



January 2010 Volume 102 Number 1

Journal

Recent years have seen the arrival of an array of early warning systems for the continuous on-line detection of anomalies relating to water security and quality. Numerous methodologies and criteria have been suggested to determine the efficacy of these methods in real-world scenarios. The authors contend that many of these evaluation techniques are deficient in evaluating all of the factors that would be essential in defining a system for deployment for the protection of key installations such as military bases. They suggest specific criteria as a means of comparing various technologies and propose a new method for determining receiver operating characteristic curves. Similar considerations should come into play in the evaluation of systems for civilian uses.

Methods for evaluating water distribution network early warning systems

The ability to detect and act on changes in water quality is a critical component in the drive to protect US drinking water supplies from intentional or accidental contamination. The distribution system represents the last analytical frontier in the water quality industry. The monitoring of source water and treatment plant processes has progressed to a level at which water suppliers can be confident that they are providing good quality water from the plant to the distribution system. Once the water reaches aging distribution systems, however, assurance of its continued integrity is limited. From a historical perspective, most monitoring in the distribution system has been relegated to the occasional snapshot provided by grab sampling for a few limited parameters or the infrequent regulatory testing required by such mandates as the Total Coliform Rule.

Several studies conducted since 9/11 have shown that bulk monitoring of basic water quality parameters has the potential to indicate the presence of many harmful agents in water at the levels of interest (Hall et al, 2007; Kroll, 2006; USEPA, 2005). The realization of the potential of bulk parameter monitoring as a practical tool to detect terrorism-related acts has led to the development of numerous sensor packages designed for deployment in the distribution system.

Several communities have installed these systems in various locations throughout their distribution networks. These continuous on-line systems have recorded large streams of data (at some sites for several years) relevant to the water quality in the distribution systems in which they have been deployed. These data streams are quite complex, and it becomes difficult to differentiate normal background noise and fluctuations attributable to everyday occurrences from changes that are indicative of a deviation in water quality deserving fur-

DAN KROLL AND KARL KING

ther attention. If useful decision trees are to be built from this type of monitoring, then computer-aided data interpretation is essential.

Intelligent algorithms should be capable of detecting the subtle changes in bulk parameter readings that are indicative of an incursion into the system without burdening the operators with a constant stream of false or trivial alarms. They should also be capable of differentiating the unique pattern of responses that are elicited by different classes of agent. These differences may be enough to narrow down the cause of unusual occurrences and possibly fingerprint the class of disturbance that caused the anomaly.

The first priority for any early detection system is that it detects contaminants in an effective manner that ensures the protection of people and infrastructure.

Over the past several years, numerous algorithms intended for this use have been under development by private industry, government programs, universities, and national labs. The question then becomes how to evaluate the effectiveness of these event detection system (EDS) solutions (sensors plus algorithms) at a given site. Although several research studies (McKenna et al, 2008; Umberg, 2008; McKenna et al, 2006) and programs such as USEPA's Environmental Technology Verification program (Battelle, 2005) have attempted to set criteria and means for evaluating such systems, many important functional criteria for such systems have been overlooked in past studies. In the following section, the authors consider the factors that are mandatory for the success of any such system.

CRITERIA FOR SYSTEM SUCCESS

Dual-use value. The first priority for any EDS is that it detects contaminants in an effective manner that ensures the protection of people and infrastructure. The ability to detect contamination is essential, but if a system is to be capable of offering a likely return on investment and achieving widespread deployment, it would also be advantageous if the EDS can provide information that would be useful on a day-to-day basis. It is presumed that an EDS spends very little time in an actual alarm condition. What is the unit doing for the rest of the time? Other functions that can help justify the cost of an EDS include

- optimizing daily operations within the system monitored,
- providing alarms for operational occurrences not related to contamination threats,

- replacing grab sampling for compliance purposes with continuous monitoring,
- documenting system operation anomalies to assist with planning maintenance activities or planning and justifying major system upgrades (line replacement), and
- building consumer confidence by continuously documenting system water quality.

Such capabilities can reduce the cost of operations for the system being monitored and possibly provide information that can be helpful to those people who use the monitored system.

Detection class requirements. There are several different classification levels for early warning systems that relate to their effectiveness and are commonly used to categorize systems for military use. Though not in widespread use within the water industry, these classifications offer a useful guide for categorizing the EDS. The following detection classes are possible:

Detect to treat. These systems have a very high confidence level (nearly 100%), which allows treatment to proceed on those who have been exposed to contaminants. Response time is slow, cost high.

Detect to protect. These systems have a high confidence level (> 99%), which allows for protection by limiting contaminant exposure without the need for confirmatory testing. Response time is fast, cost medium.

Detect to warn. This class has a presumptive confidence level (< 99%), which allows for protection by limiting exposure to contaminants while confirmatory tests are run. Response time is fast, cost low.

Confidence levels usually are a function of analysis time and cost, with higher confidence levels increasing both. Because an EDS will need to respond and alert rapidly in order for the utility response to be effective yet must be relatively low in cost to provide wide coverage, this article assumes that the first deployed defense against contamination will be systems classified as "detect to warn."

Coverage characteristics. *Cost.* Cost may not be significantly limiting when selecting an EDS for an iconic or high-profile event or facility; however, if the coverage required is large (such as a major metropolitan area), budgets may be constrained, and the degree of coverage becomes a function of the cost per point monitored. Therefore, cost becomes an issue.

Area of protection. Effective coverage may be a function of the hydraulics of the distribution system within the geographical setting. Even when a deployment site may be ideal from a protection standpoint, it may not be capable of being used because of system requirements, site noise functions, and other logistics. Although "protection of all" may be a laudable goal, reality may constrain the degree of implementation, forcing tradeoffs.

Communication. Monitoring of multiple points in a geographic area immediately raises the need for communication at least from the remote points to some centralized facility where the data can be interpreted and

actions taken. The EDS can accomplish highly sophisticated interpretations of the local data, but it cannot take actions in a complex situation affecting possibly millions of people; human interaction and analysis will be required. The information from the multiple points must be communicated to an analysis and command center. An effective EDS must be structured for secure communication. Both the instrument and the network must have tools in place to ensure a high level of security so that information cannot be blocked or false information transmitted on the network.

Operational characteristics. The design of a successful system must incorporate several essential operational characteristics.

Ease of use. A system that may be absolutely critical in a crisis must not be difficult to use. User interface operation should be intuitive enough that even minimally skilled operators can obtain necessary information without resorting to an operator's manual.

Full automation. The system must normally operate without requiring the presence of a human. Human intervention should be needed only for service or maintenance.

Continuous operation. The system must run without any long gaps in analysis that enable contamination of significant duration to slip by undiscovered. The maximum time of nonobservation should be relatively short, i.e., on the order of minutes as opposed to hours. Longer response times allow more contaminated water to pass, thus exposing more people to the hazard before remedial actions can be taken.

Reliability. A nonworking system is an opportunity for exploitation.

Cost-effectiveness. The system's amortized cost per day should be comparable to or less than the labor, travel, equipment, and reagents for an existing grab-sample program. In evaluating an EDS by this criterion, it is important to factor in the overall better picture of operations afforded by continuous monitoring versus discrete measurements.

Performance characteristics. Detection of a broad spectrum of contaminant classes. There are numerous contaminants that could cause serious harm if introduced into a drinking water distribution system. However, even chemicals within a given category may have fundamentally different chemical characteristics. The commonality found within a category does not preclude different responses of analytical sensors used in an EDS. The ability to detect chemicals with significant specificity is dependent on viewing a chemical in many dimensions. Thus, single-parameter systems have inherent limitations.

The optimum system should be able to detect essential categories of contaminants including organophosphates, carbamates, and other pesticides; herbicides; nematicides; animal poisons; petroleum products; heavy metals; infectious agents; warfare agents (tested against real threat agents); toxic industrial chemicals; poisons;

cyanides; toxins; and drugs and pharmaceuticals both legal and illicit.

Rapid response time. The response time of any detection method used on a flowing stream is actually a measure of the delivery rate of harmful contaminants, assuming an effective response. A simple way to view this relationship is given by Eq 1:

$$\text{Lethal Doses Delivered } (n) = \frac{\text{Stream Flow Rate } (v/t) \times \text{Response Time } (t) \times \text{Concentration } (m)}{\text{Lethal Dose } (m/v)} \quad (1)$$

in which n is the number of doses, v is volume, t is time, and m is mass. It is readily apparent that a rapid response time is favorable so that appropriate and effective action can be taken. In light of the flow rates of pipes involved in most deployment scenarios, short detection times (on the order of minutes) would be beneficial in initiating a rapid response and limiting exposure.

Specificity. Specificity of contaminant identification can be obtained in two ways: analysis for a specific compounds or pathogens or general analysis across multiple orthogonal dimensions via mathematical analysis of the data from multiple sensors. Detectors of both types exist, which allows comparison of the two approaches. Given the range of possible contaminants that could be put into a water distribution system, looking for specific molecules would require an immense number of sensors. This approach has so many problems associated with it that it quickly becomes untenable.

The second approach of using a manageable set of orthogonal sensors (nonlinearly correlated or independent) faces only the difficulty of obtaining sufficient information to apply pattern recognition methods that can differentiate among contaminants or classes of contaminants. One advantage of using an orthogonal set of different sensors is that it becomes much more difficult to find a contaminant that can pass by all sensors unnoticed.

Reproducibility. An EDS must be reproducible to be trustworthy. Reproducibility can be demonstrated via testing with actual contaminants (e.g., aldicarb, anthrax culture, cyanide, fluoroacetate, nicotine, ricin, sarin, VX).

Low occurrence of false-negative and false-positive results. A system that is blind to certain classes of chemicals (e.g., those not visible in the ultraviolet 254 nm spectrum) represents an opportunity for the system to be contaminated without triggering an alarm. Systems that have multiple types of sensors provide a certain level of redundancy as well as multiple chances to detect, so this type of system normally will produce some kind of alarm on contamination, and false-negatives will be less likely.

Two factors that are major contributors to false alarms are random system noise (i.e., electronic noise

from instrumentation combined with environmental noise from normal water quality fluctuations) and insufficient information during occurrence data analysis. Given that systems are never noise-free and that there is always some degree of insufficient information, false-positives are inevitable. The ultimate question is the determination of the number of false-positives per time in a given installation (because noise is site-specific) and the acceptable frequency. False-positives caused by noise can be reduced by the proper choice of alarm threshold according to receiver operating characteristic (ROC) curves.

False-positives caused by imperfect information occur in an EDS using an inferential method. Given that a multiparameter system implies an inferential method that can give false-positives and that actual contamination is a rare occurrence, there will be more false-positives than instances of contamination. Nevertheless, any detection must be considered presumptive until followup testing either verifies or denies the detection.

Qualitative capability. Contaminants can be named if the EDS is specific, but the individuals using the information may not have sufficient training to recognize from the name the nature of the contaminant found. For clarification, the system should assign a general classification category along with the specific names, i.e., organophosphate or biological agent, rather than just the specific organism(s) or compound(s).

Quantitative capability. Ideally, an EDS would provide sufficient information to provide some quantitative information about the concentrations of any contaminant

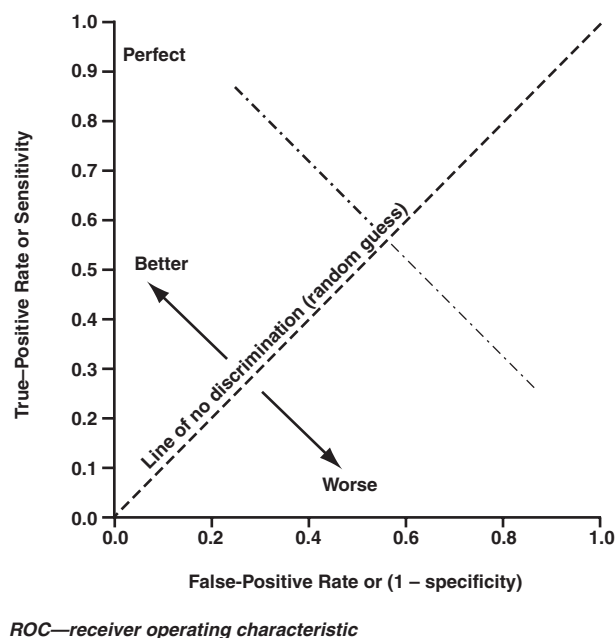
presumably found in the water system. Such information could be useful in determining the threat level during a given episode and could be essential if treatment of people or infrastructure becomes necessary.

Sensitivity. Clearly an EDS must be sensitive to contaminants present in harmful amounts, but quoting a simple minimum detection limit can be misleading. Contaminants change water chemistry, and those changes can be analyzed as part of a process that generates a trigger signal or alarm. Changes are seen against a background of noise, or natural fluctuations in measured parameters. Accordingly, one cannot address sensitivity unless it is relative to a site's noise properties. Fortunately, military needs have produced a useful tool for simply stating a detector's capability: the ROC curve. ROC curves allow an operator to select a detection alarm threshold or sensitivity level according to local noise characteristics. This allows an operator to increase sensitivity, accepting the higher probability of a false alarm when it is suspected that contamination is more likely. Figure 1 shows a representation of ROC space.

The ROC curve is a graphical representation of the tradeoff between the false-negative and false-positive rates for every possible cutoff or threshold. Equivalently, the ROC curve is the representation of the tradeoffs between sensitivity (S_n) and specificity (S_p). By tradition, the plot shows the false-positive rate on the x axis and $1 -$ the false-negative rate on the y axis. This can also be described as a plot with $1 - S_p$ on the x axis and S_n on the y axis. A good diagnostic test is one that has low rates of false-positives and false-negatives across a reasonable range of cutoff values. A bad diagnostic test is one in which the only cutoffs that keep the false-positive rate low have a high false-negative rate and vice versa.

An evaluator is usually satisfied when the ROC curve climbs rapidly toward the upper left-hand corner of the graph. This means that $1 -$ the false-negative rate is high and the false-positive rate is low. On the other hand, a ROC curve that follows a diagonal path from the lower left-hand corner to the upper right-hand corner signifies that every improvement in the false-positive rate is matched by a corresponding decline in the false-negative rate. The rate at which the ROC curve rises to the upper left-hand corner can be quantified by measuring the area under the curve. The larger the area, the better the diagnostic test. If the area is 1.0, the test is ideal because it achieves both 100% sensitivity and 100% specificity. If the area is 0.5, the test effectively has 50% sensitivity and 50% specificity; in other words, the test is no better than flipping a coin. In practice, a diagnostic test will fall somewhere between these two extremes. The closer the area under the curve is to 1.0, the better the test, and the closer the area is to 0.5, the worse the test. The following sections focus on the development and real-world use of an alternative ROC curve method that is more suitable to continuous monitoring situations.

FIGURE 1 Classic ROC curve space



CURVE DEVELOPMENT

A new method for determining ROC curves. The classical ROC curve is plotted parametrically; in other words, from the false alarm rate (FAR) as a function of trigger level threshold (sensitivity) versus hit rate (HR; i.e., detection rate), as a function of trigger level threshold. In the classical ROC curve for this technology, the trigger level chosen is the minimum concentration consistently detected against background and instrument noise. In deployment scenarios, the percentage of detection acceptable is left to the operator. An example is shown in Figure 2.

This format works well for discrete measurements (such as in test kits) but may not be optimized for continuous monitoring technologies. With this in mind, the authors, in collaboration with US military personnel at the US Army Corps of Engineers Construction Engineering Research Laboratory (CERL) and the Edgewood Chemical Biological Center in Aberdeen Proving Ground, Md., developed an alternative presentation that is more useful when the operational analysis is continuous in time. This new presentation makes it easy for the system operator to select the trigger threshold to balance trigger sensitivity against the frequency of false alarms caused by system noise. Although ROC curve methodology can be generalized, all ROC curves of either type, i.e., the classical ROC curve or the alternative ROC presented here, are specific to sensor location and contaminant type. Decisions must be based on a family of ROC curves.

In the alternative presentation, the FAR function is expressed as the mean time between false alarms versus trigger threshold. The HR function is translated as the

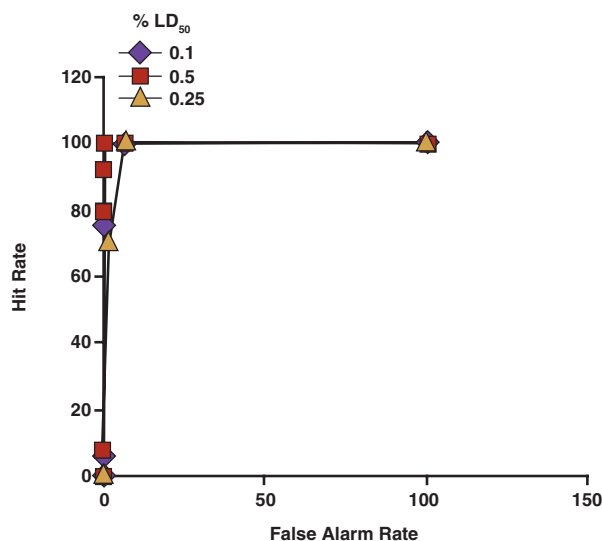
amount of agent, expressed as the percentage of the median lethal dose (LD_{50}) for a 70-kg male drinking 1 L of water, required to provide a 100% HR versus trigger threshold in the absence of other abnormal occurrences in the system such as sensor or communication failure (see example in Figure 3). This format allows the user to select the trigger threshold on the basis of desired trigger sensitivity versus the acceptable time between false alarms attributable to process noise. The blue markers in the figure are points of different trigger threshold values (increasing to the right).

A plot of the classical ROC curve requires curves for the FAR versus trigger threshold and the HR versus trigger threshold. A plot of the alternative form of ROC curve requires curves for the mean time between false alarm versus trigger threshold and the trigger amount (percentage of LD_{50}) versus trigger threshold.

HR versus trigger threshold curve. In the method used by the authors, the HR curve for the classical ROC curve format is derived from a Monte Carlo simulation with a run of 1,000 trials for each point on the HR curve. An agent is selected, which also defines the LD_{50} concentration for the agent. The lab test data for that agent at a known concentration are entered into the spreadsheet. These data are the five deviations from baseline (one deviation per water quality parameter for the system tested) produced when that concentration of an agent is introduced into drinking water of the same type as that used to define the FAR curve.

The method requires a statistical characterization of the deviations from baseline for the process water of

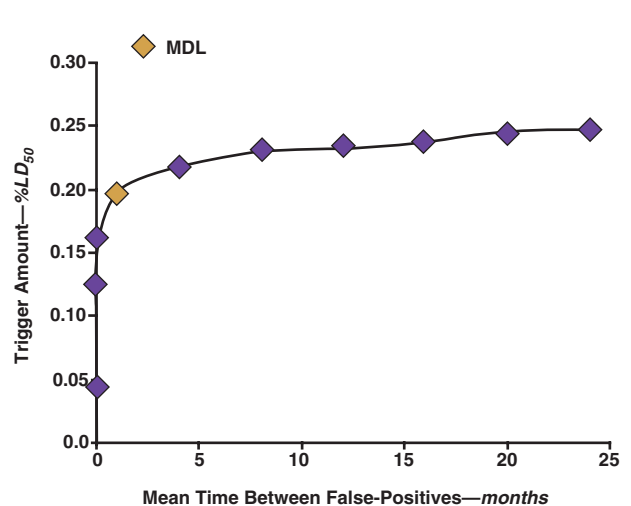
FIGURE 2 Classic ROC curve format for sodium fluoroacetate



LD_{50} —median lethal dose, ROC—receiver operating characteristic

The agent data are derived from laboratory beaker tests, and the water data are from the distribution system of interest.

FIGURE 3 ROC curve format for cyanide for a continuous system



LD_{50} —median lethal dose, MDL—minimum detection limit, ROC—receiver operating characteristic

interest. A representative data set from the process (e.g., one to three weeks without abnormal occurrence content) is processed to find baseline values and deviations from baseline. The parameter measurements used contain parameter variation plus measurement noise from the sensors themselves. The 3-sigma value for each parameter deviation set is then entered into the sheet. Next, the resolution of the measurement devices is defined, and a trigger threshold level and dose value are entered. The spreadsheet then calculates 1,000 independent points at which the process has been dosed and noise added via independent random number generators. In this example, the probability density function for each parameter noise component was selected pessimistically to be a flat distribution between the extreme values of the noise component. The flat distributions used could be considered worst-case representations of the noise.

Measurement values are then adjusted to represent the quantization noise or rounding error found in the analog-to-digital conversion electronics of the system. Pessimistic values for quantization are used. These multiparameter vector values are processed by the selected trigger algorithm to produce a trigger signal for each vector. Values are then compared with the given trigger threshold values in the Monte Carlo simulation to obtain the number of hits (detections) per 1,000 trials. For a given threshold, HR can be calculated as in Eq 2:

$$HR (\%) = 100 \times \text{Number of Hits}/1,000 \quad (2)$$

If the calculated distance measured for a given instance exceeds the threshold, it is called a hit. Monte Carlo

simulations performed for a number of trigger levels generate an HR curve. Figure 4 shows a sample curve for HR versus trigger threshold.

FAR (real-world operational data). The FAR curve for the system was generated from 5,000 data points obtained from a local water plant during a period when operation was normal and no anomalies were present. The data for the five sensed parameters (pH, conductivity, free chlorine, turbidity, and total organic carbon) input to the event monitor were processed by the trigger algorithm to derive 5,000 points of trigger signal versus time. A frequency analysis was performed in a spreadsheet to derive the probability density function and FAR curve. The data were not erratic and showed a one-sided exponential decay function. The mathematical model for the FAR in this case is shown in Eq 3:

$$FAR = e^{-13.4945 \times \text{Trigger Level}} \quad (3)$$

Thus, at a trigger level of zero, any process deviation triggers the alarm (falsely). If the trigger level is set very high (> 1.5), the system rarely triggers and FAR ≈ 0. With the HR data and the FAR data, a ROC curve for the selected agent concentration can then be plotted.

Other concentrations can be selected and the process can be repeated to obtain a family of ROC curves at different agent concentrations. Table 1 shows typical data with this approach. The ROC curves can then be plotted for the agent at the given concentrations. Figure 5 shows an example for the agent sodium fluoroacetate at three concentrations. As shown in the figure, most of the markers for the 1% curve are covered by those for the 0.5% curve.

FIGURE 4 Positive trigger versus threshold for 0.7% LD₅₀ nicotine

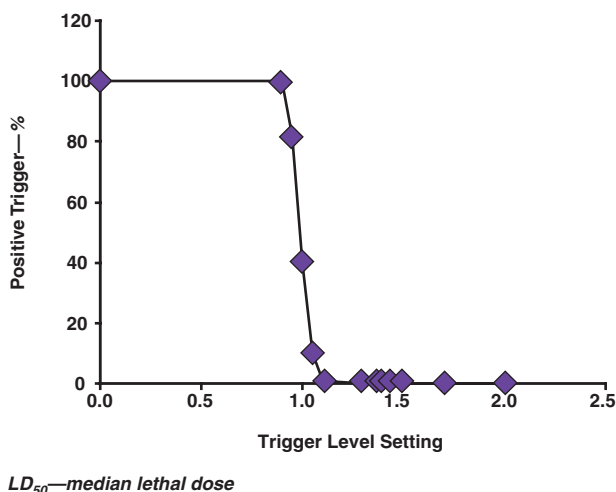
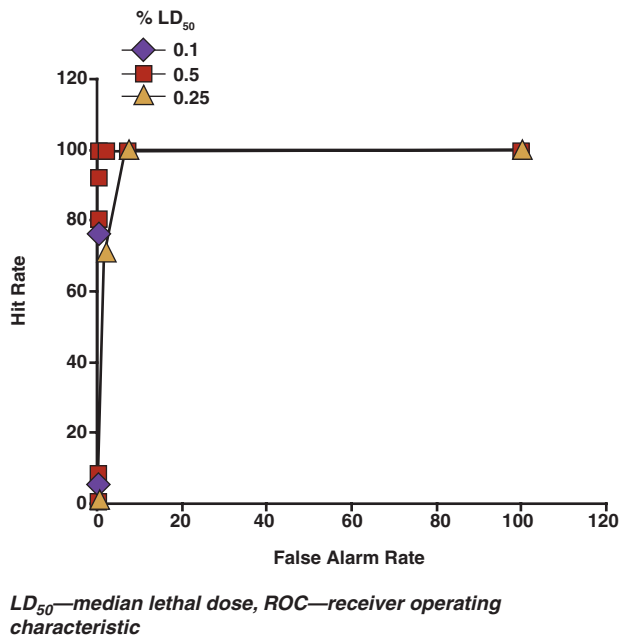


FIGURE 5 ROC curve for sodium fluoroacetate



Trigger amount (percentage of LD₅₀) versus trigger threshold curve. For the modified format ROC curve, the trigger threshold is set at a selected value. The dose is then adjusted to find the dose value at which the HR becomes 100%. A new trigger threshold value is selected and the process is repeated. This process defines the dose versus trigger threshold curve (Figure 6). This is plotted parametrically with the mean time between false-positives versus threshold curve to generate the modified format ROC curve (Figure 7).

EVALUATION OF THE MODIFIED ROC CURVE

Determining ROC curves for data from sites on military bases. Under a cooperative research and development agreement, CERL and the authors worked for the past several years to evaluate early warning systems in real-world deployment scenarios on military bases. In addition to helping to develop theories of operation and response, these deployments resulted in a large body of data from a variety of sites for which ROC curves have been calculated.

In the system evaluated for this study, the EDS algorithm processed five water quality parameter readings (pH, conductivity, chlorine, turbidity, and total organic carbon) once each minute. The algorithm calculated the deviations of these parameters from the process baseline, numerically scaled the five deviations, and then calculated a single trigger signal via a distance measure. The trigger signal was compared with a user-selectable threshold level. If the trigger signal was greater than the threshold, the system was said to be in alarm, indicating significant deviation from baseline conditions in the water. (This method for classification of the deviations in a process is patented under US Patent 6,999,898, assigned to Hach Co., Loveland, Colo.)

The EDS algorithm was tested at three sites: site 1, a military base in the southeastern United States; site 2, a military installation on the east coast; and site 3, a municipal site in a large midwestern city. Data were collected at all three sites for many months. In addition to providing the on-line monitoring function, the EDS at each site provided data that could be used for two parallel analyses: first, ROC curve analysis via a Monte Carlo simulation in a spreadsheet and second, simulation of agent additions to the water at each site via superposition of agent deviation data on the site data, followed by playing the resultant data files

through the EDS algorithm. The results from the Monte Carlo simulation were compared with the simulation that used actual site data over time.

Monitoring sites and their characteristics. The water quality at site 3 (the municipal location) was very consistent, with low variance in the water quality; as a result, data from the site were easy to analyze. This site was included in the study because it represented data that should provide the best EDS performance. Analytical results from these data demonstrated the upper limit of performance that can be expected from the algorithm.

The water quality at site 2 was more variable but still within the range observed by the authors at various sites throughout the United States. Because the site was representative of water quality seen at dozens of different locations, it was classified as typical. The water quality at site 1 was significantly more variable than what the authors' experience had shown to be typical, with extremely erratic pH and chlorine levels. In addition, at this site, there was some blending of waters from two sources. These changing conditions made for a more challenging analysis. Analytical results from these data demonstrated lower performance, as expected.

Analyzing data from sites that could be characterized as easy, medium, and difficult provided a more complete picture of the performance range delivered by the EDS algorithm. Analyzing only easy or difficult data would not have presented a broad picture of system capability.

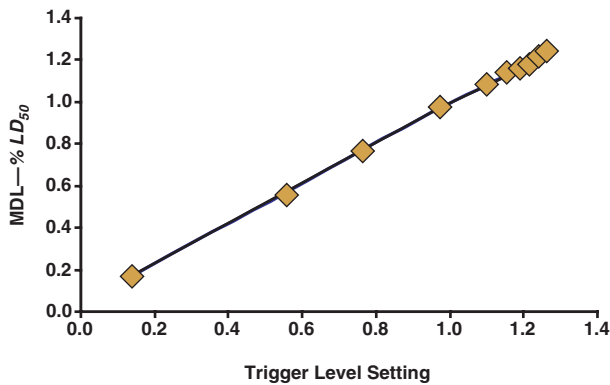
Site data statistics. The ROC curve analysis used in this study requires statistical analysis of the noise content of the parameter signals. Months of data were available from

TABLE 1 FAR and HR for various trigger levels

Trigger Level	FAR	HR at x Percentage of LD ₅₀		
		25%	5%	1%
0	100	100	100	100
0.2	6.727948	100	100	100
0.3	1.745115	100	100	70
0.4	0.452653	100	100	0
0.5	0.11741	100	100	0
0.55	0.059797	100	92	0
0.6	0.030454	100	80	0
0.7	0.007899	100	8	0
0.8	0.002049	100	0	0
1	0.000138	100	0	0
1.2	9.27E-06	76	0	0
1.3	2.41E-06	6	0	0
1.4	6.24E-07	0	0	0
1.5	1.62E-07	0	0	0
2	1.9E-10	0	0	0
3	2.62E-16	0	0	0

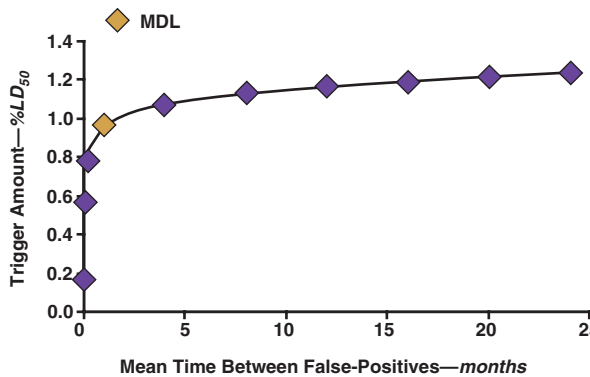
FAR—false alarm rate, HR—hit rate, LD₅₀—median lethal dose

FIGURE 6 MDL % LD₅₀ sodium fluoroacetate versus trigger threshold



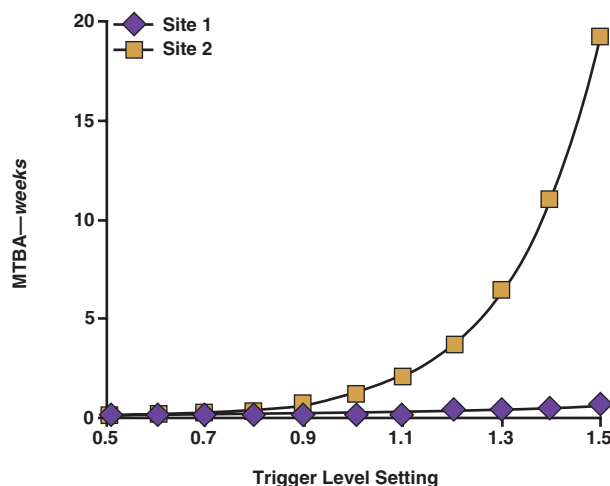
LD₅₀—median lethal dose, MDL—method detection limit

FIGURE 7 ROC curve for sodium fluoroacetate in 1 µg/L fluoroacetate



LD₅₀—median lethal dose, MDL—minimum detection limit
ROC—receiver operating characteristic

FIGURE 8 MTBA versus threshold for sites 1 and 2



MTBA—mean time between alarms

each site; from these data, nine continuous days of representative data were selected for each site. It was important to select data sets that did not contain any abnormal occurrences because these data sets were used to derive the ROC curves. The background data set for ROC curve derivation must not contain extremely improbable or rare maintenance or calibration occurrences; if it does, the statistics will be distorted, resulting in an incorrect presentation. Normal variation and common occurrences such as pumps turning on and off are acceptable.

The parameter signals for all the sensors at a site were statistically characterized over the selected data set, and those statistical values were input to the ROC curve analysis spreadsheet. The parameter values included not only the variation in the water but also any noise or errors contributed by the sensors in normal operation.

FAR analysis of the trigger signals. A statistical characterization of the trigger signal values for the individual baseline data sets was needed for ROC curve analysis. The data collected from the sites provided the trigger signal values associated with the raw parameter data.

The trigger signal data were used to produce curves of FAR versus threshold setting. This is the alarm rate attributed to process and instrument noise with no agents or abnormal occurrences present. For the nonclassical ROC method used in this study, the alarm rate was converted to a more useful measure: mean time between alarms (MTBA). For example, if there were six alarms in twelve months, the MTBA would be two months per alarm.

Figure 8 shows MTBA curves for sites 1 and 2. Site 1 shows a much lower MTBA than site 2 because of site 1's higher parameter noise content and more variable parameter signals.

Monte Carlo analysis. A spreadsheet program was used to perform a Monte Carlo analysis to generate ROC curves. This method of analysis requires less time to perform than simulation runs with site data. It also permits "what if" analyses such as exploring the benefits of modifying operating procedures (e.g., chlorine booster) to decrease the variability for parameters such as pH and chlorine.

In the Monte Carlo analysis, the deviations of the process parameters caused by an agent addition are combined with random noise from the process and sensors. In the analysis, the amount of added agent can be varied. The resultant trigger signal is calculated and compared with the selected threshold value to determine whether the system would be in alarm.

The current study analyzed addition of three plausible threat agents: cyanide and the pesticides oxamyl and aldicarb. Dose concentrations in the study were expressed as a percentage of LD₅₀, the dose that would kill 50% of a typical population (with the actual dose based on that for a 70-kg individual). The LD₅₀ values in mg/kg/L were 6.4 mg/kg/L for cyanide, 5.4 mg/kg/L for oxamyl, and 1 mg/kg/L for aldicarb. To simulate a statistically large sample, 10,000 samples of random noise according to

site statistics were added to the parameter deviations, and the trigger signals were calculated.

In a scenario in which the amount of agent added brings the trigger signal close to the threshold value, it may be that because of the noise, the EDS alarm is triggered 4,391 times and is not triggered 5,609 times at the given dose and threshold setting. The agent dose for the simulation could be increased until 10,000 alarm incidents occur. It can then be stated that the probability of detection at that dose is at least 99.99%—the worst case being 10,000 alarm incidents out of 10,001 trials. In this study, detection was considered a failure if the detection rate was not near 100%. Therefore, the dose value that overcomes all of the noise in the system to generate an alarm 100% of the time could be determined.

By performing this analysis at various threshold values, the values for a curve detection concentration versus threshold could be determined. Figure 9 shows detection concentration versus threshold in the case of cyanide at site 1. The ROC curve was plotted parametrically from this curve and the curve of MTBA versus threshold. The ROC curve for cyanide at site 1 is shown in Figure 10.

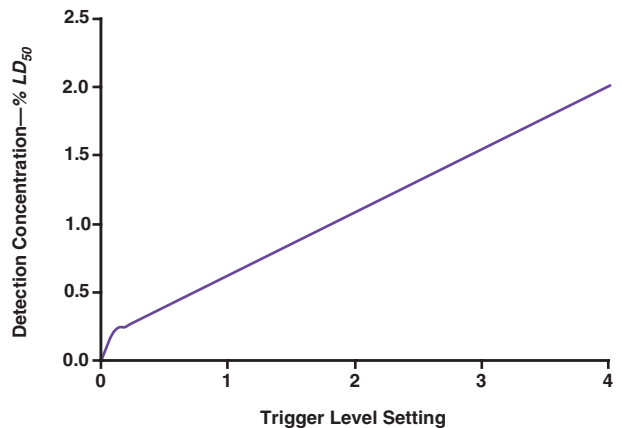
ROC curves of this type are of more practical use than the classical probability-based curves. With this form of the ROC curve, the threshold setting of the EDS can be set according to easily understood parameters: concentration represented as percentage of LD_{50} and desired MTBA. The ROC curve defines the system sensitivity at a given MTBA. Managers can opt for a long MTBA or increased sensitivity. The corresponding minimum detection level (MDL) can be read from the ROC curve. The curve of MDL concentration versus threshold setting then provides the associated threshold setting for the EDS. Figure 11 shows the aldicarb ROC curves for sites 1 and 2. The figure shows that for the same agent, the larger signal variation at site 1 pushes up the MDL, compared with the MDL for the agent at site 2.

The Monte Carlo analysis can also provide the curve for detection rate versus concentration at a constant threshold. Figure 12 shows the curve for aldicarb at sites 1 and 3. The threshold in this case is 1. The difference in slope is an indication of the difference in the noise content of the parameter signals at the two sites. The slope of the curve from the site 1 data is lower because that site has more noise.

Playing a dosed simulation file through the EDS algorithm.

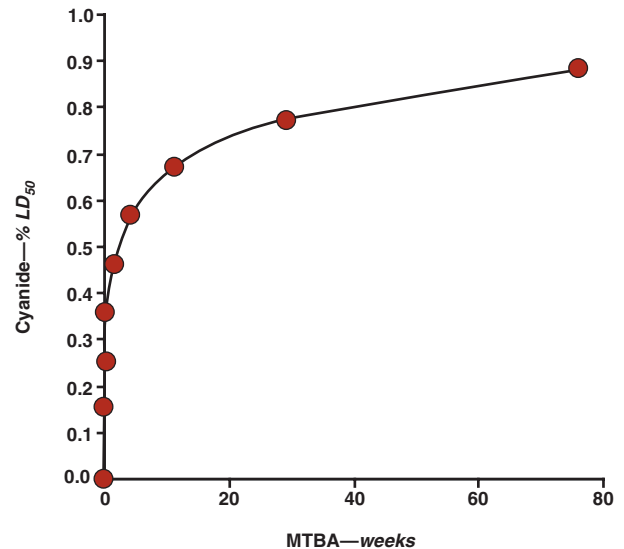
Agent doses can be arithmetically superimposed across the site data files (157 doses per file, 20-min dose duration), and those files can be played through the EDS to assess whether results obtained from the EDS algorithm match those predicted by Monte Carlo analysis. Before playing the files through the EDS algorithm, the EDS can be “tuned” to the noise content of a site’s water and sensors by playing through the raw data set from the site without dosing. In this study, the data were played through in the simulation mode, which runs 60 times

FIGURE 9 Detection concentration versus threshold



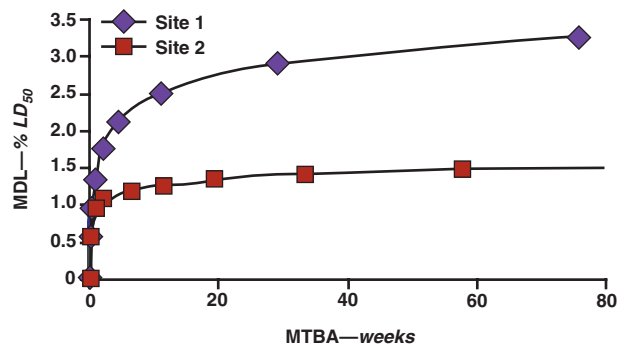
LD_{50} —median lethal dose

FIGURE 10 ROC curve for cyanide at site 1



LD_{50} —median lethal dose, MTBA—mean time between alarms, ROC—receiver operating characteristic

FIGURE 11 ROC curves for aldicarb at sites 1 and 2



LD_{50} —median lethal dose, MDL—minimum detection level, MTBA—mean time between alarms, ROC—receiver operating characteristic

FIGURE 12 Hit rate for alarm curve for aldicarb at sites 1 and 3

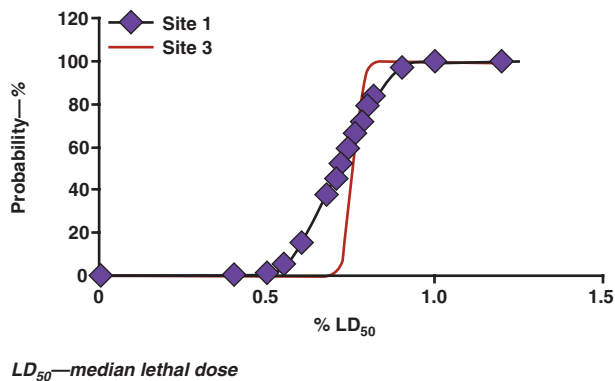
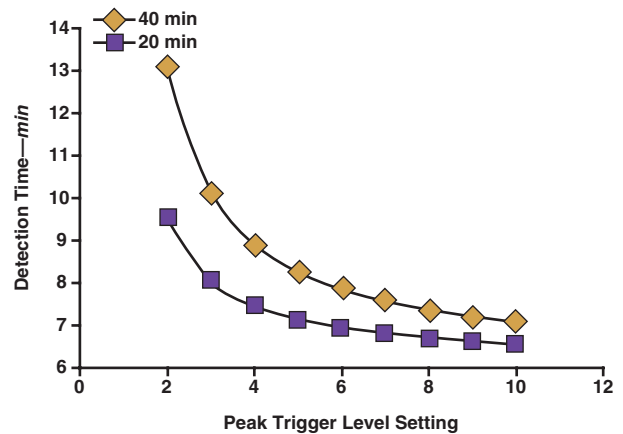


FIGURE 13 Detection time versus injection peak value



faster than real time to speed the process. In actual on-line operation, the EDS analyzes data as they arrive in real time to perform the tuning.

Comparison of analyses. The Monte Carlo analysis can predict a trigger signal peak value for a given agent dose at a given site. The same dose applied to the site data also gives a mean value for the peak trigger signal over multiple injections. Comparison of the two methods indicated that both provided the same expected result. All values for differences were < 1.5%.

Table 2 compares the MDL from the Monte Carlo analysis to the MDL found in the comma-separated value dosing study. Units are percentage of LD₅₀ for the given agent. The Monte Carlo analysis was conservative, predicting slightly higher MDL values than those found in the dosing study.

Detection time analysis. The nature of the EDS detection algorithm and the sensors used for parameter measurement in the current study allowed for calculation of the detection time after contaminated water arrives at the sensor package. Detection time is a function of the response time of the sensor package, the time for the dose curve to reach its maximum value, and the peak trigger value of the dose. Figure 13 shows results for the sensor set used in the studies and two reasonable dose-response times of 20 and 40 min. The detection time is shown as a function of the peak trigger value. For any agent contamination of consequence (trigger signal > 4), the time to detect will be 7–9 min. Detection times longer than 15 min unnecessarily add to response time.

TABLE 2 Comparison of MDLs from the Monte Carlo study and CSV dosing study

Site	Agent	ROC MDL Percentage of LD ₅₀	CSV MDL Percentage of LD ₅₀	Difference
Site 1	Aldicarb	0.974	0.971	0.003
	Cyanide	0.609	0.498	0.111
	Oxamyl	0.258	0.229	0.029
Site 2	Aldicarb	0.972	0.959	0.013
	Cyanide	0.6	0.527	0.073
	Oxamyl	0.257	0.229	0.029
Site 3	Aldicarb	0.81	0.779	0.031
	Cyanide	0.523	0.499	0.024
	Oxamyl	0.219	0.217	0.002

CSV—comma-separated value, EDS—event detection system, LD₅₀—median lethal dose, MDL—method detection limit, ROC—receiver operating characteristic

Water quality at site 3 was the least variable and was included in the study because it represented data that should provide the best EDS performance.

CONCLUSION

Several projects have been initiated to develop and/or evaluate EDSs. Earlier studies focused on the ability of common sensors to detect noticeable changes in water quality when a contaminant was present (Hall et al, 2007; Kroll, 2006; Battelle, 2005; USEPA, 2005). After these early studies verified the efficacy of bulk parameter monitoring, later research focused on the development and testing of event detection algorithms and their use in interpreting the generated data streams (Umberg, 2008; Kroll & King, 2007; McKenna et al, 2007; Yang et al, 2006). Although these studies used some of the criteria for evaluation summarized in this article (including ROC curves), none concentrated on all of the criteria. Even those studies that did take into account such

factors as ROC, FARs, and time to response failed to establish meaningful goals that would indicate that these systems were ready for real-world deployment. And even when they were evaluated, factors such as concentration needed to detect, time to alarm, and detection of the full contaminate range were not held to specifications that would be protective of human health.

The systems deployed in the CERL study were found to perform acceptably under all of the evaluation criteria set forth at the three specific sites evaluated and with the limited number of contaminants tested. They demonstrated dual use by detecting a number of operational and nonsecurity anomalies including dead ends, contamination by aviation fuel, and pumping schedule problems. Results indicated that the EDSs under examination can detect low levels of threat agents in a few minutes at 100% probability of alarm. The systems studied were detect-to-warn-type systems and lived up to that definition in speed (time to respond was as short as 3 min) and cost criteria (compared with grab-sampling protocols). Coverage characteristics were found to be adequate, and communications were determined to be simple and secure.

All operational criteria, including continuous operation and reliability, were verified. The systems were under a service contract with the manufacturer, which performed routine maintenance and calibrations. Operational characteristics of the systems were deemed to meet the criteria. The calculation of ROC curves using the new methodology detailed here for the specific sites indicated that the equipment is capable of detecting likely threats at levels and with a time to alarm activation that would be needed by an effective early warning system. Furthermore, the improved ROC curve method can be used by operators to set threshold alarm levels to minimize unknown alarms while still maintaining the desired level of sensitivity. The current research also demonstrated that Monte Carlo

analysis for the determination of ROC curves and MDLs at a site closely matches results of simulations with agent-superimposed data sets.

Early warning systems for water distribution systems have been shown to be an effective means of enhancing the security of water supplies. In addition, the new ROC curve method proved to be an effective means of validating these systems and should be useful to operators in running and tuning these systems.

ABOUT THE AUTHORS



Dan Kroll is chief scientist for Homeland Security Technologies and principle investigator for the Advanced Technology Group at Hach Homeland Security Technologies, 5600 Lindberg Dr., Loveland, CO 80539; DKROLL@hach.com. In his 20 years working at Hach, he has developed both advanced and simplified methods for a variety of crucial water parameters, and his simplified arsenic-testing method is used throughout the world as the standard field method for screening. He holds bachelor's degrees in genetics and microbiology and a master's degree in water resource management and environmental engineering from Iowa State University in Ames. Karl King is a principal scientist and chief technologist at Hach Homeland Security Technologies.

Date of submission: 02/20/09

Date of acceptance: 09/02/09

JOURNAL AWWA welcomes comments
and feedback at journal@awwa.org.

REFERENCES

- Battelle, 2005. Multi-parameter Water Monitors for Distribution Systems. EPA Environmental Technology Verification Rept. www.battelle.org/PRODUCTS/CONTRACTS/etv/verifications.aspx#W12.
- Hall, J.; Zaffiro, A.D.; Marx, R.B.; Kefauver, P.C.; Krishnan, E.R.; Hought, R.C.; & Herrmann, J.G., 2007. On-Line Water Quality Parameters as Indicators of Distribution System Contamination. *Jour. AWWA*, 99:1:66.
- Kroll, D., 2006. Safeguarding the Distribution System: On-Line Monitoring for Security and Enhancing Operational Performance. *Jour. New England Water Works Assn.*, June.
- Kroll, D. & King, K., 2007. Operational and Laboratory Verification Testing of a Heuristic On-Line Water Monitoring System for Security. *Intl. Jour. High Speed Electronics & Systems*, 17:4:631.
- McKenna, S.A.; Wilson, M.; & Klise, K.A., 2008. Detecting Changes in Water Quality Data. *Jour. AWWA*, 100:1:74.
- McKenna, S.A.; Hart, D.B.; Klise, K.A.; Cruz, V.A.; & Wilson, M.P., 2007. Event Detection From Water Quality Time Series. Proc. ASCE World Envir. & Water Resources Congress, Tampa, Fla.
- McKenna, S.A.; Klise, K.A.; & Wilson, M.P., 2006. Testing Water Quality Change Detection Algorithms. Proc. 8th Ann. Water Distribution Analysis Symp., Cincinnati.
- Umberg, K., 2008. Evaluation of Water Quality Event Detection Systems Deployed at the First Water Security Initiative Pilot Utility. Proc. AWWA Water Security Congress, Cincinnati.
- USEPA (US Environmental Protection Agency), 2005. Technologies and Techniques for Early Warning Systems to Monitor and Evaluate Drinking Water Quality: A State-of-the-Art Review. Research Rept., Ofce. of Research and Development, National Homeland Security Research Center, Cincinnati.
- Yang, J.; Hought, R.C.; Hall, J.; Goodrich, J.A.; & Hasan, J., 2006. Adaptive Monitoring to Enhance Water Sensor Capabilities for Chemical and Biological Contaminant Detection in Drinking Water Systems. Proc. Optics and Photonics in Global Homeland Security II (T. Saito and D. Leffeld, editors). SPIE Publications, Bellingham, Wash.