On Practical Realization of Evasion Attacks for Industrial Control Systems

Erba, Alessandro; Murillo, Andres F.; Taormina, Riccardo; Galelli, Stefano; Tippenhauer, Nils Ole

DOI
10.1145/3689930.3695213

Publication date
2024

Document Version
Final published version

Published in
RICSS '24

On Practical Realization of Evasion Attacks for Industrial Control Systems

Alessandro Erba*
alessandro.erba@kit.edu
KASTEL Security Research Labs,
Karlsruhe Institute of Technology
Karlsruhe, Germany

Andres F. Murillo
andres.murillo@fujitsu.com
Fujitsu Research of Europe
Slough, United Kingdom

Riccardo Taormina
r.taormina@tudelft.nl
Delft University of Technology
Delft, Netherlands

Stefano Galelli†
galelli@cornell.edu
Cornell University
Ithaca, United States of America

Nils Ole Tippenhauer
tippenhauer@cispa.de
CISPA Helmholtz Center for
Information Security
Saarbrücken, Germany

Abstract

In recent years, a number of evasion attacks for Industrial Control Systems have been proposed. During an evasion attack, the attacker attempts to hide ongoing process anomalies to avoid anomaly detection. Examples of such attacks range from replay attacks to adversarial machine learning techniques. Those attacks generally are applied to existing datasets with normal and anomalous data, to which the evasion attacks are added post-hoc. This represents a very strong attacker, who is effectively able to observe and manipulate data from anywhere in the system, in real-time, with zero processing delay, and no computational constraints. Prior work has shown that such strong attackers are theoretically difficult to detect by most existing countermeasures. So far, it is unclear if such an attack could be practically realized, and if there are challenges that would impair the attacker. In this work, we systematically discuss options for an attacker to mount evasion attacks in real-world ICS, and show the constraints that result from those options. To validate our findings, we design and implement a framework that allows the realization of evasion attacks and anomaly detection for ICS emulation. We demonstrate practical constraints that arise from different settings, and their effect on attack performance. For example, we found that network packet replay might trigger network errors, which will result in unexpected spoofing patterns.

CCS Concepts

• **Computer systems organization** → *Embedded and cyber-physical systems*; • **Security and privacy** → *Systems security*; *Intrusion detection systems.*

*Part of this work was done while Alessandro Erba was with CISPA Helmholtz Center for Information Security and Saarbrücken Graduate School of Computer Science, Saarland University.

†Part of this work was done while Stefano Galelli was with the Pillar of Engineering Systems and Design, Singapore University of Technology and Design.

RICSS '24, October 14–18, 2024, Salt Lake City, UT, USA

ACM ISBN 979-8-4007-1226-5/24/10
https://doi.org/10.1145/3689930.3695213

Keywords

Industrial Control Systems, Anomaly Detection, Evasion Attacks, Dataset

ACM Reference Format:
Alessandro Erba, Andres F. Murillo, Riccardo Taormina, Stefano Galelli, and Nils Ole Tippenhauer. 2024. On Practical Realization of Evasion Attacks for Industrial Control Systems. In *Proceedings of the 2024 Workshop on Re-design Industrial Control Systems with Security (RICSS '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3689930.3695213

1 Introduction

Industrial Control Systems (ICS) are core components of critical infrastructure such as power grids, manufacturing systems, and transportation systems. ICS rely on Programmable Logic Controllers (PLCs), industrial network protocols, and increasingly leverage network standards such as Ethernet/IP. Correct and efficient operations of ICS are essential for societies' well-being and commercial success. In general, systems that use networked digital control for physical processes are also called Cyber-Physical Systems (CPS).

Unfortunately, ICS are threatened by adversarial manipulations and general cyber attacks. For example, attackers could manipulate insecure local network traffic in an ICS to change reported sensor values, e.g., leading to incorrect control actions that result into physical damage to the process or danger to human operators. As security solutions (such as updates) are often not available for industrial legacy devices in this domain, complementary monitoring solutions were proposed in prior work to monitor correct operations of the cyber components (i.e., network and hosts) and physical processes (e.g., based on process data) [47].

As a reaction to such *process-aware* anomaly and attack detection systems, more advanced manipulation strategies were proposed [19] that hide ongoing anomalies or attacks in the physical process from detectors. We broadly call such attacks evasion attacks if they do not impact the physical process itself, and instead just manipulate sensor data as observed by the attack detector. To evaulate the attacks' success in evading detection by various detectors, pior work commonly applies evasion attacks to existing ICS datasets, e.g., the SWaT [34] or BATADAL [46]. Design and implementation

of evasion attacks to manipulate traffic in real-time is limited in prior work [19]. Tools such as the DHALSIM [38] ICS emulator are so far unable to directly run evasion attacks, as they do not provide features to manipulate ICS traffic in realtime. Thus, it remains unclear if attacks in prior work can actually be practically realized.

In this work, we address this problem by a) designing and implementing a framework to perform evasion attacks on typical ICS traffic (by extending DHALSIM [38]); b) using this framework to verify whether prior work attacks could be realized in practice; and c) for schemes that we were able to implement, we evaluate whether the realization achieves expected performance, and d) we record and share resulting datasets that include (for the first time for artificial datasets) both process and traffic data.

Our main contributions are as follows:

- We evaluate the options for practical implementations of traffic manipulation attacks in ICS, and present a framework to realize evasion attacks for real-world systems through real-time manipulation of actual industrial protocols (in our case, Ethernet/IP).
- We use the framework to realize prior work attacks (where possible) and present an evaluation of the impact that practical constraints have on the realizability of these attacks.
- We produce a new dataset based on the BATADAL dataset, where the process data are complemented with network traffic. In the dataset, we included a number of attacks from BATADAL and several new evasion attacks.
- We discuss how prior work detectors could be extended to leverage network features, and what impact this would have on the attacker.

Reproducibility. The paper is complemented with a fully functional artifact and dataset to enable reproducibility. The dataset is available at https://doi.org/10.5281/zenodo.13692004, and the evaluation scripts at https://github.com/Critical-Infrastructure-Systems-Lab/Practical-Evasion-Attacks.

2 Realizing Evasion Attacks on Process-Aware Anomaly Detection

2.1 System and Attacker Model

We assume an Industrial Control System, organized according to Purdue architecture (see Figure 1). We assume that a process-based anomaly detection system is deployed on the system to detect process anomalies. An attacker intrudes to disrupt the physical process by manipulation. Moreover, the attacker aims to remain undetected from the anomaly detector by launching an evasion attack. Different attackers can be modelled according to the attack location from which the process evasion is executed. We practically envision three possibilities for evasion attack locations (see Figure 1). The first is an attacker present at level 3 that can execute unconstrained evasion attacks (i.e., intercept and manipulate industrial traffic to spoof arbitrary sensor values to hide the anomaly). The second attacker is present at level 2 and can manipulate the information coming to the SCADA area from the lower layer and from the other areas that interact with the targeted area for control purposes. Finally, the third attacker is located at level 1 and manipulates the sensor readings outgoing from the PLC.

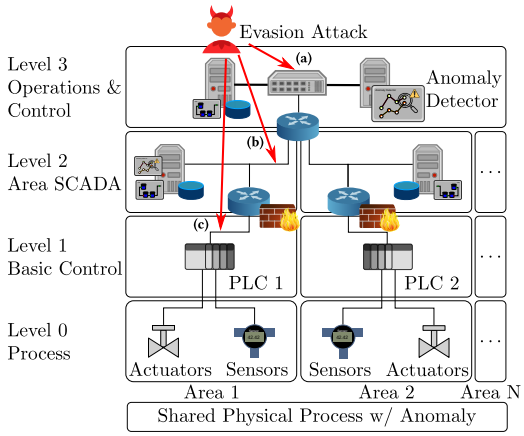


Figure 1: Purdue architecture, system model, based on [20]. An evasion Attack is performed by an attacker to hide an ongoing process anomaly. The attack can occur within the system depending on the attacker's location in the network (a/b/c).

Dataset	Simul./Phy. Process	System	Network	Evasion
SWAT [34]	Physical	WT	Yes	No
WADI [27]	Physical	WD	No	No
BATADAL [46]	Simulated	WD	No	Yes†
DHALSIM [37, 38]	Simulated	WD	Yes	No
This work	Simulated	WD	Yes	Yes

Table 1: Comparison of the proposed dataset with related anomaly detection datasets in the water sector. WD: Water Distribution, WT: Water Treatment. †Implemented strategies are replay and polyline.

2.2 Use case and Motivation

In practical scenarios, there are a number of possibilities for attacker placement in the network topology. Consider the following scenario (based on Figure 2): the attacker is located at level 1 of the Purdue architecture (i.e, (c) in Figure 1) i.e., the attacker can read and manipulate the incoming and outgoing traffic from PLC 1 (e.g., towards PLC2 and SCADA), but cannot read or manipulate traffic between PLC 2, PLC 3 and the SCADA. The attacker wants to hide a physical anomaly in the process that affects sensors attached to PLC 1 and PLC 2. Prior work in the field was conducted mainly over pre-recorded datasets, assuming an attacker to operate on all data received at the SCADA system. That implies that this approach from prior work cannot be implemented as intended in the given scenario, or that a constrained implementation would have potentially worse performance. In particular, prior work does not capture the effect of evasion traffic manipulation on industrial components such as PLC 2 and PLC 3. The difference in capabilities between a theoretical attacker and a more practical attacker make necessary to have practical frameworks to evaluate evasion attacks. Such frameworks should provide network emulation capabilities so as to

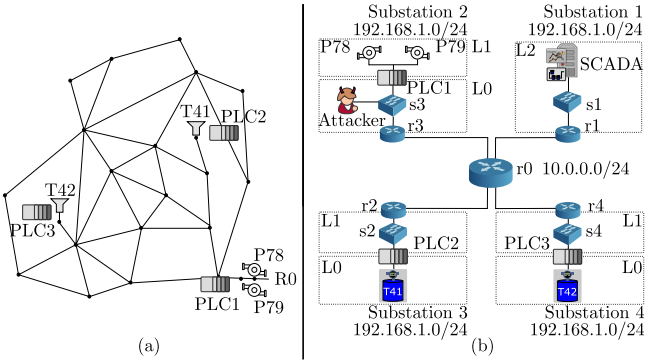


Figure 2: Use case scenario, (a) physical topology of the water distribution network. (b) Network topology of the water distribution network.

fully represent the constraints and challenges that a real attacker would face when attempting to not only compromise an ICS, but also to conceal their activities. Finally, the evaluation framework should also be a flexible and re-configurable environment to allow easy configuration of new experiments and variations of previous experiments. Such variations would allow researchers to test different attack, data manipulation, and data concealment strategies. In Table 1, we summarize the difference of the resulting data collected with our proposed framework, compared to prior datasets.

The focus this work is not the evasion per se, or the evasion of a specific detector (invariant-based, physics-based or reconstruction-based discussed in prior work [5, 19–21, 28, 32, 49]). Instead, we explore the practical challenges posed by the network during the execution of evasion attacks, a topic not addressed in prior work.
Research Questions. Based on the research gap and challenges identified, in this work, we address the following two research questions. *RQ1.* How can prior work evasion attacks be practically realized? *RQ2.* What challenges are introduced by real-world requirements?

3 Realizing Prior Evasion Attacks

3.1 Replay Attacks

In a replay attack [35], the attacker attempts to cover the traces of a physical anomaly by replaying the sensor values as occurred in the past. The message replay requires the attacker to be able to eavesdrop and spoof the traffic. No knowledge about the target anomaly detection or the target anomalies is required.
Realizability Challenges. No realizability challenges are identified for such attack. We implement it in our evaluation.

3.2 Gray-box JSMA

Anthi et al. [5] proposed a framework for evasion attacks based on the JSMA attack [41] algorithm against a multi-layer perceptron. The proposed method assumes that anomaly detection is trained on anomalous and malicious classes for attack detection. Models are trained and tested by splitting the same dataset into two parts (containing the same physical anomalies). The authors propose adversarial training to defend the classifier and achieve robustness. No details are provided about the computational time of the method.
Realizability Challenges. To launch an evasion attack, the attacker is assumed to collect a dataset to train a surrogate model of the detector. The attacker is assumed to record sensor readings from the system both under attack and under normal conditions. This assumption is not realistic in practical scenarios, since launching an attack on the system would trigger the anomaly detection system. Furthermore, if a prediction-based approach is employed to collect anomalous data, we believe it would be challenging to obtain such data unless the attacker has knowledge of the system's physical processes. The rationale is that data-driven methods are unlikely to accurately predict system behavior outside the normal operating ranges.

3.3 ConAML

The paper by Li et al. [32] proposes an attack against anomaly detectors for ICS based on linear constraints. The attacker is assumed to know such linear constraints operating the ICS and collects a meaningful set of anomalous and non-anomalous sensor readings to train a surrogate model used by the attacker to perform the perturbations. The authors also assume the attacker is constrained to not know the values of uncompromised sensor readings (which is in contradiction with the assumption that the attacker collects sensor readings to train a surrogate model). Data generation for the surrogate attack model follows the same generation used by the defender, which resulting in perfect knowledge of the anomalies.
Realizability Challenges. In addition to the previously discussed limitations, understanding the system's linear constraints is highly challenging and impractical for an attacker, as it requires perfect knowledge of the target CPS. Similar to Anthi et al. [5], the attacker must use anomalous sensor readings to train the surrogate model for the attack. However, this contradicts the evasion objective, as the attacker would be detected while collecting anomalous data.

3.4 Gradient and Genetic Algorithms

The work by Jia et al. [28], focuses on realizing evasion attacks that are effective against both residual-based detectors and invariant checking; assuming training solely on normal data. The authors propose a white-box framework consisting of two subsequent attack strategies: the first is a gradient-based method to evade the residual-based detector, the second uses genetic algorithms to evade the rule-checking algorithm. The proposed method is effective in evading both the considered detectors. No details about the time required to compute the adversarial examples is provided.
Realizability Challenges. The anomaly detection system uses on a sliding window. To compute the prediction at timestep t, the recurrent neural network uses the past observations $[x_{t-1}, x_{t-2}, \ldots, x_{t-n}]$. To evade the CUSUM detector, the attacker must minimize the difference between the actual sensor value x_t and its prediction $\hat{x}_t = f([x_{t-1}, x_{t-2}, \ldots, x_{t-n}])$. The attacker is assumed to perturb the observations $[x_{t-1}, x_{t-2}, \ldots, x_{t-n}]$. Applying this attack would is challenging for two alternative reasons: a) *the attacker knows the future* If the attacker wants to evade the detector at time $t+n$, he must know in advance the anomalous sensor reading x_{t+n} at time x_t and start perturbing the sensor readings $[x_t, x_{t+1}, \ldots, x_{t+n}]$. b)

attacker alters the past alternatively, the attacker must perturb the values stored in the historian, leading to database inconsistency.

3.5 Attacks against Image Based Detection

In this work by Niazazari et al. [39], the anomaly detection problem in ICS is formulated as an image classification problem. The sensor readings are pre-processed to obtain an image where each pixel represents the evolution of the sensor's multivariate time series over time *samples*×*sensors*. The sensor readings are fed in the image format to a convolutional neural network for anomaly detection. To perform the evasion, the authors propose to use white-box FGSM and JSMA against an attacker-trained surrogate model.
Realizability Challenges. The attacker perturbation is unconstrained within the `samples`, while the attacker is constrained in the number of `sensors` that can be manipulated. This leads to the two challenges identified in the previous section, *the attacker knows the future*, *the attacker alters the past*.

3.6 Iterative and Learning-based Attacks

In this work by Erba et al. (A) [19] authors propose white-box and black-box evasion techniques to conceal anomalies against reconstruction-based anomaly detectors. The attacker is assumed first to collect sensor readings without anomalies before an attack. Both techniques can be applied in real-time, in constrained and unconstrained settings (i.e., based on which features the attacker can observe and spoof). The proposed white-box technique uses an iterative coordinate descent algorithm to minimize the Mean Squared Error between the input and output of the target anomaly detector. The black-box technique operates by employing an autoencoder to compute the adversarial perturbations.
Realizability Challenges. In the white-box case the attacker is assumed to have a white-box knowledge of the target detector. We do not expect any particular challenge to practically implement such manipulations in practice. We demonstrate the integration of the black-box attack.

3.7 Generic Concealment Attacks

In this work by Erba et al. (B) [20], the authors propose the so-called *generic concealment attacks*. The method proposes evasion techniques to evaluate which properties of the CPS are learned by the anomaly detection system. The attacker spoofs the sensor readings to hide the anomalies from the process-based detectors. The applied spoofing techniques involve constrained replay [35], random replay, and stale data attacks [30]. Those attacks are demonstrated to evade a wide range of detectors proposed in prior work. Those evasion patterns can be pre-computed by the attacker (black-box) and applied over the sensor readings to override the anomalies.
Realizability Challenges. The attacker is assumed first to collect sensor readings without anomalies before an attack. Then, the attacker applies the spoofing of sensor readings to evade detection. We do not expect particular challenges to implement such manipulations in practice. We implement such attacks in our evaluation.

3.8 White-box Evasion

In this work by Erba et al. (C) [21], the authors proposes a method to perform white-box attacks on anomaly detectors for CPS. The proposed technique accounts for two challenges in evading CPS detectors 1) the attacker is constrained to perturb only the current sensor readings, which is related to the challenges we identified before (*the attacker knows the future*, *the attacker alters the past*). 2) not all the anomaly detectors from prior work are differentiable, i.e., gradient-based methods may not be applicable.

The proposed technique is applied to multiple anomaly detectors from prior work. The proposed method evades the detectors while being constrained to perturb only the latest process sensor readings.
Realizability Challenges. The attacker is assumed to have a white-box knowledge of the target detector. We do not expect any particular challenge to practically implement such manipulations in practice as they already account for the related challenges, *the attacker knows the future*, *the attacker alters the past*.

3.9 L_0 Optimisation and Prediction Attacks

The work by Zizzo et al. [49], focuses on attacking autoregressive deep neural network models for anomaly detection in CPS. The proposed method assumes the attacker to have a precise and accurate model of the physical system. This model can be used by the attacker to predict how the system will react to attacks. By using such a model the attacker can start perturbing the sensor readings to conceal anomalies that will launch on the system.
Realizability Challenges. In practice, the attacker, to avoid altering the past (*the attacker alters the past*), starts the evasion before an attack starts by relying on its knowledge about the future (*the attacker knows the future*). We believe that precise knowledge of the system behavior under anomalous conditions is challenging to obtain, as the this assumes that the attacker knows how the system physically responds to the process manipulation.

4 Evasion Attack Framework

As discussed in Section 2, online attackers face additional constraints and challenges compared to offline attackers. Our framework considers these challenges and constrains by launching the evasion attacks by emulating realistic network traffic. Our framework was developed with the following design goals:
Industrial Network emulation capabilities. The differences between theoretical and practical attackers only appear with an actual implementation of the communication protocols used in CPS. Without a network implementation, the experiments are carried out by doing offline modification of the experiment datasets.
Attack, manipulation, and concealment capabilities. The framework objective the evaluation of different attacker strategies and capabilities. These can be modelled as a combination of different types of network or device attacks, manipulation strategies (ways in which an attacker can change the value of sensors/actuators), and concealment capabilities. Moreover, it is important to design the framework in a way that supports extension of framework features.
Easy configuration capabilities. The framework also needs to provide a configuration interface to allow researchers to consistently control the attack start and stop, as well as the attack parameters.

4.1 Framework Design

Our proposed evasion framework is shown in Figure 3. In the framework, a co-simulation environment is used in which the physical process is simulated and the PLCs take measurements from this physical simulation through their sensors, use industrial communication protocols to exchange these measurements in an emulated network, send the system state to the SCADA server, apply the respective control rules, and finally change the status of the simulation actuator according to the decisions of the control rules. In this framework, an attacker launches MiTM attacks to intercept the messages being exchanged between the PLCs and the SCADA server. Attackers modify these messages using two types of data to manipulate and conceal their attack. Data manipulation is aimed at the PLCs with the goal of disrupting the normal operation conditions, by triggering wrongful actuators activation. The evasion data is aimed at the SCADA server with the goal of concealing the manipulation and effects of the disruption. Multiple concealment techniques can be used and are described in Section 4.3. We assume the detector is installed in the same location as the SCADA server, as this would provide a global view of the system to the detector.

4.2 Framework Implementation

We extended DHALSIM [37] to implement our evasion framework. Before our extension, DHALSIM did not provide any concealment capabilities. We thus extended DHALSIM by creating additional attacks and modifying the previous attack scripts to enable the implementation of the concealment strategies presented in Section 4.3. Moreover, in the case of learning-based concealment, we created a new type of attack that compromises the communication between PLCs and SCADA in order to capture all traffic forwarded to the SCADA server. Finally, the extension included adapting the learning-based concealment [19] mechanism to work online with DHALSIM. This adaptation required extending the DHALSIM synchronization mechanism [38] to allow for the simulation to pause and wait for the attacker to receive and potentially modify all necessary SCADA packets. The result of this extension of DHALSIM is a more diverse attack framework that enables researchers to evaluate not only the impact of different attack strategies, as presented by Murillo et al. [37], but also diverse strategies for attack concealment. The framework provides the following evasion techniques: value replay, network packet replay, and learning-based concealment.

While our framework is applied to water distribution systems, it is inherently adaptable for other CPS. Adapting DHALSIM would require replacing the physical system model (e.g., using a model of a water treatment plant) or substituting the network emulator.

4.3 Evasion Techniques

MiTM value replay attack. The attack consists of three phases: capture, idle, and replay. In the capture phase, the attacker parses the CIP/ENIP package to extract the original payload (sensor reading) of the message. The attacker then stores this value in a vector. When the replay phase starts, The attacker parses the CIP/ENIP package again to replace the original payload with a previously-captured value. The capture phase and replay phase have the same duration. Additionally, the attacker can replay one or multiple tags. Finally, between the capture and replay phase, there might be an interval where the attacker is only forwarding packets without eavesdropping them or modifying them. Providing support for multiple tag replay required extending the DHALSIM parsing capabilities to maintain the state of the CIP/ENIP transaction sessions exchanged between PLCs and SCADA (as the CIP/ENIP responses provided by MiniCPS), which contains the tag value as its payload.

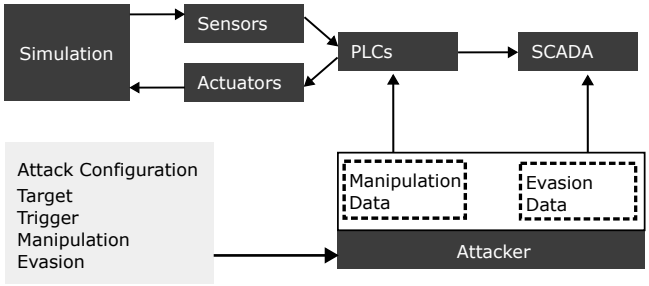


Figure 3: Attack Evasion Framework. The Attacker manipulates the process to cause a physical anomaly and spoofs the data towards the SCADA system to evade detection.

MiTM network packet replay attack. This attacks is almost identical to a value replay attack, except that during the capture and replay phases, the attacker does not parse the CIP/ENIP package to get the payload of the message. Instead, the attacker captures the entire CIP/ENIP structure of the package and replays it.

Learning-based concealment. This technique uses the work presented by Erba et al. A [19] to conceal the attack values. We specifically implemented the 'unconstrained black box attack'. The attacker also parses the CIP/ENIP package to replace the payload with a value calculated by the adversarial machine learning model. Although we only integrated this learning-based concealment method, the evasion framework could be easily extended to integrate other learning-based concealment modules. Implementing this concealment technique required modifying DHALSIM synchronization mechanism [38] to allow the concealment module to receive all values sent to SCADA in an iteration and calculate the concealment values. This module also required enhancing the performance of the SCADA server in handling the values received from the PLCs.

4.4 Creating a Dataset with Traffic Data

We used DHALSIM and our framework to realize prior attacks against industrial control systems. As DHALSIM was developed to execute cybersecurity experiments in water distribution systems (WDS), our experiments were also conducted in a WDS, specifically C-Town. We selected C-Town mainly for two reasons: 1) it is a WDS that is complex enough to have multiple control rules, targets (PLCs, sensors, and actuators), and a well-known behaviour, 2) C-Town was also used to generate a well-known dataset for cyber-security experiments in ICS, the BATADAL dataset [46]. Nevertheless, that dataset did not include ICS network traffic data. We used this framework to extend the BATADAL dataset.

We ran our experiments with DHALSIM v0.6.0 using the C-town network and generated two data types: normal and attack data. Normal operating conditions were simulated for 51 weeks. The collected physical features comprise 39 sensors and actuators.

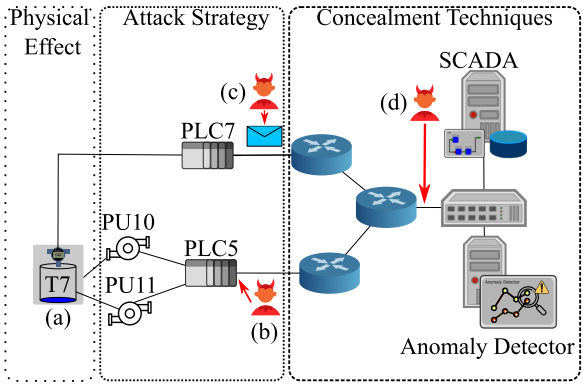


Figure 4: Relationship between physical effects, attack strategies, and concealment techniques. The same physical effect (a) can be achieved through multiple strategies; through a PLC attack or a MiTM attack (b, c). The attack can be concealed using multiple concealment strategies (d), such as value replay, network replay, or learning-based concealment.

Each week was run separately, with random initial tank levels, demand patterns, and network conditions (e.g., network packet loss and delays) generated for each week. The initial tank levels were generated within C-Town's normal operating conditions. This eliminates the system's transitory phase and avoids sudden pump or valve activation (Similar to the approach followed by Murillo et al. [38]). The attacks were inspired by the BATADAL dataset, and we implemented the majority of these attacks using various techniques. For some of these attacks, we also used different evasion techniques.

4.5 Physical Effects, Attack Strategies, and Concealment Techniques

Figure 4 explains the details of our attacks and the relationships between physical effects, attack strategies, and concealment techniques. We use the term "physical effect" to explain the objective of an attacker affecting the physical system. For example, the BATADAL attack 1 has the aims to achieve low levels in Tank 7 ((a) in Figure 4). This objective can be achieved through multiple attack strategies. In C-Town, Tank 7 level is controlled by pumps PU10 and PU11, and, in our topology scenario, Tank 7 is measured by PLC9; PLC9 sends the tank level readings to PLC5 that applies programmed control rules to activate PU10 and PU11 to maintain T7 level in the desired range. This means that with DHALSIM we can achieve this objective using different attack strategies. For example, we could launch a device attack on PLC5 that maliciously turns off pumps PU10 and PU11, causing low levels on Tank 7 ((b) in Figure 4). Another strategy would be a MiTM attack that spoofs the value of sensor T7, causing PLC5 to turn off the pumps ((c) in Figure 4). Spoofing the value could be done either by assigning a fixed value to T7 (referred to as a value attack) or by adding an offset to the real T7 value (referred to as an offset attack). Some attacks are implemented using a combination of PLC attacks and MiTM attacks. Finally, the attacker can try to conceal the attack using the concealment strategies described in 4.3 ((c) in Figure 4).

Physical Anomaly				Concealment Technique	
BTDL ID	Dur. (h)	Physical Effect	Strategy	ID	Technique
1	31	Low levels T7	PLC5 attack, turn off PU10/PU11	1	Value replay T7
				2	Network replay T7
				44	Concealment [19]
			MiTM PLC9 → PLC5	3	Value replay T7
				4	Network replay T7
				5	Concealment [19]
2	31	Low levels T7	PLC5 attack, turn off PU10/PU11	6	Value replay T7, PU10, PU11
				7	Network replay T7, PU10, PU11
			MiTM PLC9 → PLC5	8	Value replay T7, PU10, PU11
				9	Network replay T7, PU10, PU11
				10	Concealment [19]
3	31	Overflow T1	MiTM PLC2 → PLC1	11	Value replay T1
				12	Network replay T1
				13	Concealment [19]
4	94	Overflow T1	MiTM PLC2 → PLC1	14	Value replay T1, PU1, PU2, J269
				15	Network replay T1 PU1, PU2, PJ269
9	94	Overflow T2	PLC2 attack, turn on V2	25	Value replay T2
				26	Network replay T2
12	100	Overflow T2	PLC2 attack, turn on V2	27	Value replay T2, V2, J14, J422
				28	Network replay T2, V2, PJ14, PJ422
				29	Concealment [19]
14	30	Overflow T4	MiTM PLC6 → PLC3	34	Value replay T4
				35	Network replay T4
			PLC3 attack, turn on PU6/PU7	36	Value replay T4
				37	Network replay T4

Table 2: Concealment attack dataset. It is inspired by the BATADAL [46] dataset, but uses multiple attack strategies to cause the same effect on the physical system. For some of these attacks, multiple concealment methods are used. BTDL stands for BATADAL. The fifth column reports the attack identifier that we gave to the concealment attack. The last column describes the concealment technique.

5 Experiments and Evaluation

Using our framework, we collect a dataset consisting of eight physical anomalies, of which three are obtained through two different attack strategies. The physical anomalies collected are inspired by the attacks contained in the BATADAL dataset [46] (we use the same attack identifiers for the physical anomalies). We did not implement all attacks in the BATADAL for mainly two reasons: 1) the need for additional DHALSIM extensions (e.g., for pump speed and control rule thresholds), and 2) the attacks currently implemented

provide enough data to answer our research questions. For each of those anomalies, we perform two to six different evasion attacks, totaling twenty-five concealment attacks. The dataset consists of physical values (`.csv`) and the related network traffic generated (`.pcap`). Table 2 provides an overview of the attacks included in the dataset (see Appendix B for a detailed textual description).

5.1 Evaluation Results

We evaluate how the different attack and concealment strategies implemented evade process-based detectors. Table 3, summarizes the results described in this section. We plot the effects of the attacks at the three levels identified in Figure 4 (the plots are presented in the appendix). We employ the autoencoder-based detection proposed by Taormina et al. [45]. We rely on it as it is open source and widely adopted for security evaluations [19, 20, 49]. The autoencoder detector was tuned with the following parameters, window = 3 and theta = 0.002358. We note that our goal is not to evaluate the robustness of the target anomaly detector (which was extensively investigated by prior work [19, 20]) but to evaluate how the different attack and concealment strategies affect the concealment efficacy.

We start evaluating the efficacy of the different attack strategies to achieve the same physical evasion. Then, we evaluate the efficacy of the different concealment strategies. We consider the following standard metrics: Accuracy, F1 score, Precision, Recall, and False Positive Rate. Specifically, as done in prior work [19], we evaluate the evasion performance in terms of Recall score reduction. A lower Recall score after an attack indicates more successful concealment.

5.1.1 Effects of Attack strategies. **Attack strategy 1 vs 1.2.** Despite the different strategies implemented, the resulting physical effect is the same (see Appendix Figure 5a and Figure 5b). Strategy 1 directly turns off the pumps to cause low levels in T7 (and as a consequence also in T6). While strategy 1.2 spoofs a high value for T7 causing the PU10 to turn off and drain the water tank. In Table 3, we compare the detection results of those two attack strategies, before the concealment. As we can see, Anomaly 1 is detected with a lower recall compared to Anomaly 1.2. The offset induced by Anomaly 1.2 triggers the alarm as the level of T7 goes out of range.
Attack strategy 3 vs 4. The two attacks implement the same strategy but attack 3 lasts less time than attack 4. Attack 3 does not bring T1 to the overflow level (see Appendix Figure 6a and Figure 6b). In Table 3, we compare the detection results of those two attack strategies (without concealment). Due to the attack duration Anomaly 3 is undetected while Anomaly 4 is detected.
Attack strategy 9 vs 12. The water tank T2 overflows as valve V2 is maliciously kept open. The attacks use the same strategy resulting in the same physical effect and detection scores (Numerical scores are not exactly the same because of the non-deterministic initial tank conditions) (see Table 3 and Appendix Figure 8a and Figure 8b).
Attack strategy 14 vs 14.2. Similarly to what we observed earlier, the resulting physical effect is the same despite the different strategies implemented. Strategy 14 spoofs the value of T4 to cause PU6 and PU7 to turn ON. Strategy 14.2 turns on PU6 and PU7 to cause T4 to overflow (see Appendix Figure 9a and Figure 9d) . In Table 3, we can compare the results of the detection. Similar to what we observed before for Anomaly 1 and 1.2, the offset attack is detected with a higher detection probability.

BTDL ID	Att. ID	Acc.	F1	Prec.	Rec.	FPR
1	(38)	0.91	0.70	0.60	0.86	0.09
1	attack 01	0.90	0.69	0.59	0.85	0.09
1	attack 02	0.90	0.70	0.59	0.85	0.09
1	attack 44	0.90	0.67	0.58	0.81	0.09
1.2	(45)	0.92	0.77	0.63	1.00	0.09
1.2	attack 03	0.90	0.70	0.59	0.85	0.09
1.2	attack 04	0.90	0.69	0.58	0.85	0.09
1.2	attack 05	0.90	0.66	0.57	0.80	0.09
2	attack 06	0.91	0.70	0.60	0.86	0.09
2	attack 07	0.90	0.69	0.59	0.83	0.09
2.2	attack 08	0.91	0.70	0.59	0.86	0.09
2.2	attack 09	0.90	0.68	0.58	0.83	0.09
2.2	attack 10	0.89	0.66	0.57	0.80	0.09
3	(39)	0.84	0.00	0.00	0.00	0.03
3	attack 11	0.84	0.01	0.03	0.01	0.04
3	attack 12	0.84	0.01	0.03	0.01	0.03
3	attack 13	0.84	0.02	0.06	0.01	0.03
4	(40)	0.71	0.49	0.82	0.35	0.05
4	attack 14	0.76	0.60	0.85	0.47	0.05
4	attack 15	0.68	0.39	0.76	0.26	0.05
9	(41)	0.68	0.37	0.76	0.25	0.05
9	attack 25	0.71	0.48	0.82	0.34	0.05
9	attack 26	0.71	0.49	0.83	0.34	0.05
12	(42)	0.67	0.37	0.75	0.25	0.05
12	attack 27	0.70	0.46	0.81	0.32	0.05
12	attack 28	0.60	0.09	0.42	0.05	0.05
12	attack 29	0.62	0.19	0.60	0.11	0.05
14	(43)	0.97	0.90	0.82	1.00	0.03
14	attack 34	0.96	0.85	0.80	0.92	0.04
14	attack 35	0.96	0.85	0.80	0.91	0.04
14.2	(46)	0.97	0.88	0.80	0.97	0.03
14.2	attack 36	0.96	0.85	0.79	0.92	0.04
14.2	attack 37	0.96	0.85	0.78	0.92	0.04

Table 3: Evaluation results: The table summarizes the attack detection results over the different attack strategies and concealment techniques. The IDs in the parenthesis refer to the anomaly identifier in our dataset.

5.1.2 Effects of Concealment techniques. We test the different concealment techniques against the Autoencoder-based Detector [45].
Concealment attacks 1, 2, 44, 3, 4, 5. We can compare how the concealment attacks 44, and 5 (although they are both learning base concealment) differ based on the spoofed value received before the concealment is applied (see Figure 5c and Figure 5d blue line). Regarding the detection performance, the replay concealment (attacks 1,2,3,4) provides little hiding performance (considering that only 1 out of 39 features is replayed) in the case of Anomaly 1, while it reduces the recall by ≈ 0.15 in the case of Anomaly 1.2.

The learning-based replay reduces the recall by 0.05 in the case of anomaly 1, while it reduces it by 0.2 in the case of anomaly 1.2

Concealment attacks 6,7,8,9,10. For attacks 6, 7, 8, 9 the replay attack is launched on multiple features (see Table 2 and Appendix Figure 7a and Figure 7b), either with a value replay or a network packet replay. We note that the network packet replay (attacks 7 and 9) on multiple features fails to replay all the features. We can observe this phenomenon by comparing PU10 in attack 6 and attack 7 in Appendix Figure 7a: the green line should overlap the orange line for a successful replay of PU10. In practice, we experimentally observe that replaying multiple packets over the network causes an assertion error over the underlying CPPPO library, resulting in stale data behavior [30]. This causes a mixed behavior over features (some are replayed, and some are affected by stale data). This unpredictable behavior arises from the practical realization of concealment attacks. In Table 3, we observe that the mixed attacks 7 and 9 reduce the recall, compared to the replay attack (6 and 8), while the learning-based attack 10 has the highest evasion rate.

Concealment attacks 11, 12, 13. The concealment of this anomaly which is undetected, resulted in raising the alarm by 1%. (See Table 3 and Appendix Figure 6c).

Concealment attacks 14, 15. The replay attack 14 increases the recall of the detector, by inducing inconsistencies among the features (consistently with results from [19]). The network packet replay over multiple features (attack 15) results in certain features being stale. The mixed (stale and replay attack) results in the lowest detector recall. (See Table 3 and Appendix Figure 6d).

Concealment attacks 25, 26. The replay attack in this case results in higher detector recall (consistently with previous results from [19]). (See Table 3 and Appendix Figure 8c).

Concealment attacks 27, 28, 29. Attack 27 is detected with a higher than original recall rate (although it replays more features than attacks 25 and 26), while attack 28 results in a mixed stale, and replay behavior reduces the recall of the detector from 0.25 to 0.05. Finally, the learning-based concealment reduces the recall to 0.11. (See Table 3 and Appendix Figure 8d).

Concealment attacks 34, 35, 36, 37. The replay of feature T4 reduces the recall of the detector between 0.05 and 0.08. Overall the concealment techniques of the two attack strategies gave comparable results. (See Table 3 and Appendix Figure 9c and Figure 9d).

5.2 Summary of findings

By implementing the attacks with our framework, we show that the attacker has a number of possibilities to launch an attack on the system. First we show that the same physical effect can be achieved with different attack strategies, and this impacts the detection performance (for example as in anomaly 1 vs 1.2). Moreover, by implementing the concealment techniques with actual network traffic, we found that launching concealment techniques may result in network errors. As we found in the case of the network packet replay with multiple replayed tags (attacks 7, 9, 15 and 28), the attack might not work as expected in the network traffic causing unexpected and surprising results. The different concealment strategies result in different evasion performance at the anomaly detector.

6 Discussion

Our framework enables practical implementation of evasion attacks for CPS with network and simulated physical process. With our framework we capture for the first time the interplay of cyber and physical components involved in an evasion attack.

Implementing evasion attacks for CPS in practice is challenging for multiple reasons; in Section 3 we identified some realizability challenges that make some of the prior evasion attacks impractical. For the attacks that we implemented in Section 5, the practical implementation of evasion attacks highlights a number of options for the attacker that lead to different concealment outcomes, compared with a theoretical offline attacker. For example, we observe that multi-feature network packet replay fails due to some network checks, causing some features to get stale, leading to a different than expected behavior at the anomaly detection system. Such kind of errors might arise when launching evasion attacks on ICS networks.

Overall, our proposed framework and dataset enable a deeper understanding of evasion attacks in the field of CPS and foster the research in the field. The framework can be used to test novel evasion techniques and verify their feasibility in a realistic emulation environment. Moreover, the availability of network data can be leveraged for network intrusion detection and to design evasion attacks with the goal of evading both process and network detectors.

7 Conclusions

In this work we investigate the practical realization of evasion attacks in ICS. To address RQ1, we reviewed prior work evasion attacks and identified which attacks can be practically realized in an ICS environment. We then implemented a framework to practically simulate such attacks and collect both network and physical features. By using our framework we show that there are multiple options for an attacker to launch an evasion attack. To answer RQ2, we evaluate the effectiveness of such manipulations at three levels, namely the physical effect, the attack strategy, and the concealment technique. With our evaluation, we show the practical limitations introduced by real-world settings. For example, we show that the network packet replay of multiple PLC tags is challenging because it triggers errors in the network.

With this original research contribution we close the research gap about the practical feasibility of evasion attacks in ICS. Our proposed framework, available open source, can be further extended to evaluate the effectiveness novel evasion techniques.

Acknowledgments

The authors gratefully acknowledge funding by the Helmholtz Association (HGF) within topic "46.23 Engineering Secure Systems." This research has been partially supported by Singapore's National Satellite Of Excellence, Design Science and Technology for Secure Critical Infrastructure (NSoE DeST-SCI) through the project "LEarning from Network and Process data to secure Water Distribution Systems (LENP-WDS)" (Award No. NSoE_DeST-SCI2019-0003).

References

[1] Marshall Abrams and Joe Weiss. 2008. Malicious control system cyber security attack case study–Maroochy Water Services, Australia. *McLean, VA: The MITRE Corporation* (2008).

[2] Sridhar Adepu and Aditya Mathur. 2016. Using Process Invariants to Detect Cyber Attacks on a Water Treatment System. In *ICT Systems Security and Privacy Protection*, Jaap-Henk Hoepman and Stefan Katzenbeisser (Eds.). Springer International Publishing, Cham, 91–104.

[3] Chuadhry Mujeeb Ahmed, Martin Ochoa, Jianying Zhou, Aditya P. Mathur, Rizwan Qadeer, Carlos Murguia, and Justin Ruths. 2018. NoisePrint: Attack Detection Using Sensor and Process Noise Fingerprint in Cyber Physical Systems. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security* (Incheon, Republic of Korea) *(ASIACCS '18)*. Association for Computing Machinery, New York, NY, USA, 483–497. https://doi.org/10.1145/3196494.3196532

[4] Chuadhry Mujeeb Ahmed, Jianying Zhou, and Aditya P. Mathur. 2018. Noise Matters: Using Sensor and Process Noise Fingerprint to Detect Stealthy Cyber Attacks and Authenticate sensors in CPS. In *Proceedings of the 34th Annual Computer Security Applications Conference* (San Juan, PR, USA) *(ACSAC '18)*. Association for Computing Machinery, New York, NY, USA, 566–581. https://doi.org/10.1145/3274694.3274748

[5] Eirini Anthi, Lowri Williams, Matilda Rhode, Pete Burnap, and Adam Wedgbury. 2021. Adversarial attacks on machine learning cybersecurity defences in Industrial Control Systems. *Journal of Information Security and Applications* 58 (2021), 102717. https://doi.org/10.1016/j.jisa.2020.102717

[6] Daniele Antonioli and Nils Ole Tippenhauer. 2015. MiniCPS: A Toolkit for Security Research on CPS Networks. In *Proceedings of the First ACM Workshop on Cyber-Physical Systems-Security and/or PrivaCy* (Denver, Colorado, USA) *(CPS-SPC '15)*. Association for Computing Machinery, New York, NY, USA, 91–100.

[7] Wissam Aoudi, Mikel Iturbe, and Magnus Almgren. 2018. Truth Will Out: Departure-Based Process-Level Detection of Stealthy Attacks on Control Systems. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security* (Toronto, Canada) *(CCS '18)*. Association for Computing Machinery, New York, NY, USA, 817–831. https://doi.org/10.1145/3243734.3243781

[8] G. Apruzzese, H. S. Anderson, S. Dambra, D. Freeman, F. Pierazzi, and K. Roundy. 2023. "Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE Computer Society, Los Alamitos, CA, USA, 339–364. https://doi.org/10.1109/SaTML54575.2023.00031

[9] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. Evasion Attacks against Machine Learning at Test Time. In *Machine Learning and Knowledge Discovery in Databases*, Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 387–402.

[10] Defense Use Case. 2016. Analysis of the cyber attack on the Ukrainian power grid. *Electricity information sharing and analysis center (E-ISAC)* 388, 1-29 (2016), 3.

[11] Y. Chen, C. M. Poskitt, and J. Sun. 2018. Learning from Mutants: Using Code Mutation to Learn and Monitor Invariants of a Cyber-Physical System. In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 648–660. https://doi.org/10.1109/SP.2018.00016

[12] Seungoh Choi, Jeong-Han Yun, and Byung-Gil Min. 2021. Probabilistic Attack Sequence Generation and Execution Based on MITRE ATT&CK for ICS Datasets. In *Proceedings of the 14th Cyber Security Experimentation and Test Workshop* (Virtual, CA, USA) *(CSET '21)*. Association for Computing Machinery, New York, NY, USA, 41–48. https://doi.org/10.1145/3474718.3474722

[13] Mauro Conti, Denis Donadel, and Federico Turrin. 2021. A survey on industrial control system testbeds and datasets for security research. *IEEE Communications Surveys & Tutorials* 23, 4 (2021), 2248–2294.

[14] Jonathan Crussell, Thomas M. Kroeger, David Kavaler, Aaron Brown, and Cynthia Phillips. 2019. Lessons Learned from 10k Experiments to Compare Virtual and Physical Testbeds. In *12th USENIX Workshop on Cyber Security Experimentation and Test (CSET 19)*. USENIX Association, Santa Clara, CA. https://www.usenix.org/conference/cset19/presentation/crussell

[15] William Danilczyk, Yan Sun, and Haibo He. 2019. ANGEL: An Intelligent Digital Twin Framework for Microgrid Security. In *2019 North American Power Symposium (NAPS)*. IEEE, Wichita, KS, USA, 1–6. https://doi.org/10.1109/NAPS46351.2019.9000371

[16] Marietheres Dietz, Manfred Vielberth, and Günther Pernul. 2020. "Integrating Digital Twin Security Simulations in the Security Operations Center". In *Proceedings of the 15th International Conference on Availability, Reliability and Security* (Virtual Event, Ireland) *(ARES '20)*. Association for Computing Machinery, New York, NY, USA, Article 18, 9 pages.

[17] Matthias Eckhart and Andreas Ekelhart. 2018. Towards Security-Aware Virtual Environments for Digital Twins. In *Proceedings of the 4th ACM Workshop on Cyber-Physical System Security* (Incheon, Republic of Korea) *(CPSS '18)*. Association for Computing Machinery, New York, NY, USA, 61–72. https://doi.org/10.1145/3198458.3198464

[18] Matthias Eckhart, Andreas Ekelhart, David Allison, Magnus Almgren, Katharina Ceesay-Seitz, Helge Janicke, Simin Nadjm-Tehrani, Awais Rashid, and Mark Yampolskiy. 2023. Security-enhancing digital twins: Characteristics, indicators, and future perspectives. *IEEE Security & Privacy* 21, 6 (2023), 64–75.

[19] Alessandro Erba, Riccardo Taormina, Stefano Galelli, Marcello Pogliani, Michele Carminati, Stefano Zanero, and Nils Ole Tippenhauer. 2020. Constrained Concealment Attacks against Reconstruction-Based Anomaly Detectors in Industrial Control Systems. In *Annual Computer Security Applications Conference* (Austin, USA) *(ACSAC '20)*. Association for Computing Machinery, New York, NY, USA, 480–495.

[20] Alessandro Erba and Nils Ole Tippenhauer. 2022. Assessing Model-free Anomaly Detection in Industrial Control Systems Against Generic Concealment Attacks. In *Proceedings of the Annual Computer Security Applications Conference (ACSAC)*. ACM, Austin, USA, 412–426. https://doi.org/10.1145/3564625.3564633

[21] Alessandro Erba and Nils Ole Tippenhauer. 2023. White-Box Concealment Attacks Against Anomaly Detectors for Cyber-Physical Systems. In *Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer Nature Switzerland, Cham, 111–131.

[22] Cheng Feng, Venkata Reddy Palleti, Aditya Mathur, and Deeph Chana. 2019. A Systematic Framework to Generate Invariants for Anomaly Detection in Industrial Control Systems.. In *NDSS*. 1–15.

[23] Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. 2018. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732* (2018).

[24] Jonathan Goh, Sridhar Adepu, Marcus Tan, and Zi Shan Lee. 2017. Anomaly detection in cyber physical systems using recurrent neural networks. In *2017 IEEE 18th International Symposium on High Assurance Systems Engineering (HASE)*. IEEE, 140–145.

[25] Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel. 2017. Adversarial examples for malware detection. In *Computer Security–ESORICS 2017: 22nd European Symposium on Research in Computer Security, Oslo, Norway, September 11-15, 2017, Proceedings, Part II 22*. Springer, 62–79.

[26] Amin Hassanzadeh, Amin Rasekh, Stefano Galelli, Mohsen Aghashahi, Riccardo Taormina, Avi Ostfeld, and M Katherine Banks. 2020. A review of cybersecurity incidents in the water sector. *Journal of Environmental Engineering* 146, 5 (2020), 03120003.

[27] iTrust, Centre for Research in Cyber Security, Singapore University of Technology and Design. 2017. WADI datatset. https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/, Last accessed on: 2020-06-15.

[28] Yifan Jia, Jingyi Wang, Christopher M Poskitt, Sudipta Chattopadhyay, Jun Sun, and Yuqi Chen. 2021. Adversarial attacks and mitigation for anomaly detectors of cyber-physical systems. *International Journal of Critical Infrastructure Protection* 34 (2021), 100452.

[29] Moshe Kravchik and Asaf Shabtai. 2021. Efficient cyber attack detection in industrial control systems using lightweight neural networks and pca. *IEEE transactions on dependable and secure computing* 19, 4 (2021), 2179–2197.

[30] Marina Krotofil, Alvaro A Cárdenas, Bradley Manning, and Jason Larsen. 2014. CPS: driving cyber-physical systems to unsafe operating conditions by timing DoS attacks on sensor signals. In *Proceedings of the 30th Annual Computer Security Applications Conference*. ACM, 146–155.

[31] Olav Lamberts, Konrad Wolsing, Eric Wagner, Jan Pennekamp, Jan Bauer, Klaus Wehrle, and Martin Henze. 2023. SoK: Evaluations in Industrial Intrusion Detection Research. *arXiv preprint arXiv:2311.02929* (2023).

[32] Jiangnan Li, Yingyuan Yang, Jinyuan Stella Sun, Kevin Tomsovic, and Hairong Qi. 2021. Conaml: Constrained adversarial machine learning for cyber-physical systems. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*. 52–66.

[33] Qin Lin, Sridha Adepu, Sicco Verwer, and Aditya Mathur. 2018. TABOR: A graphical model-based approach for anomaly detection in industrial control systems. In *Proceedings of the 2018 on asia conference on computer and communications security*. 525–536.

[34] Aditya Mathur and Nils Ole Tippenhauer. 2016. SWaT: A Water Treatment Testbed for Research and Training on ICS Security. In *Proceedings of Workshop on Cyber-Physical Systems for Smart Water Networks (CySWater)*. https://doi.org/10.1109/CySWater.2016.7469060

[35] Yilin Mo and Bruno Sinopoli. 2009. Secure control against replay attacks. In *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*. IEEE, 911–918.

[36] Alvaro Cárdenas Mora, Simin Nadjm-Tehrani, Edgar Weippl, and Matthias Eckhart. 2022. Digital twins for cyber-physical systems security (Dagstuhl Seminar 22171). (2022).

[37] Andres Murillo, Riccardo Taormina, Nils Ole Tippenhauer, and Stefano Galelli. 2022. High-fidelity Cyber and Physical Simulation of Water Distribution Systems. Part 2: Enabling Cyber-Physical Attack Localization. *Journal of Water Resources Planning and Management* 149, 5 (2022).

[38] Andres Murillo, Riccardo Taormina, Nils Ole Tippenhauer, Davide Salaorni, Robert van Dijk, Luc Jonker, Simcha Vos, Maarten Weyns, and Stefano Galelli.

2022. High-fidelity Cyber and Physical Simulation of Water Distribution Systems. Part 1: Models and Data. *Journal of Water Resources Planning and Management* 149, 5 (2022).
[39] Iman Niazazari and Hanif Livani. 2020. Attack on grid event cause analysis: An adversarial machine learning approach. In *2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. IEEE, 1–5.
[40] Dionysios Nikolopoulos, Georgios Moraitis, Dimitrios Bouziotas, Archontia Lykou, George Karavokiros, and Christos Makropoulos. 2020. Cyber-Physical Stress-Testing Platform for Water Distribution Networks. *Journal of Environmental Engineering* 146, 7 (2020), 04020061.
[41] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 372–387.
[42] Ali Pinar, Thomas Tarman, Laura Painton Swiler, Jared Gearhart, Derek Hart, Eric Vugrin, Gerardo Cruz, Bryan Arguello, Gianluca Geraci, Bert Debusschere, Seth Hanson, Alexander Outkin, Jamie Thorpe, William Hart, Meghan Sahakian, Kasimir Gabert, Casey Glatter, Emma Johnson, and She'ifa Punla-Green. 2021. *Science and Engineering of Cybersecurity by Uncertainty quantification and Rigorous Experimentation (SECURE) (Final Report)*. Technical Report. Sandia National Laboratories. https://doi.org/10.2172/1821322
[43] R. Taormina and S. Galelli and H.C. Douglas and N.O. Tippenhauer and E. Salomons and A. Ostfeld. 2019. A Toolbox for Assessing the Impacts of Cyber-physical Attacks on Water Distribution Systems. *Environmental Modelling I& Software* 112 (2019), 46 – 51.
[44] Luis Salazar, Sebastian Castro, Juan Lozano, Keerthi Koneru, Emmanuele Zambon, Bing Huang, Ross Baldick, Marina Krotofil, Alonso Rojas, and Alvaro Cardenas. 2024. A tale of two Industroyers: It was the season of darkness. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 162–162.
[45] Riccardo Taormina and Stefano Galelli. 2018. Deep-Learning Approach to the Detection and Localization of Cyber-Physical Attacks on Water Distribution Systems. *Journal of Water Resources Planning and Management* 144, 10 (2018), 04018065. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000983
[46] Riccardo Taormina, Stefano Galelli, Nils Ole Tippenhauer, Elad Salomons, Avi Ostfeld, Demetrios G Eliades, Mohsen Aghashahi, Raanju Sundararajan, Mohsen Pourahmadi, M Katherine Banks, et al. 2018. Battle of the Attack Detection Algorithms: Disclosing Cyber Attacks on Water Distribution Networks. *Journal of Water Resources Planning and Management* 144, 8 (2018), 04018048.
[47] David I. Urbina, Jairo A. Giraldo, Alvaro A. Cardenas, Nils Ole Tippenhauer, Junia Valente, Mustafa Faisal, Justin Ruths, Richard Candell, and Henrik Sandberg. 2016. Limiting the Impact of Stealthy Attacks on Industrial Control Systems. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)* (Vienna, Austria). 1092–1105. https://doi.org/10.1145/2976749.2978388
[48] Sharon Weinberger. 2011. Computer security: Is this the start of cyberwarfare? *Nature* 174 (June 2011), 142–145.
[49] Giulio Zizzo, Chris Hankin, Sergio Maffeis, and Kevin Jones. 2020. Adversarial attacks on time-series intrusion detection for industrial control systems. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 899–910.

A Background

A.1 Anomaly Detection in ICS

The security of ICS is challenging as industrial protocols often do not implement security features, or the field devices do not support security features such as authentication and encryption. Attacks against ICS occurred in the past, with the goal of disrupting the physical process [26]. Examples of such attacks are Stuxnet [48], Maroochy [1], Black energy [10], Industroyer [44].

Anomaly detection was proposed to detect abnormal physical process data, which are symptom of an ongoing attack (referred to as process-based anomaly detection). A number of detection techniques were proposed to identify ongoing process anomalies. Such anomaly detection systems operate in two phases. In the first phase, the detector is trained over sensor readings collected during normal operating conditions. In the second phase, the learned physical patterns are compared to the observed sensor readings and the detector reports if the sensor readings are anomalous or benign. Examples of such anomaly detection systems are deep learning-based detectors [24, 29, 45], Invariant-based detectors [2–4, 22, 33], Machine Learning [7, 11], and control invariants [47].

We note that the focus of this work is the evasion of process based anomaly detection for ICS, other approaches like network intrusion detection are out of scope.

A.2 Anomaly Detection Datasets

To conduct ICS process-based anomaly detection research, a number of datasets are available, collected from various ICS sources, e.g, water, gas and energy systems. Some are collected at real-world testbeds while others are collected via numerical simulation. The works by Conti et al. [13] and Lamberts et al. [31] provide a comprehensive overview of available datasets for CPS security research. Table 1 reports a summary of available datasets in the water sector and compares them with the dataset we make available with this paper. Our dataset is a collection of physical and network data containing evasion attacks. Moreover, by using our framework, more data can be collected under different attack configurations.

A.3 Digital Twins for Security Research

Digital Twins are of outmost importance for ICS cyber security research [36]. Digital twins enable security design and testing of variuos aspects of CPS [18]. Over the years, a number of digital twins architectures for CPS have been proposed. In this section we provide a summary of related digital twin research.

DHALSIM was inspired by epanetCPA [43], which uses a simple implementation of actuators and sensors that does not offer any network emulation capabilities. RISKNOUGHT [40] was also inspired by epanetCPA [43]; this framework offers limited network emulation capabilities by offering ACK signals. Eckhart et al. [17] present a framework that relies on AutomationML to standardize digital twin generation, operation, and experimentation. The framework is also presented with a Digital Twin proof of concept that uses MiniCPS. ANGEL [15], is a Digital Twin to evaluate the security of microgrids. ANGEL uses Matlab Simulink to offer physical simulation capabilities, but does not emulate an industrial network. SCEPTRE [42] is a digital twin for electrical systems that employs the virtual machine orchestration platform Minimega [14] to offer industrial network emulation capabilities and supports multiple power simulator. The work by Dietz et al. [16] integrates MiniCPS [6] with a Security Information and Event Management (SIEM) system. The integration of such SIEM tools offers a standardised way of experimentation by also combining threat intelligence sources such as MITRE ATT&CK [12].

A.4 Evasion Attacks

An evasion attack [9] is defined as follows. Given a classifier f that classifies if the sensor readings x from the ICS are 'anomalous', an attacker launches an evasion attack by finding a perturbation $x' = x + \delta$ such that $f(x')$ = 'benign'. Evasion attacks have been demonstrated over various machine-learning tasks, ranging from image classification [9] to malware detection [25].

A.5 Related Work

In Section 3, we discussed related work that proposed evasion attacks against ICS anomaly detectors. We now complement by

Attack	Attacker knowledge	Attacker Constraints	Realizability Challenge
Replay attack [35]	Black-box	Spoof subset of the features	-
Anthi et al. [5]	Gray-box	Subset of the features (random) & perturbation budget	Attacker knows the anomaly's physical effect
Li et al. [32]	White-box	Linear contraints, random subset of the features	Attacker knows the anomaly's physical effect
Jia et al. [28]	White-box	Attack only sensors	Attacker knows the future & alters the past
Niazazari et al. [39]	White-box	Subset of features	Attacker knows the future & alters the past
Erba et al. A [19]	White&black-box	Subset of features (based on error & topology)	-
Erba et al. B [20]	Black-box	Subset of features	-
Erba et al. C [21]	White-box	Subset of features	-
Zizzo et al. [49]	White-box	Subset of features (based on error)	Attacker knows future

Table 4: Comparison of prior work evasion attackers. Some of the attacks present some realizability challenge, thus preventing the implementation of such attacks. The remaining attacks are implemented in our framework.

discussing related works that discuss the real world feasibility and practicality of evasion attacks.

Apruzzese et al [8] highlighted the gap between adversarial machine learning research in academia and industry. The work offers a comprehensive overview and case studies with real-world adversarial examples, which are unlikely to be computed with adversarial machine learning techniques (which are computationally expensive for the purpose). Instead, an attacker may look for cheaper options, which work as well against the target system and do not require optimizations. This is consistent with Erba et al. [20] findings, who showed evasion of a wide range of ICS detectors by using pre-computed patterns that were not optimized against the target model. Gilmer et al. [23] discuss how adversarial examples are often far from being practical, and discusses realistic threat models.

B Attack description

Table 2 summarizes each attack contained in the dataset. For each attack, we describe the physical anomaly, the attack strategy, and the concealment techniques we applied.

Physical Anomaly 1. This anomaly aims to cause low levels of water level in the water tank T7. As described in the example in Section 4.5, we achieve the goal with two different attack strategies. We refer to the two alternative strategies with 1 and 1.2 (this is valid also for the other anomalies that are implemented with two attack strategies, i.e., 2 and 14). We then apply a number of concealment strategies to hide the anomaly from the anomaly detection system.

For the first anomaly we launch three concealment strategies.

- Concealment 1. The value of T7 is spoofed to launch a value replay attack.
- Concealment 2. The value of T7 is spoofed to launch a network packet replay attack.
- Concealment 44. The values are spoofed by applying the learning-based concealment from [19].

For the second anomaly we launch three concealment strategies.

- Concealment 3. The value of T7 is spoofed towards the SCADA via a value replay attack.
- Concealment 4. The value of T7 is spoofed towards the SCADA to launch a network packet replay attack.
- Concealment 5. The values are spoofed by applying the learning-based concealment from [19].

Physical Anomaly 2. The second anomaly is implemented as the first anomaly, again with two alternative attack strategies. Differently from Anomaly 1, we implement a set of different evasion attacks strategies.

For the first anomaly we launch three concealment strategies.

- Concealment 6. The value of T7, PU10 and PU11 are replayed towards the SCADA with a value replay attack.
- Concealment 7. The value of T7, PU10 and PU11 are replayed towards the SCADA with a network packet replay attack.

For the second anomaly, we launch three concealment strategies.

- Concealment 8. The values of T7, PU10 and PU11 are spoofed to launch a value replay attack.
- Concealment 9. The values of T7, PU10 and PU11 are spoofed to launch a network packet replay attack.
- Concealment 10. The values are spoofed by applying the learning-based concealment from [19].

Physical Anomaly 3. The goal of the anomaly 3 is to overflow the water tank T1, this is achieved by launching a MiTM attack on PLC2 by spoofing a low level of T1 which will force the pumps on. We apply the following concealment towards the SCADA:

- Concealment 11. The value of T1 is spoofed to launch a value replay attack.
- Concealment 12. The value of T1 is spoofed to launch a network packet replay attack.
- Concealment 13. The values are spoofed by applying the learning-based concealment from [19].

Physical Anomaly 4. This physical anomaly achieves the same effects of Anomaly 3, but it lasts for a longer period of time. We apply the following concealment strategies towards the SCADA:

- Concealment 14. The values of T1, PU1, PU2, PJ269 are spoofed to launch a value replay attack.
- Concealment 15. The values of T1, PU1, PU2, PJ269 are spoofed to launch a network packet replay attack.

Physical Anomaly 9. The goal of this physical anomaly is to cause a Tank 2 overflow. In order to do so, the attacker spoofs the value of T2 to a constant low level causing a tank overflow.

Two concealment attacks are launched in this case:

- Concealment 25. The value of T2 is spoofed to launch a value replay attack.
- Concealment 26. The value of T2 is spoofed to launch a network packet replay attack.

Physical Anomaly 12. This physical anomaly achieves the same effects as Anomaly 9 and uses the same attack strategy. To hide the anomaly the following concealment attacks are tested.

Two concealment attacks are launched in this case:

- Concealment 27. The values of T2, V2, PJ14, PJ422 are spoofed to launch a value replay attack.
- Concealment 28. The values of T2, V2, PJ14, PJ422 are spoofed to launch a network packet replay attack.
- Concealment 29. The values are spoofed by applying the learning-based concealment from [19].

Physical Anomaly 14. This anomaly overflows tank 4 (In the original BATADAL dataset the reported target tank was T6, but since T6 is not controlled in C-Town we target T4). To do so, we implement two attack strategies. The first is a MiTM attack between PLC6 and PLC3 to spoof low values of T4 and let PU6 and PU7 open. We conceal this attack via:

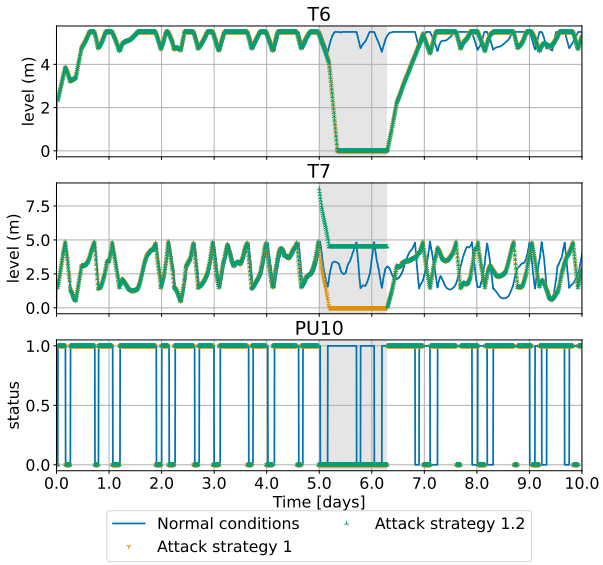
- Concealment 34. The value of T4 is spoofed to launch a value replay attack.
- Concealment 35. The value of T4 is spoofed to launch a network packet replay attack.

The second is a PLC3 attack to keep open pumps PU6 and PU7.

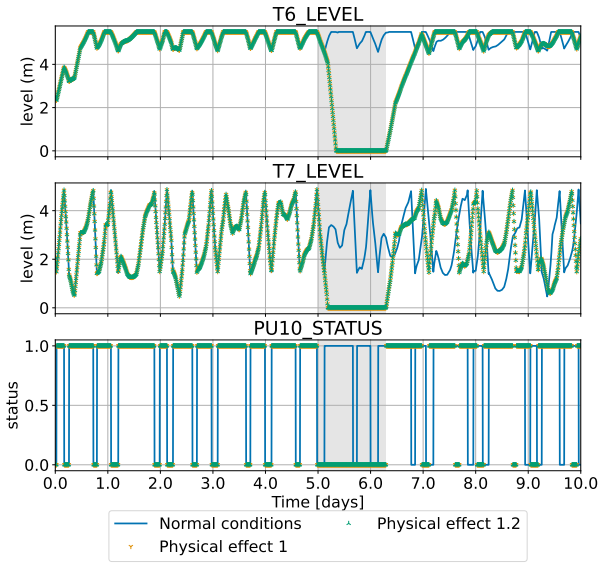
- Concealment 36. The value of T4 is spoofed to launch a value replay attack.
- Concealment 37. The value of T4 is spoofed to launch a network packet replay attack.

C Attacks Visualization

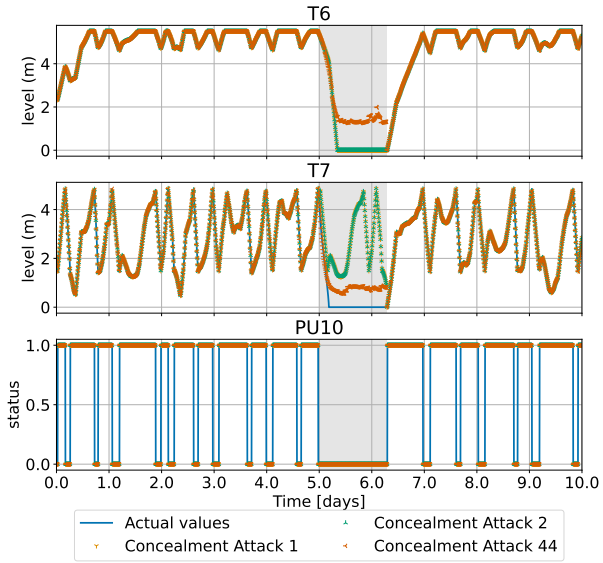
We provide additional plots to visualize the effect of the attacks at the three levels, physical process, SCADA and concealment.



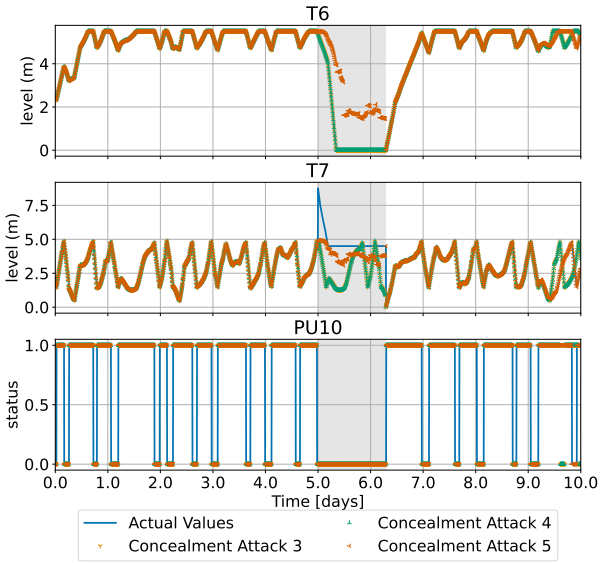
(a) Attack strategy at the SCADA, comparison between anomaly 1 and anomaly 1.2 strategies and normal operating conditions.



(b) Physical Effect, comparison between anomaly 1 and anomaly 1.2 attack strategies and normal operating conditions

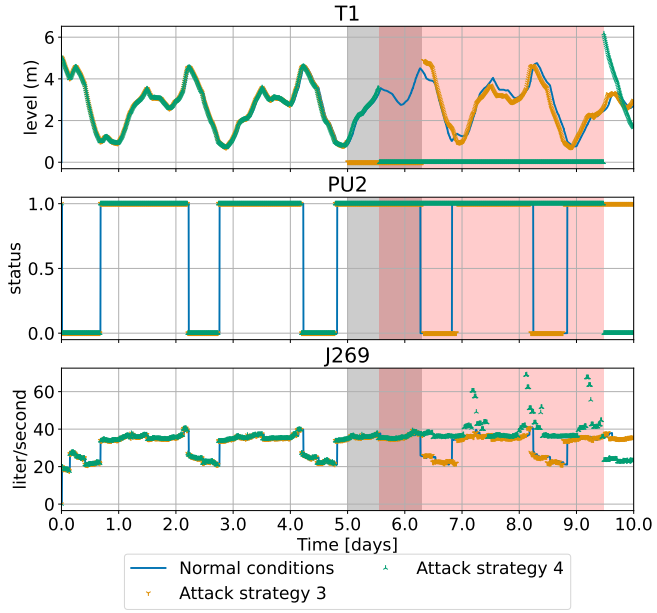


(c) Concealment attacks strategy at the SCADA, comparison between Concealment attack 1, 2, 44 and the actual values before applying the concealment (Anomaly 1).

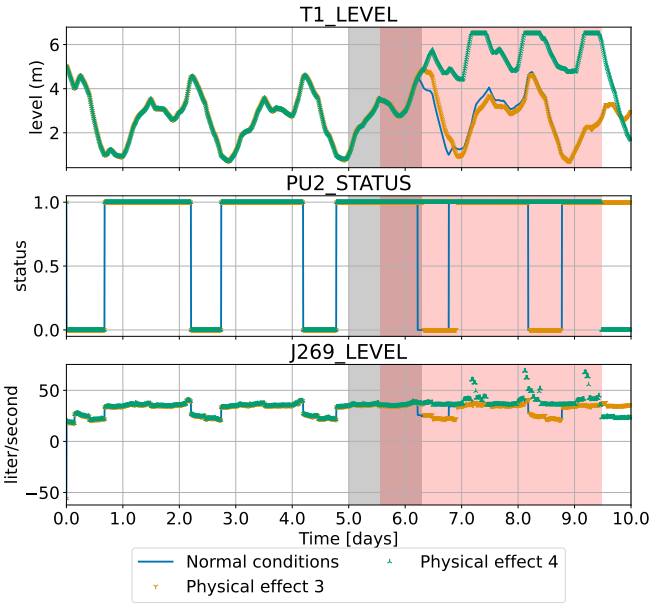


(d) Concealment attacks strategy at the SCADA, comparison between Concealment attack 3, 4, 5 and the actual values before applying the concealment (Anomaly 1.2).

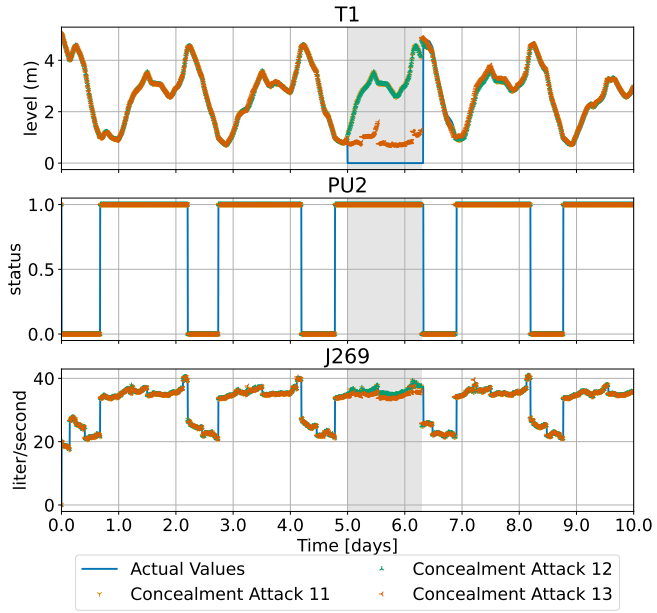
Figure 5: Visualization of Anomalies 1 and 1.2 and concealment attacks 1, 2, 44, 3, 4, 5. The gray bar indicates the moment where the physical anomaly and concealment are launched on the system.



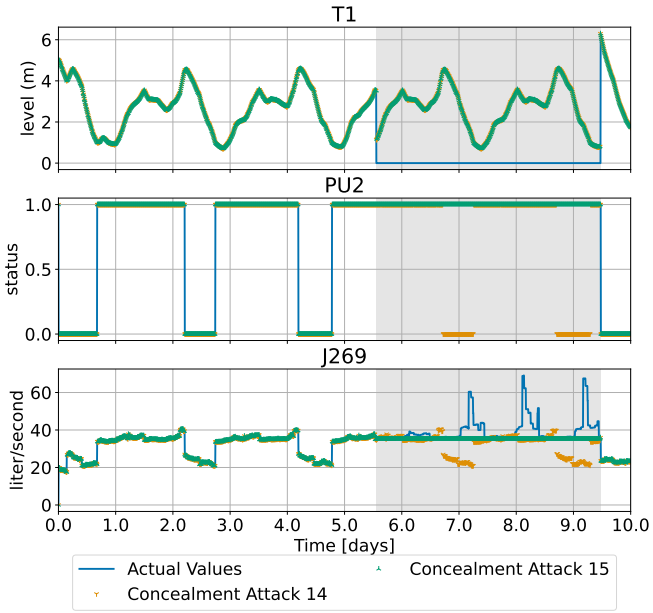
(a) Attack strategy at the SCADA, comparison between anomaly 3 and anomaly 4 attack strategies and normal operating conditions.



(b) Physical Effect, comparison between anomaly 3 and anomaly 4 attack strategies and normal operating conditions

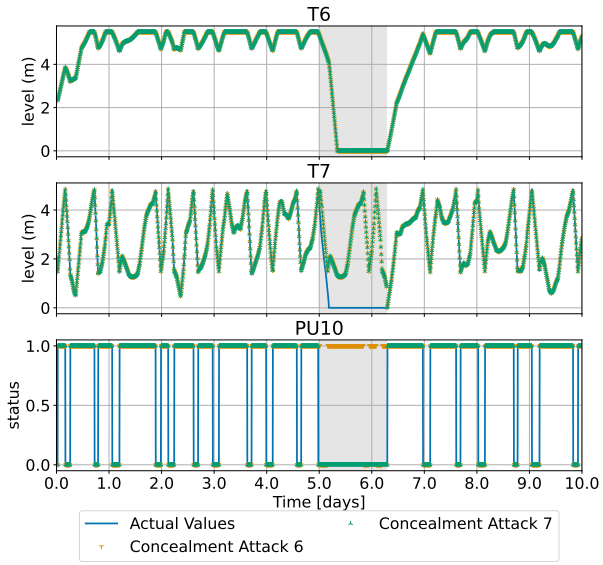


(c) Concealment attacks strategy at the SCADA, comparison between Concealment attack 11, 12, 13 and the actual values before applying the concealment (Anomaly 3).

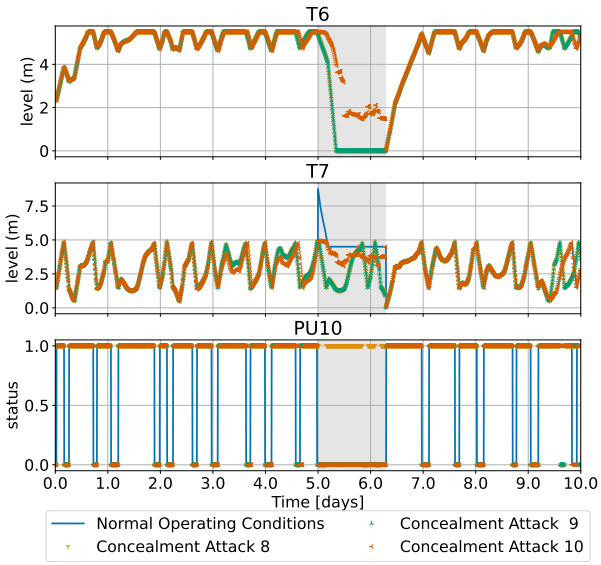


(d) Concealment attacks strategy at the SCADA, comparison between Concealment attack 14, 15 and the actual values before applying the concealment (Anomaly 4).

Figure 6: Visualization of Anomalies 3 and 4 and concealment attacks 11, 12, 13, 14, 15. The gray bar indicates the moment where the physical anomaly 3 and concealment are launched on the system. The red bar is related to attack 4

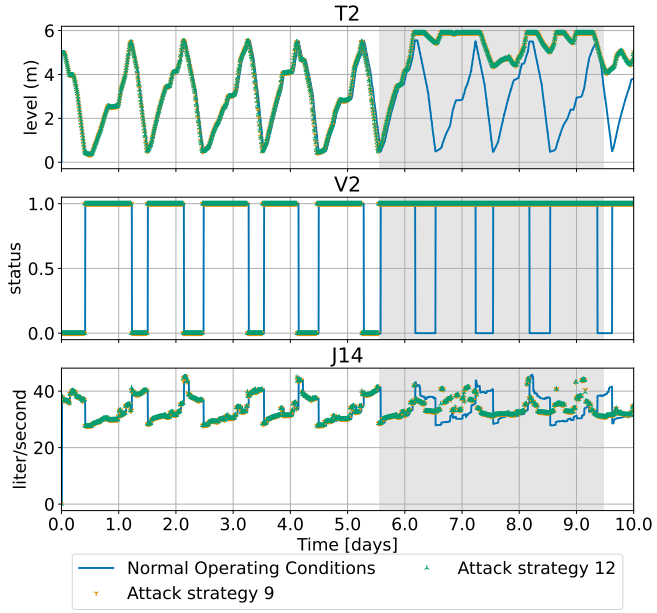


(a) Concealment attacks strategy at the SCADA, comparison between Concealment attack 6, 7 and the actual values before applying the concealment (Anomaly 1).

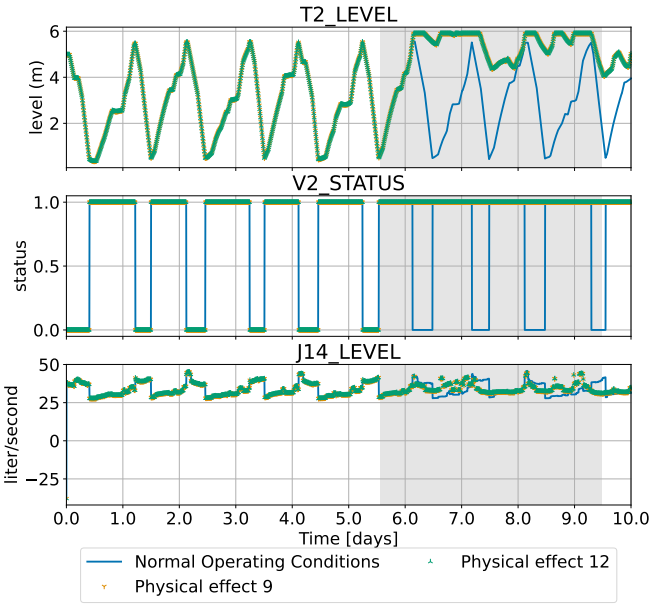


(b) Concealment attacks strategy at the SCADA, comparison between Concealment attack 8, 9, 10 and the actual values before applying the concealment (Anomaly 1.2).

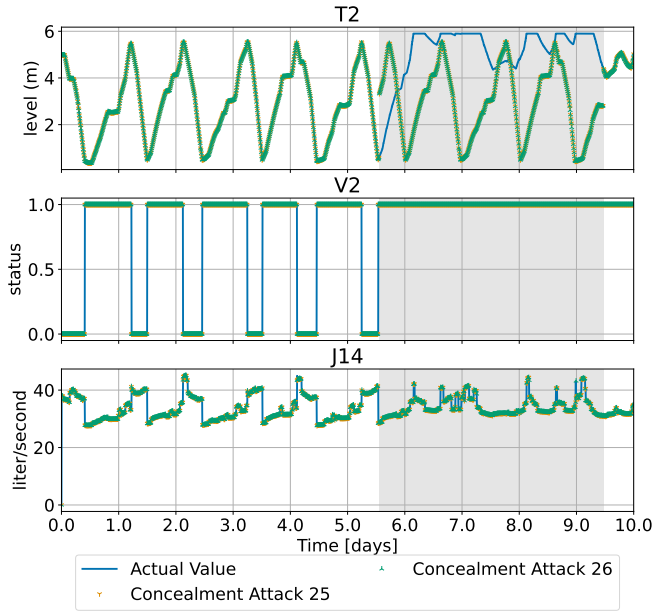
Figure 7: Visualization of concealment attacks 6, 7, 8, 9, 10. The gray bar indicates the moment where the physical anomaly and concealment are launched on the system. We can observe that despite attack 6 and 7 should achieve the same concealment effect, they differ on the feature PU10, we found that multi feature network packet replay will not always succeed, triggering some network errors, which result in stale data observed on the feature.



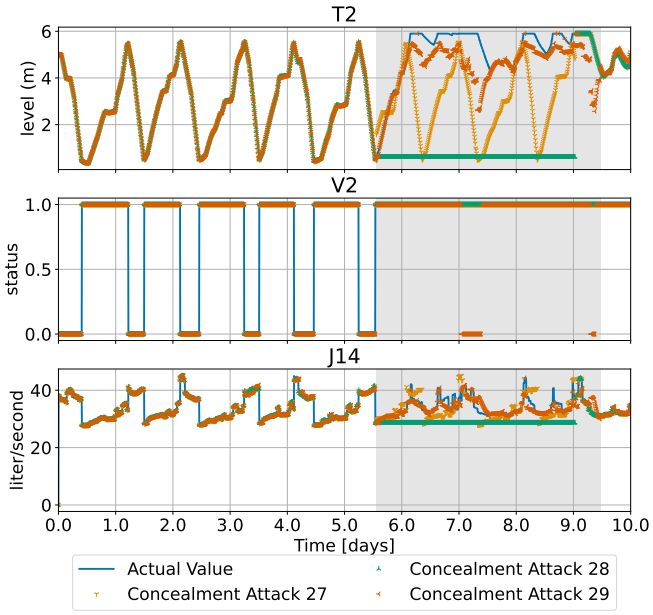
(a) Attack strategy at the SCADA, comparison between anomaly 9 and anomaly 12 attack strategies and normal operating conditions.



(b) Physical Effect, comparison between anomaly 9 and anomaly 12 attack strategies and normal operating conditions

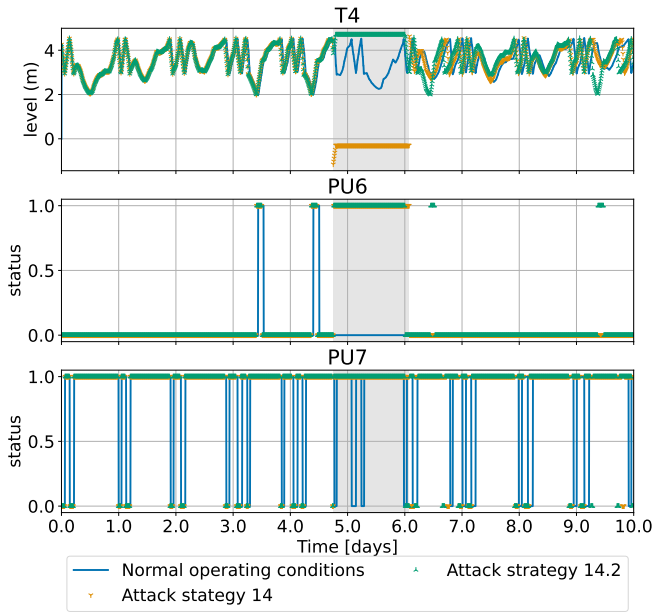


(c) Concealment attacks strategy at the SCADA, comparison between Concealment attack 25, 26 and the actual values before applying the concealment (Anomaly 9).

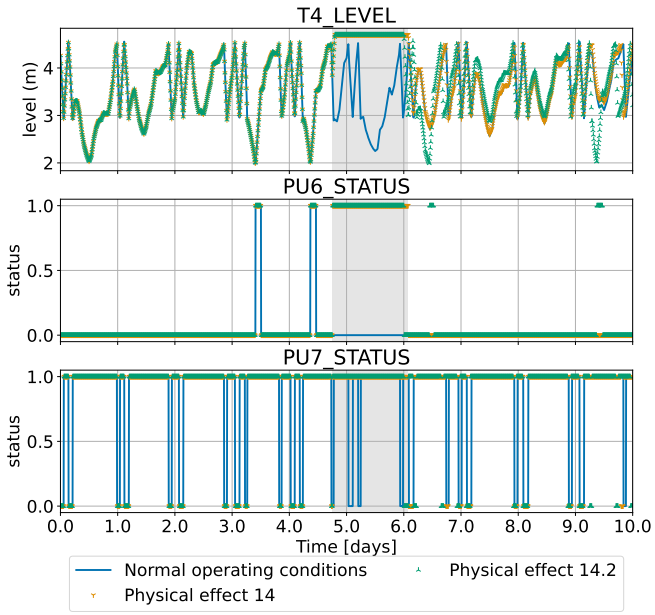


(d) Concealment attacks strategy at the SCADA, comparison between Concealment attack 27, 28, 29 and the actual values before applying the concealment (Anomaly 12).

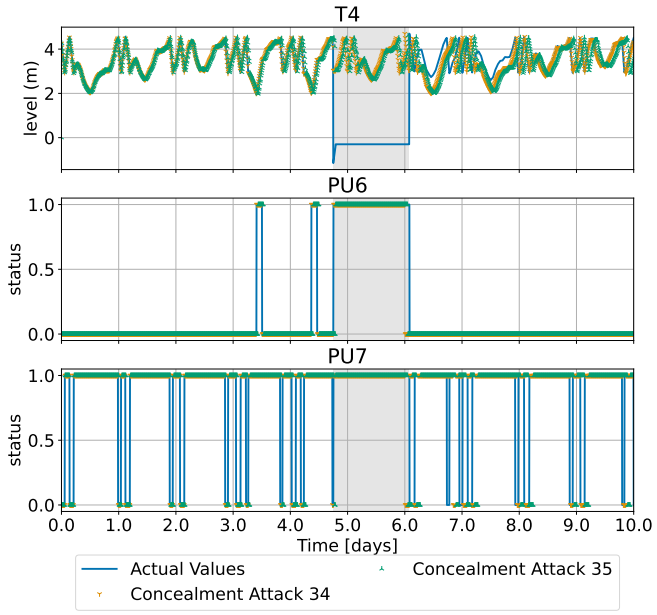
Figure 8: Visualization of Anomalies 9 and 12 and concealment attacks 25, 26, 27, 28, 29.



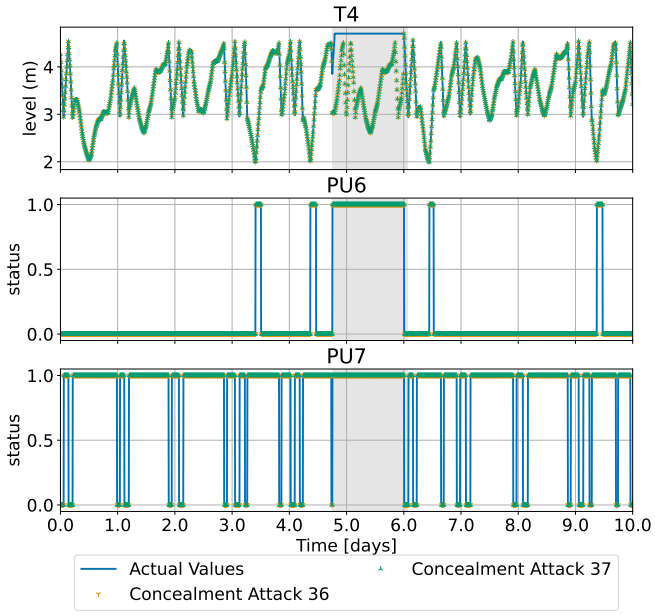
(a) Attack strategy at the SCADA, comparison between anomaly 14 and anomaly 14.2 attack strategies and normal operating conditions.



(b) Physical Effect, comparison between anomaly 14 and anomaly 14.2 attack strategies and normal operating conditions



(c) Concealment attacks strategy at the SCADA, comparison between Concealment attack 34, 35 and the actual values before applying the concealment (Anomaly 14).



(d) Concealment attacks strategy at the SCADA, comparison between Concealment attack 36, 37 and the actual values before applying the concealment (Anomaly 14.2).

Figure 9: Visualization of Anomalies 14, and 14.2 and concealment attacks 34, 35, 36 and 37.