | **20** | | | | | **30** | | | | |
| | ***history window*** | | | | | | | | | | | | | | | | | | | |
| | **360** | **720** | **1080** | **1440** | **2160** | **360** | **720** | **1080** | **1440** | **2160** | **360** | **720** | **1080** | **1440** | **2160** | **360** | **720** | **1080** | **1440** | **2160** |
| | **Total Alarms (Detected Events)** | | | | | | | | | | | | | | | | | | | |
| **0.85** | 121<br>(14) | 70<br>(14) | 58<br>(14) | 62<br>(14) | 58<br>(14) | 83<br>(14) | 52<br>(14) | 49<br>(14) | 53<br>(14) | 51<br>(14) | 72<br>(10) | 43<br>(10) | 43<br>(10) | 45<br>(10) | 41<br>(8) | 53<br>(4) | 36<br>(5) | 34<br>(4) | 36<br>(4) | 34<br>(4) |
| **1** | 93<br>(14) | 53<br>(14) | 50<br>(14) | 49<br>(14) | 53<br>(14) | 73<br>(14) | 47<br>(14) | 46<br>(14) | 46<br>(14) | 46<br>(14) | 67<br>(10) | 39<br>(10) | 40<br>(10) | 39<br>(10) | 36<br>(8) | 59<br>(4) | 34<br>(6) | 29<br>(4) | 31<br>(4) | 29<br>(4) |
| **1.15** | 52<br>(14) | 41<br>(14) | 41<br>(14) | 43<br>(14) | 46<br>(14) | 49<br>(14) | 39<br>(14) | 38<br>(14) | 41<br>(14) | 42<br>(14) | 43<br>(10) | 33<br>(9) | 33<br>(10) | 34<br>(10) | 32<br>(8) | 36<br>(4) | 28<br>(3) | 24<br>(3) | 25<br>(4) | 24<br>(4) |
| **1.4** | 47<br>(14) | 37<br>(14) | 36<br>(14) | 36<br>(14) | 40<br>(14) | 45<br>(14) | 35<br>(14) | 34<br>(14) | 34<br>(14) | 37<br>(14) | 36<br>(8) | 30<br>(10) | 29<br>(10) | 29<br>(10) | 28<br>(8) | 31<br>(4) | 22<br>(6) | 20<br>(3) | 20<br>(3) | 22<br>(4) |

*BED, binomial event discriminator*; T&E, U.S. EPA Testing and Evaluation Facility

**Table 25: Total Alarms Reported by CANARY for 01/01/2008 to 08/31/2008 for PUB Stations**

| | | ***BED window*** | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 6 | | | | | | 8 | | | | | | 12 | | | | | |
| | | ***history window*** | | | | | | | | | | | | | | | | | |
| **Facility** | ***outlier threshold*** | 144 | 288 | 432 | 576 | 864 | 2016 | 144 | 288 | 432 | 576 | 864 | 2016 | 144 | 288 | 432 | 576 | 864 | 2016 |
| PUB1 | 0.85 | 507 | 232 | 173 | 149 | 112 | 87 | 443 | 197 | 137 | 125 | 99 | 70 | 365 | 159 | 114 | 100 | 73 | 53 |
| | 1 | 330 | 132 | 86 | 74 | 58 | 41 | 294 | 117 | 65 | 65 | 44 | 31 | 242 | 84 | 52 | 48 | 32 | 23 |
| | 1.15 | 233 | 84 | 50 | 34 | 24 | 21 | 201 | 64 | 38 | 24 | 21 | 19 | 156 | 46 | 21 | 16 | 6 | 11 |
| PUB2 | 0.85 | 392 | 204 | 163 | 135 | 107 | 95 | 374 | 187 | 153 | 125 | 103 | 90 | 336 | 167 | 135 | 113 | 95 | 80 |
| | 1 | 284 | 154 | 119 | 97 | 76 | 71 | 262 | 144 | 112 | 91 | 75 | 70 | 237 | 133 | 104 | 85 | 70 | 63 |
| | 1.15 | 225 | 120 | 95 | 80 | 69 | 58 | 205 | 115 | 88 | 76 | 68 | 55 | 185 | 107 | 80 | 68 | 60 | 47 |
| PUB3 | 0.85 | 780 | 521 | 438 | 394 | 340 | 268 | 716 | 482 | 407 | 357 | 314 | 244 | 641 | 423 | 355 | 325 | 283 | 217 |
| | 1 | 588 | 370 | 298 | 288 | 246 | 171 | 546 | 343 | 283 | 269 | 224 | 158 | 485 | 298 | 249 | 239 | 195 | 143 |
| | 1.15 | 476 | 282 | 222 | 205 | 181 | 146 | 432 | 265 | 214 | 194 | 172 | 131 | 384 | 235 | 187 | 163 | 147 | 111 |
| PUB4 | 0.85 | 720 | 454 | 399 | 349 | 327 | 273 | 660 | 411 | 362 | 308 | 305 | 245 | 595 | 367 | 322 | 273 | 266 | 222 |
| | 1 | 531 | 334 | 269 | 244 | 232 | 184 | 489 | 296 | 244 | 226 | 208 | 166 | 443 | 262 | 218 | 207 | 188 | 148 |
| | 1.15 | 413 | 245 | 191 | 190 | 164 | 139 | 376 | 222 | 178 | 172 | 148 | 126 | 338 | 204 | 155 | 159 | 135 | 120 |

*BED, binomial event discriminator;* PUB, Singapore Public Utility Board

EPA
United States
Environmental Protection
Agency

PRESORTED STANDARD
POSTAGE & FEES PAID
EPA
PERMIT NO. G-35

Office of Research and Development (8101R)
Washington, DC 20460

Official Business
Penalty for Private Use
$300