TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
Gallo, A. J., C. Anand, S., Teixeira, A. M. H., & Ferrari, R. M. G. (2025). Switching multiplicative watermark design against covert attacks. *Automatica*, *177*, Article 112301. https://doi.org/10.1016/j.automatica.2025.112301

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Automatica 177 (2025) 112301

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

Switching multiplicative watermark design against covert attacks☆

Alexander J. Gallo [a,1], Sribalaji C. Anand [b,*,1], Andre M.H. Teixeira [c], Riccardo M.G. Ferrari [d]

[a] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy
[b] School of Electrical Engineering and Computer Science and Digital Futures, KTH Royal Institute of Technology, Sweden
[c] Department of Information Technology, Uppsala University, PO Box 337, SE-75105, Uppsala, Sweden
[d] Delft Center for Systems and Control, Mechanical Engineering, TU Delft, Delft, Netherlands

ARTICLE INFO

Article history:
Received 21 February 2024
Received in revised form 10 January 2025
Accepted 26 February 2025

Keywords:
Network security
Networked control systems
Fault detection and isolation

ABSTRACT

Active techniques have been introduced to give better detectability performance for cyber-attack diagnosis in cyber–physical systems (CPS). In this paper, switching multiplicative watermarking is considered, whereby we propose an optimal design strategy to define switching filter parameters. Optimality is evaluated exploiting the so-called output-to-output gain of the closed-loop system, including some supposed attack dynamics. A worst-case scenario of a matched covert attack is assumed, presuming that an attacker with full knowledge of the closed-loop system injects a stealthy attack of bounded energy. Our algorithm, given watermark filter parameters at some time instant, provides optimal next-step parameters. Analysis of the algorithm is given, demonstrating its features, and demonstrating that through initialization of certain parameters outside of the algorithm, the parameters of the multiplicative watermarking can be randomized. Simulation shows how, by adopting our method for parameter design, the attacker's impact on performance diminishes.

© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

The widespread integration of communication networks and smart devices in modern control systems has increased the vulnerability of industrial systems to online cyber-attacks, e.g., Industroyer, Blackenergy, etc. (Hemsley & Fisher, 2018). To counter this, methods have been developed to improve security by achieving attack detection, mitigation, and monitoring, among others (Sandberg et al., 2022). This paper focuses on active attack diagnosis to mitigate stealthy attacks.

Active diagnosis techniques rely on the inclusion of additional moduli to control systems to alter the behavior of the system compared to information known by the attacker. For instance, the concept of additive watermarking introduced in Mo et al. (2015), where noise signals of known mean and variance are added at the plant and compensated for it at the controller. This compensation, however, is not exact, causing some performance degradation. Thus, trade-offs between performance and detectability are necessary (Zhu et al., 2023).

In encrypted control (Darup et al., 2021), the sensor data is encrypted, sent to the controller, and then operated on directly. Encrypted input signals are sent back to the plant for decryption. Although encryption is widespread in IT security, in control systems it presents some concerns, such as the introduction of time delays (Stabile et al., 2024), while it may present inherent weaknesses (Alisic et al., 2023).

In moving target defense (Griffioen et al., 2020), the plant is augmented with fictitious dynamics, known to the controller. The plant output is transmitted to the controller along with the fictitious states over a network under attack. The additional measurements then aide in the detection of attacks. This comes at the cost of higher communication bandwidth needs, which increases rapidly with the dimension of the augmented systems.

Other recently proposed works include two-way coding (Fang et al., 2019), a weak encryption technique, and dynamic masking (Abdalmoaty et al., 2023), which enhances privacy as well as security, have been shown to be effective against zero-dynamics attacks. Furthermore, filtering techniques for attack detection are proposed by Escudero et al. (2023), Hashemi and Ruths (2022), Murguia et al. (2020), while not focusing on stealthy attacks.

Multiplicative watermarking (mWM) has been proposed by the authors as a diagnosis technique (Ferrari & Teixeira, 2021). mWM consists of a pair of filters on each communication channel between the plant and its controller; the scheme is affine to weak

☆ This work has been partially supported by the Research Council of Norway through the project AIMWind (grant ID 312486), by the Swedish Research Council under the grant 2018-04396, and by the Swedish Foundation for Strategic Research. The material in this paper was partially presented at the 60th IEEE Conference on Decision and Control, December 13–15, 2021, Austin, Texas, USA. This paper was recommended for publication in revised form by Associate Editor Angelo Alessandri under the direction of Editor Thomas Parisini.

* Corresponding author.
E-mail addresses: alexanderjulian.gallo@polimi.it (A.J. Gallo), srca@kth.se (S.C. Anand), andre.teixeira@it.uu.se (A.M.H. Teixeira), r.ferrari@tudelft.nl (R.M.G. Ferrari).

[1] These authors contributed equally.

https://doi.org/10.1016/j.automatica.2025.112301
0005-1098/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

encryption, whereby "encoding" and "decoding" are done by changing signals' dynamic characteristics through inverse pairs of filters. This enables original signals to be recovered exactly, and thus does not lead to performance degradation.

One of the critical features of multiplicative watermarking is that to detect stealthy attacks, the mWM filter parameters must be switched over time. In this paper, an algorithm to optimally design the mWM parameters after a switching event is presented, enhancing detection performance, without changing the switching time.

To formalize the filter design problem, we suppose the defender is interested in optimal performance against adversaries injecting covert attacks with matched system parameters (Smith, 2015), including the mWM parameters prior to the switch. This scenario represents a worst case where malicious agents can take full control of the system while remaining undetected. Thus, the attack strategy is explicitly included within the formulation of the closed-loop system, and the mWM filters are chosen by solving an optimization problem minimizing the attack-energy-constrained output-to-output gain (AEC-OOG) (Anand & Teixeira, 2023), a variation of the output-to-output gain proposed in Teixeira, Sandberg et al. (2015). The main contributions of this paper are:

1. the problem of optimally designing the switching mWM filters is formulated as an optimization problem, with the AEC-OOG is taken as the objective;
2. the worst-case scenario of a covert attack with exact knowledge of plant and mWM filter parameters is embedded within the design problem;
3. the feasibility of the optimization problem is shown to be dependent only on stability conditions;
4. a solution scheme is proposed to promote randomization of the mWM filter parameters such that an eavesdropping adversary cannot remain stealthy.

This builds on the results of Ferrari and Teixeira (2021), where the focus was on the design of the switching protocols, rather than the parameters themselves. Compared to previous work (Gallo et al., 2021), this paper introduces an optimization problem which is always feasible (thanks to the use of AEC-OOG in the objective), while also considering a more sophisticated class of covert attacks, where the presence of watermark is known to the adversary. Moreover, this paper poses a different objective than Zhang et al. (2023); indeed, while Zhang et al. (2023) provided a design strategy to ensure certain privacy properties, in this paper we address the problem of optimal parameter design following a switching event.

The rest of the paper is organized as follows. After formulating the problem in Section 2, we propose our design algorithm in Section 3, and analyze its properties. It is then evaluated through a numerical example in Section 4, and concluding remarks are given Section 5.

2. Problem description

We consider the Cyber–Physical System (CPS) in Fig. 1. This includes plant $\mathcal{P}$, controller and anomaly detector $\mathcal{C}$, mWM filters $\mathcal{W}, \mathcal{Q}, \mathcal{G}, \mathcal{H}$, and the malicious agent $\mathcal{A}$. The mWM filters are defined pairwise, namely $\{\mathcal{Q}, \mathcal{W}\}$ and $\{\mathcal{G}, \mathcal{H}\}$ are referred to as, respectively, the output and input *mWM filter pairs*.

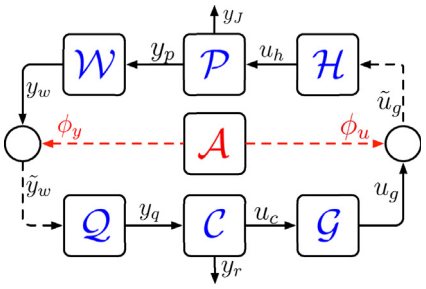


Fig. 1. Block diagram of the closed-loop CPS including the plant $\mathcal{P}$, controller $\mathcal{C}$ and watermarking filters $\{\mathcal{W}, \mathcal{Q}, \mathcal{G}, \mathcal{H}\}$. The information transmitted between $\mathcal{P}$ and $\mathcal{C}$ is altered by the adversary $\mathcal{A}$. The dashed lines represent the network affected by the adversary.

2.1. Plant and controller

Consider a linear time-invariant (LTI) discrete-time (DT) plant modeled by:

$$\mathcal{P}: \begin{cases} x_p[k+1] = A_p x_p[k] + B_p u_h[k] \\ y_p[k] = C_p x_p[k] \\ y_J[k] = C_J x_p[k] + D_J u_h[k] \end{cases} \tag{1}$$

where $x_p \in \mathbb{R}^n$ is the plant's state, $u_h \in \mathbb{R}^m$ its input, $y_p \in \mathbb{R}^p$ its measured output, and all the system's matrices are of the appropriate dimension. Furthermore, suppose a (possibly unmeasured) *performance output* $y_J \in \mathbb{R}^{p_J}$ is defined, such that the performance of the system, evaluated over the interval $[k-N+1, k]$, for some $N \in \mathbb{N}$ (Zhou et al., 1996), is given by:

$$J(x_p, u_h) = \|y_J\|^2_{\ell_2,[k-N+1,k]}.$$

Assumption 2.1. The tuples (A_p, B_p) and (C_p, A_p) are, respectively, controllable and observable pairs. ◁

Assumption 2.2. The plant $\mathcal{P}$ is stable and $x_p[0] = 0$. ◁

Assumption 2.2, necessary for the OOG to be meaningful (Teixeira, Shames et al., 2015), does not reduce generality, as stability can be ensured by a local (non-networked) controller (Hu & Yan, 2007; Lin et al., 2023), whilst $x_p[0] = 0$ can be considered because of linearity.

The plant is regulated by an observer-based dynamic controller $\mathcal{C}$, described by:

$$\mathcal{C}: \begin{cases} \hat{x}_p[k+1] = A_p \hat{x}_p[k] + B_p u_c[k] + L y_r[k] \\ u_c[k] = K \hat{x}_p[k] \\ \hat{y}_p[k] = C_p \hat{x}_p[k] \\ y_r[k] = y_q[k] - \hat{y}_p[k] \end{cases} \tag{2}$$

where $\hat{x}_p \in \mathbb{R}^n$, $\hat{y}_p \in \mathbb{R}^p$ are the state and measurement estimates, $u_c \in \mathbb{R}^m$ the control input. The matrices K and L are the controller and observer gains respectively. Finally, the term y_r in (2) is the residual output, used to detect the presence of an attack: given a threshold ϵ_r, an attack is detected if the inequality $\|y_r\|^2_{\ell_2,[0,N_r]} \leq \epsilon_r$ is falsified for any $N_r \in \mathbb{N}_+$. Note that in (1)–(2) y_q and u_h, the outputs of $\mathcal{Q}$ and $\mathcal{H}$ (to be defined), are used as the input to the controller and the plant, respectively.

2.2. Multiplicative watermarking filters

Consider mWM filters defined as follows

$$\Sigma: \begin{cases} x_\sigma[k+1] = A_\sigma(\theta_\sigma[k]) x_\sigma[k] + B_\sigma(\theta_\sigma[k]) \nu_\sigma[k] \\ \gamma_\sigma[k] = C_\sigma(\theta_\sigma[k]) x_\sigma[k] + D_\sigma(\theta_\sigma[k]) \nu_\sigma[k], \end{cases} \tag{3}$$

with $\Sigma \in \{\mathcal{G}, \mathcal{H}, \mathcal{W}, \mathcal{Q}\}$, $\sigma \in \{g, h, w, q\}$, where g, h, w, q refer to variables pertaining to $\mathcal{G}, \mathcal{H}, \mathcal{W}, \mathcal{Q}$, respectively,[2] $x_\sigma \in \mathbb{R}^{n_\sigma}$ the state of Σ, $\nu_\sigma \in \mathbb{R}^{m_\sigma}$ its input, $\gamma_\sigma \in \mathbb{R}^{p_\sigma}$ the output, and $\theta_\sigma[k]$ is a vector of parameters.

Definition 2.1 (*mWM filter parameters*)**.** The parameter $\theta_\sigma[k]$ is taken to be the concatenation of the vectorized form of all matrices $A_\sigma(\cdot), B_\sigma(\cdot), C_\sigma(\cdot), D_\sigma(\cdot)$. ◁

The parameter θ_σ is defined to be piecewise constant:

$$\theta_\sigma[k] = \bar{\theta}_\sigma[k_i], \forall k \in \{k_i, k_i + 1, \dots, k_{i+1} - 1\}$$

where $k_i, i = 0, 1, \dots \in \mathbb{N}_+$, are switching instants. In the following, with some abuse of notation, the time dependencies are dropped, with θ_σ and θ_σ^+ used to define the parameters before and after a switching instant, i.e., $\theta_\sigma = \theta_\sigma[k_i]$, $\theta_\sigma^+ = \theta_\sigma[k_{i+1}]$.

Furthermore, all filters are taken to be square systems, i.e., $m_\sigma = p_\sigma, \forall \sigma \in \{g, h, w, q\}$, and define $\nu_g \triangleq u_c$, $\nu_h \triangleq \tilde{u}_g$, $\nu_w \triangleq y_p$, $\nu_q \triangleq \tilde{y}_w$, $\gamma_g \triangleq u_g$, $\gamma_h \triangleq u_h$, $\gamma_w \triangleq y_w$, $\gamma_q \triangleq y_q$. Here, a *tilde* is used to highlight that $\tilde{u}_g, \tilde{y}_w$ are received through the insecure communication network and as such may be affected by attacks.

Remark 2.1. The objective of this paper is to *optimally* design the successive parameters of the mWM filters θ_σ^+, given their value θ_σ. It remains out of the scope of the paper to address other aspects of the switching mechanisms, such as determining the switching time, or defining the jump functions for the states. Interested readers are referred to Ferrari and Teixeira (2021). ◁

Definition 2.2 (*Watermarking pair*)**.** Two systems $(\mathcal{W}, \mathcal{Q})$ (3), are a *watermarking pair* if:

a. $\mathcal{W}$ and $\mathcal{Q}$ are stable and invertible, i.e., exists a positive definite matrix $Z_\sigma \succ 0, \sigma \in \{w, q\}$ such that

$$A_\sigma^\top Z_\sigma A_\sigma - Z_\sigma \prec 0; \tag{4}$$

b. if $\theta_w[k] = \theta_q[k]$, $y_q[k] = y_p[k]$, i.e.,

$$\mathcal{Q} \triangleq \mathcal{W}^{-1}. \quad \triangleleft \tag{5}$$

Remark 2.2. If Z_σ in (4) is the same for all $\theta_\sigma[k], k \in \mathbb{N}, \sigma \in \{g, h, w, q\}$, the mWM filters, on their own, are stable under arbitrary switching, as they all share a common Lyapunov function. ◁

Definition 2.3 (*Zhou et al. (1996, Lemma 3.15)*)**.** Define the DT transfer function resulting from the system defined by the tuple (A, B, C, D) as $G(z) = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$, and suppose that D^{-1} exists. Then

$$G^{-1}(z) = \left[\begin{array}{c|c} A - BD^{-1}C & BD^{-1} \\ \hline -D^{-1}C & D^{-1} \end{array}\right] \tag{6}$$

is the inverse transfer function of $G(z)$. ◁

Assumption 2.3. The mWM parameters are matched, i.e., $\theta_w[k] = \theta_q[k]$ and $\theta_g[k] = \theta_h[k]$, $\forall\, k \in \mathbb{N}$. ◁

2.3. Attack model

Consider the malicious agent $\mathcal{A}$ located in the CPS as in Fig. 1, capable of tampering with data transmitted between $\mathcal{P}$ and $\mathcal{C}$. Without loss of generality, the injected attacks are modeled as additive signals:

$$\tilde{u}_g[k] \triangleq u_g[k] + \varphi_u[k], \quad \tilde{y}_w[k] \triangleq y_w[k] + \varphi_y[k],$$

where $\varphi_u[k]$ and $\varphi_y[k]$ are actuator and sensor attack signals designed by the adversary $\mathcal{A}$. To properly define our design algorithm in Section 3, an explicit strategy for the attack signals φ_u and φ_y must be defined by the defender. In this paper, we focus on covert attacks (Smith, 2015), which remain undetected for passive diagnosis scheme.

The covert attack strategy, under Assumption 2.4 and 2.5, is as follows: the malicious agent $\mathcal{A}$ chooses $\varphi_u[k] \in \ell_{2e}$ freely, while $\varphi_y[k]$ satisfies:

$$\mathcal{A}: \begin{cases} x_a[k+1] = A_a(\theta^a)x_a[k] + B_a(\theta^a)\varphi_u[k] \\ y_a[k] = C_a(\theta^a)x_a[k] + D_a(\theta^a)\varphi_u[k] \\ \varphi_y[k] = -y_a[k] \end{cases} \tag{7}$$

where $x_a \triangleq [x_{h,a}^\top \;\; x_{p,a}^\top \;\; x_{w,a}^\top]^\top$ is the attacker's state, and its dynamics are the same as the cascade of $\mathcal{H}, \mathcal{P}, \mathcal{W}$, parametrized[3] by θ_σ^a.

Assumption 2.4. For all $k \in [k_{i+1}, k_{i+2}], i \in \mathbb{N}_+$, the attacker parameters $\theta_\sigma^a[k] = \theta_\sigma[k_i]$, $\sigma \in \{h, w\}$. ◁

Assumption 2.5. The attack energy is bounded and finite, i.e.,: $\|\varphi_u\|_{\ell_2}^2 \leq \epsilon_a$, with ϵ_a known to $\mathcal{C}$. ◁

Remark 2.3. Assumption 2.5 is introduced as it allows for guarantees that the algorithm proposed in Section 3 always returns a feasible solution (see Theorem 3.2). In general, while it may be that the adversary has limited energy (Djouadi et al., 2015), it is a strong assumption that the bound ϵ_a is known to the defender. Nonetheless, the attack energy bound ϵ_a may be seen as a design variable that, together with the chosen attack model (7), facilitates the definition of a systematic design of mWM filters by the defender. Further remarks regarding the consequences of Assumption 2.5 not holding are postponed to Remark 2.5, following the formal definition of the attack-energy-constrained output-to-output gain in Definition 2.4. ◁

2.4. Problem formulation

The objective of this paper is to propose a design strategy capable of optimally designing the mWM filter parameters θ^+, supposing a covert attack is present within the CPS. To formulate a metric to be used to define optimality, the closed-loop CPS dynamics must be defined. Under the attack strategy (7), the closed-loop system with the attack φ_u as input and the performance and detection output as system outputs can be rewritten as

$$\mathcal{S}: \begin{cases} x[k+1] = Ax[k] + B\varphi_u[k] \\ y_J[k] = \bar{C}_J x[k] + \bar{D}_J \varphi_u[k] \\ y_r[k] = \bar{C}_r x[k] \end{cases} \tag{8}$$

where $x = \left[x_p^\top, \; x_h^\top, \; x_g^\top, \; x_c^\top, \; x_q^\top, \; x_w^\top, \; x_a^\top\right]^\top$ is the closed-loop system state, while y_r and y_J remain the residual and performance outputs. All signals in (8) are also a function of the parameters θ^+, but this dependence is dropped for clarity. The definition of the matrices in (8) follow from (1)–(3) and (7).

The defender aims to quantify (and later minimize) the maximum performance loss caused by a stealthy and bounded-energy

[2] In the sequel whenever referring to the parameters of any one of the mWM filters, the subscript σ is used. Conversely, if referring to all parameters, θ is used.

[3] Here, and throughout the paper, a super- or subscript a is used to indicate that a variable pertains to $\mathcal{A}$.

adversary on (8). This is done by exploiting the attack-energy-constrained output-to-output gain (AEC-OOG) (Anand & Teixeira, 2023).

Definition 2.4 (*AEC-OOG*)**.** The AEC-OOG of $\mathcal{S}$ in (8) is the value of the following optimization problem:

$$\begin{aligned} \sup_{\varphi_u \in \ell_{2e}} \quad & \|y_J\|_{\ell_2}^2 \\ \text{s.t.} \quad & \|y_r\|_{\ell_2}^2 \leq \epsilon_r, \ \ \|\varphi_u\|_{\ell_2}^2 \leq \epsilon_a, \ \ x[0] = 0. \end{aligned} \tag{9}$$

where ϵ_a is the energy bound of the attack signal, ϵ_r is the detection threshold, and the value of (9) denotes the performance loss caused by a stealthy adversary. ◁

Problem 1. Given θ at some switching time k_i, $i \in \mathbb{N}_+$, find the optimal set of mWM filter parameters after a switching event θ^+, such that the AEC-OOG of the system $\mathcal{S}$ in (8) is minimized. ◁

Remark 2.4. Because of its dependence on the AEC-OOG, the solution of Problem 1 at time k_i relies explicitly on the attack parameters $\theta^a[k_i]$. Given the malicious agent's strategy outlined in Section 2.3, and Assumption 2.4, $\theta^a[k_i] = \theta^+[k_{i-1}]$ is known to $\mathcal{C}$, without any additional knowledge required. ◁

Remark 2.5. We are now ready to formally treat the violation of Assumption 2.5. To do this, let us first remark on some properties of the AEC-OOG, which follow from using finite bounds ϵ_r and ϵ_a. Firstly, as will be demonstrated in Theorem 3.2, the metric is always bounded, making it well suited for a design algorithm. Furthermore, it is explicitly related to both the H_∞ metric and the original OOG proposed in Teixeira, Sandberg et al. (2015), for increasing values of ϵ_r and ϵ_a, respectively (Anand & Teixeira, 2023, Prop. 1). Finally, we can comment on the constraint on the attack energy. Consider the value of (9) under increasing values of ϵ_a, as well the OOG as defined in Teixeira, Sandberg et al. (2015). If the OOG is finite, there is some value $\bar{\epsilon}_a$ such that the AEC-OOG is the same as the OOG for all $\epsilon_a \geq \bar{\epsilon}_a$. If there are exploitable zero dynamics, and the OOG is unbounded, $\|y_J\|_{\ell_2}^2$ grows unbounded as $\epsilon_a \to \infty$. Thus, while θ_σ^+, the solution to Problem 1, is only optimal for covert attacks satisfying $\|\varphi_u\|_{\ell_2}^2 \leq \epsilon_a$, it ensures that the effect of φ_u on y_J is in some sense minimal if the attack energy constraint is violated. ◁

3. Optimal design of filters

3.1. Design problem

As summarized in Problem 1, the objective of the parameter design is to *minimize the maximum performance loss caused by the adversary*. This can be translated, exploiting (9), to the following optimization problem

$$\inf_{\theta^+} \mathcal{L}(\theta^+, \theta^a) \tag{10}$$

$$\mathcal{L}(\theta^+, \theta^a) = \begin{cases} \sup\limits_{\varphi_u \in \ell_{2e}} \ \|y_J\|_{\ell_2}^2 \\ \quad \text{s.t. } \|y_r\|_{\ell_2}^2 \leq \epsilon_r, \ \ \|\varphi_u\|_{\ell_2}^2 \leq \epsilon_a, \\ \qquad x[0] = 0, (4), (5), (7), (8) \end{cases} \tag{11}$$

In (10), $\mathcal{L}(\theta^+, \theta^a)$ represents the value of the maximum performance loss caused by the adversary for any given pair of filters (θ^+, θ^a), and therefore is a suitable metric for the defender's objectives. The optimization problem (10) is an infinite-dimensional optimization problem in signal space. Using Anand and Teixeira (2023, Lem. 3.1, Lem. 3.2), (10) is converted to an equivalent, finite-dimensional, non-convex optimization problem in Lemma 3.1.

Lemma 3.1. *The infinite-dimensional optimization problem (10) is equivalent to the following finite-dimensional, non-convex optimization problem*

$$\begin{aligned} \inf_{P, \gamma, \gamma_a, \theta^+, Z_\sigma} \quad & \epsilon_r \gamma + \epsilon_a \gamma_a \\ \text{s.t.} \quad & R + \begin{bmatrix} \bar{C}_J^\top \bar{C}_J - \gamma \bar{C}_r^\top \bar{C}_r & \bar{C}_J^\top \bar{D}_J \\ \bar{D}_J^\top \bar{C}_J & \bar{D}_J^\top \bar{D}_J - \gamma_a I_m \end{bmatrix} \preceq 0 \\ & (4), (5), \gamma \geq 0, \gamma_a \geq 0, P \succeq 0, Z_\sigma \succ 0, \end{aligned} \tag{12}$$

where $R \triangleq \begin{bmatrix} A^\top P A - P & A^\top P B \\ B^\top P A & B^\top P B \end{bmatrix}$. □

Finding a solution to (12) solves Problem 1, as solving for θ^+ achieves the minimal worst-case impact of a covert attack satisfying Assumption 2.4. Although (12) is convex in P, γ and γ_a, it contains non-convex terms in A. As such, it cannot be easily solved via standard convex solvers (Lofberg, 2004).

3.2. Well-posedness of the impact metric (11)

Differently to our previous results (Gallo et al., 2021), using the AEC-OOG ensures that the optimization problem used for the design of the mWM parameters is always feasible, as summarized in the following.

Theorem 3.2. *Let the parameters θ^+ be chosen such that (4), (5), and Assumption 2.2 hold. Then, the value of the metric $\mathcal{L}$ in (11) is bounded if the closed-loop matrix A in (8) is Schur stable.* □

Proof. Let $\Sigma_J \triangleq (A, B, \bar{C}_J, \bar{D}_J)$ be the closed loop system from the attack input φ_u to the performance output y_J. The objective is to show that the value of (11) is bounded given Assumption 2.2, and for any given value of θ^+ that satisfies (4) and (5). To this end, start by considering the optimization problem (11) without the constraint $\|y_r\|_{\ell_2}^2 \leq \epsilon_r$. The value of the resulting optimization problem is the H_∞ gain of the system Σ_J, which is bounded, so long as Σ_J is stable. Thus, (11) is bounded, as the optimal value of any maximization problem cannot increase under additional constraints. ■

Note that the condition of the closed-loop matrix A being stable is required only at any given time $k \in \mathbb{N}$, and not under switching. The problem of ensuring A is stable under switching is addressed in Ferrari and Teixeira (2021, Thm. 3).

3.3. Filter parameter update algorithm

As mentioned previously, the optimization problem (12) is non-convex and cannot be solved exactly. One approach to solve (12) is to reformulate the problem with Bi-linear Matrix inequalities (BMI) and use some existing approaches in the literature to solve them (e.g., Dehnert et al. (2021), Dinh et al. (2011), Gallo et al. (2021), etc.), which however come with drawbacks. In light of this, here an exhaustive search algorithm, defined in Algorithm 1, is adopted, to show the main advantage of the proposed design problem (12).

The exhaustive search algorithm we propose can be sketched out as follows. Let the values of all matrices be chosen *a priori*, apart from A_σ, such that they satisfy (5). Thus, the objective is to find optimal values of A_σ minimizing (12). Furthermore, to ensure tractability, let us restrict the matrices A_σ to be diagonal. To guarantee stability of the watermark generating matrices, it is sufficient to constrain the diagonal elements to lie in $(-1, 1)$. Discretizing this set into a grid of n_s elements, sets $\mathcal{A}_h$ and $\mathcal{A}_q$ are obtained, with cardinality $n_s^{n_h}$ and $n_s^{n_q}$, respectively. Thus, the exhaustive search algorithm searches for optimal matrices A_h, A_q, under the constraint (5). The complete algorithm is summarized in Algorithm 1, where the final step provides an ordering, in case multiple parameters obtain the same optimum.

Algorithm 1. Filter parameters selection algorithm.

Initialization: $K, L, \theta, \theta^a, \gamma^* = \infty$
Result: $\theta^{+,*}$

1: Pick random matrices D_h and D_q.
While $((\mathcal{A}_h \neq \emptyset)||(\mathcal{A}_q \neq \emptyset))$, **do:**
2: Draw a matrix A_h from $\mathcal{A}_h$ and delete it from $\mathcal{A}_h$.
3: If the inverse of A_h : A_g obtained from (5) is unstable go to step **2**.
4: Draw a matrix A_q from $\mathcal{A}_q$ and delete it from $\mathcal{A}_q$.
5: If the inverse of A_q : A_w obtained from (5) is unstable go to step **4**.
6: Derive the inverse filters using (5).
7: Using the values of the watermarking filters, and θ^a, solve the convex optimization problem (12). Let us denote the value of (12) as γ_t.
8: If $\gamma_t < \gamma^*$, update $\gamma^* = \gamma_t$ and store the values of watermarking parameters, else go back to step (1).
end While

3.4. Randomizing the solution

Until now, the design of the algorithm has been purely deterministic: given the parameters θ, (12) uniquely determines the parameters θ^+. This provides optimal results, but it makes the architecture vulnerable to attacks[4] capable of identifying the mWM filter parameters, as the attacker can compute future values of θ by solving Algorithm 1. We therefore propose a method to counteract this vulnerability. Specifically, by initializing matrices D_q and D_h randomly in the first step of the algorithm, it can be shown that the resulting parameters θ^+ are also random.

Theorem 3.3. *Let $D_q[k_i], D_h[k_i]$ be the matrices defined in Step 1 of Algorithm 1 at switching times $k_i, i = 0, 1, \ldots \in \mathbb{N}_+$. It is sufficient to select $D_q[k_i] \neq D_q[k_j]$ and $D_h[k_i] \neq D_h[k_j]$ to ensure that $\theta^+[k_i] \neq \theta^+[k_j]$, $\forall i, j = 0, 1, \ldots \in \mathcal{N}_+, i \neq j$.* □

Proof. The proof follows directly from the fact that, for any two state space realizations (A_1, B_1, C_1, D_1) and (A_2, B_2, C_2, D_2) with compatible dimensions, it is sufficient for $D_1 \neq D_2$ for the resulting transfer functions $G_1(z) \neq G_2(z)$ (Chen, 1984, Thm. 4.1). As a consequence, so long as $D_h[k_i] \neq D_h[k_j]$ and $D_q[k_i] \neq D_q[k_j]$, there are no mWM parameters such that the resulting closed loop transfer functions are the same. ■

Corollary 3.3.1. *Let $D_h[k_i], D_q[k_i]$ be realizations of random variables. Then, the filter parameters $\theta^+[k_i]$ are also randomized.* □

To ensure that the parameters θ^+ remain synchronous, it is necessary for the randomized values of D_h, D_q be the same on both plant and control side. The problem of selecting variables that are synchronized and (pseudo)random is a common issue in the secure control literature, and different solutions have been found, such as Zhang et al. (2022), Zhang et al. (2023).

[4] The attacker in question is different to that defined in Section 2.3 where the attack strategy was considered as a *design choice* for the formulation of the optimization problem.

Table 1
System Parameters.

K_{lm}	1	T_{lm}	6	T_g	0.2
T_h	4	T_s	0.1	R	0.05

4. Numerical example

4.1. Plant description

Consider a power generating system (Park et al., 2019, Sec. 4) modeled by the dynamics:

$$\begin{bmatrix}\dot{\eta}_1\\ \dot{\eta}_2\\ \dot{\eta}_3\end{bmatrix} = \begin{bmatrix}\frac{-1}{T_{lm}} & \frac{K_{lm}}{T_{lm}} & \frac{-2K_{lm}}{T_{lm}}\\ 0 & \frac{-2}{T_h} & \frac{6}{T_h}\\ \frac{-1}{T_g R} & 0 & \frac{-1}{T_g}\end{bmatrix} \underbrace{\begin{bmatrix}\eta_1\\ \eta_2\\ \eta_3\end{bmatrix}}_{\eta} + \begin{bmatrix}0\\ 0\\ \frac{1}{T_g}\end{bmatrix} u \tag{13}$$

$$y_p = \underbrace{\begin{bmatrix}1 & 0 & 0\end{bmatrix}}_{C_p}\eta, \quad y_J = \underbrace{\begin{bmatrix}0 & 1 & 0\end{bmatrix}}_{C_J}\eta. \tag{14}$$

Here, $\eta \triangleq [df; dp + 2dx; dx]$, where df is the frequency deviation in Hz, dp is the change in the generator output per unit (p.u.), and dx is the change in the valve position p.u. The parameters of the plant are listed in Table 1. The Discrete-Time system matrices (A_p, B_p, C_p, D_p) are obtained by discretizing the plant (13)–(14) using zero-order hold with a sampling time $T_s = 0.1$s.

The plant is stabilized locally with a static output feedback controller with constant gain $D_c = 19$. The gains in (2) are obtained by minimizing a quadratic cost, using the MATLAB command *dlqr*, resulting in:

$$K = \begin{bmatrix}0.1986 & -0.0913 & -0.1143\end{bmatrix}$$
$$L = \begin{bmatrix}0.2735 & -0.0509 & -0.2035\end{bmatrix}^\top.$$

4.2. Initializing the mWM design algorithm

We consider a mWM filter of state dimension $n_\sigma = 2$. The mWM filter parameters are initialized as $A_q = 0.2I_2, B_q = 0.7e_{2\times1}$, $C_q = 0.1e_{1\times2}, B_h = 0.2e_{2\times1}, C_h = 0.05e_{1\times2}, A_h = 0.3I_2, D_q = 0.15$, $D_h = 0.1$ where $e_{a\times b}$ represents a unit matrix of size $a \times b$. The other mWM matrices are derived such that they satisfy (5). All unspecified matrices are zero. Following Assumption 2.4, it is assumed that the filter parameters θ are known by the adversary. To ensure randomization, as mentioned in Theorem 3.3, the parameters D_h and D_q are initialized in Algorithm 1 as random numbers within the range $[0.1, 0.15]$. We fix the parameters of all the mWM filter parameters at their initial value except for the matrix $A_\sigma, \sigma \in \{q, w, h, g\}$, i.e., our aim is to find a diagonal A_σ that minimizes the value of the AEC-OOG.

As discussed in Section 3.3, A_q and A_h are matrices whose diagonal elements take values in $(-1, 1)$. The exhaustive search is performed with a grid size of $n_s = 0.3$, and the search is initialized with $\epsilon_r = 1, \epsilon_a = 50$. Furthermore, for numerical stability, we modify the objective function of (12) to $\epsilon_r\gamma + \epsilon_a\gamma_a + \epsilon_p \text{tr}(P)$, with $\epsilon_p = 0.1$.

4.3. Result of algorithm 1

The optimal value of the matrices from the grid search are $A_q^* = -0.05I_2$ and $A_h^* = -0.65I_2$. The corresponding value of $\mathcal{L}$ is 111.03. The value of D_q and D_h were 0.1479 and 0.1482 respectively. The simulation is performed using Matlab 2021a with *Yalmip* (Lofberg, 2004) and *SDPT3v4.0* solver (Toh et al., 2012). In the remainder, we compare the results obtained by repeated

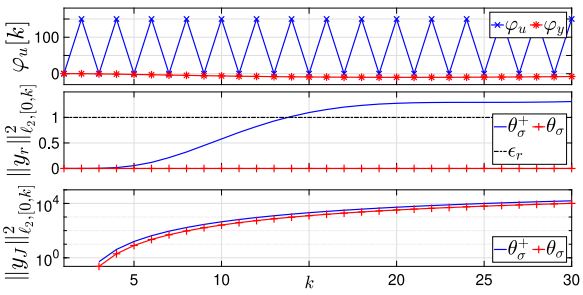


Fig. 2. (Top) The attack signal φ_u in (15) and its equivalent φ_y from (7); (Middle) $\|y_r\|^2_{\ell_2,[0,k]}$, compared to ϵ_r; (Bottom) $\|y_J\|^2_{\ell_2,[0,k]}$ before and after the mWM parameters are updated.

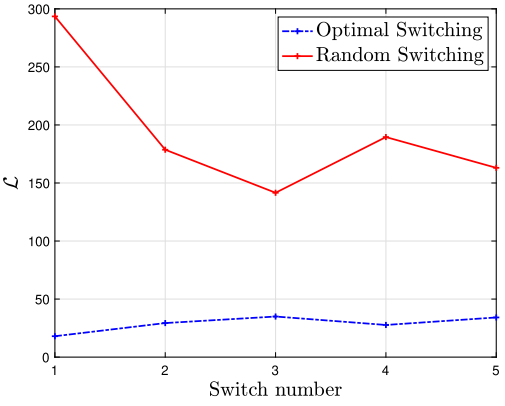


Fig. 3. The values of $\mathcal{L}$ corresponding to the optimal and random values of the watermarking parameters.

computation of Algorithm 1 compared to defining constant and random parameters. Consider an adversary injecting the signals shown in Fig. 2,

$$\varphi_u[k] = \begin{cases} 150 & \text{if } k \bmod 2 = 0 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

into the actuators, and φ_y following (7).

Comparison with no parameter switching: The performance of the attack is shown in Fig. 2, for the cases without switching and when switching happens at the attack onset with the optimal filter parameters. Without switching θ, although the performance is strongly degraded, the attack remains stealthy. Instead, if the mWM parameters are changed, it is detected after 15s.

Comparison with random parameter switching: In this scenario, we suppose the mWM parameters are updated 5 times, by running Algorithm 1, and compared against 5 random updates of A_σ — though their structure remains diagonal. The results, shown in terms of values of $\mathcal{L}$ for both cases, are shown in Fig. 3. Here, the parameters of D_h and D_q are the same as used for selecting the optimal parameters. Since the parameters are not chosen optimally, the value of $\mathcal{L}$, the performance loss, is higher.

Time complexity: To conclude, let us discuss the complexity of Algorithm 1. All mWM parameters are fixed, apart from A_σ, which is a diagonal matrix of dimension n_σ, and, for each diagonal element of A_σ, n_s points of the interval $(-1, 1)$ are searched. Given (5), only A_h and A_q must be defined, while A_w, A_g are defined algebraically; thus, define $n_\varsigma = n_q + n_h$. The complexity of the algorithm grows both in n_ς and in n_s. Specifically, the complexity is $\mathcal{O}(n_s^{n_\varsigma})$. Thus, it is exponential in the choice of n_ς and polynomial in n_s. We highlight that the average time of solution can be improved upon in two major ways. The first is via parallelization, as all SDPs can be solved independently; this provides a speed-up which depends on the number of compute nodes used to solve the problem. The second method relies on reducing the number of SDPs to be solved, by removing those values of A_h, A_q which do not lead to stable inverses, as defined by (6).

For the results presented here, a computer with an Intel Core i7-6500U CPU with 2 cores and 8 GB RAM was used. The algorithm was run both with and without parallelization (parallelization was achieved by using Matlab's `parfor` command). Without parallelization, the algorithm took 384.25s to provide a result, whilst with parallelization this was 261.65s, a 31.9% speedup.

5. Conclusion and future works

An optimal design technique for the design of the parameters of switching multiplicative watermarking filters is presented. The problem is formalized by supposing the closed-loop system is subject to a covert attack with matching parameters. We propose an optimal control problem based on a formulation of the attack energy constrained output-to-output gain. We show through a numerical example that this design improves detectability by increasing the energy of the residual output before and after a switching event. Future works include developing algorithms for optimal design and optimal switching times ensuring that mWM does not destabilize the closed-loop system under switching with mismatched parameters, and studying non-linear systems.

References

Abdalmoaty, M. R., Anand, S. C., & Teixeira, A. M. H. (2023). Privacy and security in network controlled systems via dynamic masking. *IFAC-PapersOnLine*, *56*(2), 991–996.

Alisic, R., Kim, J., & Sandberg, H. (2023). Model-free undetectable attacks on linear systems using LWE-based encryption. *IEEE Control Systems Letters*, *7*, 1249–1254.

Anand, S. C., & Teixeira, A. M. H. (2023). Risk-based security measure allocation against actuator attacks. *IEEE Open Journal of Control Systems*, *2*, 297–309.

Chen, C.-T. (1984). *Linear system theory and design*. Saunders college publishing.

Darup, M. S., Alexandru, A. B., Quevedo, D. E., & Pappas, G. J. (2021). Encrypted control for networked systems: An illustrative introduction and current challenges. *IEEE Control Systems Magazine*, *41*(3), 58–78.

Dehnert, R., Lerch, S., Grunert, T., Damaszek, M., & Tibken, B. (2021). A less conservative iterative LMI approach for output feedback controller synthesis for saturated discrete-time linear systems. In *2021 25th international conference on system theory, control and computing* (pp. 93–100). IEEE.

Dinh, Q. T., Gumussoy, S., Michiels, W., & Diehl, M. (2011). Combining convex–concave decompositions and linearization approaches for solving BMIs, with application to static output feedback. *IEEE Transactions on Automatic Control*, *57*(6), 1377–1390.

Djouadi, S. M., Melin, A. M., Ferragut, E. M., Laska, J. A., Dong, J., & Drira, A. (2015). Finite energy and bounded actuator attacks on cyber-physical systems. In *2015 European control conference* (pp. 3659–3664). IEEE.

Escudero, C., Murguia, C., Massioni, P., & Zamaï, E. (2023). Safety-preserving filters against stealthy sensor and actuator attacks. In *2023 62nd IEEE conference on decision and control* (pp. 5097–5104). IEEE.

Fang, S., Johansson, K. H., Skoglund, M., Sandberg, H., & Ishii, H. (2019). Two-way coding in control systems under injection attacks: From attack detection to attack correction. In *Proceedings of the 10th ACM/IEEE international conference on cyber-physical systems* (pp. 141–150).

Ferrari, R. M. G., & Teixeira, A. M. H. (2021). A switching multiplicative watermarking scheme for detection of stealthy cyber-attacks. *IEEE Transactions on Automatic Control*, *66*(6), 2558–2573. http://dx.doi.org/10.1109/TAC.2020.3013850.

Gallo, A. J., Anand, S. C., Teixeira, A. M., & Ferrari, R. M. (2021). Design of multiplicative watermarking against covert attacks. In *2021 60th IEEE conference on decision and control* (pp. 4176–4181). IEEE.

Griffioen, P., Weerakkody, S., & Sinopoli, B. (2020). A moving target defense for securing cyber-physical systems. *IEEE Transactions on Automatic Control*, *66*(5), 2016–2031.

Hashemi, N., & Ruths, J. (2022). Codesign for resilience and performance. *IEEE Transactions on Control of Network Systems*, *10*(3), 1387–1399.
Hemsley, K. E., & Fisher, R. E. (2018). History of industrial control system cyber incidents. http://dx.doi.org/10.2172/1505628, URL https://www.osti.gov/biblio/1505628.
Hu, S., & Yan, W.-Y. (2007). Stability robustness of networked control systems with respect to packet loss. *Automatica*, *43*(7), 1243–1248.
Lin, Y., Chong, M. S., & Murguia, C. (2023). Secondary control for the safety of LTI systems under attacks. *IFAC-PapersOnLine*, *56*(2), 965–970.
Lofberg, J. (2004). YALMIP: A toolbox for modeling and optimization in MATLAB. In *2004 IEEE international conference on robotics and automation (IEEE Cat. No. 04CH37508)* (pp. 284–289). IEEE.
Mo, Y., Weerakkody, S., & Sinopoli, B. (2015). Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs. *IEEE Control Systems Magazine*, *35*(1), 93–109.
Murguia, C., Shames, I., Ruths, J., & Nešić, D. (2020). Security metrics and synthesis of secure control systems. *Automatica*, *115*, Article 108757.
Park, G., Lee, C., Shim, H., Eun, Y., & Johansson, K. H. (2019). Stealthy adversaries against uncertain cyber-physical systems: Threat of robust zero-dynamics attack. *IEEE Transactions on Automatic Control*, *64*(12), 4907–4919.
Sandberg, H., Gupta, V., & Johansson, K. H. (2022). Secure networked control systems. *Annual Review of Control, Robotics, and Autonomous Systems*, *5*, 445–464.
Smith, R. S. (2015). Covert misappropriation of networked control systems: Presenting a feedback structure. *IEEE Control Systems Magazine*, *35*(1), 82–92.
Stabile, F., Lucia, W., Youssef, A., & Franzè, G. (2024). A verifiable computing scheme for encrypted control systems. *IEEE Control Systems Letters*.
Teixeira, A., Sandberg, H., & Johansson, K. H. (2015). Strategic stealthy attacks: the output-to-output ℓ_2-gain. In *2015 54th IEEE conference on decision and control* (pp. 2582–2587). IEEE.
Teixeira, A., Shames, I., Sandberg, H., & Johansson, K. H. (2015). A secure control framework for resource-limited adversaries. *Automatica*, *51*, 135–148.
Toh, K.-C., Todd, M. J., & Tütüncü, R. H. (2012). On the implementation and usage of SDPT3–a Matlab software package for semidefinite-quadratic-linear programming, version 4.0. In *Handbook on Semidefinite, Conic and Polynomial Optimization* (pp. 715–754). Springer.
Zhang, J., Gallo, A. J., & Ferrari, R. M. (2023). Hybrid design of multiplicative watermarking for defense against malicious parameter identification. In *2023 62nd IEEE conference on decision and control* (pp. 3858–3863). IEEE.
Zhang, K., Kasis, A., Polycarpou, M. M., & Parisini, T. (2022). A sensor watermarking design for threat discrimination. *IFAC-PapersOnLine*, *55*(6), 433–438.
Zhou, K., Doyle, J. C., Glover, K., et al. (1996). *Robust and optimal control, Vol. 40*. Prentice hall New Jersey.
Zhu, H., Liu, M., Fang, C., Deng, R., & Cheng, P. (2023). Detection-performance tradeoff for watermarking in industrial control systems. *IEEE Transactions on Information Forensics and Security*, *18*, 2780–2793.



Alexander J. Gallo received the M.Eng. in Electrical and Electronic Engineering from Imperial College London, London, UK, in 2016. He received his Ph.D. degree in Control Engineering from Imperial College London, London, UK, in 2021. From 2021 to 2024 he was a postdoctoral researcher at the Delft Center for Systems and Control, Delft University of Technology, Delft, the Netherlands. Since Sep 2024, he is a postdoctoral researcher at the Politecnico di Milano, Milan, Italy.

His main research interests include distributed cyber-security and fault tolerant control for large-scale interconnected systems, with a particular focus on energy distribution networks, as well as health-aware and fault tolerant control of wind turbines.



Sribalaji C. Anand received an M.Sc. degree in Systems and Control from Delft University of Technology, The Netherlands, in 2019, and a Ph.D. degree in Automatic Control from Uppsala University, Sweden, in 2024. He is currently a Postdoctoral Researcher at KTH Royal Institute of Technology, Stockholm, Sweden. His research interests include secure and scalable control as well as convex optimization. In 2024, he was awarded the International Postdoctoral Grant from the Swedish Research Council.



Andre M.H. Teixeira received the M.Sc. degree in electrical and computer engineering from the Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, in 2009, and the Ph.D. degree in automatic control from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2014. He is an Associate Professor at the Department of Information Technology, Uppsala University, Sweden. Before this, he was an Associate Senior Lecturer at the Department of Electrical Engineering, Uppsala University (2017–2021). From 2015 to 2017, he was an Assistant Professor at the Faculty of Technology, Policy and Management, Delft University of Technology. Dr. Teixeira was awarded a Starting Grant by the Swedish Research Council in 2019, and the Future Research Leaders grant by the Swedish Foundation for Strategic Research in 2020. He received the Lilly and Sven Thuréus prize in 2023 from The Royal Society of Sciences at Uppsala. In 2023, the Knut and Alice Wallenberg Foundation appointed him as a Wallenberg Academy Fellow. Dr. Teixeira serves as Associate Editor for Automatica. His research interests include secure and resilient control systems, distributed anomaly detection, distributed optimization, and power systems.



Riccardo M.G. Ferrari received the Laurea degree (Cum Laude and printing honours) in Electronic Engineering in 2004 and the Ph.D. degree in Information Engineering in 2009, both from University of Trieste, Italy. He received the 2005 Giacomini Award from the Italian Acoustic Society and placed second in the IFAC 2011 Competition on Fault Detection and Fault Tolerant Control for Wind Turbines. He also won an Honorable Mention for the Pauk M. Frank prize at IFAC SAFEPROCESS 2018 and an Airbus Award at IFAC 2020 for the best contribution to Aerospace Industrial Fault Detection. He is co-recipient of the O. Hugo Schuck Best Paper Award at ACC 2023.

He has held both academic and industrial R&D positions, in particular as researcher in the field of process instrumentation and control for the steelmaking sector. He is a Marie Curie alumnus and currently an Associate Professor with the Delft Center for Systems and Control, Delft University of Technology, The Netherlands. His research interests include wind power fault tolerant control and fault diagnosis and attack detection in large-scale cyber–physical systems, with applications to electric vehicles, cooperative autonomous vehicles and industrial control systems.