Received: 24 May 2022 | Revised: 14 July 2022 | Accepted: 7 August 2022

DOI: 10.1002/aic.17875

RESEARCH ARTICLE

Process Systems Engineering

AIChE JOURNAL

A control-switching approach for cyberattack detection in process systems with minimal false alarms

Shilpa Narasimhan | Nael H. El-Farra | Matthew J. Ellis

Department of Chemical Engineering, University of California, Davis, California, USA

Correspondence
Matthew J. Ellis, Department of Chemical Engineering, University of California, Davis, CA 95616, USA.
Email: mjellis@ucdavis.edu

Funding information
UC Davis College of Engineering

Abstract

The frequency of cyberattacks against process control systems has increased in recent years. This work considers multiplicative false-data injection attacks involving the multiplication of the data communicated over the sensor-controller communication link by a factor. An active detection method utilizing switching between two control modes is developed to balance the trade-off between closed-loop performance and attack detectability. Under the first mode, the control parameters are selected using traditional control design criteria. Under the second mode, the control parameters are selected to enhance the attack detection capability. A switching condition is imposed to prevent false alarms that could be triggered by the transient response induced by control mode switching. This condition is incorporated into the active detection method to minimize false alarms. The active detection method is applied to illustrative process examples to demonstrate its ability to detect attacks and minimize false alarms.

KEYWORDS
active cyberattack detection, cyberattack detectability, multiplicative false-data injection cyberattacks

1 | INTRODUCTION

Process control systems (PCSs) are industrial control systems that operate many continuous production processes, including chemical manufacturing processes. Because of the tight profit margins and inherently hazardous nature of chemical processes, PCSs may be the target of cyber-attackers aiming to disrupt operations. Recent attacks have demonstrated that cyberattackers can target control systems by circumventing the traditional information technology (IT) infrastructure-based cybersecurity measures.[1,2] This trend has inspired research focusing on the operational technology (OT) to improve the cybersecurity and cyberattack resilience of PCSs.[3,4]

Cyberattacks on PCSs can take many forms, and comprehensive studies have focused on cyberattack taxonomy.[5–8] Attackers may compromise a PCS by hijacking controller software to control the execution of a PCS computing device,[9] or by maliciously tampering with the operational data of the PCS to compromise the data integrity.[7,8,10] Attacks may be designed to alter the historical data, adversely affecting the decisions and analyses that rely on using this data (e.g., maintenance scheduling and forensic analysis for uncovering the presence of past cyberattacks).[10] Attacks may also impact online manufacturing operations by targeting PCS communication channels. Denial of service (DoS)[8] and false-data injection attacks[7] are two such attacks that target the PCS communication channels. DoS attacks prevent data from being communicated over a network by bombarding the network with spurious requests,[11] while false-data injection attacks alter data communicated over the network.[12]

To address cyberattacks from an OT perspective, incorporating cyberattack resilience into the PCS design has been another focus. Cyberattack resilience refers to the ability of the PCS to deter, detect, identify, and recover from a cyberattack. To address cyberattack resilence through control system design, several approaches have been proposed to handle different types of attacks. For example, an attacker may attempt to learn the behavior of the closed-loop process

before designing an attack that accomplishes the attacker's goals. To make it difficult for an attacker to learn the closed-loop behavior, a control design utilizing a randomized controller switching to prevent an attacker from learning the controller behavior was developed.[13] For the recovery of a process under power-constrained DoS attacks, an event-triggered communication scheme and resilient observer-based control was proposed.[11] The switching observer adapts the state estimate generation to the DoS attack. A co-design methodology for selecting the triggering parameters and the control and observer gains was presented. A machine learning approach for estimating cyberattack severity and mitigating its impact on closed-loop stability was developed for nonlinear processes.[14] To detect and recover from certain kinds of false-data injection attacks, a detection scheme using machine learning-based detectors and attack mitigation using control switching, redundant sensors, control reconfiguration, and post-cyberattack reconstruction of states were proposed.[15–17]

An important part of the cyberattack resiliency of PCSs is the ability to detect the presence of a cyberattack, and many detection schemes have been proposed.[15–26] Attack detection schemes may be broadly classified as either passive or active. Passive detection schemes monitor a process using regular operational data. Several passive detection schemes have been proposed, including schemes that use the residual (defined as the difference between the measured output and its estimate),[20–22] neural network-based schemes,[15–17] and a control barrier function-based scheme.[19]

Active detection methods utilize an external intervention or perturbation to enable cyberattack detection. For example, active detection methods utilizing secret watermarking signals added to the sensor or actuator signals[23–26] and moving target schemes[23,24] have been proposed. Under a moving target scheme, an auxiliary system with time-varying dynamics is added to the process to prevent the attacker from learning the process dynamics. Another example includes active detection methods utilizing a control system switching to probe a process for attacks.[27] Multiplicative false-data injection cyberattacks are modeled by a factor multiplied by the data communicated over a controller communication channel. These attacks require minimal process-specific knowledge for their design to evade detection. As a result, active detection methods for multiplicative attacks have received some attention.[25–27]

One technique that has been considered for handling cyberattacks is control system switching.[11,13,16,17,19,27,28] In Reference 19, control law switching was proposed for the recovery of the attacked process after an attack is detected. In References 16 and 17, the control system switches between two operational modes, (i.e., the closed-loop mode and the open-loop mode) to maintain the state within a secure set when a sensor-controller link cyberattack is detected. In the open-loop mode, measurements of the state received from the sensors are not utilized to compute control actions. For cyberattack detection, a randomized controller switching is used to probe the process for an attack and to make it difficult for an attacker to inject false state measurements while remaining undetected.[28] In previous work,[27] an active detection method utilizing occasional controller–observer parameter switching to enhance the detection capabilities of a residual-based detection scheme was presented. However, switching controller–observer parameters may excite the process dynamics and induce transients. As a result, false alarms may be generated in the detection scheme monitoring the attack-free process.

This work develops an active detection method utilizing controller–observer parameter switching with minimal false alarms. As part of the proposed active detection method, the control system operates under two modes. Under the first mode, called the nominal mode, the control system operates with controller–observer parameters selected using traditional control design criteria. Under the second mode (the "attack-sensitive" mode), the control system operates with controller–observer parameters selected to enhance the detection capability of an output and residual-based detection scheme. The active detection method manages the trade-off between closed-loop performance and attack detectability. Since switching may excite the process dynamics, generating false alarms, a state-dependent switching condition that guarantees zero false alarms is developed using a region containing the attack-free process states, called the confidence region. Practical implementation issues related to the active detection method are discussed, including the inability to ensure that the switching condition will be satisfied over the time interval it is desired to switch the control system. Switching between the nominal and attack-sensitive modes may be desirable even if the switching condition is not satisfied. A modified active detection method for minimizing false alarms that incorporates the switching condition while balancing the practical requirement to switch between modes is proposed. The application of the proposed active detection method in attack detection and minimizing false alarms is demonstrated using two illustrative process examples.

2 | PRELIMINARIES

2.1 | Notation

For an n-dimensional vector $x \in \mathbb{R}^n$, $\left\|x\right\| := \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}$ represents the Euclidean norm. For the compact set $X \subset \mathbb{R}^n$, $AX := \{Ax | x \in X\}$, where A is a matrix. The Minkowski sum of two sets, $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^n$ is represented by $X \oplus Y = \{x + y | x \in X, y \in Y\}$. The Minkowski difference of two compact and convex sets, $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^n$ is represented by $X \ominus Y = \{x - y | x \in X, y \in Y\}$. For a square matrix A, $\lambda_i(A)$ represents the i^{th} eigenvalue of A. diag(α_1, α_2, α_3, ..., α_n) represents an $n \times n$ diagonal matrix with diagonal elements α_1, α_2, α_3, ..., α_n.

2.2 | Class of processes

In this work, discrete-time linear time-invariant processes are considered:

$$x(t+1) = Ax(t) + Bu(t) + Gw(t) \tag{1}$$

where $x(t) \in \mathbb{R}^{n_x}$ is the process state vector, $u(t) \in \mathbb{R}^{n_u}$ is the manipulated input vector, and $w(t) \in W \subset \mathbb{R}^{n_w}$ is the bounded process disturbance vector. The set W is assumed to be known and described by a

convex polytope containing the origin. The measured output ($y(t)$) is subject to measurement noise and may be altered by a multiplicative sensor-controller link attack:

$$y(t) = \Lambda(Cx(t) + v(t)) \tag{2}$$

where $y(t) \in \mathbb{R}^{n_x}$ is the measured output vector and $v(t) \in V \in \mathbb{R}^{n_x}$ is the bounded measurement noise vector. The set V is assumed to be known and described by a convex polytope containing the origin. The matrix C is assumed to be invertible. The matrix $\Lambda \in \mathbb{R}^{n_x \times n_x}$ is used to model the multiplicative sensor-controller link attack on the process and is called the attack magnitude. When $\Lambda = I$, the process is attack-free. Without loss of generality, the origin of the unforced process (Equation 1 with $u \equiv 0$ and $w \equiv 0$) is assumed to the desired operating steady-state.

A Luenberger observer is synthesized to estimate the process states in Equations (1) and (2):

$$\hat{x}(t+1) = A\hat{x}(t) + Bu(t) + L(y(t) - \hat{y}(t)) \tag{3a}$$

$$\hat{y}(t) = C\hat{x}(t) \tag{3b}$$

where $\hat{x}(t) \in \mathbb{R}^{n_x}$ is the state estimate generated by the observer, $\hat{y}(t) \in \mathbb{R}^{n_x}$ is the estimated output, and $L \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ is the observer gain selected such that the eigenvalues of the matrix $A - LC$ are strictly within the unit circle. To stabilize the closed-loop process, a linear control law utilizing the state estimate is synthesized:

$$u(t) = -K\hat{x}(t) \tag{4}$$

where $K \in \mathbb{R}^{n_u} \times \mathbb{R}^{n_x}$ is the controller gain selected such that the eigenvalues of the matrix $A - BK$ are strictly within the unit circle. The estimation error is defined as the difference between the process state and the state estimate, that is, $e := x - \hat{x}$, with dynamics given by:

$$e(t+1) = L(I - \Lambda)Cx(t) + (A - LC)e(t) + Gw(t) - L\Lambda v(t) \tag{5}$$

To analyze the stability of the closed-loop process under an attack, an augmented state vector is defined as a concatenation of the state and the error vectors $\xi := [x^T\ e^T]^T$. The augmented state dynamics are described by:

$$\xi(t+1) = \underbrace{\begin{bmatrix} (A-BK) & BK \\ L(I-\Lambda)C & (A-LC) \end{bmatrix}}_{=: A_\xi(\Lambda, K, L)} \xi(t) + \underbrace{\begin{bmatrix} G & 0 \\ G & -L\Lambda \end{bmatrix}}_{=: B_\xi(\Lambda, L)} d(t) \tag{6}$$

where $d(t) := [w^T(t)\ v^T(t)]^T \in F$ and $F := \{[w^T\ v^T]^T | w \in W, v \in V\}$. Due to persistent bounded disturbances acting upon the process, the closed-loop process is continuously perturbed, and the augmented state never converges to the origin. Instead, the augmented state of the closed-loop process is ultimately bounded within a small neighborhood of the origin when $\max_i |\lambda_i(A_\xi(\Lambda, K, L))| < 1$. This neighborhood of the origin is the minimum invariant set of the process and can be expressed as the infinite Minkowski sum[29]:

$$D_\xi(\Lambda, K, L) = \oplus_{i=0}^{\infty} A_\xi(\Lambda, K, L)^i B_\xi(\Lambda, L) F \tag{7}$$

From Equation (6), the matrices $A_\xi(\Lambda, K, L)$ and $B_\xi(\Lambda, L)$ are dependent on the attack magnitude (Λ) and the controller–observer parameters (K, L). Consequently, the minimum invariant set in Equation (7) is dependent on the attack magnitude, the controller–observer parameters, and the disturbance set. For simplicity of presentation, the closed-loop process in Equation (1) operated with the control input in Equation (4) computed based on the state estimates and with the controller gain K and the observer gain L, is referred to as the closed-loop process with (K, L).

3 | CLASS OF ATTACK DETECTION SCHEMES

The detectability of an attack on the closed-loop process with (K, L) may be defined with respect to the detection scheme monitoring the process. A general class of detection schemes monitoring the process utilizing a generalized monitoring variable $\eta \in \mathbb{R}^{n_\eta}$ is considered. The generalized monitoring variable may be expressed as a weighted combination of the measured output and the estimate of the measured output:

$$\eta(t) = H_y y(t) + H_{\hat{y}} \hat{y}(t) \tag{8}$$

where H_y and $H_{\hat{y}}$ are matrices of appropriate dimensions. From Equations (2) and (3b), the measured output and its estimate may also be expressed in terms of the augmented state $\xi(t)$ and the process disturbance $d(t)$ as:

$$y(t) = \underbrace{[\Lambda C\ 0]}_{=: A_y(\Lambda)} \xi(t) + \underbrace{[0\ \ \Lambda]}_{=: B_y(\Lambda)} d(t) \tag{9a}$$

$$\hat{y}(t) = \underbrace{[C\ \ -C]}_{=: A_{\hat{y}}} \xi(t) + \underbrace{[0\ \ 0]}_{=: B_{\hat{y}}} d(t) \tag{9b}$$

Thus, Equation (8) may be re-written as:

$$\eta(t) = A_\eta(\Lambda)\xi(t) + B_\eta(\Lambda)d(t) \tag{10}$$

where $A_\eta(\Lambda) = H_y A_y(\Lambda) + H_{\hat{y}} A_{\hat{y}}$ and $B_\eta(\Lambda) = H_y B_y(\Lambda) + H_{\hat{y}} B_{\hat{y}}$.

When the closed-loop process with (K, L) is stable in the sense that all eigenvalues of the matrix $A_\xi(\Lambda, K, L)$ are strictly within the unit circle, the augmented state of the process is ultimately bounded within its minimum invariant set ($D_\xi(\Lambda, K, L)$). Furthermore, because the closed-loop process is subjected to bounded disturbances, the generalized monitoring variable is also bounded within a terminal set, denoted by $D_\eta(\Lambda, K, L)$. From Equation (10), the terminal set of the generalized monitoring variable may be computed by:

$$D_\eta(\Lambda, K, L) = A_\eta(\Lambda) D_\xi(\Lambda, K, L) \oplus B_\eta(\Lambda) F \tag{11}$$

The generalized monitoring variable is bounded within its attack-free terminal set, that is, $\eta(t) \in D_\eta(I, K, L)$ for all time $t \geq 0$

if $\xi(0) \in D_\xi(I, K, L)$ because $D_\xi(I, K, L)$ is an invariant set, that is, $\xi(t) \in D_\xi(I, K, L)$ for all time $t \geq 0$ if $\xi(0) \in D_\xi(I, K, L)$. The class of detection schemes considered in this work monitor the process for attacks by verifying the containment of the generalized monitoring variable within its attack-free terminal set:

$$h(\eta(t)) = \begin{cases} 0, & \eta(t) \in D_\eta(I, K, L) \\ 1, & \text{Otherwise} \end{cases} \tag{12}$$

where the mapping $h : \mathbb{R}^{n_\eta} \to \{0, 1\}$ returns the output of the detection scheme, with an output value of 1 being indicative of an attack detection, and an output value of 0 being indicative of a lack of attack detection. The approach adopted herein for tuning the general class of detection schemes accounts for all possible values of process disturbances and measurement noise acting on the process. As a result, the tuning approach adopted ensures a zero false alarm rate in the attack-free process.

One example of a measured variable that fits the model for the generalized detection scheme in Equation (8) is the residual, which measures the deviation of the measured output from its estimate:

$$r(t) := y(t) - \hat{y}(t) = \underbrace{\begin{bmatrix}(\Lambda - I)C & C\end{bmatrix}}_{=: A_r(\Lambda)} \xi(t) + \underbrace{\begin{bmatrix}0 & \Lambda\end{bmatrix}}_{=: B_r(\Lambda)} d(t) \tag{13}$$

Residual-based detection schemes monitor a process utilizing the residual. They are typically used for fault detection[30–32] and have also been extensively explored for attack detection.[21,22,25–27,33] From Equation (13), the residual fits within the model for the generalized monitoring variable in Equation (8), with $H_y = I$, $H_{\hat{y}} = -I$.

In Reference 27, an approach to classify attacks based on their detectability with respect to a residual-based detection scheme of the form in Equation (12) was presented. The detectability-based classification of attacks may be extended to a general class of detection schemes of the form in Equation (12) utilizing a monitoring variable of the form in Equation (10). With respect to a class of detection schemes in Equation (12) utilizing a generalized monitoring variable of the form in Equation (10), an attack is said to be detected at time t_d if $\eta(t_d) \notin D_\eta(I, K, L)$ with the output of the detection scheme $h(\eta(t_d)) = 1$. An attack is defined as a detectable attack with respect to the detection scheme in Equation (12) if the attack is detected in finite time (for all $\xi(0) \in \mathbb{R}^{2n_x}$ and $d(t) \in F$ for $t \geq 0$). An attack is defined as an undetectable attack with respect to the detection scheme in Equation (12) if the generalized monitoring variable for the attacked closed-loop process satisfies $\eta(t) \in D_\eta(I, K, L)$ for all $t \geq 0$ for all $\xi(0) \in D_\xi(\Lambda, K, L)$ and $d(t) \in F$ for all $t \geq 0$. Finally, an attack is defined as potentially detectable with respect to the detection scheme in Equation (12) if the attack is neither detectable nor undetectable.

Typically, attack detection schemes using the residual as a monitoring variable have been considered in the literature.[21,22,25–27,33] However, monitoring both the measured output and the residual may be beneficial for the detection of attacks. For example, an attack ($\Lambda \neq I$) may be undetectable with respect to a residual-based detection scheme with $D_r(\Lambda, K, L) \subseteq D_r(I, K, L)$. However, the attack may be potentially detectable with respect to an output-based detection scheme $D_y(\Lambda, K, L) \not\subseteq D_y(I, K, L)$. As a result, the attack may not be detected by the residual-based detection scheme, but the output-based detection scheme may detect the attack. Similarly, attacks that are undetectable with respect to an output-based detection scheme may be detected by a residual-based detection scheme. In the present work, a detection scheme of the form of Equation (12) monitoring the process using an output and residual-based monitoring variable defined as a concatenation of the measured output and the residual ($\chi := \begin{bmatrix} y^T & r^T \end{bmatrix}^T$) is considered. The monitoring variable $\chi(t) \in \mathbb{R}^{2n_x}$ fits the model for the generalized monitoring variable in Equation (8) with $H_y = \begin{bmatrix} I \\ I \end{bmatrix}$ and $H_{\hat{y}} = \begin{bmatrix} 0 \\ -I \end{bmatrix}$. Therefore, the detectability-based classification of attacks is valid for an output and residual-based detection scheme of the form:

$$h(\chi(t)) = \begin{cases} 0, & \chi(t) \in D_\chi(I, K, L) \\ 1, & \text{Otherwise} \end{cases} \tag{14}$$

where $D_\chi(I, K, L)$ is the terminal set of the output and residual-based monitoring variable χ for the attack-free process. $D_\chi(I, K, L)$ may be computed using Equation 11.

4 | ACTIVE DETECTION METHOD

In this section, the proposed switching-enabled active detection method is presented. A rigorous analysis is employed to develop a switching condition to minimize false alarms.

4.1 | Controller switching for active detection

From the detectability-based classification of attacks, controller–observer parameters, selected to meet standard design criteria, may mask some sensor-controller link multiplicative attacks in the sense that attacks are undetectable with respect to the detection scheme in Equation (14). The controller–observer parameters selected based on standard design criteria are called the nominal controller–observer parameters and are denoted by (K^*, L^*). Other controller–observer parameters may not mask the attacks, making the attacks potentially detectable or detectable with respect to the detection scheme. For the attack-free process, using other controller–observer parameters may lead to performance degradation relative to the closed-loop performance achieved under the nominal controller–observer parameters. Occasional switching between the nominal controller–observer parameters and other controller–observer parameters may be a way to balance the potential trade-off between closed-loop performance and attack detectability. Controller–observer parameter switching is an active detection method because switching probes for multiplicative attacks. The second set of controller–observer parameters is selected to be "sensitive" to attacks over a range of magnitudes, meaning that a range of multiplicative attacks destabilizes the closed-loop process, rendering the attacks detectable. These controller–

observer parameters are called attack-sensitive parameters and are denoted by (K_Λ, L_Λ). The dwell-time under the attack-sensitive controller–observer parameters manages the trade-off between attack detection and performance degradation and is denoted by T_c.

The terminal set of the monitoring variable under the attack-sensitive controller–observer parameters is different from the set under the nominal controller–observer parameters. To account for this difference in terminal sets, a time-dependent tuning strategy is used for the detection scheme in Equation (14):

$$h(\chi(t)) = \begin{cases} 0, & \chi(t) \in D_\chi(I, K(t), L(t)) \\ 1, & \text{Otherwise} \end{cases} \tag{15}$$

where $(K(t), L(t)) = (K_\Lambda, L_\Lambda)$ for $t \in (t_s, t_s + T_c]$, $(K(t), L(t)) = (K^*, L^*)$ otherwise, t_s denotes the time instance that the control system switches from the nominal controller–observer parameters to the attack-sensitive controller–observer parameters, and $t_s^* = t_s + T_c$ denotes the time instance that the control system switches from the attack-sensitive controller–observer parameters back to the nominal controller–observer parameters.

Remark 1. The attack-sensitive controller–observer parameters are selected such that undetectable multiplicative sensor-controller link attacks under the nominal controller–observer parameters are rendered detectable under the attack-sensitive parameters. However, finding one pair of controller–observer parameters that renders all attacks detectable may not be possible. Additionally, some attacks that are undetectable under the nominal controller–observer parameters may result in minimal performance deterioration when compared to that under attack-free conditions. Therefore, performance-based selection criteria could be employed to determine the attack-sensitive controller–observer parameters. Multiple attack-sensitive controller–observer parameter pairs may be selected and used to cover a wide range of attacks.

Remark 2. For the practical selection of the attack-sensitive controller–observer parameters, a finite set of attacks should be considered. For example, a subclass of multiplicative sensor-controller link attacks may be considered where the attack magnitude may be modeled by a diagonal matrix $(\Lambda = \text{diag}(\alpha_1, ..., \alpha_{n_x}))$ and α_i represents the magnitude of the multiplicative attack targeting the ith sensor-controller link. For this subclass of attacks, a finite set of attacks generated by considering a range of values for α_i for each i and $\alpha_j = 1$ for $j \neq i$. Knowledge of prior attacks or attacks that are critical to detect may also be employed for generating the set of attacks for the attack-sensitive parameter selection. The attack detectability under the nominal controller–observer parameters may be verified for each attack to generate a set of undetectable attacks. The resulting set of attacks may be further refined by considering a performance-based criterion. Specifically, the set of attacks may be refined to consider attacks that are such that the radius of the minimum bounding ball of the terminal set of states of the attacked process is greater than (or much greater than) the radius of the minimum bounding ball of the terminal set of states for the attack-free process, that is, $R(D_x(\Lambda, K^*, L^*)) > R(D_x(I, K^*, L^*))$ where $R(D_x(\Lambda, K^*, L^*)) := \max_{x' \in D_x(\Lambda, K^*, L^*)} \left\| x' \right\|$ and $D_x(\Lambda, K^*, L^*) = [I \;\; 0] D_\xi(\Lambda, K^*, L^*)$.

4.2 | Confidence region-based switching condition for zero false alarms

Under the proposed active detection method the control system switches between two modes of operation: the nominal mode under which the process is operated with nominal controller–observer parameters, and the attack-sensitive mode under which the process is operated with the attack-sensitive controller–observer parameters. In the attack-free process under the nominal mode, no false alarms are expected due to the tuning approach adopted for the detection scheme in Equation (14). However, switching the control system operating mode on the attack-free process may cause the augmented state to evolve outside the minimum invariant set under the controller–observer parameters for the new mode, potentially resulting in false alarms. For example, consider that the control system switches from the nominal to the attack-sensitive mode at time t_s. If $\xi(t_s) \notin D_\xi(I, K_\Lambda, L_\Lambda)$ (this occurs when $\xi(t_s) \in D_\xi(I, K^*, L^*) \setminus D_\xi(I, K_\Lambda, L_\Lambda)$), the augmented state will evolve outside $D_\xi(I, K_\Lambda, L_\Lambda)$ for some time as it converges to $D_\xi(I, K_\Lambda, L_\Lambda)$. The variable χ during this period may be outside its terminal set $(\chi(t) \notin D_\chi(I, K_\Lambda, L_\Lambda)$ for some $t \geq t_s)$, generating false alarms.

The detection objective of the active detection method is to determine if the process is under an attack, or if it is attack-free. False alarms complicate this determination. False alarms may be avoided if the control system switches when the augmented state is in the minimum invariant set under the controller–observer parameters for the new mode. However, the augmented state is not measured directly, so the exact value of the augmented state is unknown. Instead, a region in the augmented state-space containing the augmented state of the attack-free closed-loop process may be constructed to address this issue. This region is time-dependent and can be computed online from the disturbance set (F), the measured output, and the residual. The region is called the confidence region and is denoted by $\Xi(K, L, t)$, highlighting the time and the controller–observer parameter pair (K, L) dependence. Based on its definition, the vector $\chi(t)$ may be expressed in terms of the augmented state and disturbance (for the attack-free process), as:

$$\chi(t) = \underbrace{\begin{bmatrix} C & 0 \\ 0 & C \end{bmatrix}}_{=: \widetilde{C}} \xi(t) + \underbrace{\begin{bmatrix} 0 & I \\ 0 & I \end{bmatrix}}_{=: \widetilde{D}} d(t) \tag{16}$$

From Equation (16), the confidence region can be computed by:

$$\Xi(K,L,t)=\widetilde{C}^{-1}\left(\{\chi(t)\}\ominus\widetilde{D}F\right) \tag{17}$$

The matrix $\widetilde{C}$ is invertible because C is invertible.

A few properties are established to develop a switching condition that, when satisfied, leads to zero false alarms from control system switching. First, the relationship between the confidence region, the augmented state, and the minimum invariant set for the attack-free process is established.

Proposition 1. Consider the attack-free closed-loop process with (K, L). If the matrix C is invertible and $\xi(0) \in D_\xi(I, K, L)$, then the confidence region $\Xi(K, L, t)$ contains the augmented state, that is, $\xi(t) \in \Xi(K, L, t)$. Furthermore, the confidence region has a nonempty intersection with the minimum invariant set, that is, $\Xi(K, L, t) \cap D_\xi(I, K, L) \neq \emptyset$.

Proof. This proposition is proved in two parts. In the first part, the containment of the augmented state within the confidence region is considered. In the second part, the intersection of the confidence region with the attack-free minimum invariant set is considered.

Part 1: From Equations (16) and (17),

$$\begin{aligned}\Xi(K,L,t)&=\widetilde{C}^{-1}\left(\{\chi(t)\}\ominus\widetilde{D}F\right)\\&=\left(\left\{\widetilde{C}^{-1}\widetilde{C}\xi(t)\right\}\oplus\left\{\widetilde{C}^{-1}\widetilde{D}d(t)\right\}\ominus\widetilde{C}^{-1}\widetilde{D}F\right)\\&=\{\xi(t)\}\oplus\left\{\widetilde{C}^{-1}\widetilde{D}d(t)\right\}\ominus\widetilde{C}^{-1}\widetilde{D}F\end{aligned} \tag{18}$$

for the attack-free process. Because the process disturbances and measurement noise are bounded within the compact set (F) containing the origin, the origin is contained in the set $\left\{\widetilde{C}^{-1}\widetilde{D}d(t)\right\}\ominus\widetilde{C}^{-1}\widetilde{D}F$. Therefore, the right-hand side of Equation (18) contains the augmented state of the attack-free process, and the confidence region constructed at any time $t \geq 0$ contains the augmented state, that is, $\xi(t) \in \Xi(K, L, t)$.

Part 2: If the augmented state of the attack-free process at time $t = 0$ is contained within its minimum invariant set, then due to the forward invariance of the minimum invariant set, the augmented state is contained within the set for all time, that is, $\xi(t) \in D_\xi(I, K, L)$ for all $t \geq 0$. From the proof of Part 1, the confidence region constructed for the attack-free process at any time contains the augmented state ($\xi(t) \in D_\xi(I, K, L)$). Therefore, $\Xi(K, L, t)$ and $D_\xi(I, K, L)$ both contain the augmented state $\xi(t)$, and have a nonempty intersection, that is, $\Xi(K, L, t) \cap D_\xi(I, K, L) \neq \emptyset$ □.

From Equation (16), the confidence region is computed under the assumption that the process is attack-free, and therefore, the augmented state will be contained in the confidence region of the attack-free process. If the process is under a cyberattack, the confidence region does not give any information about the value of the augmented state. However, if the confidence region does not intersect the attack-free minimum invariant set, the process cannot be attack-free, because of an inconsistency between the computation of the confidence region for the attack-free process and the expected evolution of the attack-free process state within the minimum invariant set. In this regard, the confidence region may be another mechanism for detecting attacks. In particular, an attack can be declared if the confidence region and the minimum invariant set do not intersect. This is formally stated in the following proposition.

Proposition 2. Consider the closed-loop process with (K, L). Let the matrix C be invertible and $\xi(0) \in D_\xi(I, K, L)$. If the confidence region does not intersect with the minimum invariant set of the attack-free closed-loop process, that is, $\Xi(K, L, t) \cap D_\xi(I, K, L) = \emptyset$, then the process is not attack-free.

Proof. This proposition is proved by contradiction. Assume that the closed-loop process is attack-free. From the proof of Part 1 of Proposition 1, $\xi(t) \in \Xi(K, L, t)$ at any time $t \geq 0$. If the confidence region does not intersect with the minimum invariant set of the process, that is, $\Xi(K, L, t) \cap D_\xi(I, K, L) = \emptyset$, the minimum invariant set cannot contain the augmented state of the process, that is, $\xi(t) \notin D_\xi(I, K, L)$. This is a contradiction, since, for the attack-free process, the augmented state is always contained within its minimum invariant set, that is, $\xi(t) \in D_\xi(I, K, L)$ if $\xi(0) \in D_\xi(I, K, L)$. Thus, the process cannot be attack-free □.

Proposition 2 provides a confidence region-based condition that may be verified to monitor a process for attacks. However, the motivation behind constructing the confidence regions is to ensure zero false alarms from a switch between any two controller–observer parameter pairs (K_1, L_1) and (K_2, L_2). To ensure zero false alarms, the augmented state at the switching instance of the attack-free process must be within the attack-free minimum invariant sets under both controller–observer parameters. Based on this, the following theorem leverages the result of the Proposition 1 to establish a condition that, if satisfied at the time instance when the controller–observer parameters switch between (K_1, L_1) to (K_2, L_2), guarantees that zero false alarms are generated in the detection scheme in Equation (15). This further implies that any alarms generated are the result of an attack.

Theorem 1. *Consider the closed-loop process with* (K_1, L_1). *Let the matrix* C *be invertible and* $\xi(0) \in D_\xi(I, K_1, L_1)$. *Assume that a controller–observer parameter switch from* (K_1, L_1) *to* (K_2, L_2) *occurs at* t_s. *If the closed-loop process is attack-free and the confidence region satisfies* $\Xi(K_1, L_1, t_s) \cap D_\xi(I, K_1, L_1) \subseteq D_\xi(I, K_2, L_2)$, *then no alarms are generated by the detection scheme of the form in Equation (15).*

Furthermore, if there is an alarm generated by the detection scheme at some time t_d, then the closed-loop process is not attack-free.

Proof. The proof is divided into two parts. In the first part, the attack-free process is considered. In the second part, the generation of an alarm is considered.

Part 1: Because $D_\xi(I, K_1, L_1)$ is a forward invariant set for the attack-free closed-loop process with (K_1, L_1), for $t \in [0, t_s]$, the augmented state of the attack-free process is contained within $D_\xi(I, K_1, L_1)$. From Proposition 1, the augmented state of the attack-free process is contained within the intersection of the confidence region and the minimum invariant set, that is, $\xi(t) \in \Xi(K_1, L_1, t) \cap D_\xi(I, K_1, L_1)$ for $t \in [0, t_s]$ when the matrix C is invertible. If the intersection of the confidence region at t_s and the minimum invariant set with (K_1, L_1) is a subset or equal to the minimum invariant set of the attack-free process with (K_2, L_2), that is, $\Xi(K_1, L_1, t_s) \cap D_\xi(I, K_1, L_1) \subseteq D_\xi(I, K_2, L_2)$, the augmented state at t_s is contained within the minimum invariant set of the attack-free process with (K_2, L_2), that is, $\xi(t_s) \in D_\xi(I, K_2, L_2)$. For this case, $\xi(t) \in D_\xi(I, K_2, L_2)$ for $t \geq t_s$ owing to the invariance of $D_\xi(I, K_2, L_2)$.

The value of the monitoring variable $\chi(t)$ will be within the corresponding terminal set for all $t \geq 0$. In particular, $\chi(t) \in D_\chi(I, K_1, L_1)$ for $t \in [0, t_s]$ and $\chi(t) \in D_\chi(I, K_2, L_2)$ for $t \geq t_s$ by construction of the sets $D_\chi(I, K_1, L_1)$ and $D_\chi(I, K_2, L_2)$. Hence, no alarms are generated with the detection scheme in Equation 15 for the attack-free process if

$$\Xi(K_1, L_1, t_s) \cap D_\xi(I, K_1, L_1) \subseteq D_\xi(I, K_2, L_2) \tag{19}$$

Part 2: Consider the interval $[0, t_s]$ and let $\xi(0) \in D_\xi(I, K_1, L_1)$. If an alarm is generated for any $t_d \in [0, t_s]$, the value of the monitoring variable is not within its terminal set, that is, $\chi(t_d) \notin D_\chi(I, K_1, L_1)$. By construction of $D_\chi(I, K_1, L_1)$, the closed-loop process is not attack-free. The attack is detected at t_d.

The remaining part is proved by contradiction. Specifically, consider the case that no alarms are raised for all $t \in [0, t_s]$. Let a parameter switch from (K_1, L_1) to (K_2, L_2) occur at $t_s \geq 0$ when the confidence region satisfies $\Xi(K_1, L_1, t_s) \cap D_\xi(I, K_1, L_1) \subseteq D_\xi(I, K_2, L_2)$. Assume that the process is attack-free for all $t \geq 0$. Let an alarm be generated at some time $t_d \geq t_s$, implying that the value of the monitoring variable at the time t_d is not in the terminal set of the attack-free closed-loop process with (K_2, L_2), that is, $\chi(t_d) \notin D_\chi(I, K_2, L_2)$. When $\Xi(K_1, L_1, t_s) \cap D_\xi(I, K_1, L_1) \subseteq D_\xi(I, K_2, L_2)$, the process is attack-free, and $\xi(0) \in D_\xi(I, K_1, L_1)$, the monitoring variable evolves according to $\chi(t) \in D_\chi(I, K_1, L_1)$ for $t \in [0, t_s]$ and $\chi(t) \in D_\chi(I, K_2, L_2)$ for $t \geq 0$ by Part 1. Hence, no alarms can be generated. This leads to a contradiction. The closed-loop process is not attack-free when an attack is detected at any $t_d \geq 0$, $\xi(0) \in D_\xi(I, K_1, L_1)$, and $\Xi(K_1, L_1, t_s) \cap D_\xi(I, K_1, L_1) \subseteq D_\xi(I, K_2, L_2)$ □.

These results provide insight into how to design a confidence region-based switching condition. To implement the active detection method without false alarms, a switching condition can be imposed at each switch. When the control system switches from the nominal mode to the attack-sensitive mode at t_s, the confidence region should satisfy

$$\Xi(K^*, L^*, t_s) \cap D_\xi(I, K^*, L^*) \subseteq D_\xi(I, K_\Lambda, L_\Lambda) \tag{20}$$

When the control system switches from the attack-sensitive mode back to the nominal mode at $t_s^* = t_s + T_c$, the confidence region should satisfy

$$\Xi(K_\Lambda, L_\Lambda, t_s^*) \cap D_\xi(I, K_\Lambda, L_\Lambda) \subseteq D_\xi(I, K^*, L^*) \tag{21}$$

4.3 | Minimizing false alarms

In prior work,[27] an active detection method utilizing a time-triggered control system switching approach was presented. Under a time-triggered switching approach, the switching instance t_s and the dwell-time T_c are predetermined. However, process disturbances and measurement noise affect the evolution of the augmented state. At t_s and t_s^*, the desired switching conditions in Equations (20) and (21), respectively, may not be satisfied. Also, the existence of t_s and t_s^* when Equations (20) and (21) are satisfied cannot be guaranteed in general. To minimize false alarms, a state-dependent control system switching approach is utilized in the present work. Specifically, an interval of switching times is defined, over which the desired switching condition is verified. If the switching condition is satisfied, the control system switch occurs. In this sense, the switching times may be considered to be state-dependent. If the condition is not satisfied, the operator may choose to force the switch to occur or reschedule it.

For the switch from the nominal mode to the attack-sensitive mode, an interval is defined and is denoted by $[t_i, t_f]$ where $t_i \geq 0$ and $t_f > t_i$ are lower and upper bounds of the interval, respectively. Beginning at t_i, the switching condition in Equation (20) is verified at every time step. If the condition is satisfied at $t_s \in [t_i, t_f]$, the control system switches from the nominal mode to the attack-sensitive mode. If the condition is never satisfied over the interval $[t_i, t_f]$, the process operator has a few options. The operator may choose to force the switch to the attack-sensitive mode to occur at t_f or re-schedule the switch to another time. For scheduling the switch to attack-sensitive mode, several factors could be considered. For example, the interval may be chosen as the time interval when the performance degradation resulting from operating with the attack-sensitive mode is acceptable. If operational considerations allow for an unbounded implementation interval, that is, $t_f \to \infty$, the closed-loop process with (K^*, L^*) may be monitored for an appropriate switching instance over an extended period.

A similar range of switching instances is defined for switching back to the nominal mode. Denoting the minimum and maximum dwell-time under the attack-sensitive mode by $T_c^{\min}$ and $T_c^{\max}$, respectively, the range of switching instances is given by $\left[t_s+T_c^{\min}, t_s+T_c^{\max}\right]$, that is, $T_c \in \left[T_c^{\min}, T_c^{\max}\right]$ and $t_s^* \in \left[t_s+T_c^{\min}, t_s+T_c^{\max}\right]$. Starting at $t_s+T_c^{\min}$, the condition in Equation (21) is checked. If satisfied at t_s^*, the switch is performed. If the condition is never satisfied over the interval, the control system switches back to the nominal mode at $t_s+T_c^{\max}$, to minimize the performance degradation. However, false alarms are possible in this case. For the selection of the switching interval, operating the process with the attack-sensitive mode for as long as possible may be desirable from an attack detection perspective. However, limiting the dwell-time under the attack-sensitive mode may be desirable to limit performance degradation. Thus, $T_c^{\min}$ and $T_c^{\max}$ manage the trade-off between attack detection and performance degradation. For example, the minimum dwell-time $T_c^{\min}$ may be chosen as the period for which most attacks on the process in the attack-sensitive mode are detected, as demonstrated in the illustrative case study section. Similarly, the maximum dwell-time specifies a limit to the operation in attack-sensitive mode. To this end, $T_c^{\max}$ may be selected as the time of operation in attack-sensitive mode while maintaining process states within a safe set.

Under the proposed active detection method, an operator may choose to force a control system switch at a time when the zero false alarm condition in Equation (19) is not satisfied. In the event of a forced control system switch on the attack-free process, false alarms may be generated for a few time steps until the augmented state converges to the minimum invariant set under the updated controller–observer parameters. Therefore, to minimize false alarms, a modification to the detection scheme in Equation (14) may be considered. Under the modified detection scheme, alarms generated after a forced control system switch may be suppressed for a few time steps.

ALGORITHM 1 **The active detection method**

Inputs: $t_i < t_f$, $\Delta_1 < T_c^{min} \leq T_c^{max}$, Δ_2, (K^*, L^*), (K_Λ, L_Λ)
Initialization: $t = t_i$, $t_s = \infty$, $t_s^* = \infty$, $t_d = \infty$, $\Delta(t) = 0$, $(K(t), L(t)) = (K^*, L^*)$
Outputs: t_d, t_s, t_s^*

do
  Receive the measured output $y(t)$ communicated over the sensor-controller link
  Compute the residual $r(t)$ and the confidence region $\Xi(K(t), L(t), t)$ from $\chi(t) = \left[y^T(t)\ r^T(t)\right]^T$
  Monitoring logic
  if $h(\chi(t)) = 1$ *or* $z(t) = 1$ **then**
    if $\Delta(t) = 0$ **then**
      An attack is detected. Set $t = t_d$
      Activate attack identification and mitigation strategies
    else
      Suppress alarms. Set $h(\chi(t)) = 0$ and $z(t) = 0$
  Switching logic
  if $t_s = \infty$ *and* $t \in [t_i, t_f]$ **then**
    if *Eq. 20 is satisfied* **then**
      Switch to attack-sensitive mode. Set $t_s = t$ and $(K(t), L(t)) = (K_\Lambda, L_\Lambda)$
    else if $t = t_f$ **then**
      Switch to attack-sensitive mode. Set $t_s = t$, $(K(t), L(t)) = (K_\Lambda, L_\Lambda)$, and $\Delta(t) = \Delta_1$
  else if $t_s \neq \infty$, $t_s^* = \infty$, *and* $t \in [t_s + T_c^{min}, t_s + T_c^{max}]$ **then**
    if *Eq. 21 is satisfied* **then**
      Switch to nominal mode. Set $t_s^* = t$ and $(K(t), L(t)) = (K^*, L^*)$
    else if $t = t_s + T_c^{max}$ **then**
      Switch to nominal mode. Set $t_s^* = t$, $(K(t), L(t)) = (K^*, L^*)$, and $\Delta(t) = \Delta_2$
  Compute the control action $u(t)$
  Communicate the computed control action to the actuators
  Set $t \leftarrow t+1$, $(K(t+1), L(t+1)) = (K(t), L(t))$, and $\Delta(t+1) = \max\{\Delta(t) - 1, 0\}$
while $t \leq t_s^* + \Delta_2$;

This suppression of alarms is in-line with the standard industry practice of adding a delay timer to the alarm logic of a controller.[34] After the period for suppression of alarms elapses, any alarm generated in the detection scheme in Equation (14) may be considered to be indicative of the detection of an attack.

As part of the proposed active detection method, in addition to the detection scheme in Equation (15), the confidence regions are used to monitor the process for an attack (leveraging the result of Proposition 2):

$$z(t) = \begin{cases} 0, & \Xi(K(t), L(t), t) \cap D_{\xi}(I, K(t), L(t)) \neq \emptyset \\ 1, & \text{Otherwise} \end{cases} \tag{22}$$

where $z(t) \in \{0, 1\}$ is the output of the detection scheme, with an output of 1 being indicative of attack detection, and an output of 0 indicating a lack of attack detection. Algorithm 1 covers the monitoring logic, control system switching logic, and control action computation over a single cycle switching into and out of the attack-sensitive mode under the proposed active detection method.

The algorithm inputs are the time interval for switching into the attack-sensitive mode ($[t_i, t_f]$), the dwell-time range under the attack-sensitive mode ($\left[T_c^{\min}, T_c^{\max}\right]$), the alarm suppression time after a forced switch into the attack-sensitive mode (Δ_1), the alarm suppression time after a forced switch back from attack-sensitive mode (Δ_2), and the nominal controller–observer parameters (K^*, L^*), and the attack-sensitive controller–observer parameters (K_Λ, L_Λ). To perform some computations in the algorithm, additional parameters are needed ($D_\xi(I, K^*, L^*)$, $D_\xi(I, K_\Lambda, L_\Lambda)$, $D_\chi(I, K^*, L^*)$, and $D_\chi(I, K_\Lambda, L_\Lambda)$). These parameters have been omitted for simplicity of presentation. Without loss of generality, the algorithm is activated at time t_i. The algorithm terminates when the control system switches back to the nominal mode or when an attack is detected. If an attack is detected, attack identification and mitigation strategies are activated, albeit a discussion of these strategies is beyond the scope of the current work. The variable $\Delta(t)$ tracks the number of time steps from the time step t that any alarms should be suppressed. To ensure that the switch back into the nominal mode does not occur during the alarm suppression period, the alarm suppression period after a forced switch into attack-sensitive mode is chosen to be less than the minimum dwell-time, that is, $\Delta_1 < T_c^{\min}$. The algorithm outputs are the detection time and the switching instances.

When the algorithm is not active, the process is assumed to be operated and monitored under the nominal mode. The algorithm may be periodically activated, enabling routine cyberattack probing. Additionally, the algorithm may be activated multiple times using different attack-sensitive controller–observer parameters to probe for different attacks. No attacks are assumed to be detected before activating the algorithm because switching into attack-sensitive mode is not needed if an attack is detected before the algorithm is activated.

Remark 3. Considering a controller-observer parameter switch from (K_1, L_1) to (K_2, L_2), a conservative estimate of the alarm suppression time (Δ') for the detection scheme in Equation 14 is the time needed for any realization of the augmented state starting within the minimum invariant set of the process under (K_1, L_1) to converge to the minimum invariant set of the process under (K_2, L_2).

Remark 4. The set of attack magnitudes, that is, the set of values of $\Lambda \neq I$, that may be detected under a given control mode (i.e., attack-sensitive mode or the nominal mode) is the set of potentially detectable or detectable attacks. Since the process model and admissible set of process disturbances and measurement noise are fixed, this set is only dependent on the controller–observer parameters of the active control mode. The set of attack magnitudes that will be detected under a given control mode depends on the controller–observer parameters and other factors, including the dwell-time under the active mode and the realizations of the process disturbance and measurement noise. The set of attacks that may be detected can be numerically approximated by checking the detectability of attacks within a finite set of values, although the accuracy of this approximation may be limited by the number of attack magnitudes considered. However, an explicit characterization of the set of attacks that will be detected is an open problem.

5 | ILLUSTRATIVE CASE STUDIES

In this section, two illustrative processes are considered to demonstrate the application of the active detection method. All polytope computations are performed using the Multi-Parametric Toolbox (MPT 3.0).[35]

5.1 | Application to a scalar process

A scalar process consisting of a single state ($x(t) \in \mathbb{R}$), and a single measured output ($y(t) \in \mathbb{R}$) is considered:

$$x(t+1) = x(t) + u(t) + w(t)$$
$$y(t) = \Lambda(x(t) + v(t))$$

where $u(t) \in \mathbb{R}$ is the manipulated input, $\Lambda \neq 1$ is the magnitude of multiplicative sensor-controller attack, $v(t) \in V := \{v' | v' \in [-5, 5]\}$ represents the vector of bounded measurement noise corrupting the measurements of the state, and $w(t) \in W := \{w' | w' \in [-1, 1]\}$ represents the vector of bounded process disturbances. A Luenberger observer of the form in Equation (3a) is synthesized to generate estimates of states $\hat{x}(t) \in \mathbb{R}$. To stabilize the process at the origin, which is the desired operating steady-state, a linear feedback law of the form Equation (4) is used to compute the control input from the estimates of state. To analyze the stability of the closed-loop process, an

augmented state vector $\xi := [x\, e]^T$ is defined. The closed-loop process is expressed in the form of Equation (6) with

$$A_\xi(\Lambda, K, L) = \begin{bmatrix} (1-K) & K \\ L(1-\Lambda) & 1-L \end{bmatrix}, B_\xi(\Lambda, K, L) = \begin{bmatrix} 1 & 0 \\ 1 & -L\Lambda \end{bmatrix}$$

where K is the controller gain and L is the observer gain.

The nominal controller–observer parameters for the process are chosen as $K^* = 0.1$ and $L^* = 1.9$ to stabilize the attack-free closed-loop process. To detect attacks with magnitudes in the range $\Lambda \in [1.3, 4]$, the attack-sensitive controller–observer parameters for the process are chosen with $K_\Lambda = 1.7$ and $L_\Lambda = 1.5$. The range of attacks that destabilize the closed-loop process under attack-sensitive controller–observer parameters is numerically verified by checking if the value of $\max_i |\lambda_i(A_\xi(\Lambda, K_\Lambda, L_\Lambda))| > 1$ for all $\Lambda \in [1.3, 4]$, by starting at an attack magnitude equal to the lower bound of the range ($\Lambda = 1.3$), and incrementing the magnitudes by 0.01 until the upper bound of the range is reached ($\Lambda = 4$). A similar analysis performed for nominal controller–observer parameters reveals that they are not sensitive to attacks in the interval [1.3, 4]. For the attack-free process under the nominal and the attack-sensitive mode, the radii of the minimum bounding balls containing the terminal set of states are computed as $R(D_x(I, K^*, L^*)) = 15.5263$ and $R(D_x(I, K_\Lambda, L_\Lambda)) = 95$ where $D_x(I, K, L)$ denotes the terminal set of states for the attack-free closed-loop process with parameters (K, L). Defining closed-loop performance with the radius of the minimum bounding ball containing the terminal set, the closed-loop performance under the nominal parameters is better than that under the attack-sensitive parameters.

To monitor the process using a detection scheme of the form of Equation (15), invariant outer approximations of the minimum invariant sets of the attack-free process under the nominal and the attack-sensitive controller–observer parameters are computed as $D_\xi(I, K^*, L^*)$ and $D_\xi(I, K_\Lambda, L_\Lambda)$ using the method described in Reference 36. The error bound used in computing the numerical approximations is $\epsilon = 5 \times 10^{-5}$. Numerical approximations of the sets $D_\chi(I, K^*, L^*)$ and $D_\chi(I, K_\Lambda, L_\Lambda)$ are computed from Equation (11) and shown in Figure 2.

The confidence region constructed using the monitored variable $\chi(t)$ is compared with two other methods for computing the confidence region: one using the measured output and one using the residual. From the measured output, a set containing the process state may be computed by: $X_y(K, L, t) = \{y(t)\} \ominus V$ (for a given controller–observer parameter pair (K, L)). Since the augmented state of the attack-free process is bounded within its minimum invariant set, the estimation error is bounded within its terminal set, computed by: $D_e(I, K, L) := [0\ 1] D_\xi(I, K, L)$. Therefore, the sets $X_y(K, L, t)$ and $D_e(I, K, L)$ are the regions containing the process state and the estimation error. A confidence region constructed using the output alone is given by: $\Xi_y(K, L, t) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} X_y(K, L, t) \oplus \begin{bmatrix} 0 \\ 1 \end{bmatrix} D_e(I, K, L)$. Similarly, from Equation (13), the residual value for the attack-free process depends on the estimation error and the measurement noise. A set containing the estimation error values may be computed by: $E_r(K, L, t) = \{r(t)\} \ominus V$. The terminal set of states may be computed by: $D_x(I, K, L) = [1\ 0] D_\xi(I, K, L)$. Therefore, the confidence region containing attack-free states constructed from the residual alone may be computed by: $\Xi_r(K, L, t) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} D_x(I, K, L) \oplus \begin{bmatrix} 0 \\ 1 \end{bmatrix} E_r(K, L, t)$. Therefore, the confidence region computed from the output and residual-based monitoring variable may be compared with the confidence region computed from the measured output alone and that computed from the residual alone.

A simulation of the attack-free scalar process with the proposed active detection method is considered. For the active detection method, the control system switch from the nominal mode to the attack-sensitive mode is scheduled over the interval $[t_i, t_f] = [250, 400]$. Over this interval, the condition in Equation (20) is verified at each time step. The minimum and the maximum dwell-time under attack-sensitive mode are selected to be $T_c^{\min} = 100$ and $T_c^{\max} = 110$. The process disturbances and measurement noise are modeled as random variables

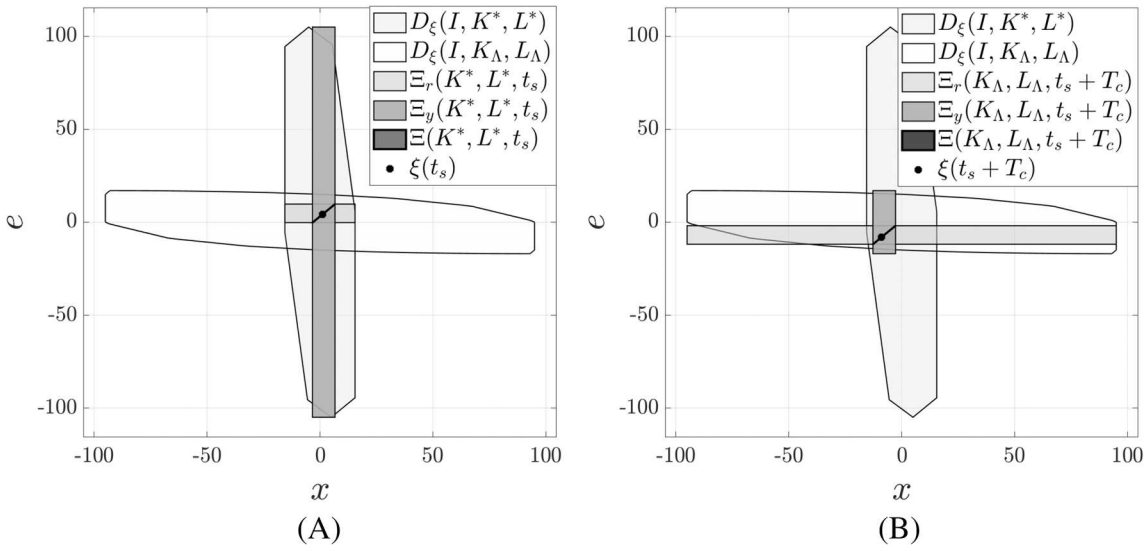

FIGURE 1 (A) Confidence regions for the attack-free process under the nominal mode at the time $t_s = 250$. (B) Confidence regions for the attack-free process under the attack-sensitive mode at the time $t_s = 350$

drawn from uniform distributions at each step and bounded between [−1, 1] and [−5, 5], respectively. The total length of the simulation is 1000 time steps, and the initial condition of the process is 0. To implement a switch, a confidence region computed using the monitoring variable $\chi(t)$ is used to check the appropriate switching condition. Over the simulation, the switch from the nominal mode to the attack-sensitive mode occurs at the time step $t_s = 250$ when the condition in Equation (20) is satisfied (Figure 1A). Similarly, the switch back to nominal mode occurs at the time step $t_s^* = t_s + T_c^{\min} = 350$ when the condition in Equation (21) is satisfied (Figure 1B). No false alarms are observed due to either switch.

For comparison, the confidence regions are computed from the residual and output at both switching instances and are depicted in Figure 1A,B. At both switching instances, the augmented state is contained within the confidence region constructed from the residual and from the output. However, the confidence region computed from the output does not satisfy Equation (19) with $\Xi_y(K^*, L^*, t_s) \cap D_\xi(I, K^*, L^*) \nsubseteq D_\xi(I, K_\Lambda, L_\Lambda)$ at the switching instance $t_s = 250$ (Figure 1A). As a result, the switch may have been prevented if the switching condition is verified based on the confidence region computed from the output. Similarly, the confidence region computed from the residual does not satisfy Equation (19) with $\Xi_r(K_\Lambda, L_\Lambda, t_s^*) \cap D_\xi(I, K_\Lambda, L_\Lambda) \nsubseteq D_\xi(I, K^*, L^*)$, and may have prevented a switch to the attack-sensitive mode at the time t_s^*. Furthermore, when compared to the confidence regions computed from the output and residual-based monitoring variable $\chi(t)$, confidence regions computed from the output or the residual alone are larger regions. Therefore, the confidence region computed from $\chi(t)$ provides a less conservative estimate of the region containing the attack-free augmented state, and is considered in the present work.

Next, the minimization of false alarms in an attack-free process with the proposed active detection method is demonstrated. Two scenarios are considered. The first scenario considers the attack-free process with the proposed active detection method. The second scenario considers the attack-free process with the active detection method, but with a time-triggered control system switching. Each scenario consists of 1000 simulations, where the bounded process disturbances and measurement noise at each time step are drawn from a uniform distribution as described previously. The same realization of the random variables is used in both scenarios to compare across simulations. The initial condition of all simulations is 0, which is contained within the attack-free minimum invariant set under the nominal controller–observer parameters. The total length of each simulation is 1000 time steps.

In the first scenario, the proposed active detection method is applied to the attack-free process. For the active detection method, the algorithm is implemented with a time interval $[t_i, t_f] = [250, 400]$ for a switch from the nominal mode to attack-sensitive mode, and a dwell-time range $T_c^{\min} = 100$ and $T_c^{\max} = 110$ for the switch back from the attack-sensitive mode to nominal mode are used. The alarm suppression period after each control system switch is chosen to be 10-time steps, that is, $\Delta_1 = 10$ and $\Delta_2 = 10$. Over numerous simulations of the attack-free process with a time-triggered switch, the augmented state converges to the minimum invariant set under the new controller–observer parameters within 10-time steps or less. The switch into attack-sensitive mode to probe for attacks is scheduled for [250, 308]. The switch back to the nominal mode is implemented over the interval [350, 412]. The switch back to nominal mode occurred when the condition in Equation (21) is satisfied in 977 of the 100 simulations. Over 23 of the 1000 simulations, the switch back to nominal mode is forced at the time $t_s^* = t_s + T_c^{\max}$ because Equation (21) is not satisfied over the implementation interval. Over the remaining 23 simulations, the augmented state converged to the minimum invariant set under the nominal controller–observer parameters in 10-time steps or less. As a result, no false alarms are observed in the detection scheme in Equation (15).

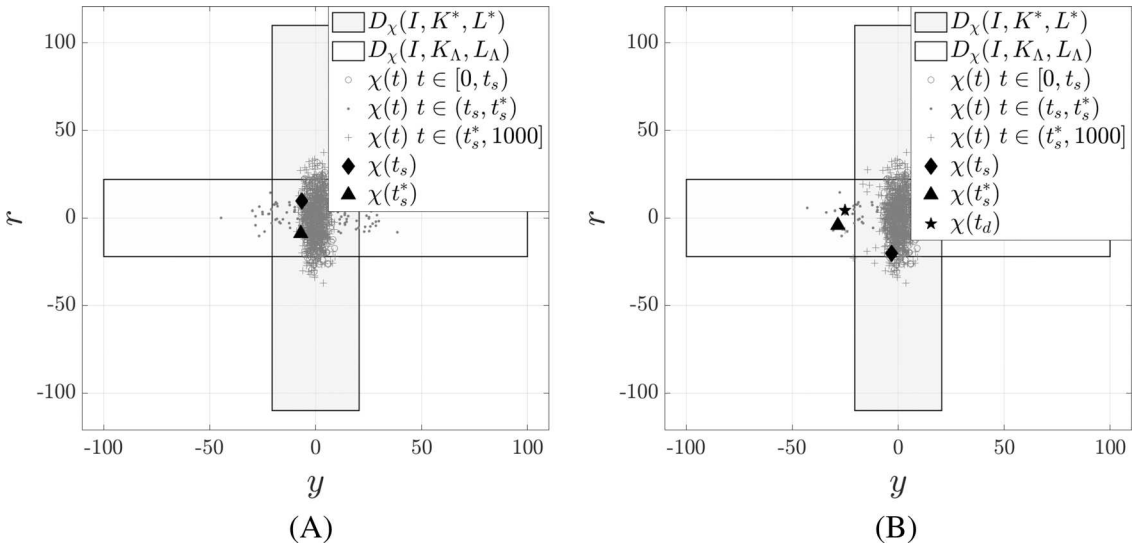

FIGURE 2 (A) Monitoring variable values for the attack-free process with the proposed active detection method. (B) Monitoring variable values for the attack-free process with a time-triggered control system switching

Figure 2A illustrates the monitoring variable values from one simulation. Over this simulation, the monitoring variable values are contained within the terminal set under nominal controller–observer parameters as indicated by the unfilled circular markers in Figure 2A. When the switch into attack-sensitive mode occurs at time step $t_s = 264$, the monitoring variable value is represented by a diamond marker in Figure 2A. After the switch into the attack-sensitive mode, the monitoring variable values are contained within the corresponding terminal set, as indicated by dot markers in Figure 2A. The switch back to nominal mode occurs at the time $t_s^* = 366$, with a monitoring value represented by a triangle marker in Figure 2A. After the switch, the monitoring variable $\chi(t)$ is contained within its corresponding terminal set, as indicated by the "plus" markers in Figure 2A.

In the second scenario, the attack-free process with an active detection method, but with a time-triggered switching strategy, is considered. The switch into attack-sensitive mode occurs at the time $t_s = 250$, and in the absence of an attack detection, a switch back to the nominal mode occurs at the time $t_s^* = 350$. In 1000 simulations of the process under the time-triggered switching strategy, no false alarms are observed after the switch from the nominal to the attack-sensitive mode. In 204 out of 1000 simulations, false alarms are generated in the detection scheme in Equation (15), after switch back to nominal mode. The monitoring variable values over one simulation are illustrated in Figure 2B. As indicated by the unfilled circular markers in Figure 2B, the monitoring variable values are contained within the terminal set under nominal controller–observer parameters until the switch into attack-sensitive mode occurs at the time step $t_s = 264$ (with monitoring variable value represented by a diamond marker in Figure 2B). After the control system switches into attack-sensitive mode, the monitoring variable values are contained within the corresponding terminal set, as indicated by dot markers in Figure 2B. No alarms are observed after switching into attack-sensitive mode at the time step $t_s = 250$. The switch back to the nominal mode occurs at the time $t_s^* = 366$, with a monitoring value represented by a triangle marker in Figure 2B. After the switch, an attack detection (false alarm) is reported by the detection scheme in Equation (14) at the time step $t_d = 351$ (indicated by the filled star marker in Figure 2B). False alarms are observed for up to 2 more time steps, after which the monitoring variable $\chi(t)$ is contained within its corresponding terminal set, as indicated by the "plus" markers in Figure 2B. With the time-triggered switching strategy, false alarms spanning 10 time steps or less are observed in 204 simulations of the process. However, no false alarms are observed over all simulations of the process with the proposed active detection method. Therefore, the proposed active detection method minimizes false alarms from a switch.

A third scenario with the process under an attack of magnitude $\Lambda = 1.3$ with the proposed active detection method is considered to demonstrate enhancement of detection capabilities. The attack is potentially detectable under nominal controller–observer parameters and detectable under attack-sensitive controller–observer parameters. For a basis of comparison, 1000 simulations of the attacked process operated exclusively under the nominal mode are performed. Over 1000 simulations, the attack is not detected by the detection scheme in Equation (14). Next, 1000 simulations of the attacked scalar process with the proposed active detection method are performed. Over 1000 simulations, the attack is detected by the scheme in Equation (15) within a maximum of 47-time steps from the switch into attack-sensitive mode. Thus, the active detection method enhances the detection capabilities of the detection scheme in Equation (14). Additionally, this case study highlights the possible use of monitoring a process using the confidence region-based detection scheme in Equation (22) because the attack is detected by the confidence region-based detection scheme in Equation (22) in all simulations.

5.2 | Application to a chemical process

A chemical process consisting of a well-mixed continuously stirred tank reactor (CSTR) where a second-order, single-phase exothermic reaction of the form $A \rightarrow B$ occurs is considered. The tank liquid may be heated or cooled. Applying standard modeling assumptions, the dynamic process model is obtained from the mass and energy balances around the CSTR liquid hold-up and is given by:

$$\begin{aligned}\frac{dC_A}{dt} &= \frac{F}{V}(C_{A0} + \Delta C_{A0} - C_A) - k_0 e^{\frac{-E}{RT}} C_A^2 \\ \frac{dT}{dt} &= \frac{F}{V}(T_0 + \Delta T_0 - T) - \frac{\Delta H k_0}{\rho C_p} e^{\frac{-E}{RT}} C_A^2 + \frac{Q}{\rho C_p V}\end{aligned} \tag{23}$$

where C_{A0} is the inlet concentration of the reactant, T_0 is the inlet temperature, C_A is the concentration of the reactant in the reactor, T is the temperature of the reactor, and Q is the heat supplied to/removed from the reactor. The definitions and values of the process parameters in Equation (23) are given in Table 1. The manipulated input is Q. The variables ΔC_{A0} and ΔT_0 represent deviations in the feed conditions from the nominal values, C_{A0} and T_0, respectively, and are considered to be bounded process disturbances. The measured variables are the reactant concentration (C_A) and temperature (T), with additive bounded measurement noise. The output matrix ($C = I$) is invertible.

TABLE 1 Model parameters for the continuously stirred tank reactor.[37]

Density	$\rho_L = 1000$ kg m^{-3}
Heat capacity	$C_p = 0.231$ kJ kg^{-1}K $- 1$
Flow rate	$F = 5.0$ m^{-3} h^{-1}
Reactor volume	$V = 1.0$ m^3
Heat of reaction	$\Delta H = -1.15 \times 10^4$ kJ mol^{-1}
Activation energy	$E = 5.0 \times 10^4$ kJ mol^{-1}
Feed temperature	$T_0 = 300.0$ K
Pre-exponential factor	$k_0 = 8.46 \times 10^6$ m^3 kmol^{-1} h^{-1}
Gas constant	$R = 8.314$ kJ mol^{-1} K^{-1}
Concentration of reactant A in the feed	$C_{A0} = 4.0$ kmol m^{-3}

The control objective of the CSTR process is to operate the process at the steady-state corresponding to $C_{As} = 1.22$ kmol m^{-3}, $T_s = 438$ K, and $Q_s = 0$ kW. A state-space model for the process is obtained using the deviation variables $x_1 = C_A - C_{As}$, $x_2 = T - T_s$, and $u = Q - Q_s$, where $x = [x_1\ x_2]^T$ are deviation variables representing the process states, and u is the deviation variable representing the manipulated input. A discrete-time linear model is needed to design the control system and analyze the attack detectability properties. The nonlinear process model is linearized about the steady-state. The resulting continuous-time linear model is discretized assuming zeroth-order hold of the inputs with a sampling period of $\Delta t = 1 \times 10^{-2}$ h. The discrete-time state-space matrices are given by:

$$A = \begin{bmatrix} 0.7364 & -0.0041 \\ 10.6953 & 1.1560 \end{bmatrix}, B = \begin{bmatrix} -9.0708 \times 10^{-8} \\ 4.6741 \times 10^{-5} \end{bmatrix},$$

$$G = \begin{bmatrix} 0.0433 & -0.0001 \\ 0.2724 & 0.0540 \end{bmatrix}$$

For the attack detectability analysis, the algorithm presented in Reference 36 is used to generate outer invariant approximations of the minimum invariant sets for the attack-free closed-loop process. The maximum error of the outer approximations of the minimum invariant sets is set to 5×10^{-5}. Outer estimates of the terminal sets of the monitoring variable for the attack-free closed-loop process are computed using the estimates of the minimum invariant sets of the process.

The nominal controller–observer parameters (K^*, L^*) are selected to stabilize the closed-loop process using pole placement by placing the poles at [0.2–0.1] to determine the controller gain and placing the poles at [0.2 0.3] to determine the observer gain. The attack-sensitive controller–observer parameters (K_Λ, L_Λ) are determined by placing the poles at [−0.2–0.3] and [−0.2–0.3] to compute the controller and observer gains, respectively. The control system with the attack-sensitive controller–observer parameters is sensitive to attacks in the set: $\{\Lambda|\text{diag}(1, \alpha)|\alpha \in [0.6, 0.9]\}$. This range of attacks is verified by checking the eigenvalues of the matrix $A_\xi(\Lambda, K_\Lambda, L_\Lambda)$ with $\Lambda = \text{diag}(1, \alpha)$ and varying α starting from $\alpha = 0.6$ and incrementing by 0.01 until a maximum value of $\alpha = 0.9$ is reached. Performing a similar analysis for the nominal controller–observer parameters found that the nominal controller–observer parameters are not sensitive to any attack over the range checked.

The theoretical analysis of this work considers linear systems of the form in Equation (1). The active detection method is applied to a nonlinear process to demonstrate its applicability to a nonlinear process, extending beyond what is considered in the theoretical analysis. The discrete-time linear control system is applied to the nonlinear process in a sample-and-hold fashion. To integrate the nonlinear ordinary differential equations in Equation (23), the explicit Euler method is used with an integration step size of 1×10^{-4} h.

Two scenarios are considered. The first scenario considers the attack-free process with the proposed active detection method that minimizes false alarms. The second scenario considers the application of the proposed active detection method to the attacked process to demonstrate the enhancement of detection capabilities of the detection scheme in Equation (15). Each scenario consists of 1000 simulations, where the bounded process disturbances in the feed concentration ΔC_{A0} and the measurement noise in the concentration sensor are modeled as random numbers drawn from two different uniform distributions on the interval [−0.01, 0.01] kmol m^{-3}. Similarly, the bounded process disturbances in the feed temperature ΔT_0 and the measurement noise in the temperature sensor are modeled as random numbers drawn from two different uniform distributions on the interval [−0.2, 0.2] K. The same realization of the random variables is used in each scenario to compare across simulations. The initial condition of all simulations is 0, which is contained within the attack-free minimum invariant set under the nominal controller–observer parameters. The total length of each simulation is 5 h.

In the first scenario, the proposed active detection method is applied to the attack-free CSTR process to demonstrate false alarm minimization. A switch into the attack-sensitive mode to probe for attacks is scheduled for $[t_i, t_f] = [50, 400]$ corresponding to a real-time interval of [0.5, 4] h. The minimum and maximum dwell-time under the attack-sensitive mode are selected to be $T_c^{\min} = 100$ (1 h in real-time) and $T_c^{\max} = 110$ (1.1 h in real-time). The alarm suppression times are chosen to span 2 time steps from a switch, that is, $\Delta_1 = 2$ and $\Delta_2 = 2$. This is because the augmented state converges to the minimum invariant set under the new controller–observer parameters in 2-time steps or less after a switch over numerous simulations of the attack-free process with a time-triggered control system switch.

No alarms are raised by the detection scheme in Equation (15) in any of the 1000 simulations of the attack-free process with the proposed active detection method. The output and residual values of the attack-free process over one simulation are illustrated in Figure 3. The measured output values (Figure 3A) and the residual values (Figure 3B) of the process under both controller–observer parameters are maintained within their corresponding terminal set. Over the simulations, the switch into the attack-sensitive mode is implemented at a time step in the interval [50, 56] ([0.5, 0.56] h). At the time instance when the control system switches from the nominal mode to the attack-sensitive mode, the condition in Equation (20) is satisfied over all simulations. As a result, this switch does not excite process dynamics. However, for the switch back to the nominal mode, the condition in Equation (21) is not satisfied over the switching interval for all 1000 simulations, and the switch back to the nominal mode is forced at the end time $t_s + T_c^{\max}$. Following this, alarms are suppressed for 2-time steps from the switch. No false alarms are observed because the augmented state converges to the attack-free minimum invariant set under nominal controller–observer parameters within 2-time steps or less from the switch. The results from one simulation are illustrated in Figure 3. In this simulation, the monitoring variable values are contained within the attack-free terminal set under the nominal mode until the control system switches from the nominal mode to the attack-sensitive mode at the time $t_s = 52$ (0.52 h). After the switch, the monitoring variable values are within the attack-free terminal sets under attack-sensitive controller–observer parameters (Figure 3A,B). As a result, no false alarms are observed. Control system switches

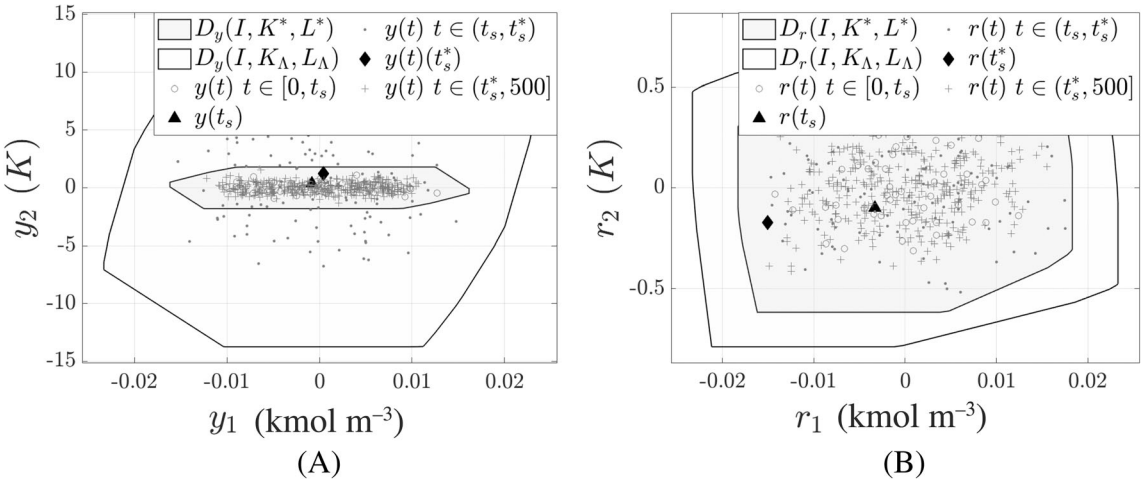

FIGURE 3 (A) The output values over a simulation of the attack-free closed-loop process with the proposed active detection method. (B) The residual values over a simulation of the attack-free closed-loop process with the proposed active detection method

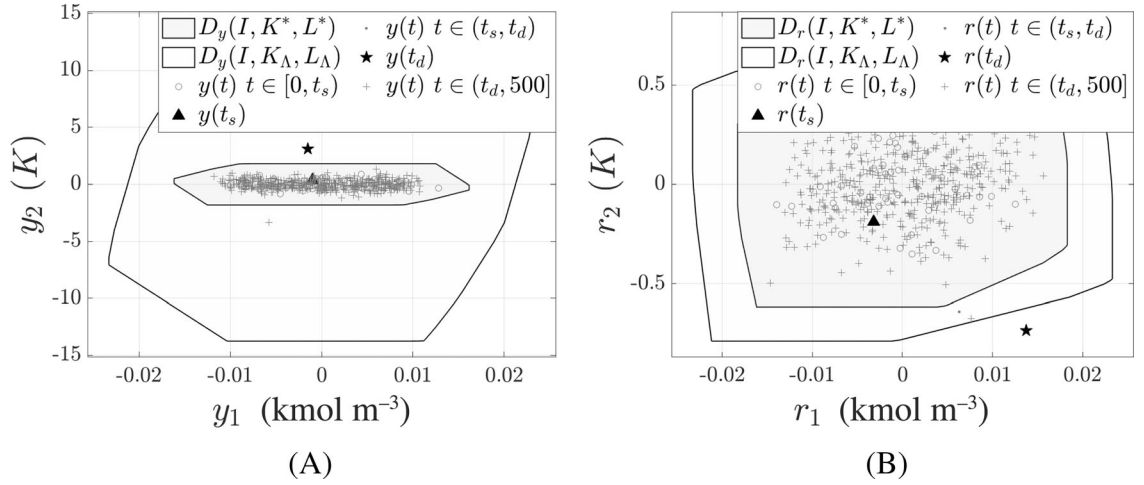

FIGURE 4 (A) The output values over a simulation of the attacked closed-loop process with the proposed active detection method. (B) The residual values over a simulation of the attacked closed-loop process with the proposed active detection method

back to the nominal mode at the time $t_s^* = 162$ (1.62 h). After the switch, the monitoring variable values are within its attack-free terminal set under nominal controller–observer parameters, and no alarms are observed.

The second scenario considers the attacked CSTR process with the active detection method to demonstrate the attack detection capabilities. A multiplicative attack of magnitude $\Lambda = \text{diag}(1, 0.85)$ is considered. The attack is potentially detectable under the nominal controller–observer parameters, and the attack is detectable under attack-sensitive controller–observer parameters. Over all simulations of the attacked process with the active detection method, the switch into attack sensitive mode is implemented over the time interval [50, 74] ([0.5, 0.74] h in real-time). The attack is detected in every simulation within 24 time steps after the switch into attack-sensitive mode. The results from one simulation are illustrated in Figure 4. In this simulation, the attack is not detected with $\chi(t) \in D_\chi(I, K^*, L^*)$ for $t \in [0, t_s]$ (Figure 4A,B). After the switch, the attack is detected at the time $t_d = 57$ (0.57 h) due to $\chi(t_d) \notin D_\chi(I, K_\Lambda, L_\Lambda)$ (Figure 4A,B). Immediately after attack detection, the control system switches back to the nominal mode to stabilize the process. After the switch, the monitoring variable is contained within its attack-free terminal set under nominal controller–observer parameters and no further alarms are observed.

For a basis of comparison, the closed-loop process is also simulated with the process operating exclusively under the nominal mode and monitored by the detection scheme in Equation (14). In this case, the attack is detected in 20 out of 1000 simulations. The attack detection times over these simulations of the attacked process under nominal mode are compared with the attack detection times for the corresponding simulations of the attacked process with the active detection method. In 4 of the 20 simulations, the attack is detected before t_i. Over the corresponding 4 simulations with the active detection method, the attack is detected at the same time as the simulations of the process exclusively under the nominal mode. Over the

remaining 16 of the 20 simulations of the process under the nominal mode, the attack is detected at a time in the interval [80, 491] ([0.8, 4.91] h). Over corresponding simulations of the process with the active detection method, the attack is detected at a time in the interval [52, 64] ([0.52, 0.64] h). Therefore, the active detection method enhances the detection capabilities of the detection scheme in Equation (14).

5.3 | Selection of a minimum dwell-time for the CSTR process

Using several simulations of the CSTR process under an attack, the choice of the minimum dwell-time of $T_c^{\min} = 1$ h is analyzed. Several scenarios are considered. Each scenario consists of 1000 simulations of the CSTR process, similar to the scenarios in the prior section. To simulate the process in the attack-sensitive mode, the simulations are initialized with the attack-sensitive controller–observer parameters, that is, for all scenarios considered, the switching time from the nominal to the attack-sensitive modes is $t_s = 0$ h. A time-triggered switching strategy with a dwell-time of $T_c = 100$ under the attack-sensitive mode is used. Process states at each simulation are initialized at 0.

First, seven different scenarios are considered to analyze if a minimum dwell-time of $T_c^{\min} = 1$ h is sufficient to allow for the detection of attacks with magnitude in the range $\{\Lambda|\text{diag}(1, \alpha)|\alpha \in [0.6, 0.9]\}$. Across scenarios, the magnitude of attack targeting the temperature sensor-controller link is varied. The first scenario considers an attack of magnitude $\Lambda = \text{diag}(1, \alpha)$, with $\alpha = 0.6$. For each of the subsequent scenarios, α is incremented by 0.05 over the range until a value of $\alpha = 0.9$ is reached for the seventh scenario. The minimum, maximum, and average time for detection of the attack are computed over each scenario, as illustrated in Figure 5A. The average time for attack detection increases with the value of α. The minimum detection time of all attacks is 0.03 h. The attack with $\alpha = 0.9$ has the maximum time for detection of $t_d = 0.23$ h. Based on this result, a dwell-time of $T_c^{\min} = 1$ h is sufficient to ensure the detection of attacks in the range $\{\Lambda|\text{diag}(1, \alpha)|\alpha \in [0.6, 0.9]\}$.

A second simulation study is conducted to analyze the impact of various dwell-times on attack detection. Several scenarios are considered for the process under an attack of magnitude with $\alpha = 0.9$. An attack with $\alpha = 0.9$ is chosen because it has the maximum detection time in the first simulation study. In total, 30 scenarios are considered. In the first scenario, a dwell-time of $T_c = 0.01$ h is chosen. Thereafter, for each scenario, the dwell-time is incremented by 0.01 h, with the last scenario considering a dwell-time of 0.3 h. Over each scenario, the total number of simulations out of 1000 simulations with an attack detection is computed (Figure 5B). As the dwell-time increases, the total attack detections also increase. Furthermore, a dwell-time of $T_c = 0.15$ h under the attack-sensitive controller–observer parameters may be sufficient to detect the attack in 97.6% of the simulations. Similarly, a dwell-time of $T_c = 0.23$ h results in the attack being detected in 100% of the simulations. Thereafter, a further increase in the dwell-time has no impact on the total attack detections. The results indicate that to limit the performance degradation in the process, a smaller dwell-time than $T_c^{\min} = 1$ h may be considered.

Remark 5. In this section, the proposed active detection method is applied to a nonlinear chemical process, extending beyond what is considered in the theoretical analysis presented in this work. From the closed-loop simulation results, the detection scheme detected the multiplicative attack, and did not raise any false alarms. Also, the augmented state is maintained within the minimum invariant set computed from the linearized process model in all cases. These results demonstrate the

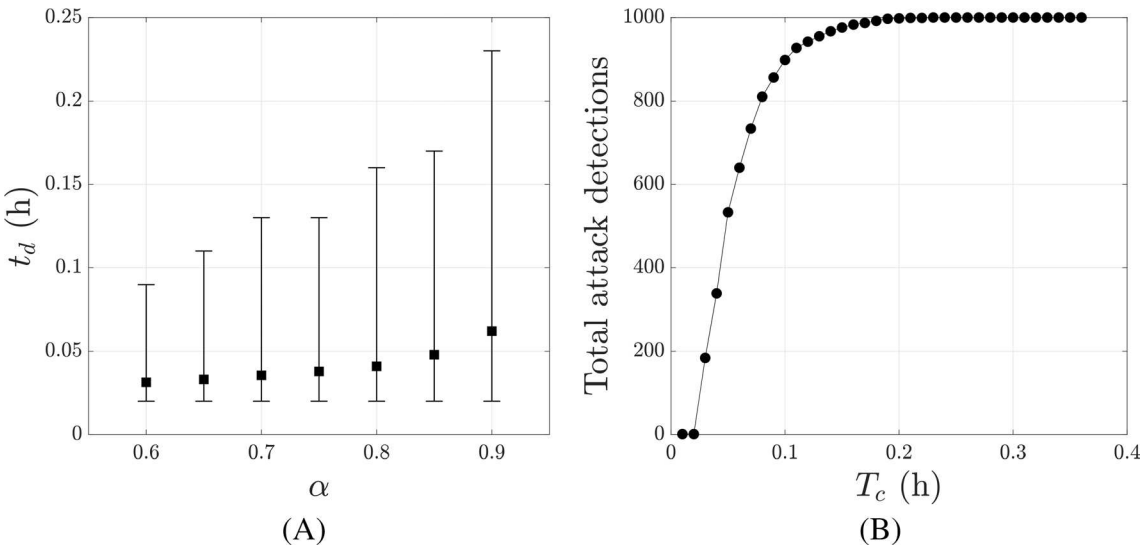

FIGURE 5 (A) The attack detection times for different attack magnitudes and a dwell-time of $T_c = 1$ h. (B) The number of attacks detected under the attack-sensitive mode for an attack of magnitude of $\Lambda = \text{diag}(1, 0.9)$ with different dwell-times

proposed active detection method's applicability to the nonlinear CSTR process. In general, it may be expected that the method will provide minimal false alarms while enhancing the detection capabilities for nonlinear processes when the augmented state is maintained in a small neighborhood of the origin such that the effect of the nonlinearities is small, that is, when the process disturbances and measurement noise are small. However, extensions of the active detection method to nonlinear processes remain an open area and are subject to future work.

6 | CONCLUSIONS

In this work, a detectability-based classification of multiplicative sensor-controller link false-data injection attacks with respect to a general class of detection schemes monitoring the process was presented. A control switching-based approach for enhancing attack detectability with respect to an output and residual-based detection scheme was proposed. To guarantee zero false alarms from switching, a confidence region for the attack-free augmented states was constructed, and a confidence region-based switching condition was developed. The switching condition was incorporated into the proposed active detection method to minimize false alarms. The application of the proposed active detection method for attack detectability enhancement and false alarm minimization was demonstrated using two illustrative processes. Future work will focus on extensions of the proposed active detection method to nonlinear processes.

AUTHOR CONTRIBUTIONS

Shilpa Narasimhan: Conceptualization (equal); formal analysis (equal); methodology (equal); software (lead); visualization (lead); writing – original draft (equal). **Nael H. El-Farra:** Conceptualization (equal); formal analysis (equal); methodology (equal); project administration (equal); supervision (equal); writing – review and editing (equal). **Matthew J. Ellis:** Conceptualization (equal); formal analysis (equal); methodology (equal); project administration (equal); supervision (equal); writing – original draft (equal); writing – review and editing (equal).

ACKNOWLEDGMENT

Financial support from the UC Davis College of Engineering is gratefully acknowledged.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Nael H. El-Farra https://orcid.org/0000-0002-1973-0287
Matthew J. Ellis https://orcid.org/0000-0003-3764-7783

REFERENCES

1. Slowik J. Evolution of ICS attacks and the prospects for future disruptive events. Technical Report. Threat Intelligence Centre; 2019.
2. Slowik J. Stuxnet to CRASHOVERRIDE to TRISIS: Evaluating the history and future of integrity-based attacks on industrial environments. Technical Report. Threat Intelligence Centre; 2019.
3. Bhamare D, Zolanvari M, Erbad A, Jain R, Khan K, Meskin N. Cybersecurity for industrial control systems: A survey. *Comput Secur*. 2020;89:101677.
4. US Department of Homeland Security. US-CERT: Chemical sector cybersecurity framework implementation guidance. Technical Report. U.S. Department of Homeland Security; 2015.
5. Xenofontos C, Zografopoulos I, Konstantinou C, Jolfaei A, Khan MK, Choo KKR. Consumer, commercial, and industrial IoT (in)security: attack taxonomy and case studies. *IEEE Internet Things J*. 2022;9:199-221.
6. Cook M, Stavrou I, Dimmock S, Johnson C. Introducing a forensics data type taxonomy of acquirable artefacts from programmable logic controllers. Proceedings of the International Conference on Cyber Security and Protection of Digital Services; Dublin, Ireland; 2020: 1-8.
7. Reda HT, Anwar A, Mahmood A. Comprehensive survey and taxonomies of false data injection attacks in smart grids: attack models, targets, and impacts. *Renew Sustain Energy Rev*. 2022;163:112423.
8. Huseinović A, Mrdović S, Bicakci K, Uludag S. A survey of denial-of-service attacks and solutions in the smart grid. *IEEE Access*. 2020;8:177447-177470.
9. Robles F, Perlroth N. 'Dangerous Stuff': Hackers Tried to Poison Water Supply of Florida Town. Online, The New York Times. 2021. https://www.nytimes.com/2021/02/08/us/oldsmar-florida-water-supply-hack.html
10. Maw A, Adepu S, Mathur A. ICS-BlockOpS: Blockchain for operational data security in industrial control system. *Pervasive Mob Comput*. 2019;59:101048.
11. Hu S, Yue D, Cheng Z, Tian E, Xie X, Chen X. Co-design of dynamic event-triggered communication scheme and resilient observer-based control under aperiodic DoS attacks. *IEEE Trans Cybern*. 2020;51(9):4591-4601.
12. Liang G, Weller SR, Zhao J, Luo F, Dong ZY. The 2015 Ukraine blackout: implications for false data injection attacks. *IEEE Trans Power Syst*. 2016;32(4):3317-3318.
13. Durand H. A nonlinear systems framework for cyberattack prevention for chemical process control systems. *Mathematics*. 2018;6(9):1-44.
14. Zedan A, El-Farra NH. A machine-learning approach for identification and mitigation of cyberattacks in networked process control systems. *Chem Eng Res Des*. 2021;176:102-115.
15. Chen S, Wu Z, Christofides PD. A cyber-secure control-detector architecture for nonlinear processes. *AIChE J*. 2020;66(5):e16907.
16. Chen S, Wu Z, Christofides PD. Cyber-attack detection and resilient operation of nonlinear processes under economic model predictive control. *Comp Chem Eng*. 2020;136:106806.
17. Wu Z, Christofides PD. *Process Operational Safety and Cybersecurity: A Feedback Control Approach*. Springer; 2021.
18. Zhang D, Wang QG, Feng G, Shi Y, Vasilakos AV. A survey on attack detection, estimation and control of industrial cyber–physical systems. *ISA Trans*. 2021;116:1-16.
19. Garg K, Sanfelice RG, Cárdenas ÁA. Control barrier function based attack-recovery with provable guarantees. *arXiv:220403077*; 2022.
20. Akbarian F, Tärneberg W, Fitzgerald E, Kihl M. A security framework in digital twins for cloud-based industrial control systems: intrusion detection and mitigation. Proceedings of the 26th IEEE International Conference on Emerging Technologies and Factory Automation; Västerås, Sweden; 2021: 1-8.
21. Hu Y, Li H, Yang H, Sun Y, Sun L, Wang Z. Detecting stealthy attacks against industrial control systems based on residual skewness analysis. *EURASIP J Wirel Commun Netw*. 2019;2019(1):1-14.

22. Ghaeini HR, Tippenhauer NO, Zhou J. Zero residual attacks on industrial control systems and stateful countermeasures. Proceedings of the 14th International Conference on Availability, Reliability and Security. Canterbury, UK; 2019: 1-10.
23. Weerakkody S, Ozel O, Griffioen P, Sinopoli B. Active detection for exposing intelligent attacks in control systems. Proceedings of the IEEE Conference on Control Technology and Applications; Hawai'i, USA; 2017: 1306-1312.
24. Ghaderi M, Gheitasi K, Lucia W. A blended active detection strategy for false data injection attacks in cyber-physical systems. *IEEE Trans Control Netw Syst*. 2020;8:168-176.
25. Na G, Eun Y. A multiplicative coordinated stealthy attack and its detection for cyber physical systems. Proceedings of the IEEE Conference on Control Technology and Applications; Copenhagen, Denmark; 2018: 1698-1703.
26. Huang T, Satchidanandan B, Kumar PR, Xie L. An online detection framework for cyber attacks on automatic generation control. *IEEE Trans Power Syst*. 2018;33(6):6816-6827.
27. Narasimhan S, El-Farra NH, Ellis MJ. Active multiplicative cyberattack detection utilizing controller switching for process systems. *J Process Control*. 2022;116:64-79.
28. Oyama HC, Durand H. Integrated cyberattack detection and resilient control strategies using Lyapunov-based economic model predictive control. *AIChE J*. 2020;66(12):e17084.
29. Kuntsevich VM, Pshenichnyi BN. Minimal invariant sets of dynamic systems with bounded disturbances. *Cybern Syst Anal*. 1996;32(1):58-64.
30. Blanke M, Kinnaert M, Lunze J, Staroswiecki M. *Diagnosis and Fault-Tolerant Control*. Springer-Verlag; 2006.
31. Isermann R. *Fault-Diagnosis Systems: an Introduction from Fault Detection to Fault Tolerance*. Springer-Verlag; 2006.
32. Venkatasubramanian V, Rengaswamy R, Yin K, Kavuri SN. A review of process fault detection and diagnosis. Part I. Quantitative model-based methods. *Comput Chem Eng*. 2003;27(3):293-311.
33. Narasimhan S, El-Farra NH, Ellis MJ. Detectability-based controller design screening for processes under multiplicative cyberattacks. *AIChE J*. 2022;68(1):e17430.
34. International Society of Automation. *ANSI/ISA-18.2-2016: Management of alarm systems for the process industries*. Standard International Society of Automation; 2009.
35. Kvasnica M, Grieder P, Baotić M. Multi-Parametric Toolbox (MPT). 2004. http://control.ee.ethz.ch/~mpt/
36. Raković SV, Kerrigan EC, Kouramas KI, Mayne DQ. Invariant approximations of the minimal robust positively invariant set. *IEEE Trans Automat Contr*. 2005;50(3):406-410.
37. Alanqar A, Ellis M, Christofides PD. Economic model predictive control of nonlinear process systems using empirical models. *AIChE J*. 2015;61(3):816-830.

How to cite this article: Narasimhan S, El-Farra NH, Ellis MJ. A control-switching approach for cyberattack detection in process systems with minimal false alarms. *AIChE J*. 2022; 68(12):e17875. doi:10.1002/aic.17875