METRICS: A Methodology for Evaluating and Testing the Resilience of Industrial Control Systems to Cyberattacks

Lennart Bader[1,2][0000−0001−8549−1344], Eric Wagner[1,2][0000−0003−3211−1015], Martin Henze[3,1][0000−0001−8717−2523], and Martin Serror[1][0000−0002−6925−5744]✉

[1] Cyber Analysis & Defense, Fraunhofer FKIE, Wachtberg, Germany
firstname.lastname@fkie.fraunhofer.de
[2] Communication and Distributed Systems, RWTH Aachen University, Aachen, Germany lastname@comsys.rwth-aachen.de
[3] Security and Privacy in Industrial Cooperation, RWTH Aachen University, Aachen, Germany henze@cs.rwth-aachen.de

Abstract. The increasing digitalization and interconnectivity of industrial control systems (ICSs) create enormous benefits, such as enhanced productivity and flexibility, but also amplify the impact of cyberattacks. Cybersecurity research thus continuously needs to adapt to new threats while proposing comprehensive security mechanisms for the ICS domain. As a prerequisite, researchers need to understand the resilience of ICSs against cyberattacks by systematically testing new security approaches without interfering with productive systems. Therefore, one possibility for such evaluations is using already available ICS testbeds and datasets. However, the heterogeneity of the industrial landscape poses great challenges to obtaining comparable and transferable results. In this paper, we propose to bridge this gap with METRICS, a methodology for systematic resilience evaluation of ICSs. METRICS complements existing ICS testbeds by enabling the configuration of measurement campaigns for comprehensive resilience evaluations. Therefore, the user specifies individual evaluation scenarios consisting of cyberattacks and countermeasures while facilitating manual and automatic interventions. Moreover, METRICS provides domain-agnostic evaluation capabilities to achieve comparable results, which user-defined domain-specific metrics can complement. We apply the methodology in a use case study with the power grid simulator Wattson, demonstrating its effectiveness in providing valuable insights for security practitioners and researchers.

Keywords: Industrial control systems · Security evaluations · Testbeds · Datasets · Resilience.

1 Introduction

The ongoing shift from local, isolated ICSs toward highly interconnected networks currently affects all areas of industrial automation, such as manufacturing systems, process control, and power grids [39]. This trend fosters enhanced

productivity, higher flexibility, and potentially better safety while reducing installation and maintenance costs [9]. On the downside, however, it increases the dependence between individual components and amplifies the harmful impact of cyberattacks. Even worse, it is largely known that ICSs exhibit significant cybersecurity deficits, mainly due to the challenges of retrofitting modern security mechanisms to long-lived legacy hardware with stringent latency and availability requirements [29]. Furthermore, ICSs are an attractive target for financially or politically motivated criminals who make use of constantly evolving attack vectors [24]. Cybersecurity research for ICSs must hence continuously adapt countermeasures and responses to keep pace with this development and even anticipate new threats when proposing preventive measures.

As a first step toward this ambitious goal, researchers and security practitioners need a profound understanding of current cyberattacks and countermeasures in ICSs and how these affect the underlying physical processes. Based on such resilience evaluations, they can identify and address existing weaknesses and, in the event of a cyberattack, select the best available response, i.e., repelling the attack while maintaining the operation of the ongoing industrial process as best as possible. Nevertheless, conducting cybersecurity research in productive ICSs is, in most cases, not a viable option due to safety concerns and the high availability requirements of the involved systems [10]. Consequently, cybersecurity researchers increasingly rely on ICS testbeds and datasets for performing resilience evaluations, e.g., the Secure Water Treatment (SWaT) testbed [23] or the HIL-based augmented ICS security (HAI) dataset [30]. However, using these tools to conduct comparable cybersecurity research remains challenging due to their heterogeneous landscape manifesting in substantial discrepancies regarding accuracy, scalability, and flexibility [12]. Moreover, the gained insights depend on made assumptions, the necessary abstractions, and the considered use cases, emphasizing the need for comparative evaluations. Hence, a general evaluation methodology for (available) ICS testbeds is missing, facilitating comprehensive and comparable resilience evaluations of such systems.

Therefore, in this paper, we propose METRICS, a combined **M**ethodology for **E**valuating and **T**esting the **R**esilience of **ICS**s to cyberattacks. Our proposed methodology facilitates automated resilience evaluations for given ICS testbed environments with defined attacker's capabilities and response mechanisms by systematically testing different options and configurations. A given ICS testbed may range from a physical setup to an entirely virtual environment (e.g., a simulator) where the respective testbed exposes its capabilities and configuration possibilities to METRICS in a cross-domain environment description format. For the evaluation, we distinguish between domain-agnostic metrics, such as the reachability of system and network components, which independently apply to every testbed, and domain-specific metrics, which individually apply to the given testbed and thus must be provided along with the testbed description. The *evaluation control* then enables users to configure distinct scenarios and facilitates manual and automated interventions in running evaluations. The evaluation results eventually converge into a *presentation layer*, providing insights and vi-

sualizations of ongoing evaluations. Moreover, METRICS retrieves datasets of each evaluation run, enabling subsequent analyses. This methodology thus helps to systematically identify weaknesses in current ICS deployments, assess the potential impact of cyberattacks, and improve the respective countermeasures.

In particular, this paper covers the following *contributions*:

- We analyze the requirements for an evaluation methodology concerning cybersecurity research for ICSs (Section 2);
- We propose METRICS, a comprehensive methodology to facilitate the resilience evaluation of ICSs to cyberattacks by providing comparable evaluation metrics (Section 3); and
- We present and discuss initial evaluation results by extensively studying a use case within the power grid simulator WATTSON [1] consisting of distinct attack vectors and countermeasures (Section 4).

Our use case evaluation demonstrates that METRICS offers valuable insights for security practitioners and researchers by facilitating a systematic iteration through possible configuration options while allowing manual and automatic interventions. Moreover, we identify the remaining challenges toward achieving universal resilience evaluation of ICSs in Section 5. In the following, we take a closer look at the fundamentals of cybersecurity research for ICSs before deriving the requirements and challenges for a comprehensive evaluation methodology.

Availability Statement. For better transparency of our conducted evaluation and enabling further research, our evaluation artifacts are publicly available: https://wattson.it/METRICS

2 Cybersecurity Research for ICSs

Productive ICSs are typically unavailable for cybersecurity research due to the high availability requirements and safety concerns [10, 12]. Security researchers and engineers thus rely on testbeds and datasets to model ICSs and conduct the evaluations in a safe environment. Figure 1 depicts the interplay between security research, testbeds, and datasets for ICSs [6]. Testbeds model real ICSs in prototypical deployments using hardware, virtual components, or a combination. Furthermore, they can provide relevant recordings of network traffic and process states in the form of datasets, which, in turn, represent specific evaluation scenarios. Both concepts are thus valuable means for security research, facilitating testing and evaluation, depending on the respective level of abstraction and the considered research questions. Several literature surveys confirm the increasing availability of ICS testbeds and datasets and, moreover, summarize the complementary benefits of the distinct concepts [16, 6]. In the following, we briefly present the methodological features of each concept in the ICS domain while putting a special focus on evaluating the resilience to cyberattacks.

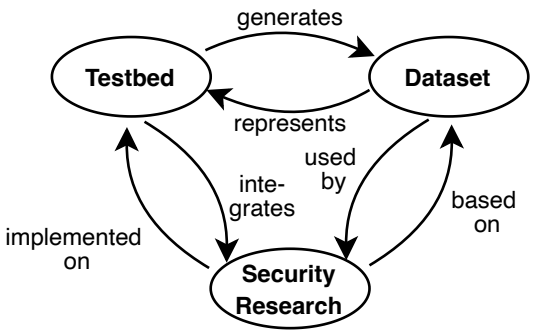

Fig. 1. Interplay of security research, testbeds, and datasets for ICSs showing that they mutually depend and benefit from each other. (Figure adapted from [6].)

2.1 Testbeds

ICS testbeds offer a protected research environment for cybersecurity research by replicating (parts of) productive ICSs. They typically consist of physical or virtual components where any combination and level of abstraction are possible [16, 6]. Thus, their concrete design depends on their individual purpose and the requirements for the considered research questions.

While testbeds relying on physical components are generally close to reality and provide high accuracy, they are typically limited in flexibility and scalability. Moreover, their deployment is costly and sometimes requires extensive maintenance. In turn, testbeds relying on virtual components, which can be realized by simulation or emulation approaches, are significantly cheaper and more flexible but sometimes do not provide real-time capabilities. Moreover, scalability must often be traded against achieved accuracy when designing a virtual testbed. Examples of the broad range of possible ICS testbeds include the Secure Water Treatment (SWaT) testbed [23], the security-focused yet universal EPS-ICS testbed [10], and the power grid co-simulator WATTSON [1].

When striving to evaluate the resilience to cyberattacks, the respective ICS testbed needs to fulfill specific requirements to assess the impact of cyberattacks and the effectiveness of possible countermeasures. These mainly refer to achieving high accuracy of the modeled ICS, i.e., a comprehensive representation of the physical processes and the underlying information and communication technologies, to also capture unanticipated side effects. Moreover, extensive traceability of the conducted experiments facilitates complex resilience analyses, where recording datasets plays a decisive role, as further explained in the following.

2.2 Datasets

ICS datasets represent specific scenarios of the considered systems, resulting from a particular configuration and a predefined measurement time. They typ-

ically include recordings of network traffic, process states, and possibly meta-information about the scenario [6]. Such recordings facilitate, on the one hand, systematically analyzing the impact of cyberattacks and countermeasures post hoc. On the other hand, recorded datasets may help to improve the prevention and detection of cyberattacks, most prominently for training and testing of intrusion detection systems (IDSs) [35]. Although desirable, ICS datasets are rarely available from productive ICSs, mainly for protecting the confidentiality of industrial processes. Therefore, their generation and provision are typically closely related to the availability of ICS testbeds.

Generally, two possibilities exist to generate ICS datasets with cyberattacks [5]. One is to perform the cyberattacks directly in an ICS testbed and record the respective data. The other is to record a scenario without cyberattacks and inject (synthetic) attack data into the recordings afterward. While the latter is also possible for datasets from productive ICSs during normal operation, there is a risk of obtaining inaccurate or inconsistent data [6]. Regardless of how the dataset was obtained, labeling normal and abnormal data within the dataset is extremely helpful, e.g., when using the data for IDSs. Examples of such ICS datasets are the HIL-based augmented ICS security (HAI) dataset [30] or the PowerDuck dataset focusing on GOOSE traffic in an electrical substation [38].

Concerning resilience evaluations, ICS datasets thus provide evidence for a detailed assessment of the countermeasures' effectiveness. Nevertheless, their full potential can only be exploited in combination with their ICS testbed, facilitating flexible adaptions of the measurement scenarios and, therefore, systematic resilience evaluations. In the following, we review related work and derive the requirements for such a comprehensive evaluation methodology.

2.3 Related Work

Evaluating and assessing system resilience, and especially the resilience of ICSs, has been identified as an important topic by both past and ongoing research [3]. Related work can be divided into research that *conducts* resilience evaluations of respective systems [1] and research proposing *evaluation methodologies* [27], where both aspects are also *combined* for certain research areas [37]. As the resilience of physical systems, e.g., buildings, railway networks, or power grids, has been an active research area for multiple decades [4], ICS-related research can seize the gained insights and transfer them into the ICS domain. For instance, Haque et al. [13, 14] adapt the well-known framework for seismic resilience by Bruneau et al. [4], defining *sub-metrics* ("the four *R*s") for ICS resilience in three dimensions (physical, organizational, technical) [13]. While their approach targets the whole ICS domain, it does not provide a concrete definition of sub-metrics, e.g., *redundancy*, as such a metric heavily depends on the concrete ICS. On the other hand, related work focusing on the resilience of a specific ICS [1] provides concrete metrics for the respective ICS without considering the transferability of results to other domains. Thus, a cross-domain methodology for comparably evaluating the resilience of ICSs is still missing.

2.4 Toward a Cross-Domain ICS Evaluation

We recognize the need for a methodology allowing systematic analyses of ICSs with comparable and reproducible results, especially concerning resilience evaluations. Such a methodology combines testbeds and datasets to facilitate the creation of accurate and safe research environments, an invaluable feature for the ICS domain. While testbeds enable modeling of ICS and realistic impact evaluations, datasets are especially useful for post hoc analyses. Moreover, we identify the following desirable design requirements for such a methodology:

Universality. It applies to diverse testbeds facilitating resilience evaluations for the entire ICS domain.

Accuracy. It supports precise representations of specific ICSs, enabling meaningful modeling of cyberattacks and countermeasures.

Assessability. It allows the integration of domain-agnostic and domain-specific metrics to promote the comparability between distinct testbeds.

Traceability. It has the ability to export datasets for retracing the evaluation results, conducting post hoc analyses, and verification by others.

Hence, the evaluation methodology must cater to the wide range of ICSs, all exhibiting distinct susceptibilities and resiliencies to various cyberattacks. Further, different countermeasures and responses might be of varying success for such systems. Thus, universally evaluating their resilience to cyberthreats remains an open challenge. Despite their differences, potential cyberattacks and countermeasures are applicable and relevant across multiple ICSs, but their actual implementations might vary. Similarly, an evaluation metric must always be defined based on domain-specific knowledge to account for the actual *impact* of attacks and countermeasures. In the next section, we propose such a comprehensive evaluation methodology while also providing details on the distinct design components and the challenges when implementing them.

3 METRICS: A Cybersecurity Evaluation Methodology for ICSs

In this section, we present METRICS, a two-layered approach for achieving an ICS domain-spanning evaluation methodology. METRICS leverages the commonalities of attack and countermeasure strategies while respecting the differences and specifics of each ICS domain to allow directly evaluating cyberattacks and responses as well as generating datasets for subsequent analyses.

Figure 2 depicts the design overview of METRICS, where we distinguish between a domain-specific *evaluation environment* and a (mostly) domain-agnostic *evaluation control*. When considering a specific ICS, a corresponding evaluation environment is required, which may range from a physical testbed over a hybrid setup to a simulation. This environment allows evaluating the desired system under test (SUT) by representing the ICS, implementing adversaries and responses, and providing insights into the system's state. In turn, the evaluation

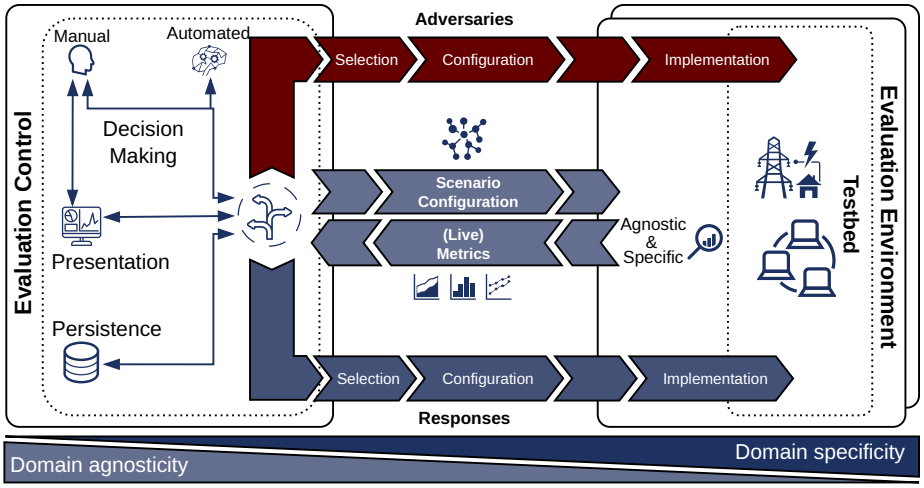

Fig. 2. METRICS' design leverages a *domain-agnostic* evaluation control which interacts with a *domain-specific* evaluation environment. This environment wraps a testbed for representing the desired ICS, implements adversaries and responses, and provides insights into their effects in the form of metrics. In evaluation control, decisions for adjustments of adversaries and responses are made based on these metrics which are further presented to the user, and persisted for later analyses.

control manages the evaluation environment by configuring the desired scenario, including adversaries and responses, and receiving reported (live) metrics. Configuration options and received metrics are visualized for user interaction. Based on metrics reported to the evaluation control, manual and automated decisions may adjust the current evaluation or schedule new ones. We now detail METRICS' components and their interactions.

3.1 Exchangeable Evaluation Environment

To fulfill the design requirements of Section 2.4, METRICS supports exchangeable, domain-specific evaluation environments in the form of physical or virtual testbeds (cf. Section 2.1). Consequently, a supported testbed must be (i) *accurate* w.r.t. its real-world equivalent, (ii) *observable* w.r.t. both the network traffic and the physical processes, and (iii) *cybersecurity-focused* to allow conducting cyberattacks and integrating individual responses. Moreover, depending on the considered use cases, there might be some additional desirable properties: (iv) *flexibility* w.r.t. the domain-specific scenarios that can be reproduced, and (v) *scalability* w.r.t. the supported network size and number of components.

The evaluation environment must expose its capabilities and configuration options for METRICS in a universally applicable *environment description file (EDF)*. This file defines available topologies, metrics, assets and their roles, as

well as adversary and response actions and their configurations. Appendix A provides an example illustrating the structure of such an EDF.

When configured for a specific scenario with potential *adversaries* and *responses*, the evaluation environment then implements the behavior of the SUT and provides insights into the state and effects of all components and their interactions. The adversaries, responses, and metrics all have domain-specific and domain-agnostic aspects. While abstract metrics, e.g., the availability of network nodes, can be applied to several domains, their concrete definition depends on the domain-specific context. Thus, we now specifically focus on the implications for adversaries and responses as well as cross-domain metrics.

3.2 Adversaries and Responses

The evaluation environment needs to represent cyberattacks *and* potential responses to *accurately* enable the resilience evaluation of ICSs. In this context, cyberattacks range from simple physical attacks, e.g., destroying or disconnecting hardware [17], over to network attacks, e.g., denial-of-service (DoS) attacks [31], up to process-aware attacks, e.g., false data injection (FDI) attacks [18]. Consequently, potential responses may include, e.g., external perimeter security systems [28], IDSs [33], or lightweight authentication schemes [22]. While most concepts of attacks and responses apply to various ICSs, their technical details, implementations, and effects differ between scenarios. Thus, we explicitly consider the resulting implications within METRICS to combine both, domain-specific implementations with domain-agnostic and generalizable concepts to comparably evaluate different ICSs. Hence, the evaluation environment provides concrete implementations for adversaries and responses, defines valid configuration options, and maps them to common concepts. To exemplify these design aspects, we now discuss them for both adversaries and responses in more detail.

Adversaries. A critical attack on ICSs is an FDI attack [26]. Here, attackers interfere with the ongoing communication to manipulate exchanged (application-layer) information, e.g., sent measurements or control commands as a machine-in-the-middle (MitM). For METRICS, this inline network payload manipulation concept is quite domain-agnostic, as such attacks apply to various ICSs. Their implementation, however, is very specific and depends on the actual ICS and its individual properties. First, establishing the technical requirements for conducting an FDI attack differ. While an ARP-spoofing attack might be appropriate for Ethernet-based networks [25], bus-based networks might require dedicated timing techniques [36], whereas base station spoofing might be applicable for wireless networks [20]. Second, the manipulation of process information depends on the used application-layer protocol as well as the use of encryption and message authentication mechanisms. Thus, successfully implementing an FDI attack depends on the domain and might differ within a given heterogeneous domain.

Responses. Like the adversary design, preventive and reactive responses follow domain-agnostic concepts but require domain- and scenario-specific realizations:

Integrity protection, encryption, or intrusion detection apply to various ICSs, while their implementation and configuration require domain-specific information. A process-aware IDS is specific to its target domain, i.e., IDSs related to manufacturing follow different approaches than, e.g., an IDS for power grid networks. Consequently, we divide adversaries and responses into a domain-agnostic (concept) selection, a concept-specific configuration, and a domain-specific implementation (cf. Figure 2). Similarly, we propose a cross-domain approach for metrics for comparative evaluations of different ICSs, as detailed in the following.

3.3 Cross-domain Metrics

Comparably assessing the impact of cyberattacks and the effectiveness of countermeasures requires appropriate metrics as desired by the *assessability* design requirement. For ICSs, defining such metrics is particularly challenging since effects can cover both the networking and the physical part of the system. The differences and specifics of each ICS further exacerbate the comparability of results across different ICSs. Thus, we propose differentiating between domain-specific and domain-agnostic metrics, similar to the adversary and response definitions.

Metric Requirements. For each ICS, the evaluation environment should provide domain- or even instance-specific metrics. Such metrics provide valuable and detailed insights into the system, allowing in-depth evaluations of system-specific characteristics and effects. However, they complicate automated decision-making when selecting (iterative and reactive) adversaries and responses, further hindering comparing certain results from different domains or instances. In METRICS, we flexibly address these challenges in three ways: (i) each evaluation environment may provide automated decision-making algorithms that enhance its domain awareness, (ii) implementations and configurations of adversaries and responses may include domain-specific metrics to adjust their behavior automatically, and (iii) each evaluation environment should provide abstract concepts for its domain-specific metrics. While the two former aspects primarily require implementation effort, the latter focuses on conceptual aspects.

The domain-specific metrics provide detailed insights into the specific system. However, their interpretation often requires specific knowledge of the SUT, which hinders comparability across domain boundaries. Therefore, we encourage the domain experts to provide domain-agnostic abstractions from these detailed metrics that follow a *normalized cross-domain specification* and allow non-domain experts to understand and interpret them.

Exemplary Cross-Domain Metric. We use a metric for *network operability* as an example. Such a metric applies to various ICSs and provides insights into potential impairments of the network's desired operation. While, for some domains, the number or fraction of operational network nodes might be well-suited to represent the network's operability, other ICSs might define this metric

based on available network paths between application layer nodes or even the number of reachable nodes from a single source.

While their abstract design enables such metrics to apply to multiple independent ICSs and allows researchers to compare them across different domains, they cannot provide the in-depth details that domain-specific metrics can. Thus, we explicitly include both domain-agnostic and domain-specific metrics in METRICS to enable cross-domain comparisons as well as in-depth evaluations.

3.4 Evaluation Control

METRICS includes an evaluation control to provide a cross-domain interface for researchers to evaluate different ICSs under different adversary and response concepts. Designed as a primarily domain-agnostic component, it allows researchers to define their desired ICS scenario, choose from adversaries, responses, or abstracted concepts of those, and control the evaluation environment. It fulfills three primary tasks, as detailed in the following.

Scenario Configuration. Before starting the evaluation, researchers must select and configure the desired ICS scenario. By selecting the targeted domain (e.g., power grids) and the domain-specific scenario (e.g., the grid's actual topology), the adversary concepts and implementations as well as responses, researchers can precisely define their evaluation parameters. For generic yet comparable cross-domain evaluations and detailed domain-specific insights, this configuration process allows both the selection of domain-agnostic adversary and response concepts and the choice of domain-specific variants based on the *environment description*. Besides this *static* (e.g., playbook-based) configuration of adversaries and responses, METRICS also considers on-demand decision-making for live interactions and adjustments. METRICS defines the *scenario description file (SDF)* analogously to the EDF to allow the configuration of distinct evaluation scenarios. In Appendix B, we provide an example of an SDF.

Decision-Making. For in-depth research, *dynamically* influencing the running evaluation represents a valuable feature. On-demand decision-making, e.g., based on live metrics, can influence the running evaluation and instantiate new adversaries and responses or re-configure existing ones. Researchers can make these decisions directly or automate them with domain-agnostic and domain-specific implementations. Examples of such automated decision-making include rule-based approaches [7] or machine learning [2]. Further, the human-in-the-loop could also re-configure the automated decision-making to follow different strategies.

The metrics provided by the evaluation environment are of particular importance for the decision-making process. While domain-specific metrics allow respective experts to choose corresponding adversaries and responses carefully, more generic and domain-agnostic metrics allow for cross-domain automation implementations, easing large-scale evaluations of multiple ICS domains. Since these insights into the ongoing evaluation are the primary input for all decisions, METRICS has to also present these insights to the human-in-the-loop.

Scenario and Result Presentation. We include a dedicated *presentation layer* within the evaluation control providing (live) insights and visualizations into the ongoing evaluation. Comparing metrics and observing their variation during the evaluation eases the human-based decision-making processes, providing a desirable feature for the evaluations. Besides the presentation and visualization of (live) metrics, the presentation layer also covers the scenario configuration, i.e., it provides insights into the domain-specific scenario, offers viable configuration options, and presents applicable adversary and response concepts to the researchers. Moreover, it allows researchers to extract datasets from completed simulation runs, thus covering the *traceability* design requirement.

Overall, all metrics of the evaluation environment are (i) used as input for decision-making, (ii) presented to and visualized for researchers, and (iii) persisted for later in-depth analyses. Thus, METRICS enables researchers to conduct individualized, in-depth evaluations of specific ICS domains and scenarios, to implement flexible yet automated evaluations of different scenarios and multiple domains, and to compare their results with analyses from other researchers with potentially different focuses. We now present a concrete use case example to emphasize the concept of METRICS and evaluate its value.

4 Use Case: METRICS for Power Grids

We apply the concepts and methodologies of METRICS to evaluate the effects of cyberattacks and respective countermeasures in a power grid network. To represent the SUT, we use Wattson [1], a co-simulator focusing on cybersecurity for power grids. We use a small medium-voltage reference grid (Cigre MV [32]) along with the corresponding information and communication technologies (ICT) network as the base scenario. Figure 3 visualizes the power grid with the corresponding ICT network. Moreover, we provide the EDF for Wattson and the SDF of the considered use case in the evaluation artifacts[4].

Our evaluation consists of multiple phases, where adversaries and responses are iteratively established or adjusted, following METRICS' basic idea of dynamic adjustments based on (live) insights into the SUT's behavior. In particular, the evaluation phases alternate between adversary and response actions.

4.1 Evaluation Phases

We now emphasize the details of each phase in our exemplary use case, the effects on both the network and the physical process, and how the phases interconnect. In Figure 4, we visualize domain-specific and domain-agnostic metrics for both the ICT network and the power grid during the evaluation.

Phase 1: Reference. The first phase is the reference phase, where the power grid operates normally without adversaries. We program the control center to

[4] https://wattson.it/METRICS

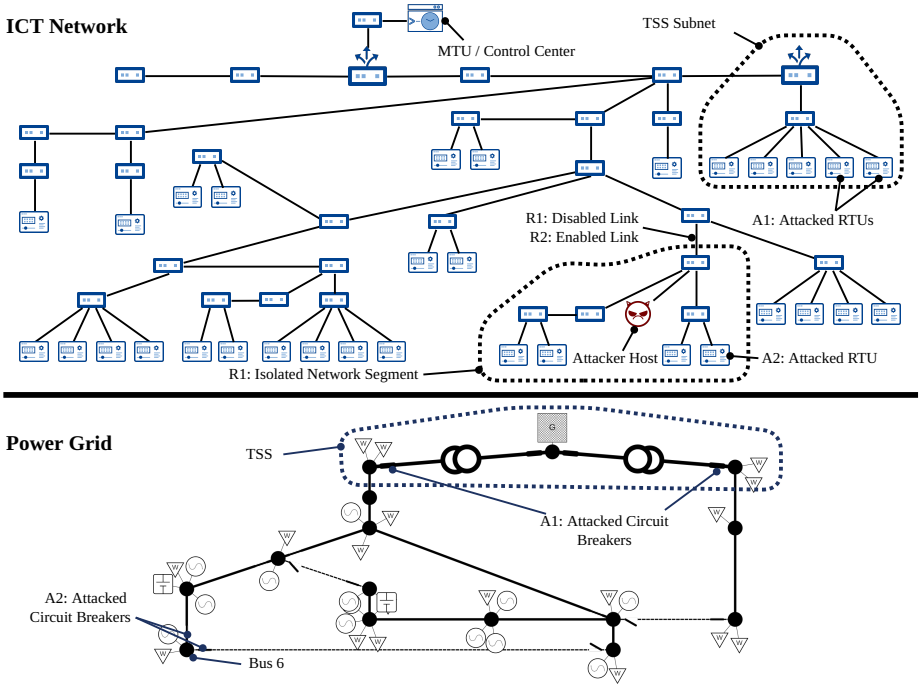

Fig. 3. The ICT network follows a tree-like topology consisting of three different subnets. The attacker host is attached to a switch within a DSS, a common attack vector for power girds [19]. In the first phase (**A1**), it connects to RTUs in the TSS and disconnects the majority of the grid. After its connection to these RTUs is blocked (**R1**) by isolating the network segment of the attacker host, the attack targets a still-reachable RTU to disconnect Bus 6 (**A2**). Finally, the operator configures all RTUs to block unauthorized connections (**R2a**) before reenabling the previously disabled link (**R2b**).

issue control commands to keep the grid connected, i.e., closing or opening circuit breakers as needed. In this phase, all RTUs are connected to the control center, and all buses in the power grid operate normally, resulting in grid availability and network operability metrics of 100 %.

Phase 2: Industroyer (A1). The first attack is conducted at 20 s into the evaluation. A new host is connected to a switch at a (remote) DSS. On this host, a variant of the infamous *Industroyer* [8] malware is executed. This malware targets power grid networks by connecting to RTUs and issuing malicious control commands. In past attacks [15], these control commands were crafted to disconnect circuit breakers at TSSs, essentially disconnecting entire parts of the power grid. During our evaluation, we follow its real-world behavior, such that the malware connects to two RTUs at the power grid's central TSS and issues control commands to open multiple circuit breakers. As a result, several

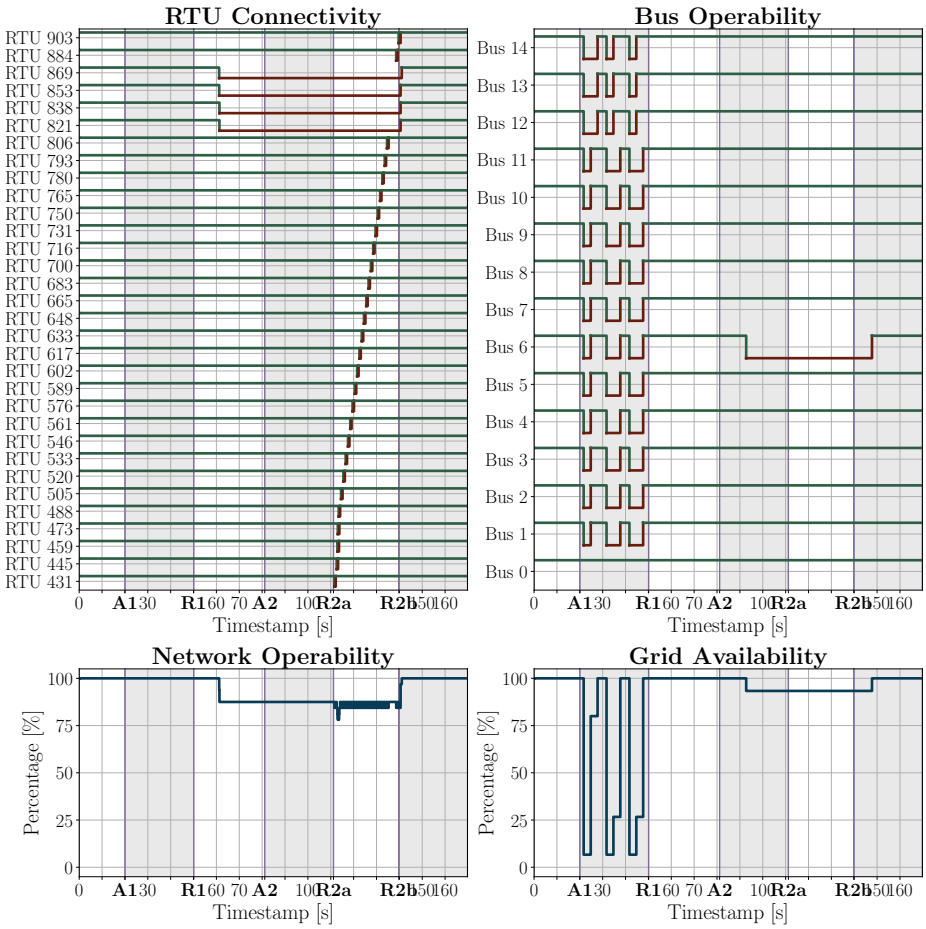

Fig. 4. The initial Industroyer attack (**A1**) repeatedly opens the circuit breakers at the transformers in the TSS. Although the control center issues respective counter commands, the grid's availability repeatedly drops significantly as most of the grid is disconnected. In contrast, the ICT network is not negatively affected during this phase. After the operator disables a link in response to the ongoing attack (**R1**), four RTUs lose their *connectivity* (domain-specific), resulting in reduced *network operability* (domain-agnostic). Since the grid operator gains back control over the previously attacked RTUs, the grid availability returns to 100 %. The second Industroyer attack (**A2**) only targets a single DSS RTU as the first response (**R1**) blocks the attacker from connecting to the TSS RTUs. As a result, Bus 6 becomes inoperable, and the grid's availability drops slightly. During the reconfiguration of all RTUs to enable client authentication (**R2a**), all previously connected RTUs shortly lose their connection to the control center. After the link is re-enabled (**R2b**), the grid is fully available again and all RTUs re-establish their connections to the control center, resulting in a network operability of 100 %. While the domain-specific *RTU connectivity* and *bus operability* metrics provides more detailed insights, the domain-agnostic *network operability* and *grid availability* metrics allow insights into the attack effects for non-experts.

buses become inoperable, significantly reducing the grid's availability. While the commands issued by the control center temporarily restore the grid availability, the malware continues to issue commands disconnecting most of the grid.

Phase 3: Preliminary Response (R1). The effects of the conducted attack are evident to the grid operator as large parts of the power grid get disconnected. At 50 s into the evaluation, the grid operator takes down a network link between the attackers' host and the attacked RTUs. However, since the precise origin of the attack is not (yet) determinable by the operator, a whole segment of the network is affected by the disabled link. As a result, four previously unaffected RTUs lose their connectivity, reducing the network operability. Since the attackers can no longer attack the RTUs in the TSS, the grid operator regains sole control over this TSS and can restore the grid's availability.

Phase 4: Industroyer Take 2 (A2). After the attackers' host lost connection to the attacked RTUs, the attackers adjust their behavior at 80 s into the evaluation. As a result of the disabled link, the Industroyer host can only reach those RTUs that are part of the disconnected network segment. Hence, the malware is reconfigured to attack an RTU within a reachable DSS to disconnect the associated bus (Bus 6), actively reducing the grid's availability. Since the affected RTU is not reachable by the control center, no immediate commands as a countermeasure are possible. Further, as no measurements from the RTU reach the control center, the second attack is not as obviously detectable as the attack of Phase 2 (A1).

Phase 5a: Client Authentication (R2a). While the preliminary response (Phase 3) reduced the impact of the attack on the grid's availability significantly, it is not sufficient to recover the reference state (Phase 1) as several RTUs are unavailable and one DSS is inoperable. Since the Industroyer malware connects as a secondary IEC 60870-5-104 client to the RTUs, this revised response enables (simple) client authentication within the network. To this end, the operator reconfigures all (reachable) RTUs to only accept connections from the IP address of the master terminal unit (MTU) in the control center. Starting at $\approx$111 s, each RTU is reconfigured individually, which resets all active connections. This process is visible in Figure 4, where these short connection losses are observable.

Phase 5b: Link Reactivation (R2b). As soon as all RTUs are reconfigured, the operator reactivates the previously disabled link at $\approx$140 s. Connections to the previously unreachable RTUs can be reestablished and the client authentication can be enabled. Therefore, the Industroyer malware, which is still active, is disconnected from the targeted RTU and can no longer establish a new connection. Consequently, the grid operator regains full control over all RTUs and can restore the grid's availability. After all network and power grid effects are averted, the malicious host can be permanently physically removed based on its position in the network.

4.2 Discussion

The evaluation of the presented use case provides valuable insights concerning the specific SUT, i.e., the cybersecurity of power grids, and the methodological approach and application of METRICS. In the following, we equally discuss these different aspects.

Cybersecurity in Power Grids. Past cyberattacks against power grid networks highlight the potentially drastic effects of such attacks and common vulnerabilities within such networks [19, 34]. Our use case evaluation highlights multiple aspects relevant to cybersecurity in power grids. First, the geographical size of such networks represents a unique challenge for securing such networks. Numerous potentially unmanned remote locations increase the risk of physical access to network assets [19]. Physical protection and the appropriate configuration of such assets are required to minimize this risk. Second, remote visibility and controllability are of paramount importance [40]. While fine-granular visibility allows identifying attacks early, controllability of network assets provides the possibility to remotely implement appropriate countermeasures to ongoing attacks. Third, the protection of process information is essential but challenging [1]. The lack of encryption and command authentication enables attackers to conduct attacks such as the presented Industroyer attack or more advanced false data injection attacks [21]. Cryptographic authentication of control commands can prevent semantic attacks that aim to manipulate the physical process over the communication network [1]. However, since power grids have stringent real-time requirements and must always ensure process safety and availability, adapted security solutions are necessary, fully adhering to these requirements.

Specific and Agnostic Metrics. In METRICS' design, we introduced both *domain-specific* and *domain-agnostic* metrics to provide insights into the SUT. For the exemplary evaluation, we follow this concept and provide a domain-specific and domain-agnostic metric for the communication network and the power grid states. The domain-specific metrics, i.e., the *Bus Operability* and the *RTU Connectivity*, provide detailed insights into the SUT. They show the number of covered assets and individually state their respective states. These insights are especially valuable for domain experts and when comparing several variants of the same scenario during an evaluation series. However, their interpretation for researchers from different domains is challenging. Consequently, we include domain-agnostic metrics for the network and the physical process: With a normalized value range (0 % – 100 %) and abstraction from the actual number of assets, these metrics offer comparability and eased interpretation for non-experts at the cost of reduced specificity. Since both variants of metrics offer valuable insights into the SUT, we assess their combined usage as favorable: While domain-agnostic metrics offer comprehensibility and comparability, detailed evaluations always require using domain-specific metrics.

METRICS' Iterative Methodology. The phase-based use case evaluation highlights the potential for METRICS' iterative evaluation methodology. While distinct phases allow us to observe the effects of each attack and response, their iterative structure enables flexible evaluation of different adversary and response behaviors. As visualized in Figure 4, we can observe the delay of certain effects (e.g., as for phase **A2**) as well as effects that span across multiple phases (e.g., multiple disconnected nodes after **R1**). Thus, METRICS provides a flexible yet structured approach for conducting cybersecurity evaluations for complex ICSs. In particular, they support grid operators in understanding the varying impact of cyberattacks on their configurations and consequently reacting more effectively in case of actual attacks.

5 Toward Cross-Domain Resilience

With METRICS, we address the demand for a cross-domain evaluation methodology regarding the resilience of ICSs against cyberattacks. Acknowledging the need for domain-specific metrics and insights as well as domain-agnostic (i.e., comparable and transferable) insights, METRICS considers individual requirements for adversary, response, and metric designs. However, deriving a comprehensive *resilience score* from metrics and evaluation results remains an open challenge. As identified by related work from the ICS domain and different research areas [4, 13], *resilience* depends on and consists of multiple aspects. While these aspects, such as robustness or redundancy, have been identified to influence the resulting system resilience, their respective definitions and weights still depend on the concrete ICS domain or even the specific instance of an ICS. Thus, we assess the derivation of concrete yet universal resilience definitions as an essential research area, which can be divided into several aspects.

First, for a concrete instance of a specific ICS, a comprehensive measure or metric for resilience has to be derived by identifying and assessing factors that influence the system's resilience. Here, *resilience* depends on the specific scenario, e.g., the tasks and features of the ICS and the presence of specific adversaries and response mechanisms. Further, multiple definitions of a system's *resilience* might be appropriate or even necessary.

Second, combining these ICS- and scenario-specific insights into an overall resilience score, i.e., a resilience measure for a specific ICS, is necessary. Since different adversaries and responses might affect various aspects of a complex ICS, weighting individual resilience measures is particularly challenging.

Third, abstracting the definitions for a specific ICS or ICS domain to enable cross-domain comparisons promises valuable and comparable insights into the strengths and weaknesses of different ICS domains. Identifying different resiliencies of distinct domains paves the way for applying successful concepts from different ICSs to strengthen the overall security and resilience of ICSs. In this context, we plan to apply METRICS to further industrial domains, starting with aquaponics [11], to identify universally applicable concepts as well as incompatibilities between acquaponics and power grid ICSs.

With METRICS, we thus foster the proposed research areas by providing a comprehensive evaluation methodology, enabling researchers to gather comparable insights into various ICSs under flexible scenarios. Based on these results, assessing the *resilience* of a specific instance, a single ICS domain, and ICSs as a whole are the next steps toward enhanced resilience of ICSs.

6 Conclusion

Spurred by the current need to improve cybersecurity in complex, interconnected ICSs, we propose METRICS, a methodology for evaluating and testing the resilience of ICSs to cyberattacks. Our approach provides a framework to integrate existing ICS testbeds while obtaining comparable evaluation results. We introduce domain-specific and domain-agnostic metrics considering the specific properties of an ICS, as well as a normalized cross-domain assessment. Security researchers and practitioners can perform systematic resilience evaluations by specifying distinct scenarios consisting of adversaries and responses and including manual and automatic interventions to influence the running evaluations.

In a preliminary case study, we demonstrate the feasibility and potential of METRICS using the power grid simulator Wattson. The results are twofold: On the one hand, they reveal the benefits of an iterative approach to understanding the impact of a cyberattack and figuring out the best possible responses. On the other hand, they help identify (recurring) weaknesses in current ICS deployments, which can be subsequently addressed to prevent actual attacks. However, leveraging METRICS' full potential requires further advances in the specification of applicable resilience metrics and, as a next step, we intend to use METRICS for performing comparative evaluations using different ICS testbeds. With this in mind, we are convinced that METRICS represents a valuable contribution toward addressing the long-neglected security deficiencies in ICSs.

Acknowledgements. This paper was supported by the EDA Cyber R&T project "Cyber Electromagnetic Resilience Evaluation on Replicated Environment (CERERE)", funded by Italy and Germany.

References

1. Bader, L., Serror, M., Lamberts, O., Sen, O., van der Velde, D., Hacker, I., Filter, J., Padilla, E., Henze, M.: Comprehensively Analyzing the Impact of Cyberattacks on Power Grids. In: European Symposium on Security and Privacy. IEEE (2023)
2. Bhattacharya, A., Ramachandran, T., Banik, S., Dowling, C.P., Bopardikar, S.D.: Automated Adversary Emulation for Cyber-Physical Systems via Reinforcement Learning. In: IEEE Int'l Conf. on Intelligence and Security Informatics (ISI) (2020)
3. Bodeau, D.J., Graubart, R.D., McQuaid, R.M., Woodill, J.: Cyber Resiliency Metrics, Measures of Effectiveness, and Scoring: Enabling Systems Engineers and Program Managers to Select the Most Useful Assessment Methods. Tech. rep., Mitre Corp Bedford Ma Bedford United States (2018)

4. Bruneau, M., Chang, S.E., Eguchi, R.T., Lee, G.C., O'Rourke, T.D., Reinhorn, A.M., Shinozuka, M., Tierney, K., Wallace, W.A., Von Winterfeldt, D.: A Framework to Quantitatively Assess and Enhance the Seismic Resilience of Communities. Earthquake Spectra **19**(4), 733–752 (2003)
5. Choi, S., Yun, J.H., Min, B.G.: Probabilistic Attack Sequence Generation and Execution Based on MITRE ATT&CK for ICS Datasets. In: Cyber Security Experimentation and Test Workshop. CSET '21, ACM (2021)
6. Conti, M., Donadel, D., Turrin, F.: A Survey on Industrial Control System Testbeds and Datasets for Security Research. IEEE Comm. Surveys & Tutorials **23**(4) (2021)
7. Deloglos, C., Elks, C., Tantawy, A.: An Attacker Modeling Framework for the Assessment of Cyber-Physical Systems Security. In: Computer Safety, Reliability, and Security. pp. 150–163. Springer International Publishing, Cham (2020)
8. ESET Research: Industroyer2: Industroyer reloaded. We Live Security (2022), https://www.welivesecurity.com/2022/04/12/industroyer2-industroyer-reloaded/
9. Galloway, B., Hancke, G.P.: Introduction to Industrial Control Networks. IEEE Communications Surveys & Tutorials **15**(2) (2013)
10. Gao, H., Peng, Y., Jia, K., Dai, Z., Wang, T.: The Design of ICS Testbed Based on Emulation, Physical, and Simulation (EPS-ICS Testbed). In: Int'l Conf. on Intelligent Information Hiding and Multimedia Signal Processing. IEEE (2013)
11. Goddek, S., Körner, O.: A fully integrated simulation model of multi-loop aquaponics: A case study for system sizing in different environments. Agricultural Systems **171**, 143–154 (2019)
12. Green, B., Lee, A., Antrobus, R., et al.: Pains, Gains and PLCs: Ten Lessons from Building an Industrial Control Systems Testbed for Security Research. In: USENIX Workshop on Cyber Security Experimentation and Test. CSET '17 (2017)
13. Haque, M.A., De Teyou, G.K., Shetty, S., Krishnappa, B.: Cyber Resilience Framework for Industrial Control Systems: Concepts, Metrics, and Insights. In: International Conference on Intelligence and Security Informatics (ISI). IEEE (2018)
14. Haque, M.A., Shetty, S., Krishnappa, B.: ICS-CRAT: A Cyber Resilience Assessment Tool for Industrial Control Systems. In: Int'l Conf. on Big Data Security on Cloud (BigDataSecurity), Int'l Conf. on High Performance and Smart Computing (HPSC), and Int'l Conf. on Intelligent Data and Security (IDS). IEEE (2019)
15. Hjelmvik, E.: Industroyer2 IEC-104 Analysis. NETRESEC AB (2022), https://www.netresec.com/?page=Blog&month=2022-04&post=Industroyer2-IEC-104-Analysis
16. Holm, H., Karresand, M., Vidström, A., Westring, E.: A Survey of Industrial Control System Testbeds. In: Secure IT Systems. NordSec'15, Springer (2015)
17. Hossain, M.J., Rahnamy-Naeini, M.: Line Failure Detection from PMU Data after a Joint Cyber-Physical Attack. In: IEEE Power & Energy Society General Meeting. PESGM (2019)
18. Kosut, O., Jia, L., Thomas, R.J., Tong, L.: Malicious Data Attacks on the Smart Grid. IEEE Transactions on Smart Grid **2**(4) (2011)
19. Krause, T., Ernst, R., Klaer, B., Hacker, I., Henze, M.: Cybersecurity in Power Grids: Challenges and Opportunities. Sensors **21**(18) (2021)
20. Lichtman, M., Rao, R., Marojevic, V., Reed, J., Jover, R.P.: 5G NR Jamming, Spoofing, and Sniffing: Threat Assessment and Mitigation. In: Int'l Conference on Communications Workshops (ICC Workshops). IEEE (2018)
21. Liu, Y., Ning, P., Reiter, M.K.: False Data Injection Attacks Against State Estimation in Electric Power Grids. ACM Trans. on Inform. & Sys. Sec. **14**(1) (2011)

22. Mahmood, K., Chaudhry, S.A., Naqvi, H., Kumari, S., Li, X., Sangaiah, A.K.: An elliptic curve cryptography based lightweight authentication scheme for smart grid communication. Future Generation Computer Systems **81**, 557–565 (2018)
23. Mathur, A.P., Tippenhauer, N.O.: SWaT: A Water Treatment Testbed for Research and Training on ICS Security. In: Int'l Workshop on Cyber-physical Systems for Smart Water Networks. CySWater (2016)
24. Miller, T., Staves, A., Maesschalck, S., Sturdee, M., Green, B.: Looking Back to Look Forward: Lessons Learnt from Cyber-Attacks on Industrial Control Systems. Int'l Journal of Critical Infrastructure Protection **35** (2021)
25. Ramachandran, V., Nandi, S.: Detecting ARP Spoofing: An Active Technique. In: Information Systems Security: First Int'l Conference. ICISS'05, Springer (2005)
26. Reda, H.T., Anwar, A., Mahmood, A.: Comprehensive Survey and Taxonomies of False Data Injection Attacks in Smart Grids: Attack Models, Targets, and Impacts. Renewable and Sustainable Energy Reviews **163**, 112423 (2022)
27. Reed, D.A., Kapur, K.C., Christie, R.D.: Methodology for Assessing the Resilience of Networked Infrastructure. IEEE Systems Journal **3**(2), 174–180 (2009)
28. Serror, M., Bader, L., Henze, M., Schwarze, A., Nürnberger, K.: Poster: INSIDE - Enhancing Network Intrusion Detection in Power Grids with Automated Facility Monitoring. In: ACM SIGSAC Conf. on Computer and Comm. Sec. CCS '22 (2022)
29. Serror, M., Hack, S., Henze, M., Schuba, M., Wehrle, K.: Challenges and Opportunities in Securing the Industrial Internet of Things. IEEE Transactions on Industrial Informatics **17**(5) (2021)
30. Shin, H.K., Lee, W., et al.: HAI 1.0: HIL-based Augmented ICS Security Dataset. In: USENIX Conf. on Cyber Sec. Experimentation and Test. CSET '20 (2020)
31. Srikantha, P., Kundur, D.: Denial of Service Attacks and Mitigation for Stability in Cyber-Enabled Power Grid. In: IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT) (2015)
32. Strunz, K., Abbasi, E., Fletcher, R., Hatziargyriou, N.D., Iravani, R., Joos, G.: Benchmark Systems for Network Integration of Renewable and Distributed Energy Resources. Cigre Task Force C **6**(04-02) (2014)
33. Upadhyay, D., Manero, J., Zaman, M., Sampalli, S.: Intrusion Detection in SCADA Based Power Grids: Recursive Feature Elimination Model With Majority Vote Ensemble Algorithm. IEEE Trans. on Network Science and Engineering **8**(3) (2021)
34. Whitehead, D.E., Owens, K., Gammel, D., Smith, J.: Ukraine Cyber-Induced Power Outage: Analysis and Practical Mitigation Strategies. In: Conference for Protective Relay Engineers (CPRE) (2017)
35. Wolsing, K., Wagner, E., Saillard, A., Henze, M.: IPAL: Breaking up Silos of Protocol-Dependent and Domain-Specific Industrial Intrusion Detection Systems. In: Int'l Symp. on Research in Attacks, Intrusions and Defenses. ACM (2022)
36. Young, C., Zambreno, J., et al.: Survey of Automotive Controller Area Network Intrusion Detection Systems. IEEE Design & Test **36**(6) (2019)
37. Yuan, X., Wang, L., Liu, T., Zhang, Y.: A Methodology for Continuous Evaluation of Cloud Resiliency. Am. Journal of Engineering and Applied Sciences **9**(2) (2016)
38. Zemanek, S., Hacker, I., Wolsing, K., Wagner, E., Henze, M., Serror, M.: PowerDuck: A GOOSE Data Set of Cyberattacks in Substations. In: Cyber Security Experimentation and Test Workshop. CSET '22, ACM (2022)
39. Zhang, X.M., Han, Q.L., Ge, X., et al.: Networked Control Systems: A Survey of Trends and Techniques. IEEE/CAA Journal of Automatica Sinica **7**(1) (2020)
40. Zhao, J., Netto, M., Huang, Z., et al.: Roles of Dynamic State Estimation in Power System Modeling, Monitoring and Operation. IEEE Transactions on Power Systems **36**(3), 2462–2472 (2020)

Appendix A Environment Description File Example

```
{"name": "power grid",
"host": "https://example.org",
"port": 443,
"topologies": ["cigre_mv"],
"devices": {
  "cigre_mv": [
    {"device-id": "1016",
    "type": "switch",
    "info": {}},
...
"links": {
  "cigre_mv": [
    {"link-id": "1003",
    "type": "digital",
    "connection": ["994","614"],
    "info": {}},
...
"adversaries": {
  "kill device": {
    "parameters": {
      "device-id":
        {"type": "string",
        "description": "The ID of the device to kill"},
        ...
```

Appendix B Scenario Description File Example

```
{   "environment": "power grid",
    "topology": "cigre_mv",
    "duration": 200,
    "adversaries": [
        {"type": "add_host",
    "start-time": 15,
    "parameters": {"name": "industroyer",
...
    "responses": [
...
      {"type": "link_action",
      "start-time": 60,
      "parameters": {
      "action": "down",
...
```