Colorado Water Watch

Colorado State University

September 2013

Colorado Water Watch

Anomaly Detection Methodology

Contents

Acknowledgement

Information of CANARY is based on the CANARY manual and webinars listed in Reference and more information is available in USEPA (2010a; 2010c).

CANARY

CANARY is a water quality event detection software that USEPA developed as part of a contamination warning systems (CWs) to secure nation's water resources and reduce the risks of drinking water contamination (USEPA, 2010a).

CANARY determines an event, the period of anomalous water quality, by using statistical and mathematical algorithms that identify normal variations in water quality (background) and water quality changes caused by contaminants.

The CANARY event detection system can:

- Continuously monitor water quality data at multiple sensor locations
- Alert periods of anomalous water quality
- Have flexibility in selecting event detection sensitivity at each monitoring station
- Ignore water quality values observed during periods of sensor malfunction or hardware alarm
- Recognize and store recurring water quality patterns that may be caused by routine operations to reduce false alarms

Real-time Data Analysis

The data of water quality signals received from the real-time sensors are transmitted to the database by the general packet radio service (GPRS). CANARY can be configured to receive water quality data from database through an online connection in near real-time and analyzes the data at a time using the data analysis tools.

The event detection system is continuously fed with new water quality data from the database until enough data are collected to fill window (number of observation or prescribed time) to start data analysis (Fig. 1).

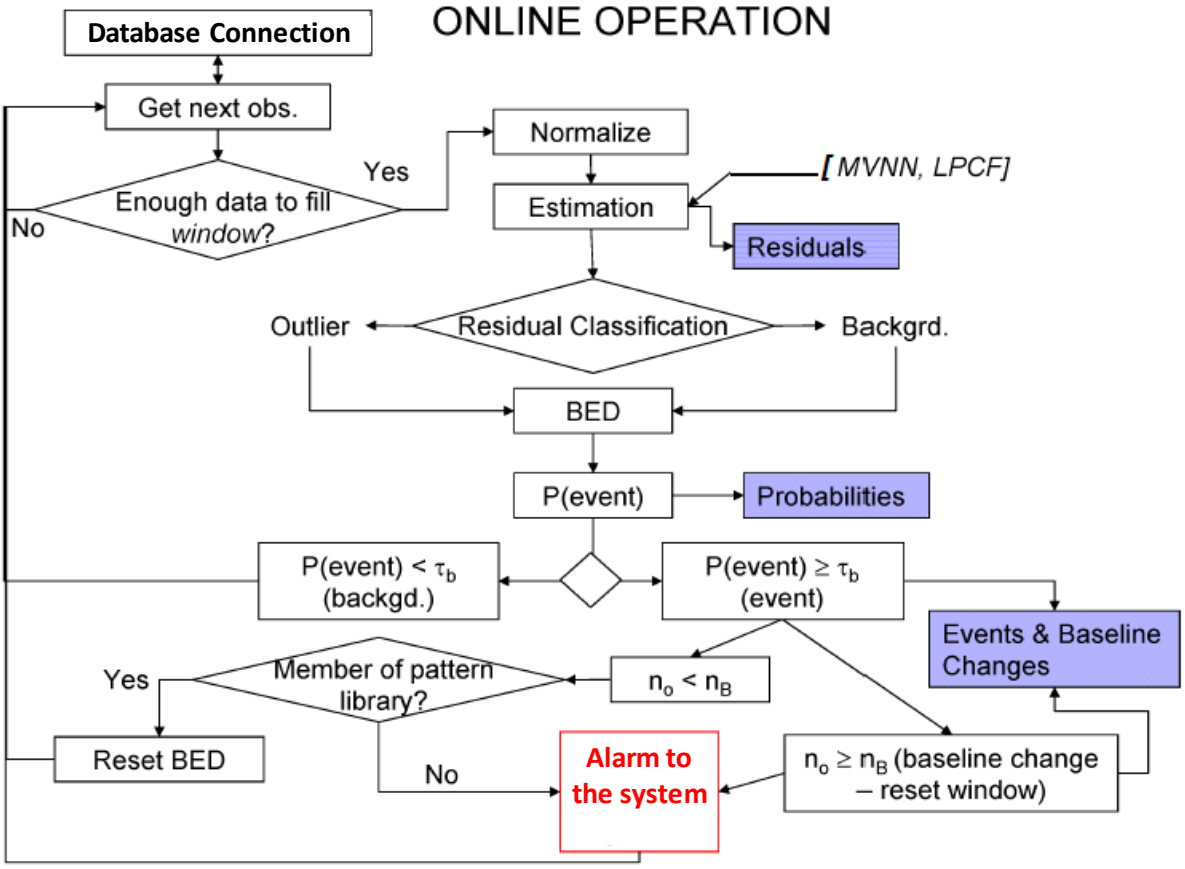

Fig. 1: CANARY flow diagram (USEPA, 2010c)

Water quality data are normalized as a mean value of zero and a standard deviation of 1 to estimate residuals, difference between the new water quality measurement and background, and the system qualifies the residual if it is an outlier (anomalous) or background (normal) using the data analysis tools: linear prediction coefficient filter (LPCF), multivariate near-neighbor (MVNN), and set-point proximity algorithm (SPPE).

As new data enter the system, the water quality background is updated, and a binomial event discriminator (BED) examines probabilities of number of outliers within a pre-set time window, and decides the onset of either an anomalous event or a change in the water quality baseline (Fig. 2).

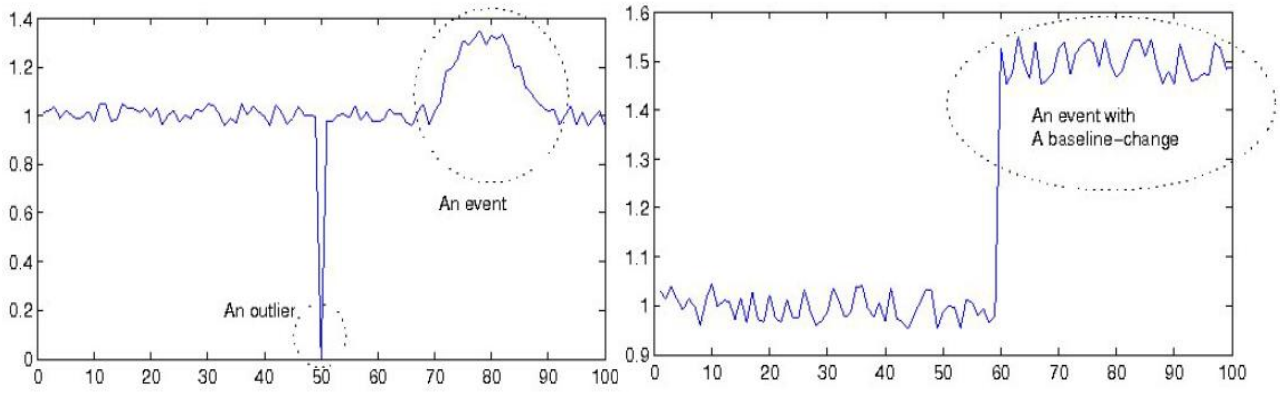

Fig. 2: Examples of an outlier, an event, and an event with a baseline-change (USEPA, 2010c)

The system also has a pattern library that allows storing patterns of water quality changes caused by noise of device operations and maintenances which might lead to false alarms in the system. CANARY is capable of pattern matching that analyzes historical data from a single monitoring station and detects reoccurring patterns (Fig. 3).

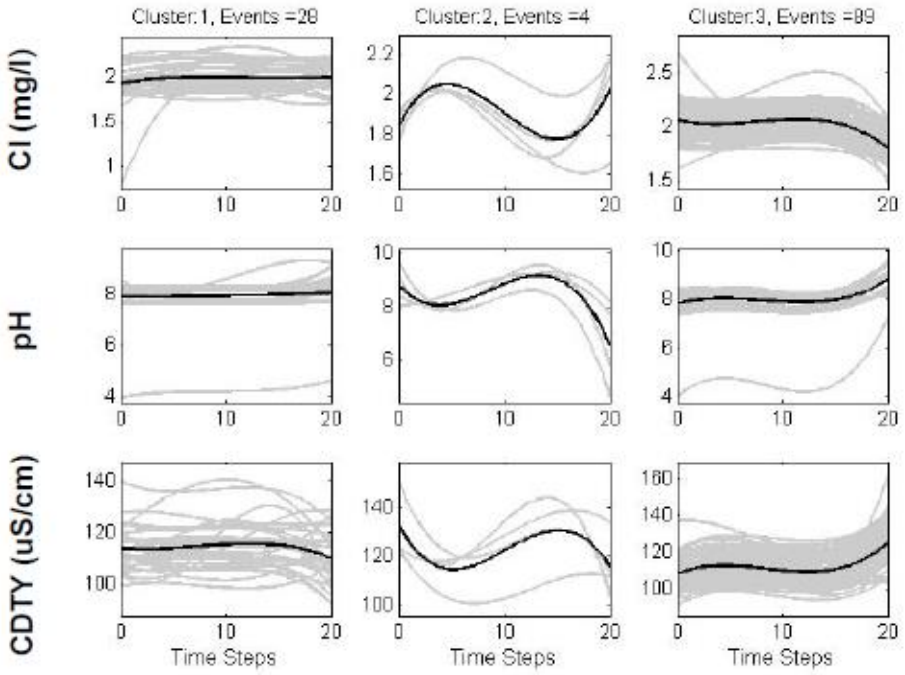

Fig. 3: Multivariate pattern library (USEPA, 2009a)

Data Analysis Tools: Event Detection Algorithms

The main event detection algorithms in the data analysis tools are LPCF, MVNN, SPPE, and BED. The LPCF algorithm or the MVNN algorithm *predicts* next water quality value based on the trends of the background data, *compares* the predicted value and the new water quality value at each time step, and *combines* the difference between the values (residual) across all water quality signals at a location to identify outliers in the data. The SPPE algorithm can be used with either the LPCF algorithm or the MVNN algorithm to look for both significant relative changes and excursions beyond set point values using consensus algorithm. The BED algorithm then *aggregates* the results across multiple time steps using probability distribution to determine the probability of an event occurring from number of outliers over a given number of time steps (Fig. 4).

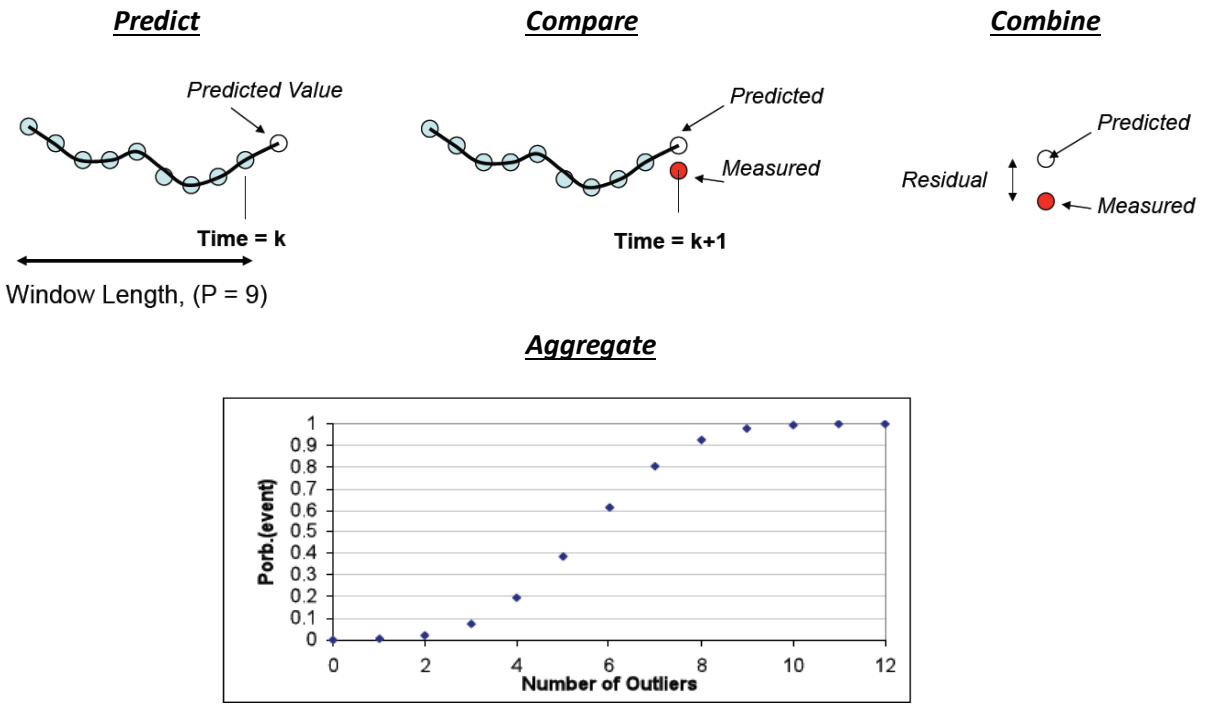

Fig. 4: Procedure of the event detection algorithms (USEPA, 2010b)

Details on the event detection algorithms can be found in: McKenna et al. (2006) (LPCF), Klise and McKenna (2006a; 2006b) (MVNN), and Hart et al. (2007) and McKenna et al. (2007) (BED).

Linear prediction coefficient filter (LPCF)

The LPCF algorithm predicts the next observation using linear coefficient estimation and digital filtering based on a weighted average of the historical data (previous values). The weight of each previous value is estimated by the linear coefficients at every time step and provided to update background water quality data in window. The formulation of the coefficient calculation gives unbiased estimates that have minimum variance. The predicted value and the measured value are compared at each time step, and the difference between the two values, residual, is calculated. The threshold of the algorithm (τ_a) is the maximum residual tolerance in units of standard deviation, and can be regulated by the analyst. Residuals greater than the predetermined threshold are considered outliers. Typical values for τ_a could be 0.5σ at stations having very stable water quality, and 1.5σ or higher at stations with unstable water quality.

Multivariate nearest-neighbor (MVNN)

The MVNN algorithm can be operated with any number of signals while the LPCF algorithm works for only one residual value. The MVNN algorithm evaluates the latest measured water quality data compared to all the data in the history window. The latest measured value is normalized and plotted against the other water quality signals in an m-dimensional multivariate space where m is the number of all signals used for the estimation. To give an example of a two-signal, a sequence of points are developed in 2-D space of the data (e.g., oxidation reduction potential vs. dissolved oxygen). The residual of the algorithm is the distance between the new point and the nearest neighbor data point in history window. The threshold (τ_a) is the maximum residual tolerance in units of standard deviation. Typical values for τ_a is 0.5σ–3σ.

Set-point proximity algorithm (SPPE)

The SPPE algorithm estimates the distance (ΔR) from any water quality signal to a predetermined minimum or maximum set point, and gives a ramped warning as any water quality value comes close to either set point value (Fig. 5).

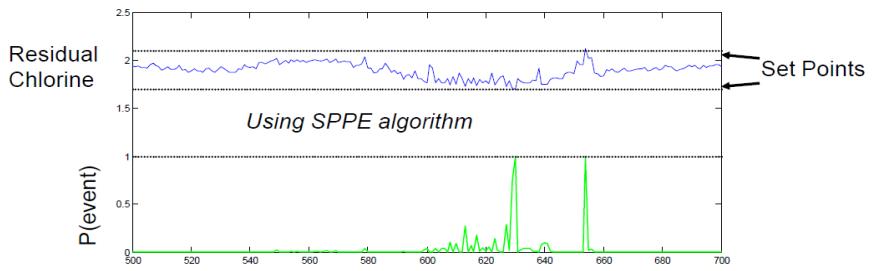

Fig. 5: A schematic diagram of the set-point proximity algorithm (SPPE) (USEPA, 2009b)

The threshold of the algorithm (τ_a) is ΔR, and the probability of an event, which starts from zero, increases as any water quality approaches to a set point. The SPPE algorithm is the simplest algorithm which does not require the previous water quality data or the binomial event discriminator to estimate the probability of an event.

Consensus algorithms: CAVE and CMAX

In real-time data analysis, CANARY can aggregate outputs of the configured algorithms (e.g., LPCF, MVNN, and SPPE) in a single report using consensus algorithms, CAVE and CMAX (Fig. 6).

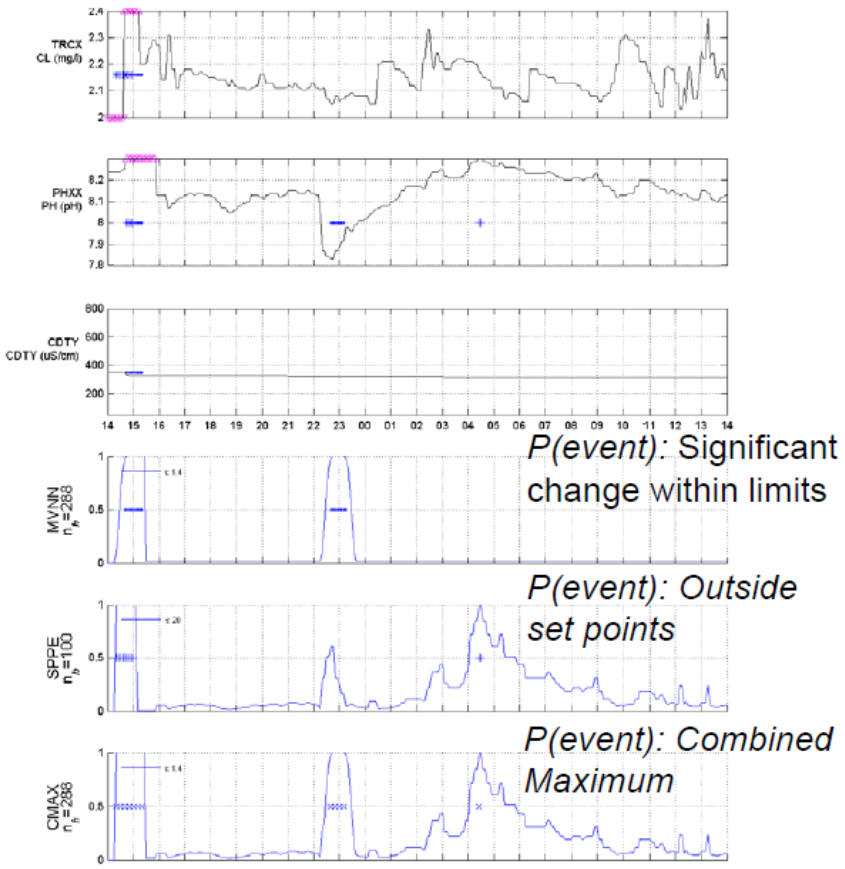

Fig. 6: Example outputs of the MVNN, the SPPE, and the CMAX data analyses; Blue dots indicate timing of identified events and pink triangles are data that is above/below max/min setting in the SPPE algorithm (USEPA, 2009a).

The CAVE algorithm gives the average of the probabilities of an event from other algorithms, and the CMAX algorithm shows the maximum probability of an event within the probabilities of

other algorithms. The algorithm which caused an alarm can be identified in a text file using duplicate outputs.

Binomial event discriminator (BED)

The BED is not an event detection algorithm that indicates whether the residual is above or below the threshold, but an algorithm which continuously provides the probability of an event that has occurred at each time step. The BED calculates the event probability based on the integration of results from other algorithms over the number of time steps to reduce false positives. The BED uses a smaller history window (n_B), separate from other algorithm's history window, to follow the number of recent outliers. P_E, the output of the binocdf(n_O, n_B, P_B) probability calculation, is calculated where n_0 is the number of outliers and P_B is the probability of an outlier occurring at each time step. CANARY will alarm an event when P_E is greater than the probability threshold (τ_B). P_E is recalculated at each time step as the latest outlier calculation is added to the window. Event time out (ETO) occurs when the event continues for multiple time steps, and CANARY notifies a baseline-change (n_{ETO}) that resets all the algorithms.

Case study

Case studies were performed using two sets of methane surrogate testing data collected on May 21, 2013 and on July 30, 2013.

Five parameters: temperature, pH, oxidation reduction potential (ORP), dissolved oxygen (DO), and conductivity were measured prior to and during the tests. The data collected prior to the tests are "background".

Before the test of May 21st, background data were collected for more than 5 days from May 16th, and methane gas was bubbled into continuously flowing water for the next 28 hours from May 21st. After the test started, dissolved methane concentration gradually increased to 4.6 ± 0.12 mg/L and reached steady state on May 22nd. The LPCF algorithm and the MVNN algorithm were performed for event detection within the data set. Event detection sensitivity was configured as following: an event threshold of 1, an event probability of 0.9, and a window of 2160 which equals 2.5 days (Fig. 7).

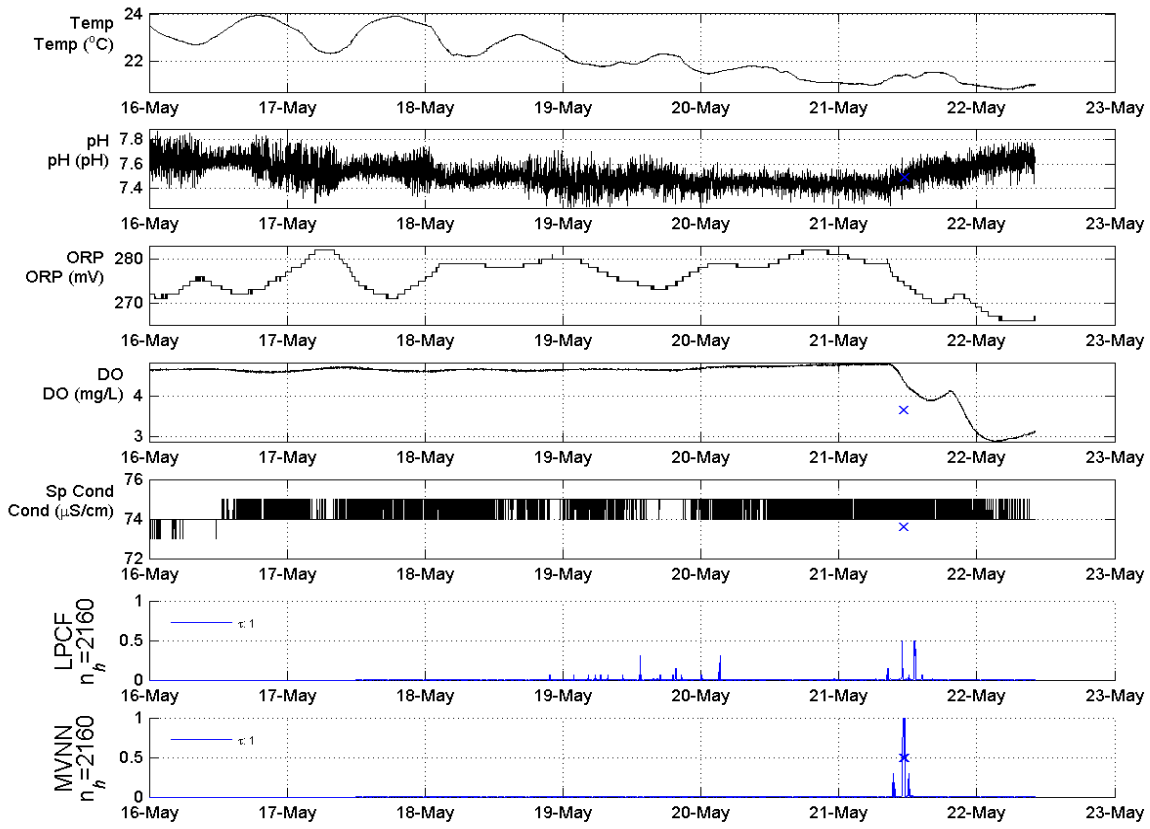

Fig. 7: Event detection using the LPCF and the MVNN algorithms configured as an outlier threshold of 1, an event probability of 0.9, and a window of 2160.

The MVNN algorithm detected an event from multivariate changes in DO, pH, and conductivity at 11:08 am on May 21st when methane concentration was 1mg/L. The detail of the event is below.

Summary results:
Run duration: 6 days 10.10 hours
Total events: 1
Event start time: 05/21/2013 11:08:00
MVNN algorithm
Signals contributed to events: Temp: 0, pH: 1, ORP: 0, DO: 1, Conductivity: 1

The SPPE algorithm was performed with the LPCF and the MVNN algorithms using the same data set and the same configurations as above (Fig. 8). Three times of standard deviations of background data were used as maximum and minimum set points.

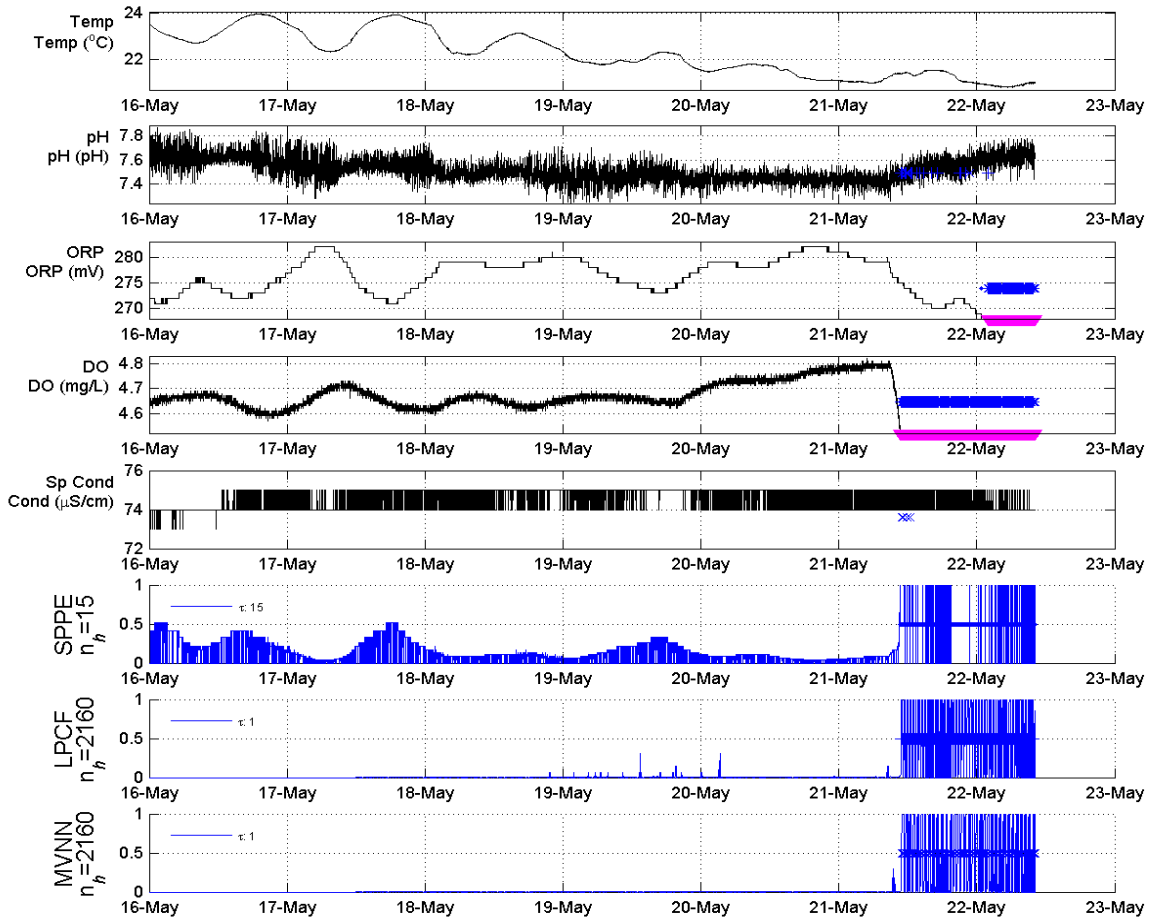

Fig. 8: Event detection using the LPCF and the MVNN algorithms, configured as an outlier threshold of 1, an event probability of 0.9, and a window of 2160, with the SPPE algorithm configured as the minimum set points (background mean - 3σ), DO (4.5 mg/L) and ORP (267 mV).

With the SPPE algorithm, the algorithms detected an event as described above as well as events from ORP and DO at 10:42 am on May 21st when dissolved methane concentration was less than 1 mg/L.

A data set collected from July 30th methane surrogate testing was also examined using the LPCF and the MVNN algorithms. Background data were collected more than 9 hours prior to the July 30th methane surrogate testing and the testing was performed for the following 14 hours. The steady-state methane concentration of the testing was 0.69 ± 0.04 mg/L.

Only ORP and DO data were used for this time to investigate events from the suggested methane surrogates. The LPCF algorithm was configured as an outlier threshold of 1, an event probability of 0.9, and a window of 480 which equals 8 hours with respect to the testing time frame (Fig. 9).

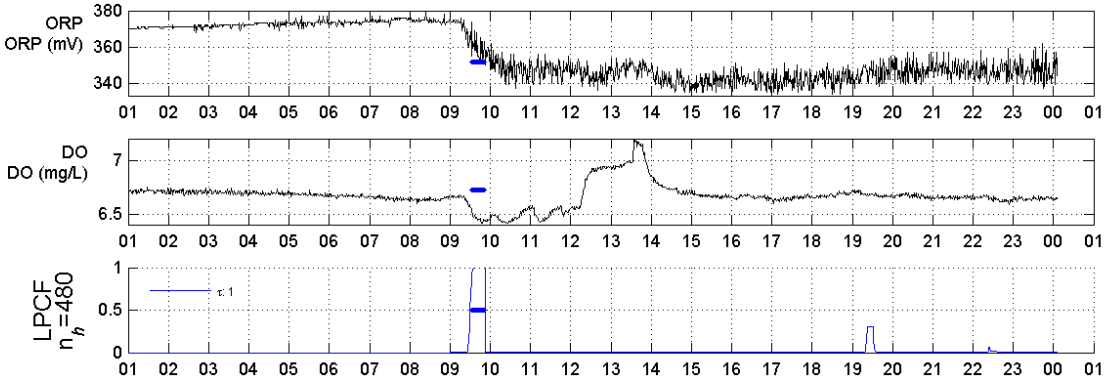

Fig. 9: Event detection using the linear prediction coefficient filter (LPCF) algorithm with an outlier threshold of 1, an event probability of 0.9, and a window of 480 (8hr).

The LPCF algorithm detected an event from ORP and DO data when methane concentration was 0.4 mg/L, which was less than 30 minutes after the test started. The detail of the event is below.

Summary results:
Run duration: 23.13 hours
Total events: 1
Event start time: 07/30/2013 09:14:00
LPCF algorithm
Signals contributed to events: ORP: 1, DO: 1

The MVNN algorithm was applied as well as the LPCF algorithm using the same data and the same configurations above (Fig. 10).

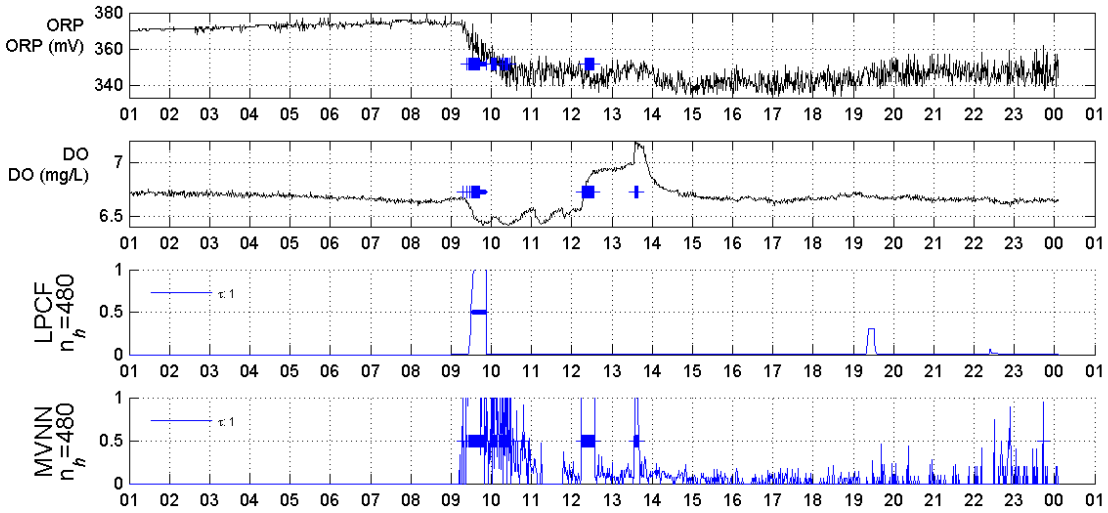

Fig. 10: Event detection using the linear prediction coefficient filter (LPCF) algorithm and the multivariate near-neighbor (MVNN) with an outlier threshold of 1, an event probability of 0.9, and a window of 480 (8hr).

The MVNN algorithm detected 17 events from multivariate changes in ORP and DO from 09:00am when methane concentration was less than 0.4 mg/L. The detail of the event is below.

Summary results:
Run duration: 23.13 hours
Total events: 17
Event start time: 07/30/2013 09:00:00
MVNN algorithm
Signals contributed to events: ORP: 14, DO: 8

From the case study, CANARY shows powerful results that prove the software is capable of detecting events from the methane surrogate testing data.

Reference

Hart, D.B., McKenna, S.A., Klise, K.A., Cruz, V.A., Wilson, M.P., 2007. CANARY: A water quality event detection algorithm development and testing tool. Proceedings of ASCE World Environmental and Water Resources Congress 2007, ASCE, Tampa, FL.

Klise, K.A., McKenna, S.A., 2006a. Multivariate applications for detecting anomalous water quality. Proceedings of the 8th Annual Water Distribution Systems Analysis (WDSA) Symposium, ASCE, Cincinnati, OH.

Klise, K.A., McKenna, S.A., 2006b. Water quality change detection: multivariate algorithms. Proceedings of SPIE Defense and Security Symposium 2006, International Society for Optical Engineering (SPIE), Orlando, FL.

McKenna, S.A., Hart, D.B., Klise, K.A., Cruz, V.A., Wilson, M.P. 2007. Event detection from water quality time series. Proceedings of ASCE World Environmental and Water Resources Congress (EWRI) 2007, ASCE, Tampa, FL.

McKenna, S.A., Klise, K.A., Wilson, M.P. 2006. Testing water quality change detection algorithms. Proceedings of the 8th Annual Water Distribution Systems Analysis Symposium (WDSA), ASCE, Cincinnati, OH.

USEPA, 2009a. CANARY webinar: algorithm overview and pattern library construction. October 19, 2009. <https://software.sandia.gov/trac/canary/downloader/download/file/15/CANARY_Webinar_3.pdf>

USEPA, 2009b. CANARY webinar: choosing algorithms and setting parameters. December 9, 2009. <https://software.sandia.gov/trac/canary/downloader/download/file/17/CANARY_Webinar_4.pdf>

USEPA, 2010a. Water quality event detection systems for drinking water contamination warning system; Development, testing, and application of CANARY. EPA/600/R-010/036, May 2010.

USEPA, 2010b. CANARY webinar: overview and advanced signals. September 22, 2010. <https://software.sandia.gov/trac/canary/downloader/download/file/34/CANARY_Webinar_6.pdf>

USEPA, 2010c. User's manual for CANARY. EPA/600/R-08/040B, September 2010.