Time Evolution Pattern Analysis for Cyber Attack Detection in a Two-tank Benchmark

Joseba Quevedo¹, Helem S. Sánchez¹, Damiano Rotondo¹,², Teresa Escobet¹ and Vicenç Puig¹,²

Abstract—This paper presents some results related to the detection and isolation of cyber attacks in a recently proposed benchmark based on a two-tank system. The benchmark proposes some attack scenarios in which a malicious attacker alters the signals of the water level sensors in the tanks, in order to remain hidden while stealing water. This paper shows the difficulty to detect cyber attacks based only on the model-based residuals calculated using the measured variables. On the other hand, by using the time evolution pattern analysis of the measured sensors, it becomes possible to detect some of these cyber attacks.

I. INTRODUCTION

In recent years, the increasing integration between control, communication and computation (the so-called triple C) has provided the ability for large numbers of interconnected sensors, actuators and computational units to interact with the physical environment [1]. The merging of cyber elements with physical processes has led to investigate a new class of systems, referred to as cyber-physical systems (CPSs) [2]. CPSs are characterized by a higher efficiency, but also by bigger vulnerabilities, which can be exploited by a malicious agent in order to perform cyber attacks, which might result in critical damage or economical loss [3], [4]. Cyber attacks are different from faults due to the fact that they do not affect only the physical layer of the CPS, but the cyber one as well. In order to make a control system resilient in face of such attacks, attack detection and secure control techniques must be developed [5], [6].

These attacks, usually motivated by terrorism, criminality or sabotage, exploit the system’s vulnerabilities and result in some kind of disturbance or damage in the physical and cyber layers. The interconnected nature of Industry 4.0-driven operations means that cyber attacks have far more extensive effects than ever before, and digital control systems, computers and their supply networks may not be prepared for this kind of risks [7], [8].

In [9], a benchmark based on a two-tank interconnected system was proposed for testing different schemes for detection and isolation of cyber attacks. The benchmark was obtained from a previously proposed fault diagnosis benchmark [10]–[12] by incorporating a malicious attacker who wants to steal water from the tanks while remaining hidden through an appropriate alteration of the measurements coming from the level sensors of the tanks.

This paper presents the application of the classical fault diagnosis approach based on analytical redundancy relations (ARRs) [13], and how an analysis of the time evolution pattern of the measured sensors can be used to improve the ability to detect cyber attacks. ARRs are analytical expressions in terms of the system’s input, output and their derivatives, which allow testing the consistency of the measured signal with the nominal model of the system. Their generation is a problem that has attracted a lot of attention since the late 1990s [14], and it is still of interest nowadays [15]–[17], due to several applications of ARRs in different fields, such as electromechanical systems [18], automotive [19] and wind turbines [20].

On the other hand, pattern analysis uses a representation of signal trends in order to extract features which allow inferring the state of a process [21], e.g., whether it is being affected or not by faults. Different approaches can be applied to this aim, such as wavelet-based methodologies [22], Fischer discriminant analysis [23], or cumulative sum charts [24], [25].

This paper is structured as follows: Section II summarizes the benchmark description while the considered scenarios are described in Section III. The results of application of the attack detection methods are presented in Section IV. Finally, in Section V, the conclusions are drawn.

II. DESCRIPTION OF THE BENCHMARK

The considered benchmark consists of two coupled water tanks, which are connected to each other through connecting pipes controlled with an automatic valve $V_b$ regulated by an On-Off controller (see Fig. 1). The first tank, denoted as $T_1$, receives water from the pump $P_1$, which is controlled by a proportional-integral (PI) controller. The second tank, denoted as $T_2$, is equipped with an outlet electro-valve $V_o$ to supply water to the consumers.

The benchmark model has been derived from the one described in [10] by incorporating a possible malicious attacker who has the goal of stealing water from the tanks while going unnoticed thanks to appropriate alterations of the outputs of the sensors, which hide the attacks. In the modified benchmark, it is assumed that the thief can extract water from the tanks using extraction pumps with flow rates $Q_{f1}$ and $Q_{f2}$, which move the water from the tanks $T_1$ and $T_2$.

*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 also by AGAUR ACCIO RIS3CAT UTILITIES 4.0 – P7 SECUTIL. This work has been also supported by the grant Juan de la Cierva-Formacion (FJCI-2016-29019).

¹The authors are with the Research Center for Supervision, Safety and Automatic Control (CS2AC) of the Universitat Politecnica de Catalunya (UPC), Spain. joseba.quevedo@upc.edu

²D. Rotondo and V. Puig are also with the Institut de Robòtica i Informàtica Industrial (UPC-CSIC), Barcelona, Spain
to the theft tanks $T_{f1}$ and $T_{f2}$, respectively. At the same time, it is assumed that the signals provided by the sensors are sent by wireless to the PI and On-Off controller, and the thief is able to hack these signals and modify them. Depending on the type of theft and the type of sensor alteration, different attack scenarios are considered, as described in Section III.

Hereafter, the model of the benchmark is described and Table I presents the model parameters. Additionally, the benchmark simulator provides complementary information about the amount of stolen water volumes $V_{12}$.

Table I presents the model parameters. Additionally, the attack scenarios are considered, as described in Section III.

### A. Model of the interconnected tanks

The variation of $V_1$ and $V_2$, which are the water volumes in $T_1$ and $T_2$, respectively, can be calculated from the balance mass equations

$$\dot{V}_i(t) = A_i h_i(t) = \sum Q_{in,i}(t) - \sum Q_{out,i}(t), \quad i = 1, 2 \quad (1)$$

where $A_i$ denotes the cross-section area of the tank $T_i$, $\sum Q_{in,i}$ is the sum of all the water inflows into the tank $T_i$ and $\sum Q_{out,i}$ is the sum of all the water outflows from the tank $T_i$.

In particular, (1) can be rewritten as

$$A_1 h_1(t) = Q_p(t) - Q_{12}(t) - Q_{f1}(t) \quad (2)$$

$$A_2 h_2(t) = Q_{12}(t) - Q_o(t) - Q_{f2}(t) \quad (3)$$

with $Q_{f1} = Q_{f2} = 0$ when no attack is performed on the system.

### B. Model of the electro-valve $V_o$

The water outflow $Q_o$ is controlled by a valve $V_o$, which is open in nominal regime, where $C_{vo}$ is the global hydraulic flow coefficient of the valve $V_o$, and $U_o \in \{0, 1\}$ is the valve position provided by the user (0 = closed, 1 = open)

$$Q_o(t) = C_{vo} \sqrt{h_2(t)} U_o(t) \quad (4)$$

### C. Model of the valve $V_b$

The water flow $Q_{12}$ through the valve $V_b$ is controlled by an On-Off controller. The flow can be calculated using Bernoulli’s law

$$Q_{12}(t) = C_{vb} U_b(t) \sqrt{(h_1(t) - h_2(t)) \sqrt{|h_1(t) - h_2(t)|}} \quad (5)$$

### D. System measurements

It is assumed that the available measurements are given by

$$y^m_x(t) = y_x(t) + \varepsilon_{y_x}(t) \quad (6)$$

where $y_x \in \{Q_p, U_p, h_1, h_2, U_b, U_o\}$ are the measured variables, and $\varepsilon_{y_x}$ denotes the corresponding measurement noise.

The values of the sensors noises of this benchmark\footnote{The benchmark is available at the URL: https://cs2ac.upc.edu/en/training-benchmarks/cyber-attacks-benchmark-simulator} are provided in the file `init.m`, located in the directory `Benchmark Program Simulation`, and are obtained as uniformly distributed signals.

### E. PI controller

The water level of the tank $T_1$, denoted as $h_1$, is regulated by a PI controller, whose output is given by

$$U^m_p(t) = K_p (h_{1,ref} - h_1(t)) + K_i \int_0^t (h_{1,ref} - h_1^m(\tau)) d\tau \quad (7)$$

where $h_{1,ref} = 0.5$ m is the set-point for $h_1^m$, while the proportional and integral gains of the controller are chosen as $K_p = 10^{-3} m^{-1}$ and $K_i = 5 \cdot 10^{-6} (m \cdot s)^{-1}$, respectively.

### F. On-Off controller

The water level $h_2$ is regulated by an On-Off controller with hysteresis with 0.09 m and 0.11 m as lower and upper switching points, respectively.

### G. Pump model

$Q_p$ is the outflow from the pump $P_1$, which is assumed to be proportional to the PI controller output $U_p$. Taking into account that the flow from the pump is limited by physical constraints, modeled as a standard saturation nonlinearity, then $Q_p$ is given by

$$Q_p(t) = \begin{cases} U_p(t) & \text{if } 0 < U_p(t) < Q_{p,max} \\ 0 & \text{if } U_p(t) \leq 0 \\ Q_{p,max} & \text{if } U_p(t) \geq Q_{p,max} \end{cases} \quad (8)$$

![Two-tank benchmark](image-url)
III. SCENARIOS OF CYBER ATTACKS

A number of attacks are considered, covering different attack policies. This section presents the different kinds of scenarios of cyber attacks.

Scenario 1 - Attackless mode: This scenario corresponds to the normal behavior of the two-tank system when nobody is stealing water.

Scenario 2 - Short-term water theft from T1: This scenario is similar to a leakage fault, the only remarkable difference being that it is cast maliciously, with the purpose of stealing water from the tank T1. In this scenario, a pump extracts a constant flow \( Q_{f1} = 10^{-4} \text{m}^3/\text{s} \) between \( t = 40 \text{s} \) and \( t = 80 \text{s} \) while adding a signal to the output of the level sensor in tank T1 so that the introduced signal hides the theft. Thanks to the introduced signal, the water level in tank T1 is kept constant, and the PI controller works as if nothing had happened. In particular, the modified value of \( h_1^m \) is given by

\[
h_1^m(t) = h_1(t) + \varepsilon h_1(t) + \frac{1}{A_1} \int_0^t Q_{f1}(\tau) d\tau \quad (9)
\]

Scenario 3 - Short-term water theft from T1 with hiding signal added to the measurement \( h_1^m \): In this scenario, the thief uses a pump to extract water with a constant flow \( Q_{f1} = 10^{-4} \text{m}^3/\text{s} \) between \( t = 40 \text{s} \) and \( t = 80 \text{s} \) while adding a signal to the output of the level sensor in tank T1 so that the introduced signal hides the theft. Thanks to the introduced signal, the water level in tank T1 remains constant, and the PI controller acts as if nothing had happened. In particular, the modified value of \( h_1^m \) is given by

\[
h_1^m(t) = h_1(t) + \varepsilon h_1(t) + \frac{1}{A_1} \int_0^t Q_{f1}(\tau) d\tau \quad (9)
\]

Scenario 4 - Long-term water theft from T1 with hiding signal added to the measurement \( h_1^m \): This scenario is similar to Scenario 3, but the theft duration is extended from 40 s to 120 s. Due to the large quantity of stolen water, the plant exhibits some physical functioning problems, since the tank T1 is emptied out, affecting the tank T2 due to the interconnection, and the consumption of water \( Q_6 \), which becomes zero.

Scenario 5 - Long-term water theft from T1 with small signal added to the measurement \( h_1^m \): In this scenario, the thief will steal water as in the previous scenarios while adding a signal that deceives the PI controller to force more water to be pumped inside the system while making harder to detect the theft. In particular, the modified value of \( h_1^m \) is given by

\[
h_1^m(t) = h_1(t) + \varepsilon h_1(t) + \frac{1}{A_1} \int_0^t Q_{f1}(\tau) d\tau \quad (9)
\]

Scenario 6 - Short-term water theft from T2: This attack scenario is similar to Scenario 2, but it affects T2 instead of T1.

Scenario 7 - Short-term water theft from T2 with hiding signal added to the measurement \( h_2^m \): This attack scenario is similar to Scenario 3, but it affects T2 instead of T1. In this case, the thief uses a pump to extract water with a constant flow \( Q_{f2} = 10^{-4} \text{m}^3/\text{s} \) while adding a signal to the output of the level sensor in tank T2, which forces the On-Off controller to act on the interconnecting valve \( V_6 \) as if nothing had happened. In particular, the modified value of \( h_2^m \) is given by

\[
h_2^m(t) = h_2(t) + \varepsilon h_2(t) + \frac{1}{A_2} \int_0^t Q_{f2}(\tau) d\tau \quad (11)
\]

Scenario 8 - Long-term water theft from T2 with hiding signal added to the measurement \( h_2^m \): This scenario is similar to Scenario 4, but the pump corresponding to \( Q_{f2} \) is used by the thief instead of the one corresponding to \( Q_{f1} \).

Scenario 9 - Long-term water theft from T2 with small signal added to the measurement \( h_2^m \): This scenario is similar to Scenario 5, but the thief steals water from the tank T2 and the introduced signal is meant to deceive the ON-OFF controller instead. In this case, the modified value of \( h_2^m \) is given by

\[
h_2^m(t) = h_2(t) + \varepsilon h_2(t) + \frac{1}{A_2} \int_0^t 0.5Q_{f2}(\tau) d\tau \quad (12)
\]

Scenario 10 - Replay attack: In this scenario, the thief steals water when the plant has reached its steady-state. However, before doing so, he/she records the measurements coming from the sensors without stealing water from the tanks. Then, in a subsequent phase of the attack, the thief steals water while replacing the data with the recorded one. More specifically, the water is stolen from \( t = 160 \text{s} \) to \( t = 200 \text{s} \), while measurements recorded in the 50 s previous to the attack are used to deceive the controller and the supervision system. At time \( t = 200 \text{s} \), the replay attack ends and the controller and the supervision system are able to see the real data coming from the system.

IV. RESULTS OF TWO METHODS FOR CYBER ATTACK DETECTION

This section presents two methods to detect the cyber attacks of this benchmark. One of the methods is based on model-based fault diagnosis using analytical redundancy relations: the well-known ARRs [13] and the second method relies on extracting some features from the time response of the sensors.

A. Results based on analytical redundancy relations

Based on the available sensors and replacing the equations of the model described in Section III, four possible analytical redundancy relations (ARRs) can be derived replacing the physical variables by the real measured signals and using the perfect matching approach [17]. These ARRs allow deriving the residual expressions as follows.
\[ r_1(t) = -Cvb U^m_b(t) \text{sign} (h^m_1(t) - h^m_2(t)) \sqrt{|h^m_1(t) - h^m_2(t)|} \]
\[ + Q^m_p(t) - A_1 \frac{dh^m_1}{dt} \]
\[ r_2(t) = Cvb U^m_b(t) \text{sign} (h^m_1(t) - h^m_2(t)) \sqrt{|h^m_1(t) - h^m_2(t)|} \]
\[ - Cvo \sqrt{h^m_2(t) U^m_o(t)} - A_2 \frac{dh^m_2}{dt} \]
\[ r_3(t) = U^m_p(t) - K_p (h_{1,ref} - h^m_1(t)) - K_I \int (h_{1,ref} - h^m_1(t)) dt \]
\[ r_4(t) = Q^m_p(t) - \begin{cases} 
U^m_p(t) & \text{if } 0 < U^m_p(t) < Q_{p,max} \\
0 & \text{if } U^m_p(t) \leq 0 \\
Q_{p,max} & \text{if } U^m_p(t) \geq Q_{p,max}
\end{cases} \]

The residuals (13)-(16) are expressed in discrete-time by applying an Euler discretization with sampling time \( T_s = 1s \).

Figs. 2-5 show the results of the residual \( r_1 \) for Scenarios 2, 3, 4 and 5 in order to detect an attack in the tank \( T_1 \). The threshold has been calculated as three times the standard deviation of the residual in Scenario 1 (i.e., the attackless scenario).

Clearly, in Scenarios 2 and 4, the attacks can be detected analyzing the residual \( r_1 \). In Scenarios 3 and 5, the residual \( r_1 \) appears to be insensitive, such that the corresponding attacks cannot be detected.

B. Results based on some features of the temporal pattern of the sensors

A possible approach to improve the detectability of the cyber attacks consists in analyzing the signals coming from the sensors in the scenarios where the analytical residuals are inefficient for this task, as Scenarios 3 and 5. Fig. 6 presents the time evolution of the measured level in tank \( T_1 \) during Scenario 1 (attackless) and the real evolution of the water level during Scenario 3. The difference between the real and measured level \( y_1 \) in the tank \( T_1 \) is due to the water stolen during this attack. On the other hand, the fast oscillation of \( h^m_1 \) without any attack is due to the fast switching On-Off of...
the interconnected pipe to maintain the level in the tank $T_2$ around 0.09 and 0.11 meters (see Fig. 7).

However, the oscillation of $h_1^m$ in Scenario 3 is much slower due to the lower difference between the levels of tanks $T_1$ and $T_2$ such that more time is necessary to refill the tank $T_2$. This feature of the signals $h_1^m$ or $h_2^m$ could be exploited to detect an attack. In particular, analyzing the frequency oscillation of $h_2^m$ in Fig. 7, the time to detect a misbehavior (an attack in this scenario) can be shortened significantly. The same situation is found in Scenario 5 (see Fig. 8) and the interesting feature is that the smaller oscillations of these signals start when the attack is developed and remains at this low oscillation frequency until the end of the scenario because of the water theft and the low level in tank $T_1$.

Table II shows the difference in the oscillation frequency of $h_2^m$ for the Scenarios 1, 3 and 5. Clearly, this feature is an interesting and complementary information that can be used in order to detect these attacks.

**TABLE II**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Freq. oscillation $h_2^m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$1/12$ Hz</td>
</tr>
<tr>
<td>3</td>
<td>$1/49$ Hz</td>
</tr>
<tr>
<td>5</td>
<td>$&lt; 1/35$ Hz</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS

This paper has presented some results related to the detection and isolation of cyber attacks using a recently proposed benchmark based on a two-tank system. The benchmark has proposed some attack scenarios in which a malicious attacker alters the signals of the water level sensors in the tanks, in order to remain hidden while stealing water. Results in five of these scenarios have been presented showing the difficulty to detect cyber attacks based only on the model-based residuals calculated using the measured variables, and how using the time evolution pattern analysis of the measured sensors, it becomes possible detect some of these cyber attacks. As future research, the remaining five scenarios will be considered.

![Fig. 6](image_url)  
**Fig. 6.** Pattern time evolution of the measured water level $h_1^m$ in Scenarios 1 and 3.

![Fig. 7](image_url)  
**Fig. 7.** Pattern time evolution of the measured water level $h_2^m$ in Scenarios 1 and 3.

![Fig. 8](image_url)  
**Fig. 8.** Pattern time evolution of the measured water level $h_2^m$ in Scenarios 1 and 5.

REFERENCES


