THE BATTLE OF THE ATTACK DETECTION ALGORITHMS

Riccardo Taormina[1], Stefano Galelli[2], Member, ASCE, Nils Ole Tippenhauer[3], Elad Salomons [4], Avi Ostfeld [5], Fellow, ASCE, Demetrios G. Eliades [6], Mohsen Aghashahi [7], Raanju Sundararajan [8], Mohsen Pourahmadi [9], M. Katherine Banks [10], B. M. Brentan [11], Enrique Campbell [12], G. Lima [13], D. Manzi [14], D. Ayala-Cabrera [15], M. Herrera [16], I. Montalvo [17], J. Izquierdo [18], E. Luvizotto Jr. [19], Sarin E. Chandy [20], Amin Rasekh [21], Zachary A. Barker [22], Bruce Campbell [23], M. Ehsan Shafiee [24], Marcio Giacomoni [25], Nikolaos Gatsis [26], Ahmad Taha [27], Ahmed A. Abokifa [28], S.M., ASCE, Kelsey Haddad [29], Cynthia S. Lo [30], Pratim Biswas [31], M. Fayzul K. Pasha [32], Bijay Kc [33], Saravanakumar Lakshmanan Somasundaram [34], Mashor Housh [35], Ziv Ohar [36]

[1]Singapore University of Technology and Design, 8 Somapah Road, Singapore, 487372.

[2]Singapore University of Technology and Design, 8 Somapah Road, Singapore 487372. E-mail: stefano_galelli@sutd.edu.sg

[3]Singapore University of Technology and Design, 8 Somapah Road, Singapore 487372.

[4]OptiWater, 6 Amikam Israel St., Haifa 3438561, Israel.

[5]Faculty of Civil and Environmental Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel.

[6]KIOS Research and Innovation Center of Excellence, University of Cyprus, 75 Kallipoleos Avenue, CY-1678, Nicosia, Cyprus.

[7]Zachry Dept. of Civil Engineering, Texas A&M Univ., College Station, TX.

[8]Dept. of Statistics, Texas A&M Univ., College Station, TX.

[9]Dept. of Statistics, Texas A&M Univ., College Station, TX.

[10]Zachry Dept. of Civil Engineering, Texas A&M Univ., College Station, TX.

[11]Universidade Estadual de Campinas, Campinas, Brazil.

[12]Universitat Politècnica de València, Valencia, Spain, Berliner Wasserbetriebe, Berlin, Germany.

[13]Universidade Estadual de Campinas, Campinas, Brazil.

[14]Universidade Estadual de Campinas, Campinas, Brazil.

[15]Irstea, Cestas, France.

[16]Univ. of Bath, Bath, U.K.

[17]Ingeniousware GmbH, Karlsruhe, Germany.

[18]Universitat Politècnica de València, Valencia, Spain.

[19]Universidade Estadual de Campinas, Campinas, Brazil.

[20]Sensus Inc., 8601 Six Forks Rd., Suite 700, Raleigh, NC 27615.

[21]Sensus Inc., 8601 Six Forks Rd., Suite 700, Raleigh, NC 27615.

[22]Sensus Inc., 8601 Six Forks Rd., Suite 700, Raleigh, NC 27615.

[23]Sensus Inc., 8601 Six Forks Rd., Suite 700, Raleigh, NC 27615.

[24]Sensus Inc., 8601 Six Forks Rd., Suite 700, Raleigh, NC 27615.

[25]Dept. of Civil and Environmental Engineering, Univ. of Texas at San Antonio, San Antonio, TX 78249.

[26]Dept. of Electrical and Computer Engineering, Univ. of Texas at San Antonio, San Antonio, TX 78249.

[27]Dept. of Electrical and Computer Engineering, Univ. of Texas at San Antonio, San Antonio, TX 78249.

[28]Dept. of Energy, Environmental, and Chemical Engineering, Washington Univ., St. Louis.

[29]Dept. of Energy, Environmental, and Chemical Engineering, Washington Univ., St. Louis.

[30]Dept. of Energy, Environmental, and Chemical Engineering, Washington Univ., St. Louis.

[31]Dept. of Energy, Environmental, and Chemical Engineering, Washington Univ., St. Louis.

[32]Dept. of Civil and Geomatics Engineering, California State Univ., Fresno, CA 93740.

[33]Dept. of Civil and Geomatics Engineering, California State Univ., Fresno, CA 93740.

[34]Dept. of Civil and Geomatics Engineering, California State Univ., Fresno, CA 93740.

[35]Faculty of Management, Dept. of Natural Resources and Environmental Management, Univ. of Haifa, Haifa, Israel.

ABSTRACT

The BATtle of the Attack Detection ALgorithms (BATADAL) is the most recent competition on planning and management of water networks undertaken within the Water Distribution Systems Analysis Symposium. The goal of the battle was to compare the performance of algorithms for the detection of cyber-physical attacks, whose frequency increased in the past few years along with the adoption of smart water technologies. The design challenge was set for C-Town network, a real-world, medium-sized water distribution system operated through Programmable Logic Controllers and a Supervisory Control And Data Acquisition (SCADA) system. Participants were provided with datasets containing (simulated) SCADA observations, and challenged with the design of an attack detection algorithm. The effectiveness of all submitted algorithms was evaluated in terms of classification performance and time-to-detection. Seven teams participated in the battle and proposed a variety of successful approaches leveraging data analysis, model-based detection mechanisms, and rule checking. Results were presented at the Water Distribution Systems Analysis Symposium (World Environmental & Water Resources Congress), in Sacramento, on May 21-25, 2017. This paper summarizes the BATADAL problem, proposed algorithms, results, and future research directions.

Keywords: Water distribution systems, Cyber-physical attacks, Cyber security, EPANET, Smart water networks, Attack detection

INTRODUCTION

The past decades witnessed the transition of water distribution systems from traditional physical infrastructures to *cyber-physical systems* that combine physical processes with computation and networking: physical assets—such as pipes, pumps, and valves—work in unison with networked devices that monitor and coordinate the operations of the entire system. These devices include Programmable Logic Controllers (PLCs), Supervisory Control And

[36]Faculty of Management, Dept. of Natural Resources and Environmental Management, Univ. of Haifa, Haifa, Israel.

Data Acquisition (SCADA) systems, Remote Terminal Units (RTUs), static and mobile sensor networks, and smart meters (Hill et al. 2014; Gong et al. 2016; Sønderlund et al. 2016). The adoption of such smart water technologies plays a pivotal role in enhancing the reliability and autonomy of water distribution systems, but simultaneously exposes them to cyber-physical attacks (Rasekh et al. 2016)—namely the deliberate exploitation of computer systems aimed at accessing sensitive information or compromising the operations of the underlying physical system. Water (and wastewater) systems represent one of the sixteen critical infrastructure sectors identified by the U.S. Department of Homeland Security (U.S. Department of Homeland Security 2017), according to which the number of reported attacks on water infrastructures has been growing steadily (ICS-CERT 2014; ICS-CERT 2015; ICS-CERT 2016)—making them the third highest targeted sector after critical manufacturing and energy (ICS-CERT 2016).

Protecting water distribution systems from cyber attacks requires (as with other cyber-physical systems) a combination of proactive and reactive mechanisms (Cardenas et al. 2008). Proactive mechanisms comprise all tools that reduce the 'attack surface' available to hackers, such as appropriate measures for traffic authentication and confidentiality protection, access control, and device hardening (Graham et al. 2016). Since it is not possible to rule out all attacks, cyber-physical systems should also be equipped with intrusion detection schemes that assist with the recovery phase (Anderson 2010). Disclosing cyber attacks—without issuing false alarms—is thus crucial. Unfortunately, this does not come without some system-specific challenges. First, the definition of anomalous behaviours should not only be related to 'outliers'—i.e., data points lying beyond some specific thresholds—since cyber-physical attacks can tamper one or multiple network components while keeping the performance characteristics within the historical bounds (Abokifa et al. 2017). This implies that detection schemes should be capable of disclosing both outliers and contextual anomalies—i.e., data points that do not conform with normal operating conditions. Second, the same hydraulic response of a water network (e.g., low water levels in a tank) can be obtained through different attacks (Taormina et al. 2017). Therefore, detection schemes should also identify the cyber components that have been attacked; a non-negligible challenge in large water networks. Third, all networked devices, including SCADA systems, represent potential targets. This means that the information provided by SCADA systems may not be fully reliable.

As the field of intrusion detection continues to grow, so too does the need of an objective comparison of attack detection algorithms for water distribution systems. The BATtle of the Attack Detection ALgorithms (BATADAL) was oragnized for this purpose. Participants were provided with datasets containing (simulated) SCADA data for a water distribution system victim of cyber attacks, and were tasked with the design of an online attack detection mechanism. The design goals of a detection algorithm were to: (1) disclose the presence of an ongoing attack in the minimum time possible, (2) avoid issuing false alarms, and (3) identify which components of the system have been compromised (optional). Seven teams, from both academia and industry, contributed with novel solutions, which were evaluated using specific evaluation criteria—i.e., time-to-detection and accuracy. BATADAL results were presented at a special session of the Water Distribution Systems Analysis Symposium (World Environmental & Water resources Congress), in Sacramento, on May 21-25, 2017.

This paper summarizes the main solutions and outcomes of the BATADAL, and proposes future research directions for event detection in the realm of cyber-physical security. The remainder of the paper describes: (1) the BATADAL problem, data, and evaluation criteria; (2) a synopsis of the proposed attack detection algorithms; (3) an analysis of the results; and (4) conclusions and future research directions.

PROBLEM DESCRIPTION

The operators of C-Town water distribution system have observed anomalous behaviors in some hydraulic components, e.g., tank overflows, reduction in pump speed, anomalous activation/deactivation of pumps. They suspect that the anomalies are attributable to cyber-physical attacks that interfered with the system operations and tampered with the readings recorded by the SCADA system. The participants' aim was to develop an attack detection mechanism that detects the presence of attacks—in the shortest amount of time—from the available SCADA data. In particular, attack detection algorithms must classify the system state as either 'safe' or 'under attack'. A summary description of C-Town is provided below, along with the development data and evaluation criteria. BATADAL rules, problem details, and data are available in the supplemental material of the paper.

C-Town Network

C-Town water distribution system is based on a real-world, medium-sized network, first introduced for the *Battle of the Water Calibration Network* (Ostfeld et al. 2011). The network consists of 429 pipes, 388 junctions, 7 storage tanks, 11 pumps (distributed across 5 pumping stations), 5 valves, and a single reservoir (see Figure 1). Water consumption is fairly regular throughout the year. These physical assets were augmented with a network of nine Programmable Logic Controllers (PLCs), which are located in proximity of pumps, storage tanks, and valves. As shown in Table 1, most of the PLCs controlling the pumps receive the information needed by the control logic from other PLCs—for instance, PLC1 controls pump PU1 and PU2 on the basis of tank T1 water level, which is monitored by PLC2. PLCs controlling pumps and valves record information on the device status (ON/OFF or OPEN/CLOSED), the flow passing through it, and the suction and discharge pressures. The cyber network includes a SCADA system, whose role is to coordinate the operations and store the readings provided by the PLCs. All information regarding the distribution system were incorporated into the EPANET2 (Rossman 2000) input file *C-Town.inp*.

Development data

Participants were provided with three datasets containing SCADA readings for 43 system variables, i.e., tank water levels (7 variables), inlet and outlet pressure for one actuated valve and all pumping stations (12 variables), as well as their flow and status (24 variables). All variables are continuous, with the exception of the valve and pumps' status, represented by binary variables. The datasets were generated via simulation with *epanetCPA*, a Matlab toolbox that allows to design a variety of cyber attacks and simulate, with EPANET2 (version 2.0.12), the hydraulic response of a water distribution network; see Taormina et al. (2017) for further details. The first two datasets, hereafter named *Training dataset 1* and *Training dataset 2*, were provided at the beginning of the competition, while the third one (*Test dataset*) was subsequently used to evaluate and rank the attack detection algorithms.

- *Training dataset 1* was generated with a simulation horizon and hydraulic time step of 365 days and one hour, respectively. A key aspect of the dataset is the absence of cyber attacks, which made it suitable for studying the operations of the water distribution system under normal operating conditions.
- *Training dataset 2* contains seven attacks, spanning over 492 hourly time steps. One attack was entirely revealed to the participants (by appropriately labelling the corresponding time steps), while the remaining attacks were either partially revealed or hidden; see Table 2 for additional details. This corresponds to a post-attack scenario, in which forensics experts carry out an investigation to determine whether, when, and where the water distribution system has been affected.
- *Test dataset* contains seven additional attacks, spanning over 407 hourly time steps (see Table 3). Naturally, no information regarding the attacks was revealed. Participants were required to run the detection algorithms on the *Test dataset* and to submit a detection report containing the following information: number of attacks detected, start and end time of each attack (in *DD-MM-YYYY hh* format), and the label of the attacked device(s) (optional).

The operations of the water system were altered through malicious activation of hydraulic actuators, change of actuator settings, and *deception* attacks—amongst the most common for cyber-physical systems (Cardenas et al. 2009). The latter were aimed at manipulating the information sent or received by sensors and PLCs, with the ultimate goal of affecting the operations of an actuator (Urbina et al. 2016). Note that deception attacks were also used to alter the information received by SCADA, therefore concealing the real, physical outcomes of the attacks. SCADA concealment was performed by either replacing actual traffic information between PLCs and SCADA with previously-recorded data (*replay attacks*) or adding an offset to the transmitted sensor readings (Urbina et al. 2016). Figure 2 illustrates attack #3 (Training dataset 2), where both pump operations and SCADA data are compromised. In this case, a deception attack manipulates Tank T1 water level readings sent by PLC2 to PLC1, resulting in an excessive use of pumps PU1 and PU2. This causes Tank T1 to overflow. A second deception attack alters the signal sent by PLC2 to SCADA by adding a time-varying offset.

Evaluation criteria

The evaluation of the attack detection algorithms was based on two scores that account for (1) the time taken to detect an attack, and (2) the algorithm classification performance. The two scores were eventually combined into an overall ranking score, as explained next.

Time-to-detection

The time-to-detection (TTD) is the time needed by an algorithm to disclose a threat. It is defined as the difference between the time t_d at which the attack is detected and the time t_0 at which the attack started:

$$TTD = t_d - t_0. \tag{1}$$

The lower the value of TTD, the better the algorithm performs. If an attack is detected, we then have:

$$0 \leq TTD \leq \Delta t, \tag{2}$$

where Δt is the total duration of the attack. If the attack is not detected while it is ongoing (or at all), we set $TTD = \Delta t$. To facilitate the comparison of all algorithms under different attack scenarios, the following performance score (S_{TTD}) was computed:

$$S_{TTD} = 1 - \frac{1}{n_a} \sum_{i}^{n_a} \frac{TTD_i}{\Delta t_i}, \tag{3}$$

where n_a is the number of attacks contained in a dataset, TTD_i the time-to-detection relative to the i-th attack, and Δt_i the corresponding duration. S_{TTD} varies between 0 and 1, with $S_{TTD} = 1$ being the ideal case in which all attacks are immediately detected, and $S_{TTD} = 0$ the case in which none of the attacks is detected.

Classification performance

We determined the accuracy of an algorithm as its ability to disclose threats without raising false alarms. In the context of binary classification problems—like BATADAL—the ability to identify threats is generally assessed with the *True Positive Rate* (TPR, also known as *recall* or *sensitivity*), which is defined as:

$$TPR = \frac{TP}{TP + FN}, \tag{4}$$

where TP and FN represent the number of True Positives and False Negatives, respectively. In other words, the True Positive Rate is the ratio between the number of time steps correctly classified as under attack and the total number of time steps during which the system is under attack.

The ability to avoid false alarms is measured with the *True Negative Rate* (TNR, or *specificity*), defined as

$$TNR = \frac{TN}{FP + TN}, \tag{5}$$

where FP and TN represent the number of False Positives and True Negatives, respectively. The True Negative Rate is thus the ratio between the number of time steps correctly classified as safe conditions and the total number of time steps during which the system is in safe conditions.

To ease the comparison across all algorithms, the True Positive and True Negative Rate were combined into a single classification performance score (S_{CLF}), defined as the mean between TPR and TNR, namely:

$$S_{CLF} = \frac{TPR + TNR}{2}. \tag{6}$$

This score, also known as *area under the curve* (Powers 2011), accounts for both correct detection and false alarms, so it is suited for binary classification problems in which the sample distribution is biased towards one of the two classes—i.e., safe conditions, in BATADAL. The value of S_{CLF} varies between 0 and 1, with 1 representing a perfect classification.

Ranking score

The time-to-detection and accuracy scores were finally merged into an overall ranking score (S), defined as:

$$S = \gamma \cdot S_{TTD} + (1 - \gamma) \cdot S_{CLF}, \tag{7}$$

where γ ($0 \leq \gamma \leq 1$) determines the relative importance of the two evaluation scores. The coefficient γ was set to 0.5 for the analysis reported below; so, early detection and accurate classification were equally weighed. Note that a naïve detection mechanism that predicts the system to be always in safe conditions gets a score S equal to 0.25 ($S_{TTD} = 0$, $S_{CLF} = 0.5$). On the other hand, flagging the system as always under attack yields a value of S equal to 0.75 ($S_{TTD} = 1$, $S_{CLF} = 0.5$). This reflects the fact that S is intrinsically biased towards attack identification, since the the consequences of failing to disclose an attack are deemed more costly than issuing false alarms.

ATTACK DETECTION ALGORITHMS

Seven teams participated in BATADAL. Here, we provide a brief description of each team's attack detection algorithm.

- Aghashahi et al. (2017) adopted a two-step approach. First, a spectral domain method was used to extract the important characteristics of the observed data and make them independent of time; then, a supervised machine learning technique (i.e., Random Forests, Breiman (2001)) was used to classify the system state as safe or under attack.
- Brentan et al. (2017) reduced the dimensionality of the problem by exploiting the division of C-Town network in District Metered Areas (DMAs). For each DMA, the authors used data on normal operating conditions to create Recurrent Neural Networks that forecast tank water levels as a function of pump flow, upstream pressure (of the corresponding pump station), and hour of the day (Díaz et al. 2016). A statistical control process was finally used to identify abrupt changes in the neural networks error time series when the latter were applied to data containing cyber attacks (Guralnik and Srivastava 1999). The rationale behind this approach is that it is plausible to expect an increase in the error time series when the system is under attack, since all neural networks are trained with data pertaining to normal operations.
- Chandy et al. (2017) developed two detection models running sequentially. The first one uses features of the SCADA data (e.g., combined flow of pump stations, volume pumped and stored) to check whether physical and/or operating rules have been violated (e.g., tank levels within the bounds, hydraulic relationships between nodes hold). The outcome of this model is a set of flagged events, which are confirmed by the second model. The latter is a Convolutional Variational Auto-Encoder—belonging to the family of deep learning methods (Kingma and Welling 2013; Doersch 2016)—that calculates the reconstruction probability of the data: the lower the probability, the higher the chance of the data being anomalous.
- Giacomoni et al. (2017) proposed two detection methods. The first one verifies the integrity of the actuator rules and SCADA data—by (1) checking whether the SCADA readings are consistent with the actuator rules defined for the water distri-

bution system, and (2) comparing the data for all variables to identify values falling below or above thresholds created by analyzing data corresponding to normal operating conditions. The second method builds on a convex optimization routine, which unveils low-dimensionality components in the available data as well as the sparse nature of anomalies, thereby facilitating the separation of anomalies from the overall data (Mardani et al. 2013). (The results reported below for Giacomoni et al. (2017) correspond to the first detection method.)

- Abokifa et al. (2017) introduced a three-stages detection method, with each stage targeting a specific class of anomalies. The first step features outlier detection techniques to find statistical outliers in the data, thereby focusing on local anomalies that affect each sensor individually. The second stage employs an Artificial Neural Network—in the form of a Multi-Layer Perceptron—to detect contextual anomalies that do not conform to normal operating conditions. The third stage targets global anomalies that simultaneously affect multiple sensors. To disclose these anomalies, the layer uses Principal Component Analysis to decompose the high-dimensional datasets of sensor measurements into two sub-spaces representing normal and anomalous conditions (Lee et al. 2013).
- Pasha et al. (2017) presented an algorithm consisting of three main interconnected modules working on control rules and consistency checks, pattern recognition, and hydraulic and system relationships. The first module checks the consistency of the data against the set of control rules characterizing the water system, while the second one uses statistical analysis to identify patterns for single hydraulic parameters and combination thereof. The idea is that patterns under cyber attacks may not follow the original ones. The anomalous behaviors detected by the first two modules are finally confirmed by the third one, which develops relationships for some physical quantities (e.g., tank levels, flows) and compares their estimates against those reported by the first two modules.

- Housh and Ohar (2017b) proposed a model-based approach that employs EPANET to simulate the hydraulic processes of the water distribution systems, and then uses the error between EPANET simulated values and the available SCADA readings to detect anomalous behaviors. The approach consists of three main steps: first, available SCADA readings are used in a Mixed-Integer Linear Program to estimate the water demand in all nodes of C-Town; second, EPANET is used to generate two sets of simulated values (i.e., with and without attacks); and third, a multi-level classification approach is implemented to classify the obtained simulation errors into outliers and normal errors. A similar approach was successfully developed by Housh and Ohar (2017a) to detect contamination events in water distribution systems.

RESULTS

Algorithms performance

Table 4 reports the values of the ranking, time-to-detection, and classification score (S, S_{TTD}, and S_{CLF}) obtained by the competing algorithms on the test dataset. The table also reports the number of attacks detected and the elements of the confusion matrix yielding the classification score (i.e., TP, FP, TN, and FN). A visual comparison of S, S_{TTD}, and S_{CLF} is given in the scatter plot of Figure 3.

Figure 3 highlights a cluster of four high-performing algorithms, all achieving a ranking score S higher than (or close to) 0.90. The group is led by the algorithm proposed by Housh and Ohar (2017b), which shows the best overall performance ($S = 0.970$). Note that this algorithm is the top scorer in terms of both time-to-detection S_{TTD} and classification score S_{CLF}. Indeed, the detection trajectory depicted in Figure 4(a) shows that all attacks were immediately detected, with the exception of the last one, which was disclosed a few hours after its starting time. The algorithm of Abokifa et al. (2017) comes a close second, with S equal to 0.949. This method was almost as quick as Housh and Ohar (2017b) one in identifying the attacks, but it was more prone to false alarms. As shown in Figure 4(b), Abokifa et al. (2017) algorithm disclosed Attack #10 and #11 as a single continuous episode, erroneously flagging the system as under attack for the period in between. The algorithm proposed by Giacomoni et al. (2017) has the same number of false positives and true negatives as that of Housh and Ohar (2017b)—meaning that both algorithms were the most successful in avoiding false alarms. However, Giacomoni et al. (2017) algorithm is less sensitive, resulting in a higher number of false negatives and minor timing errors (see Figure 4(c)) that lead to a score S equal to 0.927. With a value of S equal to 0.896, the algorithm proposed by Brentan et al. (2017) can also be regarded as a strong performer. As shown in Figure 4(d), this algorithm was able to consistently and accurately detect most of the attacks, but it failed to identify the last one.

Although outdistanced by the leading group, the contributions of Chandy et al. (2017) and Pasha et al. (2017) are still sensibly better than the naïve detection mechanisms described in Section 2. Their score S is equal to 0.802 and 0.773, respectively. As illustrated in Figure 4(e,f), these two detection algorithms appear to suffer from opposite problems. The algorithm of Chandy et al. (2017) turned out to be over-sensitive—meaning that it was able to identify most of the attack instances, but at the cost of issuing numerous false alarms. On the other hand, the algorithm of Pasha et al. (2017) issued just a few false alarms, but it lacked sensitivity, thus failing to flag the system as under attack for the entire duration of the events. Finally, the contribution of Aghashahi et al. (2017) detected only three attacks, leading to a score S equal to 0.534.

General Observations

The main insights from the results presented above can be summarized as follows:

- All algorithms but one achieved a ranking score S larger than 0.75, meaning that they performed better than naïve detection mechanisms. Yet, we observed a large variability in the algorithm performance.

- Both time-to-detection and classification score are important aspects of performance. Logically, the algorithms that performed consistently well for both metrics achieved a higher ranking score. With the exception of Brentan et al. (2017) and Pasha et al. (2017), there appears to be a strong correlation between these two metrics (see Figure 3).
- Only a few algorithms provided information on the attacked devices. Among these, the algorithms proposed by Brentan et al. (2017) and Giacomoni et al. (2017) were the most accurate.
- Interestingly, the BATADAL was won by the only model-based approach. The idea of estimating the water demands to simulate system dynamics with EPANET, and then measure the errors with respect to SCADA readings, proved successful. In this regard, it is important to note that BATADAL demand patterns were fairly regular and consistent across the three datasets. Similarly, the participants were given the same computational model of the C-Town network that was used to generate the SCADA data. Therefore, successful application of this approach in real-world settings might be hindered by the intrinsic variability of demand patterns or the unavailability of a reliable system model.
- We can probably conclude that both model-based and data-driven approaches are suitable for attack detection problems, although their performance would probably vary with the modelling context at hand.
- Detection algorithms adopting a 'multivariate' approach may be best suited than algorithms analyzing a single time series per time. The inherent interdependence of the elements in the water network should theoretically allow for the detection of anomalies, even when the adversary tries to conceal his (her) actions by altering the SCADA readings of one or a few deployed sensors.
- Most teams presented multi-stage detection methods. Comparing and confirming the

detection issued by different modules can help decrease classification errors.

- The adoption of supervised classification algorithms that learn how to classify the system state (as either safe or under attack) may not be ideal, since the number of attacks in the available data is generally limited. Supervised classification algorithms should always be combined with cross-validation schemes.
- It appears that consistency checks and the analysis of control rules should lead to the identification of the simplest attacks.

FUTURE RESEARCH DIRECTIONS

The BATADAL highlighted the following gaps that may need additional research efforts:

- *Robustness analysis.* The evaluation of BATADAL algorithms can be seen as a deterministic analysis carried out on three specific datasets, which represent only a small portion of the entire set of cyber-attacks that could threaten a water distribution system. Hence, the generation of different attacks is likely to produce different results; a limitation observed in other battles (e.g., Ostfeld et al. (2008)). To evaluate the robustness of an algorithm, it is thus advisable to generate stochastic simulation scenarios comprising varying hydraulic conditions (i.e., water demand, initial tank levels) and multiple attack sequences.
- *Use of real SCADA data.* A major limitation of the current research on cyber-security is the absence of detailed information on cyber attacks to water utilities (e.g., timing, compromised devices, hydraulic response of the system). Access to such information and to the corresponding SCADA data—perhaps, in some anonymized forms—would drastically enhance our understanding on skills and limitations of detection algorithms. Another challenge with SCADA data is that they often contain noise and measurement errors, so attack detection algorithms should be coupled with data preprocessing techniques.
- *Pressure deficient conditions and water quality problems.* A limitation of this battle

is its reliance of data generated with a demand-driven engine (Taormina et al. 2017). The range of attacks should be thus extended to include pressure-deficient conditions, water quality problems, and adversial attempts aimed at threatening emergency responses, such as firefighting operations. In the absence of real SCADA data, simulated data could be generated by combining *epanetCPA* with more sophisticated hydraulic engines (e.g., Sayyed et al. (2015)) or water quality models (e.g., EPANET-MSX, Shang et al. (2007)).

- *Sensitivity analysis.* The definition of the cut-off criteria defining outliers regulates the trade-off between True Positive and True Negative Rate for most of the algorithms, so there is a need to adopt or develop sensitivity analysis tools that draw the appropriate line between normal and anomalous data (Abokifa et al. 2017). This step should always precede the application of an algorithm to new datasets—or its deployment in a SCADA system.
- *Computational requirements and scalability to large networks.* The algorithms presented in this paper were applied to a medium-sized water distribution system comprising one SCADA system and nine PLCs. Since attack detection algorithm are meant to run in real-time, it is necessary to evaluate their computational requirements as well as their scalability to larger networks.
- *Attack localization.* To facilitate and hasten incident resolution, an ideal detection mechanism should be able to identify which components of the network are being attacked. This is a rather challenging task due to the intrinsic correlation among the hydraulic variables.
- *Integration with other fault detection mechanisms.* Since attack detection mechanisms aim to disclose outliers and contextual anomalies in the system behavior, they may accidentally disclose anomalous behaviors that are not necessarily caused by cyber attacks. Hence, there is a need to disclose the nature of each problem being identified—for example, by combining the attack detection algorithms with fault

detection mechanisms that monitor PLCs operations.

- *Cost effectiveness of attack detection.* In BATADAL, the different algorithms were evaluated based on their responsiveness and classification performance. Although these metrics provide some insights on the potential benefits of deploying an attack detection mechanism, a more comprehensive evaluation is needed. For example, one could try to estimate the cost associated to each cyber-physical attack and the corresponding cost savings guaranteed by a detection algorithm.

CLOSURE

The BATADAL is the first *battle competition* dealing with the emerging topic of cyber-physical security of water distribution systems. This battle gave an opportunity to develop, test, and compare attack detection algorithms for SCADA data. The solutions provided by seven teams suggest that timely and accurate detection can be obtained by both model-based and data-driven approaches, usually made of multiple sequential stages. While the data and algorithms presented here provide a first step towards an objective comparison of attack detection algorithms for water distribution systems, they do not represent the entire spectrum of modelling contexts that practitioners and researchers would encounter. Hence, we hope that the availability of a dedicated website (`www.batadal.net`) will help share more datasets and case studies.

SUPPLEMENTAL DATA

The supplemental data include the following files, which are available online in the ASCE Library (`www.ascelibrary.org`):

- *BATADAL rules.pdf*—competition rules, available to participants;
- *C-Town.inp*—EPANET input file, version 2.00.12, available to participants;
- *Training dataset 1.dat*—data without attacks, available to participants;
- *Training dataset 2.dat*—data with attacks and corresponding labels, available to the participants with partial labels;

- *Test dataset.dat*—data with attacks and corresponding labels, available to the participants without labels;
- *Detection reports.dat*—detection reports submitted by the participants.

Additional details about BATADAL are available at `http://batadal.net`.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation (NRF), Singapore, under its National Cybersecurity R&D Programme (Award No. NRF2014NCR-NCR001-40).

REFERENCES

Abokifa, A. A., Haddad, K., Lo, C. S., and Biswas, P. (2017). "Detection of cyber physical attacks on water distribution systems via principal component analysis and artificial neural networks." *World Environmental and Water Resources Congress 2017*, 676–691, <http://ascelibrary.org/doi/abs/10.1061/9780784480625.063>.

Aghashahi, M., Sundararajan, R., Pourahmadi, M., and Banks, M. K. (2017). "Water distribution systems analysis symposium; battle of the attack detection algorithms (BATADAL)." *World Environmental and Water Resources Congress 2017*, 101–108, <http://ascelibrary.org/doi/abs/10.1061/9780784480595.010>.

Anderson, R. J. (2010). *Security engineering: a guide to building dependable distributed systems*. John Wiley & Sons.

Breiman, L. (2001). "Random forests." *Machine Learning*, 45(1), 5–32.

Brentan, B. M., Campbell, E., Lima, G., Manzi, D., Ayala-Cabrera, D., Herrera, M., Montalvo, I., Izquierdo, J., and Luvizotto, E. (2017). "On-line cyber attack detection in water networks through state forecasting and control by pattern recognition." *World Environmental and Water Resources Congress 2017*, 583–592, <http://ascelibrary.org/doi/abs/10.1061/9780784480625.054>.

Cardenas, A., Amin, S., Sinopoli, B., Giani, A., Perrig, A., and Sastry, S. (2009). "Challenges for securing cyber physical systems." *Proceedings of Workshop on future directions in cyber-physical systems security*, Vol. 5.

Cardenas, A. A., Amin, S., and Sastry, S. (2008). "Secure control: Towards survivable cyber-physical systems." *Proceedings of Conference on Distributed Computing Systems Workshops (ICDCS)*, IEEE, 495–500.

Chandy, S. E., Rasekh, A., Barker, Z. A., Campbell, B., and Shafiee, M. E. (2017). "Detection of cyber-attacks to water systems through machine-learning-based anomaly detection in scada data." *World Environmental and Water Resources Congress 2017*, 611–616, <http://ascelibrary.org/doi/abs/10.1061/9780784480625.057>.

Díaz, S., González, J., and Mínguez, R. (2016). "Uncertainty evaluation for constrained state estimation in water distribution systems." *Journal of Water Resources Planning and Management*, 142(12), 06016004.

Doersch, C. (2016). "Tutorial on variational autoencoders." *arXiv preprint arXiv:1606.05908*.

Giacomoni, M., Gatsis, N., and Taha, A. (2017). "Identification of cyber attacks on water distribution systems by unveiling low-dimensionality in the sensory data." *World Environmental and Water Resources Congress 2017*, 660–675, <http://ascelibrary.org/doi/abs/10.1061/9780784480625.062>.

Gong, W., Suresh, M. A., Smith, L., Ostfeld, A., Stoleru, R., Rasekh, A., and Banks, M. K. (2016). "Mobile sensor networks for optimal leak and backflow detection and localization in municipal water networks." *Environmental Modelling & Software*, 80, 306–321.

Graham, J., Olson, R., and Howard, R. (2016). *Cyber security essentials*. CRC Press.

Guralnik, V. and Srivastava, J. (1999). "Event detection from time series data." *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 33–42.

Hill, D., Kerkez, B., Rasekh, A., Ostfeld, A., Minsker, B., and Banks, M. K. (2014). "Sensing and cyberinfrastructure for smarter water management: the promise and challenge of ubiquity." *Journal of Water Resources Planning and Management*, 140(7), 01814002.

Housh, M. and Ohar, Z. (2017a). "Integrating physically based simulators with event detection systems: Multi-site detection approach." *Water Research*, 110, 180–191.

Housh, M. and Ohar, Z. (2017b). "Model based approach for cyber-physical attacks detection in water distribution systems." *World Environmental and Water Resources Congress 2017*, 727–736, <http://ascelibrary.org/doi/abs/10.1061/9780784480625.067>.

ICS-CERT (2014). "NCCIC/ICS-CERT year in review: FY 2013." *Report No. 13-50369*, U.S. Department of Homeland Security – Industrial Control Systems-Cyber Emergency Response Team, Washington, D.C.

ICS-CERT (2015). "NCCIC/ICS-CERT year in review: FY 2014." *Report No. 14-50426*, U.S. Department of Homeland Security – Industrial Control Systems-Cyber Emergency Response Team, Washington, D.C.

ICS-CERT (2016). "NCCIC/ICS-CERT year in review: FY 2015." *Report No. 15-50569*, U.S. Department of Homeland Security – Industrial Control Systems-Cyber Emergency Response Team, Washington, D.C.

Kingma, D. P. and Welling, M. (2013). "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114*.

Lee, Y.-J., Yeh, Y.-R., and Wang, Y.-C. F. (2013). "Anomaly detection via online oversampling principal component analysis." *IEEE Transactions on Knowledge and Data Engineering*, 25(7), 1460–1470.

Mardani, M., Mateos, G., and Giannakis, G. B. (2013). "Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies." *IEEE Transactions on Information Theory*, 59(8), 5186–5205.

Ostfeld, A., Salomons, E., Ormsbee, L., Uber, J. G., Bros, C. M., Kalungi, P., Burd, R., Zazula-Coetzee, B., Belrain, T., Kang, D., et al. (2011). "Battle of the water calibration networks." *Journal of Water Resources Planning and Management*, 138(5), 523–532.

Ostfeld, A., Uber, J. G., Salomons, E., Berry, J. W., Hart, W. E., Phillips, C. A., Watson, J.-P., Dorini, G., Jonkergouw, P., Kapelan, Z., et al. (2008). "The battle of the water sensor networks (bwsn): A design challenge for engineers and algorithms." *Journal of Water Resources Planning and Management*, 134(6), 556–568.

Pasha, M. F. K., Kc, B., and Somasundaram, S. L. (2017). "An approach to detect the cyber-physical attack on water distribution system." *World Environmental and Water Resources Congress 2017*, 703–711, <http://ascelibrary.org/doi/abs/10.1061/9780784480625.065>.

Powers, D. M. (2011). "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation." *Journal of Machine Learning Technologies*, 2(1), 37–63.

Rasekh, A., Hassanzadeh, A., Mulchandani, S., Modi, S., and Banks, M. K. (2016). "Smart water networks and cyber security." *Journal of Water Resources Planning and Manage-*

ment, 142.

Rossman, L. A. (2000). *EPANET 2 Users Manual*. U.S. Environmental Protection Agency, Washington, D.C., EPA/600/R-00/057 edition.

Sayyed, M. A. H. A., Gupta, R., and Tanyimboh, T. T. (2015). "Noniterative application of epanet for pressure dependent modelling of water distribution systems." *Water Resources Management*, 29(9), 3227–3242.

Shang, F., Uber, J. G., and Rossman, L. A. (2007). "Modeling reaction and transport of multiple species in water distribution systems." *Environmental Science & Technology*, 42(3), 808–814.

Sønderlund, A. L., Smith, J. R., Hutton, C. J., Kapelan, Z., and Savic, D. (2016). "Effectiveness of smart meter-based consumption feedback in curbing household water use: Knowns and unknowns." *Journal of Water Resources Planning and Management*, 142(12), 04016060.

Taormina, R., Galelli, S., Tippenhauer, N. O., Salomons, E., and Ostfeld, A. (2017). "Characterizing cyber-physical attacks on water distribution systems." *Journal of Water Resources Planning and Management*, 143(5), 04017009.

Urbina, D., Giraldo, J., Tippenhauer, N. O., and Cárdenas, A. (2016). "Attacking fieldbus communications in ICS: Applications to the SWaT testbed." *Proceedings of Singapore Cyber Security Conference (SG-CRC)* (January).

U.S. Department of Homeland Security (2017). "Critical infrastructure sectors, <https://www.dhs.gov/critical-infrastructure-sectors> (September).

List of Tables

TABLE 1. Sensors and actuators (pumps, valves) monitored/controlled by the PLCs. For each PLC, we also report the corresponding controlling sensor, which provides the information needed to operate the actuators. Note that a PLC-to-PLC connection is established whenever an actuator and the corresponding control sensor are connected to two different PLCs.

PLC	Sensor	Actuators (Controlling sensor)
PLC1	-	PU1(T1), PU2(T1)
PLC2	T1	-
PLC3	T2	V2(T2), PU4(T3), PU5(T3), PU6(T4), PU7(T4)
PLC4	T3	-
PLC5	-	PU8(T5), PU9(-), PU10(T7), PU11(T7)
PLC6	T4	-
PLC7	T5	-
PLC8	T6	-
PLC9	T7	-

TABLE 2. Attacks featured in Training dataset 2.

ID	Starting time [dd/mm/YY HH]	Ending time [dd/mm/YY HH]	Duration [hours]	Attack description	SCADA concealment	Labeled [hours]
1	13/09/2016 23	16/09/2016 00	50	Attacker changes L_T7 thresholds (which controls PU10/PU11) by altering SCADA transmission to PLC9. Low levels in T7.	Replay attack on L_T7 .	42
2	26/09/2016 11	27/09/2016 10	24	Like Attack #1.	Like Attack #1 but replay attack extended to PU10/PU11 flow and status.	0
3	09/10/2016 09	11/10/2016 20	60	Attack alters L_T1 readings sent by PLC2 to PLC1, which reads a constant low level and keeps pumps PU1/PU2 ON. Overflow in T1.	Polyline to offset L_T1 increase.	60
4	29/10/2016 19	02/11/2016 16	94	Like Attack #3.	Replay attack on L_T1, PU1/PU2 flow and status, as well as pressure at pumps outlet.	37
5	26/11/2016 17	29/11/2016 04	60	Working speed of PU7 reduced to 0.9 of nominal speed causes lower water levels in T4.		7
6	06/12/2016 07	10/12/2016 04	94	Like Attack #5, but speed reduced to 0.7.	L_T4 drop concealed with replay attack.	73
7	14/12/2016 15	19/12/2016 04	110	Like Attack #6.	Replay attack on L_T1, as well as PU1/PU2 flow and status.	0

TABLE 3. Attacks featured in the Test dataset.

ID	Starting time [dd/mm/YY HH]	Ending time [dd/mm/YY HH]	Duration [hours]	Attack description	SCADA concealment
8	16/01/2017 09	19/01/2017 06	70	Attacker changes L_T3 thresholds (which control PU4/PU5) by gaining control of PLC3. Low levels in T3.	Replay attack on L_T3, as well as PU4/PU5 flow and status.
9	30/01/2017 08	02/02/2017 00	65	Attack alters L_T2 readings arriving to PLC3, which reads a low level and keeps valve V2 OPEN, leading T2 to overflow.	Polyline to offset L_T2 increase.
10	09/02/2017 03	10/02/2017 09	31	Malicious activation of pump PU3	
11	12/02/2017 01	13/02/2017 07	31	Similar to Attack #10	
12	24/02/2017 05	28/02/2017 08	100	Similar to Attack #9	Replay attack on L_T2, V2 flow and status, as well as V2 inlet + outlet pressure readings (P_J14, P_J422)
13	10/03/2017 14	13/03/2017 21	80	Attacker changes L_T7 thresholds (which control PU10/PU11) by gaining control of PLC5, causing the pumps to switch ON/OFF continuously.	Replay attack on L_T7, PU10/PU11 flow and status, as well as inlet + outlet pressure readings (P_J14, P_J422). Inlet pressure concealment terminates before that of other variables.
14	25/03/2017 20	27/03/2017 01	30	Alteration of T4 signal arriving to PLC6. Overflow in T6.	

TABLE 4. Performance of all attack detection algorithms, assessed in terms of number of attacks detected, overall ranking score (S), time-to-detection (S_{TTD}), accuracy (S_{CLF}), and number of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). The algorithms are ranked according to the their overall ranking score.

Rank	**Team**	**# Attacks detected**	S	S_{TTD}	S_{CLF}	TP	FP	TN	FN
1	Housh and Ohar	7	0.970	0.965	0.975	388	5	1677	19
2	Abokifa et al.	7	0.949	0.958	0.940	375	69	1613	32
3	Giacomoni et al.	7	0.927	0.936	0.917	341	5	1677	66
4	Brentan et al.	6	0.894	0.857	0.931	362	45	1637	45
5	Chandy et al.	7	0.802	0.835	0.768	349	541	1141	58
6	Pasha et al.	7	0.773	0.885	0.660	134	14	1668	273
7	Aghashahi et al.	3	0.534	0.429	0.640	161	195	1487	246

List of Figures

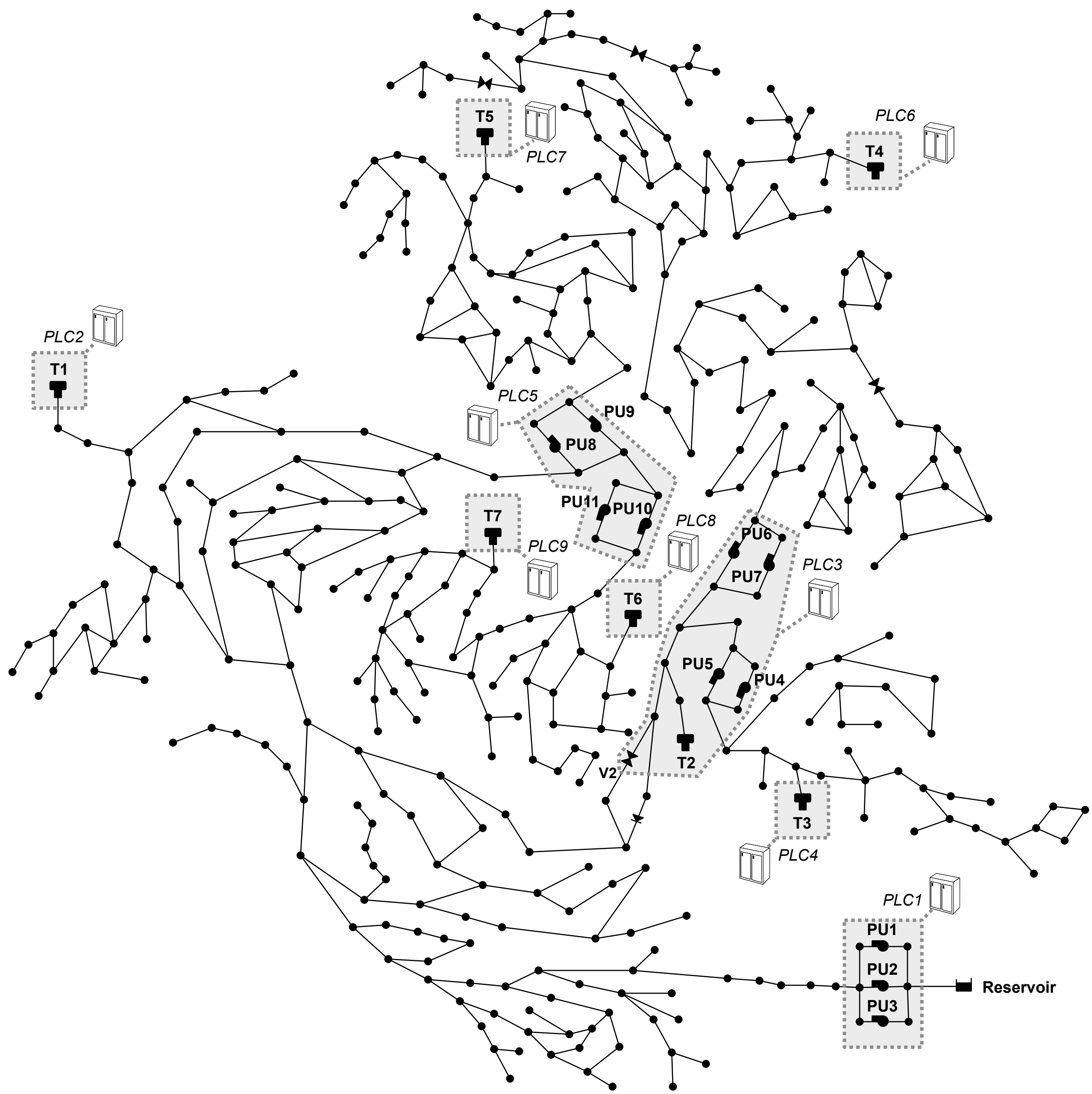

FIG. 1. Graphical representation of C-Town water distribution system (adapted from Taormina et al. 2017).

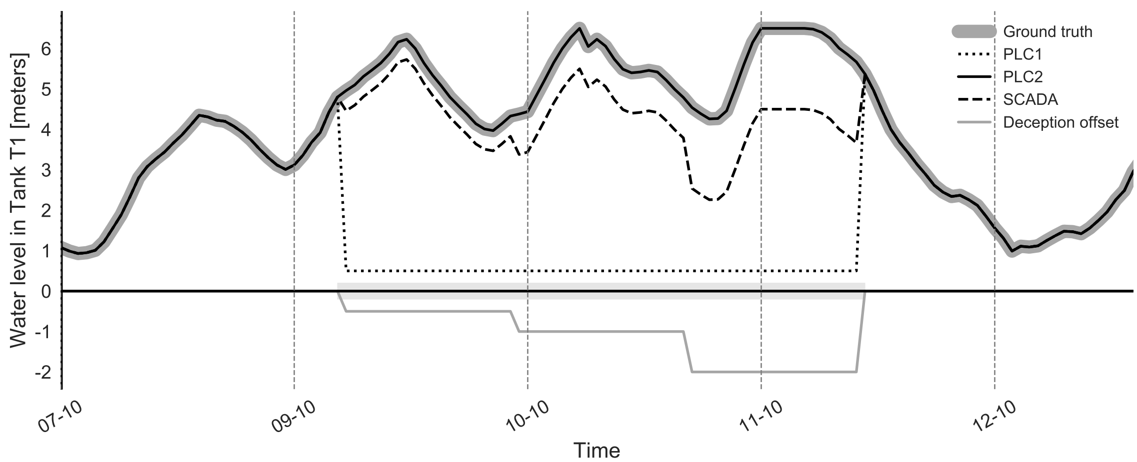

FIG. 2. Illustration of attack #3 (from Training dataset 2). The attacker alters Tank T1 water level readings (continuous black line) sent by PLC2 to PLC1, which reads a constant low level (dotted black line) and keeps Pumps PU1/PU2 ON. This causes an overflow in Tank T1 (thick gray line). To conceal the action, the attacker alters the signal sent by PLC2 to SCADA (dashed black line) by adding a time-varying offset (continuous gray line). The duration of the entire attack is highlighted by the light gray line on the horizontal axis.

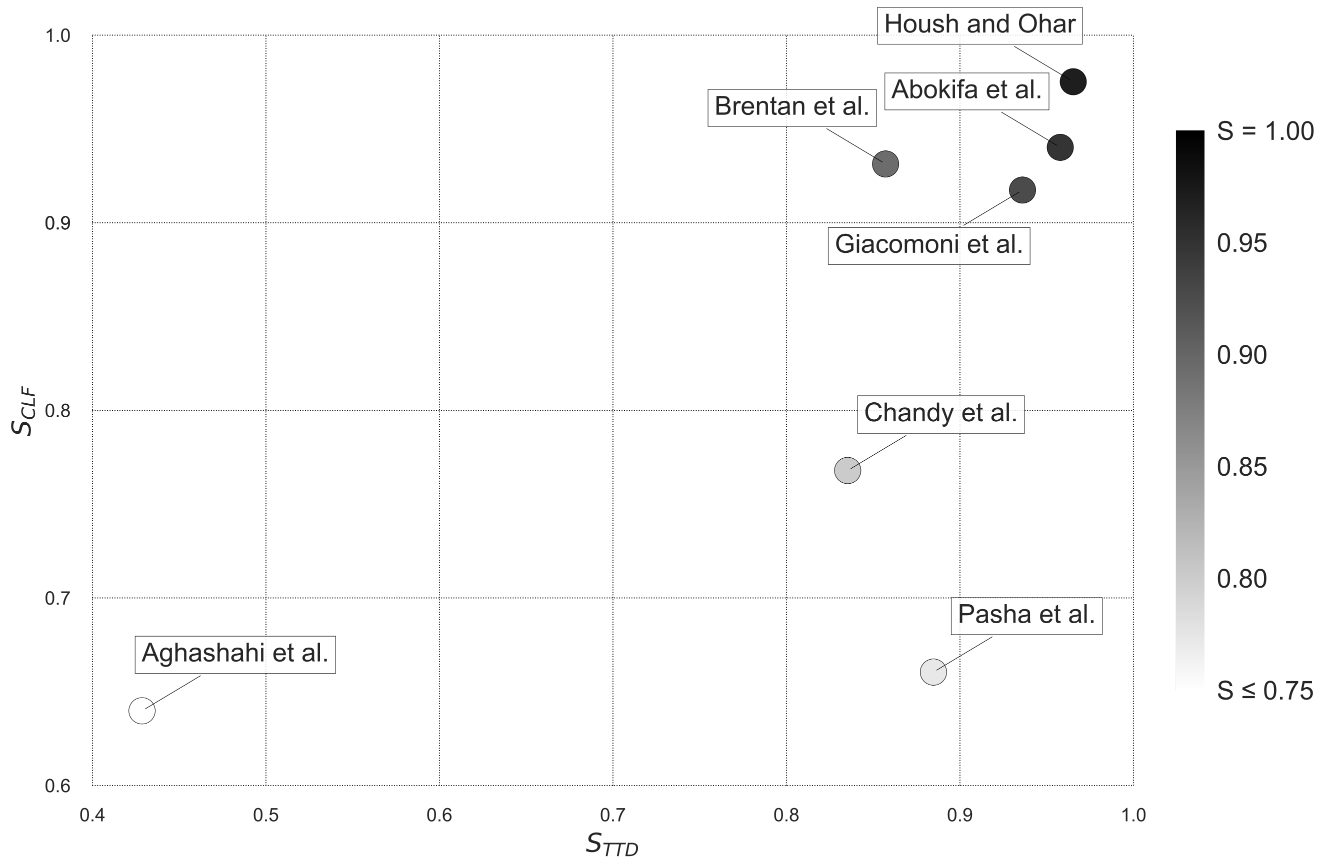

FIG. 3. Graphical representation of the algorithm performance, measured in terms of time-to-detection (S_{TTD}, horizontal axis), classification performance (S_{CLF}, vertical axis), and overall ranking score (S, color-bar).

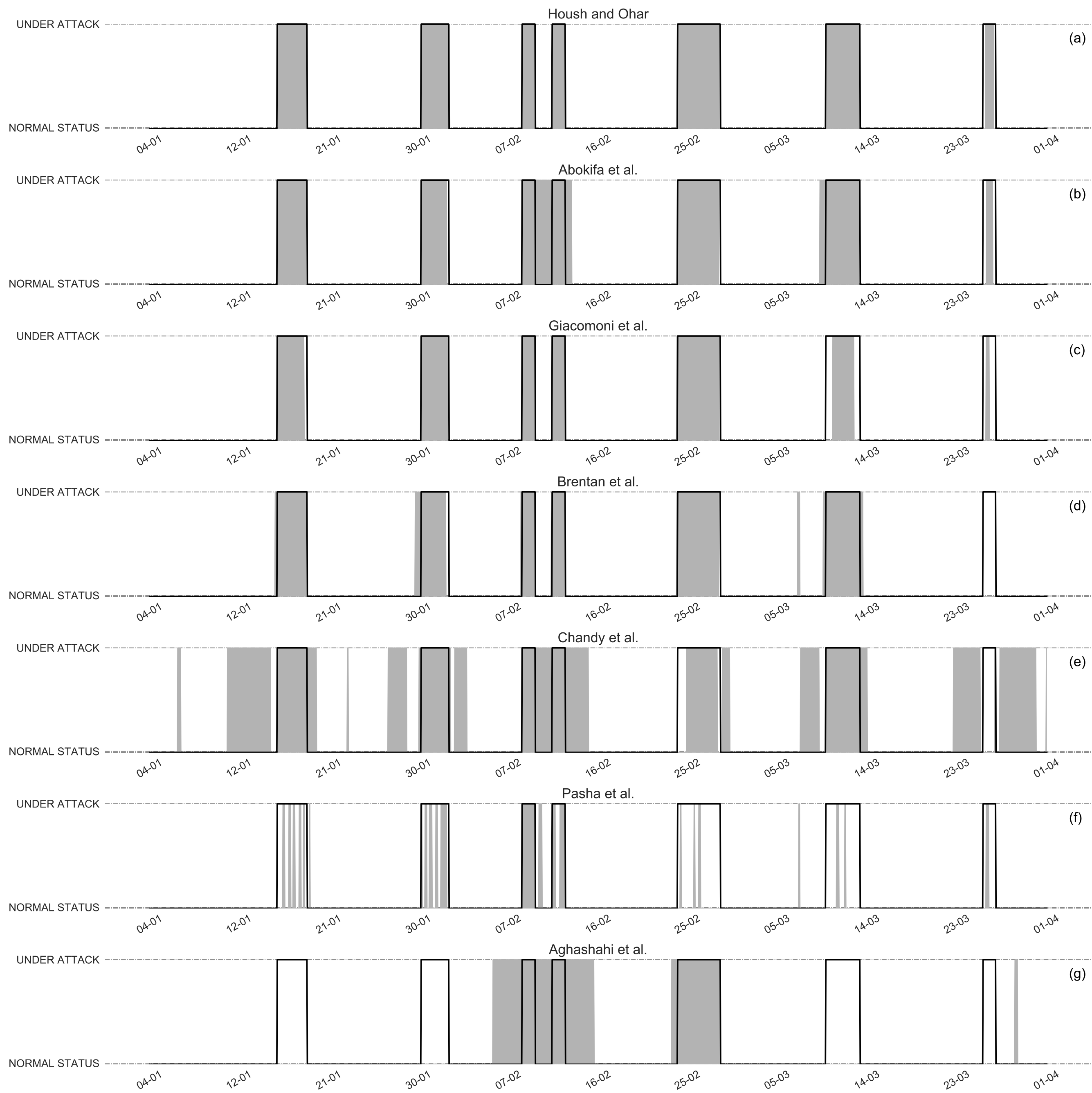

FIG. 4. Comparison between actual and detected attacks (gray area and black line, respectively) for the *Test dataset*. Each panel corresponds to a different attack detection algorithm.