2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Resilience of Cyber-Physical Systems: an Experimental Appraisal of Quantitative Measures

Giuseppina Murino, Alessandro Armando, Armando Tacchella[1]
Dipartimento di Informatica, Bioingegneria e Ingegneria dei Sistemi (DIBRIS)
Università degli Studi di Genova –
Viale Causa 13, 16145 –
Genova, ITALY
giuseppina.murino@edu.unige.it
alessandro.armando@unige.it
armando.tacchella@unige.it

Abstract: Cyber-Physical Systems (CPSs) interconnect the physical world with digital computers and networks in order to automate production and distribution processes. Nowadays, most CPSs do not work in isolation, but their digital part is connected to the Internet in order to enable remote monitoring, control and configuration. Such a connection may offer entry-points enabling attackers to gain control silently and exploit access to the physical world at the right time to cause service disruption and possibly damage to the surrounding environment. Prevention and monitoring measures can reduce the risk brought by cyber attacks, but the residual risk can still be unacceptably high in critical infrastructures or services. *Resilience* – i.e., the ability of a system to withstand adverse events while maintaining an acceptable functionality – is therefore a key property for such systems. In our research, we seek a *model-free, quantitative*, and *general-purpose* evaluation methodology to extract *resilience indexes* from, e.g., system logs and process data. While a number of resilience metrics have already been put forward, little experimental evidence is available when it comes to the cyber security of CPSs. By using the model of a real wastewater treatment plant, and simulating attacks that tamper with a critical feedback control loop, we

1 The authors wish to thank Leonardo S.p.A. for its financial support. The research herein presented is partially supported by project NEFERIS awarded by the Italian Ministry of Defense to Leonardo S.p.A. in partnership with the University of Genoa. This work received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 830892 for project SPARTA.

provide a comparison between four resilience indexes selected through a thorough literature review involving over 40 papers. Our results show that the selected indexes differ in terms of behavior and sensitivity with respect to specific attacks, but they can all summarize and extract meaningful information from bulky system logs. Our evaluation includes an approach for extracting performance indicators from observed variables which does not require knowledge of system dynamics; and a discussion about combining resilience indexes into a single system-wide measure is included.

Keywords: *cyber-physical systems security, critical infrastructure protection, situational awareness and security metrics*

1. INTRODUCTION

A cyber-physical system (CPS) intertwines physical processes, hardware, software, and communication networks [1]. Examples of CPSs include water treatment plants, power plants and distribution networks, industrial plants, transportation vehicles, and smart buildings. The number of security incidents affecting CPSs has been steadily increasing over the past few years – see, e.g., [2]. The bottom line is that CPSs connected to the Internet can be the root cause of disruption in services, damage to equipment or severe impairment of human activities. Malicious acts most often exploit the weakness of the "*red dot*" representing the virtual place of convergence between Information Technology (IT) and Operation Technology (OT): exploitation of the former provides attack vectors, while exploitation of the latter makes kinetic impacts possible. Detecting weaknesses, fixing them and monitoring critical events in CPSs are compelling and heavily investigated matters, but we must also acknowledge that, in spite of all the efforts made to secure CPSs, interconnected systems may never be fully secure.

In this scenario, the concept of *resilience* emerges as an additional target, complementary to prevention and protection from attacks, but no less important. This line of thought is pervasive in the Presidential Policy Directive 21 [3] about the security of critical infrastructure, which defines resilience as "*[...] the ability to [...] withstand and recover rapidly from disruptions. Resilience includes the ability to withstand and recover from deliberate attacks, accidents, or naturally occurring threats or incidents*". More recently, the term cyber resilience has been coined to identify specifically "*the ability to continuously deliver the intended outcome despite adverse cyber events*" [4], and this is the interpretation to which we adhere in the following. More specifically, we believe that stakeholders like CERTs (Computer

Emergency Response Teams), management authorities, regulators, and local and national government branches could be interested in a resilience evaluation framework possessing the following properties:

- *Model-free*. Accurate mathematical models of real-world scale CPSs are very difficult to obtain and maintain. Therefore, the assessment of resilience should not require a detailed description of the system dynamics, e.g., in the form of system equations or other formal models, but rather it should be possible to rely on monitored process data and events only.
- *Quantitative*. A synthetic measure (or index) must be provided that describes as faithfully as possible the amount of damage that a system can tolerate before becoming unstable or irreversibly damaged, or before exhibiting potentially dangerous behaviors.
- *General-purpose*. The way in which the resilience index is computed, starting from performance indicators, should be applicable, in principle, to as wide a class of systems as possible, in order to achieve economy of scale in the deployment of the framework.

We propose an evaluation methodology that fulfills all the requirements cited above to extract resilience indexes from, e.g., system logs, control process data, and SIEM (Security Information and Event Management) tool logs. While several proposals exist in the literature, many of them do not meet the requirements we seek and, for those that do, little or no experimental evidence about their adequacy to account for resilience against cyber attacks is available. In order to start bridging this gap, out of a literature analysis consisting of 47 research papers and surveys, we selected four indexes that can be applied to quantify resilience independently from system dynamics and structure. Using the model of a real wastewater treatment plant, and simulating attacks that tamper with a critical feedback control loop inside the plant, we compare the indexes considering different attack hypotheses on a daily basis using Monte Carlo simulations. The computation of the indexes is oblivious of specific features of the system, but critically depends on the selection of performance indicators to extract system performances out of the evolution of monitored data. Our results show that the distributions of the selected indexes across the simulation of different attacks differ in terms of behavior and sensitivity, but they all extract meaningful information from bulky system logs.

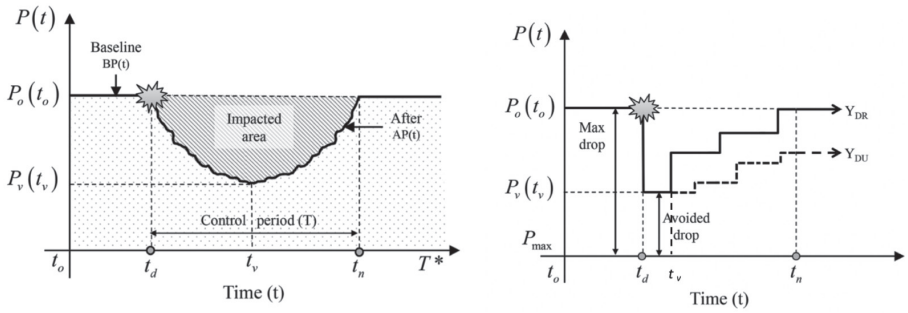
To sum up, the main contributions of the paper are:

- Comparison of four resilience indexes obtained from a thorough literature analysis involving over 40 research papers, in order to ensure model freedom and generality.

- An approach that does not require a mathematical model of system dynamics to extract performance indicators from observed variables.
- A discussion and a proposal about combining resilience indexes obtained from several process variables into a single system-wide measure.

The rest of the paper is structured as follows. In Section 2, we introduce the basic terminology. We succinctly review the related literature and we introduce the indexes we selected for evaluation, including some of the motivation behind their choice. In Section 3, we introduce our wastewater treatment facility case study and we describe the model that we devised in Matlab/Simulink® including its simulation under attack-free conditions. In Section 4, we describe the experimental models, including attack modalities, extraction of performance indicators and a discussion about the combination of resilience indexes. In Section 5, we present some results related to the case study according to the experimental setup described in Section 4. A brief discussion of the results is contained in Section 6, and we conclude the paper in Section 7 with some final remarks.

FIGURE 1: GENERIC RESILIENCE EVALUATION SCENARIO (LEFT) FOCUSING ON THE DIFFERENCE BETWEEN BASELINE PERFORMANCE BP(T) AND AFTER-IMPACT PERFORMANCE AP(T) OVER A CONTROL PERIOD T. GENERIC RESILIENCE EVALUATION SCENARIO (RIGHT) FOCUSING ON THE MAXIMUM AND AVOIDED PERFORMANCE DROPS DURING THE ADVERSE EVENT. NOTATION AND PICTURES FROM [5].



2. BACKGROUND AND RELATED WORK

The definitions and notation that we use are mostly borrowed from [5]. The plot in Figure 1 (left) is presented to describe a generic resilience evaluation scenario. The coordinates are time (x-axis) and performance (y-axis), *BP(t)* is the *Baseline Performance* and represents the performance of the system under normal conditions,

whereas *AP(t)* is the *After-impact Performance* and represents the performance of the system after the impact of some disruptive event. Such an event is assumed to happen at time t_d (*disruption time*) and end at time t_n (*return to normality time*), where $T = t_n - t_d$ is defined as the *control period* in [5]. A further point of interest is t_v (*lowest performances time*) where the system reaches the minimum level of performance after disruption. The period T^* is defined as the *observation period* and the condition $T^* > T$ holds. The plot in Figure 1 (right) introduces the notion of *maximum performance drop* (Max drop) and *avoided performance drop* (Avoided drop) which represent, respectively, how much performance can be lost before the system ceases to be functional and how much performance is left when the system reaches the minimum level of functionality after the attack and before the recovery. With reference to Figure 1 (left), the first resilience index that we consider is introduced by [6] and is defined as

$$\psi_A = \int_{t_d}^{t_n} \frac{AP(t)}{T}\, dt$$

The index ψ_A considers the area of the curve *AP(t)* normalized over the control period *T*, i.e., the residual normalized performance of the system during the disruption. Clearly, the higher the value, the closer to normal operating conditions, and the greater the resilience of the system. The advantage of this index is that it does not require establishing a baseline and it can be readily applied to any performance indicator computed on process data. The main disadvantage is that it assumes knowledge of the control period which, in the majority of cyber attacks, is not known and is difficult to estimate.

An index that overcomes such limitations, but that does require the establishment of a baseline performance, is introduced by [7], [8] and [9]. It is defined as

$$\psi_B = \frac{\int_{t_0}^{T^*} AP(t)dt}{\int_{t_0}^{T^*} BP(t)dt}$$

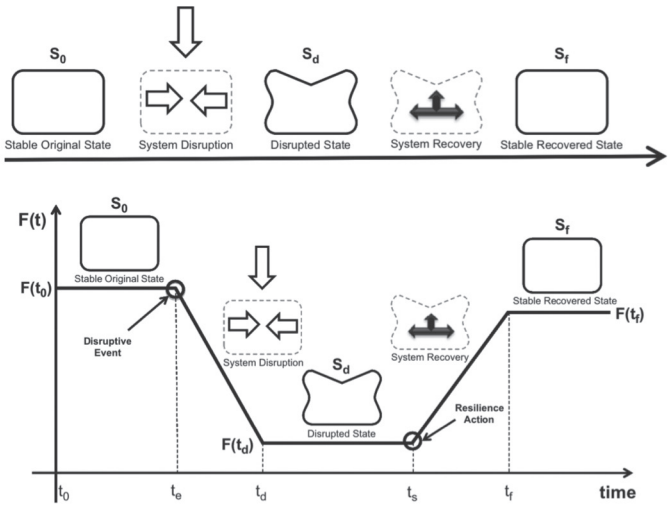
This index is the ratio of the areas enclosed by the curves $AP(t)$ and $BP(t)$. It ranges from 0 to 1, where the former is the limit case in which the disruptive event occurs at time t_0 and the system immediately loses its functionality, so that $AP(t)=0\ \forall t \in [t_0; T^*]$. The latter is the limit case in which no functionality is lost, i.e., $AP(t)=BP(t)\ \forall t \in [t_0; T^*]$. Both ψ_A and ψ_B consider the overall evolution of the system during (a subinterval of) the observation period. However, in [10] an index based on the values of max drop and avoided drop is put forward:

$$\psi_C = \frac{Avoided\ drop}{Max\ drop} = \frac{P_v(t_v)\text{-}P_{max}}{P_o(t_o)\text{-}P_{max}}$$

In this case, the evolution of the curves $AP(t)$ and $BP(t)$ are not relevant to establishing the value of the index, since only their extreme values are taken into account. While it is sufficient to consider only specific points in time to compute ψ_C, the evolution of system performances over the control period is completely disregarded.

Besides the above-mentioned contributions, our literature analysis included several other papers that we do not list here owing to a lack of space. References that are worth mentioning are [11], which helped us frame the problem of resilience evaluation, and [12], which provided us with an extensive bibliography to which we refer for further reading about the topic. Since our case study relates to wastewater treatment, we also considered a number of references related to the resilience of water/wastewater treatment plants, including [13], [14] [15] and [16], but we could not find additional candidates for evaluation that met our requirements. In particular, all the indexes proposed in the water/wastewater literature are specific to a given topology and system structure and are difficult to generalize to other plants.

FIGURE 2. PICTORIAL EVOLUTION OF THE STATE OF A SYSTEM UNDER ATTACK (TOP) AND RELATIONSHIP BETWEEN STATE EVOLUTION AND PERFORMANCE OF THE SYSTEM COMPUTED BY A FIGURE OF MERIT (FOM) FUNCTION (BOTTOM). NOTATION AND PICTURES FROM [17].



Considering the fact that the resilience indexes of our choice are based on performance indicators, the question of how to compute such indicators arises. In other words, while

it is relatively easy to monitor process variables, the performance of the system cannot always be monitored directly, and should be inferred from collected data. In Figure 2, we present two plots excerpted from [17], wherein a resilience-oriented general-purpose and model-free method to derive performance indicators from state variables is presented. The plot on the top of Figure 2 represents pictorially the evolution of the state of the system during a disruptive event. The deformed box represents the state of the system under duress, and it is meant to show that the impact on state variables can involve several of them at the same time. Nevertheless, as it is shown in the plot at the bottom of Figure 2, we must relate the evolution of state variables to some "bathtub" curve which resembles the curve $AP(t)$ of Figure 1 (left). The proposal of [17] is to introduce a *Figure of Merit* (FOM) function, i.e., a function $F:S\rightarrow\mathbb{R}$ which maps any state $s \in S$ to a corresponding performance indicator. In general, mappings such that the condition $F(s)>F(s')$ holds whenever the performance of the system in state s is better than in state *s'* should work. In [17] no details on how to derive such a function are given, because this is a system-specific process.

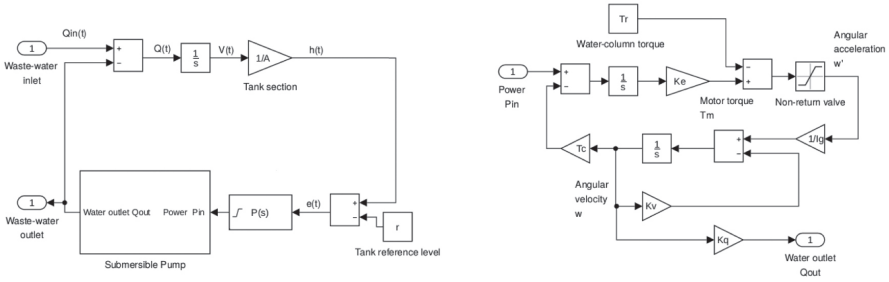
3. CASE STUDY: WASTEWATER TREATMENT FACILITY

A. Brief Description

The facility[2] performs sewage treatment using MemJet™/MemPulse™ MBR (micro-membranes) technology and ensures depollution and dumping at sea of urban wastewater produced by domestic and economic activities in an international tourist area encompassing a marine reserve. The facility handles an estimated maximum of 36,000 people, roughly equivalent to a wastewater supply of 250 liters per person, per day. The maximum output reaches up to 1,200 cubic meters/hour of purified wastewater. The plant is heavily automated: all biological, chemical and mechanical processes are controlled and monitored by a SCADA system connected through the Internet with a remote monitoring center located in the headquarters of the utility company running the plant. The plant consists of a pre-treatment compartment, responsible for filtering large solids – e.g., rags, plastics, nappies, grit and floating materials, oils and fats – before feeding a balancing reservoir. From here, the pre-treated input flow is pumped into the biological compartment where, passing through a denitrification (anoxic) process and a transition into nitrification-oxidation tanks, the oxygenated mixed liquor flows into the MBR reactor for solid-liquid separation and subsequent discharge of the effluent at sea. This is a physical-biological process, which requires precise software-based regulation in order not to wear out micro-membranes and to avoid outputting untreated liquor. The maximum mass flow rate through of MBR tanks – a reference for the whole process − is 900 cubic meters per hour.

2 Name and location of the facility cannot be disclosed for security reasons.

FIGURE 3. MATLAB/SIMULINK® MODELS OF THE NITRIFICATION-OXIDATION (NO) TANK SUBSYSTEM (LEFT) AND OF THE "SUBMERSIBLE PUMP" COMPONENT (RIGHT). THE ACTUAL UNIT IS DRIVEN BY AN ASYNCRHONOUS MOTOR WITH 15KW OF RATED POWER CONTROLLED THROUGH AN INVERTER. IN THIS SIMPLIFIED MODEL WE ASSUME THAT THE INPUT SIGNAL IS THE POWER DELIVERED TO THE MOTOR AS COMPUTED BY A PROPORTIONAL REGULATOR.



B. Modeling and Simulation

In order to achieve a realistic, yet manageable, case study, we decided to model only the main wastewater cycle. Furthermore, we focus on the nitrification-oxidation process (tank NO) which is upstream from the final purification process (tank MBR) and thus is critical for the performance of the whole cycle. In Figure 3 (left) we show the detailed Matlab/Simulink® model of the tank NO. As we can see from the diagrams, we have assumed a simplified (first order) linear model, whereby the total volume *V(t)* of fluids contained in the tank is obtained by integrating the net inlet mass flow rate *Q(t)* which, in turn, is obtained by subtracting the outlet mass flow rate *Qout(t)* from the tank inlet *Qin(t)*. While the latter is an input to the NO subsystem, the tank outlet is controlled by electrical pumps driven by a proportional regulator tracking a given set point *r* on the height of the tank. The detail of the motor/pump model is given in Figure 3 (right). Also in this case, we assumed a (second order) simplified linear model of an asynchronous drive, whereby the pump rotation generates both viscous friction and counter-motion force, which simulates the asynchronous drive frequency lag.

Two key nonlinearities in the model are *(a)* the saturation of the control signal between 0 and 15KW, which corresponds to the actual range of power within which the pump operates and *(b)* the presence of a non-return valve which does not allow the pump to reverse its operation. The goal of the regulator is to avoid the tank becoming too full, so as to avoid triggering emergency bypasses, or too empty, so as to avoid impairing the chemical process undergone in the NO tank. Both events are undesirable because bypasses dump untreated sewage liquor in the sea, whereas incomplete chemical processing of wastewater may cause failures in subsequent steps. For this reason, we decided to focus our study on this part, on the hypothesis that an attacker may gain

virtual access to the facility network and compromise this feedback loop and thus also the inlet flow to the MBR tank. As a yardstick for the calculation of resilience indexes, we simulate the plant without assuming external attack attempts in a Monte Carlo setting. To achieve this, we consider historical data made available from the managing utility to simulate regular sewage inlet. Random variates of the daily inlet profile under conditions of maximum utilization are obtained by adding (band-limited) Gaussian white noise with deviations of 20%. In the following, we call *baseline scenario* the simulation obtained by running the plant without attacks.

4. MODELING: SIMULATING ATTACKS, PERFORMANCE INDICATORS AND SYSTEM-WIDE RESILIENCE

A. Attack Scenarios

To develop attack scenarios, we must consider the effects that an attacker may induce by gaining system access. Conceptually, feedback control loops are at the core of every CPS, and an attacker gaining access to the control system can alter them in three ways: *(a)* by changing the set point, *(b)* by altering the feedback signal, and *(c)* by changing the regulator parameters. To illustrate, consider the control loop that keeps the level of the NO tank close to the desired level shown in Figure 3 (left). Here, attack *(a)* corresponds to changing the desired tank level r, attack *(b)* corresponds to altering the actual tank level feedback h, and attack *(c)* corresponds to changing the proportional gain of the regulator $P(s)$. In practice, an attacker may decide to perform all such actions and in more than one part of the system, as well as other disruptive actions – blocking the functionality of components or flooding them with requests. Some of these attacks can be prevented or detected by SIEM tools, but attacks on feedback loops can be subtle and destructive. As an example, the pump keeping the NO tank at level can be exercised more than necessary by fooling the controller about the tank level in a small, but persistent way. Such an attack pattern – similar to the one staged by the famous Stuxnet virus [18] – is very difficult to detect, but it reduces the residual life of the pump and thus it is worth evaluating its impact on resilience.

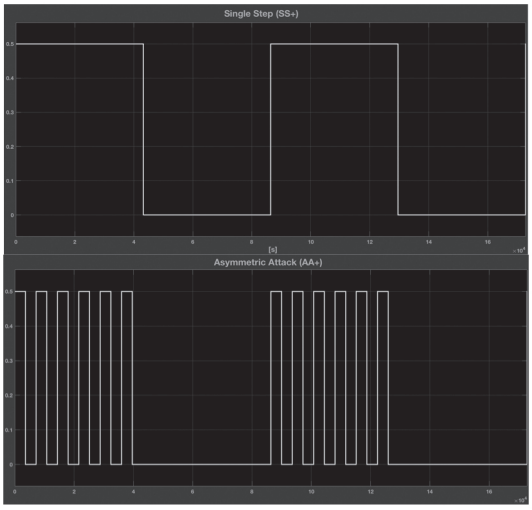
In our simulations we assume that an attacker may alter the set point of the regulator by subtracting a disturbance – attack *(a)*. Under this hypothesis and given the structure of the feedback loop, this attack is equivalent to an alteration of the feedback signal – attack *(b)*. We did not consider attack *(c)* as well as multiple or blocking attacks, but our evaluation framework is able to handle them without modifications. We can obtain several attack scenarios by changing:

- The *duration*, i.e., the control period (in seconds) $T=t_n-t_d$, as defined in Section 2.

- The a*mplitude* Δ_a, i.e., how much the reference signal is changed.
- The *frequency* f_a; when the disturbance is periodically zeroed every $1/f_a$ seconds during T.

In Figure 4 we show an example assuming T=12 hours and Δ_a=0.5 meters. The plot on top depicts the case in which the duration of the disturbance is held fixed during the attack: we call this *positive single step attack* scenario (SS+), and we foresee also a negative counterpart SS- (*negative single step attack*). The plot on the bottom depicts the case in which the attack signal has a period of two hours (f_a in the order of 10^{-4} Hertz): we call this *positive asymmetric attack scenario* (AA+) and *negative asymmetric attack scenario* (AA-) its counterpart. We also combine the two attacks in a *symmetric attack scenario* (SA), wherein the disturbance ranges from Δ_a to $-\Delta_a$ with frequency f_a. In Section 5 we report results obtained by running these scenarios with different values of T, and Δ_a.

FIGURE 4. CHANGES TO THE NO TANK REFERENCEL BROUGHT BY THE HACKER ATTACK. SINGLE STEP POSITIVE (TOP) AND ASYMMETRIC POSITIVE (BOTTOM). THE PLOTS DEPICT TWO ATTACKS LASTING 12 HOURS EACH OVER A TOTAL TIME OF 48 HOURS. THE PERIOD OF THE ASYMMETRIC ATTACK IS 2 HOURS.



B. Building Performance Indicators Through FOM Functions

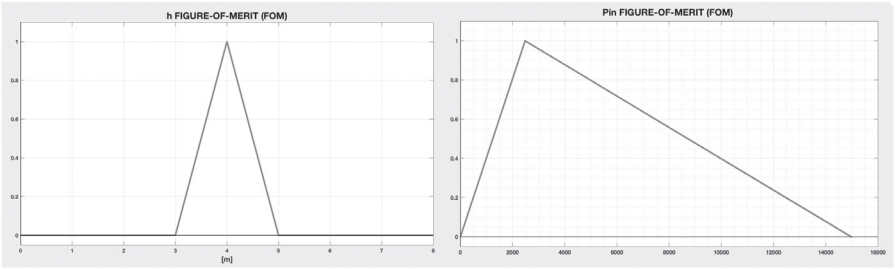
The resilience indexes presented in Section 2 rely on performance indicators, and suitable FOM functions must be provided to map observed variables to the performance space. Considering our case study, the variables that we observe are the following:

- The height of the NO tank *h*; this is the state variable whose reference point is subject to the attack, and it is thus the main focus of our investigation.
- The power delivered to the pump *Pin*; among the effects of a successful and silent cyber-attack, wearing the pump and reducing its residual life is a concrete possibility.
- The outlet mass flow rate *Qout*; the mass flow rate through membranes in the MBR tank, which is downstream from the NO tank, must be regulated precisely, lest the purification process malfunctions or even ceases to work.

As for the definition of FOMs, we can make some observations:

- FOM functions are of the form $F:D\rightarrow\mathbb{R}$, where D is the domain of the observed variable, but without loss of generality we can restrict our FOMs in the range [0;1], where 0 and 1 represent minimum and maximum performance, respectively.
- We posit that, when an observed variable x is close to some desirable value(s) x_{good}, then $F(x)\cong 1$, whereas if x is close to undesirable value(s) x_{bad}, then $F(x)\cong 0$.
- $F(x)$ should behave monotonically with respect to the distance from x_{bad} and x_{good}: it must decrease when getting close to x_{bad} and increase when getting close to x_{good} – a concept we borrow from [19].

FIGURE 5. FIGURE-OF-MERIT (FOM) FUNCTIONS FOR TWO OUT OF THREE OBSERVED VARIABLES RELATED TO THE NO TANK: TANK HEIGHT H (TOP) AND POWER SIGNAL TO THE PUMP PIN (BOTTOM). EACH FOM FUNCTION TAKES AS INPUT AN OBSERVED VARIABLE AND RETURNS AN ADIMENSIONAL FIGURE BETWEEN 0 (WORST PERFORMANCE) AND 1 (BEST PERFORMANCE).



We now consider Figure 5, where we represent FOM function for NO tank height (top) and power delivery to the pump (bottom). We do not show the one for outlet mass flow rate, but it similar to the ones shown in Figure 5. The shape of the functions is the simplest satisfying the constraints outlined above, where a linear decay in performance is assumed when variables are getting away from desirable values. More specifically, for each observed variable we identify (un)desirable values as follows:

- The reference value of tank height h is 4 meters, therefore we consider h_{good}=4; the tank can tolerate some amount of overshooting of the reference level, but heights of five meters and more may cause spilling; therefore, we set h_{bad}=5 and, symmetrically, h_{bad}=3.
- Under normal conditions, the power delivery to the motor is Pin≅3 KW, therefore we set Pin_{good} to the average value under normal operations; the pump operates within 0 to 15KW, which means that delivering power always close to 15KW reduces its residual life, whereas values close to 0 mean that the pump is switched off or works at reduced power; therefore, we set Pin_{bad}=0 and Pin_{bad}=15000.
- Under normal conditions, the outlet mass flow rate is $Qout$≅0.05 m^3/s, therefore we set $Qout_{good}$ to the average value that the variable assumes under normal daily operations. Attempting to deliver more than 0.3 m^3/s mass flow rate to the MBR tank as well as shutting down the flow completely might damage the membranes; therefore, we can set $Qout_{bad}$=0 and $Qout_{bad}$=0.3.

In Figure 5, we show FOM functions assuming linear decay of performances. We remark that this choice is arbitrary and other possibilities exist which are compatible with our assumptions, e.g., quadratic or cubic decay to penalize small changes with respect to x_{good} less than large ones, or RBF (radial basis function) profiles to smooth the decay and avoid discontinuities at the boundaries.

C. A Discussion About System-wide Resilience Indexes

The introduction of FOM functions for each observed variable h, Pin and $Qout$, enables us to compute resilience indexes related to each variable separately. In our comparison this is fine because we have a relatively limited scope of investigation – the feedback control of the NO tank – and we wish to compare the behavior and the sensitivity of the indexes we consider. However, it can be desirable to build indexes that summarize the performance of the system as a whole, instead of relying on many separate figures. This is especially true when the size of the system grows, and so does the number of observed variables. Keeping in mind that we seek a model-free and general-purpose approach, we can consider three possibilities to extend resilience indexes to a system-wide measure:

- Use a FOM function that maps all the observed variables into a single performance indicator; in our case, this would amount to devising a vector function $F(h,Pin,Qout)$ to summarize the change of the observed variables into a single performance index.
- Construct a system-wide performance indicator out of scalar FOM functions; in our case, this would amount to combining $F(h)$, $F(Pin)$ and $F(Qout)$ into a single measure, e.g., a linear combination of the three $F(h,Pin,Qout)=\alpha F(h)+$

$\beta F(Pin)+\gamma F(Qout)$, where $\alpha,\beta,\gamma \in [0;1]$ and $\alpha+\beta+\gamma=1$ are *weights* determining the contribution of FOMs.

- Finally, one may either come up with a definition of resilience that accommodates a vector as a performance indicator, or combine resilience indexes computed with scalar performance indicators on single variables; in our case, one may consider, e.g., that a worst-case estimation of the resilience of the whole system can be obtained by considering the smallest index computed according to $F(h)$, $F(Pin)$ and $F(Qout)$.

The first approach is quickly ruled out as the number of observed variables increases. As long as the definition of the FOM function relies on a manual process, defining hyper-surfaces that are meant to respect the given constraints is untenable. One may consider using optimization or machine-learning techniques in order to devise suitable $\mathbb{R}^n \rightarrow [0;1]$ mappings (n number of observed variables), but the complexity of the procedure should be factored in. The second approach provides a simplification of the first one, and it remains amenable to manual configuration as long as the number of FOM functions to combine remains small. Scalable linear optimization and relatively simple machine-learning techniques can be used when the number of variables to combine is growing, and hierarchical composition is a possibility. Also, the definition of each single FOM will remain an explainable scalar-to-scalar function. Clearly, the choice of weights to combine the FOM functions is critical for the assessment of resilience, because underestimating or overestimating impacts of a specific FOM may obscure relevant effects in the evaluation of the global resilience index. The third option shares the same issues as the first one when it comes to finding a vector-based index, whereas the combination of different resilience measures is the only approach for which some literature exists. In particular, in [6], the authors propose a method to combine different indexes based on the assumption that they are computed from independent systems. This proposal is not applicable to our case, because the indexes are part of a single feedback control loop. In this case, our proposal is to apply a "weakest link" rule, and estimate the resilience of the overall system considering the resilience index with the smallest median among the ones we compute.

5. EXPERIMENTAL ANALYSIS

We briefly recapitulate the definitions that we have introduced so far to put them in context for our experimental setup. Starting from the resilience indexes that we define in Section 2, let $F:\mathbb{R}\rightarrow 0;1$ be one of the FOM functions introduced in Section 4-B, and $x \in \{h,Pin,Qout\}$ be one of the observed variables, where $x(t)$ denotes its value under normal operations and $x_a(t)$ denotes its value under attack scenarios. We consider four resilience indexes defined as follows:

$$\psi_A = \int_{t_d}^{t_n} \frac{F(x_a(t))}{T}\, dt \qquad \psi_B = \frac{\int_0^{T^*} F(x_a(t))dt}{\int_0^{T^*} F(x(t))dt} \qquad \psi_C = \frac{\min_{t\in T^*} F(x_a(t))}{\min_{t\in T^*} F(x(t))} \qquad \psi_D = \int_0^{T^*} \frac{F(x_a(t))}{T^*}\, dt$$

The indexes ψ_A and ψ_B are exactly those defined in Section 2, under the hypothesis that t_0=0. The index ψ_C is computed assuming that the worst-case estimation of the maximum performance drop is the minimum performance of the system under normal operating conditions for a given observation period T^* and that the avoided drop is the minimum performance of the system under attack. Finally, the index ψ_D is obtained from ψ_A by changing the span of the integral from $T=t_n-t_d$ to T^*. The idea behind ψ_D is that, while in our simulations the control period T is known, in practice it might be difficult to estimate. On the other hand, the observation period T^* is always chosen by design: in all our experiments, T^*=24 hours.

As far as the attack is concerned, we consider all the scenarios defined in Section 4-A, namely single step positive and negative attacks, denoted SS+ and SS-, symmetric attack, denoted SA, and asymmetric positive and negative attacks, denoted AA+ and AA-. For each such attack, we build a factorial experiment with different levels of T, Δ_a and f_a. In particular we consider:

- T={6,12,18}, i.e., the attack always starts at midnight and lasts 6 to 18 hours.
- Δ_a={0.25,0.5,075}, i.e., the attacker can change the tank reference level from 25 to 75 centimeters.
- f_a={1/3600,1/7200,1/10800}, i.e., the attack period can be one, two, or three hours.

The main reason behind the choice of these values is to increase the probability that the attack on the system remains silent. Indeed, decreasing the period of the attack ($1/f_a$) below two hours can trigger fast oscillatory system dynamics (e.g., in the pumps) that are unusual in the normal operation of the facility and thus can be identified as anomalous. Also, attempting to change the tank reference level beyond one meter can cause overflow alarms to be triggered. Finally, the attack period is kept at a fraction of the observation period, knowing that longer attack periods imply higher chances of being uncovered. As mentioned in Section 3-B, all the scenarios are simulated on a daily basis, obtaining a different value of the performance indexes that we average over the number of days – one hundred in all of our experiments – for which the simulation runs.[3]

[3] All our experiments run on a PC equipped with an Intel 2.6Ghz Dual Core i7 CPU, 32GB of RAM and running Matlab/Simulink® ver. 2018a on Mac Os Sierra.

TABLE 1. RESILIENCE INDEXES COMPUTED FOR ALL OBSERVED VARIABLES CONSIDERING FIVE DIFFERENT ATTACK SCENARIOS.

SCENARIO $T = 6\ [h];\ \Delta_a = 0.5\ [m]$	ψ_A						ψ_B						ψ_C						ψ_D					
	h		P_{in}		Q_{out}		h		P_{in}		Q_{out}		h		P_{in}		Q_{out}		h		P_{in}		Q_{out}	
	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr
POSITIVE SINGLE STEP ATTACK (SS+)	0.4577	0.0028	0.7098	0.025	0.7557	0.0256	0.8685	3.6e-04	0.9675	4.98e-04	0.9686	0.0017	0.4586	0.0047	0	0	1.28e-11	1.71e-11	0.7915	0.0011	0.8659	0.0069	0.8734	0.0074
NEGATIVE SINGLE STEP ATTACK (SS-)	0.5793	0.0024	0.7777	0.0244	0.8227	0.0236	0.8970	0.0013	0.9635	3.49e-04	0.9626	6.13e-04	0.6197	0.0049	0	0	6.4e-20	1.91e-19	0.8176	0.0013	0.8622	0.0072	0.8680	0.0071
SYMMETRIC ATTACK (SA)	0.7134	0.0021	0.3271	0.0077	0.3709	0.0072	0.9407	9.87e-04	0.8677	0.0043	0.8675	0.0039	0.5883	0.0038	0	0	0	0	0.8573	9.32e-04	0.7768	0.0037	0.7824	0.0042
POSITIVE ASYMMETRIC ATTACK (AA+)	0.7542	0.0051	0.4408	0.0224	0.4730	0-0224	0.9522	8.64e-04	0.8971	6.7e-04	0.8945	0.0014	0.4586	0.0047	0	0	9.04e-17	3.59e-16	0.8678	0.0014	0.8027	0.0068	0.8063	0.0070
NEGATIVE ASYMMETRIC ATTACK (AA-)	0.7196	0.0033	0.4324	0.0237	0.46	0.0209	0.9427	0.0016	0.8952	6.52e-04	0.8913	9.37e-04	0.5929	0.0049	0	0	7.97e-18	2.26e-17	0.8592	0.0013	0.8011	0.0067	0.8038	0.0069

In Table 1 we show the results for T=6 and Δ_a=0.5. Each row of the table is related to an attack scenario. Columns are divided into four groups, corresponding to the resilience indexes, and each group reports the median and interquartile range (iqr) of the resilience index computed using a specific variable and related FOM function. The choice of median and iqr as measure of center and spread, respectively, are motivated by the fact that they are more robust to outliers and the presence of skewed distributions. A glance at the table reveals the following facts:

- The iqr is always at least one order of magnitude smaller than the median except when the median is 0 as in ψ_C; this indicates that indexes are not very sensitive to the random variation of the input flow.
- The index ψ_D is more conservative than ψ_A; this is to be expected, because the former averages the effects of the attack over the whole observation period.
- Considering observed variable *h*, the lowest resilience values are obtained for the SS+ attack; this is because the attack signal is *subtracted* from the reference level, and thus throughout the duration of the attack the tank is seen by the controller to be emptier than in reality; in AA+ and SA attacks this is not true, because the attack signal oscillates and, on average, the controller keeps the tank level closer to normal.
- Considering observed variables *Pin* and *Qout*, the worst figures are obtained for the SA attack because the performance of the pump and the mass flow rate output are far from 1 only during transient regimes induced by the "steps" in the attack signals; therefore, in SS+ and SS- attacks, the height of the tank remains "off balance" while the pump and the mass flow rate output stabilize to levels corresponding to normal operation.

We analyzed the data shown in Table 1 considering 36 distributions obtained with SA, AA+ and AA- attacks. We preliminary tested each distribution for normality with the Shapiro-Wilk test, and groups of distributions across attack modality for

homoskedasticity (equal variance) with the Levene test (non-parametric version). The results can be summarized as follows:

- the null hypothesis of the Shapiro-Wilk test (values being normally distributed) cannot be rejected at the 5% confidence value for all but a few distributions, e.g., ψ_A computed on state variable h for SA, and the distributions of ψ_C for state variables *Pin* and *Qout*.
- considering the distributions of single resilience indexes computed for specific state variables, and comparing them across different attacks, the null hypothesis of the Levene test (variances being equal) can be rejected at the 5% confidence value for all the groups we consider with the single exception of ψ_C for state variables h and *Pin*.

Given the above results, we compare the distributions across attack modalities with a multiple pairwise Mann-Whitney U-test (non-parametric alternative to t-test) using Bonferroni's correction for p-values. Overall, the results of this test confirm that the qualitative observations we made above hold true.

For example, in the case of ψ_A considering variable h, and attacks SA, AA+ and AA-, the null hypothesis that two samples obtained from different attacks are coming from the same distributions can be rejected at the 5% confidence level in all cases. For lack of space, data obtained with T=12,18 hours, Δ_a=0.25,0.75 meters and f_a={1/3600,1/10800} are not reported, but similar considerations apply also to these cases.

Using the rule proposed in Section 4-C, the overall resilience of the system under the various attacks can be estimated considering the minimum value for each index in a given row. For instance, if we consider ψ_A with T=6 and Δ_a=0.5, we would get a global index $\Psi_A = 0.4577$ (the value for h) in the SS+ attack, and $\Psi_A = 0.3271$ (the value for *Pin*) for the SA attack.

6. DISCUSSION

While our current results are not a ready-made tool for detecting or preventing cyber attacks, in principle some of the resilience indexes we propose could be deployed to support an intrusion detection tool, e.g., by letting the tool "learn" the baseline distribution of some resilience index during secure operation, and then relying on the tool to detect significant deviations from the baseline during normal operation. Our methodology consists of three steps:

- Identify the relevant state variables considering those available from process control logs.
- Build FOM functions considering (un)desirable values and making assumptions about the effects of variables change on system performance.
- Compute resilience indexes based on FOM functions.

We stress that any system is amenable to this analysis, therefore our methodology is general-purpose. It is also model-free, because identifying state variables does not require knowing system dynamics in detail; also, identifying (un)desirable values requires behavioral knowledge of the process being carried out by the system but does not require the mathematical model of the system. The advantage of relying on our methodology, with respect to standard intrusion detection applied to single process variables, is that our resilience indexes are built and tested to provide statistically significant deviations when anomalies affect the system, and can also be used to summarize the combined effect of several process variables at once. More generally, if a simulator of the CPS under scrutiny is available, one can test and tune resilience indexes to achieve desired properties by means of controlled experiments, and the indexes engineered through simulations will be deployable on the implemented system without further adaptations. For systems in which simulation is not an option, computing indexes is still possible by relying on process data and system logs, while testing and tuning could be performed by replaying historical data.

One key issue arising in practice is the ability of the selected indexes to tell naturally occurring faults from cyber attacks. Given our current approach, a statistically significant deviation in resilience indexes for the wastewater facility can be produced, e.g., by a faulty pump or a stuck-at-level tank sensor. However, naturally occurring faults exhibit predictable patterns, whereas cyber attacks, in general, do not. Therefore, hints about the cause of an anomaly could come from comparison between several indexes, including those obtained simulating possible faults. While we have not yet developed a procedural way to diagnose symptoms of decreasing resilience indexes, we can observe that the behavior of the system in case of SA, AA+/- attacks can hardly be traced back to a physical anomaly: a change in resilience indexes, that are known to be sensitive to those attacks, will indicate that the system is being compromised with high probability.

7. CONCLUSIONS AND FUTURE WORK

We have improved on the current state of the art in resilience evaluation by providing experimental data showing that it is possible to summarize the resilience of a system through numerical indexes that ensure model freedom and generality. Our approach,

based on FOM functions computed from observed variables, does not require a mathematical model of system dynamics, but only knowledge of (un)desired values for process variables. We have provided a discussion and preliminary experimental evidence about combining resilience indexes obtained from several process variables. Future work will include furthering our investigation into the combination of several FOM functions or resilience indexes in systems with several observed variables and more complex hierarchical structures. We plan to analyze data from logs of real systems and validate the results obtained with simulation to provide tools for security monitoring for critical infrastructure.

REFERENCES

[1] E. A. Lee, "Cyber Physical Systems: Design Challenges" in *11th {IEEE} International Symposium on Object-Oriented Real-Time Distributed Computing {(ISORC} 2008), 5-7 May 2008, Orlando, Florida, {USA}*, 2008.

[2] G. Loukas, *Cyber-physical attacks: A growing invisible threat*, Butterworth-Heinemann, 2015.

[3] B. Obama, "Presidential Policy Directive 21 (PPD21): Critical infrastructure security and resilience" *Washington, DC*, 2013.

[4] F. Björck, M. Henkel, J. Stirna and J. Zdravkovic, "Cyber Resilience-Fundamentals for a Definition." in *WorldCIST (1)*, 2015.

[5] N. Yodo and P. Wang, "Engineering resilience quantification and system design implications: a literature survey" *Journal of Mechanical Design*, vol. 138, n. 11, p. 111408, 2016.

[6] C. S. Renschler, A. E. Frazier, L. A. Arendt, G. P. Cimellaro, A. M. Reinhorn and M. Bruneau, A framework for defining and measuring resilience at the community scale: The PEOPLES resilience framework, MCEER Buffalo, 2010.

[7] D. G. Dessavre, J. E. Ramirez-Marquez and K. Barker, "Multidimensional approach to complex system resilience analysis" *Reliability Engineering \& System Safety*, vol. 149, pp. 34-43, 2016.

[8] M. Ouyang, L. Dueñas-Osorio e X. Min, "A three-stage resilience analysis framework for urban infrastructure systems" *Structural Safety*, vol. 36, pp. 23-31, 2012.

[9] A. Shafieezadeh and L. I. Burden, "Scenario-based resilience assessment framework for critical infrastructure systems: Case study for seismic resilience of seaports" *Reliability Engineering \& System Safety*, vol. 132, pp. 207-219, 2014.

[10] A. Rose, "Economic resilience to natural and man-made disasters: Multidisciplinary origins and contextual dimensions" *Environmental Hazards*, vol. 7, n. 4, pp. 383-398, 2007.

[11] Y. Y. Haimes, "On the definition of resilience in systems" *Risk Analysis: An International Journal*, vol. 29, n. 4, pp. 498-501, 2009.

[12] S. Hosseini, K. Barker and J. E. Ramirez-Marquez, "A review of definitions and measures of system resilience" *Reliability Engineering \& System Safety*, vol. 145, pp. 47-61, 2016.

[13] P. Juan-Garcia, D. Butler, J. Comas, G. Darch, C. Sweetapple, A. Thornton and L. Corominas, "Resilience theory incorporated into urban wastewater systems management. State of the art" *Water Research*, vol. 115, pp. 149-161, 2017.

[14] S. N. Mugume, D. E. Gomez, G. Fu, R. Farmani and D. Butler, "A global analysis approach for investigating structural resilience in urban drainage systems" *Water Research*, vol. 81, pp. 15-26, 2015.

[15] S. Panguluri, W. Phillips and J. Cusimano, "Protecting water and wastewater infrastructure from cyber attacks" *Frontiers of Earth Science*, vol. 5, n. 4, pp. 406-413, 2011.

[16] M. Schoen, T. Hawkins, X. Xue, C. Ma, J. Garland and N. J. Ashbolt, "Technologic resilience assessment of coastal community water and wastewater service options" *Sustainability of Water Quality and Ecology*, vol. 6, pp. 75-87, 2015.

[17] D. Henry and J. E. Ramirez-Marquez, "Generic metrics and quantitative approaches for system resilience as a function of time" *Reliability Engineering \& System Safety*, vol. 99, pp. 114-122, 2012.

[18] J. P. Farwell and R. Rohozinski, "Stuxnet and the future of cyber war" *Survival*, vol. 53, n. 1, pp. 23-40, 2011.

[19] A. Armando and A. Coletta, “Security Monitoring for Industrial Control Systems” in *Security of Industrial Control Systems and Cyber Physical Systems - First Workshop, CyberICS 2015 and First Workshop, {WOS-CPS} 2015, Vienna, Austria, September 21-22, 2015, Revised Selected Papers*, 2015.