A Monte Carlo-based exploration framework for identifying components vulnerable to cyber threats in nuclear power plants

Wei Wang, Antonio Cammi, Francesco Di Maio, Stefano Lorenzi, Enrico Zio

A Monte Carlo-Based Exploration Framework For Identifying Components Vulnerable To Cyber Threats In Nuclear Power Plants

Wei Wang , Antonio Cammi , Francesco Di Maio , Stefano Lorenzi , Enrico Zio

PII: S0951-8320(17)30862-1
DOI: 10.1016/j.ress.2018.03.005
Reference: RESS 6087

To appear in: *Reliability Engineering and System Safety*

Received date: 19 July 2017
Revised date: 20 February 2018
Accepted date: 3 March 2018

Please cite this article as: Wei Wang , Antonio Cammi , Francesco Di Maio , Stefano Lorenzi , Enrico Zio , A Monte Carlo-Based Exploration Framework For Identifying Components Vulnerable To Cyber Threats In Nuclear Power Plants, *Reliability Engineering and System Safety* (2018), doi: 10.1016/j.ress.2018.03.005

Highlights

- An exploration framework for identifying CPS vulnerabilities to cyber threats is proposed.
- Cyber attack scenarios are explored by Monte Carlo sampling.
- A safety margin estimation approach is proposed for cyber threat prioritization.
- The framework is illustrated with respect to the digital I&C system of an ALFRED simulator.

A MONTE CARLO-BASED EXPLORATION FRAMEWORK FOR IDENTIFYING COMPONENTS VULNERABLE TO CYBER THREATS IN NUCLEAR POWER PLANTS

Wei Wang[1], Antonio Cammi[1], Francesco Di Maio[1], Stefano Lorenzi[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, 91190 Gif-sur-Yvette, France

Abstract: With the extensive use of digital Instrumentation and Control (I&C) systems, Nuclear Power Plants (NPPs) are becoming Cyber-Physical Systems (CPSs). Their integrity can, then, be compromised also by security breaches (such as cyber attacks). Multiple failure modes (such as bias, drift and freezing) can occur, both due to random failures or induced by malicious external attacks. In this paper, we illustrate an exploration approach that, based on safety margins estimation, allows identifying the most vulnerable components to malicious external attacks. For demonstration, we apply the approach to the Advanced Lead-cooled Fast Reactor European Demonstrator (ALFRED). Its object-oriented model is embedded within a Monte Carlo (MC)-driven engine that injects different types of cyber attacks at random times and magnitudes. Safety margins are, then, calculated and used for identifying the most vulnerable CPS components. This allows selecting protections to make ALFRED resilient towards maliciously induced failures.

Keywords: Nuclear Power Plant; Cyber-Physical System; Cyber Threats; Safety Margins; Advanced Lead-cooled Fast Reactor European Demonstrator (ALFRED).

ABBREVIATIONS

ALFRED	Advanced Lead-cooled Fast Reactor European Demonstrator
CPS	Cyber-Physical System
CR	Control Rod
DoS	Denial of Service
FA	Fuel Assemblies
I&C	Instrumentation and Control
MC	Monte Carlo
NPP	Nuclear Power Plant
OS	Order Statistics
PI	Proportional-Integral
SG	Steam Generator
SISO	Single Input Single Output

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	Proportional gain
K_i	Integral gain
$K_{p,ref}$	Reference value of proportional gain
$K_{i,ref}$	Reference value of integral gain
t	Time
t_A	Attack time
t_M	Mission time
Δt	Sensor measuring time interval
y	Variable (safety parameter)
$y(t)$	Real value of y
$y_{set,ref}$	Reference value of controllers set point value of y
y_{set}	Controller set point value of y under cyber attack
$e(t)$	Residual between $y(t)$ and y_{set}
L_y	Lower threshold of y
R_y	Reference value of y
U_y	Upper threshold of y
$y_{sensor}(t)$	Sensor real measurement at t
$\delta(t)$	Sensor measuring errors
a	Accidental scenario

N	Number of samples of a
$y_{\max,a}$	Maximum value of y during an accidental scenario a
$y_{\max,a}^{\gamma_1}$	Specific γ_1 percentile of the distribution of the measured maximum values of y
$y_{\max,a}^{1}$	First element of N samples sorted in descending order
$\hat{y}_{\max,a}^{\gamma_1,\beta_1}$	Estimated $y_{\max,a}^{\gamma_1}$ with confidence β_1
$y_{\min,a}$	Minimum value of y during a
$y_{\min,a}^{\gamma_2}$	Specific γ_2 percentile of the distribution of the measured minimum values of y
$\hat{y}_{\min,a}^{\gamma_2,\beta_2}$	Estimated $y_{\min,a}^{\gamma_2}$ with confidence β_2
M_{U,y_a}	Safety margin of y with respect to U_y
$M_{U,y_a}^{\gamma_1,\beta_1}$	Safety margin of y with respect to U_y, given with confidence β_1 on the percentile γ_1
$M_{L,y_a}^{\gamma_2,\beta_2}$	Safety margin of y with respect to L_y, given with confidence β_2 on the percentile γ_2
$M_{T,y_a}^{\gamma_1,\gamma_2,\beta}$	Safety margin of y with respect to both U_y and L_y, given with confidence β for percentiles γ_1 and γ_2

1. INTRODUCTION

Hazards and threats are major concerns for the safety and security of modern industry (Aven, 2016; Aven and Krohn, 2014; Zio, 2016; Kriaa et al., 2015; Pièvre-Cambacédès and Bouissou, 2013). The accidents that may originate can be prevented only if they are known in advance, at least to some extent (Paté-Cornell, 2002; Paté-Cornell, 2012).

Modeling and simulation can be used to explore and understand the behavior of a system, under different, possibly uncertain conditions, including hazardous ones (Turati et al., 2017a; Turati et al., 2017b). Design-Of-Experiment (DOE) approaches have been proposed to study different operating conditions, in order to analyze the corresponding system responses with respect to specified performance criteria: safety, reliability, resilience, business continuity, etc. (Santner et al., 2013; Simpson et al., 2001; Zeng & Zio, 2017). One outcome of the analysis, which is of particular interest, is the identification of the conditions (represented by factors, parameters and variables values) that lead the system to critical conditions of failure (Zio and Di Maio, 2009;

Zio, 2016; Turati et al., 2017a; Ntalampiras, 2016).

In this paper, we consider Cyber-Physical Systems (CPSs). A CPS features a tight combination of (and coordination between) the system computational units and physical elements. The integration of computational resources into physical processes is aimed at adding new capabilities to stand-alone physical systems and realize 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)). The benefit of such self-adaptive capability is the reason why CPSs are increasingly operated in transportation, energy, 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 (IAEA, 2009). Cyber controllers have been shown to benefit from the use of information related to: (1) environmental conditions (which play an important role in affecting the system dynamics, and should be measured and adaptively integrated into the cyber real-time monitoring and control in an intelligent manner (Wang et al., 2017a)); (2) periodically updated values of parameters (for keeping up-to-date the CPS settings (Liu et al., 2014)); (3) new interaction modalities between human and system user interfaces (leading to more flexible system operability from the human perspective (Paelke and Röcker, 2015)); and (4) computer-based networks status (to enhance the network connectivity and remote control, communicate with sensing data, and coordinate over constrained environments (Ali et al., 2015)).

Cyber threats, initiated in the cyber domain and manifested in the physical domain, can be misclassified as component failures, disguising their malicious

character (Zalewski et al., 2016; Rahman et al., 2016; Wang et al., 2018). Even if they are different from components stochastic failures, they can lead to similar consequences on the system physical processes (e.g., both a stochastic failure and a cyber attack can result in sensor performance degradation (Rahman et al., 2016)).

From the perspective of security analysis, the identification of the cyber threats most affecting the system response is quite important for decision-making on optimal protection (Fang and Sansavini, 2017; Hu et al., 2017).

Other works have focused on the formulation and modeling of malicious activities to CPSs (Kriaa et al., 2015; Xiang et al., 2017; Pasqualetti et al., 2013). Besides graphical methods (such as attack graphs (McQueen et al., 2006; Sheyner and Wing, 2003; Ingols et al., 2006), attack trees (Schneier, 1999; Fovino et al., 2009), Petri nets (Mitchell and Chen, 2013)), mathematical models (such as those based on game theory (Backhaus et al., 2013; Xiang et al., 2018) and attacker-defender models (Fang et al., 2017; Yuan et al., 2014)), cyber attacks have also been simulated (Huang et al., 2009; Rahman et al., 2016; Khalid and Peng, 2016).

In particular, Monte Carlo (MC) simulation allows considering the interactions among the physical parameters of the process (e.g., temperature, pressure, flow rate, etc.), human actions, components stochastic failures, and malicious activities (Zio, 2013; Wang et al., 2017b). Attacks aiming at damaging different components of the CPSs can, thus, be explored, generating different scenarios in the physical domain which lead to different consequences (e.g., magnitude of failure). Similarly, models can be introduced for describing attack magnitudes and the attackers' adaptive/responsive behaviors, generating and exploring specific deviations caused by cyber attacks.

Specifically, in this work, we develop a general modelling and simulation framework for generating cyber attack scenarios by MC sampling, testing their effects on CPS integrity and prioritizing the most vulnerable components of the CPS. An approach is originally undertaken for processing cyber attack scenarios, based on the related estimated safety margin, the most vulnerable components are identified.

A number of non-parametric statistical methods have been used in safety analysis for safety margin estimation: the Wilk's method based on Order Statistics (OS) (Wilks, 1941; Wilks, 1942; Wald, 1943; Nutt and Wallis, 2004), Beran and Hall simple linear interpolation (Beran and Hall, 1993), Hutson fractional statistics (Hutson, 1999) and data-based bootstrap method (Efron and Tibshirani, 1986). Among these, OS is popular and consolidated because it provides relatively conservative results with a few computer code runs, for leveraging the usually expensive computational cost of simulation codes (Nutt and Wallis, 2004; Zio et al., 2010; Sanchez-Saez et al., 2017). In this study, we, thus, take a "Bracketing" OS approach for tackling the computational problem and calculating the safety margins (Nutt and Wallis, 2004; Di Maio et al., 2016a; Di Maio et al., 2016b; Di Maio et al., 2017).

Without loss of generality and for demonstration purposes, the proposed approach is illustrated with respect to cyber attack scenarios injected into a specific nuclear CPS, namely, the digital I&C system of the pool-type Advanced Lead Fast Reactor European Demonstrator (ALFRED) (Alemberti et al., 2013), whose previously developed object-oriented DYMOLA simulator (Ponciroli et al., 2014; Ponciroli et al., 2015) with a multi-loop PI control scheme (Skogestad and Postlethwaite, 2007) is utilized. Cyber attacks to the components of the digital I&C system are injected into the ALFRED simulator by MC sampling of four important safety parameters: turbine inlet steam temperature, Steam Generator (SG) pressure, lead temperature at the SG outlet (the "cold leg" temperature) and thermal power. For simplicity, but without loss of generality, no protection is taken into account, i.e., the control system for the normal operation mode remains in operation during (and after) the cyber attack. This is, thus, a "worst-case" condition, since the protections to prevent or mitigate unwanted consequences are not considered.

The paper is organized as follows. Section 2 presents the main characteristics of the ALFRED reactor, with its control scheme at full power nominal conditions, and the MC engine of cyber breaches injection for generating cyber attack scenarios. In

Section 3, safety margins are quantified with respect to sensors, actuators and controllers failure modes. Identification of the components most vulnerable to cyber attacks is illustrated in Section 4. In Section 5, conclusions are drawn.

2. THE ADVANCED LEAD-COOLED FAST REACTOR EUROPEAN DEMONSTRATOR

The ALFRED reactor with its full power mode control scheme and the MC engine of cyber breaches injection are described in Sections 2.2 and 2.3, respectively.

2.1 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). The ALFRED core is composed by wrapped Fuel Assemblies (FAs) for providing the thermal power P_{Th}, and Control Rods (CRs) systems adjust the heights of CRs h_{CR} for power regulation and reactivity swing compensation.

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 ^{o}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 ^{o}C; from there, it is delivered to eight Steam Generators (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 ^{o}C, and leaves the SGs after absorbing heat from the primary coolant, entering the turbine as steam at temperature T_{steam} equal to 450 ^{o}C. 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*). 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.

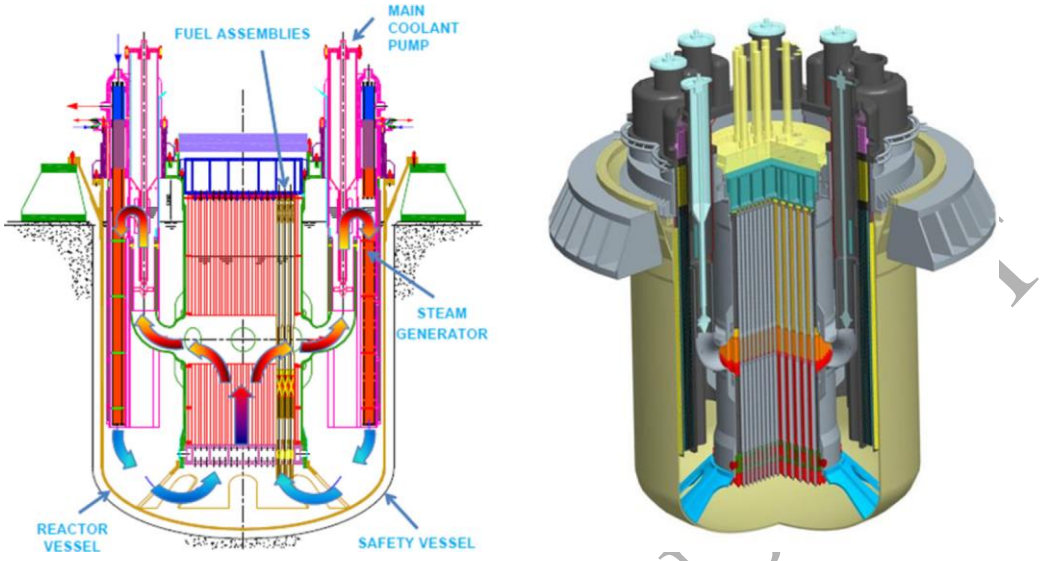

Fig. 1. ALFRED primary system layout (Alemberti et al., 2013)

A simplified schematics of the ALFRED primary and secondary cooling systems is shown in Fig. 2. The input data of the ALFRED model 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	^oC
$T_{L,cold}$	Coolant SG outlet temperature	400	^oC
Γ	Coolant mass flow rate	25984	$kg \cdot s^{-1}$
T_{feed}	Feedwater SG inlet temperature	335	^oC
T_{steam}	Steam SG outlet temperature	450	^oC
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

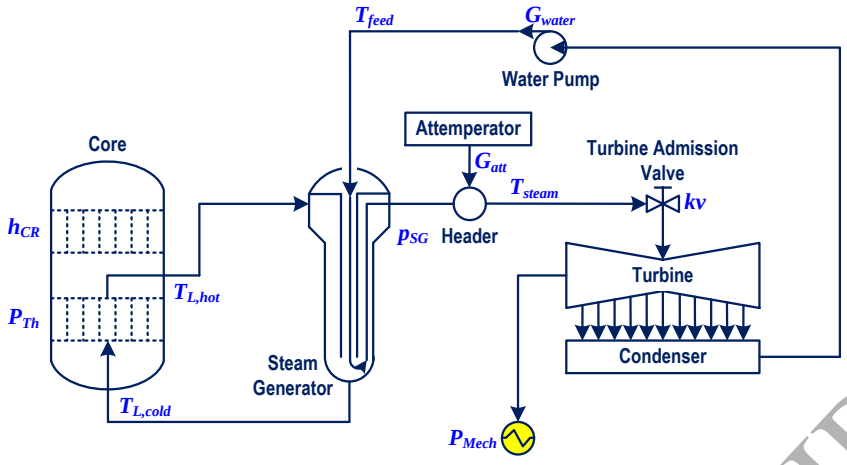

Fig. 2. ALFRED simplified schematics

2.2 Reactor Control Scheme

To design the regulators and simulate the system controlled response, an object-oriented simulator of the entire plant has been developed (Fig. 3), based on the Modelica language (Fritzson, 2010) and implemented in the Dymola environment (DYMOLA, 2015) (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 PI-based feedback control configuration employs four SISO (Single Input Single Output) control loops independent of each other (Ponciroli et al., 2015). The parameters of the PI regulators have been calibrated and are reported in Table 2.

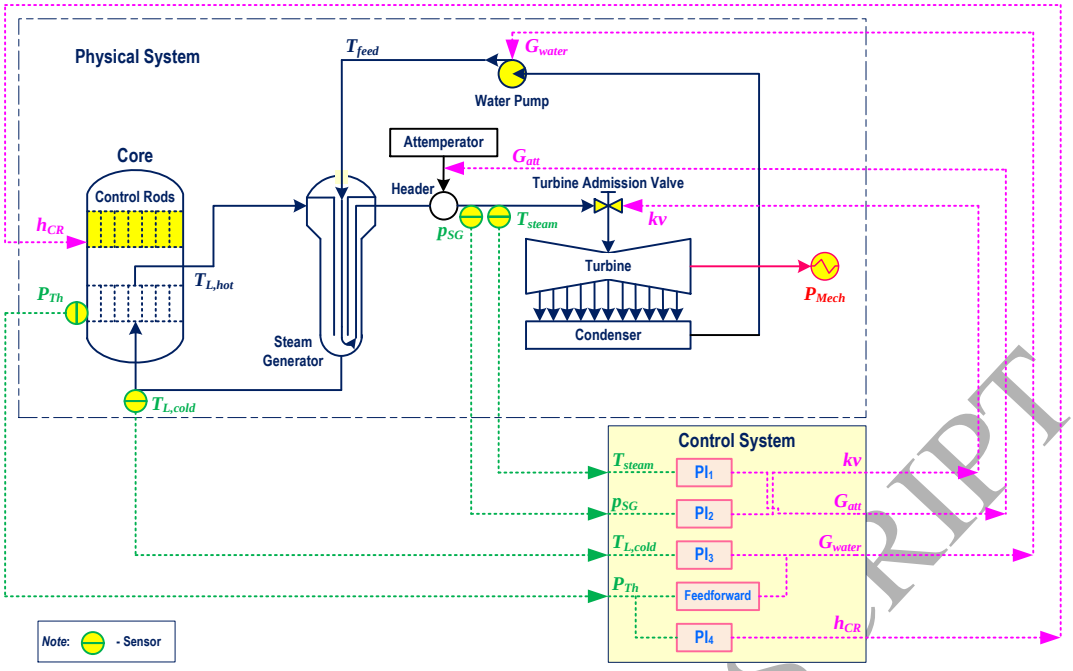

Fig. 3. ALFRED reactor control scheme

Table 2 Parameters of PI controllers

PI	Control Loop		Controller Parameters	
	Controlled variable	Control variable	$K_{p,ref}$	$K_{i,ref}$
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)	$2 \cdot 10^{-11}$	$4 \cdot 10^{-11}$

The control aims at keeping the controlled variables of the control loops approximately at the steady state values, for outputting a steady mechanical power. The values represent the optimal working conditions of the system at full power nominal conditions. The regulation of the controlled variables is of particular concern, to bring benefits to the structural materials and ensure safe NPP operation conditions. Safety thresholds for each variable, listed in Table 3, are set such that consequences of transients and accidents are limited: for example, the $T_{L,cold}$ must be kept above 350°C to avoid the embrittlement of the structural materials in aggressive environments enhanced by the fast neutron irradiation.

In Fig. 4, profiles of the controlled variables, with a mission time t_M equal to 3000s, are shown. Under the control scheme of Fig. 3, the values of the variables are kept approximately at their nominal values, at full power nominal conditions, despite

the measuring errors (white noise).

Table 3 List of reference and threshold values for safety variables

Variable, y	Reference value, R_y, at full power nominal conditions	Safety thresholds	
		Lower, L_y	Upper, U_y
T_{steam} (°C)	450	-	550
p_{SG} (Pa)	$180 \cdot 10^5$	$170 \cdot 10^5$	$190 \cdot 10^5$
$T_{L,cold}$ (°C)	400	350	-
P_{Th} (W)	$300 \cdot 10^6$	$270 \cdot 10^6$	$330 \cdot 10^6$

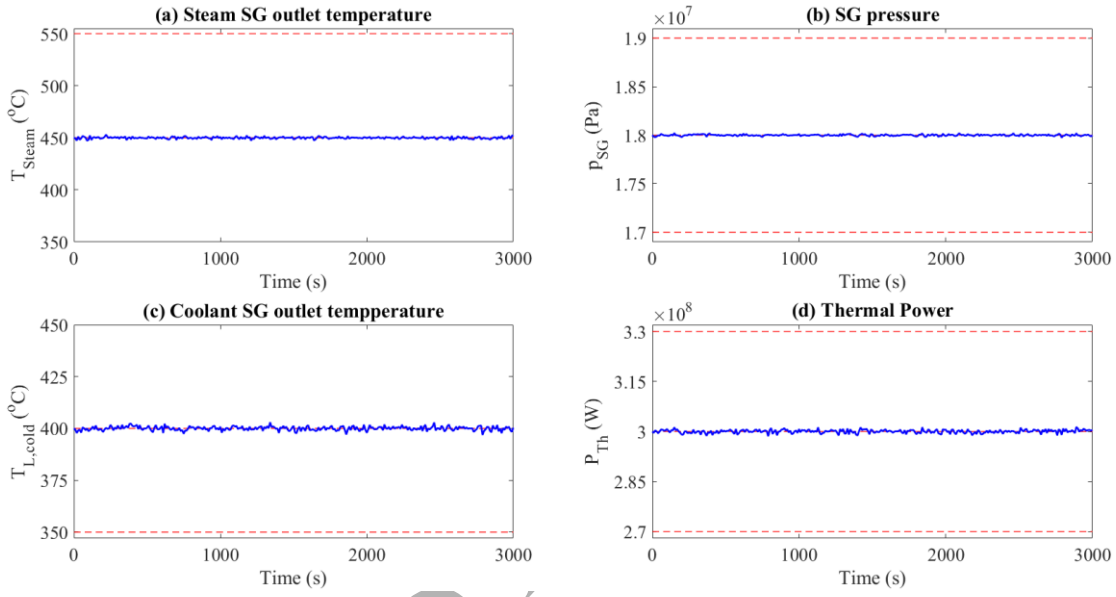

Fig. 4. Profiles of the controlled variables of the ALFRED model at full power nominal conditions: (a) Steam SG outlet temperature; (b) SG pressure; (c) Coolant SG outlet temperature; and (d) Thermal power

2.3 The Monte Carlo Engine of Cyber Breaches Injection

To test the effects of cyber attacks on system integrity, a MC engine is integrated with the ALFRED model for injecting cyber breaches at random times and magnitudes. It shall be noted that, the random time t_A of the attack occurrence only plays an illustrative role in modeling the random occurrence of a cyber attack in reality. The cyber attacks here considered are sketched in Fig. 5 and hereafter described.

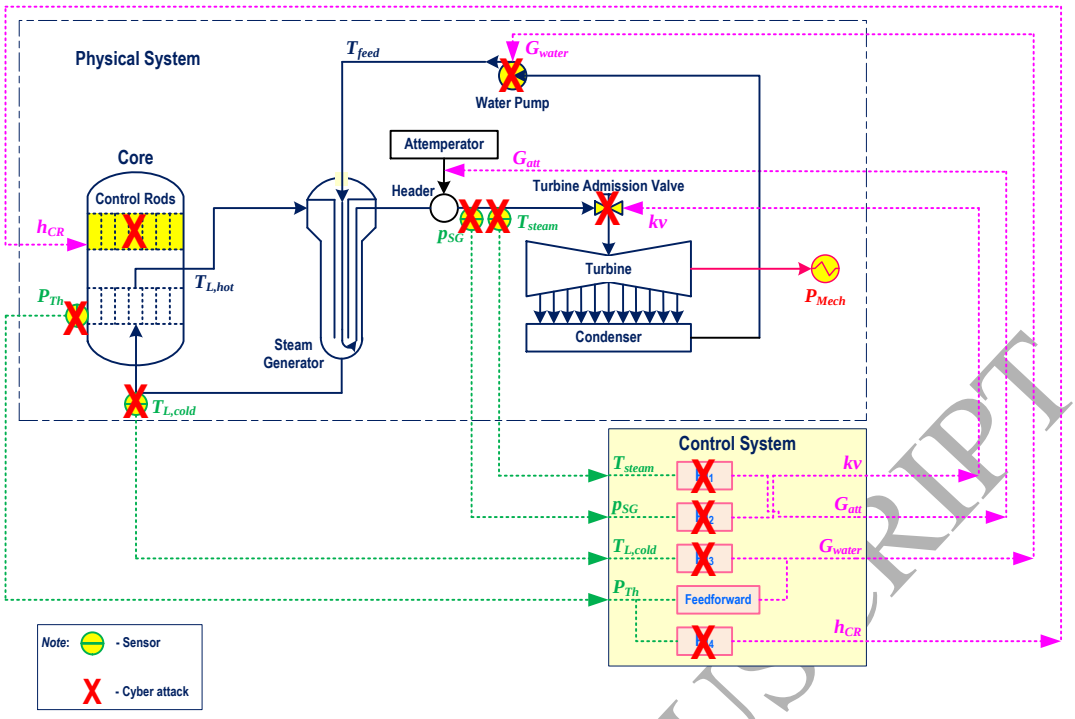

Fig. 5 Sketch of cyber attacks injected into the ALFRED system

(1) Sensors

Controlled variables of the physical system are measured by sensors, whose values are fed to the control system. Four types of cyber attacks occurring at random time t_A are considered for each sensor, preventing the controllers from receiving legitimate measurements (equivalent to typical Denial of Service (DoS) attacks (Zhang et al., 2016; Ding et al., 2016; Yuan et al., 2014; Wang et al., 2018; Zhu et al., 2014)), mimicking stochastic failures (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):

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

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 larger variance than $\delta(t)$ (α>1).

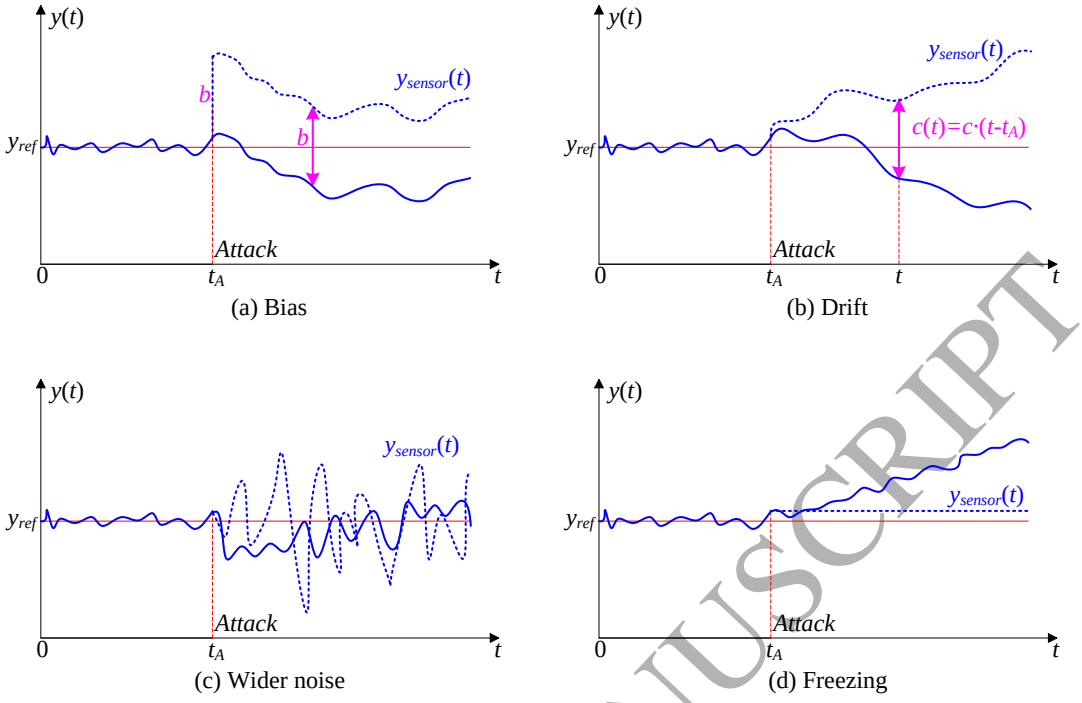

Fig. 6. Sensor failure modes: (a) bias; (b) drift; (c) wider noise; and (d) freezing. Solid lines represent the real measurements of the controlled variables, whereas dotted lines are the altered measurements of the failed sensors

Practically, the MC sampling procedure used to inject a random cyber attack to sensors at time t_A consists in sampling the uncertain parameters b, c, $\delta'(t)$ from the distributions listed in Table 4 and, then, running the ALFRED simulator for collecting the controlled variables evolution throughout the mission time t_M. Notice that Gaussian noises are typical of sensor data acquisition, leading to sensor nominal errors (column 2) and wider errors (column 5) under nominal condition and wider noise failure mode, respectively. Bias and drift (columns 3 and 4, respectively) are, instead, a-priori set from uniform distributions, to mimic sensor stochastic failures due to cyber attacks.

Table 4 Parameters of sensors

Sensor	Nominal error $\delta(t)$	Failure factors		
		Bias b	Drift c	Wider noise $\delta'(t)$
T_{steam} (°C)	$N(0,1)$	$U(-200,200)$	$U(-1,1)$	$N(0,10)$
p_{SG} (Pa)	$N(0,0.1)\cdot10^5$	$U(-100,30)\cdot10^5$	$U(-0.2,0.2)\cdot10^5$	$N(0,2)\cdot10^5$
$T_{L,cold}$ (°C)	$N(0,1)$	$U(-30,30)$	$U(-1,1)$	$N(0,5)$
P_{Th} (W)	$N(0,0.5)\cdot10^6$	$U(-300,30)\cdot10^6$	$U(-0.5,0.5)\cdot10^6$	$N(0,0.7)\cdot10^6$

(2) Actuators

Three actuators of the digital I&C system of ALFRED are considered susceptible of a malicious attack, namely: control rods that regulate the rod heights h_{CR}, water pump that regulates the feedwater mass flow rate G_{water} and turbine admission valve kv that regulates the steam inlet mass flow rate. At nominal conditions, the actuators execute the command signals of the control system to respond to the sensors measurements and accommodate disturbances, transients or accidents. On the other hand, under attack, the actuators might fail stuck to a random magnitude of actuation $A(t)$, here sampled from a uniform distribution (see Table 5): in this situation, the actuators would no longer receive proper control commands and the I&C system would not be capable of accommodating disturbances, transients or accidents, as shown in Fig. 7.

Table 5 Parameters of actuators

Actuator	Regulated control variable	Reference regulation	Failure distribution
Control rods	h_{CR} (cm)	12.3	U(0,64)
Water pump	G_{water} (kg·s^{-1})	192	U(0,300)
Turbine admission valve coefficient	kv (-)	1	U(1,1.5)

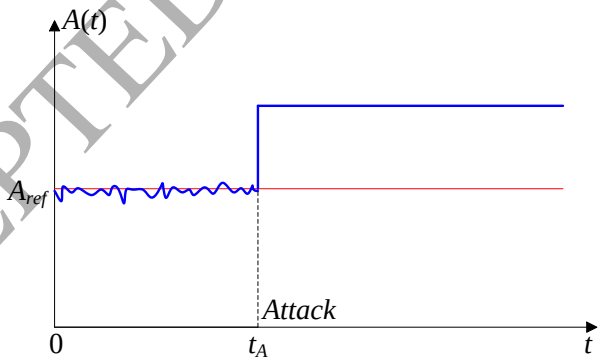

Fig. 7. Typical actuator-stuck failure

(3) PI controllers

At nominal conditions, PI gains (i.e., K_p and K_i) and controlled variables set points $y_{set,ref}$ are fixed by the control designers, to keep the physical process variables close to their nominal values. Under the cyber attack of Fig. 8, equivalent to a deception attack maliciously injecting a false message to the controller (Rahman et

al., 2016; Ding et al., 2016), PI gains and set points are randomly sampled from uniform distributions, covering all possible values (see Table 6). In terms of uniform distributions for sampling random values of PI gains (columns 6 and 7), their expectations are larger than the reference values, for increasing the possibility that the cyber attack impacts the system integrity (Di Maio et al., 2011).

Table 6 Parameters of PIs

PI	Controlled variable, y	Reference value			PI parameter upon attack		
		$K_{p,ref}$	$K_{i,ref}$	Set point, $y_{set,ref}$	K_p	K_i	Set point, y_{set}
PI_1	T_{steam}	$1 \cdot 10^{-1}$	$5 \cdot 10^{-2}$	450 (°C)	$U(1 \cdot 10^{-2}, 1)$	$U(5 \cdot 10^{-4}, 5)$	U(430,470) (°C)
PI_2	p_{SG}	$3 \cdot 10^{-7}$	$1 \cdot 10^{-8}$	$180 \cdot 10^5$ (Pa)	$U(3 \cdot 10^{-8}, 3 \cdot 10^{-4})$	$U(3 \cdot 10^{-10}, 3 \cdot 10^{-5})$	$U(170,190) \cdot 10^5$ (Pa)
PI_3	$T_{L,cold}$	$6 \cdot 10^{-1}$	$1 \cdot 10^{-2}$	400 (°C)	$U(6 \cdot 10^{-2}, 6)$	$U(1 \cdot 10^{-4}, 1)$	U(380,420) (°C)
PI_4	P_{Th}	$2 \cdot 10^{-11}$	$4 \cdot 10^{-11}$	$300 \cdot 10^6$ (W)	$U(2 \cdot 10^{-12}, 2 \cdot 10^{-7})$	$U(4 \cdot 10^{-13}, 4 \cdot 10^{-6})$	$U(285,315) \cdot 10^6$ (W)

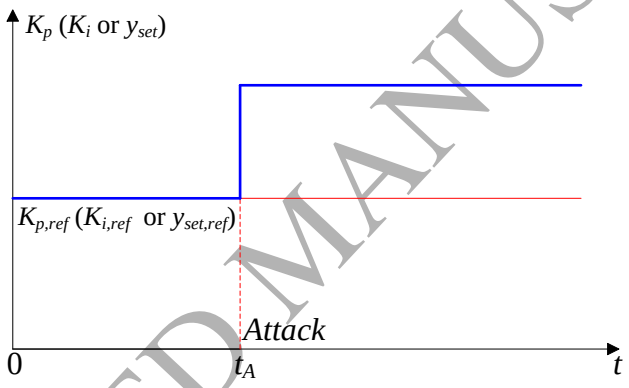

Fig. 8 PI regulator failure due to the cyber attack

It is worth mentioning that the components of the digital I&C system considered, their failure modes and cyber attack types are not intended to provide a comprehensive description of the system accidental behavior, but are only taken as exemplary for generating the dynamic accident scenarios to be processed for safety margins estimation, within the framework here proposed for the identification of the components most vulnerable to cyber threats. Moreover, we observe that an attacker is interested also in injecting "soft" failures that slowly drive the system into failure, rather than, only "hard" failures because the former is more difficult to detect and recover.

3. SAFETY MARGINS ESTIMATION FOR THE IDENTIFICATION OF THE COMPONENTS MOST VULNERABLE TO CYBER THREATS

A safety margin approach of literature (Zio et al., 2010; Di Maio et al., 2016b; Di Maio et al., 2017) is here originally used to estimate the extent of the consequences of cyber threats on the CPS components.

3.1 Estimation of Safety Margins

(1) One-sided safety margin

Considering an accidental scenario a simulated over a mission time t_M, the safety margin M_{U,y_a} of a safety parameter y, with respect to a predefined upper threshold U_y, is defined as the ratio between the computed value reached by the maximum value $y_{\max,a}$ during the accidental scenario and the design value y_{ref} (see Fig. 9) (Nutt and Wallis, 2004; Di Maio et al., 2016b; Di Maio et al., 2017):

$$M_{U,y_a} = \begin{cases} \dfrac{U_y - y_{\max,a}}{U_y - y_{ref}} & y_{ref} < y_{\max,a} < U_y \\ 0 & U_y \le y_{\max,a} \\ 1 & y_{\max,a} \le y_{ref} \end{cases} \quad (2)$$

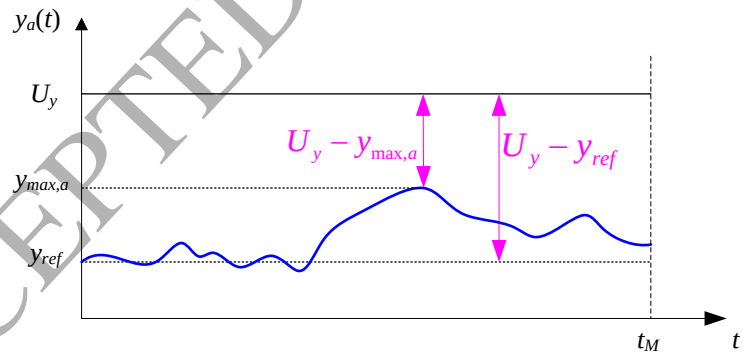

Fig. 9 One-sided safety margin M_{U,y_a}

Being M_{U,y_a} a stochastic variable, the safety margin with respect to U_y (see Fig. 10) is more rigorously defined as the difference between U_y and the value of a specific γ_1 percentile of the distribution of the measured maximum values, $y_{max,a}^{\gamma_1}$, where $\hat{y}_{max,a}^{\gamma_1,\beta_1}$ (i.e., the estimate of $y_{max,a}^{\gamma_1}$) is given with confidence β_1 (Lehmann and

Casella, 2006), viz:

$$\begin{cases} \gamma_1 = \Pr\left(y_{\max,a} < y_{\max,a}^{\gamma_1} \right) \\ \beta_1 = \Pr\left(y_{\max,a}^{\gamma_1} < \hat{y}_{\max,a}^{\gamma_1,\beta_1} \right) \end{cases} \tag{3}$$

and,

$$M_{U,y_a}^{\gamma_1,\beta_1} = \begin{cases} \dfrac{U_j - \hat{y}_{\max,a}^{\gamma_1,\beta_1}}{U_j - y_{j,ref}} & y_{ref} < \hat{y}_{\max,a}^{\gamma_1,\beta_1} < U_y \\ 0 & U_y \le \hat{y}_{\max,a}^{\gamma_1,\beta_1} \\ 1 & \hat{y}_{\max,a}^{\gamma_1,\beta_1} \le y_{ref} \end{cases} \tag{4}$$

The value $\hat{y}_{\max,a}^{\gamma_1,\beta_1}$ is estimated by a Bracketing OS approach, which allows controlling the computational cost of the simulation codes and guarantees that the first element (out of *N*) in the descending sorted sample $y_{\max,a}^{1}$ has a certain probability β_1 of exceeding the unknown true γ_1 percentile. The number *N* can be calculated by Eq. (5), when γ_1 and β_1 are predefined.

$$\beta_1 = 1 - \gamma_1^N \tag{5}$$

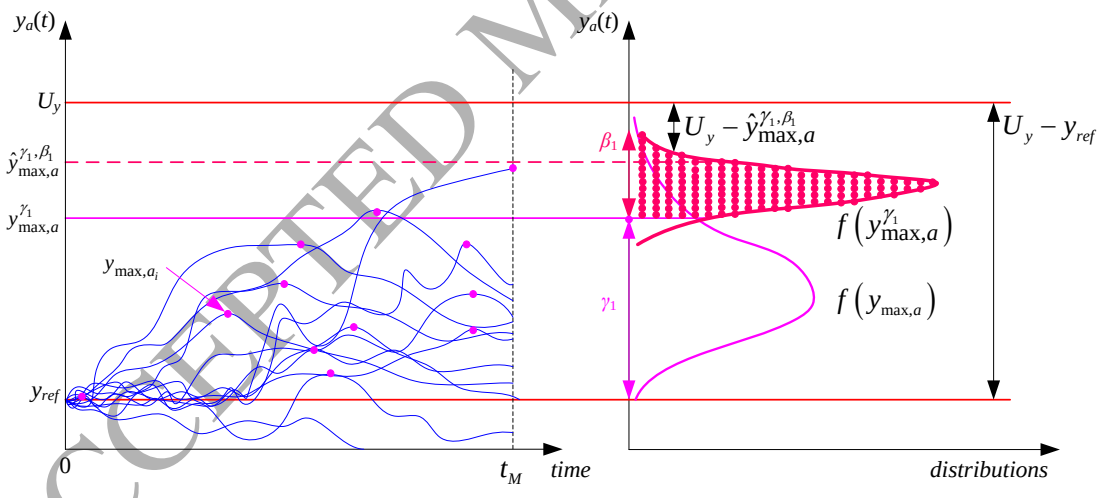

Fig. 10. $y_{\max,a_i}$ *obtained from N samples of the accidental scenario a used to estimate* $\hat{y}_{\max,a}^{\gamma_1,\beta_1}$*, and, thus, to estimate* $M_{U,y_a}^{\gamma_1,\beta_1}$

Similarly, the safety margin with respect to a lower threshold L_y becomes:

$$M_{L,y_a}^{\gamma_2,\beta_2}=\begin{cases}\dfrac{\hat{y}_{\min,a}^{\gamma_2,\beta_2}-L_y}{y_{ref}-L_y} & L_y<\hat{y}_{\min,a}^{\gamma_2,\beta_2}<y_{ref}\\ 0 & \hat{y}_{\min,a}^{\gamma_2,\beta_2}\le L_y\\ 1 & y_{ref}\le\hat{y}_{\min,a}^{\gamma_2,\beta_2}\end{cases} \tag{6}$$

where, $\hat{y}_{\min,a}^{\gamma_2,\beta_2}$ is the point estimate value of the γ_2 percentile of the distribution of the measured values $y_{\min,a}$, with a confidence β_2, and, γ_2 and β_2 are:

$$\begin{cases}\gamma_2=\Pr\left(y_{\min,a}<y_{\min,a}^{\gamma_2}\right)\\ \beta_2=\Pr\left(y_{\min,a}^{\gamma_2}>\hat{y}_{\min,a}^{\gamma_2,\beta_2}\right)\end{cases} \tag{7}$$

The number N can be calculated by Eq. (8), when γ_2 and β_2 are predefined.

$$\beta_2=1-\left(1-\gamma_2\right)^N \tag{8}$$

(2) Two-sided safety margin

The safety margin M_{T,y_a} of a safety parameter y with respect to the double-sided (both upper U_y and lower L_y) thresholds (see Fig. 11) is defined as the minimum value between $M_{U,y_a}^{\gamma_1,\beta}$ and $M_{L,y_a}^{\gamma_2,\beta}$ of Eqs. (4) and (6):

$$M_{T,y_a}^{\gamma_1,\gamma_2,\beta}=\min\left(M_{U,y_a}^{\gamma_1,\beta},M_{L,y_a}^{\gamma_2,\beta}\right) \tag{9}$$

where, the number of the scenario samples N to be sorted can be calculated, when γ_1, γ_2 and β are predefined (Nutt and Wallis, 2004), according to Eq. (10):

$$\beta=1-\gamma_1^N-\left(1-\gamma_2\right)^N+\left[\gamma_1+\left(1-\gamma_2\right)-1\right]^N \tag{10}$$

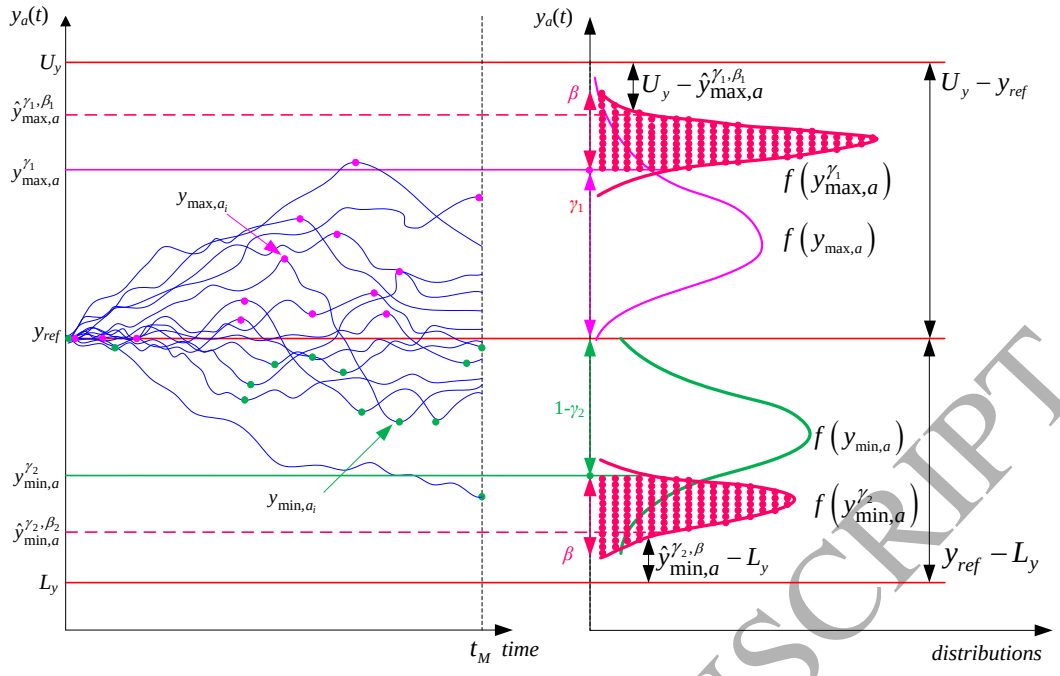

Fig. 11. N pairs of maximum and minimum values of the accidental scenario are used to estimate $\hat{y}^{\gamma_1,\beta}_{\max,a}$ *and* $\hat{y}^{\gamma_2,\beta}_{\min,a}$*, and, thus, to estimate* $M^{\gamma_1,\gamma_2,\beta}_{T,y_a}$

3.2 Cyber Threats Prioritization

The responses of ALFRED to cyber attacks to sensors, actuators and PI regulators are investigated by simulation. From the simulations outcomes, safety margins of the four controlled variables (i.e., T_{steam}, p_{SG}, $T_{L,cold}$, and P_{Th}) are estimated, to quantify the effects of the cyber attacks on the system functionalities. A total of N_1=29 runs of the ALFRED model are simulated, to satisfy the requirements of the percentiles estimations of the safety parameters by the Bracketing OS of Section 3.1, with respect to both one-sided (N=22, given *i*) γ_1= 90th, β_1=90th, or *ii*) γ_2=10th, β_2=90th) and two-sided (N=29, given γ_1,= 90th, γ_2=10th, β=90th) thresholds. Accordingly, N=22 samples are randomly taken to estimate the safety margins of T_{steam} (with respect to its upper threshold) and $T_{L,cold}$ (with respect to its lower threshold) and N=29 samples are used to estimate the safety margins of p_{SG} and P_{Th} (with respect to their two-sided thresholds).

Effects of cyber attacks on the CPS components and on the system integrity are qualitatively ranked according to a three-level risk metric (see Table 7). Table 8 shows

the quantified design safety margins when the code is run 29 times under nominal conditions (proving that the system works with ample safety margins).

Table 7 A three-level risk metric for ranking the effects of cyber attacks on the CPS

Effect	$M_{\#3}^{\#1(or\#2)}$
Negligible	[0.8, 1.0]
Medium	[0.2, 0.8)
Severe	[0.0, 0.2)

Note: 1) #1 refers to "γ_2,β_2" for $T_{L,cold}$, and to "γ_2,β" for p_{SG} and P_{Th};
2) #2 refers to "γ_1,β_1" for T_{steam}, and to "γ_1,β" for p_{SG} and P_{Th};
3) #3 refers to "U,y_a" for T_{steam}, to L,y_a for $T_{L,cold}$, and to "T,y_a" for p_{SG} and P_{Th};
4) γ_1= 90th, γ_2=10th, β_1=90th, β_2=90th, β=90th.

Table 8 Safety margins estimation of the safety parameters under normal conditions

Variable	T_{steam} (°C)	p_{SG} (Pa)	$T_{L,cold}$ (°C)	P_{Th} (W)
$\hat{y}_{\min,y_a}^{\#1}$	-	$1.7967\cdot10^7$	396.1839	$2.9819\cdot10^8$
$\hat{y}_{\max,y_a}^{\#2}$	455.3330	$1.8029\cdot10^7$	-	$3.0181\cdot10^8$
$M_{\#3}^{\#1(or\#2)}$	0.9667	0.9672	0.9237	0.9396

4. RESULTS

4.1 Sensors

Table 9 presents the results of the safety margins estimation of the four types of failure modes of the four sensors measuring the values of the controlled variables, i.e., T_{steam}, p_{SG}, $T_{L,cold}$, and P_{Th}.

Table 9 Safety margins estimation of the safety parameters of the cyber attacks to sensors

Scenario a		$M_{U,T_{steam},a}^{\gamma_1,\beta_1}$	$M_{T,p_{SG},a}^{\gamma_1,\gamma_2,\beta}$	$M_{L,T_{L,cold},a}^{\gamma_2,\beta_2}$	$M_{T,P_{Th},a}^{\gamma_1,\gamma_2,\beta}$
T_{steam} sensor	bias	0.9562	0.9626	0.9350	0.9349
	drift	0.9604	0.9654	0.9188	0.9322
	wider noise	0.9579	0.9478	0.9195	0.9330
	freezing	0.9604	0.9654	0.9203	0.9374
p_{SG} sensor	bias	0.8185	0	0.7776	0.8136
	drift	0.8701	0	0.9042	0.9335
	wider noise	0.9349	0.4098	0.9257	0.9220
	freezing	0.8988	0	0.9031	0.8600
$T_{L,cold}$ sensor	bias	0.5875	0.5787	0.5436	0.3838
	drift	0	0	0.5469	0
	wider noise	0.9138	0.9002	0.9085	0.9073
	freezing	0.2187	0.9722	0.6261	0.4707
P_{Th} sensor	bias	0.9641	0	0.2342	0

	drift	0.9539	0.9662	0.8811	0
	wider noise	0.9601	0.9649	0.9212	0.9326
	freezing	0.9657	0.9645	0.9261	0.7672

The results show that cyber attacks leading to T_{steam} sensor failures do not affect the system functioning because all safety parameters are negligibly affected. System integrity can be affected by cyber attacks to the p_{SG}, $T_{L,cold}$ and P_{Th} sensors, directly resulting in large variations of the respective variables (attacks to P_{Th} with a minor impact on the other controlled variables, whereas, cyber attacks to $T_{L,cold}$ sensor, e.g., bias, drift, or freezing, may impact the whole physical system).

As example, Fig. 12 shows the evolution of the safety parameters when the $T_{L,cold}$ sensor is affected by the freezing failure mode. In all cases, the lead temperature at the SG outlet, $T_{L,cold}(t)$ deviates from its set point equal to 400°C (Fig. 12(a)), due to the PI_3 response to the frozen value $T_{L,cold,sensor}(t)$. Then, the steam SG outlet temperature T_{steam} changes accordingly to the change of the lead temperature (Fig. 12(b)), causing the change of Thermal power P_{Th} (Fig. 12(d)). SG pressure change (Fig. 12(c)) is negligible thanks to the effective regulation of the steam mass flow rate by the turbine admission valve. These alterations are well caught by the safety margin analysis. In particular, the safety margin of T_{steam}, $T_{L,cold}$, and P_{Th} in case of $T_{L,cold}$ sensor freezing (Table 10) result to be equal to 0.2187, 0.6260, and 0.4707, respectively. This corresponds to a "medium" effect, according to the predefined risk metric of Table 7. On the other hand, SG *Pressure* is kept approximately at the nominal level with little disturbances, and, thus, "negligibly" affected by the cyber attacks.

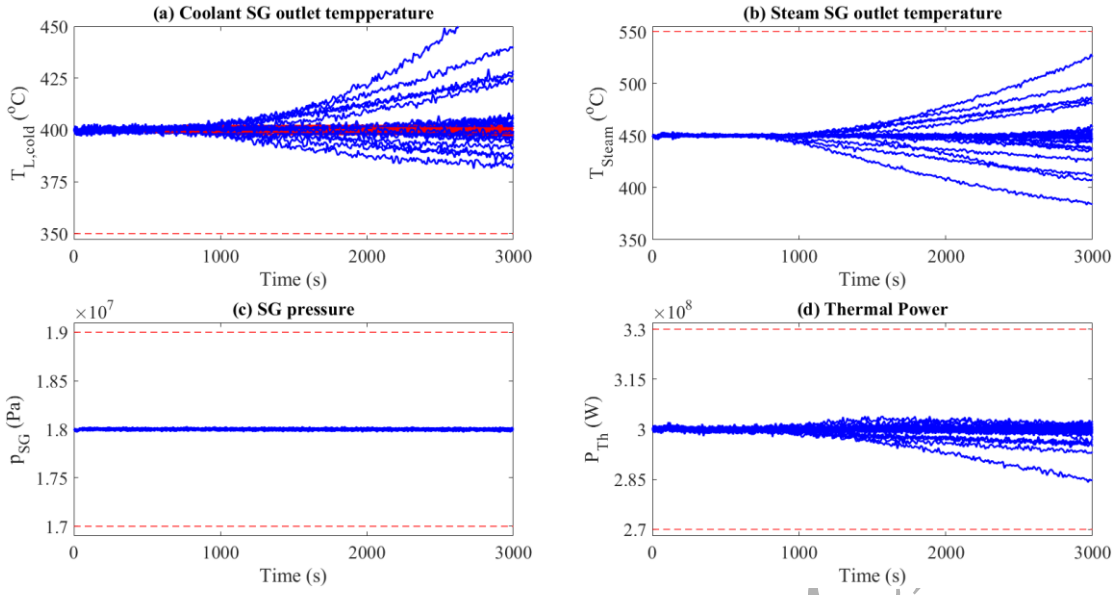

Fig. 12. Profiles of the safety parameters for N_T=29 runs, when $T_{L,cold}$ sensor is frozen: (a) evolution of lead temperature in the cold leg; (b) evolution of steam SG output temperature; (c) evolution of SG pressure; and (d) evolution of reactor thermal power

Table 10 Safety margins estimation of the safety parameters of $T_{L,cold}$ sensor freezing cyber attack scenarios

Variable	T_{steam} (°C)	p_{SG} (Pa)	$T_{L,cold}$ (°C)	P_{Th} (W)
$\hat{y}^{\#1}_{\min,y_a}$	-	$1.7972 \cdot 10^7$	381.3042	$2.8412 \cdot 10^8$
$\hat{y}^{\#2}_{\max,y_a}$	528.1336	$1.8026 \cdot 10^7$	-	$3.0391 \cdot 10^8$
$M^{\#1(or\#2)}_{\#3}$	0.2187	0.9722	0.6260	0.4707

Note: 1) *a* in this Table refers to $T_{L,cold}$ *sensor freezing*, denoting that the simulation is run to simulate the system dynamic scenario processing when the $T_{L,cold}$ sensor is attacked to freezing and, to test the effects of such cyber attacks on the system integrity.

4.2 Actuators

The results of the safety margins estimation of the three actuator failures are shown in Table 11. The cyber attacks leading to actuator-stuck failure at a random output level, severely affect the system functioning and integrity since most of the safety margins of the parameters turn out to be less than 0.2. This evidence should raise defenders' concern, because the ALFRED dynamics would be severely affected if cyber breaches are injected into these vulnerable components.

Table 11 Safety margins estimation of the safety parameters of the cyber attacks to actuators

Scenario *a*	$M^{\gamma_1,\beta_1}_{U,T_{steam,a}}$	$M^{\gamma_1,\gamma_2,\beta}_{T,p_{SG,a}}$	$M^{\gamma_2,\beta_2}_{L,T_{L,cold,a}}$	$M^{\gamma_1,\gamma_2,\beta}_{T,P_{Th,a}}$

CR height stuck	0.4895	0	0.7532	0
Water pump stuck	0	0	0.5441	0
Turbine valve stuck	0.1708	0	0.8484	0.1792

As illustrative example, Fig. 13 shows the evolution of the safety parameters when the water pump is attacked to fail stuck with a random value sampled from the uniform distribution in $U(0,300)$ mentioned in Table 5, at a random time t_A. The feedwater mass flow rate G_{water} is output at a constant value in each case and this directly affects the SG performance. As a result, the lead temperature at the SG outlet $T_{L,cold}$ (Fig. 13(a)) and steam SG outlet temperature T_{steam} (Fig. 13(b)) are strongly affected. Then, changes in T_{steam} cause transients of SG pressure (Fig. 13(c)), and, at the same time, $T_{L,cold}$ causes the CRs regulation that affects the reactor thermal power P_{Th} (Fig. 13(d)). The results are shown in Table 12. Regarding the lower threshold, the safety margin of $T_{L,cold}$ turns out to be 0.5441, classified as a "medium" effect, according to three-level risk metric of Table 7. On the other hand, all safety margins of T_{steam}, p_{SG}, and P_{Th}, result to be equal to 0, indicating that a cyber attack to the water pump-stuck would "severely" affect the system dynamics and integrity.

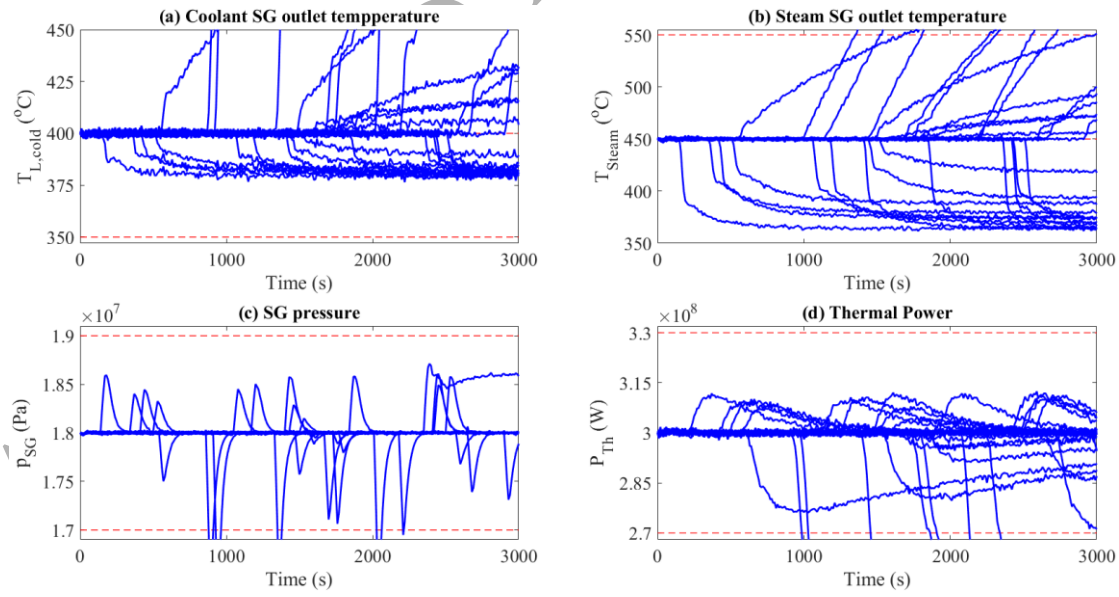

Fig. 13 Profiles of the safety parameters for N_T=29 runs, when the water pump is attacked to fail stuck with a random value: (a) evolution of lead temperature in the cold leg; (b) evolution of steam SG output temperature; (c) evolution of SG pressure; and (d) evolution of thermal power

Table 12 Safety margins estimation of the safety parameters of water pump-stuck

cyber attack scenarios

Variable	T_{steam} (°C)	p_{SG} (Pa)	$T_{L,cold}$ (°C)	P_{Th} (W)
$\hat{y}^{\#1}_{\min,y_a}$	-	$1.6089\cdot10^{7}$	377.2037	$2.0677\cdot10^{8}$
$\hat{y}^{\#2}_{\max,y_a}$	715.0707	$1.8712\cdot10^{7}$	-	$3.1242\cdot10^{8}$
$M^{\#1(or\#2)}_{\#3}$	0	0	0.5441	0

Note: 1) *a* in this Table refers to *water pump-stuck*, denoting that the simulation is run to test the system dynamic scenario processing when the water pump is attacked to get stuck in a random value and, to test the effects of such cyber attacks on the system integrity.

4.3 PI Controllers

The safety margins estimation results of cyber attacks to PI gains and set points are presented in Table 13. Cyber attacks to change of PI gain values have negligible effects on the safety parameters and on the system functionalities (except for changes of the K_p value of PI_3). This is potentially ascribed to the PI controller capability of regulating the errors of controlled variables close to zero even if the (relative small) gain values are changed to 3 or 4 orders of magnitude larger than the reference settings. On the other hand, cyber attacks changing the controllers set point values (i.e., $p_{SG,set}$, $T_{L,cold,set}$, $P_{Th,set}$) are more likely to cause system performance degradation. Such evidences demonstrate that PI gain values play a less important role, compared with the residual between the measurement and the set point value, $e(t)$.

Table 13 Safety margins estimation of the safety parameters of the cyber attacks to PI regulator value changes

Scenario *a*		$M^{\gamma_1,\beta_1}_{U,T_{steam,a}}$	$M^{\gamma_1,\gamma_2,\beta}_{T,p_{SG,a}}$	$M^{\gamma_2,\beta_2}_{L,T_{L,cold,a}}$	$M^{\gamma_1,\gamma_2,\beta}_{T,P_{Th,a}}$
PI_1	K_p	0.9676	0.9696	0.9203	0.9355
	K_i	0.9624	0.9698	0.9219	0.9232
	$T_{steam,set}$	0.9534	0.9591	0.9263	0.9295
PI_2	K_p	0.9612	0.9722	0.9304	0.9321
	K_i	0.9677	0.9684	0.9213	0.9370
	$p_{SG,set}$	0.9647	0.0213	0.9260	0.9300
PI_3	K_p	0.9451	0.7570	0.9156	0.8981
	K_i	0.9677	0.9660	0.9199	0.9414
	$T_{L,cold,set}$	0.6879	0.8840	0.5739	0.6264
PI_4	K_p	0.9623	0.9671	0.9168	0.9287
	K_i	0.9657	0.9699	0.9187	0.9343
	$P_{Th,set}$	0.9655	0.9685	0.9120	0.4628

Fig. 14 shows the evolution of the safety parameters when the reference value of

K_p of PI_1 is attacked at a random time t_A and changes to a random value distributed as U(1e-2,1) (see Table 6).

Under such circumstances, the steam SG outlet temperature T_{steam} (Fig. 14(a)) is negligibly affected. The most probable reason is that K_p plays a less important role in PI computation, compared with the residual between the measurement T_{steam} and the set point value $T_{steam,set}$, $e(t)$. The resulting negligible change of T_{steam} will not lead to any transients of SG functioning. Also, the evolutions of SG pressure p_{SG} (Fig. 14(b)), of lead temperature at the SG outlet $T_{L,cold}$ (Fig. 14(c)), and of reactor thermal power P_{Th} (Fig. 14(d)) are not altered with respect to normal conditions. Safety margins of T_{steam}, p_{SG}, $T_{L,cold}$ and P_{Th} result to be equal to 0.9676, 0.9696, 0.9203, and 0.9355, respectively.

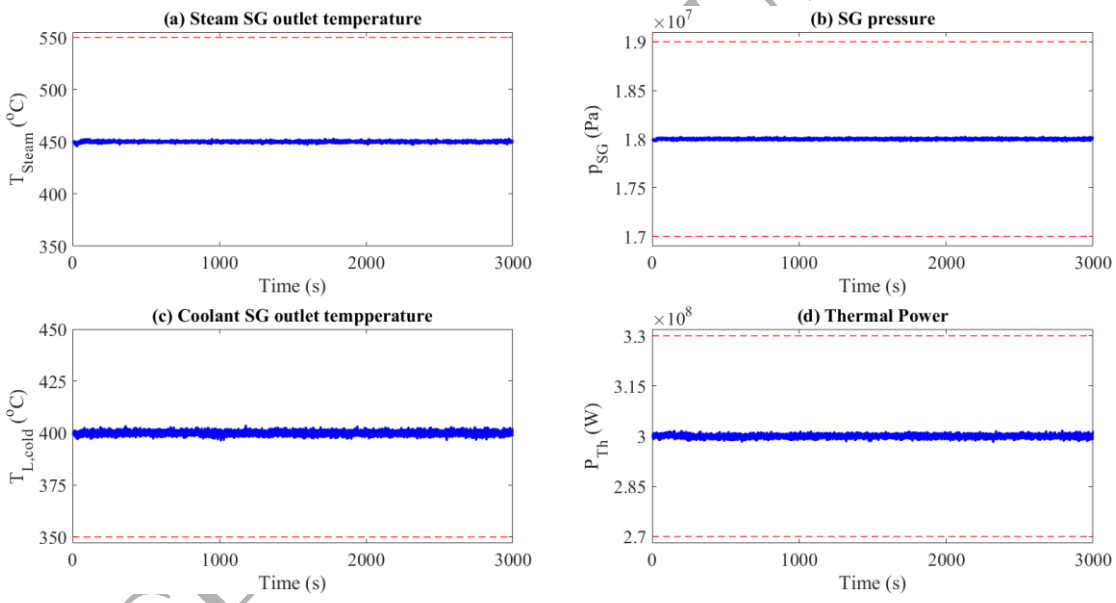

Fig. 14 Profiles of the safety parameters for N_T=29 runs, when the PI_1 K_p gain is changed to a random value: (a) evolution of steam SG output temperature; (b) evolution of SG pressure; (c) evolution of lead temperature in the cold leg; and (d) evolution of reactor thermal power

Table 14 Safety margins estimation of the safety parameters of change of K_p value of PI_1 cyber attack scenarios

Variable	T_{steam} (°C)	p_{SG} (Pa)	$T_{L,cold}$ (°C)	P_{Th} (W)
$\hat{y}^{\#1}_{\min,y_a}$	-	$1.7971\cdot10^7$	396.0125	$2.9809\cdot10^8$
$\hat{y}^{\#2}_{\max,y_a}$	453.2381	$1.8030\cdot10^7$	-	$3.0193\cdot10^8$
$M^{\#1(or\#2)}_{\#3}$	0.9676	0.9696	0.9203	0.9355

Note: 1) *a* in this Table refers to *change of K_p value of PI_1*, denoting that the simulation is run to test the system dynamic scenario processing when the K_p gain value of PI_1 is attacked to be changed to a random value and, to test the effects of such cyber attacks on the system integrity.

4.4 Multiple Cyber Attacks to PI Controllers

Coupling of malicious alteration of PI gain values and set points are hereafter considered, being expected to lead to more severe effects on the PI performance and, therefore, on the system safety, rather than the single failure modes considered in Section 4.3. In each simulation run, the PI gain value and the controlled variable set point value are both attacked at a random time t_A, within the mission time of t_M. Thus, eight types of the multiple cyber attack scenarios are considered (see Table 15 first two columns).

Results are shown in Table 15. Except for the cyber attacks to PI_1 resulting in negligible effects, the scenarios originated by attacks to PI_2, PI_3 and PI_4 would result in more severe impacts on the safety parameters and the system functionality, compared with single cyber attacks of PIs of Table 13. The evidence entails further concerns on the protection design of multiple cyber attack scenarios, for optimizing the cyber defense strategies from the perspective of the defenders.

Table 15 Safety margins estimation of the safety parameters of the sequence of multiple cyber attacks to PI regulators

Scenario *a*		$M^{\gamma_1,\beta_1}_{U,T\_steam_a}$	$M^{\gamma_1,\gamma_2,\beta}_{T,Pressure_a}$	$M^{\gamma_2,\beta_2}_{L,T\_cold\_leg_a}$	$M^{\gamma_1,\gamma_2,\beta}_{T,Th\_power_a}$
PI_1	K_p & $T_{steam,set}$	0.9562	0.9560	0.9246	0.9373
	K_i & $T_{steam,set}$	0.9590	0.9607	0.9134	0.9352
PI_2	K_p & $p_{SG,set}$	0.9670	0	0.9224	0.9203
	K_i & $p_{SG,set}$	0.9628	0.0566	0.9200	0.9342
PI_3	K_p & $T_{L,cold,set}$	0.6182	0.2687	0.6300	0.4196
	K_i & $T_{L,cold,set}$	0.7991	0.8850	0.5872	0.5840
PI_4	K_p & $P_{Th,set}$	0.8170	0.4819	0.8355	0
	K_i & $P_{Th,set}$	0.9675	0.9719	0.9195	0.5201

4.5 Comments

The results of the single failure modes of Sections 4.1, 4.2 and 4.3 suggest insightful recommendations. On one hand, the cyber attacks to actuators (control rod height, water pump and turbine coefficient valve) seem to be the most worrying for

the entire system functionality and integrity. The effect of the attacks to the p_{SG} and P_{Th} sensors are limited to the secondary and primary circuits, respectively, making them less problematic. The situation is different if the attack involves the $T_{L,cold}$ sensor, since the whole system is affected by a departure from the nominal values, underlying once again the relevance of the lead temperature control. Functionality of the ALFRED reactor will be negligibly affected, if attackers access the T_{steam} sensor database or the values of PI gains. As a last remark, it is important to point out that multiple cyber attacks originated by the coupled alteration of gain values and set points, discussed in Section 4.4, raise further concerns on protection design decision-making for counteracting cyber threats, compared with single failure modes of controllers.

We conclude by noting that, in practice, modality, timing and sequencing of cyber attacks are less predictable than stochastic failures, making the identification of the most vulnerable components to cyber threats an issue of outmost importance for protection design. Optimal protection design strategies have to be considered also on the basis of cyber threats prioritization, on one hand and, on the other hand, a trade-off between safety.

5. CONCLUSIONS

In this study, we have proposed a Monte Carlo-based exploration framework for generating cyber attack scenarios in Cyber-Physical Systems (CPSs) and accounting for multiple failure modes of attacked components of the CPSs, to test the effects of the cyber threats on the system functionality and integrity, and to prioritize the most vulnerable components for cyber security protection decision-making.

A safety margin estimation approach has been proposed for cyber threat prioritization. Safety margins of the safety parameters are estimated by a Bracketing OS approach, with respect to the one- and two-sided thresholds.

We have taken the digital I&C system of the Advanced Lead-cooled Fast Reactor European Demonstrator (ALFRED) as case study, in which cyber breach events aiming at attacking the embedded CPS components are injected by a Monte Carlo

sampling procedure, at random times and of random magnitudes. The results of the case study identify actuators as the most vulnerable CPS components, their failures leading more easily to the loss of system functionality and integrity, along with the lead temperature sensor, which is relevant component for the control of the temperature lead in the cold pool.

With due caution, in future works we seek to accommodate the notion of probabilistic safety margin assessment (Di Maio et al., 2016b; Grabaskas et al., 2015; Zio et al., 2008) to encompass the explorative characteristics of the here proposed framework and the underlying (if any) probabilistic distributions behind cyber attacks and attackers behaviors. Besides, other future works will regard, on one hand, the development of both statistical and dynamic scenario processing methods with the purpose of distinguishing between cyber attacks and stochastic failures, and, on the other hand, modeling possible attack strategies considering factors such as frequencies of occurrence, component compromise probabilities, attack costs, etc., and optimizing defense countermeasures, considering factors such as economics loss and defense costs. Both single and multiple cyber attack scenarios will be considered in these future works. In particular, a special concern is to model the effects of cyber attacks on the power grid, with reference to the therein transferred mechanical power, since the stability of plant power production plays an important role in maintaining functionality and integrity of the complex power infrastructure, where the NPP is functionally located.

REFERENCES

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.

Ali, S., Qaisar, S.B., Saeed, H., Khan, M.F., Naeem, M. and Anpalagan, A., 2015. Network challenges for cyber physical systems with tiny wireless devices: A case study on reliable pipeline condition monitoring. *Sensors*, 15(4), pp.7172-

7205.

Aven, T., 2016. Ignoring scenarios in risk assessments: Understanding the issue and improving current practice. *Reliability Engineering & System Safety*, 145, pp.215-220.

Aven, T. and Krohn, B.S., 2014. A new perspective on how to understand, assess and manage risk and the unforeseen. *Reliability Engineering & System Safety*, 121, pp.1-10.

Backhaus, S., Bent, R., Bono, J., Lee, R., Tracey, B., Wolpert, D., Xie, D. and Yildiz, Y., 2013. Cyber-physical security: A game theory model of humans interacting over control systems. *IEEE Transactions on Smart Grid*, 4(4), pp.2320-2327.

Beran, R. and Hall, P., 1993. Interpolated nonparametric prediction intervals and confidence intervals. Journal of the Royal Statistical Society. Series B (Methodological), pp.643-652.

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.

Di Maio, F., Secchi, P., Vantini, S. and Zio, E., 2011. Fuzzy C-means clustering of signal functional principal components for post-processing dynamic scenarios of a nuclear power plant digital instrumentation and control system. IEEE Transactions on Reliability, 60(2), pp.415-425.

Di Maio, F., Bandini, A., Zio, E., Alberola, S.C., Sanchez-Saez, F. and Martorell, S., 2016a. Bootstrapped-ensemble-based Sensitivity Analysis of a trace thermal-hydraulic model based on a limited number of PWR large break LOCA simulations. *Reliability Engineering & System Safety*, 153, pp.122-134.

Di Maio, F., Rai, A. and Zio, E., 2016b. A dynamic probabilistic safety margin characterization approach in support of Integrated Deterministic and Probabilistic Safety Analysis. *Reliability Engineering & System Safety*, 145,

pp.9-18.

Di Maio, F., Picoco, C., Zio, E. and Rychkov, V., 2017. Safety margin sensitivity analysis for model selection in nuclear power plant probabilistic safety assessment. *Reliability Engineering & System Safety*, 162, pp.122-138.

Ding, D., Wang, Z., Han, Q.L. and Wei, G., 2016. Security control for discrete-time stochastic nonlinear systems subject to deception attacks. IEEE Transactions on Systems, Man, and Cybernetics: Systems.

DYMOLA, 2015. Dymola (Version 2015). France: Dassault Systèmes. Retrieved from http://www.3ds.com/products-services/catia/products/dymola.

Efron, B. and Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science, pp.54-75.

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

Feng, Z., Wen, G. and Hu, G., 2017. Distributed secure coordinated control for multiagent systems under strategic attacks. IEEE transactions on cybernetics, 47(5), pp.1273-1284.

Fovino, I.N., Masera, M. and De Cian, A., 2009. Integrating cyber attacks within fault trees. *Reliability Engineering & System Safety*, 94(9), pp.1394-1402.

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

Grabaskas, D., Bucknor, M., Brunett, A. and Nakayama, M., 2015. Quantifying Safety Margin Using the Risk-Informed Safety Margin Characterization (RISMC). Proceedings of PSA, pp.26-30.

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.

Huang, Y.L., Cárdenas, A.A., Amin, S., Lin, Z.S., Tsai, H.Y. and Sastry, S., 2009.

Understanding the physical and economic consequences of attacks on control systems. International Journal of Critical Infrastructure Protection, 2(3), pp.73-83.

Hutson, A.D., 1999. Calculating nonparametric confidence intervals for quantiles using fractional order statistics. Journal of Applied Statistics, 26(3), pp.343-353.

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

Ingols, K., Lippmann, R. and Piwowarski, K., 2006, December. Practical attack graph generation for network defense. In Computer Security Applications Conference, 2006. ACSAC'06. 22nd Annual (pp. 121-130). IEEE.

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.

Khalid, H.M. and Peng, J.C.H., 2016. A Bayesian Algorithm to Enhance the Resilience of WAMS Applications Against Cyber Attacks. *IEEE Transactions on Smart Grid*, 7(4), pp.2026-2037.

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.

Lehmann, E.L. and Casella, G., 2006. Theory of point estimation. Springer Science & Business Media.

Liu, K., Lee, V.C.S., Ng, J.K.Y., Chen, J. and Son, S.H., 2014. Temporal data

dissemination in vehicular cyber–physical systems. *IEEE Transactions on Intelligent Transportation Systems*, 15(6), pp.2419-2431.

McQueen, M.A., Boyer, W.F., Flynn, M.A. and Beitel, G.A., 2006, January. Quantitative cyber risk reduction estimation methodology for a small SCADA control system. In System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on (Vol. 9, pp. 226-226). IEEE.

Mitchell, R. and Chen, R., 2013. Effect of intrusion detection and response on reliability of cyber physical systems. *IEEE Transactions on Reliability*, 62(1), pp.199-210.

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

Nutt, W.T. and Wallis, G.B., 2004. Evaluation of nuclear safety from the outputs of computer codes in the presence of uncertainties. *Reliability Engineering & System Safety*, 83(1), pp.57-77.

Paelke, V. and Röcker, C., 2015, August. User interfaces for cyber-physical systems: challenges and possible approaches. In *International Conference of Design, User Experience, and Usability* (pp. 75-85). Springer International Publishing.

Pasqualetti, F., Dörfler, F. and Bullo, F., 2013. Attack detection and identification in cyber-physical systems. *IEEE Transactions on Automatic Control*, 58(11), pp.2715-2729.

Paté-Cornell, E., 2002. Finding and fixing systems weaknesses: Probabilistic methods and applications of engineering risk analysis. *Risk Analysis*, 22(2), pp.319-334.

Paté-Cornell, E., 2012. On "Black Swans" and "Perfect Storms": risk analysis and management when statistics are not enough. *Risk analysis*, 32(11), pp.1823-1833.

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.

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.

Sanchez-Saez, F., Sánchez, A.I., Villanueva, J.F., Carlos, S., Martorell, S., 2017. Uncertainty analysis of a LBLOCA in a PWR nuclear power plant using TRACE with Wilks as compared to other non-parametric methods. Reliability Engineering & System Safety.

Santner, T.J., Williams, B.J. and Notz, W.I., 2013. The design and analysis of computer experiments. Springer Science & Business Media.

Schneier, B., 1999. Attack trees. Dr. Dobb's journal, 24(12), pp.21-29.

Sheyner, O. and Wing, J., 2003, November. Tools for generating and analyzing attack graphs. In International Symposium on Formal Methods for Components and Objects (pp. 344-371). Springer, Berlin, Heidelberg.

Simpson, T.W., Poplinski, J.D., Koch, P.N. and Allen, J.K., 2001. Metamodels for computer-based engineering design: survey and recommendations. *Engineering with computers*, 17(2), pp.129-150.

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

Turati, P., Pedroni, N. and Zio, E., 2017a. An adaptive simulation framework for the exploration of extreme and unexpected events in dynamic engineered systems. *Risk analysis*, 37(1), pp.147-159.

Turati, P., Pedroni, N. and Zio, E., 2017b. Simulation-based exploration of high-dimensional system models for identifying unexpected events. Reliability Engineering & System Safety, 165, pp.317-330.

Wald, A., 1943. An extension of Wilks' method for setting tolerance limits. The

Annals of Mathematical Statistics, 14(1), pp.45-55.

Wang, W., Di Maio, F. and Zio, E., 2017a. Hybrid Fuzzy-PID Control of a Cyber-Physical System Working Under Varying Environmental Conditions. *Nuclear Engineering and Design, under review (2nd revision)*.

Wang, W., Di Maio, F., Zio, E., 2017b. Three-Loop Monte Carlo Simulation Approach to Multi-State Physics Modeling for System Reliability Assessment. *Reliability Engineering & System Safety*, Vol. 167, pp. 276–289.

Wang, W., Di Maio F., Zio, E., 2018. A Non-Parametric Cumulative Sum Approach for Online Diagnostics of Cyber Attacks to Nuclear Power Plants. *Resilience of Cyber-Physical Systems: From Risk Modelling to Threat Counteraction*, Springer, *accepted*.

Wilks, S.S., 1941. Determination of sample sizes for setting tolerance limits. The Annals of Mathematical Statistics, 12(1), pp.91-96.

Wilks, S.S., 1942. Statistical prediction with special reference to the problem of tolerance limits. The annals of mathematical statistics, 13(4), pp.400-409.

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.

Xiang, Y., Wang, L. and Zhang, Y., 2018. Adequacy evaluation of electric power grids considering substation cyber vulnerabilities. *International Journal of Electrical Power & Energy Systems*, 96, pp.368-379.

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.

Yuan, Y., Yuan, H., Guo, L., Yang, H. and Sun, S., 2016. Resilient control of networked control system under DoS attacks: A unified game approach. IEEE Transactions on Industrial Informatics, 12(5), pp.1786-1794.

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.

Zhang, H., Cheng, P., Shi, L. and Chen, J., 2016. Optimal DoS attack scheduling in wireless networked control system. IEEE Transactions on Control Systems Technology, 24(3), pp.843-852.

Zhu, M. and Martínez, S., 2014. On the performance analysis of resilient networked control systems under replay attacks. IEEE Transactions on Automatic Control, 59(3), pp.804-808.

Zeng, Z. and Zio, E., 2017. An integrated modeling framework for quantitative business continuity assessment. *Process Safety and Environmental Protection*, 106, pp.76-88.

Zio, E., Di Maio, F., Martorell, S. and Nebot, Y., 2008. Neural networks and order statistics for quantifying nuclear power plants safety margins. In Proceedings, European Safety & Reliability Conference (ESREL).

Zio, E., 2013. The Monte Carlo simulation method for system reliability and risk analysis (Vol. 39). London: Springer.

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., Di Maio, F. and Tong, J., 2010. Safety margins confidence estimation for a passive residual heat removal system. *Reliability Engineering & System Safety*, 95(8), pp.828-836.