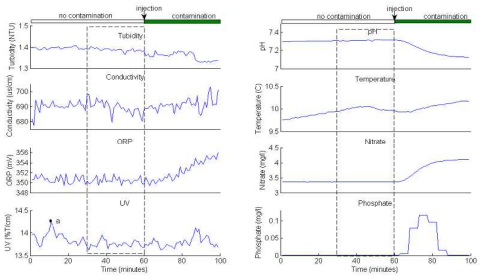
Environmental Science
Processes & Impacts

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

rsc.li/process-impacts

Correlative relationship of multiple types of conventional sensors can be used to detect contamination event.



Early warning systems are often used to detect deliberate and accidental contamination events in a water system. Conventional methods normally detect a contamination event by comparing the predicted and observed water quality values from one sensor, which suffer from high false positive and false negative rates. This paper proposes a new method for event detection by exploring the correlative relationships between multiple conventional water quality sensors, which could detect a contamination event 9 minutes after the introduction of lead nitrate solution (0.01mg/l). By implementing the proposed method, the accuracy and efficiency of an early warning system can be improved.

Contamination event detection using multiple types of conventional water quality sensors in source water

Shuming Liu[1*], Han Che[1], Kate Smith[1], Lei Chen[1],

[1]School of Environment, Tsinghua University, Beijing, 100084, China

[*]corresponding author: shumingliu@tsinghua.edu.cn, +86 1062787964

Abstract

Early warning systems are often used to detect deliberate and accidental contamination events in a water system. Conventional methods normally detect a contamination event by comparing the predicted and observed water quality values from one sensor. This paper proposes a new method for event detection by exploring the correlative relationships between multiple types of conventional water quality sensors. The performance of the proposed method was evaluated using data from contaminant injection experiments in a laboratory. Results from these experiments demonstrated the correlative responses of multiple types of sensors. It was observed that the proposed method could detect a contamination event 9 minutes after the introduction of lead nitrate solution with a concentration of 0.01mg/l. The proposed method employs three parameters. Their impact on the detection performance was also analyzed. The initial analysis showed that the correlative response is contaminant-specific, which implies that it can be utilized not only for contamination

detection, but also for contaminant identification.

Keywords

Contamination event detection, conventional sensor, correlative response, early warning system, water quality

1. Introduction

Water systems are vulnerable to contamination accidents and bioterrorism attacks because they are relatively unprotected, accessible, and often isolated[1-3]. In 2005, for example, the Songhua River was contaminated by nitrobenzene from a chemical plant explosion, which resulted in a 4-day suspension of water supply service to Harbin, China[4]. Besides this, about 1906 contamination accidents occurred per year in China between 1992-2006[5]. An emerging area of water security research involves developing methods to minimize the public health and economic impact of a large-scale accident or attack. An intense effort is currently underway to improve analytical monitoring and detection of biological, chemical, and radiological contaminants in drinking water systems as part of the overall aim of securing drinking water supplies[6]. One approach for avoiding or mitigating the impact of contamination is to establish an Early Warning System (EWS). EWS should provide a fast and accurate means to distinguish between normal variations and contamination events[7]. Ideally, it should be inexpensive, easy to maintain and integrate into network

operations and reliable, with few false positives and negatives[8].

A key part of an EWS is the detection module, which utilizes online sensors to evaluate water quality and detect the presence of contamination. Generally, there are two types of online water quality sensors. The first type refers to non-compound specific or conventional water quality sensors, which are normally used for routine water quality parameters, including pH, chlorine, total organic carbon (TOC), oxidation reduction potential (ORP), conductivity and temperature. Many commercially-available technologies for these parameters provide reliable means of detecting anomalies within water systems. The second type is compound specific water quality sensors or advanced sensors, which are capable of confirmative detection at low concentrations for a specific component and are mainly based on emerging detection technologies. Examples are Algae Toximeter for detection of the presence of toxic substances[9], Daphnia Toximeter for pesticides[10], Fish Activity Monitoring System for toxins[11], biological sensors relying on the detection of specific biomolecules (including adenosine triphosphate (ATP), enzymes, immunoassay and polymerase chain reaction (PCR) techniques), evaporative light scattering, refractive index measurement, fluorescence, and Raman spectroscopy[12-14].

Although compound specific sensors are capable of confirmative detection for contaminants at low concentration, their application in EWS for quick contamination

screening and detection is not popular since the type of potential contaminant is unknown at the time of sensor selection. It is difficult to determine which contaminant must be tested at such an early stage. In the past, conventional water quality sensors were generally used by operators in process control and regulatory compliance. In recent years, they have played a growing role in EWS. These sensors are advantageous in operational economics[15]. This type of 'dual-use' is practically attractive. For example, in the Water Security Initiative program in the United States, pH, turbidity, temperature, conductivity, TOC, and chlorine were chosen on the basis of their sustainability for long-term operation, and to provide 'dual-use' benefits to drinking water utilities, such as improved water quality management[16-17].

For an EWS with conventional water quality sensors, performance is highly dependent on the contamination detection method. Numerous publications have been devoted to discussing different ways of event detection using data from conventional water quality sensors. These mainly include statistical, artificial intelligence and data mining methods. For example, Hart et al.[18], reported two state estimation based algorithms for event detection, a linear prediction coefficient filter (LPCF) and a multivariate nearest-neighbor (MVNN) algorithm. These algorithms process the water quality data at each time step to identify periods of anomalous water quality and provide the probability of a water quality event existing at that time step. The LPCF method predicts the water quality at a future time step and evaluates the residual

between predicted and observed water quality values. The MVNN approach provides a measure of the similarity between the sampled water quality and the previously measured samples contained in the history window. Klise and McKenna[19] developed an algorithm to classify the current measurement as normal or anomalous by calculating multivariate Euclidean distance. Bucak and Kalik[20] and Bouamar and Ladjal[21] utilized artificial neural networks (ANN) and support vector machines (SVM) to classify water quality data into normal and anomalous classes. A common feature of these methods is to compare observed and predicted responses from time series data for one sensor. Meanwhile, several papers were published on optimal location of sensors for event detection and response actions[22-25]. In most of these studies, an ideal sensor is assumed. Actually, routine operation or equipment noise could result in fluctuation, which might lead to a high false alarm rate.

Several researchers have reported the phenomenon of correlative responses. For example, Hall et al.[26] reported a sensor response experiment for 9 types of contaminants and realized that more than one sensor responded to each tested contaminant. After noticing this phenomenon, researchers have attempted to develop contaminant detection methods using responses from multiple types of sensors. Yang et al.[15] explored a real-time event adaptive detection, identification and warning (READiw) methodology in a drinking water pipe. The suggested adaptive transformation of sensory measurements reduced background noise and enhanced

contaminant signals. In the method employed by Yang et al.[15], the relative value of concentrations of free and total chlorine, pH and ORP are used for contaminant classification. This allowed for contaminant detection and further classification based on chlorine kinetics. Kroll[27] reported the Hach HST approach using multiple sensors for event detection and contaminant identification. In the Hach HST approach, signals from 5 separate orthogonal measurements of water quality (pH, conductivity, turbidity, chlorine residual, TOC) are processed from a 5-paramater measure into a single scalar trigger signal. The deviation signal is then compared to a preset threshold level. If the signal exceeds the threshold, the trigger is activated[28]. Murray et al.[29] used Bayesian belief networks to detect a contamination event based on data from multiple sensors. Aliker and A. Ostfeld[30] reported the application of SVM to detect contamination based on data from multiple sensors. Perelman et al.[31] and Arad et al.[32] reported a general framework that integrates a data-driven estimation model with sequential probability updating to detect quality faults in water distribution systems using multivariate water quality time series. In particular, in Arad et al.'s work, univariate event probabilities are fused to give a unified multivariate event probability. The multivariate probability reflects the likelihood of a contamination event based on all data analyzed from all parameters[32]. A common feature of these methods is that efforts based on multivariate water quality measurements consider correlations between parameters, even if a correlation coefficient is sometimes not calculated explicitly.

In this paper, we will describe a new method for real-time contamination detection using multiple types of conventional water quality sensors for source water. The proposed method aims to achieve contamination detection by exploring the correlative relationship between responses from multiple sensors for the same type of contaminant. The proposed method is tested using data from contaminant dosing experiments in a laboratory.

2. Materials and methods

2.1 Pilot-scale contaminant injection and monitoring system

The pilot-scale system used in this study is a recirculating system simulator in the School of Environment Laboratory at Tsinghua University, Beijing, China. A process flow schematic of the pilot-scale system used for baseline establishment and single-pass contaminant tests is shown in Figure 1 and the photograph of the experiment setup (P1, supporting information). The water tank is approximately 85 cm high with a diameter of 70 cm, and has a total capacity of 300 L. The tank is linked with on line water quality sensors via a peristaltic pump at 0.5 L per minute.

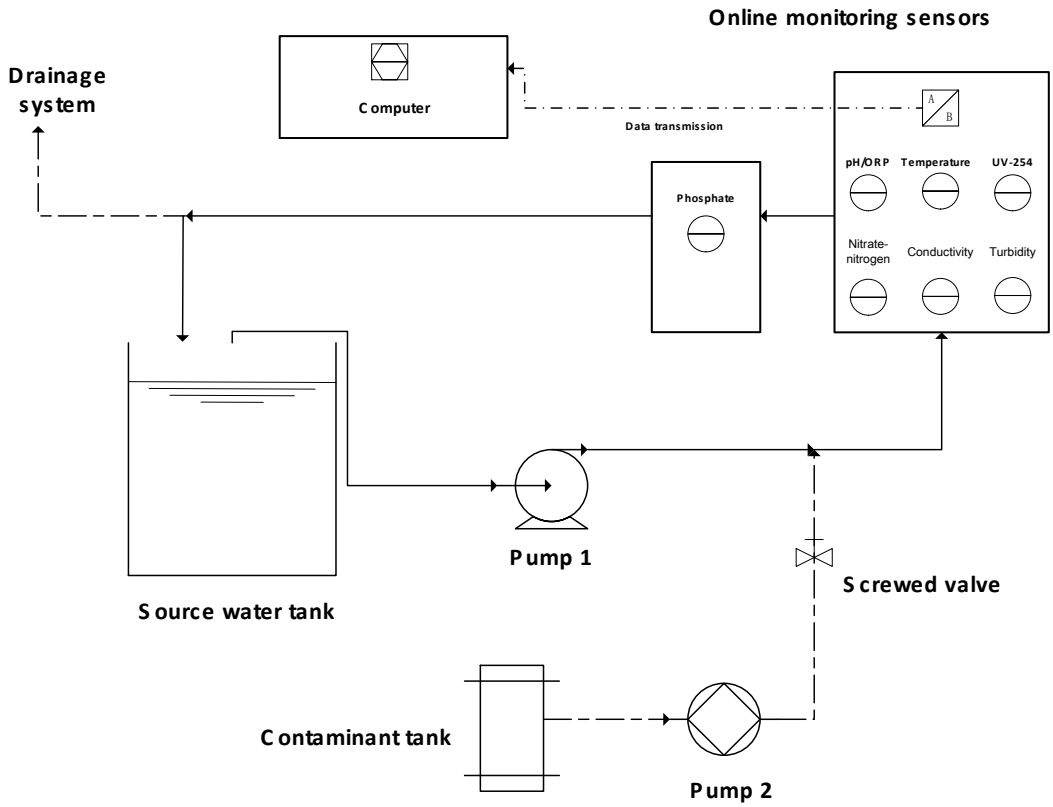


Figure 1 A process flow schematic of the pilot-scale system

The system was operated in recirculation mode for baseline establishment. In this mode, 300L source water flows through the multi-sensors and back to the tank. The characteristic of the source water is shown in Table 1. The entire volume of water in the loop is replaced every 72 hours if no contaminant test is conducted. Generally, the process of establishing baseline takes 2-3 hours before any contaminant experiments can be carried out. When operating in single-pass contaminant mode, the target contaminant is injected into the pipe connecting the tank and sensors via another peristaltic pump[33]. It is injected at a rate of 2-20 mL per minute depending concentration requirement. The water combined with contaminant flows through the sensors directly into a specific waste liquid bucket, avoiding pollution of the water in

158 the tank.

159 Table 1 Characterization of source water in current study

Parameter	Concentration	Parameter	Concentration
Temperature	10 ℃	pH	7.31
DO	12.77 mg/L	Turbidity	1.39 NTU
COD	4 mg/L	BOD_5	<2 mg/L
Conductivity	690 $\mu s / cm$	NH_3-N	0.03 mg/L
ORP	350 mV	NO_x-N	3.36 mg/L
Sulfate	170 mg/L	Total phosphorus	0.06 mg/L
Chloride	28 mg/L	Sulfide	<0.02 mg/L

160

161

162 **2.2 Sensors investigated**

163 8 sensors developed by Hach Homeland Security Technologies were utilized in this
164 study. They can measure the following 8 parameters simultaneously and continuously:
165 temperature, pH, turbidity, conductivity, oxidation reduction potential (ORP),
166 UV-254, nitrate-nitrogen and phosphate. Table 2 shows a list of the parameters and
167 the detailed information of their associated sensors.

168 Table 2 Detailed information of the parameters and sensors

Parameter	Sensor name	Measuring range	Sensitivity	Measuring interval
Temperature	DPD1R1-WDMP	-10-50℃	±0.01℃	1 min
pH	DPD1R1-WDMP	-2.00-14.00	±0.01	1 min
Turbidity	LXV423.99.10100	0.001-4000 NTU	±0.001 NTU	1 min
Conductivity	D3725E2T-WDMP	0-2000000 $\mu s / cm$	±1 $\mu s / cm$	1 min
ORP	DRD1R5-WDMP	-1500-1500 mV	±0.5 mV	1 min
UV-254	LXG418.99.20000	0.01-60 1/m	±0.01 1/m	1 min
Nitrate-nitrogen	LXG.717.99.50000	0.1-100.0 mg/L	±0.1 mg/L	1 min
Phosphate	LXV422.99.20102	0.05-15 mg/L	±0.05 mg/L	5 min

169

170 **2.3 Contaminants investigated**

171 Specific quantities of various contaminants were injected into the system simulator.

The contaminants were determined according to statistical reports on water pollution incidents in urban water supply systems in China in the past 20 years and included three types of the most commonly seen pollutants: herbicides (glyphosate), pesticides (atrazine) and heavy metals (lead nitrate, cadmium nitrate, nickel nitrate and trivalent chromium). They were also selected based on China's national standards regarding source water quality GB3838-2002. The concentration ranges were decided by the concentration limit given in the standards (Table 3).

Table 3 Concentration limits in GB3838-2002

Contaminant	Concentration limit (mg/L)
Glyphosate	0.7
Lead	0.01
Atrazine	0.003
Nickel	0.02
Chromium	0.01
Cadmium	0.001

2.4 Experimental procedure

Sensors were calibrated in accordance with the manufacturer's recommendations and were verified with a calibration check standard. Before the introduction of contaminants, the experimental system was kept running to establish a baseline. Sensor data were collected continuously and archived electronically to establish stable baseline conditions and to record sensor responses to injected contaminants. Data from the ORP, nitrate, temperature, pH, conductivity, turbidity and UV sensors were monitored and recorded every 1 minute during the test period, while the phosphate sensor was recorded every 5 minutes. After the baseline was established, a specific

concentration of contaminant was injected. Each contaminant injection lasted for over 30 minutes to reach a stabilized reading. The sensors were then supplied with uncontaminated raw water and the responses returned to the baseline. Another different concentration of the same contaminant was injected after sensor responses had returned to the baseline following the previous test.

2.5 Detection Method

In this research, it is assumed that multiple water quality sensors can respond to a contaminant simultaneously. The proposed method detects contamination by exploring the correlative relationship between responses from multiple water quality sensors. This relationship is evaluated using the Pearson correlation coefficient. The *window size* is the number of past observations used to calculate the Pearson correlation coefficient. For each sensor, a new observation enters the sliding window at every time step *t* and the oldest observation exits (i.e., first in first out).

The value of r_{xy} is between -1 and 1. In this study, a *correlation indicator* C_{xy} is calculated using

$$\begin{cases} C_{xy} = 0 & if\ |r_{xy}| < threshold_{indicator}\ or\ x = y \\ C_{xy} = 1 & if\ threshold_{indicator} \le |r_{xy}| < 1 \end{cases} \quad (1)$$

A contamination alarm will be trigged if

$$\sum_x^{\forall} \sum_y^{\forall} C_{xy} \ge threshold_{alarm} \quad (2)$$

in which $x \neq y, x \in (pH, ORP, UV \dots\dots), y \in (pH, ORP, UV \dots\dots)$. The value of $threshold_{indicator}$ and $threshold_{alarm}$ can be determined based on experimental and practical analysis.

The performance of the detection method is measured through detection time (*DT*), true positive rate (*TPR*) and false positive rate (*FPR*). *TPR* and *FPR* can be calculated by

$$TPR = \frac{TP}{TP+FN} \quad (3)$$

$$FPR = \frac{FP}{FP+TN} \quad (4)$$

where *TP* (true positive) is the detection of an actual event (alarm on); *FP* (false positive) refers to a routine operation being incorrectly classified as a contamination event (alarm on); *TN* (true negative) refers to a routine operation correctly being classified as such (alarm off); *FN* (false negative) is the situation that an actual event is not detected (alarm off). A greater *TPR* means the method is more capable to detect a real event, while a small *FPR* implies the method is less likely to classify a routine operation as an event.

DT is defined as the time difference between a contamination event taking place and when it is detected, and is evaluated by

$$DT = T_1 - T_0 \quad (5)$$

where T_0 is the time when the contamination event occurs and T_1 is the time when

the contamination event is detected. A smaller *DT* means the detection method is more effective and can detect contamination within a shorter time frame.

In this study, the calculation is based on a 1 minute step. A contaminant injection with period of *t* is assumed to be *t* contamination events. Then *TPR* is used to evaluate the performance. For example, for a contamination injection with period of 30 minutes, if a contamination event is first detected at the 10th minute and 4 contamination events are detected within the remaining 20 minutes, *DT* and *TPR* will be 10 minutes and $\frac{(1+4)}{30}$=0.17.

2.6 Evaluation of reproducibility

In this research, the reproducibility of the proposed detection method is evaluated using a concordance correlation coefficient method[34]. Taking the Pearson correlation coefficients as an example, the concordance correlation coefficient method is briefly introduced here. Let us assume that pairs of Pearson correlation coefficients of one type of sensor to the others (Y_{il}, Y_{i2}), i = 1, 2 …, n, are calculated from two independent contaminant injection experiments with means of μ_1 and μ_2 and covariance matrix

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \qquad (6)$$

The concordance correlation coefficient is calculated by

$$\rho_c = \frac{2\beta_1\sigma_2^2}{(\sigma_1^2+\sigma_2^2)+[\beta_0+(\beta_1-1)\mu_2]^2} \qquad (7)$$

in which $\beta_1 = (\sigma_1/\sigma_2)r$ and $\beta_0 = \mu_1 - \beta_1\mu_2$ represent the regression slope and intercept, respectively. r is the Pearson correlation coefficient. If $\rho_c = 1$, two

groups of Pearson correlation coefficients of one sensor from two experiments are in perfect agreement or in perfect reversed (if ρ_c =-1) agreement. For more information about the concordance correlation coefficient method, the readers can refer to Lin's work[34].

3. Experiments and Results

3.1 Correlative responses

As an example, the results from the experiment involving lead nitrate are shown in Figure 2. The experimental results for glyphosate, atrazine, cadmium nitrate, nickel nitrate and trivalent chromium can be found in the supporting information (See Figure F1, F2, F3, F4, F5, supporting information). In the experiment, lead nitrate solutions with concentrations of 0.01mg/l, 0.02mg/l, 0.04mg/l and 0.08mg/l were added in sequence. The concentrations are at the sensors, not in the contaminant tank. This is illustrated using solid green bars at the top of Figure 2. As shown in Figure 2, ORP and nitrate increase due to the presence of lead nitrate, while pH decreases. Sensor responses show correlative relationships, especially for pH, nitrate, ORP and temperature. This suggests the correlative response is caused by the introduction of contaminant and implies that this type of phenomenon can be utilized for detection of the presence of contamination. The magnitudes of the sensors' responses were related to contaminant concentrations. In order to justify the applicability of the proposed

method, the lead nitrate injection experiment with the same configuration was repeated with drinking water. The sensors' responses are shown in Figure F6 (supporting information). As shown in Figure F6, the responses of sensors are similar to the case of source water. This suggests the proposed method can also be used for contamination detection for drinking water. By comparing with results from other types of contaminants (See Figure F1, F2, F3, F4, F5, F6, supporting information), it is clear that the response curves are contaminant-specific, which implies that the correlative response could be utilized not only for contamination detection, but also for contaminant identification.

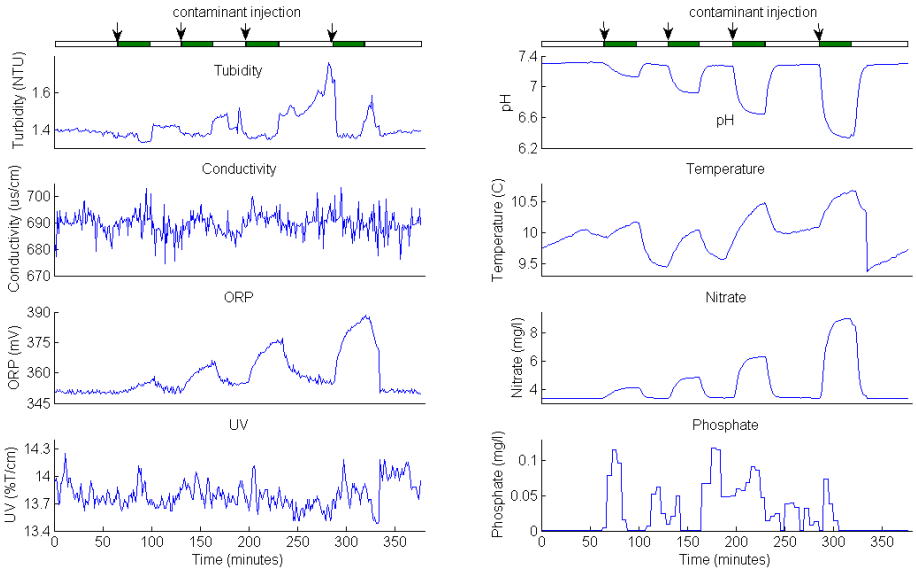


Figure 2 Sensor responses for lead nitrate (concentrations: 0.01, 0.02, 0.04, 0.08mg/l)

Table 4 summarizes the responding sensors for different contaminants. As shown in Table 4, the nitrate, conductivity and UV sensors have correlative responses to the

introduction of atrazine. For ORP, pH, conductivity and nitrate also respond simultaneously to the presence of cadmium nitrate. ORP shows response to atrazine, glyphosate and lead nitrate. Other studies have also revealed a similar phenomenon. For example, Hall et al.[26] conducted a sensor response experiment for 9 types of contaminants and realized that there was more than one sensor responding to each tested contaminant. Yang et al.[15] also reported similar findings from an experiment for 11 contaminants. Drinking water was used in the studies by Hall et al. and Yang et al., while source water is utilized in the current study. The sensor arrays adopted and manufacturers were also different in these three experiments. Therefore, comparisons are difficult to make. However, as shown in Table 4, for all tested contaminants, it was found that multiple types of sensors respond to the introduction of a contaminant and a sensor can respond to different types of contaminants. This verified the correctness of the assumption of the proposed method.

Table 4 The responding sensors to different contaminants

Contaminant	Responding sensors		
	Hall et al. (2007)	Yang et al. (2009)	Current study
	Sensor array: A, B, D, E, G, H, I	Sensor array: A, B, C, D, F	Sensor array: D, F, G, I, J, K, L, M,
	Water type: Drinking water	Water type: Drinking water	Water type: Source water
Aldicarb	A,B,D,E,I	A,B,C,D	
Arsenic trioxide	A,B,D,G,H,I		
Atrazine			D,G,I,J,K,L,M
Cadmium nitrate			D,F,G,I,J,K,M
Colchicine		A,B,C,D	
Dicamba		D,F	
E. coli	A,B,E,H,I	A, B, C, D, F	

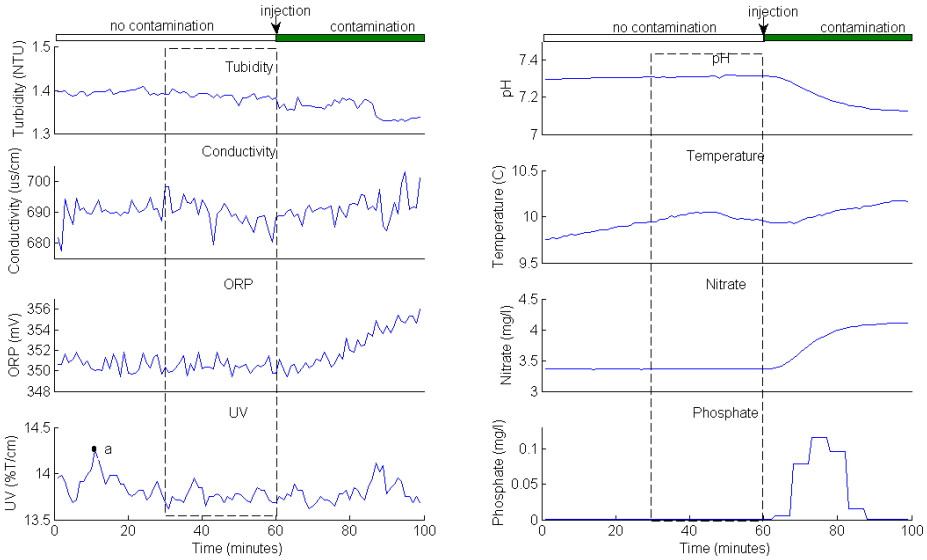
Glyphosate	A,C,D,E	A, B, C, D, F	D,F,G,I,J,M
K Ferricyanide		B,C,D	
Lead nitrate			D,F,G,I,J,M
Malathion	A,D,E,I		
Mercuric Chloride		A,C,D,F	
Nickel Nitrate			D,F,G,I,J
Nicotine	A,B,D,E,G,H	A,B,D,F	
Nutrient broth		A,B,C,D	
Potassium Ferricyanide	A, C,D,E, G,H,		
Terrific broth	A,B,D,E,I	A,B,D,F	
Trivalent chromium			D,F,G,I,J
Typtic soy broth		A,B,C,D	

Note: A - Free chlorine; B – Total chlorine; C – Chloride; D – ORP; E – TOC; F – pH; G – Nitrate-nitrogen; H – Ammonia-nitrogen; I – Turbidity; J – Temperature; K – Conductivity; L – UV; M - Phosphate

3.2 Contamination detection

Taking lead nitrate as an example, the implementation of the proposed method is demonstrated here. Figure 3 presents the responses of 8 types of sensors before and after introduction of lead nitrate solution with concentration of 0.01mg/l (the 1st lead nitrate experiment). As shown in Figure 3, from the 1st to the 60th minute, the system was running on baseline and no contaminant was added. The lead nitrate solution was introduced at the 61st minute and lasted for 37 minutes. The values of $threshold_{indicator}$, $threshold_{alarm}$ and *window size* adopted were 0.8, 6 and 30 respectively. The Pearson correlation coefficients for each couple of sensors and the correlation indicators for 'no contamination' and 'contamination' scenarios were

318 calculated and are listed in Table 5 and Table 6 respectively. The 'no contamination'
319 and 'contamination' scenarios represent two extreme situations. In the 'no
320 contamination' scenario, the dataset containing observations from the 31st to 60th
321 minute (the baseline) were used, while in the 'contamination' scenario, data from the
322 69th to 98th minutes (lead nitrate injected) were adopted.
323



324
325 Figure 3 Sensor responses for lead nitrate (concentration: 0.01mg/l)
326
327 As shown in Table 5, the relationships between sensors' responses in the 'no
328 contamination' scenario are weak. Only turbidity and pH show a moderate negative
329 relationship with a coefficient of -0.61. All Pearson correlation coefficient values are
330 smaller than the value of $threshold_{indicator}$, and the value of the correlation
331 indicator is 0. Therefore, no contamination alarm was trigged off for the 'no
332 contamination' scenario.

333

334 Table 5 Pearson correlation coefficients and correlation indicators (no contamination)

	Turb.	pH	Cond.	Temp	ORP	Nitra.	UV	Phos.
Turb.	1.00(0)	-0.61(0)	0.34(0)	0.04(0)	0.09(0)	-0.12(0)	-0.19(0)	0.00(0)
pH	**-0.61(0)**	1.00(0)	-0.45(0)	-0.25(0)	0.14(0)	0.49(0)	0.07(0)	0.00(0)
Cond.	0.34(0)	-0.45(0)	1.00(0)	0.16(0)	0.08(0)	-0.13(0)	0.05(0)	0.00(0)
Temp.	0.04(0)	-0.25(0)	0.16(0)	1.00(0)	0.28(0)	-0.35(0)	0.20(0)	0.00(0)
ORP	0.09(0)	0.14(0)	0.08(0)	0.28(0)	1.00(0)	0.09(0)	0.17(0)	0.00(0)
Nitra.	-0.12(0)	0.49(0)	-0.13(0)	-0.35(0)	0.09(0)	1.00(0)	-0.35(0)	0.00(0)
UV	-0.19(0)	0.07(0)	0.05(0)	0.20(0)	0.17(0)	-0.35(0)	1.00(0)	0.00(0)
Phos.	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	1.00(0)
Sum of correlation indicator: 0								

335 Note: numbers outside of brackets are Pearson correlation coefficients; numbers in
336 brackets are correlation indicators.

337

338 In the 'contamination' scenario, as shown in Table 6, ORP and nitrate show strong
339 negative relationships with pH. The Pearson correlation coefficients are -0.88 (ORP),
340 and -0.99 (nitrate) respectively. ORP shows strong positive relationships with nitrate
341 (0.84), and negative relationships with phosphate (-0.84) and pH (-0.89). The value of
342 the correlation indicator was calculated to be as 18. Based on this evaluation, a
343 contamination event was confirmed at the 99th minute.

344 Table 6 Pearson correlation coefficients and correlation indicators (lead nitrate, the 1st
345 experiment)

	Turb.	pH	Cond.	Temp	ORP	Nitra.	UV	Phos.
Turb.	1.00(0)	0.59(0)	-0.14(0)	-0.68(0)	-0.75(0)	-0.50(0)	-0.38(0)	0.73(0)
pH	0.59(0)	1.00(0)	-0.26(0)	**-0.98(1)**	**-0.89(1)**	**-0.99(1)**	-0.48(0)	**0.82(1)**
Cond.	-0.14(0)	-0.26(0)	1.00(0)	0.32(0)	0.28(0)	0.25(0)	-0.07(0)	-0.27(0)
Temp.	-0.68(0)	**-0.98(1)**	0.32(0)	1.00(0)	**0.93(1)**	**0.95(1)**	0.38(0)	**-0.84(1)**
ORP	-0.75(0)	**-0.89(1)**	0.28(0)	**0.93(1)**	1.00(0)	**0.84(1)**	0.43(0)	**-0.84(1)**

Nitra.	-0.50(0)	**-0.99(1)**	0.25(0)	**0.95(1)**	**0.84(1)**	1.00(0)	0.49(0)	-0.76(0)
UV	-0.38(0)	-0.48(0)	-0.07(0)	0.38(0)	0.43(0)	0.49(0)	1.00(0)	-0.44(0)
Phos.	0.73(0)	**0.82(1)**	-0.27(0)	**-0.84(1)**	**-0.84(1)**	-0.76(0)	-0.44(0)	1.00(0)
Sum of correlation indicator: 18								

Note: numbers outside of brackets are Pearson correlation coefficients; numbers in brackets are correlation indicators.

Figure 3 shows that graphs for turbidity, conductivity and UV have a number of peaks and troughs. No significant differences before and after introduction of lead nitrate (the left and right parts of each graph) can be observed. The peaks and troughs are mainly due to equipment noises. These noises are independent and not related to other sensors' responses. This is verified by the weak Pearson correlation coefficients for turbidity, conductivity and UV in Table 5 and Table 6. This also suggests that turbidity, conductivity and UV do not respond to the presence of lead nitrate. If a detection decision were drawn in the light of these peaks or troughs, false positive and false negative errors would be obtained.

A common question for the contamination detection method is how fast the contamination event can be detected or what the *detection time* is. In a practical situation, the proposed method will calculate the Pearson correlation coefficients and correlation indicators, and make a detection decision at each time step (1 minute for the sensors used in this research) once the new readings from online sensors are received. As shown by the rectangle with a dashed line in Figure 3, the calculation

starts from the 30th minute with the time step of 1 minute. The sums of correlation indicators and detection time are listed in Table 7. It is shown that the proposed method can detect a contamination event 9 minutes after a 0.01 mg/l lead nitrate solution is added to the water, with the *window size* of 30, $threshold_{indicator}$ value of 0.8 and $threshold_{alarm}$ value of 6. Meanwhile, using Equation 3 and 4, the *TPR* and *FPR* were calculated as being 78.95% and 0, respectively.

Table 7 The correlation indicator and detection time

Time (m)	Sum of C_{xy}	*DT*(m)	Time (m)	Sum of C_{xy}	*DT*(m)	Time (m)	Sum of C_{xy}	*DT*(m)
30	2	N/T	53	0	N/T	76	6	16
31	2	N/T	54	0	N/T	77	6	17
32	2	N/T	55	0	N/T	78	6	18
33	2	N/T	56	0	N/T	79	6	19
34	2	N/T	57	0	N/T	80	6	20
35	2	N/T	58	0	N/T	81	10	21
36	2	N/T	59	0	N/T	82	10	22
37	2	N/T	60	0	N/T	83	10	23
38	2	N/T	61	0	N/T	84	8	24
39	2	N/T	62	0	N/T	85	8	25
40	2	N/T	63	0	N/T	86	10	26
41	2	N/T	64	2	N/T	87	10	27
42	2	N/T	65	2	N/T	88	12	28
43	2	N/T	66	2	N/T	89	12	29
44	2	N/T	67	4	N/T	90	12	30
45	2	N/T	68	2	N/T	91	12	31
46	2	N/T	**69**	**6**	**9**	92	12	32
47	2	N/T	70	6	10	93	12	33
48	0	N/T	71	6	11	94	12	34
49	0	N/T	72	6	12	95	12	35
50	0	N/T	73	6	13	96	14	36
51	0	N/T	74	6	14	97	16	37
52	0	N/T	75	6	15	98	18	38

Note: N/T means 'not detected'.

Meanwhile, for the contaminants examined in this study, Table 8 summarizes detection performances. As shown in Table 8, by taking the default parameter values,

the proposed method has rather good performance for all contaminants under discussion.

Table 8 Summary of detection performance for different contaminants

Contaminant	*TPR*	*FPR*	Detection time (minute)
Glyphosate	0.80	0.10	2
Cadmium	1.00	0.00	1
Atrazine	1.00	0.00	1
Nickel	0.76	0.00	3
Chromium	0.79	0.00	8
$threshold_{indicator}$= 0.6, $threshold_{alarm}$= 8 and *window size* = 30			

4. Discussion

4.1 Discrimination of equipment noise and contamination

In previous studies, a detection alarm was normally set off if the deviation between the predicted and observed sensor values was greater than a preset threshold value. However, this type of judgment is greatly dependent on the reliability and stability of a sensor. For example, as shown in Figure 3, some peaks and troughs (e.g. point *a* in the UV graph) shifted significantly from the previous reading due to equipment noise. This type of shift is difficult to predict and a big deviation between prediction and observation can be expected. If the detection decision is made based on the deviation, a false positive error would occur. Taking the autoregressive moving average method reported by Hou[35] as an example, and setting the 1% deviation of prediction and observation as the detection threshold, point *a* in Figure 3 (also in Figure F7 in the supporting information) was grouped into a contamination event, which is obviously a false positive. Meanwhile, Figure F7 shows all the deviations between the predictions

and observations. The red line is the detection threshold. The *TPR* and *FPR* were calculated as 0.28 and 0.32. In this study, the proposed method overcomes this by exploring the correlative response between sensors. As shown in Figure 3 and Table 5, although conductivity, UV and ORP show obvious fluctuations at the same time period, their correlative relationships are weak, which means the fluctuations are more related to the equipment noise than external reasons, for example, presence of contamination. By exploring the internal relationship proposed in this method, the influence of equipment reliability and stability on detection can be reduced.

In the case of ORP, as shown in Table 5 and Table 6, the Pearson correlation coefficients of ORP with other sensors are 0.09 (turbidity), 0.14 (pH), 0.08 (conductivity), 0.09 (nitrate), 0.17(UV) and 0.00 (phosphate) for the left part (no contamination) and -0.75 (turbidity), -0.89 (pH), 0.28 (conductivity), 0.84 (nitrate), 0.43 (UV) and -0.84 (phosphate) for the right part (contaminant added) respectively. This means the fluctuations in the left part are mainly due to random equipment noises, while the fluctuations in the right part are mainly due to the introduction of contaminant. The significant values of Pearson correlation coefficients in Table 6 also further indicate the correlative responses to the introduction of lead nitrate. A key to an efficient contamination detection method is being able to discriminate between these two types of fluctuations. As shown in the left part of the ORP graph, the peaks or troughs shift significantly from other readings. If a detection decision is made

based on the deviation of response from the sensor, a false positive can be expected. In the proposed method, the difference between these two types of fluctuations is evaluated and differentiated using the Pearson correlation coefficients.

4.2 Impacts of parameters

In the proposed method, there are three parameters: $threshold_{indicator}$, $threshold_{alarm}$ and *window size*. The values of these parameters might influence the performance of the detection method. In order to understand this, their impact on the performance of the detection method was investigated. To facilitate the analysis, the other two parameters were kept constant when analyzing one parameter. The default values for $threshold_{indicator}$, $threshold_{alarm}$ and *window size* were 0.8, 6 and 30 respectively. The performance of detection was evaluated using *TPR* and *FPR*.

The data used for this analysis were originally from the first lead nitrate experiment (Figure 2) with some arbitrary combinations. Four datasets were regrouped. Datasets 1, 2, 3, and 4 are the data for baseline and for lead nitrate with concentration of 0.01 mg/l, 0.02 mg/l, 0.04mg/l and 0.08 mg/l respectively. For each parameter configuration, *TPR* and *FPR* were calculated with the time step of 1 minute. The total number of calculations for each dataset for 1 parameter equals the difference between the length of dataset and *window size*. The averaged *TPR* and *FPR* over the entire time period were deemed to be the true positive rate and false positive rate associated with

this parameter configuration. The analysis results are shown in Figure 4, Figure 5 and Figure 6, in which *TRP* 1, *TPR* 2, *TPR* 3 and *TPR* 4 refer to the *TPRs* for the concentration of 0.01 mg/l, 0.02 mg/l, 0.04 mg/l and 0.08 mg/l.

4.2.1 Parameter $threshold_{indicator}$

In the analysis for $threshold_{indicator}$, its value changes from 0.6 to 0.9 with the step of 0.1 while *window size* and $threshold_{alarm}$ are kept constant. As shown in Figure 4, *FPR* and *TPRs* decrease with an increase in the $threshold_{indicator}$ value. This indicates that a small $threshold_{indicator}$ value could incorrectly classify equipment noise as a contamination event. Meanwhile, it also suggests that a high $threshold_{indicator}$ might result in the overlook of a real contamination event.

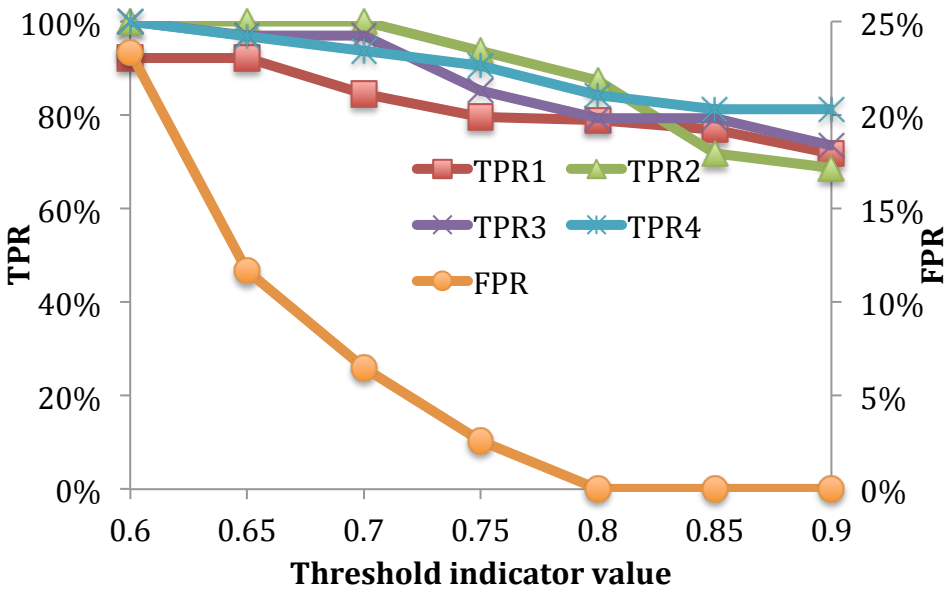


Figure 4 Impact of $threshold_{indicator}$ on detection performance

4.2.2 Parameter $threshold_{alarm}$

In the analysis for $threshold_{alarm}$, its value changes from 4 to 12 in step of 1. As shown in Figure 5, when the $threshold_{alarm}$ increases, both *FPR* and *TPRs* decrease. From $threshold_{alarm}$ = 5, *FPR* approaches 0, which means a large $threshold_{alarm}$ value can significantly reduce false positives. However, a large $threshold_{alarm}$ can also lead to false negatives and result in a low true positive rate, especially for the case of low concentrations. It also shows that the *TPR1* for dataset 1 (lead nitrate concentration at 0.01 mg/l) drops significantly with an increase in the $threshold_{alarm}$ value. This might be because the correlative responses are rather weak at low concentration. This is consistent with the graphs in Figure 2.

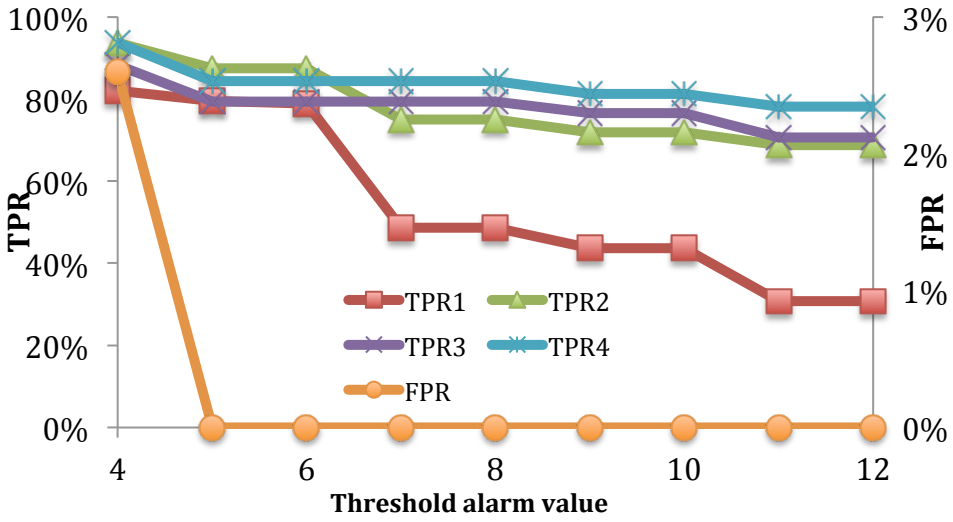


Figure 5 Impact of $threshold_{alarm}$ on detection performance

4.2.3 Parameter *window size*

Window size denotes the number of data involved in the calculation of the Pearson

correlation coefficient. In this analysis, the window size from 8 to 34 in step of 2. The maximum value of the *window size* is based mainly on the contaminant injection period. As shown in Figure 6, when *window size* is smaller than 18, *TPR*s decrease with an increase in the value of *window size*, while they keep rather flat after 18. This implies that the performance of the detection method is more sensitive to the small values of *window size*. Figure 6 also suggests that *FPR* reaches peak values with medium values of *window size*.

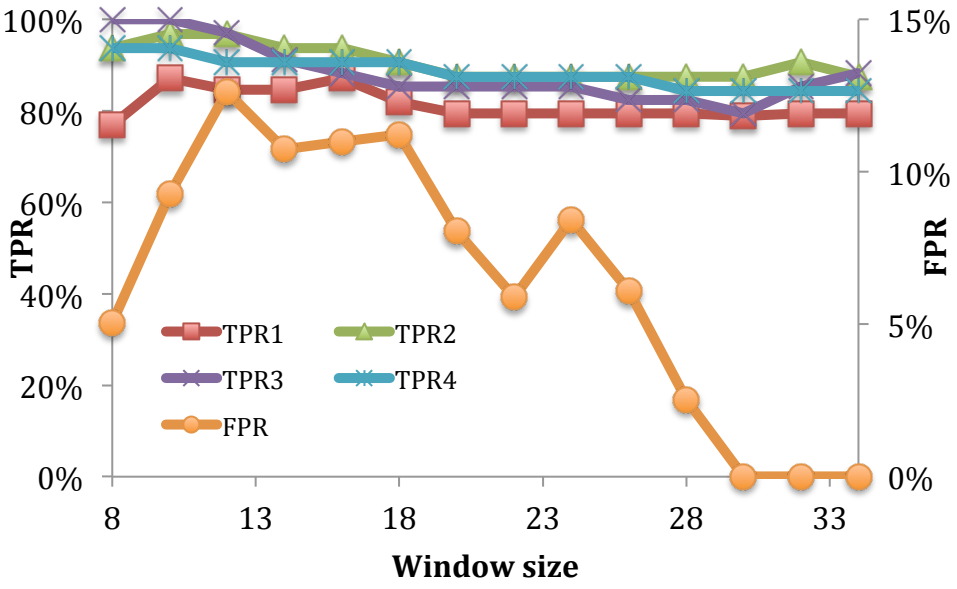


Figure 6 Impact of window size on detection performance

From Figures 4, 5 and 6, it is concluded that the values of parameters have impacts on the performance of the detection method, which suggests that they should be determined carefully to achieve a better detection performance in practical application. For example, by taking the value of $threshold_{indicator}$= 0.75, $threshold_{alarm}$= 6

and *window size* = 30, as shown in Figure 4, a 0.01mg/l lead nitrate contamination event can be detected with a true positive rate of 80% and false positive rate of 2%. The detection time is 9 minutes (shown in Table T1 in supporting information). A comprehensive sensitivity analysis would benefit the implementation of the proposed method and the optimal values of parameters should be determined for a specific contaminant through experiment and analysis.

4.3 **Reproducibility**

Data from two independent lead nitrate injection experiments were used to evaluate the reproducibility of the proposed detection method. Figure F8 shows the sensors' responses from the second lead nitrate injection with the sequence of 0.08mg/l, 0.04mg/l, 0.02mg/l and 0.01mg/l (Figure 2 is for the first injection experiment). The experiment conditions in the two experiments are the same. Similar to Table 6, Table 9 shows the Pearson correlation coefficients for the 2nd experiment, which are directly calculated from the experiment data. Each column represents the Pearson correlation coefficients of other sensors and this sensor. For example, the 2nd column lists the Pearson correlation coefficients of turbidity and other sensors. Using equation 7, the concordance correlation coefficients of data in Table 6 and Table 9 were obtained and are shown at the bottom of Table 9. As shown in Table 9, the concordance correlation coefficients for turbidity, pH, temperature, ORP, nitrate and phosphate are greater than 0.81, which suggests high agreement between the

calculation results of the data from the 1st and the 2nd lead nitrate experiments. It is also noticed the sensors with high concordance correlation coefficients consist with the one having high correlative coefficients in Table 6 and Table 9. This implies that the responses of these sensors are mostly due to injection of lead nitrate, rather than equipment noises. For conductivity and UV, the concordance correlation coefficients are low, which is consistent with their low values of Pearson correlation coefficients in Table 6 and Table 9. This suggests the responses of conductivity and VU are mostly from equipment noises. Therefore, their reproducibility is low. This will not affect the reproducibility of the proposed method since the low Pearson correlation coefficient values are not taken into account for the correlation indicator. By using the same parameter values (the default ones), the *TPR* and *FPR* were calculated as being 78.95% and 0 respectively, which are the same as the ones from the first experiment. In a summary, the proposed method has a good reproducibility.

Table 9 The Pearson correlation coefficients and the concordance correlation coefficients ρ_c (lead nitrate, the 2nd experiment)

	Turb.	pH	Cond.	Temp	ORP	Nitra.	UV	Phos.
Turb.	1.00	0.53	-0.09	-0.62	-0.73	-0.45	-0.46	0.71
pH	0.53	1.00	-0.18	-0.98	-0.88	-0.99	-0.62	0.74
Cond.	-0.09	-0.18	1.00	0.24	0.17	0.17	0.00	-0.22
Temp.	-0.62	-0.98	0.24	1.00	0.92	0.96	0.56	-0.78
ORP	-0.73	-0.88	0.17	0.92	1.00	0.83	0.61	-0.83
Nitra.	-0.45	-0.99	0.17	0.96	0.83	1.00	0.61	-0.67
UV	-0.46	-0.62	0.00	0.56	0.61	0.61	1.00	-0.53
Phos.	0.71	0.74	-0.22	-0.78	-0.83	-0.67	-0.53	1.00
ρ_c	**0.81**	**0.96**	-0.06	**0.96**	**0.86**	**0.95**	0.53	**0.88**

517

518 4.4 Future works

519 Using lead nitrate as an example, this study shows that the detection performance is
520 highly sensitive to the values of parameters. Meanwhile, in a real situation, the type of
521 contaminant is not known a priori. Therefore, the values of the parameters should not
522 be determined for a specific type of contaminant, but for a large group of
523 contaminants. A generalized set of values of parameters is more deserved. Therefore,
524 in the future, the optimal determination of the values of the parameters should be
525 conducted.

526 In an EWS, the question after detection of contamination is how to identify the type
527 of contaminant quickly. A commercially available Hach method relies on correlations
528 between responses of different types of sensors for contaminant identification. From
529 this study, as shown in the Figure F1, F2, F3, F4 and F5, the sensors' responses are
530 contaminant dependent. For a clearer view, Figure 7 shows all sensors' responses for
531 lead nitrate and atrazine using radar map. Obviously, the shapes formed by axes for
532 lead nitrate and atrazine are different. By utilizing the contaminant dependent feature,
533 it is possible to differentiate the types of the contaminants. Therefore, in the future,
534 more work should be done on how to extract features and patterns for contaminant
535 identification. Possible techniques are data mining and pattern recognition. For
536 example, by comparing the Euclidean distance between samples and classes for
537 different contaminants, it is possible to identify the type of the contaminant.

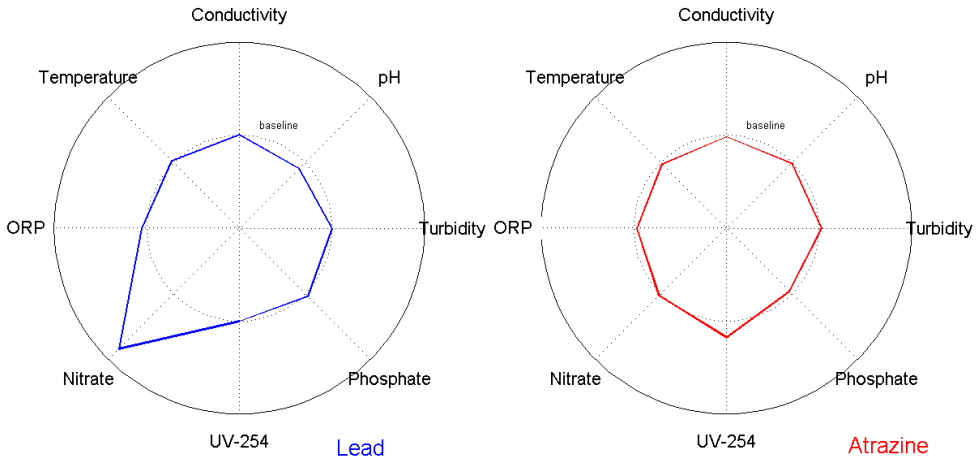


Figure 7 Radar map of sensor's responses for lead and atrazine

5. Conclusion

EWS should provide a fast and accurate means to distinguish between normal variations and contamination events. Compared with component-specific sensors, conventional water quality sensors are still widely used because they are low cost and easily maintained. For an EWS with conventional water quality sensors, a key issue is how to efficiently detect the presence of contamination. In this study, a platform with 8 types of online water quality sensors was established and utilized for contaminant injection experiment. By analyzing the results from the experiment, the following conclusions are drawn.

1) A contamination detection method utilizing the correlations between multiple types of conventional water quality sensors was proposed in this study. The results from the experiment and analysis show that the proposed method could detect a lead nitrate contamination 9 minutes after the introduction of contaminant at the concentration of 0.01mg/l using responses from online water

quality sensors. The *TRP* and *FPR* are 80% and 2% respectively.

2) It was noticed that multiple sensors responded simultaneously to the presence of contamination. Initial analysis showed that the correlative response is contaminant-specific, which implies that the correlative response could be utilized not only for contamination detection, but also for contaminant identification and even for quantification. Meanwhile, in some previous studies, an ideal sensor is assumed. In a real situation, this is not always the case. If the sensor fails to function properly, a false positive is expected. The proposed method can overcome this by utilizing the correlations between sensors.

3) The proposed method employs three parameters: $threshold_{indicator}$, $threshold_{alarm}$ and window size. From the analysis, it was concluded that these parameters have an impact on the detection performance. For a specific contaminant, the optimal values of parameters should be determined through experiment and analysis.

4) The basis of the proposed method is that multiple sensors respond to one type of contaminant simultaneously. This has been verified by experiment in this study and other studies. However, it should be envisaged that the types of contaminants previously tested are still limited. More experiments should be conducted in the future.

Acknowledgements

This work is jointly supported by National High Technology Research (2013AA065205), Development Program and Tsinghua Independent Research Program (2011Z01002) and Water Major Program (2012ZX07408-002).

Reference

1. USEPA, *Baseline threat information for vulnerability assessments of community water systems*, Washington, DC, 2002.

2. USEPA, *Planning for and responding to drinking water contamination threats and incidents*, Washington, DC, 2003.

3. W. M. Grayman, R. A. Deninger and R. M. Clark, *CE News*, 2002, 14, 34-38.

4. C. Wang, Y. J. Feng, S. S. Zhao and B. L. Li, *Chemosphere*, 2012, 88 (1), 69-76.

5. J. Yang, J. Bi, H. Y. Zhang, F. Y. Li, Zhou, J. B. Liu, *China Environmental Science*, 2010, 30(4), 571~576 (in Chinese).

6. J. Hasan, S. States, R. Deininger, *Journal of Contemporary Water Research and Education*, 2004, 129, 27-33.

7. M.V. Storey, B. van der Gaag and B. P. Burns, *Water Research,* 2011, 45 (2), 741-747.

8. M. Brussen, *On-line Water Quality Monitoring*, Sydney Water Report, Sydney, 2007.

9. C. J. de Hoogh, A. J. Wagenvoort, F. Jonker, J. A. Van Leerdam and A. C. Hogenboom, *Environmental Science & Technology*, 2006, 40, 2678-2685.

596 10. J. Jeon, Kim, J. H. Lee, B.C., S. D. Kim, *Science of the Total Environment*, 2008,
597 389, 545-556.
598 11. B. van der Gaag and J. Volz, *Real-time On-line Monitoring of Contaminants in*
599 *Water: Developing a Research Strategy from Utility Experiences and Needs*, KIWA
600 Water Research, Nieuwegein, 2008.
601 12. R. K. Henderson, A. Baker, K. R. Murphy, A. KHambly, R. M. Stuetz and S. J.
602 Khan, *Water Research*, 2009, 43 (4), 863-881.
603 13. P. R. Hawkins, S. Novic, P. Cox, Neilan, B. A. Burns, B. P. Shaw, G., W.
604 Wickramasinghe, Y. Peerapornpisal, W. Ruangyuttikarn, T. Itayama, T. Saitou, M.
605 Mizuochi and Y. Inamori, *Journal Water Supply: Research and Technology*, 2005, 54,
606 509-518.
607 14. C. P. Marshall, S. Leuko, C. M. Coyle, M. R. Walter, B. P. Burns and B. A.
608 Neilan, *Astrobiology*, 2007, 7, 631-643.
609 15. J. Y. Yang, R. C. Haught and J. A. Goodrich, *Journal of Environmental*
610 *Management*, 2009, 90, 2494–2506.
611 16. USEPA, *WaterSentinel System Architecture,* U.S. Environmental Protection
612 Agency, Office of Water, Office of Ground Water and Drinking Water, Washington,
613 D.C., 2005.
614 17. USEPA, *Water Security Initiative Cincinnati Pilot Post implementation System*
615 *Status: Covering the Pilot Period: December 2005 Through December 2007,* U.S.
616 Environmental Protection Agency, Office of Water, Office of Ground Water and

617 Drinking Water, Washington, D.C., 2008.

618 18. D. Hart, S. A. McKenna, K. Klise, V. Cruz and M. Wilson, Proceedings of the

619 World Environmental and Water Resources Congress. ASCE, Reston, VA, 1-9, 2007.

620 19. K. A. Klise and S. A. McKenna, Proceeding of the International Society for

621 Optical Engineering, Saito, 2006.

622 20. I. O. Bucak and B. Kalik, *EKOLOJI*, 2011, 20(78), 75-81..

623 21. M. Bouamar and M. Ladjal, *International Journal of Computational Intelligence*

624 *and Applications*, 2012, 11 (2), 1250013-1-1250013-14.

625 22. S. Alvisi, M. Franchini, M. Gavanelli and M. Nonato, *Journal of*

626 *Hydroinformatics,* 2012, 14(2), 345-365.

627 23. M. Guidorzi, M. Franchini and S. Alvisi, *Urban Water,* 2008, 6(2), 115–135.

628 24. A. Ostfeld and E. Salomons, *Journal of Water Resources Planning and*

629 *Management,* 2004,130(5), 377–385.

630 25. A. Preis and A. Ostfeld, *Journal of Hydroinformatics,* 2008, 10(4), 267–274.

631 26. J. Hall, A. D. Zaffiro, R. B. Marx, P. C. Kefauver, E. R. Krishnan, R. C. Haught

632 and J. G. Herrmann, *Journal – American Water Works Association*, 2007, 99(1),

633 66-77.

634 27. D. Kroll, *Securing our water supply: Protecting a vulnerable resource*, Pennwell,

635 2006.

636 28. D. Kroll, K. King, Proceeding of the 8th Annual Water Distribution Systems

637 Analysis Symposium, ASCE, Reston, VA, 2006.

638 29. S. Murray, M. Ghazali and E. A. McBean, *Journal of Water Resources Planning*
639 *and Management*, 2012, 138(1), 63- 70.
640 30. N. Aliker and A. Ostfeld, *Water Research*, 2014, 51, 234-245.
641 31. L. Perelman, J. Arad, M. Housh, M. and A. Ostfeld, Environmental Science &
642 Technology, 2012, 46, 8212-8219.
643 32. J. Arad, M. Housh, L. Perelman and A. Ostfeld, *Water Research,* 2013, 47 (5),
644 1899-1908.
645 33. J. S. Hall, J. Szabo, S. Panguluri and G. Meiners*, Distribution system water*
646 *quality monitoring: sensor technology evaluation methodology and results*, EPA
647 600/R-09/076, 2009.
648 34. L. I. Lin, *Biometrics*, 1989, 45(1), 255-268.
649 35. D. Hou, X. Song, G. Zhang, H. Zhang and H. Loaiciga, *Environmental Science*
650 *and Pollution Research*, 2013, 20, 4496-4508.