A NON-PARAMETRIC CUMULATIVE SUM APPROACH FOR ONLINE DIAGNOSTICS OF CYBER ATTACKS TO NUCLEAR POWER PLANTS

Wei Wang[1], Francesco Di Maio[1], Enrico Zio[1,2]

[1]*Energy Department, Politecnico di Milano, Via La Masa 34, 20156 Milano, Italy*

[2]*Chair on System Science and the Energy Challenge, Fondation Electricite' de France (EDF), CentraleSupélec, Université Paris-Saclay, Grande Voie des Vignes, 92290 Chatenay-Malabry, France*

Abstract: Failures and cyber attacks can both compromise the integrity of Cyber-Physical Systems (CPSs). Cyber attacks manifest themselves in the physical system and, can be misclassified as component failures, leading to wrong control actions and maintenance strategies. In this chapter, we illustrate the use of a non-parametric cumulative sum (NP-CUSUM) approach for online diagnostics of cyber attacks to CPSs. This allows for a prompt recognition of cyber attacks from component failures, and effective actions for CPSs protection. We apply the approach to the Advanced Lead-cooled Fast Reactor European Demonstrator (ALFRED) and its digital Instrumentation and Control (I&C) system. For this, an object-oriented model previously developed is embedded within a Monte Carlo (MC) engine that allows injecting into the I&C system both components (stochastic) failures (such as sensor bias, drift, wider noise and freezing) and cyber attacks (such as Denial of Service (DoS) attacks mimicking component failures).

Keywords: Cyber-Physical System; Cyber Attacks; Stochastic Failures; Diagnostics; Non-Parametric Cumulative Sum (NP-CUSUM); Nuclear Power Plant; Advanced Lead-cooled Fast Reactor European Demonstrator (ALFRED).

ABBREVIATIONS

CPS	Cyber-Physical System
NP-CUSUM	Non-Parametric CUmulative SUM
ALFRED	Advanced Lead-cooled Fast Reactor European Demonstrator
I&C	Instrumentation and Control
MC	Monte Carlo
NPP	Nuclear Power Plant
PI	Proportional-Integral
DoS	Denial of Service
PID	Proportional-Integral-Derivative
FDI	False Data Injection
SG	Steam Generator
FA	Fuel Assembly
CR	Control Rod
SISO	Single Input Single Output
DAC	Digital-to-Analog Converter
LSB	Least Significant Bit

NOMENCLATURE

P_{Th}	Thermal power
h_{CR}	Height of control rods
$T_{L,hot}$	Coolant core outlet temperature
$T_{L,cold}$	Coolant SG outlet temperature
Γ	Coolant mass flow rate
T_{feed}	Feedwater SG inlet temperature
T_{steam}	Steam SG outlet temperature
p_{SG}	SG pressure
G_{water}	Feedwater mass flow rate
G_{att}	Attemperator mass flow rate
kv	Turbine admission valve coefficient
P_{Mech}	Mechanical power
$K_{p,j}$	Proportional gain value of j-th PI
$K_{i,j}$	Integral gain value of j-th PI
t	Time
t_R	Accident time
t_M	Mission time
Δt	Sensor measuring time interval
y	Variable (safety parameter)
y^{ref}	Reference value of controller set point value of y
$y^{real}(t)$	Real value of y
$y^{sensor}(t)$	Sensor measurement
$y^{feed}(t)$	Measurement received by the computing (feeding) subsystem
$y^{monitor}(t)$	Measurement received by the monitoring subsystem
$Y(t)$	Redundant channel measure, $Y = y^{feed}$ and $y^{monitor}$
$\delta_y(t)$	Sensor measuring error
$q_y(t)$	Converter quantization error

a	Accidental scenario
b	Bias factor
c	Drift factor
$S_Y(t)$	Score function-based statistic of the collected $Y(t)$, $S_Y(t) = S_y^{feed}(t)$ and $S_y^{monitor}(t)$
h_y	Positive threshold
τ_Y	Time to alarm, $\tau_Y = \tau_y^{feed}$ and $\tau_y^{monitor}$
$\Delta\tau_y$	Delay difference between τ_y^{feed} and $\tau_y^{monitor}$
Γ_y^{ref}	Reference delay difference
c_y	NP-CUSUM parameter
ε_y	NP-CUSUM parameter
ω_y	NP-CUSUM positive weight
g_Y	Score function
Δg_Y	Score function difference value
μ_Y	Pre-change mean value of Y
θ_Y	Post-change mean value of Y
$\hat{\theta}_Y(t)$	On-line estimate of θ_Y
$\mu_{\Delta g_Y}$	Known pre-change mean value of Δg_Y
$\theta_{\Delta g_Y}$	Unknown post-change mean value of Δg_Y
α_y^h	False alarm rate
β_y^h	Missed alarm rate
$\gamma\left(\Gamma_{T_{L,cold}}^{ref}\right)$	Misclassification rate with respect to Γ_y^{ref}

1. INTRODUCTION

Cyber-Physical Systems (CPSs) feature a tight combination of (and coordination between) the system computational units and physical elements. To the benefit of safe operation, the integration of computational resources into physical processes is aimed at adding new capabilities to stand-alone physical systems, to enable functionalities of real-time monitoring, dynamic control and decision support during normal operation as well as in case of accidents. In CPSs, cyber and physical processes are dependent and interact with each other through feedback control loops (e.g., embedded cyber controllers monitor and control the system physical variables, whilst physical processes affect, at the same time, the monitoring system and the computation units by wired or wireless networks (Kim and Kumar, 2012; Lee, 2008; Alur, 2015)). The benefit of such self-adaptive capabilities is the reason why CPSs are increasingly operated in energy, transportation, medical and health-care, and other applications (Lee, 2008; Khaitan and McCalley, 2015; Bradley and Atkins, 2015). In the context of nuclear energy, the introduction of digital Instrumentation and Control (I&C) systems allows Nuclear Power Plants (NPPs) to take advantage of the new technologies in the field, for safe operation (IAEA, 2009).

In the context of CPSs, sensor measurements can be used to monitor the behavior of the systems under different operational conditions, including hazardous and malicious ones. Indeed, CPS functionality and integrity can be compromised by both hazards (safety related) and malicious threats (security related) (Piètre-Cambacédès and Bouissou, 2013; Kriaa et al., 2015; Zalewski et al., 2016). Hazards and cyber threats originate from different sources (stochastic degradations and accidental conditions, for the former, external malevolent activities that are usually less accessible and less predictable for the latter (Aven, 2009; Kriaa et al., 2015)). Distinct properties and mechanisms between them suggest different assessment methodologies for their identification.

The difficulty lies in the fact that components hazards and malicious threats can lead to similar consequences on the system (Rahman et al., 2016; Wang et al., 2017b;

Li and Huang, 2016; Kornecki and Liu, 2013). For example, in a situation where system shutdown is demanded, both failure of the shutdown of the actuator and interception of the shutdown command by an attacker result in unavailability of the safety action. In such situation, diagnosing of the failure cause would allow taking the right decision to respond to the system shutdown unavailability with the right emergency procedure (e.g., manual operation of the actuation in the case of such cyber attack).

In this sense, diagnostic of cyber attacks and component failures is important for the system protection and resilience, allowing prompt recovery from the effects of disruptive events and, thus, increasing system resilience (Obama, 2013; Moteff, 2012; Zio, 2009; Zio, 2016; Fang and Sansavini, 2017; Hu et al., 2017).

In this work, we develop a nonparametric cumulative sum (NP-CUSUM) detection approach (Qiu and Hawkins, 2003; Tartakovsky et al., 2006a; Tartakovsky et al., 2006b; Tartakovsky et al., 2013) for diagnosing cyber attacks, distinguishing them from component failures. The proposed approach is illustrated considering the possible occurrence of stochastic components failures and cyber attacks in the digital I&C system of the Advanced Lead Fast Reactor European Demonstrator (ALFRED) (Alemberti et al., 2013). An object-oriented simulator previously developed (Ponciroli et al., 2014; Ponciroli et al., 2015), and comprising a multi-loop Proportional-Integral (PI) control scheme (Skogestad and Postlethwaite, 2007), is utilized for simulating the ALFRED dynamic response to failures and cyber attacks.

The main original contribution of this work lies in prompt recognition of cyber attacks from component failures in CPSs by relying on simultaneous treatment, within the NP-CUSUM approach, of the measurements taken from redundant channels, guiding decisions for recovering CPSs from anomalies and for CPSs protection.

The chapter is organized as follows. Section 2 sets the issue of security analysis in the framework of risk assessment and, highlights the contributions of cyber attack diagnostics to overall system resilience. Section 3 presents the main characteristics of the ALFRED reactor, with the data measuring and transmission, and control schemes in the channels of its digital I&C system at full power nominal conditions, and the MC

engine for injection of component failures and cyber breaches. In Section 4, the proposed NP-CUSUM diagnostic algorithm is presented. The NP-CUSUM diagnostic method is illustrated in Section 5 and evaluated with respect to its diagnostic performances, such as false alarm, missed alarm and misclassification rates in Section 6. Conclusions are drawn in Section 7.

2. HAZARDS AND THREATS FOR CPSs

CPSs demand that in the risk analysis both safety and security aspects are considered (Eames and Moffett, 1999; Kriaa et al., 2015; Pièvre-Cambacédès and Bouissou, 2013; Zalewski et al., 2016). With respect to safety, hazards relate to components failures that can result in accidental scenarios leading to unacceptable consequences on the system physical processes; as for security, malicious attacks can impair both the physical and cyber parts of the system, possibly leading to unacceptable consequences.

2.1 Hazards

Failures of both hardware and software can compromise CPS integrity and functionality.

During operation, failures of embedded hardware components (e.g., sensors and actuators) can be induced by aging, degradation, and process and operational conditions, which modify the way components work and interact with each other, generating multiple failure modes (Wang et al., 2016). For example, sensors can degrade and fail in different modes such as bias, drift and freezing (Wang et al., 2016); actuators can fail stuck, accidentally driving the physical process to be isolated from the controlling units of the cyber domain (Zio and Di Maio, 2009; Zaytoon and Lafortune, 2013).

Components failures can lead to two types of misoperations: (1) failure on-demand, e.g., failing to trigger protections or execute proper control strategies (when demanded); (2) malfunction, e.g., spurious triggering of protections (e.g., unintentional shutdown) or incorrect execution of control actions. Failures on-demand and malfunctions of both hardware and software components have gained increasing attention in the risk

community (Aldemir et al., 2010; McNelles et al., 2016).

Resilience of CPS to failures can be granted by self-adaptiveness of control decisions on actuators, resorting to intelligent control systems that properly manipulate sensors measurements (Machado et al., 2016). For example, Proportional-Integral-Derivative (PID) controllers, typically used as feedback controller in CPS to retroact to actuators the actions to be undertaken for responding to changes of physical parameters, may suffer of software failures/errors (generated from inadequate specification, incomplete testing scope and algorithm/logic failures) that are latent and triggered only when context modifications are to be met (Aldemir et al., 2010; Jockenhövel-Barttfeld et al., 2016). In these situations, control rules adaptability to variable physical conditions is a fundamental requirement to the robustness of CPS for resilience during CPS operation.

2.2 Threats

CPSs reliance on digitalization and remote control systems increases their exposure to cyber attacks to controllers, databases, networks and human-system interfaces, that can result in the loss of system integrity and/or functionality. Malicious activities can be manifested as Denial of Service (DoS) attacks (Zargar et al., 2013; Yuan et al., 2013; Rahman et al., 2016), False Data Injection (FDI) attacks (e.g., packet/data modification) (Liang et al., 2017; Tan et al., 2017; Mohammadpourfard et al., 2017), network scan & sniffing attacks (Trabelsi and Rahmani, 2005; Rahman et al., 2016), integrity attacks (e.g., through malware contagion) (Ntalampiras, 2015; Ntalampiras, 2016) and, illegal command executions (Shin et al., 2015). They can be initiated in the cyber domain through local or remote accesses, mimicking the components failures but isolating the connectivity between cyber and physical systems, leaving the physical process uncontrolled and possibly drifting towards severe consequences.

Cyber attacks can cause serious security and privacy issues (Xiang et al., 2017). Under cyber attacks, e.g., by contagion of malware, security-related system features may result to be compromised and, the system safety potentially endangered. The

identification of the cyber threats most affecting the system response is quite important for decision-making on optimal protection and resilience, as prevention and mitigation of malicious attacks contribute to guaranteeing CPS integrity and functionality (Yuan et al., 2014; Fang and Sansavini, 2017; Hu et al., 2017; Wang et al., 2017a).

2.3 Hazards and Threats Diagnosis

From the perspective of integrated safety and security of CPSs, distinguishing cyber attacks from component failures is important for anticipating the potential impact on the system integrity and defining proper protection and mitigation actions for resilience.

Cyber threats aimed at altering the CPS normal operation can be diagnosed either by comparison of statistical estimates of occurrence probabilities from field data collected on the real CPS with reference values of failure probabilities of the CPS components (Wang et al., 2017b; Jockenhövel-Barttfeld et al., 2016; Shin et al., 2015), or by scenario processing (i.e., modeling the malicious cyber events and their manifestation on the physical domain, affecting, in turn, both cyber and physical properties of the CPS) (Debar et al., 1999; Tartakovsky et al., 2006a; Carl et al., 2006).

A variety of methods for scenario processing have been proposed, based on artificial intelligence techniques. In general terms, observations are compared with the normal conditions measurements and a deviation from the legitimate data flow is found by methods such as the Sequential Probability Ratio Test (SPRT) (Wald, 1947; Hines and Garvey, 2006), the Cumulative Sum (CUSUM) chart (Tartakovsky et al., 2006a; Tartakovsky et al., 2006b; Page, 1954), the Exponentially Weighted Moving Average (EWMA) inspection scheme (Roberts, 1959), the Reversible-jump Markov Chain Monte Carlo (RJ-MCMC) (Zio and Zoia, 2009; Zhao and Chu, 2010) and the control charts (Xie et al., 2002).

Practically, both components stochastic failures and cyber attacks occur at unknown times, leading to unpredictable changes in the distributions of physical variables that differ from the normal condition distribution. The sequential Non-Parametric CUSUM (NP-CUSUM) approach (Tartakovsky et al., 2006a) has been

shown capable of distinguishing normal from abnormal conditions. Based on this approach, we originally design a framework for early diagnostics of cyber attacks in CPSs. The novelty of the work lies in the reliance on the simultaneous treatment within the NP-CUSUM approach of the measurements taken from redundant channels.

3. THE ADVANCED LEAD-COOLED FAST REACTOR EUROPEAN DEMONSTRATOR

The ALFRED reactor and the MC scheme for injecting component failures and cyber breaches are described in Sections 3.2 and 3.3, respectively.

3.1 ALFRED Description

ALFRED is a small-size (300 MW) pool-type LFR, whose primary system configuration is shown in Fig. 1 (Alemberti et al., 2013). All components of the primary cooling system, including core, primary pumps and Steam Generators (SGs), are contained in the main reactor vessel, located in a large pool within the reactor tank. The ALFRED core providing the thermal power P_{Th} is composed by wrapped hexagonal Fuel Assemblies (FAs) with pins arranged on a triangular lattice. Control Rods (CRs) systems adjusting the heights of CRs h_{CR} have been foreseen for power regulation and reactivity swing compensation during a fuel cycle, and for scram purposes with the required reliability for a safe shutdown (Grasso et al., 2013).

At full power nominal conditions, the coolant (i.e., lead) flow coming from the cold pool enters the core at temperature $T_{L,cold}$ equal to 400 °C and, once passed through the core, it is collected in the volume of the hot collector at temperature $T_{L,hot}$ equal to 480 °C; from there, it is distributed to eight parallel pipes and delivered to as many SGs. After leaving the SGs, the coolant enters the cold pool through the cold leg and returns to the core.

The eight SGs work at pressure p_{SG} equal to 180 bar. The feedwater of the secondary cooling system flows in the SGs, at pressure p_{SG} and temperature T_{feed} equal to 335 °C, and leaves the SGs after absorbing heat from the primary coolant, entering the turbine as steam at temperature T_{steam} equal to 450 °C (at full power nominal

conditions). From a control point of view, it is worth noticing that the steam mass flow rate is considered proportional to the inlet pressure and governed by maneuvering the turbine valve admission (*kv*), not by throttling. An attemperator is foreseen between the SG outlet header and the turbine, to limit the steam temperature at the turbine inlet T_{steam}, keeping it as close as possible to its nominal value, by adjusting the attemperator mass flow rate G_{att}.

Eventually, ALFRED produces mechanical power P_{Mech} to be transformed for the power grid.

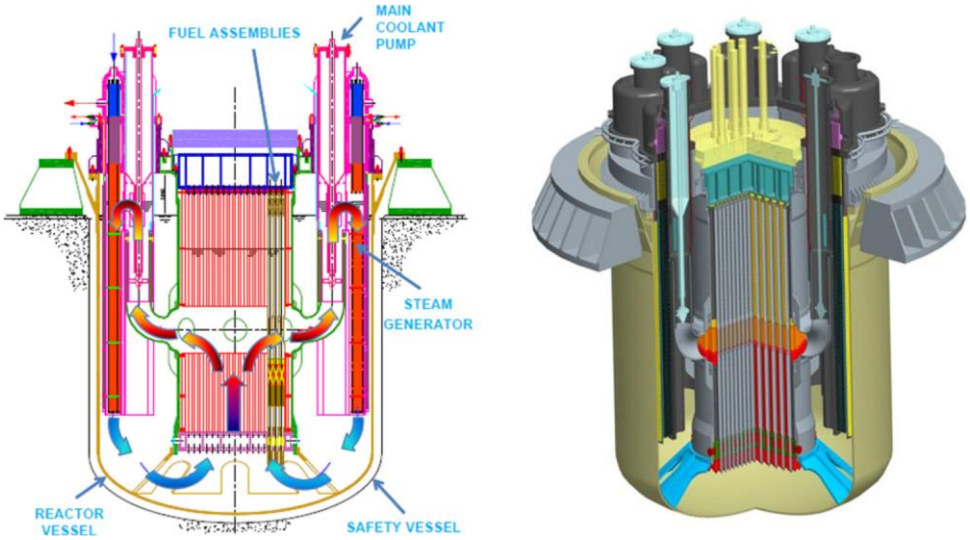

Fig. 1. ALFRED primary system layout (Alemberti et al., 2013)

A simplified schematics of the ALFRED primary and secondary cooling systems is shown in Fig. 2. The parameters specification of ALFRED at full power nominal conditions are reported in Table 1.

Table 1 ALFRED parameters values, at full power nominal conditions

Parameter	Parameter Description	Value	Unit
P_{Th}	Thermal power	$300 \cdot 10^6$	W
h_{CR}	Height of control rods	12.3	cm
$T_{L,hot}$	Coolant core outlet temperature	480	°C
$T_{L,cold}$	Coolant SG outlet temperature	400	°C
Γ	Coolant mass flow rate	25984	$kg \cdot s^{-1}$
T_{feed}	Feedwater SG inlet temperature	335	°C
T_{steam}	Steam SG outlet temperature	450	°C
p_{SG}	SG pressure	$180 \cdot 10^5$	Pa
G_{water}	Feedwater mass flow rate	192	$kg \cdot s^{-1}$
G_{att}	Attemperator mass flow rate	0.5	$kg \cdot s^{-1}$
kv	Turbine admission valve coefficient	1	-
P_{Mech}	Mechanical power	$146 \cdot 10^6$	W

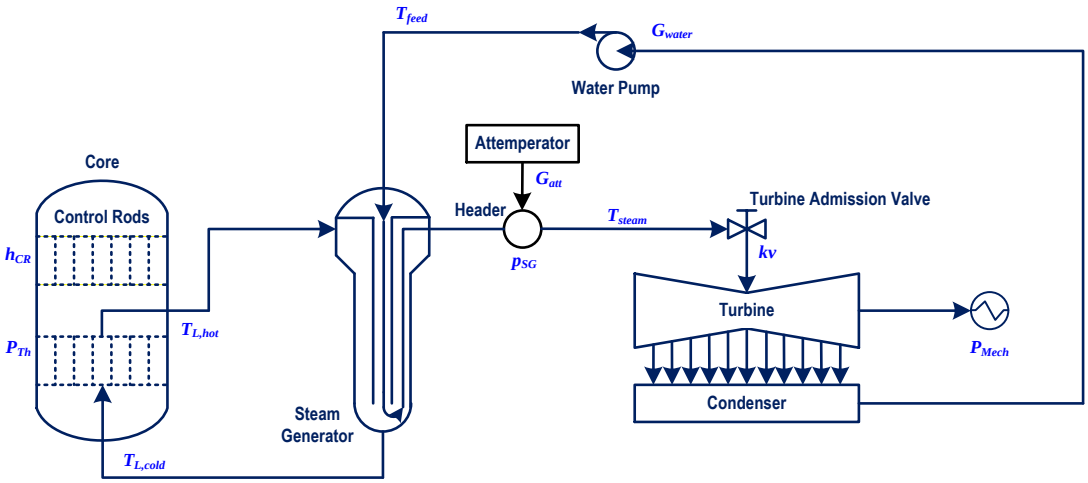

Fig. 2. ALFRED simplified schematics

3.2 The Reactor Digital Control Scheme

A multi-loop PI (Proportional and Integral) digital control scheme, i.e., a decentralized control scheme, is developed because of its simplicity of implementation and robustness to malfunctioning of the single control loops (Ponciroli et al., 2015). Indeed, it can be regarded as constituted by several redundant SISO (Single Input Single Output) control loops (Levine, 1996).

To design the regulators and simulate the system controlled response, an object-oriented model of the entire plant has been developed (Fig. 3). Based on the Modelica language (Fritzson, 2010) and implemented in the Dymola environment (DYMOLA, 2015), the corresponding simulator has been built by connecting several dedicated models for the description of the reactor (for details, see Ponciroli et al., 2014; Ponciroli et al., 2015).

Both feedback and feedforward digital control schemes are adopted for ALFRED (see Fig. 3 shadowed part). The control aims at keeping the controlled variables of the control loops approximately at the steady state values, for operating a constant mechanical power. The PI-based feedback control configuration employs four SISO control loops independent of each other (Ponciroli et al., 2015). The parameters of the PI regulators reported in Table 2 are calibrated by adopting the procedures commonly employed for the SISO systems and the tuning for each PI control loop is verified by

adopting the Bode criterion (Levine, 1996). The values represent the optimal working conditions of the system at full power nominal conditions. The overall control scheme has been verified to effectively damp disturbances due to the change of the operating conditions. The proposed feedback scheme is improved by adding a feedforward control action, thanks to which the water mass flow rate (G_{water}) is adjusted according to the value of the thermal power (P_{Th}) exchanged at the SG interface. The implemented feedforward controller allows adjusting the heat exchange conditions in the SGs and enhancing the robustness of the control system against errors on the evaluation of the time delay between the SGs and the core due to transport phenomena.

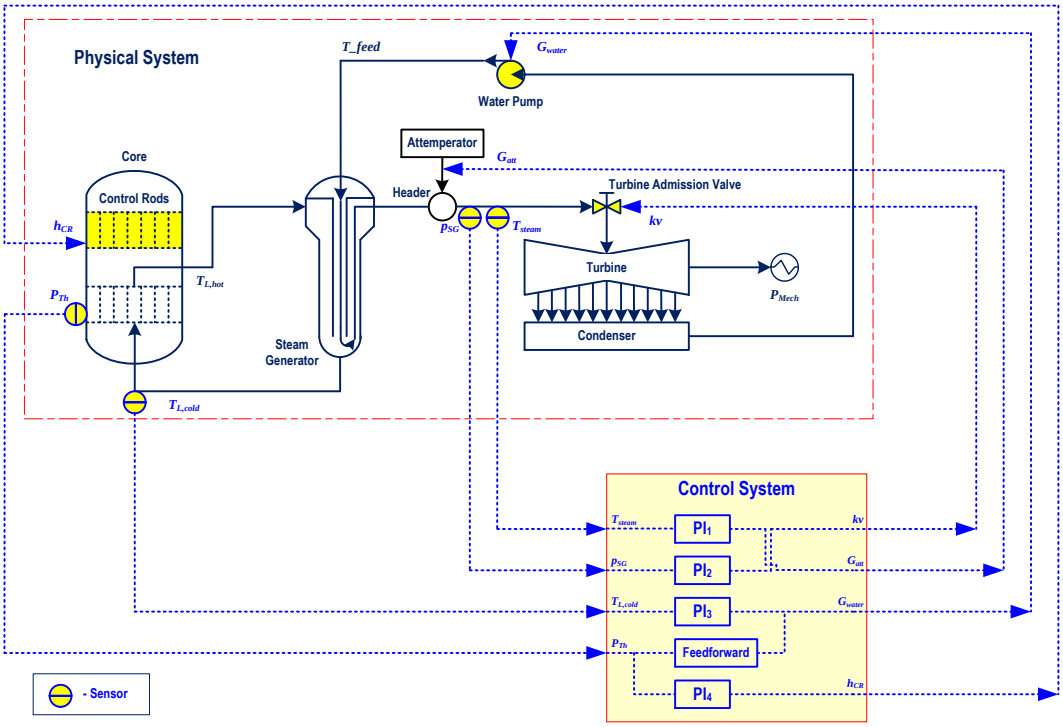

Fig. 3. ALFRED reactor control scheme

Table 2 Parameters of PI controllers

PI	Control Loop		Controller Parameters, j=1,2,3,4	
	Controlled variable	Control variable	$K_{p,j}$	$K_{i,j}$
PI_1	T_{steam} (°C)	G_{att} (kg·s^{-1})	$1 \cdot 10^{-1}$	$5 \cdot 10^{-2}$
PI_2	p_{SG} (Pa)	kv (-)	$3 \cdot 10^{-7}$	$1 \cdot 10^{-8}$
PI_3	$T_{L,cold}$ (°C)	G_{water} (kg·s^{-1})	$6 \cdot 10^{-1}$	$1 \cdot 10^{-2}$
PI_4	P_{Th} (W)	h_{CR} (cm)	$2 \cdot 10^{-11}$	$4 \cdot 10^{-11}$

Redundancy is commonly applied to sensors and signal processing units of a

digital I&C system (Authen and Holmberg, 2012). In the ALFRED digital control scheme, redundancy has been used to design each independent SISO loop.

Fig. 4 shows an example of the redundant design scheme of the $T_{L,cold}$-PI$_3$-G_{water} control loop. The real values of the coolant SG outlet temperature $T_{L,cold}(t)$ are measured by a sensor. After collected and converted to quantized (discretized) values by a data acquisition system, the measurements are duplicated by two identical digital-to-analog converters (DACs) to Subsystem 1 for computing (feeding) and 2 for monitoring, respectively. The received measurements of Subsystem 1 $T_{L,cold}^{feed}(t)$ is fed to the computational unit PI$_3$, whereas those of Subsystem 2 $T_{L,cold}^{monitor}(t)$ are taken as redundant data, for detecting anomalous conditions of the physical system.

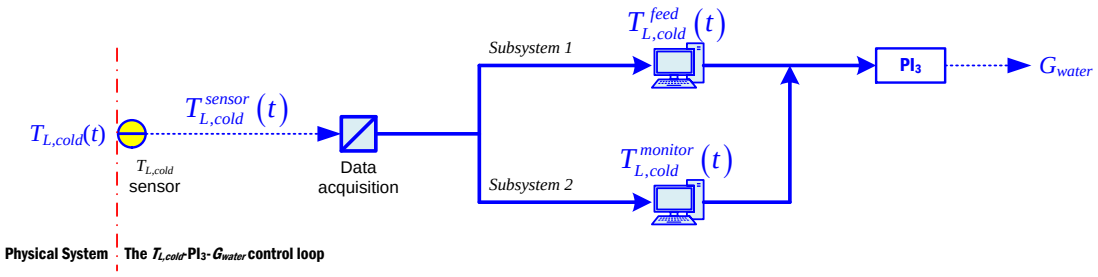

Fig. 4 The redundancy design of the $T_{L,cold}$-PI_3-G_{water} control loop

Measurements are realistically considered to be affected by two types of errors (Gray and Neuhoff, 1998; Widrow, 1961): measurement errors (assumed distributed according to a normal distribution) and quantization errors (which are rooted in the DACs and are assumed uniformly distributed between −1/2 and +1/2 Least Significant Bit (LSB)). For simplicity, but without loss of realism, Table 3 lists the reference values of the controlled variables, the distributions of sensor measurement errors and the quantization errors that each control loops is subjected to.

Table 3 List of reference parameters for safety variables

Variable, y	Reference value, y^{ref}, at full power nominal conditions	Sensor measuring error $\delta_y(t)$	Converters quantization error $q_y(t)$
T_{steam} (°C)	450	$N(0,1)$	[-0.05, +0.05]
p_{SG} (Pa)	$180 \cdot 10^5$	$N(0,0.1) \cdot 10^5$	$[-0.01,+0.01] \cdot 10^5$
$T_{L,cold}$ (°C)	400	$N(0,1)$	[-0.05, +0.05]
P_{Th} (W)	$300 \cdot 10^6$	$N(0,0.5) \cdot 10^6$	$[-0.05,+0.05] \cdot 10^6$

In Fig. 5, measurements from the four control loops of the ALFRED are shown, on a time horizon t_M equal to 1000s,: the values of the variables are kept approximately at their nominal values, at full power nominal conditions, with some measurement errors (white noise) and quantization errors.

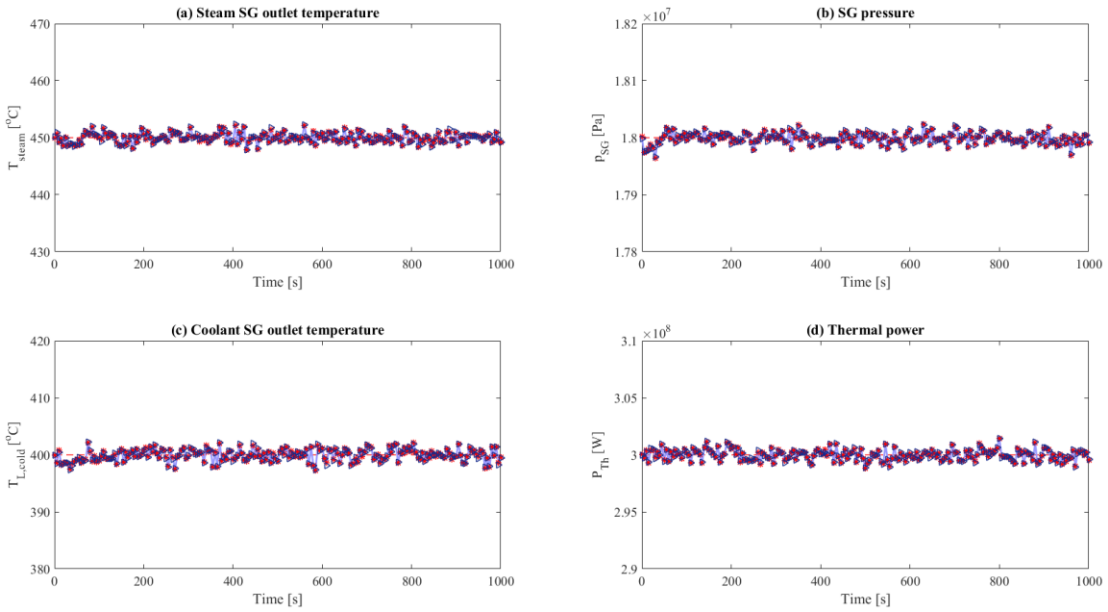

Fig. 5. Measurements from the four control loops of ALFRED at full power nominal conditions (star values for computing subsystem and triangle values for monitoring subsystem): (a) Steam SG outlet temperature; (b) SG pressure; (c) Coolant SG outlet temperature; and (d) Thermal power

3.3 Failures and Cyber Breaches

To model failures and cyber attacks, a MC sampling scheme is integrated with the ALFRED model for injecting stochastic failures of sensors and cyber breaches, at random times and of random magnitudes.

Four types of sensor failure modes that may occur at random time t_R are considered (Boskvic and Mehra, 2002): (*a*) bias, (*b*) drift, (*c*) wider noise and (*d*) freezing (see dotted lines in Fig. 6 (a), (b), (c) and (d), respectively). The occurrence of any of these failure modes results in altered sensor measurements $y^{sensor}(t)$, as in Eq. (1):

$$y^{sensor}(t)=\begin{cases} y(t)+\delta(t), & \delta(t)=N(0,\sigma),\sigma>0, \quad t\geq 0, \quad normal \\ y(t)+\delta(t)+b, & \dot{b}(t)\equiv 0, b(t_F)\neq 0, \quad t\geq t_R, \quad bias \\ y(t)+\delta(t)+c(t), & c(t)=c\cdot(t-t_F), \quad t\geq t_R, \quad drift \\ y(t)+\delta'(t), & \delta'(t)=N(0,\alpha\sigma),\alpha>1, \quad t\geq t_R, \quad wider \quad noise \\ y^{sensor}(t_R-\Delta t), & \quad t\geq t_R, \quad freezing \end{cases} \tag{1}$$

where $y(t)$ is the real value of the controlled variable y at time t, $\delta(t)$ is the nominal measuring error, distributed according to a normal distribution $N(0,\sigma)$, b is a constant bias factor, c is a constant drift factor, $\delta'(t)$ is a wider measuring error, distributed according to a normal distribution $N(0,\alpha\sigma)$ with a variance larger than $\delta(t)$ ($\alpha>1$).

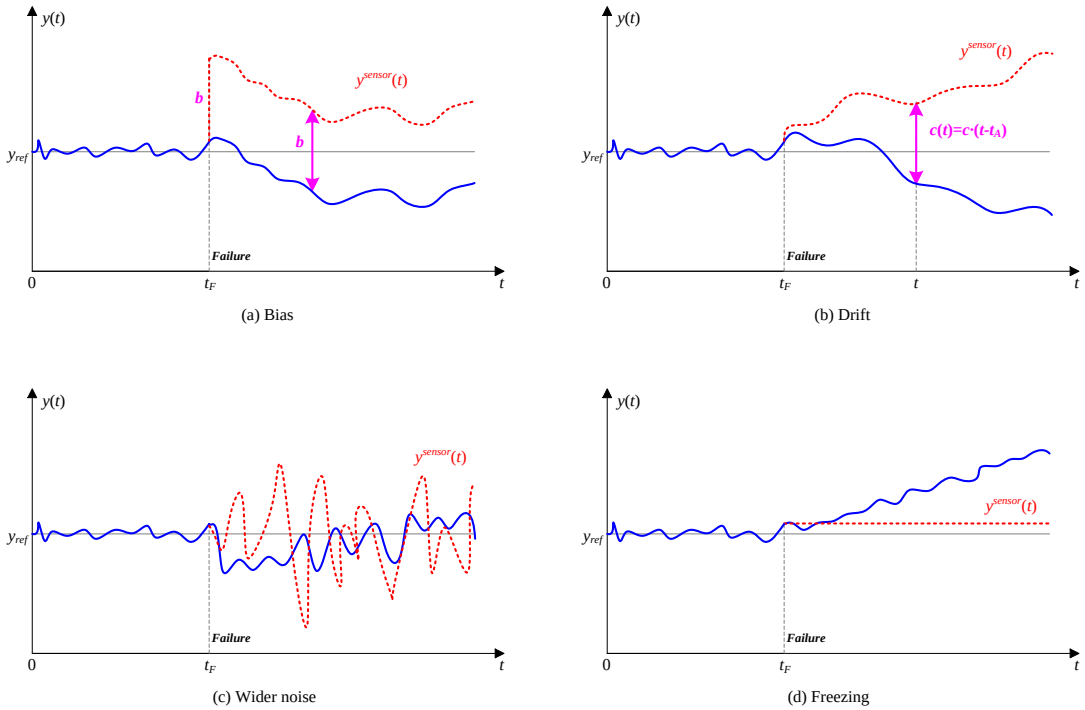

Fig. 6. Sensor failure modes: (a) bias; (b) drift; (c) wider noise; and (d) freezing. Solid lines represent the real measurements of the controlled variables, whereas dotted lines are the altered measurements of the failed sensors (for the sake of clarity, measurement and quantization errors are neglected)

Without loss of generality, only the $T_{L,cold}$ sensor (see Fig. 7) is hereafter considered (but the following discussion remains valid for any other sensor of the I&C system). Stochastic failures cause differences of the measurements $T_{L,cold}^{sensor}(t)$ from the real values of the controlled variable in the physical system. The MC sampling procedure used to inject stochastic failures to the $T_{L,cold}$ sensor at uniform random time t_R consists in sampling the uncertain parameters b, c, $\delta'(t)$ from the distributions listed in Table 4 and, then, running the ALFRED simulator for generating the controlled variables evolution throughout the mission time t_M. Erroneous measurements are, then, converted to two sets of quantized data in the data acquisition system and fed to both

the computing (feeding) and monitoring subsystems.

Table 4 Parameters of sensors

Sensor	Nominal error $\delta(t)$	Failure factors			
		Bias b	Drift c	Wider noise $\delta'(t)$	Freezing
$T_{L,cold}$ (°C)	$N(0,1)$	$U(-30,30)$	$U(-1,1)$	$N(0,5)$	0

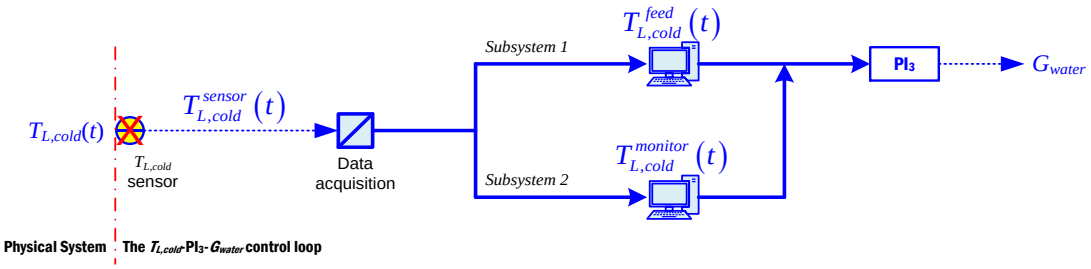

Fig. 7 Schematics of $T_{L,cold}$ sensor stochastic failures

Alternatively, a DoS attack is modelled to block a legitimate packet traffic that processes the genuine connection and is substituted by a malicious packet traffic (Tartakovsky et al., 2006a; Carl et al., 2006). Fig. 8 shows the schematics of a DoS attacks, in which the computing unit is fed by malicious packet traffic, altering the legitimate information, whereas, a legitimate packet traffic is regularly fed to the monitoring unit. DoS attacks are modelled to occur at uniform random time t_R within the time horizon t_M, and the uncertain parameters b, c, $\delta'(t)$ are sampled from the distributions of Table 4, as previously explained for the sensor failure.

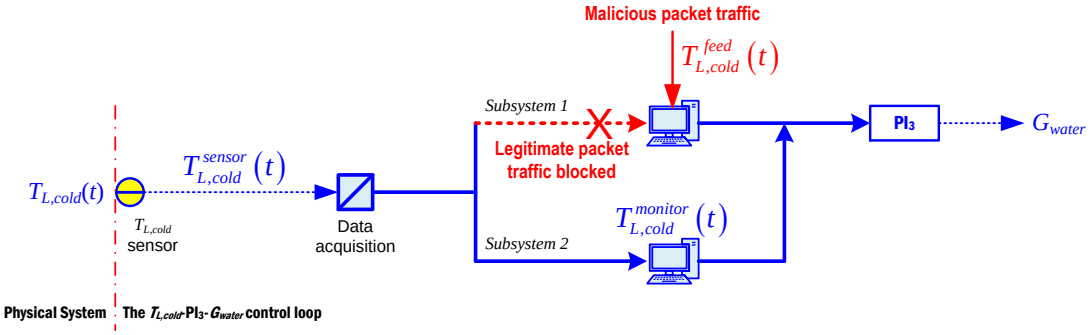

Fig. 8 Schematics of DoS attacks

4. THE NON-PARAMETRIC CUMULATIVE SUM APPROACH FOR REAL-TIME DIAGNOSTICS OF CYBER ATTACKS

The diagnostic approach is based on a NP-CUSUM algorithm of literature (Tartakovsky et al., 2006a), whose details are given in Appendix A.

4.1 The Diagnostic Approach

The diagnostics approach is here illustrated with reference to the stochastic failures and the DoS attacks described in Subsection 3.3. As shown in Fig. 9, the approach involves two main steps: (*i*) on-line collection of measurements received by the controllers, which are fed to the NP-CUSUM algorithm that is (off-line) trained on different system behaviors to set its parameters; (*ii*) an on-line application of the rules of classification of failures and cyber attacks.

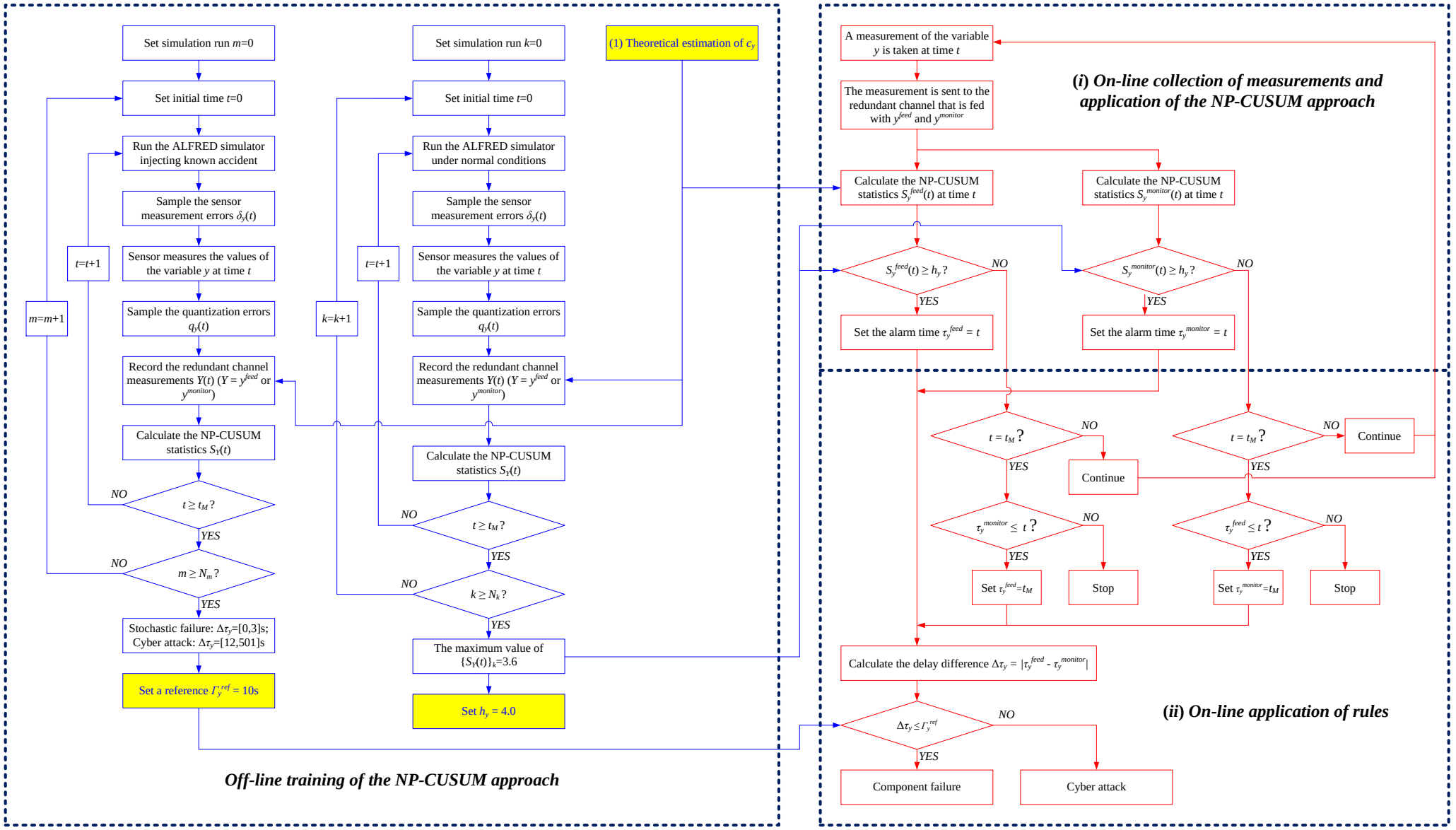

Fig. 9 Flowchart of the NP-CUSUM diagnostic approach

(*i*) On-line collection of measurements and application of the NP-CUSUM approach

The redundant channel measurements $Y(t)$, $Y=y^{feed}$ and $y^{monitor}$, where $y = T_{L,cold}$, are collected online by the subsystems as follows. At each time t,

(1) The sensor measures the values $y = T_{L,cold}$, which is affected by the sensor measurement error $\delta_y(t)$ distributed as a normal distribution of Table 3, i.e., $y^{sensor}(t) = y^{real}(t) + \delta_y(t)$;

(2) The data acquisition system collects and converts $y^{sensor}(t)$ with the quantization accuracy $q_y(t)$ of Table 3, resulting in two redundant channels of quantized measurements;

(3) The computing and monitoring subsystems receive the redundant measurements $Y(t)$;

(4) The NP-CUSUM algorithm calculates score function-based statistics $S_Y(t)$ of the collected $Y(t)$, to check whether either $S_y^{feed}(t)$ or $S_y^{monitor}(t)$ exceeds a predefined threshold h_y:

- If yes, record the time to alarm τ_Y (τ_y^{feed} or/and $\tau_y^{monitor}$, respectively, and proceed with the rule-based diagnostics at Step (*ii*));
- If either $S_y^{feed}(t)$ or $S_y^{monitor}(t)$ exceeds h_y, collect the successive measurement because the monitored component is working under normal conditions.

(*ii*) On-line application of rules

(5) If both τ_y^{feed} and $\tau_y^{monitor}$ exist, calculate the delay difference $\Delta\tau_y$ (i.e., denoting the difference between the time-to-detection delays τ_y^{feed} and $\tau_y^{monitor}$):

$$\Delta\tau_y = \left|\tau_y^{feed} - \tau_y^{monitor}\right| \quad (2)$$

Otherwise, set $\tau_y^{monitor}$ equal to t_M (when $S_y^{monitor}(t)$ has not exceeded h_y when $S_y^{feed}(t)$ does, and vice versa, respectively).

If neither exists before t_M, continue diagnostics.

(6) Compare $\Delta\tau_y$ with a predefined reference delay difference Γ_y^{ref} and take decision:

- If $\Delta\tau_y \le \Gamma_y^{ref}$, classify the event as ***Failure***;
- If $\Delta\tau_y > \Gamma_y^{ref}$, classify the event as ***Cyber Attack***.

The reference delay difference Γ_y^{ref} is estimated on a batch of $N_m = 100$ reference simulations, where, for each m-th simulation, a known component failure or cyber attack is injected. The minimum and maximum collected values of $\Delta\tau_y$ are found to be equal to 0s and 3s in case of components failures, and 12s and 501s in case of cyber attacks. Thus, we conservatively set Γ_y^{ref} equal to 10s, so that $\Delta\tau_y$ larger than 10s indicates that a cyber attack has occurred on the feeding subsystem.

(*iii*) Off-line training of the NP-CUSUM algorithm

The NP-CUSUM algorithm requires that the parameters c_y and h_y be customized to the different system behaviors, to guarantee the maximum capability of discriminating between failures and cyber attacks, in the ALFRED system.

(1) Estimation of c_y

A positive constant of c_y needs to be set in such a way to guarantee a negative mean value of $\mu_{\Delta g_Y} = \sum_t \Delta g_Y\left(Y(t)\right) \Big/ t$, $t = dt, 2dt, \ldots, t, (t<t_R)$, to hold before any anomaly (either failure or cyber attack) is detected, and a positive mean value $\theta_{\Delta g_Y} = \sum_t \Delta g_Y\left(Y(t)\right) \Big/ \left(t - t_R\right)$, $t = t_R, t_R+dt, t_R+2dt, \ldots$, to hold after the anomaly occurrence (Tartakovsky et al., 2006a), viz:

$$\mu_{\Delta g_Y} = E\left[\omega_y \cdot \left(\left|Y(t) - \mu_Y\right| - c_y\right)\right] = -\omega_y \cdot \left(\frac{2\sigma_Y}{\sqrt{2\pi}} - c_y\right) < 0 \tag{3}$$

$$\theta_{\Delta g_Y} \ge \omega_y \cdot \left(\left|\hat{\theta}_Y(t) - \mu_Y\right|_{\min} - c_y\right) > 0 \tag{4}$$

where, $\left|\hat{\theta}_Y(t) - \mu_Y\right|_{\min}$ is defined as the minimum difference between the estimated post-change mean $\hat{\theta}_{\Delta g_Y}$ and the known pre-change mean $\mu_{\Delta g_Y}$. As a result,

$$\frac{2\sigma_Y}{\sqrt{2\pi}} < c_y < \left|\hat{\theta}_Y(t) - \mu_Y\right|_{\min} \tag{5}$$

where,

$$c_y = \varepsilon_y \cdot \hat{\theta}_{Y,a} \tag{6}$$

where $\hat{\theta}_{Y,a}$ is a postulated post-change mean value for an accidental scenario *a*.

Since under normal conditions, the probability of $Y(t)$ (distributed according to a normal distribution $N(\mu_Y, \sigma_Y)$) of falling within the interval $[\mu_Y\text{-}2\sigma_Y, \mu_Y\text{+}2\sigma_Y]$ is at least equal to 0.95 (Duda et al., 1973), viz:

$$\Pr\left[\mu_Y - 2\sigma_Y \le Y(t) \le \mu_Y + 2\sigma_Y\right] \ge 0.95 \tag{7}$$

we assume an anomaly to be occurred if $\hat{\theta}_{Y,a}$ falls outside the interval $[\mu_Y\text{-}2\sigma_Y, \mu_Y\text{+}2\sigma_Y]$. Without loss of generality, we suppose that $\hat{\theta}_{Y,a} > \mu_Y$. The minimum value of $\hat{\theta}_{Y,a}$ results to be equal to $\mu_Y\text{+}2\sigma_Y$ and, thus, $\left|\hat{\theta}_{Y,a} - \mu_Y\right|_{\min}$ is equal to $2\sigma_Y$. Eqs. (5) and (6) change to:

$$\frac{1}{\sqrt{2\pi}}\left(1 - \frac{\mu_Y}{\mu_Y + 2\sigma_Y}\right) < \varepsilon_y < 1 - \frac{\mu_Y}{\mu_Y + 2\sigma_Y} \tag{8}$$

In conclusion, without loss of generality, we take a value of ε_y equal to:

$$\varepsilon_y = \frac{1}{2}\left(1 - \frac{\mu_Y}{\mu_Y + 2\sigma_Y}\right) \tag{9}$$

that, with respect to ($T_{L,cold}$ distributed as $N(400,1)$ºC) makes c_y turn out to be equal to 1.005ºC.

(2) Estimation of h_y

The threshold h_y can be set relying on a batch of N_k reference simulations under normal conditions, whose behaviors of the variable *y* without change points to failures or cyber attacks can be learnt, the NP-CUSUM statistics calculated and the parameter tailored to the simulation results. Specifically, we utilize N_k = 100 ALFRED randomly generated simulations. For each *k*-th simulation,

(a) Record the redundant channel measurements, $Y(t)$, $Y = y^{feed}$ or $y^{monitor}$, at each time t, $t = dt, 2dt, \ldots, t_M$;

(b) Calculate the corresponding NP-CUSUM statistics, $S_Y(t)$.

(c) Set the threshold h_y such that:

$$h_y > \max_{1 \le k \le N_k} \left\{ h_{y,k} \right\} \tag{10}$$

where,

$$h_{y,k} = \max_{1 \le t \le t_M} \left\{ S_Y(t) \right\}_k \tag{11}$$

and, $\left\{ S_Y(t) \right\}_k$ is the collection of the statistics for the k-th simulation.

As shown in Fig. 10 with respect to $T_{L,cold}$, the maximum value of the NP-CUSUM statistics is equal to 3.6 and, therefore, in what follows, we conservatively set $h_{T_{L,cold}}$ equal to 4.0.

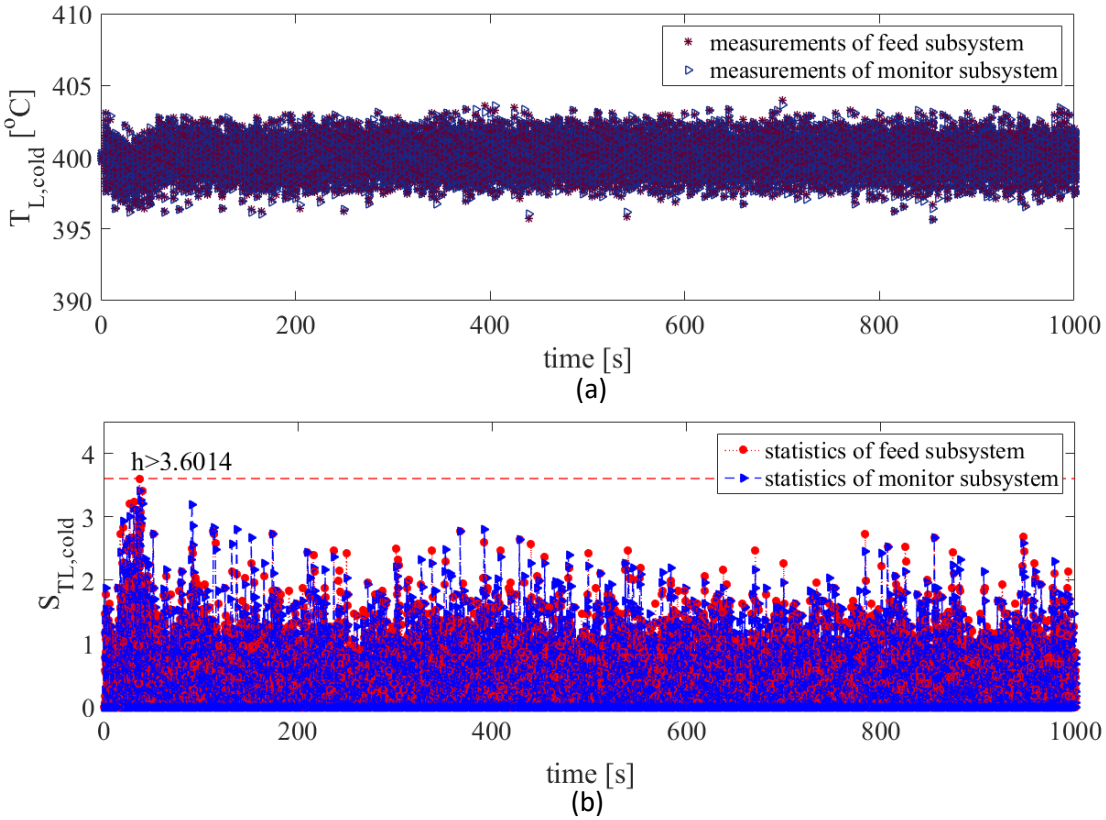

Fig. 10 Estimation of the threshold $h_{T_{L,cold}}$: (a) the received measurements of the two subsystems of the control loop; (b) the corresponding statistics calculated from the measurements

5. RESULTS

We illustrate the results of the NP-CUSUM-based diagnostic approach considering different $T_{L,cold}$ sensor failures and cyber attacks to the $T_{L,cold}$-PI$_3$-G_{water} control loop.

5.1 Bias Failure Mode

Fig. 11 presents the results of injecting bias failure at time t_R = 630s with a factor b equal to 7.569°C on $T_{L,cold}$ sensor measurements. As shown in Fig. 11(a), the $T_{L,cold}$ sensor bias failure deviates both measurements $T_{L,cold}^{feed}(t)$ and $T_{L,cold}^{monitor}(t)$ from the real values of the physical system $T_{L,cold}^{real}(t)$. Fig. 11 shows that the bias b results in very quick response of both statistics evaluated on the measurements $T_{L,cold}^{feed}(t)$ and $T_{L,cold}^{monitor}(t)$: both statistics reach quickly the threshold $h_{T_{L,cold}}$ (dotted line) and the difference $\Delta\tau_{T_{L,cold}}$ between times to alarm ($\tau_{T_{L,cold}}^{feed}$ and $\tau_{T_{L,cold}}^{monitor}$) turns out to be equal to 0 (i.e., less than Γ_y^{ref} equal to 10s) (see Fig. 11(b)), allowing for a (correct) identification of the event as a sensor failure mode and not as a cyber attack.

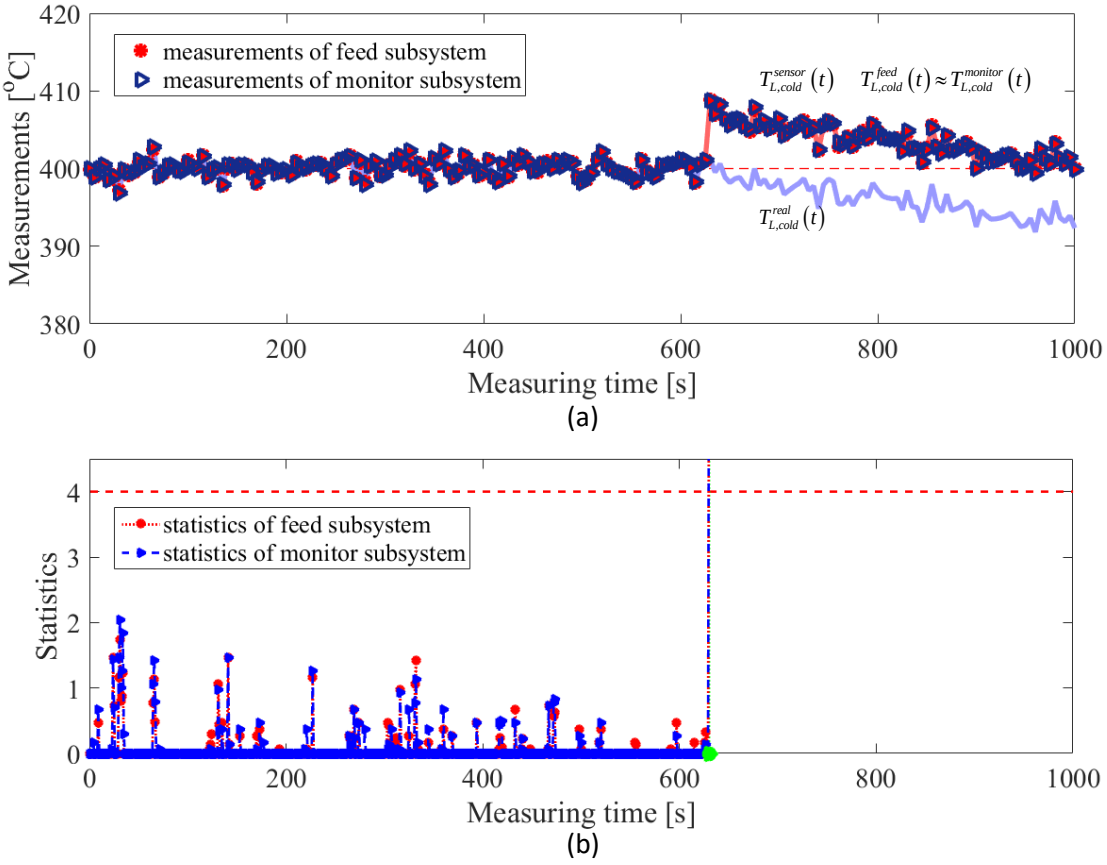

Fig. 11 $T_{L,cold}$ sensor bias failure mode: (a) the received measurements of feed and monitor Subsystems in which the bias occurs at time t_R equal to 630s; (b) the corresponding NP-CUSUM statistics for diagnosing the bias failure

Contrarily, Fig. 12(a) shows a cyber attack to the computing unit mimicking a bias failure mode at t_R=630s (with b again equal to 7.569°C): this leads $T_{L,cold}^{feed}(t)$ to deviate from $T_{L,cold}^{monitor}(t)$ (that, indeed, is the legitimate $T_{L,cold}^{sensor}(t)$ measured by the $T_{L,cold}$ sensor). The different values between the malicious and the legitimate measurements, then, lead to a delay response $\Delta\tau_{T_{L,cold}}$ equal to 66s (larger than Γ_y^{ref}) between the threshold exceedance of $S_{T_{L,cold}}^{monitor}(t)$ and $S_{T_{L,cold}}^{feed}(t)$ (see Fig. 12(b)), and allowing for a (correct) identification of the event as a cyber attack.

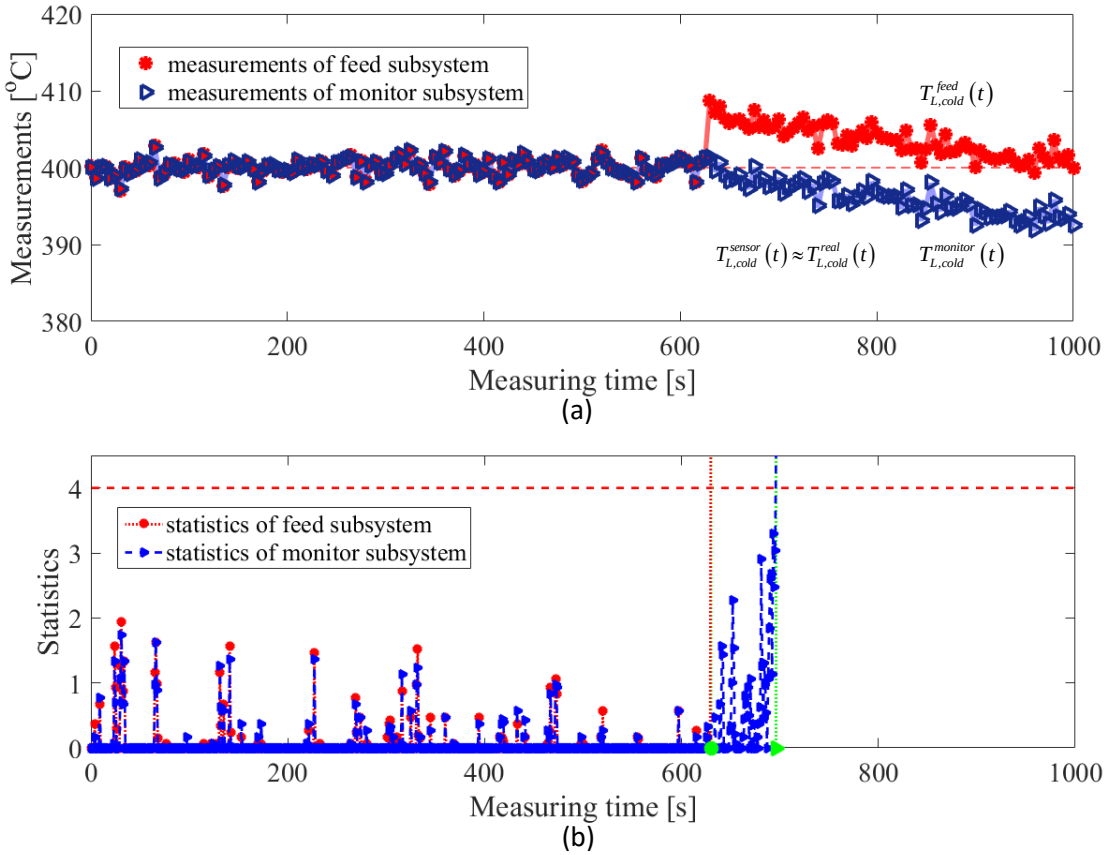

Fig. 12 Cyber attack to the computing unit mimicking a bias failure mode: (a) the received measurements of feed and monitor Subsystems in which the cyber attack occurs at time t_R equal to 630s; (b) the corresponding NP-CUSUM statistics for diagnosing the cyber attack

5.2 Drift Failure Mode

Fig. 13 presents the results of injecting a drift at time t_R = 740s, with the drift factor c equal to 0.398. The drift c results in a very quick response of both statistics evaluated

on the measurements $T_{L,cold}^{feed}(t)$ and $T_{L,cold}^{monitor}(t)$: both statistics reach quickly the threshold $h_{T_{L,cold}}$ (dotted line) and the difference $\Delta\tau_{T_{L,cold}}$ between times to alarm ($\tau_{T_{L,cold}}^{feed}$ and $\tau_{T_{L,cold}}^{monitor}$) turns out to be equal to 0 (i.e., less than Γ_y^{ref}) (see Fig. 13(b)), allowing for a (correct) identification of the event as a sensor failure.

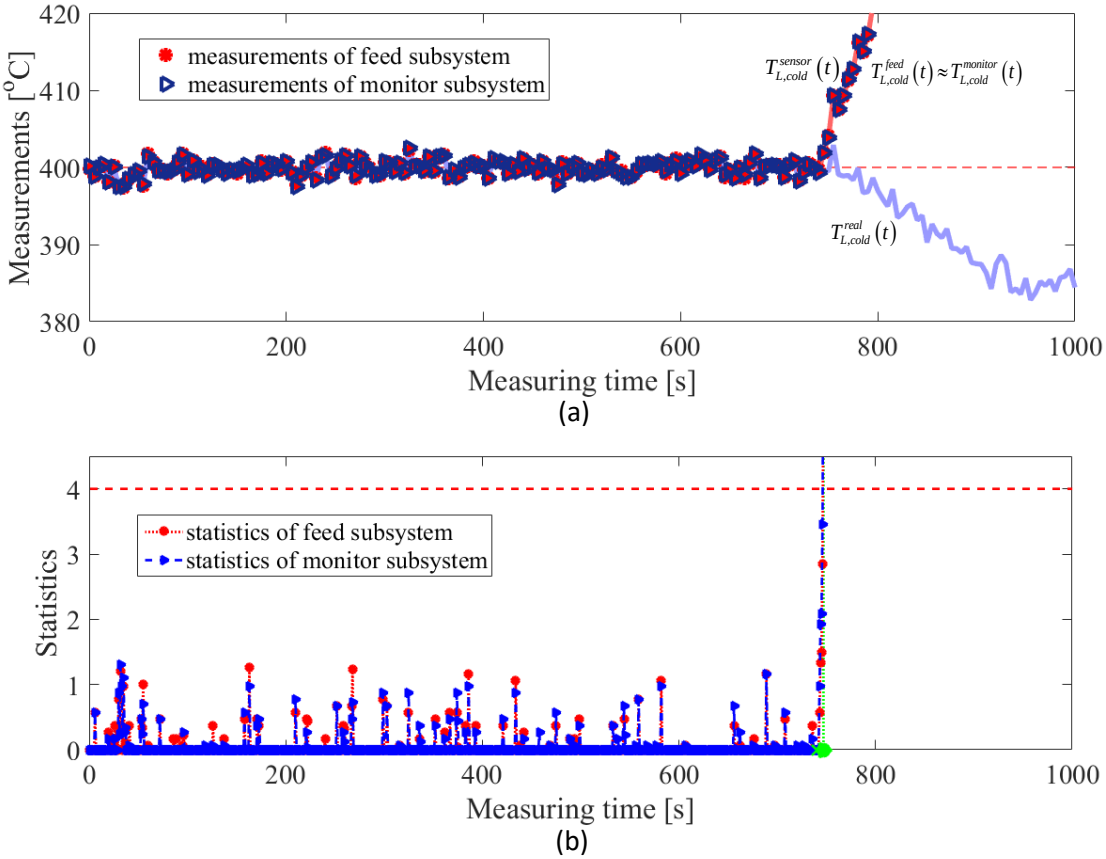

Fig. 13 $T_{L,cold}$ sensor drift failure mode: (a) the received measurements of feed and monitor Subsystems in which the drift occurs at time t_R equal to 740s; (b) the corresponding NP-CUSUM statistics for diagnosing the bias failure

Contrarily, Fig. 14(a) shows a cyber attack to the computing unit mimicking a drift failure mode at t_R=740s (with c again equal to 0.398), leading $T_{L,cold}^{feed}(t)$ to deviate from the legitimate $T_{L,cold}^{monitor}(t)$. The different values between the malicious and the legitimate measurements, then, lead to a delay response $\Delta\tau_{T_{L,cold}}$ equal to 41s (larger than Γ_y^{ref}) between the threshold exceedance of $S_{T_{L,cold}}^{monitor}(t)$ and $S_{T_{L,cold}}^{feed}(t)$ (see Fig. 14(b)), allowing for a (correct) identification of the event as a cyber attack.

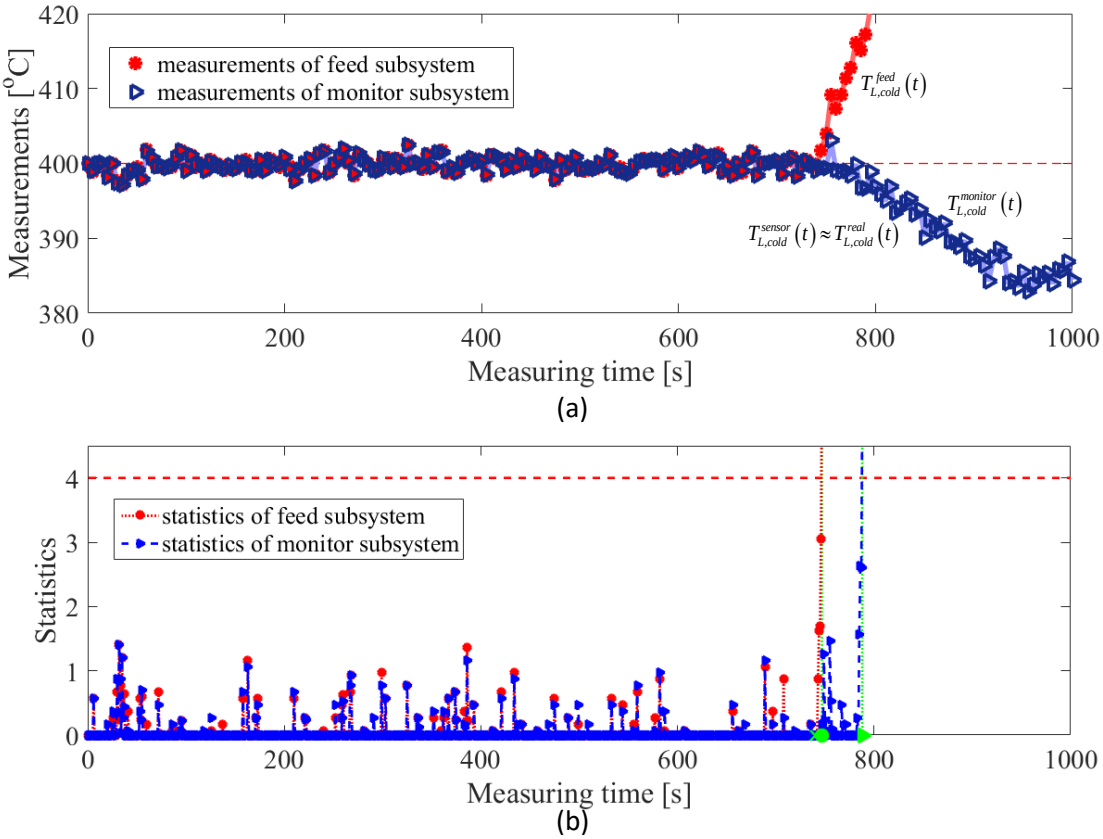

Fig. 14 Cyber attack to the computing unit mimicking a drift failure mode: (a) the received measurements of feed and monitor Subsystems in which the cyber attack occurs at time t_R equal to 740s; (b) the corresponding NP-CUSUM statistics for diagnosing the cyber attack

5.3 Wider Noise Failure Mode

Fig. 15 presents the results of injecting wider noise at time t_R = 750s. This results in a very quick response of both statistics evaluated on the measurements $T_{L,cold}^{feed}(t)$ and $T_{L,cold}^{monitor}(t)$: both statistics reach quickly the threshold $h_{T_{L,cold}}$ (dotted line) and the difference $\Delta\tau_{T_{L,cold}}$ between times to alarm ($\tau_{T_{L,cold}}^{feed}$ and $\tau_{T_{L,cold}}^{monitor}$) turns out to be equal to 0 (i.e., less than Γ_y^{ref}) (see Fig. 15(b)), allowing for a (correct) identification of the event as a sensor failure mode.

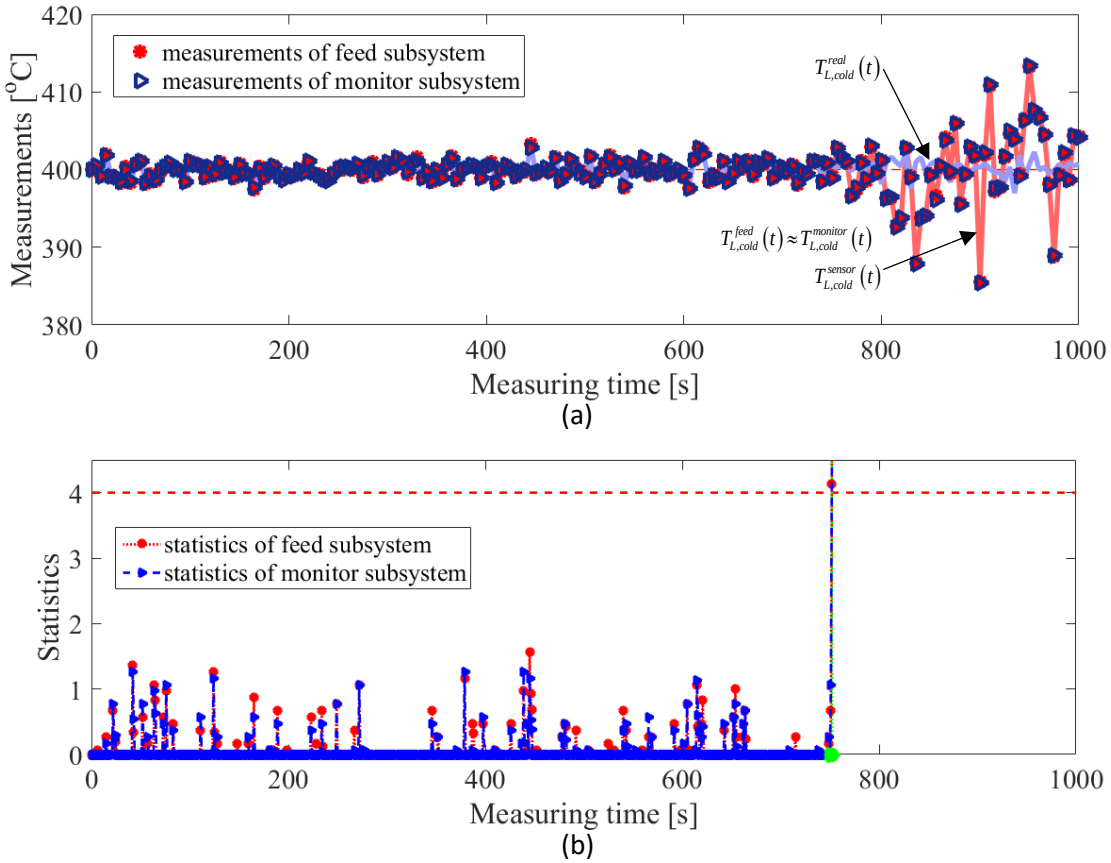

Fig. 15 $T_{L,cold}$ sensor wider noise failure mode: (a) the received measurements of feed and monitor Subsystems in which the wider noise failure occurs at time t_R equal to 750s; (b) the corresponding NP-CUSUM statistics for diagnosing the bias failure

Contrarily, Fig. 16(a) shows a cyber attack to the computing unit mimicking a wider noise failure mode at t_R=750s, leading $T_{L,cold}^{feed}(t)$ to deviate from the legitimate $T_{L,cold}^{monitor}(t)$. The different values between the malicious and the legitimate measurements, then, lead to a delay response $\Delta\tau_{T_{L,cold}}$ equal to 247s (i.e., larger than Γ_y^{ref}) at t_M (see Fig. 16(b)), allowing for a (correct) identification of the event as a cyber attack.

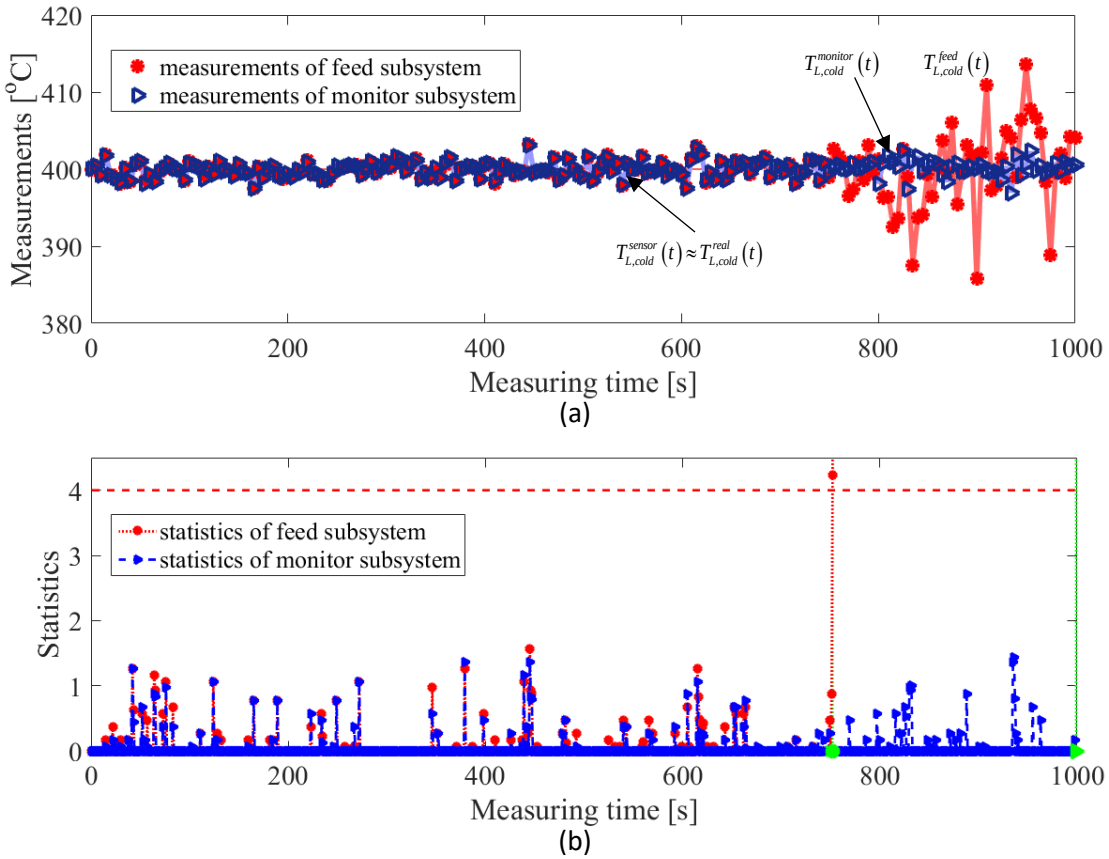

Fig. 16 Cyber attack to the computing unit mimicking a wider noise failure mode: (a) the received measurements of feed and monitor Subsystems in which the cyber attack occurs at time t_R equal to 750s; (b) the corresponding NP-CUSUM statistics for diagnosing the cyber attack

5.4 Freezing Failure Mode

Fig. 17 presents the results of injecting freezing at time t_R = 460s with the frozen $T_{L,cold}^{sensor}(t)$ equal to 402.53°C. The freezing results in a very quick response of both statistics evaluated on the measurements $T_{L,cold}^{feed}(t)$ and $T_{L,cold}^{monitor}(t)$: Both statistics reach quickly the threshold $h_{T_{L,cold}}$ (dotted line) and the difference $\Delta\tau_{T_{L,cold}}$ between times to alarm ($\tau_{T_{L,cold}}^{feed}$ and $\tau_{T_{L,cold}}^{monitor}$) turns out to be equal to 0 (i.e., less than Γ_y^{ref}) (see Fig. 17(b)), allowing for a (correct) identification of the event as a sensor failure mode.

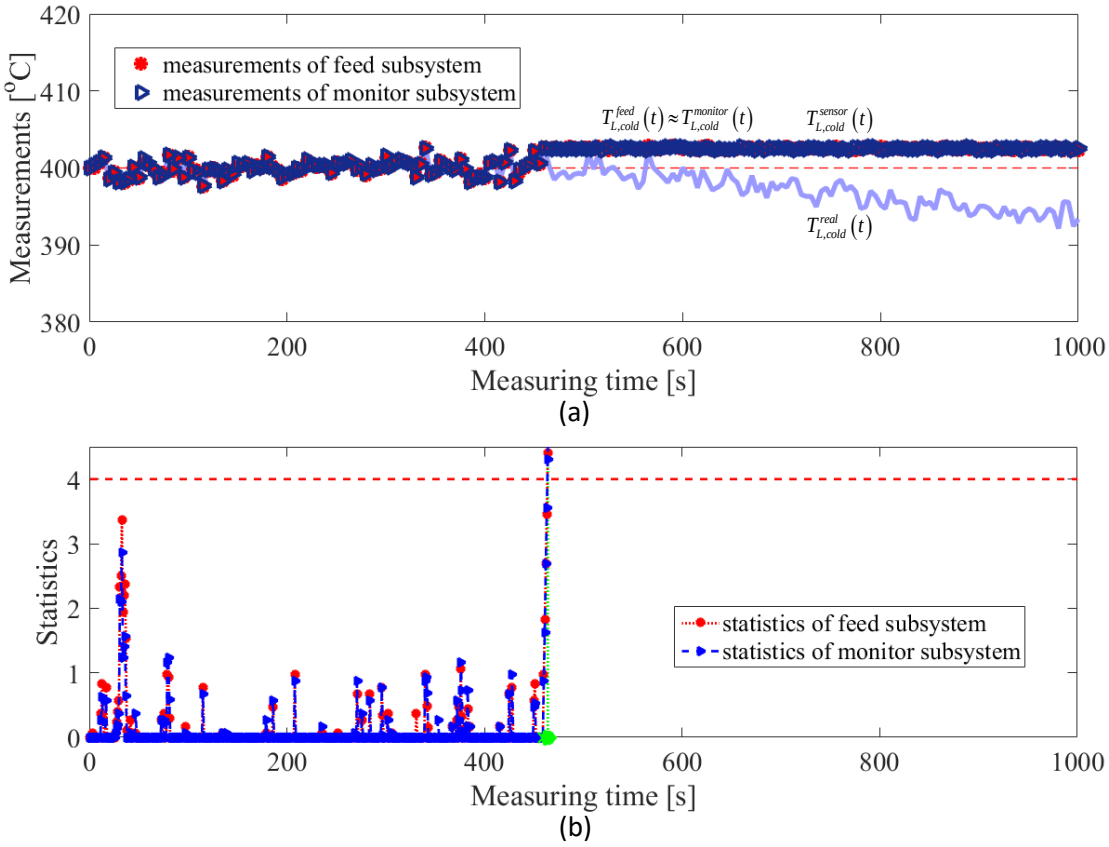

Fig. 17 $T_{L,cold}$ sensor freezing failure mode: (a) the received measurements of feed and monitor in which the freezing occurs at time t_R equal to 460s; (b) the corresponding NP-CUSUM statistics for diagnosing the bias failure

Contrarily, Fig. 18(a) shows a cyber attack to the computing unit mimicking a freezing failure mode at t_R=460s (with frozen $T_{L,cold}^{sensor}(t)$ again equal to 402.53°C), leading $T_{L,cold}^{feed}(t)$ to deviate from the legitimate $T_{L,cold}^{monitor}(t)$. The different values between the malicious and the legitimate measurements, then, lead to a delay response $\Delta\tau_{T_{L,cold}}$ equal to 187s (i.e., larger than Γ_y^{ref}) between the threshold exceedance of $S_{T_{L,cold}}^{monitor}(t)$ and $S_{T_{L,cold}}^{feed}(t)$ (see Fig. 18(b)), allowing for a (correct) identification of the event as a cyber attack.

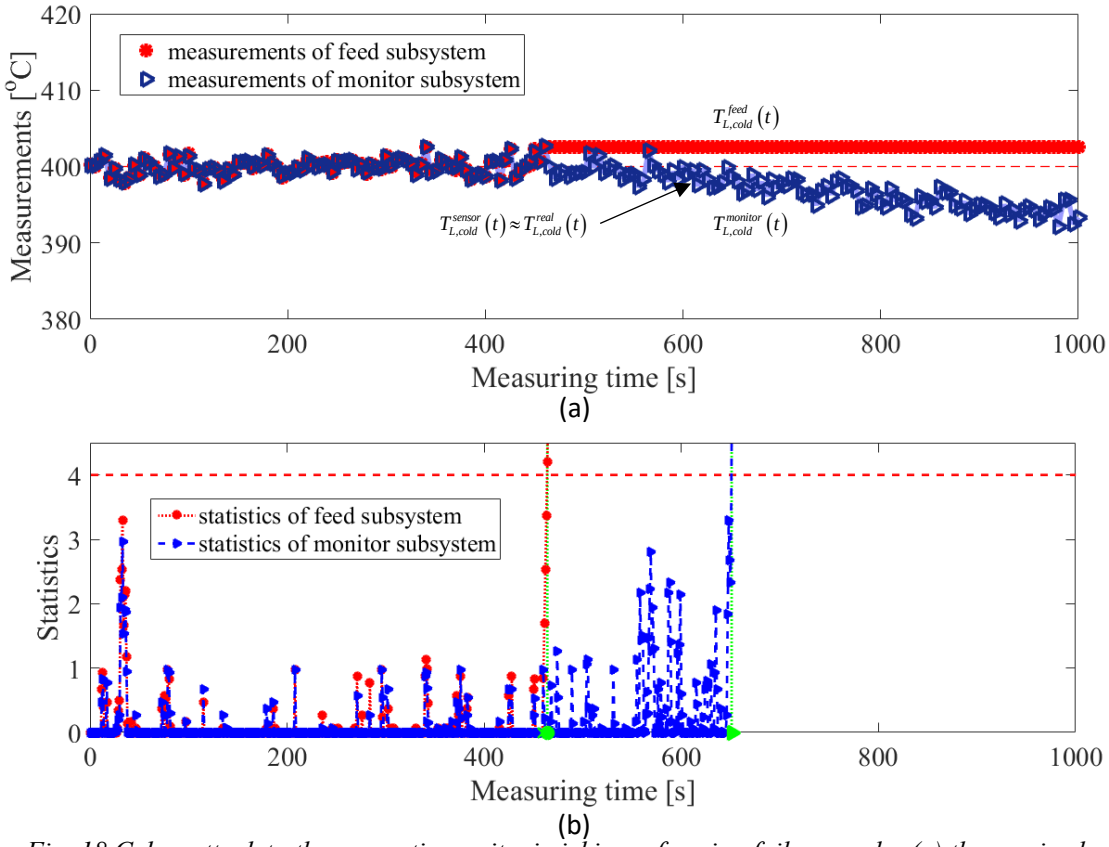

Fig. 18 Cyber attack to the computing unit mimicking a freezing failure mode: (a) the received measurements of feed and monitor Subsystems in which the cyber attack occurs at time t_R equal to 460s; (b) the corresponding NP-CUSUM statistics for diagnosing the cyber attack

The results of these illustrative examples show that the NP-CUSUM-based diagnostics approach is capable of diagnosing cyber attacks, distinguishing them from stochastic failures of components, based on the identified rules of assignments.

6. PERFORMANCE OF THE DIAGNOSTIC APPROACH

The previous examples shown in Section 5 demonstrate the effectiveness of the NP-CUSUM diagnostics approach. Since the proposed diagnostic approach may suffer from either large false alarm rate (if the threshold is set too small) or high missed alarm rate (if the threshold is set too large) (Di Maio et al., 2013), an extensive and massive test with respect to unknown sensor failures and/or unknown cyber attacks is performed for assessing its diagnostic capabilities. We calculate false alarm, missed alarm and misclassification rates with respect to 100 randomly sampled stochastic failures and 100 different cyber attacks for each failure mode (i.e., bias, drift, wider noise or freezing) (thus, a total of N_A=800 runs). At each run of the simulation: a random time t_R within

the mission time t_M=1000s and an uncertain parameter value (i.e., b for bias, c for drift, $\delta'(t)$ are sampled from the distributions listed in Table 4 for wider noise or frozen value for freezing) and used to inject a $T_{L,cold}$ sensor failure or a cyber attack to the computing unit. Then, the NP-CUSUM-based diagnostic algorithm is applied to both $T_{L,cold}^{feed}(t)$ and $T_{L,cold}^{monitor}(t)$, to calculate $S_{L,cold}^{feed}(t)$ and $S_{L,cold}^{monitor}(t)$, respectively. The diagnostic performances are measured as follows:

- False alarm rate $\alpha_{T_{L,cold}}^{h}$: the probability of either $S_{L,cold}^{feed}(t)$ or $S_{L,cold}^{monitor}(t)$ in an accidental scenario exceeding the threshold $h_{T_{L,cold}}$ before t_R.
- Missed alarm rate $\beta_{T_{L,cold}}^{h}$: the probability of neither $S_{L,cold}^{feed}(t)$ or $S_{L,cold}^{monitor}(t)$ in an accidental scenario exceeding the threshold $h_{T_{L,cold}}$ within the mission time t_M.
- Misclassification rate $\gamma\left(\Gamma_{T_{L,cold}}^{ref}\right)$: given a reference delay difference $\Gamma_{T_{L,cold}}^{ref}$, the probability of a misclassified assignment of an event.

Table 5 lists the estimates of $\alpha_{T_{L,cold}}^{h}$ and $\beta_{T_{L,cold}}^{h}$ with respect to the threshold $h_{T_{L,cold}}$ equal to 4.0, among the total of N_A=800 runs of stochastic failures and cyber attacks. The results in the Table show that the total values of $\alpha_{T_{L,cold}}^{h}$ and $\beta_{T_{L,cold}}^{h}$ are equal to 0.0313 and 0.0250, respectively, and the low values are accepted in the diagnostics of cyber attacks of the ALFRED.

Table 5 False and missed alarm rates with respect to $h_{T_{L,cold}}$

Character	Bias	Drift	Wider noise	Freezing	Total
$\alpha_{T_{L,cold}}^{h}$	8/200	8/200	1/200	8/200	25/800=0.0313
$\beta_{T_{L,cold}}^{h}$	13/200	5/200	0/200	2/200	20/800=0.0250

To analyze the effect of an improper choice of $\Gamma_{T_{L,cold}}^{ref}$ that may mistakenly ascribe an accidental scenario to inconsistent reasons and lead to misclassified diagnostics, we estimate $\gamma\left(\Gamma_{T_{L,cold}}^{ref}\right)$ among the N_A=800 scenarios, with respect to different values of

$\Gamma^{ref}_{T_{L,cold}}$. We assess the misclassification rates by defining four misclassification types (i.e., Misclassification I, II, III and IV), that differ in terms of the difference between alarm delays $\Delta\tau_{T_{L,cold}}$ to a reference value $\Gamma^{ref}_{T_{L,cold}}$ in Table 6.

Table 6 Misclassification assessment with respect to $\Gamma^{ref}_{T_{L,cold}}$

Real scenario	Comparison	Assignment	Check	False alarm of	Missed alarm of
Sensor failure	$\Delta\tau_{T_{L,cold}} \le \Gamma^{ref}_{T_{L,cold}}$	Sensor failure	Correct	-	-
	$\Delta\tau_{T_{L,cold}} > \Gamma^{ref}_{T_{L,cold}}$	Cyber attack	***Misclassification I***	Cyber attack	Sensor failure
	Neither $\tau^{feed}_{T_{L,cold}}$ nor $\tau^{monitor}_{T_{L,cold}}$	Normal condition	***Misclassification II***	-	Sensor failure
Cyber attack	$\Delta\tau_{T_{L,cold}} \le \Gamma^{ref}_{T_{L,cold}}$	Sensor failure	***Misclassification III***	Sensor failure	Cyber attack
	$\Delta\tau_{T_{L,cold}} > \Gamma^{ref}_{T_{L,cold}}$	Cyber attack	Correct	-	-
	Neither $\tau^{feed}_{T_{L,cold}}$ nor $\tau^{monitor}_{T_{L,cold}}$	Normal condition	***Misclassification IV***	-	Cyber attack

Fig. 19 shows the calculated misclassification rates $\gamma\left(\Gamma^{ref}_{T_{L,cold}}\right)$ varying with $\Gamma^{ref}_{T_{L,cold}}$ from 0 to 60. $\gamma\left(\Gamma^{ref}_{T_{L,cold}}\right)$ is calculated by summing all the misclassified assignments of the accidental scenarios, which are recorded in the way of false and missed alarm of sensor failures and of cyber attacks, respectively. Results show that the minimum misclassification rate (equal to 0.02875) can be achieved if the categorical difference $\Gamma^{ref}_{T_{L,cold}}$ is optimally equal to 8s or 9s. It is also noted that, the minimum rate being larger than $\beta^{h}_{T_{L,cold}}$ (equal to 0.025) turns out to be reasonable because the identified misclassification scenarios here include the missed alarms identified with respect to $h_{T_{L,cold}}$ equal to 4.0.

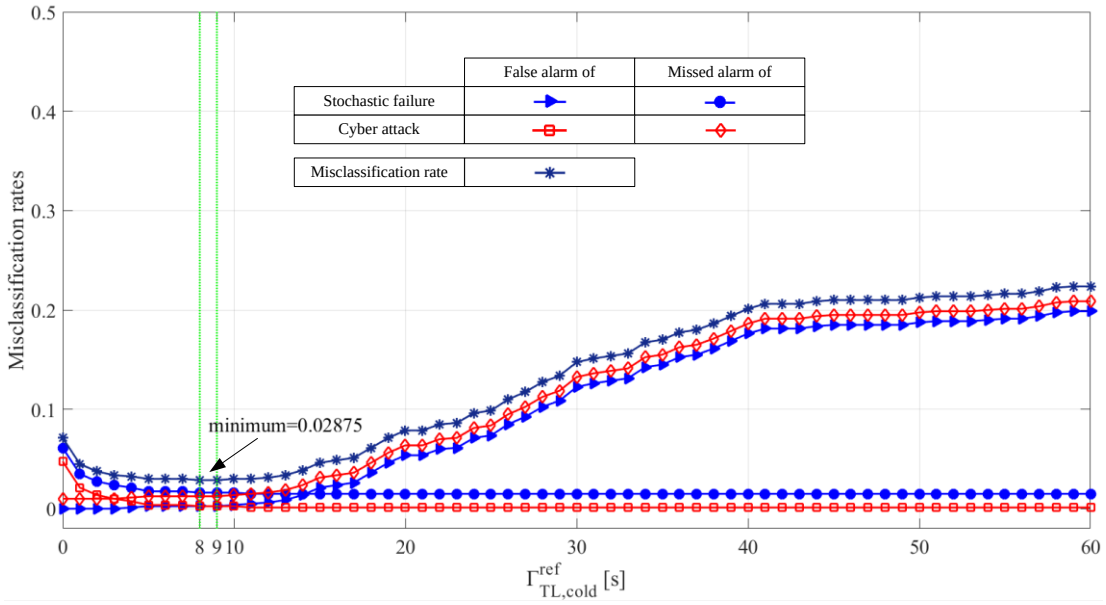

Fig. 19 The misclassification rates varying with $\Gamma^{ref}_{T_{L,cold}}$

7. CONCLUSIONS

In this chapter, we have presented a nonparametric cumulative sum (NP-CUSUM) approach for real-time diagnosing cyber attacks to Cyber-Physical Systems (CPSs). The diagnostics approach allows distinguishing between components failures and cyber attacks to the controllers, guiding decisions for recovering CPSs from anomalies.

The diagnostic performance of the approach has been analyzed by the false and missed alarm rates, with reference to a prespecified threshold, and the misclassification rates varying with the reference delay differences for identifying a cyber attack or a sensor failure.

We have applied the diagnostics approach to the digital Instrumentation and Control (I&C) system of the Advanced Lead-cooled Fast Reactor European Demonstrator (ALFRED). Cyber breach events attacking the embedded CPS controllers and sensor failures are injected by a Monte Carlo sampling procedure, at random times and with random magnitudes. Results show that the diagnostic approach is capable of identifying most of the generated failure/attack scenarios, with low false alarm rate, missed alarm rate and misclassification rate.

Future work will regard, on one hand, the optimization of the threshold setting and

the decision of the reference delay difference, for further minimizing the false and missed alarms, and, on the other hand, the development of an extended multi-variable/channel NP-CUSUM diagnostics framework (e.g., for all the four control loops in the digital I&C system of ALFRED), for localizing and recognizing the failures and/or cyber attacks.

ACKNOWLEDGEMENT

The authors are thankful to Prof. Antonio Cammi and Dr. Stefano Lorenzi of the Energy Department, Politecnico di Milano, for providing guidance and training on code simulating the ALFRED reactor.

APPENDIX A: THE NP-CUSUM ALGORITHM

Without loss of generality, let us consider an accidental scenario *a* simulated over a mission time t_M, during which a cyber attack occurs at random time t_R ($t_R<t_M$). Considering a time interval dt, we can define the pre-attack signal mean value $\mu_Y\left(Y(t)\right)=\sum_t Y(t)\Big/t$, $t = dt, 2dt, \ldots, t$, ($t<t_R$), where $Y(t)$ is the measurement Y of a controlled variable y at time t under normal operation conditions (see Fig. A.1(a), for example). Assume that DoS attacks lead to arbitrary and abrupt changes in the distributions of observations, such that the (unknown) post-attack mean value results to be $\theta_Y\left(Y(t)\right)=\sum_t Y(t)\Big/\left(t-t_R\right)$, $t = t_R, t_R+dt, t_R+2dt, \ldots$.

We define a score function $g_Y(Y(t))$ as:

$$g_Y\left(Y(t)\right)=\sum_t \omega_y\cdot\Lambda\left(Y(t)\right)=\sum_t \omega_y\cdot\left(\left|Y(t)-\mu_Y\right|-c_y(t)\right) \qquad \text{(A.1)}$$

where ω_y is a positive weight that is used for normalizing $\Lambda\left(Y(t)\right)$ and chosen equal to $1/\sigma_Y$, where σ_Y is the standard deviation of $Y(t)$, $t = dt, 2dt, \ldots$, and the parameter $c_y(t)$ depends on the past t-1 measurements as in Eq. (A.2):

$$c_y(t)=\varepsilon_y\cdot\hat{\theta}_Y(t) \qquad \text{(A.2)}$$

where ε_y is a tuning parameter belonging to the interval (0,1) and $\hat{\theta}_Y(t)$ is an estimate

of the unknown mean value $\theta_Y(Y(t))$. In practice, it is difficult to estimate $\hat{\theta}_Y(t)$ on-line. Hence, Eq. (A.1) is simplified in:

$$\Delta g_Y(Y(t)) = \omega_y \cdot (|Y(t) - \mu_Y| - c_y) \quad \text{(A.3)}$$

The score function $S_Y(t)$ adopted in the NP-CUSUM algorithm is, then, defined as:

$$S_Y(t) = \max\{0, S_Y(t-1) + \Delta g_Y(Y(t))\} \quad \text{(A.4)}$$

where, $S_Y(0) = 0$.

In practice, with respect to a stream of measurement $Y(t)$, the NP-CUSUM statistics $S_Y(t)$ remain close to zero or slightly positive under normal operation conditions, whereas, it starts drifting and increasing when a cyber attack occurs at time t_R and, ends up with exceeding a predefined positive threshold h_y (see Fig. A.1(b)). An alarm can be triggered when $S_Y(t)$ reaches h_y at the time of alarm:

$$\tau_Y = \min\{t \geq 1 : S_Y(t) \geq h_y\} \quad \text{(A.5)}$$

The detection delay $d\tau_Y$ between t_R and τ_Y depends on the choice of h_y. A good diagnostic algorithm is expected to perform with a low False Alarm Rate (FAR) and a small value $d\tau_Y$.

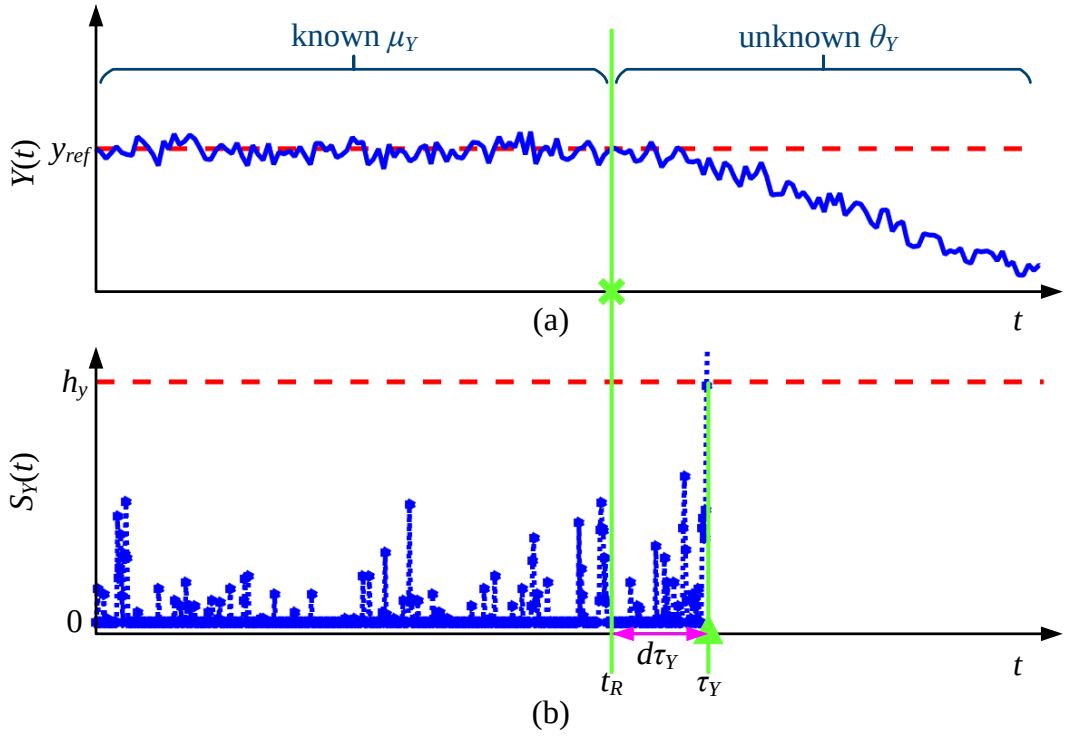

Fig. A.1 The NP-CUSUM algorithm: (a) a stream of measurement Y(t) of an accidental scenario

in which a cyber attack occurring at time t_R; (b) the corresponding NP-CUSUM statistic $S_Y(t)$ for diagnosing the cyber attack at the time to alarm τ_Y

REFERENCES

Aldemir, T., Guarro, S., Mandelli, D., Kirschenbaum, J., Mangan, L.A., Bucci, P., Yau, M., Ekici, E., Miller, D.W., Sun, X. and Arndt, S.A., 2010. Probabilistic risk assessment modeling of digital instrumentation and control systems using two dynamic methodologies. Reliability Engineering & System Safety, 95(10), pp.1011-1039.

Alemberti, A., Frogheri, M., Mansani, L., 2013. The Lead fast reactor demonstrator (ALFRED) and ELFR design. In: Proceedings of the International Conference on Fast Reactors and Related Fuel Cycles: Safe Technologies and Sustainable Scenarios (FR 13), Paris, France, March 4-7, 2013.

Alur, R., 2015. Principles of cyber-physical systems. MIT Press.

Authen, S., and Holmberg, J., 2012. Reliability analysis of digital systems in a probabilistic risk analysis for nuclear power plants. Nuclear Engineering & Technology, 44(5), 271-284.

Aven, T., 2009. Identification of safety and security critical systems and activities. Reliability Engineering & System Safety, 94(2), pp.404-411.

Boskvic, J.D. and Mehra, R.K., 2002, May. Stable adaptive multiple model-based control design for accommodation of sensor failures. In American Control Conference, 2002. Proceedings of the 2002 (Vol. 3, pp. 2046-2051). IEEE.

Bradley, J.M. and Atkins, E.M., 2015. Optimization and control of cyber-physical vehicle systems. Sensors, 15(9), pp.23020-23049.

Carl, G., Kesidis, G., Brooks, R. R., and Rai, S., 2006. Denial-of-service attack-detection techniques. IEEE Internet Computing, 10(1), 82-89.

Debar, H., Dacier, M., and Wespi, A., 1999. Towards a taxonomy of intrusion-detection systems. Computer Networks, 31(8), 805-822.

Di Maio, F., Baraldi, P., Zio, E. and Seraoui, R., 2013. Fault detection in nuclear power plants components by a combination of statistical methods. IEEE Transactions on Reliability, 62(4), pp.833-845.

Duda, R.O., Hart, P.E. and Stork, D.G., 1973. Pattern classification (pp. 526-528). Wiley, New York.

Eames, D.P. and Moffett, J., 1999, September. The integration of safety and security requirements. In International Conference on Computer Safety, Reliability, and Security (pp. 468-480). Springer Berlin Heidelberg.

Fang, Y. and Sansavini, G., 2017. Optimizing power system investments and resilience against attacks. Reliability Engineering & System Safety, 159, pp.161-173.

Fritzson, P., 2010. Principles of object-oriented modeling and simulation with Modelica 2.1. John Wiley & Sons.

Grasso, G., Petrovich, C., Mikityuk, K., Mattioli, D., Manni, F. and Gugiu, D., 2013, March. Demonstrating the effectiveness of the European LFR concept: the ALFRED core design. In Proc. of the IAEA International Conference on Fast Reactors and Related Fuel Cycles: Safe Technologies and Sustainable Scenarios.

Gray, R. M. and Neuhoff, D. L., 1998. Quantization. IEEE Transactions on Information Theory, 44(6), 2325-2383.

Hines, J. W., and Garvey, D., 2006. Development and application of fault detectability performance metrics for instrument calibration verification and anomaly detection. Journal of Pattern Recognition Research, 1(1), 2-15.

Hu, X., Xu, M., Xu, S. and Zhao, P., 2017. Multiple cyber attacks against a target with observation errors and dependent outcomes: Characterization and optimization. Reliability Engineering & System Safety, 159, pp.119-133.

IAEA, 2009. Implementing Digital Instrumentation and Control Systems in the modernization of Nuclear Power Plants. Technical Report NP-T-1.4. IAEA.

Jockenhövel-Barttfeld, M., Taurines, A. and Hessler, C., October, 2016. Quantification of Application Software Failures of Digital I&C in Probabilistic Safety Analyses. In 13th International Conference on Probabilistic Safety Assessment and Management, Seoul, Korea.

Khaitan, S.K. and McCalley, J.D., 2015. Design techniques and applications of cyberphysical systems: A survey. IEEE Systems Journal, 9(2), pp.350-365.

Kim, K.D. and Kumar, P.R., 2012. Cyber–physical systems: A perspective at the centennial. Proceedings of the IEEE, 100(Special Centennial Issue), pp.1287-1308.

Kornecki, A.J. and Liu, M., 2013. Fault tree analysis for safety/security verification in aviation software. Electronics, 2(1), pp.41-56.

Kriaa, S., Pietre-Cambacedes, L., Bouissou, M. and Halgand, Y., 2015. A survey of approaches combining safety and security for industrial control systems. Reliability Engineering & System Safety, 139, pp.156-178.

Lee, E.A., 2008, May. Cyber physical systems: Design challenges. In Object Oriented Real-Time Distributed Computing (ISORC), 2008 11th IEEE International Symposium on (pp. 363-369). IEEE.

Levine, W.S., 1996. The Control Handbook. IEEE Press.

Li, J. and Huang, X., 2016, June. Cyber Attack Detection of I&C Systems in NPPS Based on Physical Process Data. In 2016 24th International Conference on Nuclear Engineering (pp. V002T07A011-V002T07A011). American Society of Mechanical Engineers.

Liang, G., Zhao, J., Luo, F., Weller, S. and Dong, Z.Y., 2017. A review of false data injection attacks against modern power systems. IEEE Transactions on Smart Grid.

Machado, R.C., Boccardo, D.R., De Sá, V.G.P. and Szwarcfiter, J.L., 2016. Software control and intellectual property protection in cyber-physical systems. EURASIP Journal on Information Security, 2016(1), pp.1-14.

McNelles, P., Zeng, Z.C., Renganathan, G., Lamarre, G., Akl, Y. and Lu, L., 2016. A comparison of Fault Trees and the Dynamic Flowgraph Methodology for the analysis of FPGA-based safety systems Part 1: Reactor trip logic loop reliability analysis. Reliability Engineering & System Safety, 153, pp.135-150.

Mohammadpourfard, M., Sami, A. and Seifi, A.R., 2017. A statistical unsupervised method against false data injection attacks: A visualization-based approach. Expert Systems with Applications, 84, pp.242-261.

Moteff, J. D., 2012. Critical Infrastructure Resilience: The Evolution of Policy and Programs and Issues for Congress. Congressional Research Service Reports. Congressional Research Service, Library of Congress.

Ntalampiras, S., 2015. Detection of integrity attacks in cyber-physical critical infrastructures using ensemble modeling. IEEE Transactions on Industrial Informatics, 11(1), pp.104-111.

Ntalampiras, S., 2016. Automatic identification of integrity attacks in cyber-physical systems. Expert Systems with Applications, 58, pp.164-173.

Obama, B., 2013. Presidential policy directive 21: critical infrastructure security and resilience. Washington, DC.

Page, E. S., 1954. Continuous inspection schemes. Biometrika, vol. 41, no.1, pp. 100–115.

Piètre-Cambacédès, L. and Bouissou, M., 2013. Cross-fertilization between safety and security engineering. Reliability Engineering & System Safety, 110, pp.110-126.

Ponciroli, R., Bigoni, A., Cammi, A., Lorenzi, S. and Luzzi, L., 2014. Object-oriented modelling and simulation for the ALFRED dynamics. Progress in Nuclear Energy, 71, pp.15-29.

Ponciroli, R., Cammi, A., Della Bona, A., Lorenzi, S. and Luzzi, L., 2015. Development of the ALFRED reactor full power mode control system. Progress in Nuclear Energy, 85, pp.428-440.

Qiu, P. and Hawkins, D., 2003. A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions. Journal of the Royal Statistical Society: Series D (The Statistician), 52(2), pp.151-164.

Rahman, M.S., Mahmud, M.A., Oo, A.M. and Pota, H.R., 2016. Multi-Agent Approach for Enhancing Security of Protection Schemes in Cyber-Physical Energy Systems. IEEE Transactions on Industrial Informatics, pp.1-10.

Roberts, S., 1959. Control chart tests based on geometric moving averages. Technometrics, vol. 1, no. 3, pp. 239–250.

Shin, J., Son, H. and Heo, G., 2015. Development of a cyber security risk model using

Bayesian networks. Reliability Engineering & System Safety, 134, pp.208-217.

Skogestad, S. and Postlethwaite, I., 2007. Multivariable feedback control: analysis and design (Vol. 2). New York: Wiley.

Subramanian, N. and Zalewski, J., 2013, April. Assessment of safety and security of system architectures for cyberphysical systems. In Systems Conference (SysCon), 2013 IEEE International (pp. 634-641). IEEE.

Tan, R., Nguyen, H.H., Foo, E.Y., Yau, D.K., Kalbarczyk, Z., Iyer, R.K. and Gooi, H.B., 2017. Modeling and Mitigating Impact of False Data Injection Attacks on Automatic Generation Control. IEEE Transactions on Information Forensics and Security, 12(7), pp.1609-1624.

Tartakovsky, A.G., Rozovskii, B.L., Blažek, R.B. and Kim, H., 2006a. Detection of intrusions in information systems by sequential change-point methods. Statistical methodology, 3(3), pp.252-293.

Tartakovsky, A.G., Rozovskii, B.L., Blazek, R.B. and Kim, H., 2006b. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. IEEE Transactions on Signal Processing, 54(9), pp.3372-3382.

Tartakovsky, A.G., Polunchenko, A.S. and Sokolov, G., 2013. Efficient computer network anomaly detection by changepoint detection methods. IEEE Journal of Selected Topics in Signal Processing, 7(1), pp.4-11.

Trabelsi, Z. and Rahmani, H., 2005. An Anti-Sniffer Based on ARP Cache Poisoning Attack. Information Systems Security, 13(6), pp.23-36.

Wang, W., Di Maio, F. and Zio, E., 2016. Component-and system-level degradation modeling of digital Instrumentation and Control systems based on a Multi-State Physics Modeling Approach. Annals of Nuclear Energy, 95, pp.135-147.

Wang, W., Cammi, A., Di Maio F., Lorenzi, S., Zio, E., 2017a. A Probabilistic Modelling and Simulation Framework for Identifying Components Vulnerable to Cyber Threats in Nuclear Power Plants. Submitted to Reliability Engineering & System Safety.

Wang, W., Di Maio, F., Zio, E., 2017b. Estimation of Failure on-Demand Probability and Malfunction Rate Values in Cyber-Physical Systems of Nuclear Power Plants. In the 2017 International Topical Meeting on Probabilistic Safety Assessment and Analysis (PSA2017), Pittsburgh, USA, 24-28 September, 2017.

Wald, A., 1947. Sequential Analysis. New York: Wiley.

Widrow, B., 1961. Analysis of amplitude-quantized sampled-data systems. Transactions of the American Institute of Electrical Engineers Part II Applications & Industry, 80(6), 450-450.

Xiang, Y., Wang, L. and Liu, N., 2017. Coordinated attacks on electric power systems in a cyber-physical environment. Electric Power Systems Research, 149, pp.156-168.

Xie, M., Goh, T.N. and Ranjan, P., 2002. Some effective control chart procedures for reliability monitoring. Reliability Engineering & System Safety, 77(2), pp.143-150.

Yuan, Y., Zhu, Q., Sun, F., Wang, Q. and Başar, T., 2013, August. Resilient control of cyber-physical systems against denial-of-service attacks. In Resilient Control Systems (ISRCS), 2013 6th International Symposium on (pp. 54-59). IEEE.

Yuan, W., Zhao, L. and Zeng, B., 2014. Optimal power grid protection through a defender–attacker–defender model. Reliability Engineering & System Safety, 121, pp.83-89.

Zalewski, J., Buckley, I.A., Czejdo, B., Drager, S., Kornecki, A.J. and Subramanian, N., 2016. A Framework for Measuring Security as a System Property in Cyberphysical Systems. Information, 7(2), p.33.

Zargar, S.T., Joshi, J. and Tipper, D., 2013. A survey of defense mechanisms against distributed denial of service (DDoS) flooding attacks. IEEE communications surveys & tutorials, 15(4), pp.2046-2069.

Zaytoon, J. and Lafortune, S., 2013. Overview of fault diagnosis methods for discrete event systems. Annual Reviews in Control, 37(2), pp.308-320.

Zhao, X. and Chu, P.S., 2010. Bayesian changepoint analysis for extreme events

(typhoons, heavy rainfall, and heat waves): An RJMCMC approach. Journal of Climate, 23(5), pp.1034-1046.

Zio, E., 2009. Reliability engineering: old problems and new challenges. Reliability Engineering & System Safety, 94(2), pp.125-41.

Zio, E., 2016. Challenges in the vulnerability and risk analysis of critical infrastructures. Reliability Engineering & System Safety, 152, pp.137-150.

Zio, E. and Di Maio, F., 2009. Processing dynamic scenarios from a reliability analysis of a nuclear power plant digital instrumentation and control system. Annals of Nuclear Energy, 36(9), pp.1386-1399.

Zio, E. and Zoia, A., 2009. Parameter identification in degradation modeling by reversible-jump Markov Chain Monte Carlo. IEEE Transactions on Reliability, 58(1), pp.123-131.