A Framework for Attack-Resilient Industrial Control Systems: Attack Detection and Controller Reconfiguration

This paper describes an industrial control systems policy that uses estimations to provide resiliency against attacks.

By Kaveh Paridari, Niamh O'Mahony, Alie El-Din Mady, Rohan Chabukswar, Menouer Boubekeur, and Henrik Sandberg

ABSTRACT | Most existing industrial control systems (ICSs), such as building energy management systems (EMSs), were installed when potential security threats were only physical. With advances in connectivity, ICSs are now, typically, connected to communications networks and, as a result, can be accessed remotely. This extends the attack surface to include the potential for sophisticated cyber attacks, which can adversely impact ICS operation, resulting in service interruption, equipment damage, safety concerns, and associated financial implications. In this work, a novel cyber–physical security framework for ICSs is proposed, which incorporates an analytics tool for attack detection and executes a reliable estimation-based attack-resilient control policy, whenever an attack is detected. The proposed framework is adaptable to already implemented ICS and the stability and optimal performance of the controlled system under attack has been proved. The performance of the proposed framework is evaluated using a reduced order model of a real EMS site and simulated attacks.

KEYWORDS | Artifical intelligence; building management systems; cyber–physical security; energy management; industrial control; knowledge-based systems; resilient control; SCADA systems; security analytics; stability; virtual sensor

I. INTRODUCTION

Industrial control systems (ICSs) play an important role in the monitoring and control of physical and chemical processes. ICS is a general term that encompasses several types of control systems, used in industrial production, including supervisory control and data acquisition (SCADA) systems, distributed control systems (DCSs), and other smaller control system configurations such as programmable logic controllers (PLCs), often found in the industrial sectors and critical infrastructures. ICSs are commonly seen in many critical infrastructures, such as electricity generation, transmission and distribution, water treatment, manufacturing, etc. [1]. In electricity distribution grids and, especially, in residential areas, automatic control of electrical/thermal components in buildings has become a necessary task for ICSs, in order to achieve optimal performance. In this context, the ICS is often called an energy management system (EMS).

The aim of a modern EMS is to enhance the functionality of interactive control strategies leading toward energy efficiency and a more comfortable or user-friendly environment. In recent years, EMSs and, more generally, ICSs, have been connected to communication networks, allowing remote monitoring and control of the underlying processes. While this can enable significant efficiency and usability benefits, it also increases the possibility of cyber

Manuscript received March 6, 2017; revised May 29, 2017; accepted July 5, 2017. Date of publication September 4, 2017; date of current version December 20, 2017. This work was supported in part by the European Union Seventh Framework Programme (FP7/2007-2013) under Grant 608224, by the Swedish Research Council under Grants 2013-5523 and 2016-0861, and by the Swedish Civil Contingencies Agency. *(Corresponding author: Kaveh Paridari.)*

K. Paridari and **H. Sandberg** are with the ACCESS Linnaeus Center and the Department of Automatic Control, KTH Royal Institute of Technology, 11428 Stockholm, Sweden (e-mail: paridari@kth.se; hsan@kth.se).

N. O'Mahony is with Dell EMC Research Europe, Cork, Ireland (e-mail: niamh.omahony@dell.com).

A. El-Din Mady, **R. Chabukswar** and **M. Boubekeur** are with the United Technologies Research Center, Cork, Ireland (e-mail: madyaa@utrc.utc.com; chabukr@utrc.utc.com; boubekm@utrc.utc.com).

Digital Object Identifier: 10.1109/JPROC.2017.2725482

attacks by providing an access point for would-be attackers to infiltrate the system. There have been a number of high profile attacks, in recent years, highlighting the need for appropriate security measures to protect ICS infrastructure, as will be discussed in more detail in Section II.

Many studies have been done on fault-tolerant control (see, for example, [2]–[4]), which can provide tools for attack-resilient control as well. However, there are substantial differences between the fault-tolerant and attack-resilient control, when it comes to attack detection and isolation, which motivate the need for specific methodologies to address security issues in ICSs. For example, faults are considered as physical events that affect the system behavior, where the events do not act in a coordinated way, while cyber attacks may be performed over a significant number of attack points in a coordinated fashion [5], [6]. In addition, faults do not have an intent or objective to fulfill, whilst cyber attacks do have a malicious intent.

In this work, a novel cybersecurity framework for ICS, is presented. The focus of the paper is on EMSs, as an example of ICSs. However, the framework is extendable to other ICS applications by considering the domain risk assessment that identifies the critical process and the corresponding resilience policies. The framework consists of an attack detection module, which relies on data analytics to detect anomalies in the EMS data, and a resilient control policy, which maintains the physical system in a safe state, during and after an attack. The main contributions of this paper are: 1) an evaluation of the combination of domain-specific expert knowledge (physical models) and data-driven machine learning algorithms for the detection of anomalies in data measured from the sensors in the industrial control systems (Section IV-A); 2) the adoption of a novel security metric [7], dedicated to measuring cybersecurity in industrial control systems, for evaluation purposes (Section V-B); 3) adaptation of fault-tolerant control techniques (virtual sensing) in a networked control system to mitigate cyber attacks; the proposed framework is adaptable to already implemented ICSs, with no need for major re-design of the local control loops (Section IV-B); and 4) the strengthening of available theoretical results proving stability and optimal performance, and the illustration of how a small number of carefully selected trusted (protected) sensors greatly can reduce the capabilities of a man-in-the-middle attacker (Section IV-B).

The remainder of this paper is organized as follows. Section II outlines the state of the art and motivation for the work. Section III defines the scope of the work by 1) providing a detailed description of ICS hierarchy; 2) introducing a testbed, which has been used to evaluate the feasibility of the proposed cyber–physical security framework; and 3) defining the attack models under consideration. The proposed security framework is presented in Section IV, focusing on attack detection, resilient control, and attack isolation. In Section V, simulated data are described and the performance of the proposed framework, in terms of attack detection and resiliency against attacks, is evaluated. Final remarks and conclusions are discussed in Section VI.

II. BACKGROUND AND MOTIVATION

In this section, state of the art and motivation for our work is presented, from the perspectives of assessing threats and risk, detecting attacks when they occur and implementing resilient control in a system under attack.

A. ICS Cyber-Attack Threats and Risk Assessment

Applying cybersecurity specific solutions (e.g., antivirus and firewalls) on ICSs is of a great value to reduce system vulnerabilities and protect system accessibility. Most of these solutions are oriented to monitor and protect the cyber part of the ICS such as network and devices layers. However, ICS introduces more security vulnerability due the tight integration between the controlled physical environment and the cyber system [8]. Therefore, system-level security methods are required for ICS to analyse the physical system behavior to maintain system operation availability.

Failing to maintain state awareness and acceptable performance of ICSs under unexpected cyber–physical faults or attacks can have considerable consequences. The U.S./Canada Northeastern blackout in 2003, which was caused by a software bug in the alarm system, illustrated that loss of state awareness of the ICS can result in severe economical losses [9]. The StuxNet cyber attack [10], supposedly targeting a nuclear-enrichment plant by corrupting the measurements and actuator signals in Iran, and BlackEnergy malware, targeting several electricity distribution companies in Ukraine [11], are further examples of cyber attacks against ICSs. As discussed in [12], attacks on the measurement signals may lead to a poor system performance or may cause instability of the process under the control of the ICSs. The impact of such attacks can range from financial losses to equipment damage and, even, danger to life. Thus, it is crucial to make the control of ICSs resilient against cyber crime.

Existing methods for EMS cybersecurity are mainly based on running tests and benchmarks to analyse the system vulnerabilities and evaluate possible cyber attacks and their impact; see [13]. Typically, the risk assessment is carried out using a set of best practices, as follows: 1) identify information assets; 2) locate information assets; 3) classify information assets; 4) conduct a threat modeling; and 5) apply a security plan [14]. In addition, there are many formal methods that can be applied to EMS risk assessment, as described in [15].

One such method is fault tree analysis (FTA) [16], where the attack impact is evaluated based on system component dependencies. The framework proposed in this paper originated in a risk assessment based on the FTA method, combined with an empirical study that quantifies the financial and safety impact of an attack [17], [18]. Smart-grid infrastructure can control both electrical and thermal loads, where some equipment, such as combined heat and power (CHP) can be an energy source for both electrical and thermal demand. In this context, heating, ventilation, and air conditioning (HVAC) systems are considered an important contributor to energy consumption, making them a target for attacks with financial impact. In addition, attacking the EMS in a smart grid can lead to a safety risk [19], due to damage to the water transport system or heating sources (e.g., CHP and boilers).

The focus of this work is on so-called "man-in-the-middle" attacks, whereby the attacker can intercept communications between components in the system and manipulate or corrupt the values of measurements or commands being sent and received. There are studies which consider several attack scenarios where the adversary's goal is to drive the system to an unsafe state without triggering any alarm [20], [21]. These attacks are called stealthy attacks. A particular class of stealthy attacks that has raised a lot of interest in the research literature is the class of 0-stealthy attacks, which are also called undetectable attacks; see, for instance, [5] and [22]. In short, these attacks refer to an attacker who is able to corrupt measurements in a manner that they exactly correspond to a valid physical state of the system. These attacks are undetectable, in the sense that any fault or attack detector that simply detects when received measurements do not satisfy physical laws or relationships will not generate an alarm. The detector proposed in this work is designed to detect attacks even when the attacked variables are physically valid, as described in Section IV-A. 0−stealthy attacks typically require plant-wide corruption of measurements to be feasible [23]. An important property of the resilience policy proposed in this work is to ensure that such an attack cannot destabilise the system. This property is obtained by protecting a small, but carefully selected, number of sensors, as will be described in Section IV-B.

B. Attack Detection

Recently, there has been an increase in research into the limitations of the existing attack detection and identification methods [22], [24], [25]. The time during which vulnerabilities remain hidden, together with the time required to patch them, leave a window large enough for adversarial system penetration. These factors highlight the importance of detecting attacks as soon as possible, in order to minimize damage and impact.

Analytics capabilities enable quick detection of cyber attacks by checking the system behavior at application level and responding quickly to minimize the attacks' impact. As studied in [26], the operational model must go beyond the conventional focus on distribution and generation infrastructure for fault isolation, remediation and recovery, and focus on information and a new understanding of data analysis. In addition, as discussed in [27], it requires the ability to handle processing of huge amounts of data, by using new analytics and visualization techniques. In this work, a security information analytics tool will be proposed, using SCADA and EMS data for the detection of man-in-the-middle attacks.

C. Resilient Control

Once an attack is diagnosed, control policies which are resilient against the attacks, should be triggered. Since cyber attacks to ICSs also affect the physical behavior of the system, the tools used for fault-tolerant control can be applied for attack-resilient control. For example, virtual sensor concepts, as described in [2] to deal with sensor failures, can be used in the case of sensor attack, as will be shown in this work. Different approaches have been studied for increasing the system resiliency against attacks [28]–[32]. In [28], Schenato *et al.* consider the problem of control and estimation in a networked system when the communication links are subject to disturbances (corresponding to packet losses), resulting from a denial-of-service (DoS) attack, for instance. The estimation and control of linear systems, when some of the sensors or actuators are corrupted by an attacker, is studied in [29]. In that work, they propose an efficient algorithm, inspired by techniques in compressed sensing, to estimate the state of the plant despite attacks. The authors assume that the attacked nodes does not change over time. In addition, a general framework to model and analyse impact of attacks, is proposed in [5]. In [30], a method for state estimation in the presence of attacks, for systems with noise and modeling errors, is proposed. In that work, it is shown that the attacker cannot destabilize the system by exploiting the difference between the model used for state estimation and the real physical dynamics of the system. In [31], a control technique is proposed which is resilient against certain sensor attacks. In that technique, a recursive filtering algorithm, to estimate the states of the system, is implemented, taking advantage of redundancy in the information received by the controller. The resilience policy proposed in this work is designed to maintain the stability and optimal performance of the system, whilst preventing undetectable attacks.

III. ICS STRUCTURE AND ATTACK MODELS

In this section, we provide a detailed description of ICS's hierarchy, introduce a testbed being used to evaluate the feasibility of the proposed cyber–physical security framework, and define the attack models under consideration.

A. ICS Hierarchy

A general hierarchical structure of an ICS is composed of a lower layer and a supervisory layer, both of which will be described below. Our proposed attack-resilient framework is found in the supervisory layer.

1) Lower Layer: The lower layer of the ICS, known as the plant, consists of the physical interconnected infrastructure and local controllers. A schematic of the plant is illustrated in Fig. 1, showing a linear, closed-loop system. In the figure, the physical interconnected infrastructure is represented by a number N of interconnected processes (P_i, $i \in \Phi$), which are controlled by the local controllers ($K_i, i \in \Phi$), where $\Phi = \{1, \ldots, N\}$ is the index set of processes. In this system, the controllers send the control vector signal $u = [u_1 \ldots u_N]^\top$ to the processes and receive the sensor measurements $y = [y_1 \ldots y_N]^\top$ from them.

The closed-loop interconnected system evolves as

$$\begin{aligned} x(k+1) &= A^{cl}x(k) + B_1^{cl}d(k) + M\begin{bmatrix} w(k) \\ v(k) \end{bmatrix} \\ y(k) &= C^{cl}x(k) + L\begin{bmatrix} w(k) \\ v(k) \end{bmatrix}, \end{aligned} \tag{1}$$

where, at the kth instant, $x \in \mathbb{R}^{n_x}$ is the state vector of the plant, with a number n_x of state variables, $y \in \mathbb{R}^N$ is the measurement vector being sent to the controllers, and $d \in \mathbb{R}^N$ is a deterministic disturbance vector. Here, w and v are process and measurement zero-mean Gaussian white noise, respectively. The dimensions of the matrices A^{cl}, B_1^{cl}, M, C^{cl}, and L conform with the relevant vectors. Let the expectation and the covariance of w and v be given by

$$\mathrm{E}\begin{bmatrix} w(k) \\ v(k) \end{bmatrix} = 0, \ \ \mathrm{E}\begin{bmatrix} w(k) \\ v(k) \end{bmatrix}\begin{bmatrix} w(l) \\ v(l) \end{bmatrix}^\top = \underbrace{\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}}_{R}\delta_{kl} \tag{2}$$

where R_{11} and R_{22} are the covariance of w and v, respectively, and $R_{12} = R_{21}^\top$ is the cross covariance between w and v.

Note that the system described by (1) is considered to be in a normal state, with no anomaly in the received control and measurement signals, $\tilde{u}_i$ and $\tilde{y}_i$, respectively. This means that $\tilde{u}_i = u_i$ and $\tilde{y}_i = y_i$, $i \in \Phi$. Under this condition, we have the following assumption, which captures that the plant is assumed to be stable and well-configured initially.

Assumption 1: The linear closed-loop system (1) is stable, which means matrix A^{cl} is Schur stable (i.e., $\rho(A^{cl}) < 1$), and, also, the pair (A^{cl}, C^{cl}) is observable.

2) Supervisory Layer: The supervisory layer, often referred to as the supervisory controller, can be viewed as the brain of the system. A schematic of the supervisory controller is illustrated in Fig. 2, consisting of three crucial subtasks [2].

1) Attack/fault detection: determine whether an attack/fault has occurred. (See Section III-C for a description of attack models.)
2) Attack/fault isolation: identify which measurements have been manipulated by the attack/fault (e.g., $\tilde{y}_i \neq y_i$). It should be noted that, in this work, it is assumed that there is no attack on the control signals, such that $\tilde{u}_i = u_i$, $i \in \Phi$. However, the framework can be easily extended to include also the attacks on the control signals.
3) Controller reconfiguration: when an attack/fault is detected, reconfigure the associated control loops.

In our proposed framework, illustrated in Fig. 2, the security information analytics (SIA) tool in the supervisory controller is responsible for attack detection (see Section IV-A). To perform the controller reconfiguration, the resilience policy applies estimation-based methods to generate correction signals (see Section IV-B). Attack isolation is outside the scope of this paper; in Section IV-A, we briefly indicate how the proposed tools can contribute to isolation but a complete treatment of the isolation problem will be reported in future work. It should also be noted that we do not consider, in this work, cases where the supervisory layer is subject to attack, such as those discussed in [33].

B. Testbed Description

As a proof of concept for the cybersecurity framework developed in this work, an EMS which controls a small-sized smart-grid, covering several buildings at the Cork

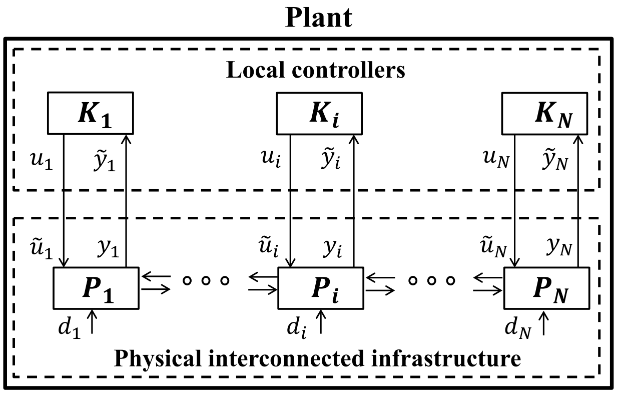

Fig. 1. *Schematic of a linear closed-loop system.*

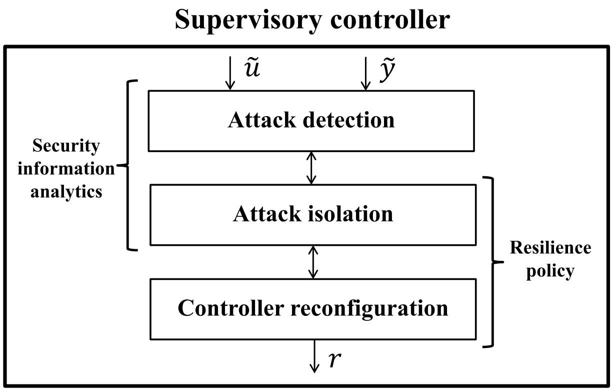

Fig. 2. *Schematic of the supervisory controller.*

Institute of Technology (CIT, Ireland), is considered. A detailed description of the grid can be found in [34]. In the following, the main components used by the EMS to control the HVAC system will be highlighted. The modeling techniques used to capture the system dynamics will also be discussed briefly.

Fig. 3 shows the HVAC system at the CIT demo site, which is considered to be the highest source of energy consumption in the building operation. The HVAC system consists of many functions that are controlled by various control elements under the EMSs, including SCADA and building management system (BMS). SCADA manages the electrical component operations (e.g., CHP) and the BMS manages the operation of thermal components (e.g., boilers). Here, the CHP and boiler heat water to a defined temperature setpoint, which is identified using a weather compensation method [35]; an on/off controller is, then, used to keep the water in the heating sources at its setpoint. Whether the boiler and/or CHP are in operation, at any given time, is dependent on the thermal load in the building, which is indicated by the return temperature. Mixing valves, controlled by a proportional–integral (PI) control algorithm, regulate the supply temperatures for each floor in each building and water is distributed across several radiators in each floor, where each radiator is controlled using an on/off controller to reach a predefined room temperature set-point.

The proposed resilience policy, which will be described in Section IV-B, requires a linearised model of the controlled HVAC system. To arrive at the linear state–space model of this system, given in (1)–(2), subspace identification, followed by a prediction error method [36], was applied. This system identification has been discussed, in detail, in [18] and [37]. Modeling the system, in this manner, results in a simple, linear, third-order system, in an innovation form.

C. Attack Models

The "man-in-the-middle" attacker, considered in this work, can secretly listen to the values being communicated between the processes and controllers in the lower layer and has the ability to manipulate or corrupt the measurement signals *yi*. For example, the attacker can modify the measurements by placing a malware in the PLC used by the BMS to control the HVAC system. Furthermore, this attacker may have knowledge of the model of the plant. The supervisory layer is assumed not to be accessible by the attacker. For example, consider that the measurement y_i is manipulated by adding an offset Δy_i. Thus, a measurements' attack vector $\Delta y = [\Delta y_1 \ldots \Delta y_N]^{\top}$ is defined, which has nonzero entries for measurements under attack and zero values for all other measurements. A general model for the measurement signals received by the local controllers and, subsequently, by the supervisory controller, can, then, be given by

$$\tilde{y} = y + \Delta y. \tag{3}$$

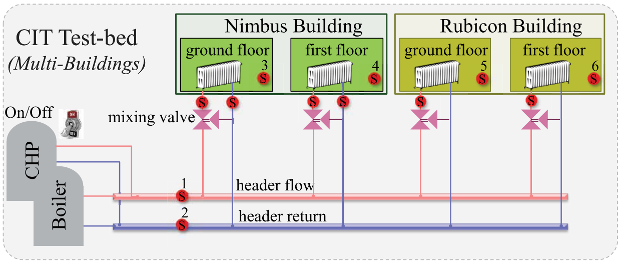

Fig. 3. ***BMS for HVAC system at the CIT demo site.***

D. Undetectable (0-Stealthy) Attacks

By assuming that $d = 0$, for an attack signal Δy to be undetectable, there must exist an initial state x_0, which results in $\tilde{y} = 0$. Existence of such a signal can easily be checked by considering the matrix pencil (Rosenbrock system matrix), a detailed discussion of which can be found in [18] and [37]. As discussed in those works, we can ensure that no undetectable attacks exist by protecting a subset of the sensors in C^{cl} (i.e., removing the corresponding elements of Δy). It should be noted that, currently, there is no systematic way, beyond enumerative schemes, to find the desired subset of the sensors; this is an interesting topic for future work.

A number of methods exist for protecting the measurements, such as, for instance, measurement signal encryption or hard wiring. For further discussions on signal protection in SCADA systems, see [23].

Definition 1: Considering that some of the measurements are protected, we classify measurements into the following four types.

1) Unprotected measurements: we assume the attacker can have access to at most a number m of sensor measurements y_j for $j \in \Gamma$. Here, $\Gamma \subset \mathbf{\Phi}$ is the index set of unprotected measurements and the cardinality of Γ is $\mathbf{card}(\Gamma) = m$.
2) Protected measurements: these measurement are not accessible by the attacker. Here, $\Gamma^C = \mathbf{\Phi} \backslash \Gamma$ is the index set of protected measurements, where $\mathbf{card}(\Gamma^C) = N - m = h$.
3) Attacked measurements: since attacking all the unprotected measurements is costly for the attacker, some of them may be unattacked by the attacker. Thus, we define the set of attacked measurements y_j for $j \in \Gamma^a$. Here, $\Gamma^a \subset \Gamma$ is the index set of attacked measurements ($0 \leq \mathbf{card}(\Gamma^a) \leq m$), and we have $\tilde{y}_i \neq y_i, \forall i \in \Gamma^a$.
4) Healthy measurements: here, $\Gamma^h = \mathbf{\Phi} \backslash \Gamma^a$ is the index set of healthy measurements, and we have $\tilde{y}_i = y_i, \forall i \in \Gamma^h$.

Note that, in principle, the attacker can observe and manipulate any or all of the unprotected measurements, given enough resources.

IV. PROPOSED ICS SECURITY FRAMEWORK

Here, we propose a strategy for attack detection and control reconfiguration, to ensure the stability and optimal performance of the controlled system under attack.

A. Attack Detection

The security information analytics (SIA) tool is responsible for detecting potential attacks. It consists of a set of anomaly detection algorithms and a web application that allows analysts to examine the results. In this work, the focus is on the algorithms and their performance; as such, the web application is not described in detail.

The SIA tool uses a combination of 1) domain-specific expert knowledge and 2) machine learning algorithms, to understand the behavior of the system under normal, stable operating conditions, and to detect any anomalies or deviations from normal behavior. These anomalies are flagged as potential attacks and the resilience policy is triggered to maintain stability of the system, while further investigation into the anomalies is carried out.

The combination of expert knowledge and data-driven machine learning approaches aims to provide a high attack detection rate, in the face of highly sophisticated attackers. The expert knowledge component exploits domain-specific physical laws and system topology, to define the explicit relationships between different variables in the system; if these relationships are not satisfied by the measured variables, it may indicate that one or more of the variables is under attack. There also exist other, less formal dependencies between variables, which, perhaps, are defined by highly complex relationships among many elements of the system or, for which formal relationships cannot be explicitly defined. Furthermore, given certain conditions, particular behaviors may be unlikely; for example, it might be improbable, but not impossible, for the heating system to be turned on in a building when the external temperature is high, or it might be unlikely for the energy consumption in a household to be constant throughout the day and night. Despite not having explicit definitions, these relationships between operating conditions and measured variables can play an important role in detecting abnormal behavior in the system and, especially, in detecting stealthy attacks. A well-designed machine learning algorithm can allow the implicit patterns and dependencies in the data to be learned, building a model that represents the normal behavior of the system. Measured data can then be compared to the model to determine if the system is operating normally or not.

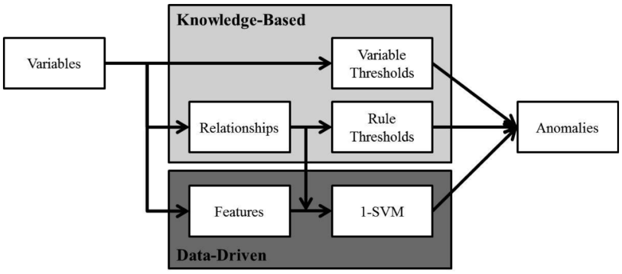

Fig. 4. ***Schematic of the SIA tool showing both KB and DD algorithms and the interaction between them.***

Fig. 4 shows the components of the SIA, illustrating the dependency of the data-driven (DD) detector on the knowledge-based (KB) components. Details of the KB and DD algorithms, implemented in the SIA tool, are described in Sections IV-A1 and IV-A2, respectively.

1) Knowledge-Based Anomaly Detection: The KB detector includes a) thresholds, which include safety limits (such as maximum allowable temperature for a boiler or current rating for an electrical element) and device specifications (such as nominal operating temperature, voltage, frequency, etc.), applied directly to the measured variables, along with b) physical rules and relationships between multiple variables (such as Ohm's law, relating voltage, current and impedance, and other similar rules governing variables measured in the system). Static rules, relating the measured room temperatures with their corresponding setpoints (i.e., the relevant thermostat setting), and comparing the measured supply and return temperatures at the boiler to the equipment specifications, were derived. Additionally, the pairwise correlation coefficient in a sliding window and pairwise difference between the temperatures in the system were calculated, e.g., between the boiler supply and return temperatures and between the temperatures in different rooms.

Dynamic rules were also derived, as follows. The estimated value of the kth sample, $\hat{y}(k)$, can be calculated, using the previously measured values, by system identification and state estimation, in a similar manner to that described in Section IV-B. The estimation residue $r_{est}(k) = y(k) - \hat{y}(k)$, where $y(k)$ is the kth measured sample, is expected to be small, assuming that the model used for estimation is a valid one, and the behavior of the system is "normal." However, if the estimation error becomes large, this indicates that the behavior of the system does not match that of the model, which may mean that an attack is taking place. Some estimation error is expected, as the system model, typically, has a lower order than the system itself, due to complexities in the system and its dynamics. Upper and lower thresholds are derived from the statistical distribution of the estimation residue for historical data. In this work, a dynamic rule was derived to estimate the current value of each variable, given the previous values, based on a system identification process.

It should be considered that a knowledgeable attacker, such as a malicious insider or an attacker who has observed the system over a very long period, could, conceivably, exploit the difference in granularity of the model, compared to the system itself, to carry out an undetected attack. The combination of KB and DD detection methods aims to address this concern. Protected measurements could be taken into account in the KB and DD detectors, by weighting more heavily the values measured by the protected variables. However, in order to maintain independence between the attack detector and the resilient policy, in this work, the protected and unprotected variables are not treated differently by the SIA tool.

A set of "healthy" data, during which no attacks occurred, was used to calculate appropriate thresholds for each variable or rule; for example, the maximum value of the difference between the room temperatures and their setpoints, or the 99th percentile of the estimation residue. The detector, then, classifies a given sample as an anomaly if one calculated metric or more exceeds the corresponding threshold. Because each metric relies on a subset of all of the variables, this detector may give an indication of which metric(s) have been manipulated by an attacker, thus contributing to the attack isolation task. Furthermore, the magnitude of the residue or the amount by which the threshold was exceeded can provide an indication of the severity of the disturbance.

It is worth noting that other, more robust approaches to KB anomaly detection do exist, such as using detailed specifications of the application being monitored and its execution, in order to compare the implementation with predicted behavior [38], or the use of a detailed vulnerability analysis, based on nonlinear constrained models for power grids, to identify conditions for constrained false data injection attacks [39]. However, in this work, which focuses on the temperature data in a HVAC system, whose behavior is influenced, not only, by system specifications but, also, by external factors, such as number of occupants in a room, degree to which windows are open, external weather conditions, and many others, a more general approach to applying domain knowledge is taken, as described above, to investigate the feasibility of such an approach, in cases where a systematic model of interdependencies between components is challenging to define.

2) Data-Driven Anomaly Detection: As mentioned previously, machine learning (ML) algorithms are employed to learn normal behavior from available data and, then, to compare measured samples to the learned models, to determine if those new samples are anomalous or not. Many ML problems involve the classification of measurements into predefined groups or classes. If a dataset exists, containing samples for which the class label is known, the problem is referred to as "supervised" learning; conversely, if no such labeled dataset exists, it is an "unsupervised" learning problem. In most anomaly detection problems, where the vast majority of available data belongs to the "normal" class or null hypothesis, with few examples of data from the "anomalous" class or alternative hypothesis, supervised learning approaches are unsuitable. Furthermore, it is important that anomaly detection systems can successfully detect previously unseen anomalies, for which no labeled data exists.

For cases where training data are available but they represent only a single class or, more generally, where the vast majority of samples can be assumed to be from a single class, with a very small number of samples from other (anomalous) classes, single-class machine learning algorithms exist. Most of these are adaptations of multiclass classifiers. In this work, a single-class support vector machine (1-SVM) is used, which is an adaptation of the popular support vector machine (SVM) [40]. A detailed description of the 1-SVM can be found in [41].

The 1-SVM learns its model from a vector of features derived from the measured data. The features may be the raw measured values, functions combining multiple values, statistics derived from the values, or other metrics. In this work, the features, which are used as inputs to the 1-SVM, include 1) the raw variable values; 2) the residues and differences calculated in the KB detector; and 3) time-domain statistics, such as mean and standard deviation, calculated in a moving window. Results for the different types of features will be shown in Section V.

The healthy data were, once again, used to train the 1-SVM to recognize what the data looks like during normal operation. New samples are, then, classified as normal or anomalous, depending on how closely they fit with the learned model.

The 1-SVM treats the features jointly and provides a binary classification (i.e., anomalous or normal); as such, it does not provide any indication regarding which specific variable(s) caused the anomaly, or regarding the severity of the disturbance. However, it is an effective method for limiting the impact of stealthy attacks.

B. Resilient Control

Once the attack has been detected and, ideally, isolated, the resilience policy guarantees that the ICS will meet the following criteria.

I. Abbreviation	Expansion
C_1	*Undetectable attack blocking:* no undetectable attack can be injected to the measurement signals;
C_2	*Performance optimality:* performance is optimal, in terms of minimum variance error of state estimation,under the abnormal state (i.e., the state in which attacks have taken place);
C_3	*Stability:* the system remains stable under abnormal states.

This section describes how the proposed resilience policy fulfills each of these criteria.

From [42], we learn that if the closed-loop transfer function from Δy_i to y_i, which is seen by the attacker, is nonzero, then the attacker can destabilize the system by injecting a Δy_i that violates the small-gain theorem's necessary and sufficient conditions. Thus, to ensure the stability of the closed-loop system, the resilience policy must eliminate the influence of the corrupted measurement from the control loop. This is achieved by means of correction vector signal r, designed in such a way that the transfer function from Δy_i to y_i becomes zero. The correction vector signal r is generated in the supervisory controller and sent to the local controllers to correct the attacked signals y_j for $\forall j \in \Gamma^a$. This results in a controller reconfiguration.

To generate r, a virtual sensor is implemented in the supervisory controller, using a Kalman filter. Since the virtual sensor is running in the supervisory level, it has access to system-wide measurements $(y_1, \ldots, y_N)$ and can estimate the states of the plant $(\hat{x}_i,\ i = 1, \ldots, n_x)$, based on the available model of the system and all the available healthy measurements, at all instants. From $\hat{x}$, $\hat{y}$ can be estimated, as shown in Fig. 5. The correction vector signal r is, then, calculated and sent to the plant by the supervisory controller. It is important to note that, since the virtual sensor and controller reconfiguration takes place in the supervisory layer, the resilience policy can be easily implemented in existing ICS, without the need for major redesign of the local controllers in the plant.

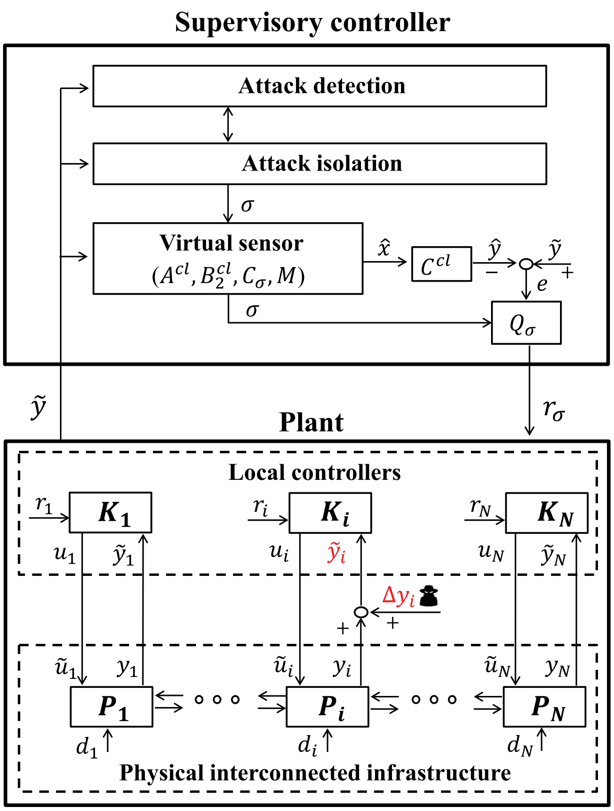

Fig. 5. ***Schematic of control system, which is resilient against the adversarial actions on the measurements.***

Since the attacker has access to a number m of unprotected measurements y_i, $\forall i \in \Gamma$, we can consider 2^m different attack scenarios. This means there are 2^m different modes, $\sigma \in \{1, \ldots, 2^m\}$, for the system operation, where each mode indicates the combination of measurements that are under attack. Note that, over time, σ can switch. In the proposed framework, σ is continuously generated by the attack isolation modulation. This means that, as the attacker varies the attack policy, the attack isolation module informs the control reconfiguration module of the current attack mode, allowing the system to react and maintain optimal performance. Thus, considering different modes of operation, the measurement attack vector and the correction vector signals are represented by $\Delta y_\sigma(k)$ and $r_\sigma(k)$, respectively, at any given time k.

Here, the system mode is $\sigma = 1$ if the system is under normal state (i.e., $\Delta y_\sigma = 0$ if $\sigma = 1$) and we have $\sigma \in \{2, \ldots, 2^m\}$ for all the possible abnormal states (i.e., $\Delta y_\sigma \neq 0$ if $\sigma \neq 1$). $\sigma = 2^m$ denotes the mode in which all unprotected measurements are attacked. Considering the different modes of operation and the previously introduced Δy_σ and r_σ, the discrete-time LTI system model (1) becomes a switched linear system (see [43])

$$x(k+1) = A^{cl}x(k) + B_1^{cl}d(k) + B_2^{cl}(\Delta y_\sigma(k) + r_\sigma(k)) + M\begin{bmatrix} w(k) \\ v(k) \end{bmatrix}$$

$$y(k) = C^{cl}x(k) + L\begin{bmatrix} w(k) \\ v(k) \end{bmatrix}. \qquad (4)$$

The switched linear model (4) is called the abnormal system model. Recall that $\sigma(k)$ can vary over time and the goal of $r_\sigma(k)$ is to replace the attacked measurement $\tilde{y}$, with $\hat{y}$. Here, for simplicity of notation, we have used the subscript σ instead of $\sigma(k)$, which is dependent on the given time k.

Under the normal and abnormal states and for different attack scenarios, the virtual sensor generates an estimate of the outputs $(\hat{y}_i(k) = C_i^{cl} x_i(k), i \in \Phi)$, based on the linear state–space model of the plant (C^{cl}) and the received measurement signals $x_i(k)$. The virtual sensor estimates the output$\hat{y}_i$, $i \in 1, \ldots, n_y$, based on all the available healthy measurements (see [2]). As such, the corrupted measurements are not used by the virtual sensor. At each time k, the virtual sensor is informed, by means of σ, that the measurements y_j, $j \in \Gamma^a$ are corrupted and should not be used for updating and predicting the state estimate $\hat{x}(k)$. To account for the fact that there can be communication delays between the plant and the supervisory controller, we make the following assumption.[1]

[1] The assumption can be relaxed at the expense of a slightly more complicated estimator.

Assumption 2: At the given time k, only the measurements until time $k-1$ are available in the supervisory controller.

Based on Assumption 2, the virtual sensor, which, here, is a switched Kalman filter (see [43]), takes the prediction step for the state of the system as

$$\hat{x}(k+1|k) = A^{cl}\hat{x}(k|k-1) + K_\sigma(k)\underbrace{\left[y_\sigma(k) - C_\sigma\hat{x}(k|k-1)\right]}_{\varepsilon(k)} \tag{5}$$

where $y_\sigma(k)$ is a vector of healthy measurements. Here, C_σ is constructed from the matrix C^{cl} by removing the rows related to the corrupted measurements and based on the operation mode σ. Note that by $\hat{x}(k|k-1)$ we mean an estimation of $x(k)$, given the measurements $y_\sigma(t)$ up until time $t = k-1$. The optimal one step ahead prediction of $y(k)$ is $\hat{y}(k) = C^{cl}\hat{x}(k|k-1)$. The time-varying Kalman gain $K_\sigma(k)$ is given by

$$\begin{aligned} K_\sigma(k) &= \left(A^{cl}P_\sigma(k)C_\sigma^\top + R_{12\sigma}\right) \times \left(C_\sigma P_\sigma(k)C_\sigma^\top + R_{2\sigma}\right)^{-1} \\ P_\sigma(k) &= A^{cl}P_\sigma(k-1){A^{cl}}^\top + R_{1\sigma} - \left(A^{cl}P_\sigma(k-1)C_\sigma^\top + R_{12\sigma}\right) \\ &\quad \times \left(C_\sigma P_\sigma(k-1)C_\sigma^\top + R_{2\sigma}\right)^{-1} \\ &\quad \times \left(A^{cl}P_\sigma(k-1)C_\sigma^\top + R_{12\sigma}\right)^\top \end{aligned} \tag{6}$$

where $P_\sigma(k)$ is the time-varying estimation error covariance matrix. In (6), we have $R_{1\sigma} = MRM^\top$, $R_{2\sigma} = L_\sigma RL_\sigma^\top$, and $R_{12\sigma} = MRL_\sigma^\top$. Here, $L_\sigma = I_\sigma\rho_\sigma L$, in which I_σ is constructed from identity matrix having zero on the ith diagonal entry, and ρ_σ is constructed from identity matrix by removing the ith row, $\forall i \in \Gamma^a$, based on the operation mode σ.

In the following assumption, the system is under the worst case attack mode, $\sigma = 2^m$, in which all of the m unprotected sensor measurements (y_j, $\forall j \in \Gamma$) are under attack and the virtual sensor only uses the protected measurements' noise for estimation.

Assumption 3: The states of the system (4) are all stabilizable from the protected measurements' noise. This means that the pair $(A^{cl} - R_{12\sigma}R_{2\sigma}^{-1}C_\sigma, R_{1\sigma} - R_{12\sigma}R_{2\sigma}^{-1}R_{12\sigma}^\top)$ is stabilizable for $\sigma = 2^m$.

Given Assumption 3, the system is stabilizable for all other modes σ. If Assumption 3 is omitted, there may be several positive–semidefinite solutions for (6) in the steady state.

Given $\hat{x}(k|k-1)$, the correction signal is, then, given by

$$r_\sigma(k) = Q_\sigma(k)\big(\hat{y}(k) - C^{cl}\hat{x}(k+1|k)\big). \tag{7}$$

Here, the matrix $Q_\sigma(k)$ is a diagonal matrix, having -1 on diagonal entries related to the measurements under attack, and 0 on the rest. In this way, the resilience policy omits the attacked measurements, and uses the estimated outputs instead.

Based on the estimated states, the supervisory controller will send the correction signal r_σ, to the plant for control reconfiguration and to improve the performance of the system under attack. The local controller receives the signal $y + \Delta y_\sigma + r_\sigma$ instead of $y + \Delta y_\sigma$. The signal $y + \Delta y_\sigma + r_\sigma$ would not be of the same quality, and may be time delayed compared to measurements of the system y under normal state. However, in this way, we make sure that the attacker cannot destabilize the system. The other advantage of this approach is that it does not require many changes in the lower level implementation of the local controllers. Next, it will be proven that the proposed scheme preserves stability and performance optimality.

Definition 2: (Completely Switched Observability [43]): The deterministic part of (4) is completely switched observable over the finite time horizon $[k_0, k_1]$ if and only if the observability matrix

$$\mathrm{D}(k_1, k_0) := \begin{bmatrix} C_\sigma(k_0) \\ C_\sigma(k_0+1)A^{cl} \\ \vdots \\ C_\sigma(k_1)(A^{cl})^{k_1-k_0} \end{bmatrix} \tag{8}$$

has full rank $\mathrm{rank}\{\mathrm{D}(k_1, k_0)\} = n_x$, for each possible switching sequence $\sigma(k_0), \ldots, \sigma(k_1)$.

However, the switched observability of the system (4) is not ensured by the assumption that for each subsystem $\sigma \in \{1, \ldots, 2^m\}$, the pair (A^{cl}, C_σ) is observable. Therefore, the following assumption is made.

Assumption 4: There exists redundancy in the sensor measurements' information, and the system is observable from the protected measurements alone y_i, $i \in \Gamma^C$.

To fulfill the criterion C_1, the following assumption on measurements' protection must also be made.

Assumption 5: The protected measurements block undetectable attacks.

The set of protected measurements should, therefore, be selected to fulfill the above assumptions, in order for the resilience policy to achieve criterion C_1.

Lemma 1: Consider the switched system (4) with the finite number of switching modes $\sigma \in \{1, \ldots, 2^m\}$. This system is completely switched observable over the finite time horizon $[k_0, k_1]$, $\forall k_1 \geq k_0 + n_x - 1$ under Assumption 4.

Proof: Based on Assumption 4, at most m measurements could be corrupted, and there exist 2^m different modes $\sigma \in \{1, \ldots, 2^m\}$, for the switched system. Recall that C_{2^m} relates to the worst case mode in which all m unprotected measurements are corrupted. It is known that C_{2^m} is a subset of C_σ, $\forall \sigma \in \{1, \ldots, 2^m\}$. In addition, based on Assumption 4, it is known that (A^{cl}, C_{2^m}) is observable ($\mathbf{Obsv}(A^{cl}, C_{2^m}) = n_x$). Thus, $\mathrm{rank}\big\{\mathrm{Obsv}(A^{cl}, C_{2^m})\big\} \leq \mathrm{rank}\{\mathrm{D}(k_1, k_0)\}$, $\forall k_1 \geq k_0 + n_x - 1$, which means that $\mathrm{rank}\{\mathrm{D}(k_1, k_0)\} = n_x$.

Lemma 2: If, for each possible switching sequence $\sigma(k_0), \ldots, \sigma(k_1)$, over the finite time horizon $[k_0, k_1]$, the pair (A^{cl}, C_σ^{cl}) is completely switched observable and the pair (A^{cl}, M) is controllable, by defining $P_\sigma(k) = \mathrm{E}(e^x(k) e^x(k)^\top)$, for an arbitrary switching sequence $\sigma(0), \ldots, \sigma(k), \forall k$, the error variance $tr(P_\sigma(k))$ of the switching state estimation is bounded.

Note that $e^x(k) = x(k) - \hat{x}(k)$ is the estimation error. *Proof:* For proof, see [43], Lemma 2 for instance.

Theorem 1: The application of the switching Kalman filter (5) yields an unbiased linear estimate $\hat{x}(k)$ of the system state $x(k)$, with minimum error variances $\forall k \geq n_x - 1$, for an arbitrary switching sequence $\sigma(0), \ldots, \sigma(k)$.

Proof: Based on Lemma 1, the switched linear system in (4) is completely switched observable over the finite time horizon $[0, k]$, $\forall k \geq n_x - 1$. By having completely switched observability, as is shown in the proof of Lemma 3 in [43], the switching Kalman filter (5) leads to the minimum error variance $\forall k \geq n_x - 1$.

Based on Theorem 1, the criterion C_2 has been fulfilled here.

Considering, under all conditions, only the protected measurement for the estimation gives an upper bound, $tr(P_\sigma(k))$, $\sigma = 2^m$, $\forall k$, on the optimal performance of the system, in the sense of variance of the estimation error $e^x(k)$. Thus, by considering all the healthy measurements, which gives us more information for the estimation, the application of the switching Kalman filter (5) yields an unbiased linear estimate $\hat{x}(k)$ of the system state $x(k)$ with minimum error variance $tr(P_\sigma(k))$, $\sigma \in \{1, \ldots, 2^m\}, \forall k$.

By designing the switched observer (5) for the stochastic switched linear system (4), for different modes of operation $\sigma_k \in \{1, \ldots, 2^m\}, \forall k$, the closed-loop dynamics of the system are given by

$$\begin{bmatrix} x(k+1) \\ e^x(k+1) \end{bmatrix} = \underbrace{\begin{bmatrix} A^{cl} & B_2^{cl} Q_\sigma C^{cl} \\ 0 & A^{cl} + B_2^{cl} Q_\sigma C^{cl} - K_\sigma(k)\, C_\sigma \end{bmatrix}}_{A_\sigma^{sp}} \begin{bmatrix} x(k) \\ e^x(k) \end{bmatrix} + \begin{bmatrix} B_1^{cl} \\ B_1^{cl} \end{bmatrix} d(k) + B_2^{cl} \begin{bmatrix} I + Q_\sigma \\ 0 \end{bmatrix} \Delta y_\sigma(k) + \begin{bmatrix} M + B_2^{cl} Q_\sigma L_\sigma \\ M + B_2^{cl} Q_\sigma L_\sigma - K_\sigma(k)\, L_\sigma \end{bmatrix} \begin{bmatrix} w(k) \\ v(k) \end{bmatrix}. \tag{9}$$

Theorem 2: For arbitrary switching sequence modes, $\sigma \in \{1, \ldots, 2^m\}, \forall k$, the closed-loop system (9) is asymptotically stable.

Proof: Here, since $(I + Q_\sigma)$ is a diagonal matrix with diagonal zero entries for the corresponding nonzero entries of $\Delta y_\sigma(k)$, then we have $(I + Q_\sigma)\Delta y_\sigma(k) = 0$. In addition, $e^x(k)$ is the input for the state $x(k)$, and the matrix $A_{\sigma_1}^{sp} \times A_{\sigma_2}^{sp} \times \cdots \times A_{\sigma_l}^{sp}$, $\forall l$, is an upper block triangular matrix. In [43], proof of Lemma 4, it is shown that the deterministic part of the error $e^x(k)$ is asymptotically stable, in the sense of Lyapunov, and vanishes for $k \to \infty$, which means $A^{cl} + B_2^{cl} Q_\sigma C^{cl} - K_\sigma(k)\, C_\sigma$ is a stable matrix. Thus, the matrix $A_{\sigma_1}^{sp} \times A_{\sigma_2}^{sp} \times \cdots \times A_{\sigma_l}^{sp}$, $\forall l$, has bounded off-diagonal block, since the matrix A^{cl} is Schur stable by Assumption 1 and $A^{cl} + B_2^{cl} Q_\sigma C^{cl} - K_\sigma(k)\, C_\sigma$ is stable. Therefore, the closed-loop system is asymptotically stable, since the effect of $e^x(k)$ on $x(k)$ vanishes, the effect of $\Delta y_\sigma(k)$ on $x(k)$ is zero, and A^{cl} is Schur stable.

Based on Theorem 2, the criterion C_3 has been fulfilled here.

V. SIMULATION RESULTS

A. Simulation Environment

Because attacks cannot be carried out on operational infrastructure, like the Nimbus smart grid, without causing, at worst, danger and, at best, inconvenience to users of the buildings, a simulation of the environment was built, in order to measure the performance of the proposed security system. The simulated environment is a reduced-order model of the system, which does not intend to capture all variables and operating conditions, but to provide a representation of a subset of the components, capturing their inter-relationships and dynamics.

To arrive at a linear state–space model of the controlled HVAC system, a subspace identification followed by a prediction error method was applied. In this identification, the external temperature (which is the disturbance to the system) is considered as the input (u), and the outputs are the header flow temperature (y_1), the return temperature (y_2), as well as ground and first floor room temperatures in the Nimbus building (y_3 and y_4, respectively) and the Rubicon building (y_5 and y_6, respectively). Each of these quantities is illustrated in Fig. 3. Simulated samples were generated at a rate of one sample per minute, to match the Nimbus BMS sampling rate.

The simulated environment was validated by replicating the external temperature variation measured by the Nimbus BMS during a 24-h period and using these values as the simulation input signal u. Fig. 6 shows the output variables y_1 to y_6, for both simulated and real data. Because the simulated environment is a simplified model of the entire BMS system, the dynamics are not expected to match perfectly in the two data sets. Furthermore, the initial conditions of the real system were not exactly replicated in the simulated data set. However, the similarity between the systems, overall, allows the performance of the proposed security measures on the simulated environment to be considered indicative of what the performance would be on a real operational system.

Attacks were simulated by modifying the value of one or more measured variables y_1 to y_6. The disturbances to the variables under attack were generated by a random walk process, in order to explore the performance of the system

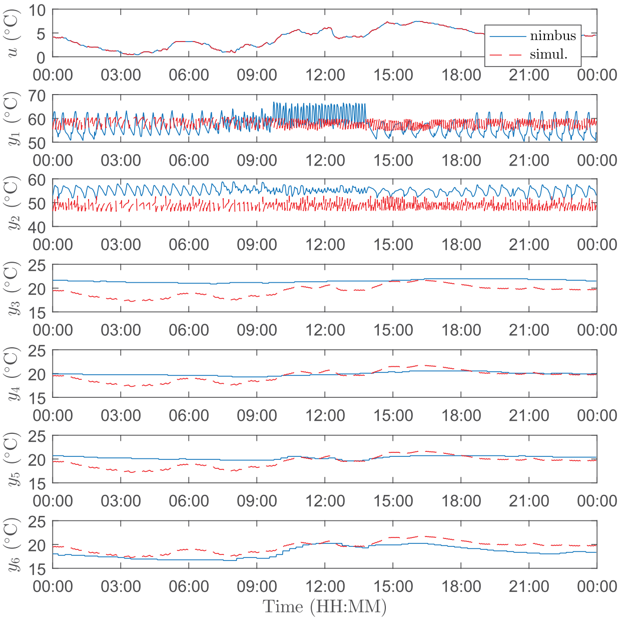

Fig. 6. ***Simulated data for "healthy" system (red) and measured data from Nimbus BMS (blue), with the same external temperature signal.***

in the face of varying levels of attack. Attacked data were simulated for three individual days, with the input signal u, following external temperature variations measured from the real system for randomly selected days in February 2016. Using the measured external temperature data, as the input, ensures that the operating conditions for simulation are realistic. Each of the three attacked datasets has a duration of 24 h, with four separate attacks occurring, each with a duration of 2 h; the system was allowed to return to normal operating conditions in between attacks to avoid making attack detection easier due to the cumulative effect of multiple attacks.

In order to train the detectors, so-called "healthy" data, with no attacks was simulated for nine days, using measured external temperature data from February 2016. It should be noted that the external temperature data for this healthy dataset were not the same as those for the attacked data, to ensure that attack detection was not simply the result of overfitting the model to the healthy training data. The training data set, referred to in Section IV-A, contained seven of the nine days of healthy data. One of the attacked data sets and data from one of the remaining two healthy days was used for tuning of the model parameters for the 1-SVM. Finally, the remaining two attacked data sets (which contain both healthy and attacked data), along with the final day of healthy data, were used to evaluate the performance of the detectors. For each attacked data set, a corresponding healthy data set was, also, generated, in order to quantify the effect of the attacks.

B. Attack Detection

Figs. 7 and 8 show binary indicators of the attack and detection status for the KB detector, based on static rules, and the DD detector, based on the raw measured variables as 1-SVM features. For the test data sets using external temperature data from February 9, 2016 and February 11, 2016, attacks occur, each lasting 2 h. During each attack, one variable (from $y_1,\cdots,y_6$) was manipulated with a disturbance, whose magnitude was generated by a random walk process. For the test data set using external temperature data from February 12, 2016, no attack was present; this data set was used to quantify the false alarm rate. The performance of the KB detectors relies heavily on the choice of appropriate rules and thresholds, as will be discussed in more detail later. For the 1-SVM detector, the target false alarm rate can be varied by tuning the model parameters. For the detectors, whose results are shown in the figures, the expected time between false alarms was found to be 75.8 and 72.0 min, for the KB detector and DD detector, respectively; thus, they are considered to have a comparable performance. The figures show that all attacks were eventually detected by both KB and DD detectors.

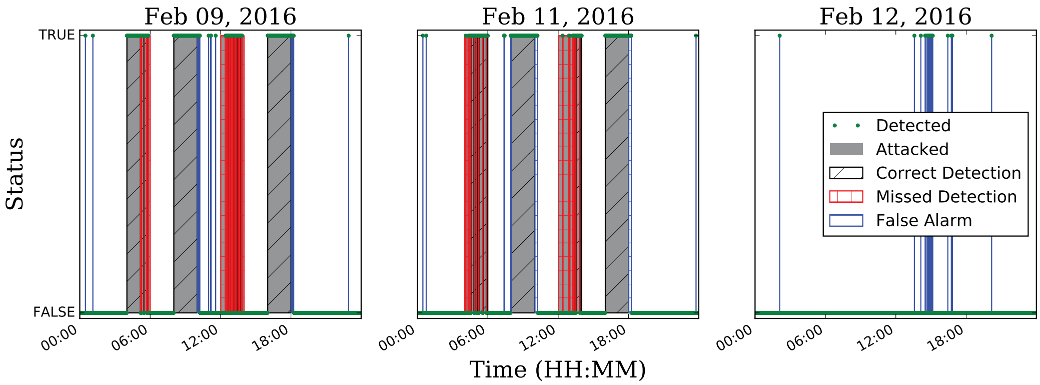

Fig. 7. ***Attack and detection status (1 = TRUE, 0 = FALSE) versus time of day for KB detector.***

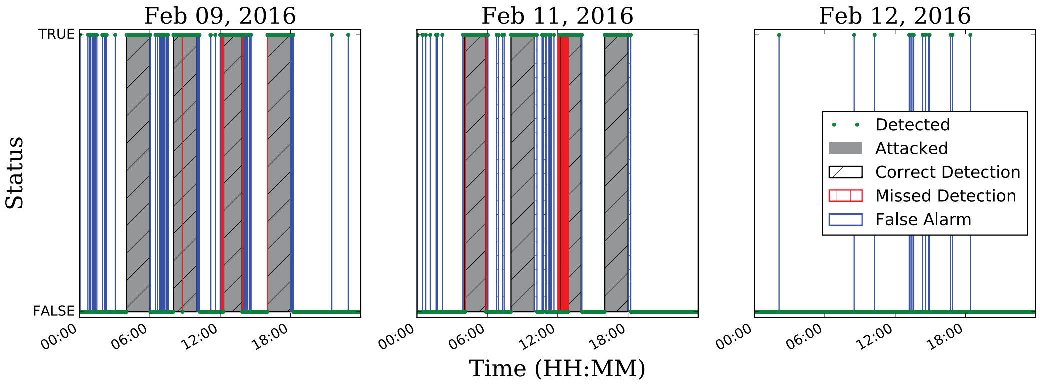

Fig. 8. ***Attack and detection status versus time of day for DD detector.***

For the DD detector, there appear to be more false alarms in between attacks, than for the KB detector. However, when it is considered that, between attacks in the simulated data set, the systems is recovering from the previous attack, it should not be assumed to be operating under “normal” conditions. Thus, these apparent false alarms are to be expected. Similarly, the so-called “false alarms” that occur between 00:00 and 04:00 on February 9, 2016 for the DD detector were found to be caused by the settling time of the simulation model; during this time, the operation of the simulated system was not in accordance with normal steady state operation. As such, although this period does not strictly represent a simulated attack, it is a positive result that the DD detector identified such behavior as abnormal.

For the KB detector, there were a greater number of attacked samples for which detection was missed, a deeper analysis of which shows that the extent to which data were manipulated for these samples was small, relative to the nominal operating conditions (i.e., injected temperature deviation of the order of 1 °C, relative to nominal temperatures around 60 °C). This highlights the limitations of the KB detector implemented here as a standalone method for detecting stealthy attacks and the sensitivity of such detectors to the specification of the underlying models. More robust methods for specifying the KB detectors rules, such as those reported in [38] and [39], would likely result in an improved performance.

The detection delay (i.e., the time between the instant when the attack began and the instant when it was first detected) varies for each attack. On average, the DD detector had a shorter detection delay than the KB detector, as shown in Table 1, which shows the amount of time, in minutes, between the onset of the attack and the first sample classified as an anomaly, for each attack in the test data sets.

In the case of the attack at 12:00 on February 11, 2016, even the DD detector did not detect the attack until the seventh sample of manipulated data. Analysing this case in more depth, to understand why the delay was greater than that for other attacks, it was seen that the degree of manipulation of the data in that case was minimal, during the initial period of the attack, with less than 0.3 °C of temperature disturbance, compared to a nominal value of 56 °C. As such, it is clear that the detection delay is dependent on the magnitude of the disturbance, which varied throughout the attacks. Urbina *et al.* [7] proposed a method of evaluating attack detectors, using the expected time between false alarms $E[T_{fa}]$, and the maximum deviation that the attacker can achieve in the attacked variable, while remaining stealthy. Here, we adopt Urbina *et al.*’s method for comparing the performance of detectors, because detection delay, alone, does not represent the impact of the data manipulation on the system under attack.

Because multiple different variables may be attacked in a smart grid or EMS, the maximum deviation could be defined in many different ways. For illustrative purposes, we use the maximum value of the sum of the absolute deviation of the header supply and return temperatures as the metric of interest for measuring the severity of the attack that can go undetected. These temperatures are considered to be relevant because loss of control over them could lead to overheating in the boiler and pose a safety risk for users of the building. However, the temperatures at other points throughout the system, such as room temperatures, could also be used as alternative metrics of interest. The maximum deviation reported in the following results is the maximum deviation that was undetected in the simulated attacks; it is not a theoretical maximum value.

Table 1 Simulated Attack Detection Delay for KB and DD Detectors

Attack start time	KB delay (mins)	DD delay (mins)
09-Feb-2016 04:00	0	0
09-Feb-2016 08:00	0	0
09-Feb-2016 12:00	24	2
09-Feb-2016 16:00	0	1
11-Feb-2016 04:00	6	0
11-Feb-2016 08:00	0	0
11-Feb-2016 12:00	22	7
11-Feb-2016 16:00	0	0

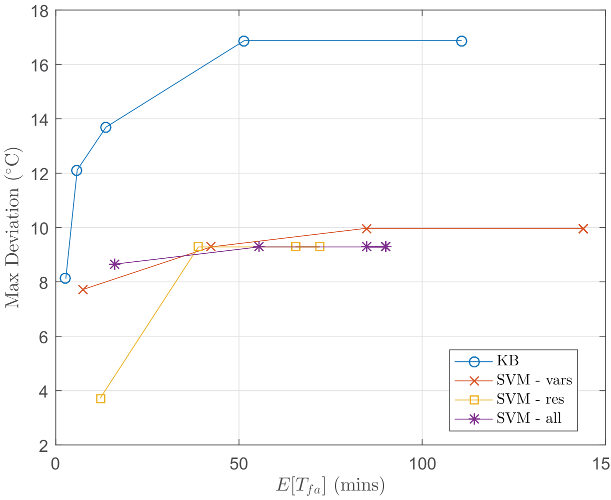

Fig. 9. ***Maximum (experimental) temperature deviation versus E[T_{fa}] for KB and SVM detectors.***

Fig. 9 shows the tradeoff between the false alarm rate and the disturbance to the system that can be achieved by an attacker when the KB and DD (1-SVM) detectors are implemented. Various DD detectors, each using a different set of features at the input, are illustrated. In the experiments, the 1-SVM detectors can be seen to outperform the KB detector, resulting in a smaller maximum undetected deviation for the same false alarm rate. Three 1-SVM detectors are shown in the figure, each with a different set of features at the input: SVM-vars uses the raw variables measured by the sensors as features; SVM-res uses the residues calculated from the KB rules as features; SVM-all uses the raw variables, the KB rule residues and time-domain statistics as features. The performance of SVM-res and that of SVM-all are seen to be marginally better than that of SVM-vars, suggesting that combining both expert knowledge and machine learning can be beneficial for attack performance.

C. Resilient Control

The performance of the proposed resilient control for the BMS is evaluated in this section. In the simulation results shown in Fig. 10, the measurements $y_1, y_2, \ldots, y_6$, as defined previously, are under consideration. In this system, we find that it is enough to protect any one of these measurements to prevent 0−stealthy attacks, and the measurement y_6 has been chosen to be protected. Results are shown for three different cases:

- healthy BMS: where no attack is carried out on the system;
- attacked BMS: where the attacker corrupts some of the measurements, but no resilience policy is in the place to recover the system from abnormal state;
- attack-resilient BMS: where the attacker corrupts some of the measurements, and the resilience policy recovers the system and returns it to normal operating conditions after the attack detection.

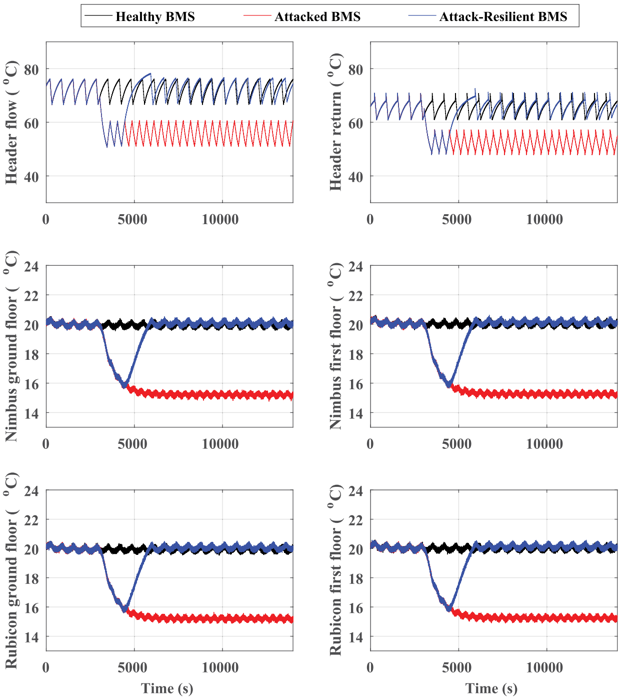

Fig. 10. ***Performance comparison of the healthy BMS, attacked BMS, and attack-resilient BMS, in the presence of delay in attack detection (the attack starts at time 3000 s and is detected at 4440 s).***

Here, for both the attacked BMS and attack-resilient BMS data sets, a combined attacks scenario, with attacks on y_1 and y_2, simultaneously, is considered to start at instant, $k' = 3000$ s. For all samples with time greater than k', the measurements of the header supply and header return temperatures y_1 and y_2, respectively, are manipulated, by adding 15 °C to each of them, and are fed to the respective controllers. This attack on y_1 and y_2 leads to high safety risk due to potential damage to CHP in the attacked BMS, as the return temperature becomes too low (below 65 °C) after the attack. In the attack-resilient BMS data set, the attack is considered to be detected by the SIA, and the resilient policy is triggered, such that the corrupted measurements are replaced with their estimates.

Some major factors, such as communication delays, the large volume of data to be processed, and time-consuming security analysis algorithms, can affect real-time attack detection. To investigate the performance of the proposed resilient control situations where a detection delay occurs, the following scenario is considered: the attack starts at time $k = 3000$ s, and is detected at $k = 4440$ s (i.e., with a 24-min detection delay, which is the maximum detection delay in Table 1). As is shown in Fig. 10, the attack-resilient BMS has the same outputs as the attacked BMS, until attack detection ($k = 4440$ s), but it can successfully recover the system and return it to normal operating conditions after detection. Other simulations, with different detection delays, have been carried out and, in all the cases, the attack-resilient BMS has successfully recovered the system and returned it to normal operating conditions after attack detection. The figure shows that, in this attack scenario, the attack-resilient BMS is robust against the multimeasurement attack and has a stable performance, similar to that of the healthy BMS, after attack detection.

VI. CONCLUSION

In this work, a framework for ICS security has been proposed, incorporating both data analytics for attack detection and a resilient control policy, based on the adaptation of virtual sensors. It is clear that implementing the framework in existing ICS does not require any major modification to the local controllers because the framework is implemented in the supervisory layer. It was shown that data-driven anomaly detection provides a promising detection improvement over the knowledge-based detectors employed in this study, with respect to the tradeoff between the expected time to false alarm and the maximum undetected deviation achieved by an attack. Future work to investigate other methods of defining the KB detector rules will be carried out, using complex system modeling approaches, and similar. When it is considered that data-driven methods are also more flexible and adaptable, in the case of changes to system configuration, for example, due to addition, removal or replacement of equipment, it suggests that such methods should be adopted for securing ICS against cyber attacks. Further work with larger and more diverse data sets and a wide range of simulated attack scenarios will be carried out to gain a deeper understanding of the combinations of anomaly detection algorithms and feature sets, including knowledge-based system models as features, that achieve the best performance. Simulation results also showed that the proposed controller reconfiguration approach can recover the system from abnormal states, even when a detection delay exists. The proposed resilience policy has, thus, been shown to be effective to ensure that stability and performance of the system are maintained, even under attack conditions.

Considering the performance of the SIA, from the point of view of the maximum deviation achieved by an attacker versus the expected false alarm rate, allows the relationship between the attack detector and the resilience policy to be understood clearly: using the protected measurements only for state estimation degrades the quality of the system, relative to the estimate based on all measurements, in the case that no attack has occurred; thus, it is vital to minimize the frequency of false alarms. However, in the case where an attack has occurred, it is important that the delay before triggering the resilience policy be as short as possible, in order that the deviation from normal, stable operating conditions is minimized. The third element of the proposed framework is the attack isolation stage and this remains to be explored in future work, whereby the SIA and resilience policy will cooperate to isolate the variable(s) which have been corrupted, allowing the virtual sensor to use as many healthy variables as possible for estimation. Another important aspect of this work, which remains to be explored in more detail in the future, is the question of determining, in a systematic way, which subset of measurements should be protected, in order to prevent undetectable attacks. ■

Acknowledgements

The authors would like to thank Prof. A. Teixeira from Delft University of Technology and Dr. T. Sadamoto from Tokyo Institute of Technology for their valuable suggestions.

REFERENCES

[1] Q. Zhu, C. Rieger, and T. Başar, “A hierarchical security architecture for cyber-physical systems,” in *Proc. 4th Int. Symp. Resilient Control Syst. (ISRCS)*, Aug. 2011, pp. 15–20.

[2] M. Blanke, M. Kinnaert, J. Lunze, and M. Staroswiecki, *Diagnosis Fault-Tolerant Control*. Berlin, Germany: Springer-Verlag, 2006.

[3] Z. Gao, C. Cecati, and S. X. Ding, “A survey of fault diagnosis and fault-tolerant techniques—Part I: Fault diagnosis with model-based and signal-based approaches,” *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3757–3767, Jun. 2015.

[4] Z. Gao, C. Cecati, and S. X. Ding, “A survey of fault diagnosis and fault-tolerant techniques—Part II: Fault diagnosis with knowledge-based and hybrid/active approaches,” *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3768–3774, Jun. 2015.

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

[6] A. Teixeira, G. Dán, H. Sandberg, and K. H. Johansson, “A cyber security study of a SCADA energy management system: Stealthy deception attacks on the state estimator,” *IFAC Proc. Volumes*, vol. 44, no. 1, pp. 11271–11277, Jan. 2011.

[7] D. I. Urbina, *et al.*, “Limiting the impact of stealthy attacks on industrial control systems,” in *Proc. 23rd ACM Conf. Comput. Commun. Secur.*, Vienna, Austria, Oct. 2016, pp. 1092–1105.

[8] A. A. Cárdenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, and S. Sastry, “Challenges for securing cyber physical systems,” in *Proc. Workshop Future Directions Cyber-Phys. Syst. Secur. (DHS)*, 2009, pp. 1–7.

[9] *U.S.-Canada Power System Outage Task Force, Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations.*, Natural Resources Canada, 2004.

[10] J. P. Farwell and R. Rohozinski, “Stuxnet and the future of cyber war,” *Survival*, vol. 53, no. 1, pp. 23–40, 2011.

[11] E. Kovacs. (2016). *BlackEnergy Malware Used in Ukraine Power Grid Attacks.* [Online]. Available: http://www.securityweek.com/blackenergy-group-uses-destructive-plugin-ukraine-attacks

[12] A. Teixeira, K. Paridari, H. Sandberg, and K. H. Johansson, “Voltage control for interconnected microgrids under adversarial actions,” in *Proc. IEEE 20th Conf. Emerg. Technol. Factory Autom. (ETFA)*, Sep. 2015, pp. 1–8.

[13] S. Gold, “The SCADA challenge: Securing critical infrastructure,” *Netw. Secur.*, vol. 2009, no. 8, pp. 18–20, Aug. 2009.

[14] G. A. Francia, III, D. Thornton, and J. Dawson, “Security best practices and risk assessment of SCADA and industrial control systems,” in *Proc. Int. Conf. Secur. Manage. (SAM), Steering Committee World Congr. Comput. Sci., Comput. Eng. Appl. Comput. (WorldComp)*, 2012, p. 1.

[15] P. A. S. Ralston, J. H. Graham, and J. L. Hieb, "Cyber security risk assessment for SCADA and DCS networks," *ISA Trans.*, vol. 46, no. 4, pp. 583–594, 2007.

[16] C. A. Ericson, II, "Fault tree analysis—A history," in *Proc. Syst. Safety Conf.*, 1999.

[17] SPARKS, "SPARKS Deliverable D2.6: Smart grid vulnerability and risk assessment," Tech. Rep., September 2016. [Online]. Available: https://project-sparks.eu/publications/deliverables/

[18] SPARKS, "SPARKS deliverable D4.4: High-level design documentation (for grid control system) and a deployment architecture for the monitoring solution," Tech. Rep., Sep. 2016. [Online]. Available: https://project-sparks.eu/publications/deliverables/

[19] H. Khurana, M. Hadley, N. Lu, and D. A. Frincke, "Smart-grid security issues," *IEEE Security Privacy*, vol. 8, no. 1, pp. 81–85, Feb. 2010.

[20] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on smart grid state estimation: Attack strategies and countermeasures," in *Proc. 1st IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Oct. 2010, pp. 220–225.

[21] H. Sandberg, A. Teixeira, and K. H. Johansson, "On security indices for state estimators in power networks," in *Proc. 1st Workshop Secure Control Syst. (SCS)*, Stockholm, Sweden, 2010.

[22] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 58, no. 11, pp. 2715–2729, Nov. 2013.

[23] O. Vukovic, K. C. Sou, G. Dan, and H. Sandberg, "Network-aware mitigation of data integrity attacks on power system state estimation," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 6, pp. 1108–1118, Jul. 2012.

[24] Y. Yuan and Y. Mo, "Security in cyber-physical systems: Controller design against known-plaintext attack," in *Proc. 54th IEEE Conf. Decision Control (CDC)*, Dec. 2015, pp. 5814–5819.

[25] J. Wurm, "Introduction to cyber-physical system security: A cross-layer perspective," *IEEE Trans. Multi-Scale Comput. Syst.*, to be published.

[26] M. Govindarasu, A. Hann, and P. Sauer, "Cyber-physical systems security for smart grid," in *The Future Grid to Enable Sustainable Energy Systems*. PSERC Publication, 2012.

[27] A. A. Cardenas, P. K. Manadhata, and S. P. Rajan, "Big data analytics for security," *IEEE Security Privacy*, vol. 11, no. 6, pp. 74–76, Nov./Dec. 2013.

[28] L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, and S. S. Sastry, "Foundations of control and estimation over lossy networks," *Proc. IEEE*, vol. 95, no. 1, pp. 163–187, Jan. 2007.

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

[30] M. Pajic, "Robustness of attack-resilient state estimators," in *Proc. ACM/IEEE Int. Conf. Cyber-Phys. Syst. (ICCPS)*, Apr. 2014, pp. 163–174.

[31] N. Bezzo, J. Weimer, M. Pajic, O. Sokolsky, G. Pappas, and I. Lee, "Attack resilient state estimation for autonomous robotic systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2014, pp. 3692–3698.

[32] F. Pasqualetti, F. Dorfler, and F. Bullo, "Control-theoretic methods for cyberphysical security: Geometric principles for optimal cross-layer resilient control systems," *IEEE Control Syst.*, vol. 35, no. 1, pp. 110–127, Feb. 2015.

[33] S. Amin, X. Litrico, S. Sastry, and A. M. Bayen, "Cyber security of water SCADA systems—Part I: Analysis and experimentation of stealthy deception attacks," *IEEE Trans. Control Syst. Technol.*, vol. 21, no. 5, pp. 1963–1970, Sep. 2013.

[34] V. Valdivia, "Sustainable building integrated energy test-bed," in *Proc. IEEE 5th Int. Symp. Power Electron. Distrib. Gener. Syst. (PEDG)*, Jun. 2010, pp. 1–6.

[35] S. Timlin, "Improving automation routines for automatic heating load detection in buildings," *J. Sustain. Eng. Design*, vol. 1, no. 2, 2012, Art. no. 2.

[36] L. Ljung, "Prediction error estimation methods," *Circuits, Syst. Signal Process.*, vol. 21, no. 1, pp. 11–21, Jan. 2002.

[37] K. Paridari, "Optimal and resilient control with applications in smart distribution grids," Licentiate dissertation, Roy. Inst. Technol., Stockholm, Sweden, 2016.

[38] M. T. Khan, D. Serpanos, and H. Shrobe, "A rigorous and efficient run-time security monitor for real-time critical embedded system applications," in *Proc. IEEE 3rd World Forum Internet Things (WF-IoT)*, Reston, VA, USA, Dec. 2016, pp. 100–105.

[39] S. Gao, L. Xie, A. Solar-Lezama, D. Serpanos, and H. Shrobe, "Automated vulnerability analysis of ac state estimation under constrained false data injection in electric power systems," in *Proc. 54th IEEE Conf. Decision Control (CDC)*, Osaka, Japan, Dec. 2015, pp. 2613–2620.

[40] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed.,M. Jordan, J. Lawless, S. Lauritzen, and V. Nair, Eds. New York, NY, USA: Springer-Verlag, 2000.

[41] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.

[42] A. Rai, D. Ward, S. Roy, and S. Warnick, "Vulnerable links and secure architectures in the stabilization of networks of controlled dynamical systems," in *Proc. Amer. Control Conf. (ACC)*, Jun. 2012, pp. 1248–1253.

[43] G. Böker and J. Lunze, "Stability and performance of switching Kalman filters," *Int. J. Control*, vol. 75, nos. 16–17, pp. 1269–1281, 2002.

ABOUT THE AUTHORS

Kaveh Paridari received the B.Sc. and M.Sc. degrees in electrical engineering (automatic control) from Sharif University of Technology, Tehran, Iran in 2009 and 2011, respectively, and the Licentiate degree from the Department of Automatic Control, Royal Institute of Technology (KTH), Stockhol, Sweden, in 2016, where he is currently working toward the Ph.D. degree at the Department of Electric Power and Energy Systems.

His research interests include distributed control, optimization, cyber–physical systems, energy systems, and smart electricity grids.

Mr. Paridari was one of the Best Student Paper Award Finalist at the IEEE International Conference on Automation Science and Engineering in 2014, and a recipient of the Best Paper Award from the IEEE International Conference on Industrial Technology in 2013.

Niamh O'Mahony received the B.E.E.E. degree from University College Cork, Cork, Ireland, in 2006 and the Ph.D. degree, under joint supervision from University College Cork and the University of Calgary, Calgary, AB, Canada, in 2010.

She is a Senior Research Scientist with Dell EMC Research Europe, based in the Centre of Excellence in Ireland. Her primary research interests include data analytics, machine learning, and digital signal processing.

Alie El-Din Mady received the M.Sc. degree in embedded-systems design from Universita della Svizzera Italiana, Lugano, Switzerland and the Ph.D. degree in computer science from University College Cork, Cork, Ireland.

He joined the United Technologies Research Center, Cork, Ireland, in 2012 where he is currently working as a Senior Research Scientist and Principal Investigator in Cybersecurity Research. His research interests include cyber–physical system, cybersecurity, model-based design, and embedded systems. He has authored and coauthored patents and journal and conference papers in the areas of system of systems, embedded systems, cyber–physical systems, and cybersecurity.

Rohan Chabukswar received the B.S. degree in engineering physics from Indian Institute of Technology Bombay, Mumbai, India, in 2008 and the M.S. and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2009 and 2014, respectively.

Since then he has been with the United Technologies Research Center, Cork, Ireland, where he currently holds the position of Senior Research Scientist. His research interests include diagnostics and decision support in aerospace, and secure control of cyber–physical systems.

Dr. Chabukswar received the UTRC's Outstanding Achievement Award for codevelopment of new technologies in 2016.

Menouer Boubekeur received the Ph.D. degree for research into formal verification of asynchronous circuits from the University of Joseph Fourier, Saint-Martin-d'Hères, France, in 2004.

He has more than 18 years of industrial and academic research and technology management experience in innovation, software/system engineering, cyber–physical systems, security, and energy. He is currently an Associate Director leading External Research Development activities at the European Research Center for UTC, high-technology products and support services provider to customers in the aerospace and building industries worldwide. He has proven records in writing and winning European proposals in various technology/application domains. He also leads a number of projects in cyber–physical system areas, including cybersecurity, Internet-of-Things (IoT), and digital factory. He has authored and coauthored patents and journal and conference papers in the areas of complex systems, embedded and real-time systems, cyber–physical systems, and cybersecurity.

Henrik Sandberg received the M.Sc. degree in engineering physics and the Ph.D. degree in automatic control from Lund University, Lund, Sweden, in 1999 and 2004, respectively.

He is Professor at the Department of Automatic Control, KTH Royal Institute of Technology, Stockholm, Sweden. From 2005 to 2007, he was a Postdoctoral Scholar at the California Institute of Technology, Pasadena, CA, USA. In 2013, he was a visiting scholar at the Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. He has also held visiting appointments at the Australian National University and the University of Melbourne, Australia. His current research interests include security of cyber–physical systems, power systems, model reduction, and fundamental limitations in control.

Dr. Sandberg was a recipient of the Best Student Paper Award from the IEEE Conference on Decision and Control in 2004 and an Ingvar Carlsson Award from the Swedish Foundation for Strategic Research in 2007. He is an Associate Editor of the *Automatica* and the IEEE Transactions on Automatic Control.