Resilience estimation of Cyber-Physical Systems via quantitative metrics

Michel Barbeau, Frederic Cuppens, Nora Cuppens, Romain Dagnas, Joaquin Garcia-alfaro

Resilience Estimation of Cyber-Physical Systems via Quantitative Metrics

MICHEL BARBEAU[1], (Member, IEEE), FRÉDÉRIC CUPPENS[2], (Member, IEEE), NORA CUPPENS[2], (Member, IEEE), ROMAIN DAGNAS[3], (Member, IEEE), AND JOAQUIN GARCIA-ALFARO[4], (Senior Member, IEEE)

[1]Department of Computer Science, Carleton University, Ottawa, ON K1S 5B6, Canada
[2]Polytechnique Montréal, Montréal, QC H3T 1J4, Canada
[3]Institut de Recherche Technologique SystemX, 91120 Palaiseau, France
[4]Institut Polytechnique de Paris, Télécom SudParis, 91000 Évry-Courcouronnes, France

Corresponding author: Joaquin Garcia-Alfaro (garcia_a@telecom-sudparis.eu)

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), and in part by the European Commission (H2020 SPARTA Project) under Grant 830892.

ABSTRACT This paper is about the estimation of the cyber-resilience of Cyber-Physical Systems (CPS). We define two new resilience estimation metrics: k-steerability and ℓ-monitorability. They aim at assisting designers to evaluate and increase the cyber-resilience of CPS when facing stealthy attacks. The k-steerability metric reflects the ability of a controller to act on individual plant state variables when, at least, k different groups of functionally diverse input signals may be processed. The ℓ-monitorability metric indicates the ability of a controller to monitor individual plant state variables with ℓ different groups of functionally diverse outputs. Paired together, the metrics lead to CPS reaching (k, ℓ)-resilience. When k and ℓ are both greater than one, a CPS can absorb and adapt to control-theoretic attacks manipulating input and output signals. We also relate the parameters k and ℓ to the recoverability of a system. We define recoverability strategies to mitigate the impact of perpetrated attacks. We show that the values of k and ℓ can be augmented by combining redundancy and diversity in hardware and software, in order to apply the moving target paradigm. We validate the approach via simulation and numeric results.

INDEX TERMS Cyber-physical systems, control theory, cyber-resilience, covert attacks, security metrics, attack remediation, recoverability, resilience estimation.

I. INTRODUCTION

Cyber-Physical Systems (CPS) integrate network and software resources to control and monitor physical components operating on different spatial and temporal scales [15]. Examples of CPS include industrial control systems for energy distribution (e.g., smart grids), autonomous vehicles, robotics and next-generation medical systems. Since physical, networked and computational components are deeply intertwined, the protection of the system as a whole highly relies on steerability and monitorability. Steerability refers to the ability of a controller to drive and maintain a Cyber-Physical System (CPS) in a desired operating point, by sending command and control input signals to the system actuators. Monitorability indicates the capability of the system to process and interpret output signals produced by the system sensors, in order to accurately deduce the internal state of the CPS. The controller follows reference signals such that the CPS ends being asymptotically stable. In the short term, slight variations of either the input or outputs of the system do not affect the stability. However, in the long term, such slight variations may affect and disrupt the system. This is the goal of a cyber-physical adversary. Given the knowledge of central controllers about the physical behavior of a CPS, i.e., steerability and monitorability, the goal of the defender is to face faults and attacks by increasing system recoverability, i.e., the system must be able to adapt and bounce back from stability disruptions, as quick as possible.

Recent studies acknowledge the vulnerability of CPS to integrity and availability attacks [2], [4], [23]. In this paper, we focus on covert attacks [28], [29], i.e., a family of cyber-physical attacks taking the form of physical aggressions against the operation of a CPS, by manipulating input signals to actuators and output signals from sensors. The approach is, however, valid for other family of integrity and availability attacks reported in the control-theoretic literature

The associate editor coordinating the review of this manuscript and approving it for publication was Azwirman Gusrialdi .

(being the family of covert attacks those reported as more ambitious to handle [29]).

In this paper, we introduce the notion of k-steerability. The parameter k corresponds to the minimum number of input signals available to act on each individual plant state variable. We also define the concept of ℓ-monitorability. The parameter ℓ reflects the minimum number of output signals that can be used to monitor each individual plant state variable. We study values of k and ℓ with respect to system resilience and the ability to recover a plant state. If due to covert attacks, h input signals are compromised, then steerability of each individual plant state is not entirely lost as long as h is lower than k. This partial steerability can be leveraged to run a covert attack mitigation plan. If due to covert attacks, g output signals are hacked, then the ability to detect the condition is not entirely lost as long as g is lower than ℓ. We discuss how k and ℓ can be determined and augmented by adding redundant and diverse hardware. k-steerability and ℓ-monitorability combine together into the (k, ℓ)-resilient metric. Our work is about resilience estimation. It is complementary to work on security risk assessment. In the sequel, we elaborate further on how to use them together. We assume that both input and output signals can be correlated into functionally diverse groups, as a complement to the traditional use of redundancy in critical systems.

The use of redundancy assume the inclusion of alternative copies, e.g., sensors, actuators and controllers, in order to guarantee system availability. If the system finds itself under a situation of attack, and the values of a group of components are not behaving as expected, then the validation of such values can be contrasted with the values of redundant replicas, assuming that there was an attack affecting the system. This technique is complementary to fault tolerance techniques, also used to address situations in which some system components are victims of failures or faults. However, the use of redundancy for security purposes may have some drawbacks. Since the replicas may be seen as identical, once an attacker has managed to compromise one of them, then the rest of the replicas can also be compromised very easily. Hence, we need to impose the use of diversity. For instance, if the replicas are geographically distributed, or the replicas compute their values using different physical phenomena, the approach can improve the way to handle attacks exploiting the physical nature of vulnerable components. Hence, our approach assumes the existence of different replicas behaving in an independent manner and with non-overlapping patterns (e.g., physical patterns) to handle the attacks.

In terms of contributions[1]:

- We propose a novel design stage approach to cyber-physical resilience thinking in terms of physical processing and intentional attacks.
- We introduce the steerability and monitorability measures paired together as the (k, ℓ)-resilience metric, k being the degree of steerability and ℓ the degree of monitorability.
- We relate (k, ℓ)-resilience metric to recoverabilty, i.e., the number of steps required to recover from attacks.
- We review example CPS and their resilience estimation.
- We validate the approach with numeric simulations and real world examples, including analysis of the proposed ideas under practical attacks with respect to performance disruption.

Sections II and III introduce CPS modeling, covert attacks and related work. The k-steerability, ℓ-monitorability and (k, ℓ)-resilient concepts are developed in Section IV. The design and evaluation of representative CPS that are (k, ℓ)-resilient are reviewed in Section V. Section VI concludes the paper. Additional details about our simulations and results are available in an appendix.

II. SYSTEM AND ADVERSARY MODELS

A CPS consists of a plant and a controller. They are distributed and communicate through a network. Several mathematical models exist for representing them [6], [21]. In the sequel, we introduce the necessary modeling background, using a CPS with fluid dynamics as an example.

A. DIFFERENTIAL EQUATION REPRESENTATION

Let us consider as a plant an individual cylindrical tank, with a single inflow and a single outflow of liquid. The tank liquid level can be modeled by the following differential equation [30]:

$$\alpha \frac{dh(t)}{dt} = F(t) - a\sqrt{h(t)} \tag{1}$$

Eq. (1) models the relationship between instantaneous changes of liquid level and difference between the inflow rate and outflow rate. As a function of time t (second), the level of the liquid in the tank is $h(t)$ (cm). Variable α represents a cross-sectional area of the tank (cm^2). The term $F(t)$ represents the inflow rate (cm^3/second). The parameter a denotes the outlet valve coefficient. The outflow rate (cm^3/second) is proportional to the product of a times the square root of the liquid level as at time t, represented by the term $h(t)$. Note that because of the square root term, the system in nonlinear.

The model represented by Eq. (1) is linearized assuming a linear inflow rate and operation around a liquid level h_0, termed the *operating point*. It is assumed the level is maintained at point $h_0 + \Delta$, with $|\Delta|$ small. Linearization is based on the observation that the expression $(1 + \epsilon)^\beta$ is approximately equal to the expression $1 + \beta\epsilon$, when $\epsilon \ll 1$. In the expression $a\sqrt{h(t)}$, substituting $h(t)$ by the sum $h_0 + \Delta$, we get a linear model for the outflow rate:

$$\begin{aligned} a\sqrt{h(t)} &= a\sqrt{h_0 + \Delta} = a\sqrt{h_0}\sqrt{1 + \Delta/h_0} \\ &\approx a\sqrt{h_0}\left(1 + \frac{\Delta}{2\,h_0}\right) \end{aligned}$$

[1] An early (short version) of this paper was presented and discussed at [1]. New material and discussions, including experimental work, have been added to this version.

Inflow rate $F(t)$ is modeled by the product $\gamma\kappa v$. The parameters γ, κ and v respectively denote the valve coefficient, pump coefficient (cm^3/V second) and voltage applied to the pump (V). The voltage v is the variable governed by the controller. The resulting linear differential equation modeling the liquid level is:

$$\alpha \frac{dh(t)}{dt} = \gamma\kappa v - a\sqrt{h_0}\left(1 + \frac{\Delta}{2h_0}\right) \tag{2}$$

Eq. (2) is an approximation that remains valid as long as Δ is relatively small, that is, at the chosen operating point h_0 there are only small level fluctuations. To maintain that condition, the inflow rate $\gamma\kappa v$ must be equal to the outflow rate $a\sqrt{h_0}$, with small fluctuation $-a\sqrt{h_0}\left(\frac{\Delta}{2\,h_0}\right) = -\frac{a\Delta}{2\sqrt{h_0}}$.

B. STATE SPACE REPRESENTATION

The state space representation of a linear CPS is as follows:

$$x_{i+1} = Ax_i + Bu_i + w_i \tag{3}$$

$$y_i = Cx_i + Du_i + v_i \tag{4}$$

Eq. (3) models the evolution of the CPS. At time i, given input u_i, state x_i is transformed into state x_{i+1}, where the index i is in $\mathbb{Z}^+$, state column vectors x_i and x_{i+1} are in $\mathcal{X} \subseteq \mathbb{R}^m$, input column vector u_i is in $\mathcal{U} \subseteq \mathbb{R}^p$, output column vector y_i is in $\mathcal{Y} \subseteq \mathbb{R}^n$ and dimensions m, n and p are in $\mathbb{Z}^+$. The transition may also be affected by random noise w_i, in $\mathbb{R}^m$. Eq. (4) represents the CPS input, state and output relation. At time i and in state x_i, the sensor measurements are y_i. The sensor measurements may be also affected by random noise represented by v_i, in $\mathbb{R}^n$. Matrices A, B, C and D are respectively called the state (m by m), input (m by p), output (n by m), and direct transmission (n by p) matrices.

For example, let us map Eq. (2) to a state-space representation. The input u_i is the voltage applied to the pump. Let t be the continuous time corresponding to the discrete time i. The state variable x_i tracks the difference between the liquid level $h(t)$ and operating point h_0, i.e., $x_i = h(t) - h_0$, which is the symbol Δ in Eq. (2). The corresponding state, input, output and direct transmission matrices are:

$$A = \left(-\frac{a}{\alpha 2\sqrt{h_0}}\right), \quad B = \left(\frac{\gamma\kappa}{\alpha}\right), \quad C = \left(1\right) \text{ and } \quad D = \left(0\right) \tag{5}$$

The state vector has one element $x_i[1]$, which is the current level difference Δ, w.r.t. the operating point h_0. The state matrix A contains one element, which is used to calculate in a transition, from time i to time $i+1$, the change in the amount of liquid leaving the tank, i.e., $\frac{a}{\alpha 2\sqrt{h_0}} \cdot x_i[1]$. Note that at the operating point, the total amount of liquid leaving the tank is the subtrahend in Eq. (2), divided by α. The input vector has a single element $u_i[1]$. The input matrix B contains one element and calculates the amount of liquid coming into the tank in one transition, i.e., the product $\frac{\gamma\kappa}{\alpha} \cdot u_i[1]$. Note that this mapping has the linearity advantage, but fidelity if limited to small fluctuations around an operating point. As we move from the operating, the effect of gravity is distorted. This degree of fidelity is although more than sufficient for the type of analysis conducted in the sequel of this paper.

Together, Equations 3, 4 and 5 model the dynamics of the plant. Integrated in a CPS, the input and output signals are transported over a network. It is reasonable to assume that input and output signals are protected with security protocols. It is also reasonable to expect that such protocols have vulnerabilities, initially unknown but eventually uncovered and exploited by an adversary, in a manner such as the one discussed in the upcoming Section II-C. As a second line of defense, it is also reasonable to believe that attack detection methods, such as the ones discussed in the upcoming Section III, are deployed. The CPS has attack protection and detection. However, this individual cylindrical tank CPS lacks alternative inputs and outputs that can be used to steer and monitor the plant when one input, one output or both are attacked. These backup inputs and outputs should ideally reflect different physical phenomena, protected by different security protocols with of course their own vulnerabilities, but possibly unlikely to be uncovered at the same time as for the main input and output. The systematic estimation of this type of resilience, which at the outset simply calls for common sense, is precisely the purpose of this article.

C. ADVERSARY MODEL

We assume adversaries perpetrating covert attacks. Covert attacks are a family of cyber-physical attacks in which the adversary perturbs the state of a CPS while succeeding to evade detection, i.e., the adversary attempts to remain invisible [32], [34], [37], [38]. It is powerful attack because it is assumed that the adversary knows the plant dynamics (matrices A, B, C and D) and that input and output signals can be spoofed. While an attack is being carried out, the perpetrator manipulates the measurements to conceal the effect of the spoofed inputs. Hence, from the point of view of an observer, responsible for detecting attacks, the measurements look normal. Using Eqs. (3) and (4), attacks are represented as follows:

$$x_{i+1} = Ax_i + B(u_i + u_i^a) + w_i \tag{6}$$

$$y_i = Cx_i + D(u_i + u_i^a) + v_i + s_i^a \tag{7}$$

The variable u_i^a, in $\mathcal{U}$, denotes the addition of the adversary to the signals to the actuators. The term s_i^a, in $\mathbb{R}^n$, represents the manipulation done by the adversary on the sensor measurements.

The adversary model succinctly captures covert attacks where an adversary has the ability to manipulate actuators and sensors. Attacks can be perpetrated by insiders, but also by outsiders due to communication channel vulnerabilities. For instance, an adversary infiltration was perpetrated on a steel mill using a spear phishing email tactic, first achieving access to the corporate network and then succeeding entering the plant network [16]. Generic covert attacks exploiting communication channel vulnerabilities have been modeled in numerous papers [29], [31], [36].

III. RELATED WORK

Methods have been devised to detect covert attacks. They all require the analysis of inputs and outputs of the plant. Rubio-Hernan *et al.* [24]–[26] have revisited challenge-response detectors via authentication techniques, initially proposed by Mo *et al.* [18], [19], [35]. Hoehn and Zhang [10] and Schellenberger and Zhang [27] developed the idea of external synthetic states that evolve in parallel and are coupled to the physical states of the CPS.

Adversaries can apply system identification [30] and machine learning [11], [30] to infer the dynamics of the plant. All detection methods acknowledge that the adversary has the ability to learn the dynamics of the CPS. However, they are all based on the important assumption that the knowledge of the adversary is not perfect. Due to this imperfect knowledge, the adversary makes errors that may be caught by the detection methods. Whether they are caught or not depends on the degree of knowledge of the adversary and the level of difficulty to avoid being detected. To make it challenging, detection methods comprise the integration of time-varying elements (inputs or states) concealed in the dynamics of the plant. Assuming the parameters of these elements are changed fast enough, the dynamics of the plant becomes a moving target for the adversary [13], [14]. In other words, the adversary does not have enough time to learn properly, makes errors and perpetrates attacks that are not covert [8], [9]. Next, we discuss in more details the concepts of challenge-response and auxiliary state.

1) CHALLENGE-RESPONSE AUTHENTICATION

Challenge-response detectors, defined in [24]–[26], revisit the authentication signal in [18], [19] to extend error detectors into cyber-physical attack detectors. The resulting scheme provides a real-time protection of the linear time-invariant models of the plant. Built upon *Kalman filters* and *linear-quadratic regulators*, the scheme produces authentication signals to protect the integrity of physical measurements communicated over the cyber and physical control space of a networked control system. It is assumed that, without the protection of the networked messages, malicious actions can be conducted to mislead the system towards unauthorized or improper actions, i.e., by disrupting the plant services.

Assume u_i^* as the output of a controller and u_i the control input that is sent to the plant, cf. Eq. (3). The idea of challenge-response authentication is to superpose to the control law u_i^* an authentication signal $\Delta u_i \in \mathbb{R}^p$ that serves to detect integrity attacks. Thus, the control input u_i is given by:

$$u_i = u_i^* + \Delta u_i \tag{8}$$

The authentication signal is a Gaussian random signal with zero mean that is independent both from the state noise (w_i) and measurement noise (v_i). The authentication signal is used by the detector to identify the malicious signals originated by the adversary. Since the control law u_i^* carries the authentication signal Δu_i, the detector (physically co-located within the controller) triggers an alarm whenever a malicious signal is observed, i.e., whenever the challenge sent by the controller over the plant is not observed within the measurements returned by the plant. Towards this end, [18], [19] propose to employ a χ^2 detector, i.e., a well-known category of real-time anomaly detectors classically used for fault detection in control systems [3], for the purpose of signaling the anomalies identified in the behavior of the plant.

Further details about some more powerful challenge-response detectors, capable of identifying adversaries which are empowered by identification tools such as ARX (autoregressive with exogenous input) and ARMAX (autoregressive-moving average with exogenous input) [20], i.e., using identification tools to evade detection, are available in [25], [26].

2) AUXILIARY STATES

The CPS can also be augmented with a synthetic auxiliary state, synthetic outputs and optionally new inputs [10], [27]. The auxiliary state has a linear time-varying dynamics that is evolved in parallel with the CPS. The dynamics is concealed to the adversary. Because it is time-varying, it becomes a moving target that is challenging to identify by an adversary, a precondition to the covert attack [7]. But, it is known to and used by the operator to detect the covert attack. The operator is in synchrony with the linear time-varying dynamics. It is therefore able to track it properly and compare the actual evolution of the auxiliary dynamics with the expected evolution. Significant discrepancies indicate the presence of anomalies, which can be used to identify the adversary.

The CPS model is extended with the auxiliary state $\tilde{x}_i$ and additional actuators and sensors ($\tilde{u}_i$ and $\tilde{y}_i$) related to the auxiliary state. The state x_i and auxiliary state $\tilde{x}_i$ are correlated. Together with the auxiliary state, the state transformation model is:

$$\begin{pmatrix} \tilde{x}_{i+1} \\ x_{i+1} \end{pmatrix} = \mathcal{A}_i \begin{pmatrix} \tilde{x}_i \\ x_i \end{pmatrix} + \mathcal{B}_i \begin{pmatrix} \tilde{u}_i \\ u_i \end{pmatrix} + \begin{pmatrix} \tilde{w}_i \\ w_i \end{pmatrix} \tag{9}$$

Together with the additional elements, the sensor measurements are:

$$\begin{pmatrix} \tilde{y}_i \\ y_i \end{pmatrix} = \mathcal{C}_i \begin{pmatrix} \tilde{x}_i \\ x_i \end{pmatrix} + \mathcal{D}_i \begin{pmatrix} \tilde{u}_i \\ u_i \end{pmatrix} + \begin{pmatrix} \tilde{v}_i \\ v_i \end{pmatrix} \tag{10}$$

with

$$\mathcal{A}_i = \begin{pmatrix} A_{1,i} & A_{2,i} \\ 0 & A \end{pmatrix}, \quad \mathcal{B}_i = \begin{pmatrix} B_i \\ B \end{pmatrix}, \quad \mathcal{C}_i = \begin{pmatrix} C_i & 0 \\ 0 & C \end{pmatrix} \text{ and } \mathcal{D}_i = \begin{pmatrix} D_i & 0 \\ 0 & D \end{pmatrix}.$$

Hidden to the adversary, the state sub-matrices $A_{1,i}$ and $A_{2,i}$, the input matrix B_i, output matrix C_i and direct transmission matrix D_i are randomized variables. According to the approach proposed by Schellenberger and Zhang [27], the actual matrices are randomly switched from time-to-time. The operator and CPS are synchronized on the switching sequence, perhaps through a switching signal. This secret

is not shared with the adversary. Sensor measurement $\tilde{y}_i$ is visible to the adversary, but changes over time in a random way. The adversary is challenged with learning the random auxiliary system state, input, output and direct transmission matrices.

We have introduced the system and adversary models and reviewed defense methods. In the next sections, we build upon that material and introduce new ideas to address resilience and state recovery.

IV. THE (k, ℓ)-RESILIENT PROPERTY

We define the k-steerability and ℓ-monitorability properties. In conjunction, they define the (k, ℓ)-resilient property.

A. INTER-VARIABLE DEPENDENCIES

To bright to light the dependency between two variables, we use Pearson correlation coefficients.

Definition 1 (Pearson correlation coefficient): Given two random variables, E and F, and n observations for each of them, their correlation coefficient is defined by

$$\rho(E, F) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{e_i - \mu_E}{\sigma_E} \right) \left(\frac{f_i - \mu_F}{\sigma_F} \right) \tag{11}$$

where $e_1, \ldots, e_n$ $(f_1, \ldots, f_n)$, μ_E (μ_F) and σ_E (σ_F) are the observations, mean and standard deviation of random variable E (F).

A correlation coefficient is a unitless value between minus one and one. When $\rho(E, F)$ is equal to one, we have perfect positive correlation between E and F. When it is minus one, we have perfect negative correlation. Intuitively, when $|\rho(E, F)|$ is between zero and 0.2, the linear correlation is from null to weak. It is moderate between 0.2 and 0.6. Above 0.6, it is strong [33]. Note that null linear correlation does not mean necessarily that variables E and F are independent. In such a case, there is no linear dependency revealed by the observations, but a nonlinear dependency is possible. For example, Eq. (1) generates nonlinear output correlated with the input. In such a case, existence of correlation can be confirmed calculating the correlation coefficient using a linearized version of the output data. Furthermore, correlation is one way to establish dependencies between variables.

B. DEPENDENCY GRAPH

Let u, x and y be respectively p-element, m-element and n-element column vectors representing the input, state and output variables of a CPS. We define correlation coefficient matrices to capture the relationships that exist between state variables and input or output variables.

Definition 2 (Input correlation coefficient matrix): The $m \times p$ input correlation coefficient matrix Q is equal to $(q_{i,j})$, where $i = 1, \ldots, m$, $j = 1, \ldots, p$. An entry $q_{i,j}$ is the correlation coefficient $\rho(x_i, u_j)$ between the state variable x_i and input variable u_j.

Definition 3 (Output correlation coefficient matrix): The $m \times n$ output correlation coefficient matrix R is equal to $(r_{i,j})$, where $i = 1, \ldots, m$, $j = 1, \ldots, n$. An entry $r_{i,j}$ is the correlation coefficient $\rho(x_i, y_j)$ between the state variable x_i and output variable y_j.

Definition 4 (Input dependency graph): The input dependency graph is a bipartite graph $G_U = (X, U, E)$ where the two sets of vertices are $X = \{x_1, \ldots, x_m\}$ and $U = \{u_1, \ldots, u_p\}$, the state and input variables. Pearson correlation is used to determine dependencies. There is an edge (x_i, u_i) in E if-and-only-if the absolute value of the correlation between variables x_i and u_i, i.e., $|q_{i,j}|$, is greater than or equal to a threshold T. Possible values for T are discussed in Section IV-A. In this article, we use strong correlation and a value of T close to one is chosen.

Definition 5 (Output dependency graph): The output dependency graph is a bipartite graph $G_Y = (X, Y, E)$ where the two sets of vertices are $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$, the state and output variables. There is an edge (x_i, y_i) in E if-and-only-if the absolute value of the correlation between variables x_i and y_i, i.e., $|r_{i,j}|$, is greater than or equal to a threshold T.

For the dependency graph G_U and a vertex x in X, let the expression $\deg(x)$ be its input degree, i.e., the number of adjacent vertices in U. Similarly, for the dependency graph G_Y and a vertex x in X, let $\deg(x)$ be its output degree, i.e., the number of adjacent vertices in Y. The ℓ-monitorability degree reflects the availability of at least ℓ sensor output signals for monitoring any state variable.

Definition 6 (ℓ-monitorability degree[2]): Let G_Y be the output dependency graph of a CPS. Let ℓ be equal to

$$\min_{x \in X} \deg(x).$$

Then, the CPS has ℓ-monitorability.

The notion of steerability is related to the control-theoretic concept of controllability. Controllability refers to the ability to drive a system to any state of its state space, under certain constraints [21]. This is consistent with our conceptualization of steerability, but the latter is a weaker and necessary condition emphasizing redundancy that can be evaluated calculating statistical correlation between state variables and inputs. The idea of monitorability is related the one of observability used in control theory. For example, in Ref. [6] a state variable is observable when it is connected to the outputs, which is consistent with our concept of monitorability. However, the exact techniques behind these two concepts are different and do not capture the same properties. In Ref. [6], observability is determined by computing the rank of an observability matrix. However, this particular technique is not universally recommended in the control literature for observability testing [22]. Our technique measures the correlations between state variables and outputs, intuitively, a necessary condition to ascertain connections between state variables or outputs. While steerability highlights redundancy in inputs, monitorability emphasizes redundancy in outputs.

We introduce the notion of k-steerability. It indicates that there are at least k actuator input signals available for acting on every single plant state variable.

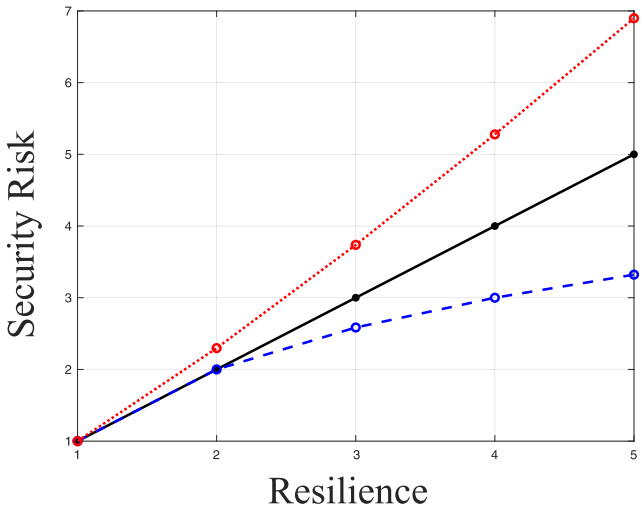


FIGURE 1. **Security risk versus resilience scenarios.**

Definition 7 (k-steerability degree[3]*):* Let G_U be the input dependency graph of a CPS. Let k be equal to

$$\min_{x \in X} \deg(x).$$

Then, the CPS has k-steerability.

Definition 8 ((k, ℓ)-resilient): A CPS with k-steerability and ℓ-monitorability is said to be (k, ℓ)-resilient.

The dependency graphs highlight the relationships that exist between the inputs and state variables, and relationships between state variables and outputs. This is essential to formally determine who can control what and who can monitor what. Besides, being (k, ℓ)-resilient means that the CPS can tolerate a maximum of $k - 1$ attacked actuators, while being able to act on every single state variable x_i, $i = 1, \ldots, m$. It also means that the CPS can withstand no more than $\ell - 1$ attacked sensors, while being able to monitor every single state variable x_i.

As in any system design exercise, there are several objectives that can conflict with each other. In the design of a CPS, security and resilience are two of them. While higher k and/or ℓ achieves higher resilience, this may also translate to more points where an adversary can try to control or monitor the plant, that is, the attack surface is augmented. In this article, we provide a methodology to estimate the resilience of a CPS. Complementing the work presented in this article, there are methodologies for CPS security risk assessment [17]. A fine balance between security risk and resilience must be achieved. Figure 1 schematically represents security risk versus resilience. The x-axis represents resilience. It can be quantified either with the resilience estimates k, ℓ or a weighted sum thereof. One can also envision a 3D model with one axis for k and another for ℓ, forming together a resilience estimation plane. In other words, resilience can examined from the point of view of the inputs, outputs or both at the same time. The y-axis represents security risk. It can be quantified with a risk assessment method [17]. Three scenarios are pictured with three different curves. The black solid line represents a case where security and resilience are in equilibrium. The red dotted line pictures the undesirable case where a growth in resilience implies a strong security risk increase. The blue dashed line shows the most desirable situation where a growth in resilience may imply a security risk increase, due to the augmentation of points where attacks can be perpetrated. Although, the increase is moderate and offset by significant growth in resilience.

We have established the principles of our approach. In the following section, we review a number of designs, explain how the (k, ℓ)-resilient property translates into possibilities of acting on and recovering the state of a plant when attacks are perpetrated. We define performance of a CPS design as the ability to maintain the plant in target state, despite the fact that there may be actuators and sensors being attacked. We compare performance of the different designs.

V. REVIEW OF (k, ℓ)-RESILIENT DESIGNS

The degree of steerability (k) of a CPS can be increased by introducing a diversity of new actuators. Adding more actuators increases the number of points for acting on a plant. Likewise, monitorability (ℓ) can be increased by introducing a diversity of new sensors. Adding new sensors provides more monitoring points for detecting anomalies and estimating the state of a plant. As discussed at the end of Section IV, we reiterate that increasing k, ℓ or both must be done in conjunction with security risk assessment. For a CPS with fluid dynamics, there is a diversity of sensor types that include flow rate, liquid level, turbidity, water leak, water pressure and gravity liquid level.

Making abstraction of noise for the sake of simplicity, we revisit the state-space representation of Eq. (5) augmented with an inflow rate sensor and an outflow rate sensor. The output column vector y comprises three entries: (1) the level difference (y_1), (2) inflow rate (y_2) and (3) outflow rate (y_3):

$$A = \left[-\frac{a}{\alpha 2\sqrt{h_0}}\right], \quad B = \left[\frac{\gamma k}{\alpha}\right], \quad C = \begin{bmatrix} 1 \\ 0 \\ \frac{a}{2\sqrt{h_0}} \end{bmatrix} \text{ and}$$

$$D = \begin{bmatrix} 0 \\ \gamma k \\ 0 \end{bmatrix} \tag{12}$$

The design comprises three outputs strongly correlated with the liquid level difference, the correlation is strong between the level difference state variable and any of the outputs. The CPS has three-monitorability, because in G_U, $\min \deg(x)$ for $x \in X$, is equal to three. With respect to Eq. (5), only the output matrix (C) and direct transmission matrix (D) have changed. When there are changes in the actuator configuration, the input matrix (B) needs to be modified. The plant dynamics, represented by the state matrix (A), does not change. Hereafter, we discuss a series of configurations, with increasing k and ℓ, i.e., increasing resilience estimation pairs. We review the different possibilities of state recovery according to their (k, ℓ)-resilient design.

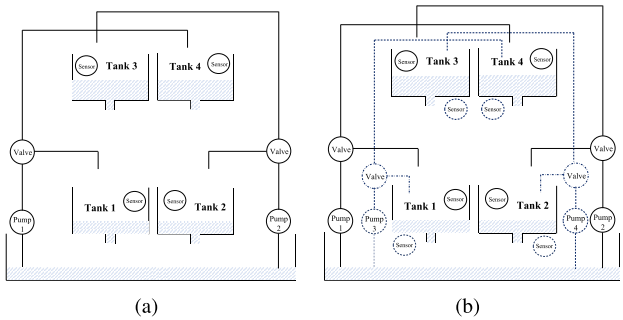


FIGURE 2. Quadruple-tank plant scenario. (a) Original scheme, based on Ref. [12]. (b) Extended (2, 2)-resilient scheme, with additional pumps and sensors.

A. SCENARIOS

We use the quadruple-tank plant of Johansson [12] as experimental testbed, cf. Fig. 2, Part (a), and supplementary material available in an online repository.[4] There are four tanks and two pumps. Each tank has an outlet at its bottom. Pump 1 pushes liquid into Tanks 1 and 4. Pump 2 pushes liquid into Tanks 2 and 3. Tank 3 is placed above Tank 1. By gravity, liquid from Tank 3 flows into Tank 1. Similarly, Tank 4 is placed above Tank 2. By gravity, liquid from Tank 4 flows into Tank 2. We examine three different designs for this CPS: (1, 1)-, (1, 2)- and (2, 2)-resilient.

(1,1)-resilient CPS — In this initial design, there are four ultrasonic sensors measuring the liquid level (one per tank) and two actuators (mechanic pumps) moving liquid into the tanks. Every pump has one liquid input and two outputs. The sensors and actuators are visible on the cyber space. The plant is observed and controlled from the cyber space. The state representation of the plant is as follows:

$$A = \begin{bmatrix} \frac{-a}{\alpha 2\sqrt{h_0}} & 0 & \frac{a}{\alpha 2\sqrt{h_0}} & 0 \\ 0 & \frac{-a}{\alpha 2\sqrt{h_0}} & 0 & \frac{a}{\alpha 2\sqrt{h_0}} \\ 0 & 0 & \frac{-a}{\alpha 2\sqrt{h_0}} & 0 \\ 0 & 0 & 0 & \frac{-a}{\alpha 2\sqrt{h_0}} \end{bmatrix},$$

$$B = \begin{bmatrix} \frac{\gamma_1 \kappa}{\alpha} & 0 \\ 0 & \frac{\gamma_2 \kappa}{\alpha} \\ 0 & \frac{(1-\gamma_2)\kappa}{\alpha} \\ \frac{(1-\gamma_1)\kappa}{\alpha} & 0 \end{bmatrix}, \quad C = I_4, \quad D = 0_{4,2} \tag{13}$$

State matrix A has four rows and four columns. The elements of row m, i.e., $a[m, \cdot]$ ($m = 1, 2, 3, 4$), determine the next value of the m-th state variable, i.e., $x_{i+1}[m]$. The individual element $a[m, n]$ of that row ($n = 1, 2, 3, 4$), determines the weight that the old state variable value $x_i[n]$ has in determining $x_{i+1}[m]$. In the input matrix B, γ_1 (γ_2) is the fraction of the liquid flow of Pump 1 (Pump 2) going to Tank 1 (Tank 2), $1 - \gamma_1$ ($1 - \gamma_2$) is going to Tank 4 (Tank 3). Since there are four level sensors, the output matrix C is the identity matrix of dimension four I_4. The direct transmission matrix D is a null matrix of dimension four by two $0_{4,2}$.

(1,2)-resilient CPS — The previous model is extended with outflow meters, one for each tank, see Fig. 2, Part (b). The output signals $y_i[2n - 1]$ and $y_i[2n]$ correspond to the level and outflow of Tank n ($n = 1, 2, 3, 4$). Output matrix C is augmented to represent readings of outflows from the tanks.

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{a}{\alpha 2\sqrt{h_0}} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{a}{\alpha 2\sqrt{h_0}} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{a}{\alpha 2\sqrt{h_0}} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & \frac{a}{\alpha 2\sqrt{h_0}} \end{bmatrix} \tag{14}$$

(2,2)-resilient CPS — This new design comprises new auxiliary actuators connected to fixed-flow Pumps 3 and 4. The fixed-flow pumps can take over the roles of Pumps 1 and 2, respectively. The input matrix of the plant is updated as follows:

$$B = \begin{bmatrix} \frac{\gamma_1 \kappa}{\alpha} & 0 & \frac{\eta_1 \lambda w_1}{\alpha} & 0 \\ 0 & \frac{\gamma_2 \kappa}{\alpha} & 0 & \frac{\eta_2 \lambda w_2}{\alpha} \\ 0 & \frac{(1-\gamma_2)\kappa}{\alpha} & 0 & \frac{(1-\eta_2)\lambda w_2}{\alpha} \\ \frac{(1-\gamma_1)\kappa}{\alpha} & 0 & \frac{(1-\eta_1)\lambda w_1}{\alpha} & 0 \end{bmatrix} \tag{15}$$

where η_1 (η_2) is the fraction of the liquid flow of Pump 3 (Pump 4) going to Tank 1 (Tank 2), $1 - \eta_1$ ($1 - \gamma_2$) is going to Tank 4 (Tank 3), λ and w_1 (w_2) respectively denote the Pump 3 (Pump 4) coefficient (cm^3/V second) and voltage (V), not controllable from the cyber space. For Pumps 3 and 4, the input signals are zero or one, corresponding to off and on. The input column vector u_i has now four rows. The first two rows are the input voltages to Pumps 1 and 2. The last two rows are the off/on (0/1) signals to Pumps 3 and 4. In the sequel, we bridge the (k, ℓ)−resilient property and plant state recoverability.

B. RESILIENCE AND STATE RECOVERABILITY

We connect (k, ℓ)-resilient estimation to behavioral properties. Building upon Refs. [5], [36], we quantify the resources needed to adapt and bounce back from disruptions. For the sake of simplicity, we make abstraction of noise. Firstly, we assume that only sensor attacks may occur. When attacks are perpetrated, we show that under certain conditions an increased number and a diversity of sensors make possible recovery of the state of a CPS. Secondly, we assume that both actuators and sensors can be attacked. While attacks are carried out, we demonstrate that it may be possible to identify which actuators are being attacked and how they are being attacked. If at all possible, these actuators can be deactivated.

[4]Cf. https://github.com/mirrored-quadruple-tank

Non-attacked actuators can be used to run a resilience plan that steers the CPS in a safe state.

1) ATTACKS ON SENSORS ONLY

Let x_i and x'_i be two states in $\mathbb{R}^m$, with corresponding length τ output sequences $y_i, \ldots, y_{i+\tau-1}$ and $y'_i, \ldots, y'_{i+\tau-1}$ resulting from the application of corresponding length τ input sequences $u_i, \ldots, u_{i+\tau-1}$ and $u'_i, \ldots, u'_{i+\tau-1}$.

Definition 9 (Recoverable state with sensor attacks): The state of a CPS is recoverable in τ steps, if for all states x_i and x'_i whenever the corresponding observed output sequences are such that $y_i = y'_i, \ldots, y_{i+\tau-1} = y'_{i+\tau-1}$, then x_i is equal to x'_i.

Theorem 1 (Recoverable state with sensor attacks): The state $x_i \in \mathbb{R}^m$ of an attacked CPS is recoverable in one step, if for $j = 1, \ldots, m$, there is at least one non-attacked sensor implementing an injective function with input state element $x_i[j]$.

Proof: Let C and D be the output and direct transmission matrices of the CPS. Let x_i and x'_i be two states. Because for $j = 1, \ldots, m$ at least one sensor implements an injective function, when Cx_iDx_i is equal to $Cx'_iDx'_i$ we have that x_i is equal to x'_i. Every state is uniquely determined by the sensor outputs. ■

A technique such as watermarking [36] can be used to determine which sensors are being attacked. Theorem 1 can be used to determine the exact state of CPS under attack.

Case 1: The state of the one-tank system modeled by Eq. (12) is recoverable in one step if only sensor level difference (y_1) or sensor outflow rate (y_3) is attacked, but not both.

Proof: It follows from the fact that both sensor types are injective functions with domain system states and co-domain length-one output traces. ■

Simulation of Case 1: The one-tank system has one level sensor and one outflow sensor. The state of this system is recoverable if the level sensor or the outflow sensor is attacked, but not both. Fig. 3 (a,b) shows that we can recover the state of the system from the outflow sensor, in case an attack is targeting the level sensor. The simulation is based on Matlab code, available on-line in a github repository.[5] Additional details about the simulation code and results are available in the appendix.

Case 2: The state of the (1, 1)-resilient system, cf. Eq. (13), is not recoverable if one sensor is attacked.

Proof: When one sensor is attacked, there are no additional points of observations (Fig. 3 (c,d)). ■

Simulation of Case 2: The (1, 1)-resilient system has only four levels sensors (one per tank). When an adversary perpetrates an attack on these sensors, the state of the system is not recoverable. Fig. 3 (c) shows the levels in each tank, when the system is not attacked. Fig. 3 (d) shows the levels when an attack is perpetrated. Since there is no non-attacked sensor type implementing an injective function on its elements, the state is not recoverable.

[5]Cf. https://github.com/mirrored-quadruple-tank/

Case 3: The state of the (1, 2)-resilient system, modeled by Eqs. (13) and (14), is recoverable in one step if, for $i = 1, 2, 3, 4$, only level sensors $y[2i-1]$ or outflow sensor $y[2i]$ is attacked, but not both (note, output column vector has format $(y[1], \ldots, y[n])^T$).

Proof: It follows from the fact that sensors $y[2i-1]$ and $y[2i]$ are injective functions of state component $x[i]$, for $i = 1, 2, 3, 4$ (note, state column vector has format $(x[1], \ldots, x[m])^T$). ■

Simulation of Case 3: Details available in the appendix.

2) ATTACKS ON ACTUATORS AND SENSORS

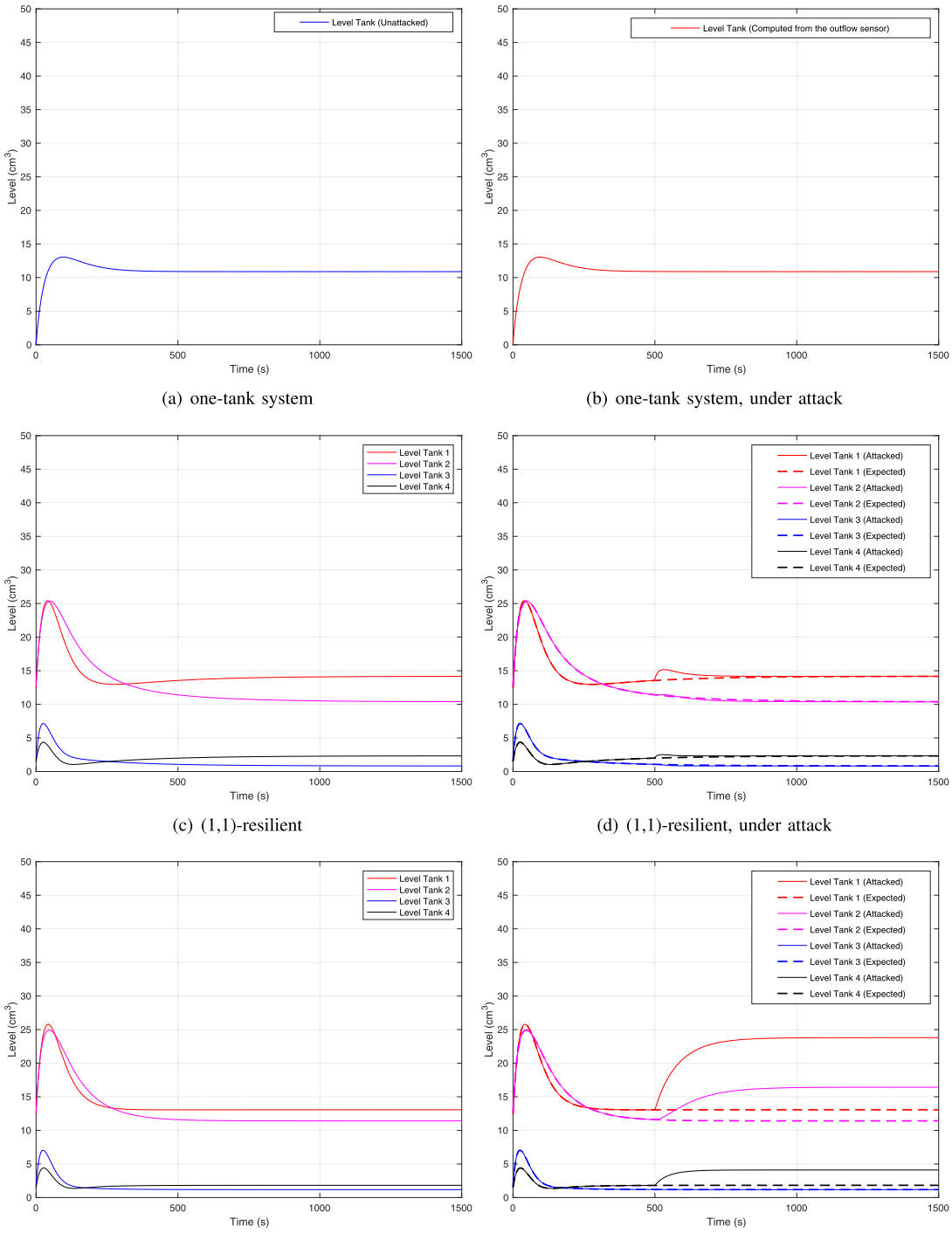
We now assume that both actuators and sensors can be disrupted by an adversary perpetrating a covert attack (cf. Section II-C). When actuators and sensors are attacked, and thanks to redundancy and diversity, it may be possible to determine the state of a CPS and which actuators are attacked. The current state can be recovered provided that the output sequence is unique, w.r.t. that state. Furthermore, we can find out which actuators are attacked and how they are attacked, provided that the output sequence is unique w.r.t. the input sequence. Hence, the entire state of the CPS is recoverable. The non-attacked actuators can be used to mitigate the attack and steer the CPS into a safe condition.

Definition 10 (Recoverable with actuator and sensor attacks): The state of a CPS is recoverable in τ steps, if for all states x_i and x'_i whenever the corresponding observed output sequences are such that $y_i = y'_i, \ldots, y_{i+\tau-1} = y'_{i+\tau-1}$, then x_i is equal to x'_i and input signals are such that $u_i = u'_i, \ldots, u_{i+\tau-1} = u'_{i+\tau-1}$.

Case 4: The state of the (2, 2)-resilient system, modeled by Eqs. (13), (14) and (15), is recoverable in one step when, for Tanks 1, 2, 3 and 4, the inflows are greater than zero, but respectively less than $\frac{\eta_1\lambda w_1}{\alpha}$, $\frac{\eta_2\lambda w_2}{\alpha}$ $\frac{(1-\eta_2)\lambda w_2}{\alpha}$ and $\frac{(1-\eta_1)\lambda w_1}{\alpha}$.

Proof: When the states x_i and x_{i+1} are recoverable according to Definition 9, the evaluation of the product Ax_i in Eq. (6) can be determined. Hence, the exact value of the product $B(u_i + u_i^a)$ can be resolved, i.e., the exact inflow for each tank, despite the presence of the adversary signal u_i^a on actuators. When for Tanks 1, 2, 3 and 4, the inflows are greater than zero, but respectively less than $\frac{\eta_1\lambda w_1}{\alpha}$, $\frac{\eta_2\lambda w_2}{\alpha}$ $\frac{(1-\eta_2)\lambda w_2}{\alpha}$ and $\frac{(1-\eta_1)\lambda w_1}{\alpha}$, there is no way to obtain such flows involving Pumps 3 or 4. It means that Pumps 1 or/and 2 have been functioning, but not Pumps 3 and 4. ■

Simulation of Case 4: Case 4 is simulated in Fig. 4. Inflows to Tanks 1, 2, 3 and 4 are shown by pump number. In Part (a), because they are variable flow, Pumps 1 and 2 can achieve inflows that are the same or below the inflows achievable by fixed-flow Pumps 3 and 4. When inflows are below what fixed-flow pumps can achieve, they can only be attributed to variable-flow pumps. When either Pumps 1 and 2 operate or Pumps 3 and 4 operate, we can tell which pair is involved. Discrimination is possible. Part (b) shows a condition where Pumps 1 and 2 are operated in ranges above what fixed-flow pumps can achieve. For example, an adversary adds voltages to signals and provokes inflow increases. Such a condition is



(a) one-tank system

(b) one-tank system, under attack

(c) (1,1)-resilient

(d) (1,1)-resilient, under attack

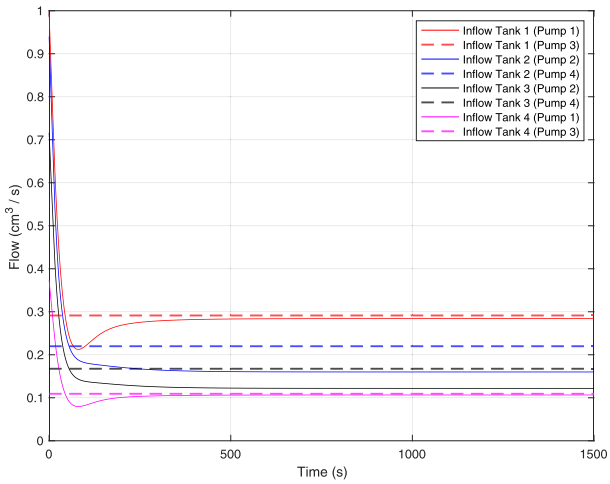
(e) (2,2)-resilient

(f) (2,2)-resilient, under attack

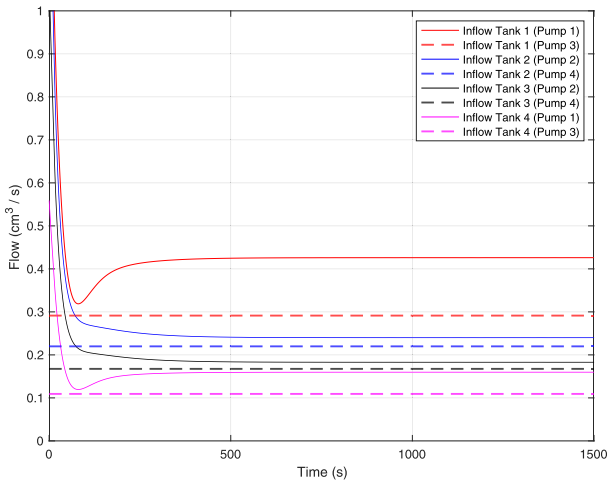
FIGURE 3. Simulation of Cases 1 and 2. Part (a) plots the level in a one tank system under normal operation (solid blue line). In Part (b), and assuming solely the ultrasonic sensor is attacked, it is possible to track the level using the outflow sensor (solid red line). In Part (c), tank levels are tracked with ultrasonic sensors in the (1, 1)-resilient system. In Part (d) an adversary spoofs actuators and manipulates sensor signals such that they look as expected (dashed lines), although actual levels (solid lines) are different. The degree of resilience does not enable state recovery. In Part (e), tanks levels are tracked with ultrasonic sensors in the (2, 2)-resilient system. In Part (f), an adversary spoofs actuators and manipulates solely ultrasonic sensor signals (dashed lines). Actual levels (solid lines) can be recovered using observations from outflow sensors.

achievable operating Pumps 1 and 2 alone, or also in combination with Pumps 3 and 4. For this example, discrimination might be impossible.

Fig. 5 (a) shows the input voltages u_1, and u_2, respectively applied to Pump 1 and Pump 2. The dashed line represents the attack signal used by an adversary. Fig. 5 (b) represents

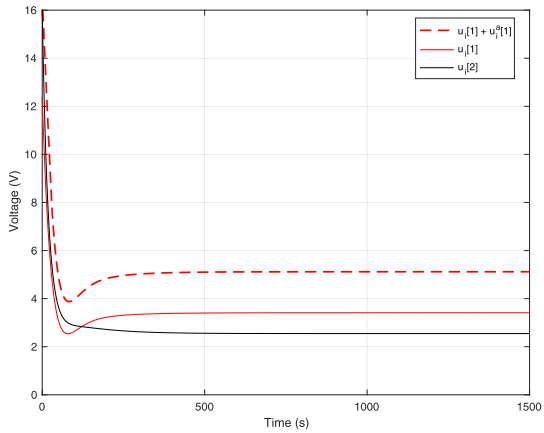


(a) (2,2)-resilient

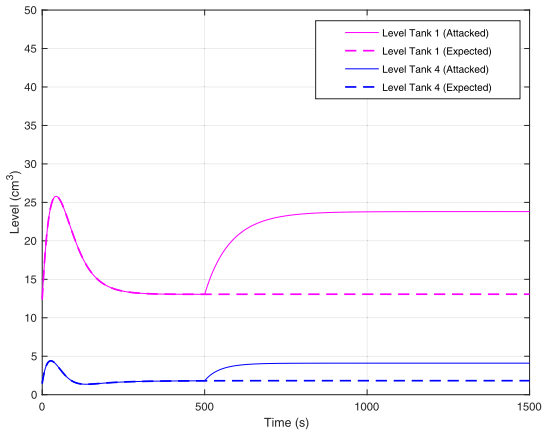


(b) (2,2)-resilient, under attack

FIGURE 4. Simulation of Case 4. Plots show inflows to Tanks 1, 2, 3 and 4, attributed to each pump. In Part (a), variable-flow pumps push liquid into tanks at rates below what fixed-flow pumps can do. In Part (b), variable-flow pumps push liquid into tanks at rates above what fixed-flow pumps can do.



(a) (2,2)-resilient



(b) (2,2)-resilient, under attack

FIGURE 5. Simulation results of Case 5. As a function of time, Part (a) shows values of input signals $u_i[1]$ (solid read), $u_i[2]$ (solid blue) and spoofed input signal $u_i^a[1]$ (dashed red). In Part (b), the adversary manipulates the ultrasonic sensor signal such that they look as expected (dashed) lines. Actual levels (solid lines) are recovered using the outflow sensors. Assuming inflows are below what fixed-flow pumps can achieved, it is possible to determine and track the values of the adversary signal $u_i^a[1]$.

the levels in Tank 1 and Tank 4, when the attack starts at $T = 500$ seconds.

Case 5: When the state x_i of the $(2, 2)$-resilient system, modeled by Eqs. (13), (14) and (15), is recoverable in one step and the action of the adversary on actuators can be determined resolving column vector u_i^a in the following equation:

$$x_{i+1} = Ax_i + BS(u_i + u_i^a) \tag{16}$$

with the selection matrix

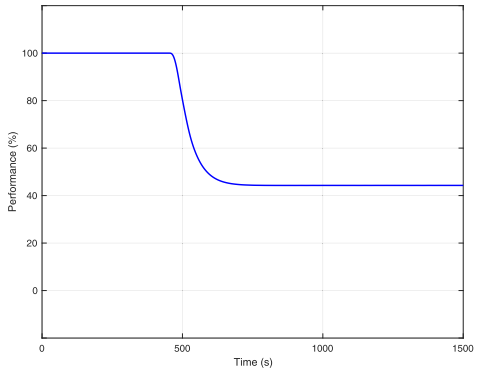
$$S = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Proof: By assumption, states x_i and x_{i+1} are recoverable. Since it is determined by the controller, the input column u_i is known. The effects of Pumps 3 and 4 have been excluded. The selection matrix S picks the inputs (voltages) applied to Pumps 1 and 2. It results into two equations, with one unknown in each of them, i.e., the adversary contributions to the actuator inputs $u_i^a[1]$ and $u_i^a[2]$. ■

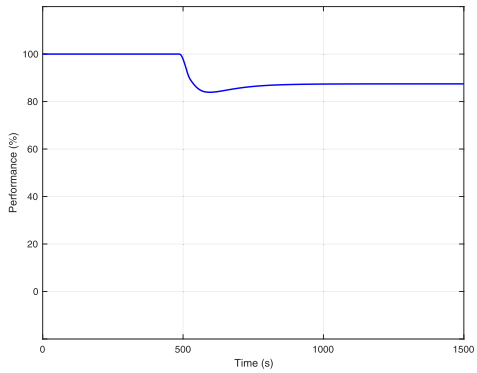
3) DISCUSSION

Fig. 6 provides an interpretation of all our simulations. We consider that the performance of a system is the capacity to maintain expected levels in tanks. Hence, the performance degradation corresponds to the deviation from the expected levels. The larger the deviation, the lower the performance. In Fig. 6, we represent these deviations in percentages.

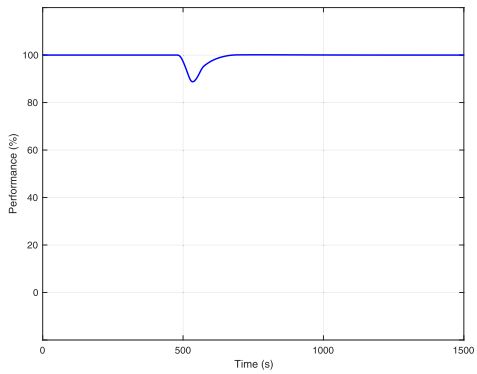
Figs. 6 (a), (b), and (c) respectively show the performance of the (1, 1)-, (1, 2)- and (2, 2)-resilient systems, when attacks are perpetrated. When a system is not attacked, performance is 100%. Attacks start at $T = 500$ seconds. The adversary manipulates inputs to drive more liquid in the Tanks 1 and 4. The consequence of the attack is a deviation from the expected system state. Quantifying this deviation,



(a) (1,1)-resilient system



(b) (1,2)-resilient system



(c) (2,2)-resilient system

FIGURE 6. Performance evolution of the (1, 1)-, (1, 2)- and (2, 2)-resilient systems, when they are confronted to a covert attack. Performance degradation corresponds to the deviation from their expected levels. The larger the deviation, the lower the performance. The (1, 1)-resilient system, with no recovery capability, experiences a performance drop. In contrast, the (1, 2)- and (2, 2)-resilient systems recover from the attack. The (1, 2)-resilient system recovers with graceful degradation, due to the absence of actuator redundancy, while the (2, 2)-resilient system fully recovers.

we obtain a percentage of performance loss. When the (1, 1)-resilient system (with no recovery capability) is under attack, it experiences a performance drop. In the (1, 2)- and (2, 2)-resilient systems, it is possible to mitigate the effects of attacks and bounce back. As shown in Figs. 6 (b) and (c), respectively, the (1, 2)-resilient system recovers with graceful degradation, due to the absence of actuator redundancy, while the (2, 2)-resilient system fully recovers.

VI. CONCLUSION

We have addressed covert attacks on CPS. We have defined the new k-steerability and ℓ-monitorability control-theoretic concepts. The k-steerability concept reflects the ability in a CPS to act on each of its individual plant state variables with at least k functionally diverse groups of input signals. In other words, it reflects the ability of the CPS to mitigate the impact of covert attacks when less than k groups of input signals are compromised, using static functional diversity. The ℓ-monitorability concept reflects the number of observations on each state variable of a CPS that can be used to identify covert attacks. Together, k-steerability and ℓ-monitorability determine the (k, ℓ)-resilient property of a CPS. If we assume that the detection process is conducted by combining strategies, such as redundancy and diversity in hardware and software techniques, the resulting (k, ℓ)-resilient concept applies the moving target paradigm, in which the CPS adapts itself to invalidate the acquired knowledge of the adversaries. We have validated our findings by conducting representative simulations. Future work will improve current results by applying dynamic functional diversity, e.g., by applying a functional diversity of components that will evolve over time.

APPENDIX. A. SUPPLEMENTARY MATERIAL TO THE SIMULATIONS

We report in this appendix the simulation of Case 3 (cf. Section V-B). An existing Matlab implementation of the quadruple-tank process for this case scenario (available online at https://github.com/karrocon/pcsmatlab), was adapted and complemented with Matlab and Simulink code, w.r.t. the resilience and adversary models presented in this paper. The resulting code is also available on-line, in our github repository (cf. https://github.com/mirrored-quadruple-tank/). The simulation of the Case 3, as in the Cases 1, 2, 4, and 5 (already reported in the main body of this paper, cf. Section V-B), implements a proportional-integral (PI) controller based on the differential equations of the quadruple-tank scenario By Johansson in Ref. [12].

Since the valves of the quadruple-tank scenario are not assumed vulnerable (e.g., we assume they cannot be attacked from the cyber space), we build the attacks assuming that the adversary is only taking control over the pumps (i.e., the adversary manages a remote access to the system, that allows manipulating the input voltages of the pumps acting as actuators of the quadruple-tank plant). Fig. 7 depicts the idea of our attack for both the original scheme of the quadruple-tank scenario in Ref. [12], and the extended (2, 2)-resilient scheme discussed in Section V-A. By attacking the voltage of the pumps, the adversary changes the inflow levels of the tanks. As depicted in Fig. 7, the adversary adds an attack signal to the input voltage of Pump 1. As a result of the attack, more liquid is pumped into Tanks 1 and 4.

According to the theorems defined in Section V-B, the adversary can also attack the sensors, in order to evade detection (i.e., by attacking both sensors and actuators, the adversary perpetrates a covert attack). The attack against

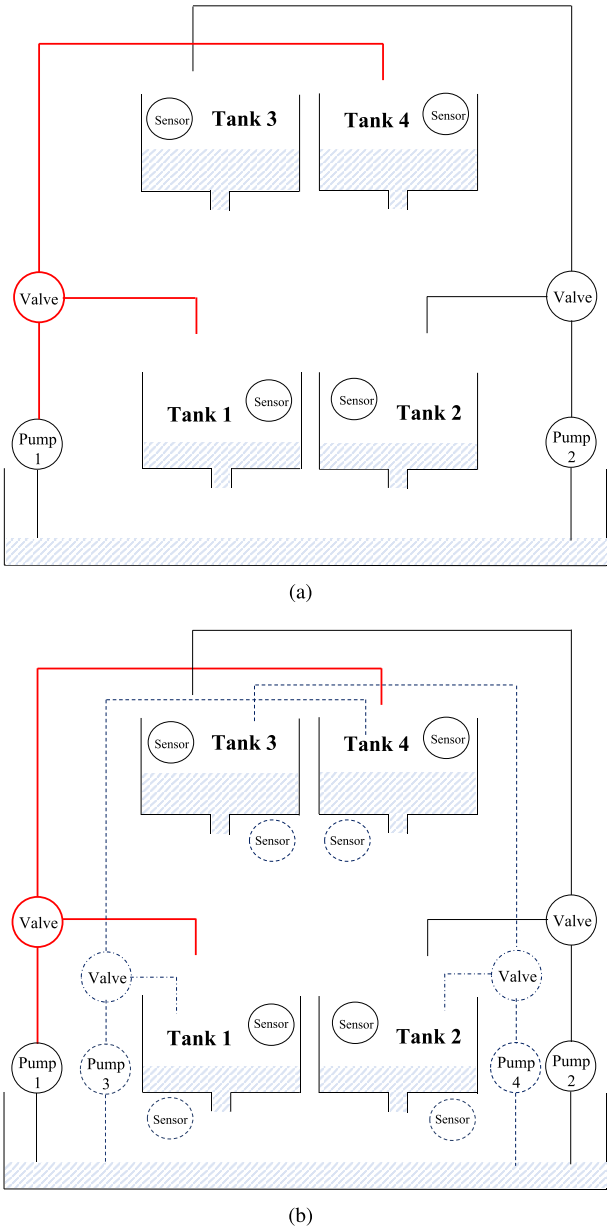


FIGURE 7. **Simplified representation of two representative attack scenarios. Red lines represent signals generated by the adversary. In Part (a), we assume an adversary perpetrating an attack against the original scheme in Ref. [12]. In Part (b), we assume an attack against the extended (2, 2)-resilient scheme discussed in Section V-A.**

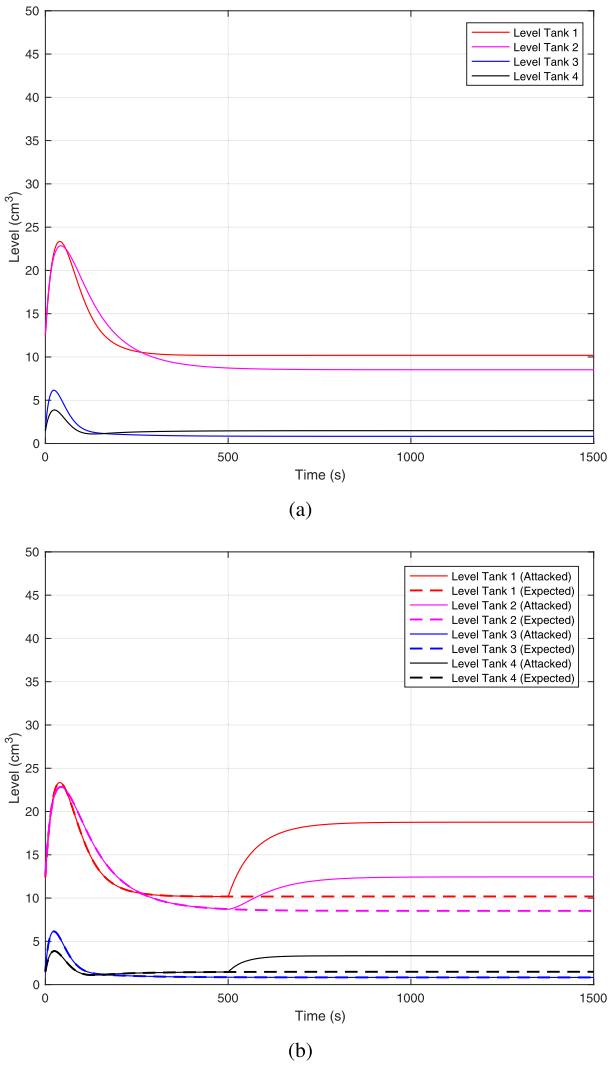


FIGURE 8. **Simulation results associated with Case 3 (cf. Section V-A), with regard to the (1,2)-resilient design. In Part (a,) we show the levels of the plant under normal operation (the ultrasonic level sensors are not under attack). In Part (b), attack mode, and assuming solely the ultrasonic sensor are attacked, we track the level using the outflow rate meters.**

the sensors consists to manipulate the measurement signals of the sensors, before reaching the controller (e.g., by means of injection, spoofing and man-in-the-middle cyber attacks, using a remote access from the cyber space). Hence, wrong measurements are provided to the controller, to conceal the detection of the attack against the actuators (i.e., the pumps). In fact, the measurement modification hides the real state of the system to the eyes of the controller. In our simulations, we can separate the processing of truthful signals, from those manipulated by the adversary. To ease the analysis, two simulations are conducted for each scenario, at the same time. The sensor signals of the second simulation are sent to the controller of the first simulation. Furthermore, and during the attack against the actuators, the adversary intercepts the truthful signals from the controller, and adds a modified input signal to the plant. This represents the disruption of the plant that is captured from the sensors of the system. Finally, the simulations assume that the attacked input voltage of the Pump 1 is increased by 50% w.r.t. its initial value, as shown in Fig. 5(a), in Section V-B. The consequence of this attack is depicted in Fig. 5(b) (also in Section V-B). As a result of the aforementioned attack simulations, with respect to the (1,2)-resilient system (cf. Section V-A), Fig. 8 shows the plant signals associated to the Case 3 (cf. Section V-B). The (1,2)-resilient system has eight sensors (four ultrasonic sensors and four outflow meters) and two actuators (Pumps 1 and 2). If only the ultrasonic level sensors (or only the outflow meters sensors) are attacked, then the state is recoverable. Fig. 8(a) shows the signals from the non-attacked

level sensors. When only one family of sensors is attacked (either the ultrasonic or the outflow meters ones), then we can appreciate the system can recover the state by using the non-attacked outflow sensors, as shown in Fig. 8(b). Notice that if we were conducting the full covert attack over the (2,2)-resilient system (cf. Fig. 7(b)), the controller will also be able to recover the system state, using the additional pumps (Pumps 3 and 4), in a more optimal way, as already indicated in the discussion of Section V-B (and shown in the interpretation of results included in Fig. 6 of Section V-B).

REFERENCES

[1] M. Barbeau, F. Cuppens, N. Cuppens, R. Dagnas, and J. Garcia-Alfaro, "Metrics to enhance the resilience of cyber-physical systems," in *Proc. IEEE 19th Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom)*, Jan. 2021, pp. 1167–1172.

[2] C. Barreto, A. A. Cárdenas, and N. Quijano, "Controllability of dynamical systems: Threat models and reactive security," in *Proc. Int. Conf. Decis. Game Theory Secur.* Fort Worth, TX, USA: Springer, Nov. 2013, pp. 45–64.

[3] B. Brumback and M. Srinath, "A chi-square test for fault-detection in Kalman filters," *IEEE Trans. Autom. Control*, vol. AC-32, no. 6, pp. 552–554, Jun. 1987.

[4] A. Chapman and M. Mesbahi, "Security and infiltration of networks: A structural controllability and observability perspective," in *Control of Cyber-Physical Systems*. Baltimore, MD, USA: Johns Hopkins Univ., 2013, pp. 143–160.

[5] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Trans. Autom. Control*, vol. 59, no. 6, pp. 1454–1467, Jun. 2014.

[6] G. Franklin, J. Da Powell, and A. Emami-Naeini, *Feedback Control of Dynamic Systems*. London, U.K.: Pearson, 2014.

[7] J. Giraldo, A. Cardenas, and R. G. Sanfelice, "A moving target defense to detect stealthy attacks in cyber-physical systems," in *Proc. Amer. Control Conf. (ACC)*, Jul. 2019, pp. 391–396.

[8] P. Griffioen, S. Weerakkody, and B. Sinopoli, "An optimal design of a moving target defense for attack detection in control systems," in *Proc. Amer. Control Conf. (ACC)*, Jul. 2019, pp. 4527–4534.

[9] P. Griffioen, S. Weerakkody, and B. Sinopoli, "A moving target defense for securing cyber-physical systems," *IEEE Trans. Autom. Control*, early access, Jun. 29, 2020, doi: 10.1109/TAC.2020.3005686.

[10] A. Hoehn and P. Zhang, "Detection of covert attacks and zero dynamics attacks in cyber-physical systems," in *Proc. Amer. Control Conf. (ACC)*, Jul. 2016, pp. 302–307.

[11] G. Horvath, "Neural networks in system identification," *Nato Sci. Sub III Comput. Syst. Sci.*, vol. 185, pp. 43–78, 2003.

[12] K. H. Johansson, "The quadruple-tank process: A multivariable laboratory process with an adjustable zero," *IEEE Trans. Control Syst. Technol.*, vol. 8, no. 3, pp. 456–465, May 2000.

[13] A. Kanellopoulos and K. G. Vamvoudakis, "Entropy-based proactive and reactive cyber-physical security," in *Proactive and Dynamic Network Defense*. Springer, 2019, pp. 59–83. [Online]. Available: https://link.springer.com/book/10.1007/978-3-030-10597-6

[14] A. Kanellopoulos and K. G. Vamvoudakis, "A moving target defense control framework for cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 65, no. 3, pp. 1029–1043, Mar. 2020.

[15] E. A. Lee, "Cyber-physical systems-are computing foundations adequate?" in *Proc. NSF Workshop Cyber-Phys. Syst.*, 2006, pp. 1–9.

[16] R. M. Lee, M. J. Assante, and T. Conway, "German steel mill cyber attack," SANS, Denver, CO, USA, Tech. Rep. 12, 2014.

[17] X. Lyu, Y. Ding, and S. Yang, "Safety and security risk assessment in cyber-physical systems," *IET Cyber-Phys. Syst., Theory Appl.*, vol. 4, no. 3, pp. 221–232, Sep. 2019.

[18] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on SCADA systems," *IEEE Trans. Control Syst. Technol.*, vol. 22, no. 4, pp. 1396–1407, Jul. 2014.

[19] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Proc. 47th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2009, pp. 911–918.

[20] H. G. Natke, "System identification: Torsten Söderström and Petre Stoica," *Automatica*, vol. 28, no. 5, pp. 1069–1071, 1992.

[21] K. Ogata, *Modern Control Engineering*. London, U.K.: Pearson, 2011.

[22] C. Paige, "Properties of numerical algorithms related to computing controllability," *IEEE Trans. Autom. Control*, vol. AC-26, no. 1, pp. 130–138, Feb. 1981.

[23] B. Ramasubramanian, M. A. Rajan, and M. G. Chandra, "Structural resilience of cyberphysical systems under attack," in *Proc. Amer. Control Conf. (ACC)*, Jul. 2016, pp. 283–289.

[24] J. Rubio-Hernan, L. De Cicco, and J. Garcia-Alfaro, "Revisiting a watermark-based detection scheme to handle cyber-physical attacks," in *Proc. 11th Int. Conf. Availability, Rel. Secur. (ARES)*, Aug. 2016, pp. 21–28.

[25] J. Rubio-Hernan, L. De Cicco, and J. Garcia-Alfaro, "Adaptive control-theoretic detection of integrity attacks against cyber-physical industrial systems," *Trans. Emerg. Telecommun. Technol.*, vol. 29, no. 7, p. e3209, Jul. 2018.

[26] J. Rubio-Hernan, L. De Cicco, and J. Garcia-Alfaro, "On the use of watermark-based schemes to detect cyber-physical attacks," *EURASIP J. Inf. Secur.*, vol. 2017, no. 1, pp. 1–25, Dec. 2017.

[27] C. Schellenberger and P. Zhang, "Detection of covert attacks on cyber-physical systems by extending the system dynamics with an auxiliary system," in *Proc. IEEE 56th Annu. Conf. Decis. Control (CDC)*, Dec. 2017, pp. 1374–1379.

[28] R. S. Smith, "A decoupled feedback structure for covertly appropriating networked control systems," *IFAC Proc. Volumes*, vol. 44, no. 1, pp. 90–95, Jan. 2011.

[29] R. S. Smith, "Covert misappropriation of networked control systems: Presenting a feedback structure," *IEEE Control Syst.*, vol. 35, no. 1, pp. 82–92, Feb. 2015.

[30] A. K. Tangirala, *Principles of System Identification: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2014.

[31] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, pp. 135–148, Jan. 2015.

[32] J. Tian, R. Tan, X. Guan, Z. Xu, and T. Liu, "Moving target defense approach to detecting stuxnet-like attacks," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 291–300, Jan. 2020.

[33] S. Tutorials. (Apr. 2020). *Pearson Correlation*. [Online]. Available: https://libguides.library.kent.edu/SPSS/PearsonCorr

[34] D. Umsonst, S. Saritas, and H. Sandberg, "A Nash equilibrium-based moving target defense against stealthy sensor attacks," in *Proc. 59th IEEE Conf. Decis. Control (CDC)*, Dec. 2020, pp. 3772–3778.

[35] S. Weerakkody, Y. Mo, and B. Sinopoli, "Detecting integrity attacks on control systems using robust physical watermarking," in *Proc. 53rd IEEE Conf. Decis. Control*, Dec. 2014, pp. 3757–3764.

[36] S. Weerakkody, O. Ozel, Y. Mo, and B. Sinopoli, "Resilient control in cyber-physical systems: Countering uncertainty, constraints, and adversarial behavior," *Found. Trends Syst. Control*, vol. 7, nos. 1–2, pp. 1–252, 2020.

[37] S. Weerakkody and B. Sinopoli, "Detecting integrity attacks on control systems using a moving target approach," in *Proc. 54th IEEE Conf. Decis. Control (CDC)*, Dec. 2015, pp. 5820–5826.

[38] S. Weerakkody and B. Sinopoli, "A moving target approach for identifying malicious sensors in control systems," in *Proc. 54th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2016, pp. 1149–1156.