Detection of replay attacks in CPSs using observer-based signature compensation

Carlos Trapiello[1,2], Damiano Rotondo[1,2], Helem Sanchez[1] and Vicenç Puig[1,2]

Abstract— **This paper presents a replay attack detection method that addresses the performance loss of watermarking-based approaches. The proposed method injects a sinusoidal signal that affects a subset, chosen at random, of the system outputs. The presence of the signal in each one of the outputs is estimated by means of independent observers and its effect is compensated in the control loop. When a system output is affected by a replay attack, the loss of feedback of the associated observer destabilizes the signal estimation, leading to an exponential increase of the estimation error up to a threshold, above which the estimated signal compensation in the control loop is disabled. This event triggers the detection of a replay attack over the output corresponding to the disrupted observer. The effectiveness of the method is demonstrated using results obtained with a quadruple-tank system simulator.**

I. INTRODUCTION

Ongoing advances in electronics and information technologies have improved the connection between computational and physical elements, allowing to shift from traditional networks to more complex cyber-physical systems (CPS), aiming for a better resource management. These systems integrate interconnected subsystems that interact through control, communication, and computation. Among the applications of CPSs, there are critical infrastructures like water and gas supply systems, smart grids or nuclear facilities, such that guaranteeing their safe operation has converted into a critical issue.

The advantages in efficiency and adaptability brought by CPSs, come at the price of new vulnerabilities and security weaknesses that could be exploited by a malicious attacker [1]. Examples like [2], [3] demonstrated the severe repercussions that attacks to CPSs could entail to society, attracting the attention of the scientific community, in an attempt to model, detect and repel possible attacks against CPSs. In the attack classification proposed in [4], attacks are characterized in a three dimensional attack space, depending on the attacker's a priori knowledge of the system model, and his/her access to the disruption and disclosure resources. Replay attacks can be easily placed within that framework, as according to [5], are used by an attacker who does not have knowledge about the system dynamics, apart from the fact that the system itself will be in steady state during the attack.

Standard replay attacks are modeled as a two phase attack: i) the attacker records data from the sensors without disturbing the system, and ii) the attacker replays back the recorded data to the monitor center, while conducting an attack on the physical system. In order to address this attack, two different strategies arise: the design of control strategies resilient to replay attacks, or the replay attack detection problem. However, the assumptions of limited energy of the attacker in the first strategy [6], [7], has brought the focus to the latter problem.

Among the most common methods for replay attack detection, there are the watermarking-based approaches, where an authentication signal is added to the control loop, at the cost of sacrificing the control performance. The used watermarking signals vary: [5] has used an independent and identically distributed (i.i.d.) Gaussian distribution to generate the signal; [8] has proposed to employ a periodic watermarking strategy; the watermark proposed by [9] aims at destabilizing the residual of the system, while preserving the stability of the main system; and in [10], a sinusoidal signal with a time-varying frequency is proposed as possible signature. Also, alternative methods that try to detect the replay attacks without injecting signals in the control input have been proposed (see [11] , [12]).

The main contribution of this paper is to present a novel strategy to detect replay attacks that addresses the problem of performance loss shared by state-of-the-art watermarking methods. This detection methodology can be framed within a state machine framework, in the sense that a mode switch from the *nominal mode* to a *system under replay attack mode* is triggered when some conditions are met. The detection method is developed for square systems, and it is based on injecting into the control loop a sinusoidal signal that affects a subset, chosen at random, of the system outputs. By means of a cascade of observers, each one fed independently with one different output, the presence of the injected signal is estimated and its effect on the control loop compensated. When a system output is affected by a replay attack, the loss of feedback of the associated observer destabilizes the estimation, boosted by the error induced after a change in the injected signal. The estimation

This work has been partially funded by the Spanish State Research Agency (AEI) and the European Regional Development Fund (ERFD) through the projects SCAV (ref. MINECO DPI2017-88403-R) and DEOCS (ref. MINECO DPI2016-76493) and AGAUR ACCIO RIS3CAT UTILITIES 4.0 – P7 SECUTIL. This work has been also supported by the AEI through the Maria de Maeztu Seal of Excellence to IRI (MDM-2016-0656) and the grant Juan de la Cierva-Formacion (FJCI-2016-29019).

[1] The authors are with the Research Center for Supervision, Safety and Automatic Control (CS2AC) of the Universitat Politècnica de Catalunya (UPC), Spain carlos.trapiello@upc.edu.

[2] The authors are also with the Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Llorens i Artigas 4-6, 08028 Barcelona, Spain.

error will increase exponentially up to a threshold, which is considered as the trigger event for the aforementioned mode switch, such that the signal compensation is disabled.

The remaining of the paper is structured as follows: Section 2 formulates the proposed detection method. In Section 3, the different modes are characterized. Section 4 presents the application to an example based on a quadruple-tank. Throughout Section 5, the results obtained in simulation are analyzed. Finally, the main conclusions are drawn in Section 6.

II. Replay attack detection

In this section, the replay attack detection method is formulated. We describe first the injection of the authentication signals and, afterwards, the signal estimation and compensation.

A. Injected signal

Let us consider a linear time-invariant (LTI) system:

$$\begin{aligned} \dot{x}_p(t) &= A_p x_p(t) + B_p u_p(t) \\ y_p(t) &= C_p x_p(t) \end{aligned} \tag{1}$$

where $x_p(t) \in \mathbb{R}^{n_x}$ is the vector of state variables, $u_p(t) \in \mathbb{R}^{n_u}$ is the control action, and A_p, B_p, C_p are matrices of appropriate dimensions. The vector $y_p(t) \in \mathbb{R}^{n_y}$, is the vector of measurements sent to a supervision center. A square system is assumed in this work, i.e. $n_u = n_y$. The system (1) is stabilized by means of a state-feedback control law

$$u_p(t) = -K x_p(t) \tag{2}$$

Hence, the closed-loop system behavior is modeled as

$$\dot{x}_p(t) = (A_p - B_p K) x_p(t) \tag{3}$$

A sinusoidal wave at a fixed frequency $d(t) = \rho^* \sin(\omega^* t)$, is injected in the control loop, such that it affects a subset of the system outputs chosen following a random pattern. In order to achieve this, a decoupler must be designed, such that the presence of the signal in a specific plant output is achieved by applying the signal in the corresponding decoupler input.

For the design of the aforementioned decoupler, the closed-loop steady state response to a persistent sinusoidal at frequency ω^* can be characterized by the system transfer matrix calculated at $s = j\omega^*$:

$$G_p(j\omega^*) = C_p (j\omega^* I - A_p + B_p K)^{-1} B_p \tag{4}$$

Then, the matrix G_d that input-output decouples the system in amplitude at the given frequency (ω^*), can be directly computed as

$$||G_p(j\omega^*)||\, G_d = I \;\rightarrow\; G_d = ||G_p(j\omega^*)||^{-1} \tag{5}$$

The inputs affected by the sinusoidal signal can be modeled as a pseudo-random generated binary code $z(t)$, with $z^i \in \{0,1\}$, whose values vary with a switching period of T_s. Hence, the elements $z^i(t)$, are piecewise functions that take binary values 0 or 1 at equally-spaced time instants $t_s^{(j)}$, $j \in \mathbb{N}$, with $t_s^{(0)} = 0$ and $t_s^{(j+1)} - t_s^{(j)} = T_s$:

$$z(t) = \left[z^1(t), \cdots, z^{n_y}(t) \right]^T \tag{6}$$

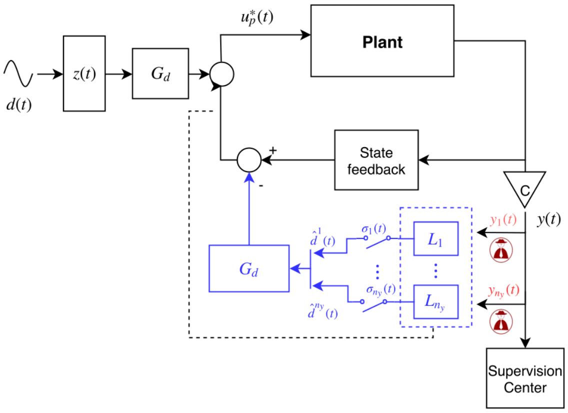

Fig. 1. Replay attack detection scheme

Knowing that the state space representation of a sinusoidal wave with frequency ω^* has the form

$$\begin{aligned} \dot{x}_d(t) &= A_d x_d(t) \\ d(t) &= C_d x_d(t) \end{aligned} \tag{7}$$

with

$$A_d = \begin{bmatrix} 0 & \omega^* \\ -\omega^* & 0 \end{bmatrix} \quad C_d = \begin{bmatrix} c_0 & c_1 \end{bmatrix} \tag{8}$$

then, the state space representation of the system plant affected by the injection of the sinusoidal signal (see Fig. 1), yields to the following switched system

$$\begin{bmatrix} \dot{x}_p(t) \\ \dot{x}_d(t) \end{bmatrix} = \begin{bmatrix} A_p - B_p K & B_p G_d z(t) C_d \\ 0 & A_d \end{bmatrix} \begin{bmatrix} x_p(t) \\ x_d(t) \end{bmatrix} \tag{9}$$

in which $z(t)$ acts as a switching signal.

B. Signal estimation

The presence of the injected signal $d(t)$ (see (7)) in each of the system outputs, is estimated analyzing each output independently. For this purpose, a cascade of state observers is designed such that the i^{th} observer is only fed by the system output $y_p^i(t)$. Thus, each observer is able to estimate only the trajectory in the subspace of the state-space which is observable with the considered system output, plus the sinusoidal signal in case it affects the output by means of the random code $z(t)$.

The input-output decoupling property, i.e. the fact that when the code $z(t)$ introduces the signal $d(t)$ through the i^{th} input of the decoupler G_d ($z^i(t) = 1$), this signal will be observable only through the i^{th} output of the plant, being impossible to estimate its states through any other output, follows from the same definition of observability [13].

According to the described procedure, in the design of the i^{th} observer, only the corresponding row (C_p^i) of the system output matrix C_p will be taken into consideration. Then, in the general case, the observability matrix $O_i = [A_p, C_p^i]$ will have $rank(O_i) = m_i < n_x$. A similarity transformation $\xi_i = T_i x_p$ can be performed [13], such that the m_i observable states, ξ_{oi} in the new base, are independent from the $(n_x - m_i)$ non-observable ones $\xi_{\bar{o}i}$. The proposed Kalman decomposition has the form

$$\begin{bmatrix}\dot{\xi}_{oi}\\ \dot{\xi}_{\bar{o}i}\end{bmatrix} = \underbrace{\begin{bmatrix}A_{oi} & 0\\ A_{21i} & A_{\bar{o}i}\end{bmatrix}}_{\tilde{A}_i}\begin{bmatrix}\xi_{oi}\\ \xi_{\bar{o}i}\end{bmatrix} + \underbrace{\begin{bmatrix}B_{oi}\\ B_{\bar{o}i}\end{bmatrix}}_{\tilde{B}_i} u_p$$
$$y = \underbrace{\begin{bmatrix}C_{oi} & 0\end{bmatrix}}_{\tilde{C}_i}\begin{bmatrix}\xi_{oi}\\ \xi_{\bar{o}i}\end{bmatrix} \qquad (10)$$

where

$$\tilde{A}_i = T_i A_p T_i^{-1} \quad \tilde{B}_i = T_i B_p \quad \tilde{C}_i = C_p^i T_i^{-1} \qquad (11)$$

Besides, taking into consideration the imposed input-output relationship of the sinusoidal signal $d(t)$, done by means of the matrix G_d, a vector t_i can be defined with the following form

$$t_i = \left[\delta_{1i}, \cdots, \delta_{ii}, \cdots, \delta_{n_y i}\right]^T \qquad (12)$$

where δ_{ij} represents the Kronecker delta. Thus, the i^{th} state observer fed with the $y_p^i(t)$ output, as shown in Fig. 1, has the form

$$\begin{bmatrix}\dot{\hat{\xi}}_{oi}\\ \dot{\hat{x}}_{di}\end{bmatrix} = \underbrace{\begin{bmatrix}A_{oi} & B_{oi}G_d t_i C_d\\ 0 & A_d\end{bmatrix}}_{A'_i}\begin{bmatrix}\hat{\xi}_{oi}\\ \hat{x}_{di}\end{bmatrix} + \underbrace{\begin{bmatrix}B_{oi}\\ 0\end{bmatrix}}_{B'_i} u + L_i(y_p^i - C_{oi}\hat{\xi}_{oi}) \quad (13)$$

By designing a matrix $L_i = [L_i^p \,|\, L_i^d]^T$ that stabilizes $(A'_i - L_i C'_i)$, then $\hat{d}^i(t) = C_d\hat{x}_{di} \rightarrow z^i(t)d(t) = z^i(t)C_d x_d(t)$.

C. Signal compensation

According to the previous discussion, after a transient stage caused by a change in the code $z(t)$ at each switching time $t_s^{(j)}$, the estimated set of signals $\hat{d}(t) = [\hat{d}^1(t), \cdots, \hat{d}^{n_y}(t)]^T$ will match the injected signal $z(t)d(t)$. Hence, the estimations can be injected into the control loop, in order to compensate for the effects of the injection of the sinusoidal signals ruled by $z(t)$.

As shown in Fig. 1, the initial control law is extended with the addition of the injected signal and the injection of the signal estimation after a decoupling block. Hence, by defining the error in the estimation of one of the signals as

$$e_i(t) = z^i(t)d(t) - \hat{d}^i(t) \qquad (14)$$

then, the new control action $u_p^*(t)$ is a function of the set of estimation errors $e(t) = [e_1(t), \cdots, e_{n_y}(t)]^T$:

$$u_p^*(t) = -Kx_p(t) + G_d z(t)d(t) - G_d\hat{d}(t) = -Kx_p(t) + G_d e(t) \qquad (15)$$

An estimation error $e_i(t)$ does not affect the estimation performed by the rest of observers as both $\hat{d}^i(t)$ and $z^i(t)d(t)$ are sinusoidal waves at frequency ω^* (when $z^i(t) = 1$). Thus, the error $e_i(t)$ is also a sinusoidal wave at frequency ω^* that enters the plant through the decoupler G_d (see (15)), and hence, only affects the output $y_p^i(t)$.

The direct injection in the control loop of a set of estimated signals $\hat{d}^i(t)$ as presented in (15), may destabilize the system when a subset of system outputs is under replay attack. In order to limit the aforementioned harmful effect, each estimated signal $\hat{d}^i(t)$ is only reinjected in the control loop if its estimation error is below a threshold α_i. Then, a set of vectors $p_i(t)$ can be defined as

$$p_i(t) = \left[\delta_{1i}\sigma_1(t), \cdots, \delta_{ii}\sigma_i(t), \cdots, \delta_{n_y i}\sigma_{ny}(t)\right]^T \qquad (16)$$

where

$$\sigma_i(t) = \begin{cases} 1, & \text{if } e_i(t) < \alpha_i. \\ 0, & \text{otherwise.} \end{cases} \qquad (17)$$

The introduction of the switching elements σ_i leads to define the control law affecting the plant as

$$u_p^*(t) = -Kx_p(t) + G_d z(t)d(t) - G_d\sum_{i=1}^{n_y} p_i(t)\hat{d}^i(t) \qquad (18)$$

that will match (15) when the compensations of all the estimated signals are enabled ($\sigma_i = 1,\ \forall i \in \{1, \ldots, n_y\}$).

III. System modes

In this section, the *nominal mode* and the *system under replay attack mode* are characterized. Also, the triggering event that indicates the transition from the first to the latter is studied.

A. Nominal mode

For the *nominal mode*, i.e. when the system is not under attack and all the estimated signals are compensated in the control loop ($\sigma_i = 1,\ \forall i \in \{1, ..., n_y\}$), the estimation errors defined in (14) tend to zero ($e(t) \rightarrow 0$). Hence, the system output is the same as the one provided by the closed loop system expressed by (3).

Taking into account the aforementioned estimation errors, the overall system dynamics for a plant with n_y outputs, can be expressed as

$$\begin{bmatrix} \dot{x}_p \\ \dot{x}_d \\ \dot{\hat{\xi}}_{o1} \\ \dot{e}_1 \\ \vdots \\ \dot{\hat{\xi}}_{on_y} \\ \dot{e}_{n_y} \end{bmatrix} = A^{\diamond} \begin{bmatrix} x_p \\ x_d \\ \hat{\xi}_{o1} \\ e_1 \\ \vdots \\ \hat{\xi}_{on_y} \\ e_{n_y} \end{bmatrix} \tag{19}$$

where the system matrix $A^{\diamond}$ has the form

$$A^{\diamond} = \begin{bmatrix} a_{11} & 0 & 0 & a_{1,k_1} & \dots & 0 & a_{1,k_{n_y}} \\ 0 & a_{22} & 0 & 0 & \dots & 0 & 0 \\ a_{j_1,1} & 0 & a_{j_1,j_1} & 0 & \dots & 0 & 0 \\ a_{k_1,1} & 0 & a_{k_1,j_1} & a_{k_1,k_1} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{j_{n_y},1} & 0 & 0 & 0 & \dots & a_{j_{n_y},j_{n_y}} & 0 \\ a_{k_{n_y},1} & 0 & 0 & 0 & \dots & a_{k_{n_y},j_{n_y}} & a_{k_{n_y},k_{n_y}} \end{bmatrix} \tag{20}$$

being $j_i = 2i+1$, $k_i = 2i+2$, and

$$\begin{aligned} a_{11} &= A_p - B_p K & a_{22} &= A_d \\ a_{1,k_i} &= B_p G_d t_i & a_{j_i,1} &= -B_{oi} K + L_i^p C_p^i \\ a_{j_i,j_i} &= A_{oi} - L_i^p C_{oi} & a_{k_i,1} &= -C_d L_i^d C_p^i \\ a_{k_i,j_i} &= C_d L_i^d C_{oi} & a_{k_i,k_i} &= C_d A_d \end{aligned} \tag{21}$$

that yields a stable system by construction.

B. Switching mode event

For the system starting in the *nominal mode*, i.e. with all the signal compensations enabled ($\sigma_i = 1$, $\forall i \in \{1,...,n_y\}$), if a malicious attacker replaces the measurements coming from the i^{th} output by a serial repetition of previously recorded measurements $\bar{y}_i$, then, the feedback loop that corrects the estimation error of the i^{th} observer is broken and the overall system is described by

$$\begin{bmatrix} \dot{x}_p \\ \dot{x}_d \\ \dot{\hat{\xi}}_{o1} \\ \dot{e}_1 \\ \vdots \\ \dot{\hat{\xi}}_{on_y} \\ \dot{e}_{n_y} \end{bmatrix} = A^{*}_{y_i} \begin{bmatrix} x_p \\ x_d \\ \hat{\xi}_{o1} \\ e_1 \\ \vdots \\ \hat{\xi}_{on_y} \\ e_{n_y} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ L_1^p \bar{y}_i \\ -C_d L_1^d \bar{y}_i \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{22}$$

where $A^{*}_{y_i}$ is obtained from $A^{\diamond}$ by replacing the terms $a_{j_i,1}$ and $a_{k_i,1}$ respectively with

$$a^{*}_{j_i,1} = -B_{oi} K \quad a^{*}_{k_i,1} = 0 \tag{23}$$

Designing the set of L_i observer gains such that they do not only stabilize each observer estimation error, but they also fulfill the condition that for the whole set $\{A^{*}_{y_i}, i \in [1,n_y]\}$ each system matrix is unstable, then, whenever a replay attack is being carried out over a system output, the corresponding disturbance estimation will become unstable, causing the propagation of the estimation error through the compensation loop. Hence, the corresponding estimation error $e_i(t)$ will increase up to reach the threshold defined in (17), leading to disable the i^{th} signal compensation, by setting $\sigma_i = 0$. This event will trigger the detection of a replay attack affecting the i^{th} output, and drives the system into the *system under replay attack mode*.

C. System under replay attack mode

The disconnection of the compensation of the i^{th} estimated signal, causes the rest of the system to become independent from the dynamics of the i^{th} observer. Hence, the dynamics of the system in this mode are ruled by a new state matrix $A^{\star}_{a_i}$ obtained from (20) by erasing the j_i and k_i rows and columns and replacing the element a_{12} with

$$a^{\star}_{12} = B^p G_d t_i \tag{24}$$

due to the presence of the non-compensated signal $z_i(t)d(t)$ affecting the i^{th} system output.

In this mode, an appropriate set of countermeasures for protecting the system against other malicious actions masked behind a replay attack (see [4]), should be developed, although this goes beyond the scope and goal of this paper. Note that the i^{th} observer could still be used in order to detect the end of the replay attack.

IV. Motivating example

In this section, the application of detection method presented in the previous sections is illustrated by considering a quadruple-tank process controlled through a wireless communication network (see [14]).

A. Quadruple-tank process

The plant model is given by [15].

$$\begin{aligned} \frac{dh_1}{dt} &= -\frac{a_1}{A_1}\sqrt{2gh_1} + \frac{a_3}{A_1}\sqrt{2gh_3} + \frac{\gamma_1 k_1}{A_1} v_1 \\ \frac{dh_2}{dt} &= -\frac{a_2}{A_2}\sqrt{2gh_2} + \frac{a_4}{A_2}\sqrt{2gh_4} + \frac{\gamma_2 k_2}{A_2} v_2 \\ \frac{dh_3}{dt} &= -\frac{a_3}{A_3}\sqrt{2gh_3} + \frac{(1-\gamma_2)k_2}{A_3} v_2 \\ \frac{dh_4}{dt} &= -\frac{a_4}{A_4}\sqrt{2gh_4} + \frac{(1-\gamma_1)k_1}{A_4} v_1 \end{aligned} \tag{25}$$

where h_i are the heights of water in each tank, A_i and a_i the cross-section areas of the tanks and the outlet holes respectively, k_i the pump constants, γ_i the flow ratios and g the gravity. The process inputs, are v_1 and v_2, the input voltages to the pumps. The list and values of the model parameters are given in Table I.

The previous nonlinear model, can be represented by a fourth order state space model, choosing the tank liquid levels h_i as the state variables x_i, $i = 1,\ldots,4$. The system is linearized for the steady-state equilibrium point $x_e = [x_{1e}, x_{2e}, x_{3e}, x_{4e}]^T = [12.26, 12.78, 1.63, 1,41]$ [cm], reached

TABLE I

MODEL PARAMETERS

Parameter	Value	Units
A_1, A_3	28	cm^2
A_2, A_4	32	cm^2
a_1, a_3	0.071	cm^2
a_2, a_4	0.057	cm^2
g	981	cm/s^2
k_1, k_2	$3.33, 3.35$	cm^3/V$_s$
γ_1, γ_2	$0.7, 0.6$	

by applying the constant input voltages $u_e = [v_{1e}, v_{2e}]^T = [3,3]^T$ [V]. Considering the deviations of the state and input from the equilibrium point Δx and Δu, the linearized system can be expressed as

$$\Delta\dot{x} = A\Delta x + B\Delta u \tag{26}$$

with

$$A = \begin{bmatrix} -0.016 & 0 & 0.044 & 0 \\ 0 & -0.011 & 0 & 0.033 \\ 0 & 0 & -0.044 & 0 \\ 0 & 0 & 0 & -0.033 \end{bmatrix} \quad B = \begin{bmatrix} 0.083 & 0 \\ 0 & 0.063 \\ 0 & 0.048 \\ 0.032 & 0 \end{bmatrix} \tag{27}$$

and a state feedback control law $\Delta u = -K\Delta x$ with the controller gain K

$$K = \begin{bmatrix} 36.16 & -36.21 & 46.79 & -79.07 \\ -94.8154 & 130.22 & -155.08 & 251.54 \end{bmatrix} \tag{28}$$

Throughout the rest of the paper, it is assumed that the liquid level of the two bottom tanks is monitored by a supervision station, leading to an output matrix

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{29}$$

B. Detector specifications

The specification of the injected signal $d(t) = \rho^* \sin(\omega^* t)$ in the application example is

$$\rho^* = 0.02 \qquad \omega^* = 10\pi\, rad/s \tag{30}$$

For the above frequency, the amplitude decoupling matrix G_d is

$$G_d = \begin{bmatrix} 377.42 & -0.97 \\ -0.69 & 500.28 \end{bmatrix} \tag{31}$$

For each of the two outputs defined by (29), an observer is designed according to (13), being the obtained gains

$$L_1 = \begin{bmatrix} 1.14 \\ -16.70 \\ 31.09 \\ 31.77 \end{bmatrix} \quad L_2 = \begin{bmatrix} 1.15 \\ 1.37 \\ 31.19 \\ 31.71 \end{bmatrix} \tag{32}$$

such that L_1 and L_2 fulfill the condition of destabilizing the matrices $A^*_{y_1}, A^*_{y_2}$ defined in Section III-B.

The considered switching period is $T_s = 150s$. The thresholds for the mode switch from *nominal* to *system under replay attack* affecting the i^{th} output are chosen as

$$\alpha_1 = \alpha_2 = 0.05 \tag{33}$$

V. SIMULATION RESULTS

In order to assess the effectiveness of the proposed method, two different simulation scenarios are considered.

A. Scenario I - Nominal mode

The first scenario simulates the system behavior in the *nominal mode*. Fig. 2 shows the injected signal $z(t)d(t)$ altogether with the corresponding estimation error that is provided by each one of the designed observers.

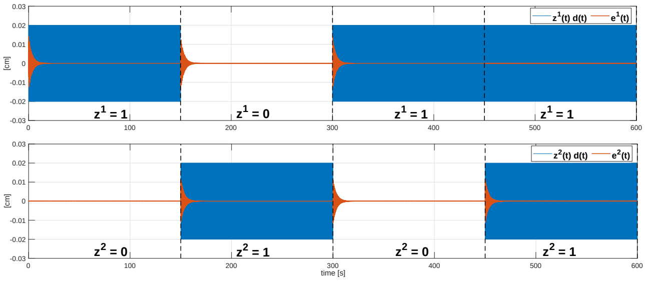

Fig. 2. System outputs - nominal mode

During the *nominal mode*, both estimated signals are compensated in the control loop ($\sigma_1 = 1, \sigma_2 = 1$). As presented in Fig. 3, the effect in the system outputs of the injected signals $z(t)d(t)$, is compensated after a transient stage induced by a change in $z(t)$.

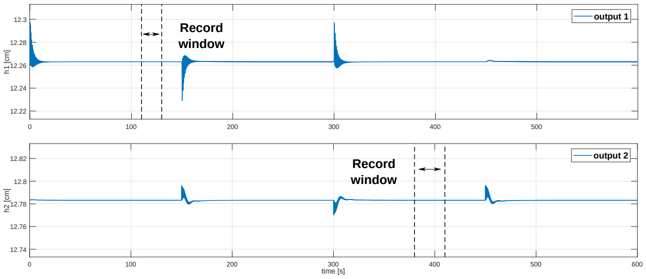

Fig. 3. System outputs - nominal mode

For the system in the *nominal mode*, it was assumed that a malicious attacker recorded intervals of system measurements during the steady state. The considered record windows are (see Fig. 3): $\bar{y}_1$ constitutes a recording of output 1 from $t^1_0 = 110s$ to $t^1_f = 130s$, and $\bar{y}_2$ constitutes a recording of output 2 from $t^2_0 = 380s$ to $t^2_f = 410s$.

B. Scenario II - System under attack

In this scenario, a replay attack against the different system outputs is simulated. The interval $\bar{y}_1$ is repeatedly replayed during the time interval $t \in [680s, 780s]$ replacing the real measurements of the first output, while the interval

$\bar{y}_2$ is replayed in the time interval $t \in [870s, 990s]$ instead of the second output measurements.

Figs. 4-5 show how for both outputs, right after the replay attack is performed, the loss of feedback of the associated observer implies that the corresponding estimation error increases up to reach the defined thresholds, instant when the signal compensation is disabled ($\sigma_i = 0$), and hence, the replay attack detected.

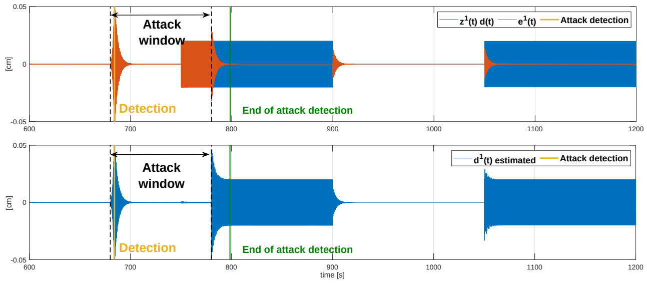

Fig. 4. Injected / estimated signal - output 1

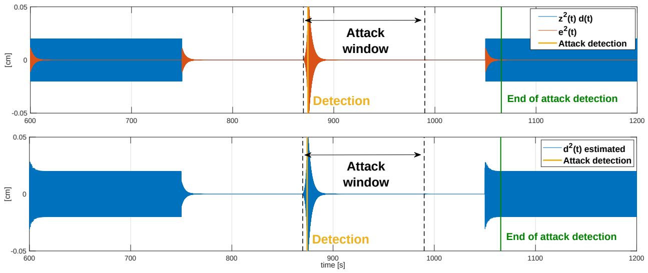

Fig. 5. Injected / estimated signal - output 2

The random presence of the injected signals imposed by $z(t)$, ensures that, under replay attack, an initial estimation error will exist after each switch in the conditions, triggering the divergence of the estimation error associated to the attacked output. Nevertheless, the presence of small errors even in the steady state, causes the immediate increase in the corresponding estimation error seen in Figs. 4-5.

As commented in Section III-C, when the system is in the *system under replay attack mode* affecting the i^{th} output, the capability of correctly estimating the injected signal $z^i(t)d(t)$ (when $z^i(t) = 1$) by the associated observer, could be used in order to detect the end of the attack. Figs. 4-5, also include (in green) the obtained end of attack detection times, computed when the corresponding estimation error $e^i(t) \approx 0$ for $z^i(t) = 1$.

The mean absolute tracking errors (MAE) in the *nominal mode* presented in the first scenario, and the detection times obtained for the proposed attacks, are shown in Table II.

TABLE II
PERFORMANCE

	MAE (cm)	**Attack** (s)	**Attack detection** (s)
output 1	$2.62 \cdot 10^{-4}$	$[680, 780]$	684.16
output 2	$2.36 \cdot 10^{-4}$	$[870, 990]$	874.39

VI. Conclusions

This work has introduced a novel method for detecting replay attacks as a consequence of the feedback loss of the observer associated to the output under attack, proving its effectiveness in a quadruple-tank system. Simulations have shown how in the *nominal mode*, the compensation of the estimated signal in the control loop reduces the performance loss of the plant. On the other hand, whenever a system output is under attack, the estimation error ends up diverging due to the appropriate choice of observer gains. The design of the observer gains such that the previous condition of instability is achieved, is a future research direction.

References

[1] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[2] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Security & Privacy*, vol. 9, no. 3, pp. 49–51, 2011.

[3] G. Liang, S. R. Weller, J. Zhao, F. Luo, and Z. Y. Dong, "The 2015 ukraine blackout: Implications for false data injection attacks," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 3317–3318, 2017.

[4] A. Teixeira, D. Pérez, H. Sandberg, and K. H. Johansson, "Attack models and scenarios for networked control systems," in *Proceedings of the 1st international conference on High Confidence Networked Systems*. ACM, 2012, pp. 55–64.

[5] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*. IEEE, 2009, pp. 911–918.

[6] M. Zhu and S. Martínez, "On the performance analysis of resilient networked control systems under replay attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 804–808, 2014.

[7] B. Chen, D. W. Ho, G. Hu, and L. Yu, "Secure fusion estimation for bandwidth constrained cyber-physical systems under replay attacks," *IEEE transactions on cybernetics*, vol. 48, no. 6, pp. 1862–1876, 2018.

[8] C. Fang, Y. Qi, P. Cheng, and W. X. Zheng, "Cost-effective watermark based detector for replay attacks on cyber-physical systems," in *Control Conference (ASCC), 2017 11th Asian*. IEEE, 2017, pp. 940–945.

[9] A. Khazraei, H. Kebriaei, and F. R. Salmasi, "A new watermarking approach for replay attack detection in lqg systems," in *Decision and Control (CDC), 2017 IEEE 56th Annual Conference on*. IEEE, 2017, pp. 5143–5148.

[10] H. S. Sánchez, D. Rotondo, T. Escobet, V. Puig, J. Saludes, and J. Quevedo, "Detection of replay attacks in cyber-physical systems using a frequency-based signature," *Journal of the Franklin Institute*, 2019.

[11] B. Tang, L. D. Alvergue, and G. Gu, "Secure networked control systems against replay attacks without injecting authentication noise," in *American Control Conference (ACC), 2015*. IEEE, 2015, pp. 6028–6033.

[12] A. Hoehn and P. Zhang, "Detection of replay attacks in cyber-physical systems," in *American Control Conference (ACC), 2016*. IEEE, 2016, pp. 290–295.

[13] K. J. Åström and B. Wittenmark, *Computer-controlled systems: theory and design*. Courier Corporation, 2013.

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

[15] K. H. Johansson, "The quadruple-tank process: A multivariable laboratory process with an adjustable zero," *IEEE Transactions on control systems technology*, vol. 8, no. 3, pp. 456–465, 2000.